Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Food Science 2 (2013) 52 – 59

Procedia Food Science

36th National Nutrient Databank Conference

ARS, USDA updates food sampling strategies to keep pace with demographic shifts

Pamela Pehrsson*, Charles Perry, and Marlon Daniel

Nutrient Data Laboratory,Beltsville Human Nutrition Research Center, Agricultural Research Service, 10300 Baltimore Avenue, Beltsville, MD 20705

Abstract

The National Food and Nutrient Analysis Program (NFNAP) was implemented in 1997 as a collaborative food composition research effort between USDA and NIH. The goal of this program is to obtain nationally representative estimates of the nutritional components of important foods consumed in the US for inclusion in the USDA National Nutrient Databank System; to date, analytical food composition data generated for over 1800 foods have vastly improved overall data quality in the database. The NFNAP sampling approach was updated in 2001 using 2000 US Census data and recently updated to use 2010 Census population estimates. This design, like the 2001 design, employs a three-stage, stratified, probability-proportional-to-size (PPS) sample selection process; 1) county selection (based on population density); 2) supermarket outlets within selected counties (based on annual sales); and 3) specific brands of foods (based on market share data). In the first stage, Census regions (4), divisions and states were used to obtain a self-weighting sample of population centres, ensuring geographic dispersion across the 48 conterminous states; 48 locations were selected, with nested subsets of 24, 12 and 6 locations. Due to demographic changes in the population and congressional redistricting it was necessary to revise the sampling scheme to reflect these changes. With the increased penetration of warehouse-type retail outlets into the grocery industry, the sampling frame must be adjusted to include these purchase locations. Food samples which are collected nationally according to a statistically rigorous sampling approach are consistent with national representativeness and allow better estimates of the mean and variability than convenience sampling or less rigorous options.

Selection and peer-review under responsibility of National Nutrient Databank Conference Steering Committee

Keywords: National sampling; food composition

1. Introduction

The US Department of Agriculture's (USDA) Nutrient Data Laboratory (NDL) develops high quality food composition databases for foods available in the US food supply. This paper describes the second revision of the National Food and Nutrient Analysis Program (NFNAP) sampling plan, implemented in 2012, for the national collection of food samples from retail outlets for nutrient analysis. A more detailed history of this program, implemented in 1997, and applications of the sampling approach and NFNAP in general are presented in earlier NDL publications (1,2,3,4). The basic objectives of NFNAP are to secure reliable estimates with known variability for the nutrient content of food and beverages consumed by the US population.

The initial and subsequent updated sampling plans are based on a stratified three-stage design using the most current population density data from the US Bureau of the Census and food sales data for retail

* Corresponding author. Tel.: +1-301-504-0693; fax: 1-301-504-0692.
E-mail address: pamela.pehrsson@ars.usda.gov

Selection and peer-review under responsibility of National Nutrient Databank Conference Steering Committee
doi:10.1016/j.profoo.2013.04.009

outlets in selected locations and product market shares, both from ACNielsen, Inc. Selection of locations (population density), retail outlets (sales), and specific brands (market share data) are selected probability-proportional-to-size, so that any county, store or brand in these three selection levels has a chance of being selected; the greater the proportion to the total, the greater the probability of being selected. The primary focus of this paper will be the first stage, selection of locations. The sampling plan provides a self-weighting nationally representative sample set of the food products.

2. Methods

2.1. Chromy's PMRPPS Procedure

Chromy's algorithm, a probability minimum replacement (PMR) probability proportional to size (PPS) sampling scheme, was again used to select a stratified sample of counties to purchase foods for nutrient analysis (5,6,7). A sequential sampling scheme considers a frame's sampling unit in a predefined order. PMR sample designs are PPS designs which allow some sampling units to be selected more than once. Chromy's procedure identifies the following elements:

n(i) = number of times unit i is selected in sample

n = sample size

S(i) = size measure for sample unit i

S(+) = sum of size measures for all units in frame

q(i) = E[n(i) = nS(i)/ S(+)]

