# American Statistical Association

# 2000
# Proceedings
*of the*
# Section on Survey Research Methods

Papers presented at the
Annual Meeting of the American Statistical Association,
Indianapolis, Indiana, August 13-17, 2000,
and the
American Association for Public Opinion Research
World Association for Public Opinion Research
Annual Conference
Portland, Oregon, May 17-21, 2000,
under the sponsorship of the
Section on Survey Research Methods

**American Statistical Association ■ 1429 Duke Street, Alexandria, VA 22314**

# A NATIONAL SAMPLING PLAN FOR OBTAINING FOOD PRODUCTS FOR NUTRIENT ANALYSIS

Charles R. Perry, Daniel G. Beckler - USDA, National Agricultural Statistics Service
Pamela Pehrsson, Joanne Holden, - USDA, Nutrient Data Laboratory
Daniel G. Beckler, USDA/NASS/RDD, Room 305, 3251 Old Lee Highway, Fairfax, VA 22030-1504

**Keywords:** Sampling, National Food and Nutrient Analysis Program, USDA Nutrient Database for Standard Reference

## 1. Introduction

This paper describes the sampling procedures used by USDA's Nutrient Data Laboratory (NDL) to select food products for their National Food and Nutrient Analysis Program (NFNAP). The goal of this program is to obtain national level estimates of the nutritional components for common foods consumed in the United States. These estimates will be imported into the USDA National Nutrient Databank System and disseminated in the USDA Nutrient Database for Standard Reference and other data sets produced by the Laboratory.

The sampling procedures provided a means to select a self-weighting nationally representative sample of foods consumed by people in the United States. It is assumed that ounces of food consumed is proportional to population and is constant across the United States. The plan was a three stage design where counties were selected at the first stage, grocery store outlets within the selected counties were selected at the second stage, and the third stage selected specific food products to be purchased and analyzed for nutrient content. In effect, this gave a sample of grocery outlets from selected geographical dispersed areas across the United States. The stages are described below.

## 2. First Stage Sample Design

The goal of the first stage selection process was to obtain a sample of counties dispersed across the United States. These counties were selected proportional to their population to comply with our desire for a self-weighting sample.

To begin, estimated 1997 population data were obtained from the U.S. Bureau of the Census for all states. Population data are available from the Census' Internet web site. States were then grouped by geography into four approximately equal regions in terms of population (the target population for each region was 66,909,015, one fourth of the United States' population). Alaska and Hawaii were excluded for logistical reasons. Final regions, with populations, are given in Table 1. Figure 1 illustrates the four regions; it should be noted that the regions do not follow those established by the Bureau of the Census.

Table 1: First Stage Regions

| Region Num. | Region Name | 1997 Estimated Population |
|---|---|---|
| 1 | Northeast | 66,492.898 |
| 2 | South | 65,245,265 |
| 3 | Great Lakes & Texas | 68,014,743 |
| 4 | Plains. Rockies & Pacific | 67.883.155 |

The next step involved selecting three Consolidated Metropolitan Statistical Areas (CMSA) from each region. However, since all counties are not included in a CMSA, Generalized Consolidated Metropolitan Statistical Areas (gCMSA) were used. The gCMSAs were defined as the standard CMSAs or individual counties for areas not in a CMSA. Two methods were explored for selecting the gCMSAs. The first method involved sorting the gCMSAs within a region in descending order by population. Once sorted, a probability proportional to size (PPS) systematic sample of size three was drawn within each region. A systematic sample consists of drawing a single unit at random and taking every x^th^ element thereafter, where x is a fixed number sufficient to give the desired sample size. Consult Cochran (1977) or Särndal, et al. (1992) for a thorough discussion of systematic sampling.

The second method involved creating three equally sized strata by population within each region and selecting one gCMSA at random per stratum. If the same gCMSA was selected in more than one stratum[1], the sample was rejected and redrawn. This method was essentially a manually implemented Chromy's Method (Chaudhuri and Vos, 1988). Chromy's Method is a sequential sampling technique that maintains exact probability proportional to size (PPS) selection for unequal sized sampling units. See Krewski et al. (1981) for a thorough discussion of the method.

