An Investigation of the Methods for Estimating Usual Dietary Intake Distributions

A Thesis Presented in Partial Fulfillment of the Requirements for the Degree of

Master of Applied Statistics

at Massey University, Albany, New Zealand

Stefan Kremenov Stoyanov

Student ID # 00038377

2008

Abstract

The estimation of the distribution of usual intake of nutrients is important for developing nutrition policies as well as for etiological research and educational purposes. In most nutrition surveys only a small number of repeated intake observations per individual are collected. Of main interest is the long-term usual intake which is defined as long-term daily average intake of a dietary component. However, dietary intake on a single day is a poor estimate of the individual's long-term usual intake. Furthermore, the distribution of individual intake means is also a poor estimator of the distribution of usual intake since usually there is large within-individual compared to between-individual variability in the dietary intake data. Hence, the variance of the mean intakes is larger than the variance of the usual intake distribution. Essentially, the estimation of the distribution of long-term intake is equivalent to the estimation of a distribution of a random variable observed with measurement error.

Some of the methods for estimating the distributions of usual dietary intake are reviewed in detail and applied to nutrient intake data in order to evaluate their properties. The results indicate that there are a number of robust methods which could be used to derive the distribution of long-term dietary intake. The methods share a common framework but differ in terms of complexity and assumptions about the properties of the dietary consumption data. Hence, the choice of the most appropriate method depends on the specific characteristics of the data, research purposes as well as availability of analytical tools and statistical expertise.

Acknowledgements

I would like to thank my supervisor Dr. Barry McDonald for sparking my interest in measurement error models and for his support and encouragement during my work on the thesis.

I am grateful to Dr. Hoffmann and Dr. Tooze for the useful comments and computer code which were of great help in completing this research.

Table of Contents

List of Tables and Graphs

Tables:

Graphs:

1. Introduction

One of the common purposes of food consumption surveys is to estimate the distribution of long-term intake of various nutrients for a target population or specific subpopulations. The information on the consumption habits of the populations of interest is collected with food consumption surveys using dietary assessment instruments such as food frequency questionnaires, 24-hour recalls or food records.

Food frequency questionnaires (FFQs) are by far the most often used method for dietary assessment because they are easy to administer, relatively inexpensive and designed to measure the long-term dietary intake which is of primary interest in food consumption surveys. However, as documented by Hoffmann et al (2002a) the FFQs have various degrees of validity, completeness and specification. The application of FFQs for assessing the usual dietary intake has been criticized by many researchers since they rely on individuals' long-term memory and capability of correct recording of the average consumption. Hence, the FFQs could have considerable bias.

In contrast, the 24-hour recall method refers to very short periods of exposure. The method relies on individuals' short-term memory since it is based on nutrients or foods consumed on one or more specific days. Thus, it could be argued that the 24-hour recall method gives more accurate and detailed picture of the daily intake in the population. Recent studies (Kipnis et al 2001, Kroke et al, 1999) document that intakes obtained by 24-hour recalls or food records are also biased compared to the measurements obtained by biomarkers. However, the size of bias and systematic measurement error are moderate compared to those of the FFQs.

An important concept used in the analysis of dietary intake data is the so called 'usual intake' which could be defined as a long-run daily average intake of a nutrient. The estimation of the distribution of usual daily intake of nutrients in a population is extremely important since it enables researchers to evaluate the general nutrition status of that population. Often, the nutrient intake is evaluated against some standard in order to assess the adequacy of the nutrient consumption and proportion of the population or subpopulation above or below a certain cut-off value. Deficient nutrient intake is a factor which contributes to many chronic diseases. The relationship between diet and diseases such as cancer and cardiovascular diseases is an area of active academic research.

Table 1 illustrates some of the most important nutrients the intake of which is regularly assessed in food consumption surveys:

Table 1: Main Nutrients

Macronutrients	Vitamins	Minerals and Trace Elements
Energy	Vitamin A	Calcium
Protein	Thiamin	Chromium
Fat	Riboflavin	Copper
n-6 fatty acids (linoleic)	Niacin	Fluoride
n-3 fatty acids (α-linolenic)	Vitamin B6	Iodine
LC n-3 fatty acids (omega-3 fats, DHA, DPA, EPA)	Vitamin B12	Iron
Carbohydrates	Folate	Magnesium
Dietary fibre	Pantothenic acid	Manganese
Water	Biotin	Molybdenum
	Choline	Phosphorus
	Vitamin C	Potassium
	Vitamin D	Selenium
	Vitamin E	Sodium
	Vitamin K	Zinc

Source: New Zealand Ministry of Health (2005)

The nutrients analysed in this research belong to the three major groups shown in Table 1 above. The New Zealand Ministry of Health (2005) has published an updated document listing the recommended dietary intake for more than 40 nutrients. In the same document some of the major definitions used to measure the adequacy of the nutrient intake are described in detail. Three of the most often used measures include:

- EAR (Estimated Average Requirement). The measure is used to derive a daily nutrient level which meets the nutrient requirements of half the healthy individuals in a particular life stage and age group. It is used to estimate the prevalence of inadequate consumption within a group.

- RDI (Recommended Dietary Intake). The measure is used to derive the average daily dietary intake that is sufficient to meet the nutrient requirements of nearly all (97-98 percent) of individuals in a particular life stage and age group. The measure is not used to assess intakes of a group.

- AI (Adequate Intake). The average daily nutrient intake level based on observed or estimated nutrient intake by a group of healthy individuals assumed to be adequate. Mean intake at or above the AI level would imply low prevalence of inadequate nutrient intake in the population.

Since most of the consumption survey data are repeated measurements data the average of infinitely many daily intake observations collected on a single individual converges to the individual's usual intake. The usual intake distribution obtained from the mean of 365 days would be the most accurate estimate of the unobserved usual intake distribution. However, due to various financial, time and resource constraints such data are impossible to collect in practice. Hence, most of the consumption surveys collect only a small number of observations per individual.

It is well documented in the research literature that the observed daily nutrient intake is closely connected to an individual's usual intake. However, since usually only a small number of repeated observations per individual are collected the observed intake is a poor estimate of individual's long-term usual intake. Some of the reasons include high positive skewness, nuisance effects which cause the consumption to differ according to the day-of-the week or month of the year, correlated repeated measurements, correlation between within-individual variance and usual intake means as well as under and over-reporting biases.

For accurate assessment of the adequacy of nutrient intake an estimate of the distribution of long-term intake needs to be obtained. One simple method is to use the mean of several days of daily consumption for each individual to derive the distribution of usual intake. However, using this method the estimates of the proportion of individuals above or below a given consumption threshold will be biased since the within-individual variability greatly inflates the variance of the distribution of individual means. It could be argued that the day to day within-individual variation is superimposed on an underlying consistent nutrient consumption pattern. The degree of variation differs according to the nutrient. Total energy intake usually has the least day-to-day variation. Macronutrients do not have a large degree of variation whereas vitamins usually have large degree of day-to-day variation. Hence, for accurate dietary assessment an adjusted distribution of usual intake which is free from the effect of within-person variation has to be obtained.

Statistical methods for estimating the usual intake distributions have been investigated extensively in the academic literature in recent years. The purpose of this research is to review and summarize the major developments in the field as well as to investigate the application of a number of statistical methods based on specific research goals and consumption data characteristics.

2. Literature Review

The statistical theory which underlines some of the most popular methods for estimating usual intake distributions is discussed in detail. All the methods investigated in the literature review section share a common framework but differ in terms of assumptions about the properties of the repeated measurements data.

The procedures are based on the classical measurement error model. They attempt to estimate the long-term average nutrient intake by decomposing the variability of the observed individual mean daily intakes into within and between-individual components. The variance of the average daily nutrient intakes is adjusted for measurement error (within-individual) variance and the adjusted variance is used to estimate the distribution of the long-term intake. In order to perform the adjustments, and estimate the distribution of usual intake, the methods require that more then one repeated measurement per individual are collected on at least a representative sample of the population. The majority of the procedures provide estimates of the long-term intake distribution which are useful for group level assessments only, and do not provide estimates of individual usual intake.

The measurement error theory, and in particular, the problem of estimating the long-term nutrient intake distributions are areas of active academic research. The recently developed methods built upon their predecessors but include a number of statistical enhancements. Some of the most often used procedures include the NRC (1986), ISU (1996), Hoffmann et al (2002) and NCI (2006) methods. These methods represent major contributions to the development of the methodology for estimating usual intake distributions.

2.1 Dodd et al (2006)

Dodd et al (2006) investigate some of the major statistical methods used for estimating the distribution of usual dietary intake. The focus of the article is on the strengths and weaknesses of each method. Special attention is given to the problems inherent in the modeling of usual intake of episodically consumed nutrients. The overall conclusion of the authors is that the major statistical methods documented in the article share a common methodology but vary in terms of statistical complexity. The differences between the methods arise from the different assumptions about the measurement characteristics of the dietary intake data. Hence, depending on the method used and underlying assumptions the estimated usual intake distributions often differ from one another.

Dodd et al (2006) argue that the 24-hour recalls and food frequency questionnaires (FFQs) are the primary instruments for collecting dietary data. The FFQs are designed to measure long-term behaviour and are relatively inexpensive compared to the 24-hour recalls. However, the FFQs have some shortcomings which introduce substantial error into the usual dietary intake estimates based on FFQs. Some of those shortcomings include limited list of foods and errors in reporting consumption over long time periods. Extensive research on the topic has been published by Flegal and Larkin (1990), Subar et al (2003), Kupnis et al (2003) and Freedman et al (2004), among others.

The 24-hour recalls provide much richer information about the types and amounts of food consumed. Since they focus on a 24-hour period the magnitude of the systematic reporting error is greatly reduced. However, the 24-hour recalls method has a number of drawbacks as well. Some of them include variability in the day-to-day individual diets, measurement errors and errors resulting from the use of standardised recipe files and consumption databases. All of the above factors contribute to the considerable within-person variability observed in the 24-hour recalls data and lead to the conclusion that the intake measured on a single day is a poor estimate of the long-term intake. Extensive research on the topic has been published by Beaton et al (1979, 1983).

Dodd et al (2006) review four of the major statistical methods for estimating the usual dietary intake distributions. The simplest method which is based on averages of several 24-hour recalls proves to be unsatisfactory since the mean of the financially and operationally feasibly achievable number of 24-hour recalls contains considerable within-person variation. Hence, the distribution of the within-person means has larger variance than the true usual intake distribution. This leads to biased estimates of the parameters of the usual intake distribution.

The statistical modeling mitigates the limitations of the 24-hour recalls data by estimating and removing the effects of the within-person variation in the dietary intake data. The methods investigated by Dodd et al (2006) follow a common framework consisting of the following steps:

A. Preparatory Step.

Some of the more complex methods include a preparatory step where initial adjustments are made to the 24-hour recalls data. All of the methods documented by Dodd et al (2006) require decomposition of the variance into between and within-person components. If between and within-person deviations are normally distributed then the usual intake distribution is also normal and could be described only by its mean and variance.

In practice, most of the dietary intake distributions are positively skewed and for some categories of nutrients have a large proportion of zero values. For this reason, often the intake data are transformed to approximate normality in the first stage of the analysis. If the transformation approximates normally distributed data the distribution of the original (untransformed) intake could be described in terms of the normal distribution and the transformation. In such cases back-transformation would be required as well since all standards used to assess the intake are expressed in terms of the units of the untransformed intake.

The form of the back-transformation depends on the assumptions about the 24-hour recalls as dietary assessment instrument. There are two main assumptions that could be used at this stage. The first one is that the 24-hour recall intake is an unbiased estimator of the usual intake on the transformed scale (assumption A). The second one is that the 24-hour recall intake is an unbiased estimator of the usual intake in the original scale (assumption B). If assumption A is used then the back-transformation is relatively straightforward and is the inverse of the original transformation. Assumption B requires more complicated back-transformation with an additional adjustment for bias. The adjustment is described in detail in the section documenting the back-transformation methods.

B. Describe the assumed relationship between the 24-hour recall measurements and individual usual intake.

The assumption is that the 24-hour recall intake is an unbiased estimator of usual intake. This assumption does not imply lack of error. The 24-hour recall could over or under estimate the true usual intake of an individual but estimation errors cancel out when the number of repeated measurements is large. The assumption that the 24-hour recall intake is an unbiased estimator of usual intake is equivalent to the assumption that 24-hour recall intake is unbiased for true single day intake. However, it should be noted that when individual intake observations at each time point are being considered the problem about estimation errors remains.

C. Partition the total variation in the 24-hour recall measurements into within and between-person components.

The individual usual intake could be expressed as the sum of group mean usual intake and some person-specific variation around the group mean. Such variation shown in the brackets below represents the between-person variation:

Individual usual intake = group mean usual intake + (individual usual intake – group mean usual intake) (1)

Since the 24-hour recall intake has significant within-person variation we could define the 24-hour recall intake as:

24-hour intake = group mean usual intake + (individual usual intake – group mean usual intake) + (24-hour recall intake - individual usual intake) (2)

The third term in the above equation represents the within-person variation. The variance components could be estimated using standard methods. However, the estimation of within-person variance requires repeated measurements. Hence, for at least some of the individuals two or more 24-hour recalls are needed.

D. Estimate the distribution of usual intake after removing the effect of within-person variation.

Given the partition of the variance described in point C above a set of intermediary values with desired variance σ_b^2 is constructed by shrinking each individual's mean towards the overall population mean using the following formula:

Intermediary value = (1-w)(overall mean) +w*(individual mean)* (3)

Where w is a shrinkage factor calculated as the square root of the ratio of between-person variance to the variance of the within-person means:

$$w = \sqrt{\frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2 / n}} \quad (4)$$

In the above equation σ_b^2 denotes the between-person variance, σ_w^2 is the within-person variance, σ_w^2 / n is the variance of the average of n recalls per individual and $\sigma_b^2 + \sigma_w^2 / n$ is the variance of the empirical distribution of the within-person means.

The empirical distribution of the intermediary values has a mean equal to the grand mean of the 24-hour recalls. The percentage of a group of individuals with an intake less than some standard could be estimated by the percentage of intermediary values which are less than the relevant threshold value. However, the intermediary values are not suitable for the estimation of individual usual intake. They could only be used to describe the distribution of the long-term intake in the population.

From equation (4) above it could be concluded that when between-person variance is small or n is small w is close to zero and each intermediary value is close to the grand mean. In contrast, when the within-person variation is small or n is large the intermediary values are close to the individual means.

2.1.1 Statistical Methods

The methods investigated in Dodd et al (2006) article are as follows:

1) National Research Council/Institute of Medicine Methods (1986).

Both methods are consistent with assumption A. The methods could be applied to 24-hour recalls data obtained from a simple random sample of individuals. The initial adjustment is usually a power or log transformation of the dietary intake data to approximate normality. The methods are fairly robust to mild departures from normality. Variance components are estimated on the transformed scale. The set of intermediary values are estimated by shrinking the individual means of the transformed 24-hour recalls to the overall mean of the transformed 24-hour recalls intake.

The inverse of the original transformation is applied to the shrunken means in order to estimate the distribution of usual intake in the original scale. The National Research Council (1986) method is reviewed in detail in a separate section of the literature review.

2) Iowa State University Method (1996).

The method is consistent with assumption B and could be applied to 24-hour recalls data from large surveys. The method is based on a complex procedure consisting of two-stage transformation of the data in order to obtain normality. Various additional transformations could also be performed in order to adjust for nuisance bias. For these reasons the method requires fairly large samples of approximately several hundred individuals out of which at least 50 must have at least two or more 24-hour recalls. Furthermore, the flexibility of the method requires the intermediary values to be based on theoretical quantiles from a normal distribution instead of individual means. The back-transformation requires a bias-adjustment in order to be consistent with assumption B. The ISU method is reviewed in detail in a separate section of the literature review.

3) Best Power Method (1996).

The method is developed by the same team who developed the ISU method and represents a simplified alternative to the ISU method. The method could also be applied to complex surveys but uses only one-stage power or log transformation of the data to approximate normality. It does not allow the within-person variance to vary among individuals. The simple form of the initial transformation leads to simplified adjustment for transformation induced bias in the back-transformation stage. A simulation study performed by Nusser et al (1996) documents that although the ISU method is superior, the differences between the ISU and best power methods are very small in practice. The best power method is not evaluated further in this research. A method proposed by Hoffmann et al (2002a) which is similar to the best power method is reviewed in detail in a separate section of the literature review. A modified version of the best power method is described in the discussion chapter of the thesis.

4) Iowa State University Method for Episodically Consumed Dietary Components (ISUF 1997).

All of the three methods above are designed for usual intake data which could be transformed to approximate normality. However, some nutrients investigated in this research such as vitamin D are episodically consumed and a large proportion of individuals in the sample have zero consumption. Hence, the observed distribution has a large proportion of zero values in the left tail. Nusser et al (1997) proposed a method for modeling the distributions of such nutrients which treats the zero observations separately from the positive observations.

The simple n-day within-person mean is expressed as a product of two parts:

$$\text{Total intake of nutrient/n} = (\text{k/n}) * (\text{total intake of nutrient/k}) \qquad (5)$$

where k is the number of days on which the nutrient is consumed. Therefore, for large values of n the above equation models the long-term average consumption as the product of two components – the probability of consumption (k/n) and long-term average amount consumed on consumption days $(\text{total intake of nutrient/k})$.

The method assumes that the consumption probability and consumption day intake are independent. The distribution of usual consumption-day intake is modeled using the ISU method. Hence, the assumption is that the non-zero 24-hour recalls are unbiased for usual consumption day intake on the original scale. The distributions of consumption probability and usual consumption-day intake are combined to obtain the estimated distribution of usual intake.

The method is not investigated further in this research. A method proposed by Tooze et al (2006) which is the latest development in the field and overcomes some of the major drawbacks of the ISUF (1997) method is reviewed in detail in a separate section of the literature review.

2.1.2 Assumptions and Bias Correction

The individual's usual dietary intake τ_i is the mean of infinitely many single day intakes and could be expressed as:

$$\tau_i = E\left[T_{ij} \mid i\right] \tag{6}$$

where T_{ij} denotes the true intake for individual i on day j.

The above approximation should hold for complete 365 days of data which for most consumption surveys is not feasibly achievable. However, it will be a good approximation if the number of sampling days is sufficiently large. Some of the methods for optimal sample size estimation for dietary surveys are discussed in detail by Volatier et al (2002).

If we denote the 24-hour recall for individual i on day j as R_{ij} then in general R_{ij} measures T_{ij} with error:

$$R_{ij} = T_{ij} + \varepsilon_{ij} \tag{7}$$

where $E\left[\varepsilon_{ij} \mid i\right] = \mu_{\varepsilon}$ (8)

The transformation $g(.)$ of the 24-hour intake is such that the following normal theory applies:

$$r_{ij} = g\left(R_{ij}\right) = E\left[r_{ij} \mid i\right] + \left\{r_{ij} - E\left[r_{ij} \mid i\right]\right\} = b_i + w_{ij} \tag{9}$$

where $g(.)$ is one-to-one transformation with inverse transformation $h(.) = g^{-1}(.)$, $b_i \sim N\left(\mu, \sigma_b^2\right)$ and $w_{ij} \sim N\left(0, \sigma_w^2\right)$.

Under assumption A $b_i = g\left(\tau_i\right)$ and therefore $\tau_i = h\left(b_i\right)$. Estimates of the unknown parameters μ, σ_b^2 and σ_w^2 are obtained using standard component of variance analysis applied to the transformed 24-hour intakes.

Under assumption B the 24-hour recall is unbiased for true usual intake in the original scale:

$$\tau_i = E\left[T_{ij} \mid i\right] = E\left[R_{ij} \mid i\right] = E\left[h\left(r_{ij}\right) \mid i\right] = E\left[h\left(b + w\right) \mid b = b_i\right] \tag{10}$$

where w denotes the within-person variation in the transformed 24-hour intakes. However, if $h(.)$ is non-linear τ_i does not in general equal $h\left(b_i\right)$.

If an approximation to the expectation of a random variable is used than according to the Taylor approximation:

$$E\left[h(b+w)\,|\,b=b_i\right] \approx h\left(E\left[b+w\,|\,b=b_i\right]\right)+(1/2)h^{"}\left(E\left[b+w\,|\,b=b_i\right]\right)Var\left[b+w\,|\,b=b_i\right]= \\ =h(b_i)+(1/2)h^{"}(b_i)Var[w]=h(b_i)+(1/2)h^{"}(b_i)\sigma_w^2 \quad (11)$$

where $h^{"}(b_i)$ is the second derivative of the function $h(.)$ evaluated at b_i.

Hence, the back-transformed intermediary values:

$$\tau_i^* = h(b_i^*)+(1/2)h^{"}(b_i^*)\sigma_w^2 \quad (12)$$

have a distribution which is approximately the same as that of τ_i. In the above equation b_i^* are a set of intermediary values which have the same mean and variance and hence the same normal distribution as the b_i. The second term on the right-hand side of the equation is the bias-adjustment term.

Table 2 below lists some of the bias-adjustment terms used in the best power method developed by Nusser et al (1996) for some of the most common choices of normality transformation $g(R)$ with inverse function $h(r)$.

Table 2 : Dodd et al (2006). Data Transformations and Bias Adjustments

Transformation	$g(R)$	$h(r)$	Bias-adjustment term
Logarithm	$r=\log(R)$	$R=\exp(r)$	$(1/2)\exp(b_i^*)\sigma_w^2$
Power	$r=R^{\frac{1}{\lambda}}$	$R=r^{\lambda}$	$(1/2)\lambda(\lambda-1)(b_i^*)^{\lambda-2}\sigma_w^2$
Box-Cox	$r=\lambda\left(R^{\frac{1}{\lambda}}-1\right)$	$R=\left(\frac{r}{\lambda}+1\right)^{\lambda}$	$(1/2)\left[(\lambda-1)/\lambda\right]\left[b_i^*/\lambda+1\right]^{\lambda-2}\sigma_w^2$

Dodd et al (2006) recommend the application of methods consistent with assumption B whenever possible. It would ensure that the estimate of the group mean usual intake will coincide with the overall average of the 24-hour recalls and is consistent with the practice of tracking averages over time. Furthermore, Dodd et al (2006) argue that the 24-hour recalls have the tendency to underestimate true usual intake and the bias adjustments used in ISU, best power and ISUF methods partially offset this tendency.

However, it should be noted that there is some evidence indicating that neither assumption A nor assumption B truly hold in practice. For example, such evidence could be found in the research published by Freedman et al (2004). The authors document that both FFQs and 24-hour recall instruments exhibit substantial bias and within-person variation compared to estimates obtained with biomarkers. The 24-hour recalls showed substantially smaller bias than the FFQs and greater within-person variation. After adjustment for within-person variation the distributions estimated from 24-hour recalls agreed with the biomarker derived distributions more closely than those derived from FFQs.

Table 3 : Dodd et al (2006). Comparison of the Model Steps

	Method			
	NRC	**ISU**	**BP**	**ISUF**
Step 0: Apply initial data adjustments	Power or log transformation to approximate normality.	1. Adjustment for nuisance effects such as season, day-of-week, and interview sequence. 2. Two-stage transformation involving power transformation and polynomial regression to approximate normality.	1. Adjustment for nuisance effects such as season, day-of-week, and interview sequence. 2. Power or log transformation to approximate normality.	1. Estimate the distribution of the probability to consume on a given day, based on the relative frequency of nonzero 24-hour recalls. 2. Adjust the consumption-day 24-hour recalls for nuisance effects such as season, day-of-week and time-in sample. 3. Construct a two-stage transformation involving power transformation and polynomial regression so that the distribution of the transformed adjusted consumption day 24--hour recalls is approximately normal.
Step 1: Assumed relationship between the individual 24-hour recall measurements and individual usual intake	A transformed 24-hour recall is unbiased for transformed usual intake. (Assumption A)	A 24-hour recall is unbiased for usual intake in the untransformed scale. (Assumption B)	A 24-hour recall is unbiased for usual intake in the untransformed scale. (Assumption B)	1. Usual intake is the probability to consume on a given day multiplied by the usual intake amount for the day the food is consumed. 2. A 24-hour recall measures zero consumption exactly. 3. A nonzero 24-hour recall is unbiased for usual consumption-day intake in the untransformed scale. (Assumption B)
Step 2: Partition the total variation in the 24-hour recall measurements into within- and between-person components	Within-person variance is the same among individuals.	Within-person variance can vary among individuals.	Within-person variance is the same among individuals.	Within-person variance can vary among individuals.
Step 3: Estimate the usual intake distribution accounting for within-person variation	1. Set of intermediary values that retain mean and between-person variance of the transformed 24-hour recalls is constructed. 2. Back Transformation: inverse of initial power or log. 3. The empirical distribution of backtransformed intermediary values is the estimated usual intake distribution.	1. Set of intermediary values that retain the mean and average between-person variance of the transformed 24-hour recalls is constructed. 2. Back Transformation: inverse of the initial two-stage normality transformation in conjunction with bias-adjustment. 3. The empirical distribution of the original scale intermediary values is the estimated usual intake distribution.	1. Set of intermediary values that retain the mean and average between-person variance of the transformed 24-hour recalls is constructed. 2. Back Transformation: inverse of the initial power of log transformation in conjunction with bias-adjustment. 3. The empirical distribution of the original scale intermediary values is the estimated usual intake distribution.	1. Back Transformation: inverse of the initial two-stage normality transformation in conjunction with bias-adjustment. 2.Combine the estimated consumption day usual intake distribution with the estimated distribution of consumption probability to obtain a set of intermediary values that represent usual intake,assuming that consumption probability and consumption-day intake are statistically independent. 3. The empirical distribution of original-scale intermediary values is the estimated usual intake distribution.

Source: Dodd et al (2006)

Table 4: Dodd et al (2006). Strengths and Weaknesses of the Models

	Method			
	NRC	**ISU**	**BP**	**ISUF**
Strengths	1. Simple transformations to obtain approximate normality. 2. Robust to mild departures from normality. 3. Easy to apply to real data.	1. Can be used for data from complex surveys. 2. Software (SIDE, C-SIDE, PC-SIDE) has been developed to implement the ISU method. 3. Software produces standard errors for estimated parameters of the usual intake distributions. 4. Preliminary data adjustments for nuisance effects such as interview sequence or day-of-week.	1. Simple transformations to obtain approximate normality. 2. Robust to mild departures from normality. 3. Can be used for data from complex surveys. 4. Software (SIDE) has been developed to implement the BP method. 5.Software produces standard errors for estimated parameters of usual intake distributions. 6. Preliminary data adjustments for nuisance effects such as interview sequence or day-of-week.	1. Two-part model that can be applied to datasets with many observed zero intakes. 2. Can be used for data from complex surveys. 3. Software (C-SIDE) has beem developed to implement the ISUF method. 6. Preliminary data adjustments for nuisance effects such as interview sequence or day-of-week in one of the two parts of the model.
Weaknesses	1. Application to datasets from complex surveys is less straightforward. 2. Published computer code does not produce standard errors for estimated parameters of the usual intake distributions. 3. Can not adjust for nuisance effects such as interview sequence or day-of-the-week. 4. Cannot be applied to datasets with many observed zero intakes.	1. Suitable for larger sized datasets. 2. Adjustment for nuisance effects is not done in the context of a unified model. 3.Complex two-stage transformations may fail to obtain exact normality and the subsequent steps of the method rely heavily on the (unsatisfied) normality assumption. 4. Cannot be applied to datasets with many observed zero intakes. 5. Does not allow the use of covariates to model the usual intake distribution.	1. Adjustment for nuisance effects is not done in the context of a unified model. 2. Cannot be applied to datasets with many observed zero intakes.	1. The two parts of the model are estimated independently ignoring the correlation between probability to consume and usual amount consumed. 2. Adjustment for nuisance effects is not done in the context of a unified model. 3. Complex two-stage transformations may fail to obtain exact normality and the subsequent steps of the method rely heavily on the (unsatisfied) normality assumption. 4. The currently available software does not produce standard errors for the estimated parameters of the usual intake distributions. 5. Does not allow the use of covariates to model the usual intake distribution.

Source: Dodd et al (2006)

2.2 National Cancer Institute Method (2006)

The method developed by Tooze et al (2006) is designed to estimate the usual intake of episodically consumed foods by separating the probability of consumption from consumption day amount using a two-part model. Two or more 24-hour recalls are needed for the estimation. The NCI (2006) method allows for correlation between probability of consumption on a single day and consumption day amount. Furthermore, with the NCI (2006) method the relationship between covariates such as sex, age, race, etc. and usual intake could also be modeled.

The NCI (2006) method assumes that the 24-hour recall is an unbiased instrument for usual intake of episodically consumed foods. This assumption implies that the 24-hour recall does not misclassify the individual's food consumption and is an unbiased measure of the food consumed on consumption day. Hence, for intake on consumption day the method is consistent with assumption B described in the Dodd et al (2006) article.

Since the method uses parametric regression analysis to derive the distribution of usual intake the usual assumptions underlying the parametric regression are valid as well. In particular, the method assumes that after transformation the amount of food consumed on a consumption day is approximately normally distributed.

The method is similar to the ISUF (1997) method since it represents the usual intake as the product of probability to consume a food on a particular day and consumption day amount. However, in contrast to the ISUF method it allows for correlation between consumption probability and amount consumed as well as the use of covariates to model the usual intake.

The method consists of two major parts. The first is a statistical model which estimates consumption probabilities, describes the relationship between usual intake and covariates and estimates the within and between-person variability. The second part uses the outputs from the first stage to derive the distribution of usual intake using Monte-Carlo simulations.

2.2.1 Statistical Model

- **Step 1**

In the first step of the statistical model the probability that a day is a consumption day is estimated using logistic regression with a person-specific random effect. The mixed model could incorporate covariates to model the effect of covariate information on probability to consume. The person-specific random effect allows the individual consumption probability to differ from the population level. It could be interpreted as the individual's personal tendency to consume food. The first step of the statistical model could be written as:

$$Logit = Intercept_j + Slope_i \, x \, Covariate_i + Person\text{-}specific\ Effect_i \qquad (13)$$

where Logit for consumption probability p is defined as $\mathrm{logit}(p) = \log(p/1-p)$. The intercept, slope and person-specific effect are the parameters of the model. The model allows for multiple covariates.

- **Step 2**

The second step of the statistical model specifies the transformed consumption day amount as a function of covariates, person-specific effect and within-person variability due to day-to-day variation in an individual's intake and other sources of random error:

$$\textit{Transformed 24-hour recall} = \textit{Intercept}_j + \textit{Slope}_j \textit{ x Covariate}_j + \textit{Person-Specific Effect}_j + \textit{Within-Person Variability}_j \quad (14)$$

The model is developed on the transformed scale where the person-specific effect and within-person random variability are normally distributed. The subscript j indicates that the model parameters differ from those derived in Step 1.The model allows for multiple covariates.

In contrast to the ISUF (1997) method where the two steps above are assumed independent and are modeled separately, the Tooze et al (2006) method fits both steps simultaneously. Hence, the consumption probability and the consumed daily amount could be associated in two ways. First, the two person-specific effects are modeled as correlated random variables. Second, some covariates might be included in both steps which will also induce correlation between them. Furthermore, the Box-Cox transformation parameter used to transform the data towards normality is computed as a part of the mixed model likelihood maximization procedure thus allowing the 24-hour recall normality transformation to depend conditionally on the covariates in the model.

2.2.2 Estimating the Distribution of Usual Intake

In order to obtain the distribution of usual intake the estimated parameters of the statistical model are used to simulate a population with the same characteristics and between-person variability as the sample on which the model was fit. The within-person variability estimated in Step 2 above is not included since by definition it does not contribute to long-term usual intake. The simulation procedure involves the following steps:

1) The parameters for the intercept and covariates are used to obtain an estimate for each individual in the sample.

2) To the above an estimate of the person-specific effect is added. The estimate is generated using normal distribution with mean zero and variance estimated from fitting the statistical model.

3) To improve the accuracy of the estimated usual intake distribution for each person in the sample an additional 100 pseudo-persons are generated each with the same covariate values but different simulated person-specific effect.

4) Since the model parameters are estimated using transformed 24-hour recall data it is necessary to transform the data back to the original scale before estimates of the usual intake distribution are obtained. The back-transformation is similar to the approach used in the ISU (1996) best power method. As described by Dodd et al (2006), an adjustment term is added to correct for bias.

5) The mean, standard deviation and percentiles of the usual intake distribution are estimated empirically from the simulated population.

Tooze et al (2006) argue that the importance of using covariates to model the usual intake distribution depends on the purpose of the analysis. If the aim of the research is to estimate the distribution of usual intake in a population or subpopulation then it is not important how the between-person variability is partitioned between the part explained by the covariates and the unexplained part captured by the person-specific effect. The focus in such cases should be on how well the variability of the person-specific effect and within-person random error could be transformed to normality. Including covariates in the model might make the normality more realistic and improve the accuracy of the estimated distribution, especially in the tails.

When the distribution of usual intake is estimated in subpopulations characterized by sex, age, race, income, education or other covariates then the inclusion of those covariates in the model could lead to substantial improvements in terms of efficiency of the estimation compared to stratification. The advantage of this efficiency is expected to increase as the size of the subpopulation decreases.

The covariate information could help reduce the unexplained between-person variation of nutrient intake when the goal of the research is to predict individual usual intake and relate it to health outcomes, for example. The inclusion of covariates in the estimation of usual intake distribution results in an improved ability to make inferences regarding the effects of the covariates on the probability to consume and the daily amount consumed. Furthermore, the NCI (2006) method allows for separation of the effects of covariates on the probability of consumption and daily amount consumed. The method could also be used to estimate the usual intake of foods or nutrients that are consumed on a daily basis by nearly all individuals.

In terms of drawbacks, it should be noted that the model never predicts a true zero intake since the logistic regression used to model the probability of consumption does not predict a zero value. Furthermore, to obtain reliable estimates the method requires a substantial proportion of the individuals in the sample to have at lest 2 recalls of nutrient consumption. Since the NCI (2006) method is based on the assumption that the 24-hour recall is an unbiased instrument for measuring usual food intake the problems with misreporting mentioned in the paragraph describing Dodd et al (2006) article are applicable to the NCI (2006) method as well.

2.3 The National Academy of Sciences (2003)

The open book publication of the Academy of Sciences “Dietary Reference Intakes: Applications in Dietary Planning” (2003) documents that the individual’s actual intake varies considerably from day to day and it is the usual or long-term intakes that are of major interest in planning dietary programs to ensure nutrient adequacy for individuals or groups. The assessment of nutrient adequacy is facilitated by statistical procedures which perform data adjustments to estimate the distribution of usual intakes from the observed intakes as long as there is more than one day of intake data collected for a representative sample of a group of individuals. These procedures are not designed to estimate the usual intake of a specific individual in the group but rather the adjusted distribution which is used in the analysis of the prevalence of inadequate or excess intakes in the subpopulation of interest.

The publication of the Academy of Sciences discusses in detail two of the major statistical procedures for estimating the distributions of usual intake – the National Research Council (NRC, 1986) method and the Iowa State University (ISU, 1996) method. The open book argues that both methods are based on a common conceptual foundation but the ISU method includes a number of statistical enhancements which make it more suitable for the analysis of large population surveys. The NRC method is simpler and may be more suitable for use with small to medium sized samples. However, it should be noted that both methods have certain limitations.

2.3.1 The National Research Council Method (1986)

According to the publication, the usual intake can not be inferred from the observed intake without error. For any individual the observed intake could be expressed with the formula:

$$Observed\ Intake = Usual\ Intake + Measurement\ Error \quad (15)$$

The observed variance of a distribution of intakes for a group of individuals derived from more than one day of intake per individual is the sum of the variance of the usual intakes of the individuals included in the group (also called between-person variance) and the error in the measurement of the individuals' true intakes. The error is a result of the variation in individuals' intakes from one day to the next as well as random error in the measurement of intakes on any given day. It is also referred to as within-person or day-to-day variance. Hence, the observed variance could be expressed as:

$$V_{observed} = V_{between} + V_{within} + V_{misreporting} \quad (16)$$

The observed distribution of intakes is wider and flatter than the true distribution of usual intakes as a result of the presence of within-person variance. However, if we assume that the within-person variation is random, the estimate of the mean intake of a group will not be affected by this variance. If multiple days of intake data per individual are averaged and the distribution of intakes in a group is constructed from the means of each individual's multiple intakes, then the within-person variance will diminish as a function of days of intake data per individual. Hence, as the number of days of intake data per individual increases the distribution of the observed mean intakes over the days of data collection becomes a better and better approximation of the true distribution of usual intake in the group.

