A REVISED SAMPLING PLAN FOR OBTAINING FOOD PRODUCTS FOR NUTRIENT ANALYSIS

Charles R. Perry, Jr., USDA-NASS, Pamela R. Pehrsson and Joanne Holden, USDA-ARS
Charles R. Perry, USDA-NASS, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

KEY WORDS: Controlled Sampling, Chromy's PMRPPS Procedure, Kolmogrov'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 available in the United States. This paper describes the revised National Food and Nutrient Analysis Program (NFNAP) sampling plan, which will be implemented in the Fall of 2003, for the collection of food samples from retail outlets for nutrient analysis.

In 1997, NDL inaugurated the NFNAP, the main goal of which is to obtain reliable estimates with known variability for the nutrient content of food and beverages consumed by the U.S. population (Perry, *et al.*, 2000; Pehrsson, *et al.*, 2000; Haytowitz, *et al.*, 2002). The first objective of NFNAP was to identify one thousand key foods contributing critical nutrients in 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 food 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. Under NFNAP, the sampling plan was to reflect the most current population information available; hence, the revision described in this paper.

The sampling plan used to collect food samples for analysis was based on a stratified three-stage design using 1997 population projections from the U.S. Bureau of the Census and food product market share data from A.C. Nielsen, Inc. Counties were selected at the first stage, grocery store outlets within the selected counties were selected at the second stage, and specific food products to be purchased for nutrient analyses were selected at the third stage. The plan provided self-weighting nationally representative samples of the food products consumed by a typical person in the United States for nutrient analysis.

To the extent possible, the strata boundaries followed standard U.S. Census regions. To obtain nearly equal population size strata, which were desired to facilitate self-weighting, balanced data analysis, it was necessary to shift one or more states from the southern region to the central region. Texas was shifted because the agricultural industry and population diversity of Texas suggest it fits equally well in either region. The change resulted in a central region containing noncontiguous states. Figure 1 gives a graphical display of the regions used with the initial sampling plan. The locations of the counties that contained the outlets where food products were purchased are shown on the map in red.

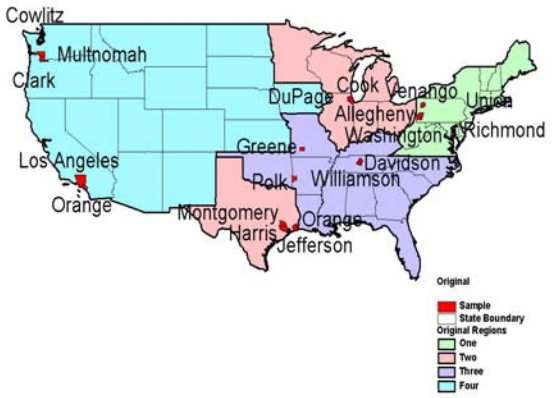

Figure 1: Original NFNAP Regions and County Sample

The revised design, like the initial design, uses a three-stage, self-weighting selection process where counties are selected at the first stage, grocery store outlets are selected at the second stage, and specific food products to be purchased for nutrient analyses are selected at the third stage. The revised design, unlike the initial design, incorporates the 2000 Census Bureau regions, divisions, and states into the first stage sample selection process.

Under the revised design, the selected counties are not only geographically dispersed across the nation and regions according to 2001 Census Bureau projected state population sizes, but are also statistically representative with respect to both the county sizes and the Consolidated Metropolitan Statistical Areas (CMSAs) of the nation and regions. Incorporating the Census regions, divisions, and states into the sample selection process as implicit stratifiers facilitates analyses at different Census geographic levels. A CMSA is an urban area with population of at least one million that satisfies several other requirements (U.S. Bureau of the Census, 2000).

Section 2 describes Chromy's probability minimum replacement (PMR) probability proportional to size (PPS) sample selection procedure (Chromy, 1979; Williams and Chromy, 1980; Chromy, 1981). Section 3 describes the original sampling plan. Section 4 describes the objectives for the revised sampling plan. Section 5 describes the revised county sampling plan. Section 6 describes revised outlet and product sampling plan. Section 7 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
$q(i) = \mathrm{E}[n(i)] = nS(i)/S(+)$

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 2 illustrates the procedure for a hypothetical case where a sample of size five is to be drawn from eight available sampling units.

