Chapter 1: Principles of Research Design

Overview

- ➲ Major considerations at the outset of research
- ➲ Implications of research design
- ➲ Considerations for data analysis
- ➲ Preparation of data prior to analysis

Objectives

This introductory chapter serves to cover the essential stages through which the researcher must pass prior to analysing their data. To do justice to each of these potentially complex issues is beyond the scope of this course, focusing as it does upon data analysis. Therefore, most will be discussed only briefly. However, such brief coverage should not distract from the fact that these issues will have considerable impact upon subsequent data analyses and hence should not be overlooked when planning research.

The main issues covered in this chapter include:

- early consideration of the main research goals, hypotheses and the population to which the research is applicable;
- choosing an appropriate research design;
- ensuring that a suitable sampling method is selected;
- consideration of the analyses to be performed when generating the items that will provide the data.

1.1 Basic Elements of Research

All forms of research can be broken down into a number of discrete components which should be carried out in sequence in order that the specific research goals may be fully realised. Some of these points may seem rather obvious, but it is surprising how often some of the most basic principles are overlooked, potentially resulting in data that is impossible to analyse with any confidence. The main categories may be summarised as follows:

1. Specify exactly the aims and objectives of the research along with the main hypotheses;

2. Define the population;

3. Choose a method of data collection, design the research and decide upon an appropriate sampling strategy;

4. Collect the data;

5. Prepare the data for analysis;

6. Analyse the data;

7. Report the findings.

In this introductory chapter the first four topics will be covered briefly. Although this is primarily a statistical course, you should be aware that statistics and research design are highly interconnected disciplines and you should really have a thorough grasp of both before embarking upon any form of research. It should be pointed out, however, that this introductory chapter merely skims the surface of the issues involved in research design and for more detailed coverage of these topics, you should refer to a research design text.

The first chapters deal with data preparation whilst the remaining chapters cover aspects of data analysis, data presentation or data manipulation.

1.2 Research Aims & Objectives

Surprisingly enough, this is one of the most commonly-overlooked stages of the research process. Quite often, a researcher or team of researchers will be asked to investigate a particular phenomenon, but the specific aims and objectives are not addressed. It could be that those commissioning the research do not know exactly which questions they would like answered. This rather vague approach can be a recipe for disaster and may result in a completely wasted opportunity as the most interesting aspects of the subject matter under investigation could well be missed. One may end up with a huge data file containing information on a large sample of the population, but if you failed to ask the crucial question in the correct way, the entire research will have been a wasted effort.
This approach may require extra work at the very start of the research programme, but the effort will be worth it. For example, you may be asked to ‘find out about attitudes

to drink-driving'. Rather than jumping straight in and designing a questionnaire based around your own knowledge of the subject, a bit of background reading and/or seeking of specialist advice may uncover a more specific hypothesis that is of more interest to your organisation than the vague objective stated above. In this example, you come up with a number of very specific research questions, such as:

"What proportion of people admit to driving whilst above the legal alcohol limit?";

"What demographic factors (e.g. age/sex/social class) are linked with a propensity to drink-driving?";

"Does having a conviction for drink-driving affect attitudes towards driving whilst over the legal limit?";

From these three questions alone, it is possible to recognise some of the items that will need to be included on the questionnaire. Additionally, the research questions will affect the sampling strategy adopted by highlighting the characteristics of people that should be included in the survey. For example, the third question above requires that a proportion of the sample have a conviction for drink-driving and, given that a relatively small proportion of people will have such a conviction, a simple random sample may not yield enough cases for reliable analyses to be performed. In this case, it may be necessary to "disproportionately" sample this particular group of people.

It should be clear, therefore, that it is essential to formally state the main aims and objectives at the outset of the research. The subsequent stages, such as defining the population and designing the research, may then be carried out with these specific questions in mind.

1.3 Defining the Population

The vast majority of research is carried out on **samples** from **populations**. This is because it is usually impractical to obtain information about everyone in a particular population. Imagine how many interviews would need to be conducted to discover the voting intentions of all UK citizens who are eligible to vote. This impracticality is generally not a problem because we can obtain sufficiently precise information from a subset of that population and, providing that the sampling procedure is appropriate, use this information to infer the situation in the population as a whole.

