Probability Sampling Methods

Simple Random Sampling

everyone in the population is equally likely to be sampled
done when there is little information available that would drive a choice to stratify the population
primary interest is in multivariate relationships within the population that are obtained from regression analysis, e.g. linear, logistic, Poisson regressions
there is almost never a true simple random sampling in practice

Stratified Random Sampling

population divided into strata that are mutually exclusive and take simple random sample from each stratum
strata are usually subgroups of interest that you want to make sure are represented in sample
e.g., regions of the country, race/ethnicity groups, gender
elements of the same strata tend to be more similar to each other
***for greatest precision, individual elements within each stratum should have similar values, but stratum means should vary as much as possible

<u>Pros</u>
1) make sure minority groups are well-represented in sample
2) decrease costs by sampling in the most efficient way for each stratum
3) lower variance and increase precision
variance within a stratum is lower than variance between strata

<u>Sample Size Selection</u>
goal is select the number in each strata to minimize the total variance
proportionate sample = sample size per strata is proportional to their relative fraction of the population
disproportionate sample = sampling faction is not the same for each strata, done if variation isn't the same within all strata

Cluster Sampling

divide population into smaller groups, primary sampling cluster, secondary sampling cluster, etc.
take a simple random sample within each cluster and continue down levels of clusters
***for greatest precision, individual elements within each cluster should be as different as possible, but cluster means should be similar to each other

Systematic Sample

randomly select a starting point
include that unit in the sample and every k^{th} unit after
can be seen as a special case of cluster sampling

Multistage Sampling

multiple sampling strategies in sequence
can use different sampling methods at each stage

Pros
increases feasibility of obtaining a representative sample of a large population
reduces number of survey locations
increases efficiency of data collection
decreases cost and time needed

Cons
takes a lot more time to plan

Sampling Weights

$$w_i = \frac{1}{\pi_i} - \frac{1}{probability\ unit\ i\ is\ sampled}$$

weights are the inverse of the probability a unit is selected for a sample, which depends on the proportion their group is of a population
in more complex sampling schemes than simple random sampling, weights are used to create unbiased estimators

Simple Random Sampling

$$\pi_i = \frac{n}{N}$$

$$w_i = \frac{1}{\pi_i} = \frac{1}{\frac{n}{N}} = \frac{N}{n}$$

weights do not have a significant impact on estimates as long as n is relatively small compared to N

Stratified Sampling

$$\bar{x}_h = \frac{1}{n_h} \sum_j x_{hj}$$

h indexes the stratum
j indexes the individual

$$\bar{x} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{x}_h$$

N_h = population size in stratum h

$$\pi_{hj} = \frac{n_h}{N_h}$$

$$w_{hj} = \frac{N_h}{n_h}$$

$$\bar{x} = \frac{\sum_{h=1}^{H} \sum_j w_{hj} x_{hj}}{\sum_{h=1}^{H} \sum_j w_{hj}}$$

Cluster Sampling

$$\bar{x} = \frac{\sum_{i=1}^{N} \sum_j w_{ij} x_{ij}}{\sum_{i=1}^{N} \sum_j w_{ij}}$$

$$w_{ij} = \frac{1}{P(SSU\ j\ of\ PSU\ i\ is\ sampled)}$$

$$P(SSU\ j\ of\ PSU\ i\ is\ sampled) = P(i^{th}\ PSU\ is\ sampled) \times P(j^{th}\ SSU\ is\ sampled | i^{th}\ PSU\ is\ sampled) = \frac{n}{N} \times \frac{m_i}{M_i}$$

Accounting for Sampling Weights

sample weights adjust for oversampling certain groups
clustering adjusts for individuals in the same cluster and correlation in the data
weights and adjustments must be made for sampling design, when calculating a population parameter, e.g. prevalence, mean
do not need to weight case-control data when it's estimating an association, e.g. linear regressions, odds ratios
always adjust for clustering

Missing Data

Missing Completely At Random (MCAR)
probability that an individual value will be missing doesn't depend on the outcome, any collected variables, variables not collected, or the survey design
missing data is truly due to random chance and has no pattern to it
missing values are randomly dispersed through variables and observations
unform non-response

Missing At Random (MAR)
probability that an individual value is missing is independent of the outcome and unobserved values, but depends on the covariates in the model
the response rate only depends on the observed data and nothing else
there is a pattern but there is enough data to determine the pattern and predict the missing values
e.g. men are less likely to respond to survey

Non-Ignorable Missing Data
probability that an individual value is missing depends on unobserved variables that information hasn't been collected on
cannot be completely explained by collected variables
no way to do estimators on missing data because no information available on other variables
e.g. missing data is due to whether subject likes chocolate

Lost to Follow-Up
subjects don't want to participate anymore
move residences
illness or death

Handling Missing Data

Prevention

design survey to prevent nonresponse
best method

Complete Cases

omit any subjects with missing data
most common, but worst approach because it ignores missing data issues
assumes MAR/MCAR, but results can be biased if remaining data isn't a representative sample due to non-ignorable missing data
reduces sample size and power

All Available Cases

include all observations without missing values for each variable statistic
number of observations may change between variables
assumes MAR/MCAR

Imputation

estimate missing values using information from other observations using stratification or regressions
requires MAR/MCAR assumption that if true, can provide adequate, reliable, unbiased estimates
allows use to incomplete data without limiting sample size

<u>Stratified Imputation</u>

Step 1	Divide data into homogenous strata
Step 2	Determine variables to be imputed
Step 3	For each stratum, calculate variable means by using mean value for each observation within strata or impute mean values for missing observations.
Step 4	Perform analysis on full data

less information causes systematically underestimation of variance/covariance

<u>Multiple Imputation</u>

Step 1	Select variables with missing data to impute
Step 2	Select explanatory variables that may be associated with missing values
Step 3	Software predicts most likely value for each missing observation, similar to a regression model giving predicted values
Step 4	Repeat multiple times to obtain a measure of variability