American Journal of Epidemiology

Vol. 144, No. 1
Printed in U.S.A.

Survey Inference for Subpopulations

Barry I. Graubard[1] and Edward L. Korn[2]

One frequently analyzes a subset of the data collected in a survey when interest focuses on individuals in a certain subpopulation of the sampled population. Although it may seem natural to eliminate from the data set all data from individuals outside the subpopulation before analysis, this procedure may yield incorrect standard errors and confidence intervals. The authors give two examples of this using data from the 1987 National Health Interview Survey and the 1986 National Mortality Followback Survey. The correct method of analysis is described, as well as a simple condition that, when satisfied, ensures that the elimination approach yields identical answers to the correct method. *Am J Epidemiol* 1996;144:102–6.

epidemiologic methods; statistics

Interest frequently focuses on inference for a subpopulation of interest (domain) of a sampled population. For example, a researcher may be interested in addressing hypotheses involving children using a survey that sampled all ages. In the nonsurvey setting, one would simply eliminate the data from individuals outside the subset of interest from the data set before beginning the statistical analysis. For survey data, however, the situation is more complicated. The simple elimination approach is appropriate for estimation of parameters (e.g., means or relative risks) but may not work for calculation of standard errors and confidence intervals. This fact, well known to survey statisticians (1–4), may surprise other data analysts. The effect of inappropriately subsetting the data before analysis can be substantial.

The purpose of this paper is to 1) demonstrate that subsetting the data before analysis of survey data can lead to incorrect confidence intervals, 2) give a simple condition for when it is sufficient to subset the data before analysis, and 3) describe appropriate analyses regardless of whether the simple condition holds. The next section displays selected variance formulas for survey data from which one can see the effect of subsetting the data. This is followed by two applications involving the 1987 National Health Interview Survey and the 1986 National Mortality Followback Survey in which subsetting the data yields incorrect inferences. We conclude with a discussion of some remaining issues.

ESTIMATION AND VARIANCE ESTIMATION WITH COMPLEX SURVEY DATA

Two aspects of data acquired from a survey must be considered when complex survey data are analyzed. The first is that surveys frequently sample individuals in the population with unequal probabilities of selection. In this situation, the sample weights effectively represent the number of individuals in the population that each sampled individual represents. The sample weight associated with an individual is the inverse of that individual's probability of being included in the sample, adjusted, if necessary, for nonresponse. There is often an additional poststratification to ensure that the sum of the sample weights equals known population values for various subgroups, e.g., age/race/sex subgroups. Weighted estimators, which are weighted by the sample weights, are approximately unbiased for their corresponding population quantity (5), whereas unweighted estimators that ignore the sampling design can be badly biased (6). For multistage designs, the clusters sampled at the first stage of sampling are typically known as primary sampling units (PSUs). A general expression for the weighted mean of a characteristic Y for a sampling design involving a stratified selection of PSUs from L strata is

Received for publication May 15, 1995, and accepted for publication October 13, 1995.

[1] Biometry Branch, National Cancer Institute, Bethesda, MD.
[2] Biometric Research Branch, National Cancer Institute, Bethesda, MD.
Reprint requests to Dr. Barry I. Graubard, National Cancer Institute, Biometry Branch EPN-344, Bethesda, MD 20892.

This paper was prepared under the auspices of the US Government and is therefore not subject to copyright.

Downloaded from http://aje.oxfordjournals.org/ by guest on December 3, 2013

$$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}y_{hij}}{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}},$$

where there are k_h PSUs sampled from the hth stratum, a total of n_{hi} individuals sampled from the ith-sampled PSU of the hth stratum, and y_{hij} is the observed value of Y for the jth-sampled individual from the ith-sampled PSU of the hth stratum. The sample weights, w_{hij}, reflect the probabilities of inclusion of the PSUs into the sample, which may differ based on PSU-level characteristics, as well as different sampling rates associated with additional stages of stratification within the PSUs.

