Chapter 9: Data collection and sampling methods

- Data collection
  - primary data (collect the data yourself)
  - secondary data (use data collected by others)

- Examine secondary sources first
  - paper sources
  - electronic sources (becoming much more prevalent)

Secondary data sources

- Most common way to obtain and use data
- No control over methods of collection, etc.
- May have to compromise on the research questions that can be answered with the data
- Need to be aware of traps in using such data…

Make sure you collect the right data!

- Seems obvious, but…
  - real or nominal data?
  - covers England, GB or UK?
  - use wage rate or earnings?
  - measure wage per hour, per week, per…?
  - full-time or part-time workers?
- These are just a few examples of issues.

Check definitions of variables

- From the ONS Subnational Population Projections:

2.3.1. Internal Migration
For subnational population projection purposes this is defined as migration within England only[5].

5 For other uses internal migration will include flows to and from Wales, Scotland and Northern Ireland, termed cross-border flows. However, the propensity to migrate model used to project internal migration requires a full matrix of flows out of and into each LA by single year of age and sex. This level of detail is not required to project cross border flows and therefore projections add these flows to other international migration flows.

Try to get up to date figures

- Data series are often revised in the light of newer and better information

- UK balance of payments 1970, as published in:

1971	1972	1973	1974	1975	1976	1977	1978	. . .	1986
579	681	692	707	735	733	695	731	. . .	795

Latest estimate is £819m

- Work *backwards* through publications (i.e. most recent first) to ensure you have latest revision of data.

- Economic theory might provide some guidance, e.g. there is often a preference for real values rather than nominal

Keep a record of your data sources

- Got a useful table? Need to update it?
- Then you need to know where you found it originally!!
- Keep a record (comments in spreadsheets are useful).
- Record publication, date, volume, page or table number.
- Keep URL of electronic data source.

Check your data

- Boring but vital. Garbage in, garbage out.
- Drawing graphs of the data can be a useful way to spot errors, outliers, etc.
- Correct errors, think about outliers. Are the latter relevant to your enquiry or just a nuisance, for example?

Electronic data sources

- Can be an easy, quick and cheap way to collect data.
- Usually easy to find *some* data but the precise data you need might be more difficult.
- Usually easiest to download in Excel format.
- Data in pdf format is more tricky, but can be ‘cut and pasted’ using the text selection tool

- Electronic data still needs checking! Even from ‘official’ web sites.
- From ONS on-line - The price of cinema admission:

1963	1964	1965	1966	1967
37.00	40.30	45.30	20.60	21.80

This cannot be right! Why the break in 1966?

Collecting primary data

- You have control over the questions asked and the sample size (hence confidence intervals of resulting estimates)
- Some form of random sampling important to ensure results are not misleading.
- Random sampling: each member of the population should have an equal or known probability of selection

How not to do a survey

- 1936: *Literary Digest* tried to predict the up-coming US election by sending out *10 million* questionnaires. Got the answer wrong…
- Surveyed its own (well-off) readers
- Used lists of car and telephone owners to sample.
- Magazine went out of business soon after.

Types of random sample

- Simple random sample
- Stratified random sample
- Cluster random sample
- Multi-stage sample
- Quota sample

Simple random sample

- Every possible sample has an equal chance of occurring.
- E.g. draw sample from names in a hat
- Can be expensive to carry out (e.g. might have one sample observation from the north of Scotland, one from Cornwall in the west, etc.)

Stratified sampling

- Attempts to avoid ‘bad’ (i.e. unrepresentative) samples.
- E.g. rules out sample with only men. Ensures representative numbers of men and women.
- Gender is here a stratification factor
- Sample would be 50:50 men and women

Example of stratification

- Population consists of 20% older people, 50% middle aged, 30% young.
- Sample should reflect this. Simplest solution would for sample to reflect this, i.e. a sample of 100 people would have:

Class	Old	Middle-aged	Young	Total
Number in sample	20	50	30	100

- Not essential to have proportional sampling (as above). It is better to:
  - sample cheaper strata more heavily (increases overall sample size, given your budget)
  - sample more diverse strata more heavily
- Stratification is most useful when
  - there are differences *between* strata
  - there is little variation *within* strata.

