**DATA ANALYSIS: A BRIEF INTRODUCTION by LEE HARVEY**

Originally 1990, modified 2017

Available at <http://www.qualityresearchinternational.com/Harvey%20papers/>

1. DATA ANALYSIS

1.1 COMPETENCY WITH DATA

We live in a data filled world and competence in dealing with numerical data is as important as literacy. In addition, ability to handle computers is becoming an important social skill. This monograph is a general introduction to data analytic techniques, outlining what, in principle, the techniques are about and provide a limited number of examples and exercises. Whilst it is useful to know how to compute the various statistics, as this gives a clearer insight into their meaning, it is more important to understand what they show and how they may be used.

Information technology enables speedy computations through programs, such as Statistical Package for the Social Sciences (SPSS), and obviates both the need to remember computational formulas and to spend considerable amounts of time in computational processes.

Data analytic skills are important from a research perspective in particular in the following ways:

1. interpreting research done by others

2. evaluating existing research

3. designing one's own research

4. maximizing the effectiveness of the data collected.

An ability to handle data allows you to be more critical of research and to be more conscious of what is involved in, for example, talking of causes in social science. The more you know about statistical techniques the less likely you are to be confused or bamboozled by them when critiquing social research.

1.2 WHAT IS DATA ANALYSIS ?

Data analysis is sometimes referred to as statistics and this is reasonable in the social sciences provided it is accepted that the precision of mathematical statistics does not apply to the social scientific situation. However, to avoid confusion between published statistics and statistical techniques the term data analysis will be used for the latter.

There are two types of data analysis *descriptive* and *inferential*.

Descriptive procedures provide a summary picture of a set of numerical information or data. These may be averages, tabulations, graphs, and so on.

Inferential procedures attempt to draw some inferences about the group studied rather than merely describe it. Such inferences would be, for example, to see if one group is different from another, or whether one variable is related to another within a group, and so on.

2. PROBLEMS OF DESIGN IN RESEARCH

2.1. MEASUREMENT.

How does one measure social scientific concepts? This is a perennial problem. Certain aspects of measurement should be kept in mind, they are:

a: [reliability](http://www.qualityresearchinternational.com/socialresearch/reliability.htm)

b: [validity](http://www.qualityresearchinternational.com/socialresearch/validity.htm)

c: [bias](http://www.qualityresearchinternational.com/socialresearch/bias.htm)

d: [accuracy](http://www.qualityresearchinternational.com/socialresearch/accuracy.htm)

2.2. REPRESENTATIVENESS

In conducting social scientific research, the data should relate to the population being studied, the sample used should be representative of it. Bias results when this is not the case.

2.3. CONTROL

Some social scientists, notably psychologists, attempt research that shows the effect of a stimulus on an individual or group. To assess the impact of such an administered stimulus it is necessary to have some [control group](http://www.qualityresearchinternational.com/socialresearch/control.htm) as a comparison. This leads to problems of representativeness and matching.

3. TYPES OF RESEARCH STUDY

3.1. EXPERIMENTATION

This is limited for most social science it tends to be used more (in an impure way) in psychology. While [experiments](http://www.qualityresearchinternational.com/socialresearch/experiment.htm) (in their ideal form) control the environment, they do so only through imposing artificial conditions. Rarely, can experiments in social science be regarded as representative, never 'natural' and their validity and reliability is suspect.

3.2. SAMPLES SURVEYS

As populations are usually too big to deal with, social scientists take [samples](http://www.qualityresearchinternational.com/socialresearch/sampling.htm). Samples can be selected so as to be 'representative', (i.e. everyone has an equal or known [probability](http://www.qualityresearchinternational.com/socialresearch/probability.htm) of being in the sample). This is known as probability or random sampling. In the main, social scientific research rarely has really random samples although it often strives towards them.

Analysis of causal relationships within samples is done by identifying possible causal variables and seeing whether they consistently relate to observed effects. (The processes are discussed in detail below).

3.3 SECONDARY DATA ANALYSIS

[Secondary data](http://www.qualityresearchinternational.com/socialresearch/secondarydata.htm) (sometimes referred to as 'unobtrusive data') is data not directly collected by the researcher, but is initially collected or produced for other purposes which range from the government census to betting slips. Researchers do not have to worry about the impact they have had on the respondent in the data collection process. (Although the respondent may have been previously affected by the collection procedure, or the implications of the collected data, e.g. tax returns). However, the researcher has no control over the data collection nor the criteria for measurement. Such data tends to be indicative rather than precise. Representativeness can be a problem.

3.4. OBSERVATIONAL STUDIES

These do not usually involve much in the way of numerical data other than very simple counts. Essentially they are 'qualitative' rather than 'quantitative' and observational approaches rarely involve attempts at representativeness, control or even measurement (in any precise numerical way). Observational research is usually concerned with subject’s meanings, how the subject(s) interpret or understand the world (see [Researching the Real World, Section 3](http://www.qualityresearchinternational.com/methodology/RRW3pt1Introduction.php)).

4. VARIABLES, VALUES AND CASES.

4.1 VARIABLES

Data analysis measures and manipulates variables. A [variable](http://www.qualityresearchinternational.com/socialresearch/variable.htm) is a theoretical concept defined in such a way that it can be operationally measured. For example, age can be defined operationally as number of years since birth. Income could be defined as the gross annual take-home pay. Social class might be defined using occupation and measured according to the Registrar General's Social Class (see Implications of changes in the UK social and occupational classifications in 2001 for vital statistics, (*Population Trends*, 107, Spring 2002) available at <http://www.ons.gov.uk/ons/rel/population-trends-rd/population-trends/no--107--spring-2002/index.html>(accessed 17 October 2017))

4.2 VALUES

Each variable can take two or more values. For example, the variable 'sex' takes two values 'male' and 'female'. Age takes a range of values from 0 to the age of the oldest person in the sample. Age values may be in whole years or they may be further broken down into years and days. In some circumstances age may be measured in hours (if for example the study involved new born children).

4.3 CASES

Each individual in a sample is referred to as a case. In social science a case is usually an individual and data is recorded for each individual. Thus if we were interested in the age, sex and income of a sample of 100 people then we would need to find the values of the three variables, age, sex and income, for the 100 cases in the sample.

4.4 CODING DATA

The distinction between variables and values is important when coding data. Coding data is the process of recording the value of a variable for each case. This is a straightforward, if laborious process, best handled by constructing a grid with each row of the grid representing a case and each column a variable. The value of the variable for each case is then written into the grid.

5. DESCRIPTIVE DATA ANALYSIS

5.0 AGGREGATES AND PERCENTAGES

Data analysis relies on being able to categorise or measure variables. The first analytic procedure you would normally do is to count the number of cases that fall into different categories. For example, in a survey, the numbers who give each of the alternative answers to your questions. (Or to be more precise, the numbers who fall into each value of your variables).

This provides you with aggregates for each variable. The simplest way to record these for reference (and possibly presentation) is in the form of frequency tables.

