

Stratified Weighting of Sample Data for Descriptive Purposes¹

David L. Alderton, Ph.D.²

*HIV/AIDS Surveillance Branch, National Center for HIV, STD, and TB Prevention
Centers for Disease Control and Prevention (CDC), Atlanta, GA*

Abstract

The effect of reweighting a sample to a population is that the marginal distribution of the strata (e.g., sex, race/ethnicity, and age) will be the same in the sample as in the source population or another sample. The logic for solving this problem is similar to that for selecting a stratified sample.³ Reweighting a sample to look like another sample or population is straightforward and the technique can be used to descriptively compare groups on outcome measures without resorting to methods of statistical control such as partial correlations in regression models, and the advantage is that many outcome variables can be simultaneously compared. Several weights will be discussed and techniques for accomplishing the weighting will be explained.

Introduction

There are very well established statistical sampling methods for relating a sample back to a population which are primarily used in the survey research literature (e.g., Kalton, 1983). In essence, the population (technically, the sampling frame) is partitioned (stratified) along dimensions which may influence dependent measures. The probability structure mapping individuals selected into the sample back to the population is then estimated (often corrected for biases such as non-response, under- or over-sampling, clustering, and other adjustments); the reciprocal of this selection probability is used as the mapping weight. For example, if a white female between 18 and 25 years of age has a probability of selection of 5.518×10^{-5} , then that person receives an analysis (or mapping) weight of 18,123; that is, she represents 18,123 people in the larger population. In analyses for point estimates, the weight would be used in SAS® WEIGHT statements to adjust for the person's representation in the broader population. Because the weight was derived based on a stratification of the population, the marginal distributions of the sample using the weight statement will be nearly identical to that in the population for all variables included in the stratification: in this example, sex, race, and age group. An important benefit of deriving weights is that analyses can be done on any variable in the dataset, using the weight statement, without having to adjust each dependent measure for the stratification variables. Say there are two binary response questions and you wanted to compare the proportion endorsing each question in two separate years. To do so without weights would require separate logistic

regressions controlling for the stratification variables. In contrast, the proportions from Proc Freq or Means could be directly compared by using a weight variable.

Two datasets will be used to in this example. The first is the National HIV/AIDS Reporting System (HARS) used by all state and local health departments to report (without names) cases of Acquired Immune Deficiency Syndrome (AIDS) to the Centers for Disease Control and Prevention (CDC). A second database is from the Adult and Adolescent Spectrum of Disease (ASD) project, located in 11 major metropolitan areas. When an AIDS case is reported in any of these 11 locations, local ASD project staff then collect an extensive medical history on the patient. The purpose is to collect detailed information on a subset of HARS AIDS cases. Ideally, ASD locations would represent the nation and we would know the probability of a case being selected into ASD so that statistical sampling methods could relate ASD cases to HARS. Unfortunately, this is practically impossible since states had to volunteer to participate in ASD, and while some sites accept all HARS cases, others sample cases, and still others sample cases with specific modes of HIV exposure.⁴ Thus, there is an unknown probability structure mapping an ASD case to the HARS population. Nonetheless, there is a superset-subset relationship between HARS and ASD and HARS can serve as a good standardization population.

As a practical example, we want to look at the prevalence⁵ of three opportunistic illness⁶, *Pneumocystis carinii* pneumonia (PCP), Kaposi sarcoma (KS), and disseminated tuberculosis (TB). Because of changes in ASD and the AIDS epidemic, recent and historical data are not comparable in terms of demographics and other variables that influence prevalence. For example, Kaposi sarcoma (KS) results from a sexually transmitted virus found almost exclusively among men who have sex with men (MSM). Since the percentage of MSM (of all AIDS cases) has dropped over the years, a casual view of the yearly prevalence of KS would falsely reveal a decreasing trend. We could perform separate analyses for each opportunistic illness, controlling for demographic variables that may influence prevalence estimates that may have changed across time, but this becomes unwieldy as the number of dependent measures increases. Moreover, results from logistic regression must be transformed to answer the question "what percentage

of people had disease X in year Y." In contrast, the dataset could be standardized to the demographics in the national HARS data (or the aggregate ASD data). Once the data have been standardized or weighted to a referent dataset, each dependent measure can be compared directly, in Proc Freq or Mean output. This is the technique I'll describe.

