Pioneering Futures Since 1898

UEL-DS-7006 Quantitative Data Analysis (QDA)

WEEK 1- Reading Material

Quantitative Research Processes; Relationship with Qualitative Research; Mixed Mode Approaches.

Table of Contents

Learning Objectives	1
Quantitative Data Analysis and the Research Process	1
Theory	2
Operationalization of Concepts	4
Selection of respondents or subjects	5
Setting up a research design	6
Collect data	7
Analyze data	7
Findings	8
Causality and research design	10
Establishing causality	11
Causality and experimental designs	13
Survey design and causality	19
Conclusion	23
References	24

University of East London

Pioneering Futures Since 1898

UNICAF

UEL-DS-7006 Quantitative Data Analysis (QDA)

WEEK 1- Reading Material

Quantitative Research Processes; Relationship with Qualitative Research; Mixed Mode Approaches.

Learning Objectives

By the end of this week, you will be able to;

- Be aware of how to deal with quantitative data
- Understand theories, hypothesis and operationalization in social sciences
- Demonstrate understanding of selection of respondents or subjects in survey investigation.
- Know how to set up a research design
- Understand how to collect and analyze data
- Be able to report research findings

Quantitative Data Analysis and the Research Process

In this section, the way in which quantitative data analysis fits into the research process – specifically the process of quantitative research – will be explored. As we will see, the area covered by this module does not solely address the question of how to deal with quantitative data, since it is also concerned with other aspects of the research process that impinge on data analysis.

Figure 1.1 provides an illustration of the chief steps in the process of quantitative research. Although there are grounds for doubting whether research always conforms to a neat linear sequence (Bryman, 1988a, 1988b), the components depicted in Figure 1.1 provide a useful model. The following stages are delineated by the model.

Theory

The starting point for the process is a theoretical domain. Theories in the social sciences can vary between abstract general approaches (such as functionalism) and low-level theories to explain specific phenomena (such as voting behavior, delinquency, aggressiveness). By and large, the theories that are most likely to receive direct empirical attention are those which are at a fairly low level of generality. Merton (1967) referred to these as theories of the middle range, to denote theories that stood between general, abstract theories and empirical findings. Thus, Hirschi (1969), for example, formulated a ‘control theory’ of juvenile delinquency which proposes that delinquent acts are more likely to occur when the child’s bonds to society are breached. This theory in large part derived from other theories and also from research findings relating to juvenile delinquency.

