**Sinara L Rossato[I,II,III]**

**Sandra C Fuchs[I,II]**

# Handling random errors and biases in methods used for short-term dietary assessment

# Manejo de erros aleatórios e vieses em métodos de avaliação de dieta de curto período

## ABSTRACT

Epidemiological studies have shown the effect of diet on the incidence of chronic diseases; however, proper planning, designing, and statistical modeling are necessary to obtain precise and accurate food consumption data. Evaluation methods used for short-term assessment of food consumption of a population, such as tracking of food intake over 24h or food diaries, can be affected by random errors or biases inherent to the method. Statistical modeling is used to handle random errors, whereas proper designing and sampling are essential for controlling biases. The present study aimed to analyze potential biases and random errors and determine how they affect the results. We also aimed to identify ways to prevent them and/or to use statistical approaches in epidemiological studies involving dietary assessments.

**DESCRIPTORS: Diet Records. Data Analysis, methods. Eating. Food Consumption. Diet Surveys, methods.**

[I]   Hospital de Clinicas de Porto Alegre. Porto Alegre, RS, Brasil

[II]  Programa de Pós-Graduação em Epidemiologia. Faculdade de Medicina. Universidade Federal do Rio Grande do Sul. Porto Alegre, RS, Brasil

[III] Department of Nutrition. Harvard School of Public Health. Boston, Massachusetts, USA

**Correspondence:**
Harvard School of Public Health
Department of Nutrition – Room 300
677 Huntington Avenue
Boston, Massachusetts 02115
E-mail: sinara.lrossato@gmail.com

## RESUMO

Estudos epidemiológicos têm evidenciado o efeito da dieta na incidência de doenças crônicas, mas a precisão e a acurácia de dados de ingestão alimentar requerem planejamento, delineamento e modelagem estatística. A estimativa da ingestão alimentar usual na população por métodos de avaliação de curto período, como recordatórios alimentares de 24 horas ou diários alimentares, é influenciada por erros aleatórios e vieses inerentes ao método. Para o manejo de erros aleatórios, utilizam-se a modelagem estatística e o apropriado delineamento e amostragem, cruciais para controle de vieses. O objetivo deste artigo é analisar potenciais vieses e erros aleatórios, suas influências nos resultados e como prevenir e/ou tratá-los estatisticamente em estudos epidemiológicos de avaliação de dieta.

**DESCRITORES: Registros de Dieta. Análise de Dados, métodos. Ingestão de Alimentos. Consumo de Alimentos. Inquéritos sobre Dietas, métodos.**

## INTRODUCTION

The assessment of food consumption and nutrient intake involves systematic and random errors that are inherent to the method used for data collection, which can be obtained either by a 24-h food record (R24h) or by maintaining a food diary (FD). Information obtained from a single R24h or FD does not represent the usual food intake. Proper representation of the usual food intake depends on the cooperation of the participant and on the number of reported days. Nevertheless, means obtained from several replicate observations may display high variability that could lead to errors in the portion of the population that reports unusual food intake.[2] Thus, data obtained from a single day or several days are susceptible to errors, which can be minimized using a proper statistical approach and by adequate sampling.

When the error originates from variations in individual food choices, which may simply differ from one day to another, the error is characterized as random and is common to all individuals in a population. However, apart from individual characteristics, other factors affect the variability in food consumption, including the level of development of the country where the study is being performed, specific characteristics of the population, and methods used for data collection. When these factors affect the results, the event is referred to as bias and is no longer referred to as a random error.[6] Examples of biases include differences in calorie intake in the summer versus that in the winter or calorie intake on weekdays versus that on weekends and also when obese individuals under-report food consumption. In addition, biases can be related to study outcomes; in case-control studies, individuals included as cases may report food intake differently from those included as control.[3]

Both random and systematic errors may affect data analysis and the interpretation of results.

The objective of this study was to analyze potential biases and random errors as well as their effect on the results. In addition, we aimed to identify methods to prevent them and/or use statistical approaches in epidemiological studies involving dietary assessments.

