Methods Supplementary Lecture 1: Survey Sampling and Design

Department of Government
London School of Economics and Political Science

1 Populations

- Representativeness
- Sampling Frames
- Sampling without a Frame
- Cluster Sampling
- Weights

Inference Population

- We want to speak to a population
- But what population is it?

Inference Population

- We want to speak to a population
- But what population is it?
- Example: "The UK population"

Population Census

- All population units are in study

Population Census

- All population units are in study
- History of national censuses
  - Denmark 1769–1970 (sporadic)
  - U.S. 1790 (decennial)
  - India 1871 (decennial)

Population Census

- All population units are in study
- History of national censuses
  - Denmark 1769–1970 (sporadic)
  - U.S. 1790 (decennial)
  - India 1871 (decennial)
- Other kinds of census
  - Citizen registry
  - Commercial, medical, government records
  - "Big data"

Advantages and Disadvantages

- Advantages
- Disadvantages

Advantages and Disadvantages

- Advantages
  - Perfectly representative
  - Sample statistics are population parameters
- Disadvantages

Advantages and Disadvantages

- Advantages
  - Perfectly representative
  - Sample statistics are population parameters
- Disadvantages
  - Costs
  - Feasibility
  - Need

Representativeness

- What does it mean for a sample to be representative?

Representativeness

- What does it mean for a sample to be representative?
- Different conceptions of representativeness:
  - Design-based: A sample is representative because of how it was drawn (e.g., randomly)
  - Demographic-based: A sample is representative because it resembles in the population in some way (e.g., same proportion of women in sample and population, etc.)
  - Expert judgement: A sample is representative as judged by an expert who deems it "fit for purpose"

Obtaining Representativeness

- Quota sampling (common prior to the 1940s)

Obtaining Representativeness

- Quota sampling (common prior to the 1940s)
- Simple random sampling

Obtaining Representativeness

- Quota sampling (common prior to the 1940s)
- Simple random sampling
- Advanced survey designs

Convenience Samples

- What is a convenience sample?

Convenience Samples

- What is a convenience sample?
- Different types:
  - Passive/opt-in
  - Sample of convenience (not a sample per se)
  - Sample matching
  - Online panels

Convenience Samples

- What is a convenience sample?
- Different types:
  - Passive/opt-in
  - Sample of convenience (not a sample per se)
  - Sample matching
  - Online panels
- "Purposive" samples (common in qualitative studies)

1 Populations

- Representativeness
- Sampling Frames
- Sampling without a Frame
- Cluster Sampling
- Weights

Sampling Frames

- Definition: Enumeration (listing) of all units eligible for sample selection
- Building a sampling frame
  - Combine existing lists
  - Canvass/enumerate from scratch (e.g., walk around and identify all addresses that people might live in)
- There might be multiple frames of the sample population (e.g., telephone list, voter list, residential addresses)
- List might be at wrong unit of analysis (e.g., households when we care about individuals)

Coverage: A Big Issue

- Coverage: any mismatch between population and sampling frame
  - *Undercoverage*: the sampling frame does not include all eligible members of the population (e.g., not everyone has a telephone, so a telephone list does not include all people)
  - *Overcoverage*: the sampling frame includes ineligible units (e.g., residents of a country are not necessarily citizens so a list of residents has overcoverage for the population of residents)
- Coverage of a frame can change over time (e.g., residential mobility, attrition)

Multi-frame Designs

- Construct one sample from multiple sampling frames
- E.g., “Dual-frame” (landline and mobile)
- Analytically complicated
  - Overlap of frames
  - Sample probabilities in each frame

1 Populations

- Representativeness
- Sampling Frames
- Sampling without a Frame
- Cluster Sampling
- Weights

Sampling without a Sampling Frame

- Sometimes we have a population that can be sampled but not (easily) enumerated in full

Sampling without a Sampling Frame

- Sometimes we have a population that can be sampled but not (easily) enumerated in full
- Examples
    - Protest attendees

Sampling without a Sampling Frame

- Sometimes we have a population that can be sampled but not (easily) enumerated in full
- Examples
  - Protest attendees
  - Streams (e.g., people buying groceries)

Sampling without a Sampling Frame

- Sometimes we have a population that can be sampled but not (easily) enumerated in full
- Examples
  - Protest attendees
  - Streams (e.g., people buying groceries)
  - Points in time

Sampling without a Sampling Frame

- Sometimes we have a population that can be sampled but not (easily) enumerated in full
- Examples
  - Protest attendees
  - Streams (e.g., people buying groceries)
  - Points in time
- Population is the sampling frame

Rare or “hidden” populations

- Big concern: coverage!