The second aspect of survey data that must be considered is that complex designs can induce a correlation structure among the observations. Treating the observations as if they were from a simple random sample can therefore lead to incorrect confidence intervals. Fortunately, there are variance estimation techniques to correctly account for the sample design (7). If the sampling at the first stage is done with replacement, or if the sampling fractions are small enough at the first stage so that this is a reasonable approximation, then the variance of $\bar{y}$ can be estimated in terms of the variability of PSU-level aggregated data. In particular, Taylor series linearization yields

$$\hat{var}(\bar{y}) = \frac{1}{(\sum_{h=1}^{L}\sum_{i=1}^{k_h} W_{hi})^2}\sum_{h=1}^{L}\frac{k_h}{k_h-1}\sum_{i=1}^{k_h}\left[W_{hi}(\bar{y}_{hi}-\bar{y}) - \frac{1}{k_h}\sum_{s=1}^{k_h} W_{hs}(\bar{y}_{hs}-\bar{y})\right]^2, \quad (1)$$

where $W_{hi} = \sum_{j=1}^{n_{hi}} w_{hij}$, and $\bar{y}_{hi}$ is the weighted mean of the sampled observations in the ith-sampled PSU of the hth stratum (8).

With this background, we are now in a position to examine inference for the mean of Y for only individuals in a subset of interest, D. The correct analysis uses formulas that assign sample weights of zero to those sampled individuals outside D, whereas sampled individuals within D retain their original sample weights. These assignments are intuitively reasonable if one thinks of a sample weight as the number of individuals in the target population represented by a sampled individual. Weighted estimates can be calculated using all of the sampled data set with these modified sample weights. These weighted estimates are equal to those that would be obtained by subsetting the data set before estimation to include only those individuals in the domain. For example, for the mean we have

$$\bar{y}_D = \frac{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}y_{hij}I[(hij)\epsilon D]}{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j=1}^{n_{hi}} w_{hij}I[(hij)\epsilon D]},$$

where $I[(hij)\epsilon D]$ is 1 (0) if the jth-sampled individual from the ith-sampled PSU of the hth stratum is in (out of) the subset D.

For variance estimation, when the individuals outside D are assigned sample weights zero, expression 1 becomes

$$\hat{var}(\bar{y}_D) = \frac{1}{(\sum_{h=1}^{L}\sum_{i=1}^{k_h} W_{hiD})^2}\sum_{h=1}^{L}\frac{k_h}{k_h-1}\sum_{i=1}^{k_h}\left[W_{hiD}(\bar{y}_{hiD}-\bar{y}_D) - \frac{1}{k_h}\sum_{s=1}^{k_h} W_{hsD}(\bar{y}_{hsD}-\bar{y}_D)\right]^2, \quad (2)$$

where $W_{hiD} = \sum_{j=1}^{n_{hi}} w_{hij}I[(hij)\epsilon D]$, and $\bar{y}_{hiD}$ is the weighted mean of the sampled observations in D and in the ith-sampled PSU of the hth stratum. This can be compared with the incorrect formula that is obtained by applying expression 1 to the data set for which individuals outside D have been removed:

$$\tilde{var}(\bar{y}_D) = \frac{1}{(\sum_{h=1}^{L}\sum_{i=1}^{k_h} W_{hiD})^2}\sum_{h=1}^{L}{}''\frac{k_h'}{k_h'-1}\sum_{i=1}^{k_h'}\left[W_{hiD}(\bar{y}_{hiD} - \bar{y}_D) - \frac{1}{k_h'}\sum_{s=1}^{k_h}{}' W_{hsD}(\bar{y}_{hsD}-\bar{y}_D)\right]^2, \quad (3)$$

where the primed sum (Σ') is over those PSUs where there is at least one sampled observation in D, the double primed sum (Σ'') is over those strata that have at least two sampled PSUs with sampled observations in the domain, and k_h' equals the number of sampled PSUs in stratum h that have at least one observation in D. Comparing expressions 2 and 3, we see that the variance estimators will be equal unless there is at least one stratum in which some of the sampled

Downloaded from http://aje.oxfordjournals.org/ by guest on December 3, 2013

PSUs have no observations in D and other sampled PSUs have some observations in D. The correct variance estimator can always be obtained by using expression 2.

We now give a hypothetical example to show the potential of expression 3 to give misleading results. Suppose that in the population, one-half the PSUs in each stratum have all individuals in D, and the other half have no individuals in D. Additionally, assume that the sample sizes and sample weights are all equal ($n_{hi} \equiv n_0$, $w_{hij} \equiv w_0$) and that the observations are constant within each stratum, e.g., $Y \equiv 10$ for stratum 1, $Y \equiv 8$ for stratum 2, and so forth. Then, $\bar{y}_D$ will vary over repeated sampling of the population because the proportions of sampled PSUs in each stratum with observations in the domain will vary (around one-half). The variance estimator 2 will appropriately reflect this variability, whereas the estimator 3 will be zero. Notice that this difference holds even if the number of sampled clusters is large.

