Data - sources, survey errors, sampling methods & properties

BEA140 Quantitative Methods - Module 2

UNIVERSITY *of* TASMANIA

Sources of data

The first question usually asked when seeking out data is "do we have to collect this data ourselves, or has somebody already collected it?".

(i) **primary data source:** original collector of the data; and

(ii) **secondary data source:** a compiler or subsequent user of the data.

Primary sources of data

Types of primary data sources:

(i) **Survey:** interviews, questionnaires - people are asked about their beliefs, behaviours, other characteristics. Skill and planning are required for design and interpretation;

(ii) **Observation:** observing and recording behaviour as it happens. Can be done in a time series or cross-sectional manner. E.g. pedestrian/bicyclist traffic studies and focus groups; and

(iii) **Experiment:** use of experimental and control groups, where at least one (assumed) casual variable is manipulated. Often used in product testing and the use of appropriate experimental designs is important.

The 'garbage in garbage out' (GIGO) concept - the value/usefulness of results heavily relies on the quality of: design, application and analysis, garbage in leads to garbage coming out. Depending on the circumstances, data collection can range from being quite simple through to being very complex. Entire books have been written on topics such as experimental and survey design.

Secondary sources of data

Typically we come across secondary data when it is:

(i) collected and published for purposes other than the problem at hand. E.g. ABS, trade journals, RBA, ANZ job ads series, etc.. Note that data from secondary sources is not necessarily free, a large amount of data can be purchased through business research bureaus and can be quite expensive to purchase, though is still usually cheaper than collecting it ourselves; or

(ii) collected as a by-product of administrative processes. E.g. births, deaths, vehicle registrations, rates etc.).

Note: Searching for secondary sources of data is usually more cost efficient than collecting primary data, so is a good place to start. Secondary sources may provide exactly the information needed, though even if not adequate may assist with defining the problem/population and help to plan primary collection.

Survey errors

Although called survey errors, some of the following sources of error can occur with respect to observational and experimental primary sources of data too:

- Coverage error/selection bias - can occur when a sample is inaccurate or excludes parts of the population. e.g. In telephone surveys we should remember that people with phones are not necessarily representative of the general population (possible exclusions include people who are highly mobile or financially less fortunate);
- Non-response error/bias - when those who do not respond have different views / characteristics to those that do. Non-response follow up can help identify whether such error/bias exists;
- Sampling error - results depend on random composition of sample. Expect (slightly) different results between different samples. Sampling error can be quantified and reduces as sample size increases; and
- Measurement error - measured response varies from true value. Can arise from poor questionnaire design (e.g. ambiguity), interviewer's skill/manner/bias, respondent being unwilling or unable to provide information (prestige (eg. halo effect), privacy etc.), perceived incentives to respond in particular ways, and so on.

Sampling methods

Sampling methods can be divided in to two kinds:

(i) **probability/random sampling:** samples where the laws of probability dictate the composition of the sample (i.e. the elements selected); and

(ii) **non-probability/non-random sampling:** samples that do not make use of a process of randomisation (generally used for indicative purposes only).

Sampling methods can also be classified as to whether they sample from the whole population or a sub-divided population.

Samples from the whole population

- **Convenience/chunk sampling:** easy to observe (non-probability);
- **Judgement sampling:** items believed to be representative are selected (non-probability, prone to selection-bias errors);
- **Simple random sampling:** items selected independently with each element of the population having the same chance of being selected (probability); and
- **Systematic sampling:** first item chosen randomly, and then every kth item thereafter (probability, provided items in the population are ordered randomly).

Samples from a subdivided population

- **Quota sampling:** a number to be selected from each group is specified, then a selection is taken from the population by convenience (non-probability). Quota might be based on age, gender, etc.;
- **Stratified random sampling:** population divided into strata, items randomly selected from each group (probability); and
- **Cluster sampling:** population divided into groups (or clusters), clusters randomly selected and all members of selected clusters sampled (probability).

Sampling in BEA140

- Unless otherwise specified, all sample data referred to in this unit should be assumed to have been obtained through a simple random sampling process;
- The formulas in video lectures for sample statistics (e.g. sample mean, sample standard deviation, etc.) assume that the sample is drawn from the whole population. When a sample is drawn from a subdivided population (e.g. using stratified random sampling, cluster sampling, or more complex sampling schemes) weighted sample statistics are required, the formulas for which are more involved and not covered in this unit.

