Introduction to Sampling for Non-Statisticians

Dr. Safaa R. Amer

Overview

Part I
- Introduction
- Census or Sample
- Sampling Frame
- Probability or non-probability sample
- Sampling with or without replacement
- Some random sampling techniques

Part II
- Sample Size
- Post-sampling steps
- Sampling versus non-sampling errors
- References
Questions to Ponder Upon!

• Pre-sampling
  • Nature of study: exploratory, descriptive, analytical
  • Variables of interest
  • Target population & sub-populations
  • Data collection mode
  • Is sampling appropriate?

• During sampling
  • Availability of population listing
  • Error that can be tolerated
  • Sampling technique
  • Number of units sampled

Questions to Ponder Upon!

• Post-sampling
  • Use of weights
  • Impact of non-response
  • Standard errors for study estimates
### Census or Sample

- **Census** is “a complete enumeration of the population”

- **Sampling** is “the act, process, or technique of selecting a suitable sample, or a representative part of a population for the purpose of determining parameters or characteristics of the whole population”

- **Goal** is to determine a population’s characteristics by directly observing only a portion of the population.

### Benefits of Sampling:

- **Reduced Cost**
- **Faster results**
- **Increase Precision ➔ Even over Census sometimes!**
Sampling Frame

- A list of all elements in the study population of interest:
  - Names of individuals
  - Telephone numbers
  - House addresses
  - Census tracts

- Target population:
  - From which the sample is drawn
  - To which the sample data will be generalized

Probability or Non-Probability Sample

- Probability/Random Sample:
  - Allows a known probability that each elementary unit will be chosen
  - Type of sampling that is used in lotteries and raffles
  - Used to estimate population parameters with accuracy

- Non-Probability Sample:
  - Convenience sample, judgment sample, snowball sample, quota sample
  - Useful for pilot studies, case studies, qualitative research, and for hypothesis development
Sampling With or Without Replacement

- **Sampling with replacement:**
  - A unit is selected at random from the population and it is returned to the main lot before the second unit is selected
  - The two sample values are independent
  - What we get on the first one doesn’t affect what we get on the second

- **Sampling without replacement:**
  - A unit is selected at random from the population and it is not returned to the main lot
  - The two sample values aren’t independent
  - What we got on the first one affects what we can get for the second one

Some Random Sampling Techniques

- **Sampling Design**
  - Simple
  - Complex (CRS)

- **Number of stages**
  - Single
  - Multiple
**Simple Random Sample (SRS)**

- Each unit in the population has an equal chance of being selected
- Simple but requires a complete listing of the population of interest

**Systematic Sample**

- Proxy for SRS when no list of the population exists or the list is in roughly random order
- Identify the length of sampling interval
- Selecting the first unit on a random basis from an interval
- Selecting additional elementary units at evenly spaced intervals
Some Random Sampling Techniques:

**Systematic Sample**

- Similar to SRS but easier in the field setting
- Need to make sure that the sampling frame is not sorted according to a cycle coinciding with the sampling interval (step size)
- Sampled units are dispersed across the entire geographic spread of the population
- \( k = \frac{N}{n} \) is the step size
  \[ 1 \leq R \leq k \text{ where } R \text{ random number} \]
  \[ \Rightarrow \text{Sample } R, R+k, R+2k, \ldots, R+(n-1)k \]

Some Random Sampling Techniques:

**Stratified Sample**

- Group study population into non-overlapping strata
- Independently selecting a separate simple random sample from each stratum
- Used to:
  - Ensure proportional representation for each stratum
  - Decrease sampling variability
  - Yield sufficient number of a subpopulation in the sample
Some Random Sampling Techniques: *Stratified Sample*

- Distribution of sampling units across strata:
  - Proportional allocation
  - Equal allocation
  - Oversampling from a stratum
  - Optimal allocation
- Homogenous strata ➔ Increase precision

Some Random Sampling Techniques: *Cluster Sample*

- Selecting groupings/clusters from the population on the basis of simple random sampling
- Take a census of units within each selected cluster ➔ reduces travel due to proximity of units in cluster
- Units in cluster are highly correlated ➔ loss of precision
  ➔ Increasing number of clusters increases the precision
Some Random Sampling Techniques:  

*Cluster Sample*

- Used when listing of clusters is available while a list of all population units is not available

![Diagram of Cluster Sample]

Some Random Sampling Techniques:  

*Multi-Stage Sample*

- Simple version: Two-stage sampling
  - Select clusters as Primary Sampling Units (PSU)
  - Select members within the selected clusters as Secondary Sampling Units (SSU)

![Diagram of Multi-Stage Sample]
Some Random Sampling Techniques: 

**Multi-Stage Sample**

- **Advanced Version**
  - Stage 1: Sample counties within region
  - Stage 2: Sample segments
  - Stage 3: Sample neighborhoods
  - Stage 4: Sample households