The selected samples from the two methods were compared and revealed no practical differences. Although the sample from the first method was used operationally, if another first stage sample is drawn in the future, we

[1] This could occur when a stratum boundary divided a gCMSA's population.

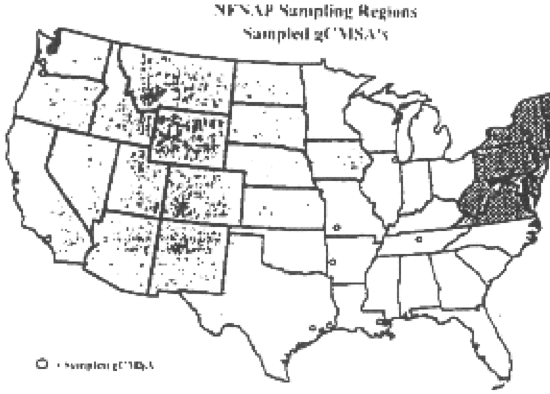

**Figure 1**

recommend using Chromy's method.

Once the gCMSAs were selected, the counties that made up each gCMSA were sorted in descending order by their urbanicity (Goodall, et al., 1998) and a systematic sample of size two was selected. Urbanicity is a measure of how urban a county is, based on the population of the largest cities/towns in the county. Using such a measure ensures that counties bordering major cities (e.g., New York City, Washington, DC) are treated more list the major city than like the area on the outskirts of the metropolitan area. Sorting counties within gCMSAs by urbanicity ensured that the sample contained both more urban and lesser urban areas. For those gCMSAs made up of only a single county, the county was selected twice. Table 2 contains the selected gCMSAs and counties.

After the initial counties were selected, the list of grocery store outlets described under **Second Stage Design** was reviewed to ensure each selected county had at least ten outlets available to sample from. Ten was chosen arbitrarily, but it was felt that outlets in areas with very few stores (i.e., less than 10) would not give a representative sample of grocery stores and may not carry the wide range of food products NFNAP will sample. The product collection contractor charged by the number of stores visited, so in order to keep costs down, the number of stores visited was tried to be kept to a minimum. As indicated by Table 2, four additional counties were included with Polk County, AR to ensure that the area contained the minimum number of grocery store outlets.

### 3. Second Stage Design

The goal of the second stage was to select a sample of grocery store outlets from within each county selected in the first stage. To accomplish our desire for a self-weighting sample, outlets were selected proportional to their value of annual sales; larger outlets, in terms of annual sales, had higher probabilities of selection.

The list of grocery store outlets was obtained from **Trade** *Dimensions* (Wilton, CT), a private market research company specializing in the retail grocery industry. The **Trade** *Dimensions* outlet list contained grocery stores in the counties selected in the first stage with annual value of sales of \$2.0 million or more. Stores with less than \$2.0 million of annual sales were excluded because it was NDL's opinion that such stores could not be expected to carry the diversity of products which NFNAP will sample. The outlet list contained 2,411 stores and provided contact information (store name, address, telephone number) in addition to annual value of sales.

The selection process consisted of drawing a systematic sample of size one, proportional to the outlets' annual value of sales, from each selected county. The procedure gave higher probabilities of selection to larger outlets. Two outlets were drawn in counties where only a single county made up the gCMSA. Two outlets were also drawn in the group of five counties that made up the Polk County, AR gCMSA. Alternate outlets were drawn in each county in case the primary selected outlets were inaccessible or products were unavailable. Table 3 contains the primary outlet sample.

### 4. Third Stage Design

The goal of the third sampling stage was to select specific food products (brands and package sizes) for nutrient analysis from food types (e.g., cheese pizza, chili with beans) identified by NDL. The intent was to purchase the same food products from each of the sampled grocery outlets and send them to the laboratory for nutrient analysis.

This section describes two types of product samples. The first type, which we will consider the main sample, was designed to test foods to find nutrient means for composited samples of the products. A composited sample is a homogeneous mixture of several packages of a specific food. It is important to note that results from nutrient analyses obtained from composited samples pertain to an average serving from the homogenized product, not to a typical serving [2].

[2] A typical serving generally comes from a single package of a food product, not from a mixture of several packages. It is not possible to separate individual serving variation from composites of several packages.

