A Revised Sampling Plan for Obtaining Food Products for Nutrient Analysis under the National Food and Nutrient Analysis Program: Selection of Sampling Locations

Charles R. Perry, Jr.[1], Marlon G. Daniel[1], Pamela R. Pehrsson[1]
[1]USDA, Agricultural Research Service, Nutrient Data Laboratory
Room 105, Building 005, BARC-West, 10300 Baltimore Avenue
Beltsville, MD 20705

Abstract

The Nutrient Data Laboratory (NDL) of the United States Department of Agriculture (USDA) implemented the National Food and Nutrient Analysis Program (NFNAP) in 1997. The goal of this program was to obtain nationally representative estimates of the nutritional components of common foods consumed in the United States for inclusion in the USDA National Nutrient Databank System. In 2001 the initial design was updated, employing a three-stage, stratified, probability-proportional-to-size (PPS) sample selection process. Using this method resulted in a self weighting sample of population centers, ensuring geographic dispersion across the 48 contiguous states.

With demographic shifts in the population it was necessary to revise the county sampling methodology to reflect these changes in 2010. With the increased penetration of warehouse-type retail outlets into the grocery industry, it was also necessary to update the grocery store sampling methodology to include warehouse type retail purchases. These updates will ensure that the estimates of the nutrient means and variability derived from food samples collected under the new sample design are representative of the foods consumed in the U.S.

A sample of 24 counties was derived by first selecting a large number of county samples using data from the 2010 U.S. Population Census. Then a highly representative sample was selected using several goodness of fit criteria. The final sample of 24 counties simultaneously ensured that Kolmogorov's D statistics and the relative mean difference between the population quantiles and sample quantiles was less than 5% for five distribution objectives.

Key Words: Controlled Sampling, Chromy's PMRPPS Procedure, Kolmogorov's D

1. Introduction

The U.S. Department of Agriculture's Nutrient Data Laboratory (NDL), a division of the Agricultural Research Service (ARS), develops databases and methodologies to evaluate and disseminate composition data on foods consumed in the United States (U.S.). This paper describes the revised National Food and Nutrient Analysis Program (NFNAP) sampling plan, which will be implemented in the fall of 2013, for the collection of food samples from retail outlets for nutrient analysis. Data that is collected from this program

are used by researchers, nutrition public policy developers, the food industry and a large consumer base.

In 1997, NDL inaugurated the NFNAP; the main goal of which was to obtain nutrient estimates with known variability for foods and beverages consumed by the U.S. population (Perry, et al., 2000; Pehrsson, et al., 2000; Haytowitz, et al., 2002). The program was built on five primary objectives which provide the framework for the continued sampling and analysis of key foods in the food supply. The first objective of NFNAP was to identify one thousand key foods contributing critical nutrients to the U.S. food supply. The second objective was to evaluate the quality of existing data on these foods and nutrients. The third objective was to develop a sampling plan for the collection of a representative sample of the foods consumed by the U.S. population. The fourth objective was to conduct nutrient analysis on the collected food samples under USDA-supervised contracts. The fifth objective was to disseminate the results from these analyses after quality reviews. This resulted in the selection of 24 counties within the contiguous United States (See Figure 1). The sampling plan used to collect food samples for analysis was based on a stratified three-stage design using the most current population projections (1997) from the U.S. Bureau of the Census and food product market share data from A.C. Nielsen, Inc.

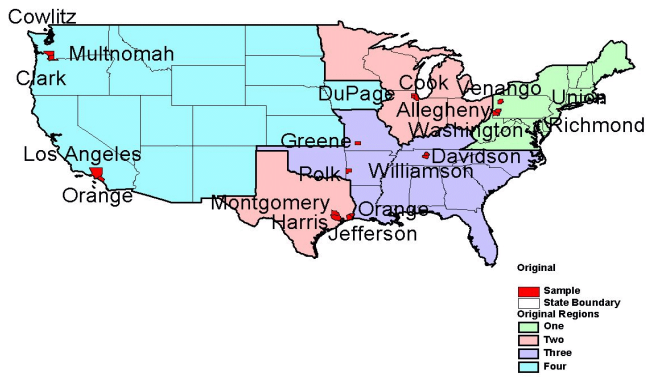

Figure 1: Original NFNAP Regions and County Sample (1997 design)

In 2000, the sampling plan was revised. The new procedure used a controlled sampling methodology and incorporated Chromy's Probability Minimum Replacement Population Proportional to Size (PMRPPS) sampling procedure (Chromy, 1979; Williams and Chromy, 1980; Chromy, 1981). This revision, which was similar to the initial design, used a three-stage, self-weighting selection procedure where counties were selected at the first stage, grocery store outlets are selected at the second stage, and specific food products to be purchased for nutrient analyses were selected at the third stage. Figure 2 shows the counties that were selected in 2000. Unlike the initial design, this revision

incorporated the 2000 Census Bureau regions, divisions, and states into the first stage sample selection process.

