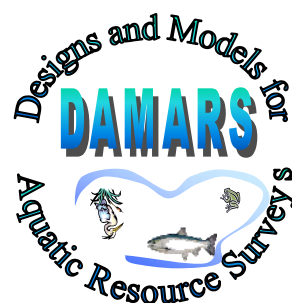


Ignorable Nonresponse Adjustment Procedures and Algorithms

Leigh Ann Harrod and Virginia Lesser
Oregon State University, Department of Statistics

March 2, 2006



The research described in this publication has been funded by the U.S. Environmental Protection Agency through the STAR Cooperative Agreement CR82-9096-01 National Research Program on Design-Based/Model-Assisted Survey Methodology for Aquatic Resources at Oregon State University. It has not been subjected to the Agency's review and therefore does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

TABLE OF CONTENTS

1	INTRODUCTION	1
	1.1 Introduction to nonresponse.....	1
	Table 1: Nonresponse classifications.....	3
	1.2 Determining the nonresponse classification	4
2	CASE STUDY DESCRIPTION.....	6
	2.1 Oregon Coho Salmon Case Study.....	6
	Figure 1: Five Monitoring Areas for monitoring Coho salmon in Oregon’s coastal watersheds.....	6
	Table 2: Definitions of ODFW site survey status and target population.....	7
	Table 3: Status of sample survey sites by survey year.....	7
	2.2 New Mexico Elk Hunter Questionnaire Case Study.....	8
3	DISTINGUISHING BETWEEN NONINFORMATIVE AND INFORMATIVE NONRESPONSE.....	9
	3.1 Case Study 1: ODFW Coho Salmon Survey	9
	Table 4: Number and proportion of unsurveyed and surveyed sites by Monitoring Area and the results of the Chi-squared test of association....	10
	Table 5: Sample statistics for the 1999 ODFW spawner survey	10
	Table 6: Chi-squared test of association between number of landowners for surveyed and unsurveyed sites	11
	3.2 Case study 2: NMDGF elk hunter questionnaire.....	11
	Table 7: Levels of potential covariates for modeling the response probability of elk hunters.....	11
	Table 8: Chi-squared test of association between NMDGF response rates and weapon type	12
	Table 9: Chi-squared test of association between NMDGF response rates and landowner type.....	12
	Table 10: Chi-squared test of association between NMDGF response rates and age class	12
	Table 11: Chi-squared test of association between NMDGF response rates and residency	12
	Table 12: Chi-squared test of association between NMDGF response rates and gender.....	13
4	DISTINGUISHING BETWEEN MAR AND NMAR.....	14
5	MAR DATA ADJUSTMENT METHODS.....	15
	5.1 Identifying weighting adjustment variables.....	15
	5.1.1 Case study 1: ODFW Coho salmon survey	15
	Table 14: ANOVA of spawner AUC by landowner group.....	16
	5.1.2 Case study 2: NMDGF elk hunter questionnaire.....	16
	Table 15: ANOVA of harvest rates by weapon type	16
	Table 16: ANOVA of harvest rates by age class	16
	Table 17: ANOVA of harvest rates by residency	17
	Table 18: ANOVA of harvest rates by gender.....	17

Table 19: ANOVA of harvest rates by landowner type.....	17
5.2 Weighting adjustment methods.....	17
Table 20: Weighting adjustment assumptions.....	18
5.2.1 Weighting class adjustment estimator.....	18
5.2.1.1 Case study 1: ODFW Coho salmon survey.....	19
Table 21: Unadjusted estimates and weighting class adjustment estimates of Coho salmon spawners by survey year and adjusted for numbers of landowners.....	19
5.2.1.2 Case study 2: NMDGF elk hunter questionnaire.....	20
Table 22: Weighting class adjustment calculations by age class.....	20
Table 23: Variance calculation of weighting class adjustment estimator of total harvested elk.....	21
5.2.2 Poststratification adjustment estimator.....	21
5.2.2.1 Case study 1: ODFW Coho salmon survey.....	21
5.2.2.2 Case study 2: NMDGF elk hunter questionnaire.....	22
Table 24: Calculation of poststratification adjustment estimator of total harvested elk.....	22
Table 25: Variance calculation of poststratification adjustment estimator of total harvested elk.....	23
5.3 Formulas to calculate weighting adjustment estimators.....	23
5.3.1 Simple random sampling.....	24
5.3.1.1 Weighting class adjustment.....	24
5.3.1.2 Poststratification adjustment.....	24
5.3.2 Stratified random sampling.....	24
5.3.2.1 Weighting class adjustment.....	24
5.3.2.2 Poststratification adjustment.....	25
5.3.3 Two-stage random sampling.....	25
5.3.3.1 Weighting class adjustment.....	25
5.3.3.2 Poststratification adjustment.....	26
6 CONCLUSIONS.....	27
7 REFERENCES.....	27
8 ACKNOWLEDGMENTS.....	27
APPENDIX A: New Mexico Department of Game and Fish elk hunt licensee questionnaire.....	28
APPENDIX B: General S-PLUS/R functions for weighting adjustment programs.....	29
APPENDIX C: Functions to compute weighting class adjustments for finite population surveys and simple random sampling in R, S-PLUS, and SAS.....	31
APPENDIX D: Functions to compute poststratification adjustments for finite population surveys and simple random sampling in R, S-PLUS, and SAS.....	33
APPENDIX E: Functions to compute weighting class adjustments for finite population surveys and stratified random sampling in R, S-PLUS, and SAS.....	35

APPENDIX F: Functions to compute poststratification adjustments for finite population surveys and stratified random sampling in R, S-PLUS, and SAS37

APPENDIX G: S-PLUS/R Function to compute weighting class adjustments for finite population surveys and two-stage cluster sampling with simple random sampling at both levels39

APPENDIX H: S-PLUS/R Function to compute poststratification adjustments for finite population surveys and two-stage cluster sampling with simple random sampling at both levels40

1 INTRODUCTION

Many environmental scientists encounter missing data when conducting probability surveys. Sampled sites may be difficult to locate or access, landowners may refuse to participate, instruments fail, databases are damaged, and survey questionnaires may be misplaced or damaged. In these cases, the response rate, defined as the proportion of sampled units for which a response is obtained, is less than one. Whenever missing data occurs, nonresponse error occurs. Lohr (1999) warns that bias associated with nonresponse may significantly affect estimates of the parameters of interest if the unsurveyed units are different from the units for which information was obtained.

There are two types of nonresponse: *unit nonresponse* and *item nonresponse*. *Unit nonresponse* occurs when the response of an entire sampled unit cannot be obtained (Lohr, 1999). For example, when an entire site representing an independent sampling unit is not visited, unit nonresponse has occurred. *Item nonresponse* is encountered when an incomplete response is obtained from a sampled unit (Lohr, 1999). For instance, if a subset of measurements from a surveyed site is not collected due to instrument failure, then the measurements are missing. This manual will

address methods used to correct for unit nonresponse in probability surveys. Item nonresponse will not be addressed in this document. Imputation methods commonly used to account for item nonresponse are discussed in Lessler and Kalsbeek (1992), Lohr (1999), and Rubin (1986). Assume for the remainder of this document that complete responses are obtained for surveyed units.

1.1 Introduction to nonresponse

When considering nonresponse, the population of interest is often described as consisting two subpopulations: the subpopulation of respondents and the subpopulation of nonrespondents (Lessler and Kalsbeek, 1992). A random sample chosen from the entire population may represent each subpopulation. If the chosen measure of the variable of interest differs significantly between the respondent and nonrespondent subpopulations, then estimates derived from respondent information may be biased if the objective is to represent the entire population. For example, if fish abundance is higher in accessible sites than in inaccessible sites, then the estimate of total abundance calculated from accessible sites may significantly overestimate the true abundance.

The methods discussed in this report address the bias introduced from nonresponse.

The process determining whether or not a unit responds (or data are missing) is called the *response mechanism*. The methods of dealing with nonresponse discussed in this report depend on the type of the nonresponse mechanism. Three types of nonresponse mechanisms may generate the missing data. These classifications that depend on the nature of the missingness are mechanisms that generate data that are completely missing at random (MCAR), data that are missing at random but depend on the covariates (MAR), and data that are not missing at random (NMAR). Data that are MCAR are missing but the cause of the missingness is not associated with the variable of interest, any auxiliary data, or the survey design. For example, if an observer becomes ill and cannot survey a site, the site is probably missing due to factors that are completely unrelated to the survey. When data are MAR, the nonresponse may be dealt with using auxiliary information that is associated with the response mechanism. For example, an inaccessible survey site may be missing due to

site gradient or distance from the nearest road. The MCAR and MAR response mechanisms are considered *noninformative nonresponse* mechanisms because the response does not differ significantly between the missing and measured units.

When data are NMAR, the nonresponse depends on the value of the response variable and cannot be explained fully by the auxiliary variables. This situation is known as *informative nonresponse* because the value of the response differs between surveyed and unsurveyed units. For example, if inaccessible sites are found in areas where the site gradient is high and fish prefer these areas, then estimates of total fish abundance from surveyed sites will underestimate the true total. Methods to determine if the nonresponse is informative are difficult and generally handled on a case by case basis. Therefore, these methods are not discussed in this report. Methods discussed here will mainly address testing to distinguish between MCAR or MAR nonresponse. The nonresponse classifications are outlined in Table 1

Table 1: Nonresponse classifications

Type of missingness	Definition	Noninformative or informative?
Missing-completely-at-random (MCAR)	Response mechanism does not depend on the variable of interest, any covariates, or the survey design	Noninformative
Missing-at-random (MAR)	Response mechanism depends on related covariates but not the variable of interest	Noninformative
Not-missing-at-random (NMAR)	Response mechanism depends on the variable of interest	Informative

Environmental scientists may estimate population means, totals, and proportions with the Horvitz-Thompson estimator (1952), which extrapolates the variable of interest for each sampled unit by the proportion of the population it represents. The Horvitz-Thompson estimators are unbiased for their respective parameters and are useful in studies in which inclusion probabilities are not constant for all members of the population (Thompson, 1992).