Random variables

The phenomena/characteristics observed when collecting data are known as **random variables**. I.e. they are variable, have a range of values and are random (in as much as we do not know what value will occur in advance). E.g. result of flipping a coin, car accidents per day, height and income are examples of random variables.

Categorigal/qualitative variables

- **Nominal scale:** possible measurement/data values are classified into distinct categories with no implicit order. E.g. (degree:arts, business, economics, law, science, etc.), (gender; male, female). Arithmetic is limited to counting (frequencies, relative frequencies) to determine mode; and
- **Ordinal scale:** same as nominal except the categories are ordered/ranked. E.g. (mark: NN, PP, CR, DN, HD), family size and GDP. We can do the same statistics as with nominal, plus positional measures (in particular the median).

Numerical/quantitative variables

- **Interval scale:** possible measurement/data values are ordered and the difference between measurements is a meaningful quantity, i.e. equal differences between values represent the same differences in the characteristic, and a value of zero is arbitrary. E.g. temperature (the difference between 4°C and 6°C is the same as between 6°C and 8°C, but 8°C is not *twice as hot* as 8°C). Frequency and positional measures are appropriate, as are mean and standard deviation; and
- **Ratio scale:** same as interval except there is a true zero. E.g. 100kg is twice as heavy as 50kg. All statistical measures can be applied.

Numerical/quantitative variables can be further subdivided into continuous (e.g. time) and discrete (e.g. family size). Rounding rules will often need to be applied with continuous variables. In most business environments, the sub classification of numerical into *interval* and *ratio* scales is usually not particularly relevant - most data can be thought of as being on a ratio scale.

Data types and information content/value

As one moves from nominal to ordinal to numerical data, the information content/richness/value of the data increases, as does the number of things you can do with the data.

For example, consider student results for a unit:

- Outcome - nominal - {successful, unsuccessful};
- Award - ordinal - {NN, PP, CR, DN, HD};
- Mark - numerical - $\{x : 0 \leq x \leq 100\}$.

In effect, the mark tells us more than the award, and the award tells us more than the outcome.

Note: It is not always possible that more than one data type might be appropriate to a situation, often there will be no choice.

Examples of variable types & scales of measurement

Variable	Example value	Type	Scale
country of birth	Australia	categorical	nominal
judo belt	blue	categorical	ordinal
mortgage size	$125,000$	(continuous) numerical	ratio
class size	353	(discrete) numerical	ratio

Legal and/or ethical issues in data collection

There are many legal and/or ethical issues in data collection. Some topical issues include:

- "Sugging", selling under the guise of research (`http://en.wikipedia.org/wiki/Sugging`);
- "Push polling", using research as a vehicle to give a false impression, most commonly employed during political campaigning (`http://en.wikipedia.org/wiki/Push_poll`);
- Privacy issues (`https://www.oaic.gov.au/`);and
- Ownership of data & intellectual property, especially when third parties engage in data collection - who owns the data? who owns the collection methodology? etc..

The above and other legal/ethical issues will undoubtedly pop up in other units.

Examples of how value systems may affect data

- **Suicide rates**. Famous sociologist Emile Durkheim claimed to have 'demonstrated' that suicide rates were lower amongst Catholics than Protestants. However, in what was an otherwise careful study, he made the mistake of assuming that suicides were correctly recorded and classified. In reality, many Catholic suicides were being recorded by sympathetic doctors as 'accidental deaths'. A starting point for further reading is `http://en.wikipedia.org/wiki/Total_social_fact`;
- **Premature births in Australia**. The recorded rate of premature births in Australia has fallen dramatically in Australia over the last 100 years. Many ascribe this to improvements in medicine and standards of living. However the reliability of data is also not constant over time, and as social standards change the types of cases that sympathetic doctors tend to misclassify also change. In particular, often births were recorded as premature (no shame in this) whereas in reality they were full term births following shotgun weddings and conception out of wedlock (not very socially acceptable at the time).

Examples of how value systems may affect data

- **Milikan's oil drop experiment**. 'Cargo Cult Science', written by Richard Feynman (who's a very famous physicist), gives some excellent insight about social influences on so-called objective measurements.

. . . that's it for now!