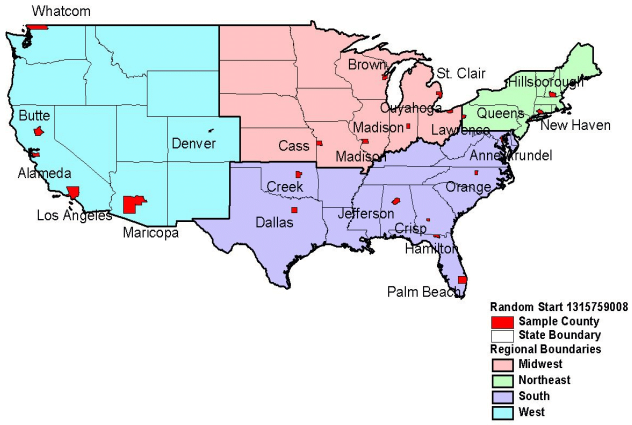

Figure 2: NFNAP Regions and County Sample (2000 design)

With demographic shifts in the population it was necessary to revise the county sampling methodology to reflect these changes in 2010 (Pehrsson et al 2013). With the increased penetration of warehouse-type retail outlets into the grocery industry, it was also necessary to update the grocery store sampling methodology to include warehouse type retail purchases. These updates will ensure that the estimates of the nutrient means and variability derived from food samples collected under the new sample design are representative of the foods consumed in the U.S.

In the 2010 design, the goal was not only to have the selected counties geographically dispersed across the nation and regions according to the 2010 Census Bureau summary file data (2010 Census Summary File 1), but also to be statistically representative with respect to both the county sizes and the Core Based Statistical Areas (CBSAs) of the nation and regions. A CBSA is a statistical geographic entity consisting of the county or counties associated with at least one core (urbanized area or urban cluster) of at least 10,000 individuals, plus adjacent counties having a high degree of social and economic integration with the core as measured through commuting ties with the counties containing the core (Office of the Management and Budget, 2013). Incorporating the Census regions, divisions, and states into the sample selection process as implicit stratifiers will facilitate analyses at different geographic Census levels.

Section 2 describes Chromy's PMRPPS sample selection. Section 3 describes the objectives for the revised sampling plan. Section 4 describes the revised county sampling plan. Section 5 describes the revised outlet and product sampling plan. Section 6 provides summary comments and conclusions.

2. Chromy's PMRPPS Procedure

Chromy's algorithm, a sequential, probability minimum replacement sampling scheme, was used to select a stratified sample of counties in which to purchase foodstuffs for nutrient analysis for the NFNAP. A sequential sampling scheme considers a frame's sampling units in a predefined order. PMR sample designs are PPS designs that allow some sampling units to be selected more than once. Let:

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*(+)

The Chromy procedure divides the ordered frame into n zones of size S(+)/n. One sampling unit is selected from each zone with probability proportional to size. Associated with each unit i is a line segment of length q(i), which either falls entirely within one sampling zone or overlaps two or more zones. Figure 3 illustrates the procedure for a hypothetical case where a sample of size five is to be drawn from eight available sampling units.

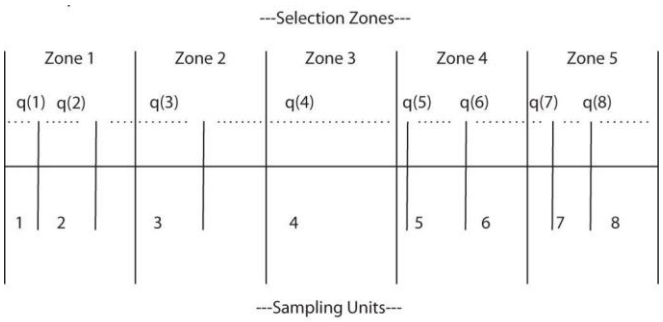

Figure 3: Chromy's PMRPPS Sampling Procedure

If q(i) exceeds one, then sampling unit i covers one or more zones completely and is known as a self-representing unit (e.g., unit 4 in Figure 2). Such units are guaranteed to appear in the sample at least once. If a unit is in part of two adjoining sampling zones but is not self-representing (units 3 and 6 in Figure 2), then it can be selected in one of the two zones but not both. Selecting a single unit from each zone, ensures that the sample is implicitly stratified by the frame ordering. The variance is reduced as long as units in close proximity are more homogeneous than those in the population at large. The frame is ordered using control variables highly correlated with the quantity being measured so that neighboring units are similar.