The NRC (1986) method is typically applied to data containing multiple days of intake observations per individual. Ideally, there should be an equal number of observations per individual. The NRC (1986) method partitions the observed variance into between and within-person components and then shifts each individual's mean intake by a function of the ratio of the square root of the between-person variance and observed variance. Therefore, the method attempts to remove the effect of within-person variation on the observed distribution. The variance of the adjusted distribution should be an estimate of the between-person variation.

The NRC method could be summarized with the following four steps:

1) Transformation of data to normality.

Since in most cases the observed intakes data are positively skewed the NRC (1986) method uses some of the common transformation methods for approximating normality (or at least transforming to a symmetric distribution) such as the square root, logarithm or cubed root transforms.

2) Estimation of the within and between-person variance.

The between-person variance is estimated using well-known procedures such as ANOVA as follows:

$$\hat{V}_{between} = \left(MSM - MSE \right) / k \quad (17)$$

where MSM is mean square of the model, MSE is the mean square of the error and k is the mean number days of intake data per individual in the sample. $\hat{V}_{between}$ represents the 'true' variance of the distribution of usual intakes.

It could be assumed that the daily intakes X_{ij} follow the model:

$$X_{ij} = x_i + u_{ij,} \quad i = 1,...,n,\ j = 1,....,k \tag{18}$$

where $x_i \sim NI\left(\mu_x, \sigma_x^2\right)$ is the unobservable usual intake, $u_{ij} \sim NI\left(0, \sigma_u^2\right)$ is the measurement error, i denotes the ith individual and j denotes the jth repeated measurement.

Table 5 illustrates the derivation of the sum of squares and expected mean squares in the case of one way ANOVA for balanced data:

Table 5: ANOVA Table: One-way Classification with Balanced Data

Source of Variability	Sum of Squares	Degrees of Freedom	Expected Mean Squares
Individual	$k\sum_{i=1}^{n}\left(\bar{X}_{i.} - \bar{X}_{..}\right)^2$	$n-1$	$\sigma_u^2 + k\sigma_x^2$
Error	$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_{i.}\right)^2$	$n\left(k-1\right)$	σ_u^2
Total	$\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_{..}\right)^2$	$nk-1$	

Source: A Technical Guide to SIDE (1996)

The usual estimators shown in Table 5 are defined as:

$$\hat{\mu}_x = \bar{X}_{..} = \left(nk\right)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij} \tag{19}$$

$$\bar{X}_{i.} = k^{-1}\sum_{j=1}^{k} X_{ij} \tag{20}$$

$$\hat{\sigma}_u^2 = \left[n\left(k-1\right)\right]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_{i.}\right)^2 \tag{21}$$

$$\hat{\sigma}_x^2 = k^{-1}\left[\left(n-1\right)k\sum_{i=1}^{n}\left(\bar{X}_{i.} - \bar{X}_{..}\right)^2 - \hat{\sigma}_u^2\right] \tag{22}$$

3) Adjustment of the individual's mean intake to estimate the distribution of usual intakes.

Each individual's mean intake is adjusted using the following formula:

$$\text{Estimated adjusted Intake} = \left[\text{estimated individual's mean- estimated group mean}\right] \text{x} \left(\hat{SD}_{between} / \hat{SD}_{observed}\right) + \text{estimated group mean} \tag{23}$$

where $\hat{SD}_{between}$ and $\hat{SD}_{observed}$ are the square roots of the estimates of $\hat{V}_{between}$ and $\hat{V}_{observed}$, respectively. $\hat{V}_{observed}$ is the estimated variance of the observed mean daily intake defined in equation (16) above.

If the data were transformed in Step 1 prior to partitioning the variance in Step 3 the above equation is applied to the individual's means and group mean calculated from the transformed data. If there weren't any data transformations the adjusted intakes represent the estimated distribution of usual intakes. The adjustment procedure could also be applied to ratios of dietary intakes.

4) Back-transformation of the adjusted data to the original scale.

If the data were transformed to approximate normality the adjusted data need to be transformed back to the original scale prior to using the adjusted distribution for nutrient assessment. The back-transformation is the inverse function of the original transformation.

There are several features of the NRC (1986) method that need to be noted. Probably the most important one is the requirement for normally distributed data. The statistical validity of the NRC (1986) method hinges on the assumption of normally distributed intake data. If the transformation to normality is completely achieved then the means and medians of the original and adjusted intake distributions will be equivalent. However, if the transformation only approximates normality then the means and the medians of the original and adjusted usual intake distributions will differ. It should be noted that the observed distributions of some nutrient intakes such as vitamins are highly skewed or might contain a large proportion of zero values and achieving normality with simple transformations is almost impossible. In such cases the NRC (1986) method and the resultant adjusted long-term intake distributions should be used cautiously for nutrient assessment or avoided altogether.

The NRC (1986) method does not take into account the systematic effects on the within-person variance and does not account for autocorrelation in the observed intakes. It estimates the adjusted distribution of usual intakes for a sample only since the method does not incorporate sampling weights.

If some individuals have missing data the denominator in the calculation of the $V_{between}$ should be adjusted accordingly. In larger samples the decomposition of the variance should be performed separately for particular subgroups of the population. If the intake data have been collected across more than one life stage, gender, income or education level strata then the adjustment of intake data for both NRC (1986) and ISU (1996) methods should be performed separately for each stratum.

Carriquiry (2003) documents that although the NRC (1986) method is simple to implement it has a major flaw since it may introduce bias into the estimator of the usual intake distribution by applying a simple inverse transformation to the adjusted mean usual intakes. The author notes that the mean of a nonlinearly transformed variable is typically not equal to the transformation of the mean of the variable. Hence, if the NRC (1986) method uses a transformation and its inverse on the adjusted means it will result in biased estimates of the usual intake distribution parameters in the original scale.

2.3.2 The Iowa State University Method (1996)

The method is described in detail in Paragraph 2.5. However, an overview and discussion of some of the limitations of the method are included as an introduction to the ISU (1996) procedure. The method is implemented through a software packages called SIDE or C-SIDE. It can be used to adjust the observed dietary intakes in large surveys as long as two nonconsecutive or three consecutive days of individual's intake observations are included in the data. Compared to the NRC (1986) method the ISU (1996) method represents a major advancement since it includes a number of statistical enhancements designed to improve the accuracy of the estimated usual intake distribution and its parameters.

The ISU (1996) method transforms the intakes for a nutrient to the standard normal distribution using complicated techniques which are a combination of power transformation and grafted polynomial. The distribution of usual intakes is estimated from the distribution of the transformed intakes and then back-transformed to the original scale using bias-adjusted back transformation. Furthermore, the method is designed to adjust for nuisance effects such as day of the week, time of the year and sequence of observation that may introduce systematic effects on the observed distribution of usual intake. The ISU (1996) method also accounts for the correlation between observations on consecutive days and for heterogeneous within-person variance. The method incorporates sampling weights and could be used to estimate the adjusted distribution of usual intake not only in a sample but in a target population as well.

However, the ISU (1996) method might not be the appropriate method to use with small samples. The greater complexity of the method requires larger samples to attain acceptable levels of reliability during the various transformation stages. In contrast, the NRC (1986) method could be used with small samples since the adjustment to normality step of the method is more simplistic.

An important difference between NRC (1986) and ISU (1996) methods is that the ISU method is typically applied to the distribution of usual intakes on day one of the data collection. In contrast, the NRC method is applied to the multiple-day mean of observed intake. The underlying assumption of the NRC method is that all days have equivalent validity. If there is a sequence effect and the reported nutrient intake declines systematically across multiple days of data collection the adjustment of intakes to day-one data will result in a higher usual intake estimate compared to the one derived using multiple-day means (Guenter et al 1997). However, if the day-one intake is more accurately reported than the intake on subsequent days then the day-one data adjustment will result in less biased estimates. At present there are no reliable methods to establish the validity of the self-reported intakes and it is difficult to determine whether day-one data or multiple-day means data will lead to better estimates of true intake.

2.4 Hoffmann et al (2002a, 2002b)

Hoffmann et al (2002a) document the research undertaken on the availability and efficacy of the appropriate statistical methods for estimating the usual dietary intake distributions of as part of the EFCOSUM 2000/2001 project. The project European Food Consumption Survey Method is part of the EU Programme on Health Monitoring. The aim of the EFCOSUM was to establish a method for monitoring dietary intakes using nationally representative samples of all age/sex categories in Europe in a comparable way. A total of 23 European countries participated in this project. The conclusions of the EFCOSUM project are that the 24-hour recall interview is the most suitable method for estimating of the population usual intake means and distributions. The number of repeated measurements and participants in dietary surveys are a compromise between theoretical considerations and practical constraints but a minimum sample size of 2000 adults in each European country will be needed to identify trends in mean intakes of nutrients and foods with the desired precision of 5%. For presentation purposes the usual intake distribution parameters of interest are the mean, median, quartiles as well as the distribution tail estimates such as P5 and P95 percentiles. (EFCOSUM, 2002). The EFCOSUM proposes the use of the simplified ISU method developed by Hoffmann et al (2002a) for estimating the distributions of usual dietary intake of nutrients and food groups.

Hoffmann et al (2002a) document that there are two different ways to assess usual dietary intake. One of them is to use dietary assessment methods such as food frequency questionnaires (FFQs). The other alternative is to use repeated short-term measurements such as 24-hour diet recalls or food records. In the latter case the variance of the reported intake is inflated by additional day-to day variation of individual dietary intake (Willet 1998). Hence, to estimate long-term dietary intake the so-called intra-individual or within-subject variability of the data must be removed by appropriate statistical methods.

Hoffmann et al document that the use of FFQs for assessing and comparing usual dietary intakes have been criticized by many academic researchers. Some examples include the work of Kushi (1994), Liu (1994), Sempos (1992) and Briefel et al (1992). The main disadvantages of FFQs pointed out by a number of researchers include the reliance of the FFQs on the subjects' long-term memory and capability of correct averaging. The FFQs are country-specific and are based on a selection of closed questions with predetermined food frequencies. Hence, dietary intake data obtained with different FFQs are often not directly comparable and could have considerable bias. Some researchers such as Kipnis et al (2001), Kroke et al (1999), Sawaya et al (1996) and Martin et al (1996) document that data collected by 24-hour recalls also have bias compared to biomarker measurements. However, Hoffmann et al (2002a) argue that the size of the systematic measurement error on a group level should be moderate compared to that of the FFQs.

Hoffmann et al (2002a) investigate 6 different statistical methods for estimating the distribution of usual intake using short-term measurements. The investigated methods include the methods proposed by Slob (1993), Wallace et al (1994), Buck et al (1995), Nusser et al (1996), Guy (2000) and the simplified ISU method developed by Hoffmann et al (2002a).

According to Hoffmann et al (2002a) the methods share the common goal of eliminating the intra-individual variability of the short-term measurements or at least reducing its impact on the estimated usual intake distribution. Hoffmann et al (2002a) summarise the main characteristics of the six methods as follows:

Table 6: Hoffmann et al (2002a). Comparison of Methods for Estimating Usual Intake Distribution

Method	Assumed Distribution	Data Transformation	Statistical Approach	Software Package
Slob method (Slob, 1993)	Log-normal	Logarithmic	Variance decomposition	ANOVA (SAS,SPSS)
Wallace method (Wallace et al,1994)	Log-normal	Logarithmic	System of equations	Small programme
Buck method (Buck et al, 1994)	Symmetric	None	Variance decomposition	ANOVA (SAS,SPSS)
ISU method (Nusser et al, 1996)	Normal after transformation	Power/grafted polynomial	Variance decomposition	SIDE (SAS), C-SIDE (UNIX)
Guy method (Guy,2000)	Normal after transformation	Power	Variance simulation	ANOVA, simulation programme
S-Nusser method (Hoffmann et al,2002)	Normal after transformation	Two-parameter Box-Cox	Variance decomposition	ANOVA,SAS macro

Hoffmann et al (2002a) document that the intra-individual variance component could be reduced if individual means of repeated measurements are used instead of single measurements. The intra-individual variability of the individual means decreases as the number of repetitions increases and vanishes as the number of repetitions reaches its maximum when it covers the whole period of interest. Since it is impossible and impractical to have such a large number of repeated measurements the method proposed by Hoffmann et al (2002a) estimates the intra-individual variance component and subtracts this component from the observed variance of the nutrient intake.

The inter-individual variance can be estimated using the formula:

$$\hat{\sigma}^2_{usual} = \hat{\sigma}^2_{\bar{X}} - \frac{1}{k}\hat{\sigma}^2_{\varepsilon} \qquad (24)$$

where k is the number of repetitions, $\hat{\sigma}^2_{\bar{X}}$ is the observed variance of the individual's mean intake and $\hat{\sigma}^2_{\varepsilon}$ is the estimated average intra-individual variance.

If ANOVA is used for variance decomposition then the above variance formula could also be written as:

$$\hat{\sigma}^2_{usual} = \frac{1}{k}\left(MSM - MSE\right) \tag{25}$$

where MSM and MSE are the model and error mean squares, respectively with a subject identifier as the only independent model variable.

If there are different numbers k_i of repetitions for the n subjects, then the denominator in equations (24) and (25) above must be replaced by a denominator calculated using the following formula:

$$k^* = \left[\left(\sum k_i\right)^2 - \sum k_i^2\right] / \left[(n-1)\sum k_i\right] \tag{26}$$

in order to derive an unbiased estimator (Donner,1986). In the context of dietary surveys the conditions under which the above estimator is unbiased are simple random sampling of individuals and days as well as constant variability across individuals for the specific nutrient or food element.

Hoffmann et al (2002a) propose to estimate the mean and variance of the usual intake distribution by the sample mean and the estimator given in equation (25) above. The authors argue that although in general a distribution can not be constructed on the sole basis of mean and variance alone in the special case of normally distributed data both parameters uniquely determine the whole distribution and all percentiles can be calculated from the mean and variance using a simple equation.

Buck et al (1995) propose a method which shrinks the observed variance to the estimated inter-individual variance and afterwards applies the simple percentile equation which is accurate only if the sample distribution is nearly normal or at least nearly symmetric. In order to preserve the grand mean $\overline{X}_{..}$ Hoffmann et al (2002a) modify the Buck et al method (1995) so that the percentiles of the usual intake distribution are calculated from the variables:

$$\hat{Z}_i = \frac{\hat{\sigma}_{usual}}{\hat{\sigma}_{\overline{X}}}\left(\overline{X}_i - \overline{X}_{..}\right) + \overline{X}_{..} \tag{27}$$

where $\overline{X}_i$ is the individual mean intake of the ith subject.

Some of the other methods included in Table 6 such as those of Slob (1993) and Nusser et al (1996) also reduce the variance of the individual means by eliminating the intra-individual component and include data transformation steps to obtain approximate normality. Slob (1993) uses logarithmic transformations whereas Nusser et al (1996) use a more complicated family of transformations which adjust for nuisance effects and improve the transformation to normality with grafted polynomials. In contrast, Guy (2000) also uses data transformations but estimates the variance of the usual intake distribution using simulation of a 52-week sample. Wallace et al (1994) use a completely different approach based on the assumptions that inter and intra-individual components of the variance are independent and log-normally distributed. Wallace et al (1994) estimate the parameters of the usual dietary intake distribution by a number of equations for the geometric mean, arithmetic mean and geometric standard deviation calculated for each combination of sampling days. However, Slob (1996) argues that the Wallace et al (1994) procedure yields estimates which are very unstable compared to the estimates based on the ANOVA procedure.

Hoffmann et al (2002a) evaluate the six methods for estimating the usual daily intake distributions using data on vegetables and total fat consumption collected by repeated 24h diet recalls. The choice of target characteristics is motivated by the EFCOSUM project since they both belong to the relevant nutrition indicators. Furthermore, Hoffmann et al (2002a) investigate the performance of the six methods with an increasing number of sampling days. Four, eight and twelve sampling days per individual are formed by one, two and three days per season whereas two sampling days per individual are formed by one sampling day in winter and summer, respectively. The results indicate that the standard deviation decreases as the number of repetitions increases. The decrease in the standard deviation is combined with an increase in the lower percentiles (P5, P10 and P25) and decrease in the upper percentiles (P75, P90 and P95). The arithmetic mean is constant over the different data sets. The estimated distributions using the 5 methods are very different from the sample distribution of the 2-day means and more similar to the sample distribution of the 12-day means.

Hoffman et al (2002a) document that the reduction of the standard deviation of the sample distribution with the increase of the number of sampling days is more significant for the vegetable than the total fat intake. They argue that this is as a result of the higher ratio of intra to inter-individual variance for the vegetable intake. Furthermore, the increase in the lower percentiles and decrease in the upper percentiles is more pronounced in the case of the vegetable intake.

Hoffmann et al (2002a) document that the most flexible and efficient method is the ISU (1996) method. The transformation procedures which are part of the method guarantee in most cases that the distribution of the transformed data is near normality. As drawbacks of the method Hoffmann et al (2002a) point out the considerable computation effort and the need for special software packages SIDE or C-SIDE. However, the software packages allow for more sophisticated analyses such as initial adjustments for nuisance effects and incorporation of sampling weights. Furthermore, they allow for correlation in case of consecutive sampling days. The software could be used for the analysis of foods and food groups which are not consumed daily, assuming that the probability of consumption is independent of the amount consumed (Nusser et al, 1997).

The methods proposed by Slob (1993) and Buck (1994) need only an ANOVA procedure which is available in most statistical software packages. The implementation of the Wallace (1994) method is time consuming but does not need a special software package. The Gay (2000) method needs software for the simulations and these simulations are not described in detail in his paper. Hoffmann et al (2002a) document that the Slob (1993) method has the drawback of a missing correction for intra-individual variation in the back transformation with the serious implication of producing biased usual intake distribution estimates.

The other four methods give similar results for the estimated usual intake distributions with the exception of the skewness estimates. The Buck et al (1995) method always reproduces the skewness of the sample distribution of the used data set whereas the Wallace (1994), ISU (1996) and the proposed S-Nusser (Hoffman et al 2002a) methods derive reduced skewness estimates. However, the distribution of the 12-day means of fat intake is almost symmetric and all methods investigated by Hoffmann et al (2002a) seem to overestimate the skewness of this distribution. The Wallace et al method (1994) failed to produce a positive-valued solution for the intra-individual variance of the vegetable intake distribution and the distribution could not be estimated using this method. For vegetable intake the Slob (1993) method derives an unrealistic estimated distribution with standard deviation and arithmetic mean which are too small as well as negative skewness. Hoffmann et al (2002a) argue that the failure of the two methods could be explained by violation of the assumption concerning the underlying distribution since the vegetable intake data are not log-normally distributed. The log-normal distribution assumption is fulfilled only by the total fat intake data but only approximately. In contrast, the Buck et al (1995) method proves to be markedly robust against departures from normality.

Hoffmann et al (2002a) conclude that ISU (1996) and the S-Nusser methods (Hoffmann et al 2002a) are suitable for estimating the usual intake distributions of a broad class of normally and non-normally distributed food groups and nutrients. The Buck et al (1995) method is a robust procedure applicable to intake data in which the intra and inter-individual variations have similar degrees of non-normality. The methods of Slob (1993) and Wallace et al (1994) fail seriously if the assumption of log-normality is violated.

The authors document that the results indicate similar percentile estimates for 2 and 7 survey days. Guy (2000) also calculates the 95% confidence intervals for estimated percentiles of usual intake for 2, 4 and 7 sampling days and documents some evidence that the widths of the confidence intervals of the distribution parameter estimates do not differ significantly. Hence, Hoffmann et al (2002a) argue that the knowledge that only 2 days per individual are necessary to estimate usual intake distribution could help reduce the costs of the food consumption surveys. The authors note that it is important that sampling days cover all seasons and all days of the week. The best sampling design is based on a simple random selection procedure of the sampling days. If the sample size is sufficiently large, the randomly selected days should be representative of the period of interest. However, if a simple random sample is not feasible and produces an unequal distribution of seasons and days the data should be adjusted for significant nuisance effects by linear regression (Nusser et al, 1996) or should be weighted to achieve equal sums of weights for each season and each day of the week (Gay,2000).

It should be noted that Hoffmann et al (2002a) document that 2 days of nutrient intake are applicable only to the problem of estimating the usual intake distribution. If the research has different objectives such as to assess the individual's usual intake or to rank individuals by their usual intake, then a higher number of daily measurements is required (Volatier et al 2002, Hartman et al 1990, Nelson et al 1989). In general, the minimum days required for estimating individual usual intake ranges between 3 and 10 days for energy and macronutrients. On average 20-50 days are needed for food components with large day-to-day variation such as cholesterol or vitamins A and C (Buzzard, 1998).

Hoffmann et al (2002a) argue that sampling designs based on non-consecutive days in different seasons should be preferred. Sampling designs based on adjacent days can lead to overestimation of the population variance of usual intake. Tarasuk and Beaton (1992) find some evidence that mean intake estimates derived from samples on consecutive days are less reliable and more likely to be biased compared to those derived using randomly selected non-consecutive days. Furthermore, Hartman et al (1990) showed that consecutive-day intakes are more highly correlated than non-consecutive day intakes and hence the correlation with true usual intake is reduced in surveys based on consecutive days.

Hoffmann et al (2002a) note that another approach to handle intake data of non-consumers would be to identify them based on additional information obtained from a question as to whether the individual is a non-consumer of a specific food or food group included as a part of a short questionnaire. The additional information will help to separate consumers from non-consumers and a slightly modified S-Nusser method (Hoffmann et al 2002a) can be applied. However, it should be noted that it is not possible to identify non-consumers by zero intakes only, without any additional information because such identification depends on the number of sampling days. It is well documented in the research literature that the percentage of individuals with at least one consumption day increases with the increase of the number of sampling days (Lowik et al 1999, Lambe and Kearny 1999, Lambe at al 2000).

Based on the results from the investigation of the five methods Hoffmann et al (2002a) propose a new, simplified ISU (1996) method. The simplified method is a three-step procedure which could be implemented in most standard statistical packages. The proposed simplified version is not optimal but requires considerably less computational effort. The steps of the proposed method are as follows:

1) The observed intakes of a particular nutrient R_{ij} are transformed to obtain a distribution which is at least approximately normal. The two-parameter Box-Cox (1964) family of transformations is used and it could be defined as:

$$g(R) = g_{\tau,\omega} \begin{cases} \left((R+\omega)^{\tau} - 1 \right) \tau^{-1}, \tau \neq 0 \\ \ln(R+\omega), \qquad \tau = 0 \end{cases} \tag{28}$$

Hoffmann et al (2002a) use only transformations with power parameter $\tau = 0$ or the inverse of positive integers. This restriction simplifies greatly the back transformation to the original scale since an exact formula could be used. However, due to the above restrictions the power parameter τ and the shift parameter ω can not be estimated by the common maximum likelihood method and a grid search procedure which maximises the Shapiro-Wilk statistics is used.

The subsequent steps of the procedure are based on the assumption that the classic measurement error model with normally distributed components holds for the transformed dietary intake.
If we denote the transformed measurements as $X_{ij} = g(R_{ij})$ then the measurement error model could be written as:

$$X_{ij} = T_i + \varepsilon_{ij} \tag{29}$$

where T_i denotes the true usual intake of the ith individual on the transformed scale. 'Usual' denotes a long-term daily average. In most cases the long-term daily average is over a one year period.

The error term ε_{ij} includes both day-to-day variation and the random measurement error. Both T_i and ε_{ij} are assumed to be independent and normally distributed with expectations μ and 0 and variances σ_T^2 and σ_ε^2, respectively. Hence, the distribution of X_{ij} is normal with expectation and variance given by:

$$X_{ij} \sim N\left(\mu, \sigma_T^2 + \sigma_\varepsilon^2\right) \tag{30}$$

Furthermore, the average $\overline{X}_i$ for the ith individual has the same distribution as X_{ij} but with variance is reduced to:

$$\sigma_{\overline{X}}^2 = \sigma_T^2 + \frac{\sigma_\varepsilon^2}{k} \tag{31}$$

where k denotes the number of replicates per individual.

The variable $\hat{T}_i$ defined as:

$$\hat{T}_i = \frac{\sqrt{\sigma^2_{\bar{X}} - \frac{\sigma^2_\varepsilon}{k}}}{\sigma_{\bar{X}}}\left[\bar{X}_i - \mu\right] + \mu \qquad (32)$$

has the same distribution as the transformed usual intake T_i. Hence, if we use the standard estimators for the unknown parameters in the above equation, T_i could be estimated as:

$$\hat{T}_i = \frac{\sqrt{\hat{\sigma}^2_{\bar{X}} - \frac{\hat{\sigma}^2_\varepsilon}{k}}}{\hat{\sigma}_{\bar{X}}}\left[\bar{X}_{i.} - \bar{X}..\right] + \bar{X}.. \qquad (33)$$

where $\hat{\sigma}^2$ denotes the corresponding empirical variance and $\bar{X}..$ denotes the grand mean.

As seen from formula (33) above, $\hat{T}_i$ is a shrinkage estimator which shrinks the individual mean $\bar{X}_{i.}$ to the grand mean thus removing the intra-individual variation in the individual means. In rare cases when the quantity under the square root is negative the variance component of T_i should be estimated using the non-negative minimum biased invariant estimator of Hartung (1981).

In the third step of the method the estimated usual intake $\hat{T}_i$ is back-transformed to the original scale of reference measurements by integrating the inverse function $g^{-1}\left(t+\varepsilon\right)$ over the normal distribution of the error term ε. In contrast to the approximation used in Nusser et al (1996) an exact solution exists provided the power parameter used in step 1 is zero or the inverse of positive integers. The explicit formula ensures that the arithmetic means of the original and back transformed data are equal, provided that the distribution of the transformed data is normal or at least symmetric. Furthermore, the back transformation formula improves the accuracy of the estimations and reduces the computational effort needed for the analysis. The explicit formula could be written as follows:

$$g_{back}\left(t\right) = \int_{-\infty}^{\infty} g^{-1}\left(t+\varepsilon\right)\varphi\left(\varepsilon\right)d\varepsilon \qquad (34)$$

where t is any value on the transformed scale. Since t is measured with an error ε the inverse function g^{-1} of the Box-Cox transformation (28) is applied to the term $t+\varepsilon$ and the integration in the general formula (34) above is over the distribution of the error term ε. In equation (34) φ is the density of the normal distribution with mean 0 and variance σ^2_ε.

In the special case where $g(R)=\ln(R+\omega)$ Hoffmann et al (2002a) obtain the well known result:

$$\begin{aligned} g_{back}(t) &= \int_{-\infty}^{+\infty} \exp\{t+\varepsilon\} \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\varepsilon^2\right\} d\varepsilon - \omega \\ &= \exp\left\{t+\frac{1}{2}\sigma_\varepsilon^2\right\} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi\sigma_\varepsilon}} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2}\left(\varepsilon-\sigma_\varepsilon^2\right)^2\right\} d\varepsilon - \omega \\ &= \exp\left\{t+\frac{1}{2}\sigma_\varepsilon^2\right\} - \omega \end{aligned} \tag{35}$$

Furthermore, in cases when the inverse power p of the transformation function g is a positive integer Hoffmann et al (2002a) calculate the back transformation using the binomial formula:

$$\begin{aligned} g_{back}(t) &= \int_{-\infty}^{\infty} \left(p^{-1}(t+\varepsilon)+1\right)^p \varphi(\varepsilon)\, d\varepsilon - \omega \\ &= p^{-p} \sum_{r=0}^{p} \binom{p}{r} (t+p)^{p-r} \int_{-\infty}^{\infty} \varepsilon^r \varphi(\varepsilon)\, d\varepsilon - \omega \\ &= p^{-p} \sum_{s=0}^{\mathrm{int}\left(\frac{p}{2}\right)} \binom{p}{2s} (t+p)^{p-2s} \sigma_\varepsilon^{2s} (2s-1)!! - \omega \end{aligned} \tag{36}$$

where $(2s-1)!!$ denotes the product of all uneven integers from 1 to x with the exception of $(-1)!!=1$. Hence, it could be rewritten as $(2s-1)!! = \dfrac{(2s-1)!}{2^{s-1}(s-1)!}$.

Hoffmann et al (2002a) argue that the empirical results indicate that the above formulae hold also for symmetric non-normal error distributions. This could be interpreted as a robustness property of the proposed S-Nusser method. If some individuals are known to be non-consumers and have zero daily intake measurements, then they have zero intake by definition, and the estimation of the usual intake distribution is based on the pooled data of the back transformed usual intakes of consumers and the zero usual intakes of non-consumers.

Since the major drawback of the FFQ assessment instrument is the lack of comparability between data collected using different FFQs, Hoffmann et al (2002b) propose a method for standardization of the dietary intake measurements collected by different FFQs which removes this deficiency. The proposed method is a combination of the S-Nusser method (Hoffmann et al 2002a) and a nonlinear calibration procedure. The method could be used for pooled dietary intake data collected in multicenter and multiethnic studies. Furthermore, the method could also be used in studies with repeated standardized reference measurements that refer to different time periods. The method is also suitable for short-term reference assessment methods such as 24-hour recalls, diet records and biomarkers.

2.5 Nusser et al (1996)

Nusser et al (1996) suggest a methodology for estimating usual intake distributions which allows for varying degrees of departure from normality and consists of four major steps that could be summarized as follows:

1) The original data is standardized in order to adjust for nuisance effects such as day of the week, month, interview sequence, etc.

2) The daily intake data are transformed to normality using a combination of power and grafted polynomial transformations.

3) The distribution of the usual intakes is constructed using the transformed data and normal components-of-variance model.

4) The transformed data is converted back to the original scale.

2.5.1 Data Adjustments

The adjustment process is implemented using standard OLS regression techniques. Hence, the consumption data are regressed on variables representing nuisance effects using linear regressions. For example, dummy variables representing month or day of the week are regressed on the dietary intake data. In the case of 1985 CFII data used by Nusser et al (1996) least squares methods are used to investigate whether day of the week, month, interview mode and sequence are statistically important. The authors found weekday and interview sequence effects statistically significant and the data are adjusted to remove those nuisance effects. Regression models are also used for ratio adjustments to the first-day mean.

Since the consumption data are often skewed the following formula is used in a grid-search procedure to compute the power γ used to power-transform the observed intake towards normality:

$$\sum_{i=1}^{n} w_i \sum_{j=1}^{k} \left(U_{ij} - \beta_0 - \beta_1 W_{0ij}^{\gamma} \right)^2 \qquad (37)$$

In the above formula w_i is the sampling weight for the ith individual. W_{0ij}^{γ} denote the observed intake for the ith individual on the jth day in the interview sequence plus a constant equal to 0.0001 times the sample mean for the nutrient. The constant is added to avoid subsequent computational problems in procedures which depend on the derivative of the power of the data. The coefficients β_0 and β_1 are estimated in the grid search procedure for each value of γ. U_{ij} is the Blom's (1958) normal score for the $ijth$ observation computed using the following formula:

$$U_{ij} = \Phi^{-1}\left[\left(\frac{s_{ij} - 3}{8} \right) / \left(\frac{nk+1}{4} \right) \right] \qquad (38)$$

where Φ is the standard normal distribution function and s_{ij} is the rank of the $ijth$ observation.

Once an optimal power is selected a model containing dummy variables for the day of the week, month, interview mode and interview sequence is fitted by weighted least squares to the power transformed observed intake. The weights in the regression are the sampling weights. The adjustment method could be illustrated with the following equation:

$$Z_{ij} = \hat{Z}_{0ij}^{-1}\overline{Z}_{01}Z_{0ij} \tag{39}$$

where Z_{0ij} are the power-transformed observed intake values and $\overline{Z}_{01}$ is the mean of the observed power-transformed intakes for the first interview day. $\hat{Z}_{0ij}$ are the predicted intakes from the weighted least squares regression with nuisance effects for the ith individual on the jth day.

The data are adjusted to the mean of the first interview day because it is assumed that the first interview day data are more accurate compared to the data collected on subsequent days. Furthermore, in order to reduce the time-in-sample effects (Bailar 1975) the sample variance of the transformed values is standardized to the sample variance observed on the first day. Hence, if we take the inverse power of the transformed observations we could define the adjusted observations in the original scale as:

$$Y_{ij}^{*} = \left[\hat{\mu}_1 + S_j^{-1}S_1\left(Z_{ij} - \hat{\mu}_j\right)\right]^{\frac{1}{\lambda}} \tag{40}$$

where:

$$S_j^2 = \left(n-1\right)^{-1}\sum_{i=1}^{n}\left(Z_{ij} - \hat{\mu}_j\right)^2 \tag{41}$$

$$\hat{\mu}_j = n^{-1}\sum_{i=1}^{n} Z_{ij} \tag{42}$$

In equation (40) $\hat{\mu}_1$ and S_1 are the mean and standard deviation observed on the first day.

The ISU (1996) method could also be used for the analysis of data collected in surveys using complex sampling designs because it allows for sampling weights in the analysis. If weights are included in the data the ISU method derives an equal-weight sample from the adjusted unequal-weight sample. The adjusted equal-weight intakes Y_{ij} are referred to as daily intakes. In such cases in equation (37) $w_i = 1$ and W_{0ij} is replaced by Y_{ij}. However, since the power transformation defined by equation (37) in most cases does not produce satisfactory results a further transformation is required to improve the transformation of the data to normality. A semiparametric regression is used to fit a grafted cubic polynomial to the $\left(U_{ij}Y_{ij}^{\lambda}\right)$ pairs. The number of joint points is chosen to be the minimum number required to make the Anderson-Darling (Anderson and Darling, 1952) test statistics for normality less than or equal to 0.58.

2.5.2 Estimating the Usual intake Distribution in the Normal Scale

Nusser et al (1996) use a measurement error model for estimating the distribution of usual intake in the normal scale. Hence, following the notation used in the paper the transformed daily intakes could be defined as:

$$X_{ij} = x_i + u_{ij} \tag{43}$$

where $x_i \sim NI\left(\mu_x, \sigma_x^2\right)$, $u_{ij} \sim NI\left(0, \sigma_{ui}^2\right)$ and $\sigma_{ui}^2 \sim \left(\mu_A, \sigma_A^2\right)$. In the above equation x_i is the unobservable usual intake value for individual i in the normal scale and u_{ij} is the unobservable measurement error for individual i on day j in the normal scale. It is assumed that u_{ij} as well as x_i are independent for all i and j. Since normal scores are used to transform the daily intakes to normality the transformed daily intakes have $\mu_x = 0$ and $\sigma_x^2 = 1$. The individual means $\overline{X}_i = k^{-1}\sum_{j=1}^{k} X_{ij}$ are independent $\left(0, \sigma_{\overline{X}}^2\right)$ random variables where, similarly to equation (31) $\sigma_{\overline{X}}^2 = \sigma_i^2 + k^{-1}\mu_A$. Here k denotes the number of replicates per individual.

Similarly to equations (19), (20), (21) and (22) we could also define the estimators of the moments as:

$$\hat{\mu}_r = n^{-1}\sum_{i=1}^{n} \overline{X}_i \tag{44}$$

$$\hat{\sigma}_X^2 = \left(n-1\right)^{-1}\sum_{i=1}^{n}\left(\overline{X}_i - \hat{\mu}_r\right)^2 \tag{45}$$

$$\hat{\mu}_A = \left[n\left(k-1\right)\right]^{-1}\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \overline{X}_i\right)^2 \tag{46}$$

$$\hat{\sigma}_i^2 = \hat{\sigma}_X^2 - k^{-1}\hat{\mu}_A \tag{47}$$

$$\hat{\sigma}_A^2 = n^{-1}\left[1 + 2\left(k-1\right)^{-1}\right]\sum_{i=1}^{n}\left(A_i^2 - \overline{A}^2\right) \tag{48}$$

where

$$A_i = \left(k-1\right)^{-1}\sum_{j=1}^{k}\left(X_{ij} - \overline{X}_i\right)^2 \tag{49}$$

and

$$\overline{A} = \sum_{i=1}^{n} A_i \tag{50}$$

2.5.3 Estimating the Usual Intake Distribution in the Original Scale

The transformation used to transform the adjusted observed intakes Y to normality is denoted by g. Then, if we denote with y_i the usual intake on the original scale it can be shown that:

$$y_i = E\{Y \mid x = x_i\} = E\{g^{-1}(x+u) \mid x = x_i\} = h(\dot{x}_i) \quad (51)$$

The distribution of the measurement error u has mean zero and variance with an estimated mean $\hat{\sigma}_u^2$.