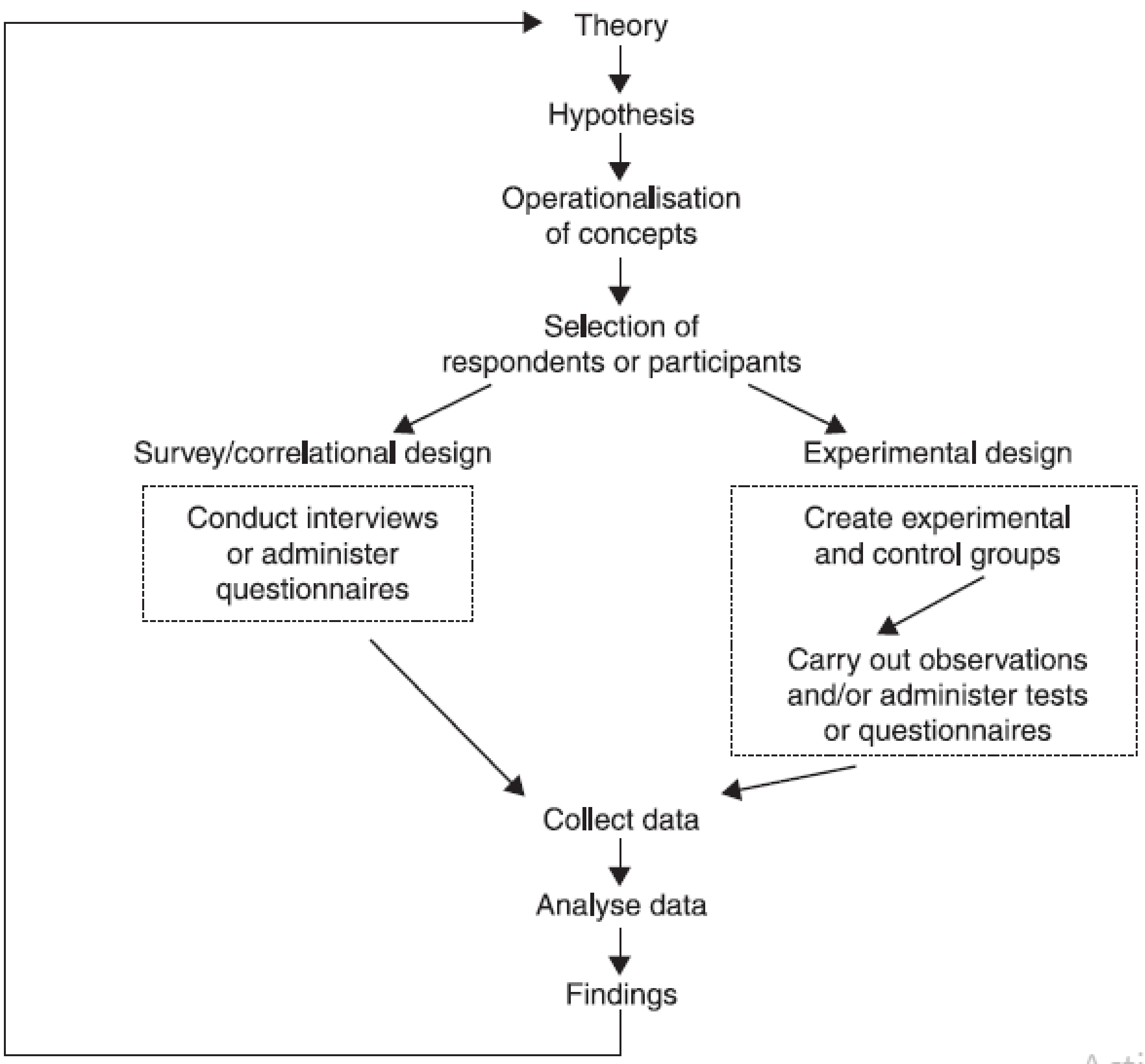

Figure 1.1 The research process (Bryman and Cramer, 2008, p.31)

Hypothesis

Once a theory has been formulated, it is likely that researchers will want to test it. Does the theory hold water when faced with empirical evidence? However, it is rarely possible to test a theory as such. Instead, we are more likely to find that a hypothesis, which relates to a limited facet of the theory, will be deduced from the theory and submitted to a searching enquiry. For example, Hirschi, drawing upon his control theory, stipulates that children who are tied to conventional society (in the sense of adhering to conventional values and participating or aspiring to participate in conventional values) will be less likely to commit delinquent acts than those not so tied. Hypotheses very often take the form of relationships between two or more entities – in this case commitment to conventional society and juvenile delinquency. These 'entities' are usually referred to as 'concepts', that is, categories in which are stored our ideas and observations about common elements in the world. Although hypotheses have the advantage that they force researchers to think systematically about what they want to study and to structure their research plans accordingly, they exhibit a potential disadvantage in that they may divert a researcher's attention too far away from other interesting facets of the data he or she has amassed.

Operationalization of Concepts

In order to assess the validity of a hypothesis it is necessary to develop measures of the constituent concepts. This process is often referred to as *operationalization*, following expositions of the measurement process in physics (Bridgman, 1927). In effect, what is happening here is the translation of the concepts into variables, that is, attributes on which relevant objects (individuals, firms, nations, or whatever) differ. Hirschi operationalized the idea of commitment to conventional society in several ways. One route was through a question on a questionnaire asking the children to whom it was to be administered whether they liked school. Delinquency was measured in one of two ways, of which one was to ask about the number of delinquent acts to which children admitted (i.e. self-reported delinquent acts). In much experimental research in psychology, the measurement of concepts is achieved through the observation of people, rather than through the administration of questionnaires. For example, if the researcher is interested in aggression, a laboratory situation may be set up in which variations in aggressive behavior are observed. Another way in which concepts may be operationalized is through the analysis of existing statistics, of which Durkheim's (1952 [1898]) classic analysis of suicide rates is an example.

Selection of respondents or subjects

If a survey investigation is being undertaken, the researcher must find relevant people to whom the research instrument that has been devised (e.g. self-administered questionnaire, interview schedule) should be administered.

Hirschi, for example, randomly selected over 5,500 school children from an area in California. The fact of random selection is important here because it reflects a commitment to the production of findings that can be generalized beyond the confines of those who participate in a study. It is rarely possible to contact all units in a population, so that a *sample* invariably must be selected.