Chromy's PPS PMR sample selection procedure can also be visualized by observing how the cities are ordered in the list and then thinking of them as strung out proportional to their size on a ruler having 24 equal size units. For example, on a two foot ruler which has 24 one inch long units, each unit interval represents 1/24 of a population (see Figure 3). Then one spot/location on the ruler is randomly selected from each of the 24 inch units on the ruler. The 24 randomly selected spots, one from each one inch interval on the ruler, correspond to one of 24 sample locations/cities. Using Chromy's selection

procedure ensures that any location that crosses one or more of the one inch boundaries is selected no less or no more often than once from its rounded interval length over all and its probability of selection is proportional to its population size. Thus, it should be clear that Chromy's PPS PMR sample procedure is just a generalization of PPS systematic sampling that allows a PPS random sample to be selected in each interval with the sampling units falling on one or more interval boundaries to be selected the appropriate number of times overall.

3. Objectives for Selection of Counties

As in the previous sample design cycle, the first stage of the NFNAP sampling plan was updated. An extensive list of options was explored in order to ensure the sample is well dispersed nationally and regionally. The revised county sample selection procedures resulted in a PPS sample of counties that satisfied, to the extent possible, each of the following five criteria.

1. The states containing sample counties should be geographically well dispersed regionally (over the four U.S. Census regions) and nationally (over the 48 contiguous states). That is, when the states are sorted in the usual serpentine Census order, the cumulative proportion of the sample counties approximate the cumulative proportion of the population at any point along the ordering.
2. The gCBSAs containing sample counties should be well dispersed when the gCBSAs are sorted by size regionally. That is, the cumulative proportion of sample counties approximated the cumulative proportion of the population at any point along the ordering.
3. The sample counties should be well dispersed when the counties are sorted by size regionally.
4. The gCBSAs containing sample counties are well dispersed when the gCBSAs are sorted by size nationally.
5. The sample counties are well dispersed when the counties are sorted by size nationally.

The counties were first ordered by the size of the gCBSA containing the county. Then, a sample of counties was chosen using any of a number of PPS sequential zonal sampling procedures. The resulting sample was highly representative with respect to gCBSA sizes. However, there is no guarantee that the sample will be geographically well-dispersed across the U.S. or representative with respect to U.S. county population sizes. The goal was to satisfy multiple criteria concurrently but the sorting procedure only allows us to control for one criterion at a time. To obtain a sample that approximately satisfied all of the criteria, a random search was performed. This was done by first drawing a large number of samples that satisfied one criterion. Each of these candidate samples was then compared to a set of ideal samples, using goodness of fit measures, to find a sample that nearly satisfied the remaining criteria. In many respects, the objective of this approach was the same as re-weighting using generalized regression or calibration that is commonly employed to ensure that the weighted sample represents the population with respect to a set of control variables. However, one advantage of controlled sampling procedures versus re-weighting is under many circumstances, controlled sampling can be used to produce a highly efficient self-weighting sample satisfying multiple criteria.

4. Current County Sampling Plan

This section describes the implementation of the second method discussed in section 3. The method allows all five of the criteria described in that section to be approximately met simultaneously in a self-weighting sample.

Candidate samples satisfying criterion 1 were obtained as follows:

1. The counties were sorted by Census region, within region by division, within division by state, within state serpentinely by gCBSA population size, and within gCBSA serpentinely by urbanicity, and;
2. Chromy's method was used to draw candidate samples of size 24. Each candidate sample satisfied criterion 1. That is, since Chromy's method divides the counties along the serpentine ordering into equal population size zones (implicit strata) and selects one county from each zone with probability proportional to size, the cumulative proportion of the sample counties at any point along the ordering is approximately the same as the cumulative proportion of the population. This sampling procedure ensures that the counties of each candidate sample are geographically well dispersed across regions, divisions, and states.

To evaluate how well each candidate sample met the other four criteria, an "ideal" sample of size 24 counties was constructed for each of the four remaining criteria. Each ideal sample was obtained by sorting the population of counties to induce an implicit stratification to meet one of the four criteria:

1. The sort for criterion 2 was by region, population size of gCBSA serpentinely within region, and urbanicity of county serpentinely within gCBSA;
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.

