SAMPLING AND ANALYSIS PLAN FOR A NATIONAL FLUORIDE STUDY

Charles R. Perry, Jr., Michael E. Bellow, USDA-NASS
Pamela R. Pehrsson, USDA-ARS
Michael E. Bellow, USDA-NASS, 3251 Old Lee Hwy., Room 305, Fairfax, VA 22030

KEY WORDS: Probability minimum replacement, mixed effects model, time replicates

1. INTRODUCTION

The U.S. Department of Agriculture's Nutrient Data Laboratory (NDL), a division of the Agricultural Research Service (ARS), develops reliable databases and state of the art methodology to evaluate and disseminate composition data on foods available in the United States. In 1997, NDL in cooperation with the National Heart Lung and Blood Institute (part of the National Institutes of Health) inaugurated the National Food and Nutrient Analysis Program (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 (Pehrsson et al., 2000). Toward this objective, highly representative probability-based food and beverage samples are selected and the resulting nutrient datasets analyzed. NFNAP has already achieved major improvements to NDL's National Nutrient Databank (NNDB) through a comprehensive revision of scientific concept and technical approach. USDA's National Agricultural Statistics Service (NASS) provides technical support to NDL in the development of unique sampling plans for specific retail foods and nutrients.

In 2000, NDL began planning a nationwide study aimed at evaluating the mean concentration and variability of fluoride in the U.S. food and water supply. Of particular interest were drinking water from municipal supplies throughout the country and those beverages and foods that are the chief contributors to dietary fluoride in the U.S. The results of the study will be critical to the national fluoride database to be developed by NDL. In preparation, NDL carried out two preliminary studies of municipal water supplies and carbonated beverages to examine the concentration of fluoride as well as other mineral elements. Results of the preliminary studies were used to determine the sample sizes required for the larger study.

Sampling requirements for the main study were influenced by a number of factors including: 1) variability of fluoride in foods and beverages; 2) sources of fluoride variation (e.g., geography, season, production plants); 3) product brands having a significant market share; 4) product distribution patterns; 5) desired level of confidence in the final estimates; and 6) costs associated with data collection.

This paper describes the sampling and analysis plan for the national fluoride study. Section 2 covers the sampling frame development. Section 3 describes the sample selection procedures. Data collection issues are covered in Section 4, along with a brief description of the chemical analysis. Section 5 discusses plans for statistical analysis of the survey data.

2. SAMPLING FRAME

The basic sampling framework for NFNAP divides the United States into four regions (first stage strata), with communities sampled within each region. From each community, a preset number of locations are sampled further, with food categories selected from these locations. Within food categories, food types are sampled. Subsamples consisting of specific food brands are drawn from the food types. The factors that affect the distribution of fluoride concentration are geography and sources of fluoride. The geographical factors are regions, communities nested within regions, and locations nested within communities.

The sampling frame for the fluoride survey, which contains one record for each county in the U.S., was developed as follows. First, estimated population data for all states were obtained from the U.S. Bureau of the Census web site (www.census.gov). The standard Census regions, i.e., Northeast, Midwest, South and West, were used. The next step involved selecting generalized Census Consolidated Metropolitan Statistical Areas (gCMSAs) within each region. The gCMSA concept is based on the Consolidated Metropolitan Statistical Area (CMSA), i.e., an urban area with population at least one million and satisfying several other requirements (U.S. Bureau of the Census, 1999). Since most U.S. counties are not part of a CMSA, we need a more general term. All CMSAs are defined to be gCMSAs as well. A given county is defined to be a gCMSA if and only if it is not part of a CMSA.

Each record in the frame contains the following data: county name and FIPS (Federal Information Processing Standards) code, state name and FIPS code, population (2000 Census), gCMSA name and code, local (within gCMSA) urbanicity index, and Census region. The

urbanicity index, a measure of urban character, is based on the populations of the largest cities and towns in a county (Goodall et al., 2000). The use of this index ensures that counties bordering a major city are treated more like that city than the area on the outskirts of the gCMSA.