Rare or "hidden" populations

- Big concern: coverage!
- Solutions?

Rare or "hidden" populations

- Big concern: coverage!
- Solutions?
  - Snowball sampling
  - Informant sampling
  - Targeted sampling
  - Respondent-driven sampling

Inference from Sample to Population

- We want to know population parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

Inference from Sample to Population

- We want to know population parameter θ
- We only observe sample estimate $\hat{\theta}$
- We have a guess but are also uncertain

- What range of values for θ does our $\hat{\theta}$ imply?
- Are values in that range large or meaningful?

How Uncertain Are We?

- Our uncertainty depends on sampling procedures (we'll discuss different approaches shortly)
- Most importantly, *sample size*
  - As $n \to \infty$, uncertainty $\to 0$
- We typically summarize our uncertainty as the *standard error*

Standard Errors (SEs)

- Definition: "The standard error of a sample estimate is the average distance that a sample estimate ($\hat{\theta}$) would be from the population parameter (θ) if we drew many separate random samples and applied our estimator to each."

What affects size of SEs?

- Larger variance in x means smaller SEs
- More unexplained variance in y means bigger SEs
- More observations reduces the numerator, thus smaller SEs
- Other factors:
  - Homoskedasticity
  - Clustering
- Interpretation:
  - Large SE: Uncertain about population effect size
  - Small SE: Certain about population effect size

Ways to Express Our Uncertainty

1. Standard Error
2. Confidence interval
3. p-value

Confidence Interval (CI)

- Definition: Were we to repeat our procedure of sampling, applying our estimator, and calculating a confidence interval *repeatedly* from the population, a fixed percentage of the resulting intervals would include the true population-level slope.
- Interpretation: If the confidence interval overlaps zero, we are uncertain if β differs from zero

Confidence Interval (CI)

- A CI is simply a range, centered on the slope
- Units: Same scale as the coefficient ($\frac{y}{x}$)
- We can calculate different CIs of varying *confidence*
  - Conventionally, $\alpha = 0.05$, so 95% of the CIs will include the β

p-value

- A summary measure in a hypothesis test
- General definition: “the probability of a statistic as extreme as the one we observed, if the null hypothesis was true, the statistic is distributed as we assume, and the data are as variable as observed”
- Definition in the context of a mean: “the probability of a mean as large as the one we observed . . . ”

The p-value is not:

- The probability that a hypothesis is true or false
- A reflection of our confidence or certainty about the result
- The probability that the true slope is in any particular range of values
- A statement about the importance or substantive size of the effect

Significance

1. Substantive significance
2. Statistical significance

Significance

1. Substantive significance
   - Is the effect size (or range of possible effect sizes) *important* in the real world?
2. Statistical significance

Significance

1. Substantive significance
   - Is the effect size (or range of possible effect sizes) *important* in the real world?
2. Statistical significance
   - Is the effect size (or range of possible effect sizes) larger than a predetermined threshold?
   - Conventionally, $p \leq 0.05$

Simple Random Sampling (SRS)

- Advantages
  - Simplicity of sampling
  - Simplicity of analysis
- Disadvantages
  - Need sampling frame and units without any structure
  - Possibly expensive

Sample Estimates from an SRS

- Each unit in frame has equal probability of selection
- Sample statistics are unweighted
- Sampling variances are easy to calculate
- Easy to calculate sample size need for a particular variance

Sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

where y_i = value for a unit, and
n = sample size

$$SE_{\bar{y}} = \sqrt{(1-f)\frac{s^2}{n}} \tag{2}$$

where f = proportion of population sampled,
s^2 = sample (element) variance, and
n = sample size

Sample proportion

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3}$$

where y_i = value for a unit, and
n = sample size

$$SE_{\bar{y}} = \sqrt{\frac{(1-f)}{(n-1)}p(1-p)} \tag{4}$$

where f = proportion of population sampled,
p = sample proportion, and
n = sample size

Estimating sample size

- Imagine we want to conduct a political poll
- We want to know what percentage of the public will vote for which coalition/party
- How big of a sample do we need to make a relatively precise estimate of voter support?