The Horvitz-Thompson estimator of the population total (τ) is given by:

$$\hat{\tau} = \sum_{i=1}^n \frac{y_i}{\pi_i},$$

where y_i is the measurement of the variable of interest for sample unit i and π_i is the inclusion probability of the i^{th} unit. The *inclusion probability* of a population unit is the probability that the unit will be selected in the sample. Inclusion probabilities depend on the sampling design and are necessary in calculating

design-based estimates for probability samples. The *sampling weight* of a population unit is simply the inverse of the inclusion probability and represents the amount by which the sample value is inflated to represent unsampled units in the population. The sampling weight of the i^{th} unit is then calculated as $w_i = \frac{1}{\pi_i}$. For example, in a simple random sample of n units from a population of N units, the inclusion probability of each unit in the population is $\pi_i = \frac{n}{N}$. The weight of each unit in the sample is $w_i = \frac{1}{\pi_i} = \frac{N}{n}$, the factor by which each sampled unit represents the population.

Two unbiased estimates of the variance of the Horvitz-Thompson estimator of the total are given by (Shao, 1999):

$$\text{Var}_1(\hat{\tau}) = \sum_{i \in s} \frac{1 - \pi_i}{\pi_i^2} y_i^2 + 2 \sum_{i \in s} \sum_{j \in s, j > i} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j \pi_{ij}} y_i y_j$$

and

$$\text{Var}_2(\hat{t}) = \sum_{i \in s} \sum_{j \in s, j > i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2,$$

where π_{ij} is defined as the joint inclusion probability of units i and j .

1.2 Determining the nonresponse classification

Distinguishing between the three nonresponse mechanisms is necessary to choose the appropriate method for handling the nonresponse. Ideally, a random subsample of nonrespondents would provide information to compare the variable of interest for respondents and nonrespondents. However, a nonresponding site may not always be subsampled, especially when observer safety is the reason for the nonresponse.

When data are MCAR, the missing data are caused by a response mechanism that is completely unassociated with the variable of interest, any auxiliary information, or the survey design. This implies that the response mechanism causing the nonresponse is a completely random event. For example, if data from a survey site are missed because of a malfunction in survey equipment, the missing data are most likely MCAR. If data are MCAR, the respondent data are representative of the selected sample (Lohr, 1999). In this case, the respondent data is representative of the

population. The effective sample size is reduced to account for the nonresponse. Then the data from the surveyed units are summarized and no further analyses are needed.

However, one must carefully consider all possible correlations of survey factors with the response mechanism. For instance, if the survey equipment malfunctions under certain environmental circumstances such as low temperatures or high elevations, the missing data may systematically bias the estimates from a sample that excludes certain members of the population. In this case, the missing data are MAR and response probabilities may be modeled from temperature, elevation data, or other associated covariates. To make the appropriate adjustments, auxiliary data, such as temperature, collected on at least the respondents.

To determine whether or not data are MCAR, the response rates may be examined by subgroups of covariates of interest that are available for both surveyed and unsurveyed units. Using the covariates, the original sample is divided into mutually exclusive and nonoverlapping subgroups (Lessler and Kalsbeek, 1992). The observational units within each subset are assumed to have similar values of the variable of interest and have equal probabilities of responding. A chi-square test of observed counts is used to detect significant

differences in response rates among subgroup levels. If differences exist, this would imply the response rates among the subgroups are not the same. This would indicate that the nonresponse is not MCAR because the response rates are related to the covariate used to create the subgroups.

If the response probability shows a strong correlation with one or more auxiliary variables, then the data are not considered MCAR and are either MAR or NMAR. Data may be considered MCAR if one can logically conclude that the response mechanism is unrelated to the

sample in any way. One may not necessarily be confident that the data are MCAR if no covariates can be found to model the response probability. When the MCAR assumption can be made, the observed data are treated as a random sample of the population with the attained sample size used for estimation and confidence interval calculation instead of the target sample size. For example, if the intended sample size was 50 sampling units and only 45 of those were surveyed, the effective sample size of 45 units would be used to calculate unbiased estimates.

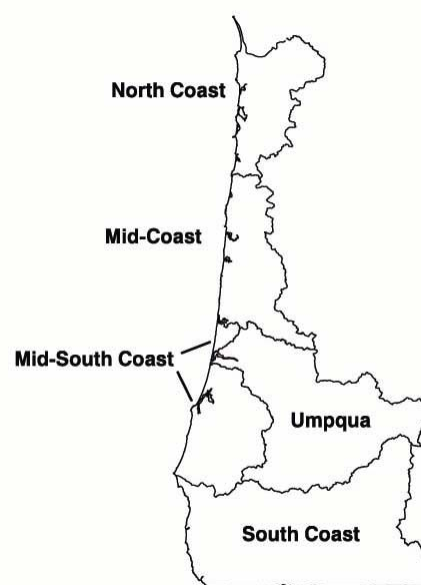
2 CASE STUDY DESCRIPTION

Two case studies will be explored in this document. In both of the case studies examined, data are missing and the estimation procedure needs to account for this missingness. One study involves an environmental survey of spawning Coho salmon sites and the other study is a questionnaire of elk hunters.

2.1 Oregon Coho Salmon Case Study

The first case study implements data from the Oregon Department of Fish and Wildlife (ODFW) annual survey of spawning Coho salmon (*Oncorhynchus kisutch*) in Oregon's five major coastal watersheds (Figure 1). These five Monitoring Areas (MA) act as strata in the probability sample of sites randomly selected each year. Trained observers count spawning salmon and collect auxiliary information at randomly selected sites along Oregon rivers and streams. The response of interest is an estimate of spawner abundance at each site that is calculated from multiple visits to a site. This response is referred to as the *area under the curve* and is called the "AUC." The AUC is calculated from measurements of abundance that are taken no less than 10 days apart and in water with low turbidity for acceptable visibility. For more information on this response measure, see English, *et al.* (1992).

Figure 1: Five Monitoring Areas for monitoring Coho salmon in Oregon's coastal watersheds



Each year, a subset of sites is not surveyed due to various reasons. In some cases, the site is regarded as outside the target population frame because the site does not meet the habitat requirements of spawning Coho salmon (*e.g.*, lacks gravel of the appropriate size, located above a waterfall, etc.). For some sites on private land, the landowners deny ODFW access to survey. Other sites are not surveyed due to physical inaccessibility caused by extreme terrain and vegetation. The status of each survey site is recorded by ODFW and a summary of status types is given in Table 2. When a site is denied access, the landowner has

not provided ODFW with permission to enter his/her private land to measure the response variable. In this case, whether or not the site meets the requirements of the target population cannot be assessed. A status of “Dropped” or “Not Set Up” was given to sites that met the requirements of the target population but time or workload restraints prevented the site from being prepared for surveying or required the

survey load to be reduced. A site given a status of “Discard” or “Zero” was determined to not possess the qualities of Coho salmon spawning habitat and was deemed outside the target population frame. Inaccessible sites were not surveyed due to difficulty in reaching the site location. Table 3 displays the number of sites in the Coho spawner survey by site survey status for survey years 1998-2001.

Table 2: Definitions of ODFW site survey status and target population

ODFW Site Status	Is site in target population?	Description
Denied access	Not evaluated	Permission to access the site was not granted.
Dropped	Yes	Site was dropped because of time or workload constraints.
Discard	No	Site did not meet requirements for population of spawning sites.
Inaccessible	Not evaluated	Site could not be reached safely or was too remote.
No AUC	Yes	Turbidity of water prevented counts from being recorded for more than 10 consecutive days.
Not Set Up	Yes	Site was dropped because of time or workload constraints.
Surveyed	Yes	Survey completed.
Zero	No	Site did not meet requirements for population of spawning sites.

Table 3: Status of sample survey sites by survey year (target population only)

Year	Unsurveyed sites					Surveyed	Total Target Sites in Sample
	Denied Access	Dropped	Inaccessible	No AUC	Not Set Up		
1998	33	18	72	25	1	456	605
1999	43	27	80	22	12	443	627
2000	47	16	51	12	16	478	620
2001	27	16	95	66	3	436	643

2.2 New Mexico Elk Hunter Questionnaire Case Study

The second illustration employs an annual census conducted by New Mexico Department of Game and Fish (NMDGF) of elk hunters in New Mexico. The census of elk hunters does not require that licensees return their completed questionnaire. Elk harvest success rates may differ between survey respondents and nonrespondents if the survey response

mechanism is associated with hunter success. The questionnaire includes questions regarding the elk hunt licensee's hunt activity, success, effort, and location (see Appendix A). This study was selected for illustration due to the large data set and the availability of many auxiliary variables.

3 DISTINGUISHING BETWEEN NONINFORMATIVE AND INFORMATIVE NONRESPONSE

If response rates within levels of auxiliary variables are significantly different, then evidence exists to assume that responses are not MCAR. Chi-square tests are applied in the two case studies to test whether or not data are MCAR. In the Chi-square test of association, the sampled units are classified into groups indicating survey response or nonresponse. The Chi-square test of association tests the null hypothesis that the response probability does not differ for levels of a selected auxiliary variable. The assumptions of Chi-square tests include random sampling, independent sampling units, and the classification of each sampled unit into only one of the two classes. Furthermore, for robust testing, expected values for each cell of the Chi-square test table should total to at least 5 units. Expected values for each cell are calculated as the product of the row and column totals for that cell divided by the total number of units in the survey. If the test is significant, then we conclude that the response rates do depend on the auxiliary variable and data are not MCAR.

3.1 Case study 1: ODFW Coho salmon survey

Response rates for the Coho salmon spawning sites during the surveys in 1998 through 2001 were 75%, 71%, 77%, and 68%, respectively. The large proportion of missing sites should be accounted for in the estimation of the population total. Assessment of the nonresponse mechanism must be made to determine the appropriate estimation method. If AUC, the variable of interest, is not related to the physical factors that affect the response probability of a site, then the MCAR assumption may be made. Given the potential for highly different AUC values in sites where landowner access was not granted, the sites that were denied access by private landowners will not be included in the analysis covered in this report. However, the remaining missing target sites will be examined for the MAR assumption.

Consider the 1999 spawner survey data. As shown in Table 4, a Chi-squared test of association shows no significant difference between the numbers of surveyed and unsurveyed sites among Monitoring Areas (p -value = 0.4521). Therefore, the missing data classification and weighting variables will be explored over all Monitoring Areas combined.