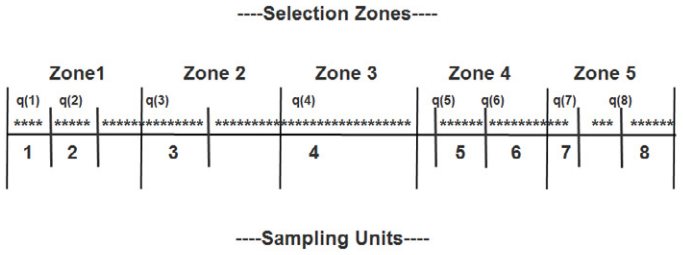

Fig.1.Chromy's PMRPPS Sampling Procedure Designed to Choose 5 Sampling Zones from Among 8 Population Units (Cities or Counties)

This procedure divides the ordered frame into *n* zones of size *S*(+)/*n*. One sampling unit is selected PPS from each zone and each unit I has a line segment of length *q(i)* associated with it; each line segment either falls entirely within one sampling zone or overlaps two or more zones. Figure 1 illustrates an example where a sample of size five is to be drawn from eight available sampling units. If *q(i)* is greater than one, then sampling unit i completely covers one or more zones and is considered a self-representing unit (unit 4 in Figure 1). These units will appear in the sample at least one time. If a unit is in part of two adjoining zones but is not self-representing, it can be selected in one of the two sampling zones but not both (units 3 and 6 in Figure 1). When a single unit is selected within each zone, the sample is stratified by the ordering of the frame; the ordering considers control variables highly correlated with the quantity being measured so that conterminous units are similar. In other words, the variance is reduced as long as units in close proximity are more homogenous than units in the population at large.

2.2. Objectives for County Selections

In the 2012 design using 2010 data similar to previous designs, the selected counties are not only geographically dispersed across the nation and regions but are statistically representative with respect to

county size and the CBSAs (Core based Statistical Areaa) for the nation and regions. According to the OMB a CBSA is a collective term for both metro and micro areas. A metro area contains a core urban area of 50,000 or more population, and a micro area contains an urban core of at least 10,000 (but less than 50,000) population. Each metro or micro area consists of one or more counties and includes the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core (8). For the purposes of the NDL, generalized CBSAs (gCBSAs) were employed. For counties in a CBSA the gCBSA is defined as the CBSA. For counties not in a CBSA, the gCBSA is the county itself. The sampling approach is to sample in three stages. The first stage is to sample counties. To achieve this counties are sorted by region, within region by division, within division by state, within state serpentinely by gCBSA population size and within gCBSA serpentinely by urbanicity). At the second stage, a list of at least 10 grocery stores, each having annual sales of at least four million dollars will be created for each selected county. The third stage of the design will involve selecting product samples based on market share of the targeted commercially available food or distribution data.

The sampling design once again utilized is the probability minimum replacement (PMR) probability-proportional-to-size (PPS) sample selection based on Chromy's methodology. This methodology will attempt to satisfy the 5 criteria identified as necessary to ensure proper national representativeness of the sample. These criteria are:

1. The states containing sample counties should be geographically well dispersed regionally;
2. The gCBSA containing sample counties should be well dispersed when the gCBSAs are sorted by size regionally;
3. The sample counties should be well dispersed when the counties are sorted by size regionally;
4. The gCBSAs containing sample counties should be well dispersed when the gCBSAs are sorted by size nationally; and
5. The sample counties should be well dispersed when the counties are sorted by size nationally

To accomplish the sampling, the Census 2010 summary file was obtained (11). Some of the linkages required block level linkings to create the analytical dataset. Summary population values were used to validate that all linkages were correct. After construction of the dataset, Chromy's methodology was used to draw multiple samples of 24 counties using 500,000 iterations of the US population at the county level. To evaluate how well each candidate sample met the other four criteria, an "ideal" sample of size 24 was constructed for each of the four remaining criteria. Each ideal sample was constructed by sorting the population of counties to induce an implicit stratification to meet one of the four criteria listed below:

1. The sort for criterion 2 was by region, population size of gCBSA serpentinely within region, and urbanicity of county[b] serpentinely within gCBSA (10);
2. The sort for criterion 3 was by region and population size of county serpentinely within region;
3. The sort for criterion 4 was by population size of gCBSA and urbanicity of county serpentinely within gCBSA; and
4. The sort for criterion 5 was by population size of county.