- Stratification improves the precision of the estimates, i.e. reduces the sampling variance of the estimates.
- This results in smaller confidence intervals.

Cluster sampling

- Much cheaper than other methods, hence allows bigger sample size.
- Cluster methodology intrinsically leads to less efficient estimates (bigger confidence intervals, for a *given* sample size) but the larger sample size can offset this.

- Population divided into clusters, e.g. regions of the country
- Only *some* of the clusters sampled. This reduces cost, possibly substantially.
- Clustering method benefits most when
  - clusters are similar to each other
  - there is plenty of variation within each cluster

(Note contrast with stratification)

Multi-stage sampling

- Combines the other types of sampling, e.g.
  - Stage 1: cluster sampling of counties (could be stratified to ensure (e.g.) different regions fairly represented)
  - Stage 2: simple random sample of districts within the selected counties
  - Stage 3: stratified (e.g. according to age) sample of individiuals with selected districts.

Quota sampling

- Non-random but extremely cheap
- E.g. sampling passers-by in the street
- Can make some attempts to ensure it is representative, e.g. ensure similar numbers of men and women; appropriate numbers of different age groups, etc.
- Political polling often like this.

How big should a survey be?

- Depends upon the desired accuracy. Suppose we want the average to be measured to within ± 20, with 95% confidence.
- The 95% CI is given by $\bar{x} \pm 1.96 \times \sqrt{s^2/n}$
- hence $20 = 1.96 \times \sqrt{s^2/n}$

$$\Rightarrow n = \frac{1.96^2 \times s^2}{20^2}$$

- In general, the required sample size is given by:

$$n = \frac{1.96^2 \times s^2}{p^2}$$

where p is the desired accuracy.

- s^2 may be estimated, possibly from earlier or similar surveys.

The sampling frame

- A list from which to choose the sample observations is required. This is called the sampling frame.
- Ideally it should contain the whole population (e.g. the Post Office Address File contains all the addresses in the UK)
- Sampling should be done randomly from this list.

Interview techniques

- Interviewers should not ‘lead’ the subjects towards an answer:
  - “Do you agree with me that …?” would be a bad way to ask the question
- The ordering of questions might matter:
  - Do you know how many were killed by the atomic bomb on Hiroshima?
  - Do you believe in nuclear deterrence?
- Non-response bias is a potential problem.

Example: the Expenditure and Food Survey

- A three stage, rotating, stratified random sample!

- Stage 1:
  - Country divided into 150 strata (groups of local authorities), stratified by geographic area, urban/rural character, prosperity
  - Each quarter of the year, one new authority chosen from each stratum. It remains in the sample for one year, then is replaced.

- Stage 2: four ‘wards’ (smaller administrative areas) in each local authority are selected.
- One ward is used in each of the four quarters.
- Stage 3: 16 addresses chosen at random in each ward.

- The sampling frame is the Postcode Address File.
- Some of these are business, rather than household, addresses so are not used.
- About 60% of households agree to participate in the survey, giving about 6,500 households as the sample size.

- Interviewed households keep a diary of all their expenditures (to avoid a recall bias, where some items are easily remembered, others not)
- Participants are paid a small sum (around £10) for participating.
- Confidence intervals are difficult to calculate because of complicated design. They are wider than with simple random sampling of the same sample size, but the design permits a much larger sample.

Summary

- There are a few useful things to remember about collecting and using data, such as to always, always check your data.
- There are different types of sample design. There is a trade-off between accuracy and cost.
- Important factors in collecting data are the desired accuracy of the results, the sample frame and the interview techniques.