Survey Analysis Workshop

Statistical Notes[1]

[Last updated 13 July 2013]

[NB: The books referred to in the text are those in use at the time the notes were written (1981-1988). For a more up-to-date list see the SPSS Textbooks page.]

Introduction

The following notes (originally drafted by Jim Ring) derive from teaching hands-on, professional practice oriented courses (postgraduate, part-time, evenings) in survey data collection, data management, computer processing and statistical analysis to (mostly public or not-for profit, but some private sector) researchers, graduate students and to graduates (some of whom were unemployed) looking to enter a social research career. The courses were given at the Polytechnic of North London from 1976 to 1992, and represent a continuation of the computing and statistical elements of the Summer Schools in Survey Methods offered by the then SSRC Survey Unit from 1970 to 1976.

These notes represented an attempt to fill a gap in the textbook provision for students who found computers and statistics daunting, and were mostly written before the appearance of the SPSS Guide to Data Analysis (Norusis, M., 1987). They were not intended as a replacement, and should be used in conjunction with the recommended texts.

We are extremely grateful to previous students of the Survey Analysis Workshop and of the BA Applied Social Studies (Social Research Option) and BSc Social Science (Research Pathway) on whom earlier versions were tested and for whom they were written.

We are also particularly grateful to our colleague **David Phillips** whose lecture notes for his introductory statistics course provided the basis for the whole of section 1 and about two thirds of sections 3 and 4.

[1] **[NB** Most of these notes were written on a Vax mainframe and printed up on a line-printer long before PCs and Windows came out. The text was in Courier 12 fixed width font, but conversion to another font with proportional spacing (eg Arial or Times New Roman) is extremely tedious and time-consuming so the original font and printing format is retained for SPSS output.

Appendix 2 contains sample output for the same examples in SPSS for Windows. To date these are from SPSS11 (full screenshots before I learned how to copy SPSS output properly) but they will be extended to include output from SPSS 18.

In places the font and layout seems to have gone haywire, most likely as a result of transfers across several machines and editions of word-processing software. Some small bits seem to be missing. These are currently being sought or reconstructed.]

Section 1 - Statistical Methods in Social Research

Basic Reading:

ROWNTREE **Statistics without Tears**: Chapter 1 or

LOETHER & MCTAVISH **Descriptive Statistics for Sociologists**: Chapter 1 or

MUELLER et al **Statistical Reasoning in Sociology**:
Chapter 12, Section 1 and Chapter 11, Section 1.

1.1 What Are `Statistics'?

The word `statistics' can be used in at least four different senses:

1. It can indicate a whole subject or discipline: the study of statistics;

2. It may refer to the methods used to collect, process or interpret information i.e. data): statistics as a set of research methods;

3. It may be applied to the actual data collected: the statistical records themselves;

4. It may refer to certain specially calculated figures that characterise such a collection of data: averages, percentages etc.

The meaning to be emphasised in this course is the second of these: statistics as a set of methods of enquiry. The first meaning, statistics as a subject of study, is really more to do with mathematics, and will not be covered in this course. The third will also be used in the first part of the course, but only describe summaries taken from the data. It is very rare nowadays to refer to `statistics' in terms of the data itself. The fourth meaning refers to a statistic (singular), one of many specially calculated figures whose detailed definition will be introduced later in the course.

1.2 Quality and Quantity

Statistics differs from most other methods in social research in that it is quantitative rather than qualitative: extensive rather than intensive. For example, ntensive methods such as ethnography use a qualitative approach, examining a large amount of very structured information about a limited number of subjects.
Conversely, statistics - an extensive method - uses a quantitative approach, examining a limited amount of information about a large number of subjects.

However, it would be wrong to associate statistics with quantitative research only (as many people mistakenly do) as the ethnographer very often needs to use such techniques in order to make sense of his/her data too. Another very common mistake is for people to see these methods as being totally separate and often as standing in opposition to each other. But rather than taking an either (quantitative) or (qualitative) approach, it is better to regard such methods as being complementary to each other - the ethnographer providing an in-depth subjective level of information lacking in the more formal `factual' or `objective' statistical survey, whilst such factual information can help the ethnographer better to place his/her results in the wider social and economic context.

1.3 When Are Statistics Used?

At the wider level, you'd have to be fairly blind not to notice that statistics invade our daily lives at every level. In the media and the newspapers, etc. we find for example that the issues of the day - the economy, unemployment, prices, factory closures, wage rates, etc. - are all presented in statistical terms; likewise such issues as crime and its causes, abortion, immigration, etc. statistics, then, are used to provide us with information about what's happening in society. They comprise one of the means by which we can come to understand and evaluate changing social trends and the differing explanations offered to account for them.

However, we need always to keep in mind that statistics are not just numbers. If you take an issue like `Does unemployment cause crime?', some sociologists would tend to agree while others would argue that any correspondence was merely coincidental. It is not, then, just a matter of collecting figures on crime on the one hand and unemployment on the other.

