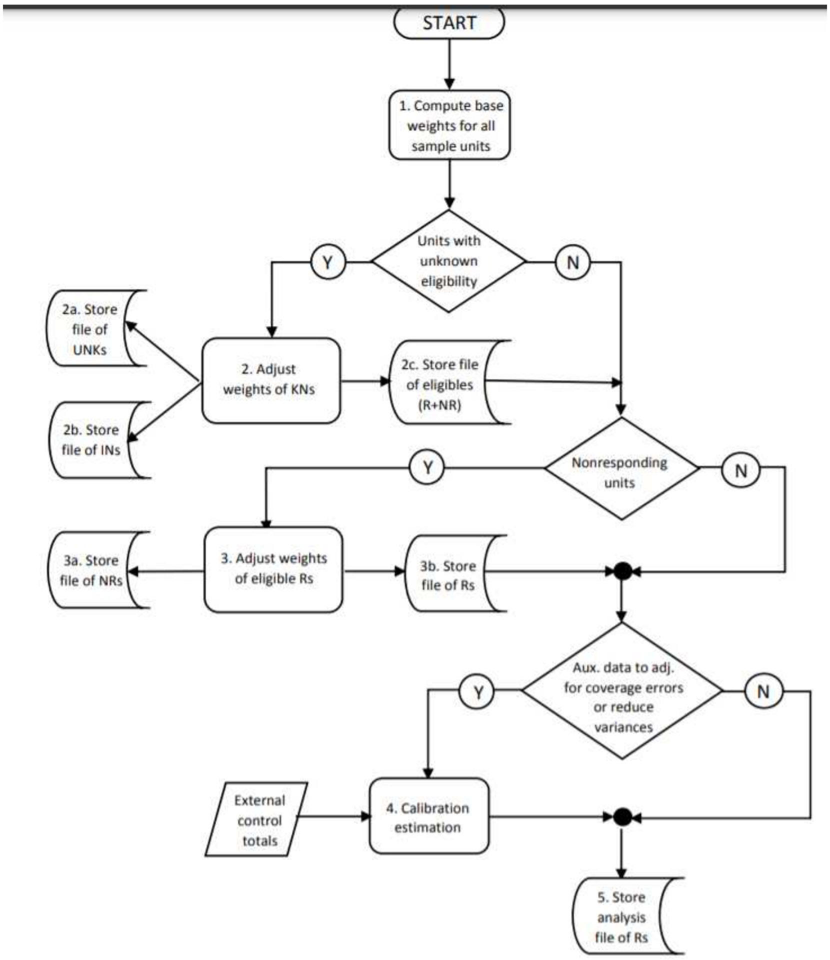
# W2 Specific Steps

**Overview**

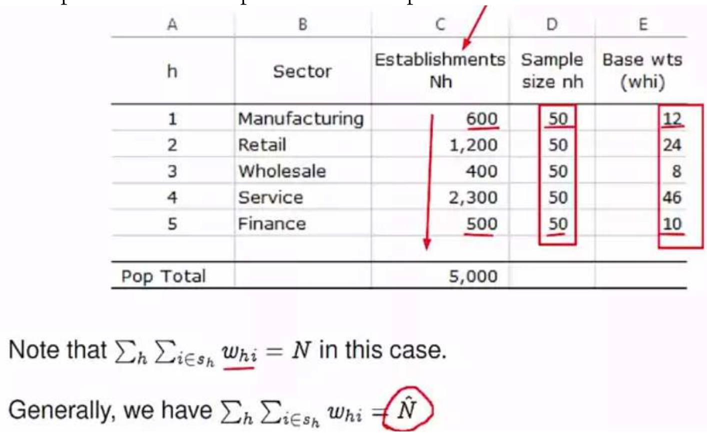
Probability Samples: Four Steps in weighting
- compute base weights ➔ Inverse of selection probabilities; we keep track of those
- adjust base weights to account for units with unknown eligibility (if any)
- adjust for nonresponse
- calibrate to pop control totals


START
1. Compute base weights for all sample units
Units with unknown eligibility
Y
N
2a. Store file of UNKs
2. Adjust weights of KNs
2c. Store file of eligibles (R+NR)
2b. Store file of INs
Nonresponding units
Y
N
3a. Store file of NRs
3. Adjust weights of eligible Rs
3b. Store file of Rs
Aux. data to adj. for coverage errors or reduce variances
Y
N
External control totals
4. Calibration estimation
5. Store analysis file of Rs