5.1 FREQUENCY TABLES

In general a frequency table looks like this:

**Table 5.1 Name of variable: POLITICAL PARTY**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Category Label | Code | Absolute frequency | Relative frequency % | Adjusted frequency % | Cumulative  Frequency % |
| Conservative | 1 | 40 | 20.0 | 28.6 | 28.6 |
| Labour | 2 | 60 | 30.0 | 42.9 | 71.5 |
| Liberal D/SDP | 3 | 25 | 12.5 | 15.6 | 87.1 |
| Others | 4 | 15 | 7.5 | 12.9 | 100.0 |
| Don't know | 8 | 30 | 15.0 | MISSING |  |
| Refuse | 9 | 30 | 15.0 | MISSING |  |
| **Total** |  | **200** | **100.0** |  |  |

This is the standard format from programs such as SPSS.

From this you can see how many people fall into each category, (3rd column headed 'Absolute Frequency') what percentage that is of the whole sample (4th column) and what percentage it is of the whole sample excluding those cases you have labelled as MISSING (5th column). The final column gives you a cumulative frequency (based on the percentages in column 5), this has limited use in this case.

The cumulative frequency can be helpful in other situations, for example, when you collect ages and you construct a frequency table with a range from say 18 through to around 50, you will get a very long table. You may want to break this down into, say 4 age blocks. Using the cumulative frequency, you make the first block equal to the ages which cover the first 25% (approximately), the second block equal to the ages which cover the next 25% and so on. This gives you approximately equal numbers of people in each group. (N.B. There may be reasons for dividing the respondents up in a way that is irrespective of the distribution of replies, in which case you would not adopt the above.) Alternatively, you may, for example want to find the age of the age of a given proportion of the sample, for example the minimum age of the oldest third and the cumulative frequency will readily provide that.

For presentation or calculation purposes, frequency tables are often reduced to the relevant columns, this may be just the adjusted frequencies, as, for example, in the presentation of opinion poll data in the press.

You can, of course, construct frequency tables by hand but it is much quicker and easier using a data analysis package such as SPSS (or JASP the free alternative), SAS (or the free replacement DAP).

5.2 RECODING DATA

Sometimes you will find that frequency tables are very long and unwieldy. Frequency tables on ages when each year has a separate row are sometimes difficult to interpret and to get a clear overview of the data and you may want to group the data into a smaller number of categories. For example, you could construct the frequency table for four age group: 18 to 29, 30 to 49, 50 to 65, and over 65. The decision on the grouping of categories is up to you and depends on the purposes for which the data was collected in the first place.

Grouping data involves recoding data entries. When using a computer program you need to identify the new categories so that the program generates the data in the way you want. Using the example of four grouped categories above, the under 30s would be recoded with the value 1, those between 30 and 39 have the value 2, and so on. Once you have recoded data you would normally want to give the recoded values a label so that you remember how you have grouped the data. These are called value labels.

6. GRAPHICAL (or PICTORIAL) REPRESENTATION

Sometimes you may want to present your data in a way that relies on an optical impression such that the numbers are not the primary focus of attention. The use of various forms of graphical or [pictorial presentation](http://www.qualityresearchinternational.com/socialresearch/pictorialrepresentation.htm) can help. Imagination on your part is crucial here! Presentation in graphical or visual form is greatly enhanced by the use of computer graphics packs.

The following broad types can be distinguished and are illustrated in Appendix 3.

6.1 GRAPHS

Usually these are line graphs in which a series of points are joined to imply a continuous development : e.g. a temperature graph. Misleading impressions can be given by altering scales on the axes.

6.2 BAR CHARTS

These are charts that indicate frequencies by the length of the horizontal bars or vertical columns. Often bar charts include pictorial representations (e.g. rows of figures, cars, pound notes etc.). Cumulative or stacked bar charts can be used to show the breakdown of some total into component parts .

6.3 PIE CHARTS

Such breakdowns can also be presented as pie charts. The relative frequencies are pictured as a circular 'pie' cut into pieces. The use of ‘exploded’ pie charts allows you to highlight one section.

6.4 HISTOGRAMS

These are like vertical bar charts but are for continuous data. That is, the columns are placed next to each other the end of one being the start point of the next.

6.5 TIME SERIES

These can be represented in a number of ways, either as histograms, or line graphs (such as temperature graphs for hospital patients), or as picturegrams . Time series are represented with a chronological measure on the horizontal axis.

6.6 PICTUREGRAMS

A plethora of other ways of presenting data that rely on putting data into pictorial form for comparison purposes.

7. AVERAGES

An [averages](http://www.qualityresearchinternational.com/socialresearch/average.htm) (or measure of central tendency) is a summary of a data set and provides information on the most 'representative' value of a variable. There are three commonly used averages in social science

7.1 ARITHMETIC MEAN (commonly called ‘the average’)

The Arithmetic Mean is the sum total of the values of a variable in a sample divided by the number of people in the sample.

Arithmetic Mean = (Total X)/N   
  
 Where X = value of the variable   
 N = sample size  
 / = divide by

7.2 MEDIAN

The median is the middle value of a variable in a sample when ranked in order.

7.3 MODE

The mode is most frequently occurring value in a sample. It is the value of X with the highest frequency.

7.4 SELECTING THE MOST APPROPRIATE MEASURE OF CENTRAL TENDENCY

Which measure of central tendency is appropriate depends on several things, not least, what the average is being used for.

The only 'objective' criterion is the scale of the data. This is explained in the next section.

8. LEVELS OF MEASUREMENT

In social science there are effectively three levels of measurement (or [measurement scales](http://www.qualityresearchinternational.com/socialresearch/measurementscales.htm)) of data.

Interval where items on a scale are ranked in order and each unit gap is equal. (E.g. time)

Ordinal where items are ranked in order (E.g. preference)

Nominal where items can be categorised but where ordering is not feasible (E.g. gender, political party)

Each of the three measures of central tendency relate directly to each of these scales. It makes no sense to talk of the mean political party or mean gender. Nor is it meaningful to talk about the middle gender or middle party when ranked in order (from highest to lowest). Similarly, it is incorrect to compute the mean of a set of ranked categories as the mean requires interval data.

For example, a sample of 15 people were asked whether they agreed with the proposition that no mineral exploration should take place in Antarctica. Each was asked to say whether they 'strongly agreed, agreed, disagreed, or strongly disagreed' Strongly agree was given a value of 1 and strongly disagreed a value of 4. The results were as follows:

1, 1, 1, 1, 1, 1, 2, 2, 3, 3, 3, 3, 3, 4, 4

The arithmetic mean is 33/15 = 2.2. The median is 2. The mode is 1

Although the mean looks to be O.K. it misleading. It assumes that the difference between each value is the same. The difference between the numerical values are the same but it is doubtful that the difference between what the values stand for, i.e. the difference between 'strongly agree' and 'agree' and between 'agree' and 'disagree', etc., are identical. And beside what does a value of 2.2 mean?