Food Frequency Questionnaires (FFQ) usually rely on the use of R24h and FD as standard assessment tools, and the strategies used in these questionnaires determine the accuracy and precision of the method. It is important that the investigator, at the time of sample planning, recognizes the variability in food consumption for a given individual and the need to use more than one tool for characterizing the routine diet. This will minimize potential biases and ensure the statistical power of the study.[6] In this case, the investigator needs to calculate the proper sampling size and determine the number of observations to be obtained by an individual on the basis of the ratio between the values calculated for intra- and inter-individual variations for specific nutrients.[1,5] One of the methods used to calculate the number of days required to estimate the usual food intake is based on the correlation between the expected and usual intake [$d = [r^2/ (1 - r^2)] \sigma_w/\sigma_b$], where $d$ is the number of data collection days per individual, $r$ is the expected correlation between usual and observed values, and $\sigma_w/\sigma_b$ is the ratio between the intra- and inter-individual variation. The higher the $r$ value, the greater is the proportion of individuals that are correctly classified; in contrast, the lower the ratio between the variations, the lower is the number of days required for proper classification of the individuals.[5]

A second method is based on the calculation of the confidence level of estimations of food intake, expressed as percentages $[d = (Z_a CV_w/D_o)^2]$, where $d$ is the number of days required by an individual that, when normal, assumes the value of 1.96; $CV_w$ is the coefficient of intra-individual variation calculated by dividing the intra-individual variation by the mean food intake; and $D_o$ is the specified level of error (confidence level) that could vary between 10.0% to 30.0%.[5] When the calculation is not performed, the interpretation of the no significant results can be confirmed by estimating the statistical power, obtained by the number of replicate observations.

The estimation of the sampling size can be obtained from results in studies performed with similar populations. For example, in adult Japanese women, the number of days required for obtaining reliable food intake data varied between 3 and 10 days when R24h was used to estimate the intake of energy and macronutrients. The study of nutrients with high variability of intake, such as cholesterol and vitamins A and C, may require 20 to 50 records. Assuming that the error in the estimation of intake varies between 10.0% and 20.0%, the number of assessment days would be as follows: 10 and three days for energy intake; 91 and 23 days for cholesterol intake; 118 and 30 days for zinc intake.[7] Basiotis et al[1] studied 13 men and 16 women during one year while evaluating the difference between the number of days required to evaluate usual diet between groups, individually and for different nutrients, considering the expected statistical precision. These authors demonstrated that the number of days required to evaluate nutrient intake varies according to the nutrient and from person to person. Compared with vitamin A, fewer days were required to evaluate energy intake because energy was consumed by all individuals. Although both energy and vitamin A intakes differ between individuals, the energy variation is considerably lower than vitamin A variation (14 days for energy in men and women; for vitamin A, these numbers corresponded to 115 days in women and 152 days in men). To reach a statistical precision of 10.0% for each individual, a greater number of days was required, whereas the number of replicate observations was considerably lower for the whole population. The authors concluded that the sample size and number of replicate observations are essential for increasing the statistical precision of the study.[1]

## INFLUENCE OF RANDOM ERRORS AND STATISTICAL MODELLING

A random error often leads to misinterpretations. According to Dood et al,[2] random errors increase the scope of the results, as demonstrated by comparing the scope of the dietary assessment based on data collected from a single R24h with those obtained from two or more R24h assessments. With regard to the intake of fruits and vegetables, for example, the number of individuals with an intake corresponding to less than one daily serving varied from 9.3% (based on estimation from a single R24h) to 0.4% (based on a mean of two R24h assessments). The second common error is related to the interpretation of hypothesis tests. The excessive variability leads to a loss in the statistical power, which makes statistical tests an invalid resource.[2]

Based on the assumption that food intake data are free of biases, statistical modeling can attenuate the inherent variability.[2] The method proposed by the National Research Council (1986) generated at least six other methods: the Slob method (1993), Wallace (1994), original and modified Buck methods (1995), Nusser (2000), Gay (2000), and N-Nusser;[4] more recently, other methods have been proposed. The table below describes different statistical modeling methods used to adjust the variability in food intake in a step-by-step manner. This table is based on the original work published by Dodd et al;[2] however, it is also supplemented with information from the Statistical Program to Assess Dietary Exposure (SPADE) and Multiple Source Method (MSM).

