Sampling Techniques

Department of Political Science and Government
Aarhus University

September 22, 2014

Assignment for this week

1. Form groups of 3 (or so)
2. Discuss the sampling plans for the surveys you identified online
3. Select one of those from your group to present to the class
4. Think about: coverage, representativeness, sample size

Ideas from last week

1. What is a population?

Ideas from last week

1. What is a population?
2. What is a sampling frame?

Ideas from last week

1. What is a population?
2. What is a sampling frame?
3. What is a sample?

Ideas from last week

1. What is a population?
2. What is a sampling frame?
3. What is a sample?
4. How do we construct a sampling frame?

Activity!

- Work in pairs
- Pick one of the two populations
- Develop two sampling frames/sampling strategies for a population
- Share with class and discuss

Ideas from last week

1. What is a population?
2. What is a sampling frame?
3. What is a sample?
4. How do we construct a sampling frame?

Ideas from last week

1. What is a population?
2. What is a sampling frame?
3. What is a sample?
4. How do we construct a sampling frame?
5. What is the process of determining necessary sample size for a study?

Questions about sampling strategies?

Total Survey Error

- What sources of survey error have we discussed so far?

Total Survey Error

- What sources of survey error have we discussed so far?
- Now we also think about sampling error

Sampling Error

- Definition?

Sampling Error

- Definition?
- Unavoidable!

Sampling Error

- Definition?
- Unavoidable!
- Sources of sampling error:
  - Sampling
  - Sample size
  - Unequal probabilities of selection
  - Non-Stratification
  - Cluster sampling

Readings for this week

1. Walter and Enticott

Readings for this week

1. Walter and Enticott
2. Reinisch et al.

Readings for this week

1. Walter and Enticott
2. Reinisch et al.
3. AAPOR Report

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?
- How are online panelists recruited?

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?
- How are online panelists recruited?
- How good is the coverage for an online panel? How would we evaluate it?

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?
- How are online panelists recruited?
- How good is the coverage for an online panel? How would we evaluate it?
- How are panelists recruited into studies?

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?
- How are online panelists recruited?
- How good is the coverage for an online panel? How would we evaluate it?
- How are panelists recruited into studies?
- Does stratified sampling of panelists solve concerns about representativeness?

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?
- How are online panelists recruited?
- How good is the coverage for an online panel? How would we evaluate it?
- How are panelists recruited into studies?
- Does stratified sampling of panelists solve concerns about representativeness?
- How do we assess response rates for an online panel?

Online panels/Non-Probability Surveys

- What are the major issues raised in the AAPOR Report?
- How are online panelists recruited?
- How good is the coverage for an online panel? How would we evaluate it?
- How are panelists recruited into studies?
- Does stratified sampling of panelists solve concerns about representativeness?
- How do we assess response rates for an online panel?
- How long should someone be eligible to be in an online panel?

Purposive and Quota Sampling

- What is purposive sampling?
- What is quota sampling?
- What concerns do we have about these methods?
- When are they appropriate?

Questions about non-probability sampling?

Simple Random Sampling (SRS)

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

Stratified Sampling

- What is it?
- Why do we do?

Stratified Sampling

- What is it?
- Why do we do?
- 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 strataum-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

Questions about stratified sampling?

1 Assignment

2 Review of Last Week

3 New Material to Cover
- Total Survey Error
- Readings
- Online Panels
- Stratified Sampling
- An Extended Example
- Cluster Sampling

4 Preview of Next Week

Example Setup

- Interested in individual-level rate of crime victimization in Denmark
- 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 $\bar{Y}_i mm = 6$ (thus $\bar{Y}_{pop} = 4.24$)
- Assume: $Var(Y_{native}) = 1$ and $Var(Y_i mm) = 3$ 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

Questions about stratified sampling?

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

Example: Burnham et al.

- What is the research question?

Example: Burnham et al.

- What is the research question?
- What are the population and unit of analysis?

Example: Burnham et al.

- What is the research question?
- What are the population and unit of analysis?
- What is the sampling strategy? Why?

Example: Burnham et al.

- What is the research question?
- What are the population and unit of analysis?
- What is the sampling strategy? Why?
- What do they find?

Questions about cluster sampling?

Assignment for next week: Task

- What is your research topic/question?
- What is your population?
- What is your sampling frame? How does it over-cover or under-cover your population?
- How do you plan to sample?
- How big of a sample do you need?

Assignment for next week: Procedure

- Present it in-class next week
- Email me your assignment (by Saturday night)
- Meet with me tomorrow or Wednesday

Next week's agenda

- Cluster sampling
- Concept definition and operationalization
- Opinion questions and factual questions
- Practice developing questions