

Paper 206-29

Using SAS[®] Procedures to Make Sense of a Complex Food Store Survey

Jeff Gossett, University of Arkansas for Medical Sciences, Little Rock, AR
Pippa Simpson, University of Arkansas for Medical Sciences, Little Rock, AR
Renée Hall, University of Arkansas for Medical Sciences, Little Rock, AR
Carol Connell, University of Southern Mississippi, Hattiesburg, MS
Kathy Yadrick, University of Southern Mississippi, Hattiesburg, MS
Margaret Bogle, USDA ARS, Delta NIRI Project, Little Rock, AR

ABSTRACT

A case study of a community food store survey for the Lower Mississippi Delta Nutrition Intervention Research Initiative (Delta NIRI*) is given. Survey design and analysis issues are discussed with references to the SAS survey software. In this paper we address the estimation of the cost of a market basket of food items. Basket items, and therefore prices, are not available at all stores. We discuss the imputation of missing food item prices using PROC MI[®]. Since we are sampling from a finite population, most commonly used SAS procedures are not appropriate for the estimation of correct standard errors. Therefore, we combine PROC MI with PROC SURVEYMEANS[®]. For each of 5 sets of imputed data, we estimate the mean cost of the market basket using PROC SURVEYMEANS. We then use PROC MIANALYZE[®] to combine estimates and to estimate the standard error of the cost estimates.

INTRODUCTION

The Lower Mississippi Delta Nutrition Intervention Research Initiative (Delta NIRI) Food Store Survey (FSS) is a cross-sectional survey used to collect data on food store characteristics, food availability, styles and package types, food quality, and food prices in the Lower Mississippi Delta counties of Arkansas, Louisiana, and Mississippi. The survey was designed to provide estimates of market basket cost at the county level. We briefly discuss the development of the sampling frame. The focus of this paper is analysis of food basket pricing using SAS survey procedures and multiple imputation techniques. Food basket items, and therefore prices, are not available at all stores. We discuss the imputation of missing prices using PROC MI. Since we are sampling from a finite population, most commonly used SAS procedures are not appropriate for the estimation of correct standard errors. Therefore, we combine PROC MI with PROC SURVEYMEANS. For each of five sets of imputed data, we estimate the mean market basket cost using PROC SURVEYMEANS. We discuss how PROC SURVEYMEANS output must be transformed for input into PROC MIANALYZE, which is used to combine cost estimates and to estimate their standard errors.

METHODS**SAMPLING FRAME DEVELOPMENT**

A list of stores for the 18 counties was compiled. Using onsite screening of these stores for store type and location accuracy, we developed a sampling frame of 558 stores. There are three types of stores: convenience store, small/medium grocery, and supermarket. One survey objective was to obtain county level estimates of the market basket price for each store type. Within each county, the goal was to sample five stores of each type, where available. In many counties there were fewer than 5 supermarkets to sample, in which case we sampled all of them.

CALCULATING SAMPLE WEIGHTS

Base weights are calculated as the inverse of the probability of selection. The Food Store Survey weights are based on the number of stores surveyed relative to the number in the population within county and store type. Non-response adjustments to sampling weights to account for differential rates of refusals are sometimes appropriate. If there are systematic differences between stores that participate and stores that do not participate, one can sometimes make adjustments to the weights to reflect probabilities of response. Unfortunately, our information on stores' non-participation is very limited. We are assuming that the stores that participated are interchangeable with the ones that refused.

The numbers of stores surveyed (recruited) are summarized in Table 1. A total of 226 stores were surveyed. Since some refusals at the time of survey were expected, up to 7 stores were recruited in each county and store type. The strata variables for this design are the county and store type. All of the stores within a county and store type are assumed to have equal probability of selection. Within each stratum, the goal was to sample up to 5 stores. The sampling weight, calculated as the inverse of the probability of selection, for example, for when there are 10 stores available and 5 surveyed, is $10/5 = 2$. Thus, the sum of the weights equals the number of stores in the sampling frame (i.e., 558).