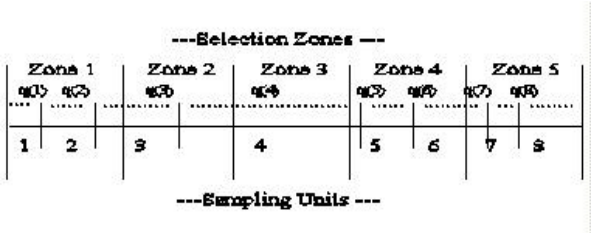

Figure 2: Chromy's PMRPPS Sampling Procedure

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. By ensuring that a single unit is selected from each zone, 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.

3. OVERVIEW OF INITIAL SAMPLING PLAN

The objective of both the original and revised sampling plans is to collect representative samples of specific food products identified by NDL. This is accomplished by purchasing, to the extent possible, the same food products (brands and package sizes) from each of the sampled outlets for nutrient analysis.

As outlined above, the original sampling plan used a three-stage stratified design. At the first stage, generalized CMSAs (gCMSAs) were used. The gCMSA concept is based on the Census Bureau's CMSA. For counties in a CMSA, the county's gCMSA is defined as the CMSA; for counties not in a CMSA the county's gCMSA is the county itself. Within each stratum, the gCMSAs were sorted in decreasing order by population size and then a (PPS) systematic sample of size three was selected. Then the counties within each selected gCMSA were sorted in descending order by their urbanicity and a PPS systematic sample of two counties was selected from each gCMSA. The urbanicity index measures urban character based on the populations of a county's largest cities and towns (Goodall, *et al.*, 1998). Sorting the counties within gCMSAs by urbanicity ensured that the sample contained both more urban and less urban counties. For those gCMSAs made up of only a single county, the county was selected twice.

At the second stage, for each of the selected counties, a list of outlets with sales greater than two million dollars per year was obtained from Trade Dimensions™. Each list was reviewed to ensure that each county contained at least 10 retail outlets. For counties having fewer than 10 outlets, adjacent counties were added to ensure that the area contained a minimum of 10 outlets. Then a PPS sample of outlets with size proportional to the outlet's annual value of sales was drawn from each selected county. Where a single county made up the gCMSA, two outlets were drawn from the county. Likewise, where multiple counties were aggregated to ensure 10 or more outlets, two outlets were drawn from the aggregation. Otherwise, one outlet was drawn from each county. Alternate outlets were drawn for each county in case the primary selected outlets were inaccessible or products were unavailable.

At the third stage, two types of samples were selected. The main sample was drawn to support the estimation of the mean nutrient content of an average serving from composited samples of a product. A composited sample is a homogeneous mixture of several packages of a specific food. Nutrient analyses obtained from composited samples pertain to an average serving from the homogenized product, not to a typical serving. Estimates of critical nutrients were expected to have wide

variability. For some critical nutrients, prior data on variability were limited or nonexistent. Compositing reduced variability by increasing the effective sample size.

The secondary sample was drawn to support the estimation of the variability of the nutrient content of a typical serving of a product. Nutrient analyses obtained from individual packages of food products provided variability estimates between typical servings of the product. Secondary samples were drawn only for critical foods because of the substantial cost of the associated nutrient analyses. The secondary samples were taken from the product samples selected for the composited sample. The secondary samples were also used to develop models for the prediction of the serving-to-serving variance of nutrient content from the variance of composited samples.

For the composited sample, food products were selected proportional to the amount of the product sold nationally using information obtained from AC Nielsen Market Research SCANTRACK™ data, which are collected from checkout price scanners. Products were chosen by selecting a sequential sample using Chromy's Method, which was described in section 2. The selection was done with probability proportional to the market share adjusted for the package size (in ounces). The number of samples chosen for each product was based on the desired statistical reliability and the number of nutrient analyses NDL could afford to perform. Designated food products were purchased from each selected outlet.