Estimating sample size

$$Var(p) = (1 - f)\frac{p(1-p)}{n-1} \tag{5}$$

Given the large population:

$$Var(p) = \frac{p(1-p)}{n-1} \tag{6}$$

Need to solve the above for n.

$$\tag{7}$$

Estimating sample size

$$Var(p) = (1 - f)\frac{p(1-p)}{n-1} \tag{5}$$

Given the large population:

$$Var(p) = \frac{p(1-p)}{n-1} \tag{6}$$

Need to solve the above for n.

$$n = \frac{p(1-p)}{v(p)} = \frac{p(1-p)}{SE^2} \tag{7}$$

Estimating sample size

Determining sample size requires:

- A possible value of p
- A desired precision (SE)

If support for each coalition is evenly matched ($p = 0.5$):

$$n = \frac{0.5(1 - 0.5)}{SE^2} = \frac{0.25}{SE^2} \tag{8}$$

Estimating sample size

What precision (margin of error) do we want?

- $+/$- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \qquad (9)$$

Estimating sample size

What precision (margin of error) do we want?

- $+/-$ 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{9}$$

- $+/-$ 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{10}$$

Estimating sample size

What precision (margin of error) do we want?

- +/- 2 percentage points: $SE = 0.01$

$$n = \frac{0.25}{0.01^2} = \frac{0.25}{0.0001} = 2500 \tag{9}$$

- +/- 5 percentage points: $SE = 0.025$

$$n = \frac{0.25}{0.000625} = 400 \tag{10}$$

- +/- 0.5 percentage points: $SE = 0.0025$

$$n = \frac{0.25}{0.00000625} = 40,000 \tag{11}$$

Important considerations

- Required sample size depends on p and SE

Important considerations

- Required sample size depends on p and SE
- In large populations, population size is irrelevant

Important considerations

- Required sample size depends on p and SE
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled

Important considerations

- Required sample size depends on p and SE
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled
- In anything other than an SRS, sample size calculation is more difficult

Important considerations

- Required sample size depends on p and SE
- In large populations, population size is irrelevant
- In small populations, precision is influenced by the proportion of population sampled
- In anything other than an SRS, sample size calculation is more difficult
- Much political science research assumes SRS even though a more complex design is actually used

Sampling Error

- Definition? Reasons why a sample estimate may not match the population parameter

Sampling Error

- Definition? Reasons why a sample estimate may not match the population parameter
- Unavoidable!

Sampling Error

- Definition? Reasons why a sample estimate may not match the population parameter
- Unavoidable!
- Sources of sampling error:
  - Sampling
  - Sample size
  - Unequal probabilities of selection
  - Non-Stratification
  - Cluster sampling

Simple Random Sampling (SRS)

- Advantages
  - Simplicity of sampling
  - Simplicity of analysis
- Disadvantages
  - Need complete sampling frame
  - Possibly expensive

Stratified Sampling

- What is it? Random samples within “strata” of the population
- Why do we do? To reduce uncertainty of our estimates

Stratified Sampling

- What is it? Random samples within “strata” of the population
- Why do we do? To reduce uncertainty of our estimates
- Most useful when subpopulations are:
  1. identifiable in advance
  2. differ from one another
  3. have low within-stratum variance

Stratified Sampling

- Advantages

Stratified Sampling

- Advantages
  - Avoid certain kinds of sampling errors
  - Representative samples of subpopulations
  - Often, lower variances (greater precision of estimates)

Stratified Sampling

- Advantages
  - Avoid certain kinds of sampling errors
  - Representative samples of subpopulations
  - Often, lower variances (greater precision of estimates)
- Disadvantages

Stratified Sampling

- Advantages
  - Avoid certain kinds of sampling errors
  - Representative samples of subpopulations
  - Often, lower variances (greater precision of estimates)
- Disadvantages
  - Need complete sampling frame
  - Possibly (more) expensive
  - No advantage if strata are similar
  - Analysis is more potentially more complex than SRS

Outline of Process

1. Identify our population
2. Construct a sampling frame
3. Identify variables we already have that are related to our survey variables of interest
4. Stratify or subset or sampling frame based on these characteristics
5. Collect an SRS (of some size) within each stratum
6. Aggregate our results

Estimates from a stratified sample

- Within-strata estimates are calculated just like an SRS
- Within-strata variances are calculated just like an SRS

- Sample-level estimates are weighted averages of stratum-specific estimates
- Sample-level variances are weighted averages of stratum-specific variances

Design effect

- What is it?

Design effect

- What is it?
- Ratio of variances in a design against a same-sized SRS

Design effect

- What is it?
- Ratio of variances in a design against a same-sized SRS
- $d^2 = \frac{Var_{stratified}(y)}{Var_{SRS}(y)}$

Design effect

- What is it?
- Ratio of variances in a design against a same-sized SRS
- $d^2 = \frac{Var_{stratified}(y)}{Var_{SRS}(y)}$
- Possible to convert design effect into an *effective sample size*:
- $n_{effective} = \frac{n}{d}$

How many strata?

