Week 2 (what we wanted from Friday Week1)

-Elaina Hyde

Review of statistics, examples of file cleaning and dealing with empty spaces, data formats and conversions

Part 1 of review: Additional Statistics

Main Issues

Why you will collect certain data
What data you will collect
From where you will collect it
When you will collect it
How you will collect it
How you will analyse it

Table 4.1 Methodologies associated with the main paradigms

Positivism ⟵	⟶ Interpretivism
Experimental studies	Hermeneutics
Surveys (using primary or secondary data)	Ethnography
Cross-sectional studies	Participative inquiry
Longitudinal studies	Action research
	Case studies
	Grounded theory
	Feminist, gender and ethnicity studies

Qualitative Data

Research data rather than the method of collection that can be described as quantitative (numerical) or qualitative (non-numerical)

Data can also be described as primary data (generated from an original source) or secondary data (existing data)

Qualitative Data

Interpretivists are interested in collecting qualitative data, which they will analyse using interpretive methods

Positivists sometimes collect some qualitative data, which they usually quantify before analysing it using statistical methods

You need to describe and justify your method(s)

Triangulation

Triangulation is the use of multiple sources of data, different research methods and/or more than one researcher to investigate the same phenomena in a study

Term comes from surveying where several reference points are taken to check the location of an object

Can reduce bias in data sources, methods and investigators (see Jick, 1979)

Main types of triangulation (Easterby-Smith, Thorpe and Jackson, 2012)

Triangulation of theories – A theory is taken from one discipline (eg psychology) and used to explain phenoma in another (eg marketing)

Data triangulation – Data are collected at different times or from different sources in the same study

Investigator triangulation – Different researchers independently collect data on the same phenomenon and compare the results

Can lead to greater validity and reliability if all the researchers reach the same conclusions (Denzin, 1978)

Methodological triangulation – More than one method (from the same paradigm) is used to collect and/or analyse the data

Content analysis

Content analysis is 'a method by which selected items of qualitative data are systematically converted to numerical data for analysis' (Collis and Hussey, 2014, p. 166)

Normally a document is examined

Eg Interview transcripts

It can also be used to analyse other forms of communications

Eg Newspapers, broadcasts, audio recordings of interviews and video recordings of non-participant observations and focus groups

Procedure for content analysis

If you have a large amount of secondary data, determine basis for selecting a sample (eg certain sections)

Identify the coding units

Construct a coding frame and record the frequency of occurrence of each code

Positivists analyse the data using statistical methods

Interpretivists prefer to analyse the data for emerging themes and patterns and/or use linguistic analysis to explore the semantics, syntax and context

Coding

Positivists use a priori (predetermined) codes representing the variables you have identified in your theoretical framework

But this implies that these are the only important variables and you may discard relevant research data because your coding scheme does not recognise them

Interpretivists develop codes from the research data

But this is very time consuming and other important factors may be missing from your data

Or combine these approaches by using a priori codes and indentifying additional codes from the data

Main issues in collecting quantitative data

Quantitative data are normally precise, and can be captured at various points in time and in different contexts

In a positivist study this usually leads to results with a high degree of reliability, but the validity may be low

-Reliability refers to the accuracy and precision of the measurement and absence of differences in the results if the research were repeated

-Validity is the extent to which the findings reflect the phenomena under study

Overview of data collection in a positivist study

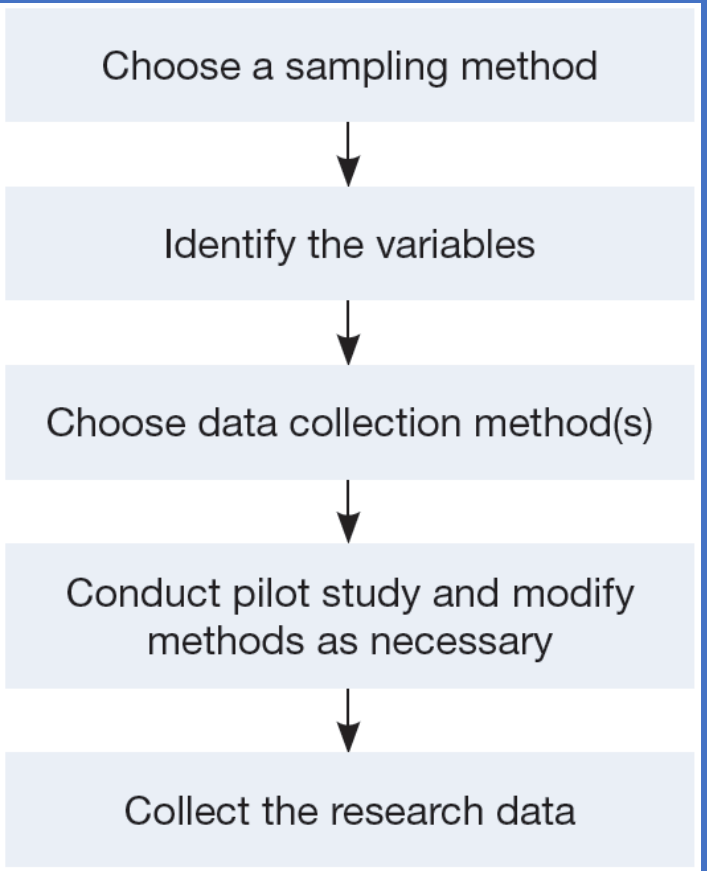

Figure 10.1 Overview of data collection for a positivist study

Selecting a sample in a positivist study

You will need to identify a sampling frame

A sampling frame is a record of the population from which a sample can be drawn

A population is a body of people or collection of items under consideration for statistical purposes