The median is sensible as it is the middle value when all the answers are ranked in order, i.e. the sample 'agree' on average.

The mode makes sense but it is a weak measure, especially with small samples as just one or two values can change it drastically.

To compute an arithmetic mean you need interval data.  
To compute the media you need at least ordinal data.   
To compute the mode you need only nominal data.

9. MEASURES OF DISPERSION

There are various measures of [dispersion](http://www.qualityresearchinternational.com/socialresearch/dispersion.htm) used in social science, they include the range, the quartile deviation (semi-inter quartile range) and the standard deviation.

Measures of dispersion provide information on how spread out a set of data is. Averages provide you with some idea of the most representative measure of a group, but give you no idea of the spread. (E.g. a sample may have an average age of 35 but that tells you nothing about how spread the ages are.)

9.1 RANGE

The range is simply the difference between the highest and lowest value.

For example, with discrete data (whole number data) such as the number of bedrooms in a sample of households that ranges from 1 to 6, the range is (6-1)=5.

If the data is continuous, for example, the age range of a sample in which the youngest is 20 years old and the oldest is 60, the age variable has a range of 41 years inclusive (from 20 and no days to 60 and 364 days).

9.2 QUARTILE DEVIATION

The quartile deviation (QD) is the range of values that covers the middle half of a distribution, ignoring extremes. The data is put in order and then divided into four parts (the quartiles) the difference between the value of the first and third quartile is the inter quartile range.

The quartile deviation is half the IQR. It measures the average deviation of each of the two other quartiles (Q1 and Q3) from the median (Q2).

9.3 STANDARD DEVIATION

The standard deviation is the only one of the three measures that takes into account all the values in a distribution. It is based on the deviation of each value of the variable (X) from the arithmetic mean.

The standard deviation is the square root of the variance. The variance is the mean squared deviation of each value of X from the mean.

Variance = Total(X - Mean)2 / N

What that means is that to work out the standard deviation one computes the mean, then calculates the difference between the mean and each value of the variable, then square these differences (they all become positive), add them up and then divide by the number in the sample, to get the average squared deviation from the mean (the average squared deviation of X from the mean is called the variance)

The standard deviation is the square root of the variance.

This appears to be rather complicated just to measure the spread of a distribution, and conceptually it is a bit cumbersome. However, as you will see later, it is an important measure of dispersion. What you need to know, for now, is that the standard deviation is a measure of spread around the mean and is suitable for interval scale data.

A useful formula to compute the variance (reorganising the formula above) is the following:

Variance = (Total X2)/N - (Mean of X)2

9.4 WORKED EXAMPLE

The following table shows the number of children in 12 families. The mean, median, mode, range, quartile deviation and standard deviation will be computed.

**Table 9.1 No of children per family**

Family No. of children   
 x x2  
 A 0 0  
 B 0 0  
 C 0 0  
 D 1 1  
 E 1 1  
 F 1 1  
 G 2 4  
 H 2 4  
 I 2 4  
 J 2 4  
 K 4 16  
 L 6 36

Totals 21 71

Mode =2

There are more families with two children than with any other number

Median = 1.5

Six families have 1 or fewer children, six families have 2 or more. The mid-point is 1.5

Mean = 21/12 = 1.75

The total number of children divided by the total number of families

Range = 0 to 6 = (6-0) = 6

Quartile deviation = (2-0.5)/2 = 1.5/2 =0.75

The third quartile is 2 (the value of the 9th and 10th families when ranked in order).

The first quartile is mid-way between the third and fourth families, i.e. 0.5.

The interquartile range (IQR)is thus 2-0.5 = 1.5.

The Quartile Deviation is half the IQR, i.e. 1.5/2 = 0.75.

Variance = 71/12 - (1.75) 2 = 5.9167 - 3.0625 = 2.8542

Standard deviation = square root of the variance,

i.e. sq. root of 2.8542 = 1.689 (or 1.7 to one decimal place.)

Exercise 9.1  
  
1. Write out the data on family size above, Table 9.1, as a frequency table.   
  
2. The answer to the question   
'How many professional football matches have you attended this season?'   
is shown in a table 9.2, (for a hypothetical sample of 1000 people).  
  
What proportion of the sample have not been to any matches this season?  
  
Calculate the mean, median, mode, range, quartile deviation and standard deviation of the sample data.   
  
[Note, when calculating the measures you need to take into account the frequencies]

**Table 9.2 Football match attendance**

|  |  |
| --- | --- |
| No. of matches | Frequency |
| x | f |
| 0 | 700 |
| 1 | 80 |
| 2 | 70 |
| 3 | 50 |
| 4 | 30 |
| 5 | 30 |
| 6 | 20 |
| 7 | 10 |
| 8 | 10 |
| **Total** | **1000** |

ANALYTIC OR INFERENTIAL TECHNIQUES

10. CROSSTABULATION

10.1 INTRODUCTION

An important analytic technique in the social sciences is the use of [crosstabulation](http://www.qualityresearchinternational.com/socialresearch/tabulation.htm#crosstabulation). It allows you to break down data. Instead of simply providing a table for the entire sample of who votes for a particular party, crosstabulation would permit you to break down voting preferences by other variables, e.g., gender.

Crosstabulations are typically given in the following way:

Three numbers appear in each 'cell' of the table, thus:

Frequency (i.e. actual numbers in each category)  
Row percent  
Column percent

This is the usual way, for example, that SPSS crosstabulations are presented (although the default on some versions is frequency only).

Crosstabulation, then, links one variable (say the number of football matches attended) with another variable or series of other variables (e.g. gender).

10.2 WORKED EXAMPLE

Taking the answers to the question on football attendance summarised in Table 9.1, above, and breaking them down (i.e. crosstabulating) by gender, we get a table like this:

**Table 10.1 Have you been to a football match this year?**

|  |  |  |  |
| --- | --- | --- | --- |
| Frequency Row % Column% | **No** | **Yes** | **Totals** |
| **Male** | 400 | 200 | 600 |
|  | 67 | 33 |  |
|  | 57 | 67 |  |
|  |  |  |  |
| **Female** | 300 | 100 | 400 |
|  | 75 | 25 |  |
|  | 43 | 33 |  |
|  |  |  |  |
| **Totals** | 700 | 300 | 1000 |

Interpreting crosstabulations can be misleading, and it is important that when interpreting such tables one makes inferences in terms of the independent variable. That is, one takes into account the different numbers of people in the independent variable category (in this case gender).

Consider the following possible interpretations.

1. Using the totals (or raw data).

Twice as many males go to football as females: 200 compared to 100.

2. Using column percentages.

Again, twice as many males attend football matches as females, 67% compared to 33%.

Both these interpretations are misleading because they ignore the fact that there are more males than females in the sample.

3. Using row percentages

33% per cent of males in the sample attend football matches as opposed to only 25% of females.