Technique

The first step in creating analysis weights is to determine which variables will be used to adjust the dataset; we selected: sex, race/ethnicity, age group, mode of HIV transmission, and year of death (between 1990 and 1995). A variable is then created that uniquely represents the cross-classification of all levels. If the coded values for each level of each variable is short, a simple technique for creating the stratification variable is to concatenate the values from each variable for each observation. The SAS® statement we used to create the strata variable (after defining it as \$7) is:

```
strata=trim(sex)||trim(nrace)||
trim(agegp)||trim(mode)||
trim(put(dthyr, $2.));
```

This variable needs to be created in both datasets, the superset (HARS) and subset (ASD). We also need the counts for the strata from each dataset which can be obtained from Proc Freq. For the ASD dataset the code looks like:

```
proc freq data=ASD;
table strata / out=ASDcnt noprint ;
run;
```

The same procedure is used for the HARS dataset except that year of death, the variable of focus, is ignored. Since corresponding strata between HARS and ASD must be merged, the counts from HARS are obtained in several steps. First, the distinct strata including year of death are found, then a strata variable that ignores year of death is created (stratal). Counts for stratal are found and merged back to the dataset with the unique strata such that each strata (strata) that includes the year of death receives the same weight as the corresponding strata excluding year of death (stratal).

```
proc sql noprint;
create table HARScnt0 as
select distinct strata from HARS ;
quit;
```

```
data HARScnt0;
set HARScnt0;
length stratal $5 ;
stratal=trim(substr(strata, 1, 5));
run;
```

```
proc freq data=HARS;
table stratal / out=HARScnt1 noprint;
run;
```

```
data HARScnt;
merge HARScnt0
HARScnt1;
```

```
by stratal;
run;
```

The two frequency tables are combined, eliminating strata in HARS not present in ASD (e.g., young black females in 1995 infected by transfusion). Assuming ASDcnt and HARScnt are the frequency tables from ASD and HARS respectively, the SAS® code is:

```
data FreqAll;
merge ASDcnt (in=in1 rename =
(count = ASD_grp) drop = percent)
HARScnt (rename = (count =
HARS_grp) drop = percent stratal);
by strata;
if in1;
run;
```

The combined frequency table has three variables, strata, the strata count from ASD (ASD_grp) and the strata count from HARS (HARS_grp).

A decision must now be made among several weights for adjusting the data. In traditional surveys, the weight for a case reflects how many observations it represents in the population. This would be created by dividing HARS_grp by ASD_grp. If this weight is used and merged back into the ASD dataset, the sum of the weights will equal the number of observations in the HARS dataset. If we were trying to generalize the data to the HARS population this would be the appropriate weight. However, our goal is to standardize ASD to HARS and then look at the data within ASD. In this context, a weight that sums to the number of observations in ASD would be better and more meaningful. Survey statisticians have long made "poststratification" adjustments to data to correct demographic differences between a sample and population such as the US Census data. These weights are formed by dividing the proportion of individuals in a stratum in the population by the proportion of individuals in the same stratum from the sample (e.g., Henry, 1990, p. 130). An interesting benefit of using this weight is that after merging the weight into the sample dataset (ASD), the sum of the weights is the same as the number of observations and it has a mean of 1.0. To form this weight from the combined frequency table (FreqAll above), we need to get the sum of the strata counts for the ASD and HARS data first. An efficient way of doing this is to use SAS® SQL to get the counts and place them in macro variables. The SAS® code for getting the sum of the counts of ASD_grp and HARS_grp from FreqAll is:

```
proc sql noprint;
select sum(ASD_grp), sum(HARS_grp)
into :ASD_N, :HARS_N from FreqAll ;
quit;
```

Following this, the actual weight (HARS_wgt) can be easily calculated in a data step:

```
data FreqAll ;
set FreqAll ;
if HARS_GRP=. then HARS_wgt=0;
else HARS_wgt =
(HARS_grp/&HARS_N) /
```

```
run;
(ASD_grp/&ASD_N) ;
run;
```

Next the weights are merged into the dataset:

```
data ASD;
merge ASD FreqAll;
by strata;
run;
```

Now, analyses can be conducted using the WEIGHT statement and the results will be directly comparable. For example, the prevalence of the three opportunistic illnesses (PC, KS, and TB) can be found:

```
proc freq data=ASD;
table pc ks tb ;
weight HARS_wgt ;
run;
```

Table 1 shows a comparison of three analyses. The first (under the column heading "HARS"), shows the overall count and several marginal distribution percentages for key demographic variables among individuals who died between 1990 and 1995 in HARS. The second column shows the unweighted results from ASD patients who died between 1990 and 1995. The last column shows ASD results using HARS_wgt in the calculations; this is the weighted analysis. Notice that the marginal percentages are nearly identical to those from the HARS column with the exception of year of death which was collapsed across. The last three rows of the table show estimates for the prevalence of three diseases, first from the unweighted ASD data then the weighted values.

		HARS	Unweighted ASD	Weighted ASD
N		251,799	10,198	10,198
Sex	Male	86.2	90.1	86.1
Race-Ethnicity	White	49.0	47.6	49.1
	Black	35.2	39.3	35.7
	Hispanic	14.6	11.9	14.4
Age Group	13-24	3.5	4.6	3.4
	25-39	58.7	66.1	58.7
	40+	37.9	29.3	37.9
Mode of HIV Transmission	Male Homosexual (MSM)	54.7	59.5	56.2
	Injecting Drug User (IDU)	26.4	13.5	26.6
	Heterosexual Transmission	8.9	5.4	8.8
	Undetermined	0.9	8.1	0.9
Year of Death	1990	12.3	9.2	16.3
	1991	14.3	18.4	16.7
	1992	16.2	19.3	16.7
	1993	17.7	19.9	16.8
	1994	19.1	18.3	16.8
	1995	20.4	14.8	16.7
Disease Prevalence	Pneumocystis carinii Pneumonia	—	52.5	52.4
	Kaposi Sarcoma	—	19.7	18.0
	Disseminated Tuberculosis	—	4.5	5.1

Table 1. Comparison of HARS values with weighted and unweighted ASD data.