Conceptually, to draw an ideal sample by gCBSA within regions, the gCBSA are sorted by Census region and within regions the gCBSA are sorted serpentinely by population size. If the gCBSAs of a region were sorted in increasing order, the gCBSAs of adjacent regions were sorted in decreasing order and vice versa. Within gCBSAs, the counties were sorted serpentinely by urbanicity. The county containing the 24 quantile centers were selected as the ideal sample. Thus, the ideal sample corresponds to the centers, with respect to cumulative population size, of the 24 zones for Chromy's PMRPPS zonal sample. The other ideal samples were drawn in a similar manner. Figure 4 displays the location of the final sample of 24 counties in the four census regions of the contiguous United States.

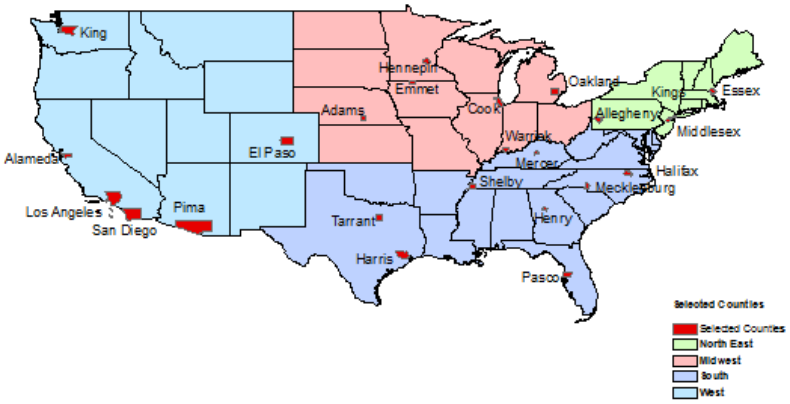

Figure 4: Regions and Revised NFNAP County Sample (2010 design)

To determine how well a candidate sample came 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. 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). 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 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 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.

1. Kolmogorov's D
2. R^2 values
3. Relative Mean Differences between the eCDFs
4. Subject matter expertise

The QQ plot in Figures 5-8 compare the revised sample to the ideal samples associated with criterion 2-5.

Figure 5 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.

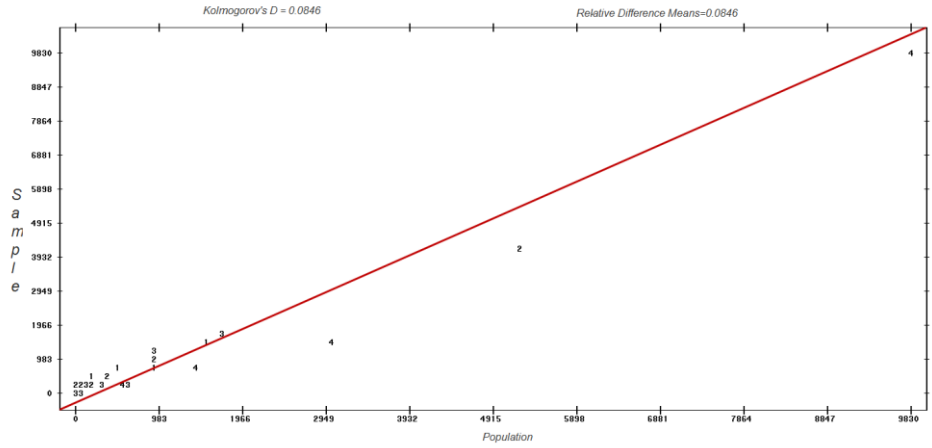

Figure 5: QQ Plot of Sample vs Ideal Sample for gCBSA by Regions

Figure 6 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.

Figure 7 indicates that when the sample and population are sorted by gCBSA size the quantiles of the sample and the centered quantiles of the ideal sample associated with criterion 4 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 gCBSAs containing sample counties are well dispersed over the population when the gCBSAs are sorted by size.