A population can be defined as *"the aggregate of all cases that conform to some designated set of specifications"* (Chein, 1981). In the example described above, our "designated set of specifications" would be 'UK citizens' and those 'eligible to vote'. Hence, we should exclude those under 18 years of age, those not in possession of UK citizenship and those not meeting all other legal requirements for voting eligibility from our sample.

It is important to be aware of the population under investigation as this is the only one to which we can generalise the findings from our analyses. Using the same example, we would only be able to make inferences from our sample about voting intentions of the population of UK voters and not, for example, all UK residents.

1.4 Research Design

With specific research goals and a target population in mind, it is then possible to begin the design stage of the research. There are many things to consider at the design stage and the most important will be discussed in this section.

Principal types of research

The first design consideration involves the type of research that will be most appropriate to the research aims and objectives. The main alternatives are survey research and experimental research. The specific data recording techniques that may be adopted are many and varied, but can be divided into two main categories: **objective** and **subjective**. The former would include things such as physiological measures (e.g. heart-rate) or computer-based techniques (e.g. to record reaction times) whilst the latter may include observational techniques or questionnaire surveys.

You will probably find that the research goals will only lend themselves to one particular form of research, although there are cases where more than one technique may be used. For example, a questionnaire survey would be inappropriate if the aim of the research was to test the effectiveness of different levels of a new drug to relieve high blood pressure. This type of work would be more suited to a tightly-controlled experimental study in which the levels of the drug administered could be carefully controlled and objective measures of blood pressure could be accurately recorded. Not surprisingly, this type of laboratory-based work would not be a suitable means of uncovering people’s voting intentions.

Research designs

The classic experimental design consists of two groups: the **‘experimental group’** and the **‘control group’**. They should be equivalent in all respects other than those in the former group are subjected to an effect or treatment whilst the latter are not. Therefore, any differences between the two groups can be directly attributed to the effect of this treatment. There are many other research designs, but most are more elaborate variations on this basic theme.

In survey research, you will rarely have the opportunity to implement such a rigorously-controlled design. However, the same general principles will apply to many of the analyses you wish to perform, particularly the concept of having a control group against which it is possible to measure the effect of another variable. These treatment variables are usually referred to as **independent** variables, whilst the quantity being measured is the **dependent** variable.

Dependent and independent variables

The best way to think of the distinction is to ask yourself which variable is likely to influence the other? For example, does it make more sense to think of education level influencing age or age influencing education level? Hopefully, you feel that the second option makes most sense (you are as old as you are irrespective of how much education you had) so, in this case, age would be the independent variable and education level the dependent variable.

Generating the items

Having settled upon research aims and objectives, the target population, the type of research and the research design; it is then important to consider the exact nature of the information you will be recording. This involves noting all the variables (i.e. the characteristics) you wish to measure and how best to collect the information. Whatever type of research you are involved with, this vital stage will influence the type of analyses you can perform and, ultimately, whether or not you will be able to provide answers to the main research questions.

With all types of research, it is essential that you obtain information on the full range of factors that may influence the main variables in which you have an interest. This stage usually requires some background reading to establish methods and questions adopted by previous researchers. It is also vital that you pre-test the data collection techniques to ensure that the methodologies adopted work correctly and that the final data set will contain useable information (i.e. respondents answer in an appropriate manner, missing values are minimised etc.). In general, the more preparation and pre-testing you are able to do at this stage, the better. However, the pressure to obtain the ‘answer’ yesterday may mean that this stage cannot be completed in as much detail as would be ideal. Whatever the time constraints, you should always aim to carry out a dry run or ‘pilot’ version of the research to establish that everything is working according to your requirements (e.g. are all the questionnaire items being completed, how long does the questionnaire take to complete etc.).

Questionnaire-based research is particularly vulnerable to lack of due consideration at this design stage, and unfortunate researchers have found that one inappropriately-worded key question has ruined the whole survey.

