SO5041 Unit 6: Sampling

Brendan Halpin, Sociology

Autumn 2019/0

Sample not Census

- Most quantitative research proceeds by using samples: rather than ask the entire electorate every month or so who they would vote for, ask a random sample
- If these individuals are chosen at random, their responses will approximate those of the whole electorate
- Random here does not mean completely without control, rather that each individual in the reference population has an equal chance of selection

Simple Random Sampling

- The basic sampling strategy is the simple random sample
  - every subject in the population has the same chance of being selected
  - every possible sample of that size has the same chance of being selected
- How to select a random sample:
  - Acquire a sampling frame: a list of all individuals in the reference population
  - Number the individuals from 1 to N, the size of the population
  - Draw n (sample size) numbers from a random number table or a computer, in the range 1 to N
  - Interview the corresponding individuals

Random numbers

- In the past random number tables were widely used:

Line/Col	(1)	(2)	(3)	(4)
01	96252	48687	59641	99460
02	71334	10218	07459	20339
03	35932	10229	83514	76461
04	33568	36397	92080	98430
05	31535	29990	89596	77529
06	16483	30849	18676	63225
07	26374	60736	14522	66096
08	12040	90130	91860	08280

- Now computers do the job very conveniently: for instance in Excel typing `=rand()` in a cell gives you a random number between 0 and 1.

Probability sampling

- Simple random samples are an example of probability sampling
- Probability sample has the enormously important theoretical advantage of representativity
- If properly conducted, the sample is guaranteed to be *representatitive*
- That is, all sample characteristics *approximate* those of the population – even characteristics you haven't thought of
- Non-probability sampling does not have these advantages but is sometimes used for other reasons

Non-probability sampling

- There are many forms of non-probability sampling:
  - volunteer sampling (self-selection)
  - "streetcorner interview"
  - quota sampling
- Volunteer sampling is popular in the media: questionnaires in magazines, phone-in lines (vote for/against TV programme issue)
- Sometimes the sample is very large
- However, because respondents are self-selecting serious bias can enter

Volunteer sample – dodgy inference!

- Agresti/Finlay quote several examples (pp 20–21): e.g., TV phone vote on "should UN stay in US?" had a response of 186,000 with 67% voting for "No"
- A simultaneous random sample of 500 had 28% voting no: *much* smaller but far more reliable
- TV phone-in subject to bias:
  - Who is watching the program? Time of day, interest, etc.
  - Who is motivated to phone? Those with a strong opinion on the subject
- Twitter polls a particularly bad example of volunteer samples ("retweet for a bigger sample!")

"Opportunistic" samples

- "Streetcorner interview" sampling simply interviews whoever comes along: random in that sense
- However, where and when the interviewing is done will affect who is likely to be interviewed: under-represent workers if during the working day, under-represent country people if done in the city, over-represent shoppers if done in shopping area, etc.{}
- Quota sampling is a form of sampling that avoids these bias problems by using quotas (of age group, sex, employment status, income group, region etc.) – people passing by at "random" are interviewed only if they fit the quota
- Not really probability sampling, but often a very effective and efficient way of "faking" a true random sample, especially for well-understood purposes like opinion polling

Samples are *representative*

- The characteristics of a random sample approximate those of the reference population by virtue of randomisation
- For instance, mean income in a sample (the "*sample statistic*") will be more-or-less close to the mean income for the population from which it is drawn (the "*population parameter*")
- The sample statistic is a more-or-less good estimator of the population parameter
- This is in general true for any statistic: proportion answering yes, distribution of qualifications, average working hours, joint distributions (e.g., crosstabs) etc.{}

Samples are *approximate*

- But because it comes from a sample, the sample statistic will not be an exact estimate of the population parameter
- The sample statistic depends on chance: the exact contents of the sample
- Repeating the exercise with a different sample will give a different sample statistic – still approximating the population parameter, of course
- In principle we can think of there being a distribution of possible sample statistic values, all falling more or less close to the true value of the population parameter

Sampling error

- This difference between the statistic and the true value is called sampling error
- If the true proportion of UL students with part-time jobs is 65% and a sample reports 72% we have a sampling error of $72 - 65 = 7\%$
- How useful a sample is depends on the size of sampling error
  - The bigger the sample the smaller the sampling error
  - The more variability in the population the larger the sampling error
- How do we know how big the sampling error is? We don't know the true value!
- With certain assumptions about the range of the true value and information about the dispersion of the sample we can estimate it – e.g., with a sample of 1,000 sampling error for proportions (%age working etc.) is around 3–4%

Systematic error

- Sampling error is due to sampling variability: always present
- Other sources of error can arise – systematic error is a problem
- Error can arise from sampling problems
  - Undercoverage – e.g., homeless people, geographically mobile people
  - Bad sampling frame – e.g., out-of-date electoral register
- Error can also arise from response problems
  - Outright refusal – "non-response"
  - Refusal to answer certain questions – "item non-response"
  - Bias or deception in answers

Survey design and error

- Sampling error is part of the design
- Non-sampling error is a problem and can invalidate the conclusions we wish to make – bias the sample away from representativity of the reference population
- Undercoverage or non-response introduces the possibility that the sample we get differs systematically from the population (by differing from the part we don't get)
- For instance, if we are interested in financial wellbeing and our survey underrepresents homeless people or people who move very often the people in the sample may be systematically better off than those we fail to trace
- If UL students working part-time are too busy to respond to a questionnaire our sample will under-estimate the average hours spent working part-time

More sampling strategies

- We have discussed several types of sampling strategy
  - Simple random sampling
  - Volunteer sampling and "street-corner interviewing"
  - Quota sampling
- Random sampling is probability sampling; the others are not
- Other forms of probability sampling also exist:
  - Systematic sampling
  - Stratified sampling
  - Cluster sampling
  - Multistage sampling

Systematic sampling

- Systematic sampling is used where the sampling frame has a more-or-less random order
- Pick a name at random near the start of the sampling frame
- Skip k cases, pick the next name, and continue until finished
- k, the size of the skip, is calculated as N/n, so that you arrive at the full size of the desired sample
- This works as a simple random sample, on the assumption that there is no relationship between where you are in the list and the relevant characteristics you have
- If there is any order to the sampling frame (like people being grouped together, for instance, or periodicity) this will not yield a random sample

Stratified sampling

- Stratified sampling works by dividing the population into strata and collecting random samples within the strata
- If the groups or strata are defined according to variables important to the study (e.g., age, gender, occupational group) this method yields statistically more efficient samples (i.e., sampling error is less than with a simple random sample)
- Where we know the population distribution of age or gender (e.g., from a census) our sample will automatically have the right proportions
- This method also allows over-sampling of small groups: ethnic minorities, people on FÁS schemes, the unemployed – this allows comparisons small groups and others that would not be possible in a simple random sample

Cluster sampling

- Cluster sampling divides the population into groups called clusters
- These are often geographic areas, or organisationally based
- A random sample of clusters is drawn
- Within each cluster, simple random samples of individuals are drawn

Cluster sampling – pros and cons

- Advantages:
  - Cost – much less interviewer travel time
  - Can draw better random samples within cluster than from national sampling frame (e.g., identify all households in cluster first, then sample: better than out-of-date list of addresses)
- Disadvantage: statistically less efficient – larger sample needed for same accuracy

Multistage sampling

- Multistage samples use several of these methods in combination
- For instance, using electoral wards as clusters, take a simple random sample of clusters; within each ward, treat streets as clusters and sample again; within each street identify every separate address and do a systematic sample of every nth household

Reading

- A&F Chapter 2
- Bryman Ch 8 (see also Ch 7)