AN EXAMPLE OF ESTIMATION PROBLEMS WHEN DATA UNIQUE TO MULTIPLE SOURCES ARE COLLECTED FOR EACH SAMPLED UNIT

Steven L. Botman, National Center for Health Statistics

Section 1: Introduction

Sometimes survey specifications dictate a data set for each sampled unit that cannot be obtained from a single source (or a single respondent). The survey estimator in this case must reflect data from multiple sources. Special imputation or post-survey adjustment is necessary to compensate for missing data. The 1980 National Natality Survey (NNS), conducted by the National Center for Health Statistics (NCHS), is an example of survey requiring use of multiple sources. Two techniques are suggested to compensate for NNS missing data; however, these create special problems in estimation.

In part, the special imputation or other post-survey adjustment is required because although for each sampled unit several sources may be surveyed, each source may or may not respond. If a source did not respond, every item from that source was missing. A middle ground in such a survey exists between item nonresponse unit nonresponse--source or chunk nonresponse.

To compensate for "source" or "chunk" nonresponse, the survey designer has several options. One option classifies all sampled cases with one or more nonresponding sources as nonresponding "units" (i.e., as complete nonrespondents to the survey). This then requires adjustment of the sampling weights. Another option imputes on an item-by-item basis to compensate for missing data.

Estimation problems result from either option. Classifying survey units as nonrespondents discards "good" responses of survey sources, which decreases the sample size. The use of an extensive item-by-item imputation scheme for source nonresponse allows exploitation of all survey data, but complicates the approximation of the sampling errors. The latter option or technique was used for the NNS. The technique used to compensate for missing data and the resulting problems are described within the context of 1980 NNS.

The next Section describes the NNS purpose and design. Section 3 describes NNS response and imputation strategy. Section 4 provides additional information on response. Section 5 reexamines response according to adjustment strata for imputation. Even though we have reservations about the procedure used, Section 6 outlines how the survey sampling variances were adjusted to reflect the increased sampling variability due to the imputation strategy. Section 7 summarizes the problems associated with the collection of data from multiple sources for each sampled unit.

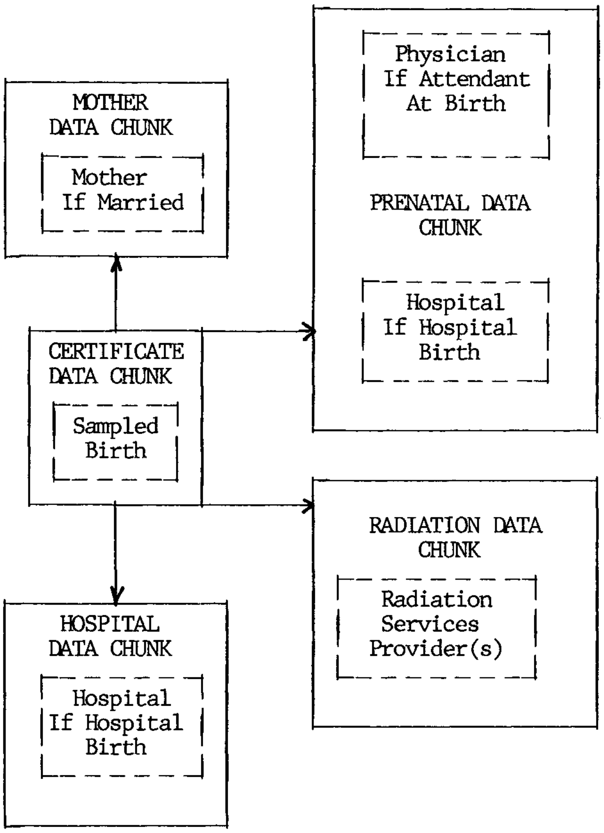
Section 2: NNS Purpose and Design

The purpose of the 1980 NNS is to develop uniform and comparable data beyond that available on the birth certificates themselves. To develop these data, the NNS "followed back" a sample of 9941 birth certificates. State Vital Record registrars selected these sampled certificates according to NCHS specifications. The NNS encompassed births to residents of the United States in 1980. Low-weight births were oversampled. Other details of the sampling plan are described elsewhere.[1]

The NNS employed a complex data collection scheme. For each sampled birth certificate, the NNS sought information on prenatal care, mothers, hospital care, and other health-related areas. To obtain such information the NNS surveyed several sources: the married mother [2]; the hospital where the birth occurred; the non-hospital staff physician attendant at birth; and providers of radiation health care.