In order to be able to generalize to a wider population, a *representative sample*, such as one that can be achieved through random sampling, will be required. Moreover, many of the statistical techniques to be covered in this module are *inferential statistics*, which allow the researcher to demonstrate the probability that the results deriving from a sample are likely to be found in the population from which the sample was taken, but only if a random sample has been selected.

Setting up a research design

There are two basic types of research design that are employed by psychologists and sociologists. The former tends to use *experimental* designs in which the researcher actively manipulates aspects of a setting, either in the laboratory or in a field situation, and observes the effects of that manipulation on experimental subjects. There must also be a 'control group' which acts as a point of comparison with the group of subjects who receive the experimental manipulation. With a *survey/correlational* design, the researcher does not manipulate any of the variables of interest and data relating to all variables are collected simultaneously. The term *correlation* also refers to a technique for analyzing relationships between variables but is used in the present context to denote a type of research design. The researcher does not always have a choice regarding which of the two designs can be adopted. For example, Hirschi could not *make* some children committed to school and others less committed and observe the effects on their propensity to commit delinquent acts. Some variables, like most of those studied by sociologists, are not capable of manipulation. However, there are areas of research in which topics and hypotheses are addressed with both types of research design (e.g. the study of the effects of participation at work on job satisfaction and performance – see Bryman, 1986; Locke and Schweiger, 1979). It should be noted that in most cases, therefore, the nature of the research design – whether experimental or survey/correlational – is known at the outset of the sequence signified by Figure 1.1, so that research design characteristics permeate and inform several stages of the research process. The nature of the research design has implications for the kinds of statistical manipulation that can be performed on the resulting data. The differences between the two designs are given greater attention in the next section.

Collect data

The researcher collects data at this stage, by interview, questionnaire, observation, or whatever. The technicalities of the issues pertinent to this stage are not usually associated with a book such as this. Readers should consult a textbook concerned with social and psychological research methods if they are unfamiliar with the relevant issues.

Analyze data

This stage connects very directly with the material covered in this module. At a minimum, the researcher is likely to want to describe his or her subjects in terms of the variables deriving from the study. For example, the researcher might be interested in the proportion of children who claim to have committed no, just one, or two or more delinquent acts. However, the analysis of a single variable is unlikely to suffice, and the researcher will probably be interested in the connection between that variable and each of a number of other variables, that is, *bivariate analysis*. The examination of connections among variables can take either of two forms. A researcher who has conducted an experiment may be interested in the extent to which experimental and control groups differ in some respect. For example, the researcher might be interested in examining whether watching violent films increases aggressiveness. The experimental group (which watches the violent films) and the control group (which does not) can then be compared to see how far they differ. The researcher may be interested in relationships between variables – are two variables connected with each other so that they tend to vary together? For example, Hirschi (1969: 121) presents a table which shows how liking school and self-reported delinquent acts are interconnected. He found that whereas only 9 per cent of children who say they like school have committed two or more delinquent acts, 49 per cent of those who say they dislike school have committed as many delinquent acts. Very often the researcher will be interested in exploring connections among more than two variables, that is, *multivariate analysis*. The distinction between studying differences and studying relationships is not always clear-cut. We might find that boys are more likely than girls to commit delinquent acts. This finding could be taken to mean that boys and girls differ in terms of propensity to engage in delinquent acts or that there is a relationship between gender and delinquency.

Findings

If the analysis of data suggests that a hypothesis is confirmed, this result can be fed back into the theory that prompted it. Future researchers can then concern themselves either with seeking to replicate the finding or with other ramifications of the theory. However, the refutation of a hypothesis can be just as important in that it may suggest that the theory is faulty or at the very least in need of revision. Sometimes, the hypothesis may be confirmed in some respects only. For example, a multivariate analysis may suggest that a relationship between two variables pertains only to some members of a sample, but not others (e.g. women but not men, or younger but not older people). Such a finding will require a reformulation of the theory. Not all findings will necessarily relate directly to a hypothesis. With a social survey, for example, the researcher may collect data on topics whose relevance only becomes evident at a later juncture.