We need to consider how the statistics are interpreted and evaluated, and this involves an interaction between statistical `facts' and social theory, which in the end is often a political question.

Statistics, therefore, can never be seen as just a neutral presentation of numerical `facts'. Different interpretations of the same statistical data can produce very different results, often varying according to the ideological sympathies of the researcher and/or social commentators concerned - even when we come ourselves to evaluate such research according to our own beliefs and prejudices.

1.4 Some Basic Concepts Used In Statistics

Let's go on to define a few basic concepts. Suppose we are carrying out some research using statistics. Where do the statistical methods actually come in?

Population

At a very early stage, we need to define exactly who (or what) we are looking at - the subject of the study. This is known in statistics as the POPULATION - the whole group of people, institutions (or whatever) which forms the subject of the research.

Sample

But there aren't usually the resources to look at the whole population. Instead we draw a SAMPLE from the population and look at this instead. The sample, then, consists of those members of the population who are actually interviewed. Later, we deal with the whole process of selecting a sample. For now, you need only know that the sample should be REPRESENTATIVE of the population - that it represents a smaller image of it.

Data Matrix

The next stage is to collect the information from the sample: to prepare a questionnaire, to interview the subjects and to transfer the information onto a computer. Here the information is usually referred to simply as DATA - the actual answers recorded from the sample - or more precisely as a DATA MATRIX. We go on to look at this in detail in the next section.

Descriptive Statistics

This is the point at which statistics comes in - in the summarising and making sense of a mass of data. You'll be able to use it to group observations in various ways. For example, how many people live in this kind of housing, or comparing people according to age, income, political preferences etc. You'll also be able to identify patterns and spot trends in the data - points of similarity, points of difference, things that are common to most, things that are unusual or out of the way, etc. This is **DESCRIPTIVE STATISTICS** - the ordering, grouping and summarising of data so that patterns, trends and possible relationships can be discerned.

Inferential Statistics

A second range of statistical procedures, known as **INFERENTIAL STATISTICS**, involves taking the basic data and attempting to go beyond it - to generalise on the basis of your relatively small set of observations and derive conclusions of wider applicability. Generally, this kind of statistics produces findings and conclusions which are expressed in terms of estimates or approximations - how likely it is that the things you've found from your data apply to a wider population.

1.5 A Word Of Warning!

There's a well known saying - `There are three kinds of lies: lies, damned lies and statistics' - which is indicative of the general mistrust most people have of statistics. There may well be a lot of truth in this. Statistics have often been used to `bend' or `slant' data in particular ways, and there are the odd instances when statistics are used to confuse and mystify people - actually to produce false results in some cases. But the main reason for this mistrust is because people wrongly think of statistics as producing an exact and absolutely accurate result, when in fact it can only indicate probabilities and other statistical predictions.

So, for example, an opinion poll may give Labour a 3% point lead over the Conservatives - and imply that Labour will win the election - when the actual figure could be, say, a 2% Conservative lead - because of the margin for error which always exists in any statistical prediction. Statistics are about probability and prediction - not about absolute certainties. Keep this in mind and you will be using them wisely and as they should be used - forget it and you will be unnecessarily led astray by them.

Section 2 - The Nature of Statistical Information

Basic Reading:

ROWNTREE **Statistics without Tears**: Chapter 2 or

LOETHER & MCTAVISH **Descriptive Statistics for Sociologists**: Chapter 2.

2.1 The Data Matrix

Let's look first at the nature of the information (or **DATA**) with which we shall be dealing. It will be composed, first of all, of a number of individual records (or **CASES**), one for each of the individuals, or organisations (or whatever) which make up our **SAMPLE**. Each of these cases will contain a number of items of information, each one representing a general social characteristic (or **VARIABLE**) common to them all. It is the responses to all of these variables in each of the cases which will go to make up the raw data, or **DATA MATRIX** of our survey. The responses themselves are often referred to as **VALUES**. For example, if we were to do a study of students on this course, each case (for each individual student) would contain information on, say, age, sex, qualification, previous occupation etc., and we would get a data matrix which looked something like this:

```
                         VARIABLES:

               <-------------------------------->

     CASES   | Sex | Age | Qualification | Last occupation    |
       |
       v    --------------------------------------------------
     Jane    |  F  |  19 | GCE `A' level | At school          | < V
     Bill    |  M  |  25 | ONC           | Unemployed         | < A
     Colin   |  M  |  30 | GCE `O' level | Skilled manual     | < L
     Mary    |  F  |  21 | CSE           | Skilled non-man.   | < U
     Sheila  |  F  |  28 | None          | No paid employmt.  | < E
     etc.    |     |     |               |                    | < S
```

Table 2.1

2.2 Levels Of Measurement

When you think of measuring things, you generally think of something like a tape measure or weighing machines. But social life is more complex than this and it is not so easy to think of ways of `measuring' it. To begin with, social variables are rarely quantitative in nature, being on the contrary usually non-numerical things like social class, religious denomination, or attitudes to divorce for example. In fact, a lot of the work in quantitative research involves thinking about ways of measuring - devising `yardsticks' - for characteristics which are, by their very nature, difficult to measure. Let's look at a few social variables and see how they can be measured:

(a) Age

You can simply take the actual age - 29, 35, 84 etc. - which would give you a lot of different values, each representing an actual interval in time. Such variables have an **INTERVAL** scale of measurement - each value represents a quantity, a certain number of things.

