QS101: Introduction to Quantitative Methods in Social Science

Week 8: Sampling – Who, What, Where, When

Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression

November 25, 2014

Why Sample?

Reasons to Sample

- Cost
- Time
- Possibility of Generalisation

A Few New Terms

New Terms

- Population
  - The group we want to draw conclusions about
  - This population must be clearly defined (!)
- Sample
  - Sub-group of the population
- Representative Sample
  - Is a sample "in which every major attribute of the larger population from which the sample is drawn is present in roughly the proportion or frequency with which those attributes occur in the larger population." (Brians et al., 2014, p. 139)
  - "a smaller, but accurate model of the larger population"
  - Necessitates awareness of *all* sampling parameters

Sampling Fiascos

- Literary Digest: Presidential Election Polls
- Correctly predicted results in 1920. 1924, 1928 and 1932
- 1936: F.D. Roosevelt versus Alf Landon

Sampling Fiascos (contd.)

- LD poll had Landon winning by 59% to 41%

Sampling Fiascos (contd.)

- LD poll had Landon winning by 59% to 41%
- Roosevelt won with 61%

Sampling Fiascos (contd.)

- LD poll had Landon winning by 59% to 41%
- Roosevelt won with 61%
- Sample Size: 2 million

Sampling Fiascos (contd.)

- LD poll had Landon winning by 59% to 41%
- Roosevelt won with 61%
- Sample Size: 2 million
- Sampling Frame: List of telephone subscribers and car owners

Sampling Fiascos (contd.)

- LD poll had Landon winning by 59% to 41%
- Roosevelt won with 61%
- Sample Size: 2 million
- Sampling Frame: List of telephone subscribers and car owners
- BUT: Depression in 1936

Sampling Fiascos (contd.)

- LD poll had Landon winning by 59% to 41%
- Roosevelt won with 61%
- Sample Size: 2 million
- Sampling Frame: List of telephone subscribers and car owners
- BUT: Depression in 1936
- Roosevelt voters were the economically disadvantaged

Probability Sampling

Probability Sampling

- All potential cases *must have* the same probability of entering the sample
- This allows for generalisability

Random Samples

- "[The] sample must be chosen in such a manner that each and every individual or case in the entire population has an equal opportunity to be selected for analysis."
- "[The] sample must be chosen in such a manner that each and every possible combination of n cases, where n is simply the number of cases in the sample, has an equal opportunity to be selected for analysis." (Brians et al., 2014, pp. 141-142)
- In practice: random number generator