Table 1: NUMBER OF STORES SURVEYED (RECRUITED) BY STATE, COUNTY AND STORE TYPE

STATE	COUNTY	CONVENIENCE	SMALL/MEDIUM	SUPERMARKET	TOTAL
AR	Chicot	5(7)	5(6)	3(3)	13
	Crittenden	5(9)	5(6)	5(6)	15
	Lincoln	5(5)	2(2)	2(2)	9
	Phillips	5(7)	5(6)	4(5)	14
	St. Francis	5(7)	5(7)	5(5)	16
	Woodruff	4(4)	5(6)	2(2)	11
AR TOTAL		29	28	21	78
LA	Avoyelles	5(7)	5(7)	4(4)	14
	Catahoula	5(5)	2(2)	2(2)	9
	Concordia	5(8)	2(2)	3(3)	10
	Franklin	5(7)	2(2)	4(4)	11
	Pointe Coupee	5(8)	5(7)	4(4)	15
LA TOTAL		25	17	17	59
MS	Coahoma	5(8)	5(7)	3(3)	13
	Humphreys	5(5)	5(5)	1(1)	11
	Leflore	5(8)	5(6)	5(6)	15
	Panola	5(8)	5(8)	5(5)	15
	Sunflower	5(9)	5(10)	5(6)	15
	Tunica	2(2)	4(4)	1(1)	7
	Washington	5(5)	5(7)	5(8)	15
MS TOTAL		32	34	25	91
DELTA NIRI TOTAL		86	76	64	226

MARKET BASKET

The market basket is a set of 11 food items denoted in Table 2. The surveyors were asked to select the item with the lowest price per unit. If the quantity selected differed from the hypothetical basket, the price per unit was multiplied by the quantity specified in the basket, thus yielding the price per quantity specified.

TABLE 2: MARKET BASKET DEFINITION

FOOD ITEM	QUANTITY
Fruit Drink	128 oz
Milk (cheapest of varieties)	1 gallon
Sugar granulated	4 lb
White or Wheat Bread	24oz
Corn Bread / Muffin mix	16 oz
Microwave popcorn	10.5 oz package
Cooking oil	32 oz
Pork and Beans (canned)	15 oz
Tuna (canned)	6 oz
Green Beans (canned)	14 oz
Rice	16 oz

MULTIPLE IMPUTATION

PROC MI assumes that data are missing at random. Before we can use PROC MI to impute missing values we have to consider whether it is a reasonable assumption for missing price data to be missing at random. Unfortunately, items in the market basket are not available in all stores, possibly because the item was out of stock or the store did not stock the item. It's possible that a surveyor could not locate an item that was available.

Missing values can be classified as either ignorable or non-ignorable. Ignorable missing values are missing at random for reasons unrelated to the (unobserved) true values. Whether or not a unit responds (i.e. price is available) is independent of the values of analysis and demographic variables (e.g.. store type, county, state). Non-ignorable non-response missing values

are not missing at random. The reason the value is missing is related to the (unobserved) true value of the variables of interest and related demographic variables.

Our situation is somewhat typical. We probably have missing values from a combination of ignorable and non-ignorable factors. A store owner will try to select food items that maximize her profits, constrained by availability of items from suppliers. Within each store type, each of the food items is available. We cannot assess for example, why an item is available at one convenience store but not at another.

As seen in table 3, 62 out of the 63 supermarkets surveyed had all of the items in the basket. Approximately half (40/77) of the small/medium grocery stores carried all of the items. Only 18 of the 86 convenience stores surveyed had all of the items. It is no surprise that larger stores tended to carry more of the market items.

TABLE 3: CROSS TABULATION OF MARKET ITEMS MISSING BY STORE TYPE. THE MARKET BASKET CONTAINS 11 ITEMS.

Store Type	Number of missing basket items								Total
	0	1	2	3	4	5	6	7	
Supermarket	62	1	0	0	0	0	0	0	63
Small/Medium	40	18	7	5	3	3	1	0	77
Convenience	18	17	12	13	15	6	3	2	86
Total	120	36	19	18	18	9	4	2	226

We consider several issues to resolve before imputing missing values.