As suggested above, the sequence depicted in Figure 1.1 constitutes a model of the research process, which may not always be reproduced. None the less, it does serve to pinpoint the importance to the process of quantitative research of developing measures of concepts and the thorough analysis of subsequent data. One point that was not mentioned in the discussion is the *form* that the hypotheses and findings tend to assume. One of the main aims of much quantitative research in the social sciences is the demonstration of *causality* – that one variable has an impact upon another. The terms *independent variable* and *dependent variable* are often employed in this context. The former denotes a variable that has an impact upon the dependent variable. The latter, in other words, is deemed to be an effect of the independent variable. This causal imagery is widespread in the social sciences and a major role of multivariate analysis is the elucidation of such causal relationships (Bryman, 1988a). The ease with which a researcher can establish cause and effect relationships are strongly affected by the nature of the research design and it is to this topic that we shall now turn.

Causality and research design

As suggested in the last paragraph, one of the chief preoccupations among quantitative researchers is to establish causality. This preoccupation in large part derives from a concern to establish findings similar to those of the natural sciences, which often take a causal form. Moreover, findings which establish cause and effect can have considerable practical importance: if we know that one thing affects another, we can manipulate the cause to produce an effect. In much the same way that our knowledge that smoking may cause a number of illnesses, such as lung cancer and heart disease, the social scientist is able to provide potentially practical information by demonstrating causal relationships in appropriate settings. To say that something causes something else is not to suggest that the dependent variable (the effect) is totally influenced by the independent variable (the cause). You do not necessarily contract a disease if you smoke and many of the diseases contracted by people who smoke afflict those who never smoke. 'Cause' here should be taken to mean that variation in the dependent variable is affected by variation in the independent variable. Those who smoke a lot are more likely than those who smoke less, who in turn are more likely than those who do not smoke at all, to contract a variety of diseases that are associated with smoking. Similarly, if we find that watching violence on television induces aggressive behavior, we are not saying that only people who watch televised violence will behave aggressively, nor that only those people who behave aggressively watch violent television programmers. Causal relationships are invariably about the likelihood of an effect occurring in the light of levels of the cause: aggressive behavior may be more likely to occur when a lot of television violence is watched and people who watch relatively little television violence may be less likely to behave aggressively.

Establishing causality

In order to establish a causal relationship, three criteria have to be fulfilled. First, it is necessary to establish that there is an apparent relationship between two variables. This means that it is necessary to demonstrate that the distribution of values of one variable corresponds to the distribution of values of another variable. Table 1.1 provides information for ten children on the number of aggressive acts they exhibit when they play in two groups of five for two hours per group. The point to note is that there is a relationship between the two variables in that the distribution of values for number of aggressive acts coincides with the distribution for the amount of televised violence watched – children who watch more violence exhibit more aggression than those who watch little violence. The relationship is not perfect: three pairs of children – 3 and 4, 6 and 7 and 9 and 10 – record the same number of aggressive acts, even though they watch different amounts of television violence. Moreover, 8 exhibits more aggression than 6 or 7, even though the latter watch more violence. None the less, a clear pattern is evident which suggests that there is a relationship between the two variables. Second, it is necessary to demonstrate that the relationship is *nonspurious*. A spurious relationship occurs when there is not a 'true' relationship between two variables that appear to be connected. The variation exhibited by each variable is affected by a common variable. Imagine that the first five

TABLE 1.1 Data on television violence and aggression

Child	*Number of hours of violence watched on television per week*	*Number of aggressive acts recorded*
1	9.50	9
2	9.25	8
3	8.75	7
4	8.25	7
5	8.00	6
6	5.50	4
7	5.25	4
8	4.75	5
9	4.50	3
10	4.00	3

children are boys and the second five are girls. This would suggest that gender has a considerable impact on both variables. Boys are more likely than girls both to watch more television violence *and* to exhibit greater aggressiveness. There is still a slight tendency for watching more violence and aggression to be related for both boys and girls, but these tendencies are far less pronounced than for the ten children as a whole. In other words, gender affects each of the two variables. It is because boys are much more likely than girls both to watch more television violence and to behave aggressively that Figure 1.2 illustrates the nature of such a spurious relationship.

Third, it is necessary to establish that the cause precedes the effect, that is, the *time order* of the two related variables. In other words, we must establish that aggression is a consequence of watching televised violence and not the other way around. An effect simply cannot come before a cause. This may seem an extremely obvious criterion that is easy to demonstrate, but as we will see, it constitutes a very considerable problem for non-experimental research designs.