To determine how nearly a candidate sample comes to satisfying any one of the criteria 2-5, the distribution of the candidate sample was compared to the distribution of the corresponding ideal sample. After exploring several alternatives, a version of Kolmogorov's D statistic based on centered quantiles was chosen to measure the similarity between the distribution of each candidate sample and that of each of the ideal samples.

Kolmogorov's D quantifies the similarity between two cumulative distribution functions (CDFs). Since the population was known, both distributions (the one for the candidate sample and the one for the population) were described by empirical CDFs (ECDFs). Note the ideal samples were precisely the population center quantiles used to define the ECDF of each ordering. The equivalent quantiles of the candidate sample were found by sorting it in the same order as the population was sorted to draw the ideal

---

[a] In previous papers we referred to Consolidated Metropolitan Statistical Area (CMSA) In 2003, the Office of the Management and Budget (OMB) created a new designation named Core Based Statistical Area (CBSA, reference 8).For more information on the differences between the CMSA and CBSA refer to the Office of Management and Budget "Standards for Defining Metropolitan and Micropolitan Statistical Areas" and the paper by Hamilton and Thrall (9).

[b] Urbanicity is a measure developed by Goodall, Kafadar and Tukey (1998, reference 10) to attempt to quantify the urban quality of an area. In this paper, we used urbanicity as a sort factor in counties and CBSAs to rank how urban or rural a particular place (city, town) is in a county

sample to which it is being compared. The two ordered samples were then paired and the absolute value of difference of the sample cumulative gCBSA (county) populations at each pair of observations was computed. The maximum of this set of absolute differences was used as the D statistic.

The overall D that was associated with each candidate sample was the maximum of the Kolmogorov's D statistics for the four individual criteria, which indicates the worst fit of the candidate sample to any of the four ideal samples. Since at any point along the serpentine ordering associated with criterion 1 the cumulative proportion of sample counties approximates the cumulative proportion of the population, the states containing the sample counties are geographically well dispersed regionally and nationally according to population size. The criteria used to determine how to make a decision on the sample are listed below in order of importance

1. Kolgomorov D
2. Relative Mean Differences between the ECDFs
3. Subject matter expertise
4. R2 values

3. Results

The QQ plot in Figure 2 compares the revised sample to the ideal sample associated with criteria 2-5. Figure 4 indicates that when the sample and population are sorted serpentinely by region according to gCBSA size the quantiles of the sample and the centered quantiles of ideal sample associated with criterion 2 lie approximately along the 1-1 line. Thus, it follows that the cumulative proportion of sample counties approximates the cumulative proportion of the population at any point along the ordering. Thus, ensuring the gCBSAs containing sample counties are well dispersed over the population when the gCBSAs are sorted by size regionally.