- Should weights be considered in the imputation? Our weights vary from 1 to 7.4. In our survey, the weights are equal within strata defined by county and store type. In hot deck imputation, for example, the sample weights of the potential donor individuals can be incorporated into this choice, by preferentially selecting donors with larger sample weights (Cox, 1980).
- For arbitrary missing data, PROC MI has only one option available Markov Chain Monte Carlo (MCMC). Alternatively, one could use MCMC to produce a monotone missing data set followed by another method. For monotone missing data, PROC MI offers regression and propensity score methods. The data in this study are arbitrary missing.
- What variables should be included in the imputation model?
 - The MCMC method assumes that the model is multivariate normal.
 - How do we handle categorical variables (i.e., by statement or indicator variables)?
 - The imputation model should be rich and include variables that may be used in later analyses (e.g., store type and state).
- The missing data problem is not uniform across store types.
- Should we exclude some stores that have too many missing items?
- Should we use PROC MI or data step programming? Korn and Graubard (1999) discuss mean imputation and hot-deck imputation as popular methods for use in surveys. These could be programmed using SAS data steps.

We include the individual prices of the 11 food items and indicator variables for state. We use the “by” statement to request separate imputations for each store type.

```
PROC MI DATA=FSS1 OUT=FSS2 SEED=37851;
  BY STORETYPE;
  VAR I_AR I_LA BREAD CORNBREAD FRUITDRINK GREENBEAN MILK OIL POPCORN PORKBEANS RICE
  SUGAR TUNA;
RUN;
```

By default, PROC MI uses the MCMC method, which is appropriate for arbitrary missing data, with a single chain to create five imputations. We use the BY statement to ask for separate computations for each store type. The first two variables I_AR (state="AR") and I_LA (state="LA") are indicator variables for store location (Arkansas, Louisiana, or Mississippi). The price data for the eleven items in the market basket are the other variables included on the VAR statement. PROC MI creates an output data set “FSS2” with a variable named `_IMPUTATION_` with values 1 to 5. The total price, `TOTALPRICE`, of the market basket is calculated using the sum function in a data step.

```
DATA FSS3;
  SET FSS2;
  TOTALPRICE=SUM(BREAD CORNBREAD, FRUITDRINK, GREENBEAN, MILK, OIL, POPCORN,
  PORKBEANS, RICE, SUGAR, TUNA);
RUN;
```

It's not a bad idea to create indicator variables to have a record of the missing values prior to imputation in a SAS data step (e.g. `BREAD_MISSING=(BREAD=.)`). Then you can compare the distributions of the imputed values to known values as a check using, for example, PROC MEANS[®] (e.g. `PROC MEANS; VAR BREAD; CLASS BREAD_MISSING;`).

ANALYSIS OF MEAN BASKET COST

The market basket price data can be analyzed using either PROC SURVEYMEANS or PROC SURVEYREG[®]. In this paper, we use PROC SURVEYMEANS. Using the survey procedures requires some creativity in PROC MIANALYZE.

Our sample design is stratified with different population totals in the strata (county and store type). STRATATOT is a dataset containing the strata identifiers (COUNTYN and STORETYPE), population totals (_TOTAL_), and imputation indicator (_IMPUTATION_).

```
DATA STRATATOT;
  SET FSS3(RENAME=(SAMPLED=_TOTAL_));
  KEEP COUNTYN STORETYPE _TOTAL_ _IMPUTATION_;
  RUN;
PROC SORT DATA=STRATATOT;
  BY _IMPUTATION_ COUNTYN STORETYPE ;
  RUN;
PROC SORT DATA=FSS3;
  BY _IMPUTATION_ COUNTYN STORETYPE ;
  RUN;
```

We use PROC SURVEYMEANS to calculate the mean cost of the market basket for each store type (supermarket, small/medium, and convenience).

```
PROC SURVEYMEANS DATA=FSS3 N=STRATATOT MEAN DF STDERR VAR NOBS SUMWGT;
  BY _IMPUTATION_;
  DOMAIN STORETYPE;
  VAR TOTALPRICE;
  STRATA COUNTYN STORETYPE / LIST;
  WEIGHT STORE_WT;
  ODS OUTPUT DOMAIN=MEANS(DROP=TOTALPRICE);
  RUN;
```