Non-probability samples
- No base weights in probability sampling sense
- Identify ineligible units
- No nonresponse in probability sampling sense
- Maybe compute a “pseudo-inclusion” probability and use inverse as a “base weight”
- Calibrate to population control totals

**Base Weights**

Example: Stratified simple random sample

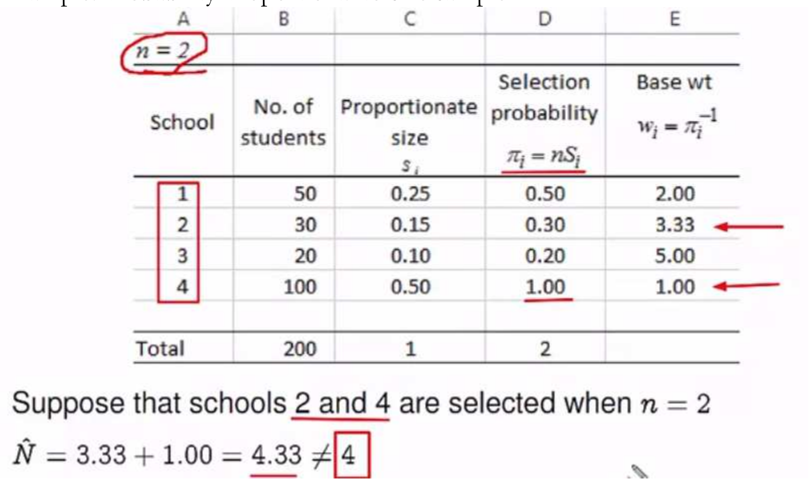


| h | Sector | Establishments Nh | Sample size nh | Base wts (whi) |
|---|---|---|---|---|
| 1 | Manufacturing | 600 | 50 | 12 |
| 2 | Retail | 1,200 | 50 | 24 |
| 3 | Wholesale | 400 | 50 | 8 |
| 4 | Service | 2,300 | 50 | 46 |
| 5 | Finance | 500 | 50 | 10 |
| Pop Total | | 5,000 | | |

Note that $\sum_h \sum_{i\in s_h} w_{hi} = N$ in this case.

Generally, we have $\sum_h \sum_{i\in s_h} w_{hi} = \hat{N}$

Example: Probability Proportional to size sample



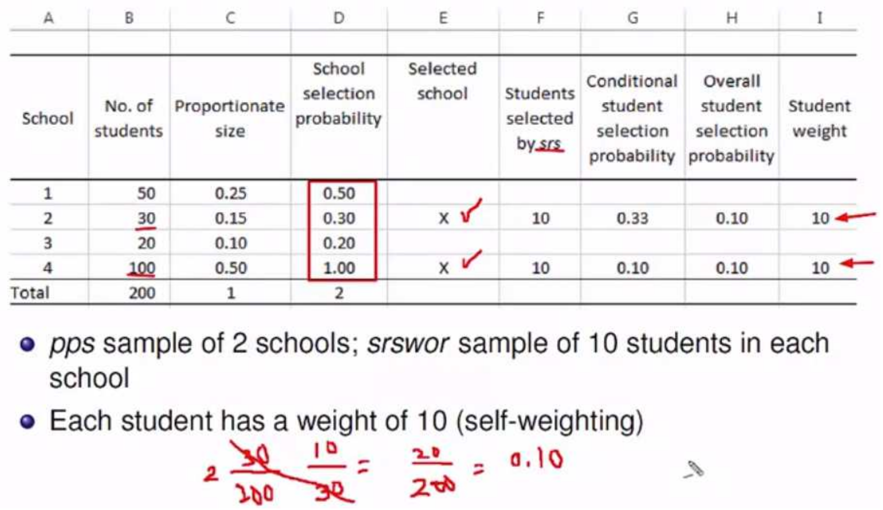
$n = 2$

| School | No. of students | Proportionate size $S_i$ | Selection probability $\pi_i = nS_i$ | Base wt $w_i = \pi_i^{-1}$ |
|---|---|---|---|---|
| 1 | 50 | 0.25 | 0.50 | 2.00 |
| 2 | 30 | 0.15 | 0.30 | 3.33 |
| 3 | 20 | 0.10 | 0.20 | 5.00 |
| 4 | 100 | 0.50 | 1.00 | 1.00 |
| Total | 200 | 1 | 2 | |