Figure 1 shows that the sampled birth certificate in survey data collection played a central role. In fact, the birth certificate itself identifies the name of the mother, place of birth, the physician attendant at birth, and other pertinent information.

Figure 1.--Data Collection Scheme of 1980 National Natality Survey Data Chunks and Sources.



Generally, several eligible sources for each sampled birth were surveyed. Each eligible source, however, may or may not have responded.

Since imputation was used to compensate for missing data, only a few inscope sample births in the NNS were classified as unit nonrespondents (i.e., as complete nonrespondents to the survey). These corresponded to the inscope sample births where the physician's signature on the birth certificate was illegible or the mother questionnaire was returned by the Postal Service as non-deliverable. For this paper, these cases are ignored. Additional information on the imputation strategy and response is presented in the next section.

Section 3: Response and Imputation Strategy

The NNS data collection strategy affected the level of survey response. For example, married mothers were surveyed first and asked to sign a form authorizing medical providers to release information to the NNS. These signed authorizations, when provided, were then sent along with the questionnaires to medical providers. The response rate for the medical provider chunk was greater among the providers receiving a signed authorization than among those providers who were not provided a signed authorization.[3]

Although the NNS source (or chunk) data sets were related, one source data set minimally overlapped any other. Consistency of the data from the different sources for a sampled birth was not forced during editing, for the correct response could not be determined when congruent data were collected from two or more sources for the same sampled birth. For example, the age of the mother reported on a few sampled birth certificates differed with the age reported by the mother.

To impute for missing data, the NNS used a hot deck procedure. Missing data from "item" as well as source nonresponse were generally imputed on an item-by-item basis. For each missing data item was substituted a response for that item from the same weighting stratum. The weighting strata were defined by data on the birth certificate for mother age, marital status, and for the child's race, birth weight and birth order.

The NNS imputation strategy was motivated by the NNS data collection plan, as well as the NNS analysis plan. The NNS analysis plan was not limited to data obtained from an individual chunk. If any statistic was based on data only from a single source, then imputation on an item-by-item basis for missing data could be avoided, since chunk nonresponse could have been addressed in the weighting scheme--the data for each source could have been weighted separately.

Instead, many statistics would be obtained from data from two or more noncertificate data chunks. Since the certificate chunk data were available for each sampled birth, the certificate data chunk did not require imputation. Also since patterns of response help us to understand the effect of missing data, these patterns should be examined.

The next Section discusses patterns of response at the national level for different questionnaire combinations; the Section 5 discusses patterns of response by imputation strata.

Section 4: Further Discussion of Response

Section 4.1 Response to Individual Data Chunks

Table 1 presents the number and percent distribution of the births eligible for the individual survey chunks according to response status. For example, 79.5 percent of the eligible sample mothers responded; 77.9 percent of eligible individual hospital sources responded; and 87.6 percent of eligible medical providers of prenatal care responded. These chunk "response rates" are comparable with the response rates in other surveys.

Section 4.2 Response to Pairs of Data Chunks

Table 2 shows that the number of sampled births simultaneously eligible for each pair of survey sources vary, since every sampled birth is not eligible for every data source. For example, the smallest totals in Table 2 correspond to mother/prenatal pair and the mother/hospital pair, since unmarried mothers were not surveyed for the NNS although their medical sources were surveyed. These patterns of response affect precision of survey statistics.

Let us look at a statistic derived from both the mother and the hospital chunks: suppose a user estimated the total number of smoking married mothers who gave birth in a hospital through use of a Caesarean section. This statistic in the NNS is derived from information in both the mother and the hospital data sources. We, therefore, must examine patterns of response of both the married mothers and hospital sources.

For most births of married mothers and occurring in hospitals, both the mother and the hospital sources responded to the survey. In some cases, either the mother or the hospital, or both sources, did not respond. For the survey, missing data were compensated through imputation. In fact, the occurrence of complete source nonresponse (e.g., hospital) motivated the NNS imputation strategy.

Detailed information on response to the hospital/mother sources is presented in Table 2. Although 65.6 percent of the sampled births eligible for both the hospital and mother chunks had the data obtained from response from both sources, 93.2 percent of the sampled births had response from at least one of the two sources. This response pattern is similar to the patterns for the other pairs of chunks.