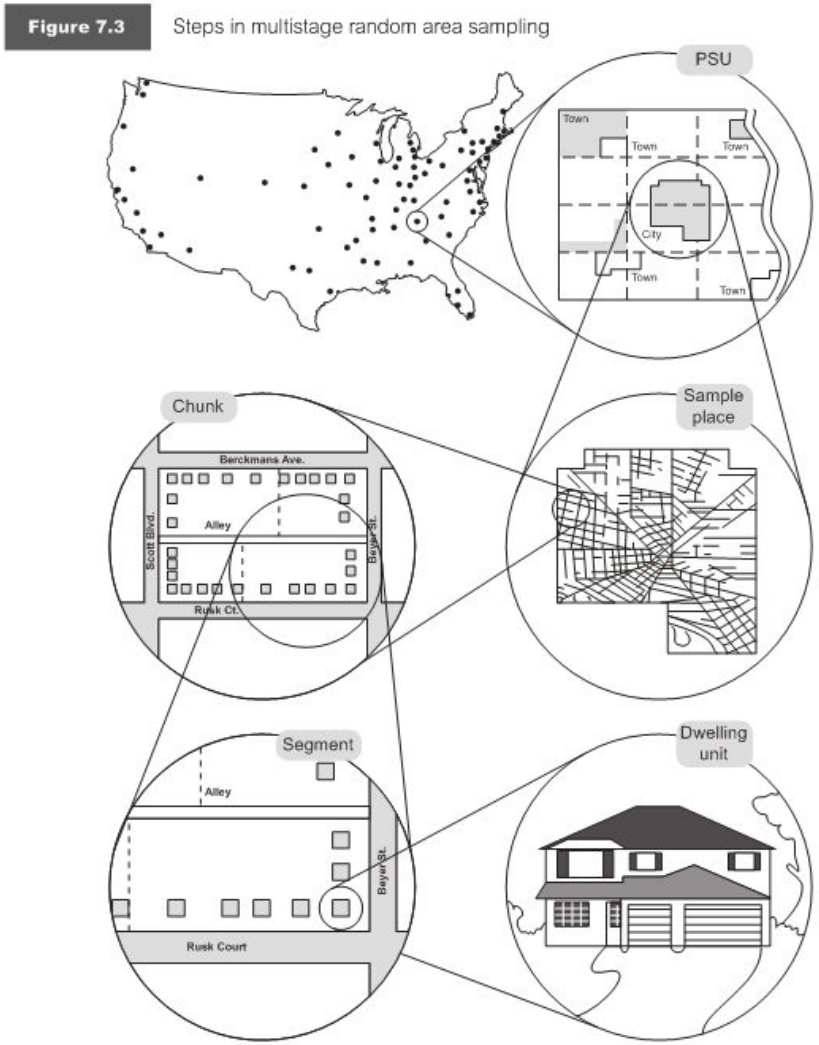
Systematic Random Samples

- Used for quite large populations
- Selection from lists, such as telephone books, student directories, etc.
- Count the number of total cases, and divide by the number desired in the sample (select every k^{th} case)
- Randomised first draw
- Risk of systematic bias
- Sometime practical problems

Cluster Sampling

- AKA multistage random area sampling
- "[Rather] than identifying members of a sample as individuals, we identify them as residents of particular housing units." (Brians et al., 2014, p. 145)
- Units remain fixes, individuals are more mobile, but generalisable to individuals nonetheless

Steps in Cluster Sampling (Brians et al., 2014, p. 147)



Stratified Random Sample

- Used if a sub-group of the population is too small to permit detailed analysis thereof
- For example for ethnic minorities, etc.
- Sample is divided into *two* samples
- Subgroups therefore need to be known in advance
- Employed as a second-order method, as it is not a replacement for random sampling methods (bias!)

Non-Probability Sampling

Non-Probabailty Sampling

- NOT selected randomly
- Greater chance for bias and distortion
- Still used sometimes, as random techniques are often infeasible
- Awareness of potential shortcomings

Convenience Samples

- Use participants that are readily available

Convenience Samples

- Use participants that are readily available
- YOU!

Convenience Samples

- Use participants that are readily available
- YOU!
- Could work for basic psychological or physiological attributes

Volunteer Samples

- Mostly participation on the basis of a material incentive
- Self-selection makes these studies tricky with regards to representative conclusions

Purposive Samples

- Select participants who share a common experience
- For example: Shopping experience in a mall
- Representative for a very specific target population only

Snowball Samples

- For some studies it might be hard to identify subjects
- E.g.: Study on homosexuality in sub-Saharan Africa

Quota Samples

- If samples need to be balanced by a certain characteristic
- E.g.: sex, age, race, etc.
- Individuals displaying the specific traits are selected in proportion to their share in the general population
- Ensures that all combinations are present in the sample

What is the right Sample Size?

Warning!

- THIS is only an INTRODUCTION
- More next week: Sampling Distributions

Homogeneity

- Two extremes
  - Absolute homogeneity: One person is sufficient
  - Absolute heterogeneity: A census of all is necessary
- Reality is somewhere in between
- The more categories we are interested in, the larger the sample must be

Sampling Error

- Sample served to *estimate* the characteristics of the population
- This naturally entails some error
- Therefore, the more accuracy we want, the larger the sample must be

Confidence Interval

- Within how many percentage points above or below the true distribution of the attribute we are interested in is the measurement of the sample
- For example: In a survey 65% of Warwick students like quantitative methods
- With a confidence interval of ± 4 percent, the true percentage of quantsy students would be between 61% and 69%

Confidence Level

- This indicates how confident we are that the sample is in fact representative of the larger population

Confidence Level

- This indicates how confident we are that the sample is in fact representative of the larger population
- Assume we draw 100 samples.

Confidence Level

- This indicates how confident we are that the sample is in fact representative of the larger population
- Assume we draw 100 samples.
- Common level: 95 out of these must be an accurate representation of the true population

Confidence Level

- This indicates how confident we are that the sample is in fact representative of the larger population
- Assume we draw 100 samples.
- Common level: 95 out of these must be an accurate representation of the true population
- Naturally, if we want to go higher (say: 99 samples), the sample size must increase