Causality and experimental designs

A research design provides the basic structure within which an investigation takes place. While a number of different designs can be found, a basic distinction is that between experimental and non-experimental research designs of which the survey/correlational is the most prominent. In an experiment, the elucidation of cause and effect is an explicit feature of the framework. The term *internal validity* is often employed as an attribute of research and indicates whether the causal findings deriving from an investigation are relatively unequivocal. An internally valid study is one which provides firm

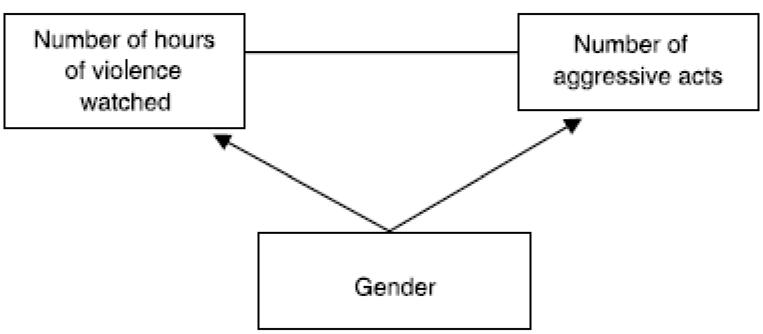

Figure 1.2 A spurious relationship (Bryman and Cramer, 2008, p.38)

evidence of cause and effect. Experimental designs are especially strong in respect of internal validity; this attribute is scarcely surprising in view of the fact that they have been developed specifically in order to generate findings which indicate cause and effect.

Imagine that we wanted to establish that watching violence on television enhances aggression in children, we might conceive of the following study. We bring together a group of ten children. They can interact and play for two hours, during which the number of aggressive acts committed by each child is recorded by observers, and the children are then exposed to a television programmed with a great deal of violence. Such exposure is often called the experimental treatment. They are then allowed a further two-hour period of play and interaction. Aggressive behavior is recorded in exactly the same way. What we have here is a sequence which runs

Obs1 Exp Obs2 where Obs1 is the initial measurement of aggressive behaviour (often called the *pre-test*), Exp is the experimental treatment which allows the independent variable to be introduced, and Obs2 is the subsequent measurement of aggression (often called the *post-test*). Let us say that Obs2 is 30 per cent higher than Obs1, suggesting that aggressive behaviour has increased substantially. Does this mean that we can say that the increase in aggression was caused by the violence? We cannot make such an attribution because there are alternative explanations of the presumed causal connection. The children may well have become more aggressive over time simply as a consequence of being together and becoming irritated by each other. The researchers may not have given the children enough food or drink and this may have contributed to their bad humor. There is even the possibility that different observers were used for the pre- and post-tests who used different criteria of aggressiveness. So long as we cannot discount these alternative explanations, a definitive conclusion about causation cannot be proffered.

Anyone familiar with the natural sciences will know that an important facet of a properly conducted experiment is that it is controlled so that potentially contaminating factors are minimized. In order to control the contaminating factors that have been mentioned (and therefore to allow the alternative explanations to be rejected), a *control group* is required. This group has exactly the same cluster of experiences as the group which receives the first treatment – known as the *experimental group* – but it does not receive the experimental treatment. In the context of our imaginary television study, we now have two groups of children who are exposed to exactly the same conditions, except that one group watches the violent films (the experimental group) and the second group has no experimental treatment (the control group). This design is illustrated in Figure 1.3. The two groups' experiences have to be as similar as possible, so that only the experimental group's exposure to the experimental treatment distinguishes them. It is also necessary to ensure that the members of the two groups are as similar as possible. This is achieved by taking a sample of children and *randomly assigning* them to either the experimental or the control group. If random assignment is not carried out, there is always the possibility that differences between the two groups can be attributed to divergent personal or other characteristics. For example, there may be more boys than girls in one group, or differences in the ethnic composition of the two groups. Such differences in personal or background characteristics would mean that the ensuing findings could not be validly attributed to the independent variable, and to that factor alone.

Let us say that the difference between Obs1 and Obs2 is 30 per cent and between Obs3 and Obs4 is 28 per cent. If this were the case, we would conclude that the difference between the two groups is so small that it appears that the experimental treatment (Exp) has made no difference to the increase in aggression; in other words, aggression in the experimental group would probably have increased anyway. The frustration of being together too long or insufficient food or drink or some other factor probably accounts for the Obs2–Obs1 difference. However, if the difference between Obs3 and Obs4 was only 3 per cent, we would be much more prepared to say that watching violence has increased aggression in the experimental group. It would suggest that around 27 per cent of the increase in aggressive behavior in the experimental group (i.e. $30 - 3$) can be attributed to the experimental treatment.