Section 4.3 Response and Imputation for Triplets of Data Chunks

Table 3 presents the number and the percent distribution of sampled births simultaneously eligible for the three survey data chunks according to response status. Although for 96.2 percent of these births the NNS had response from at least one of the three sources, only 65.8 percent of the births had response from all three sources.

Section 5: Response by Imputation Strata

The response rate for each of the individual data chunks (i.e., mother, hospital, and prenatal) differed in the adjustment strata used for imputation. For example, 55.2 percent of the mothers in adjustment cell 11 in Table 4 responded; Table 4 shows 79.5 percent of all sampled mothers responded.

The actual sample size supporting survey estimates is smaller than the size of the sample on the file, because the file contains imputation for chunk missing data. Generally the response rates for the data chunks corresponding to white births were higher than the response rates for births of other races; mothers of sampled first-borns generally had higher response rates than mothers of higher order births. These differences in data chunk response rates by strata introduce another dimension to the problem resulting when data unique to different sources are collected for each sampled unit.

Section 6: Addressing the Increased Sampling Variability Due to Imputation

The first approximation to the NNS sampling variances for aggregates was developed using a balanced-repeated-replication (BRR) procedure, which was run against the NNS survey file, including the imputed data. NCHS then produced a generalized variances for the NNS by fitting curves to the survey estimates and their corresponding BRR variances. The least square technique was used to produce these curves for domains defined by birth weight and other characteristics.

The sampling errors for estimated percentages were approximated by multiplying the corresponding simple random sampling error for the percentage by the survey design effect derived from the BRR variances for the base population for the percentage.

Although independent imputation within each replicate sample would allow for better approximation of the sampling variance, the imputation procedures were not independently executed in each replicate sample. There were two primary reasons for this decision. First, the imputation scheme required numerous computer runs. Secondly, in order to allow for independent imputation of missing data in each replicate sample, the length of the survey data record would have to be increased by nearly a factor 20. This increased record length would then cause operational problems. Accordingly, the imputed value used in the full survey was used as appropriate in each replicate half sample.

In order to avoid misleading users of NNS statistics about their precision, the sampling errors had to be adjusted. A simple adjustment was needed. That adjustment was made by applying multiplicative factors to the sampling errors in the generalized error curves. These adjustment factors were derived by assuming that nonresponse to each data chunk was random. Different factors were applied for variances of estimates based on different data chunks and different combinations of data chunks.

Each curve was adjusted to reflect that only a portion of the sample size supporting the base of the percentage was obtained from response--the remainder of the sample data supporting the base was obtained using imputed values for data missing due to nonresponse. That is, the design effect for proportion estimates was modified by a multiplicative constant.

Two alternate techniques were proposed that might better reflect the sampling variance without independently imputing in each replicate sample:

(1) Repeated application of the imputation procedures on the entire sample [4], and

(2) Classification of all sample births with the data from one or more nonresponding data chunks as nonrespondents for the survey and the subsequent re-weighting of the file, as well as application of a variance approximation procedure.

Each technique had drawbacks. The first is time consuming to implement. The last overestimates the sampling variances, because considerable data available from partial response will not be used in estimation.

Section 7: Summary

Although most surveys address the bulk of nonresponse through post-survey weighting adjustments, the National Natality Survey addressed the bulk of nonresponse through item-by-item imputation. This imputation strategy was implemented since for each sampled unit multiple sources were surveyed.

Accordingly, the estimated sampling errors for this survey should be adjusted to reflect the component of error associated with the extensive item-by-item imputation strategy. This paper describes the response rates to the NNS and some of the difficulties encountered when estimating sampling errors.

References and Notes

[1] Keppel, Kenneth G. et al. Methods and Response Characteristics: 1980 National Natality Survey and 1980 National Fetal Mortality Survey. NCHS Vital and Health Statistics Series 2, in preparation for 1984 publication.

[2] In fact, two versions of the mother questionnaire were employed--one was an abbreviated telephone version and the other was the full-length mail version. This means that only a subset of the "mother" source data items (those defined by the full-length mailed mother questionnaire) was obtained from all married sample mothers--this reduced data set was defined by the intersection of the abbreviated telephone mother questionnaire and the full-length mailed questionnaire. There was, therefore, additional chunk nonresponse among those sampled mothers who were surveyed by telephone.

[3] Simpson, G. The Characteristics of Women Providing Consent Statement on the Response Rates of their Medical Providers. For presentation at the October 19-21, 1983, meeting of the Southern Regional Demographic Group in Chattanooga, Tenn. (based on her Master's thesis on the same topic).