(b) Marital Status

You can't `measure' this in the same way as age - marital status is really a `quality' variable, and you can only give each possible answer - or **CATEGORY** - a label, such as `married', `single', `divorced' etc. Variables like marital status are called **NOMINAL** scale variables (meaning `name').

(c) Social Class

This is somewhere between an interval variable like age and a nominal variable like marital status. You can't simply give it a number - say `Managerial and Professional' = 1, but on the other hand, there is a definite sequence, or ranking from the highest social class (managerial and professional) to the lowest (unskilled or economically inactive). These variables are called **ORDINAL**, since the categories can be ranked in order, but there is no such thing as an interval or distance between categories.

Each of these levels has different statistical properties - and different implications for the sort of statistical analysis that can be performed on variables at that level:

	Properties ---->			Examples
	Labels	Order	Equal intervals	
NOMINAL	Yes	No	No	Religious denomination, tenure, sex
ORDINAL	Yes	Yes	No	Social class, highest qualification
INTERVAL	Yes	Yes	Yes	Age, I.Q., no. of children

Table 2.2

You often find that nominal and ordinal variables are referred to as **QUALITATIVE** variables and interval variables as **QUANTITATIVE**. Also, such quantitative variables are often subdivided into **DISCRETE** (or `counting') and **CONTINUOUS** (`measuring'). This is not so useful for the kind of data found in social research, since we rarely come across truly continuous variables.

A final word on levels of measurement:

To simplify matters a little, we won't actually be dealing with ordinal variables in this course - just nominal and interval. What we do is to look at a variable which is really ordinal and see if we can treat it either as nominal or as interval, depending on the circumstances. You aren't really supposed to treat an ordinal variable like an interval variable, but we find that it is sometimes convenient to do so - provided we are careful about interpreting the results.

2.3 Problems Of Measurement

The trouble with a lot of social information is that it is not readily converted into statistical data. There are a number of reasons for this, but they all stem from the fact that this kind of information can't always be easily fitted into neat categories. Even an apparantly basic variable like social class can present enormous problems: Do you take the last occupation for students? Supposing they came straight from school, do you take their parents' occupation then? What about the unemployed - do you take their spouse's occupational status? And so on. There are particular kinds of problems which we ought to recognise right from the start:

(a) Identification problems

There is obviously a limit to the amount of categories which we can use to label each variable and this can mean that our labels end up being too over-simplified. As a result we might find that we end up missing important relationships between them. Suppose we were looking at religion, for example, and we grouped all `methodist' denominations together under one category, and then we decided that we needed to look for a relationship with race. We would miss the fact that people of West Indian origin were overwhelmingly predominant in particular forms of `fundamentalist' Methodist denominations.

(b) Response problems

Another type of problem occurs when there is more than one possible answer to the same question. (Statistical data only allows one response for each a variable.) Thus, we could get people who are both `married' and `divorced' (from a previous spouse). It would be logical to choose the current status (i.e.married), but we would then under-estimate the true extent of divorce amongst our sample.

(c) Reliability

A reliable measure is one that is consistent - one that can be relied upon to give the same or similar results each time that it is used. If the same question were asked to the same kind of people in the same circumstances, would they give the same answer? This is the determinant of reliability. Ambiguously worded questions, insufficiently prepared interviews, poor quality transcription etc. are some of the many factors which can affect reliability and which must be taken into account.

(d) Validity

For a variable to be valid, it must be measuring the social concept which you are actually interested in. I.Q., for example, is a very useful predictor of educational performance, but does it actually measure `intelligence' (or just the ability to jump academic hurdles). Validity, then, is about ensuring that what is being measured is what you actually want to measure, and problems can arise when you find that you've ended up measuring something else instead. Contrasting these last two problems - reliability and validity - we can say that validity problems result from `imperfect' measurement, whereas problems of reliability occur when valid variables are measured imperfectly. Validity is more crucial, since it leads to systematic error which can't be corrected later, while reliability leads to random errors - which only affect the accuracy of the results.

Section 3 - Describing Univariate Data

Basic Reading:

ROWNTREE **Statistics without Tears**: pp. 38-39 or

LOETHER & MCTAVISH **Descriptive Statistics for Sociologists**: Chapter 3, Section 2.