Four separate methods for sorting the counties in the sampling frame were considered: 1) NFNAP regions; 2) Census regions; 3) Census divisions; and 4) Census states. The regions used in previous NFNAP surveys were roughly equal in population but non-standard. There were also anomalies, such as the grouping of Texas with states in the Great Lakes region. The NFNAP regions method specifies that counties be sorted first by gCMSA size within NFNAP regions, then serpentinely by urbanicity within gCMSAs (decreasing in urbanicity in one state if increasing in the previous one, and vice versa). The serpentine manner of sorting ensures that the transition of gCMSAs across state boundaries will be smooth relative to their sizes. This method leads to a good representation of county sizes but not necessarily of Census regions, divisions and states.

In the Census regions method, the counties are sorted first by Census region, then serpentinely by size within regions, and finally serpentinely by urbanicity within gCMSAs. The result is generally a good representation of county sizes and Census regions, but not necessarily of states or Census divisions. The Census divisions method sorts the counties first by Census region, then by Census division within regions, then serpentinely by gCMSA size within divisions. The last step is to sort the counties serpentinely by urbanicity within gCMSAs. This method leads to a good representation of Census regions and divisions, but not necessarily of states or county sizes. Finally, the Census states method specifies that the counties be sorted first by Census region, then by Census division within regions, then by states within divisions. Within states, counties are sorted serpentinely by population size and then serpentinely by urbanicity within gCMSAs. This method leaves gCMSAs in odd numbered states decreasing in size and those in even numbered states increasing in size. In addition, the method leaves the urbanicity index decreasing in odd numbered gCMSAs and increasing in even numbered gCMSAs.

An advantage of the Census states method is that the number of sample counties falling within a state, Census division, or Census region is within one of the correct proportioning of the 72 samples to the states, regions, and divisions with respect to their population size. The result is a good proportioning of the sample of counties to the states, Census divisions and Census regions, and a good geographical distribution of the sampled counties to the same. The limitation is that, since the final sort is not by gCMSA size within Census regions or divisions, this proportioning may not exactly match their proportional representation with respect to size.

After a careful review of the four sorting methods, the Census states sampling method was judged to be the best for yielding the information sought on the national distribution of fluoride in the U.S. food supply. Therefore, this method was used to sort the frame.

Carbonated beverages, juices, bottled water, etc. were picked up from retail outlets and selected with probability minimum replacement (PMR) from comprehensive national listings of retail outlets purchased from Trade Dimensions®, a company that provides this service. Brand name market share data were purchased from A.C. Nielsen®, Inc. to determine market distribution. Selection of brand names was PMR (where market share was computed as grams/pounds consumed) for each food or beverage item. Therefore, brand names which comprise most of the market for a particular food or beverage were usually selected for pickup and analysis. For most retail samples, a random selection of one from each of 36 pairs formed by the serpentine ordering of the 72 original selected retail outlets was used. These outlets will also be used in future sampling. Since many of the beverages had a wide distribution of brand names with no clear market leaders, this method will allow for inclusion of more brands in the sampling.

For sampling of tap water, lists of residents in the 72 sampled counties were requisitioned. To ensure adequate alternate households for non-participants and potential nonrespondents, 100 households per county are to be randomly selected. Within each county, two participants will be secured; since fluoride variability of the tap water is expected (based on the 1999 pilot study), it was taken into account in this larger sampling. Unlike the retail outlets, all households are weighted equally, i.e., there is no reason to assume that one should be weighted more than another. Participants will complete a one-page survey that addresses details of the source of their tap water.

3. SAMPLE SELECTION PROCEDURES

The sampling procedures in the preliminary fluoride studies supported a self-weighting nationally representative stratified sample of municipal water consumed in the United States. This approach is consistent with the one used during the previous three years under NFNAP (Perry et al., 2000). Water samples were to be collected from 24 locations. The amount of beverage consumed was assumed to be proportional to the population size and constant across the country. The purpose was to obtain initial estimates of location-to-location variability contributing to a quantifiable national variability. In addition to sample selection from various locations, samples were picked up once every four months for three pickups to measure possible seasonal effects.

The plan was a three-stage design, where counties within gCMSAs were selected at the first stage, grocery store outlets within counties at the second stage, and specific food products purchased and analyzed for nutrient content at the third stage. In effect, this procedure leads to a sample of grocery outlets from geographically dispersed areas across the U.S. The intent was that the nutrient data obtained from the study would be (approximately) self-weighting and therefore could be treated as if they came from a simple random sample. The volume of food or beverage consumed was assumed to be proportional to the population and constant across the country.