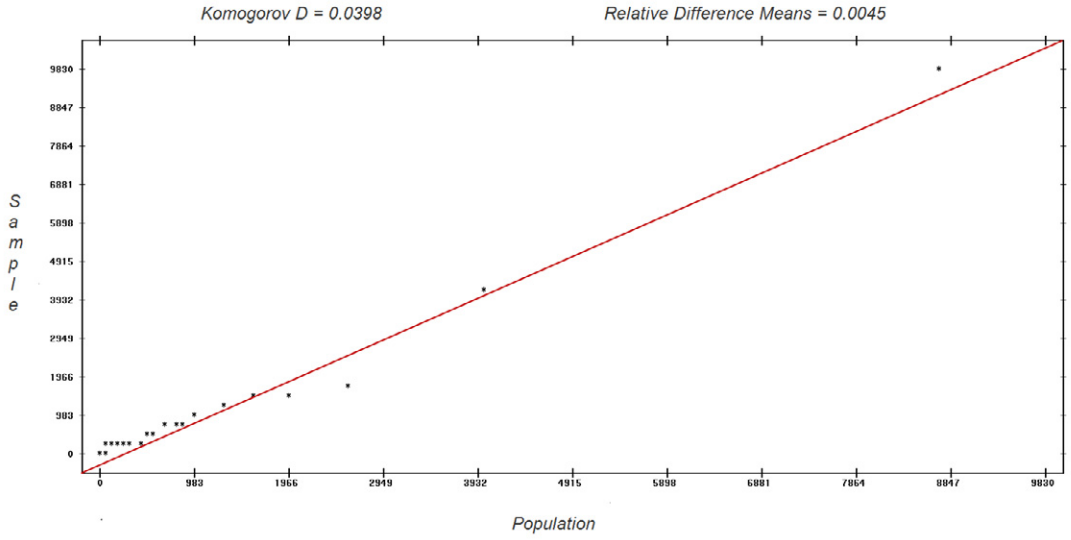

Fig. 2.QQ Plot of Sample vs. Ideal Sample by Regions for gCBSA Size

Figure 3 indicates that when the sample and population are sorted serpentinely by region according to county size the quantiles of the sample and the centered quantiles of ideal sample associated with criterion 3 lie approximately along the 1-1 line. Thus, it follows that the cumulative proportion of sample counties approximates the cumulative proportion of the population at any point along the ordering which means that the sample counties are well dispersed over the population when the counties are sorted by size regionally. After a small number of candidate samples were selected (5), the sample that had the lowest overall D (best fit) along with low relative mean differences and subject matter expert opinion was chosen as the revised NFNAP county sample. This was argued to be the most geographically well dispersed sample.

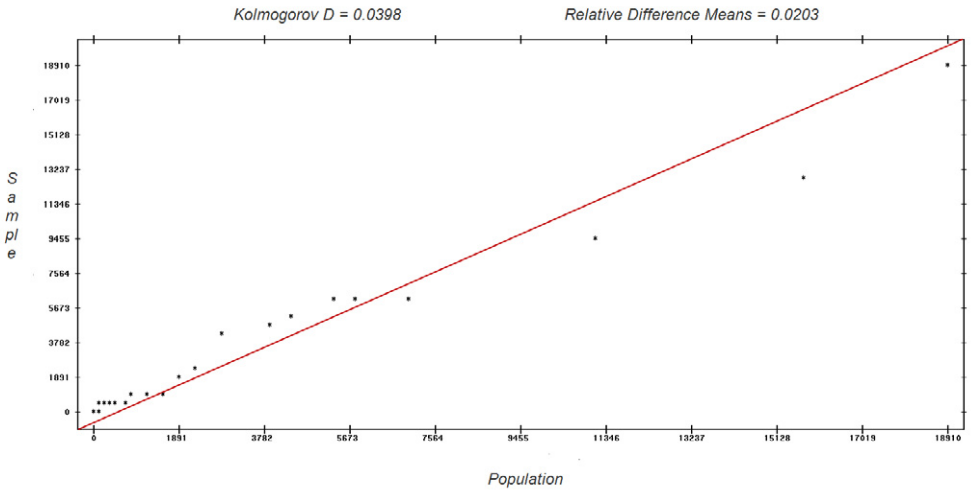

Fig.3.QQ Plot of Sample vs. Ideal Sample by Regions for County Size

With this systematic and methodologically rigorous approach foods collected from this sampling scheme will be representative of foods that are consumed by individuals in America. (The first stage of this NFNAP sampling plan, like previous plans, 24 counties will be selected by PPS). At the second stage of the revised plan, like at the second stage of the initial plan, a list of at least 10 grocery stores (outlets), each having sales of at least four million dollars per year, will be developed for each of the 24 selected counties. For counties having fewer than 10 outlets, adjacent counties will be added sequentially until the area contained a minimum of 10 outlets. Then two outlets, a primary and an alternate outlet, were selected PPS without replacement from each county's outlet list with size equal to the outlet's annual value of sales. During data collection, the alternate outlet for a county should be used when the primary is inaccessible or when a product is unavailable at the primary outlet.

At the third stage, as in previous cycles, two types of product samples will be selected. The primary product sample was drawn to support the estimation of the mean nutrient content of an average serving from composited samples of a product. The secondary product sample will be selected from the primary product sample to support the estimation of the variability of the nutrient content of a typical serving of a product. The secondary sample will also be used to develop models for the prediction of the serving-to-serving variability of nutrient content from the variability of composited samples.