TABLE 4: ESTIMATES OF TOTAL PRICE FROM PROC SURVEYMEANS

IMPUTATION	STORETYPE	DF	N	SUMWGT	MEAN	STDERR (MEAN)	VAR (MEAN)
1	Convenience	68	86	274	23.2254	0.4541	0.2062
2	Convenience	68	86	274	22.7861	0.4660	0.2172
3	Convenience	68	86	274	22.9609	0.4530	0.2052
4	Convenience	68	86	274	22.8354	0.4715	0.2223
5	Convenience	68	86	274	22.9289	0.4479	0.2006
1	Small/Medium	59	77	205	18.7983	0.3115	0.0970
2	Small/Medium	59	77	205	18.9696	0.3488	0.1217
3	Small/Medium	59	77	205	18.6827	0.2901	0.0841
4	Small/Medium	59	77	205	18.7536	0.2950	0.0870
5	Small/Medium	59	77	205	18.5814	0.2873	0.0825
1	Supermarket	45	63	79	10.7958	0.0826	0.0068
2	Supermarket	45	63	79	10.7976	0.0827	0.0068
3	Supermarket	45	63	79	10.7972	0.0827	0.0068
4	Supermarket	45	63	79	10.7913	0.0823	0.0068
5	Supermarket	45	63	79	10.7948	0.0825	0.0068

To use PROC MIANALYZE with PROC SURVEYMEANS, you have to create a TYPE=COV data set in a Data Step, and you will be able to do this only for a single variable at a time. For each imputation set, a MEAN, standard error of the MEAN (STDERR), variance of the MEAN (VAR), number of observations (NOBS), and sum of the weights (SUMWGT) are output.

The following code may be used to construct a COVARIANCE type dataset. PROC MIANALYZE assumes that the standard error of the mean can be calculated as the standard deviation divided by the square root of the number of observations. Unfortunately, this is incorrect for data from a stratified weighted sample for which the correct standard error of the mean is calculated as the standard deviation of the mean divided by the square root of the sum of the weights. When we build the covariance data set, we substitute the sum of the weights as our "N" variable. To construct the variance, we multiply the variance of the mean by the sum of the weights. PROC MIANALYZE will then calculate the standard error of the mean by dividing the "VARIANCE" variable by the "N" variable.

```
DATA MEANS2 (TYPE=COV) ;
  SET MEANS;
  LENGTH _TYPE_ $ 8;
  _TYPE_ = 'COV' ; TOTALPRICE=TOTALPRICE_VAR*TOTALPRICE_SUMWGT ;
  _NAME_ = 'TOTALPRICE' ; OUTPUT ;
  _TYPE_ = 'MEAN' ; TOTALPRICE=TOTALPRICE_MEAN ; OUTPUT ;
  _TYPE_ = 'STD' ; TOTALPRICE=( TOTALPRICE_STDERR*SQRT(TOTALPRICE_SUMWGT) ) ;
  OUTPUT ;
  _TYPE_ = 'N' ; TOTALPRICE=TOTALPRICE_SUMWGT ; OUTPUT ;
  DROP TOTALPRICE_DF TOTALPRICE_MEAN TOTALPRICE_N TOTALPRICE_STDDEV
  TOTALPRICE_STDERR TOTALPRICE_SUMWGT TOTALPRICE_VAR DOMAINLABEL ;
RUN ;
```

With TYPE=COV, PROC MIANALYZE reads sample means from observations with _TYPE_ = 'MEAN', sample size n from observations with _TYPE_ = 'N', and covariance matrices for variables from observations with _TYPE_ = 'COV'. We used the EDF parameter to specify the degrees of freedom obtained from PROC SURVEYMEANS (see Table 4) for each parameter estimate. This required a separate PROC MIANALYZE call for each store type. The EDF option is used to specify the complete sample degrees of freedom for an estimate. We used three PROC MIANALYZE calls since we wanted to specify different degrees of freedom for each store type.

```
TITLE CONVENIENCE ;
PROC MIANALYZE DATA=MEANS2 (WHERE=( STORETYPE="CONVENIENCE" )) EDF=68 ;
  VAR TOTALPRICE ;
RUN ;
TITLE SMALL/MEDIUM ;
PROC MIANALYZE DATA=MEANS2 (WHERE=( STORETYPE="SMALL/MEDIUM" )) EDF=59 ;
  VAR TOTALPRICE ;
RUN ;
TITLE SUPERMARKET ;
PROC MIANALYZE DATA=MEANS2 (WHERE=( STORETYPE="SUPERMARKET" )) EDF=45 ;
  VAR TOTALPRICE ;
RUN ;
```