Sampling requirements for water, beverages and foods include specifying the number of gCMSAs, number of locations within gCMSAs, whether different brands or forms of the product need to be represented, number of pickups over time (to capture seasonal variation), and total number of samples required at each pickup. In addition, there was a need to carry out the study as efficiently and economically as possible, utilizing existing archived samples where available.

In NFNAP, foods and beverages are composited across sampling locations, with the goal of assessing the best estimate of a nationally representative mean nutrient value in a cost effective way. Since an objective of the study is to assess the variability of fluoride, the high priority beverages were analyzed individually by location rather than compositing them into single samples. Because there is less concern about the variability of fluoride in medium and low priority foods, they will be sampled nationally using the NFNAP method but composited for analysis to estimate mean fluoride.

All county samples were drawn using Chromy's zonal sampling method (Chromy, 1971). This procedure is probability minimum replacement (PMR), i.e., the number of times that a given county can be selected depends on its size and is limited to two values that are consecutive integers (e.g., 0 or 1). The county sizes were obtained from Census county population files for the year 2000 available on the Census web site. Each sample contained 72 counties for use in picking up two water samples per county. One sample was drawn from each of 72 zones in a sorted data set with probability proportional to size. These counties were used for national selection of retail outlets in the retail beverage sampling and will be used for the national selection of residents in the tap water sampling.

The number of gCMSAs, number of locations within gCMSAs, and number of pickups over time depend on the following factors: 1) number of plants producing ready-to-drink beverage; 2) number of brands that hold a significant market share and distribution patterns across the U.S.; 3) level of national variability in earlier NDL or other published studies for that beverage; and 4) desired level of confidence and percent error around the mean.

In order to assess the contribution of covariances to the variance estimate, the fluoride levels in tap water can be represented as a mixed effects model. The region is modeled as a fixed effect. There are three random effects: gCMSAs nested within regions, samples (counties) within gCMSAs, and time replicates within samples. The mixed effects model for fluoride levels can be written as:

$$y_{ijkp} = \mu + r_i + g(r)_{ij} + c(g)_{ijk} + t(c)_{ijkp} + \varepsilon_{ijkp}$$

where:

y_{ijkp} = fluoride level measured in region i, gCMSA j, county k at time p

μ = overall mean fluoride level

r_i = fixed effect of region i (i=1,...,4)

$g(r)_{ij}$ = random effect of gCMSA within region

$c(g)_{ijk}$ = random effect of county within gCMSA

$t(c)_{ijkp}$ = random effect of time replicate within county

ε_{ijkp} = random error (normal distribution assumed)

Letting *n* be the number of gCMSAs, *m* the number of counties per gCMSA, and *q* the number of time replicates per county, the variance of the sample mean of all measurements can be written as:

$$\mathrm{var}(\bar{y}) = \sigma_g^2 / n + \sigma_c^2 / nm + \sigma_t^2 / nmq + \sigma_\varepsilon^2 / nmq$$

where:

σ_g^2 = variance of gCMSA effect

σ_c^2 = variance of county effect

σ_t^2 = variance of time replicate effect

σ_ε^2 = variance of random error

Table 1 shows estimates of the variance for the three random effects as well as the residuals, computed from the

Also given in the table are ratios of the variance estimates to the residual variance, approximate standard errors of the variance estimates, and p-values for an approximate Wald Z-test that the variance is significantly different from zero. The p-values show that each random effect variance is significantly different from zero at the ten percent level.

Table 1: Statistics for Random Effects

Effect	Var. Estim.	Ratio	Std. Error	p-Value
gCMSA	0.166	127.5	0.104	0.055
County	0.057	43.4	0.032	0.039
Pickup (time)	0.064	48.9	0.013	<0.0001
Residual	0.001	1.0	0.002	<0.0001

Based on the cost function given by Cochran (1977) for two stage sampling when travel costs between units are insignificant, the total sampling cost can be expressed as:

$$C = nc_1 + nmc_2 + q(nc_1 + nmc_2)$$

where:

n = number of gCMSAs
m = number of sampled counties per gCMSA
c_1 = cost of visiting a gCMSA
c_2 = added cost of visiting a sampled county
q = number of time replicates per sampled county