This interpretation takes into account the difference between the numbers of males and females in the sample.

Exercise 10.1: Provide an interpretation of Table 10.2

**Table 10.2 Voting preference by gender**

|  |  |  |  |
| --- | --- | --- | --- |
| Count  Row %  Column % | Male | Female | **Total** |
| Conservative | 600 | 600 | **1200** |
|  | 50 | 50 |  |
|  | 30 | 20 |  |
|  |  |  |  |
| Labour | 800 | 1200 | **2000** |
|  | 40 | 60 |  |
|  | 40 | 40 |  |
|  |  |  |  |
| Others | 600 | 1200 | **1800** |
|  | 33 | 67 |  |
|  | 30 | 40 |  |
|  |  |  |  |
| **Totals** | **2000** | **3000** | **5000** |

You can, of course, construct crosstabulations using a data analysis package.

11 ASSOCIATION

There are various ways of measuring the relationship between variables, that is, the degree of [association](http://www.qualityresearchinternational.com/socialresearch/association.htm) between them. Conventionally, all measures are designed to provide a measurement of the association on a scale of 0 to 1. A score of zero means that there is no observed relationship at all between the variables. A score of 1 means that there is a perfect relationship between them. In practice, relationships in social science fall somewhere between these extremes.

11.1 MEASURING ASSOCIATION IN CROSSTABULATED DATA

There are many different measures that will provide information on the degree of association apparent between crosstabulated data. Different situations call for different tests and alternatives are indicated below. The measures listed are all generated by the SPSS program.

*Two by two tables*(In the case of simple dichotomies in both variables, the scale of the data is irrelevant)   
  
Phi   
Tau-b   
  
Phi is particularly for 2x2,

Tau-b is more general but the best test in these circumstances. The majority of the remainder below would also be suitable, but not as good as tau-b   
  
*Other than two by two tables*  
Y and X nominal   
Lambda\*   
Uncertainty Coefficient\*   
Cramer's V   
Contingency Coefficient [square tables]  
  
Y and X ordinal   
Tau-b [for square tables]\*\*   
Tau-c [for rectangular tables]\*\*   
Gamma   
Somer's D\*\*   
  
Y interval and X nominal   
Eta

\* These measures also have tests of significance (see below).

\*\* Three coefficients are usually calculated when using computer software, (such as SPSS) two asymmetric ones, for X or Y dependent, (the machine lets you choose the dependent variable) and a kind of average of these, the symmetric version.

Essentially, measures of association measure the extent to which row categories relate to column categories. If, for example, in the football match attendance example we looked at earlier (Table 10.1) all males attended at least one match and all females had attended none at all, then there would have been a high (in fact perfect) relationship between sex and football match attendance (and the top right hand cell of Table 10.1 would have had 600 people in it and the bottom left hand cell would have had 400 people in it).

11.2 WORKED EXAMPLE

Assume we want to measure the degree of association between gender and football match attendance (Table 10.1). The table is reprinted below without the percentage figures

**Table 11.1 Have you been to a football match this year?**

No Yes Totals

Male 400 200 600

Female 300 100 400

Totals 700 300 1000

We have a two by two table. A suitable measure would be the Phi coefficient.

Phi = (ad-bc)/√(klmn) (where ad means a x d and klmn means k x l x m x n and √ means square root)

where the letters in the formula refer to cells in the 2x2 table or totals, thus:

Table 11.2 Have you been to a football match this year?

No Yes Totals

Male a b k

Female c d l

Totals m n 1000

In this example, then

phi = ((400x100)-(200x300))/√(600x400x700x300)

= (40000-60000)/√(50400000000)

= -20000/224499

= -.089

This shows that there is very little association between gender and football match attendance on the basis of the crosstabulated data. (The minus sign is irrelevant, if we had reversed the order of the rows and put females above males the result would have been +.089).

Exercise 11.1: Taking the data in Table 10.2 for Labour and Conservative voters only. Is there any association between voting preference and gender? (Work this out by hand or input into a statistical package).

12 PARAMETRIC AND NON-PARAMETRIC MEASURES

There are two kinds of measure of association. Those that use parameters to define relationships and those that do not. A [parameter](http://www.qualityresearchinternational.com/socialresearch/parameter.htm) is a single measure that summarises a distribution of values. The mean and the standard deviation are both parameters as the first is a single measure of centrality and the second is a single measure of dispersion.

Non-parametric measures of association and ones that attempt to show the relationship between variables by various attempts to match 'patterns' of two or more variables. Non-parametric measures thus attempt to relate variables by comparing entire distributions of values.

We have seen (in section 11) that there are a large number of techniques that use crosstabulation in an attempt to measure the degree of association between variables (crosstabulation usually involves nominal or ordinal scale data). For example, if one wanted to see if gender affected voting preference, a crosstabulation of vote and gender could be drawn up for a sample and the cells inspected to see if the pattern of voting preference coincided for both gender groups. If not, then there may be some association between gender and voting preference. Various techniques, as we have noted, exist for measuring the degree of these relationships. Although in such circumstances the exact nature of the relationship is not specified. Specifying the underlying nature of an observed relationship is known as regression analysis (and is discussed in section 13).

An alternative non-parametric approach for ordinal data is to compare the individual rankings of values for two variables and assess the extent to which the rankings match. For example, an assessment of introversion and conservatism scores (both ordinal scale data) could be attempted by ranking each individual in a sample on the basis of the scores attained on tests of introversion and conservatism and see to what extent the ranks correspond. This approach to measuring association on the basis of individual cases is known as correlation (and is discussed in section 13). When the correlation is done on ranks then this is known as rank order correlation.

Like regression, most correlation analysis uses interval scale data and relates the interval values (rather than rank orders) for related variables for each individual. This approach is known as product moment correlation.

Both regression and product moment correlation are parametric techniques because they measure the degree of association and the nature of underlying relationships by using the mean and the standard deviation. (There are some obscure parametric attempts to measure association and underlying relationships using the median and the quartile deviation but these are not dealt with here).

13 CORRELATION AND REGRESSION

Correlation and regression procedures are used to measure the relationship between two or more variables, usually when the data is of an interval scale. This is something of a problem for social science given that so much data is not of an interval scale.

However, when interval scale data is available grouping the data into categories (via the RECODE command) in order to construct readable crosstabulations loses the precision of the original data. Grouped categories help us to get a picture of interval scale data (such as age data, as mentioned above) but for analytic purposes you get a much better idea of the relationship between two interval scale variables (such as age and income) if you compare the original data rather than the grouped categories. Correlation and regression analysis uses the original ungrouped data and are thus more 'powerful' than measures of association applied to crosstabulated data.

Correlation and regression techniques attempt to show to what extent one variable is dependent upon or associated with one or more other variables. Rather than relate cells in a table, regression and correlation techniques operate on individual cases. Regression analysis also attempts to specify the nature of the relationship between variables rather than simply assess whether there is some relationship or not.

