FUNDAMENTAL FLAWS IN DESIGNING A FOOD CONSUMPTION SURVEY

Peng T. Liu, Division of Mathematics, CFSAN, FDA
200 C St. SW Washington DC 20204

Key Words: Cross-sectional Survey, Food Intake, Nutrition Intake, Exposure Estimation, Percentile

When the food consumption data obtained from a survey of each sampled individual covers only a short period of time, then little or no information is available to measure any given individual's real long-term food consumption. This study points out that erroneous conclusions could be drawn when a nutrition study or risk assessment depends on the percentiles of a "synthetic" distribution pieced together from a cross-sectional population survey. Under certain conditions, such a survey may provide an unbiased estimate of the population mean, but the percentile estimates will be unreliable.

THE PROBLEM

Certain food consumption databases are among the most important sources of information that enable the Center for Food Safety and Applied Nutrition, FDA to carry out its mandated duties. They provide the data vital to many scientists from different disciplines, who assess risks and make regulatory decisions, to ensure food safety and nutritional quality.

Various food consumption surveys have been routinely conducted by USDA and other agencies. Unfortunately, information collected through the "cross-sectional" population surveys contains little individual dietary information. The study of eating habits' relationship to nutrition, obesity, diabetes, hypertension, stoke, heart disease, and cancers, requires long-term individual consumption information. Such information cannot be abstracted from the current food consumption databases. For example, suppose we conduct a one-day food consumption survey and record every "ounce" of food intake 100% accurately for each and every member of the U.S. population. Even though we collect about 270 million sets of records, the survey provides few clues for any given individual's food consumption information because it is difficult to predict any person's eating patterns based on one-day information. In other words, if someone eats steak on the date of the survey, it does not imply that he eats "steak" seven times per week, thirty times per month or three hundred sixty five times per year. The lack of individual's long-term consumption information is a fundamental deficiency for current food consumption surveys.

If the individual food consumption information is not available then the "percentiles" of the individual consumption distribution, which is used for various regulatory purposes, is unobtainable. Erroneous conclusions could be drawn by using the "percentiles" of the one-day cross-sectional "synthetic" distribution to replace the "percentiles" of individuals consumption.

THE ERRONEOUS CONCLUSIONS

The following examples will demonstrate how erroneous conclusions could be drawn from the current food consumption survey.

<u>Example 1</u>

Suppose that everybody in the population eats a certain food three times per year and the eating occasions are randomly distributed over the calendar year. What conclusion could be drawn from the typical food consumption survey?

A one-day survey would tell you that 99% ((365-3)/365) of the population are non-eaters, and 1% of the population eat that food item daily. A three-day survey would tell you that 97% of the population are non-eaters and 3% of the population eat that food item once within three days (calculated via a binomial distribution with n=3, and p=0.01 (i.e., Bin. (n=3, p=0.01)). A fourteen-day survey would tell you that 87% of the population are non-eaters, 12% of the population eat that food item once within fourteen days, and 1% of population eat twice or more within fourteen days (i.e., Bin.(n=14, p=0.01)).

Note how the sample distributions are inconsistent with the "true" distribution wherein everybody consumes the food but only three times a years. The "mesh size" is inappropriate for capturing the information we seek: the survey should use as the basic measure units (sampling units) individuals throughout a reasonable length of observation. In other words, the statistical inferences based on these types of surveys may seriously distort reality.

One may argue that this example is too extreme. The number of eating occasions for most foods may be greater than three times per year. Therefore, the second example is given as follows.

Example 2

Suppose that everybody in the population eats a certain food 183 times per year and that the eating occasions are randomly distributed over the calendar year.

A one-day survey would tell you that 50% of the population are non-eaters and the other 50% of the population eat daily. A three-day survey would tell you that 12.5% of the population are non-eaters, while 37.5% of the population eat once within three days, 37.5% of the population eat twice within three days, and 12.5% of the population eat three times within three days (i.e., Bin.(n=3, p=0.5)). A fourteen-day survey would tell you that about 1% of the population eat less than three times within fourteen days, 2% of the population eat 4 times within fourteen days, 5% of population eat five times within fourteen days,.....etc. (i.e., Bin.(n=14, p=0.5)).

As in the prior examples, the scope of the survey is not over a long enough time period. Surveys covering individuals over just a few days or weeks will never provide answers to the percentile questions that are routinely asked in nutritional studies and risk assessments.

UNBIASED ESTIMATES

When you conduct a survey and collect some data, obviously, it should provide you some sort of information. What information can we expect from the cross-sectional population survey? If the eating occasions are randomly and uniformly distributed over the calendar year, the cross-sectional survey provides an unbiased estimate of the population mean and an unbiased estimate of the variance of the population mean.

From Example 1, where the true frequency of eating occasions is 3 times per year, it is easy to demonstrate numerically that the estimate for the average eating occasions is also three times per year, regardless of whether we conduct a one-day survey, a three-day survey, or a fourteen-day survey. Similarly, from Example 2, where the true number of eating occasions for each member is 183 times per year, the estimate for the average number of eating occasions per year is also 183, regardless of whether one uses a one-day survey, or a three-day survey, or a fourteen day survey.

VARIANCE AND RELATED CONCERNS

As part of the survey re-design process, at least two types of variability should be considered: 1) the variability from person to person within a population, and 2) the variability from day to day within an individual. Due to the different magnitude of "intra-" and "inter-" variabilities and due to the potential inter-correlation, these variabilities cannot be treated equally. In other words, it is inappropriate to apply the "person-day" concept in food consumption estimation. It is clear that the consumption estimated from one person observed for 365 days could differ significantly from the consumption estimated from 365 persons for one day. Appropriate algebraic notation may help to express the survey design problem more concisely. What does the user of a survey mean by "consumption" of a given food?