The selected products 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 were formed by product sample across locations. This resulted in one data point for each product sample for statistical analysis. Performing the analysis in this manner provided individual product data for major brands and overall results for the particular type of food product. Forming composites by brand across locations allows easy updating of data for a single brand without affecting the estimates for other brands of that product. It also allows changes in weighting to reflect changes in market share.

The secondary sample was to be used for estimation of serving-to-serving variation. Note that outlets 1 and 2 were in the same gCMSA, outlets 3 and 4 were in the same gCMSA, and so on. The sampling plan consisted of the following steps. First, randomly select one product sample from each pair of consecutive product samples. Second, randomly select two gCMSAs at a time without replacement and associate each sample of two gCMSAs with a product sample. Third, randomly select one outlet from each selected gCMSA. The resulting sample design was similar to an incomplete latin square experimental design. Although specific outlets were selected for the secondary sample, data collectors were instructed to pick up extra primary samples in both outlets in selected gCMSAs as a cautionary measure to avoid missed units.

4. OBJECTIVES FOR SELECTION OF COUNTIES

The first stage of the NFNAP sampling plan is being revised to bring the sampling in line with the latest available U.S. population projections based on the 2000 Census. After extensive discussion, in order to ensure the sample is well dispersed nationally and regionally, it was decided that the revised county sample selection procedures should result in a PPS sample of counties that satisfies, to the extent possible, each of the following 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, we would like the cumulative proportion of the sample counties to approximate the cumulative proportion of the population at any point along the ordering.
2. The gCMSAs containing sample counties should be well dispersed when the gCMSAs are sorted by size regionally. That is, we would like the cumulative proportion of sample counties to approximate 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 gCMSAs containing sample counties should be well dispersed when the gCMSAs are sorted by size nationally.
5. The sample counties should be well dispersed when the counties are sorted by size nationally.

Clearly, there does not exist a simple sample selection procedure which ensures that all of the above criteria are simultaneously satisfied. For example, consider the following procedure. First, order the counties by the size of the gCMSA containing the county. Then, choose a sample of counties using any of a number of PPS sequential zonal sampling procedures.

The resulting sample will be highly representative with respect to U.S. gCMSA 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. In essence, we have multiple criteria that we would like to satisfy 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 stochastic 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 a

goodness of fit measure, to find a sample that nearly satisfied the remaining criteria

In many respects, the objective of this approach is 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 the above controlled sampling procedures over after-the-fact re-weighting is that under some circumstances controlled sampling can be used to produce a self-weighting sample, which is often desirable.

5. REVISED COUNTY SAMPLING PLAN

This section describes an 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. First, the counties were sorted by Census region, within region by division, within division by state, within state serpentinely by gCMSA population size, and within gCMSA serpentinely by urbanicity. Then 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 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 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.

- The sort for criterion 2 was by region, population size of gCMSA serpentinely within region, and urbanicity of county serpentinely within gCMSA.
- The sort for criterion 3 was by region and population size of county serpentinely within region.
- The sort for criterion 4 was by population size of gCMSA and urbanicity of county serpentinely within gCMSA.
- The sort for criterion 5 was by population size of county.

For example, to draw the ideal sample by gCSMA within regions, the gCSMAs were sorted by Census region and within regions the gCSMAs were sorted serpentinely by population size; that is, if the gCMSAs of a region were sorted in increasing order, the gCMSAs of adjacent regions were sorted in decreasing order and vice versa. Within gCMSAs, the counties were sorted serpentinely by urbanicity. The counties containing the 24 quantile centers were selected as the ideal sample. Thus, the ideal sample correspond to the centers, with respect to population cumulative population size, of the 24 zones for Chromy's PMRPPS zonal sample. The other ideal samples were drawn in a similar manner.