3.1 Descriptive Statistics

Recall that we use descriptive statistics for summarising and making sense of a mass of data. One way of doing this is to group the data in various ways. Typically, you could - for example - find out how many people live in different kinds of housing, or compare different groups according to age, income, political preferences etc. You could also identify patterns and spot trends in the data - points of similarity, points of difference, things that are common to most, things that are unusual or out of the way, etc. This, then, is descriptive statistics - the ordering, grouping and summarising of data so that patterns, trends and possible relationships can be discerned.

We begin with the summarizing of a single variable (called **UNIVARIATE** data). Consider this simple example:

Say we have a group of 16 people, and - among other things we're interested in is their marital status. We ask them an appropriately-worded question and we get a series of replies showing us, as we can see, that the first person is married, the 4th divorced, the 11th single, etc.

1	Married	5	Single	9	Divorced	13	Separated
2	Single	6	Married	10	Widowed	14	Married
3	Single	7	Single	11	Single	15	Married
4	Divorced	8	Separated	12	No response	16	Separated

Table 3.1

3.2 Frequency Distribution

One very obvious thing we can do is to simply count up the numbers of people who give a particular reply and then group the data into the various different **CATEGORIES**. We then count the number of times each category occurs, ending up with the following set of figures:

Single	5
Married	4
Widowed	1
Separated	3
Divorced	2
No response	1

Table 3.2

When we group the information in this simple way - in terms of the number of responses in each category - we are listing what is called - rather grandly - a **FREQUENCY DISTRIBUTION**, but all it means is this simple grouping together of similar responses into their respective categories and then counting up how many times (how frequently) they occur - how many replies there are in each category in other words. This is an important and very common way of summarising univariate (single variable) data. This column of frequencies is usually referred to by a small letter `f', and it is common practice to include a **TOTAL** at the bottom. This is generally referred to by the capital letter

`N'. Now let's look at the frequency distribution for marital status again, just as it would look, say, on a page of computer print-out:

```
Marital Status              f
--------------              -
  Single                    5
  Married                   4
  Widowed                   1
  Separated                 3
  Divorced                  2
  No response               1
  -----------              --
  TOTAL (N)                16
```

Table 3.3

3.3 Percentage Distribution

The frequencies we have been dealing with so far are **ABSOLUTE FREQUENCIES** - the actual number of cases. But it's often useful to convert these frequencies into **RELATIVE FREQUENCES** - into the **PERCENTAGES** of cases in each category. These percentages give you an idea of the proportion of cases in each category, irrespective of the number in the sample. Almost all reports (in newspapers, magazines and learned journals), give summaries in this way: e.g. Party popularity in terms of percent voting (or saying they might vote) Conservative, Alliance or Labour. Let's look at marital status once again - this time with the extra column for percentages:

```
Marital Status              f          %
--------------              -         --
  Single                    5         31
  Married                   4         25
  Widowed                   1          6
  Separated                 3         19
  Divorced                  2         13
  No response               1          6
  -----------              --        ---
  TOTAL (N)                16        100%
```

Table 3.4

Note: Calculating Percentages You work out your percentage by multiplying your frequency by 100 and then dividing the result by the total number of cases. E.g. for `single' you multiply 5 by 100 (= 500) and divide it by 16 (= 31%). You can check that you've converted them properly by checking that they add up to 100% as they should (but see comment (a) below). One or two things to watch out for when using percentages: (a) It is normal when working out your percentages to ROUND them up or down to the nearest whole even number. Thus, in our example above the percentage for the`single' category actually worked out at 31.2% so we rounded it down to 31%. If it had been 31.5% we would have rounded it up to 32%, the nearest whole even number (likewise 30.5% down to 30%). (b) As a result of this, percentages sometimes DON'T add up to exactly 100%, but it is very unlikely for this factor (known as ROUNDING ERROR) to make a difference of more than one percent - i.e. to produce a total of less than 99% or more than 101%. In such cases just total it as if it was 100%, making a note saying `allowing for rounding error'. (c) When the total number of cases is small - as with our example - the percentages can be rather misleading. You should really stick to absolute frequencies unless you have a sample of at least 40 or 50 cases.

3.4 Allowing For Non-response

You often find another column in the frequencies distribution called ADJUSTED PERCENTAGES. This column takes into account those respondents who haven't really given an answer - they are NON-REPONSES. Thus, in our example, one person came into this category, and we put him/her into the distribution under `no response'. The `non-responses' should really have been put into one of the other categories, but we didn't know which one. So we include this column to give us a more accurate idea of the proportion of respondents in each category. So this is what our frequencies distribution for marital status looks like now:

			Adjusted
Marital Status	f	%	%
Single	5	31	33
Married	4	25	27
Widowed	1	6	7
Separated	3	19	20
Divorced	2	13	14
No response	1	6	-
TOTAL (N)	16	100%	100%

Table 3.5

Sometimes people have given an answer like `don't know' or `it all depends', for example, when asked a question such as `Who would you vote for?'. Such responses should not be treated as non-response, since they have actually given an answer, even if it didn't fit neatly into our scheme. So it is only when we don't know (ourselves) what the response was that we treat it as a non-response. Some common words used in relation to non-response: the non-respondents themselves are known as **MISSING CASES**, while those who have answered are called **VALID CASES**. Similarly, the category denoting non-response is called a **MISSING VALUE**.