Table 4: Number and proportion of unsurveyed and surveyed sites by Monitoring Area and results of the Chi-squared test of association

Monitoring Area	Unsurveyed sites	Surveyed sites	Total sites
1-NC	34 (25%)	103 (75%)	137
2 MC	36 (28%)	94 (72%)	130
3-MS	33 (26%)	94 (74%)	127
4-UMP	26 (19%)	114 (81%)	140
5-SC	21 (26%)	59 (74%)	80
TOTAL	150 (24%)	464 (76%)	614

$$\chi^2 = 3.67, df = 4, p=0.4521$$

Table 5 displays the sampling statistics for the 1999 spawner survey. Notice that the mean response appears very different for the three classes of landowner number, indicating that the

number of landowners may be an appropriate variable with which to apply weighting adjustments.

Table 5: Sample statistics for the 1999 ODFW spawner survey

Number of landowners	N (stream miles)	Sample size (1-mile reaches)	Surveyed sites	Mean AUC	Sample variance of AUC
1	1889	303	207	11.91	779.31
2	1133	183	142	7.42	190.18
≥3	762	128	115	3.29	40.99
Total	3784	614	464		

If only respondents were analyzed, then the mean and standard error of the mean would be calculated as follows:

Mean AUC for surveyed sites:

$$\sum_{i=1}^n y_i = (207*11.91) + (142*7.42) + (115*3.29) = 3897.36$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{3897.36}{464} = 8.40$$

SE of Mean AUC for surveyed sites:

$$\begin{aligned} \sum_{i=1}^n y_i^2 &= \sum_{h=1}^H [(n_h - 1)s_h^2 + n\bar{y}_h^2] \\ &= [(206)(779.31) + (207)(11.91)^2] \\ &\quad + [(141)(190.18) + (142)(7.42)^2] \\ &\quad + [(114)(40.99) + (115)(3.29)^2] \\ &= 230451.44 \end{aligned}$$

$$s^2 = \frac{\sum_{i=1}^n y_i^2 - n\bar{y}^2}{n-1}$$

$$= \frac{[230451.44 - (464)(8.40^2)]}{463} = 427.02$$

$$SE_{\bar{y}} = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}}$$

$$= \sqrt{\left(1 - \frac{464}{3784}\right) \frac{427.02}{464}} = 0.90$$

A Chi-squared test of association indicates that the probability that a site is surveyed is associated with the number of landowners ($\chi^2 = 23.17$, $df = 2$, $p < 0.0001$), as shown in Table 6. Therefore the null hypothesis that response probabilities among the three landowner classes are equal is rejected and the data are assumed to be not MCAR.

Table 6: Chi-squared test of association between number of landowners for surveyed and unsurveyed sites

Number of landowners	Unsurveyed sites	Surveyed sites	Total sites
1	96 (32%)	207 (68%)	303
2	41 (22%)	142 (78%)	183
≥3	13 (10%)	115 (90%)	128
TOTAL	150 (24%)	464 (76%)	614

$\chi^2 = 23.17$, $df = 2$, $p < 0.0001$

3.2 Case study 2: NMDGF elk hunter questionnaire

Several auxiliary variables from the NMDGF questionnaire may be correlated with response rates and should all be examined. These variables are weapon type, landowner type, age class, state residency, and gender. These potential covariates have the following discrete

classes (or levels) given in Table 7. Hunts are available for rifle, muzzle-loader, and bow weapon types as well as a separate class of hunts for physically-impaired hunters that is treated as a distinct weapon type. The age of the hunter is classified into one of five age categories. Landowner status is classified as public or private.

Table 7: Levels of potential covariates for modeling the response probability of elk hunters

Variable	Levels
Weapon type	Rifle, muzzle-loader, bow, impaired
Landowner type	Public or private
Age class (in years)	<18, 18-34, 35-49, 50- 64, ≥ 65
State residency	Yes or no
Gender	Male or Female

First, a Chi-squared test of association is used to determine if proportions of respondents differ between levels of each variable. If a test is significant, the variable may be associated with the response mechanism and therefore

appropriate for model further consideration in adjusting for nonresponse. Tables 8 - 12 display the results of the Chi-squared tests of association.

Table 8: Chi-square test of association between NMDGF response rates and *weapon type*

Weapon	Nonrespondents	Respondents	Total
Rifle	15198 (70%)	6394 (30%)	21592
Bow	6468 (73%)	2403 (27%)	8871
Muzzle-loader	5156 (69%)	2319 (31%)	7475
Impaired	131 (48%)	142 (52%)	273
All Licensees	26953 (71%)	11258 (29%)	38211

$$\chi^2 = 99.87, df = 3, p < 0.0001$$

Table 9: Chi-square test of association between NMDGF response rates and *landowner type*

Land Type	Nonrespondents	Respondents	Total
Private land	6949 (71%)	2777 (29%)	9726
Public land	20004 (70%)	8481 (30%)	28485
All Licensees	26953 (71%)	11258 (29%)	38211

$$\chi^2 = 5.20, df = 1, p = 0.022$$

Table 10: Chi-square test of association between NMDGF response rates and *age class*

Age Class (years)	Respondents	Nonrespondents	Total
< 18	1376 (79%)	375 (21%)	1751
18 to 34	6793 (80%)	1684 (20%)	8477
35 to 49	11337 (72%)	4419 (28%)	15756
50 to 64	6158 (63%)	3667 (37%)	9825
≥65	1287 (54%)	1112 (46%)	2399
All Licensees	26953 (71%)	11258 (29%)	38211

$$\chi^2 = 1066.87, df = 4, p < 0.0001$$

Table 11: Chi-square test of association between NMDGF response rates and *residency*

Resident of NM?	Nonrespondents	Respondents	Total
Nonresident	7494 (65%)	4104 (35%)	11598
Resident	19459 (73%)	7154 (27%)	26613
All Licensees	26953 (71%)	11258 (29%)	38211

$$\chi^2 = 281.08, df = 1, p < 0.0001$$

Table 12: Chi-square test of association between NMDGF response rates and gender

Gender	Nonrespondents	Respondents	Total
F	1578 (70%)	669 (30%)	2247
M	25375 (71%)	10589 (29%)	35964
All Licensees	26953 (71%)	11258 (29%)	38211

$\chi^2 = 0.11, df = 1, p=0.7394$

The Chi-squared tests of association indicate that all variables except gender have different response rates, i.e. the response rates appear to

be related to weapon type, landowner type, age class, and residency. Therefore, the data are not MCAR and are either MAR or NMAR.

4 DISTINGUISHING BETWEEN MAR AND NMAR

Distinguishing between MAR and NMAR data is more difficult because the correlated variable is not available for both surveyed and unsurveyed units unless further sampling of the nonrespondent subpopulation is conducted. If the response mechanism can be reasonably attributed to a factor that is associated with the response, then the data can be assumed NMAR. For example, if a site is not surveyed because a private landowner refused access to the property on which the survey site is located, then a worst-case scenario might be that the variable of interest on the landowner's property is different from the measure at other sites. When people are given the option of involvement in the response process, Lessler and Kalsbeek (1992) propose that negative factors such as feeling an invasion of privacy, hostility toward the interview sponsor, and subject sensitivity may reduce the response rate. If these feelings are related to the response taken on the landowner's property, then the estimates computed from samples of respondents may be biased due to nonresponse.

The theory of nonresponse employs the concept that the population is comprised of two subpopulations: those who would respond if sampled and those who would not respond (Lessler and Kalsbeek, 1992). Two-phase

sampling (also called double sampling) may be used to obtain an estimate for the nonrespondent portion of the population. For this sampling procedure, a random subsample of the nonrespondents is chosen and additional effort is made to obtain responses for this subsample (Lohr, 1999). The random subsample may be used in a weighted estimator developed by Hansen and Hurwitz (1946) to obtain an unbiased estimate for the population. Two-phase sampling requires added expense and sampling effort and may not be possible depending on the nature of the missingness. For example, if a Coho salmon spawner site was not surveyed due to access issues or lack of landowner permission, then the site will be missing regardless of the structure of the survey.

When data are NMAR and a subsample is not available, methods to assess and estimate nonresponse bias are complex and difficult to generalize. Research is underway to develop explicit methodology for adjusting data NMAR. The scope of this manual will be limited to the adjustment of data that are MAR. However, research by Munoz, Smith, and Lesser (2003) presented at the Joint Statistical Meetings addresses nonresponse models for informative missing data in a Bayesian context.

5 MAR DATA ADJUSTMENT METHODS

When the nonresponse mechanism is MAR, an adjustment can be made to account for the nonresponse. The Chi-square test of association can be used to identify correlated covariates. Techniques that adjust the inclusion probability to account for nonresponse, called *weighting adjustments*, are appropriate in the case that data are MAR. One or more weighting class variables are necessary for the application of weighting adjustments. Kalsbeek and Lessler (1992) suggest that weighting class variables are ideally mutually uncorrelated but highly correlated with the variable of interest. They further discuss the need for adjustment within only a few covariate levels, balancing reduced variance from a few classes against bias reduction from a large number of classes.

The original sample (consisting of both respondents and nonrespondents) is divided into exclusive and nonoverlapping subsets called *adjustment cells*. Members within each cell are assumed to have similar response values with equal response probabilities. Within each class, the inclusion probability is weighted by the observed response probability so that the observed results will reflect the survey respondents and nonrespondents (Lessler and Kalsbeek, 1992).

5.1 Identifying weighting adjustment variables

Analysis of variance (ANOVA) may be used to test the null hypothesis that the levels of an auxiliary variable related to response rates have the same mean level. If the mean response differs significantly by the levels of the auxiliary variable, then that auxiliary variable might be an appropriate variable to consider as a weighting class variable. Assumptions of ANOVA, including independence between units, normally-distributed errors, and constant variance, should be checked before applying the analysis technique.

5.1.1 Case study 1: ODFW Coho salmon survey

To assess the significance of the landowner number variable for weighting adjustments, ANOVA was used to compare the response variable, AUC, across landowner numbers. If AUC means vary significantly across sites grouped by the number of landowners, then the number of landowners may be a suitable variable to consider as a weighting class variable. The analysis of variance, shown in Table 14, indicates that the mean AUC differs significantly by landowner number. Since the response rate and mean response both

demonstrate association with the number of landowners, this variable will be used as the

weighting class adjustment variable.