The fluoride sampling plan uses 288 samples divided among four regions: 18 gCMSAs per region, two sampled counties per gCMSA, and two time replicates per sampled county. Therefore n=72, m=2, and q=2 so the cost is:

$$C = 72c_1 + 144c_2 + 2[72c_1 + 144c_2]$$

$$= 216c_1 + 432c_2$$

The approximate cost of visiting a gCMSA was found to be \$20, and the additional cost of visiting a county and taking a sample is \$32. Thus the total cost is:

$$C = 216(20) + 432(32) = 18{,}144$$

The objective is to find sample allocations, i.e., sets of values (n, m, q), that minimize the variance of the sample mean subject to maximum cost \$18,144 and certain other restrictions. This goal is accomplished by evaluating the cost function for all allowable integer values of n, m and q, discarding the sets of values leading to total cost greater than C, computing the variance associated with each remaining set of values, and selecting the one with the smallest variance.

An expression for the variance of the sample mean was provided earlier in this Section. Since the variance components are not known, their estimates (denoted by the ^ symbol) from Table 1 can be used. The random error variance is negligible and can be ignored, so the estimated variance of the sample mean is given by:

$$\hat{\sigma}^2(\bar{y}) = \hat{\sigma}_g^2 / n + \hat{\sigma}_c^2 / nm + \hat{\sigma}_t^2 / nmq$$

$$= 0.166 / n + 0.057 / nm + 0.064 / nmq$$

The cases (sets of restrictions) that were considered are as follows:

1. Equal number of gCMSAs in each region (n a multiple of four), two or more samples per gCMSA, two time replicates.

2. Equal number of gCMSAs in each region, any number of samples per gCMSA, two time replicates.

3. Equal number of gCMSAs in each region, two or more samples per gCMSA, one time replicate.

4. Any number of samples per gCMSA, one time replicate, and a) equal number of gCMSAs in each region, or b) any number of gCMSAs

Only case 1 above allows for estimation of all components of variance. Case 2 allows estimation of variance across time replicates but not counties. Case 3 allows estimation of variance across counties but not time replicates. Case 4b does not allow computation of any variance components, but is most cost efficient for estimating the mean. Table 2 shows the results of optimizations performed for each case. The sets of values (n*, m*, q*) that led to minimum variance are shown, as well as the variance itself and the cost.

Table 2: Optimal Allocations

Case	n*	m*	q*	Var.	Cost($)
1	72	2	2	.029	18,144
2	116	1	2	.022	18,096
3	108	2	1	.0021	18,144
4a	172	1	1	.0017	17,888
4b	174	1	1	.0016	18,096

If one is interested in estimating the components of variance, the best way to allocate the sample is to use the original sampling plan (case 1), i.e., 72 gCMSAs per region, two samples per gCMSA, and two time replicates per sample. If one is only interested in estimating the

mean, then the best strategy is to use only one observation per gCMSA. Table 3 shows the minimum sample size (n*) required to have 90 percent confidence that the estimated mean is within ten percent of the true mean, for cases 1, 2, 3 and 4b above. Here, the cost is no longer constrained to be less than or equal to $18,144.

Table 3: Sample Size (n*) Required to Achieve 10% Error Bound on Mean with 90% Confidence

Case	n*	m*	q*	Cost ($)
1	252	2	2	63,504
2	306	1	2	47,736
3	271	2	1	45,528
4b	344	1	1	35,776

4. DATA COLLECTION AND CHEMICAL ANALYSIS

Sampling of municipal (tap) water in residential homes presents different challenges than the retail sampling. The water sampling includes a total of 288 national samples. Preparations for assurance of water sample integrity included development of: 1) shipping protocols for water collection bottles to minimize sample loss and contamination; 2) protocols to assure sample integrity and adequacy; and 3) water collection kits and pickup strategies with collection/survey kit delivery agents from Superior Pickup, Inc. Steps in the development of a participant . survey to secure information on the household water supply and plumbing included: 1) devising participant recruitment procedures; 2) contract placement with a company for a residential phone listing by counties (randomly ordered); 3) survey development, and 4) processing of the survey distribution application through USDA's survey approval office and the U.S. Office of Management and Budget (OMB).