3.5 Coding The Categories

Another feature of frequency distributions as produced by the computer is the CODE NUMBER of each category. Since computers work best with numbers, it is common to allocate a number to each category (e.g. 1 for `single', 2 for `married', etc.). These code numbers have no quantitative meaning at all: they are simply tags used by the computer to identify each category. When codes are added into the frequency distribution, the final result looks like this:

				Adjusted
Marital Status	Code	f	%	%
Single	1	5	31	33
Married	2	4	25	27
Widowed	3	1	6	7
Separated	4	3	19	20
Divorced	5	2	13	14
No response	9	1	6	-
TOTAL (N)		16	100%	100%

Table 3.6

3.6 Grouping Categories

We have shown how cases can be grouped according to categories in order to summarise the data in the form of a frequency distribution. But we can group the categories too, combining two or more categories into a single new category. For example, consider marital status again. We might want to combine - say - `separated' and `divorced' into one category, ignoring the distinction between them.

The distribution would then be a little simpler:

				Adjusted
Marital Status	Code	f	%	%
Single	1	5	31	33
Married	2	4	25	27
Widowed	3	1	6	7
Div./Sep.	4	5	31	33
No response	9	1	6	-
TOTAL (N)		16	100%	100%

Table 3.7

There are advantages and disadvantages in this approach. On the positive side, the distribution becomes clearer to understand, and the main differences more apparent. On the other hand, any combination of this sort is bound to lead to some LOSS OF INFORMATION which might turn out to be crucial at some later stage in our research. Others may also want to check the difference between two combined categories - and this would be impossible if only the combined figures are presented.

Section 4 - Measures of Central Tendency

Basic Reading:

ROWNTREE **Statistics without Tears**: pp. 43-50 or

LOETHER & MCTAVISH **Descriptive Statistics for Sociologists**: Chapter 5, Section 3.

4.1 Averages So far we've seen how we can summarise univariate (single variable) data by grouping the responses together into their different categories, in the form of **FREQUENCY DISTRIBUTIONS**. Now this is O.K. as far as it goes, but it's a somewhat limited and laborious way of describing data, especially when you're likely to have a lot of information and/or be interested in a lot of different variables - as is the case in most research. In order then to compress - or summarise - the data even further, the obvious next step is to work out an **AVERAGE** for it (known statistically as a measure of **CENTRAL TENDENCY** or of **LOCATION**). This `average' will then give you a single figure to represent - or summarise - all of the data. You will have heard, for example, of statements like the `average level of income' or the `average number of children per family' (often quoted as being 2.4) and, although a somewhat artificial figure (as you can see), nonetheless such figures do give us some real idea of what our data is telling us. The term `average' is actually rather a misleading one because there are in fact at least 3 different Measures of Central Tendency, and the term itself doesn't make it clear which one of them is being referred to. The 3 basic Measures of Central Tendency are then: (i) The Arithmetic **MEAN** - (`mean' for short) - what most people mean by `average'. (ii) The **MEDIAN**, and (iii) The **MODE**. You can use any of these measures for **INTERVAL** variables, but only the **MODE** for **NOMINAL** ones - the reason for this will become more obvious as we go along.

4.2 The Mean - The `Average' Value

Let's start by looking at how we would work out a **MEAN**. Suppose we take the incomes of 10 people (rounding them to the nearest thousand):

Case number	Income	Case number	Income
	£		£
1	4,000	6	7,000
2	5,000	7	2,000
3	7,000	8	4,000
4	8,000	9	10,000
5	12,000	10	7,000

To get the mean - we simply add up all of their incomes - giving us a total of £66,000 - and then divide by the total number of cases (N = 10), giving us a figure of £6,600. This is the **MEAN** - what most people think of when they use the term `the average'. Now, the advantage of the mean is that it is sensitive to every response - each single case makes a contribution to the final figure. The mean is therefore a very `democratic' measure of central tendency. Every case has just one `vote' and it contributes equally to the calculation. But there are times when this can actually be a disadvantage - particularly if you have a distribution with one or two extreme values. Let's look at what happens when we have such a situation. Suppose the last case was was £70,000 instead of £7,000, i.e.:

Case number	Income	Case number	Income
	£		£
1	4,000	6	7,000
2	5,000	7	2,000
3	7,000	8	4,000
4	8,000	9	10,000
5	12,000	10	70,000

Then the total for all of the cases would be £129,000 - and the mean £12,900 pounds - giving us an `average' which is higher than 9 out of 10 of the cases. This is clearly a very un-typical `average', distorted as it is by just one extreme and atypical response.