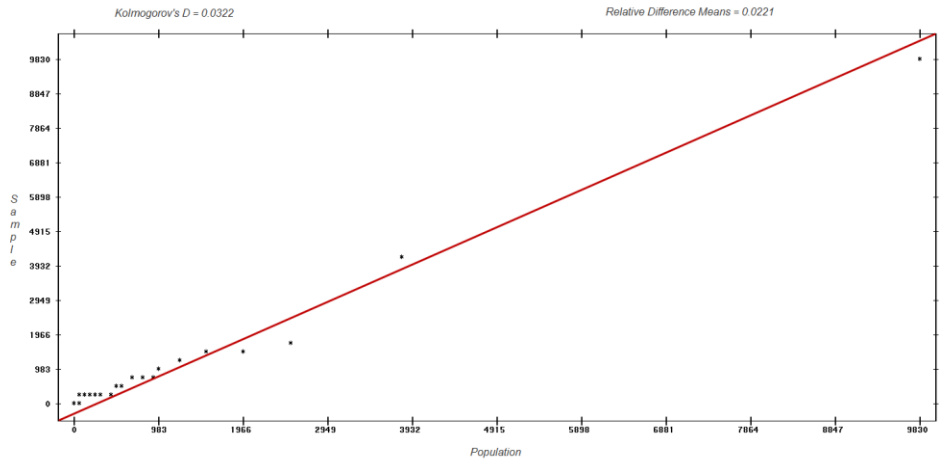

Figure 6: QQ Plot of Sample vs. Ideal Sample for County Size by Regions

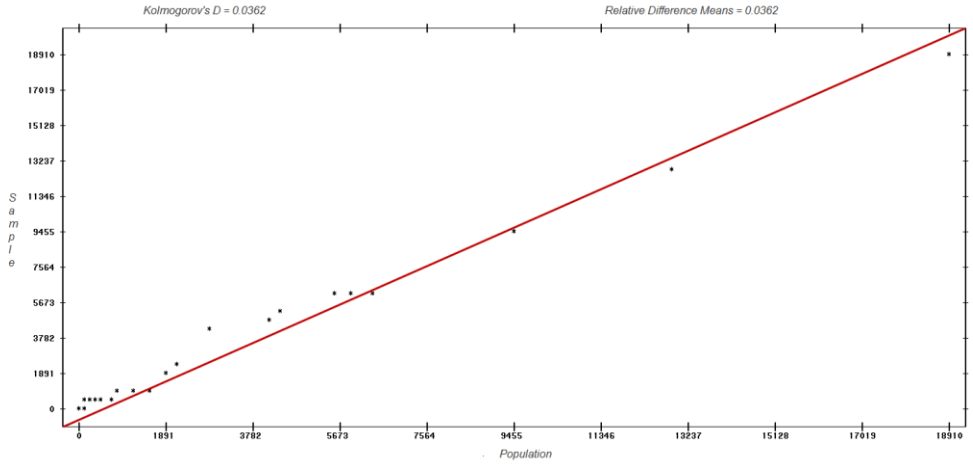

Figure 7: QQ Plot of Sample vs. Ideal Sample for gCBSA Size overall

Figure 8 indicates that when the sample and population are sorted by county size the quantiles of the sample and the centered quantiles of the ideal sample associated with criterion 5 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 population size.

Thus, it follows that the cumulative proportion of sample counties approximates the cumulative proportion of the population at any point along the ordering, ensuring the sample counties are well dispersed over the population when the counties are sorted by size regionally.

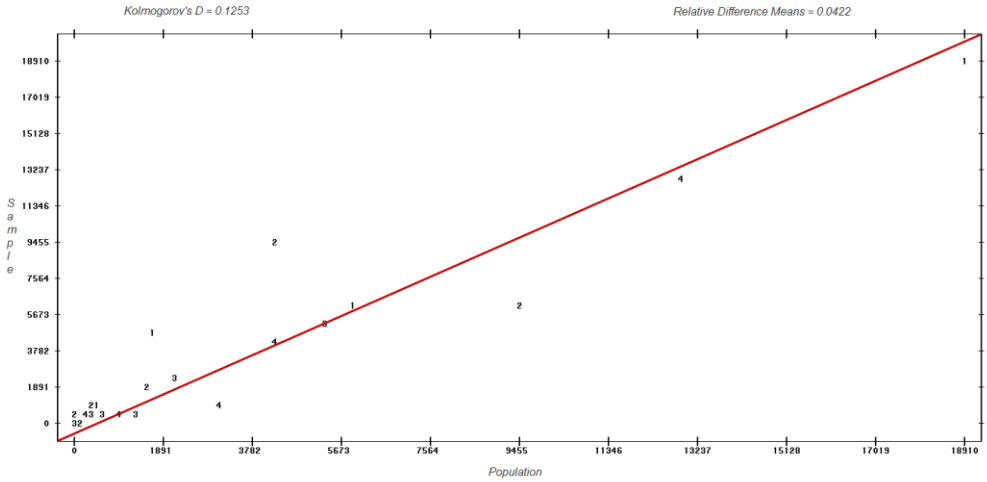

Figure 8: Overall County QQ Plot of Sample vs. Ideal

When the sample and population are sorted by gCBSA size, the quantiles of the sample and the centered quantiles of the ideal sample associated with criterion 4 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 gCBSAs containing sample counties are well dispersed over the population when the gCBSAs are sorted by size.