Adult individuals, who are considered in a position of responsibility for the household and have agreed to participate, will complete a questionnaire which focuses on their source of drinking/cooking water and any treatment of the water (e.g., water softening or purification systems) They will fill two 250 ml bottles with tap water from the kitchen faucet. The bottles and survey will be shipped in a prepaid, prelabeled package (provided to the consumers) to Virginia Polytechnic Institute and State University for sample preparation. To ensure the confidentiality of individual participants, fluoride data and household information on water source and treatment will be attached to a consumer code and reported only in table format. The data will not be used to assess the quality of a family's socioeconomic status or any other characteristic of that individual's home. Each participant will be awarded an incentive at the time of water collection and survey completion in order to maximize compliance with the study.

In addition to municipal (tap) water sampling, a frame was developed for sampling of beer, wine, and retail beverages such as fruit juices. The design consists of the same counties. For counties where beverages containing alcohol are not sold in retail outlets, a listing of state distribution centers was used. The state distribution centers nearest to a selected retail outlet were targeted for pickup of beer and wine in those states or counties. Since information on sales and distribution of alcohol-containing beverages was not available through Nielsen® or other market tracking companies, NDL gathered that data from the trade associations. A subset of 36 counties was randomly selected for the sampling of beer and most other retail non-alcoholic beverages. From this subset of 36, a further subset of 18 counties was randomly selected and used for the sampling of wine.

Once samples are collected, they are composited and homogenized according to specific work plans through an accredited university laboratory. Quality Control (QC) materials are introduced into the sample stream to validate the analytical method on a routine basis. The selection of the appropriate analytical lab, valid analytical methods, and appropriate QC materials (i.e., similar to the matrix being analyzed) were crucial steps in the process. Fluoride analysis is conducted either by the direct read or micro-diffusion method. Those samples deemed to be high relative contributors of fluoride to the diet are analyzed individually, while the other samples are composited at various levels. Once the data are generated, they pass through a rigorous QC evaluation process for validation of accuracy and precision.

5. STATISTICAL DATA ANALYSIS

Following chemical analysis, the individual results by location will be averaged to determine a national mean and standard deviation. Additionally, the data will be categorized as either fluoridated or non-fluoridated and means and variability estimated. For beverages and foods with a relatively small contribution to fluoride intake compared with water, samples will be composited to national samples prior to analysis. In such cases, variability estimates are not possible. In the final data analysis, components of variance will be evaluated using a mixed model similar to the design used in developing the sampling approach: within gCMSAs, county-to-county, within county, and pickup to pickup (over time). Means and standard errors will be determined, as well as upper

and lower error bounds based on 90% confidence limits. Once evaluated, these data will be disseminated to the University of Minnesota's Nutrition Coordinating Center (a collaborator on this project) and to the USDA's National Nutrient Databank Standard Reference (www.nal.usda.gov/fnic/foodcomp), in the 2003-04 time frame.

6. CONCLUSIONS

The fluoride database resulting from the national study will provide values on the fluoride content of tap water and other fluoride-contributing beverages and foods, support important research on the analytical methodology for fluoride, and be of considerable value to USDA and other investigators in the US dental health research community. Although the sampling approach developed for this research does not give the best distribution of the sampled counties with respect to the size of the counties, it does give a good proportioning of the sample of counties to the states, Census Divisions and Census Regions and a good geographical distribution of the sampled counties to the same. These properties were deemed most important to the appropriateness of the data for database use as well as dental and health research.

REFERENCES

Chromy, J.R. (1971), "Sequential Sample Selection Methods", *1971 Proceedings of the American Statistical Association, Section on Survey Research Methods*, American Statistical Association, pp. 401-406.

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

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", *Journal of Food Composition and Analysis*, Vol. 15, No. 2, pp. 183-194.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996), *SAS System for Mixed Models*. Cary, N.C.: SAS Institute, Inc.

Perry, C.R., Beckler, D.G., Pehrsson, P., and 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, pp. 267-72.

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", *Journal of Food Composition and Analysis*, Vol. 12, pp. 379-89.

U.S. Bureau of the Census (1999), "Decennial Management Division Glossary", available on web at *www.census.gov/dmd/www/glossary.htm*.