Samples and Weights - The Concepts and an Example[1]

In a random sample, each case has an equal chance of being selected. In a non-random sample, the likelihood of being sampled varies depending on the criteria being used in the sample design. Non-random samples may be used to increase the number of members of small groups that are of particular interest in the study, or for some other cost-saving reason. A non-random sample may not represent the general population. However, it is possible to use the statistical technique of weighting to approximate a representative sample.

A Non-Random Sample Design

Suppose that there is a population of 100,000 people, and there is enough money in the grant to collect data from 1,000 people. The population is divided into two regions, (A and B). The percentage of indigenous people in the total population is 20%. In addition to comparing indigenous to non-indigenous people, the researchers want to be able to do some separate analysis within each group. If they conduct a random sample for the entire population, they could expect to get approximately 200 indigenous people, and they would prefer to have more in their sample. The indigenous population is not distributed evenly between the two regions. Region A has 25,000 people, 50% of whom are indigenous. Region B has 75,000 people, and 10% are indigenous. If the researchers draw a higher percentage of the sample from Region A, they can expect to get more indigenous people in their sample.

They decide on the following strategy. Instead of pooling all 100,000 people together and choosing random 1% of all of the people from this pool, they create a pool for each region. They will choose a 2% sample of people (n=500) from Region A. They can only interview a total of 1,000 people, so they must reduce the percentage of people that they take from Region B from 1% to .66667% (n = 500). With this sampling scheme, they can expect to get 250 indigenous people from Region A, and 50 indigenous people from Region B, for a total of 300 indigenous people. (This is the expected value - actual value will probably not be exactly the same.)

This type of sample creates a problems for the researcher who wants to analyze Regions A and B together. The sample will not be representative of the country's population. If an adjustment is not made for the non-random sample, a researcher would conclude that 30% (instead of 20%) of the population is indigenous. The two regions might differ socioeconomically and culturally, so incorrect conclusions about other variables could also be reached. Weighting is a solution to this problem.

Calculating Sampling Weights: Sampling weights are the inverse of the likelihood of being sampled.   The likelihood of a person in Region A being selected is 500/25,000. Each person in Region A represents 50 people (25,000/500 = 50). The sampling weight for people in Region A would be 50. The chance of a person in Region B being selected in 500/75,000. Each person in Region B represents 150 people (75,000/500 = 150). The sampling weight for the people in Region B would be 150. Note that the weight for people in Region A are lower than those in Region B. People in Region A are over-represented in the sample, and people in Region B are under-represented in the sample. The sampling weights inflate the impact of those under-represented, and deflate the impact of those who ore over-represented so that the original population is approximated.

Total Country Population 100,000
Sample n = 1,000.

Region A	Region B
25,000 People 500 of 25,000 Sampled 500/25000 = .02 2% of Population Sampled 25000/500 = 50 Each Sampled Person represents 50 People. Weight = 50	75,000 People 500 of 75,000 Sampled 500/75000 = .006667 .67% of Population Sampled 75000/500 = 150 Each Sampled Person represents 150 People. Weight = 150

[1] Prepared by Patty Glynn, University of Washington. 8/9/05 C:\all\help\helpnew\samples_weights.wpd