- How many strata can we have in a stratified sampling plan?

How many strata?

- How many strata can we have in a stratified sampling plan?
- As many as we want, up to the limits of sample size

How do we allocate sample units to strata?

- Proportional allocation
- Optimal precision
- Allocation based on stratum-specific precision objectives

Example Setup

- Interested in individual-level rate of crime victimization in some country
- We think rates differ among native-born and immigrant populations
- Assume immigrants make up 12% of population
- Compare uncertainty from different designs ($n = 1000$)

SRS

- Assume equal rates across groups ($p = 0.10$)
- Overall estimate is just $\frac{Victims}{n}$
- $SE(p) = \sqrt{\frac{p(1-p)}{n-1}}$
- $SE(p) = \sqrt{\frac{0.09}{999}} = 0.0095$

SRS

- Assume equal rates across groups ($p = 0.10$)
- Overall estimate is just $\frac{Victims}{n}$
- $SE(p) = \sqrt{\frac{p(1-p)}{n-1}}$
- $SE(p) = \sqrt{\frac{0.09}{999}} = 0.0095$
- SEs for subgroups (native-born and immigrants)?

SRS

- Assume equal rates across groups ($p = 0.10$)
- Overall estimate is just $\frac{Victims}{n}$
- $SE(p) = \sqrt{\frac{p(1-p)}{n-1}}$
- $SE(p) = \sqrt{\frac{0.09}{999}} = 0.0095$
- SEs for subgroups (native-born and immigrants)?
- What happens if we don't get any immigrants in our sample?

Proportionate Allocation I

- Assume equal rates across groups
- Sample 880 native-born and 120 immigrant individuals
- $SE(p) = \sqrt{Var(p)}$, where
  - $Var(p) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{p_h(1-p_h)}{n_h-1}$
  - $Var(p) = (\frac{0.09}{879})(.88^2) + (\frac{0.09}{119})(.12^2)$
  - $SE(p) = 0.0095$
- Design effect: $d^2 = \frac{0.0095^2}{0.0095^2} = 1$

Proportionate Allocation I

- Note that in this design we get different levels of uncertainty for subgroups
- $SE(p_{native}) = \sqrt{\frac{p(1-p)}{879}} = \sqrt{\frac{0.09}{879}} = 0.010$
- $SE(p_{imm}) = \sqrt{\frac{p(1-p)}{119}} = \sqrt{\frac{0.09}{119}} = 0.028$

Proportionate Allocation IIa

- Assume different rates across groups (immigrants higher risk)
- $p_{native} = 0.1$ and $p_{imm} = 0.3$ (thus $p_{pop} = 0.124$)
- $Var(p) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{p_h(1-p_h)}{n_h - 1}$
- $Var(p) = (\frac{0.09}{879})(.88^2) + \frac{0.21}{119})(.12^2))$
- $SE(p) = 0.01022$

Proportionate Allocation IIa

- $SE(p) = 0.01022$
- Compare to SRS:
  - $SE(p) = \sqrt{\frac{0.124(1-0.124)}{n-1}} = 0.0104$
- Design effect: $d^2 = \frac{0.01022^2}{0.0104^2} = 0.9657$
- $n_{effective} = \frac{n}{sqrt(d^2)} = 1017$

Proportionate Allocation IIa

- Subgroup variances are still different
- $SE(p_{native}) = \sqrt{\frac{p(1-p)}{879}} = \sqrt{\frac{.09}{879}} = 0.010$
- $SE(p_{imm}) = \sqrt{\frac{p(1-p)}{119}} = sqrt\frac{.21}{119} = 0.040$

Proportionate Allocation IIb

- Assume different rates across groups (immigrants lower risk)
- $p_{native} = 0.3$ and $p_{imm} = 0.1$ (thus $p_{pop} = 0.276$)
- $Var(p) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{p_h(1-p_h)}{n_h - 1}$
- $Var(p) = (\frac{0.21}{879})(.88^2) + \frac{0.09}{119})(.12^2))$
- $SE(p) = 0.014$

Proportionate Allocation IIb

- $SE(p) = 0.014$
- Compare to SRS:
  - $SE(p) = \sqrt{\frac{0.276(1-0.276)}{n-1}} = 0.0141$
- Design effect: $d^2 = \frac{0.014^2}{0.0141^2} = 0.9859$
- $n_{effective} = \frac{n}{sqrt(d^2)} = 1007$

