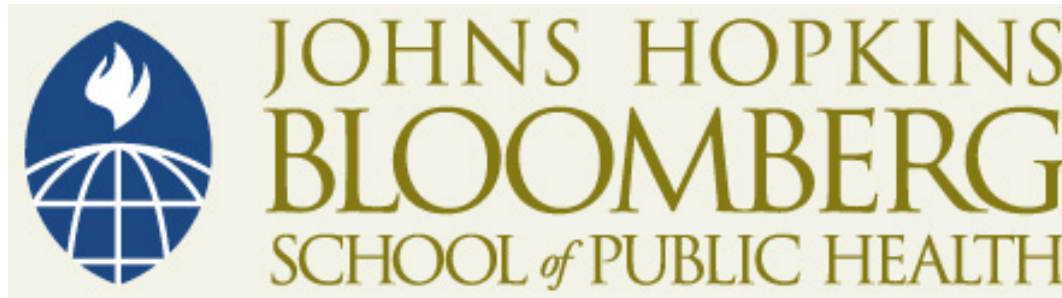


This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



Copyright 2009, The Johns Hopkins University and Saifuddin Ahmed. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Methods in Survey Sampling

Biostat 140.640

weighting

Saifuddin Ahmed, PhD
Dept. of Biostatistics

Sample Weighting

- The purpose of weighting sample data is to improve **representativeness** of the sample in terms of:
 - **size**
 - **Distribution, and**
 - **characteristics** of the study population.
- By introducing sampling weight, the estimation is carried out in such a way that the **estimates** reflect the actual distribution/characteristics of the population.

Population total estimate from sample:

$$T(y) = N * \bar{y},$$

where \bar{y} is the estimated mean

We may rewrite this as:

$$T(y) = \frac{N}{n} * n * \bar{y}$$

By multiplying the “inverse of sample selection probability” with a “sample” estimate, we may derive a “population” level estimate.

An example

- $N = 300,000$
- $n = 300$
- Sampling fraction (f) = $300/300,000 = 0.1$
- $y_{\text{bar}}(\text{mean}) = 0.5$ (50% of children are immunized)
- Total $Y = (1/f) * (n * y_{\text{bar}}) = (1/0.1) * 300 * 0.5$
 $= 150,000$

- The inverse of the selection probability ($w=1/f=N/n$) is used as a *weight* to make the estimate equal to the population
- Even when the sample is selected with different selection probabilities, the same principle could be extended, and the **weight is inversely proportional to the unit selection probabilities.**
- The weights as measured from the inverse of the selection probabilities are called **design weights.**

**Inverse of the *sample selection probability*
is the design weight.**

**The sum of sampling weights is equal to
the population size N**

ta v005

weight	Freq.	Percent	Cum.
199788	326	5.15	5.15
203687	107	1.69	6.84
342747	91	1.44	8.27
352024	312	4.93	13.20
473247	262	4.14	17.33
571152	248	3.91	21.25
726240	267	4.21	25.46
728423	128	2.02	27.48
792283	81	1.28	28.76
836419	187	2.95	31.71
845095	294	4.64	36.35
851062	138	2.18	38.53
907765	384	6.06	44.59
930111	242	3.82	48.41
967833	134	2.12	50.53
979842	391	6.17	56.70
1005824	348	5.49	62.19
1026552	67	1.06	63.25
1068896	145	2.29	65.54
1083215	496	7.83	73.37
1095179	79	1.25	74.62
1106982	120	1.89	76.51
1224149	417	6.58	83.09
1312089	76	1.20	84.29
1465329	884	13.95	98.25
1608275	111	1.75	100.00
Total	6,335	100.00	

When weighting is needed?

- Different selection probabilities
- Need of improving the variance estimation in case of:
 - High non-response and coverage errors
 - Departures from representative(probability) sampling
 - Small sample sizes

Why do we weight?

- To improve representativeness of the sample in terms of **size, distribution** and **characteristics** of the study population.
- to ensure that the estimates are **simple unbiased estimates**.

***Weighting, however, is not
without problems.***

Advantages of such self-weighted sampling designs

- Weighting increases complexity of the survey operations
- Haphazard weights which are not related to population variances increase variance of the survey estimates.
- Weighting may reduce the flexibility and ease with which the same sample may be used for diverse purposes and different surveys.
- Repeated sampling from the same list is straight forward with self-weighting, but the selection probabilities become more complex if previous selections from the frame were not with equal probabilities.
- Self-weighted samples are more readily understood and accepted by the non-statistical and the general public.
- Moderate departures from self-weighting have small effect on variances, and weighting is only suggested when the estimates involves large departures with self weighted samples.

When to weight?

- For appropriate representativeness of smaller domain (e.g., residence, geographical territories, race, sex)
- Fixed sample size for different geographical areas
- Defect in sampling frame, errors in selection, high non-response

Disadvantages

- Increased complexity
- Inconvenience
- Increased variance with haphazard weighting
- Analysis /statistical programming
- Cost
- Higher possibilities of error
- Increased bias with incorrect weight

Major reasons of weighting in practice

- Departures from self-weighting
- Frame defects
- Nonresponse

Region = A $N=50,000$ $n=500$ $P=0.5$ (# immunized= $500 \cdot .5$ $=250$)	Region = B $N=15,000$ $n=500$ $P=0.6$ (# immunized= $500 \cdot .6$ $=300$)
Region = C $N=30,000$ $n=500$ $P=0.7$ (# immunized= $500 \cdot .7$ $=350$)	Region = D $N=5,000$ $n=500$ $P=0.8$ (# immunized= $500 \cdot .8$ $=400$)

Then, sample based immunization rate in the country is:
 $(250+300+350+400)/(500+500+500+500)= 1300/2000 = 65\%$

$$w_h = \frac{N_h}{n_h}$$

$$w_A = \frac{50,000}{500} = 100$$

$$w_B = \frac{15,000}{500} = 30$$

$$w_C = \frac{30,000}{500} = 60$$

$$w_D = \frac{5,000}{500} = 10$$

$$t = \sum_{h=1}^H \sum_{i=1}^{n_h} w_h y_{ih}$$

$$\bar{y} = \frac{t}{\sum w_h}$$

. di (100*.5+30*.6+60*.7+10*.8)/ (100+30+60+10)

.59

. di (100*250+30*300+60*350+10*400)/ (50000+15000+30000+5000)

.59

Weighted based “unbiased estimate” = 59%