Table 14: ANOVA of spawner AUC by landowner group

Source	df	SS	MS	F	p-value
Between groups	2	7250.72	3625.36	8.49	0.0002
Within groups	461	196856.22	427.02		
Total	463	204106.94			

5.1.2 Case study 2: NMDGF elk hunter questionnaire

ANOVA was used to examine response rates for levels of several auxiliary variables of interest from the NMDGF elk hunter questionnaire. Harvest rates were found to be significantly different within levels of weapon type (p-value < 0.0001), landowner type (p-value < 0.0001), age class (p-value < 0.0001), and residency (p-

value < 0.0001) but not significantly different between men and women (p-value = 0.5945). These results match the Chi-squared tests of associations for the response rates, indicating that weapon type, landowner type, age class, and residency could be used as weighting adjustment variables.

Table 15: ANOVA of harvest rates by weapon type

Source	df	SS	MS	F	p-value
Weapon	3	83.37	27.79	122.89	<0.0001
Residuals	11254	2544.89	0.23		
Total	11257	2628.26			

Table 16: ANOVA of harvest rates by age class

Source	df	SS	MS	F	p-value
Age	4	6.65	1.66	7.13	<0.0001
Residuals	11253	2621.61	0.23		
Total	11257	2628.26			

Table 17: ANOVA of harvest rates by residency

Source	df	SS	MS	F	p-value
Resident	1	91.26	91.26	404.90	<0.0001
Residuals	11256	2537.00	0.23		
Total	11257	2628.26			

Table 18: ANOVA of harvest rates by gender

Source	df	SS	MS	F	p-value
Gender	1	0.07	0.07	0.28	0.5945
Residuals	11256	2628.19	0.23		
Total	11257	2628.26			

Table 19: ANOVA of harvest rates by landowner type

Source	df	SS	MS	F	p-value
Landowner	1	238.72	238.72	1124.47	<0.0001
Residuals	11256	2839.54	0.21		
Total	11257	2628.26			

To determine which variables among several should be used, ANOVA models incorporating all significant variables could be used with the extra-sum-of-squares F-test, respectively, to examine interactions and reduce the model to only significant variables. Alternatively, the weighting adjustments could be calculated for each of the significant variables and the variable or combination of variables that provided the most precise estimates could be chosen. Generally, adjusted estimates for different variables differ very little. For simplicity in the example, age class will be used as the weighting adjustment variable.

5.2 Weighting adjustment methods

Two types of weighting adjustment strategies to deal with MAR nonresponse are the *weighting class adjustment* and the *poststratification adjustment*. These methods are based on assumptions of a MAR missing data mechanism, equal response probabilities within weighting classes, and a sample including at least one respondent within each weighting class (Oh and Scheuren, 1983). The weighting class adjustment requires the additional assumption that class membership is known for all sample members. However, for poststratification the weighting class membership is not necessary for each nonrespondent. Table 20 provides a synopsis of the assumptions of the two types of weighting adjustments.

Table 20: Weighting adjustment assumptions

Assumption	Weighting class adjustment	Poststratification adjustment
MAR data	X	X
Equal response probabilities within weighting classes	X	X
At least one respondent within each weighting class	X	X
Class membership known for entire sample	X	
Population subgroup totals known		X

In order to estimate a mean or total, the Horvitz-Thompson estimator accounts for the inclusion probability of each unit in the sample. Inclusion probabilities are defined as the probabilities that sampled units are included in the sample. This probability can be partitioned into the probability of selecting the unit into the sample and the probability that a response for the unit is obtained. Weighting adjustment estimates adjust these inclusion probabilities within each weighting class to obtain an estimator of the Horvitz-Thompson form that accounts for incomplete samples. For the weighting class adjustment, the subgroup totals within each weighting class are unknown and response rates within each weighting class must be estimated, resulting in an additional source of variation in the estimate. The poststratification adjustment can employ a direct calculation of the response rate within a weighting class because the subgroup totals are known.

5.2.1 Weighting class adjustment estimator

Let y_{hi} denote the measurement of the variable of interest for the i^{th} unit in class h ,

$$\hat{\phi}_h = \frac{m_h}{n_h} = \text{the observed response rate in weighting class } h$$

n_h = the total sample size in class h , and
 m_h = the responding sample size in class h .

The weighting class adjustment estimate of the population total is calculated as:

$$\begin{aligned} \hat{\tau}_{WC} &= \sum_{h=1}^H \frac{y_{hi}}{\pi_i \hat{\phi}_h} = \sum_{h=1}^H \frac{w_i y_{hi}}{\hat{\phi}_h} \\ &= \sum_{h=1}^H \frac{N n_h y_{hi}}{n m_h} \end{aligned} \quad (1)$$

One expression of the variance of the weighting class adjustment estimator under finite sampling for large m_h given by Oh and Scheuren (1983), is:

5.2.1.1 Case study 1: ODFW Coho salmon survey

$$\begin{aligned} \text{Var}(\hat{\tau}_{wc}) &= \left(\frac{N}{n}\right) \sum_{h=1}^H \frac{n_h N}{n} (\bar{y}_h - \bar{y})^2 \\ &+ \sum_{h=1}^H \left(\frac{n_h N}{n}\right)^2 \left(1 - \frac{n \hat{\phi}_h}{N}\right) \frac{\hat{V}_h}{m_h} \\ &+ \sum_{h=1}^H \left(\frac{n_h N}{n}\right)^2 \left(1 - \frac{m_h}{n}\right) \frac{\hat{V}_h}{m_h^2} \end{aligned} \quad (2)$$

where

N = the total population size,

n = the total sample size,

$$\bar{y}_h = \frac{\sum_{i=1}^{m_h} y_{hi}}{m_h},$$

$$\bar{y} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} y_{hi}}{n}, \text{ and}$$

\hat{V}_h = the estimated variance of y_{hi} for all units in class h .

Earlier analysis indicated that the number of landowners for each site is associated with the probability that a site is surveyed. Weighting class adjustments were applied to the data from each year between 1998 through 2001, inclusive, using Monitoring Areas as strata. The weighting class adjustment estimates for surveys conducted from 1998 through 2001 are provided in Table 21. Adjustments added from 5,000 to 35,000 spawners to the unadjusted estimates. For 1998-2000, the confidence intervals of unadjusted and adjusted estimates generally overlap to a small degree. The adjustment for 2001 is very large, reflecting a high nonresponse rate combined with mean spawner counts that vary greatly between weighting classes.

Table 21: Unadjusted estimates and weighting class adjustment estimates of Coho salmon spawners by survey year and adjusted for number of landowners

Year	Unadjusted Estimates			Weighing Class Adjustment Estimates		
	Estimate	SE	95% - CI	Estimate	SE	95% - CI
1998	15929	1405	(13175, 18682)	20615	2342	(16025, 25205)
1999	21417	1707	(18070, 24763)	28523	2771	(23092, 33954)
2000	34483	2436	(29708, 39258)	39697	3195	(33435, 45959)
2001	81263	5693	(70105, 92421)	116722	9682	(97745, 135699)

5.2.1.2 Case study 2: NMDGF elk hunter questionnaire

Since the entire elk hunt licensee population was censused, each unit has the same weight, $\pi_i = \pi = 1$, for all units. Therefore, the inclusion probability and weight of each member of the

sample is 1 (*i.e.* $w_i = 1$ for all i). Since age class was determined from earlier analysis to be correlated with response rates, this variable was used to create weighting cells for the weighting class adjustment. Table 22 displays the calculations for the weighting class adjustment of the total elk harvested.

Table 22: Weighting class adjustment calculations by age class

Age class	n_h (Number of sampled licensees)	m_h (Number of responding licensees)	y_h (Number of licensees reporting a harvested elk)	$\hat{\phi}_h = \frac{m_h}{n_h}$	$\frac{w_i y_{hi}}{\hat{\phi}_h}$
< 18	1751	375	121	0.21	565
18 to 34	8477	1684	656	0.20	3302
35 to 49	15756	4419	1729	0.28	6165
50 to 64	9825	3667	1319	0.37	3534
≥ 65	2402	1113	356	0.46	768
TOTAL	38211	11258	4181		14334

The weighting class adjustment estimate of the total number of elk harvested by nonrespondents is estimated as:

$$\hat{\tau}_{WC} = \sum_{h=1}^H \frac{w_i y_{hi}}{\hat{\phi}_h} = 14334.$$

Given that $N = n = 38211$, $m = 11258$, and the grand mean (subsample harvest rate)

$$\bar{y} = \frac{4181}{11258} = 0.37, \text{ the variance components}$$

are estimated as given in Table 23.

Table 23: Variance calculation of weighting class adjustment estimator of total harvested elk

Age Class (years)	n_h	m_h	y_h	$\hat{\phi}_h$	\bar{y}_h	V_h	$\frac{N}{n} \left(\frac{n_h N}{n} \right) (\bar{y}_h - \bar{y})^2$	$\left(\frac{n_h N}{n} \right)^2 \left(1 - \frac{n \hat{\phi}_h}{N} \right) \frac{\hat{V}_h}{m_h}$	$\left(\frac{n_h N}{n} \right)^2 \left(1 - \frac{m_h}{n} \right) \frac{\hat{V}_h}{m_h^2}$
< 18	1751	375	121	0.21	0.32	0.22	4.16	1407.96	4.73
18 to 34	8477	1684	656	0.20	0.39	0.24	2.80	8136.40	5.76
35 to 49	15756	4419	1729	0.28	0.39	0.24	6.23	9629.81	2.68
50 to 64	9825	3667	1319	0.37	0.36	0.23	1.34	3801.03	1.50
≥65	2402	1113	356	0.46	0.32	0.22	6.38	605.73	0.98
TOTAL	38211	11258	4181		0.37		20.91	23580.93	15.65

Then, from equation (2), the variance of the estimated total number of elk harvested is:

$$\begin{aligned} \text{Var}(\hat{\tau}_{wc}) &= 20.91 + 23580.93 + 15.65 \\ &= 23617.49 \end{aligned}$$

A 95%-confidence interval on the total number of elk harvested is (14033, 14635).

