Redistribution of Weights for Missing Employment Data from the Ohio Family Health Survey

June 2000

Analysis Conducted by
Stacy Hoshaw-Woodard, Ph.D.
Biostatistics Program
The Ohio State University

This analysis was conducted under contract with the Center for Public Health Data and Statistics, Ohio Department of Health

I. Overview of the Problem

In the first round of the sampling for the Ohio Family Health survey, 1859 respondents who indicated that their employer did not offer health insurance, or did not know whether their employer offered insurance, were erroneously skipped out of a sequence of questions that asked about:

1. Number of hours worked per week
2. Industry of primary employer
3. Number of employees working for primary employer
4. Class of employment (Private, Government, or Self-Employed).

1213 of these respondents were successfully recontacted, leaving 646 with missing values. To illustrate the problem, Table 1 provides an abbreviated list of the sample size and sum of the statistical weight of the respondents within each county for those who have missing and non-missing employment values. The statistical weight is the inverse of the probability of selection and can be thought of as the number of people in the population that a sampled individual represents.

Table 1: Illustration of Missing Values

County Number	County Name	Respondents with Non-Missing Values		Respondents with Missing Values	
		Sum of the Statistical Weight	*Sample Size*	*Sum of the Statistical Weight*	*Sample Size*
1	Adams	2444.8	43	1306.3	21
3	Allen	2086.6	3	8703.3	6
5	Ashland	2504.7	4	3092.2	4
7	Ashtabula	6367.1	32	3267.7	16
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
171	Williams	4370.6	7	614.7	1
173	Wood	10877.7	50	1479.2	10
175	Wyandot	1261.7	5	278.8	1
	TOTAL	644160.5	1213	391752.5	646

II. Objective of this Analysis

The statistical weights of the 646 respondents with missing values will be redistributed to the 1213 respondents who are not missing this information, so that the employment information is weighted to comprise the correct population size. For example, the 21 respondents in Adams County with missing values represent a total of 1306.3 residents. These statistical weights will be redistributed among the 43 other respondents in Adams County who have complete employment values, making those 43 respondents representative of 3751.1 residents (See Table 1). This method was chosen over other

imputation methods due to the large percentage of missing values. (Note that the Gallup Organization tried different imputation methods and failed).

The redistribution will be accomplished using the following steps:

1. Create subgroups based on demographic and socioeconomic variables for which both groups have complete data.
2. Assign each of the respondents to the appropriate subgroup.
3. Sum the statistical weights in each subgroup for the respondents with missing data.
4. Distribute the total weight of the respondents with missing data equally among the respondents in the corresponding subgroup who have complete data (i.e. divide the total weight of the missing respondents by the number of respondents with complete data and add this amount to the original weight of the respondents with complete data).

Subgroups will be defined based on differences in demographic and socioeconomic characteristics between the missing and non-missing groups. For variables where there are no differences found between the two groups, it is assumed that random allocation would adequately redistribute the weights. This method assumes that respondents of similar demographic and socioeconomic characteristics will also have similar employment variables.

III. Analysis

All of the analyses were conducted using STATA Statistical Software: Release 6.0. (Stata Corporation, College Station, TX). The first step in this analysis was to investigate differences between respondents in the two groups (with missing and complete employment data) with respect to demographic and socioeconomic variables. The variables included: race, ethnicity, reason for no phone coverage, number of adults in the household, education level, age, gender, family income as a percentage of the Federal Poverty Level (FPL), marital status, number of children in the family, employment status of the spouse or partner, currently has some form of health insurance, chronic health condition, smoking status, and the number of doctor visits in the last year. Table 2a provides the weighted percentage (and sample size) for the categorical variables comparing the two groups and the p-value from a design-based F-test for independence. The standard Pearson's Chi-Square statistic is turned into an F-statistic to account for the survey design. Table 2b provides the mean (and standard deviation) of each of the continuous variables for the two groups along with the p-value from an adjusted Wald test. The adjusted Wald test is based on an approximate F-statistic. The design-based F and Wald tests are the default tests in STATA for the analysis of complex survey data. (See the STATA reference manuals for more information on the calculation of these tests.)