When the sample and population are sorted by county size, the quantiles of the sample and the centered quantiles of the ideal sample associated with criterion 5 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 population size. Therefore, the revised NFNAP sample satisfies criterion 1 and simultaneously approximately satisfies criteria 2-5.

Then a small number of candidate samples (5) were selected; that is, samples having the lowest overall Ds (best fit) and lowest relative mean differences. The final NFNAP county sample was selected from the 5 candidate samples by NDL nutrition researchers based on their subject area expertise as the most geographically well dispersed sample, relative to a number of socio-demographic criteria. Some of the criteria were age, race-ethnicity and income that were not included in the formal statistical sampling procedures.

The process of sub-sampling the 24 counties, can be conceptualized as an extension of the visual representation of the initial sampling process: 1) first one spot is randomly selected in each of the 24 one inch intervals, which have been grouped into 12 consecutive pairs, and then 2) one of the two spots in each consecutive pairs is randomly selected. The 12 spots correspond to the 12 sub-samples. Since one could have equivalently obtained the sub-sample of 12 cities by randomly locating a spot in each of the 12 consecutive interval pairs, it is clear the selected sub-sample has essentially the

same characteristics as the initial Chromy's sample. This process can also be used to obtain a sub-sample of size 6.

5. Revised Outlet and Product Sampling Plan

In the current sampling design, 24 counties were selected PPS. This was similar to previous sampling plans executed at NDL. At the second stage, a list of grocery stores (outlets), each having sales of at least $4 million dollars per year, was developed for each of the 24 selected counties. (See Table 1.)

In 2000, 70.9% of food sales in the American home came from conventional supermarkets while 7.2% of food sales came from warehouse club stores and supercenters. In 2010, 64.4% of food sales came from conventional supermarkets while 16.1% came from warehouse clubs and supercenters. With the changes in the purchasing patterns of the American household, it becomes imperative to include warehouse clubs and supercenters in our current sample of outlets.

Table 1. Listing of Counties and Outlet Types (2010 design)

Location	Outlet Type
Adams County, NE	Supermarket-Conventional
Alameda County, CA	Supermarket-Conventional
Allegheny County, PA	Supermarket-Conventional
Cook County, IL	Supermarket-Limited Assortment
El Paso County, CO	Supermarket-Conventional
Emmet County, IA	Supermarket-Conventional
Essex County, MA	Supermarket-Conventional
Halifax County, NC	Supermarket-Conventional
Harris County, TX	Supermarket-Conventional
Hennepin County, MN	Supermarket-Conventional
Henry County, GA	Supermarket-Conventional
King County, WA	Supermarket-Conventional
Kings County, NY	Supermarket-Conventional
Los Angeles County, CA	Supermarket-Conventional
Mecklenburg County, NC	Supermarket-Conventional
Mercer County, KY	Supercenter
Middlesex County, NJ	Supermarket-Conventional
Oakland County, MI	Supermarket-Conventional
Pasco County, FL	Supermarket-Conventional
Pima County, AZ	Supercenter
San Diego County, CA	Supermarket-Conventional
Shelby County, TN	Supermarket-Conventional
Tarrant County, TX	Supercenter
Warrick County, IN	Supermarket-Conventional

After selecting the 24 counties, a sample of primary outlets, one from each county, was selected. Five PPS samples were drawn using different random starts. The marginal distributions with respect to the outlet types were then computed for each sample. The sample that contained the distribution most similar to USDA Economic Research Service (ERS) data on food sales in the United States was selected as the primary sample. ERS data shows that 16.0% of household food sales came from warehouse clubs

and supercenters in 2011. (Table 2.)

Table 2. Comparison of Selected Sample Outlet type distribution and Economic Research Service Estimates

Outlet type		Selected Sample	ERS Data 2011	ERS Data 2001	ERS Data 1991
Supermarket-Conventional		79.17%	63.8%	70.1%	63.5%
Warehouse Clubs and Supercenters		20.83%	16%	8.9	2.1%
Other Stores		-	20.2%	21.0%	34.4%

Based upon the trends in household shopping, assuming the rate of change remains constant, we estimated that household food items purchased at supercenters and warehouse clubs will increase by 0.8% per year. By projecting out to halfway through the next sampling cycle (5 yrs), it was estimated that food sales from supercenters and warehouse clubs would be about 4% higher at that point in time, or about 20%.