5.2.2 Poststratification adjustment estimator

The poststratification adjustment estimate of the population total for finite population sampling is calculated as:

$$\begin{aligned} \hat{\tau}_{PS} &= \sum_{h=1}^H \frac{Y_{hi}}{\pi_i \phi_h} \\ &= \sum_{h=1}^H \frac{w_i Y_{hi}}{\phi_h} \quad (3), \\ &= \sum_{h=1}^H \frac{N_h Y_{hi}}{m_h} \end{aligned}$$

where ϕ_h is the true response weighting in class h . One form of the variance estimate of the poststratification adjustment estimator for finite population sampling, as given by Oh and Scheuren (1983), is:

$$\begin{aligned} \text{Var}(\hat{\tau}_{PS}) &\doteq \sum_{h=1}^H N_h^2 \left(1 - \frac{m_h}{N_h} \right) \frac{\hat{V}_h}{m_h} \\ &+ \sum_{h=1}^H N_h^2 \left(1 - \frac{m_h}{n} \right) \frac{\hat{V}_h}{m_h^2}. \end{aligned} \quad (4)$$

5.2.2.1 Case study 1: ODFW Coho salmon survey

Given that the number of owners cannot be identified for every mile of stream in the spawning habitat population, the poststratification adjustment cannot be used to estimate the total abundance of Coho spawners.

5.2.2.2 Case study 2: NMDGF elk hunter questionnaire

As before, the inclusion probability of each sampled licensee is the same for all licensees in the sample and is 1 (*i.e.* $w_i = 1$ for all i) because the licensee population was censused. Because the total number of licensees for each age class is known from the statewide licensee database,

the poststratification adjustment may be applied to the survey data.

The licensee numbers are distributed over the age classes as shown in Table 24, which also includes the calculations for the poststratification adjustment of the total elk harvested.

Table 24: Calculation of poststratification adjustment estimator of total harvested elk

Age class	N_h (Number of licensees in population)	n_h (Number of sampled licensees)	m_h (Number of responding licensees)	y_h (Number of licensees reporting a harvested elk)	$\phi_h = \frac{m_h}{N_h}$	$\frac{w_i y_{hi}}{\phi_h}$
< 18	1751	1751	375	121	0.21	565
18 to 34	8477	8477	1684	656	0.20	3302
35 to 49	15756	15756	4419	1729	0.28	6165
50 to 64	9825	9825	3667	1319	0.37	3534
≥ 65	2402	2402	1113	356	0.46	768
TOTAL	38211	38211	11258	4181		14334

The estimate of the total number of elk harvested is estimated as:

$$\hat{\tau}_{PS} = \sum_{h=1}^H \frac{w_i y_{hi}}{\phi_h} = 14334.$$

Notice that because the licensee population was censused, the poststratification and weighting class adjustments provide identical estimates.

The calculations for the variance components are given in Table 25. The variance estimates are also identical for the poststratification and weighting class adjustments because the population was censused so the population sizes within weighting classes did not need to be estimated.

Table 25: Variance calculation of weighting class adjustment estimator of total harvested elk

Age class	N_h	n_h	m_h	y_h	$\hat{\phi}_h$	\bar{y}_h	V_h	$\left(\frac{n_h N}{n}\right)^2 \left(1 - \frac{n \hat{\phi}_h}{N}\right) \frac{\hat{y}_h}{m_h}$	$\left(\frac{n_h N}{n}\right)^2 \left(1 - \frac{m_h}{n}\right) \frac{\hat{y}_h}{m_h^2}$
< 18	1751	1751	375	121	0.21	0.32	0.22	1407.96	4.73
18 to 34	8477	8477	1684	656	0.20	0.39	0.24	8136.40	5.76
35 to 49	15756	15756	4419	1729	0.28	0.39	0.24	9629.81	2.68
50 to 64	9825	9825	3667	1319	0.37	0.36	0.23	3801.03	1.50
≥65	2402	2402	1113	356	0.46	0.32	0.22	605.73	0.98
TOTAL	38211	38211	11258	4181		0.37		23580.93	15.65

Then, by (4), the poststratification variance of the estimated total number of elk harvested is:

$$\text{Var}(\hat{\tau}_{ps}) = 23580.93 + 15.65 = 23596.58$$

and a 95%-confidence interval on the total number of elk harvested is (14033, 14635). In general, because the subgroup totals are known for the poststratification adjustment, this estimator is more precise than the weighting class adjustment estimator. The need to estimate subgroup totals for the weighting class adjustment estimator requires the additional variance component and increases the confidence interval width.

5.3 Formulas to calculate weighting adjustment estimators

Estimators are examined for simple random, stratified random, and two-phase samples for finite population sampling. The form of the

inclusion probability will be shown to depend on the sampling design. Functions to calculate these estimates are given in the appendices; functions necessary to run these programs are given in Appendix B. Programs in SAS will yield larger variances than the programs provided in S-Plus and R. The S-Plus and R functions incorporate the variances given by Oh and Scheuren (1983). The programs in SAS employ a Taylor series expansion method to estimate the variance as given by Woodruff (1971). This method employs a first-order linear approximation based on the variance of the PSU's and ignoring the variance at other hierarchical levels. This method tends toward conservatism, overestimating the true variance if there is greater variation between clusters than within clusters (as is often the case). Therefore, the authors recommend conducting analysis with the provided S-Plus and R functions rather than the SAS programs if at all possible.

5.3.1 Simple random sampling

Indices: $h = 1, \dots, H$

Inclusion probability: $\pi_i = \frac{n}{N}$

Sampling weight: $w_i = \frac{N}{n}$

5.3.1.1 Weighting class adjustment

Estimator of the total:

$$\hat{\tau}_{wc} = \sum_{h=1}^H \frac{Nn_h \tilde{y}_h}{nm_h}, \text{ where } \tilde{y}_h = \sum_{i=1}^{m_h} y_{hi}$$

Variance estimator (finite population sampling):

$$\begin{aligned} \text{Var}(\hat{\tau}_{wc}) &\doteq \left(\frac{N}{n}\right) \left(\frac{N-n}{N-1}\right) \sum_{h=1}^H \frac{n_h N}{n} (\bar{y}_h - \bar{y})^2 \\ &+ \sum_{h=1}^H \left(\frac{n_h N}{n}\right)^2 \left(1 - \frac{n\bar{m}_h}{n_h N}\right) \frac{\hat{V}_h}{\hat{m}_h} \\ &+ \sum_{h=1}^H \left(\frac{n_h N}{n}\right)^2 \left(1 - \frac{\bar{m}_h}{n}\right) \frac{\hat{V}_h}{\hat{m}_h^2} \end{aligned}$$

See Appendix C for the SAS, S-Plus, and R code to compute this estimate and variance.

5.3.1.2 Poststratification adjustment

Estimator of the total: $\hat{\tau}_{PS} = \sum_{h=1}^H \frac{N_h \tilde{y}_h}{m_h}$

Variance estimator (finite population sampling):

$$\begin{aligned} \text{Var}(\hat{\tau}_{PS}) &\doteq \sum_{h=1}^H N_h^2 \left(1 - \frac{n\bar{m}_h}{n_h N}\right) \frac{\hat{V}_h}{\hat{m}_h} \\ &+ \sum_{h=1}^H N_h^2 \left(1 - \frac{\bar{m}_h}{n}\right) \frac{\hat{V}_h}{\hat{m}_h^2} \end{aligned}$$

See Appendix D for the SAS, S-Plus, and R code to compute this estimate and variance.

5.3.2 Stratified random sampling

Inclusion probability: $\pi_{hij} = \frac{n_j}{N_j}$

Sampling weight: $w_{hij} = \frac{N_j}{n_j}$

5.3.2.1 Weighting class adjustment

Estimator of the total:

$$\hat{\tau}_{wc} = \sum_{j=1}^J \sum_{h=1}^H \frac{Nn_{hj} \tilde{y}_{hj}}{nm_{hj}}, \text{ where } \tilde{y}_{hj} = \sum_{i=1}^{m_{hj}} y_{hij}$$

Variance estimator (finite population sampling):

$$\begin{aligned} \text{Var}(\hat{\tau}_{wc}) &\doteq \sum_{j=1}^J \left[\left(\frac{N}{n}\right) \left(\frac{N-n}{N-1}\right) \sum_{h=1}^H \frac{n_{hj} N}{n} (\bar{y}_{hj} - \bar{y}_j)^2 \right. \\ &+ \sum_{h=1}^H \left(\frac{n_{hj} N}{n}\right)^2 \left(1 - \frac{n\bar{m}_{hj}}{n_{hj} N}\right) \frac{\hat{V}_{hj}}{\hat{m}_{hj}} \\ &\left. + \sum_{h=1}^H \left(\frac{n_{hj} N}{n}\right)^2 \left(1 - \frac{\bar{m}_{hj}}{n}\right) \frac{\hat{V}_{hj}}{\hat{m}_{hj}^2} \right] \end{aligned}$$

See Appendix E for the SAS, S-Plus, and R code to compute this estimate and variance.

5.3.2.2 Poststratification adjustment

Estimator of the total:

$$\hat{\tau}_{PS} = \sum_{j=1}^J \sum_{h=1}^H \frac{N_{hj}}{m_{hj}} \tilde{y}_{hj}, \text{ where } \tilde{y}_{hj} = \sum_{i=1}^{m_{hj}} y_{hij}$$

Variance estimator (finite population sampling):

$$\begin{aligned} \text{Var}(\hat{\tau}_{wc}) \doteq & \sum_{j=1}^J \left[\sum_{h=1}^H \left(\frac{n_{hj}N}{n} \right)^2 \left(1 - \frac{n\bar{m}_{hj}}{n_{hj}N} \right) \frac{\hat{V}_{hj}}{\hat{m}_{hj}} \right. \\ & \left. + \sum_{h=1}^H \left(\frac{n_{hj}N}{n} \right)^2 \left(1 - \frac{\bar{m}_{hj}}{n} \right) \frac{\hat{V}_{hj}}{\hat{m}_{hj}^2} \right] \end{aligned}$$

See Appendix F for the SAS, S-Plus, and R code to compute this estimate and variance.

5.3.3 Two-stage random sampling

This sampling scheme is used when sampling units that occur in clusters. In a two-stage random sample, a sample of primary sampling units (PSU's) is randomly selected, then random subsamples of secondary sampling units (SSU's) are selected within each selected PSU. These results may be generalized to any number of successive sampling stages with a variety of sampling schemes used at each stage. For illustration, we will assume that two stages are used, that there is no missingness at the PSU level, and that simple random samples are selected at both stages. SAS programs to compute weighting adjustments for two-stage

samples are currently not available because the weights cannot be adjusted by weighting classes in SAS. For two-stage samples, it is recommended that analysis be conducted in S-Plus or R.