Table 2: Primary Outlet Sample Locations

| Region Number | gCMSA | County | City |
|---|---|---|---|
| 1 | NY, NY; Northern NJ, Long Island, NJ | Union County, NJ | Springfield, NJ 07081 |
| 1 | NY, NY; Northern NJ, Long Island, NJ | Richmond County, NY | Staten Island, NY 10306 |
| 1 | Pittsburgh, PA | Allegheny County, PA | Pittsburgh, PA 15220 |
| 1 | Pittsburgh, PA | Washington County, PA | Canonsburg, PA 15317 |
| 1 | Venango County PA | Venango County, PA | Franklin, PA 16323 |
| 2 | Venango County PA | Venango County, PA | Franklin, PA 16323 |
| 2 | Nashville, TN | Davidson County, TN | Hermitage, TN 37076 |
| 2 | Nashville, TN | Williamson County, TN | Franklin, TN 37064 |
| 2 | Springfield, MO | Green County, MO | Springfield, MO 65803 |
| 2 | Springfield, MO | Green County, MO | Springfield, MO 65802 |
| 2 | Polk County, AR | Polk County, AR | Mena, AR 71953 |
| 2 | Polk County, AR | Sevier County, AR | De Queen, AR 71832 |
| 2 | Chicago, IL; Gary, IL; Kenosha, WI | Cook County, IL | Bartlett, IL 60103 |
| 3 | Chicago, IL; Gary, IL; Kenosha, WI | Dupage County, IL | Wheaton, IL 60187 |
| 3 | Houston, TX; Galveston, TX; Brazoria, TX | Harris County, TX | Houston, TX 77057 |
| 3 | Houston, TX; Galveston, TX; Brazoria, TX | Montgomery County, TX | Conroe, TX 77301 |
| 3 | Beaumont, TX; Port Arthur, TX | Jefferson County, TX | Beaumont, TX 77706 |
| 3 | Beaumont, TX; Port Arthur, TX | Orange County, TX | Orange, TX 77630 |
| 3 | LA, CA; Riverside, CA; Orange County CA | Los Angeles County, CA | Los Angeles, CA 90016 |
| 4 | LA, CA; Riverside, CA; Orange County CA | Orange County, CA | Laguna Niguel, CA 92677 |
| 4 | Portland, OR; Salem, WA | Multnomah County, OR | Gresham, OR 97030 |
| 4 | Portland, OR; Salem, WA | Clark County, WA | Vancouver, WA 98682 |
| 4 | Cowlitz County WA | Cowlitz County, WA | Longview, WA 98632 |
| 4 | Cowlitz County WA | Cowlitz County, WA | Longview, WA 98632 |

The second type of sample was designed to obtain serving-to-serving variation for a particular food product. Nutrient analyses were conducted on individual packages of food products and provided variability estimates of a typical serving of the product. Serving-to-serving samples were drawn only for critical foods because of the substantial cost of the associated nutrient analyses. The serving-to-serving samples were drawn from the same products selected for the main sample.

For the main sample, food products were selected proportional to the number of ounces of each product sold nationally. For the initial wave of samples, this information was obtained from Nielsen Market Research SCANTRACK data. SCANTRACK data are obtained from checkout price scanners and, consequently, exclude products sold in stores without such equipment. Files obtained from Nielsen contained product name, package size and national market share. To maximize the likelihood of having selected products available in all outlets, we restricted the product sampling universe to products with at least one percent market share.

In future sampling waves, similar market share information will come from Information Resources Inc. (IRI). This should only increase the accuracy of market shares since IRI uses the same methodology as Nielsen Market Research and they cover more stores.

Actual product selection was done proportional to the number of ounces of the product consumed in the United States. This was operationalized by selecting a sequential sample by Chromy's Method (Chaudhuri and Vos, 1988) proportional to the product of market share and package size. The number of samples chosen for each product was based on the desired statistical results and the number of nutrient analyses NDL could afford to perform. Once specific food products were identified as in the sample, they were purchased from each of the sampled outlets.