Since very specific food items are to be sampled, it is likely that some food items may not be available in all primary food outlets. Thus, two alternate outlets were selected for each primary outlet to minimize non-response. The procedure used to select the alternate outlets in each county was: 1. Calculate the distance between each outlet and the primary outlet, 2. Sort the outlets in ascending order by the distance from the primary outlet, 3. Select the two outlets with the most similar sales volume within the county to the primary outlet. The closest outlet to the primary outlet was designated the first alternate and the second closest outlet was designated the second alternate outlet. This can be visualized by considering concentric circles around the primary outlet. After an outlet chain (brand) was selected, further outlets from that chain were not selected where possible. The alternate outlets are used when the food items are not carried by the primary outlet. In cases where a county only had two outlets, all possible outlets were selected.

6. Food Sample Compositing

To reduce the program costs, the individual food samples are combined into a small number of composites for nutrient analysis. Using composites allows a smaller number of food samples to be analyzed while still maintaining the ability to compute appropriate variance estimations. To obtain variance estimates, the food samples are randomly grouped into composites. Putting the food samples randomly into composites allows one to obtain estimates for both the variance of the mean and of individual servings.

The food samples collected from the selected locations are randomly grouped into a small number of composites as described below for laboratory analysis of the nutrient content of an average serving of the product. Two samples are taken from each of the composite blends. The first sample from each composite blend is used for laboratory analyses to determine the mean nutrient content of the food and prediction of serving to serving variability of the nutrient content of the sampled product. The second sample is stored as a backup for any additional future laboratory analyses that might be required.

Assuming perfect blending of each composite, the random group method has one major advantage over other methods of combining the individual samples into composites for

nutrient analysis. First, regardless of how the composites are formed, so long as each one contains an equal number of individual food samples, the sample mean of the composite nutrient analyses provides an estimate of the population nutrient mean for the food and its standard error can be computed. Second, unless the random group method is used to form the composites, it is not possible to compute an estimate of the serving-to-serving standard deviation of the nutrient content of the food. However, when the random group method is used to form the composites, one can also obtain a rough estimate of the serving-to-serving standard deviation of the nutrient content of the food for an average serving by multiplying the estimated standard error of the mean by the square root of the random group size. (See Appendix 1 on calculations and proofs.) These estimates will then allow for variance estimation of individual foods when outlets are randomly sub-sampled to 12 units and 6 for purposes of compositing.

7. Summary and Conclusions

Several summary and analysis options are available under this new design. For every food sampled under this design, composite nutrient means will be determined for each brand across locations. For every food, the standard errors of the composite means will provide estimates of the variability of nutrients among brands. For each food that is a significant contributor of nutrients of public health interest, secondary (non-composited) samples were used to determine between serving (serving-to-serving) nutrient variability. This process allowed within product variability to be factored out using unbalanced nested mixed model analysis of variance models (Littell, et al., 2006).

In summary, the procedure described in this paper resulted in a self-weighting set of sample locations that are geographically dispersed with respect to state population size, gCBSA population size, and to county population size, both overall and within census regions. This approach allows NDL to determine reliable estimates of the mean nutrient content of the most important foods consumed in the U.S. In addition, the revised sampling plan provides information on nutrient variability associated with health, and an efficient, cost-effective model for continuing sampling on a multi-year basis.

8. References

Chromy, JR. (1979), “Sequential Sample Selection Methods,” 1979 Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association: pp. 401- 406.

Chromy, JR. (1981), “Variance Estimators for a Sequential Sample Selection Procedure,” in Krewski, D., Current Topics in Survey Sampling, Academic Press, pp. 329-347.

Goodall, CR, Kafadar, K. Tukey, JW. (1998), “Computing and Using Rural versus Urban Measures in Statistical Applications,” American Statistician, Vol. 52, No. 2, pp. 101-111.

Haytowitz DB, Pehrsson PR, Holden JM. (2002), “The Identification of Key Foods for Food Composition Research,” J Food Comp Anal, Vol.15, No. 2, pp.183-194.

Hansen, MH, Hurwitz, WN, Madow, WG. (1953), Sample Survey Methods and Theory: Volume I and II, NewYork, NY: John Wiley and Sons, Inc.

Littell, RC, Miliken, GA, Stroup, WW, Wolfinger, RD, Shabenberger O (2006), SAS System for Mixed Models, Cary, NC: SAS Institute, Inc.