Inclusion probability: $\pi_{hi} = \frac{nm_i}{NM_i}$

Sampling weight: $w_{hij} = \frac{NM_i}{nm_i}$

5.3.3.1 Weighting class adjustment

Estimator of the total:

$$\begin{aligned} \hat{\tau}_{wc} &= \sum_{i=1}^n \sum_{h=1}^H \sum_{j=1}^{m_{Rhi}} \frac{NM_i m_{hi} \tilde{y}_h}{nm_i m_{Rhi}}, \\ \text{where } \tilde{y}_h &= \sum_{i=1}^{m_h} y_{hi} \end{aligned}$$

Variance estimator (finite population sampling):

$$\text{Var}(\hat{\tau}_{wc}) = \sum_{i=1}^n N(N-n) \frac{\hat{V}_\tau}{n} + \frac{N}{n} \sum_{i=1}^n \text{Var}(\hat{\tau}_{wc-i}),$$

where \hat{V}_τ is the variance between PSU estimates and

$\text{Var}(\hat{\tau}_{wc-i})$ is the weighting class variance of the i^{th} PSU estimate.

See Appendix G for the S-Plus and R code to compute this estimate and variance.

5.3.3.2 Poststratification adjustment

Estimator of the total:

$$\hat{\tau}_{PS} = \sum_{i=1}^n \sum_{h=1}^H \sum_{j=1}^{m_{Rhi}} \frac{NM_{hi} \tilde{y}_h}{nm_{Rhi}},$$

$$\text{where } \tilde{y}_h = \sum_{i=1}^{m_h} y_{hi}$$

Variance estimator (finite population sampling):

$$\text{Var}(\hat{\tau}_{PS}) = \sum_{i=1}^n N(N-n) \frac{\hat{V}_\tau}{n} + \frac{N}{n} \sum_{i=1}^n \text{Var}(\hat{\tau}_{PS-i}),$$

where \hat{V}_τ is the variance between PSU estimates and

$\text{Var}(\hat{\tau}_{PS-i})$ is the poststratification variance of the i^{th} PSU estimate.

See Appendix H for the S-Plus and R code to compute this estimate and variance.

6 CONCLUSION

Missing data is a potentially serious problem and no methodology can improve on a complete data set. However, techniques are available to produce unbiased estimates of population parameters when data are missing. When the

missing data mechanism is MAR and associated covariates are available, weighting adjustments are an appropriate tool to adjust inclusion probabilities to account for the missing responses.

7 REFERENCES

- Cochran, W. G. (1977). *Sampling Techniques*. 3rd edition. New York: Wiley.
- English, K. K., R. C. Bocking, and J. R. Irvine. 1992. A robust procedure for estimating salmon escapement based on the area-under-the-curve method. *Canadian Journal of Fisheries and Aquatic Sciences* 49:1982-1989.
- Horvitz, D.G. and D.J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663-685.
- Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley.
- Lohr, S. (1999). *Sampling: Design and Analysis*. New York: Duxbury.
- Munoz, B., G. Lesser, and R. Smith (2003). "Model-Based Approaches for Handling the Non-ignorable Missing Data Mechanism for Inference in Environmental Surveys," *Proceedings of the American Statistical Association Joint Statistical Meetings*.
- Oh, H.L. and Scheuren, F.J. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, W.G. Madow, I. Olkin, and D.B. Rubin (eds), 143-184. New York: Academic Press.
- Shao, Jun (1999). *Mathematical Statistics*. Springer-Verlag: New York.
- Thompson, S.K. (1992). *Sampling*. New York: Wiley.
- Woodruff, R.S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association* 66: 411-414.

8 ACKNOWLEDGEMENTS

The authors would like to thank the Oregon Department of Fish and Wildlife for their 1998-2001 Coho salmon spawner survey data and the

New Mexico Department of Game and Fish for the use of their 2001-02 elk harvest survey data for manual examples.

APPENDIX A: NEW MEXICO DEPARTMENT OF GAME AND FISH ELK HUNT LICENSEE QUESTIONNAIRE

Please return your report to: New Mexico Department of Game & Fish, P0 Box 9254, Santa Fe, NM 87504-9734 within 15 days of your hunt.

Hunter Report ***1999-2000 Hunt Season***

or Complete your report on the Internet

www.ushunter.com

1. Did you hunt NM **Elk** In 1999-2000? **YES** **NO** (if NO, please stop and mail today)
2. In which Game Unit (GMU) did you hunt MOST? DAYS hunted in that unit?
3. Did you harvest an **Elk** ? **YES** **NO** (if NO, skip to question 6.)
4. What was the sex of your harvest? Male Female Sex Unknown
5. In which Game Unit (GMU) did you harvest?
6. How can we improve your hunt? *Please respond on back of report.*
7. Do you have any suggestions for improving the report process? *Please respond on back of report.*

Thank you for your time and cooperation.

APPENDIX B: GENERAL S-PLUS/R FUNCTIONS FOR WEIGHTING ADJUSTMENT PROGRAMS

```
Classes.fcn <- function(mat,col)
{
    classmat <- data.frame(mat[,seq(4,col)]) # Data frame of weighting
class variables
    classes <- unique.data.frame(classmat) # Table of unique
levels of the WC vars
    colclass <- dim(classes)[2]
    rowclass <- dim(classes)[1]
    if(colclass==1) classes <- matrix(sort(unlist(classes)),length(unlist(classes)),1)
    if(colclass==2) classes <- classes[order(classes[,1],classes[,2]),]
    if(colclass==3) classes <- classes[order(classes[,1],classes[,2],classes[,3]),]
    if(colclass==4) classes <- classes[order(classes[,1],classes[,2],classes[,3],classes[,4]),]
    if(colclass==5) classes <-
classes[order(classes[,1],classes[,2],classes[,3],classes[,4],classes[,5]),]
    if(colclass==6) classes <-
classes[order(classes[,1],classes[,2],classes[,3],classes[,4],classes[,5], classes[,6]),]
return(classes)
}

ClassesPS.fcn <- function(mat,col)
{
    classmat <- data.frame(mat[,seq(5,col)]) # Data frame of weighting
class variables
    classes <- unique.data.frame(classmat) # Table of unique
levels of the WC vars
    colclass <- dim(classes)[2]
    rowclass <- dim(classes)[1]
    if(colclass==1) classes <- matrix(sort(unlist(classes)),length(unlist(classes)),1)
    if(colclass==2) classes <- classes[order(classes[,1],classes[,2]),]
    if(colclass==3) classes <- classes[order(classes[,1],classes[,2],classes[,3]),]
    if(colclass==4) classes <- classes[order(classes[,1],classes[,2],classes[,3],classes[,4]),]
    if(colclass==5) classes <-
classes[order(classes[,1],classes[,2],classes[,3],classes[,4],classes[,5]),]
    if(colclass==6) classes <-
classes[order(classes[,1],classes[,2],classes[,3],classes[,4],classes[,5], classes[,6]),]
return(classes)
}
```

```

GetMath.fcn <- function(classi, mat, col)
{
  classno <- length(classi)

  if (classno==1) matH <- mat[mat[,4]==classi[1],]
  if (classno==2) matH <- mat[(mat[,4]==classi[1])&(mat[,5]==classi[2]),]

  if (classno==3) matH <- mat[(mat[,4]==classi[1])&(mat[,5]==classi[2])&(mat[,6]==classi[3]),]
  if (classno==4) matH <-
mat[(mat[,4]==classi[1])&(mat[,5]==classi[2])&(mat[,6]==classi[3])&(mat[,7]==classi[4]),]
  if (classno==5) matH <-
mat[(mat[,4]==classi[1])&(mat[,5]==classi[2])&(mat[,6]==classi[3])&(mat[,7]==classi[4])&(mat[,8]==classi
[5]),]
  if (classno==6) matH <-
mat[(mat[,4]==classi[1])&(mat[,5]==classi[2])&(mat[,6]==classi[3])&(mat[,7]==classi[4])&(mat[,8]==classi
[5])&(mat[,9]==classi[6]),]
  return(matH)
}

```

```

GetMathPS.fcn <- function(classi, mat, col)
{
  classno <- length(classi)

  if (classno==1) matH <- mat[mat[,5]==classi[1],]
  if (classno==2) matH <- mat[(mat[,5]==classi[1])&(mat[,6]==classi[2]),]
  if (classno==3) matH <- mat[(mat[,5]==classi[1])&(mat[,6]==classi[2])&(mat[,7]==classi[3]),]
  if (classno==4) matH <-
mat[(mat[,5]==classi[1])&(mat[,6]==classi[2])&(mat[,7]==classi[3])&(mat[,8]==classi[4]),]
  if (classno==5) matH <-
mat[(mat[,5]==classi[1])&(mat[,6]==classi[2])&(mat[,7]==classi[3])&(mat[,8]==classi[4])&(mat[,9]==classi
[5]),]
  if (classno==6) matH <-
mat[(mat[,5]==classi[1])&(mat[,6]==classi[2])&(mat[,7]==classi[3])&(mat[,8]==classi[4])&(mat[,9]==classi
[5])&(mat[,10]==classi[6]),]
  return(matH)
}

```

APPENDIX C: FUNCTIONS TO COMPUTE WEIGHTING CLASS ADJUSTMENTS FOR FINITE POPULATION SURVEYS AND SIMPLE RANDOM SAMPLING IN R, S-PLUS, AND SAS