So,

* **correlation techniques measure the degree of association between variables**,
* **regression techniques specify the nature of the relationship**.

For example, a simple relationship may be posited between height and weight. A sample of 100 mature males may be selected at random and their heights and weights measured. A pattern may emerge where the taller people tend to weigh more. However, it is unlikely that there are no exceptions to the general tendency, doubtless any such sample will contain short fat men as well as tall thin ones. Nonetheless there may still be evident a general trend that shows weight to be dependent on height.

Correlation and regression procedures presuppose some sort of underlying relationship, then attempt to

a. identify exactly what this relationship is (regression)

b. assess the extent to which the observed data 'fits' the underlying relationship (correlation).

In our example, if there was a perfect relation where for every inch increase in height individuals were two pounds heavier (e.g. someone 65 inches tall weighed 130 pounds, someone 66 inches tall weighed 132 pounds, someone 67 inches tall 134 pounds, etc) then the underlying relationship would be immediately evident, i.e. a straight line (or linear relationship), and there would be perfect correlation, i.e. no deviation from the linear relationship.

In such an ideal case the sample would generate a correlation coefficient (the measure of association) equal to 1.

The relationship could also be specified as a regression line, which would take the form of a straight line

Y=2X

(where Y = weight and X= height)

When attempting to measure relationships between variables it is usual to identify the independent variable and label it X (when there are several independent variables they are labelled, X1, X2, X3 etc.).

The dependent variable is labelled Y. The dependent variable is the one that (if there is any relationship at all) depends on the other variable(s). E.g. weight may depend on height (and not *vice versa*), income may depend on gender (despite legislation!). Height and gender would be independent in these relationships, weight and income dependent.

Regression analysis, in attempting to measure the underlying relationship, and thus providing a basis for measuring the degree of relationship between the dependent and independent variables, must first determine the nature of the underlying relationship. The normal recourse is to assume that the underlying relationship is linear. (Although techniques exist to determine curvilinear relationships, these are not widely used in social science and will not be discussed here).

In order to specify regression equations, it is necessary to have interval scale data. (There are some relatively obscure attempts to create regression equations for ordinal data but these will not be dealt with).

13.1 BI-VARIATE ANALYSIS

Consider a simple two variable relationship (bi-variate), between height and weight. A sample of 100 people will probably show some tendency that would suggest height and weight are related. However, under normal conditions the extent of the relationship and the exact specification of the relationship will not be clear.

[Regression](http://www.qualityresearchinternational.com/socialresearch/regression.htm) analysis assumes that there is a relationship and (usually) that it is linear. It then attempts to locate a specific relationship (in the form of a straight line) that best represents the data, a sort of 'average' relationship. (Much in the same way that charts of height and weight which show 'normal' weights for each height attempt to provide a profile of the 'average' weight/height relationship).

There are various methods of determining this average or best underlying relationship. These range from the 'subjective' assessment of the researcher to more 'objective' measurements which determine the unique line which minimises variation.

When the data is interval, and the minimum variation is measured using the standard deviation this is known as the *least squares regression line of best fit*, and is the primary approach to regression analysis.

[Correlation](http://www.qualityresearchinternational.com/socialresearch/correlation.htm) techniques then show the extent to which the data fits the underlying relationship. When the correlation coefficient is near to 1 then the data closely fits the underlying relationship that has been identified. When the correlation coefficient is close to zero then the data does not fit closely to the underlying relationship and there is little or no association observable between the two variables (despite being the best fit to the data).

13.1.1 Worked Example

Table 13.1 shows the number of years in full-time teaching and the salary (in thousands of pounds) of 10 teachers. The least squares regression line and the Pearson's product moment correlation coefficient will be calculated to show the underlying (linear) relationship between service and salary (given the sample data) and the extent to which underlying relationship fits the data.

**Table 13.1 Years of service and salary (£000s)**

|  |  |  |
| --- | --- | --- |
| Teacher | Years | Salary |
| A | 1 | 6 |
| B | 2 | 7 |
| C | 3 | 7 |
| D | 4 | 8 |
| E | 5 | 12 |
| F | 6 | 8 |
| G | 7 | 8 |
| H | 8 | 14 |
| I | 9 | 7 |
| J | 10 | 13 |

We need to identify the dependent and independent variables. If there is any relationship then salary is dependent on years as a teacher (rather than vice versa). Years as a teacher precedes salary.

In general, the regression line takes the form:

Y = a + bX

(where Y refers to the values of the dependent variable, X refers to values of the independent variable, a is the point at which the straight line intersects the Y axis and b is the gradient of the straight line).

In order to specify the nature of the line we need to compute a and b

The formula for b is:

n(sum of XY)-(sum of X multiplied by sum of Y)

b= -----------------------------------------------------

n(sum of X squared) - (sum of X)squared

This is usually written with the Greek letter sigma (∑) as a short hand for 'sum of'.

n(∑XY)-(∑X.∑Y)   
b= --------------------  
 n(∑X2) - (∑X)2

Note that (∑X2) is not the same as (∑X)2. The first involves squaring all the values of x and then adding them up. The second involves adding up all the X values and then squaring the result.

The formula for a is:

a= Mean of Y - b(Mean of X)

The formula for the Pearson's Product Moment Correlation Coefficient (r) is:

n(sum of XY)-(sum of X multiplied by sum of Y)

r= ----------------------------------------------------------------

√[(n(sum of X2)-(sum of X)2).(n(sum of Y2) - (sum of Y)2)]

Again this is usually written using the ∑ notation:

n(∑XY)-(∑X.∑Y)   
r= -----------------------------------------  
 √[(n(∑X2)-(∑X)2).(n(∑Y2) - (∑Y)2)]

These look horrific but they are not too bad really. What we need to calculate are the following:-

Sum of X ∑X Sum of X squared ∑X2

Sum of Y ∑Y Sum of Y squared ∑Y2

Sum of X multiplied by Y ∑XY

Thus:

Table 13.2 **Years of service and salary (£000s) calculation of regression and correlation coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Teacher | Years | Salary |  |  |  |
|  | X | Y | XY | X2 | Y2 |
| A | 1 | 6 | 6 | 1 | 36 |
| B | 2 | 7 | 14 | 4 | 49 |
| C | 3 | 7 | 21 | 9 | 49 |
| D | 4 | 8 | 32 | 16 | 64 |
| E | 5 | 12 | 60 | 25 | 144 |
| F | 6 | 8 | 48 | 36 | 64 |
| G | 7 | 8 | 56 | 49 | 64 |
| H | 8 | 14 | 112 | 64 | 196 |
| I | 9 | 7 | 63 | 81 | 49 |
| J | 10 | 13 | 130 | 100 | 169 |
| **Totals (∑)** | **55** | **90** | **342** | **385** | **884** |

n= sample size = 10

n(∑XY)-(∑X.∑Y)   
b= --------------------  
 n(∑X2) - (∑X)2

(10x542) - (55x90)

b= -----------------------

(10x385) - (55x55)