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, pp. 379-89.

Pehrsson, PR, Perry, CP, Daniel MG (2013) *Proceedings from 36th National Nutrient Databank Conference: ARS, USDA Updates Food Sampling Strategies to Keep Pace with Demographic Shifts*. Boston, MA.

Perry, CR, Pehrsson PR, Holden, J (2003), "A Revised Sampling Plan for Obtaining Food Products for Nutrient Analysis," Proceedings of the American Statistical Association, Section on Survey Research Methods, San Francisco, CA: American Statistical Association: pp. 3270-77.

Perry, CR, Beckler, DG, Pehrsson PR, Holden, J (2000), "A National Sampling Plan for Obtaining Food Products for Nutrient Analysis," Proceedings of the American Statistical Association, Section on Survey Research Methods, Alexandria, VA: American Statistical Association: pp. 267-72.

Office of Management and Budget (2013), "OMB Bulletin NO. 13-01," available on web at http://www.whitehouse.gov/sites/default/files/omb/bulletins/2013/b13-01.pdf.

Sarndal, C, Swenson, B, Wretman, J (1992), Model Assisted Survey Sampling, New York, NY: Springer-Verlag.

U.S. Department of Commerce: Economics and Statistics Administration. U.S. Census Bureau (2012). *2010 Census Summary File 1: Technical Documentation*. Washington, DC.

Williams, RL, Chromy, JR (1980), "SAS Sample Select MACROs," Proceedings of the Fifth Annual SAS Users Group International Conference, Cary, NC: SAS Institute, Inc.: pp. 392-396.

Appendix 1

Variance Formulas for Random Group Compositing.

Theorem:
Suppose a simple random sample of size $n = km$ is drawn from a very large population, and the individual samples, $x_1, x_2, \ldots, x_i, \ldots, x_n$, are randomly divided into *k* groups of size m for compositing. Suppose further that nutrient analysis of the *k* composites results in *k* nutrient values, $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_i, \ldots, \bar{x}_k$, having a grand mean $\bar{x}$ with estimate variance, $var(\bar{x})$:

$$\bar{x} = \frac{1}{k}\sum_{i=1}^{k} \bar{x}_i \quad \text{and} \quad \text{var}(\bar{x}) = \frac{1}{k(k-1)}\sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2.$$

Under these random grouping conditions, and the further assumption that each of the *k* composites is a perfect mixture of its components, the serving to serving variance, the estimated variance between the individual samples, is obtained by

multiplying the estimated variance of the grand mean by the sample size $n = km$. That is:

$$\text{var}(x_i) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = km(\text{var}(\bar{x})) = \frac{m}{k-1}\sum_{i=1}^{k}(\bar{x}_i - x)^2.$$

Proof:
By definition

$$\text{var}(\bar{x}) = \text{var}(\frac{1}{k}\sum_{i=1}^{k}\bar{x}_i).$$

Because the variance of a constant times a random variable is the constant squared times the variance of the random variable. Because the *k* groups are randomly created from a random sample from a very large population the individual groups can be considered to being independent simple random samples from the population, it follows that the variance of the sum is equal to the sum of the variances. That is:

$$\text{var}(\frac{1}{k}\sum_{i=1}^{k}\bar{x}_i) = \frac{1}{k^2}\text{var}(\sum_{i=1}^{k}\bar{x}_i) = \frac{1}{k^2}\sum_{i=1}^{k}\text{var}(\bar{x}_i).$$

Expanding the variance for each of the independent random groups yields:

$$\frac{1}{k^2}\sum_{i=1}^{k}\text{var}(\bar{x}_i) = \frac{1}{k^2}\sum_{i=1}^{k}\left(\frac{1}{m(m-1)}\sum_{i=1}^{m}(x_i - \bar{x}_i)^2\right) = \frac{1}{km}\left(\frac{1}{k}\sum_{i=1}^{k}\text{var}(x_i)\right).$$

Since each *k* independent serving-to-serving variance estimates, $var(x_i)$, provides an unbiased estimate of the serving-to-serving variance for the overall population, their mean also provides an unbiased estimate of the serving-to-serving variance for the overall population:

$$\frac{1}{km}\left(\frac{1}{k}\sum_{i=1}^{k}\text{var}(x_i)\right) = \frac{1}{km}\overline{\text{var}(x_i)}.$$

Therefore, it follows from the first and last equality above that:

$$\text{var}(x_i) = km(\text{var}(\bar{x})) = \frac{m}{k-1}\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2.$$