R/S-PLUS program

```

# Program: WCAdjSRS.fcn
# Input columns: variable of interest, weight, survey indicator (0 or 1), class variables
# Sample command: WCAdjSRS.fcn(WCAdjSRSCoho)
# Spawn AUC for 1998 and Monitoring Area 1 (North Coast) adjusted by number of landowners

WCAdjSRS.fcn <- function(mat)
{
  n <- dim(mat)[1] # Number of rows (independent
sampling units)
  N <- sum(mat[,2]) # N estimated as sum of all weights
  col <- dim(mat)[2] # Number of columns (variables of
interest)
  classes <- Classes.fcn(mat,col) # Obtain matrix of unique classes
  colclass <- dim(classes)[2] # Number of class variables
  H <- dim(classes)[1] # Number of distinct classes

  #Initialize
  phi <- rep(0,H); Vh <- rep(0,H); lambda1H <- rep(0,H); Nhhat <- rep(0,H)
  nh <- rep(0,H); mh <- rep(0,H); mbarh <- rep(0,H); YbarH <- rep(0,H)
  Yh <- rep(0,H)

  Ybar <- sum(mat[mat[,3]==1,1])/sum(mat[,3]) # Mean response for surveyed
sites

  for (i in 1:H) {
    matH <- GetMatH.fcn(unlist(classes[i,]), mat, col)
    nh[i] <- dim(matH)[1] # nh = all
rows in class h
    mh[i] <- sum(matH[,3]) # mh =
surveyed rows in class h
    phi[i] <- mh[i]/nh[i] # observed
response rate
    Nhhat[i] <- N*nh[i]/n # Nhhat =
estimated
    lambda1H[i] <- mh[i]/nh[i] # response rate
    mbarh[i] <- n*Nhhat[i]*lambda1H[i]/N # Expectation of mh
    Vh[i] <- ifelse(mh[i]<2,0,var(matH[matH[,3]==1,1])) # Sample variance
    # Weighted mean of response
    YbarH[i] <- ifelse(mh[i]==0,0,sum(matH[,1]*matH[,3])/mh[i])
    Yh[i] <- sum(matH[matH[,3]==1,1]) # Sum of responses
  }
  # Remove classes with no responses
  nh <- nh[mh!=0]
  phi <- phi[mh!=0]
  Nhhat <- Nhhat[mh!=0]
  lambda1H <- lambda1H[mh!=0]
  mbarh <- mbarh[mh!=0]
  Vh <- Vh[mh!=0]
  YbarH <- YbarH[mh!=0]
  Yh <- Yh[mh!=0]
  mh <- mh[mh!=0]

  Y <- sum(Nhhat*Yh/mh) # Estimate of
total
  # PS variance
  PSVarC <- sum((Nhhat^2)*Vh*((1/mbarh)+(1/(mbarh^2))-(1/(mbarh*n))-(1/Nhhat)))

```

```

# Additional variance component for estimating Nh
VarY <- (N/n)*sum(Nhhat*((YbarH-Ybar)^2))+PSVarC
answer <- c(Y, VarY, sqrt(VarY), Y-1.96*sqrt(VarY), Y+1.96*sqrt(VarY))
names(answer) <- c("Estimate of Total", "Var", "SE", "Lower 95% CI bound", "Upper 95% CI
bound")
return(round(answer,0))
}

```

SAS program

Instructions: Change the library directory in the second line of code to reflect the directory where the input data are stored. For analyzing data sets other than the test data set, the levels of the adjustment variables and the weights will need to be changed to reflect those of the new data set. See the output from the PROC FREQ call to obtain the data for adjusting the weights.

```

title 'MAR Adjustments - Weighting Class Adjustment for SRS';
libname mar 'C:\SCHOOL\STAR\MARMANUAL';
PROC IMPORT OUT= mar.wcadjsrscoho
DATAFILE= "C:\SCHOOL\STAR\MARMANUAL\WCAdjSRSCoho.xls"
DBMS=EXCEL2000 REPLACE;
GETNAMES=YES;
RUN;
proc freq data=mar.wcadjsrscoho;
    tables RESPONDED*ADJVAR / norow nocol;
run;
data mar.wcadjsrscohowts;
set mar.wcadjsrscoho;
if RESPONDED=1;
if AdjVar=1 then w=Wt*98/90;
if AdjVar=2 then w=Wt*25/21;
if AdjVar=3 then w=Wt*10/7;
run;
proc surveymeans data=mar.wcadjsrscohowts sum;
    var AUC;
    weight w;
run;

```

APPENDIX D: FUNCTIONS TO COMPUTE POSTSTRATIFICATION ADJUSTMENTS FOR FINITE POPULATION SURVEYS AND SIMPLE RANDOM SAMPLING IN R, S-PLUS, AND SAS

R/S-PLUS program

```

# Program: PSAdjSRS.fcn

# Input columns: variable of interest, weight, sample indicator (0 or 1), survey
# indicator (0 or 1), class variables
# Sample command: PSAdjSRS.fcn(PSAdjSRSElk)
# Elk harvests as if drawn by simple random sample and adjusted by age class

PSAdjSRS.fcn <- function(mat)
{
  N <- dim(mat)[1] # Number of rows (independent
sampling units)
  n <- sum(mat[,3]) # N = sum of all weights
  col <- dim(mat)[2] # Number of columns (variables of
interest)
  classes <- ClassesPS.fcn(mat,col) # Obtain matrix of unique classes
  colclass <- dim(classes)[2] # Number of class variables
  H <- dim(classes)[1] # Number of distinct classes
  #Initialize
  phi <- rep(0,H); Vh <- rep(0,H); Nhhatnh <- rep(0,H); nh <- rep(0,H)
  Nh <- rep(0,H); mh <- rep(0,H); nbarh <- rep(0,H); mbarh <- rep(0,H)
  YbarH <- rep(0,H); Yh <- rep(0,H)

  Ybar <- sum(mat[mat[,4]==1,1])/sum(mat[,4]) # Mean response for surveyed
units
  for (i in 1:H) {
    matH <- GetMatHPS.fcn(unlist(classes[i,]), mat, col)
    nh[i] <- sum(matH[,3]) # nh = all rows in
class h
    mh[i] <- sum(matH[,4]) # mh = surveyed
rows in class h
    Nh[i] <- dim(matH)[1] # Population size
in class h
    phi[i] <- mh[i]/nh[i] # Observed sampling
rate of class h
    nbarh[i] <- n*Nh[i]/N # exp value of n in
class h
    mbarh[i] <- nbarh[i]*phi[i] # exp value of m in class
h
    Vh[i] <- ifelse(mh[i]<2,0,var(matH[(matH[,3]==1)&(matH[,4]==1),1])) # sample
var
    YbarH[i] <- ifelse(mh[i]==0,0,sum(matH[,1]*matH[,3])/mh[i]) # mean of responses
    Yh[i] <- sum(matH[matH[,4]==1,1]) # sum of responses
  }
  # Remove classes with no responses
  nh <- nh[mh!=0]
  Nh <- Nh[mh!=0]
  phi <- phi[mh!=0]
  mbarh <- mbarh[mh!=0]
  Vh <- Vh[mh!=0]
  YbarH <- YbarH[mh!=0]
  Yh <- Yh[mh!=0]
  mh <- mh[mh!=0]
}

```

```

        Y <- sum(Nh*Yh/mh) # Estimate of
total
        VarY <- sum((Nh^2)*Vh*((1/mbarh)+(1/(mbarh^2))-(1/(mbarh*n))-(1/Nh)))
        # PS variance
        answer <- c(Y, VarY, sqrt(VarY), Y-1.96*sqrt(VarY), Y+1.96*sqrt(VarY))
        names(answer) <- c("Estimate of Total", "Var", "SE", "Lower 95% CI bound", "Upper 95%
CI bound")
        return(round(answer,0))
}

```

SAS program

Instructions: Change the library directory in the second line of code to reflect the directory where the input data are stored. For analyzing data sets other than the test data set, the levels of the adjustment variables and the weights will need to be changed to reflect those of the new data set. See the output from the PROC FREQ call to obtain the data for adjusting the weights.

```

title 'MAR Adjustments - Poststratification Adjustment for SRS';
libname mar 'C:\SCHOOL\STAR\MARMANUAL';
PROC IMPORT OUT= mar.example4b
DATAFILE= "C:\SCHOOL\STAR\MARMANUAL\PSAdjSRSE1k.xls"
DBMS=EXCEL2000 REPLACE;
GETNAMES=YES;
RUN;
proc surveyfreq data=mar.example4b;
    tables SURVEYED*RESPONDED*ADJVAR/chisq;
    weight WT;
run;
proc freq data=mar.example4b;
    tables RESPONDED*ADJVAR / norow nocol;
run;
data mar.example4bwts;
set mar.example4b;
if SURVEYED=1;
if AdjVar=1 then w=Responded*1751/375;
if AdjVar=2 then w=Responded*8477/1684;
if AdjVar=3 then w=Responded*15756/4419;
if AdjVar=4 then w=Responded*9825/3667;
if AdjVar=5 then w=Responded*2402/1113;
run;
proc surveymeans data=mar.example4bwts sum;
    var Response;
    weight w;
run;

```

APPENDIX E: FUNCTIONS TO COMPUTE WEIGHTING CLASS ADJUSTMENTS FOR FINITE POPULATION SURVEYS AND STRATIFIED RANDOM SAMPLING IN R,S-PLUS, AND SAS

R/S-PLUS program

```
# Program: WCAdjStRS.fcn

# Input columns: stratum, variable of interest, weight, survey indicator (0 or 1), class
# variables
# Sample command: WCAdjStRS.fcn(spawnAUC98[,c(4,1:3,5)])
# Spawn AUC for 1998 and by Monitoring Area strata adjusted by number of landowners

WCAdjStRS.fcn <- function(mat)
{
  strata <- sort(unique(mat[,1]))
  s <- length(strata)
  ests <- matrix(rep(0,6*s),s,6)
  col <- dim(mat)[2]

  for (i in 1:s) {
    str <- strata[i]
    matS <- mat[mat[,1]==str,2:col]
    ests[i,] <- round(c(str,WCAjSRS.fcn(matS)),0)
  }

  totalest <- sum(ests[,2])
  totalvar <- sum(ests[,3])
  totalSE <- sqrt(sum(ests[,3]))
  CIlo <- totalest-(1.96*totalSE)
  CIhi <- totalest+(1.96*totalSE)
  ests <- rbind(ests, round(c(0, totalest, totalvar, totalSE, CIlo, CIhi),0))
  dimnames(ests)[[2]] <- c("Stratum", "Estimate of Total", "Var", "SE", "Lower 95% CI bound",
"Upper 95% CI bound")
  return(ests)
}
```

