**W1 General Steps in Weighting**

**Introduction**

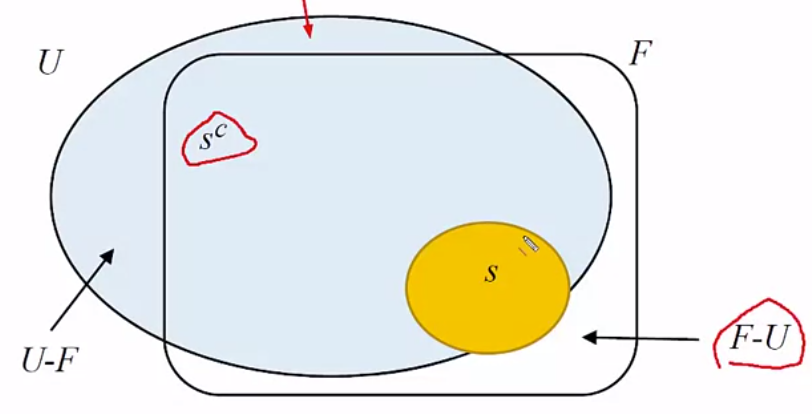
Purpose of weights

• expand a sample to a full population

• correct for “coverage problems” in sample or frame

• Use auxiliary data to create unbiased and more precise estimators

• Weights can be used for both estimating descriptive statistics and estimating model parameters.



U: Universe

F: Sampling Frame

The Frame often misses U-F and also includes F-U which should not have been included

We use sample S and **expand** it so that it includes S^c

🡺 Samples can simultaneously under- and over-cover a target population.

Weights and Estimators

• The scale of weights

- weights can be scaled to estimate population totals, or

- to sum up to the sample size

• Weights scaled up to sum up to sample size are called “normalized” weights

- partly a holdover from days when software for analyzing data was not available

- if df reported as 

• We will deal with weights that are scaled to estimate pop totals

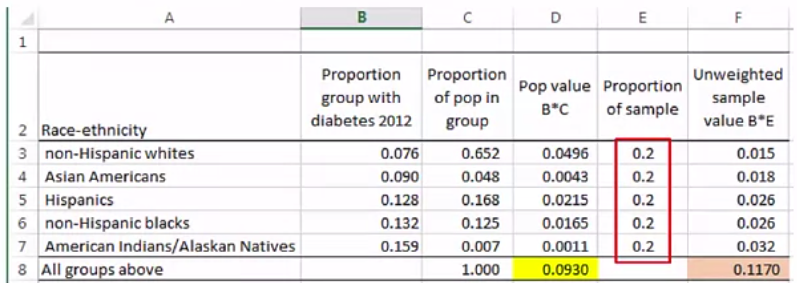
Why use the weights at all

• Unweights estimated can be biased

• An example-estimate the prevalence of diabetes across a set of ethnic groups

• Suppose a sample produces unbiased estimates for each ethnic group but equal size samples are selected from each group

• Race-ethnic groups have much different sizes in the US pop



**Quantities to Estimate**

Totals

• Totals

- no. of persons on a public assistance program

- no. of days without a job

- no. of visits to doctor in last year

• A total can be written as 

Where s is the set of sample units

r is the set of nonsample units

• Estimating the total amounts to predicting the nonsample sum

• An estimated total usually has the form 

Means

• Means

- average income

- average no. of years of schooling

- students’ average score on a standardized test



Proportions, Quantiles

• Proportions (percentages): % of persons who plan to vote for a candidate, unemployment rate

• Quantiles (medians, 1st and 3rd quartiles): median household income, median age at first marriage, 97.5th percentile of blood lead level in children age 1-5.

Algorithm:

- sort file by y (low to high)

- cumulate weights until desired percent of total weight reached (50% of median)

- record value of y for that unit

Ratio and other Combinations

• Ratios

- ratio of women’s average income to men’s average income

- odds ratios

- ratio of the odds of having diabetes for African Americans to the odds for all others

• Regression model parameter estimates

Subgroups

• compute estimate within each group

• Proportion of males, age 18-34, who watched a live sports event on TV

• SEs may need to account for random sample size in a subgroup unless it is controlled by design

**Goals of Estimation**

Population or Census Values

• Population value: the value that would be obtained if a census were one of the target population

• To describe what you are estimating, explain what the census value would be

- forces you think about what the target population is and what you can actually make an estimate for

• Even with a census, it may not be definite what the “pop value” is because of measurement issues

Unambiguous cases (maybe)

• No. of persons living in Washington DC on January 1, 2016

• No. of persons with high diastolic blood pressure (> 90mm Hg)

This seems clear as long as BP can be measured accurately

• No. of full-time employees during the week that includes 12th of September, 2015

Ambiguous Cases

• No. of persons who say they will vote in next presidential election

• No of persons who favor tighter gun control

• No. of persons in labor force

- To be in labor force, a person either must have a job or be “actively” looking for one

- what does “active” mean

• Consumer price index

- “quality changes” are accounted for (e.g. faster processor in a laptop than last year)

- what value do we place on a quality change

**Statistical Interpretation of Estimates**

Interpretation of Estimates

• A weighted estimate needs to have a statistical interpretation in order to be justified

• Interpretation can be in terms of repeated sampling (in case of probability samples) or in terms of models (in case of non-probability samples)

Probability Sampling

• An estimator is **unbiased** if, over all the random samples that could be selected, its values average out to the census value

• An estimator is **consistent** if, as the sample size gets large, the estimator gets closer and closer to the census value

• Even for complicated quantities like medians or quartiles, we want these properties to hold

Types of Probability Samples

• Various types were covered in Course 4: Sampling People & Records

• some examples

- simple random sampling

- stratified simple random sampling

- stratified systematic random sampling

- two-stage stratified sampling

- multi-stage stratified sampling

- single-stage sampling with probabilities proportional to some measure of size

Non-Probability samples

• Unbiasedness and consistency have to be with respect to a model

• We need to be able to estimate the population model from the sample

• If sample has serious holes in coverage, estimators can be biased and inconsistent for the desired target population

- Example: a volunteer web panel that has no African-American women over 70 years old

- If those women have different characteristics (follow a different model), than the volunteers, you cannot estimate for them

Types of non-probability samples

• Not all non-probability samples are equally good at representing a target population

• A convenience sample (e.g. students in an Intro Psych class)

• A quota sample of persons recruited door-to-door until a specified number of persons in a set of age groups are willing to cooperate

• A panel of persons recruited from those who visit a particular website

• A river sample which recruits potential respondents from individuals visiting one of many websites where survey invitations have been placed

• Some probability samples have so much Non response that they begin to look like non-prob samples

Interpretation when there are Coverage Errors

• With under- or over-coverage, we calibrate the weights and estimates with auxiliary data

• Target population control totals needed for each covariate used

• If sample can be projected to the target pop using the covariates, then estimates will have a model-based interpretation