SO5041 Unit 2: Surveys, Questionnaires and Sampling

Brendan Halpin, Sociology

Autumn 2020/1

SO5041 Unit 2

Survey research

Surveys and Survey Research

- Social survey research is very widespread:
  - political opinion polls and market research
  - EU-wide Labour Force Survey & CSO's Quarterly National Household Survey (LFS plus)
  - EU Eurobarometer
  - European Social Survey
  - International Social Survey Programme
  - Household Budget Survey
  - UK Family Expenditure Survey, General Household Survey
  - Growing up in Ireland, TILDA
  - Slán, Irish Study of Sexual Health and Relationships
  - surveys of business opinion, of inventories etc.
  - many emanating from ESRI or government

Surveys: representativity

- The key principle of survey research is representativity: because the sample is random, summaries of the sample's characteristics can be imputed to the relevant population
- Sometimes we end up with too few cases of a subgroup to analyse – e.g., ethnic minorities; over-sampling or specially targetted surveys may help

Longitudinal surveys

- Longitudinal surveys are a special case
  - Panel surveys the same sample at regular intervals (e.g., European Community Household Panel, US Panel Study of Income Dynamics, German Socio-Economic Panel, British 'Understanding Society' Panel Study)
  - Retrospective studies ask respondents to report complete life histories retrospectively (Irish Mobility Study, UK Family and Working Lives Survey, German Life History Study, etc.)
  - Cohort studies take a group of subjects and follow them forward (e.g., the Growing Up in Ireland study, The Irish Longitudinal Study on Ageing)
- Taking time into account makes these in many ways much richer data sources

Questionnaire design 1/3

- The questionnaire is the linchpin of the survey
- Must elicit right information with minimum of ambiguity or suggestion, minimum inconvenience to the respondent
- Question design is a black art, since small changes of phrasing may cause different results

Questionnaire design 2/3

- Extensive reliance on standardised questions, or standardised forms of questions (e.g., the typical five point answer scale: strongly agree, agree, neutral, disagree, strongly disagree)
- Standard schedules exist for certain purposes, e.g., the General Health Questionnaire
  - See e.g., https://www.iser.essex.ac.uk/bhps/documentation/volb/wave18/rindresp15.html (search for "ghq")

Questionnaire design 3/3

- V important to minimise “open” questions: much cheaper to pre-code answers (but allow an “other, specify” answer)
- Very important to test questionnaires in a pilot survey, to trap ambiguities and other problems, and to help pre-code questions

Access

- Lots of survey data is available to the public, or to researchers
- Via data archives (e.g., the Irish Social Science Data Archive: http://www.ucd.ie/issda)
- Via government, EU, organisations like OECD
- Via website like European Social Survey: http://www.europeansocialsurvey.org/

SO5041 Unit 2

Other forms of data

Other forms of data

- Administrative data
- Online data (e.g., Twitter)
- Topic-specific data such as about Covid-19
- "Big data": commercial data, online activity trackers, mobile phone records, Fitbit, traffic records

Admin data

- Huge amounts of relevant administrative data is available
- Not survey: collected as a byproduct of the operation of the state
- Vital statistics: Births, marriages, deaths
- Censuses
- Tax, employment/unemployment, benefits, education, business
- Irish Central Statistics Office puts lots online at https://statbank.cso.ie
- See also OECD, Eurostat et alia.

"Big data" increasingly important

- Matters more and more
- Requires different skills
- Sometimes threatens to replace conventional sources
  - Quicker, cheaper
  - But as accurate??
- But really big problems of representativity

SO5041 Unit 2

Inference

Inference, sampling and statistics

- The basis of survey research is inference: the projection of the characteristics of the sample onto the population
- This requires a random (or quasi-random) sample: each member of the population of interest has an equal chance of being selected
- Consider a simple summary: the mean

$$\bar{x} = \frac{\sum x_i}{n}$$

Calculations on samples

- We could calculate the mean for the entire population but that would be expensive
- We can calculate the mean for a random sample: much cheaper and the answer approximates the true answer in a probabilistic way
- That is, if we were to draw a large number of samples and calculate all the sample means, these sample means would be distributed about the true (population) value with an approximately normal distribution
- In only drawing one sample, we infer that the true mean lies probabilistically around the sample mean, with a large sample giving a tighter distribution that a small one

Error, but maybe not too much

- We can also calculate from the data how wide this probability distribution is: this is the basis for “significance”, a measure of how much the sample summary can be trusted
- It is usually the case that a sample a tiny fraction the size of the population will provide good estimates
- The same reasoning applies to means, frequency distributions, correlation coefficients, cross-tabulations, regression coefficients etc.

Multi-variable techniques

- Multi-variable techniques are the key to causal analysis: take account of the effects of many variables simultaneously to estimate the net effect of each
- Regression analysis is the most commonly used multi-variable technique, and is similar in principle to most multi-variable techniques - It is used with a dependent variable that is continuous (i.e., like age, time, income etc., rather than religion, sex, nationality which are *categorical*)