Table 2a: Differences in Demographic and Socioeconomic Values between Respondents with Missing and Non-Missing Values (Categorical Variables)

Variable		NOT MISSING *Weighted %*	*(Sample size)*	MISSING *Weighted %*	*(Sample size)*	F- Test
Race	1=White	0.90	(1107)	0.80	(556)	0.001
(AWGTRACE)	2=Black	0.09	(66)	0.18	(66)	
	3=Asian	0.003	(17)	0.005	(13)	
	4=Other	0.004	(23)	0.01	(11)	
Hispanic	1=Yes	0.03	(37)	0.02	(22)	0.361
(AETHNIC)	2=No	0.97	(1171)	0.97	(621)	
	3=Don't Know	0.0002	(2)	0.002	(3)	
	4=Refused	0.0003	(3)	0	(0)	
Reason for	1=Other	0.005	(16)	0.03	(18)	0.000
No Phone	2=Don't Know	0.002	(2)	0.001	(1)	
(AREASNOP)	3=Refused	0	(0)	0.004	(1)	
	6=Non-Payment	0.025	(31)	0.11	(45)	
	7=Moved	0.009	(15)	0.01	(14)	
	9=No phone	0.004	(4)	0.001	(1)	
	12=Weather	0.03	(35)	0.01	(13)	
	13=Downed Lines	0.025	(37)	0.008	(13)	
	. = Legitimate Skip	0.90	(1073)	0.82	(540)	
Educational	1=< 1st grade	0.002	(1)	0.006	(1)	0.211
Level	2=1-8th grade	0.009	(20)	0.025	(12)	
(AEDUC)	3=Some HS	0.11	(129)	0.16	(94)	
	4=HS Grad	0.46	(554)	0.44	(296)	
	5=Some College	0.21	(249)	0.18	(106)	
	6=Assoc. Degree	0.08	(86)	0.07	(56)	
	7=4 yr. Degree	0.10	(126)	0.08	(53)	
	8=Advanced Degree	0.03	(45)	0.03	(23)	
	98=Don't Know	0.0002	(1)	0.0007	(1)	
	99=Refused	0.0008	(2)	0.007	(4)	
Gender	1=Male	0.38	(439)	0.48	(295)	0.009
(AGENDER)	2=Female	0.62	(774)	0.52	(351)	
Income % of FPL	1=< 63%	0.09	(109)	0.19	(76)	0.000
(AINC_POV)	2=64-100%	0.08	(82)	0.08	(62)	
	3=101-133%	0.06	(74)	0.09	(70)	
	4=134-150%	0.02	(45)	0.03	(27)	
	5=151-200%	0.12	(137)	0.09	(65)	
	6=201-300%	0.17	(225)	0.16	(107)	
	7=>300%	0.39	(454)	0.26	(193)	
	8=Refused	0.07	(87)	0.10	(46)	

Table 2a (continued)

Variable		NOT MISSING Weighted % (Sample size)		MISSING Weighted % (Sample size)		F- Test
Marital Status (AMARITAL)	1=Married	0.58	(746)	0.45	(310)	0.019
	2=Divorced	0.07	(115)	0.11	(87)	
	3=Widowed	0.01	(31)	0.006	(9)	
	4=Separated	0.02	(27)	0.01	(17)	
	5=Never Married	0.26	(249)	0.35	(187)	
	6=Unmarried Couple	0.05	(40)	0.06	(32)	
	7=Don't Know	0.0004	(1)	0.002	(1)	
	8=Refused	0.005	(4)	0.006	(3)	
Spouse Works (ASPSEWRK)	1=Yes	0.53	(675)	0.41	(259)	0.013
	2=No	0.09	(107)	0.10	(78)	
	3=Don't Know	0.003	(1)	0.004	(1)	
	4=Refused	0	(0)	0.001	(1)	
	. =Legitimate Skip	0.37	(430)	0.49	(307)	
Current Insurance (ACURHI)	1=Yes	0.68	(846)	0.43	(301)	0.000
	2=No	0.32	(367)	0.57	(345)	
Chronic Condition (ACHRONIC)	1=Yes	0.31	(405)	0.30	(193)	0.992
	2=No	0.69	(803)	0.69	(451)	
	3=Don't Know	0.003	(4)	0.004	(2)	
	4=Refused	0.00004	(1)	0	(0)	
Current Smoker (ASMOKNOW)	1=Yes	0.285	(388)	0.41	(256)	0.001
	2=No	0.215	(239)	0.13	(106)	
	. =Legitimate Skip	0.51	(586)	0.46	(284)	

Table 2b: Differences in Demographic and Socioeconomic Values between Respondents with Missing and Non-Missing Values (Continuous Variables)

Variable	Non Missing Mean (SD)	Missing Mean (SD)	Wald Test
Age (AAGE)	37.16 (0.63)	32.77 (0.75)	0.000
# Adults in Household (AHHDADLT)	1.57 (0.05)	1.59 (0.08)	0.835
# Children in Household (AFAMCHLD)	0.99 (0.05)	0.99 (0.07)	0.933
# Doctor Visits (ADOCVSIT)	8.16 (2.45)	10.52 (4.18)	0.626