An additional disadvantage of subsetting the data and using expression 1 (i.e., expression 3) is that surveys frequently sample only a small number of PSUs per stratum, e.g., two per stratum. In these surveys, subsetting the data may yield strata that have only one PSU, implying that standard variance estimators like expression 1 cannot be directly applied. For example, for an analysis of the subpopulation of Hispanics sampled in the 1987 National Health Interview Survey, 29 of the 112 strata have exactly one PSU with any sampled Hispanics. (An additional 15 of the 112 strata have no PSUs with any sampled Hispanics, but this is not an issue for the variance estimation inasmuch as such strata do not affect expressions 2 or 3.) A common procedure for managing this problem is to collapse these strata with "neighboring" strata so that there are at least two PSUs with Hispanics in each (new) stratum. However, this can bias the variance estimator (9). Using expression 2 requires no collapsing provided that the original sample design sampled at least two PSUs (with any sampled individuals) in each stratum.

Although only sample means have been discussed to this point, many parameters of interest can be expressed as explicit or implicit functions of means, e.g., regression and logistic regression coefficients. The results above hold for these other parameters.

EXAMPLES

We present two examples here that utilize the data from the Epidemiology Study in the 1987 Cancer Risk Factor Supplement of the 1987 National Health Interview Survey (1987 NHIS) and the 1986 National Mortality Followback Survey (1986 NMFS) to estimate digestive cancer death rates (see (10) and (11) for details concerning these surveys). Although these examples should be viewed as illustrations of subpopulation analysis and not as substantive analyses, related substantive analyses have been conducted (12, 13).

The target population of the 1987 NHIS is the civilian noninstitutionalized population of the United States living at the time of the interview in 1987. The target population of the 1986 NMFS is adults aged 25 years or more who died in the United States in 1986. The design of the 1987 NHIS can be approximated by the stratified selection of 248 PSUs from 112 strata for a total sample size of 22,080. The design of the 1986 NMFS can be approximated by the stratified selection of 16,598 deaths from 18 strata; each observation is its own PSU. These surveys can be used together to estimate the death rates, where the numerator of the rates are estimated from the 1986 NMFS and the denominators are estimated from the 1987 NHIS. To perform this analysis, the surveys are analyzed jointly by pooling the strata so that the design can be approximated by the stratified selection of 16,846 PSUs from 130 strata. Analyzing surveys together always requires some care as the target populations may not be identical, the modes may be different (e.g., mail vs. household interview), and the questionnaires will generally be different.

Our first example examines the annual death rate from digestive cancer for whites and nonwhites for civilian noninstitutionalized adults aged 25 years or more. The two subpopulations of interest are 1) civilian noninstitutionalized whites aged ≥25 years who, if died, died from digestive cancer, and 2) civilian noninstitutionalized nonwhites aged ≥25 years who, if died, died from digestive cancer. The data sets corresponding to these subpopulations are denoted D_{WHITE} and D_{NONWHITE}, respectively, and are described in table 1. To calculate the rates, the DESCRIPT procedure in the computer program SUDAAN (14) was used on the whole data set D_{ALL} with the SUBPOPN statement that allows the specification of a subset of interest, e.g., D_{WHITE}. This analysis effectively assigns sample weights of zero to individuals outside the subset of interest. The rates (± standard error) are estimated to be 69.01 ± 3.24 and 75.71 ± 5.98 per 100,000 for whites and nonwhites, respectively; see table 2. If one instead uses the data sets D_{WHITE} for the analysis of the whites, the estimated rate is the same (69.01) but the standard error is underestimated as ±0.89. A similar underestimation of the standard error occurs if one uses D_{NONWHITE} to estimate the rate for nonwhites.