Nusser et al (1996) use a nine-point approximation to the distribution of the measurement error u to define the usual intake on the original scale as:

$$y_i = \sum_{l=-4}^{4} w_l g^{-1}(x_i + c_l) \quad (52)$$

where x_i is the ith value in the normal scale, c_l and $w_l\,(l = -4, -3 \ldots 3, 4)$ are the values and weights $\sum w_l = 1$ for the nine-point approximation to the distribution of u. The weights w_l and the nine points c_l are defined so that the first five moments of the descrete nine-point distribution match the first five moments of the distribution of $\dot{x} + u \mid x$. A sample of 400 values of $\dot{x}$ is selected to derive the approximation. A grafted cubic polynomial created from the 400 pairs $(\dot{y}_i, \dot{x}_i)$ denoted by $h(\ )$ is the transformation of the transformed usual intake $\dot{x}$ into the usual intake $\dot{y}$ on the original scale.

However, as noted by Hoffmann et al (2002a) the above back transformation method is only an approximation of the integral of the inverse function $g^{-1}(x+u)$. In the same paper Nusser et al (1996) propose a simplified best power method which uses a Taylor series approximation to adjust for bias when the usual intakes are back-transformed to the original scale after an initial nonlinear transformation to approximate normality. The back-transformation is reviewed in detail in the section describing the Dodd et al (2006) article.

The estimates of the first five moments of the usual intake distribution (y) are obtained from the representative sample $\{y_i\}_{i=1}^{400}$. The estimate of the cumulative distribution function is obtained using the equation:

$$\hat{F}(y_i) = \hat{\mathrm{P}}(y \le y_i) = \frac{i - 3/8}{400.25} \quad (53)$$

where $y_1 < y_2 < \ldots\ldots < y_{400}$. Linear interpolation of $\hat{F}$ is used to compute the quantiles and cumulative distribution function values.

For any fixed point A the estimated probability that $x \le A$ is:

$$\hat{F}(A) = \Phi\left(\hat{\sigma}_x^{-1}\left(A - \hat{\mu}_x\right)\right) \quad (54)$$

where $\Phi(.)$ is the standard normal density and x is assumed to be $N\left(\mu_x, \sigma_x^2\right)$. The usual estimators of μ_x and σ_x^2 are shown in equations (19) and (22), respectively.

A Taylor expansion is used to derive the standard errors of the percentiles. If we use a Taylor expansion for the square root function at the point σ_x^2 we obtain:

$$\hat{\sigma}_x \approx \sigma_x + \left(4\hat{\sigma}_x^2\right)^{-1/2}\left(\hat{\sigma}_x^2 - \sigma_x^2\right) \quad (55)$$

Using the above equation, the Taylor expansion about the point $\left(\mu_x, \sigma_x\right)$ and the fact that μ_x and σ_x^2 are independent we obtain:

$$Var\left\{\hat{F}(A)\right\} = \phi^2\left[\hat{\sigma}_x\left(A - \hat{\mu}_x\right)\right]\left\{\hat{\sigma}_x^{-2} Var\left\{\hat{\mu}_x\right\} + \left(\sigma_x^2\right)^{-3}\left(A - \hat{\mu}_x\right)^2 Var\left\{\hat{\sigma}_x^2\right\}\right\} \quad (56)$$

The estimated standard error of $\hat{F}(A)$ is the square root of the equation (56) above. $Var\left\{\hat{\mu}_x\right\}$ and $Var\left\{\hat{\sigma}_x^2\right\}$ are defined as:

$$Var\left\{\hat{\mu}_x\right\} = \frac{1}{n}\hat{\sigma}_x^2 + \frac{1}{nk}\hat{\sigma}_u^2 \quad (57)$$

$$Var\left\{\hat{\sigma}_x^2\right\} = k^{-2}\left\{2\left(\hat{\sigma}_u^2 + k\hat{\sigma}_x^2\right)^2\left(n-1\right)^{-1} - 2\hat{\sigma}_u^4\left[n\left(k-1\right)\right]^{-1}\right\} \quad (58)$$

Nusser et al (1996) use a Monte Carlo study to compare the usual intake distribution estimates obtained using the ISU method described above, a simplified best power method and 2-day means method. The authors document that the ISU method generally has smaller standard errors than the best power procedure, especially in the tails of the distribution. It is also uniformly superior to the best power method with respect to the MSE of all percentiles calculated in the study. The ISU and best power methods produce distribution estimates which are close to each other. The usual intake distribution estimated using the individual means method is comparable to the estimates obtained with the ISU and best power methods only for percentiles near the mean of distribution.

2.6 Wallace and Williams (2005)

The article validates the method developed by Wallace et al (1994) using two randomly selected single-day distributions for two exposure elements to predict long-term exposure to fine particles from short-term measurements. Although the application described in the Wallace et al (1994) article is different the method could also be used in the context of dietary intakes to estimate the distribution of usual intake.

The Wallace et al (1994) method requires at least two measurements on a group of individuals whose exposures to a target agent are drawn from a log-normal distribution. Similar to Hoffman et al (2002a) Wallace and Williams (2005) argue that provided that the measurements meet certain criteria, two measurements per person are sufficient to determine the parameters of the long-term exposure distribution. Wallace and Williams (2005) document that most of the distributions tested in their study failed the chi-squared test for log-normality. However, the Wallace et al (1994) method appears to be robust to deviations from the assumption of log-normally distributed data.

The method assumes a multiplicative model in which the measurement X is the product of a long-term component X_{LT} and a short-term component X_{ST}. The logarithms of the long-term population exposures are normally distributed:

$$\log X_{LT} \sim N\left(\mu, \sigma_{LT}^2\right) \tag{59}$$

The logarithms of the short-term fluctuations are also normally distributed around zero:

$$\log X_{ST} \sim N\left(0, \sigma_{\varepsilon}^2\right) \tag{60}$$

Hence, the geometric standard deviation of a distribution of exposures averaged over T sampling trips could be expressed as:

$$GSD_T = \exp\left\{\left[\sigma_{LT}^2 + \ln\left(1 + \left\{\exp\left[\sigma_{\varepsilon}^2\right] - 1\right\} / T\right)\right]^{0.5}\right\} \tag{61}$$

For the first sampling trip $\left(T = 1\right)$ the above equation reduces to:

$$GSD_1 = \exp\left\{\left[\sigma_{LT}^2 + \sigma_{\varepsilon}^2\right]^{0.5}\right\} \tag{62}$$

Hence, the GSD in the first sampling trip contains both the long and short-term variations. As the number of sampling trips increases $\left(T \to \infty\right)$ the short-term component disappears and the long-term GSD is given by the equation:

$$GSD_{LT} = \exp\left(\sigma_{LT}^2\right) \tag{63}$$

Therefore, the observed GSD declines with the increase in the number of sampling trips and approaches its long-term value. The equation for GSD_T is fitted by non-linear regression analysis to the observed geometric standard deviations and σ_{LT}^2 and σ_{ε}^2 are estimated.

The equation for the geometric mean of a distribution of exposures averaged over T sampling trips is defined as:

$$GM_{LT} = \exp\left\{\nu + \frac{\sigma_{\varepsilon}^2}{2} - \frac{1}{2} Ln\left[1 + \exp\left(\sigma_{\varepsilon}^2\right) - 1\right] / T\right\} \tag{64}$$

where ν is the natural logarithm of the geometric mean of the distribution observed on the first sampling trip $(T=1)$:

$$GM_1 = \exp(\nu) \tag{65}$$

The parameter ν could also be estimated from the equation for the arithmetic mean (67) given below once σ_{LT}^2 and σ_{ε}^2 are estimated using non-linear regression analysis.

For large T the above equation indicates that the geometric mean will approach long-term value of:

$$GM_{LT} = \exp\left(\nu + \frac{\sigma_{\varepsilon}^2}{2}\right) \tag{66}$$

Hence, the geometric mean increases with an increase in the number of sampling trips. However, the arithmetic mean is constant over all sampling trips T, numerous enough to account for seasonal and other periodic variations:

$$AM = \exp\left[\nu + \left(\sigma_{LT}^2 + \sigma_{\varepsilon}^2\right)/2\right] \tag{67}$$

The authors document that the precision of the method in estimating the long-term geometric mean and geometric standard deviation of sulfur measurements is in the order of 10%. Precision is defined as the estimated mean divided by the standard deviation. For the same element there was a negative bias of 1% for the geometric mean and 8% for the geometric standard deviation. Wallace and Williams (2005) define bias as (estimated mean-observed mean)/observed mean. The method underestimated the 99th percentile of the sulfur measurements by about 19%.

2.7 Gay (2000)

Gay (2000) proposes a statistical method to correct for the distortion of the population distribution of habitual nutrient intakes caused by day-to-day variation. The author documents that population distributions of habitual nutrient intake could be accurately estimated using 4-day weighed diary data and that the method proposed in the paper might be successfully applied using as little as 2 day weighed diary data. The method also corrects extreme nutrient intakes for the biases induced by an uneven representation of day of the week and within-person variance.

Gay (2000) argues that the effect of the within-person variation is twofold. It affects the precision of the estimate of the centre of the between-person intake distribution and pushes out the extremes of the distribution further away from the centre of the distribution. Furthermore, most of the nutrient intake distributions are positively skewed and hence the normality assumption rarely holds.

The statistical method proposed by Gay (2000) which could be used to characterise nutrient intake distributions consists of the following steps:

1) The mean intake and within-person standard deviation are estimated for each respondent and then log-transformed. The within-person standard deviation estimates are regressed on the mean estimates.

2) Subtract the estimated regression slope from unity to derive value of λ.

3) Transform the daily intakes for each nutrient using transformation of the form $Y = X^{\lambda}$ which stabilizes the within-person variance and often simultaneously achieves approximate normality.

4) On the transformed scale separate between and within-person components of the variance and estimate the size of the fixed effects of interest such as day of the week using ANOVA.

5) Back-transform the results to original units for presentation purposes.

6) Use the fitted model to simulate food records of sufficient duration as to eliminate any unwanted within-person variation. Hence, only between-person variation will be left in the simulated data. The percentiles of the intake distribution are derived from the simulated data.

The power transformation step of the model could only be applied to non-zero intakes. Hence, simulation of nutrients with zero intakes would require a more complex model and is not investigated further in the study. For the estimation of the components of variance in Step 4 the model assumes that the within-person variation is constant on consecutive days and across individuals.

Gay (2000) uses simulations in conjunction with non-parametric and parametric procedures to derive estimates of the extreme percentiles (2.5th and 97.5th) of the intake distribution. However the simulation procedures are poorly documented and hence difficult to replicate (Hoffmann et al, 2002a).

The author also suggests some refinements to the proposed model. If the within-person variance is not constant across individuals but varies in a predictable way this information could be incorporated in the simulation step of the method. Furthermore, additional fixed effects such as a seasonal component or adjustment factors for days of sickness could also be included in Step 4 of the proposed method.

2.8 Slob (1993)

The author proposes a method for modeling long-term exposure which consists of the following steps:

1) Transform the data to achieve approximate normality. Slob (1993) uses logarithmic transformations for the dietary intake data described in the article.

2) Perform regression analysis of log-intake on the independent variable of interest (age in the case of Slob (1993) paper) by fitting polynomial regression or any other function which fits the data.

3) Calculate the regression residuals.

4) Estimate the within-subject variance from the residuals using nested analysis of variance.

5) Subtract the within-subject variance from total variance to obtain between-subject variance.

Slob (1993) assumes the following model:

$$\log\left[Y_{ij}(t)\right] = f(t) + \varepsilon_i + \delta_{ij} \qquad (68)$$

The above model is formulated for a logarithmically transformed intake $\log(Y)$. $Y_{ij}(t)$ is the intake of individual i having age t on day j. $f(t)$ is the regression function fitted to the $\log(Y)$ with respect to age t. The value of $f(t)$ at age t is assumed to represent 'typical' or long-term average intake by 'typical' individual at age t. The 'typical' long-term intake of any given individual will deviate from $f(t)$ by an amount ε which is the between-subject effect. Furthermore, the transformed intake on any given day will deviate from an individual's typical intake by a certain amount δ which is the within-subject effect. Hence, in the above model ε_i is the deviation of the individual i log transformed intake compared to $f(t)$ and δ_{ij} is the deviation of individual i log-intake on day j compared to that person's typical log-intake. A nested analysis of variance is performed using the regression residuals.

The within-subject variance is estimated by the ANOVA within-subject MS. The between-subject variance is obtained after subtracting this estimate from the MS total estimate.

It is assumed that both variances are homogenous, have normal or at least symmetric distributions with zero means. It should be noted that back-transformation of means of symmetric distributions results in medians on the original scale. Therefore, the 'typical' individual is represented by the mean individual on the log-scale but by the median individual on the original scale.

To keep the model as simple as possible Slob (1993) relies on a number of assumptions. For example, he assumes that intakes do not vary systematically with the day of the week or week of the year. Furthermore, to estimate the percentiles of the usual intake distribution on the original scale Slob (1993) assumes that the distribution of the between-subject effect ε is normal and uses the following formula:

$$Q_{1-\alpha}(t) = \exp\left[f(t) + q_{1-\alpha}\sigma_{\varepsilon} \right] \qquad (69)$$

where $q_{1-\alpha}$ denotes the $(1-\alpha)$ percentile of the standard normal distribution and σ_{ε} is the square root of the between-subject variance.

Slob (1993) documents that the proposed model is suitable for both absolute and relative intakes since the ratio of two lognormally distributed variables is also lognormally distributed. In a subsequent paper Slob (1996) reviews two statistical models for estimation of inter-individual variation in long-term exposure. Those methods are the regression approach of Wallace et al (1994) and the standard ANOVA approach described in Slob (1993). Both models adopt a similar approach since they assume log-normal distribution of the intakes and that two levels of variation are distinguished - between and within-individual variances.

The simulation results documented by Slob (1996) indicate that the ANOVA method performs much better compared to the Wallace et al (1994) method in estimating the variation in short-term exposure within individuals. Both methods perform reasonably well in estimating variation in long-term exposure between individuals. Slob (1996) argues that the Wallace et al method (1994) needs more exposure observations per individual in order to derive stable estimates. In contrast, with the ANOVA method even if only two exposure observations per individual are available it will not lead to estimation problems. Furthermore, the ANOVA method appears to be a more flexible method of estimation since age, sex, seasonal and other effects could be incorporated in the model. Hence, Slob (1996) concludes that when repeated measurements data on the same individual are available the ANOVA method is preferable to the Wallace et al (1994) method. The ANOVA method is more flexible and more stable. However, Slob (1996) also notes that in situations where data originate from different studies, using different monitoring periods and different individuals the Wallace et al (1994) method is still applicable whereas the ANOVA method is not applicable.

2.9 Buck et al (1995)

In a study of exposure to environmental pollutants Buck et al (1995) propose a method for estimating long-term exposure from short-term measurements which could be described as follows:

The daily exposure values X_{ij} for the ith individual in the population sample on the jth day that the individual is sampled could be broken into additive components:

$$X_{ij} = \mu_i + \tau_{ij, i=1,.....,n,\ j=1,.....,m} \tag{70}$$

where μ_i is the true daily average exposure for the ith individual and τ_{ij} is the deviation of individual i from the mean μ_i on day j. Hence, μ_i is a random variable with distribution G, mean μ_p and variance σ_p^2. Therefore, G is the distribution of long-term average daily exposures in the population and μ_p is the long-term average exposure for the population. The error term τ_{ij} has a distribution F with a mean of 0 and variance σ_T^2. The distribution G and the mean and variance of μ_i are the parameters of interest and are used to define the level of exposure to pollutants in the environment.

Buck et al (1995) make the following key assumptions in order to develop the described exposure model:

I. There is no measurement error
II. Every individual has the same distribution of daily exposure F and variance σ_T^2
III. The random variables μ_i and τ_{ij} are independent
IV. The random variables τ_{ij} are independent across individuals
V. For the ith individual the random variables τ_{ij} are uncorrelated
VI. There is no trend in the model i.e. there is no systematic X_{ij} change in exposure over time

The distribution of X_{ij} could also be defined as follows:

$$X_{ij} = F * G\left(\mu_p, \sigma_p^2 + \sigma_T^2\right) \tag{71}$$

where $F * G$ is the distribution which arises when two independent random variables with distributions F and G are added together.

If we define $\hat{\mu}_i$ as the long-term estimated average exposure calculated from the repeated measurements for the individual i then its distribution could be defined as:

$$F_m * G\left(\mu_p, \sigma_p^2 + \sigma_T^2 / m\right) \tag{72}$$

where F_m is the distribution of the estimate $\hat{\mu}_i$ based on m estimates of the variables $\tau_{ij}, j = 1,.....,m$. Buck et al (1995) simplify the notation by letting $H = F_m * G$ and $\sigma_H^2 = \sigma_p^2 + \sigma_T^2 / m$.

Hence, the random variables X_{ij} and $\hat{\mu}_i$ are the variables observed when measuring daily and average daily exposure, respectively. From those two variables the distribution, mean and variance of the long-term exposure is estimated.

If we define $\hat{\mu}_p$ as the mean long-term exposure for the population, then it could be shown that:

$$\hat{\mu}_p = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} X_{ij} / m \tag{73}$$

and $\hat{\mu}_p$ is an unbiased estimator of μ_p. The variance of $\hat{\mu}_p$ is σ_H^2 / n. It should be noted that σ_H^2 / n is a biased estimate for σ_p^2 because $\sigma_H^2 = \sigma_p^2 + \sigma_T^2 / m$. However, if we define:

$$\hat{\sigma}_p^2 = \hat{\sigma}_H^2 - \hat{\sigma}_T^2 / m \tag{74}$$

then $\hat{\sigma}_p^2$ is an unbiased estimate of σ_p^2. The authors note that the above estimate is dependent on getting an estimate of σ_T^2. Hence, unless there is some information from other sources such an estimate could only be obtained from repeated measurements of individuals in the sample.

Buck et al (1995) document that in addition to means and variances researchers often are interested in characterizing the upper tails of exposure distributions. For example, in order to investigate high-end exposure estimates of the 95th and 98th percentiles of the distribution need to be obtained. The estimation of the percentiles depends on whether the distributions F and G are known. If the distributions F and G are known, percentiles could be estimated as long as means, variances and other essential parameters could be estimated from the data. However, typically the distributions are not known and nonparametric techniques can be used to estimate the percentiles of the long-term exposure distribution.

In order to infer the long-term exposure distribution from the data, the sample must be regarded as having been drawn from the distribution H which is the distribution of long-term average exposure estimated from a limited number of repeated observations per individual. The sample should not be regarded as having been drawn from the distribution G which is the distribution of true daily average exposure. A sample from the distribution G would be possible only if the temporal distributions for each individual are completely known. This would be possible only if the collected data contains 365 days per year. The distribution H contains temporal contribution to its variance $\left(\sigma_T^2 / m\right)$ which is decreased but not eliminated by increasing the number of repeated observations m per individual.

If we define q_a as the quantile for the 100α percentile of the distribution G then:

$$\hat{q}_a = k_a \sigma_p / \sigma_H \tag{75}$$

where k_a is the estimated quantile using ordered values of the estimates $\hat{\mu}_i$ from the distribution H.

Buck et al (1995) document the following properties of the two estimates $\hat{q}_a$ and k_a:

1) Both estimates are biased estimates of q_a in most situations but $\hat{q}_a$ is less biased estimate of q_a than k_a. Similar results hold for the estimation of the percentiles.

2) The estimate of $\hat{q}_a$ depends only on having estimates of σ_p^2 and σ_T^2. Hence, the number of repeated measurements per individuals does not need to be large. However, the bias in both $\hat{q}_a$ and k_a decreases as the number of repeated measurements per individual m increases. As σ_T^2 increases the bias in the estimate k_a increases faster than the bias in the estimate $\hat{q}_a$.

3) If F and G are symmetric or have similar skewness and/or kurtosis then $\hat{q}_a$ will be nearly unbiased for all values of α. This result underlines the limitations in effectively estimating the median of the population $(\alpha = 0.5)$.

4) If F and G both have densities which go smoothly to zero in the upper tails of the distribution then $\hat{q}_a$ will be nearly an unbiased estimate when α is an upper percentile.

The results documented in the last three points above underline the robustness of $\hat{q}_a$ as an estimate of q_a. It should be noted that the robustness property of $\hat{q}_a$ depends on the nature of the distributions F and G. Furthermore, the above results are based on the shape of the F and G distributions as determined by their skewness and kurtosis and not necessarily on the family of the distributions.

Buck et al (1995) conclude that the estimates of σ_p^2 and σ_T^2 are needed to minimize the bias in the estimates of percentiles and quantiles. Hence, a minimal number $m = 2$ of repeated measurements per individual will be sufficient if the distribution F and its variance σ_T^2 are the same for all individuals in the sample.

2.10 Slater et al (2004)

The authors propose a methodology for estimating the proportion of individuals with inadequate nutrient intake. The methodology is based on the EAR (Estimated Average Requirement) reference standard and the NRC (1986) method. EAR is an appropriate reference estimate for assessing inadequate nutrient intake and is defined as the nutrient intake that corresponds to the estimated average need for a given stage of life and gender.The authors argue that the proposed methodology minimizes the error in the calculation of the prevalence of inadequate nutrient intake since it takes into consideration the random characteristics of the diet.

Slater et al (2004) define two types of diets in their research: habitual diet and present diet. The habitual diet is defined as average food consumption over a long period of time over which a dietary pattern is maintained. Present diet is defined as average food consumption over a short period of time around the present day.

According to the authors, the habitual diet could be explained by the following model:

$$Y = \mu + individual_i + \varepsilon \quad (76)$$

where Y is the nutrient intake, μ is the true average intake, $individual_i$ is the effect of variation between individuals and ε is the error term.

According to the above model, the habitual intake of an individual consists of true average μ influenced by the effect of each individual in the population. This effect is measured by the variance between the individuals denoted by S_b^2. However, since this function can not be measured directly the true average intake can not be measured free of error. The variations in the habitual diet as a result of day-to-day fluctuations in the food intake could be measured by the variance within individuals S_w^2 which represents an individual's variability around his own average evaluated through multiple observations. The third type of variation ε is the measurement error of the instrument used. Hence, it could be defined as the difference between the observed and true intake.

The proposed methodology removes the day-to-day variability due to variation in within-person consumption. Hence, the resulting distribution reflects only the variation which exists between the individuals in a group. In order to be able to perform adjustments to the distribution of dietary intakes at least two independent measurements are needed from at least one representative sample of individuals on non-consecutive days. Slater et al (2004) document that measurement error is intrinsic to any method for assessing dietary intakes and there are no methods for measuring dietary intakes which are free of measurement error.

After transforming the habitual intake data to normality using standard data transformations Slater et al (2004) use ANOVA to estimate within-person S_w^2 and between-person S_b^2 variance components. The following formulae follow the methodology of the NCR (1986) method to perform the components of the variance analysis:

$$RMS_w = S_w^2 \quad (77)$$

$$RMS_b = S_w^2 + kS_b^2 \quad (78)$$

$$S_b^2 = \left(RMS_b - S_w^2\right)/k \quad (79)$$

The total variance of an observed distribution is given by the sum of the within and between-person variances:

$$S_{obs}^2 = S_w^2 + \left(S_b^2\right)/k \quad (80)$$

Hence:

$$S_{obs}^2 / S_b^2 = \left(S_b^2 + \left(S_w^2/k\right)\right)/S_b^2 = \left(1 + S_w^2/kS_b^2\right) \quad (81)$$

After taking the square root we get:

$$S_{obs} / S_b = \left(1 + S_w^2 / kS_b^2\right)^{\frac{1}{2}} \quad (82)$$

To remove the within-person variance the reciprocal of the above equation is used in the equation proposed by the US National Academy of Science Subcommittee on Criteria for Dietary Evaluation (NCR, 1986):

$$X_{adj} = average + \left(x_i - average\right) * \frac{S_b}{S_{obs}} \quad (83)$$

where X_{adj} is the adjusted value for the nutrient, average is the average value for the group and x_i is the value observed for each individual.

The average is not affected by the within-person variance S_w^2 and hence the adjusted distribution and the unadjusted distribution have the same average although the dispersion of the adjusted distribution is expected to be less.

The proportion of people with inadequate consumption is the area under the curve calculated using the formula:

$$Z = \frac{\left(EAR - average\right)}{SD} \quad (84)$$

where EAR is the estimated average dietary requirement, $average$ is the adjusted average for the group and SD is the standard deviation of the adjusted distribution.

It should be emphasized that the above calculations are valid for adjusted distributions which are normal or at least symmetric. The corrections are appropriate for estimates concerning groups of individuals and they can not be used for identification of individuals with dietary consumption below certain threshold.

Slater et al (2004) also derive the expectations of the components of variance which are summarized in Table 7 below:

Table 7: Slater et al (2004). Expectations of the Components of Variance

Source	Degrees of Freedom	Root Mean Square (RMS)	Expected Root Mean Square (ERMS)
Between-person	$n-1$	RMS_b	$S_w^2 + kS_b^2$
Within-person	$n\left(k-1\right)$	RMS_w	S_w^2

The authors did not discuss the back-transformation of the transformed towards normality nutrient intakes to the original scale. As pointed out by Dodd et al (2006) the back-transformation should contain a bias adjustment term.

2.11 Chang et al (2001)

Chang et al (2001) propose the application of the overdispersed exponential family of distributions to estimate the distribution of usual nutrient intake. The authors apply the adjustments developed by Liu et al (1978) to adjust the variance of the intake distribution. The proposed method has the advantage of working on the original scale and is easy to implement. The adjustment of the variance is implemented by dividing the variance into within-individual and between-individual components. The adjusted variance is used to estimate the distribution of the usual daily intakes. The method could be used for surveys with complex sampling designs. Chang et al (2001) incorporate sampling weights in all computations except in the estimation of the ratio of within to between-individual variance.

The authors argue that the method proposed by Nusser et al (1996) has two major shortcomings. First, too many transformations which depend on the specific data values are used. A slight change in data point might result in different distributions. Second, the Nusser et al (1996) method could produce negative between-individual variance after transformation. Chang et al (2001) report that after the application of the Nusser et al (1996) method the within-individual variance of some of the nutrients in their research data became larger than the total variance. Hence, if the within-individual variance is subtracted from the total variance the between-individual variance becomes negative.

Chang et al (2001) propose a method which is based on the overdispersed exponential family of distributions and combines some of the procedures developed by Nusser et al (1996) to estimate the distribution of usual daily intake. The major advantages of the method could be summarized as follows:

I. Working on the original scale

II. Taking the large intra-individual variation into account

III. Ease of implementation

However, one of the drawbacks of the method proposed by Chang et al (2001) is the need for an external set of data which is used to estimate the ratio of within to between individual variance. The estimated ratio has to be representative for the population of interest. In their research Chang et al (2001) use a dataset consisting of observations on 69 individuals aged between 18-29 years. The data in the external dataset are collected from nutrition major students using dietary record method. The students are asked specifically not to change their usual intake patterns. The collected data take into account the variation due to day of the week and season. However, Chang et al (2001) do not provide information if the estimated within and between-individual variances represent accurately the population estimates. Estimates computed using external data sets might differ from the true parameters in the population study leading to bias when those external estimates are used in the primary study to estimate the usual intake distribution. Hence, further research is needed to investigate in detail if estimates obtained from external samples could be used to adjust nutrient intake distributions. Jahns et al (2004) have published some research on the topic. In their article the authors also note that one drawback of the Chang et al (2001) research is that the issue of possible impact of using external variance estimates to adjust usual intake distribution is not addressed.

According to Chang et al (2001) since large within-individual variation exists in the usual daily intakes their distributions are overdispersed. Two major approaches are used to accommodate the extra variability in the exponential family. The first approach is to use a mixture of exponential family densities as sampling densities. The second approach is to use an additional scale parameter in the exponential family form. The authors prefer the second approach since it is known that the total variability consists of within and between-individual variances and their ratio could be estimated.

Chang et al (2001) argue that if we wish to estimate the distribution of nutrient intake $g_{\mu,n}(y)$ an ordinary exponential distribution of the form:

$$g_{\mu,n}(y) = \exp\left\{n\left[y\eta - \psi(\mu)\right]\right\}\left[dG_n(y)\right] \tag{85}$$

could be used. In the above equation μ is the expectation of the nutrient Y, y is the observed value, η is the natural parameter, n is the sample size and $\psi(\mu)$ is the normalising function which makes the density integrate to 1. The mean and variance of the distribution are given by

$E(Y_i) = \psi'(\mu_i)$ and $\text{var}(Y_i) = \psi''(\mu_i)/n_i \equiv \text{var}(\mu_i)/n_i$, respectively.

However, when the distribution of a nutrient has extra variation it becomes an overdispersed distribution. A simple equation such as equation (85) above fails to accurately describe such overdispersed distribution.

Chang et al (2001) propose to use double exponential functions documented by Efron (1986) to describe an overdispersed distribution. Given equation (85), the double exponential density could be defined as:

$$\tilde{f}_{\mu,\phi,n}(y) = c(\mu,\phi,n)\phi^{1/2}\left\{g_{\mu,n}(y)\right\}^{\phi}\left\{g_{y,n}(y)\right\}^{1-\phi}\left[dG_n(y)\right] \tag{86}$$

where the constant $c(\mu,\phi,n)$ is defined to make the density integrate to 1 and ϕ is the dispersion parameter. Efron (1986) describes many properties of the above distribution. The most useful for the estimation of usual intake distribution are the following:

a) the constant $c(\mu,\phi,n)$ is nearly equal to 1
b) the density $\tilde{f}_{\mu,\phi,n}(y)$ describes approximately the same probability distribution as equation (85) with n changed to $n\phi$

Hence, we could write that:

$$\int_A \tilde{f}_{\mu,\phi,n}(y) \approx \int_A g_{\mu,n\phi}(y)\, dG_{n\phi}(y) \tag{87}$$

The mean and variance of the overdispersed exponential family are approximately μ and $\dfrac{V(\mu)}{(n\phi)}$.

Double exponential family distributions and their properties have attracted the attention of numerous researchers. Expressions similar to the equations (86) and (87) above have been derived via a variety of different approaches. Some of those include the work of Albert and Pepple (1989) who used a quasi-likelihood approach with a new scale parameter added to the exponential likelihood to derive an expression similar to equation (87). Commenges and Jacqmin-Gadda (1997) who investigated a correlated random effects model to obtain score statistics for testing heterogeneity also derived an expression similar to equation (87) for exponential families. Other relevant research on the topic of overdispersion includes the work of West (1985), Ganio and Schafer (1992) and Dean (1992).

Similar to other researchers Chang et al (2001) define the model for the observed usual intake as:

$$Y_{ij} = y_i + u_{ij} \tag{88}$$

where y_i is the true unobservable usual intake value for individual i and u_{ij} is the departure from the true mean for individual i on the day j. It is assumed that the u_{ij} are independent given i. If we denote with $\overline{Y}_i$ the average intake of j days for individual i then the distribution of the observed usual intake has a mean $\overline{Y}$ and variance $\sigma^2_{\overline{Y}} = \sigma^2_y + \frac{1}{k}\sigma^2_u$ where σ^2_y is the between-individual variance, σ^2_u is the within-individual variance and k is the number of replicates per individual. There are two types of errors in the dietary survey data- day to day variation in the amount eaten by an individual and differences between the amount eaten and the amount recorded for the individual. Based on results from previous research published by Willett (1990), Chang et al (2001) assume that the day-to-day within individual variation is much larger that the reporting error.

The authors re-write the relationship $\sigma^2_{\overline{Y}} = \sigma^2_y + \sigma^2_u$ as:

$$\hat{\sigma}^2_{\overline{Y}} = \hat{\sigma}^2_y (1+p) \tag{89}$$

where $p = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_y}$ is the ratio of within to between-individual variances. Hence, the dispersion parameter ϕ equals to:

$$\phi = (1+p) \tag{90}$$

Liu et al (1978) use the ratio $p = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_y}$ of within to between-individual variances to estimate the correlation between dietary factor and risk. If any of the variables has within-individual variance the estimated correlation coefficient $\hat{r}$ must be multiplied by an error term $\sqrt{\left(\frac{1}{1+p}\right)}$ to estimate the true correlation.

Similarly, Chang et al (2001) use the ratio $p = \frac{\hat{\sigma}^2_u}{\hat{\sigma}^2_y}$ of within to between-individual variances to divide the observed variance by $(1+p)$ to obtain the inter-individual variance. The authors construct the distribution of the usual intake using the original mean and reduced variance.

The method proposed by Chang et al (2001) could be summarized with the following steps:

1) The parameters of the distribution based on the observed data are estimated.

2) The ratio of within to between individual variance is estimated using an external dataset.

3) The estimated ratio is applied to adjust the variance of the distribution.

4) The adjusted distribution of usual daily intake is constructed and the parameters of the distribution are estimated.

The method is applied separately to subpopulations with distinct characteristics i.e males and females. The authors argue that since most of the distributions of usual intake are skewed to the right gamma distribution could be fit to the nutrient intake data. The gamma distribution is defined as:

$$g_{\mu,v}(y) = \frac{1}{\Gamma(v)}\left(\frac{v}{\mu}\right)^{v} \exp\left(-\frac{y^{v}}{\mu}\right) y^{v-1}, 0 < y < \infty \tag{91}$$

where v is a scale parameter and

$$Var(Y) = \frac{\mu^{2}}{v} \tag{92}$$

When applying the expression for an overdispersed exponential family to the gamma family, the overdispersed gamma distribution has the same density function as in (91) with v replaced by $v\phi$, where ϕ is the dispersion parameter. The overdispersed gamma distribution has mean μ and variance:

$$Var(Y) = \frac{\mathrm{var}(\mu)}{(v\phi)} \tag{93}$$

Hence, the parameters obtained from the observed data are used to obtain the parameters of the adjusted distribution by dividing the total variance by the dispersion parameter $\hat{\phi} = (1 + \hat{p})$ where $\hat{p}$ is the estimated ratio of within to between individual variance. The ratio is estimated using the following equations:

$$\hat{\sigma}_{u}^{2} = \left[n(k-1)\right]^{-1} \sum_{i=1}^{n}\sum_{j=1}^{k}\left(Y_{ij} - \overline{Y}_{i}\right)^{2} \tag{94}$$

where $\hat{\sigma}_{u}^{2}$ is the within-individual variance, Y_{ij} is the amount of nutrient taken by individual i on day j and $\overline{Y}_{i}$ is the average nutrient taken by individual i between day 1 and day k calculated using the formula:

$$\overline{Y}_{i} = \frac{1}{k}\sum_{j=1}^{k} Y_{ij} \tag{95}$$

The between-individual variance is calculated using the well-known formula $\hat{\sigma}_y^2 = \hat{\sigma}_{\bar{Y}}^2 - \frac{1}{k}\hat{\sigma}_u^2$ where

$$\hat{\sigma}_{\bar{Y}}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\bar{Y}_i - \hat{\mu}_y\right)^2 \tag{96}$$

and

$$\hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i \tag{97}$$

Therefore, the parameters of the new adjusted overdispersed gamma distribution could be written as:

$$E\left(Y\right) = \mu \tag{98}$$

and

$$\mathrm{var}\left(Y\right) = \frac{\mathrm{var}\left(\mu\right)}{v\left(1+p\right)} \tag{99}$$

2.12 Carroll et al (1995, 2006)

The second edition of Carroll et al's (1995) excellent book 'Measurement Error in Nonlinear Models' was published in 2006. The 2006 edition includes some new material as a result of the active research in the area of measurement error models and hence is more up-to-date with the modern developments in the field. The review in this section is based on the second edition of the book.