Additional details about the development of methods included in the National Research Council/Institute of Medicine, Iowa State University (ISU), Best-Power, Iowa State University Foods (ISUF),[4] MSM, and SPADE can be obtained from the specific references (Table). Other methods have been described, adapted, or remodeled. The Slob method showed disadvantages with regard to the correction of intra-individual variability losses, affecting the mean at the lower percentiles. The Buck method reproduced the asymmetry found in the original data.[4] Consequently, the statistical software Age-mode was improved in 2006[4] (readapted to generate the SPADE software) to estimate the usual food intake (Table). Unlike other models, SPADE describes food intake as a direct correlation with age, showing differences in the scope of results for children when compared with the ISU method. The MSM method can be used to estimate sporadic food intake for QFA and for food propensity questionnaires. However, this approach also showed some issues associated with remains from regression models that are not normally distributed. This model is also being improved.

## FINAL CONSIDERATIONS

Food intake data are susceptible to random errors and should be subjected to statistical modeling for obtaining precise estimations and for a proper interpretation of the results. For most studies, the choice of methods may not have a significant effect on the results; however, more current methods such as ISUF, MSM, and SPADE can be used. The MSM method is the preferred choice for evaluating the sporadic intake of food or nutrients. An

**Table.** Statistical models used to derive usual food intake on the basis of R24h and FD.

| NCR/IOM | ISU | BP | ISUF | MSM | SPADE |
|---|---|---|---|---|---|
| **Step 0: Initial data adjustment** | | | | | |
| Subject the R24h data to *Power* or *log* transformation until the data approach the normal distribution. | Adjust the observed R24h to no individual bias such as seasons of the year, days of the week, and effect of sampling. Build a two-stage transformation so that the modified R24h data approach the normal distribution. | Adjust the observed R24h to no individual bias such as seasons of the year, days of the week, and effect of sampling. Subject the R24h data to *Power* or *log* transformation until the data approach the normal distribution. | Estimate the distribution of the probability of intake for a given day on the basis of the relative frequency of R24h values that are different from zero. Place the R24h zero values aside and adjust the observed R24h to no individual bias such as seasons of the year, days of the week, and effect of sampling. Build a two-stage transformation so that the modified R24h data approach the normal distribution. | Apply Box-Cox transformation so that data approach the normal distribution. | Apply Box-Cox transformation so that data approach the normal distribution. |
| **Step 1: Description of the relationship between individual R24h data and usual food intake** | | | | | |
| There is no bias in the estimation of transformed usual intake on the basis of R24h data (assumption A). | There is no bias in the estimation of usual intake in the no transformed scale on the basis of R24h data (assumption B). | There is no bias in the estimation of usual intake in the no transformed scale on the basis of R24h data (assumption B). | Usual intake corresponds to the probability of consumption in a given day multiplied by the total usual intake for a given day. One R24h measures the intake exactly equal to zero. There is no bias in the estimation of usual intake in the no transformed scale on the basis of R24h data (assumption B). | Estimate the probability of intake using logistic regression and the total daily intake using linear regression. | Assemble a fractional polynomial model for no transformed data. |

Continue

Continuation

| Step 2: Separation of the total variation of the R24h data into intra- and inter-individual variations | | | | |
|---|---|---|---|---|
| The intra-individual variation is the same for all individuals. | The intra-individual variation is the same for all individuals. | The intra-individual variation may vary among individuals. | The intra-individual variation may vary among individuals. | Transformed remains are used to estimate the inter- and intra-individual variations, which are then used to convert the mean intake of an individual to an overall mean.<br><br>Obtain a mixed-effects fractional polynomial model to separate the inter- and intra-individual variability on the basis of age. |