Notice that because sampled observations in the 1986 NMFS are sampled PSUs, there are many sam-

Downloaded from http://aje.oxfordjournals.org/ by guest on December 3, 2013

TABLE 1. Various possible data sets for analysis of death rates from digestive cancers based on the 1987 National Health Interview Survey (1987 NHIS) and the 1986 National Mortality Followback Survey (1986 NMFS)

Name	Description	Sample size
D_{ALL}	All individuals in both surveys	
	1987 NHIS	22,080
	1986 NMFS	16,598
	Total	38,678
D_0	1987 NHIS: age ≥25 years	19,240
	1986 NMFS: digestive cancer death, civilian, and noninstitutionalized	785
	Total	20,025
D_{WHITE}	D_0 and white	
	1987 NHIS	16,138
	1986 NMFS	561
	Total	16,699
$D_{NONWHITE}$	D_0 and nonwhite	
	1987 NHIS	2,666
	1986 NMFS	203
	Total	2,869

TABLE 2. Death rates from digestive cancer per 100,000 individuals based on data from the 1987 National Health Interview Survey and the 1986 National Mortality Followback Survey

Subpopulation	Rate	Standard error calculated	
		Correctly*	Incorrectly†
Whites	69.01	3.24	0.89
Nonwhites	75.71	5.98	2.81

* The data set D_{ALL} was analyzed; see table 1 and text.
† For whites, the data set D_{WHITE} was analyzed, and for nonwhites, the data set $D_{NONWHITE}$ was analyzed; see table 1 and text.

pled PSUs that do not contain any observations in the domains of interest, e.g., because the individuals died of something other than digestive cancer. Therefore, from the previous discussion, we should not expect the standard errors to be the same when analyzing D_{WHITE} and $D_{NONWHITE}$ as when correctly analyzing D_{ALL}.

The second example is a logistic regression analysis of digestive cancer deaths with the independent variables being sex, race, age (four groupings), smoking (three categories), and drinking (three categories); see table 3 for definitions of the categories. For this analysis, 134 of the 785 digestive cancer deaths and 1,516 observations from the 19,240 relevant alive individuals were missing drinking or smoking information. We imputed their smoking and drinking categories by sampling the information from similar individuals in the survey based on age, sex, race, drinking status (when available), and smoking status (when available); imputation procedures are discussed elsewhere (15). The regression coefficients and their standard errors were calculated correctly using the procedure LOGISTIC in the computer program SUDAAN (14)

TABLE 3. Logistic regression analysis of deaths from digestive cancer based on data from the 1987 National Health Interview Survey and the 1986 National Mortality Followback Survey

Effect	Relative risk	95% confidence interval for relative risk calculated	
		Correctly*	Incorrectly†
Sex			
Male	1.00		
Female	0.94	0.78–1.13	0.78–1.13
Race			
White	1.00		
Nonwhite	1.35	1.11–1.65	1.19–1.53
Age (years)			
25–44	1.00		
45–64	14.76	11.50–18.95	13.74–15.85
65–84	61.69	48.42–78.58	56.19–67.71
85+	143.48	96.86–212.51	116.88–176.12
Smoking‡			
0–19	1.00		
20–11,999	0.99	0.80–1.23	0.80–1.23
12,000+	1.03	0.82–1.30	0.82–1.29
Drinking§			
0–52	1.00		
53–365	1.60	1.27–2.02	1.27–2.02
365+	2.36	1.88–2.98	1.88–2.98

* The data set D_{ALL} was analyzed; see table 1 and text.
† The data set D_0 was analyzed; see table 1 and text.
‡ Lifetime number of packs of cigarettes smoked.
§ Number of drinks consumed per year as an adult.

on the whole data set D_{ALL} with the SUBPOPN statement specifying D_0 as described in table 1. These were converted to relative risks and confidence intervals by exponentiation in the usual manner. If one instead subsetted the data and used the data set D_0, then the incorrect confidence intervals given in table 3 would result. Interestingly, the amount of error in using the incorrect confidence intervals changes from variable to variable.

DISCUSSION

If there is not a sampled observation in the subset of interest in each sampled PSU, then a standard survey analysis using the subsetted data will yield incorrect standard errors. For some analyses, some computer software will allow one to specify analyses to be performed on subsets of interest, e.g., SUDAAN (14). If such software is not available for the problem at hand, and if the sampling at the first stage is with replacement (or if one is ignoring the finite-population correction factor at the first stage), then one can perform standard survey variance estimation using the whole data set with the sample weights of observations outside the domain set to zero. Care must be taken when using computer software, however, for some software may subset out observations with zero sam-

Downloaded from http://aje.oxfordjournals.org/ by guest on December 3, 2013

ple weights rather than retaining them in the variance estimation. In these cases, one can assign very small weights (e.g., 0.001) to observations outside the domain to obtain correct variance estimates.