5420 - 4950

b= ----------------

3850 - 3025

b= 470/825

**b= 0.57**

a= Mean of Y - b(Mean of X)

a= mean of Y - .57 (mean of X)

a= 90/10 - .57(55/10)

a= 9 - (.57)(5.5)

a= 9 - 3.135

**a= 5.9** to one decimal place

The least squares regression line of best fit is, therefore,

**Y = 5.9 + .57 X**

So, for example, when X=10 (10 years as a full-time teacher) then the predicted Y (salary) would be:

Y = 5.9 + (.57)(10)

Y = 5.9 + 5.7

Y = 11.6 (thousand pounds)

We can draw a scatter diagram of the original data and add to it the least squares line, thus:

Scatter diagram:

Salary

Y

15

14 x

13 x

12 x

11

10

9

8 x x x

7 x x x

6 x

0 1 2 3 4 5 6 7 8 9 10 X Years

From the regression formula:

when X=0 Y=5.9

when X=10 Y=11.6

Plot these two points and join them up to show the regression line.

The Pearson's Product Moment Correlation Coefficient (r)

n(∑XY)-(∑X.∑Y)   
r= -----------------------------------------  
 √[(n(∑X2)-(∑X)2).(n(∑Y2) - (∑Y)2)]

(10x542) - (55x90)

r= ---------------------------------------

√[(10x385 - 55x55) (10x884 - 90x90)]

5420 - 4950

r= -------------------------------

√[(3850 - 3025)(8840 - 8100)]

470

r= -----------------

√[(825) (740)]

r= 470/ √[610500]

r= 470/781.3

r= 0.6

A correlation coefficient of 0.6 shows a reasonably strong relationship between years in teaching and salary.

13.1.2 Coefficient of determination

r-squared (r2) is known as the coefficient of determination and shows the percentage of any change in Y which is attributable to a change in X.

In the worked example above (13.1.1) r2 is 0.6x0.6 = 0.36.

Thus, 36% of any change in salary is attributable to length of time in the profession. Other factors account for 74% of the change in salary (based on this small sample).

Exercise 13.1   
Table 13.3 shows the *per capita* income and the public expenditure per student by a sample of 10 states in the USA. Calculate the regression line of expenditure on income (either by hand or using a computer program). To what extent is expenditure correlated with income? If a state has a *per capita* income of $95000 what would be the best estimate of the likely expenditure on state education? How much of the expenditure on education is accounted for by per capita income across all 10 states?

**Table 13.3 Per capita income and public education expenditure ($0000)**

|  |  |  |
| --- | --- | --- |
| State | Per capita income ($0000) | Public Education Expenditure per Student ($0000) |
| Arkansas | 4 | 4 |
| California | 12 | 9 |
| Kansas | 5 | 5 |
| Michigan | 7 | 8 |
| New York | 8 | 10 |
| Oregon | 6 | 5 |
| Rhode Island | 7 | 6 |
| South Carolina | 5 | 3 |
| South Dakota | 6 | 5 |
| Wyoming | 5 | 6 |

*Data is hypothetical*

13.1.3 Correlation coefficient for ordinal data

Pearson's Product Moment Correlation Coefficient, as we have said, requires interval scale data. However, it is possible to compute a correlation coefficient for ordinal scale data without using crosstabulation.

Both Spearman's Rank Order Correlation Coefficient and Kendall's Tau statistic enable us to do this. However, only Spearman's coefficient will be demonstrated here.

Spearman's Rank Order Correlation Coefficient operates like Pearson's coefficient but instead of using the original data it ranks the original data in order and then computes the degree of association between the two sets of ranks. It thus requires that both variables are at least of an ordinal scale.

For example, a group of ten students were given an 'authoritarianism' test and an IQ test to see if there was any relationship between intelligence quotient and tendency towards anti-authoritarianism. The authoritarian test provides a score of between 1 (authoritarian) to 20 (anti-authoritarian). Table 13.4 shows the original scores and the rank order of the students on each scale.

**Table 13.4 Authoritarian score and IQ score**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student | Authoritarian Score | IQ Score | Authoritarian Rank order | IQ Rank order |
| A | 1 | 95 | 1 | 3 |
| B | 3 | 99 | 2 | 4 |
| C | 4 | 120 | 3 | 7 |
| D | 5 | 89 | 4 | 2 |
| E | 9 | 87 | 5 | 1 |
| F | 11 | 100 | 6 | 5 |
| G | 12 | 121 | 7 | 8 |
| H | 14 | 130 | 8 | 10 |
| J | 15 | 125 | 9 | 9 |
| K | 18 | 118 | 10 | 6 |

Clearly there is not a perfect correlation, IQ and authoritarianism are not perfectly matched although there is a tendency for those with the highest IQ to be more anti-authoritarian.

Spearman's rs gives an indication of the extent of this correlation. The formula for calculating Spearman's Rank Order Correlation Coefficient (rs) is as follows:

6∑d2  
rs = 1 - --------  
 n(n2-1)

Where d is the difference between ranks for each pair of observations, n is the number of observations and ∑ means 'sum of'.

So to calculate rs you need to work out the differences in the ranks, square them, and add them up. Then substitute the figure in the formula. Thus:

**Table 13.5 Authoritarian score and IQ score, calculation of rs**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Student | Authoritarian Rank order | IQ Rank order | d | d2 |
| A | 1 | 3 | 2 | 4 |
| B | 2 | 4 | 2 | 4 |
| C | 3 | 7 | 4 | 16 |
| D | 4 | 2 | 2 | 4 |
| E | 5 | 1 | 4 | 16 |
| F | 6 | 5 | 1 | 1 |
| G | 7 | 8 | 1 | 1 |
| H | 8 | 10 | 2 | 4 |
| J | 9 | 9 | 0 | 0 |
| K | 10 | 6 | 4 | 16 |
|  |  |  |  | ∑d2 = 66 |

6(66) 6(66) 396

rs = 1 - --------- = 1 - ------ = 1 - ------- = 1 - 0.4

10(102-1) 10(99) 990

rs = 0.6

There is thus a reasonably strong correlation of 0.6 between IQ and authoritarianism for the sample of students.

Again, computer programs will do the computation quickly but it is important to know what the computer is doing when it generates statistics and what these statistics tell you and when they are appropriate.

14. MULTIVARIATE ANALYSIS

Multivariate analysis extends the bi-variate case to deal with situations in which Y is dependent upon more than one independent variable. In the social world it is unlikely that a simple association between one independent variable and a dependent variable can be established. For example, income may be dependent on occupation, but that alone is insufficient, as age, gender, geographic location and so on, also effect income levels.

What the multiple linear regression line shows then, is the relationship between a single dependent variable and any number of independent variables.

Thus, whereas the bi-variate linear regression line takes the general form

Y = a + bX

the multivariate linear regression line takes the form

Y = a + b1X1 + b2X2+ b3X3 +....+ bnXn

where there are n independent variables and each has a coefficient (b1, b2 etc).