A sample is a subset of a population

Selecting a sample in a positivist study

If you want to generalise results from the sample to the population, you need to select a random sample

A random sample is an unbiased subset of a population that is representative of the population because every member has an equal chance of being selected

Size matters (eg minimum 80 for a population of 100, but 384 for a population of ≥ 1 million)

Importance of selecting a random sample

'In a positivist study, it is vital to obtain a random sample to get some idea of variation... To build general conclusions on ... limited data is a bit like a lazy evolutionary biologist finding a few mutant finches ... in a population on day one of a field outing then returning home to claim that all finches of this species display the same properties' (Alexander, 2006, p. 20)

Sampling methods to avoid bias

To obtain a random sample, allocate a number to every member and use computer-generated random numbers or the random number table on p. 218 to select a sample

For a systematic random sample, divide population by the sample size required (n) and take every n^{th} member

For a stratified random sample, identify the number of members in each stratum (eg number of companies in each industry) and select a random or systematic random sample from each

Importance of theory in a positivist study

Under positivism, research is deductive and you develop a theoretical framework from the literature

A theory is a set of interrelated variables, definitions and propositions about relationships between the variables

You then develop one or more hypotheses that can be tested for association or causality against empirical evidence (your research data)

Each variable is a characteristic of the phenomenon under study that can be observed or measured

Non-numerical observations are quantified by allocating a numerical code

Four levels of precision at which variables are measured…

Measurement levels (in decreasing order of precision)

1. A ratio variable is measured on a mathematical scale with equal intervals and a fixed zero point

Eg Journey time from London to Brussels was 2 hours in 2006, but 1½ hours in 2009 (so 25% faster)

2. An interval variable is measured on a mathematical scale with equal intervals and an arbitrary zero point

Eg It was 5^{o}C yesterday, but 10^{o}C today (so it was warmer by 5^{o}C, but not twice as warm because 0^{o}C does not mean there is no temperature)

As they are measured on a mathematical scale, ratio and interval variables are quantitative variables (Continued)

Measurement levels (continued)

3. An ordinal variable uses numerical codes to identify the order or rank of each category

Eg Order of preference (1st , 2nd, 3rd)

Rating scales (eg where 5 = strongly agree, 4 = agree, 3 = neutral, 2 = disagree and 1 = strongly disagree) can be treated as ordinal or interval variables

4. A nominal variable uses numerical codes to identify named categories

Eg Geographical location where 1 = England, 2 = Wales, 3 = Scotland, 4 = Northern Ireland

As they are not measured on a mathematical scale, ordinal and nominal variables are categorical variables

Other distinctions

A quantitative variable (ie all ratio and interval variables) can be:

A continuous variable where the data can take any value within a given range (eg time = 7 or 7½ hours)

Or a discrete variable where the data can take only one of a range of distinct values (eg Employees = 7 but not 7½)

A dichotomous variable has two groups and can be:

A categorical dichotomous variable with two categories (eg gender might be coded 1 if female and 0 if not)

Or a quantitative dichotomous variable known as a dummy variable (coded 1 if characteristic is present and 0 if not)

Dependent and independent variables

You need to identify the variables you will need to test your hypotheses before you start collecting data

The dependent variable (DV) is the variable whose values are influenced by one or more independent variables (IVs)

Conversely, an independent variable (IV) is a variable that influences the value of the dependent variable (DV)

In some tests it is more appropriate to refer to the DV as the outcome variable and each IV as a predictor variable

Part 2 of review: examples of file cleaning

- .replace(‘/n’, ‘’) #replace line ending with nothing
- assert myvar == ‘int’ #use basic assert to check data type
- Use dict{} to extract relevant data
- .split()
- .append()
- .sort()

Part 3 of review: Dealing with empty spaces

Inputting Missing Data: Playing with Fire

- **Do Nothing:** If you run models on data with systematic missing values, you will get biased estimates.
- **Weight the Data:** Weighting to the national demographic characteristics (such as using ABS data in Australia) is useful if your population of interest is the population of an entire country**.**
- **Inputting Missing Values:** Creating data where it does not exist is something that should never be taken lightly.

- For more see:
- https://www.linkedin.com/pulse/imputing-missing-data-playing-fire-jehan-gonsal

Part 4 of review: data formats

- .txt .dat .csv .xlsx
- Int, float, string, byte, bytearray
- List, dict, tuple

Part 5 of review: conversions

- assert myvar == ‘int’ #use basic assert to check data type
- .replace(‘name’, ‘new_name’) #replace line ending with nothing

- >>> list('Mary') # list of characters in 'Mary'

['M', 'a', 'r', 'y']