Suppose that schools 2 and 4 are selected when $n = 2$

$\hat{N} = 3.33 + 1.00 = 4.33 \neq \boxed{4}$

Example: Two stage sample



| School | No. of students | Proportionate size | School selection probability | Selected school | Students selected by srs | Conditional student selection probability | Overall student selection probability | Student weight |
|---|---|---|---|---|---|---|---|---|
| 1 | 50 | 0.25 | 0.50 | | | | | |
| 2 | 30 | 0.15 | 0.30 | x ✓ | 10 | 0.33 | 0.10 | 10 |
| 3 | 20 | 0.10 | 0.20 | | | | | |
| 4 | 100 | 0.50 | 1.00 | x ✓ | 10 | 0.10 | 0.10 | 10 |
| Total | 200 | 1 | 2 | | | | | |

- *pps* sample of 2 schools; *srswor* sample of 10 students in each school
- Each student has a weight of 10 (self-weighting)

$2 \, \frac{\cancel{30}}{200} \, \frac{10}{\cancel{30}} = \frac{20}{200} = 0.10$

➔ Self weighting sample

If you think there is no reason to think certain student is worth more

• In a stratified simple random sample, every sample unit in a given stratum **cannot** have a different base weight.

**Nonresponse Adjustments**

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Nonignorable NR (NINR)

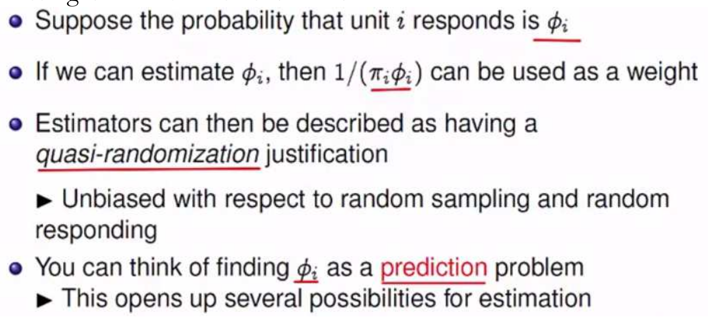
Whether a unit responds or not is treated a random event when response is categorized as one of these Responding could be considered deterministic, i.e. a unit is guaranteed to respond or not. But, random or "stochastic" response is the formulation behind the NR adjustments used.

Missing Data Mechanisms-defined

- MCAR
  - every unit has same probability of responding ➔ responding is just an extra stage of Bernoulli sampling
  - no weight adjustment needed for means; one overall adjustment needed for totals
- MAR
  - probability of responding depends on covariates
  - adjustment possible if covariates known for both Rs and NRs
  - e.g. Suppose that it is known that response rate in a survey of schools depends on a measure of the socioeconomic level of the neighborhood where the school is located and that this level is known for every school in the frame. This is an example of MAR

- NINR
  - probability of responding depends on analytic variables (y's) and possibly covariates
  - adjustment difficult or impossible
  - A survey of schools will be done to assess the extent to which computers are used in teaching mathematics. Suppose that the response rate in the survey depends on both a measure of the socioeconomic level of the neighborhood where the school is located and on whether a school has computers available for students. This is an example of NINR.

Missing data mechanisms-defined



- Suppose the probability that unit $i$ responds is $\phi_i$
- If we can estimate $\phi_i$, then $1/(\pi_i\phi_i)$ can be used as a weight
- Estimators can then be described as having a *quasi-randomization* justification
  - ▶ Unbiased with respect to random sampling and random responding
- You can think of finding $\phi_i$ as a prediction problem
  - ▶ This opens up several possibilities for estimation