Proportionate Allocation IIa

- Subgroup variances are still different
- $SE(p_{native}) = \sqrt{\frac{p(1-p)}{879}} = \sqrt{\frac{.21}{879}} = 0.0155$
- $SE(p_{imm}) = \sqrt{\frac{p(1-p)}{119}} = sqrt\frac{.09}{119} = 0.0275$

Proportionate Allocation IIc

- Look at same design, but a different survey variable (household size)
- Assume: $\bar{y}_{native} = 4$ and $\bar{Y}_imm = 6$ (thus $\bar{Y}_{pop} = 4.24$)
- Assume: $Var(Y_{native}) = 1$ and $Var(Y_imm) = 3$ and $Var(Y_{pop}) = 4$
- $Var(\bar{y}) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{s_h^2}{n_h}$
- $SE(\bar{y}) = \sqrt{\frac{1^2}{880}(.88^2) + \frac{3^2}{120}(.12^2)} = 0.0443$

Proportionate Allocation IIc

- $SE(\bar{y}) = 0.0443$
- Compare to SRS:
  - $SE(\bar{y}) = \sqrt{\frac{s^2}{n}} = \sqrt{4/1000} = 0.0632$
- Design effect: $d^2 = \frac{0.0443^2}{0.0632^2} = 0.491$
- $n_{effective} = \frac{n}{sqrt(d^2)} = 1427$

Proportionate Allocation IIc

- $SE(\bar{y}) = 0.0443$
- Compare to SRS:
  - $SE(\bar{y}) = \sqrt{\frac{s^2}{n}} = \sqrt{4/1000} = 0.0632$
- Design effect: $d^2 = \frac{0.0443^2}{0.0632^2} = 0.491$
- $n_{effective} = \frac{n}{sqrt(d^2)} = 1427$
- Why is d^2 so much larger here?

Disproportionate Allocation I

- Previous designs obtained different precision for subgroups
- Design to obtain stratum-specific precision (e.g., $SE(p_h) = 0.02$)
- $n_h = \frac{p(1-p)}{v(p)} = \frac{p(1-p)}{SE^2}$
- $n_{native} = \frac{0.09}{0.02^2} = 225$
- $n_{imm} = \frac{0.21}{0.02^2} = 525$
- $n_{total} = 225 + 525 = 750$

Disproportionate Allocation II

- Neyman optimal allocation
- How does this work?
  - Allocate cases to strata based on within-strata variance
  - Only works for one variable at a time
  - Need to know within-strata variance

Disproportionate Allocation II

- Assume big difference in victimization
- $p_{native} = 0.01$ and $p_{imm} = 0.50$ (thus $p_{pop} = 0.0688$)
- Allocate according to: $n_h = n\frac{W_h S_h}{\sum_{h=1}^{H} W_h S_h}$
- $\sum_{h=1}^{H} W_h S_h = (0.88 * 0.0099) + (0.12 * 0.25) = 0.0387$
- $n_{native} = 1000\frac{0.0087}{0.0387} = 225$
- $n_{imm} = 1000\frac{0.03}{0.0387} = 775$

Disproportionate Allocation II

- $SE(p_{native}) = \sqrt{\frac{p(1-p)}{225}} = \sqrt{\frac{0.0099}{225}} = 0.00663$
- $SE(p_{imm}) = \sqrt{\frac{p(1-p)}{775}} = \sqrt{\frac{.25}{775}} = 0.01796$
- $Var(p) = \sum_{h=1}^{H} (\frac{N_h}{N})^2 \frac{p_h(1-p_h)}{n_h - 1}$
- $Var(p) = (\frac{0.0099}{225})(.88^2) + (\frac{0.25}{775})(.12^2)$
- $SE(p) = 0.00622$

Disproportionate Allocation II

- $SE(p) = 0.00622$
- Compare to SRS:
  - $SE(p) = \sqrt{\frac{0.0688(1-0.0688)}{n-1}} = 0.008$
- Design effect: $d^2 = \frac{0.00622^2}{0.008^2} = 0.6045$
- $n_{effective} = \frac{n}{sqrt(d^2)} = 1286$

Final Considerations

- Reductions in uncertainty come from creating homogeneous groups
- Estimates of design effects are variable-specific
- Sampling variance calculations do not factor in time, costs, or feasibility

Cluster Sampling

- What is it?
- Why do we do?