SAS program

Instructions: Change the library directory in the second line of code to reflect the directory where the input data are stored. For analyzing data sets other than the test data set, the levels of the adjustment variables and the weights will need to be changed to reflect those of the new data set. See the output from the PROC FREQ call to obtain the data for adjusting the weights. For estimates by stratum, remove the comment punctuation from the “By Stratum” call in Proc SurveyMeans.

```
title 'MAR Adjustments - Weighting Class Adjustment for StRS';
libname mar 'C:\SCHOOL\STAR\MARMANUAL';
PROC IMPORT OUT= mar.wcadjstrs
DATAFILE= "C:\SCHOOL\STAR\MARMANUAL\WCAdjStrSCoho.xls"
DBMS=EXCEL2000 REPLACE;
GETNAMES=YES;
RUN;
proc surveyfreq data=mar.wcadjstrs;
    tables RESPONDED*ADJVAR/chisq;
    STRATA Stratum;
    weight WT;
run;
proc freq data=mar.wcadjstrs;
    tables Stratum*RESPONDED*ADJVAR / norow nocol;
run;
data mar.wcadjstrswts;
set mar.wcadjstrs;
if RESPONDED=1;
if Stratum=1 & AdjVar=1 then w=RESPONDED*Wt*98/90;
if Stratum=1 & AdjVar=2 then w=RESPONDED*Wt*25/21;
if Stratum=1 & AdjVar=3 then w=RESPONDED*Wt*10/7;

if Stratum=2 & AdjVar=1 then w=RESPONDED*Wt*91/68;
if Stratum=2 & AdjVar=2 then w=RESPONDED*Wt*24/23;
if Stratum=2 & AdjVar=3 then w=RESPONDED*Wt*12/12;

if Stratum=3 & AdjVar=1 then w=RESPONDED*Wt*83/63;
if Stratum=3 & AdjVar=2 then w=RESPONDED*Wt*37/31;
if Stratum=3 & AdjVar=3 then w=RESPONDED*Wt*16/13;

if Stratum=4 & AdjVar=1 then w=RESPONDED*Wt*72/48;
if Stratum=4 & AdjVar=2 then w=RESPONDED*Wt*32/23;
if Stratum=4 & AdjVar=3 then w=RESPONDED*Wt*20/16;

if Stratum=5 & AdjVar=1 then w=RESPONDED*Wt*30/17;
if Stratum=5 & AdjVar=2 then w=RESPONDED*Wt*26/23;
if Stratum=5 & AdjVar=3 then w=RESPONDED*Wt*24/23;

run;
proc surveymeans data=mar.wcadjstrswts sum;
    var Response;
    STRATA Stratum;
    %*By Stratum;*%
    weight w;
run;
```

APPENDIX F: FUNCTIONS TO COMPUTE POSTSTRATIFICATION ADJUSTMENTS FOR FINITE POPULATION SURVEYS AND STRATIFIED RANDOM SAMPLING IN R,S-PLUS, AND SAS

R/S-PLUS program

```
# Program: PSAdjStRS.fcn

# Input columns: stratum, variable of interest, weight, survey indicator (0 or 1), class
# variables
# Sample command: PSAdjStRS.fcn(PSAdjStRSTest)
# Elk harvests by landowner type strata and adjusted by age class

PSAdjStRS.fcn <- function(mat)
{
  strata <- sort(unique(mat[,1]))
  s <- length(strata)
  ests <- matrix(rep(0,6*s),s,6)
  col <- dim(mat)[2]

  for (i in 1:s) {
    str <- strata[i]
    matS <- mat[mat[,1]==str,2:col]
    ests[i,] <- round(c(str,PSAdjSRS.fcn(matS)),0)
  }

  totalest <- sum(ests[,2])
  totalvar <- sum(ests[,3])
  totalSE <- sqrt(sum(ests[,3]))
  CIlo <- totalest-(1.96*totalSE)
  CIhi <- totalest+(1.96*totalSE)
  ests <- rbind(ests, round(c(0, totalest, totalvar, totalSE, CIlo, CIhi),0))
  dimnames(ests)[[2]] <- c("Stratum", "Estimate of Total", "Var", "SE", "Lower 95% CI bound",
"Upper 95% CI bound")
  return(ests)
}
```

SAS program

Instructions: Change the library directory in the second line of code to reflect the directory where the input data are stored. For analyzing data sets other than the test data set, the levels of the adjustment variables and the weights will need to be changed to reflect those of the new data set. See the output from the PROC FREQ call to obtain the data for adjusting the weights. For estimates by stratum, remove the comment punctuation from the “By Stratum” call in Proc SurveyMeans.

```
title 'MAR Adjustments - Poststratification Adjustment for StRS';
libname mar 'C:\SCHOOL\STAR\MARMANUAL';
PROC IMPORT OUT= mar.psadjstrs
DATAFILE= "C:\SCHOOL\STAR\MARMANUAL\PSAdjStRSElk.xls"
DBMS=EXCEL2000 REPLACE;
GETNAMES=YES;
RUN;
proc surveyfreq data=mar.psadjstrs;
    tables SURVEYED*RESPONDED*ADJVAR/chisq;
    STRATA Strata;
    weight WT;
run;
proc freq data=mar.psadjstrs;
    tables Strata*SURVEYED*RESPONDED*ADJVAR / norow nocol;
run;
data mar.psadjstrswts;
set mar.psadjstrs;
if SURVEYED=1;
if Strata=1 & AdjVar=1 then w=RESPONDED*1206/277;
if Strata=1 & AdjVar=2 then w=RESPONDED*4536/925;
if Strata=1 & AdjVar=3 then w=RESPONDED*8154/2228;
if Strata=1 & AdjVar=4 then w=RESPONDED*5953/2167;
if Strata=1 & AdjVar=5 then w=RESPONDED*1743/797;
if Strata=2 & AdjVar=1 then w=RESPONDED*221/30;
if Strata=2 & AdjVar=2 then w=RESPONDED*2319/438;
if Strata=2 & AdjVar=3 then w=RESPONDED*4395/1211;
if Strata=2 & AdjVar=4 then w=RESPONDED*1772/653;
if Strata=2 & AdjVar=5 then w=RESPONDED*164/71;
if Strata=3 & AdjVar=1 then w=RESPONDED*320/68;
if Strata=3 & AdjVar=2 then w=RESPONDED*1606/315;
if Strata=3 & AdjVar=3 then w=RESPONDED*3129/941;
if Strata=3 & AdjVar=4 then w=RESPONDED*1999/788;
if Strata=3 & AdjVar=5 then w=RESPONDED*421/207;
if Strata=4 & AdjVar=1 then w=0;
if Strata=4 & AdjVar=2 then w=RESPONDED*16/6;
if Strata=4 & AdjVar=3 then w=RESPONDED*78/39;
if Strata=4 & AdjVar=4 then w=RESPONDED*101/59;
if Strata=4 & AdjVar=5 then w=RESPONDED*74/38;

run;
proc surveymeans data=mar.psadjstrswts sum;
    var Response;
STRATA Strata;
    weight w;
run;
```

APPENDIX G: S-PLUS/R FUNCTION TO COMPUTE WEIGHTING CLASS ADJUSTMENTS FOR FINITE POPULATION SURVEYS AND TWO-STAGE CLUSTER SAMPLING WITH SIMPLE RANDOM SAMPLES AT BOTH LEVELS

```

# Program: WCAAdj2Stage.fcn
# Input columns: PSU, var of interest, weightPSU, weightSSU, survey indicator (0 or
# 1), class variables
# Sample command: WCAAdj2Stage.fcn(WCAAdj2StageElk)
# Estimated elk harvest for small private-land rifle hunts, adjusted by age class

WCAAdj2Stage.fcn <- function(mat)
{
  PSUs <- unique(mat[,1])
  PSUwt <- unique(mat[,c(1,3)])
  n <- length(PSUs) # n = number of
PSUs in sample # N = sum of
  N <- sum(PSUwt[,2])
weights of PSUs # Number of columns
  cols <- dim(mat)[2]
in input matrix
  estvarmat <- matrix(rep(0, 2*n),n,2)
  for (i in 1:n) { # PSU loop
    PSUmat <- mat[mat[,1]==PSUs[i],]
    estvarmat[i,] <- WCAAdjSRS.fcn(PSUmat[,c(2,4,5:cols)])[1:2]
  }
  totalest <- N*sum(estvarmat[,1])/n # Estimate of total
  s2t <- var(estvarmat[,1]) # Variance of PSU total
estimates
  varest <- ((N^2)*(1-(n/N))*s2t/n)+(N*sum(estvarmat[,2])/n) # Two-stage variance
  ests <- c(totalest, varest)
  answer <- round(c(ests, sqrt(ests[2]), ests[1]-1.96*sqrt(ests[2]),
ests[1]+1.96*sqrt(ests[2])),0)
  names(answer) <- c("Estimate of Total", "Var", "SD", "Lower 95% CI bound", "Upper 95%
CI bound")
  return(answer)
}

```

APPENDIX H: S-PLUS/R FUNCTION TO COMPUTE POSTSTRATIFICATION ADJUSTMENT FOR FINITE POPULATION SURVEYS AND TWO-STAGE CLUSTER SAMPLING WITH SIMPLE RANDOM SAMPLES AT BOTH LEVELS

```

# Program: PSAdj2Stage.fcn

# Input columns: PSU, var of interest, weightPSU, weightSSU, sample indicator (0 or 1),
# survey indicator (0 or 1), class variables
# Sample command: PSAdj2Stage.fcn(PSAdj2StageElk)
# Estimated elk harvest for small private-land rifle hunts, adjusted by age class

PSAdj2Stage.fcn <- function(mat)
{
  PSUs <- unique(mat[mat[,5]==1,1])
  n <- length(PSUs) # n = number of
PSUs in sample
  N <- length(unique(mat[,1])) # N = number of PSUs in
population
  cols <- dim(mat)[2] # Number of columns
in input matrix
  estvarmat <- matrix(rep(0, 2*n),n,2)
  for (i in 1:n) { # PSU loop
    PSUmat <- mat[mat[,1]==PSUs[i],]
    estvarmat[i,] <- PSAdjSRS.fcn(PSUmat[,c(2,4,5,6:cols)])[1:2]
  }
  totalest <- N*sum(estvarmat[,1])/n # Estimate of total
  s2t <- var(estvarmat[,1]) # Variance of PSU total
estimates
  varest <- ((N^2)*(1-(n/N))*s2t/n)+(N*sum(estvarmat[,2])/n) # Two-stage variance
  ests <- c(totalest, varest)
  answer <- round(c(ests, sqrt(ests[2]), ests[1]-1.96*sqrt(ests[2]),
ests[1]+1.96*sqrt(ests[2])),0)
  names(answer) <- c("Estimate of Total", "Var", "SD", "Lower 95% CI bound", "Upper 95%
CI bound")
  return(answer)
}

```