The primary sample of food products (brands, varieties, etc.) for a particular food type are selected using Chromy's method from a list of all products that had been sorted in descending order by the amount of each product sold nationally. This would include brand market share for commercially prepared foods and restaurant foods, or other measures of proportion of the market such as cultivar, variety, etc. for commodity level foods. The number of samples chosen for each product shall be based on the desired statistical reliability and the number of nutrient analyses NDL could afford to perform. The selected products will be purchased from each of the primary outlets unless the product is not available or the primary outlet is inaccessible. In that case, the product will be purchased from the alternate outlet. Figure 4 and Table 1 show the county locations of the most recently developed NFNAP sampling design, which is based on the 2010 US Census data.

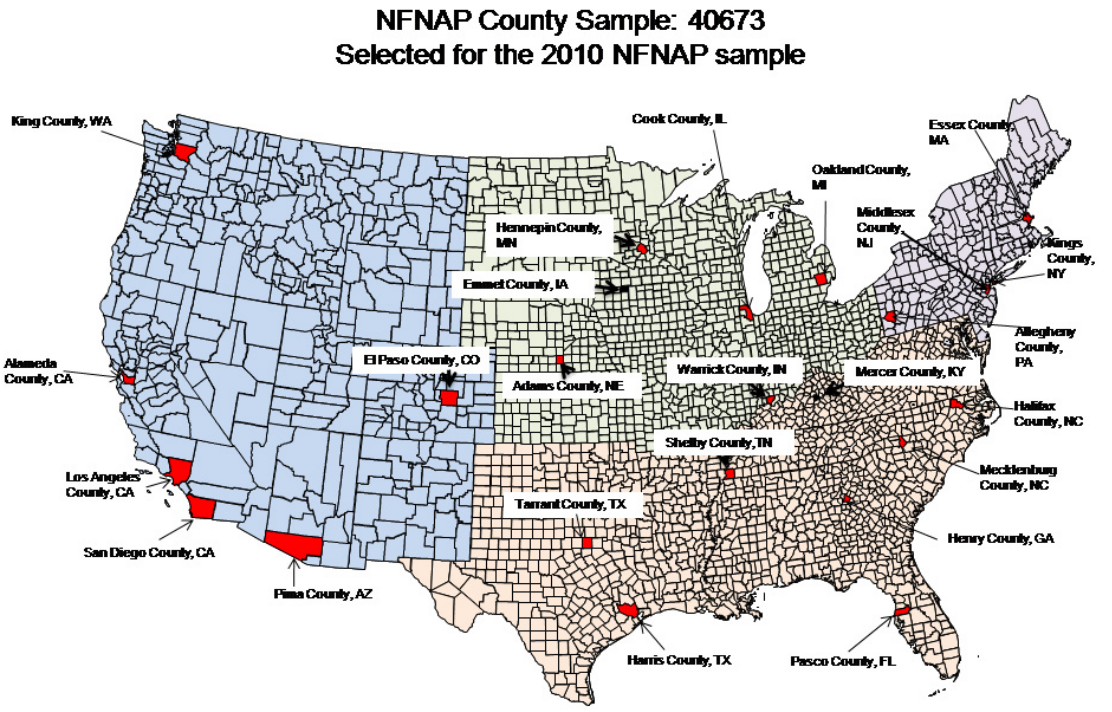

Fig.4. Selected NFNAP Counties

The selected products (brands) for a particular food item can be thought of as a matrix with the selected outlets (locations) as columns and the product samples (which correspond roughly to brands) as the rows. Composites are formed by product sample across locations as shown in Figure 1. Forming the composites in this manner will provide individual product (i.e., brand) data for major brands which permits future updates when brand composition changes. It is important to note that results from composites pertain to an average serving from the homogenized food product, not to a typical serving. A more detailed discussion of compositing options are defined by Perry et al. (1,2).