The size of the coefficient provides an indicator of the relative importance of the individual independent variables (the Xs) in determining the dependent variable (Y).

E.g.:

If Y = income

X1 = occupation

X2 = age

X3 = gender

X4 = educational attainment score

X5 = introversion score

and the regression line takes the following form

Y = .3 + .34X1 + .28X2 + .15X3 + .2X4 + .002X5

then we can see that the most important factors in determining income (Y) are probably occupation, then age, then educational attainment then gender and finally introversion.

Indeed introversion has such a small coefficient that it might be excluded and the regression line recalculated using X1 to X4 only.

N.B. the 0.3 at the beginning of the equation is a 'constant' which has little use in interpreting the regression equation.

What does this example tell us, then?

It points to the relative importance of the various factors in determining income. However, you must be careful in making assertions on the basis of regression equations. The following must be borne in mind:

1. the equation is (usually) based on a sample and thus the results are prone to sampling variation (this is discussed in more detail below).

2. the equation shows a mathematical relationship between those items you choose as independent variables and the one you chose as dependent. It therefore is not a definitive statement of causal relationships. Why?

a. because you may have left out an important variable (such as geographic region in the example above)

b. your assumption of a deterministic relationship between Y and the selection of Xs may be entirely false.

3. the equation says nothing about the extent of the association between the dependent and independent variables.

Calculating multiple regression equations by hand is a lengthy and time-consuming process; for that reason no worked example is given here. The computation involves much the same procedures as the computation of bi-variate regression lines. A statistical program can be used to compute the coefficients for a multiple regression.

14.1 MULTIPLE CORRELATION

A multiple correlation coefficient is a measure of the relationship of Y with the combined Xs. It is an indicator of the extent to which the multiple regression equation fits the data. As in the bi-variate case, the multiple correlation coefficient varies between 1 and 0, the nearer to 1, the stronger the degree of association.

The multiple correlation coefficient (usually Pearson's Multiple Product Moment Correlation Coefficient, (R)) allows us to specify the relationship between Y and the associated Xs.

The normal procedure when calculating a multiple regression line and multiple correlation coefficient is to identify likely factors that have a bearing on Y and (on the basis of available data) proceed in, what is called, a STEPWISE fashion.

The procedure works by entering each independent variable into the equation one at a time starting with the one that has the highest correlation with the dependent variable. Then each subsequent variable is entered on the basis of the which has the highest correlation with the dependent variable, allowing for the influence of the independent variables that have already been entered into the equation. (Any variable already entered whose relationship with the dependent variable disappears when other variables are entered into the equation is automatically removed from the equation by some computer programs such as SPSS). At each step a new regression line and multiple correlation coefficient is generated.

In the example above suppose the multiple regression coefficients for each step were as follows.

Y with X1 R = 0.2

Y with X1 and X2 R = 0.5

Y with X1 and X2 and X3 R = 0.9

Y with X1 and X2 and X3 and X4 R = 0.95

Y with X1 and X2 and X3 and X4 and X5 R = 0.952

This would show that X1, X2 and X3 together were closely correlated with Y (R = 0.9). Adding X4 (educational attainment score) increased the correlation coefficient a little but was of marginal importance. Introversion (X5) had virtually no effect on the multiple correlation and could easily be excluded from the analysis.

Note: the value of R will always be increased a little whatever variable is included, even a set of random values will lead to a marginal increase in R in its raw form. Usually, computer programs such as SPSS, which will compute this stepwise process for you, will provide an adjusted value of R which takes account of the number of different Xs included. The adjusted score will then sometimes fall as additional items of no consequence are added to the list of variables to be included. In the above example, if the adjusted values were:

Y with X1 Adjusted R = 0.2

Y with X1 and X2 Adjusted R = 0.495

Y with X1 and X2 and X3 Adjusted R = 0.895

Y with X1 and X2 and X3 and X4  Adjusted R = 0.935

Y with X1 and X2 and X3 and X4 and X5 Adjusted R = 0.925

Then including X5 is quite clearly a negative step and the multiple relationship would include only X1 to X4 (with X4 being of relatively little importance).

14.2 PARTIAL CORRELATION

Partial correlation is a procedure that allows you to measure the effects of one X on Y, while controlling for the effects of all the other independent variables (known as the control variables).

It operates through statistical manipulation, which has a basic assumption that there are linear relationships between all the variables (as in the linear (or first order) multiple regression equation). Once the linear relationship between the independent, the dependent and the control variables is known, it is possible to remove the effect of the control variables. This is done by predicting the values of the independent and dependent variable (separately) from the control variable(s) (on the basis of the correlation between the control variable(s) and X and the control variable(s) and Y).

Partial correlation is a useful tool in clarifying the relationships between three or more variables. While the multiple correlation coefficient gives an overall picture of the relationship between Y and the combined Xs, the partial correlation coefficients provide the basis for examining the relationship between each X and Y when all the influence of the other factors is removed.

For example, income might be seen to be a function of education, and it would be possible to assess the extent of this relationship (assuming that education can be measured). We could simply undertake a linear bi-variate analysis and show the degree of association between education (X) and income (Y). Such a correlation coefficient might be computed to be 0.45, for example.

This simple linear relationship would probably be unsatisfactory, however, as it ignores other factors, such as age, gender, and so on. A multiple analysis may then be preferable. We could compute a multiple linear regression equation such as:

Y = .2 + 0.53X1 + 0.24X2 + 0.1X3

where X1 = educational attainment score, X2 = age, X3 = gender

Suppose the bi-variate correlation matrix is as follows (this is the correlation of each variable with all the other variables, thus, in the example, the correlation between education and income is r=0.45):

income education age gender

income 1.00

education 0.45 1.00

age 0.30 0.05 1.00

gender 0.15 0.19 0.05 1.00

This indicates that educational attainment is the most important of the various factors in effecting income. A multiple correlation coefficient may be calculated. Suppose, for the example above, it equalled 0.6, this would show that the three factors (weighted and combined linearly) correlated moderately well with income, (and better than the simple correlation of X and Y, where, for the example, the best correlation = 0.45). However, the multiple correlation of 0.6 would by no means provides a complete explanation of variations in income.

The partial correlation coefficients show the relationship of any one of the independent variables with Y, whilst controlling for the effect of the others. So, it would be possible to see what effect X1 (educational attainment score) had on Y whilst controlling for age and gender. (This would be the partial correlation of X1 on Y, controlling for X2 and X3).

This procedure is different from just calculating the bi-variate relationship of X1 on Y. What the partial correlation does, in effect, in this case is to show what the relationship between income and educational attainment is, irrespective of the influence of age and gender. The computation acts as though you had an enormous sample and took out all the males of a given age and analysed the correlation between income and educational attainment, then did it for all the males of another age group, then again for the next age group, and so on, then repeated the whole process for all the females in each age group, then worked out an average of all the correlations between income and educational attainment for all the different age/gender groups. The partial correlation is in effect what that average would be. However, because of the statistical predictive procedures used, one does not need such a large sample.