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 sample to which it is being compared. The two ordered samples were then paired and the absolute value of difference of the sample cumulative gCMSA (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. The sample that had the lowest overall D (best fit) was chosen as the revised NFNAP county sample, which is geographically displayed in Figure 3 below. 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.

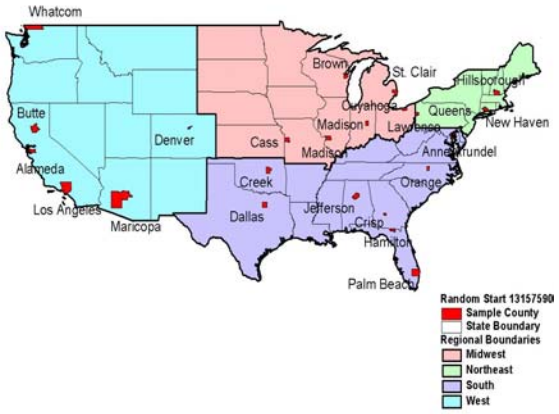

Figure 3: Regions and Revised NFNAP County Sample

The QQ plot in Figure 4-7 compare the revised sample to the ideal samples associated with criterion 2-5. Figure 4 indicates that when the sample and population are sorted serpentinely by region according to gCMSA 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 gCMSAs containing sample counties are well dispersed over the population when the gCMSAs are sorted by size regionally.

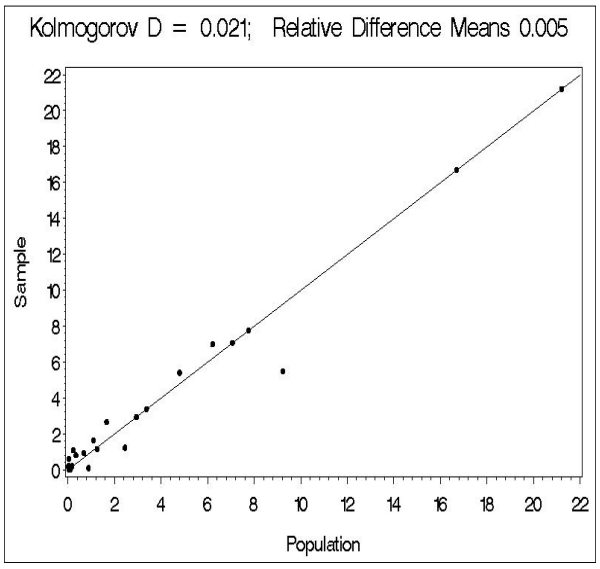

Figure 4: QQ Plot of Sample vs Ideal Sample for gCMSA Size by Regions

Figure 5 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.

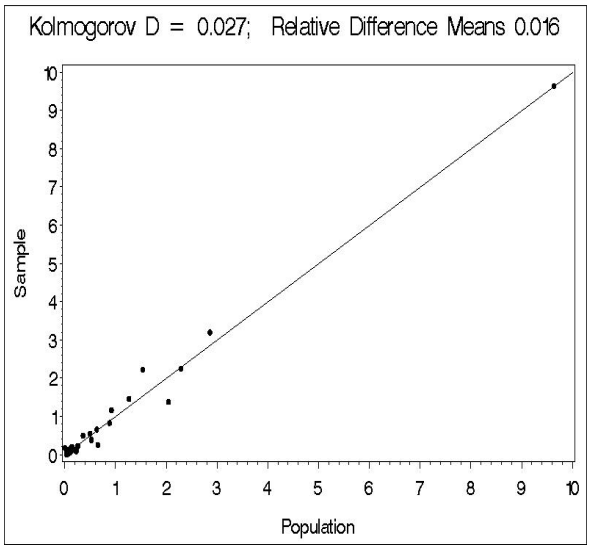

Figure 5: QQ Plot of Sample vs Ideal Sample for County Size by Regions

Figure 6 indicates that when the sample and population are sorted by gCMSA size the quantiles of the sample and the centered quantiles of 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 gCMSAs containing sample counties are well dispersed over the population when the gCMSAs are sorted by size.