Suppose we sample 1,000 persons and measure daily the diet of this cohort for a reasonably long period of time, say one year. Let $X_{i,j}$ denote the food consumption for i^{th} person on j^{th} day, where i = 1,2,3, ...,1000 and j=1,2,3,....,365. The "daily" food intake (i.e., the average food intake per day) for the i^{th} person is $\bar{x}_{i.} = \sum X_{i,j}/365$. We then rank the 1,000 $\bar{x}_i$'s in ascending or descending order. From the resulting ordered listing, the user can simply pluck out the desired "percentile" or can determine the % of the population (based on the representative sample of 1,000) that exceed some critical consumption level.

It doesn't matter whether the food consumption is expressed in daily, weekly or annual terms because it can be converted from one unit to the others. That is, if a person eats steak once a month, it is equivalent to eating steak 0.03 times per day, or 12 times per year. The unit conversion becomes meaningful if the consumption information obtained is from a reasonably long period of observation. Thus, using one-day information to predict monthly or yearly outcome is unacceptable.

A REMEDY

The inappropriateness of the time-frame of current survey designs could be rigorously remedied by extending the period of observation. It is understandable that conducting a food consumption survey is very costly. Actually monitoring individuals for more than one year is almost impossible. However, we may change the contents and scope of the questionnaire in order to obtain the relevant long-term consumption information. Such a slight perturbation in the design would not be expected to raise costs very much. For example, the individual could be asked: How often did you eat steak throughout the whole past year? Of course, this proposed solution does entail the problem of recall error. However, at least this suggestion addresses the issue of proper time frame with minimal additional expenses.

THE SAMPLE SIZE PROBLEMS

Another key issue in survey design is the appropriate "sample size" – the number of individuals to be included in your food survey. The sample size tables based on the precision of estimating a proportion, adjusted by the "design effect", are provided by USDA and other statistical textbooks. However, the applicability of these tables could be limited. The essential variables in a food consumption survey are continuous variables rather than a "dichotomous". In addition, the 90^{th} or 95^{th} percentile of distribution is the most common parameter used for regulatory purposes. The sample size should be determined by the precision of estimating a percentile instead of the precision of estimating a mean or proportion. Moreover, these surveys are usually designed for multiple purposes to answer a wide range of questions. To achieve these multiple objectives, the surveys require selection of a wide range of samples from different groups in the population. The sample size for such a complex survey cannot be determined by a single equation to achieve a single objective.

A FOOD CONSUMPTION DISTRIBUTION

A bell-type unimodal curve is frequently used to describe the distribution of food consumption, food contamination or chemical exposure. However, the adequacy of the model is questionable. There are enormous varieties of food on the market. It is unlikely for a person to try all possible food items, even over a lifetime. Particularly, during a short time period, only a small number of food items could be consumed. A large proportion of respondents may not eat specific food items on the date of the survey. This results in an unequal number of observations per food item. Based on our observation, the consumption for each food item seems to be a zero-truncated positive-skewed, long-tailed distribution. Obviously, the total consumption is a sum of different food items. For example, suppose we want to estimate the pesticide exposure for apples and corn. According to the eating status, the respondents can be classified into four mutually exclusive and exhaustive groups; eat apples only, eat corn only, eat both apples and corn, and eat neither apples nor corn. Even if the pesticide exposure for both apples and corn are distributed as bell-shaped normal curves, the distribution of total exposure by combining these four different sub-groups could be a bimodal or trimodal distribution. The best way to describe this model is a mixture distribution (e.g., $f_{a,b}(x) = (\neg a \cap \neg b)\, N(0,0) + (a \cap \neg b)\, N(\mu_a, \sigma_a^2) + (\neg a \cap b)\, N(\mu_b, \sigma_b^2) + (a \cap b)\, N((\mu_a + \mu_b, \sigma_a^2 + \sigma_b^2)$, where we assume that the variables A and B are independent). When we combine n food items, there are 2^n distinct combinations. It is no surprise the total consumption has an irregular shape. The bell-shaped unimodal model for describing food consumption distribution is unrealistic.

CONCLUSIONS

Great efforts have been made in current food consumption survey designs to select "representative" samples and to achieve minimum variance unbiased estimation. However, we cannot blindly use these databases without fully understanding the sampling scheme involved and the weighting procedures employed. It is most important that we have to judge whether the distribution derived from the databases is suitable to answer the questions raised in a nutrition study or a risk assessment. For example, we cannot assume that the distribution obtained from a one-day consumption survey is equivalent to "daily" food intake. Typically the 95^{th} percentile of lifetime exposure has been estimated from the 95^{th} percentile of daily exposure multiplying by 365 (days) and by years of life expectancy. We don't need to be a "rocket scientist" to find out how wrong this approach is. The units of daily, weekly, monthly, and yearly exposure are mathematically convertible. However, the conversion becomes reliable only when we observe the food consumption over a reasonably long period of time. The monthly or yearly consumption predicted by one-day observation is unreliable.

A cross-section population survey of food consumption is analogous to a cross-cut of a football, which loses one dimension. Based on the cross-sectional plane, we may conclude that the shape of a football is like a soccer ball or a saucer.

REFERENCE

National Food Consumption Survey (NFCS) 1977-78, Report No. 1-2, Consumer Nutrition Division, Human Nutrition Information Survey, USDA. 1978.