According to the authors the measurement error in covariates has three major effects:

I. It causes bias in parameter estimation of statistical models

II. It leads to a loss of power, sometimes considerable, for detecting relationships among variables

III. Makes graphical model analysis difficult since it masks features of the data

Most of the statistical methods investigated in the book are concerned with solutions to the first problem i.e. finding different ways to correct for bias in the estimated parameters caused by measurement error.

2.12.1 Types of Measurement Error Models

Carroll et al (2006) discuss two types of measurement error models - classical and Berkson (1950) measurement error models. The classical measurement error model assumes that truth is measured with an additive error, in most cases with constant variance. Hence, the classical measurement error model could be written as:

$$W_i = X_i + U_i \tag{100}$$

Where W_i is an unbiased measure of X_i, therefore U_i must have mean zero:

$$E\left(U_i \mid X_i\right) = 0 \tag{101}$$

The error U_i could be homoscedastic or heteroscedastic but most often it is assumed to be homoscedastic. Thus, we could write the distribution of U_i as:

$$U_i \mid X_i \sim Normal\left(0, \sigma_u^2\right) \tag{102}$$

From the above equations it could be concluded that the variability of the observed measurements is greater that the variability of the true measurements. In contrast, the Berkson (1950) measurement error model assumes that the true measurement is equal to the observed measurement plus measurement error. Hence, we could write the Berkson (1950) model as:

$$X_i = W_i + U_i \tag{103}$$

and

$$E\left(U_i \mid W_i\right) = 0 \tag{104}$$

2.12.2 Extensions of the Classical Measurement Error Model

Carroll et al (2006) argue that it is crucial to understand the differences between the two error models since an incorrect model would result in erroneous inferences. The effect of the type of measurement error is particularly important in the computing of statistical power. For example, wrongly assuming a Berkson (1950) error model when the measurement error is classical would lead to gross overstatement of the variance of the true measurement and hence grossly optimistic overstatement of the statistical power of detecting important effects.

The authors document that the measurement error models could be more complex than the classical additive measurement error model specified in equation (100) or the classical Berkson (1950) error model specified in equation (103) above.

With regards to data collected by food frequency questionnaires (FFQs) Carroll et al (2006) argue that the FFQ data are not very good measures of daily intakes since usually there is bias in the data. A more reasonable measurement error model is the one proposed by Kipnis et al (1999, 2001, 2003) which allows for bias as well as variance components:

$$W_{ij} = \gamma_0 + \gamma_1 X_{ij} + U_{ij} \tag{105}$$

$$U_{ij} = r_i + \varepsilon_{ij} \tag{106}$$

where $r_i \sim Normal\left(0, \sigma_r^2\right)$ and $\varepsilon_{ij} \sim Normal\left(0, \sigma_\varepsilon^2\right)$. The structure of the measurement error random variables U_{ij} is with components – a shared component r and a random component ε .

Kipnis et al (1999, 2001, 2003) call the shared component person-specific bias which reflects the idea that two people who eat exactly the same food will report intakes differently when given multiple FFQs. And if $\gamma_0 = 0, \gamma_1 = 1, r \equiv 0$, then we have the standard error measurement model. Carroll et al (2006) also argue that for studies which use data collected with instruments based on self-reports more complex models incorporating biases are required. They document that for such cases the general classical error model framework is appropriate:

$$W = \gamma_0 + \gamma_x^t X + \gamma_z^t Z + U, \quad E\left(U \mid X, Z\right) = 0 \tag{107}$$

where Z is an observed predictor and X is an unobserved predictor. It could be argued that the relationship between W and X also depends on the observed predictor Z .

As mentioned earlier, Carroll et al (2006) stipulate that a fundamental prerequisite for the analysis of a measurement error problem is the specification of a model for the measurement error process.

According to the authors there are two major types:

I. Error models, including the classical measurement error models where the conditional distribution of W given $\left(X, Z\right)$ is modeled. Here X denotes covariates measured with error and Z covariates measured without error.

II. Regression calibration models, including Berkson (1950) error models where conditional distribution of X given $\left(Z, W\right)$ is modeled.

2.12.3 Linear Regression Attenuation, Calibration and SIMEX method

Carroll et al (2006) document that the ordinary least squares estimator is typically biased as a result of measurement error and the direction and magnitude of the bias depends on the regression model, measurement error distribution and the correlation between the true predictor variables. Hence, the measurement error distribution determines the effects of the measurement error on the regression estimates. These effects can range from attenuation of the slope estimate in the direction of zero, hidden real effects, reversed signs of regression coefficients and relationships which are not present in the error-free data.

If $W = X + U$ and $U \sim Normal\left(0, \sigma_u^2\right)$ then ordinary least squares regression of Y on error-prone predictor W produces an estimator that is attenuated toward zero with an attenuating factor (bias):

$$\lambda = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\mu^2} < 1 \tag{108}$$

which implies that the estimated slope is:

$$\hat{\lambda}\hat{\beta}_x = \beta_x \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\mu^2}, 0 < \lambda \leq 1 \tag{109}$$

The residual variance of the regression could be written as:

$$\text{var}\left(Y \mid W\right) = \sigma_\varepsilon^2 + \left(1 - \lambda\right)\beta_x^2 \sigma_x^2 \tag{110}$$

If the reliability ratio λ were known then we could obtain an unbiased estimate of β_x by dividing the estimated OLS slope coefficient $\hat{\beta}_{x^*}$ by the reliability ratio. However, without information about σ_u^2 we can not estimate β_x. The most efficient way to get information about σ_u^2 is to observe X on a subset of data. The next best way is via replication. For example, if we have two or more replicates $W_1 = X + U_1$ and $W_2 = X + U_2$ then σ_μ^2 could be estimated via components of variance analysis. The third and least efficient method is to use instrumental variables which are related to X and independent of U. In other words, they are surrogates for X.

The differences between the ordinary least squares estimators and those under the measurement error model are summarized in Table 8 below:

Table 8: Carroll et al (2006). Comparison between Ordinary OLS and Measurement Error Model Estimators

Regression	Intercept	Slope	Residual Variance
OLS	β_0	β_x	σ_ε^2
Measurement Error Model Estimators	$\beta_0 + (1-\lambda)\beta_x\mu_x$	$\lambda\beta_x$	$\sigma_\varepsilon^2 + (1-\lambda)\beta_x^2\sigma_x^2$

The table above illustrates that the measurement error not only causes an attenuation of the regression slope but also the data are noisier with an increased variance around the regression line.

In multiple regression where some of the covariates are measured with and some without error the coefficient of the error-prone covariate W which is unbiased for X (the unobserved true predictor) has an attenuating factor:

$$\lambda_1 = \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2} \tag{111}$$

where $\sigma_{x|z}^2$ is the residual variance of the regression of X on Z. It should be noted that the coefficient of Z is also biased unless Z is independent of X.

There are various ways to correct for bias under measurement error. Fuller (1987) proposed a method-of-moments method. Another popular method is regression calibration. The basic idea of regression calibration is to replace X by the regression of X on (Z,W). Hence, X is replaced by an estimate $E(X \mid Z,W)$ which depends only on the known (Z,W). The method works well for generalized linear models but is often not appropriate for highly nonlinear problems. The regression calibration algorithm could be summarized with the following three steps:

I. Develop a model for the regression of X on (Z,W)

II. Replace X and perform the analysis under consideration

III. Obtain the standard errors by the bootstrap or 'sandwich' methods

Carroll et al (2006) discuss various techniques which could be used to estimate the regression calibration function. Some of those include the use of internal validation data, the method of Rosner et al (1990) as well as the use of replication data.

The simulation extrapolation method (SIMEX) is ideally suited to problems with additive measurement error where the measurement error is not too large in order for the approximation to be sufficiently accurate. The method is restricted to the classical measurement error model. No assumptions are made about the true X values. The basic idea of SIMEX is to add additional measurement error to the measured data in a resampling-like stage and recalculate the statistic. Hence, it establishes a trend of measurement error-induced bias versus the variance of the measurement error. The trend is extrapolated back to the case of no measurement error. The SIMEX method could be summarized with the following four steps:

I. Simulation step: add additional measurement error to the variable measured with error.

If θ controls the amount of additional measurement error then σ_u^2 is increased to $(1+\theta)\sigma_u^2$. Hence, in each simulation step data sets with successively larger measurement error variances are generated.

II. Estimation step: recalculate the regression estimate. The recalculated estimate is called pseudo estimate. The simulation and estimation steps are repeated a large number of times. The average values of the pseudo estimates for each level of θ are calculated.

III. Plot the average values of the pseudo estimates versus each level of θ. A regression method is used to fit an extrapolant function to the average values of the pseudo regression estimates.

IV. Extrapolate θ back to $\theta = -1$ which corresponds to the case of no measurement error.

A regression where the dependent variable is the average pseudo estimate $\hat{\beta}_{x,m}$ calculated during each estimation step m and the independent variable is θ_m will asymptotically have the following mean function:

$$E\left(\hat{\beta}_{x,m} \mid \theta_m\right) = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + (1+\theta)\sigma_\mu^2}, \theta \geq 0 \qquad (112)$$

If we extrapolate back to $\theta = -1$ the above equation will be equal to β_x. There are various extrapolation methods used in the research literature such as linear, quadratic and rational linear.

Carroll et al (2006) document that when the choice is between classical and Berkson (1950) error models then the classical model should be chosen if the error-prone covariate is measured uniquely to an individual and most importantly if the measurement can be replicated. In the classical error model the error U is independent of W or at least $E(U \mid X) = 0$. Furthermore, $Var(W) > Var(X)$ for classical measurement error model whereas $Var(X) > Var(W)$ for Berkson (1950) measurement error model. All of the methods investigated in this research are based on the classical measurement error model framework.

3. Methods

Chapter 3 documents details about the data used in the research, the initial data preparation steps, preliminary analyses as well as sampling methodology. Since the majority of the methods investigated in Chapters 4 and 5 are based on an ANOVA type variance decomposition it is important to investigate if the assumptions behind the univariate ANOVA analysis are fulfilled in order to obtain results which are valid from a statistical point of view.

The sampling methodology followed the recommendations given by Hoffmann et al (2002a). According to the authors the best sampling design is based on a simple random selection of the sampling days. Furthermore, Hoffmann et al (2002a) document that the randomly selected days should be representative of the period of interest. Hence, it is important that the sampling days cover all seasons and all days of the week. Hoffman et al (2002a) also noted that non-consecutive sampling days in different seasons should be preferred. A sampling design based on adjacent days can result in over-estimation of the population variance of usual intake and under-estimation of the intra-individual variance if dietary intake varies considerably across seasons. For the above reasons it was decided to use month of pregnancy to sample randomly the repeated measurements.

The analysis in Chapter 5 is based on a sample of two repeated measurements per individual. For each individual in the sample the repeated measurements were sampled from the 4^{th} and 7^{th} month of pregnancy. The month of pregnancy indicator guarantees that the two repeated measurements are at least 3 months apart and hence they were not taken on consecutive days of the same week. It is assumed that such sampling design would reduce the correlation between the repeated measurements.

Sampling designs which attempt to reduce the influence of the nuisance effects as well as the correlation between the repeated measurements allow the application of simpler and more robust statistical methods suitable for small to medium sized samples. The approach of including covariates for nuisance effects in the models would require the use of more complex procedures suitable for larger sized samples. Furthermore, the preliminary data analysis indicated that the day-of-the-week effect is not that pronounced for nutrients such as protein and carbohydrates which are consumed on a daily basis. As illustrated in Tables 25-27 when day-of-the-week adjustments are included in the models then the estimates of the percentiles of the usual intake distribution are slightly higher compared to the estimates obtained without including adjustments for nuisance effects in the models.

3.1 The Data

The data used in the research were supplied by Massey University. It consisted of 8 days of weighed diet records collected during 4^{th} and 7^{th} month of pregnancy on a total of 197 healthy, mostly European pregnant women. The patients attended a city anti-natal clinic in the lower North Island of New Zealand. The individuals in the sample were volunteers. There were no sampling weights included in the data. More details about the data are given in Watson and McDonald (2007).

The required 8-day weighed record was split into two 4-day periods which were eight days apart in order to improve subject cooperation (Bingham, 1987) and ensure that each day of the week was represented. The consumed food details and weights were recorded by the participants in the study in 4-day diet record books. All food and drink consumed was weighed on calibrated Salter Microtonic Electronic Scale with taring facility. The nutrient intake for each woman was calculated using the New Zealand Food Composition database FOODfiles maintained by New Zealand Institute for Crop and Food Research.

The collected data contained measurements on 42 nutrients which could be grouped into three major categories such as macronutrients, vitamins as well as minerals and trace elements. The data contained also information on alcohol consumption and various flags identifying the sequence of the weighed record and time of measurement.

3.2 Data Preparation

The coding, data entering and double error checking of the data used in this research have already been performed as part of the research documented by Watson and McDonald (2007). However, standard data exploratory techniques such as frequencies, descriptive statistics and graphs were applied to check the integrity of the data.

The available data were examined for missing and erroneous values. Individuals with missing dietary intake data were removed from the dataset. Although there is a formula (26) which adjusts for unequal number of repeats per individual only individuals with an equal number of repeats were left in the sample in order to simplify the analysis. Hence, individuals with less than 8 repeats per observation month (4^{th} and 7^{th} month of pregnancy) were removed from the dataset. The clean master dataset contained dietary data on 151 pregnant women with 16 repeats per individual. Some of the nutrient intakes contained zero values. Vitamin D intake contained the highest percentage of zero values which was around 12.6%. Zero intake was assumed to be genuine non-consumption. Some of the procedures described in the subsequent sections add a small fraction to the zero observations in order to avoid instability of the numerical results.

The available sample for analysis is rather small since it contains nutrient intake data on only 151 individuals. Moreover, it does not contain any covariate information which could be used to model the usual intake distribution. Hence, the analysis in this research is concerned primarily with the estimation of usual intake distribution of nutrients from small sample sizes without the use of covariate information.

3.2.1 Sampling

A number of researchers such as Beaton et al (1979), Karkeck (1987), Nelson et al (1989) as well as Willett (1998) have investigated the optimal number of repeated measurements per individual recommended to be collected in dietary surveys. The documented results indicate large variability in the optimal number of consumption days based on the type, size or budget of the survey, purpose of the analysis as well as the foods or nutrients analysed. The discussion which follows is based on the the overview of the existing academic literature provided by Volatier et al (2002).

Beaton et al (1979) provide a formula which could be used to estimate the number of repeated measurements needed to estimate individual long-term intake at the 95% confidence level:

$$k = \left(1.96 CV_w / r\right)^2 \tag{113}$$

where k is the repeated measurements needed per person, CV_w is the within-person coefficient of variation and r is the relative error accepted.

Beaton et al (1979) also note that cost considerations could be included in the design of large scale consumption surveys. The authors proposed a formula for the estimation of the optimum number of repeated interviews per individual which takes into account the costs associated with consumption surveys:

$$k = R\sqrt{\frac{C_1}{C_2}} \tag{114}$$

where k is the optimum number of repeated measurements per individual, R is the ratio of within to between-individual variance, C_1 is the cost of including an additional individual in the survey process and C_2 is the cost of conducting and analyzing a single dietary interview for an individual included in the sample.

The required number of repeated measurements per individual needed to classify 80% of the population into tertiles of dietary intake documented by Karkeck (1987) are summarised in the table below:

Table 9: Karkeck (1987). Number of Repeated Measurements Required

Nutrient	Number of Repeated Measurements
Carbohydrates	2-4
Calcium	3-5
Energy	3-7
Protein	5-7
Total Fat	5-9
Fibre	5-10
Iron	12-19

As expected, for regularly consumed foods such as carbohydrates and calcium a smaller number of repeated measurements per individual are required compared to the number of repeated measurements needed for occasionally consumed foods and nutrients such as fibre or iron.

Volatier et al (2002) comment that the number of individuals and the number of repeated measurements per individual included in the sample should be chosen so that the confidence intervals of the estimated long-term intakes distribution parameters are as small as possible. Willet (1989) documents that the lengths of the confidence intervals are not systematically affected by the number of repeated measurements per individual but decrease as the number of individuals included in the sample is increased. Hence, Volatier et al (2002) suggest that the best number of repeated measurements per individual is two but the number of individuals included in the sample should be maximized. Gay (2000) also notes that the lengths of the 95% confidence intervals of the estimated percentiles of the usual intake distribution did not differ significantly for 2, 4 and 7 repeated measurements per individual. Hoffmann et al (2002a) document that two repeated measurements per individual are sufficient to remove the effects of the within-person variation in the dietary intake data and estimate usual intake distribution.

Detailed discussion about the sample size estimation principles could also be found in the Volatier et al (2002) article. According to the authors, the number of individuals included in the sample depends on the type of the survey (national, international), allocation and sampling methodologies, list of parameters of interest, desired precision, etc.

The data used in the research could be used to investigate if the number of repeated measurements per individual has significant influence on the lengths of the confidence intervals of the estimated percentiles of the usual intake distribution. For this purpose random samples of two, four, six and eight repeated measurements per individual are generated and used in the analyses documented in Chapter 4. Following the recommendations found in the existing academic literature the comparison of the methods for estimating usual intake distribution shown in Chapter 5 is based on a sample of two repeats per individual.

The following general guidelines described by Hoffmann et al (2002a) were followed when the samples of different number of repeats per individual were selected:

1) The optimal sampling is based on a simple random selection of the sampling days.

2) If the sample size is not too small, the randomly selected days should be representative of the period of interest.

3) The sampling size should cover all seasons and all days of the week.

4) Non-consecutive sampling days in different seasons are preferred. Harman et al (1990) showed that sampling designs based on adjacent days can result in highly correlated repeated measurements. Hence, the correlation with true usual intake is reduced in surveys in which the data are collected on consecutive days. Such sampling designs could also over-estimate the population variance of the usual intake and under estimate intra-individual variance if dietary intake varies across seasons. If the consecutive days belong to the same season the seasonal variation of individual intake is not observable. Tarasuk and Beaton (1992) document that mean intake estimates obtained from samples on consecutive days are less reliable and more likely to be biased than those based on randomly selected non-consecutive days using the complete set of 365 days of data.

5) If an unequal distribution of seasons and days of the week is evident then the data should be adjusted for significant nuisance effects by linear regression (Nusser et al, 1996) or weighted to achieve equal sums of weights for each season and each day of the week (Gay 2000).

The suppled data did not contain a seasonal indicator but had indicators for day of the week and month of pregnancy. The dietary intake of the participants was recorded during the 4^{th} and 7^{th} month of pregnancy and it was assumed that the two months of pregnancy cover all seasons of the year.

The data were separated into four data sets containing 2, 4, 6 and 8 repeated measurements per individual using simple random sampling without replacement. The sampling was based on the master data set containing 16 repeated measurements per individual. Hence, the observations included in one of the data sets could be included in the other data sets. An equal number of repeats were selected from the two observation months (4^{th} and 7^{th} month of pregnancy). An attempt was made to minimize the correlation between repeated observations per individual by randomly selecting consumption measurements which are far apart from each other in terms of time.

For example, the data set which contains two repeated observations per individual has one observation drawn from the intake recorded during the 4^{th} month of pregnancy and one drawn from the intake recorded during the 7^{th} month of pregnancy. Table 10 documents that the sample with eight repeats per individual has similar number of observations across all days of the week for the two months of pregnancy. Similar results are obtained for the datasets containing two, four and six repeats per individual. Hence, it could be assumed that the bias arising from uneven coverage of the days of the week has been avoided.

Table 10: Sample Counts for Day of the Week and Month of Pregnancy. Eight Repeats per Individual

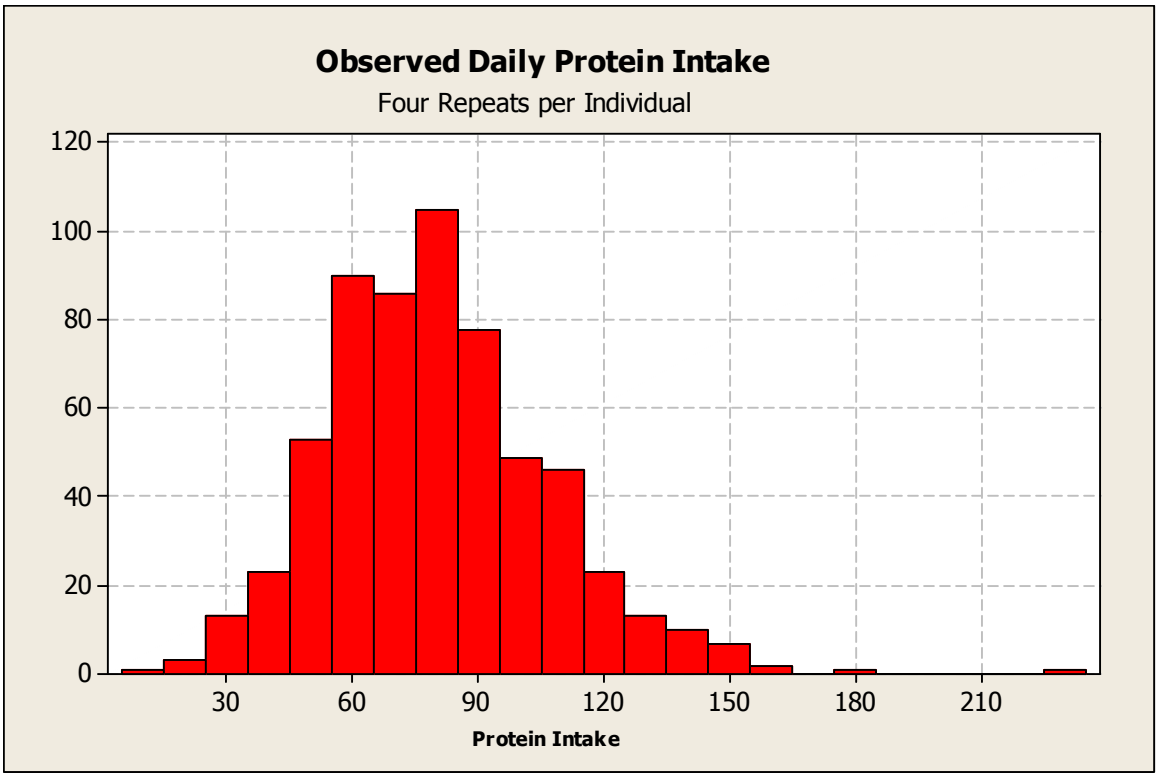
	Month of Pregnancy			
Day of the Week	**4th**	**7th**	**Total**	**Percent Total**
1	89	88	177	14.65%
2	85	100	185	15.31%
3	88	92	180	14.90%
4	78	66	144	11.92%
5	97	107	204	16.89%
6	86	77	163	13.49%
7	81	74	155	12.83%
Grand Total	604	604	1208	

3.3 Preliminary Analysis

Some of the most often used methods for estimating dietary intake distributions are based on an ANOVA type variance decomposition. Since the sample contains repeated measurements data it is important to investigate the covariance structure of the data prior to performing the analysis of variance. According to Littell et al (1998) if the repeated measurements have equal variance at al times and if the pairs of measurements on the same individual are equally correlated, regardless of the time lag between the measurements than the univariate ANOVA is valid from a statistical point of view and yields an optimal method of analysis. The condition required for validity of the univariate ANOVA tests is known as Huynh-Feldt condition (Huynh-Feldt, 1970).

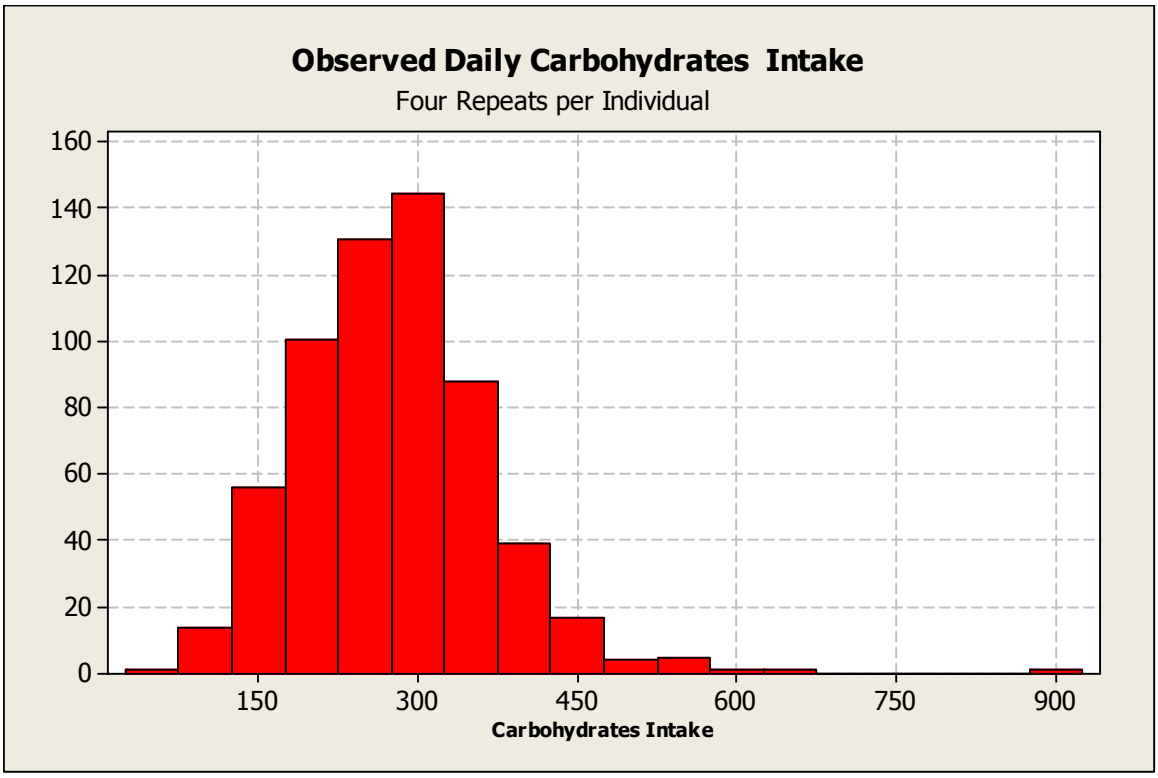
It was decided to use protein, carbohydrates, and vitamin E nutrient intake data for the investigation of the methods for estimating usual intake distributions. The distributions of the daily intakes of these nutrients exhibited most of the typical characteristics of the nutrient intake data. Graphs of the observed daility intakes of the three nutrients analyzed in this research are shown below. The graphs illustrate that the observed daily intake distributions of the three nutrients have positive skewness with some extreme values in the right tail of the distribution. The distribution of the Vitamin E intake is highly positively skewed which is typical for vitamins and nutrients that are not consumed regularly on daily basis.

Graph 1: Observed Daily Protein Intake



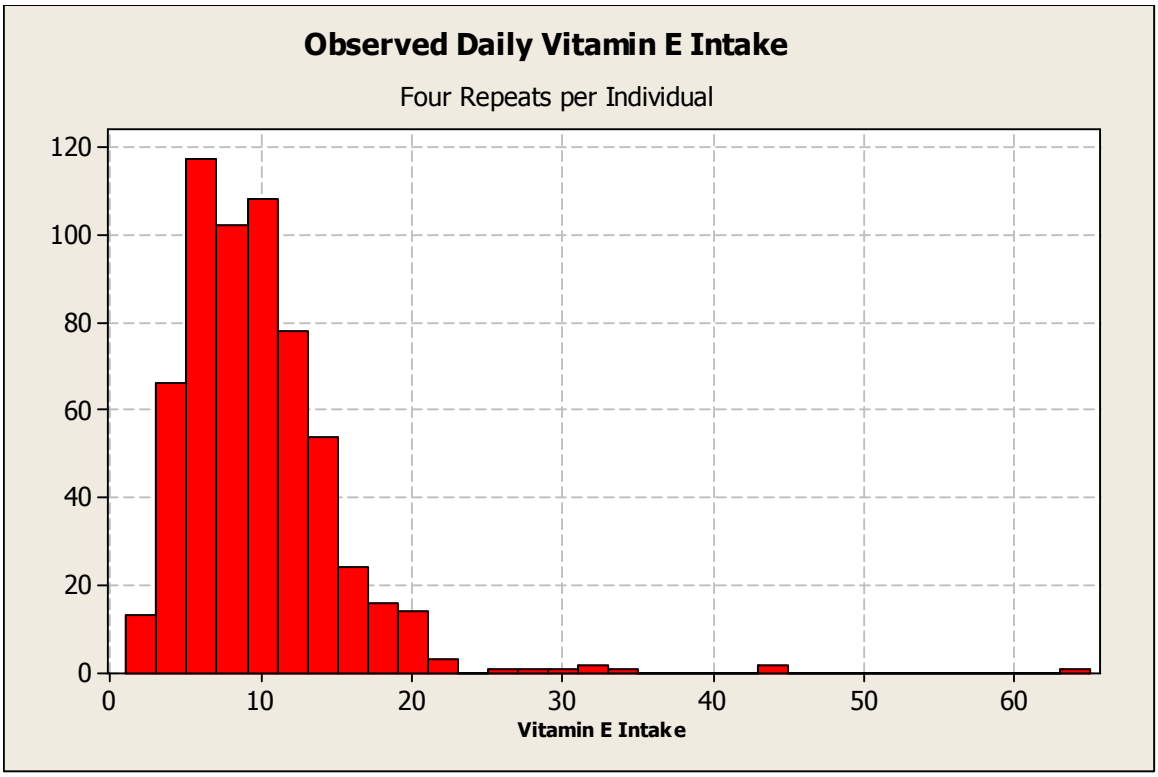
Note: Protein intake is measured in grams

Graph 2: Observed Daily Carbohydrates Intake



Note: Carbohydrates intake is measured in grams

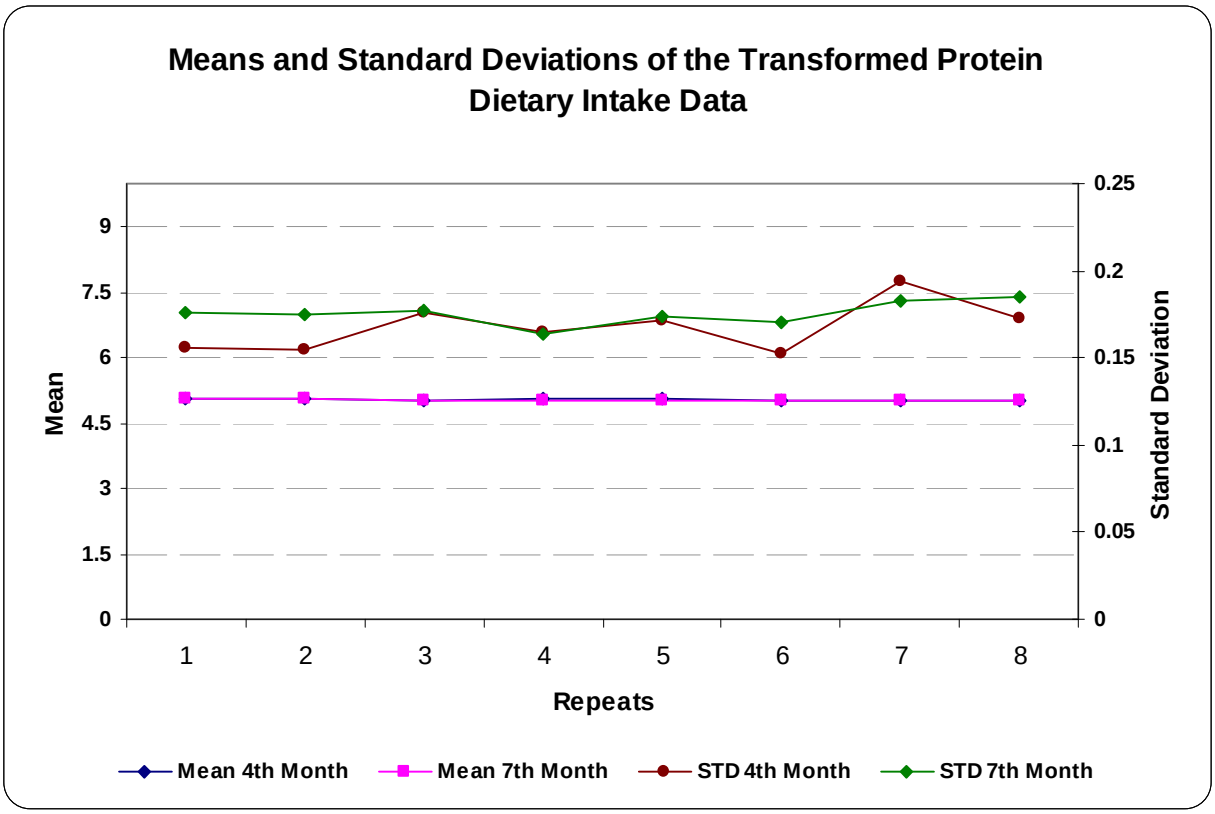
Graph 3: Observed Daily Vitamin E Intake



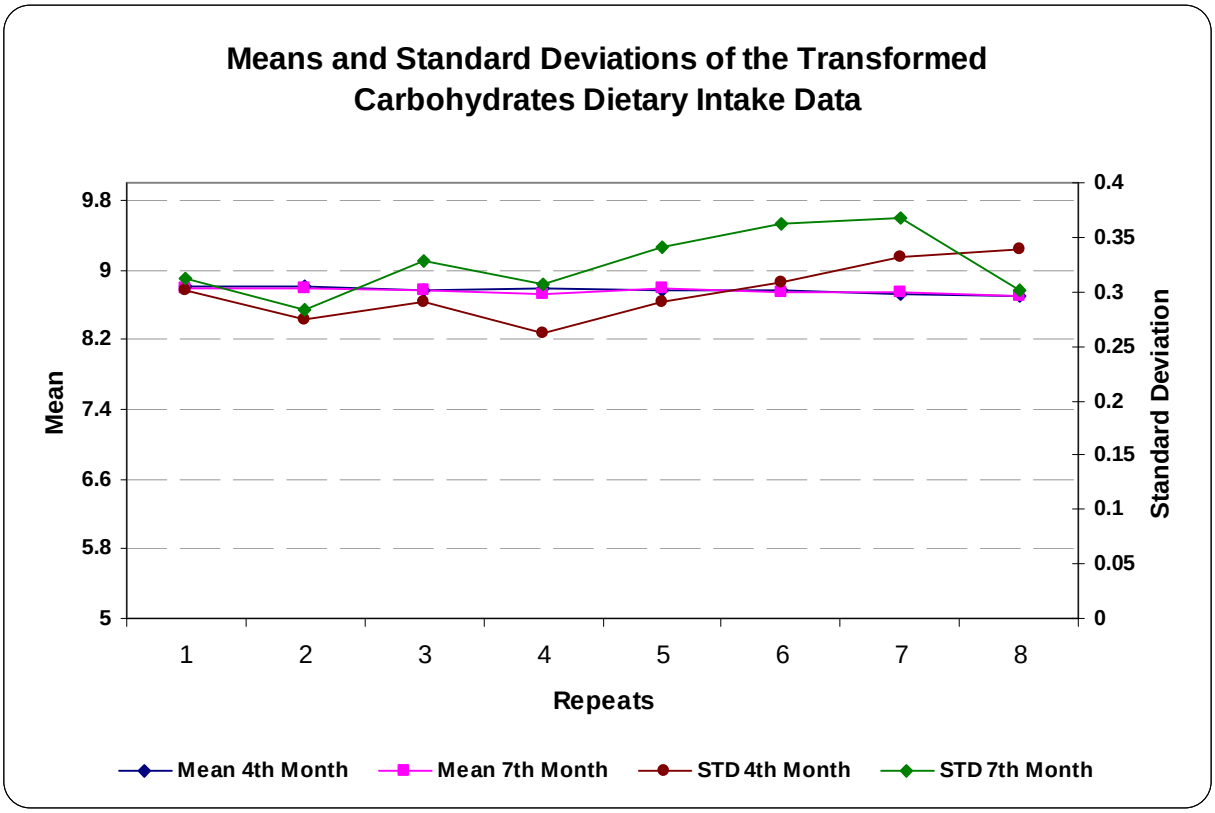
Note: Vitamin E intake is measured in milligrams

Prior to performing ANOVA variance decomposition the nutrient intakes were transformed to normality or at least symmetry using an optimal transformation algorithm developed by Hoffmann et al (2002a). The transformation has a variance stabilizing effect on the data as well. The method is described in detail in Section 4. After the transformation of the data the intake means and standard deviations were plotted and the graphs are shown below. The analysis indicates that the assumption of equal variance at all repeated measurements is reasonably accurate. Furthermore, the graphs indicate that there is no trend in the mean long-term nutrient intakes.