Figure 7 indicates that when the sample and population are sorted by county size the quantiles of the sample and the centered quantiles of 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.

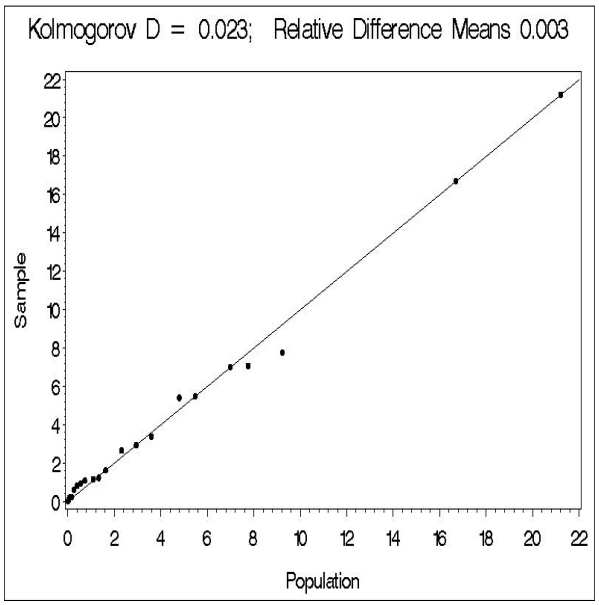

Figure 6: QQ Plot of Sample vs Ideal Sample for gCMSA Size Overall

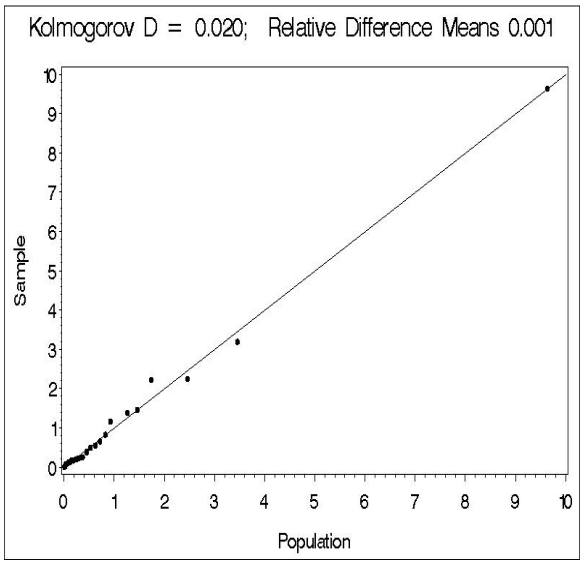

Figure 7: Overall County QQ Plot of Sample vs Ideal

Therefore, the revised NFNAP sample satisfies criterion 1 and simultaneously approximately satisfies criteria 2-5.

6. REVISED OUTLET AND PRODUCT SAMPLING PLAN

Since at the first stage of the revised NFNAP sampling plan, like at the first stage of the initial plan, 24 counties were selected PPS, the last two stages of the initial plan required only minor changes to bring them into line with the first stage of the revised plan.

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 two million dollars per year, was developed for each of the 24 selected counties. For counties having fewer than 10 outlets, adjacent counties were 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 is used when the primary is inaccessible or when a product is unavailable at the primary outlet.

At the third stage of the revised plan, like at the third stage of the initial plan, two types of product samples were 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 was 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 was also 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 was 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. The number of samples chosen for each product was based on the desired statistical reliability and the number of nutrient analyses NDL could afford to perform. The selected products were purchased from each of the primary outlets unless the product was not available or the primary outlet was inaccessible. In that case, the product was purchased from the alternate outlet. If the product was unavailable at either outlet, then a substitute product was purchased.

The selected products 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 were formed by product sample across locations as shown in Table 1. Forming the composites in this manner provided 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.