An Example: Following is a SAS program that creates a sample from a fictional population of 100,000 that has the characteristics described above, and creates a sample as described above. A variable named “score” is created with different means for Regions A and B. Sample weights are created, and weighted and unweighted means are calculated. Indigenous people are coded 1, and others are coded 0.

```
* SAMPLE.SAS ;
TITLE1 'SAMPLE.SAS' ;
* DEMONSTRATION OF SAMPLING AND WEIGHTS ;
data pop1 ;
do i = 1 to 100000 ;
 id = i ;
 if i le 25000 then region = 'A' ;
 if i gt 25000 then region = 'B' ;
output ;
end ;
data TOTAL ; set pop1 ;
indig = 0 ;
 if id le 12500 then indig  = 1   ;
 if id gt 92500 then indig  = 1   ;

* CREATE A RANDOM VARIABLE ;
rand = ranuni(100) ;
** CREATE "SCORE" SO THAT IT WILL HAVE DIFFERENT VALUES ;
** BETWEEN THE REGIONS - FOR DEMONSTRATION PURPOSES ;
** BE SURE TO USE A DIFFERENT SEED THAN THAT USED FOR THE VARIABLE RAND ;
if region = 'A' then score = (RANUNI(88) * 100 ) ;
if region = 'B' then score = (RANUNI(88) * 500) ;
proc sort; by region rand ;

data SAMPLE ; set TOTAL ; by region ;
** CREATE A SEQUENTIAL VARIABLE RANGING 1 TO N, WITHIN EACH REGION ;
if first.region then count = 0 ;
count + 1 ;
keepcase = 0 ;
** SELECT THE FIRST 500 CASES IN EACH REGION ;
** THEY ARE ALREADY SORTED IN RANDOM ORDER WITHIN REGION, SO THIS WILL ;
** BE A RANDOM SAMPLE ;
if region = 'A' and count le 500 then keepcase = 1 ;
if region = 'B' and count le 500 then keepcase = 1 ;
if keepcase = 1 ;
** CREATE SAMPLING WEIGHTS - INVERSE OF LIKELIHOOD OF BEING SAMPLED ;
if region = 'A' then perwt = (25000/500) ; * =  50 ;
if region = 'B' then perwt = (75000/500) ; * = 150 ;
* Adjust the weights to the mean of weights is 1 ;
* SAS makes mistakes when calculating standard deviation in proc means, otherwise ;
* SEE http://staff.washington.edu/glynn/adjustsa.pdf ;
adjwt = perwt / 100 ;

TITLE2 'TOTAL POPULATION, BEFORE SAMPLE IS DRAWN' ;
PROC means DATA = TOTAL ; var INDIG SCORE ; RUN ;
TITLE2 'UNWEIGHTED MEAN, SAMPLE POPULATION POPULATION' ;
proc means DATA = SAMPLE ; var indig score ; RUN ;
TITLE2 'WEIGHTED, SAMPLE POPULATION' ;
proc means DATA = SAMPLE ; var indig score ; weight adjwt ; RUN ;
```

TOTAL POPULATION: 20% INDIGENOUS, AND THE MEAN FOR THE VARIABLE “SCORE”: IS ABOUT 200.

```
                                 TOTAL POPULATION, BEFORE SAMPLE IS DRAWN

      Variable           N            Mean         Std Dev         Minimum         Maximum
      ------------------------------------------------------------------------------------
      indig         100000       0.2000000       0.4000020               0       1.0000000
      score         100000     200.3140144     152.7302935       0.0041060     499.9934530
```

SAMPLE POPULATION, UNWEIGHTED: % INDIGENOUS AND MEAN DOES NOT REFLECT TOTAL POPULATION

```
                                      SAMPLE POPULATION, UNWEIGHTED

       Variable          N            Mean         Std Dev         Minimum         Maximum
       -----------------------------------------------------------------------------------
       indig          1000       0.3200000       0.4667096               0       1.0000000
       score          1000     150.1200127     144.5380073       0.0902095     499.4960101
```

SAMPLE POPULATION, WEIGHTED: % INDIGENOUS AND MEAN IS CLOSER TO TOTAL POPULATION

```
                                       SAMPLE POPULATION, WEIGHTED

       Variable          N            Mean         Std Dev         Minimum         Maximum
       -----------------------------------------------------------------------------------
       indig          1000       0.2060000       0.4046328               0       1.0000000
       score          1000     200.3041834     152.8837295       0.0902095     499.4960101
```

There has been discussion by statisticians about whether weighting in multivariate analysis is appropriate. You may read “Sampling Weights and Regression Analysis” by Christopher Winship, Larry Radbill, in Sociological Methods and Research, Sage Periodical Press, Vol 23, Number 2, November 1994, pages 230 to 257.

Please read http://staff.washington.edu/glynn/adjustsa.pdf and http://staff.washington.edu/glynn/adjspss.pdf for information about “normalizing weights”.