[4] Rubin, Donald B. "Multiple Imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse." 1982 ASA Survey Research Methods Section Proceedings.

ACKNOWLEDGEMENT

Kenneth G. Keppel, in the Division of Vital Statistics of NCHS, provided comments and a computer program that was modified by the author to develop the tables in this paper.

Table 1.--Number and Percent Distribution of Response for Eligible Sample Births by the Individual Data Chunks: 1980 National Natality Survey Provisional[1] Data File.

Response Status for Chunk	NNS Chunk By Type					
	Mother Chunk		Hospital Chunk		Prenatal Chunk	
	Number	Percent	Number	Percent	Number	Percent
Total sample eligible for chunk	7825	100.0	9855	100.0	9811	100.0
Response	6223	79.5	7504	76.1	8346	85.1
Nonresponse	1602	20.5	2351	23.9	1465	14.9

[1] These data were obtained by counting sampled births for which imputation was done for one or more data items in the data chunk, since each data item, but not each data chunk, was flagged to indicate if the data for the item was imputed; this may have introduced error in the above tabulation. Also, in constructing this table it was assumed if a hospital responded it provided information on prenatal care; this was not always the case.

Table 2.--Number and Percent Distribution of Response for Eligible Sampled Births to Paired Data Chunks[1]: 1980 National Natality Survey Provisional Data File.

Response Status for Chunk Pairs	NNS Chunk Pair By Type					
	Mother--Hospital [C_1] [C_2]		Hospital--Prenatal [C_1] [C_2]		Mother--Prenatal [C_1] [C_2]	
	Number	Percent	Number	Percent	Number	Percent
Total responding sample of births eligible for Chunk Pair	7765	100.0	9729	100.0	7774	100.0
Response from both C_1 and C_2	5009	64.5	7437	76.4	5559	71.5
Response from C_1 but not C_2	1175	15.1	0	0.0	638	8.2
Response from C_2 but not C_1	1015	13.1	880	9.0	1199	15.4
No response from C_1 and C_2	566	7.3	1412	14.5	378	4.9

[1] These data were obtained by counting sampled births for which imputation was done for one or more data items in the data chunk, since each data item, but not each data chunk, was flagged to indicate if the data for the item was imputed; this may have introduced error in the above tabulation. Also, in constructing this table it was assumed if a hospital responded it provided information on prenatal care; this was not always the case.

Table 3.--Number Eligible and Percent Distribution of the Responding Sample Births According to Response or Nonresponse to Mother, Hospital, and Prenatal Data Chunks: 1980 National Natality Survey Provisional[1] Data File.

Response or Nonresponse Status for Inscope Births Simultaneously Eligible for Mother, Hospital, and Prenatal Chunks				
Mother	Hospital	Prenatal	Number	Percent
	T O T A L		7715	100.0
Response	Response	Response	4993	64.7
Response	Response	Nonresponse	0	0.0
Response	Nonresponse	Response	548	7.1
Response	Nonresponse	Nonresponse	618	8.0
Nonresponse	Response	Response	1002	13.0
Nonresponse	Response	Nonresponse	0	0.0
Nonresponse	Nonresponse	Response	190	2.5
Nonresponse	Nonresponse	Nonresponse	364	4.7

[1] These data were obtained by counting sampled births for which imputation was done for one or more data items in the data chunk, since each data item, but not each data chunk, was flagged to indicate if the data for the item was imputed; this may have introduced error in the above tabulation. Also, in constructing this table it was assumed if a hospital responded that it provided information on prenatal care; this was not always the case.

Table 4: Number of Sample Births Eligible for Mother, Hospital, and Prenatal Data Chunk and Percent Responding According to Data Chunk: 1980 NNS Preliminary Data File