Table 1: Matrix of Sampled Products

County Pairs	Sampled Counties 1	2	...	12	
Primary Sample	**1**	**2**	**3**	**4**	**...**	**23**	**24**	
1	→							X_1
2	→							X_2
⋮	⋮		⋮				⋮	⋮
12	→							X_{12}

The secondary sample was selected from the primary sample using essentially a replicated incomplete latin square design. Table 2 displays an example of six secondary samples of size two chosen from the primary samples displayed in Table 1 above. At step one, primary samples 1, 3, 5, 8, 10 and 12 were selected. At step two, counties pairs 1 and 10 were selected from primary sample 1, pairs 3 and 9 were selected from the remaining county pairs for primary sample 2, and so forth until two county pairs were selected without replacement for each primary sample. At step 3, county 2 was selected from county pair 1and county 19 was selected from county pair 10 and so forth until one county was selected from each county pair. The resulting secondary sample contains two counties for each selected primary sample chosen such that no two counties were in the same county pair as illustrated in Table 2.

Table 2: Serving-to-Serving Sample Example

	County Pairs 1	2	3	4	5	6	7	8	9	10	11	12
Primary Sample	**Selected County 2**	**3**	**6**	**8**	**9**	**12**	**14**	**16**	**17**	**19**	**22**	**24**
1	×									×		
3			×						×			
5					×						×	
8		×						×				
10							×					×
12				×		×						

7. SUMMARY COMMENTS AND CONCLUSIONS

Several summary and analysis options are available under the revised design. For every food sampled under this design, composite nutrient means were determined for each brand across locations. Two types of variability were measured. For every food, the standard errors of the composite means provided 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.*, 1996).

For all foods, compositing individual samples by brand across locations is cost-effective under the scope and funding of this project. However, at the cost of brand name estimates, it would have been more efficient to randomize the composites across brands to obtain lower variances. Under the revised design, the brand data in a weighted composite of brand values can be redetermined and replaced to reflect changes in product ingredients and market share distributions, updating the entire composite for a generic food profile. The additional serving-to-serving analysis developed under NFNAP is especially useful for critical nutrients whose variability is expected to be wide (based on similar foods and nutrients) or where information on a nutrient's variability is limited or nonexistent.

These serving-to-serving variability estimates are important to the USDA nutrient database for several reasons. They are crucial to accurate assessment of nutrient intakes on a national level. They provide necessary information for individuals with nutrient sensitivities or intolerances. They are important in understanding new foods or modified versions of existing foods, and they can be used to determine sample size for nutrient monitoring of the food supply.

Comparing the serving-to-serving variability estimates will allow development of variance models for predicting the between serving variance from the between composite variance for a given nutrient. In cases where differences in nutrient variability are expected or observed within food types depending on the matrix (that is, cooked or processed version versus raw), this is a cost-effective and time-efficient process for developing patterns of variability and projecting variance in future sampling.

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, gCMSA population size, and to county population size, both overall and within census regions. This approach allows NDL to determine representative estimates of the mean nutrient content of the most important foods in the U.S. food supply as consumed. 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, J.R. (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, J.R. (1981), "Variance Estimators for a Sequential Sample Selection Procedure," in Krewski, D., *Current Topics in Survey Sampling,* Academic Press, pp. 329-347.

Goodall, C.R., Kafadar, K. and Tukey, J.W. (1998), "Computing and Using Rural versus Urban Measures in Statistical Applications," *American Statistician*, Vol. 52, No. 2, pp. 101-111.

Haytowitz D.B., Pehrsson P.R., and Holden J.M. (2002), "The Identification of Key Foods for Food Composition Research," *J Food Comp Anal,* Vol.15, No. 2, pp.183-194.

Littell, R.C., *et al.*, (1996), *SAS System for Mixed Models,* Cary, NC: SAS Institute, Inc.

Pehrsson, P.R., Haytowitz, D.B., Holden, J.M, Perry, C.R., and Beckler, D.G. (2000), "USDA's National Food and Nutrient Analysis Program Food Sampling," *J Food Comp Anal,* Vol. 12, pp. 379-89.

Perry, C.R., Beckler, D.G., Pehrsson P.R., and 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.

U.S. Census Bureau (2000), "Decennial Management Division Glossary," available on web at *www.census.gov*.

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