Weight variables in the General Social Survey (GSS): how should they be used?

(With acknowledgements to Chuck Humphrey, of the University of Alberta Data Library)

Total population vs sample size

There are two reasons why a weight variable is used. First, you may want results based on the population figure instead of the sample size. For example, the individual Public Use Microdata File (PUMF) for the 1991 Census of Canada has a weight variable that, when used, will produce results based on a population estimate of 27 million instead of 809,654 (which is the sample size of the individual PUMF.)

Adjusting for sampling methods

The second reason to use a weight variable is to adjust for sampling methods. If every member of a population has an equal probability of being selected in a sample, each case would carry the same weight in an analysis, namely, the weight of 1 (think of this as multiplying 1 times the value of each variable, which doesn't change anything.)

To correct for any other sampling method, the weight variable adjusts for the differing probabilities that cases have of being selected in a sample. In other words, cases don't all have an equal probability of being selected, i.e., not every case has a weight of 1. Using the weight variable permits making generalizations to the population from which the sample was drawn.

GSS sampling methodology

The Canadian General Social Survey employs a complicated sampling method and thus requires the use of a weight variable. However, the weight variables in the General Social Survey not only adjust for the sampling method but also provide population estimates for Canada. This produces results with a count of 20+ million (i.e., the number of Canadians 18 years of age or older.)

This can cause problems for analysts who want to perform traditional inferential statistical tests. Large N's (anything over a few hundred) will generate significant test results by the very nature of inferential statistics. One way to compensate for the scale of the weight variables used by Statistics Canada is to re-base the weight variable to the sample size.

Procedure to re-basing the weight variable to the sample size

1. Run frequencies on a simple variable (e.g. sex) with the data UNWEIGHTED.
2. Make a note of the total number of cases (including missing cases).
3. Weight the data using the WEIGHT BY command and the name of the weighting variable that comes with the data.

4. Rerun frequencies on the simple variable and make a note of the total number of cases (including the missing cases).
5. Place the following commands at the BEGINNING of your syntax file:
    a. COMPUTE NWGT = wgtvar*(tnw/twt).
    b. WEIGHT BY NWGT.

        (where: wgtvar is the name of the weighting variable that comes with the data; tnw is the number of cases that you wrote down at step 2 above; and twt is the number of cases that you wrote down at step 4.)

Example of how to re-base the weight variable to the sample size

Below is an example using the individual PUMF from the 1991 Census, where the weight variable is re-based to the sample size. This ensures that adjustments for sampling methods are retained, and also the N is maintained at the sample size rather than the population estimate.

The following SPSS code re-bases the weight variable for the Alberta sub-sample of the 1991 individual PUMF.

```
compute wt=weightp*(75506/2516864).

weight by wt.
```

The re-basing is the result of dividing 75,506 (the sample N for Albertans in the PUMF) by 2,516,864 (the population estimate of Alberta using the weight variable provided with the PUMF.) The new weight (wt) now has a sample N of 75,506 (which still is HUGE by inferential statistics standards.)

An SPSS job was run to determine the two numbers used in the re-basing. This must be done before the new weight variable can be created. Below is the SPSS setup to do this. See if you can find the numbers from the output below.

```
>  get file='/afs/ualberta.ca/dept/business/data/census91/census91ind.syst

>     keep=provp sexp marstlp totincp weightp.

>  select if provp eq 48.

>  frequencies variables=sexp.



SEXP      Sex

                                                                Valid     Cum

Value Label                Value  Frequency  Percent  Percent  Percent
```

```
Female                          1     37617     49.8     49.8     49.8
Male                            2     37889     50.2     50.2    100.0
                                     -------  -------  -------
                          Total      75506    100.0    100.0

Valid cases   75506      Missing cases      0

>  weight by weightp.
>  frequencies variables=sexp.

SEXP      Sex

                                                        Valid     Cum
Value Label                Value  Frequency  Percent  Percent  Percent

Female                          1   1253899     49.8     49.8     49.8
Male                            2   1262965     50.2     50.2    100.0
                                    -------  -------  -------
                          Total     2516864    100.0    100.0

Valid cases 2516864      Missing cases      0
```

Another example showing the impact of using weight variables

Below is the frequency distribution for the marital status from the 1991 General Social Survey. This table really is made up of three frequency distributions for the variable DVCURMS2: the unweighted frequencies from the raw data file [unweighted], the frequencies applying the weight variable FWGHT (final weight) [weighted 1], and the frequencies applying the re-based variable WT [weighted 2]. The variable WT was created using the SPSS command: compute wt=fwght*(13495/20525561).

```
DVCURMS2   RESPONDENT'S CURRENT LEGAL MARITAL STATUS
```

		[unweight]		[weighted 1]		[weighted 2]	
Value Label	Value	Freq.	%	Freq.	%	Freq.	%
MARRIED	1	6759	50.1	11277210	54.9	7414	54.9
WIDOWED	2	1500	11.1	1124533	5.5	739	5.5
MARRIED BUT SEPARATE	3	528	3.9	598178	2.9	393	2.9
DIVORCED	4	975	7.2	1281271	6.2	842	6.2
SINGLE	5	3622	26.8	6124539	29.8	4027	29.8
NOT STATED	9	111	.8	119830	.6	79	.6
		7	5	8	5	5	5
	Total	13495	100.0	20525561	100.0	13495	100.0

The total number of cases in the raw data file is 13,495. There are exactly 6,759 cases with the value 1 (i.e., married) for the variable DVCURMS2 in this file, 1,500 for widowed, 528 for married but separated, etc. But as raw frequencies, they do not adjust for the sampling method and cannot be used to generalize to the Canadian population.

The middle figures for WEIGHTED 1 are based on applying the weight variable contained in the GSS 91 file, namely, FWGHT. Notice two things about applying the weight variable. First, the total N becomes 20,525,561 or the total number of Canadians in 1991 who were 18 years of age or older. Secondly, notice that the percentages differ between the unweighted and weighted distributions. This is due to the adjustment for the sampling methodology. When sampling, STC oversampled widows and divorced to ensure capturing people in these categories in the study. The weight variable corrects for this bias. Thus, 54.9% of Canadians in 1991 were married (not 50.1% as reflected in the unweighted frequency distribution.)

Finally, the re-based weight variable returns the overall N to 13,495 (i.e., the size of the sample.) Notice however, the percentages are maintained between STC's weight variables (FWGHT) and the re-based weight variable (WT). In other words, the weighted sample using WT corrects for the sampling method but also allows working with an N equal to the original sample size. The one advantage of working with the smaller N is that some researchers prefer using inferential statistical tests that simply have little meaning with the population N produced with the STC weight variable.

The bottom line: doing statistics without using one of the two weight variables produces biased results that prevents one from making generalizations to the full population. In other words, you need to apply a weight variable when doing statistical analysis with the 1991 GSS. Now, which weight variable? It doesn't really matter whether you use FWGHT or WT. Either corrects for the sampling methodology. The choice is more one of ease in working with statistical tests.

University of Regina, Systems Support, Main Library
http://uregina.ca/datalibrary/weight.html
Updated October 27, 1999
Comments or suggestions? Page Master