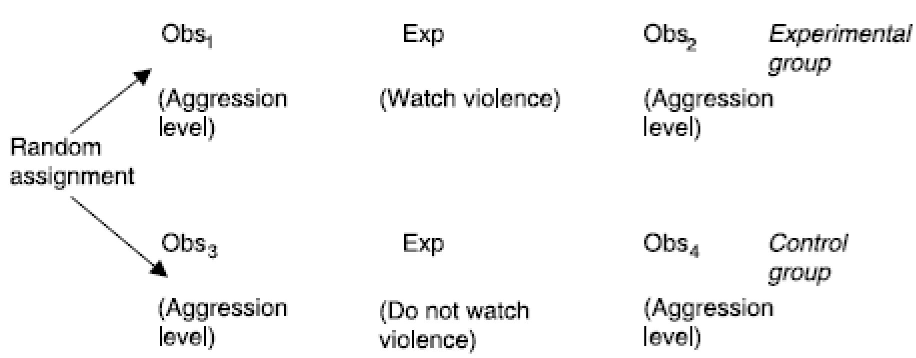

Figure 1.3 An experiment (Bryman and Crammer, 2008, p.40)

Differences between experimental and control groups are not usually as clear-cut as in this illustration, since often the difference between the groups is fairly small. Statistical tests are necessary in this context to determine the probability of obtaining such a difference by chance.

In this imaginary investigation, the three criteria of causality are met, and therefore if we did find that the increase in the dependent variable was considerably greater for the experimental group than the control group we could have considerable confidence in saying that watching television violence caused greater aggression. First, a relationship is established by demonstrating that subjects watching television violence exhibited greater aggression than those who did not. Second, the combination of a control group and random assignment allows the possibility of the relationship being spurious to be eliminated, since other factors which may impinge on the two variables would apply equally to the two groups. Third, the time order of the variables is demonstrated by the increase in aggressive behavior succeeding the experimental group's exposure to the television violence. Precisely because the independent variable is manipulated by the researcher, time order can be easily demonstrated, since the effects of the manipulation can be directly gauged. Thus, we could say confidently that Watching television violence → Aggressive behavior since the investigation exhibits a high degree of internal validity. There is a variety of different types of experimental design. These are briefly summarized in Figure 1.4. In the first design, there is no pre-test, just a comparison between the experimental and control groups in terms of the dependent variable. With the second design, there is a number of groups.

This is a frequent occurrence in the social sciences where one is more likely to be interested in different levels or types of the independent variable rather than simply its presence or absence. Thus, in the television violence context, we could envisage four groups consisting of different degrees of violence. The third design, a *factorial* design, occurs where the researcher is interested in the effects of more than one independent variable on the dependent variable.

The researcher might be interested in whether the presence of adults in close proximity reduces children's propensity to behave aggressively. We might then have four possible combinations deriving from the manipulation of each of the two independent variables. For example, Exp1+A would mean a combination of watching violence and adults in close proximity; Exp1+B would be watching violence and no adults in close proximity.