Significant differences (at the 0.05 level) were found between the missing and non-missing respondents on the variables of race, reason for no phone, age, gender, percentage of poverty level, marital status, whether spouse works, current health insurance, and smoking. Specifically, a greater percentage of blacks have missing values than whites; more respondents in the missing group have had no phone at some point during the year; the missing group are, on average, 4 years younger than the non-missing group; those with missing employment values have income at a lower percentage of the FPL; a greater percentage of males have missing values than females; fewer of the respondents in the missing group are married, which also explains why fewer of the respondents in the missing group have spouses that work; a lower percentage of those in the missing group have health insurance; and more of the respondents in the missing group are smokers.

In the redistribution of the weights, the respondents should be matched on the variables in which the two groups differ. The sample size, however, is not large enough to match on all of these variables. The relationships between significant variables were investigated in an attempt to decrease the number of variables needed to be matched on. The continuous variables and the variables with numerous categories were re-categorized to decrease the number of possible subgroups. Re-categorized variables were created as follows: NAGE ("1" = 18-34, "2" = 35-54, "3" = 55+), NMARITAL ("1" = married or unmarried couple, "2" = not married), NEWRACE ("1" = black, "2" = non-black), NINCPOV ("1" = income less than or equal to 133% of the FPL, "2" = greater than 133% of the FPL), NSPSWRK ("1" = Spouse works, "2" = spouse doesn't work, "3" = not married), NPHONE ("1" = had no phone during part of the year, "2" = continuous phone service), and NSMOKE ("1" = current smoker, "2" = never smoked or quit). The variables were re-categorized using either standard groupings, the categories used in the original weighting, or in a manner that maximized the differences between the missing and non-missing groups. Table 3 provides the p-values from the design-based F-tests of the cross-tabulations of the significant variables to investigate relationships among the variables. Current insurance, a working spouse, and percentage of poverty level are highly interrelated and also related to most of the other variables. Those respondents with insurance are more likely to have a working spouse and to have higher income. Respondents with a working spouse are married by definition, tend to be females, and have higher income. Thus, having insurance can be considered a surrogate for a working spouse and therefore also a surrogate for marital status, gender, and income. Age is not significantly related to current insurance, but it is

Table 3: P-values of Cross-tabulations

	NPHONE	NAGE	AGENDER	NINCPOV	NMARITAL	NSPSWK	ACURHI	NSMOKE
NRACE	0.5457	0.051	0.709	0.000	0.000	0.000	0.000	0.008
NPHONE	.	0.685	0.795	0.000	0.120	0.179	0.001	0.014
NAGE	.	.	0.281	0.015	0.000	0.000	0.123	0.019
AGENDER	.	.	.	0.934	0.000	0.000	0.000	0.121
NINCPOV	.	.	.	.	0.000	0.000	0.000	0.000
NMARITAL	.	.	.	.	.	0.000	0.000	0.448
NSPSWK	.	.	.	.	.	.	0.000	0.018
ACURHI	.	.	.	.	.	.	.	0.000

related to marital status, income, and smoking. Race is also related to income and smoking. A lapse in phone service is also related to income. Given these relationships, matching on age, race, and whether the respondent currently has insurance should provide information on rest of the variables. Due to the small number of blacks in the population, it is important to match on race so that blacks are not under represented. The analyses showed more differences in age between the youngest group (18-34) as compared to the 2 older groups, so the age categories were further combined into two groups with the variable NEWAGE ("1" = 18-34, "2" = 35+). The statistical weights were redistributed within each county to preserve the weighting of the sample to the county populations. To conduct the redistribution of the weights, subgroups based on county and the two levels each of NEWAGE, NEWRACE, and ACURHI were created and the statistical weights were redistributed within these subgroups.

In matching within counties, two counties were found to be problematic. Noble County has only one respondent with missing data and one respondent with complete data. Thus, when the statistical weight of the missing respondent is redistributed to the complete respondent, it leaves the stratum with only one PSU (primary sampling unit). Although a point estimate can be obtained for this county, a standard error cannot be calculated (STATA gives an error). In Hancock County, there were three respondents with missing data and no respondents with complete data. Therefore, the statistical weights cannot be redistributed to anyone in that county. These problems were resolved by combining these two counties with adjacent oversampled counties of a similar county category. The county categories classify the counties based on urbanicity, geographic region, and the percentage of employment in manufacturing. Noble County was combined with Guernsey County and Hancock was combined with Putnam County. The variable NSTRATA was created with to incorporate the merger of these counties.