							Individual Data Chunk					
Strata	Strata Composition					Number of Births Sampled	Mother		Hospital		Prenatal	
#	Birth Weight in Grams	Mother Marital Status (M, U)	Race of Child	Age of Mother	Live Birth Order		# Eligible	% Resp.	# Eligible	% Resp.	# Eligible	% Resp.
All	- - - -	- - - -	- - - -	- - - -	- - - -	9941	7825	79.5	9855	76.1	9811	85.1
1	<2500	M	White	<20	All	138	138	64.0	138	76.1	136	82.4
2	<2500	M	White	20-24	1st	205	205	80.0	202	82.7	202	89.1
3	<2500	M	White	20-24	2+	186	186	71.0	185	70.8	183	82.5
4	<2500	M	White	25-29	1st	140	140	83.6	140	75.0	138	90.6
5	<2500	M	White	25-29	2nd	137	137	81.8	137	83.9	134	84.3
6	<2500	M	White	25-29	3+	102	102	77.5	101	76.2	102	86.3
7	<2500	M	White	30-34	1-2	104	104	90.4	103	77.7	104	88.5
8	<2500	M	White	30-34	3+	97	97	72.2	96	75.0	95	84.2
9	<2500	M	White	35+	All	82	82	78.0	82	78.0	82	86.6
10	<2500	M	Other	<20	All	29	29	51.7	29	62.1	28	82.1
11	<2500	M	Other	20-24	All	87	87	55.2	84	69.0	82	75.6
12	<2500	M	Other	25+	All	182	182	65.9	182	71.4	179	79.9
13	<2500	U	White	<20	All	108	---	--	106	76.4	97	81.4
14	<2500	U	White	20-24	All	76	---	--	74	73.0	67	83.6
15	<2500	U	White	25+	All	71	---	--	70	72.9	66	77.3
16	<2500	U	Other	<20	All	173	---	--	171	64.3	159	69.8
17	<2500	U	Other	20-24	All	161	---	--	160	67.5	153	74.5
18	<2500	U	Other	25+	All	101	---	--	101	65.3	96	71.9
19	+2500	M	White	<18	All	110	110	68.2	107	83.2	110	90.9
20	+2500	M	White	18-19	1st	296	296	71.3	296	78.7	295	88.1
21	+2500	M	White	18-19	2+	106	106	75.5	105	79.0	104	85.6
22	+2500	M	White	20-24	1st	909	909	83.2	907	79.2	906	89.6
23	+2500	M	White	20-24	2nd	716	716	79.9	712	81.0	715	91.5
24	+2500	M	White	20-24	3+	237	237	75.5	235	77.0	233	86.7
25	+2500	M	White	25-29	1st	709	709	86.2	703	78.1	708	88.0
26	+2500	M	White	25-29	2nd	778	778	87.1	777	79.7	777	89.6
27	+2500	M	White	25-29	3rd	358	358	82.4	354	74.3	355	87.0
28	+2500	M	White	25-29	4+	139	139	73.4	132	75.0	137	83.9
29	+2500	M	White	30-34	1st	226	226	89.8	222	75.2	226	86.3
30	+2500	M	White	30-34	2nd	343	343	87.8	338	80.2	343	88.9
31	+2500	M	White	30-34	3rd	256	256	88.3	253	81.8	256	89.8
32	+2500	M	White	30-34	4+	171	171	79.5	168	79.8	168	90.5
33	+2500	M	White	35+	1-3	136	136	83.1	136	81.6	136	90.4
34	+2500	M	White	35+	4+	116	116	78.4	113	79.6	115	85.2
35	+2500	M	Other	<20	All	54	54	64.8	54	74.1	54	83.3
36	+2500	M	Other	20-24	1st	96	96	67.7	96	75.0	95	77.9
37	+2500	M	Other	20-24	2+	145	145	58.6	144	65.3	145	71.0
38	+2500	M	Other	25-29	1-2	136	136	68.4	136	69.9	136	80.1
39	+2500	M	Other	25-29	3+	84	84	66.7	84	72.6	83	77.1
40	+2500	M	Other	30+	1-2	110	110	78.2	109	70.6	109	78.9
41	+2500	M	Other	30+	3+	105	105	66.7	105	68.6	103	75.7
42	+2500	U	White	<18	All	141	---	--	141	74.5	136	86.0
43	+2500	U	White	18-19	All	142	---	--	139	71.9	139	80.6
44	+2500	U	White	20-24	1st	148	---	--	145	76.6	144	84.7
45	+2500	U	White	20-24	2+	113	---	--	112	76.8	111	84.7
46	+2500	U	White	25+	All	179	---	--	172	70.9	175	76.0
47	+2500	U	Other	<18	All	128	---	--	128	69.5	128	76.6
48	+2500	U	Other	18-19	All	155	---	--	155	72.3	151	80.1
49	+2500	U	Other	20-24	All	271	---	--	268	67.5	267	73.4
50	+2500	U	Other	25+	All	149	---	--	148	70.3	148	77.7