Applications

With racial and ethnic minorities comprising about 35% of the US population, and 92% of population growth over past decade attributed to these populations, many cities are "majority" minority (11). For example, Latinos comprise 16.3%, African Americans 12.2%, Asian/Asian Indian Americans, 4.8% and American Indians/Alaska Natives, others about 1.9%. For this reason, sampling plans for foods consumed by unique populations with specific cuisines are modeled after the NFNAP sampling plans. Similar approaches have been developed in the past decade for sampling foods on American Indian and Alaska Native reservations, using self-enrollment data from the Bureau of Indian Affairs. (www.bia.gov)

Table1. Selected Counties for 2012–2022 NFNAP Sampling

County Name	County Population
Essex County, MA	743,159
Middlesex County, NJ	809,858
Kings County, NY	2,532,645
Allegheny County, PA	1,223,348
Cook County, IL	5,217,080
Warrick County, IN	59689
Oakland County, MI	1,202,362
Emmet County, IA	10,302
Hennepin County, MN	1,152,425
Adams County, NE	33,185
Pasco County, FL	464,697
Henry County, GA	203,922
Halifax County, NC	57370
Mecklenburg County, NC	944.373
Mercer County, KY	9725
Shelby County, TN	927,644
Tarrant County, TX	1,809,034
Harris County, TX	4,140,894
Pima County, AZ	980,263
El Paso County, CO	622,263
Los Angeles County, CA	9,818,605
Alameda County, CA	1,510,271
San Diego County, CA	3,095,313
King County, WA	1,931,249

4. Conclusions

To date, around 1800 foods have been sampled and analyzed under the NFNAP, representing almost 100,000 values in the USDA National Nutrient Database for Standard Reference. These data are available at the Nutrient Data Lab's website: http://www.ars.usda.gov/nutrientdata. In summary, NFNAP data are current and nationally representative largely due to the sampling design. USDA food composition data, which include means and estimates of variability for nutrients and serving sizes as well as other supportive information, are used by a broad spectrum of researchers, consumers, members of the food industry and nutrition policy makers. In research, they improve the ability to detect etiologic relationships, delineate biologic mechanisms, assess time trends in nutrient intakes and define populations at risk for poor nutritional status.

References

[1] Perry CR, Beckler DG, Pehrsson P, Holden J. 2000. A National Sampling Plan for Obtaining Food Products for Nutrient Analysis. 2000 Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association: p. 267-272.

[2] Perry CR, Pehrsson PR, Holden J, 2003. A Revised Sampling Plan for Obtaining Food Products for Nutrient Analysis. 2003 Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association, CD-ROM.

[3] Pehrsson PR, Haytowitz DB, Holden JM, Perry CR, Beckler DG, 2000. "USDA's National Food and Nutrient Analysis Program: Food Sampling", J Food Comp Anal, vol. 12: p. 379-389.

[4] Pehrsson, P., Perry, C., Cutrufelli, R., Patterson, K., Wilger, J., Haytowitz, D., Holden, J., Day, C., Himes, J., Harnak, L., Levy, S., Wefel, J., Heilman, J., Phillips, K., Rasor, A. 2006. Sampling and initial findings for a national study of fluoride in drinking water. J Food Comp Anal, vol. 19:p. S45-S52.

[5] Chromy JR, 1979. "Sequential Sample Selection Methods". 1979 Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association: p. 401-406.
[6] Williams RL, Chromy JR, 1980. "SAS Sample Select MACROs", Proceedings of the Fifth Annual SAS Users Group International Conference, Cary, NC: SAS Institute, Inc.: p. 392-396.
[7] Chromy JR, 1981. "Variance Estimators for a Sequential Sample Selection Procedure", in Krewski D, Current Topics in Survey Sampling. Academic Press: p. 29-347.
[8] Office of Management and Budget. "Standards for Defining Metropolitan and Micropolitan Statistical Areas." Federal Register 249, no. 65 (2000): 82,228.
[9] Hamilton B, Thrall GI. "Getting to the Core of CBSAs", Geospatial Solutions, vol 14: p. 48- 51
[10] Goodall CR, Kagadar K, Tukey JW, 1998. "Computing and Using Rural versus Urban Measures in Statistical Applications", American Statistician, vol. 52:p 101-111.
[11] US Census Bureau, 2010 Summary File 1 Technical manual, available at: http://www.census.gov/prod/cen2010/doc/sf1.pdf, accessed on October 2, 2011.

Presented at NNDC (March 25-28, 2012 – Houston, TX) as Paper #3, Session 2 "Food Composition Databases"