Of the 88 counties in Ohio, 25 were oversampled in this survey. In the oversampled counties, an average of 55 sampled respondents in each county have employers who do not offer insurance, or did not know whether their employer offered insurance; and in the non-oversampled counties, an average of 7.6 sampled respondents in each county have employers who do not offer insurance, or did not know whether their employer offered insurance. Given the small number of respondents in the non-oversampled counties, it would be difficult to match the respondents on all three of the variables (these 3 variables create 8 subgroups within each county). Analyses within each county found that 89% of the non-oversampled counties were all white. Due to these reasons, race will not be used as a matching variable in the non-oversampled counties. Hence, the respondents in the 25 oversampled counties were matched on NSTRATA, NEWAGE, ACURHI, and NEWRACE yielding 244 subgroups; and the 61 remaining non-oversampled counties were matched on NSTRATA, NEWAGE, and ACURHI yielding 200 additional subgroups. Note that not all of the subgroups will contain respondents. Forty-four of the subgroups contained respondents with missing data and no corresponding respondents with complete data. These respondents were reassigned to similar subgroups (subgroups with a difference in one variable), where there were respondents with complete data. The variable to be changed was chosen

on the availability of matching on the two other variables. If there was more than one available subgroup to change to, the variables of poverty level and whether the spouse worked were considered in the reassignment.

IV. Results

The successfulness of the redistribution of the statistical weights was assessed by showing that the redistribution of the weights did not significantly alter the frequencies of the demographic, socioeconomic, and health related variables. This was done by comparing the statewide frequencies of variables calculated using the original weights (FINALWT) and strata (AWGTCNTY) of the 1829 respondents to the frequencies obtained using the new weights (NEWWT) and new strata (NSTRATA) for the 1213 respondents with complete data. 95% confidence intervals for the estimates were calculated and whether the intervals from the original weights and new weights overlapped was observed. Tables 4a and 4b provide the estimates and confidence intervals calculated using the two sets of weights, for a subset of the variables. Table 4a lists the frequencies for the categorical variables and Table 4b presents the means of the continuous variables. Overall, the redistribution worked very well. All of the confidence intervals overlapped, indicating no significant differences, with the majority of the estimates differing by at most one to two percentage points. Note that even matching on all of the significant demographic and socioeconomic variables, ignoring county, did not produce perfect correspondence with the original estimates.

This analysis was conducted on the subset of data that included only those respondents whose employer did not offer insurance, or did not know whether their employer offered insurance. The new weights and strata variables were reintegrated into the full dataset through the creation of the following variables. The variable MISSING is an indicator of whether the respondent had missing employment information ("1" = if missing, "0" = not missing, "." = legitimate skip, i.e. employer offers insurance). REDIST is an indicator variable as to whether the respondent was included in the dataset in which the redistribution of weights was done (i.e. those 1859 whose employer does not offer insurance). The new weights are given in the variable EMPLOYWT. This variables contains (1) the new weights from the 1213 respondents whose employer does not offer insurance, (2) missing values for the 646 respondents who are missing the employment data, and (3) the original weights for the respondents whose employer offers insurance. A new stratification variable called ESTRATA contains (1) the new strata for those respondents whose employer does not offer insurance (i.e. Noble county is incorporated into Guernsey and Hancock is incorporated into Putnam), and (2) the original strata for those respondents whose employer's offer insurance. The ESTRATA and EMPLOYWT variables should be the strata and corresponding weights used whenever the four employment variables are included in an analysis.