A remaining question for statistical inference is the number of degrees of freedom that should be associated with variance estimators. This question did not arise in the examples presented above because the numbers of sampled PSUs (and therefore the degrees of freedom) were so numerous that normal cutpoints could be used for the confidence intervals, e.g., 1.96 for two-sided 95 percent confidence intervals. When smaller numbers of PSUs are sampled, the degrees of freedom are usually taken to be (the number of sampled PSUs) − (the number of strata), which can be justified under strong homogeneity assumptions (16, 17). When some sampled PSUs have no sampled observations in the subset of interest, these assumptions are violated. For analyses restricted to a subset D, we recommend that the degrees of freedom be taken to be (the number of sampled PSUs with sampled observations in D) − (the number of strata with sampled observations in D). This should be a conservative approach inasmuch as it assumes that PSUs with no sampled observations in D do not contribute to the variance estimation when they actually do.

ACKNOWLEDGMENTS

The authors thank Dr. Donna Brogan for her helpful comments.

REFERENCES

1. Kish L. Survey sampling. New York, NY: John Wiley, 1965: 132–6.
2. Cochran WG. Sampling techniques. 3rd ed. New York, NY: John Wiley, 1977:142–4.
3. Yates F. Sampling methods for censuses and surveys. 4th ed. London, England: Charles Griffin, 1981:199–203.
4. Särndel C-E, Swensson B, Wretman J. Model assisted survey sampling. New York, NY: Springer-Verlag, 1992:390–4.
5. Kish L, Frankel MR. Inference from complex surveys (with discussion). J R Stat Soc Ser B 1974;36:1–37.
6. Korn EL, Graubard BI. Examples of differing weighted and unweighted estimates from a sample survey. Am Stat 1995; 49:291–5.
7. Wolter KM. Introduction to variance estimation. New York, NY: Springer-Verlag, 1985.
8. Shah BV, Folsom RE, LaVange LM, et al. Statistical methods and mathematical algorithms used in SUDAAN. Research Triangle Park, NC: Research Triangle Institute, 1993.
9. Rust K, Kalton G. Strategies for collapsing strata for variance estimation. J Off Stat 1987;3:69–81.
10. National Center for Health Statistics. Design and estimation for the National Health Interview Survey, 1985–94. Hyattsville, MD: US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, 1989. (Vital and health statistics. Series 2: Stat evaluation and methods research. No. 110) (DHHS publication no. (PHS) 89-1384, 1989).
11. National Center for Health Statistics. Development, methods, and response characteristics of the 1986 National Mortality Followback Survey. Hyattsville, MD: US Department of Health and Human Services, Public Health Service, Centers for Disease Control, National Center for Health Statistics, 1993. (Vital and health statistics. Series 1: Programs and collection procedures. No. 29) (DHHS publication no. (PHS) 93-1305, 1993).
12. Weinkam JJ, Rosenbaum WL, Sterling TD. Computation of relative risk based on simultaneous surveys: an alternative to cohort and case-control studies. Am J Epidemiol 1992;136: 722–9.
13. Sterling TD, Rosenbaum WL, Weinkam JJ. Analysis of the relationship between smokeless tobacco and cancer based on data from the National Mortality Followback Survey. J Clin Epidemiol 1992;45:223–31.
14. Shah BV, Barnwell BG, Hunt PN, et al. SUDAAN user's manual (release 6.0). Research Triangle Park, NC: Research Triangle Institute, 1992.
15. Little RJA, Rubin DB. Statistical analysis with missing data. New York, NY: John Wiley, 1987:60–7.
16. Shah BV, Holt MM, Folsom RE. Inference about regression models from sample survey data. Bull Int Stat Inst 1977;47: 43–57.
17. Korn EL, Graubard BI. Simultaneous testing of regression coefficients with complex survey data: use of Bonferroni t-statistics. Am Stat 1990;44:270–6.

Downloaded from http://aje.oxfordjournals.org/ by guest on December 3, 2013