Cluster Sampling

- What is it?
- Why do we do?
- Most useful when:
    1. Population has a clustered structure
    2. Unit-level sampling is expensive or not feasible
    3. Clusters are similar

Cluster Sampling

- Advantages

Cluster Sampling

- Advantages
  - Cost savings!
  - Capitalize on clustered structure

Cluster Sampling

- Advantages
  - Cost savings!
  - Capitalize on clustered structure
- Disadvantages

Cluster Sampling

- Advantages
  - Cost savings!
  - Capitalize on clustered structure
- Disadvantages
  - Units tend to cluster for complex reasons (self-selection)
  - Major increase in uncertainty if clusters differ from each other
  - Complex to design (and possibly to administer)
  - Analysis is much more complex than SRS or stratified sample

Cluster Sampling

- Number of stages
  - One-stage sampling
  - Two- or more-stage sampling
- Number of clusters
- Sample size w/in clusters
- Everything depends on variability of clusters

Sampling Variance for Cluster Sampling

- Sampling variance depends on *between*-cluster variation:
  $Var(\bar{y}) = (\frac{1-f}{a})(\frac{1}{a-1})(\sum_{\alpha=1}^{a}(\bar{y}_\alpha - \bar{y})^2)$
- When *between*-cluster variance is high, *within*-cluster variance is likely to be low
  - "Cluster homogeneity"

Design Effect for Cluster Sampling

- Cluster samples almost always less *statistically* efficient than SRS

- Design Effect depends on cluster homogeneity:
  - $d^2 = \frac{Var_{clustered}(y)}{Var_{SRS}(y)}$
  - $d^2 = 1 + (n_{cluster} - 1)roh$

- *roh* (*intraclass correlation coefficient*):
  - Proportion of unit-level variance that is between-clusters
  - Generally positive and small (about 0.00 to 0.10)

Goal of Survey Research

- The goal of survey research is to estimate population-level quantities (e.g., means, proportions, totals)
- Samples estimate those quantities with uncertainty (sampling error)
- Sample estimates are unbiased if they match population quantities

Realities of Survey Research

- Sample may not match population for a variety of reasons:
  - Due to constraints on design
  - Due to sampling frame coverage
  - Due to intentional over/under-sampling
  - Due to nonresponse
  - Due to sampling error

Realities of Survey Research

- Sample may not match population for a variety of reasons:
  - Due to constraints on design
  - Due to sampling frame coverage
  - Due to intentional over/under-sampling
  - Due to nonresponse
  - Due to sampling error
- Weighting is never perfect
  - Limited to work with observed variables
  - Rarely have good knowledge of coverage, nonresponse, or sampling error
  - Weighting can increase sampling variance

Three Kinds of Weights

- Design Weights
- Nonresponse Weights
- Post-Stratification Weights

Design Weights

- Address design-related unequal probability of selection into a sample
- Applied to *complex survey designs*:
  - Disproportionate allocation stratified sampling
  - Oversampling of subpopulations
  - Cluster sampling
  - Combinations thereof

Design Weights: SRS

- Imagine sampling frame of 100,000 units
- Sample size will be 1,000
- What is the probability that a unit in the sampling frame is included in the sample?

Design Weights: SRS

- Imagine sampling frame of 100,000 units
- Sample size will be 1,000
- What is the probability that a unit in the sampling frame is included in the sample?
- $p = \frac{1000}{100{,}000} = .01$

Design Weights: SRS

- Imagine sampling frame of 100,000 units
- Sample size will be 1,000
- What is the probability that a unit in the sampling frame is included in the sample?
- $p = \frac{1000}{100{,}000} = .01$
- Design weight for all units is $w = 1/p = 100$
- SRS is *self-weighting*

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Native-born & 10,000 Immigrants
- Sample size will be 1,000 (proportionate allocation)
  - 900 Native-born & 100 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Native-born & 10,000 Immigrants
- Sample size will be 1,000 (proportionate allocation)
  - 900 Native-born & 100 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{native} = \frac{900}{90{,}000} = .01$
  - $p_{Imm} = \frac{100}{10{,}000} = .01$

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Native-born & 10,000 Immigrants
- Sample size will be 1,000 (proportionate allocation)
  - 900 Native-born & 100 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{native} = \frac{900}{90{,}000} = .01$
  - $p_{Imm} = \frac{100}{10{,}000} = .01$
- Design weight for all units is $w = 1/p = 100$
- Proportionate allocation is *self-weighting*

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Native-born & 10,000 Immigrants
- Sample size will be 1,000 (disproportionate allocation)
  - 500 Native-born & 500 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Native-born & 10,000 Immigrants
- Sample size will be 1,000 (disproportionate allocation)
  - 500 Native-born & 500 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{Native} = \frac{500}{90{,}000} = .0056$
  - $p_{Imm} = \frac{500}{10{,}000} = .05$