Levels of measurement

Whenever you gather data, you are collecting information or observing some phenomena. The term ‘levels of measurement’ refers to the assignment of codes or numbers to particular observations. Many statistical techniques are only appropriate for data measured at particular levels, or combinations of levels and if you wish to conduct certain tests, it is important that you have measured the information in a manner appropriate for those tests. Therefore, wherever possible, you should aim to determine the analyses you will be using before deciding upon the level of measurement for each of your variables. For example, if you wish to establish the mean age of your sample, you will need to ask their age directly rather than asking them to choose an age range into which their age falls.

There are four levels of measurement (which combine into two main types: categorical and continuous) distinguishable according to the properties they are said to possess. Each successive one can be said to contain the properties of the preceding types and to record information at a higher level.

Categorical Data:

Nominal Variables:

In which numbers are used simply to distinguish between different properties. Hence, if each data value is only a group identifier or label, then the variable is categorical and at the nominal level or categorical variable. An example of a nominal level variable would be marital status or the type of organisation worked in. Each category represents a choice with no meaningful order to the list and, effectively, the numerical coding system is arbitrary.

Ordinal Variables:

In which the numerical values serve to place categories in some meaningful order. Therefore, variables containing an ordered set of responses, or ranks, are classified as ordinal. For example, you may ask about an opinion on a particular issue in which a respondent offers a code of 1 (‘strongly agree’) indicating that they agree more than a respondent who chooses the category 2 (‘agree’). Therefore, although still a nominal code, there is an implicit order in the coding of the categories.

Continuous Data:

Interval Variables:

The name ‘interval’ derives from the property that the ‘distance’ between adjacent points is the same throughout the scale, unlike ordinal scales. For example, with a variable such as age in years, the difference between 20 and 21 (1 unit, i.e. ‘year’) is equal to the difference between 45 and 46. In other words, they have equal intervals between points on the scale.

Ratio Variables:

Ratio data have all the properties possessed by interval data. However, on a ratio scale, the zero point represents a complete absence of the property being measured. If a coded value of 0 means ‘nothing there’, then ratios of numbers (e.g. twice as great etc.) are interpretable. For example, temperature measured in degrees Celsius is measured on an interval scale, but the zero point does not represent an absence of the quality ‘temperature’. This variable is therefore said to be measured at the interval level. However, a variable such as number of visits to the theatre per year is ratio data as ‘0’ indicates no visits (i.e. an absence of the quality ‘going to the theatre’) and ratios are now calculable (e.g. ‘4’ visits to the theatre represents twice as many visits as ‘2’).

NB: It can be argued that dichotomous variables, containing two possible responses (often coded 0 and 1), fall into all bar the ratio category. This flexibility allows them to be used in a wide range of statistical procedures.

This distinction between the four types is summarised in Figure 1.1.

Level of Measurement	***Property***			
	Categories	Ranks	Equal Intervals	True¶ Zero Point
Nominal	✓	*x*	*x*	*x*
Ordinal	✓	✓	*x*	*x*
Interval	✓	✓	✓	*x*
Ratio	✓	✓	✓	✓

Figure 1.1 Properties held by each Level of Measurement

With some variables (e.g. gender or hair colour), you have no choice regarding the level of measurement. Where there is a choice, you should design the questions (or other data recording techniques) to correspond with the analyses you have in mind.

Different researchers have differing opinions about phrasing of particular questions. The classic example concerning age and whether it should be asked directly (ratio level) or whether the respondent should be presented with the chance of choosing an age range (ordinal level). Some people feel that the former method may result in a large number of missing cases and therefore favour the grouping method. The problem with this is that, by reducing the level of measurement to ordinal, you lose information and the number of statistical procedures in which this variable may be included becomes more restricted. As a general rule of thumb, you should aim to record information at the highest level possible as this will enable you to perform a wider range of analyses. Only if you have good reason to suspect that this will be counter-productive should you select a 'lower' level of measurement. An additional consideration is that it is possible to change from a higher to a lower level of measurement (e.g. from a ratio to an ordinal level) but not from a lower to higher. Therefore, the higher the level, the wider the scope for analyses.

Once you have established the level of measurement for each of your variables, the next stage is to draw up a coding scheme that will determine which values will be used to code each response. This will ensure that the coding of responses remains consistent throughout the data entry stage.