4.3 The Median – The ‘Middle' Value.

The second measure of central tendency is the MEDIAN. This is the middle value - the one which divides the sample into two EQUAL halves, with 50% of the cases on either side of it. Let's see how to go about getting a median figure. We start by putting the sample in order of income, with the lowest value first(i.e. £2,000) and the highest value last (i.e. £70,000), then simply divide the sample down the middle, putting half (5 in this case) on one side and half on the other. Here's the data again, re-arranged into the two halves:

Case number	Income £	Case number	Income £
7	2,000	6	7,000
1	4,000	4	8,000
8	4,000	9	10,000
2	5,000	5	12,000
3	7,000	10	70,000

The figure dividing the two halves - the value `in the middle' (or middle value) is the median. In this case it is £7,000.

[**NB:** People often come unstuck here. The important thing to remember is to ignore the relative size of the value. What you're interested in is which value is `in the middle' - regardless of its actual `size'. Sometimes we are not so lucky, and the highest value on one side isn't the same as the lowest value on the other. What we do in this case is to take a value mid-way between - i.e. split the two values down the middle. Thus if these two values were £6,000 and £7,000 the MEDIAN would have been £6,500.]

The MEDIAN then, as we can see, gets us over the problem of extreme values very nicely. The size of such atypical values doesn't influence our results at all.

4.4 The Mode - The `Most Common' Value

The third measure of central tendency - the MODE - means simply the most commonly occurring value or group - the response which occurs most often in other words. Thus, in our income data the value 7,000 occurred most often (3 times in all), and so represents the `modal' value for income. Although the mode can be used for all kinds of data, because we can't use the other methods (the mean or the median) for NOMINAL variables the most important use of the MODE is as a measure of central tendency for this kind of data in particular. Let's look at an example of a nominal variable - say marital status - and see how we get the mode. Here is the frequency table (copied from the last section):

				Adjusted
Marital Status	Code	f	%	%
Single	1	5	31	33
Married	2	4	25	27
Widowed	3	1	6	7
Separated	4	3	19	20
Divorced	5	2	13	14
No response	9	1	6	-
TOTAL (N)		16	100%	100%

Here the most common value - the value with the highest frequency - is code 1 (or `single'), with a frequency of 33%. Thus the mode, or modal value, is `single' and the modal frequency is 33%. Note the distinction between the mode itself and the modal frequency.

4.5 Grouped Interval Scales

When we have an interval variable - such as age - which has been grouped into a number of ranges, e.g. 15-29 years etc., we must calculate the mean in a slightly different way. We work out the mid-point of each range first. Then we multiply each mid-point by the absolute frequency (f). Now we can add this total and continue as before.

Here's how it's done:

Age Group	f	mid-point	f x mid-pt
15 to 29	24	22.5	540
30 to 44	36	37.5	1350
45 to 59	78	52.5	4095
60 to 74	42	67.5	2835
75 to 89	20	82,5	1650
TOTAL (N)	200		10470

Thus the mean is 10470 divided by 200, or 52.35 years of age. Similarly, the median is often calculated in a slightly different way. There are 200 cases in our sample, so the `middle' range must be `45 to 59 years': the range in which the 100th case lies. Now if we split the sample exactly in half, 40 of the 78 cases in this range would go into the lower half, and 38 in the upper half. So we take a sort-of average within this range to take account of this (slightly) uneven division:

$$\text{median} = \frac{(\text{lower limit x } 40) + (\text{upper limit x } 38)}{78} = 52.4$$

When you use the computer, though, you don't have to go through all these calculations. You just take the age variable (ungrouped) and ask for a frequency distribution requesting the relevant statistics for mean, median and - if you want - mode. **4.6 Ordinal Scales** So far we have looked at only nominal and interval variables. We can get `averages' for ordinal data also - but only for the mode and the median. Here's a summary table to show which measures are appropriate for each scale type:

SCALE	MODE	MEDIAN	MEAN
NOMINAL	Yes	No	No
ORDINAL	Yes	Yes	No
INTERVAL	Yes	Yes	Yes

Part 5 - Measures of Dispersion

5.1 DISPERSION DEFINED

Just as central tendency indicates a single measure for the most likely or the most typical value for a variable, so dispersion gives a single measure for the `spread' of the distribution about this value. A high value for the dispersion indicates that the distribution has a wide range of more-or-less equally likely values, where a small value for the dispersion indicates that the distribution is `peaked' about a narrow band of possible values. Note also that, in general each measure of dispersion is associated with just one measure of central tendency, but each measure of central tendency may have many measures of dispersion (or none at all).

