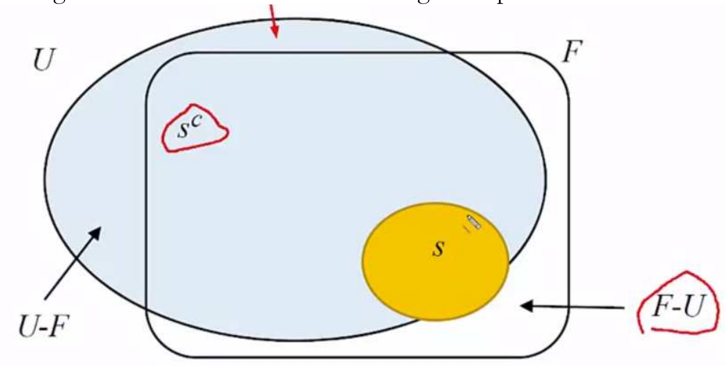
W1 General Steps in Weighting

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 $\sum_{i\in s} w_i - p$, normalized weights lead to $n - p$
• 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

	A	B	C	D	E	F
1						
2	Race-ethnicity	Proportion group with diabetes 2012	Proportion of pop in group	Pop value B*C	Proportion of sample	Unweighted sample value B*E
3	non-Hispanic whites	0.076	0.652	0.0496	0.2	0.015
4	Asian Americans	0.090	0.048	0.0043	0.2	0.018
5	Hispanics	0.128	0.168	0.0215	0.2	0.026
6	non-Hispanic blacks	0.132	0.125	0.0165	0.2	0.026
7	American Indians/Alaskan Natives	0.159	0.007	0.0011	0.2	0.032
8	All groups above		1.000	0.0930		0.1170

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 $t = \underline{\sum_{i\in s} y_i} + \underline{\sum_{i\in r} y_i}$
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 $\hat{t} = \sum_{i\in s} w_i y_i$

Means
• Means
 - average income
 - average no. of years of schooling
 - students' average score on a standardized test
$\hat{\bar{y}} = \sum_{i\in s} w_i y_i / \underline{\sum_{i\in s} w_i}$

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

Coverage Problems

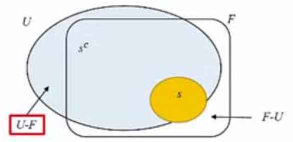
Types of coverage errors
• Either under- or overcoverage can happen (or both)
• Overcoverage example
  - frame of businesses that includes out-of-business units
  - list of organization members that contains persons who have dropped out
In both examples any ineligible units in the sample will not be included in estimates
• Undercoverage
  - volunteer panel that omits elderly women
  - Business frame that does not include recently formed businesses
In these examples, either (a) the target universe must be re-defined, or (b) a statistical adjustment is made to expand the **covered** population to the **target** population

Estimating a total when there are coverage errors

- When we have under- and overcoverage a pop total is

$$\underline{t} = \boxed{\sum_{i \in U \cap F} y_i} + \boxed{\sum_{i \in U - F} y_i}$$

- $U - F$ is the under-covered part of the target pop U
- Note that we ***do not*** want to estimate for the part of frame outside the target pop, i.e., $F - U$ which is the part of the frame that causes over-coverage



- Sample units must be projectable to full target population (sample follows same model as population)

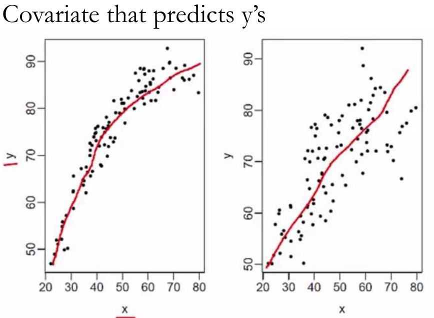
⇒ The sample s must be used to estimate for the nonsample units in the frame ($U \cap F - s$) and the part of the universe that is outside the frame ($U - F$):

$$\hat{t} = \sum_{i \in s} y_i + \sum_{i \in (U \cap F) - s} \hat{y}_i + \sum_{i \in (U - F)} \hat{y}_i$$

- Using auxiliary data may help correct coverage problems
- Auxiliaries (covariates) used to predict for nonsample units
- Accurate population totals must be known

Improving Precision

Covariate that predicts y's



Categorical covariates

• Categorical covariates (Hispanicity and Age) are related to percent of persons receiving Medicaid in this example

	Age group (years)				
Hispanicity	under 18	18–24	25–44	45–64	65+
Hispanic	32.2	10.7	7.6	11.0	27.2
Non-Hispanic White	12.6	6.6	3.8	3.1	3.7
Non-Hispanic Black and other race/ethnicity	31.3	12.7	8.8	6.4	16.5

Effects on Weighting on Standard Errors

Design features affecting Standard Errors(SE)s

• Sample design features
  - Stratification
  - clustering
  - varying probabilities of selection (in probability samples)

• Stratification
  - An efficient allocation can reduce SEs of full pop estimates
  - can be used to control sample size and precision of stratum estimates

• Clustering ➔ usually increases SEs

• Effects can be different for
  - full pop estimates and subgroup estimates
  - different y variables
  - different statistics: totals, means, model parameter estimates

• Weighting adjustments can increase or decrease SEs

• Non response (NR) adjustments often increase SEs

• Calibration to pop controls can decrease SEs
  - covariates used in calibration need to be predictors of y's