**Response Propensities**

General Procedure

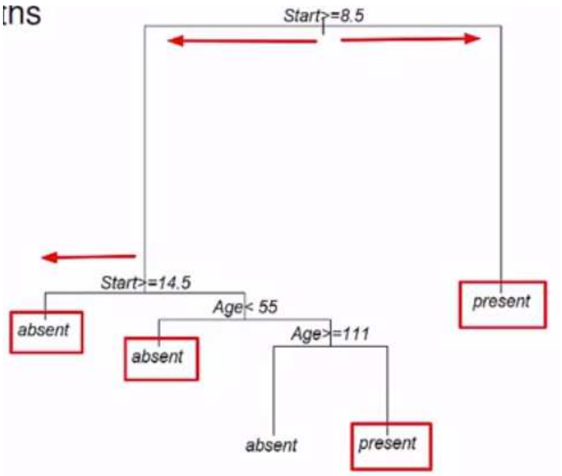
- Estimate response probabilities (propensities) for each R and NR
  - regress binary response variable on covariates
  - logistic regression typically used
- Form groups (cells) for NR adjustment
  - Sort Rs and NRs from low to high by estimated propensity
  - Divide file into groups
  - 5 groups is popular but more can be used, especially if sample is large
- Use of NR adjustment within each group
  - A single adjustment smooths out effects of any extreme propensities produced by the binary regression
  - options within a cell are unweighted RR, weighted RR, average propensity, median propensity
  - options will be similar if range of propensities in a cell is not large

**Tree Algorithms**

General Description
- Use a tree algorithm to model response probabilities
  - regress binary response variable on covariates
- Algorithm selects covariates to use
- Series of data splits leads to a set of terminal nodes
- Terminal nodes used as adjustment cells or used to give estimated response propensities for each unit

Particular Algorithms
- Classification and Regression Trees (CART)
  - Successively splits data into two parts based on covariate that maximizes log-likelihood for a binary variable
  - At each step a different covariate may be selected or a previously used covariate is further split

CART Tree
Kyphosis data from R rpart package
Model whether abnormal curvature still present after corrective surgery
Start = no. of topmost vertebra operated on
Age = age in months



➔ Terminal nodes or bends down at the end would be our non-response adjustment cells

Advantages of trees
- Easy to understand and explain (you can draw pictures)
- Selection of which covariates and interactions of covariates to include in handled automatically
- Variable values, whether categorical or continuous, are grouped automatically

Disadvantages of trees
- Other methods make more accurate predictions
- The other methods cut down on prediction variance associated with trees

More advanced methods
- Random Forests
  - fits many trees to bootstrapped training samples
  - for each tree a random subset of the covariates is picked to reduce correlation among trees
  - average predicted probability across trees is used
- Boosting
  - combines a large number of trees (like random forests does)
  - beginning with a training sample, boosting fits a small tree and then gradually adds to it
  - average predicted probability across trees is used
- For both random forests and boosting a single tree cannot be drawn
  - but a measure of the importance of a covariate in the model can be computed

**Calibration**

General Description

- use covariate data to correct for coverage errors and reduce standard errors
- Population (census) totals for the covariates must be available
- Individual values of covariates need to be known only for sample Rs – not for nonresponding or nonsample units
  - this allows a larger set of covariates for calibration than for NR adjustment
- Covariates should be related to likelihood of being covered by frame or to analytic variables or both

Types of calibration estimators

- Post-stratification
  - classify Rs into groups (post-strata) then adjust weights so that they sum to pop totals of covariates (auxiliaries)
  - age x Race-ethnicity x Gender in a household survey
  - age x Race-ethnicity x Gender census counts needed for time period close to survey period
  - number of post-strata: PRODUCT of category numbers in each covariate (e.g. Suppose that a household survey of persons is poststratified by cross-classification of 5 age categories, 2 genders, and 4 education levels. The number of poststrata is 40)
- Raking
  - similar to post-stratification but only marginal census counts needed
  - age, race-ethnicity, Gender census counts needed
  - total number of control totals: SUM of category numbers in each covariate (e.g. Suppose that a household survey of persons is raked to the marginal population totals for 5 age categories, 2 genders, and 4 education levels. The number of control totals required is 11)

Types of calibration estimators

- General regression estimation (GREG)
  - both qualitative and quantitative variables can be used
  - school survey: student counts, percentage of students receiving free or reduced-price lunch, and grade-range indicators (grades 6-9, 9-12, 10-12 etc.) for covariates