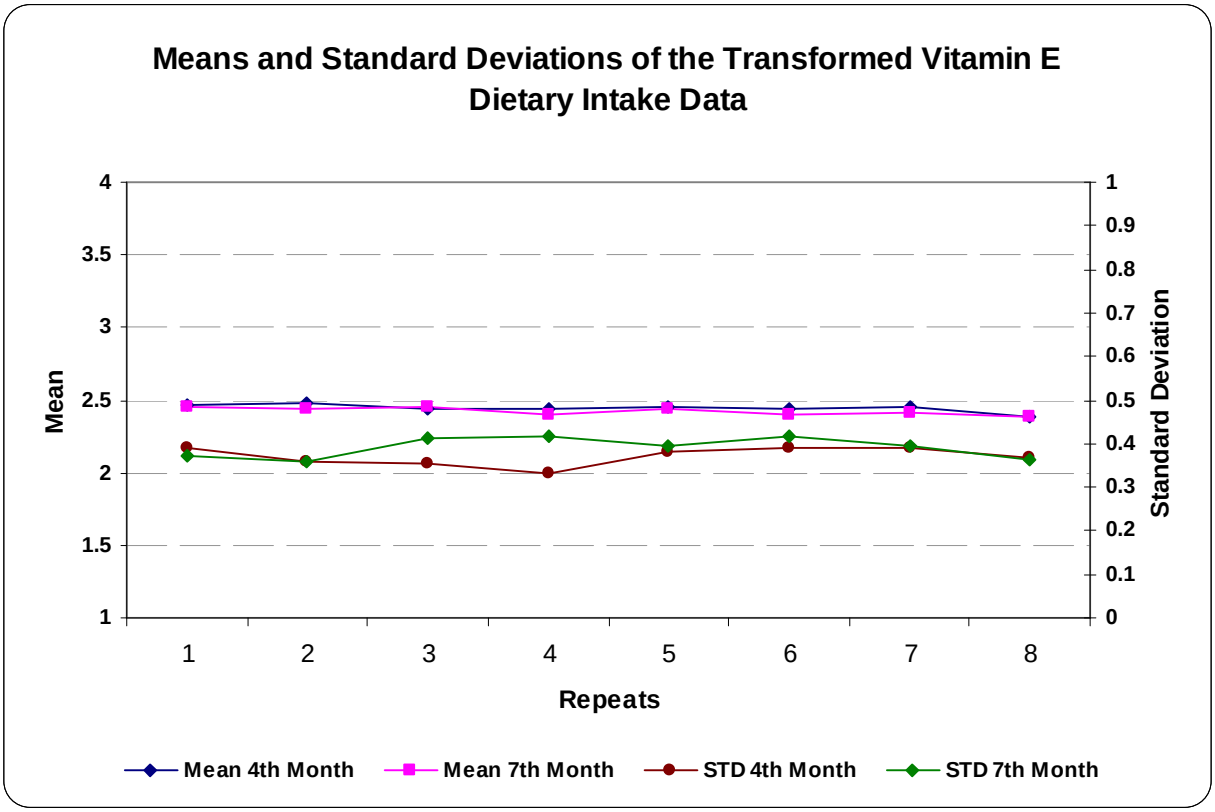
Graph 4: Means and Standard Deviations of the Transformed Protein Intake



Graph 5: Means and Standard Deviations of the Transformed Carbohydrates Intake



Graph 6: Means and Standard Deviations of the Transformed Vitamin E Intake



Since random samples of repeats on non-consecutive days from the 4th and 7th month of pregnancy are used in the analysis it could be assumed that the correlation in the repeated measurements data is minimized.

Carriquiry et al (1995) document the day-to-day correlations for a large number of dietary intakes based on the 1989-1991 CSFII data. Table 11 illustrates the correlations among the repeated measurements computed from the datasets with an increasing number of observations per individual. The correlations are computed assuming compound symmetry of the covariance structure.

Littell et al (1998) note that there are two aspects of the correlation between observations on the same individual. The repeated measurements are correlated because they share common contributions from the individual. This is due to variation between individuals. Furthermore, measures on the same individual close in time are often more highly correlated than measures far apart in time. This is covariation within individuals. The compound symmetry structure of the covariance assumes that measurements on the same individual have equal variance at all times and all pairs of measurements on the same individual have equal correlation. Since the estimated correlations are not large in magnitude it is assumed that they will not have a significant effect on the statistical results. The correlation between repeated measurements in not investigated further in this research.

Table 11: Dietary Components Correlations

Dietary Intake	Carriquiry et al (1995) correlation coefficient estimate for females =>20 years of age	Estimated correlation coefficient for 4 repeats per individual	Estimated correlation coefficient for 6 repeats per individual	Estimated correlation coefficient for 8 repeats per individual
Protein	0.10	0.08	0.06	0.07
Carbohydrates	0.178	0.16	0.13	0.08
Vitamin E	0.099	-0.05	0.03	0.03

The relationship between the means and standard deviations of the dietary intakes data was also investigated during the preliminary analysis stage. For each individual in the sample the mean and standard deviation of the transformed towards normality daily intakes were estimated. A linear regression consisting of a constant and linear term was fitted to the resulting paired data. In the regression the standard deviation was the dependent variable and the mean was the independent variable. Table 12, Table 13 and Table 14 below document the linear term coefficients and t-values obtained from the regressions:

Table 12: Regression results: Relationship Between Means and Standard Deviations. Four Repeats per Individual

Dietary Intake	Linear term	t-value	p-value
Protein	0.03	0.64	0.52
Carbohydrates	0.02	0.40	0.69
Vit E	0.01	0.37	0.71

Table 13: Regression results: Relationship Between Means and Standard Deviations. Six Repeats per Individual

Dietary Intake	Linear term	t-value	p-value
Protein	0.02	0.49	0.62
Carbohydrates	-0.03	-0.98	0.33
Vit E	-0.05	-1.62	0.11

Table 14: Regression results: Relationship Between Means and Standard Deviations. Eight Repeats per Individual

Dietary Intake	Linear term	t-value	p-value
Protein	0.05	1.46	0.15
Carbohydrates	-0.01	-0.17	0.86
Vit E	-0.03	-0.83	0.41

The above results indicate that the assumption of independence between means and standard deviations of the repeated measurements data is valid.

If the within-person variance is not constant across individuals it will introduce bias in the estimates of the usual intake distribution parameters. Table 15 documents the t-values from the tests of heterogeneity of within-person variance. Since the t-values are below 2 it could be concluded that the assumption of homogenous within-person variance is valid.

Table 15: T-values for Tests of Heterogenous Within-Person Variance

Dietary Intake	Four repeats t-values	Six repeats t-values	Eight repeats t-values
Protein	0.09	1.02	1.01
Carbohydrates	1.20	0.79	0.59
Vit E	1.06	0.99	1.16

Tests for nuisance effects using a mixed general linear model were performed in order to investigate if any adjustments of the data would be necessary. The data were tested for day-of-the-week, weekend, season and interview sequence nuisance effects. Month of pregnancy was used as a proxy for season.

The results indicated some evidence of day-of-the-week effect for carbohydrates and Vitamin E. There was also some evidence of weekend effect in the vitamin E data. For the three nutrients investigated in this research no evidence of seasonal or interview sequence effects was found. Table 16, Table 17, Table 18 and Table 19 document the F-values and in brackets the corresponding p-values obtained from the tests of equality of means:

Table 16: F-values for Day of the Week Effect

Dietary Intake	Four repeats F-values	Six repeats F-values	Eight repeats F-values
Protein[(a)]	1.62 (0.140)	1.32(0.247)	0.47(0.832)
Carbohydrates	1.23(0.288)	2.28(0.034)*	2.06(0.055)
Vit E	2.71(0.013)*	2.74(0.012)*	2.38(0.027)*

(a) The values in the brackets are the p-values for the respective F values and degrees of freedom
* Indicates significance at the 95% confidence level
** Indicates significance at the 99% confidence level

Table 17: F-values for Weekend Effect

Dietary Intake	Four repeats F-values	Six repeats F-values	Eight repeats F-values
Protein[(a)]	1.25(0.264)	2.08(0.150)	0.80(0.371)
Carbohydrates	2.35(0.126)	2.56(0.110)	0.89(0.346)
Vit E	7.93(0.005)**	7.48(0.006)**	5.08(0.025)*

(a) The values in the brackets are the p-values for the respective F values and degrees of freedom
* Indicates significance at the 95% confidence level
** Indicates significance at the 99% confidence level

Table 18: F-values for Seasonal Effect

Dietary Intake	Four repeats F-values	Six repeats F-values	Eight repeats F-values
Protein[(a)]	0.04(0.835)	0.06(0.810)	0.03(0.867)
Carbohydrates	1.54(0.216)	0.86(0.353)	0.65(0.420)
Vit E	0.24(0.621)	0.21(0.650)	0.19(0.661)

(a) The values in the brackets are the p-values for the respective F values and degrees of freedom

Table 19: F-values for Interview Sequence Effect

Dietary Intake	Four repeats F-values	Six repeats F-values	Eight repeats F-values
Protein[(a)]	1.32(0.262)	1.02(0.394)	2.31(0.57)
Carbohydrates	0.58(0.678)	1.18(0.318)	1.56(0.184)
Vit E	1.01(0.402)	1.29(0.272)	1.20(0.310)

(a) The values in the brackets are the p-values for the respective F values and degrees of freedom

The results from the preliminary analysis indicated low correlation between the repeated measurements on the same individual. Furthermore, no relationship between the individual means and standard deviations and no heterogeneity of the within-person variances were found. There was some evidence of nuisance effects, in particular day-of-the-week and weekend effects. Several of the methods described in Section 4 allow for corrections for correlated repeated observations, heterogeneity of variance and nuisance effects. As a result of the preliminary analysis conclusions it was decided that only corrections for nuisance effects will be included in some of the methods for estimating usual intake distributions.

4. Results

Four different methods for estimating usual intake distributions are evaluated and compared in this chapter. The comparison includes an overview of the methodologies and the theoretical assumptions underlying the methods as well as their application to real data containing dietary intakes. The procedures investigated in this paragraph include the Hoffman et al (2002), ISU (1996), Chang et al (2001) and NCI (2006) methods.

Hoffman et al (2002) developed a procedure which is a simplified version of the ISU (1996) method. The use of this procedure for estimating the long-term intake distribution was recommended by the EFCOSUM (2000/2001) project. The method is based on a standard ANOVA type variance decomposition and uses initial data transformation to normality for which and exact back transformation formula exists. The shrinkage estimator of individual usual intake used in the Hoffmann et al (2002) method has optimal statistical properties only if the distribution of the observed daily intakes is normal. Furthermore, the estimator is affected by the degree of day-to-day variability in the nutrient intake measurements. Small day-to-day nuisance variance will result in an adjusted usual intake distribution which is close to the observed distribution of the individual nutrient intake means. Large day-to-day nuisance variance will result in an estimated usual intake distribution with a spike at the grand mean of the observed group nutrient intake.

In contrast, the NCI (2006) procedure is a further enhancement of the ISU (1996) method. Both the NCI and ISU methods are complex, multi-stage procedures. The ISU method is also based on an ANOVA type variance decomposition but includes a number of additional data transformation steps. These are needed in order to transform the repeated measurements towards normality and perform adjustments for nuisance effects, heterogeneous within-individual variances and correlated repeated measurements.

The NCI (2006) method is based on a two-part mixed effects model which could be used for estimating the usual intake distribution from samples containing large number of nonconsumption days. It separates the probability of consumption from the consumption-day amount. Furthermore, the method allows for correlation between the probability of consuming a food on a single day and the consumption-day amount. The main advantage of the method is the ability to use covariates to model both the probability of consumption and the amount of nutrients consumed.

The Chang et al (2001) method is considered an alternative to the mixture distribution approach. The procedure is based on the overdispersed exponential family of distributions and the generalized linear models framework. The main advantage of the method is that no transformation of the data towards normality is needed since all estimations are performed on the original scale.

4.1 Hoffman et al Method (2002)

The first step in the application of the Hoffmann et al (2002) method is the transformation of the dietary intake data to normality or at least symmetry. The optimal transformation is derived using two-parameter family of Box-Cox transformations defined by equation (28). The power parameter τ is restricted to be zero or the inverse of a positive integer. The restriction is imposed so that an exact back transformation formula could be used. However, because of this restriction the common maximum likelihood method can not be applied and a grid search procedure which maximises the Shapiro-Wilk statistics is used instead. In the iterative procedure τ varies over the grid 1, 1/2, 1/3, 1/4……1/10, 0 and the shift (location) parameter ω varies over the same grid multiplied by the mean of the original data. If some of the dietary intakes contain zeros then the zero values are replaced by a small constant which is derived as 0.0001 multiplied by the original sample mean of the dietary intake prior to performing the transformation to normality.

Table 20 documents the transformation parameters and some output statistics for the set of nutrient intakes. For the three nutrients the transformation procedure only approximates normality.

Table 20: Hoffann et al Method (2002). Transformation Parameters and Output Statistics for a Set of Nutrient Intakes

Data	Statistics	Protein	Carbohydrates	Vitamin E
Four Replicates	τ	0	1	1
	ω	0	0	0
	skewness	-0.7614	0.8860	3.1114
	kurtosis	2.3285	3.6451	20.9494
	Shapiro-Wilk	0.9709	0.9640	0.7968
Six Replicates	τ	0	1	1
	ω	0	27.570	0.9267
	skewness	-0.9944	0.9132	3.6275
	kurtosis	3.5427	3.3746	28.2796
	Shapiro-Wilk	0.9573	0.9615	0.7694
Eight Replicates	τ	0	1	1
	ω	0	0	0.9099
	skewness	-1.0428	0.8162	3.3451
	kurtosis	3.7759	2.6086	26.1876
	Shapiro-Wilk	0.9523	0.9684	0.7944

The adjusted usual intakes on the transformed scale are back-transformed to the original scale using equation (34). The back-transformation takes into account that the usual intakes t in the transformed scale are measured with error ε which is normally distributed with mean zero and variance σ_{ε}^{2}. The inverse transformation is applied to the compounded term $t+\varepsilon$ and integrated over the error distribution.

Table 21 below lists the derivation of the mean in the original scale for a range of power transformation parameters.

Table 21: Hoffmann et al Method (2002). Expected Value in the Original Scale

Power Parameter τ	$E(X)$
0	$\exp\left(\mu+\frac{1}{2}\sigma_{\varepsilon}^{2}\right)-\omega$
$\frac{1}{4}$	$\left(\frac{1}{4}\mu+1\right)^{4}+\frac{3}{8}\sigma_{\varepsilon}^{2}\left(\frac{1}{4}\mu+1\right)^{2}+\frac{3}{256}\sigma_{\varepsilon}^{2}-\omega$
$\frac{1}{3}$	$\left(\frac{1}{3}\mu+1\right)^{3}+\frac{1}{3}\sigma_{\varepsilon}^{2}\left(\frac{1}{3}\mu+1\right)-\omega$
$\frac{1}{2}$	$\left(\frac{1}{2}\mu+1\right)^{2}+\frac{1}{4}\sigma_{\varepsilon}^{2}-\omega$
1	$(\mu+1)-\omega$

The Hoffmann et al (2002) method is a simplified procedure for estimating usual intake distributions. Since the data transformation stage is not that complicated and the method does not include corrections for nuisance effects or correlated repeated measurements it could be applied to small sample sizes. The Hoffmann et al (2002) method does not incorporate sampling weights or the use of covariates to model the usual intake distribution. However, it should be noted that further developments of the method in order to include corrections for nuisance effects or covariate information are relatively simple to implement. The method does not provide by default the estimates of the standard errors of the usual intake distribution parameters or confidence intervals of the distribution percentiles.

If a large proportion of individuals in the sample are genuine non-consumers and have zero daily intakes then the Hoffmann et al (2002) method could be further modified in order to account for zero consumption. In such cases the estimation of the usual intake distribution will be based on pooled data of the back-transformed usual intakes of the consumers and the zero usual intakes of the non-consumers.

Initially, the Hoffmann et al (2002) method is applied to the protein consumption data consisting of 4 repeated observations per individual. The preliminary data analysis results indicated that the protein consumption data do not have statistically significant nuisance effects and the correlation between repeated measurements on the same individual is low. Furthermore, since the sample size of the available data is small the Hoffman et al (2002) method is expected to produce robust results.

The results shown in Table 22 document that the Hoffmann et al (2002) method produces estimates which are closer to the observed distribution parameters estimated from 16 repeats per individual. It is assumed that the estimates obtained from the observed distribution of 16 recalls per individual are closest to the estimates obtained from the unobserved sample distribution of 365-day means. As expected, the estimate of the standard deviation obtained from the adjusted usual intake distribution is reduced since the effect of the within-person variance has been removed.

Graph 7: Hoffmann et al. Method (2002). Protein Usual Intake Distribution.

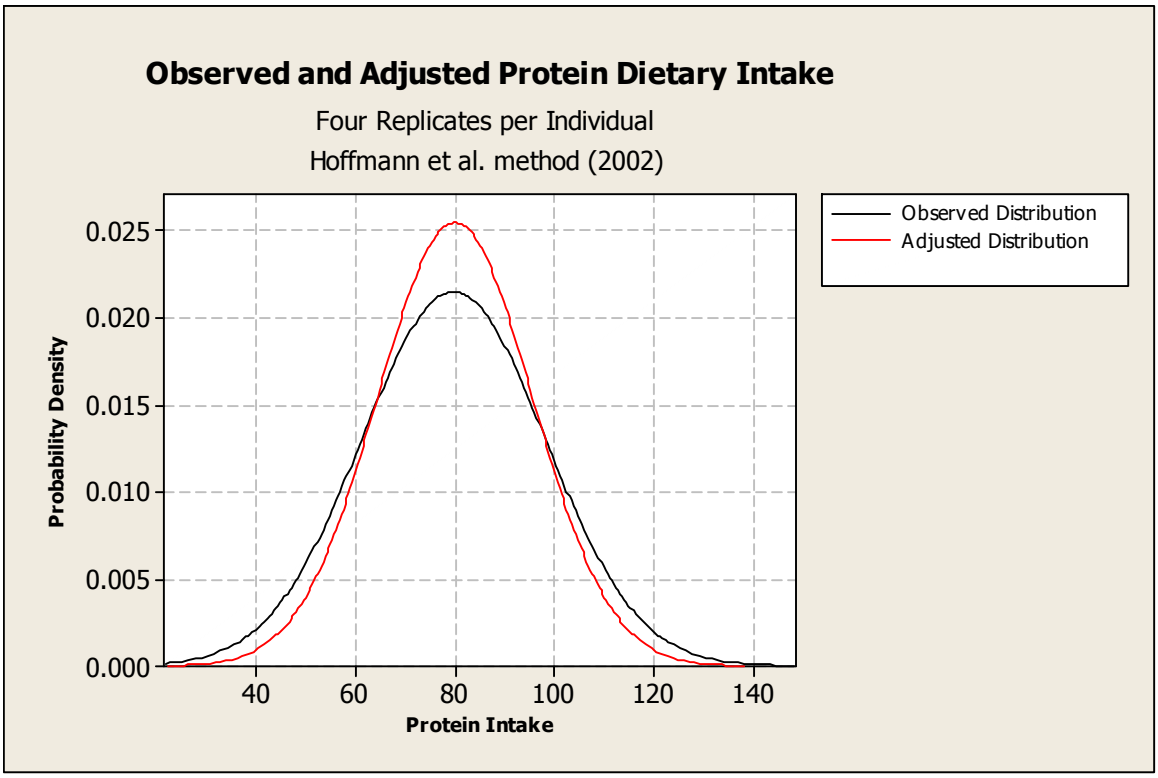


Table 22: Hoffmann et al. Method (2002). Comparison Between Observed and Estimated Distribution of Usual Protein Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	46.00	55.50	61.50	76.00	94.00	107.00	114.5	78.62	21.47
	4 recalls	49.25	54.75	69.25	78.25	91.25	102.75	110.25	79.69	18.52
	6 recalls	49.33	53.83	68.17	78.00	90.33	101.17	103.83	78.72	16.60
	8 recalls	48.88	54.25	67.13	78.00	88.50	97.88	103.13	77.83	16.27
	16 recalls	51.50	57.31	68.63	78.81	89.00	96.88	105.38	78.26	16.02
Estimated Usual Intake Distribution	2 recalls	59.58	64.96	69.46	79.04	89.07	96.34	100.06	79.37	13.33
	4 recalls	54.40	58.05	71.14	79.32	89.53	98.67	106.72	79.97	15.62
	6 recalls	52.11	56.41	70.45	79.75	89.90	97.40	102.04	79.07	14.92
	8 recalls	50.29	56.62	68.71	78.84	88.30	97.44	102.16	78.25	15.20
95% CI[b]	2 recalls	(49.07,63.03)	(60.68,67.27)	(67.85,71.96)	(76.36,82.54)	(85.56,92.23)	(93.96,99.94)	(97.18,104.37)	(77.34,81.40)	(11.98,15.13)
	4 recalls	(49.09,57.10)	(54.68,63.02)	(68.54,73.59)	(77.46,81.66)	(86.21,94.24)	(96.43,106.35)	(99.13,113.44)	(77.21,82.23)	(13.89,17.78)
	6 recalls	(50.12,55.74)	(52.46,61.45)	(66.02,73.69)	(76.83,82.51)	(86.22,92.75)	(94.89,101.04)	(99.29,103.75)	(76.49,81.22)	(13.49,16.94)
	8 recalls	(47.20,55.05)	(51.58,62.02)	(64.94,72.50)	(76.50,82.15)	(85.60,90.09)	(91.86,99.74)	(97.90,106.14)	(75.52,80.34)	(13.53,17.18)

(a) Protein intake is measured in grams
(b) Bootstrap estimates based on 1000 samples
(c) Arithmetic mean

Table 22 compares estimates of the usual intake distribution of protein obtained from different number of repeated measurements per individual. The observed standard deviation decreases when the number of repetitions increases. The estimated standard deviations are smaller than the observed standard deviations since they have been adjusted for the effect of within-person variance. Table 23 and Table 24 document the estimated usual intake distribution parameters for carbohydrates and vitamin E, respectively. For protein and vitamin E the estimates obtained from 8 recalls per individual are closer to the observed estimates obtained from 16 recalls per individual. Since by default the Hoffmann et al (2002) method does not produce standard errors and confidence intervals of the estimated usual intake distribution parameters, bootstrapping with 1000 samples is used to derive the 95% confidence intervals of the estimates.

Table 23: Hoffmann et al. Method (2002). Comparison Between Observed and Estimated Distribution of Usual Carbohydrates Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	151.50	179.50	217.00	272.50	330.00	378.00	407.5	276.39	78.94
	4 recalls	178.25	194.5	236.25	274.75	314.75	361.50	382.00	276.78	64.16
	6 recalls	168.33	199.33	229.33	271.33	327.83	357.67	380.83	275.70	65.44
	8 recalls	173.88	199.63	221.25	273.50	315.63	350.00	375.38	271.12	63.67
	16 recalls	167.25	189.44	230.19	271.00	324.31	352.69	375.94	273.69	63.58
Estimated Usual Intake Distribution	2 recalls	174.15	192.16	234.99	278.20	321.67	356.33	379.74	277.89	61.04
	4 recalls	194.73	208.26	243.03	275.09	308.40	347.33	364.40	276.78	53.43
	6 recalls	179.87	207.54	234.32	271.80	322.23	348.86	369.54	275.70	58.41
	8 recalls	182.10	205.67	225.47	273.30	311.86	343.33	366.56	271.12	58.29
95% CI[b]	2 recalls	(160.50,188.79)	(174.92,212.24)	(222.67,251.67)	(265.89,293.40)	(303.84,330.19)	(340.62,377.56)	(362.32,403.05)	(267.96,288.40)	(55.18,67.80)
	4 recalls	(169.05,207.63)	(198.89,216.59)	(225.12,252.79)	(264.68,286.96)	(300.95,324.32)	(330.88,361.30)	(352.43,391.05)	(268.02,285.56)	(48.22,60.48)
	6 recalls	(158.90,205.75)	(180.91,214.98)	(221.00,244.43)	(258.42,292.63)	(304.46,330.27)	(337.55,366.27)	(351.09,392.52)	(266.20,286.54)	(52.86,65.88)
	8 recalls	(160.41,198.46)	(187.13,212.42)	(216.77,235.77)	(252.82,287.04)	(299.85,324.85)	(327.08,360.04)	(347.40,399.06)	(261.75,281.50)	(52.94,64.97)

(a) Carbohydrates intake is measured in grams
(b) Bootstrap estimates based on 1000 samples
(c) Arithmetic mean

Table 24: Hoffmann et al. Method (2002). Comparison Between Observed and Estimated Distribution of Usual Vitamin E Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	4.00	4.50	6.50	9.00	12.00	14.00	18.00	9.40	4.21
	4 recalls	4.25	5.00	6.75	9.25	11.25	13.25	15.00	9.31	3.64
	6 recalls	4.17	5.17	6.67	9.17	11.17	13.67	15.00	9.27	3.45
	8 recalls	4.50	5.25	6.50	8.88	11.25	13.25	14.13	9.10	3.18
	16 recalls	4.44	5.31	6.75	9.44	11.25	13.06	14.31	9.23	3.11
Estimated Usual Intake Distribution	2 recalls	5.88	6.20	7.51	9.14	11.10	12.40	15.01	9.40	2.75
	4 recalls	5.47	6.04	7.37	9.27	10.78	12.30	13.63	9.31	2.76
	6 recalls	5.14	5.95	7.16	9.19	10.80	12.83	13.90	9.27	2.79
	8 recalls	5.19	5.82	6.89	8.91	10.93	12.63	13.38	9.10	2.71
95% CI[b]	2 recalls	(4.90,6.20)	(5.88,6.53)	(6.86,8.00)	(8.49, 9.79)	(10.44,11.75)	(11.75,14.03)	(12.73,16.44)	(9.00, 9.89)	(2.42,3.24)
	4 recalls	(4.81,5.95)	(5.47,6.61)	(6.80,8.10)	(8.70,9.83)	(10.21,11.26)	(11.54,13.44)	(12.87,15.53)	(8.90,9.83)	(2.37,3.69)
	6 recalls	(4.60,5.82)	(5.68,6.36)	(6.76,7.84)	(8.24, 9.59)	(10.40,11.54)	(11.88,13.90)	(12.96,15.93)	(8.83,9.79)	(2.48, 3.39)
	8 recalls	(4.81,5.61)	(5.29,6.46)	(6.57,7.42)	(8.16,9.76)	(10.34,11.51)	(11.89,13.38)	(12.84,15.50)	(8.69,9.62)	(2.47,3.12)

(a) Vitamin E intake is measured in milligrams
(b) Bootstrap estimates based on 1000 samples
(c) Arithmetic mean

4.2 ISU Method (1996)

The ISU (1996) method is a complex multi-stage procedure for estimation of usual intakes distribution. It allows corrections for nuisance effects as well as correlated repeated measurements. The method includes numerous data transformations and hence is suitable for larger sample sizes. The ISU (1996) method could be used for the analysis of survey data collected with complex sampling designs since sample weights could be incorporated in the analysis. The method is based on the assumptions of the classical measurement error theory. The most important assumptions of the theory are normal distribution and equal intra-individual variances. The preliminary analysis indicated that those assumptions are approximately fulfilled on the transformed scale. However, the ISU (1996) method could also be used in the case of heterogeneous within-individual variances remaining after the data transformations to normality. The ISU method does not allow the use of covariates to model the usual intake distribution. The method provides by default estimates of the standard errors of the usual intake distribution percentiles. The ISU (1996) method also provides estimates of the individual usual intakes for the individuals in the input data set.

If some of the individuals in the sample are non-consumers and have zero daily intakes then a modified ISU (1996) method proposed by Nusser et al (1997) could be used to model the usual intake distribution. The modified method treats the zero observations separately from the positive observations. The long-term average consumption is a product of two components – the probability of consumption and the long-term average amount consumed on consumption days. The Nusser et al (1997) method is described in detail in the literature review section.

Initially, the ISU (1996) method is applied to the protein consumption data consisting of 4 repeated observations per individual using the SIDE software. The results are shown in Graph 8 and Table 25 below:

Graph 8: ISU Method (1996). Protein Usual Intake Distribution.

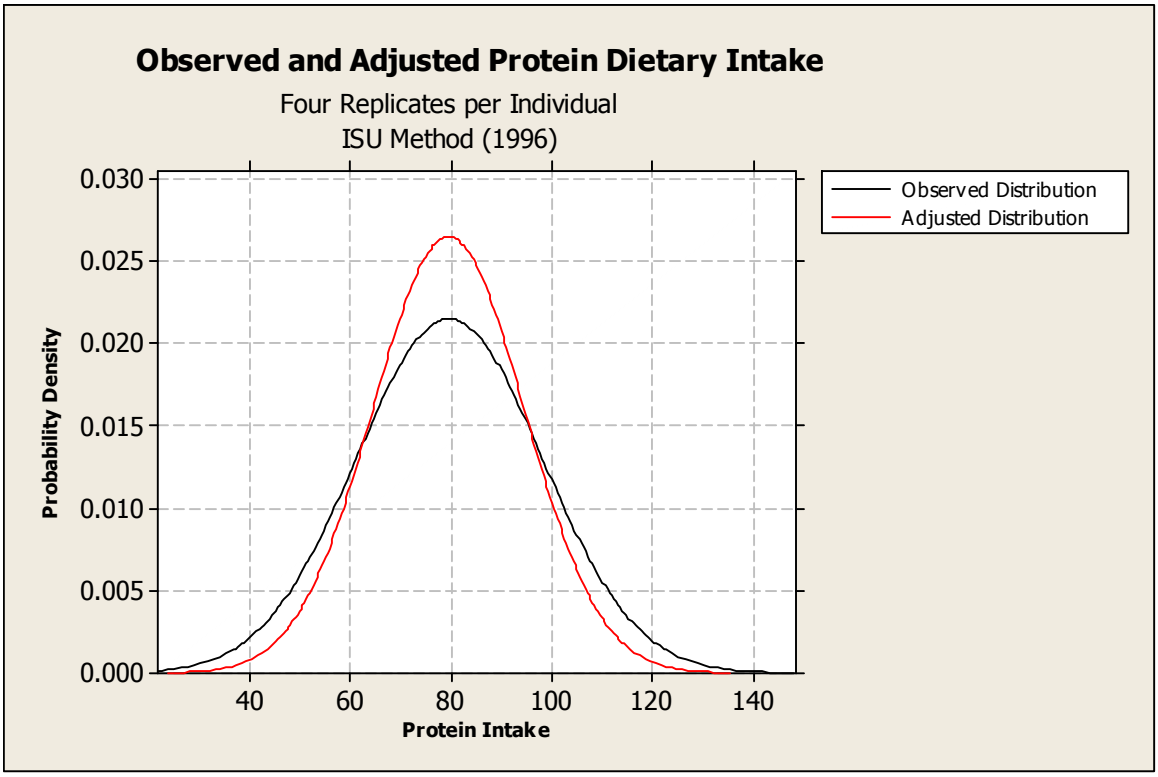


Table 25: ISU Method (1996). Comparison Between Observed and Estimated Distribution of Usual Protein Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[d]	SD
Observed Intake Distribution	2 recalls	46.00	55.50	61.50	76.00	94.00	107.00	114.5	78.62	21.47
	4 recalls	49.25	54.75	69.25	78.25	91.25	102.75	110.25	79.69	18.52
	6 recalls	49.33	53.83	68.17	78.00	90.33	101.17	103.83	78.72	16.60
	8 recalls	48.88	54.25	67.13	78.00	88.50	97.88	103.13	77.83	16.27
	16 recalls	51.50	57.31	68.63	78.81	89.00	96.88	105.38	78.26	16.02
Estimated Usual Intake Distribution[c]	2 recalls	59.59(3.87)	63.94(3.34)	71.50(2.56)	80.31(2.21)	89.57(2.84)	98.27(4.03)	103.64(4.90)	80.80	13.40
	4 recalls	57.51(2.63)	61.93(2.40)	69.73(2.08)	78.99(2.00)	88.90(2.39)	98.38(3.10)	104.31(3.65)	79.70	14.27
	4 recalls adj.[b]	57.63(2.66)	62.09(2.42)	69.93(2.10)	79.23(2.01)	89.13(2.39)	98.57(3.09)	104.45(3.63)	79.90	14.26
	6 recalls	58.05(2.48)	62.47(2.30)	70.17(2.05)	79.18(2.00)	88.66(2.26)	97.57(2.78)	103.08(3.18)	79.70	13.70
	8 recalls	58.42(2.45)	62.86(2.26)	70.41(1.98)	79.12(1.93)	88.34(2.22)	97.30(2.82)	103.02(3.27)	79.70	13.60

(a) Protein intake is measured in grams
(b) Adjusted for nuisance effects
(c) The estimated standard errors are shown in brackets
(d) Arithmetic mean

Tables 25, 26 and 27 document that the estimated standard errors of the percentiles closer to the centre of the distribution are smaller compared to the estimated standard errors of the percentiles in the tails of the distribution. The estimates of the usual intake distribution percentiles derived when adjustments for nuisance effects are used are slightly higher compared to the estimates obtained without nuisance effects adjustments.

For protein intake the parameters of the usual intake distributions estimated from 2 and 8 recalls per individual are close to each other. For carbohydrates and vitamin E the estimates obtained from 2 recalls per individual are closer to the observed estimates obtained from 16 recalls per individual. If we compare the lower (p5,p10) and upper (p90,p95) percentile estimates to the observed estimates obtained from 16 recalls per individual it could be concluded that the ISU (1996) method might not provide very accurate estimates of the tails of the distribution. The estimated adjusted standard deviations of protein usual intake vary between 13.40 and 14.27. As mentioned by Chang et al (2001), the ISU (1996) method could produce unstable variance estimates. This is illustrated by the estimate of the adjusted standard deviation of vitamin E usual intake distribution obtained from 8 recalls per individual. The estimate is larger than the observed standard deviation obtained from 8 recalls per individual. Chang et al (2001) correctly noted that the ISU (1996) method involves too many data transformations which are greatly dependent on the specific data values. A slight change in the data might result in considerably different distribution estimates. Furthermore, the back transformation algorithm used to transform the data back to its original scale is only an approximation.