Sampling

As previously noted, in most cases it will not be possible to obtain data from all members of a specified population and therefore some kind of selection process, or **sampling**, must be performed. If sampling is done with care, the results from that sample should reflect closely those that would be obtained from the population as a whole. With a perfectly-representative sample, it should begin to resemble the population after data from a relatively small number of cases has been obtained. There are two basic forms of sampling: **probability** and **non-probability**.

Probability Sampling

Probability, or random, sampling gives all members of the population an equal chance of being included in the sample and does not depend upon previous events in the selection process. For example, when tossing a coin, heads and tails are equally likely to occur (or be selected) on any single coin toss and this is independent of the outcome of previous coin tosses.

Many statistical techniques assume that the sample was selected on a random basis and this assumption is essential if you wish to infer something about the population from your sample statistics. It also serves to check for unconscious biases. In practice, many potential biases may creep into the design and you will probably never know whether a design was completely bias-free. The key is to ensure that biases that do occur are not systematic (i.e. in a certain direction). By randomly sampling, it is possible to assume that non-systematic biases (or random errors) are normally distributed such that the extreme values at one end of the spectrum 'cancel out' those at the other end.

There are various forms of probability sampling techniques, the most commonly-used being: simple random; systematic; stratified; and clustering. However, a detailed discussion of these is beyond the scope of this course.

Non-Probability Sampling

Non-probability sampling procedures are much less desirable as they will almost certainly contain sampling biases. However, in some circumstances, use of such methods are unavoidable. The classic example is in laboratory-based research, which is entirely dependent upon people **volunteering** to take part. This is known as 'volunteer bias' as it is possible that volunteers will differ in some systematic way from non-volunteers (e.g. they may be more extroverted, have an interest in the research etc.). If an experiment is over-subscribed, it is possible to randomly select from the volunteers, but in practice the reverse is usually true and researchers must accept all who volunteer.

Quota sampling is a form of non-probability sampling that is frequently used. This is where quotas within subgroups are set beforehand, usually to match known population distributions, and the interviewers then select people to meet these criteria. This does have the advantage that notoriously unwilling responders (e.g. young males) are represented in the sample. Problems include the judgmental element to selection of individuals and reliance upon the accuracy of the original data source upon which the quotas were calculated. The importance of sampling cannot be underestimated, as it determines to whom the results of your research will be applicable. All too often, researchers succumb to the temptation to generalise their results to a much broader range of people than those from whom the data was originally gathered. This is poor practice and you should always aim to adopt an appropriate sampling strategy, preferably using a random sample.

Problems of Missing Data

Missing data is a generally unavoidable aspect of data collection, but care must be taken over the nature of these missing values. Providing there is no systematic way in which these missing values occur, there should not be a problem. However, when the values are missing in a way that is not random (e.g. if a high proportion of males have failed to answer a series of attitude items) it may be inappropriate to draw conclusions about your population from these findings. It is therefore good practice to try and uncover any systematic biases within your data.

1.5 Data collection

Data collection techniques are also many and varied and, to a large extent, the one you choose will be dependent upon practical considerations. Each has its own advantages and disadvantages. For example, a postal questionnaire survey will be more time efficient than an interviewer-administered survey. However, with the former technique it is not possible to be certain who completed the questionnaire and hence answers to knowledge-based questions cannot be relied upon.

It could be that the data set to answer your specific research questions already exists. Naturally, this will save you an enormous amount of time, effort and cost, but you should ensure that it is fully appropriate to your research requirements.

1.6 Preparing the data for analysis

Having collected the data, you need to prepare it for analysis. Thus using the coding scheme (referred to in section 3) for each of the variables and then systematically entering the data into an appropriate software package. SPSS Data Entry, SPSS for Windows, Excel, Lotus 1-2-3 and ASCII text files are some of the more common methods of doing this. In the next few chapters, therefore, you will be introduced to the fundamentals of getting data into SPSS for Windows.

Summary

In this introductory chapter we looked at

- Specifying Research Objectives
- Defining the Population
- Selecting a Sampling Strategy
- Levels of Measurement
- Collecting Data