References:
1. Rowntree, **Statistics without Tears**: Chapter 3;
2. Loether & McTavish, **Descriptive Statistics for Sociologists**: Section 5.3;
3. Blalock, **Social Statistics**: Chapter 5;

5.2 NOMINAL VARIABLES

There is no really valid measure of dispersion for nominal variables. A statistic known as the `index of dispersion' is sometimes used, but it is not available within SPSS, and has a fairly complex formula. It is more useful to look at the general distribution, and, for example, compare the mode with the category with the next highest frequency. If this difference is small, then we can expect a relatively high dispersion, while if it was low (i.e. one large category and possibly several smaller ones), then the dispersion would be low.

5.3 ORDINAL VARIABLES

Recall that we can use either mode or median for ordinal variables, and we see from the section on nominal variables above that there is no real measure of dispersion related to the mode. The median, on the other hand, does have a measure of dispersion associated with it - the inter-quartile range. It is derived from the quartiles of the distribution, which are the three values which divide the distribution into four equal quarters, just as the median split the distribution into two equal halves. In fact, the 2nd quartile (i.e. the middle one) is the median. The inter-quartile range is then the difference between the 1st and the 3rd quartiles.

Let us look at the variable `left-right scale' (see Fig. 5.1 below).

Just as we determined the median by looking at the cumulative frequency distribution, so we can pick out the three quartiles. These split the sample at the 25%, 50% and 75% points. We look down the cumulative frequency column until we find the category which gives a percentage higher than 25%. This is the category for the first quartile.

Then we look for the category topping 50%, etc. Thus, in the case of the left-right scale, the median (i.e. the 2nd quartile) is code 3, and the inter quartile range stretches from code 2 to code 4. (Note that we have not defined this range by a single number, since `interval' is not defined for ordinal variables. This is the only case where we do not use a single value or category to denote `dispersion'.)

```
V216      Q.24 Left or Right Politically[2]
```

[2] The data source is the Fifth Form Survey conducted in a North London 11-18 co-educational comprehensive school in Dec 1981. See Appendix for details.

```
                                Relative  Adjusted    Cum
                      Absolute    freq      freq     freq
Category label  Code    freq      ( % )     ( % )    ( % )

Left              1.         5       3.5       9.4       9.4
                  2.        15      10.6      28.3      37.7 ←1st
                  3.        14       9.9      26.4      64.2 ←2nd
                  4.        14       9.9      26.4      90.6 ←3rd
                  5.         3       2.1       5.7      96.2
Right             6.         2       1.4       3.8     100.0
                 -1.        89      62.7    Missing    100.0
                       ------    ------    ------
               Total     142     100.0     100.0

Mean       3.019      Median     2.964      Mode       2.000
Std dev    1.232      Variance   1.519

Valid cases      53      Missing cases    89
```

Figure 5.1: Using left-right scale to show inter-quartile range (output from old version of SPSS: arrows added for clarity)

5.4 INTERVAL VARIABLES

Since the median is defined for interval variables, we can use the inter-quartile range as for ordinal variables. We can also make use of the mean, which is only defined for interval variables. There are, in fact several measures of dispersion associated with the mean, but we shall only introduce two closely related ones here. Both depend on the concept of `deviation from the mean': that is, they are measures which use the distances between the value of each case and the mean value. The variance is calculated from the squares of the deviations from the mean: it is defined as the mean (or average) of the square of the deviations.

The squares are used because if we were to take just the average deviations, we would always get a result of zero. With the squares we would get a positive number (and only zero on a very special occasion: when all the values were the same). The variance is extremely important in the analysis of causal relations between variables.

The other measure of dispersion associated with the mean, the standard deviation, is simply the square-root of the variance. It is more useful than the variance as a descriptivestatistic since it is measured in the same units as the variable itself. It can be represented on the histogram as a range of values about the mean. If the mean was used as the measure of central tendency, then the standard deviation is the most appropriate measure of deviation.

5.5 SCORES

We now consider distributions of variables which have been derived or calculated from other variables in some way. First, a score is an interval variable calculated by summing or counting over a set of similar variables.

Take, for example, the set of variables defining gender attitudes in the fifth form study (See facsimile Q. 33 in Appendix).

It consists of 14 statements measuring opinions about women, some negative, some positive, with which pupils can agree or disagree on a 4-point scale.

We can devise a `sexist' scale from the set of variables since each of the variables measures - apart from anything else - the level of sexism inherent in the question. This `sexist' score can be calculated in a number of ways, depending on the level of measurement of the component variables:

1. Treating the variables as nominal: we choose as our target category `strongly agree' in the case of 'sexist' variables and `strongly disagree' for `non-sexist' variables. We then calculate the score value by counting the number of cases with the target category.