Design Weights: Stratified Sample

- Imagine sampling frame of 100,000 units
  - 90,000 Native-born & 10,000 Immigrants
- Sample size will be 1,000 (disproportionate allocation)
  - 500 Native-born & 500 Immigrants
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p_{Native} = \frac{500}{90{,}000} = .0056$
  - $p_{Imm} = \frac{500}{10{,}000} = .05$
- Design weights differ across units:
  - $w_{Native} = 1/p_{Danish} = 178.57$
  - $w_{Imm} = 1/p_{Imm} = 20$

Design Weights: Cluster Sample

- Imagine sampling frame of 1000 units in 5 clusters of varying sizes
- Sample size will be 10 each from 3 clusters
- What is the probability that a unit in the sampling frame is included in the sample?
  - $p = n_{clusters}/N_{clusters} * 1/n_{cluster} = \frac{3}{5} * 1/n_{cluster}$

Design Weights: Cluster Sample

- Imagine sampling frame of 1000 units in 5 clusters of varying sizes
- Sample size will be 10 each from 3 clusters
- What is the probability that a unit in the sampling frame is included in the sample?
    - $p = n_{clusters}/N_{clusters} * 1/n_{cluster} = \frac{3}{5} * 1/n_{cluster}$

Design Weights: Cluster Sample

- Imagine sampling frame of 1000 units in 5 clusters of varying sizes
- Sample size will be 10 each from 3 clusters
- What is the probability that a unit in the sampling frame is included in the sample?
    - $p = n_{clusters}/N_{clusters} * 1/n_{cluster} = \frac{3}{5} * 1/n_{cluster}$
- Design weights differ across units:
    - Clusters are equally likely to be sampled
    - Probability of selection within cluster varies with cluster size
- Cluster sampling is rarely *self-weighting*

Nonresponse Weights

- Correct for nonresponse
- Require knowledge of nonrespondents on variables that have been measured for respondents
- Requires data are *missing at random*
- Two common methods
  - Weighting classes
  - Propensity score subclassification

Nonresponse Weights: Example

- Imagine immigrants end up being less likely to respond[1]
  - $RR_{Native} = 1.0$
  - $RR_{Imm} = 0.8$

[1] This refers to a lower RR in this particular survey sample, not in general.

Nonresponse Weights: Example

- Imagine immigrants end up being less likely to respond[1]
  - $RR_{Native} = 1.0$
  - $RR_{Imm} = 0.8$
- Using weighting classes:
  - $w_{rr,Native} = 1/1 = 1$
  - $w_{rr,Imm} = 1/0.8 = 1.25$
- Can generalize to multiple variables and strata

[1] This refers to a lower RR in this particular survey sample, not in general.

Post-Stratification

- Correct for nonresponse, coverage errors, and sampling errors

Post-Stratification

- Correct for nonresponse, coverage errors, and sampling errors
- Reweight sample data to match population distributions
    - Divide sample and population into strata
    - Weight units in each stratum so that the weighted sample stratum contains the same proportion of units as the population stratum does

Post-Stratification

- Correct for nonresponse, coverage errors, and sampling errors
- Reweight sample data to match population distributions
    - Divide sample and population into strata
    - Weight units in each stratum so that the weighted sample stratum contains the same proportion of units as the population stratum does
- There are numerous other related techniques

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native, Female	.45	.5		
Native, Male	.45	.4		
Immigrant, Female	.05	.07		
Immigrant, Male	.05	.03		

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native, Female	.45	.5	Over	
Native, Male	.45	.4	Under	
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native, Female	.45	.5	Over	0.900
Native, Male	.45	.4	Under	
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native, Female	.45	.5	Over	0.900
Native, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native, Female	.45	.5	Over	0.900
Native, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	0.714
Immigrant, Male	.05	.03	Under	

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification: Example

- Imagine our sample ends up skewed on immigration status and gender relative to the population

Group	Pop.	Sample	Rep.	Weight
Native, Female	.45	.5	Over	0.900
Native, Male	.45	.4	Under	1.125
Immigrant, Female	.05	.07	Over	0.714
Immigrant, Male	.05	.03	Under	1.667

- PS weight is just $w_{ps} = N_l / n_l$

Post-Stratification

- Should only be done after correcting for sampling design
- Strata must be large ($n > 15$)
- Need accurate population-level stratum sizes
- Only useful if stratifying variables are related to key constructs of interest

Post-Stratification

- Should only be done after correcting for sampling design
- Strata must be large ($n > 15$)
- Need accurate population-level stratum sizes
- Only useful if stratifying variables are related to key constructs of interest
- This is the basis for inference in non-probability samples
  - Probability samples make design-based inferences
  - Non-probability samples post-stratify to obtain descriptive representativeness

Weighted Analyses

- We can analyze data that *should be* weighted without the weights, but they are no longer mathematically representative of the larger population
- Using the weights is the way to make population-representative claims from survey data
- Most statistics can be modified to use weights, e.g.:
  - Unweighted mean: $\frac{1}{n}\sum_{i=1}^{n} x_i$
  - Weighted mean: $\frac{1}{n}\sum_{i=1}^{n} x_i * w_i$

Response Rates

- Why do we care?