Table 4a: Comparison of Frequencies from the Original Weights and New Weights

Variable		% from Original Weight (95% CI)		% from New Weight (95% CI)	
Race	1=White	0.866	(0.84, 0.89)	0.871	(0.83, 0.90)
(AWGTRACE)	2=Black	0.125	(0.10, 0.15)	0.112	(0.08, 0.15)
	3=Asian	0.003	(0.002, 0.007)	0.007	(0.004, 0.01)
	4=Other	0.006	(0.003, 0.01)	0.009	(0.005, 0.02)
Hispanic	1=Yes	0.028	(0.02, 0.04)	0.036	(0.02, 0.06)
(AETHNIC)	2=No	0.971	(0.96, 0.98)	0.964	(0.94, 0.98)
Educational	1=< 1st grade	0.004	(~0, 0.02)	0.002	(~0, 0.01)
Level	2=1-8th grade	0.015	(0.008, 0.03)	0.009	(0.005, 0.02)
(AEDUC)	3=Some HS	0.128	(0.10, 0.16)	0.100	(0.07, 0.13)
	4=HS Grad	0.450	(0.41, 0.49)	0.443	(0.40, 0.49)
	5=Some College	0.202	(0.17, 0.23)	0.234	(0.19, 0.28)
	6=Assoc. Degree	0.077	(0.06, 0.10)	0.076	(0.05, 0.11)
	7=4 yr. Degree	0.089	(0.07, 0.11)	0.109	(0.08, 0.15)
	8=Advanced Degree	0.031	(0.02, 0.05)	0.027	(0.02, 0.04)
Gender	1=Male	0.415	(0.38, 0.45)	0.387	(0.34, 0.44)
(AGENDER)	2=Female	0.585	(0.55, 0.62)	0.612	(0.56, 0.66)
Income % of	1=< 63%	0.125	(0.10, 0.16)	0.082	(0.06, 0.11)
FPL	2=64-100%	0.079	(0.06, 0.10)	0.082	(0.06, 0.11)
(AINC_POV)	3=101-133%	0.071	(0.06, 0.09)	0.064	(0.05, 0.09)
	4=134-150%	0.026	(0.02, 0.04)	0.017	(0.01, 0.03)
	5=151-200%	0.108	(0.08, 0.14)	0.134	(0.10, 0.18)
	6=201-300%	0.168	(0.14, 0.19)	0.178	(0.15, 0.22)
	7=>300%	0.340	(0.31, 0.38)	0.351	(0.31, 0.40)
	8=Refused	0.083	(0.06, 0.11)	0.091	(0.06, 0.13)
Marital Status	1=Married	0.529	(0.49, 0.57)	0.521	(0.47, 0.57)
(AMARITAL)	2=Divorced	0.083	(0.07, 0.10)	0.080	(0.06, 0.11)
	3=Widowed	0.011	(0.007, 0.02)	0.020	(0.01, 0.04)
	4=Separated	0.015	(0.009, 0.02)	0.016	(0.01, 0.03)
	5=Never Married	0.297	(0.26, 0.33)	0.289	(0.24, 0.34)
	6=Unmarried Couple	0.058	(0.04, 0.08)	0.066	(0.04, 0.10)
Current Insurance	1=Yes	0.588	(0.55, 0.62)	0.619	(0.57, 0.67)
(ACURHI)	2=No	0.412	(0.38, 0.45)	0.380	(0.33, 0.43)
General Health	1=Poor	0.016	(0.01, 0.03)	0.013	(0.01, 0.02)
(AGH1)	2=Fair	0.088	(0.07, 0.11)	0.091	(0.07, 0.12)
	3=Good	0.260	(0.23, 0.29)	0.224	(0.19, 0.27)
	4=Very Good	0.336	(0.30, 0.37)	0.323	(0.28, 0.37)
	5=Excellent	0.299	(0.26, 0.36)	0.348	(0.30, 0.39)
Chronic Condition	1=Yes	0.305	(0.27, 0.34)	0.285	(0.25, 0.33)
(ACHRONIC)	2=No	0.692	(0.66, 0.72)	0.712	(0.67, 0.75)
Exercise	1=Yes	0.672	(0.64, 0.71)	0.689	(0.64, 0.73)
(AEXRCISE)	2=No	0.321	(0.29, 0.36)	0.307	(0.26, 0.35)

Table 4b: Comparison of Means from Original Weights and New Weights

Variable	Original Weight		New Weight	
	Mean	*(95% CI)*	*Mean*	*(95% CI)*
Age (AAGE)	35.49	(34.54, 36.45)	35.99	(34.72, 37.26)
# Adults in Household (AHHDADLT)	1.58	(1.49, 1.66)	1.45	(1.35, 1.54)
# Children in Household (AFAMCHLD)	0.99	(0.90, 1.06)	1.01	(0.89, 1.12)
# Doctor Visits (ADOCVSIT)	9.05	(4.75, 13.36)	8.55	(3.31, 13.78)

V. Conclusions

The method of redistribution of statistical weights appears to have worked reasonably well in preserving the original demographic and socioeconomic characteristics of the population while providing the means to obtain population based estimates of the employment variables. This method is a form of imputation, and therefore it is as not accurate as having the actual data from the respondents. We assume that those respondents of similar demographic and socioeconomic characteristics also have similar employment values. There is no way to know the type of bias this redistribution of weights may introduce into the analyses of the employment variables.