Table 26: ISU Method (1996). Comparison Between Observed and Estimated Distribution of Usual Carbohydrates Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[d]	SD
Observed Intake Distribution	2 recalls	151.50	179.50	217.00	272.50	330.00	378.00	407.50	276.39	78.94
	4 recalls	178.25	194.50	236.25	274.75	314.75	361.50	382.00	276.78	64.16
	6 recalls	168.33	199.33	229.33	271.33	327.83	357.67	380.83	275.70	65.44
	8 recalls	173.88	199.63	221.25	273.50	315.63	350.00	375.38	271.12	63.67
	16 recalls	167.25	189.44	230.19	271.00	324.31	352.69	375.94	273.69	63.58
Estimated Usual Intake Distribution[c]	2 recalls	183.13(13.00)	202.12(11.23)	234.77(8.31)	271.45(6.56)	308.29(8.41)	343.84(13.49)	367.54(17.97)	272.83	56.40
	4 recalls	200.35(9.88)	217.56(8.98)	248.02(7.66)	284.28(7.23)	323.05(8.73)	360.08(11.60)	383.23(13.82)	287.08	55.74
	4 recalls adj.[b]	208.63(10.13)	226.32(9.18)	257.53(7.81)	294.57(7.34)	334.04(8.83)	371.62(11.72)	395.06(13.95)	297.25	56.80
	6 recalls	195.32(9.89)	231.95(9.02)	246.58(7.69)	284.50(7.08)	324.35(8.45)	363.47(12.02)	389.40(15.19)	287.39	59.46
	8 recalls	196.95(9.30)	215.04(8.59)	246.92(7.52)	284.30(7.08)	323.78(8.35)	362.22(11.44)	387.32(14.12)	287.19	58.24

(a) Carbohydrates intake is measured in grams
(b) Adjusted for nuisance effects
(c) The estimated standard errors are shown in brackets
(d) Arithmetic mean

Table 27: ISU Method (1996). Comparison Between Observed and Estimated Distribution of Usual Vitamin E Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[d]	SD
Observed Intake Distribution	2 recalls	4.00	4.50	6.50	9.00	12.00	14.00	18.00	9.40	4.21
	4 recalls	4.25	5.00	6.75	9.25	11.25	13.25	15.00	9.31	3.64
	6 recalls	4.17	5.17	6.67	9.17	11.17	13.67	15.00	9.27	3.45
	8 recalls	4.50	5.25	6.50	8.88	11.25	13.25	14.13	9.10	3.18
	16 recalls	4.44	5.31	6.75	9.44	11.25	13.06	14.31	9.23	3.11
Estimated Usual Intake Distribution[c]	2 recalls	5.09(0.52)	5.85(0.47)	7.25(0.38)	9.06(0.36)	11.26(0.57)	13.73(1.0)	15.50(1.36)	9.51	3.26
	4 recalls	5.33(0.41)	6.07(0.40)	7.50(0.39)	9.37(0.40)	11.56(0.53)	13.96(0.84)	15.70(1.14)	9.80	3.27
	4 recalls adj.[b]	5.61(0.43)	6.39(0.42)	7.87(0.40)	9.79(0.40)	12.00(0.53)	14.39(0.82)	16.08(1.11)	10.18	3.28
	6 recalls	5.27(0.39)	6.00(0.37)	7.42(0.36)	9.29(0.40)	11.56(0.52)	14.07(0.82)	15.90(1.15)	9.78	3.40
	8 recalls	5.25(0.38)	6.00(0.37)	7.44(0.37)	9.31(0.39)	11.55(0.52)	14.03(0.80)	15.84(1.07)	9.77	3.34

(a) Vitamin E intake is measured in milligrams
(b) Adjusted for nuisance effects
(c) The estimated standard errors are shown in brackets
(d) Arithmetic mean

4.3 Chang et al Method (2001)

Chang et al (2001) method is based on the parametric assumption that the overdisperesed exponential family of distributions could be used to estimate the usual dietary intake distributions. The extra variation which exists in the usual intake distribution due to the large within-person variation makes the distribution overdispersed. Two approaches could be used to model such an overdispersed distribution. The first one is to use a mixture of exponential family densities as sampling densities and use the mixture distribution methodology to estimate the usual intake distribution. The second approach is to use an additional scale parameter in the exponential family form to account for the overdispersion. Chang et al (2001) prefer the second approach since the total variance of the usual intake distribution consisting of within and between-person components and their ratio could be easily obtained from the data. The authors apply adjustments developed by Liu et al (1978) to minimize the effect of the within-person variance on the variance of the usual intake distribution. Similar to the other methods discussed in the thesis the adjustment of the variance is carried out by dividing the total variance into within-person and between-person components.

One of the major advantages of the method is that all computations are performed on the original scale. Hence, complex data transformation and back transformation procedures are avoided. The method could be used for the analysis of survey data since sample weights could be incorporated in the estimation of the usual intake distribution. Chang et al (2001) method could also be used to estimate the usual intake distribution of specific groups in the population such as age, education, or gender groups. Some future enhancements of the method that could be recommended include corrections for nuisance effects, correlated repeated measurements and extensions for modeling nutrient intake data with a large proportion of zero consumption.

A slightly modified version of the Chang et al (2001) method is used to estimate the distribution of usual intake. The estimates of the ratio of within to between-individual variance are obtained using the standard analysis of variance method rather than Liu et al (1978) formulae. As described in the literature review section, in order to obtain an estimate of the proportion of within to between-individual variance which is representative of the population of interest Chang et al (2001) use an external data set. In the analyses which follow external data sets containing two, four, six and eight repeated measurements per individual are generated and used to estimate the ratio of within to between-individual variance. These data sets contain nutrient consumption data which were not included in the samples used to estimate the usual intake distribution parameters. The external data sets are generated using the same sampling methodology as the one described in the data preparation section. They are equal in terms of size to the datasets used to estimate the usual intake distribution. However, as already noted in the literature review section estimates computed using external data sets might differ from the true parameters in the population study leading to bias when those external estimates are used in the primary study to derive usual intake distribution. Table 28 documents the estimated ratios of within to between-individual variance. Those ratios are applied in the adjustments of the usual intake distribution parameters estimated using overdispersed gamma distribution.

The main reason for choosing the gamma distribution is the right skewness of the observed distributions of the three nutrients investigated in this research. The Chang et al (2001) method does not provide by default estimates of the standard errors of the usual intake distribution parameters. The confidence interval estimates shown in Table 30, Table 31 and Table 32 are obtained by bootstrap.

Table 28: Chang et al Method (2001). Estimated Ratios of Within to Between-Individual Variance

Nutrient	2 Repeats	4 Repeats	6 Repeats	8 Repeats
Protein	1.28	1.30	1.43	1.45
Carbohydrates	1.03	1.08	1.05	1.06
Vitamin E	1.18	1.23	1.21	1.19

Table 29: Chang et al Method (2001). Anderson-Darling Goodnes of Fit Test. Gamma Distribution

Nutrient	A-D test Statistics. Two Repeats per Individual.	P-value. Two Repeats per Individual.	A-D test Statistics. Four Repeats per Individual.	P-value. Four Repeats per Individual.
Protein	0.487	0.23	0.656	0.09
Carbohydrates	0.342	0.25	0.256	0.50
Vitamin E	0.291	0.50	0.582	0.14

The results shown in Table 29 support the parametric assumption that gamma distribution could be used to model the usual intake distribution.

Graph 9: Chang et al Method (2001). Protein Usual Intake Distribution.

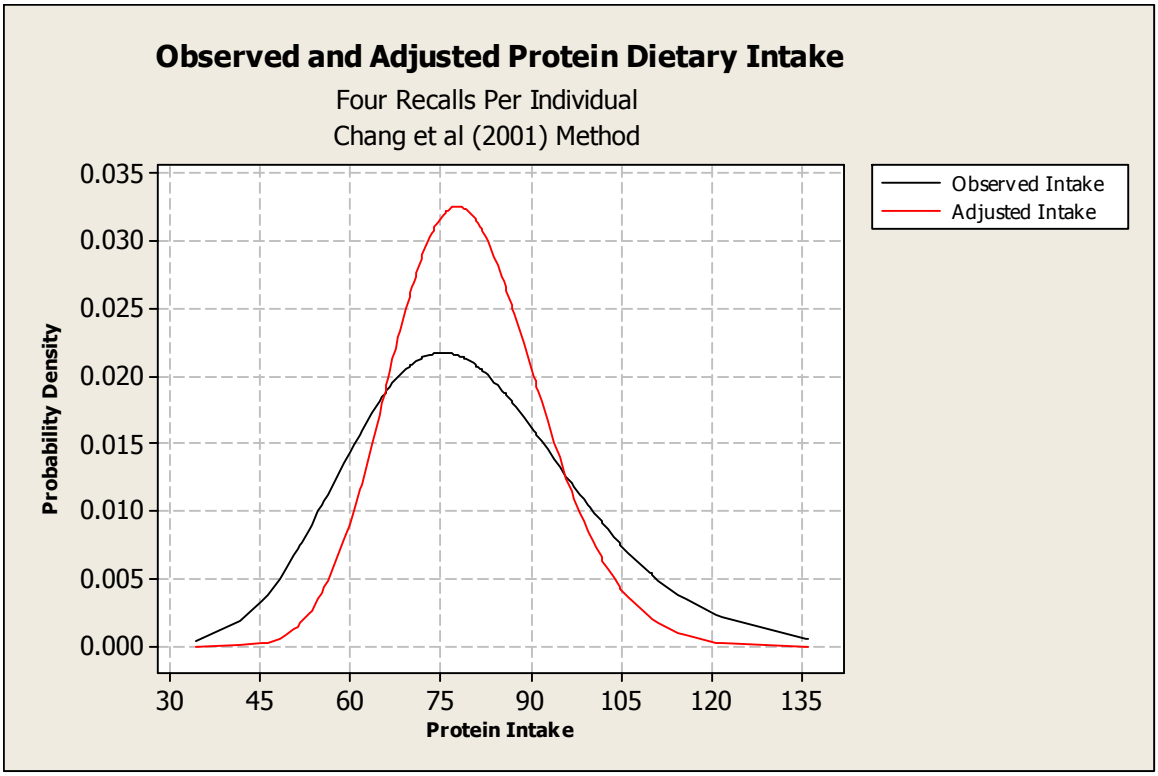


Table 30: Chang et al. Method (2001). Comparison Between Observed and Estimated Distribution of Usual Protein Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	46.00	55.50	61.50	76.00	94.00	107.00	114.5	78.62	21.47
	4 recalls	49.25	54.75	69.25	78.25	91.25	102.75	110.25	79.69	18.52
	6 recalls	49.33	53.83	68.17	78.00	90.33	101.17	103.83	78.72	16.60
	8 recalls	48.88	54.25	67.13	78.00	88.50	97.88	103.13	77.83	16.27
	16 recalls	51.50	57.31	68.63	78.81	89.00	96.88	105.38	78.26	16.02
Estimated Usual Intake Distribution	2 recalls	57.00	61.00	68.50	79.00	88.50	97.50	104.00	78.62	14.35
	4 recalls	60.50	64.75	71.75	79.25	88.75	96.00	101.75	79.69	12.38
	6 recalls	62.67	65.17	71.00	79.17	87.00	93.50	99.17	78.72	11.06
	8 recalls	61.13	64.63	71.00	77.63	84.75	92.38	96.50	77.83	10.67
95% CI[b]	2 recalls	(53.63,60.04)	(57.24,64.08)	(66.67,73.16)	(76.39,82.80)	(84.76,91.94)	(92.88,100.98)	(99.06,108.74)	(75.55,81.68)	(13.57,15.91)
	4 recalls	(57.29,65.17)	(61.49,67.05)	(68.76,75.07)	(75.93,82.12)	(85.28,93.02)	(91.81,99.46)	(97.74,106.62)	(76.78,82.58)	(11.58,13.97)
	6 recalls	(59.70,65.77)	(61.59,67.88)	(67.99,73.70)	(76.66,82.60)	(84.63,90.42)	(90.61,96.13)	(96.17,103.80)	(76.15,81.28)	(10.31,12.56)
	8 recalls	(59.29,64.76)	(61.16,67.17)	(68.31,73.90)	(74.66,80.04)	(81.69,87.39)	(88.12,94.23)	(92.80,99.95)	(75.32,80.31)	(9.94,12.11)

(a) Protein intake is measured in grams
(b) Bootstrap estimates based on 500 samples
(c) Arithmetic mean

Table 31: Chang et al Method (2001). Comparison Between Observed and Estimated Distribution of Usual Carbohydrates Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	151.50	179.50	217.00	272.50	330.00	378.00	407.5	276.39	78.94
	4 recalls	178.25	194.5	236.25	274.75	314.75	361.50	382.00	276.78	64.16
	6 recalls	168.33	199.33	229.33	271.33	327.83	357.67	380.83	275.70	65.44
	8 recalls	173.88	199.63	221.25	273.50	315.63	350.00	375.38	271.12	63.67
	16 recalls	167.25	189.44	230.19	271.00	324.31	352.69	375.94	273.69	63.58
Estimated Usual Intake Distribution	2 recalls	197.00	208.00	237.00	273.00	312.00	352.00	375.00	276.39	55.85
	4 recalls	212.00	222.25	247.75	274.50	305.75	336.00	360.00	276.78	44.97
	6 recalls	204.17	218.33	244.17	273.50	307.83	336.83	357.67	275.70	46.42
	8 recalls	202.50	216.00	240.25	271.63	301.75	331.25	349.50	271.12	44.76
95% CI[b]	2 recalls	(184.36,213.14)	(193.63,218.27)	(222.03,247.18)	(256.28,285.61)	(294.38,325.30)	(327.75,371.74)	(353.25,391.65)	(263.48,289.20)	(53.16,61.35)
	4 recalls	(201.74,225.84)	(210.59,230.58)	(238.00,259.23)	(262.96,283.79)	(293.77,315.81)	(320.94,347.68)	(346.35,379.70)	(266.68,287.05)	(42.80,49.34)
	6 recalls	(191.15,214.03)	(204.91,227.72)	(233.01,254.54)	(261.09,283.57)	(295.90,321.80)	(322.62,349.25)	(341.99,373.68)	(265.39,286.32)	(44.33,50.77)
	8 recalls	(192.08,210.94)	(204.09,226.37)	(227.42,250.91)	(262.82,282.97)	(291.81,313.69)	(315.24,345.11)	(334.30,363.82)	(261.13,281.49)	(42.84,48.56)

(a) Carbohydrates intake is measured in grams
(b) Bootstrap estimates based on 500 samples
(c) Arithmetic mean

Table 32: Chang et al Method(2001). Comparison Between Observed and Estimated Distribution of Usual Vitamin E Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	4.00	4.50	6.50	9.00	12.00	14.00	18.00	9.40	4.21
	4 recalls	4.25	5.00	6.75	9.25	11.25	13.25	15.00	9.31	3.64
	6 recalls	4.17	5.17	6.67	9.17	11.17	13.67	15.00	9.27	3.45
	8 recalls	4.50	5.25	6.50	8.88	11.25	13.25	14.13	9.10	3.18
	16 recalls	4.44	5.31	6.75	9.44	11.25	13.06	14.31	9.23	3.11
Estimated Usual Intake Distribution	2 recalls	5.50	6.00	7.50	9.50	11.50	13.50	16.50	9.40	2.86
	4 recalls	6.00	6.50	7.75	9.25	11.00	12.50	13.75	9.31	2.38
	6 recalls	6.00	6.67	7.67	9.17	10.83	12.33	13.67	9.27	2.31
	8 recalls	5.88	6.50	7.88	9.00	10.50	12.00	13.00	9.10	2.16
95% CI[b]	2 recalls	(4.90,6.16)	(5.17,6.42)	(6.79,8.05)	(8.93,10.38)	(10.72,12.51)	(12.41,14.73)	(15.03,20.07)	(8.76,10.09)	(2.79,3.10)
	4 recalls	(5.52,6.62)	(5.92,6.92)	(7.13,8.32)	(8.68,9.86)	(10.34,11.85)	(11.36,13.43)	(12.52,14.88)	(8.75,9.90)	(2.27,2.65)
	6 recalls	(5.52,6.53)	(6.22,7.24)	(7.07,8.14)	(8.57,9.70)	(10.23,11.54)	(11.40,13.09)	(12.82,14.79)	(8.74,9.81)	(2.22,2.52)
	8 recalls	(5.35,6.22)	(5.98,6.94)	(7.50,8.56)	(8.46,9.52)	(9.84,11.09)	(11.14,12.66)	(12.11,13.75)	(8.61,9.60)	(2.10,2.31)

(a) Vitamin E intake is measured in milligrams
(b) Bootstrap estimates based on 500 samples
(c) Arithmetic mean

The reduction in variance of the observed distribution depends on the estimated ratio of within to between-individual variance. Greater ratios lead to larger reductions in variance. Table 28 documents that protein intake has the biggest within to between individual variance ratios. For the protein intake data with 4 recalls per individual the ratio is 1.30 and it moves the 5^{th} percentile from 49.25g to 60.50g and the 95^{th} percentile from 110.25g to 101.75g. The standard deviation of the distribution is reduced from 18.52g to 12.38g. If we compare the lower (p5,p10) and upper (p90,p95) percentile estimates to the observed estimates obtained from 16 recalls per individual it could be concluded that the Chang et al (2001) method might not provide very accurate estimates of the tails of the distribution, especially when the assumptions of the method are violated.

Chang et al (2001) use bootstrap to investigate the mean and 95% confidence intervals of the ratio of within to between individual variance. Their results indicate that the ranges of the bootstrap confidence intervals are wide. The authors conclude that the large confidence intervals suggest that external data sets with small sample sizes might not provide stable estimates of the within-individual variance. External data sets with large sample sizes should produce better estimates of the ratio of within to between-individual variance and thus better estimates of the distribution of usual dietary intake.

4.4 National Cancer Institute Method (2006)

The method could be used for a wide array of applications and has some distinct advantages over the methods discussed in the previous sections since it includes a number of enhancements. The NCI (2006) method allows the inclusion of covariates such as sex, age, race, as well as nuisance effects such as day-of-week effect to model the usual intake distribution using a mixed model regression. Furthermore, it could also accommodate a large number of non-consumption days by separating the probability of consumption from the consumption day amount. This feature of the model is further enhanced with the ability to allow for correlation between the probability of consuming a food on a single day and the consumption-day amount. Hence, the NCI (2006) method is very flexible and represents a new development in the field. The method consists of two parts. The first part of the method estimates the probability of consuming a food and the consumption amount. However, the first part of the model could also be used to estimate the usual intake distribution of nutrients consumed on a daily basis.

The second part of the model uses Monte Carlo simulations to derive the distribution of usual intake. An estimate of the person specific random effect is added to the estimates obtained from the mixed model. The within-person variation is not included since by definition it does not contribute to the long-term intake. In order to improve the precision of the estimated usual intake distribution a large number of pseudo-persons for each person included in the sample are generated. The number of generated pseudo-persons is specified by the researcher. The method does not provide by default estimates of the standard errors of the usual intake distribution parameters or confidence intervals of the distribution percentiles. The confidence intervals shown in Tables 33-35 are computed using Monte Carlo simulations where 1000 pseudo-persons are generated for each individual in the sample. The usual intake distribution parameters and the respective confidence intervals are estimated empirically from the simulated population.

During the model fitting process the repeated measurements data are transformed using Box-Cox type of transformation. Hence, when back-transformation is used to transform the consumption data to the original scale a bias-adjustment term is added in order to make the mean of the back-transformed intake the same as the mean of the original intake.

Covariate information could be used in both parts of the model. The method also allows for correlated person-specific random effects. Unfortunately, no covariate information which could be used for estimating the usual intake distributions is available in the supplied data. It is expected that the covariate information will improve the accuracy of the estimates of the long-term intake distributions, especially of the tails of the distributions.

Compared to the ISU (1996) method, the NCI (2006) method produces slightly lower estimates of the protein usual intake distribution mean and percentiles whereas the estimated standard deviation is slightly higher. The pattern is similar for the carbohydrates usual intake distribution. For vitamin E intake the pattern is similar for the estimates of the usual intake distribution mean and percentiles. However, the NCI (2006) method produces more stable and accurate estimates of the standard deviation of the long-term vitamin E intake distribution.

Graph 10: NCI Method (2006). Protein Usual Intake Distribution.

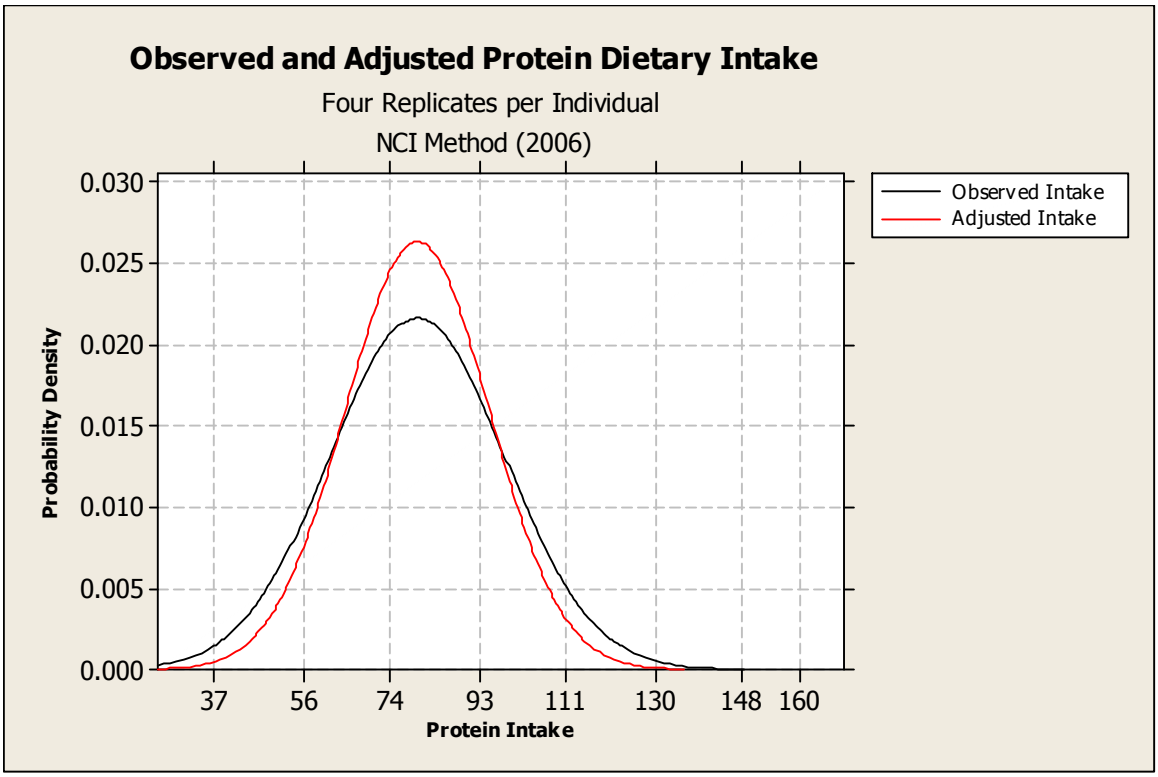


Table 33: NCI Method (2006). Comparison Between Observed and Estimated Distribution of Usual Protein Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	46.00	55.50	61.50	76.00	94.00	107.00	114.5	78.62	21.47
	4 recalls	49.25	54.75	69.25	78.25	91.25	102.75	110.25	79.69	18.52
	6 recalls	49.33	53.83	68.17	78.00	90.33	101.17	103.83	78.72	16.60
	8 recalls	48.88	54.25	67.13	78.00	88.50	97.88	103.13	77.83	16.27
	16 recalls	51.50	57.31	68.63	78.81	89.00	96.88	105.38	78.26	16.02
Estimated Usual Intake Distribution[b]	2 recalls	57.17	61.61	69.29	78.23	87.55	96.23	101.49	78.64	13.50
	4 recalls	56.26	60.89	69.08	78.93	89.44	99.56	105.95	79.71	15.15
	6 recalls	56.13	60.65	68.62	78.07	88.11	97.70	103.70	78.74	14.49
	8 recalls	55.08	59.60	67.63	77.14	87.28	96.98	103.04	77.84	14.61
95% CI[b]	2 recalls	(57.08, 57.25)	(61.53, 61.68)	(69.23, 69.34)	(78.17, 78.30)	(87.48, 87.62)	(96.14, 96.32)	(101.37, 101.61)	(78.60, 78.70)	(13.47, 13.53)
	4 recalls	(56.20, 56.33)	(60.83, 60.95)	(69.03, 69.13)	(78.88, 78.98)	(89.38, 89.50)	(99.49, 99.65)	(105.86, 106.04)	(79.67, 79.75)	(15.12, 15.17)
	6 recalls	(56.08, 56.18)	(60.60, 60.69)	(68.58, 68.66)	(78.03, 78.10)	(88.07, 88.15)	(97.64, 97.75)	(103.64, 103.78)	(78.71, 78.77)	(14.47, 14.51)
	8 recalls	(55.03, 55.12)	(59.57, 59.64)	(67.60, 67.66)	(77.11, 77.18)	(87.25, 87.32)	(96.93, 97.03)	(102.98, 103.11)	(77.82, 77.87)	(14.59, 14.63)

(a) Protein intake is measured in grams
(b) Monte Carlo estimates based on 1000 simulated values for each individual in the sample
(c) Arithmetic mean

Table 34: NCI Method(2006). Comparison Between Observed and Estimated Distribution of Usual Carbohydrates Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	151.50	179.50	217.00	272.50	330.00	378.00	407.5	276.39	78.94
	4 recalls	178.25	194.5	236.25	274.75	314.75	361.50	382.00	276.78	64.16
	6 recalls	168.33	199.33	229.33	271.33	327.83	357.67	380.83	275.70	65.44
	8 recalls	173.88	199.63	221.25	273.50	315.63	350.00	375.38	271.12	63.67
	16 recalls	167.25	189.44	230.19	271.00	324.31	352.69	375.94	273.69	63.58
Estimated Usual Intake Distribution[b]	2 recalls	183.21	201.72	234.34	273.61	315.30	355.29	380.48	276.53	60.06
	4 recalls	191.87	208.41	237.90	273.65	312.19	349.44	372.98	276.85	55.26
	6 recalls	183.62	201.26	233.06	271.79	314.08	355.53	381.91	275.80	60.54
	8 recalls	179.81	197.32	228.88	267.22	309.16	350.14	376.10	271.18	59.93
95% CI[b]	2 recalls	(182.85, 183.57)	(201.42, 202.01)	(234.07, 234.62)	(273.36, 273.88)	(314.99,315.61)	(354.86, 355.70)	(379.92, 381.03)	(276.32, 276.75)	(59.91, 60.21)
	4 recalls	(191.62, 192.11)	(208.20, 208.62)	(237.72, 238.06)	(273.46, 273.82)	(312.01, 312.38)	(349.17, 349.72)	(372.62, 373.34)	(276.71, 276.98)	(55.16, 55.36)
	6 recalls	(183.42,183.83)	(201.09, 201.44)	(232.91, 233.21)	(271.63, 271.95)	(313.90,314.25)	(355.28, 355.77)	(381.58, 382.22)	(275.68, 275.92)	(60.46, 60.63)
	8 recalls	(179.64, 179.99)	(197.16, 197.47)	(228.75, 229.02)	(267.08, 267.36)	(308.99,309.31)	(349.95, 350.36)	(375.83, 376.40)	(271.08, 271.29)	(59.86, 60.00)

(a) Carbohydrates intake is measured in grams
(b) Monte Carlo estimates based on 1000 simulated values for each individual in the sample
(c) Arithmetic mean

Table 35: NCI Method (2006). Comparison Between Observed and Estimated Distribution of Usual Vitamin E Intake

	Data[a]	P5	P10	P25	P50	P75	P90	P95	AM[c]	SD
Observed Intake Distribution	2 recalls	4.00	4.50	6.50	9.00	12.00	14.00	18.00	9.40	4.21
	4 recalls	4.25	5.00	6.75	9.25	11.25	13.25	15.00	9.31	3.64
	6 recalls	4.17	5.17	6.67	9.17	11.17	13.67	15.00	9.27	3.45
	8 recalls	4.50	5.25	6.50	8.88	11.25	13.25	14.13	9.10	3.18
	16 recalls	4.44	5.31	6.75	9.44	11.25	13.06	14.31	9.23	3.11
Estimated Usual Intake Distribution[b]	2 recalls	4.92	5.65	7.06	8.97	11.27	13.73	15.41	9.41	3.26
	4 recalls	5.10	5.77	7.06	8.84	11.02	13.42	15.09	9.30	3.13
	6 recalls	4.93	5.61	6.95	8.78	11.03	13.51	15.23	9.26	3.22
	8 recalls	4.89	5.55	6.86	8.63	10.81	13.20	14.85	9.09	3.11
95% CI[b]	2 recalls	(4.90, 4.93)	(5.63, 5.66)	(7.04, 7.07)	(8.96, 8.98)	(11.25, 11.29)	(13.71, 13.76)	(15.37, 15.45)	(9.40, 9.42)	(3.25, 3.27)
	4 recalls	(5.09, 5.11)	(5.76, 5.78)	(7.05, 7.07)	(8.83, 8.85)	(11.01, 11.04)	(13.41, 13.44)	(15.07,15.12)	(9.29, 9.31)	(3.12, 3.14)
	6 recalls	(4.92, 4.94)	(5.60, 5.62)	(6.94, 6.96)	(8.77, 8.79)	(11.02, 11.04)	(13.49, 13.53)	(15.21, 15.26)	(9.25, 9.26)	(3.22, 3.23)
	8 recalls	(4.88, 4.90)	(5.54, 5.56)	(6.85, 6.87)	(8.62, 8.64)	(10.80, 10.82)	(13.18,13.21)	(14.83, 14.87)	(9.08, 9.10)	(3.10,3.12)

(a) Vitamin E intake is measured in milligrams
(b) Monte Carlo estimates based on 1000 simulated values for each individual in the sample
(c) Arithmetic mean

The results documented in Section 4 indicate that all methods investigated in this research are based on a common theoretical framework. The methods assume classical measurement error model where the observed daily intake consists of long-term intake plus additive measurement error. Hence, the variance of the observed daily intake is larger than the variance of the usual intake. The obtained results support the evidence that the widths of the confidence intervals of the estimated usual intake distribution percentiles are not systematically affected by the number of repeated measurements per individual.

All of the methods require repeated measurements data. The repeated measurements are necessary in order to be able to obtain an estimate of the intra-individual variance which is the link between the observed daily nutrient intake and the unobserved usual nutrient intake. The investigated methods attempt to estimate the measurement error (intra-individual variance) and use this estimate to reduce the variance of the observed daily intake in order to derive the distribution of the long-term nutrient intake. Hence, in all graphs and tables shown in Section 4 the long-term intake has smaller variance compared to the variance of the observed daily intake.

The estimates of the usual nutrient intake distribution are suitable for group-level assessment only where a reference value is used to carry out the assessment. For accurate assessment robust estimates of the tail areas of the usual intake distribution need to be obtained. For example, a target intake for a group of individuals can be estimated using the EAR reference value and the variance of the intake. Usually, when setting a target value for the mean group intake the objective is the majority of the individuals (97% or 98%) to meet the nutrient requirement. In such cases the target group's mean intake must be at least two intake standard deviations above the EAR in order less than 2% to 3% of the individual intakes to fall below the EAR. However, because the standard deviation is often related to the magnitude of the intake, the coefficient of variation is used to calculate the target group mean intake. According to Beaton (1994), the formula used to derive a target group mean intake assuming normal distribution of the intake could be defined as:

$$\text{Target mean intake for a group} = \frac{\text{EAR}}{\left(1-\left[2 \text{ x } \text{CV}_{\text{Intake}}\right]\right)} \qquad (115)$$

Table 36 documents the estimates of the coefficient of variation for protein usual intake obtained from two repeated measurements per individual:

Table 36: Usual Protein Intake Coefficients of Variation

Method	CV
Hoffmann et al (2002)	0.168
ISU (1996)	0.166
NCI (2006)	0.172
Chang et al (2001)	0.183

As illustrated in Table 36 the difference between the largest and the lowest estimated coefficients of variation derived using the different methods is around 10%. The estimates obtained using the Hoffmann et al (2002) and ISU (1996) methods are similar. Thus, it could be concluded that although the methods investigated in Section 4 differ in terms of complexity they produce similar group-level assessment results when applied to moderate sized samples of 200 to 400 individuals.

Tables 37 and 38 below provide a summary as well as comparison of three of the methods investigated in the chapter and the Simplified Power method described in Chapter 5 in terms of modeling process steps, strengths and weaknesses.

Table 37: Comparison of the Modeling Process Steps

Method				
	NCI	**Hoffman et al**	**Simplified Power**	**Chang et al**
Step 0: Apply initial data adjustments	1. A Box-Cox transformation to approximate normality as part of the likelihood maximization procedure. The repeated measurements are transformed to normality conditionally on the covariates in the model. 2. Adjustments for nuisance effects such as day-of-the-week.	1. A Box-Cox transformation with power parameter restricted to be zero or the inverse of a positive integer.	1. A Box-Cox transformation to approximate normality.	1. No initial transformation needed. All calculations are performed on the original scale.
Step 1: Assumed relationship between the individual 24-hour recall measurements and individual usual intake	A 24-hour recall is unbiased for usual intake in the untransformed scale. (Assumption B)	A 24-hour recall is unbiased for usual intake in the untransformed scale. (Assumption B)	A 24-hour recall is unbiased for usual intake in the untransformed scale. (Assumption B)	A 24-hour recall is unbiased for usual intake in the untransformed scale. (Assumption B)
Step 2: Partition the total variation in the 24-hour recall measurements into within- and between-person components	Within-person variance is the same among individuals.	Within-person variance is the same among individuals.	Within-person variance is the same among individuals.	Within-person variance is the same among individuals.
Step 3: Estimate the usual intake distribution accounting for within-person variation	1. For foods or nutrients consumed every day the usual intake is estimated using nonlinear mixed effects model fitted to the 24-hour recalls data. For episodically consumed foods and nutrients the usual intake is estimated using a two-part mixed effects model that defines the distribution of reported intake on a given day as the probability of consumption multiplied by the conditional distribution of the amount consumed on a consumption day. 2 The parameters of the mixed effects model are used in a Monte Carlo simulation to estimate the distribution of usual intake. The estimated consumption-day amount data are back-transformed to original scale using an inverse transformation similar to the approach used in the ISU (1996) method which includes a bias adjustment term. 3. The empirical distribution of the original scale simulated values is the estimated usual intake distribution.	1. Set of intermediary values that retain the mean and average between-person variance of the transformed 24-hour recalls is constructed. 2. Back Transformation: An exact formula exists due to the restrictions imposed on the power parameter of the initial Box-Cox transformation. 3. The empirical distribution of the original scale intermediary values is the estimated usual intake distribution.	1. Set of intermediary values that retain the mean and average between-person variance of the transformed 24-hour recalls is constructed. 2. Back Transformation: Taylor series approximation which includes a bias adjustment term as described in the Dodd et al (2006) article. 3. The empirical distribution of the original scale intermediary values is the estimated usual intake distribution.	1. The overdispersed exponential family of distributions and the generalized linear models framework are used to estimate the distribution of usual nutrient intake. 2. Divide the estimated variance of the nutrient intake into within-individual variance and between-individual variance. Apply the adjustments developed by Liu et al (1978) to derive the variance of the usual intake distribution. 3. Use the estimated variance to construct the usual intake distribution.

Table 38: Comparison of the Strengths and Weaknesses of the Models

	Method			
	NCI	**Hoffmann et al**	**Simplified Power**	**Chang et al**
Strengths	1. Accommodates the large number of nonconsumption days that occur with episodically consumed foods, nutrients and trace elements by separating the probability of consumption from the consumption-day amount using a two-part mixed effects model. 2. Allows for correlation between the probability of consuming a food on a single day and the consumption-day amount. 3. Allows the use of covariates to model the probability of consumption on a single day as well as the consumption-day amount.	1. Simple transformations to obtain approximate normality. 2. Robust to mild departures from normality. 3. An exact back transformation formula exists. 4. Suitable for small to medium sized samples.	1. Simple transformations to obtain approximate normality. 2. Robust to mild departures from normality. 3. Provides more accurate estimates of the tails of the usual intake distribution if the transformation to normality is further refined by using a spline with knots. 4. Suitable for small to medium sized samples. 5. More flexible than the Hoffmann et al (2002) method.	1. Working on the original scale of data. 2. Implementation convenience. 2. Can be used for data from complex surveys.
Weaknesses	1. Application to datasets from complex surveys is a future development. 2. Suitable for larger sized datasets. In order to utilize all strengths of the model the data should contain relevant covariate information. 3. The method never produces zero intakes because the logistic regression that is used to model the probability of consumption does not predict zero values. 3. Does not produce standard errors for the estimated parameters of the usual intake distribution. 4. The application of the method for estimating individual usual intake is a future development.	1. Application to datasets from complex surveys is less straightforward. 2. Does not derive standard errors for the estimated parameters of the usual intake distribution. 4. The application to data with many observed zero intakes requires additional modeling steps. 5. Does not allow the use of covariates to model usual intakes distribution. 6. Does not allow adjustments for nuisance effects such as interview sequence or day-of-the-week.	1. Application to datasets from complex surveys is less straightforward. 2. Does not derive standard errors for the estimated parameters of the usual intake distribution. 4. The application to data with many observed zero intakes requires additional modeling steps. 5. The use of covariates to model usual intake distribution requires the application of a complex mixed effects model. 6. The adjustments for nuisance effects such as interview sequence or day-of-the-week are future development. 7. The application of the method for estimating individual usual intake is a future development.	1. The parametric assumption that the usual intake could be modeled using the overdispersed exponential family of distributions might not be fulfilled. 2. An estimate of the ratio of the within to between-individual variance obtained from an external dataset is used to derive the variance of the usual intake distribution. 3. Does not produce standard errors for the estimated parameters of the usual intake distribution. 4. Cannot be applied to datasets with many observed zero intakes. 5. Does not allow the use of covariates to model usual intake distribution. 6. Does not allow adjustments for nuisance effects such as interview sequence or day-of-the-week.