The resulting group of selected products for a particular food item (e.g., 70% fat margarine) can be thought of as a matrix with outlets numbered one through 24 across the top and sample numbered one through 12 (assume 70% fat margarine had 12 samples) running vertically. Compositing took place by sample number. For example, product samples Number One from all 24 outlets were homogenized then a quantity sufficient for nutrient analysis removed. This process was done for each sample number, resulting in 12 individual data points for statistical analysis. Performing the analysis in this manner provided individual product (i.e., brand) data for major brands and overall results for the particular food product (e.g., 70% fat margarine). It is important to note that these results pertained to an average serving from the homogenized food product, not to a typical serving.

To obtain serving-to-serving variation (i.e., variation in a typical serving) for a particular food item, a supplemental sampling plan was used. The associated nutrient analyses were based on individual packages of the food item and provided variability estimates for a typical serving of the food product. The 70% fat margarine main sample (drawn as described above) will be used to illustrate the supplemental plan.

Suppose we chose 12 margarine samples from each of 24 outlets in the main plan. Table 3 illustrates the resulting matrix of products. We also note that the market share decreases (or stays the same) from sample One to sample 12 and that outlets from each gCMSA are consecutive. That is, outlets One and Two were in the same gCMSA, outlets Three and Four were in the same gCMSA, and so on.

The steps for the supplemental sampling plan consisted of: (1) Randomly selected one of main sample numbers One and Two, randomly selected one of main sample numbers Three and Four, and so on until one was randomly selected from main sample numbers 11 and 12. (2) Randomly selected two gCMSAs from each selected main sample so that no gCMSA was selected twice. (3) Randomly selected one outlet from each selected gCMSA. Table 3 illustrates the resulting matrix.

Table 4 shows that at step (1) primary samples One, Three, Five, Eight, 10 and 12 were selected. At step (2) gCMSAs One and 10 were selected from primary sample One, gCMSAs Three and Nine were selected from the remaining gCMSAs, and so on. Finally, at step (3), outlet Two was selected from gCMSA One and outlet 19 was selected from gCMSA 10 and so forth until we selected one outlet from each selected gCMSA. The result is that two outlets were chosen from each selected primary sample in such a way that no two outlets fell in the same gCMSA, which is illustrated by the ×'s in Table 4.

Although specific outlets were selected for the supplemental sample (see Table 4), as a cautionary measure to avoid missed units, data collectors were instructed to pick up extra primary samples in both outlets in selected gCMSAs. For instance, extra primary samples were collected in outlet number Two and in the other outlet in the gCMSA. Likewise, extra primary samples were collected in outlet number 19 and in the other outlet in the gCMSA.

Once obtained, these additional 12 packages were each homogenized and sufficient quantities removed for analysis. This yielded 12 data points for each nutrient and allowed serving-to-serving means and variability to be estimated.

**Table 3:** Matrix of Sampled Products

|  | **Outlets (same gCMSAs are consecutive)** |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| **Primary** | **gCMSA 1** |  | **gCMSA 2** |  |  | **gCMSA 11** |  | **gCMSA 12** |  |  |
| **Sample** | **1** | **2** | **3** | **4** | **...............** | **21** | **22** | **23** | **24** |  |
| 1 | —— | —— | —— | —— | —— | —— | —— | —— | ——→ | $\bar{x}_1$ |
| 2 | —— | —— | —— | —— | —— | —— | —— | —— | ——→ | $\bar{x}_2$ |
| 3 | —— | —— | —— | —— | —— | —— | —— | —— | ——→ | $\bar{x}_3$ |
| ⋮ | ⋮ |  | ⋮ |  | ⋮ | ⋮ | ⋮ |  | ⋮ | ⋮ |
| 11 | —— | —— | —— | —— | —— | —— | —— | —— | ——→ | $\bar{x}_{11}$ |
| 12 | —— | —— | —— | —— | —— | —— | —— | —— | ——→ | $\bar{x}_{12}$ |

**Table 4:** Serving-to-Serving Sample Example

|  | gCMSAs |  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| Main | Selected Outlets |  |  |  |  |  |  |  |  |  |  |  |
| Sample | 2 | 3 | 6 | 8 | 9 | 12 | 14 | 16 | 17 | 19 | 22 | 24 |
| 1 | × |  |  |  |  |  |  |  |  | × |  |  |
| 3 |  |  | × |  |  |  |  |  | × |  |  |  |
| 5 |  |  |  |  | × |  |  |  |  |  | × |  |
| 8 |  | × |  |  |  |  |  | × |  |  |  |  |
| 10 |  |  |  |  |  |  | × |  |  |  |  | × |
| 12 |  |  |  | × |  | × |  |  |  |  |  |  |