PROC MIANALYZE calculates a total variance that is a combination of within imputation variance (U) and between imputation variance (B). U is calculated as the mean of the variance estimates from the 5 imputations. B is calculated as the variance of the Mean estimates. The total variance, T , is a linear combination of B and U . The coefficient for B is $(1 + 1/m)$, where m is the number of imputations. Hence, with 5 imputations the coefficient for B is $6/5$. Results are shown in Table 5.

TABLE 5: SUMMARY OF PROC SURVEYMEANS AND PROC MIANALYZE ESTIMATES

SURVEYMEANS RESULTS					MIANALYZE RESULTS			
Imputation	Store type	Mean	SE of Mean	Variance Of Mean	Within Variance (U)	Between Variance (B)	Total Variance (U+6/5B)	SE sqrt(T)
1	Convenience	23.23	0.45	0.2062	0.2103	0.0291	0.2452	0.495
2	Convenience	22.79	0.47	0.2172				
3	Convenience	22.96	0.45	0.2052				
4	Convenience	22.84	0.47	0.2223				
5	Convenience	22.93	0.45	0.2006				
1	Small/Medium	18.80	0.31	0.097	0.0945	0.0208	0.1195	0.346
2	Small/Medium	18.97	0.35	0.1217				
3	Small/Medium	18.68	0.29	0.0841				
4	Small/Medium	18.75	0.30	0.087				
5	Small/Medium	18.58	0.29	0.0825				
1	Supermarket	10.80	0.083	0.0068	0.0068	6.44E-06	0.0068	0.083
2	Supermarket	10.80	0.083	0.0068				
3	Supermarket	10.80	0.083	0.0068				
4	Supermarket	10.79	0.082	0.0068				
5	Supermarket	10.79	0.083	0.0068				

The amount of missing information is greatest for convenience stores and smallest for supermarkets. The between variance "B" estimates reflect that uncertainty.

CONCLUSIONS

We welcome the addition of imputation procedures in SAS. Missing data is a difficult problem, particularly in the context of complex surveys. Currently, the procedures are somewhat difficult to use with the survey data analysis procedures. Our calculations were done using SAS version 8.02. It is our understanding that SAS 9.0 has many enhancements. For example, SAS 9.0 MIANALYZE offers a CLASS statement.

REFERENCES

Cox, B.G. (1980). The weighted sequential hot deck imputation procedure. American Statistical Association 1980 Proceedings of the Section on Survey Research Methods, 721-726.

Korn, E.L. and Graubard, B.I. (1999) Analysis of Health Surveys. New York: Wiley.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, D.B. and Schenker, N. (1986). Multiple Imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of American Statistical Association 81, 366-374.

Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association 47, 635-646.

Rob Agnelli, Technical Support Statistician, SAS.

SAS Institute Inc. (1999), *SAS/STAT User's Guide, Version 8*, Cary, NC: SAS Institute Inc.

SAS Institute Inc., SAS OnlineDoc®, Version 8, Cary, NC: SAS Institute Inc., 1999.

ACKNOWLEDGEMENTS

This work was funded under the Lower Mississippi Delta Nutrition Intervention Research Initiative, USDA ARS grant # 6251-53000-003-00D.

CONTACT INFORMATION

Contact the author at:

Jeff Gossett
UAMS Pediatrics / Section of Biostatistics
1120 Marshall St
Slot 512-43
Little Rock, AR 72202
Work Phone: (501) 364-4960
Work Fax: (501) 364-1552
Email: GossettJeffreyM@uams.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies. □