5. Discussion and Suggestions for Future Research

The results obtained from the statistical methods investigated in Chapter 4 indicate that procedures such as the ISU (1996) and NCI (2006) methods require larger sized samples and additional covariates information in order to obtain robust estimates of the usual intake distribution parameters. The ISU method involves a number of data transformations heavily dependent on the specific data values. The polynomials used to refine the transformation to normality may be inaccurate at the tails of the nutrient intake distribution. However, the extreme percentiles of the long-term nutrient intake distribution are places of critical interest to nutrition researchers. Furthermore, the back transformation to the original scale used in the ISU method is only an approximation over the distribution of the measurement error term.

The NCI method requires covariate information in order to obtain accurate estimates of the usual intake distribution, especially of the tails of the distribution. It is difficult to asses whether the ISU and NCI methods are underfitting or overfitting the data since very little diagnostic graphs and statistics are produced. Hence, the application of the ISU and NCI methods as well as the interpretation of their output would require the help of a statistician.

The Hoffmann et al (2002) procedure is a simplified version of the ISU (1996) method. It illustrates that the procedures for estimating usual nutrient intake distributions could be simplified along the lines of the transformation to normality. The method is not optimal but requires considerably less computational effort and could be applied without the help of a statistician. However, overly simplified procedures might produce biased distribution parameter estimates, especially for nutrients with highly skewed intake distributions. For such nutrients the overly simplified methods are expected to produce poor estimates of the tails of the nutrient intake distribution, in particular in cases where the transformation to normality fails.

The purpose of the research is to investigate methods which could be applied to moderate sized samples of 200-400 individuals and be of practical use to nutrition researchers working with nutrient intake data which were not collected using large scale surveys with complicated designs. Ideally, such methods should retain the simplicity of the Hoffmann et al (2002) procedure and demonstrate robustness to departures from the normality assumption. Furthermore, they should provide accurate estimates of the extreme percentiles of the usual intake distribution and have the flexibility of the complex multi-stage methods in order to allow for nuisance effects adjustments or the use of covariates information.

An attempt to develop such method is the simplified best power procedure described in detail below. Similar to the Hoffmann et al (2002) method it is based on a Box-Cox transformation to normality and an ANOVA type variance decomposition. However, in cases where the transformation to normality is not optimal it could be further refined using a spline with knots in order to provide better estimates of the tails of the usual intake distribution. In its most complex form the simplified best power method allows the use of covariates information by applying a mixed effects model. The back-transformation to the original scale is based on a Taylor series approximation with a bias adjustment term. Hence, if only a simple Box-Cox transformation is used in the first stage of the procedure to transform the daily intakes towards normality, then an exact formula could be applied to back transform the estimated usual intakes to the original scale. However, if the transformation to normality is further refined using a spline with knots, then the back transformation is complicated and may not be explicitly expressed. In such cases an approximation of the back transformation function is estimated from the data.

5.1 Simplified Power Method

The method is based on the classical measurement error theory and is similar to the best power method proposed by Nusser et al (1996). The simplest form of the method involves only one-stage power transformation towards normality. The procedure is not optimal but is robust and simple to implement.

The simplified power method is also similar to the method proposed by Hoffmann et al (2002a). Both methods are suitable for small to moderate sized samples and are relatively simple to use. The main differences are in the transformation to normality and back-transformation to the original scale. Hoffman et al (2002a) use a grid search procedure to find an optimal transformation. The power transformation parameter is restricted to be zero or the inverse of a positive integer in order to derive an exact formula for the back-transformation. In contrast, the simplified power method uses standard Box-Cox transformation and the back-transformation is based on a Taylor series approximation which includes a bias adjustment term.

In cases where the transformation to normality needs to be further refined a second step involving smoothing with regression spline is used. The adjustment for within-individual variance is carried out on the transformed scale by means of the standard ANOVA method. The back-transformation to the original scale is performed using Taylor series approximation shown in equation (12) which includes a bias adjustment term. The simplified method is very flexible and could be further enhanced by including adjustments for nuisance effects or covariates to model the long-term intake of nutrients. It does not require the purchase of specialized software and could be performed with most standard statistical packages.

The simplest version of the method is the case where only a power transformation is sufficient to transform the nutrient intake to normality and no covariate information is available to model the long-term nutrient intake. In such cases the method could be applied as follows:

5.1.1 One-Stage Simplified Power Method Without Covariate Information

1) Power-transform the observed nutrient intake towards normality using a Box-Cox transformation. Since in most cases the nutrient intake data is highly positively skewed the goal of the transformation is to transform the data towards normality or at least symmetry. In cases of high positive skewness normality might be achieved only for the middle 90% of the data. This transformation also has a variance stabilizing effect.

2) Use the estimates obtained from the standard components of variance analysis applied to the transformed nutrient intakes to partition the variance into between and within-subject components. Shrink the individual means using the formulae shown in the National Research Council or Hoffmann et al (2002a and b) sections of the literature review. The shrunken means have the mean and variance of the usual intake distribution on the transformed scale.

3) Back-transform the shrunken means to the original scale using the Taylor series approximation which includes a bias adjustment term as described in the Dodd et al (2006) section of the literature review. Bias adjustment is necessary since nonlinear transformation is applied to the observed intakes to transform them towards normality.

4) Estimate empirically the parameters of the usual intake distribution from the back-transformed shrunken means.

5.1.1.1 Power Transformation of the Data

The initial transformation towards normality is chosen from the Box-Cox (1964) family of transformations which could be defined as:

$$Y = p_\lambda(X) = \begin{cases} \left(X^\lambda - 1\right)/\lambda, & \lambda \neq 0 \\ \ln X, & \lambda = 0 \end{cases}$$

where X is a positive random variable and λ is a transformation parameter. It is assumed that there exists an λ such that Y has a normal distribution. However, as noted by Freeman and Modarres (2003) it could be shown that Y has an exact normal distribution only when $\lambda = 0$. For $\lambda \neq 0$ the domain of Y is not the entire real line. Hence, the researchers generally assume that Y has an approximate normal distribution.

For the protein intake data with four recalls per individual the optimal power transformation towards normality is square root as illustrated in Graph 11 below:

Graph 11: Box-Cox Transformation of Protein Intake

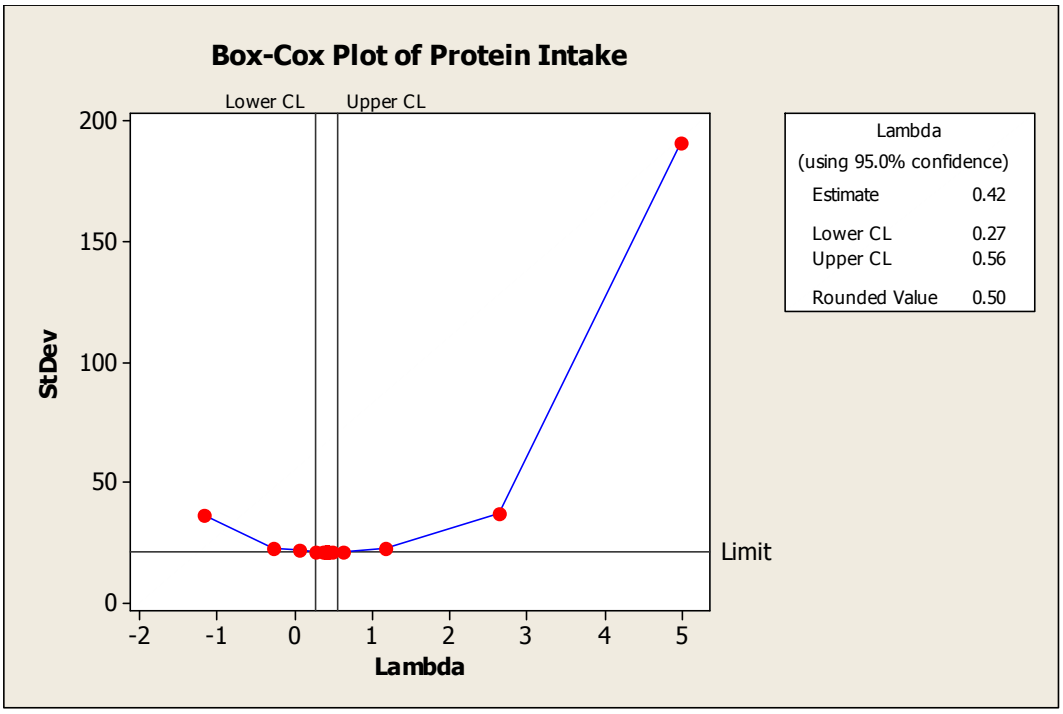


Table 39 illustrates that the simple one-stage power transformation to normality produces acceptable results for protein and carbohydrates intake. In such cases no further refinement of the transformation to normality is considered necessary and the application of the simplified power method could proceed with steps 2-4 described above.

Table 39: Simple Power Method. Anderson-Darling Test of Normality

Nutrient	A-D Statistics and p-values. Two recalls per Individual	A-D Statistics and p-values. Four recalls per Individual
Protein	0.498 (0.22)	0.505 (0.21)
Carbohydrates	0.769 (0.05)	0.673 (0.08)

Since the observed distributions of some nutrients such as vitamins exhibit high positive skewness and large deviations from normality, a two-stage transformation procedure might be applied in order to provide maximum flexibility in the transformation to approximate normality. Overly simplified models could produce inaccurate estimates of highly skewed distributions, especially of the tail percentiles. For this reason a second transformation could be applied to the power transformed observed nutrient intake in order to further refine and smooth the results from the initial power transformation.

5.1.2 Two-Stage Simplified Power Method Without Covariate Information

1) Power-transform the observed nutrient intake towards normality using a Box-Cox transformation. Since in most cases the nutrient intake data is highly positively skewed the goal of the transformation is to transform the data towards normality or at least symmetry. In cases of high positive skewness normality might be achieved only for the middle 90% of the data. This transformation also has a variance stabilizing effect.

2) Further refine and smooth the transformation to normality by using spline with knots or any other polynomial regression function that fits the data. Blom's (1958) normal scores are computed and used as a predictor variable in the smoothing procedure. Alternatively, the Blom's (1958) normal scores could be used as a dependent variable to transform the daily intakes to standard normal variables as described by Nusser et al (1996). In such cases the measurement error adjustments are performed on the $\mu_X = 0$ and $\sigma_X^2 = 1$ scale.

3) Use the estimates obtained from the standard components of variance formulae applied to the transformed nutrient intakes to partition the variance into between and within-subject components. Shrink the individual means using the formulae shown in the National Research Council or Hoffmann et al (2002) sections of the literature review. The shrunken means have the mean and variance of the usual intake distribution on the transformed scale.

4) Back-transform the shrunken means to the original scale using Taylor series approximation which includes a bias adjustment term as described in the Dodd et al (2006) section of the literature review. Bias adjustment is necessary since nonlinear transformation is applied to the observed intakes to transform them towards normality.

   In most cases the two-stage transformation is a complicated procedure and the back transformation function may not be explicitly expressed. Hence, the Taylor series approximation has to be estimated from the data. An approach similar to the method described by Nusser et al (1990) could be used to derive the Taylor series approximation and back transform the shrunken means to the original scale.

If we define the transformed towards normality observed dietary intakes as X_{ij} and the original observed dietary intakes as Y_{ij} then the transformation function which performs the transformation could be written as:

$$X_{ij} = g\left(Y_{ij}\right) \tag{116}$$

Hence, the inverse transformation $h(\ )$ could be defined as:

$$Y_{ij} = h\left(X_{ij}\right) \tag{117}$$

The back transformation $h(\ )$ is usually complicated and may not be explicitly expressed. Thus, the back transformation is derived using a Taylor series approximation over the measurement error distribution $h\left(x_i + u_{ij}\right)$ where x_i is the unobservable usual intake for the individual i and u_{ij} is the unobservable measurement error for individual i on day j. The Taylor series approximation requires expressions for $h\left(x_i\right)$ and $h''\left(x_i\right)$ to be estimated. Following Nusser et al (1990) approach $h\left(x_i\right)$ is estimated from the $\left(X_{ij}, Y_{ij}\right)$ pairs via grafted polynomial function with linear end segments. The second derivative $h''\left(x_i\right)$ is locally approximated for each individual by fitting a simple quadratic function. These estimators are used to construct the Taylor series approximation and convert the adjusted mean usual intakes to the original scale with the following formula:

$$y^* = \hat{h}\left(x_i^*\right) + 2^{-1}\hat{h}''\left(x_i^*\right)\hat{\sigma}_\varepsilon \tag{118}$$

where $\hat{h}\left(x_i\right)$ and $\hat{h}''\left(x_i\right)$ are the estimates of $h\left(x_i\right)$ and $h''\left(x_i\right)$, respectively, x_i^* is the adjusted mean usual intake for individual i and $\hat{\sigma}_\varepsilon^2$ is the estimated within-individual variance. The back transformation formula (118) is the same as formula (12) but the derivatives are approximated from the data. The second term on the right-hand side of equation (118) is the bias-adjustment term.

5) Estimate empirically the parameters of the usual intake distribution from the back-transformed shrunken means.

5.1.2.1 Selection of the Optimal Number and Location of Knots

Since complex spline functions are used to further refine the transformation to normality in small to moderate sized samples overfitting should be avoided. In general, any polynomial regression function which fits the data could be used to smooth the transformation to normality. The overfitting of the data could be avoided by using algorithms for selection of an optimal number and position of the knots used in the polynomial regressions.

There are a number of algorithms which could be applied to automatically select an optimal number and position of the knots. The selection of knots in polynomial regressions is investigated in the research published by Friedman (1991), Smith and Kohn (1996) and Hansen et al (2003), among others. Further examples of algorithms which could be used for knot selection are the Gauss-Newton or Stone and Koo (1985) algorithms. Ruppert et al (2003) document that the number of knots should be enough to model the essential structure in the underlying regression function without overfitting the data.

According to Ruppert et al (2003), a reasonable rule of thumb is to ensure that there are a fixed number of unique observations (at least 4 or 5) between each knot. For large data sets the maximum recommended number of knots is 20-40 in total.

Ruppert et al (2003) propose several algorithms for default knot selection which do not use any information in the data except the sample size. For example, default knot locations could be chosen using the following formulae:

$$k_k = \left(\frac{k+1}{K+2} \right) th\ sample\ quantile\ of\ unique\ x_i \text{ for } k = 1,.....K \tag{119}$$

or

$$K = \min\left(\frac{1}{4}\ number\ of\ unique\ x_i, 35 \right) \tag{120}$$

Ruppert et al (2003) also document some algorithms for automatic selection of the number of knots which use the information contained in the data. Those include the myopic and the full-search algorithms. The myopic algorithm searches a sequence of trial values of K and stops when there is no improvement in the generalized cross-validation (GCV) criterion. Only values of K which are less than $n_{unique} - p - 1$ are used in the search. In the algorithm n_{unique} denotes the number of unique x_i. The algorithm is called myopic because it does not continue the search beyond the value of K where it stops. The generalized cross-validation criterion is defined as:

$$GCV(\lambda) \equiv \frac{RSS(\lambda)}{\left\{1 - n^{-1} df_{fit}(\lambda)\right\}^2} \tag{121}$$

The full-search algorithm computes GCV minimized over λ searching the entire sequence of trial K values and uses the one which minimizes GCV. The main difference is that the myopic algorithm takes far less computational time but might stop the search for optimal knots prematurely.

5.1.2.2 Natural and Cubic Splines

The natural or restricted cubic spline is similar to the cubic splines but is linear beyond the boundary knots which is achieved by removing the quadratic and cubic terms. Hence, it requires only $n+2$ parameters, including the intercept:

$$Y = \beta_0 + \beta_1 x + \beta_2 (x - k_1)^3 + \beta_3 (x - k_2)^3 .....\beta_n (x - k_n)_+^3$$
$$where \tag{122}$$
$$(u)_+ = u, u > 0,\ 0 \text{ otherwise}$$

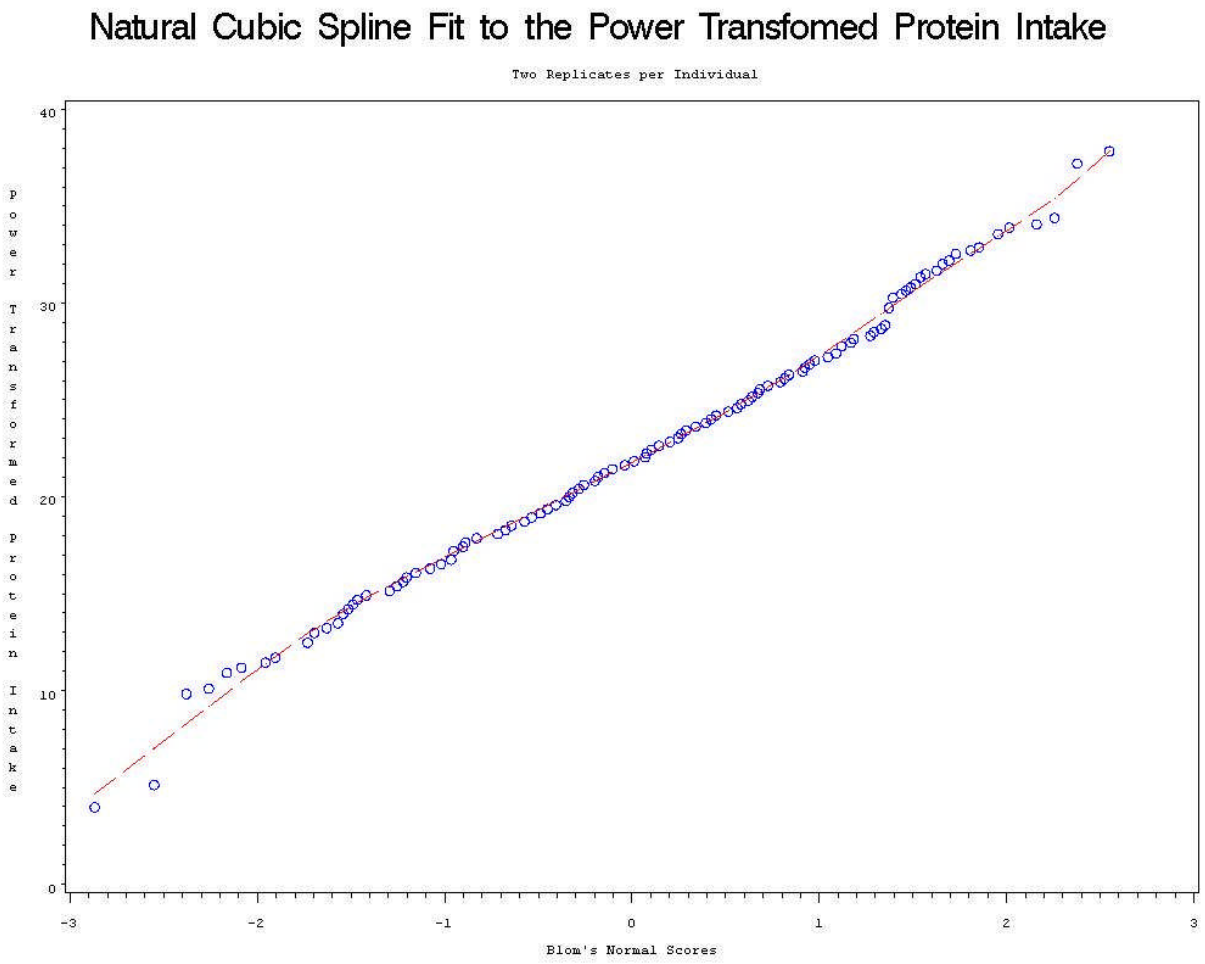
The cubic spline is defined as a set of third degree polynomials, joined at the knot points. At the knot points the polynomials are continuous in the second derivative which makes the spline very smooth:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - k_1)^3 + \beta_5 (x - k_2)^3 .....\beta_n (x - k_n)_+^3$$
$$where \tag{123}$$
$$(u)_+ = u, u > 0,\ 0 \text{ otherwise}$$

The cubic spline continues as a cubic polynomial beyond the range of data. If there are n knots the spline would require $n+4$ parameters, including the intercept. In the above equations Y is the power-transformed nutrient intake and k_n denotes the position of the knots.

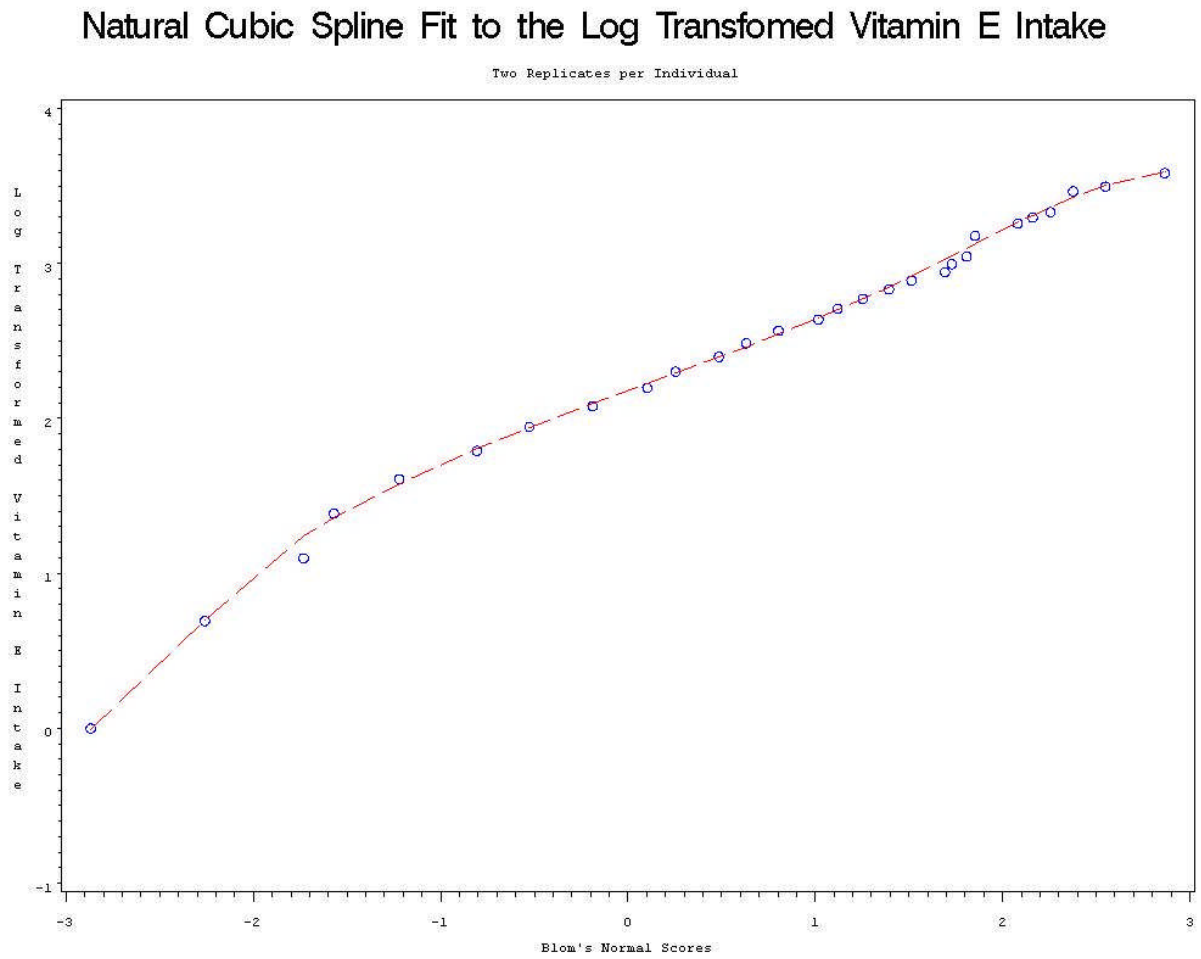
For the transformed protein and vitamin E intakes data with two replicates per individual natural splines with knots at the tail percentiles have a good fit since they adequately reveal the essential structure in the underlying data, especially in the tails of the distributions.

Graph 12: Natural Cubic Spline Fit to the Power Transformed Protein Intake



Graphs 12 and 13 illustrate that since the polynomials used in the transformations have continuous second derivatives and are constrained to be linear beyond the boundary points they might not pass through the extreme observations. Although all individual observations are transformed towards normality in practice the interest is focused on the normality of the individual means since their distribution is closely related to the distribution of usual intake.

Graph 13: Natural Cubic Spline Fit to the Log Transformed Vitamin E Intake



As noted by Tooze et al (2006) the use of covariates might improve the transformation to normality of the nutrient intake of interest and result in more accurate estimates of the tail percentiles of the usual intake distribution. Examples of such covariates include age, sex, education or race of the individuals in the sample. Unfortunately, the available data did not contain any covariate information which could be used for modeling of the long-term intake distribution. For those reasons one additional modification of the simplified power method is proposed but not investigated further in this research.

5.1.3 Simplified Power Method With Covariates Information

If covariate data are available an approach similar to the methods described by Slob (1993) and Waijers et al (2006) could be used to model the usual intake distribution. Both methods use age of the respondent as a covariate. The simplified power method could be modified as follows:

1) Transform the observed nutrient intake data towards normality using a Box-Cox transformation. If necessary apply two-stage transformation to the observed daily intakes in order to refine the transformation to normality.

2) Perform polynomial regression of the transformed daily intakes on the covariates. Any other regression function which fits the data could be used. For example, Waijers et al (2006) use fractional polynomial regression. The fractional polynomial regression could be described by the equations:

$$y_i = a + b\left(x_i\right)^p + c\left(x_i\right)^q + \varepsilon_i \left(i = 1, 2, ...n,\ p \neq q\right)$$

or (124)

$$y_i = a + b\left(x_i\right)^p + c\left(x_i\right)^q \ln\left(x\right) + \varepsilon_i \left(i = 1, 2, ...n,\ p = q\right)$$

where y_i is the transformed intake and x_i is the ith value of the covariate x.

3) Since we have repeated measurements data with at least two repeats per individual, refit the estimated fractional polynomial from point 2 above using a mixed effects model in order to obtain estimates of the inter and intra-individual variances. In this way, each person is treated as group with two or more observations. Equation (124) could be reformulated as:

$$y_{ij} = a + \gamma_i + b\left(x_i\right)^p + c\left(x_i\right)^q + \varepsilon_{ij} \quad (125)$$

where y_{ij} is the transformed intake for individual i on day j, $\gamma_i \sim N\left(0, \sigma^2\right)$ with σ^2 being the inter-individual variance and $\varepsilon_{ij} \sim N\left(0, \tau^2\right)$ with τ^2 denoting the intra-individual variance. The residuals ε_{ij} are assumed to be normally distributed with constant variance over the values of the covariate.

4) Simulate a large number of pseudo-persons using the estimates obtained from the mixed effects model shown in step 3. The simulated population should have the same characteristics in terms of covariates and between-person variability as the sample on which the mixed model is fit. Similar to the simulation described by Tooze et al (2006) at least 100 pseudo-persons for each person present in the sample are generated. For those pseudo-persons initial usual intake values are estimated using the mixed model intercept and covariate parameter estimates. In the next step a person-specific random effect drawn from the estimated between-individual variance distribution is added. The within-person variation is not included since by definition it does not contribute to the long-term nutrient intake.

5) Back-transform the simulated values to the original scale using Taylor series approximation with bias adjustment. The estimation of the Taylor series approximation depends on whether one or two stage transformation is used to transform the observed daily intakes to normality.

6) Estimate empirically the parameters of the usual intake distribution from the back-transformed individual mean intakes.

5.2 Empirical Comparison of the Methods for Estimating Usual intake Distributions

Table 40 : Comparison Between the Protein Usual Intake Distribution Estimated by Different Statistical Methods and the Observed Distribution of Individual 2-Day Means

Method	Data	P5	P10	P25	P50	P75	P90	P95	AM [(c)]	SD
2-day Means [(a) (b)]	2 recalls	46.00	55.50	61.50	76.00	94.00	107.00	114.5	78.62	21.47
Hoffmann et al Method (2002)	2 recalls	59.58	64.96	69.46	79.04	89.07	96.34	100.06	79.37	13.33
ISU Method (1996)	2 recalls	59.59	63.94	71.5	80.31	89.57	98.27	103.64	80.80	13.40
Chang et al Method (2001)	2 recalls	57.00	61.00	68.50	79.00	88.50	97.50	104.00	78.62	14.35
NCI Method (2006)	2 recalls	57.17	61.61	69.29	78.23	87.55	96.23	101.49	78.64	13.50
Simplified Power Method (One Stage)	2 recalls	58.42	64.03	68.11	77.49	88.34	96.10	100.49	78.62	13.29
Simplified Power Method (Two Stages)	2 recalls	57.24	62.41	67.43	76.95	88.54	96.03	100.82	78.08	13.75

(a) Protein intake is measured in grams
(b) Estimates obtained from two recalls per individual
(c) Arithmetic mean

Table 41 : Comparison Between the Carbohydrates Usual Intake Distribution Estimated by Different Statistical Methods and the Observed Distribution of Individual 2-Day Means

Method	Data	P5	P10	P25	P50	P75	P90	P95	AM [(c)]	SD
2-day Means [(a) (b)]	2 recalls	151.50	179.50	217.00	272.50	330.00	378.00	407.5	276.39	78.94
Hoffmann et al Method (2002)	2 recalls	174.15	192.16	234.99	278.20	321.67	356.33	379.74	277.89	61.04
ISU Method (1996)	2 recalls	183.13	202.12	234.77	271.45	308.29	343.84	367.54	272.83	56.40
Chang et al Method (2001)	2 recalls	197.00	208.00	237.00	273.00	312.00	352.00	375.00	276.39	55.85
NCI Method (2006)	2 recalls	183.21	201.72	234.34	273.61	315.30	355.29	380.48	276.53	60.06
Simplified Power Method	2 recalls	176.29	195.44	231.96	275.02	318.07	352.98	377.12	276.40	60.12

(a) Carbohydrates intake is measured in grams
(b) Estimates obtained from two recalls per individual
(c) Arithmetic mean

Table 42: Comparison Between the Vitamin E Usual Intake Distribution Estimated by Different Statistical Methods and the Observed Distribution of Individual 2-Day Means

Method	Data	P5	P10	P25	P50	P75	P90	P95	AM [(c)]	SD
2-day Means [(a) (b)]	2 recalls	4.00	4.50	6.50	9.00	12.00	14.00	18.00	9.40	4.21
Hoffmann et al Method (2002)	2 recalls	5.88	6.20	7.51	9.14	11.10	12.40	15.01	9.40	2.75
ISU Method (1996)	2 recalls	5.09	5.85	7.25	9.06	11.26	13.73	15.5	9.51	3.26
Chang et al Method (2001)	2 recalls	5.50	6.00	7.50	9.50	11.50	13.50	16.50	9.40	2.86
NCI Method (2006)	2 recalls	4.92	5.65	7.06	8.97	11.27	13.73	15.41	9.41	3.26
Simplified Power Method (One Stage)	2 recalls	4.55	5.56	7.28	9.06	11.53	13.09	15.11	9.43	3.22
Simplified Power Method (Two Stages)	2 recalls	4.35	5.21	7.11	9.18	11.54	13.79	14.92	9.31	3.23

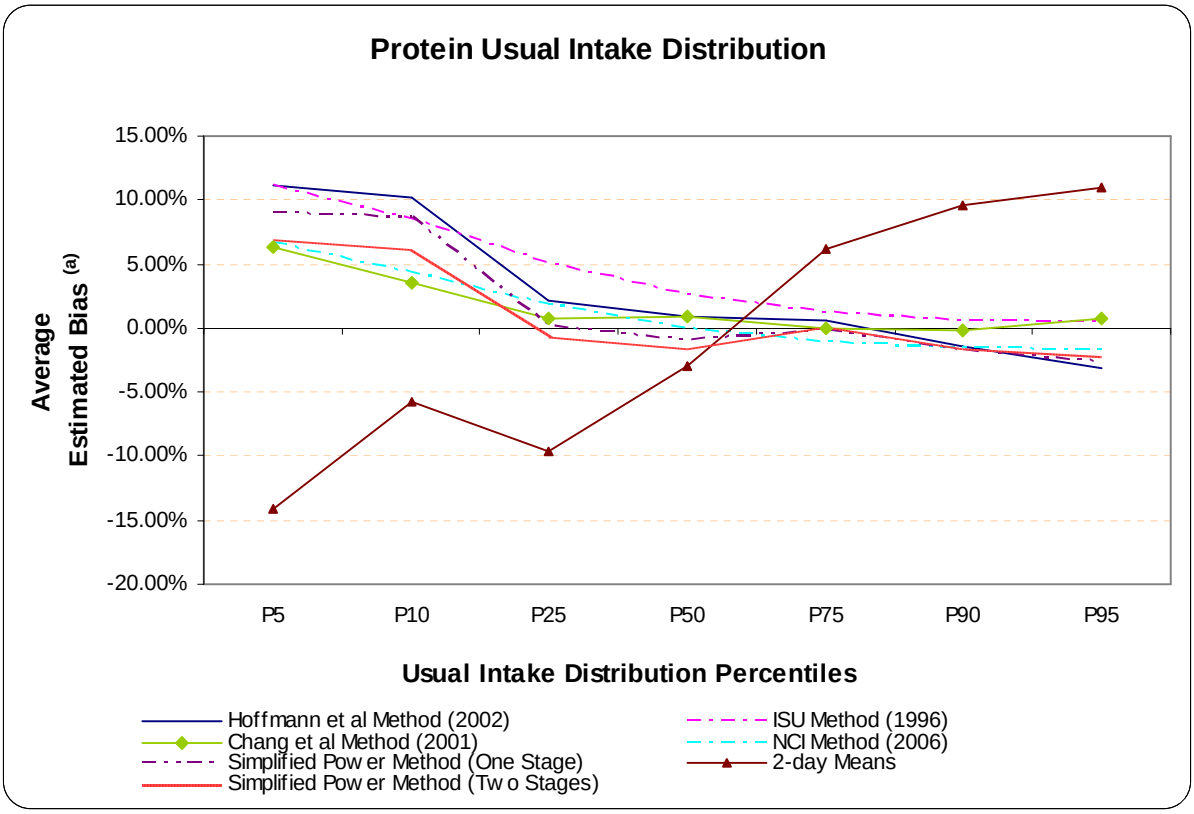
(a) Vitamin E intake is measured in milligrams
(b) Estimates obtained from two recalls per individual
(c) Arithmetic mean

Tables 40-42 above document the observed sample distributions of the individual means for 2 and 16 recalls per individual. The standard deviations decrease as the number of sampling days is increased but the arithmetic mean remains similar over the different data sets. In most cases this is accompanied by an increase in the lower percentiles (p5, p10) and decrease in the upper percentiles (p90 and p95).

In the lower part of the tables the distributions of usual protein, carbohydrates and Vitamin E intakes are estimated using five different methods and only two repeats per individual. The estimated distributions are somewhat closer to the observed distributions of the 16-day means. Furthermore, the estimated usual intake distributions are close to each other. Those estimates could be formally considered estimates of the sample distributions of the 365-day means.

Graphs 14-16 illustrate the comparison between the bias in the estimates obtained using the five methods investigated in this research and the naïve estimates obtained from two-sample days means. The bias is estimated as the difference between the percentile estimates obtained with the different methods and observed percentiles from an assumed true usual intake distribution generated using Monte-Carlo simulations. The simulated distribution is based on distribution parameters obtained from 16 recalls per individual adjusted for measurement error.