## 5. Estimation

By construction, the nutrient analysis data obtained under this plan will be [approximately] self-weighting and, consequently, will be treated as if it came from a simple random sample (SRS). In effect, the sampling plan has been used to obtain a well designed sample that approximately represents the foods eaten by the United States' population. Thus, the following formulas for nutrient means and standard errors of the means will be used for data from the main sample.

Let $x_1, x_2, x_3, \ldots, x_n$ represent a set of nutrient values obtained from the NFNAP sample for a specific food item. These n values are the result of nutrient analyses performed on the composited samples described under **Third Stage Design**. A reasonable estimate of the nutrient mean is given by equation (1) and an estimate of its standard error is given by equation (2) (ignoring the finite population correction factor).

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

$$SE(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n(n-1)}} \qquad (2)$$

A confidence interval for the estimated mean can be based on a *t* distribution with df = n - 1 degrees of freedom.

Due to a necessary number of simplifications in our sampling plan, we lost the ability to summarize the data in a statistically rigorous manner. However, Equation (1) still provides a reasonable estimate of the population mean nutrient value under the mild constraints and Equation (2) provides a reasonable estimate of the standard error of the mean.

## 6. Conclusion

The sampling plan described above provides a close approximation to what may be considered the ideal plan. An ideal plan would involve selecting geographically dispersed land areas across the United States, selecting retail food outlets in those area, and finally, selecting food products from a carefully constructed listing of all foods sold in those outlets. Such a plan was suggested and described by Nusser and Carriquiry (1998). Perhaps with abundant resources (staffing, money and time), such an ambitious plan could be implemented; however, this was not possible under NDL's current resources. Instead, the plan presented here achieved a [approximately] self-weighting, geographically dispersed sample across the United States that can be used to provide a means of selecting foods consumed. The basic notions of the plan suggested by Nusser and Carriquiry have been carried out in a cost and time effective, reasonable manner. However, there are a few shortcomings to the plan, as described below.

First, the sample is not *exactly* self weighting because the four regions in the first stage were not *exactly* equal in population. Second, the sample was also not *exactly* self weighting because in the second stage we used national level market share; product market share in the areas where our samples were drawn may be quite different. Third, Nielsen (or IRI) data cover only grocery stores that have automated price scanners; omission of food products sold from other outlets in the third stage may add bias and further detract from an exactly self-weighted sample. Fourth, results may be biased because all places to purchase food were not eligible for sampling in the third stage (we only included stores with $2.0 million or more annual value of sales, convenience stores were excluded, the list obtained from **Trade** *Dimensions* was doubtfully 100% complete and up-to-date). The results may also be biased because we used a cut off of one percent of market share for the individual food products. And fifth, standard error estimates may be somewhat conservative since our summary formula is based on the assumption of independent identically distributed (iid) observations.

## 7. References

Chaudhuri, A. and J. W. E. Vos (1988 ) Unified Theory and Strategies of Survey Sampling. NHC: New York, NY.

Chromy, J. R. (1981) "Variance estimators for a sequential sample selection procedure," in D. Krewski, R. Platek and J. N. K. Rao (Eds) *Current Topics In Survey Sampling*, Academic Press: New York, NY.

Cochran, W. G. (1977) *Sampling Techniques*, John Wiley and Sons: New York, NY.

Goodall C. R., K. Kafadar and J. W. Tukey (1998) "Computing and using rural versus urban measures in statistical applications." *The American Statistician*, 52:2, pp 101-111.

Nusser, S. M. and A. L. Carriquiry (1998) "Sampling approaches for nutrient data bases." Unpublished working paper from contract between USDA, Nutrient Data Laboratory and Iowa State University.

Särndal, C. E., B. Swennson and J. Wretman (1992) *Model Assisted Survey Sampling*. Springer-Verlag: New York, NY.