Suppose, in the example, that the partial correlation of X1 (education) on Y controlling for X2 and X3, was small, say 0.1. This would suggest that there was no direct correlation between income and educational attainment, that once the effect of gender and age were removed there was a more or less random variation between X1 and Y.

Partial correlation analysis, then, is useful in clarifying relationships. It can indicate spurious relationships by uncovering preceding or intervening variables.

A [spurious](http://www.qualityresearchinternational.com/socialresearch/spuriousness.htm) relationship is one in which Y appears to be associated with X1 (as indicated by say the simple correlation of Y and X1, but that X1 merely varies with some other preceding independent variable(s), X2 (X3 etc) which is the real predictor of Y. When X2 is controlled for, the partial correlation coefficient for X1 on Y (controlling for X2) will be small if such a spurious relationship exists.

This is illustrated in the above example where the partial correlation of X1 on Y drops to 0.1. In such a case, the variation in X1 is reflected in the variation in age and gender (which precede educational attainment). What this would mean, then, is that, in the example, when one controls for age and gender, income is independent of (i.e. not related to) education.

Partial correlation can also be used to suggest causal chains of intervening variables between Y and X1. In other words, a relationship observed between Y and X1 (a simple correlation coefficient) may be spurious in the sense that Y is dependent on some intervening variable(s) that is dependent on X1.

Take for example, the relationship between social class (as measured by some socio-economic index based on income and occupation) of parents and of children. A simple correlation might reveal a high correlation coefficient, suggesting (possibly) a substantial degree of transfer of wealth from one generation of a family to the next. However, a second thesis might suggest that the relationship is not that simple (that such a correlation is misleading) that, in fact, the social class of the parents provides a basis for the generation of conditions by which the children attain the same class. This might be through the mediation of education, occupation etc., which the parents effect and which, in turn effect the social class of the child (in later life).

For the computation and interpretation of the partial correlation coefficient, the same procedures operate for intervening variables as in the case of preceding variables, only the theoretical model (suggesting a causal chain, rather than the dissolution of a causal link) is different.

So, generally, if the partial correlation between X1 and Y (controlling for X2, (X3 etc.)) drops below the simple correlation of X1 and Y, then the relationship between X1 and Y can be said to be likely to be spurious. The nearer the partial correlation approaches zero, the clearer is the spurious nature of the original observed relationship between X1 and Y. In such a case, X1 should be removed from the relationship and Y considered as a function of X2 (X3 etc).

Sometimes, however, the partial correlation coefficient of X1 and Y (controlling for X2 etc.) is larger than the simple correlation of X1 and Y.

Suppose that, in the example, correlating income with educational attainment, age and gender, the partial correlation of Y on X1 (controlling for X2 and X3) was 0.8, then this would suggest that, when the effect of age and gender was removed, there was a strong association between income and educational attainment, (stronger than the simple correlation which was 0.45). In effect, the role of gender and age in the initial regression equation, while necessary as controls, was to conceal the direct effect of X1 on Y. In such a case, controlling for other relevant variables has served to increase one’s confidence in relationship between X and Y (in the example, between education and income).

14.3 MULTIVARIATE ANALYSIS USING CROSSTABULATIONS

It is possible to analyse the relationship between two variables taking into account other factors when the data is crosstabulated (i.e. of a nominal or ordinal scale). This is done by constructing n-way crosstabulations (i.e. crosstabulations that are more than simple two-way crosstabulations. The approach is to create crosstabulations of X by Y controlling for a third (or any number) of other variables. So, for example, using sample data, one might construct a crosstabulation of voting preference by social class.

For example:

**Table 14.1 Crosstabulation Vote by Class (**Count)

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 125 | 75 | 200 |
| Labour | 75 | 125 | 200 |
| Totals | 200 | 200 | 400 |

This is a simple two-way crosstabulation. It shows a relationship between class and voting preference.

One may, however, think that this is too simplistic and want to take other variables into account. For example, gender may play a part in voting preference, thus one may want to control for gender. A three way crosstabulation, controlling for gender, would further breakdown the initial crosstabulation into the relationship between voting preference and social class for men and for women. I.e. create two tables, thus:-

**Table 14.2a Crosstabulation Vote by Class (Count), Controlling for gender. Gender = male (n=200)**

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 55 | 45 | 100 |
| Labour | 45 | 55 | 100 |
| Totals | 100 | 100 | 200 |

**Table 14.2b Crosstabulation Vote by Class (Count), Controlling for gender. Gender = female (n=200)**

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 75 | 25 | 100 |
| Labour | 25 | 75 | 100 |
| Totals | 100 | 100 | 200 |

These two tables could be further broken down using a second control variable, such as age. If we divide age into two groups ('old' as over 40 and 'young' those between 18 and 40), then we would create four tables of vote by class, one for young males, one for young females, one for old males and one for old females.

For example:

**Table 14.3a Crosstabulation Vote by Class (Count), Controlling for gender and age. Gender = male, age = young (18-40 years) (n=90)**

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 20 | 20 | 40 |
| Labour | 30 | 20 | 50 |
| Totals | 50 | 40 | 90 |

**Table 14.3b Crosstabulation Vote by Class (Count), Controlling for gender and age. Gender = male, age = old (Over 40 years) (n=110)**

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 35 | 25 | 60 |
| Labour | 15 | 35 | 50 |
| Totals | 50 | 60 | 110 |

**Table 14.3c Crosstabulation Vote by Class (Count), Controlling for gender and age. Gender = female, age = young (18-40 years) (n=110)**

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 30 | 20 | 50 |
| Labour | 20 | 40 | 60 |
| Totals | 50 | 60 | 110 |

**Table 14.3d Crosstabulation Vote by Class (Count), Controlling for gender and age. Gender = female, age = old (Over 40 years) (n=90)**

|  |  |  |  |
| --- | --- | --- | --- |
| Party | Middle Class | Working Class | Totals |
| Conservative | 45 | 5 | 50 |
| Labour | 5 | 35 | 40 |
| Totals | 50 | 40 | 90 |

The big problem with this way of analysing multivariate relationships is that as we increase the number of control variables, the total numbers in each table decline. There is then, a practical limit to the number of controls one can introduce for any given sample size (otherwise one ends up with very small sub-samples which tell you nothing).

What does the above example tell us about specifying the relationship between vote and class, when we control for gender and age?

There was an initial relationship observed between vote and class. When we took account of gender, Tables 14.2a and 14.2b reveal a stronger relationship between vote and class for women than for men. When this was further broken down, in using age, Tables 14.3(a-d), it was evident that the strongest relationships existed amongst the older voters, both male and female. (The female voters, however, exhibiting a stronger relationship than the males in both age categories). This would lead us to suppose that while class and vote are associated, the relationship is mediated by age and gender.