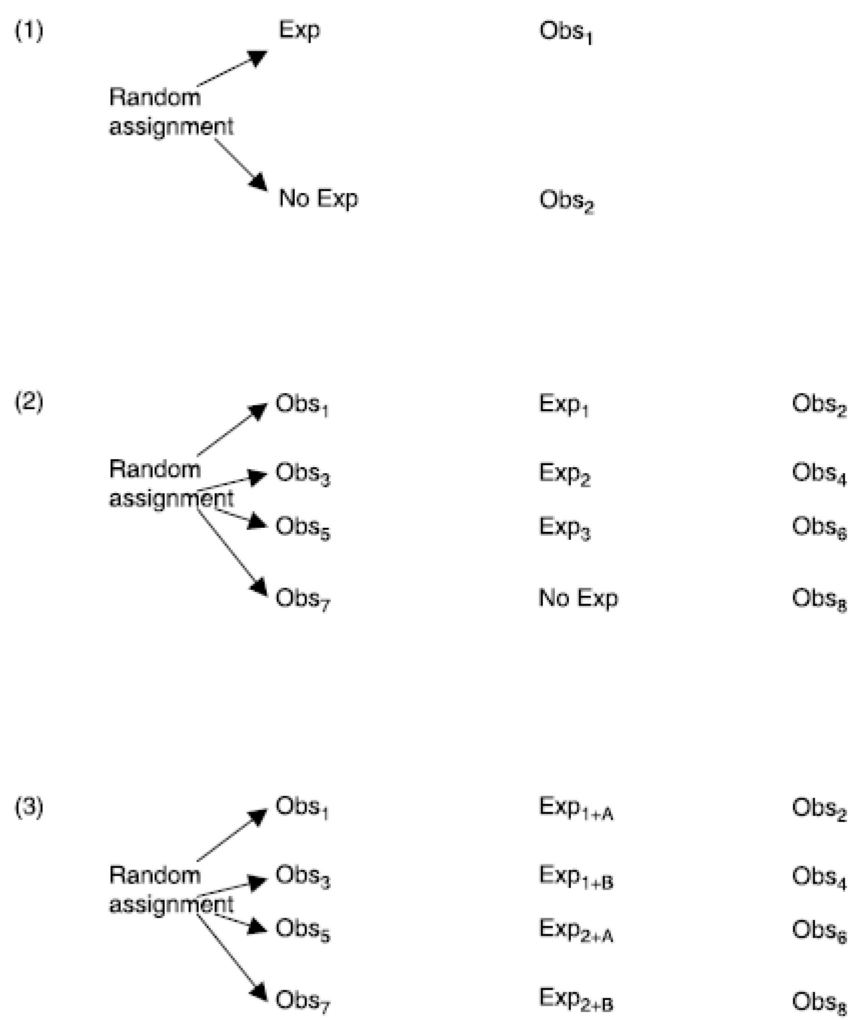

Figure 1.4. Three types of experimental design (Bryman, and Cramer, 2008)

Survey design and causality

When a social survey is carried out, the nature of the research design is very different from the experiment. The survey usually entails the collection of data on a number of variables at a single juncture. The researcher might be interested in the relationship between people's political attitudes and behavior on the one hand, and a number of other variables such as each respondent’s occupation, social background, race, gender, age, and various non-political attitudes. But none of these variables is manipulated as in the experiment. Indeed, many variables cannot be manipulated and their relationships with other variables can only be examined through a social survey. We cannot make some people old, others young, and still others middle-aged and then observe the effects of age on political attitudes. Moreover, not only are variables not manipulated in a social survey study, data on variables are simultaneously collected so that it is not possible to establish a time order for the variables in question. In an experiment, a time order can be discerned in that the effect of the manipulated independent variable on the dependent variable is directly observed. These characteristics are not solely associated with research using interviews or questionnaires. Many studies using archival statistics, such as those collected by governments and organizations, exhibit the same characteristics, since data are often available in relation to a number of variables for a particular year.

Survey designs are often called *correlational* designs to denote the tendency for such research to be able to reveal relationships between variables and to draw attention to their limited capacity in connection with the elucidation of causal processes. Precisely because in survey research variables are not manipulated (and often are not capable of manipulation), the ability of the researcher to impute cause and effect is limited. Let us say that we collect data on manual workers’ levels of job satisfaction and productivity in a firm.

We may find, through the kinds of techniques examined later, that there is a strong relationship between the two, suggesting that workers who exhibit high levels of job satisfaction also have high levels of productivity. We can say that there is a relationship between the two variables (see Figure 1.5), but as we have seen, this is only a first step in the demonstration of causality. It is also necessary to confirm that the relationship is non-spurious. For example, could it be that workers who have been with the firm a long time are both more satisfied and more productive (see Figure 1.6)?

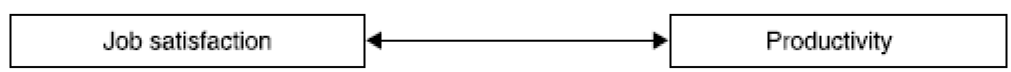

Figure 1.5 A relationship between two variables

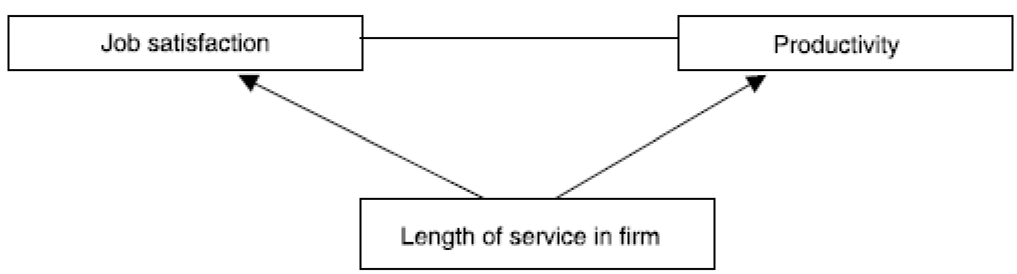

Figure 1.6 Is the relationship spurious?