Response Rates

- Why do we care?
- Survey Error
  - Variance
  - Bias

Response Rates

- Why do we care?
- Survey Error
  - Variance
  - Bias
- Sample size calculations (and design effects) are based on completed interviews

Response Rates

- Why do we care?
- Survey Error
  - Variance
  - Bias
- Sample size calculations (and design effects) are based on completed interviews
- Cost, time, and effort

Response Rates

- Imagine we need $n = 1000$
- How many attempts to obtain that sample:

Response Rate	Needed Attempts
1.00	1000
0.75	1333
0.50	2000
0.25	4000
0.10	10,000

Response Rate

- Interviews divided by eligibles
- $RR = \frac{I}{E}$
- Challenges
  - Unknown eligibility
  - Partial interviews
  - Non-probability samples
  - Complex survey designs
- Cooperation Rate (I's divided by contacts)

Disposition Codes

Every attempt to interview someone needs to be categorized into a “disposition code”. The usual codes fall into four broad categories:

- Interviews
- Refusals
- Unknowns
- Ineligibles

Disposition Codes

- Complete Interview (I)
- Partial Interview (P)
- Non-interviews
  - Refusal (R)
  - Non-contact (NC)
  - Other (O)

What is a refusal?

- How do categorize a respondent as a refusal?

What is a refusal?

- How do categorize a respondent as a refusal?
- When can we try to convert an apparent refusal?

What is a refusal?

- "I don't want to participate."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."
- "My husband can do it if you call back."

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."
- "My husband can do it if you call back."
- "How did you get my number?"

What is a refusal?

- "I don't want to participate."
- "I'm too busy to do this right now."
- "What do I get for my time?"
- (Hang-up phone without saying anything.)
- "Okay, but I only have 5 minutes."
- "My husband can do it if you call back."
- "How did you get my number?"
- "Go f' yourself."

Disposition Codes

- Complete Interview (I)
- Partial Interview (P)
- Non-interviews
  - Refusal (R)
  - Non-contact (NC)
  - Other (O)

Disposition Codes

- Complete Interview (I)
- Partial Interview (P)
- Non-interviews
  - Refusal (R)
  - Non-contact (NC)
  - Other (O)
- Unknowns (U)
- Ineligibles

Eligibility

- Why would an ineligible unit be in our sample?

Eligibility

- Why would an ineligible unit be in our sample?
- How do we determine ineligibility?

Eligibility

- Why would an ineligible unit be in our sample?
- How do we determine ineligibility?
- What do we do with “unknowns”?

Response Rates[2]

Without accounting for eligibility of unknowns:

- $RR1 = \frac{I}{(I+P)+(R+NC)+U}$
- $RR2 = \frac{I+P}{(I+P)+(R+NC)+U}$

Accounting for eligibility of unknowns:

- $RR3 = \frac{I}{(I+P)+(R+NC)+(e*U)}$
- $RR4 = \frac{I+P}{(I+P)+(R+NC)+(e*U)}$

- e is estimated Pr(eligible) among unknowns

[2]Note: Simplified slightly

Refusal Rates

- Related to response rate
- Numerator is refusals
- E.g., $REF1 = \frac{R}{(I+P)+(R+NC)+U}$

Complex Survey Designs

- Stratified Sampling (unequal allocation)
  - Sums of codes weighted by $\frac{1}{p}$
  - p is probability of selection
  - May want to report stratum-specific rates
- Multi-stage sampling (e.g., cluster sampling)
  - RR is product of cluster cooperation and within-cluster response rate

Internet Surveys

- For *probability-based samples*, RR is a product of:
  - Recruitment Rate (RR for panel enrollment)
  - Completion Rate (RR for specific survey)
  - Profile Rate (in some cases)
  - E.g., if Recruitment Rate is 30% and Completion Rate is 80%, $RR = 0.3 * 0.8 = 24\%$
- For *non-probability samples*, RR is undefined
  - No sampling involved (so no denominator)
  - If from panel, report Completion Rate
  - If fully opt-in, there's nothing you can do