**Complex Design Effect**

![Graph showing Complex Design Effect](image)

- Serial
- SRS variance

Sample Size
Questions & Answers

Quick Comparison

<table>
<thead>
<tr>
<th></th>
<th>SRS / Systematic</th>
<th>Stratified</th>
<th>Cluster</th>
<th>Multi-stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frame</td>
<td>Listing</td>
<td>Listing with stratifying variable</td>
<td>Listing of clusters</td>
<td>Listing of primary sampling units</td>
</tr>
<tr>
<td>Benefits</td>
<td>Simple Self-weighted</td>
<td>Improve Efficiency Subpopulation analysis</td>
<td>No element listing needed</td>
<td>No element listing needed Reduce travel time Improves efficiency</td>
</tr>
<tr>
<td>Costs</td>
<td>Information to stratify Weighting</td>
<td>Increase sampling error</td>
<td>Increase sampling error</td>
<td></td>
</tr>
</tbody>
</table>
Sample Size

• Sample size can be determined by various constraints:
  
  - Funding/Cost
  - How heterogeneous a universe is sampled
  - Desired precision of the estimate(s)
  - Tolerable error of the estimates (power of analysis)
  - Nature of the analysis to be performed
  - Kind and number of comparisons that will be made
  - Number of variables that have to be examined simultaneously

Sample Size

- **Infinite Population**
  
  \[ n_0 = \frac{Z^2 \sigma_f^2}{e^2} \]

- **Finite Population**
  
  \[ n = n_0 \frac{N}{N + n_0} \]

- E.g. Estimating a proportion, 95% CI \((Z_{1.96}^2 = 1.96)\), 3% margin of error, \(S^2 = 0.25\) \(\Rightarrow n_0 = 1067\)
  
  With \(N = 1251\) then \(n = 576\)

- **Note:** Adjust sample size to account for expected non-response
Efficient Sample Size

- Efficient sample size is *the sample size required to achieve a certain precision of the estimate from the sample (reducing variability)*

- Efficient sample size assumes an SRS
  - Sampling variability
    - increases with cluster sampling
    - Decreases with stratified sampling

Design Effect

- Design Effect ($deff$) is used if the sample deviates from SRS
  \[
  deff = \frac{\text{Variance}(\text{CRS})}{\text{Variance}(\text{SRS})}
  \]

- The $deff$ impacts efficient sample size needed
For stratified samples:
- deff is expected to be <1
- Depends on variability between strata and homogeneity within stratum

For cluster samples:
- deff is expected to be > 1
- Depends on: difference between cluster mean and overall mean, heterogeneity of the clusters, number of clusters selected

Post-Sampling Steps

Use of weights
- Needed when sample not selected with equal probabilities
  \[ \text{weight} = \frac{1}{\text{Probability of Selection}} \]

- To adjust for sampling bias
- Depends on the unit of analysis
Post-Sampling Steps

- To post-stratify
  \[
  \text{weight} = \frac{\text{Population proportion}}{\text{Sample proportion}}
  \]
- Analyze the difference between weighted & un-weighted results.

- Non-Response Evaluation & Adjustment
  - Non-response creates non-sampling bias due to omission from sample
  - Study non-response if feasible
  - Adjust for non-response if deemed adequate

Post-Sampling Steps

- Standard error (SE)
  - needed for descriptive and analytical results
  - SE are measures of variability \( \Rightarrow \) determine precision
  - Requires complex calculations in many cases (beyond scope)
Sampling & Non-sampling Error

Total Survey Error

- Sampling
- Non-sampling/Systematic
- Coverage
- Measurement
- Non-Response

Sampling Error

• Sources:
  • Chance/luck of the draw when choosing a sample
    – Likely impact of sampling error usually quantified using the standard Error (SE)
    – The SE can be estimated using
      – sample design
      – sample data
  • Poor sampling plan
Non Sampling Error:
1. Coverage

- Some members in the population do not have a known non zero chance of being included in the sample
- If members included in the sample are different from excluded
- Survey mode not providing adequate coverage (e.g. Phone, internet)
- Frame error (e.g. incomplete frame, duplications, contaminated)

Non Sampling Error:
2. Measurement

- If respondent answer is inaccurate or imprecise
- Result of questionnaire design and wording
- Mode and instructions/training
- Interviewer effect
Non Sampling Error:
3. Non-Response

• Unit non-response: Not everyone in the sample responds (unreachable, refusal, etc.)
  - Respondents differ from non-respondents
  - Design of implementation system (reach respondents, encourage them to respond)

Non Sampling Error:
3. Non-Response

• Item non-response: Those who respond to the survey do not answer all questions
  - Incomplete data
  - Questionnaire design and interviewer training
References


Questions & Answers
E-mail: Amer-Safa@norc.org

Thank You!

Insight for informed decisions™