However, the third hurdle – establishing that the putative cause precedes the putative effect – is extremely difficult. The problem is that either of the two possibilities depicted in Figure 1.7 may be true. Job satisfaction may cause greater productivity, but it has long been recognized that the causal connection may work the other way around (i.e. if you are good at your job you often enjoy it more). Because data relating to each of the two variables have been simultaneously collected, it is not possible to arbitrate between the two versions of causality presented in Figure 1.7. One way of dealing with this problem is through a reconstruction of the likely causal order of the variables involved. Sometimes this process of inference can be uncontroversial.

For example, if we find a relationship between race and number of years spent in formal schooling, we can say that the former affects the latter. However, this modelling of likely causal connections is more fraught when it is not obvious which variable precedes the other, as with the relationship between job satisfaction and productivity. When such difficulties arise, it may be necessary to include a second wave of data collection in relation to the same respondents in order to see, for example, whether the impact of job satisfaction on subsequent productivity is greater than the impact of productivity on subsequent job satisfaction. Such a design is known as a *panel design* (Cramer, 1996), but is not very common in the social sciences. The bulk of the discussion in this book about non-experimental research will be concerned with the survey/correlational design in which data on variables are simultaneously collected.

The procedures involved in making causal inferences from survey data are examined in later section in the context of the multivariate analysis of relationships among variables. The chief point to be gleaned from the preceding discussion is that the extraction of causal connections among variables can be undertaken with greater facility in the context of experimental research than when survey data are being analyzed.

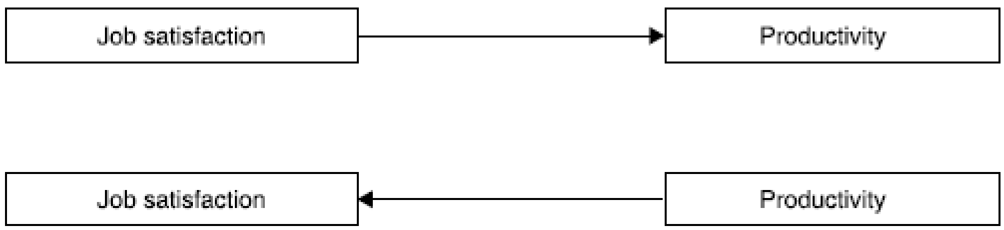

Figure 1.7 Two possible causal interpretations of a relationship

Conclusion

This section was intended to introduce students to quantitative data analysis and research processes. It covered the theory, hypothesis, operationalization of concepts, and how-to setup a research design. The section also explains causality and research design in great depth. We also explained the mixed method of research in great details. In week 2, we shall discuss the sources of data and how to handle large datasets.

REFERENCES

Bryman, A. (2008) *Social Research Methods*. Oxford University Press, Oxford

Cramer, D. (2003) *Advanced Quantitative Data Analysis*. Open University Press, Maidenhead.

Bryman, A. (1988a) *Quantity and Quality in Social Research*, London: Routledge.

Bryman, A. (1988b) 'Introduction: "inside" accounts and social research in organizations', in A. Bryman (ed.), *Doing Research in Organizations*, London: Routledge.

Bryman, A. (1989) *Research Methods and Organization Studies*, London: Routledge.

Bryman, A. and Cramer, D. (1990) *Quantitative Data Analysis for Social Scientists*. London: Routledge

Cramer, D. (1996) 'Job satisfaction and organizational continuance commitment: a two-wave panel study', *Journal of Organizational Behavior* 17: 389–400.

Cramer, D. (1998) *Fundamental Statistics for Social Research: Step-by-Step Calculations and Computer Techniques Using SPSS for Windows*, London: Routledge.

Cronbach, L.J. and Meehl, P.E. (1955) 'Construct validity in psychological tests', *Psychological Bulletin* 52: 281–302.

Davis, J.A. (1985) *The Logic of Causal Order*, Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 55, Beverly Hills, CA: Sage.

Durkheim, E. (1952 [1898]) *Suicide: A Study in Sociology*, London: Routledge & Kegan Paul.