The goal we were striving for was to adjust the data such that estimates of the prevalence for *Pneumocystis carinii*, Kaposi sarcoma, and disseminated tuberculosis could be compared across years and diseases. While the values in the last three rows of Table 1 show that the aggregate prevalence estimates are quite similar, Table 2 shows substantial differences when prevalence estimates are displayed by year. Moreover, important differences in conclusions would be drawn dependent upon whether the weighted or unweighted temporal trends were used. For example, the unweighted data

for disseminated tuberculosis shows a clear decrease across time whereas the weighted data show that there really was no change in the prevalence. Additionally, Kaposi sarcoma shows a much stronger decreasing trend in the unweighted data than in the weighted data, while just the opposite occurs for *Pneumocystis carinii* pneumonia. Using the weighted analyses we can view the data across diseases and years and make direct comparisons, uncontaminated by other variables, in a clear, simple, unambiguous way. While the current example focused only on three opportunistic illnesses

(dependent variables) we often look at 26 of these illnesses simultaneously where this weighting

technique accrues even greater benefits in simplicity and clarity.

Weighted ASD Data		1990	1991	1992	1993	1994	1995
Pneumocystis Carinii Pneumonia		61.5	57.5	54.9	51.8	46.2	40.8
Kaposi Sarcoma		19.7	17.8	18.8	19.7	16.9	15.1
Disseminated Tuberculosis		5.4	5.9	5.6	5.0	5.0	5.0
Unweighted ASD Data		1990	1991	1992	1993	1994	1995
Pneumocystis Carinii Pneumonia		60.2	57.9	55.3	52.2	48.3	42.8
Kaposi Sarcoma		24.3	21.1	20.1	20.2	17.7	16.1
Disseminated Tuberculosis		5.1	4.8	6.1	4.6	4.4	3.8

Table 2. Comparison of Weighted and Unweighted ASD prevalence estimates across year of death.

Conclusion

In conclusion, weighting a dataset for variables that are typically controlled for statistically is a valuable and in some cases preferable way to evaluate across a number of dependent variables for descriptive purposes. The qualifier "descriptive" is important because this weighting technique has altered the degrees of freedom and forced non-random intercorrelations among error terms, thus statistical results from weighted data sets, such as Type I error levels ("p" values), cannot be used directly from SAS® output (even if we knew the probability structure between the sample and population). However, this procedure can greatly enhance our ability to "visualize" and display complex data in a fair and simple format, and can be used in a variety of situations.

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries.

® indicates USA registration.

References

- Boudreaux, D. & Cranford, K. (1995). Simple random sampling and subsetting strategies using SAS® software. *Observations (Third Quarter)*, 4(4), 34-40.
- Henry, G. T. (1990). *Practical Sampling*. Sage University Series on Applied Social Research Methods Series (No. 21). Beverly Hills, CA: Sage.
- Kalton, G. (1983). *Introduction to survey sampling*. Sage University Series on Quantitative Applications in the Social Sciences (No. 07-035). Beverly Hills, CA: Sage.

¹ I wish to acknowledge the past ASD project director (and my current supervisor), Jeffrey L. Jones, M.D., for allowing me to use ASD data. Debra Hanson also provided guidance in accessing the extensive ASD dataset as well as technical

assistance along the way. Finally, I need to thank Ian Whitlock of Westat who has responded to my pleas for help through SAS-L and provided SAS technical assistance that is reflected in the program in this paper (and in nearly every program I've written in recent years).

² Street: Mail Stop E-47, 1600 Clifton Road, Atlanta, GA 30333; Phone: 404/639-2053; Email: DLA5@CDC.gov

³ See Boudreaux and Cranford (1995) for a description and techniques for selecting random samples from populations, including stratified populations. Some of their techniques provide the logical basis for what follows.

⁴ *Mode of HIV Exposure or Transmission* refers to the most likely route by which the case was infected with HIV, the virus that causes AIDS. HIV is a blood borne virus transmitted through sexual contact or the receipt of HIV infected blood or blood products, such as sharing contaminated needles among injection drug users or transfusions with contaminated blood products.

⁵ *Prevalence* refers to how frequent or widespread the illness is, usually measured as the proportion in a group who have the disease or a standardized measure is used such as the rate of illness per 100,000 people.

⁶ An *opportunistic illness* is a disease whose manifestation is thought to be the result of a suppressed immune system.