Graph 14: Protein Usual Intake Distribution. Average Estimated Bias

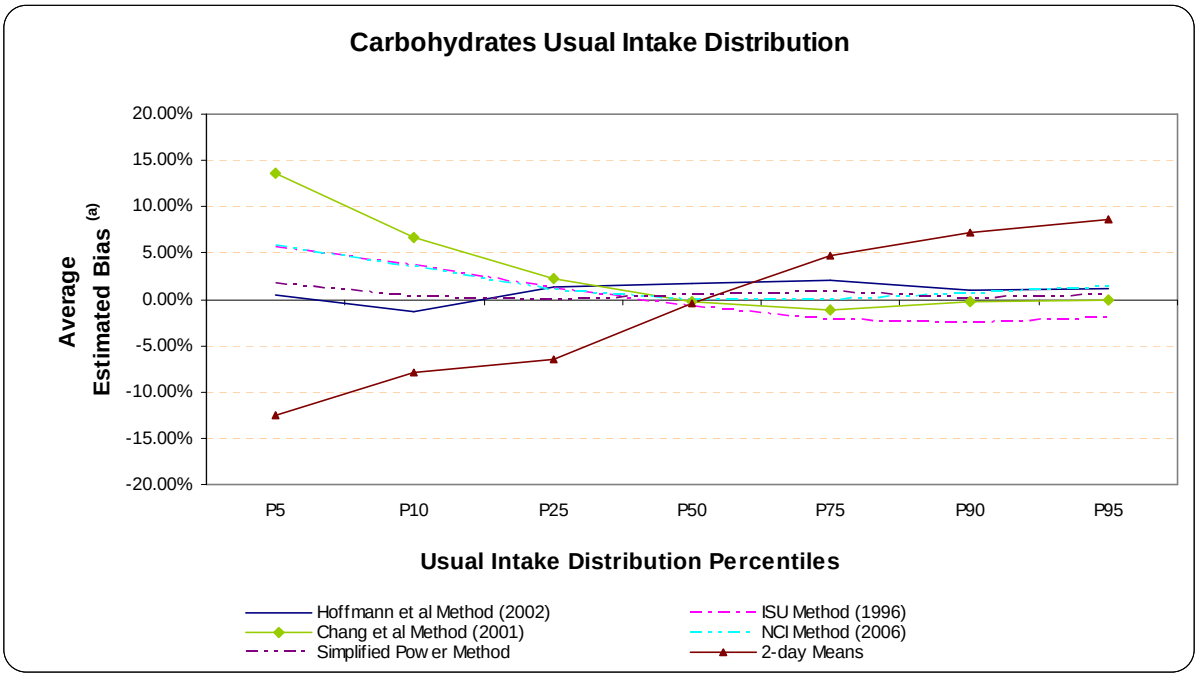


(a) Bias = (estimated percentile-observed percentile)/observed percentile

Graphs 14-16 document that the naïve estimates obtained from the two-sample days' means are very biased, especially in the tails of the distribution as a result of the presence of within-person variation. As shown in Graph 14 the Hoffmann et al (2002) and the simplified power methods provide similar estimates for protein usual intake. The same graph illustrates that complex procedures such as the ISU (1996) and NCI (2006) methods provide more accurate estimates of the usual intake distribution, especially of the upper tail percentiles of the distribution. The Chang et al (2001) method also shows very good performance in terms of estimates of the upper tails of the distribution. The two stage simplified power method produces less biased estimate of the lower percentile (p5) and only marginally better estimate of the upper percentile (p95) of the distribution when compared to the one stage version.

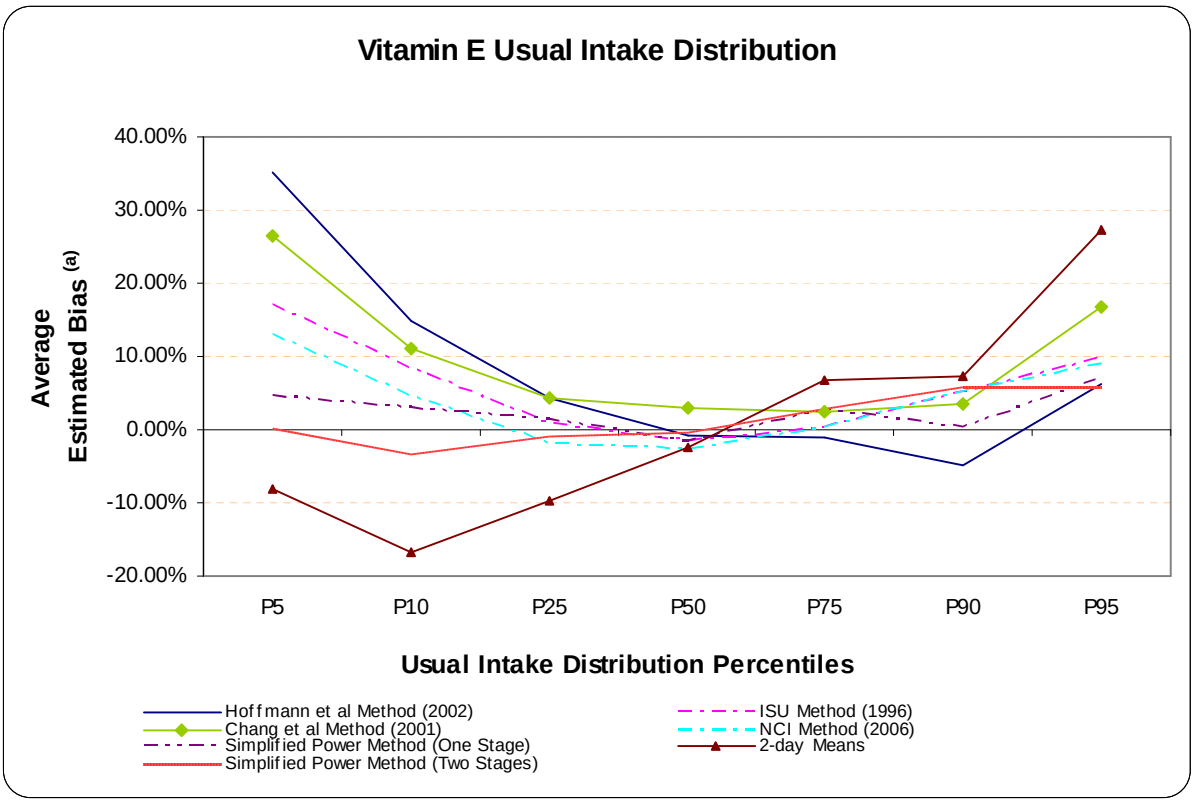
Graph 15 shows the results for the usual intake distribution of carbohydrates. The two simple procedures – Hoffmann et al (2002) and simplified power methods show very good performance which is fully comparable to the performance of the more complex methods. It could be argued that the least biased estimates are obtained with the simplified power method. Similar to the usual intake distribution of protein, the Chang et al (2001) method produces estimates of the lower percentiles of the distribution (p5,p10) with relatively large bias but very accurate estimates of the upper percentiles (p90, p95) of the distribution with almost no bias. The performance of the ISU (1996) and NCI (2006) methods is very similar, although the NCI (2006) method provides slightly better estimates of the upper percentiles (p90, p95) of the usual intake distribution of carbohydrates. In general, for the three nutrients investigated in this research the most accurate estimates are obtained for the usual intake distribution of carbohydrates. This might indicate that the carbohydrates intake has the smallest within-person variation compared to the protein and vitamin E daily intake distributions.

Graph 15: Carbohydrates Usual Intake Distribution. Average Estimated Bias



(a) Bias = (estimated percentile-observed percentile)/observed percentile

Graph 16: Vitamin E Usual Intake Distribution. Average Estimated Bias



(a) Bias = (estimated percentile-observed percentile)/observed percentile

Graph 16 illustrates that when compared to the one stage version, the two stage version of the best power method provides less biased estimate of the lower tail percentile (p5) of the vitamin E distribution and only marginally better estimate of the upper tail (p95) of the vitamin E distribution. In general, the one stage simplified power method shows very good performance even when applied to highly skewed distributions which could be interpreted as a robustness property of the method. Furthermore, on the vitamin E data the one stage simplified power method shows better performance when compared to the Hoffmann et al (2002) method as well as ISU (1996) and NCI (2006) methods. One possible explanation could be the fact that the complex methods are applied to a relatively moderate sized sample containing only 151 individuals with two repeated measurement per individual. Since the complex methods involve advanced data transformation and estimation procedures they are more suitable for large sized samples. Large bias is observed in the estimates of the lower tail of the long-term Vitamin E intake distribution obtained with the Chang et al (2001) and Hoffmann et al (2002) methods.

One of the most important survey design issues is the number of dietary records collected per individual. The results support the evidence documented by Gay (2000), Volatier et al (2002) and Hoffmann et al (2002a) that the methods could be successfully applied to data containing only two repeats per individual. However, the sample used to estimate the usual intake distribution should be representative of the period of interest with an equal distribution of seasons and days of the week. Furthermore, as noted by Hoffmann et al (2002a), two repeats per individual might be sufficient to produce reliable estimates of the usual intake distribution but if there are other objectives in the study such as ranking of individuals by long-term intake then a higher number of daily measurements is required. The authors give some general recommendations based on previous research documenting that minimum 3 to 10 daily measurements are required for the estimation of individual usual intake for energy and macronutrients whereas for food components with large day-to-day variation such as vitamins minimum 20-50 daily measurements are needed.

Buck et al (1995) noted that the minimum number of two repeated measurements per individual will be sufficient to minimize the bias in the estimates of the distribution percentiles if the assumption that the individuals in the sample have the same distribution of the measurement error holds. Both the shape and variance of the measurement error distribution affect the estimation of the long-term intake distribution. In general, the bias in the estimates of the distribution percentiles depends on the shapes of the long-term daily intake and measurement error distributions as determined by their skewness and kurtosis rather than on the families from which the distributions come. This conclusion is supported by Hoffmann et al (2002a) who wrote that the Buck (1995) method could be applied to asymmetric distributions in cases where intra and inter-individual variations have similar skewness and kurtosis values. It seems that this conclusion could be extended to the simplified methods for estimating usual intake distributions investigated in this research since they are similar to the Buck et al (1995) method.

The ISU (1996), Hoffmann et al (2002) and Simplified Power methods are suitable for modeling usual nutrient intake where the distributions of the single-day intakes can be transformed at least approximately to normality. Hence, these methods could be used to estimate the long-term intake distributions of most nutrients and commonly consumed food groups.

However, special foods, food groups or nutrients which are consumed rarely or only by part of the population have large proportion of days with zero intake. The distributions of such intakes have clumps of zero observations in the left tail and it is almost impossible to transform such distributions towards normality. Dodd et al (2006) note that the normality for such distributions is not attainable since the transformations towards normality preserve the clump at zero. Hence, improved methods need to be used in order to estimate the usual intake distributions of episodically consumed foods and nutrients.

Two such methods which explicitly account for clumping at zero are discussed in detail in the literature review section. The Iowa State University Method for Episodically Consumed Dietary Components (ISUF 1997) was developed by Nusser et al as an extention of the ISU method (1996). Nusser et al (1997) recommended a procedure where at first the distribution of food consumption probabilities is estimated from the observed numbers of zero and non-zero intakes. Secondly, the usual intake distribution for consumption days only is estimated from the observed positive intake values. This eliminates the problem of transforming towards normality nutrient intake data with large proportion of zero values. In the third step of the method the usual intake distribution for all days is estimated from the joint distribution of consumption day usual intake and individual consumption probabilities. The procedure assumes that the amount of consumed food is independent of the frequency with which the food is consumed. The major drawback of such method is that it does not allow for correlation between the consumption probability and the amount of food consumed. As noted by Dodd et al (2006), in such cases the ISUF (1997) method may introduce bias by overestimating the amount consumed by individuals with estimated low probability of consumption and underestimating the amount consumed by individuals with estimated high probability of consumption.

The NCI (2006) method is particularly suitable for estimating the usual intake distributions of episodically consumed foods and nutrients since it separates the probability of consumption from the consumption-day amount using a two-part mixed effects model. The NCI procedure allows for correlation between the probability of consumption on a single day and the consumption-day amount. Furthermore, covariates could be used to model both daily consumption probability and the consumption-day amount.

The simpler statistical methods suitable for small to medium sized samples investigated in this research could also be modified in order to model the usual intake distributions of trace elements and other food items with occasional rather than regular consumption. For example, if the correlation between probability of consumption on a single day and consumption-day amount is low then the Hoffmann et al (2002) method could be applied to the observations with positive consumption only. In such cases the estimation of the usual intake distribution will be based on pooled data of the back transformed usual intakes of the consumers and the zero usual intakes of the non-consumers.

Based on the above results no conclusive evidence that complex procedures such as ISU (1996) and NCI (2006) methods systematically produce more accurate estimates of the percentiles of the usual intake distribution compared to simpler methods was found. Possible reasons include the simple survey design used to collect the data, the small sample sizes as well as the lack of covariates in the data used in the research. From the two complex methods the NCI (2006) method seems to produce more accurate estimates when compared to the ISU (1996) method. For the NCI (2006) and ISU (1996) methods the trend in the estimates of the distribution percentiles, and respectively the bias, seem similar. For the two complex methods the bias is larger in the estimates of the lower tails of the distributions (p5, p10) and smaller in the estimates of the upper tails (p90, p95) of the distributions.

For the protein data the average bias in the estimates of the distribution percentiles obtained with the one stage simplified power method is 3.3%. When the two stage version is applied the average bias is reduced to 2.8%. For the vitamin E data the same figures are 2.9% and 2.7%, respectively. However, the two stage version is more complex and requires an estimation of the approximation of the back-transformation function. Hence, the marginal increase in the accuracy of the estimated usual intake distribution parameters should be weighed against an increase in the complexity of the modeling process. In general, the one-stage simplified power method seems to have a robustness property since it produces relatively accurate results even when applied to the non-normal distribution of vitamin E daily intake.

The two simple methods – Hoffmann et al (2002) and the one stage simplified power method produce similar estimates for protein and carbohydrates usual intake distributions. For the vitamin E usual intake distribution the simplified power method produces more accurate estimates. Hence, there is some evidence that the back transformation formula proposed by Hoffmann et al (2002a) holds for symmetric but moderately non-normal distributions. However, it needs to be investigated further if the back transformation formula could be applied to highly skewed daily intake distributions or cases where the measurement error distribution is highly non-normal. If even in such cases the Hoffmann et al (2002) method produces relatively accurate estimates of the usual intake distribution parameters then it should be interpreted as a robustness property of the method.

In general, the estimates of the usual intake distributions of protein and carbohydrates derived by the Chang et al (2001) method are compatible to the estimates obtained by the other four methods. The Chang et al (2001) method produces estimates of the lower percentiles (p5, p10) with relatively large bias but the estimates of the upper percentiles (p90, p95) of the usual intake distributions of protein and carbohydrates are very accurate. The estimates of the usual intake distribution of vitamin E have relatively large bias in both the lower and upper tails of the distribution. One possible explanation of the observed results could be the fact that the Chang et al (2001) method is based on a parametric assumption. To derive the results shown in Graphs 14-16 an overdispersed gamma distribution is assumed. However, for some of the nutrients this might not be an optimal distribution assumption and other distributions might produce more accurate results. Furthermore, the estimates of the ratio of within to between-individual variance obtained from external datasets have a large influence on the results derived with the Chang et al (2001) method. Hence, if the estimated ratio is not representative of the population of interest the long-term intake distribution estimates obtained with the Chang et al (2001) method will be biased.

The Chang et al (2001) method is considered an alternative to the mixture distribution framework. Instead of modeling the nutrient intake data as a mixture of distributions an additional scale parameter is added to the exponential family form to account for overdispersion.

Future research could further investigate the use of mixture distributions for modeling long-term dietary intake. A mixture distribution is composed of several statistical distributions and arises as a result of inhomogeneous population consisting of various subpopulations with different probability density functions in each subpopulation. A finite mixture has a finite number of components. As noted by Bohning and Seidel (2003) the popularity of the mixture distributions is due to the fact that they offer natural models for the unobserved population heterogeneity.

When modeling long-term nutrient intake the standard assumptions are that the population is homogenous which implies that single density could be used to model the distribution of the usual intake of the nutrient. However, the extreme skewness of some of the nutrient intake distributions might indicate that those standard assumptions are violated. Hence, there could be subpopulations with distinctly different nutrient intake compared to the rest of the population. For example, the consumption of some nutrients might vary considerably within ethnic or age groups. Furthermore, dietary intake may have a normal or lognormal distribution within subpopulations but not when they are all put together. Mixture models could be used to model such complexities.

The mixture density has a density function which is a sum of the density functions specific to the subpopulations and could be expressed as:

$$g(x) = \pi_1 f_1(x) + \pi_2 f_2(x) + ... + \pi_k f_k(x) \qquad (126)$$

where $g(x)$ is the mixture density, $f_k(x)$ is the density specific to the $k\,th$ subpopulation and π_k are the mixing weights or mixing proportions for each subpopulation. The mixing weights are proportional to the subpopulation size.

Another suggestion for future research is the use of nonparametric and semiparametric algorithms to estimate the usual intake distributions. Their empirical performance has not yet been investigated in detail in the research literature. These methods assume that the distribution of the variable measured with error is a convolution of the distribution functions of the unobserved long-term intake and the measurement error. Hence, the estimation of the usual intake distribution is considered a deconvolution problem. Kernels and nonparametric maximum likelihood estimators have been proposed by many authors. Some examples include the work of Diggle and Hall (1993), Stefanski et al (1990) and Goutis (1997), among others. Other examples include the work of Chen et al (2003) who suggested semiparametric spline estimators of the density of a variable measured with error and the recent research by Staudenmayer at al (2007) who proposed a method which uses penalized mixture of B-splines to model the long-term dietary intake distributions.

6. Conclusion

The results from this study support the evidence documented by previous research that the dietary intake data have large within-person variance which affects the ability to precisely estimate the centre and the extremes of the long-term intake distribution. It has been shown that it is possible to draw valid statistical inferences from dietary intake data containing as little as two daily intakes per individual. However, it should be noted that with such samples it is likely to obtain more accuare estimates of the long-term intake of nutrients which are consumed more or less regularly compared to the estimates of the distributions of the episodically consumed nutrients or trace elements. The episodically consumed foods or nutriens contain a large proportion of zero intakes and for such data the recommendation is to use the NCI (2006) or ISUF (1997) methods and samples with higher number of repeated measurements per individual. Furthermore, as noted by Hoffmann et al (2002a) two daily intakes per individual are sufficient only for the estimation of the usual intake distribution. If there are other research objectives then higher number of daily repeated measurements per individual are required. Furthermore, Hoffmann et al (2002a) document that the sampling days should cover all days of the week and all seasons of the year.

The methods investigated in this research assume that the main objective of the study is to estimate the distribution of usual intake in a population. Five methods for estimating usual intake distributions are examined in detail. The results indicate that the methods could be applied to a broad class of symmetrical or skewed distributions and allow the estimation of the unknown parameters of the population usual intake distribution.

Various enhancements have been included in the methods in order to address a number of important issues that might be encountered in the analysis of dietary intakes data. Some of the most important enhancements include:

1) Ability to estimate the distribution of usual intake from data collected with complex sampling designs containing sampling weights.

2) Adjustments for non-normal distribution of the individual intakes.

3) Adjustments for nuisance effects.

4) Adjustments for heterogeneous within-individual variance.

5) Adjustments for correlation among repeated measurements.

6) Ability to estimate usual intake distribution of special foods and nutrients which are consumed rarely or only by part of the population. Such data contain large proportion of zero consumption.

7) Ability to allow for correlation between the propensity to consume specific food or nutrient and the amount consumed.

8) Ability to use covariate information to model the long-term intake distribution.

Simpler procedures such as the Hoffmann et al (2002) and the simplified power methods could be used for small to moderate sized samples of 100-400 individuals. Most often such samples do not contain sampling weights, nuisance effects or any other features that might require the application of more complicated procedures. The simpler methods are easier to implement and nutrition researchers who do not have strong statistical background could be able to use them and make statistical inferences without the help of a statistician.

If the nutrient intake samples are large, collected with complex sampling methods and various adjustments for nuisance effects or correlated daily intake measurements are needed then more advanced procedures such as the ISU (1996) method have to be used. However, this method could be implemented only if the SIDE or C-SIDE software packages are available.

The NCI method (2006) is a new development in the field. It is a very flexible procedure and represents an enhancement of the ISU (1996) method. In contrast to the ISU (1996) method it allows the use of covariates to model the usual intake distribution. The NCI (2006) method is particularly suitable for the analysis of special foods and nutrients which are not consumed daily or are consumed by only some segments of the population. Such data contain large proportion of zero dietary intake consumption. Furthermore, the NCI (2006) method allows the probability of consumption and amount consumed to be modeled separately. A correlation between the probability to consume a nutrient on a single day and the consumption day amount could also be included in the analysis. The method is implemented using a set of SAS macros and hence does not require the purchase of a specialized software.

The application of the simplified power method illustrated that some of the complex procedures could be simplified along the lines of the transformation to normality. The further refinement of the simple power transformation to normality by using polynomial regressions with knots resulted in some improvement in the accuracy of the estimates of the percentiles of the usual intake distribution.

7. Bibliography

1. Albert J., Pepple P., (1989). “A Baysian approach to some overdisperion models.” Canadian Journal of Statistics 17:333-344.

2. Anderson, T.W. and Darling D. A., (1952). “Asymptotic theory of certain ‘goodness-of-fit’ criteria based on stochastic processes” Annals of Mathematical Statistics, 23:193-212.

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

4. Beaton G.H., Milner J., Corey P., McGuire V., Cousins M., Stewart E., de Ramos M., Hewitt D., Grambsch P.V., Kassim N,. Little J.A.,(1979). “Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation.” Am. J. Clin. Nutr. 1979;32: 2546-2559.

5. Beaton G.H., Milner J., McGuire V., Feather T.E., Little J.A., (1983). “Source of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. Carbohydrate sources, vitamins, and minerals.” Am. J. Clin. Nutr. 1983;37:986-995.

6. Beaton G.H., (1994). “Criteria of an adequate diet”. In: Shils R.E., Olson J.A., Shike M., eds. Modern Nutrition in Health and Disease, 8th edition. Philadelphia: Lea & Febiger. p. 1491-1505.

7. Bingham S.A., (1987). “The dietary assessment of individuals; methods, accuracy, new techniques and recommendations.” Nutr. Abstr. Rev. 57, 705-742.

8. Biro G., Hulshof K., Ovesen L.and Amorim Cruz J.A., (2002). “Selection of methodology to assess food intake” European Journal of Clinical Nutrition 56:Suppl2, S25-S32.

9. Blom G., (1958). “Statistical estimates and transformed beta variables.” New York: John Wiley& Sons. Inc.

10. Box G.and Cox D.R., (1964). “An analysis of transformations.” J. R. Stat.Soc. Ser. B 26, 211-252.

11. Bohning D., Seidel W., (2003).”Recent developments in mixture models.” Computational Statistics and Data Analysis, 41:349-357.

12. Briefel R.R,. Flegal K.M.,Winn D.M., Loria C.M., Johnson C.L., and Sempos C.T., (1992). “Assessing the nation’s diet: limitations of the food frequency questionnaire.” J. Am. Diet. Assoc. 92, 959-962.

13. Brussaard J.H., Lowik M.H., Steingrimsdottir L.,Møller A., Kearney J., De Henauw S. and Becker W., (2002). “A European food consumption survey method-conclusions and recommendations.” European Journal of Clinical Nutrition 56: Suppl 2, S89-S93.

14. Buck R.J., Hammerstrom K.A. and Ryan P.B., (1995). “Estimating long-term exposures from short-term measurements.” J. Expos. Anal. Environ. Epidemiol. 5, 359-373.

15. Buzzard M., (1998). “24-hour dietary recall and food record methods.” In Nutritional Epidemiology, 2nd ed. pp 50-73. New York: Oxford University Press.

16. Carriquiry, A.L., Fuller W., Goyeneche J. and Jensen H., (1995). “Estimated correlations among days for the combined 1989-91CSFI”. Dietary Assessment Research Series 4. CARD Staff Report 95-SR 77. Center for Agricultural and Rural Development, Iowa State University, Ames.

17. Carriquiry A.L., (2003). “Estimation of usual intake distributions of nutrients and foods.” Journal of Nutrition 133:601S-608S.

18. Chang H., Suchindran C., Pan W., (2001). “Using the overdispersed exponential family to estimate the distribution of usual daily intakes of people aged between 18 and 28 in Taiwan.” Statist.Med.20:2337-2350.

19. Chen C., Fuller W., Breidt F.J., (2003). “Spline estimators of the density function of a variable measured with error.” Communications in Statistics-Simulation and Computation, 32:1, 73-86.

20. Commenges D., Jacqmin-Gadda H., (1997). “Generalized score test of homogeneity based on correlated random effects models.” Journal of the Royal Statistical Society, Series B 59:157-171.

21. Dean B., (1992). “Testing for overdispersion in Poisson and binomial regression models.” Journal of the American Statistical Association 87:451-457.

22. Donner A., (1986). “A review of inference procedures for the intraclass correlation coefficient in the one-way random effect model.” Inst.Stat.Rev. 54, 67-82.

23. Dodd, K.W., Guenther P.M., Freedman L.S., Subar A.F.,Kipnis V., Midthune D.,Tooze J. A., Krebs-Smith S. M., (2006). “Statistical methods for estimating usual intake of nutrients and foods: A review of the theory.” J. Am. Diet. Assoc. 2006;106:1640-1650.

24. Diggle P. J., Hall P., (1993). “A fourier approach to nonparametric deconvolution of a density estimate.” Journal of the RoyalStatistical Society Ser. B 55:523-531.

25. Efron B., (1986). “Double exponential families and their use in generalized linear regression.” Journal of the American Statistical Association 81:709-721.

26. Flegal K.M., Larkin F.A.,(1990). “Partitioning macronutrient intake estimates from a food frequency questionnaire.” Am. J. Epidemiol. 1990:131:1046-1058.

27. Freedman L.S., Midthune D., Carroll R.J., Krebs-Smith S., Subar A.F., Troiano R.P., Dodd K., Schatzkin A., Ferrari P., Kipnis V.,(2004). “Adjustments to improve the estimation of usual dietary intake distributions in the population.” J. Nutr. 2004;134:1836-1843.

28. Freeman J., Modarres R., (2003). “Analysis of censored environmental data with Box-Cox transformations.” Technical Report, U.S. EPA.

29. Friedman J.H., (1991). “Multivariate adaptive regression splines (with discussion).” Annals of Statistics 19:1-141.

30. Fuller W A., (1987). “Measurement error models.” New York: John Wiley&Sons.

31. Ganio M., Schafer W., (1992). “Diagnostics for overdispersion.” Journal of the American Statistical Association 87:795-804.

32. Gay C., (2000). “Estimation of population distributions of habitual nutrient intake based on a short-run weighed food diary.” Br. J. Nutr. 83, 287-293.

33. Goutis, C., (1997). “Nonparametric estimation of a mixing density via the kernel method.” Journal of the American Statistical Association 92:1445-1450.

34. Guenther P.M., Kott P.S. and Carriquiry A.L., (1997). “Development of an approach for estimating usual nutrient intake distributions at the population level.” J. Nutr. 127, 1106-1112.

35. Hansen M.H., Huang Z., Koopenberg C., Stone C.J., Truong Y.K.,(2003).”Statistical modeling with spline functions: Methodology and theory.” New York: Springer-Verlag.

36. Hartman A.M., Brown C.C., Palmgren J., Pietinen P., Verkasalo M., Myer D. and Virtamo J., (1990). “Variability in nutrient and food intakes among older middle-aged men.” Am. J. Epidemiol. 132, 999 -1012.

37. Hartung J., (1981). “Non-negative minimum biased invariant estimation in variance component models.” Ann. Stat. 9:278-292.

38. Hoffmann K., Boeing H., Dufour A., Volatier JL., Telman J., Virtanen M., Becker W., Henauw S., (2002a). “Estimating the distribution of usual dietary intake by short-term measurements.” European Journal of Clinical Nutrition 56: Suppl 2:S53-S62.

39. Hoffmann K., Kroke A., Klipstein-Grobush K., Boeing H., (2002b). “Standardization of dietary intake measurements by nonlinear calibration using short-term reference data.” American Journal of Epidemiology 156:862-870.

40. Jahns L., Arab L., Carriquiry A. and Popkin B., (2004).”The use of external within-person variance estimates to adjust nutrient intake distributions over time and across populations.” Public Health Nutrition, 8(1), 69-76.

41. Karkeck J.M., (1987): “Improving the use of dietary survey methodology.” J. Am. Diet. Assoc. 87, 869- 871.

42. Kipnis V., Subar A.F., Midthune D., Freedman L.S., Ballard-Barbash R., Troiano R.P., Bingham S., Schoeller D.A., Schatzkin A., Carroll R.J.,(2003). “Structure of dietary measurement error: results of the OPEN biomarker study.” Am. J. Epidemiol. 2003;158:14-21.

43. Kipnis V.,Midthune D., Freedman L.S., Bingham S., Schatzkin A., Subar A. and Carroll R.J., (2001). “Empirical evidence of correlated biases in dietary assessment instruments and its implications.” Am. J. Epidemiol.153, 394- 403.

44. Kroke A., Klipstein-Grobusch K., Voss S.,Moseneder J., Thielecke F., Noack R. and Boeing H., (1999). “Validation of a self-administered food-frequency questionnaire administered in the European prospective investigation into cancer and nutrition (EPIC) study: comparison of energy, protein, and macronutrient intakes estimated with the doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods.” Am. J. Clin. Nutr. 70, 439-447.

45. Kushi L.H., (1994). “Gaps in epidemiologic research methods: design considerations for studies that use food-frequency questionnaires.” Am. J. Clin. Nutr. 59(Suppl), 180S-184S.

46. Lambe J. and Kearney J., (1999). “The influence of survey duration on estimates of food intakes-relevance for food-based dietary guidelines.” Br. J. Nutr. 81: Suppl 2, S139-S142.

47. Lambe J., Kearney J., Leclercq C., Zunft H.F., De Henauw S., Lamberg-Allardt C.J., Dunne A. and Gibney M.J., (2000). “The influence of survey duration on estimates of food intakes and its relevance for public health nutrition and food safety issues.” European Journal of Clinical Nutrition 54, 166-173.

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

49. Liu K., (1994). “Statistical issues related to semiquantitative food frequency questionnaires.” Am. J. Clin. Nutr. 59: Suppl, 262S-265S.

50. Liu K., Stamler J., Dyer A., McKeever J., McKeever P., (1978). ”Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol.” Journal of Chronic Disease 31:399-418.

51. Lowik M.R.H., Hulshof K., Brussaard J.H. and Kistemaker C., (1999). “Dependence of dietary intake estimates on the time frame of assessment.” Regul. Toxicol. Pharmac. 30: Suppl, S48-S56.

52. Martin L.J., Su W., Jones P.J., Lockwood G.A., Tritchler D.L. and Boyd N.F., (1996). “Comparison of energy intakes determined by food records and doubly labeled water in women participating in a dietary intervention trial.” Am. J. Clin. Nutr. 63, 483-490.

53. National Academy of Sciences (2003). “Adjustment of observed intake data to estimate the distribution of usual intakes in a group.” Washington, DC. National Academy Press.

54. National Academy of Sciences (2003). “Dietary reference intakes: applications in dietary planning”,http://www.nap.eclu/openhook/0309088534/htin1/208.html.

55. Nelson M., Black A.E., Morris J.A. and Cole T.J., (1989). “Between- and within subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision.” Am.J. Clin.Nutr. 50, 155-167.

56. NRC (1986). “Nutrient adequacy: assessment using food consumption surveys.” Washington, DC. National Academy Press.

57. Nusser S.M., Carriquiry A.L., Jensen H.H., Fuller W.A., (1990). “A transfomation approach to estimating usual intake distributions.” Research Report, Center for Agricultural and Rural Development, Iowa State University, Ames, Iowa 50011.

58. Nusser S.M., Fuller W.A. and Guenther P.M., (1997). “Estimating usual dietary intake distributions: adjusting for measurement error and nonnormality in 24-hour food intake data.” In Survey Measurement and Process Quality, eds. L Lyberg, M Collins, E DeLeeuw, C Dippo, W Schwartz & D Trewn, pp 689-709. New York: Wiley.

59. Nusser S.M., Carriquiry A. L., Dodd K.W., Fuller W.A., (1996a). “A semiparametric transformation approach to estimating usual daily intake distributions.” Journal of the American Statistical Association 91:1440-1449.

60. Rosner B., Spiegelman D. and Willett W. C., (1990). “Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error.” American Journal of Epidemiology, 132:734-745.

61. Ruppert D., Wand M.P., Carroll R.J., (2003).”Semiparametric regression.” Cambridge University Press.

62. Sawaya A.L., Tucker K., Tsay R., Willett W., Saltzman E., Dallal G.E. and Roberts S.B., (1996). “Evaluation of four methods for determining energy intake in young and older women: comparison with doubly labeled water measurements of total energy expenditure.” Am. J. Clin. Nutr. 63, 491-499.

63. Sempos C.T., (1992). “Some limitations of semiquantitative food frequency questionnaires.” Am. J. Epidemiol. 135, 1127-1132.

64. Slater B., Marchioni L. D., Fisberg M R., (2004). “Estimating prevalence of inadequate nutrient intake.” Rev Saude Publica 38:4.

65. Slob W., (1993). “Modelling long-term exposure of the whole population to chemicals in food.” Risk Anal. 13, 525-530.

66. Slob W., (1996). “A comparison of two statistical approaches to estimate long-term exposure distributions from short-term measurements.” Risk Anal. 16, 195-200.

67. Slater B., Morchioni D., Voci S.,(2007). “Use of linear regression for correction of dietary data.” Rev. Saude Publica, 41(2).

68. Smith M., Kohn R., (1996).”Nonparametric regression using Bayesian variable selection.” Journal of Econometrics 75:317-44.

69. Stone C.J., Koo C.Y., (1985). “Additive splines in statistics.” Proc. Stat. Comp. Sect. Am. Statist. Assoc., pp. 45-8.

70. Stefanski L. A., Carroll R. J., (1990). “Deconvoluting kernel density estimators.” Statistics 21:249-259.

71. Staudenmeyer J., Ruppert D., Buonaccorsi J., (2007). “Density estimation in the presence of heteroskedastic measurement error.” http://www.math.umass.edu/~jstauden/

72. Subar A. F., Kipnis V., Troiano R.P., Midthune D., Schoeller D. A., Bingham S., Sharbaugh C.O., Trabulsi J., Runswick S., Ballard-Barbash R., Sunshine J. and Schatzkin A., (2003). “Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: The OPEN study.” Am. J. Epidemiol. 2003;158:1-13.

73. Tarasuk V., Beaton G.H., (1992). “Statistical estimation of dietary parameters: implications of patterns in within-subject variation -a case study of sampling strategies.” Am. J. Clin. Nutr. 55, 22-27.

74. The EFCOSUM Group (2002). “Summary - European food consumption survey method.” European Journal of Clinical Nutrition, 56: Suppl 2, S1-S3.

75. Volatier J.L., Turrini A. and Welten D., (2002). “Some statistical aspects of food intake assessment.” European Journal of Clinical Nutrition 56: Suppl 2, S46-S52.

76. Wallace L.A., Duan N. and Ziegenfus R., (1994). “Can long-term exposure be predicted from short-term measurements?” Risk Anal. 14, 75-85.

77. Wallce L. and Williams R., (2005). “Validation of a method for estimating long-term exposures based on short-term measurements.” Risk Anal. 25, No.3:687-694.

78. Watson P., McDonald B., (2007). ”Seasonal variation of nutrient intake in pregnancy: effects on infant measures and possible influence on diseases related to season of birth.” European Journal of Clinical Nutrition 61, 1271 -1280.

79. West M., (1985). “Generalized linear models: scale parameters, outlier accommodation and prior distribution.” In Bayesian Statistics 2. North-Holland, Amsterdam.

80. Willett W., (1998). “Nutritional epidemiology”, New York:Oxford University Press.