| Step 3: Estimation of the distribution of usual intake taking intra-individual variation into account | | | | |
|---|---|---|---|---|
| Assemble a group of intermediate values, which retain the variability of transformed R24h data among individuals.<br><br>Inverse transformation: apply the inverse function of the initial value to each intermediate value.<br><br>The inverse of the empirical distribution corresponds to the distribution of usual intake. | Assemble a group of intermediate values, which retain the variability of the transformed R24h data among individuals.<br><br>Inverse transformation: use the inverse function of the initial *Power* or *log* transformation in parallel to adjusting for bias, and correct each intermediate value in a normal scale to obtain the original scale.<br><br>The inverse of the empirical distribution corresponds to the distribution of usual intake. | Inverse transformation: apply the inverse function of the two-stage transformation, in parallel to adjusting biases; concomitant to bias adjustment, mathematically describe the original distribution of the usual daily intake.<br><br>Mathematically combine the distribution of the daily intake with the estimated distribution of the probability of intake to obtain the group of intermediate values that represent usual intake, while assuming that usual intake and daily intake are statistically independent variables.<br><br>The inverse of the empirical distribution corresponds to the distribution of usual intake. | Inverse transformation: integrate nonnegative whole values of the Box-Cox parameters.<br><br>The estimation of usual intake is obtained by multiplying the probability of intake and the total daily intake estimated by regression models. | Identify discrepant values using the Grubbs method.<br><br>Test residual normality and data distribution by the Kolmogorov-Smirnov test using the statistical model S-plus.<br><br>Check $\lambda$ distribution.<br><br>Identified discrepant values are eliminated, and previous steps are repeated.<br><br>Inverse transformation: apply inverse transformation with a quadratic Gaussian function (Monte Carlo Simulations). |

improved version of this method will soon be available. A proper study design and sample selection can help minimize biases. It is important that selected characteristics such as nutritional and health status, days of the week, and seasons of the year are proportional and heterogeneous to avoid sampling-related systematic errors. The number of replicate observations of R24h and the sample size can be estimated on the basis of the variability in the nutrient intake among individuals. For example, nutrients that are present in most food types, such as macronutrients, require a lower number of replicate observations because of less variability among these observations. When the purpose of the study is to evaluate the overall food intake of a population, larger samples with a lower number of replicate observations may be sufficient to generate reliable data. However, in validation studies, where the variability among individuals is critical because it serves as the reference to evaluate data validity, the use of a higher number of replicate observations is preferred.

## REFERENCES

1.  Basiotis PP, Thomas RG, Kelsay JL, Mertz W. Sources of variation in energy intake by men and women as determined from one year's daily dietary records. *Am J Clin Nutr*. 1989;50(3):448-53.

2.  Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D, et al. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. *J Am Diet Assoc*. 2006;106(10):1640-50. DOI:10.1016/j.jada.2006.07.011

3.  Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional studies. *J Natl Cancer Inst*. 2011;103(14):1086-92. DOI:10.1093/jnci/djr189

4.  Hoffmann K, Boeing H, Dufour A, Volatier JL, Telman J, Virtanen M, et al. Estimating the distribution of usual dietary intake by short-term measurements. *Eur J Clin Nutr*. 2002;56(Suppl 2):S53-62. DOI:10.1038/sj/ejcn/1601429

5.  Nelson M, Black AE, Morris JA, Cole TJ. 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*. 1989;50(1):155-67.

6.  Willett WC. Nutrition epidemiology. 3.ed. New York: Oxford University Press; 2013.

7.  Tokudome Y, Imaeda N, Nagaya T, Ikeda M, Fujiwara N, Sato J, Kuriki K, Kikuchi S, Maki S, Tokudome S. Daily, weekly, seasonal, within- and between-individual variation in nutrient intake according to four seasons consecutive 7 day nutrient diet records in Japanese female dietitians. *J Epidemiol*. 2002;12:85-92.

8.  Department of Epidemiology of the German Institute of Human Nutrition Postdam-Rehbrucke, Versão 1.0.1. Disponível em: https://nugo.dife.de/msm

9.  Waijers PMCM et al. The potential of AGE_MODE, an age-dependent model, to estimate usual intake and prevalence of inadequate intakes in a population. *J Nutr*. 2006;136:2916-20.