2. For ordinal variables we can choose a range of categories for our target response: e.g. `strongly agree' or `agree' for sexist variables. (see Fig. 2 for example).

```
SEXISM[3]
                                 Relative  Adjusted    Cum
                     Absolute      freq      freq      freq
               Code    freq       ( % )     ( % )     ( % )

                 0.          4       2.8       3.6       3.6
                 1.         23      16.2      20.7      24.3
                 2.         18      12.7      16.2      40.5
                 3.         16      11.3      14.4      55.0
                 4.         19      13.4      17.1      72.1
                 5.          6       4.2       5.4      77.5
                 6.          5       3.5       4.5      82.0
                 7.         11       7.7       9.9      91.9
                 8.          6       4.2       5.4      97.3
                 9.          3       2.1       2.7     100.0
                -1.         31      21.8   Missing     100.0
                        ------    ------    ------
              Total        142     100.0     100.0
```

Figure 5.2a: Frequency of Sexism Scale
(SPSS output - source: Fifth Form Survey 1981)

[3] Fig 5.2a is a copy of lineprinter output from older versions of SPSS. The data source is the same Fifth Form Survey . See Appendix for details.:

```
SEXISM[4]

       Code
            :
         0.  ***** (     4)
            :
         1.  ************************ (    23)
            :
         2.  ******************* (    18)
            :
         3.  ***************** (    16)
            :
         4.  ******************** (    19)
            :
         5.  ******* (     6)
            :
         6.  ****** (     5)
            :
         7.  ************ (    11)
            :
         8.  ******* (     6)
            :
         9.  **** (     3)
            :
            :.........:.........:.........:.........:.........:
            0         10        20        30        40        50
            Frequency

       Mean       3.559      Median     3.156     Mode   1.000
       Std dev    2.400      Variance   5.758

  Valid cases     111      Missing cases    31
```

Figure 5.2b: Histogram of Sexism Scale

(output from old version of SPSS. For Windows version 11 see Appendix 2)

From this you can see that SEXISM is positively skewed (i.e. the tail is pulled out towards the higher values)

3. Finally, as interval variables, we can use numerical values for each category (e.g. 1 for 'strongly disagree', 2 for 'disagree' etc.). We then make all non-sexist variables negative in value, and sum the values for all valid variables. We must, however, take care to correct the resulting sum for the number of missing categories: for example if three variables have missing values for a particular case, the score must be multiplied by the fraction 11/(11-3) or 11/8, where 11 is the number of variables in the set.

Scores are not only of interest of themselves: they are also very useful because of their distribution. They have the same characteristics as interval variables, and can be considered as continuous variables.

5.6 GROUPED DISTRIBUTIONS

Finally, we consider the distribution of a variable that has been grouped into fewer categories, or a continuous variable that is grouped into a finite number of categories.

[4] Fig 5.2b is a copy of lineprinter output from older versions of SPSS. The data source is the same Fifth Form Survey . See Appendix for details.:

The first consideration for such distributions is grouping error. This is a result of reducing the amount of information used to calculate the statistics of the distribution. The mean for a grouped distribution, for example , may vary widely from the mean of the original. In particular, the variance of a grouped distribution will always be less than the original variance, since the grouped distribution can make no estimate of the dispersion of a case within a particular group - the `within-group variance'.

Secondly, when grouping a variable that is inherently continuous, like age, a correction is often necessary to account for the difference between the value at the mid-point of each group and the value coded for the group itself. E.g. age is coded according to `age at last birthday', so the mid-point value for each `group' is 0.5 plus the code value.

Section 6 - Distributions of two variables

References:

1. Rowntree, **Statistics without Tears**: pages 150 to 155;
2. Loether & McTavish, **Descriptive Statistics for Sociologists'**: Chapter 6;
3. Blalock, **Social Statistics**: Section 15.4.

6.1 From one variable to two.

Having examined the distributions of each variable separately (univariate distributions), we can now look at the relationship between two variables (bivariate distributions). Take the example of the `sexism' scale derived earlier. Suppose we were to include the variable `gender' (or sex), in addition to the sexism scale. We could present this first as two separate univariate distributions: one for boys and the other for girls. These two distributions are called conditional distributions because they define the distribution of sexism conditional on gender.

The CROSSTABS procedure enables us to combine the two conditional distributions into one table (see Figure 1). Each column of the table represents one conditional distribution.

The rows correspond to the categories of the first variable (sexism), the columns to the categories of the second (gender). Each cell represents the simultaneous occurence of one category from each variable: e.g. girls who are not at all sexist (score 0) define a cell with 5 cases and 10.0% of all girls. At the right of the table is a column headed `row total', consisting of the univariate distribution of `sexism'. Similarly the `column total' at the bottom of the table is the univariate distribution of `gender'. The extra row and column are called the marginal distributions. Finally, the count of the total valid cases appears in the bottom right of the table.

6.2 DEPENDENT AND INDEPENDENT VARIABLES

For most purposes, we are interested in a possible causative effect within the relationship: if sexism is effected by gender, then the corresponding variables should show a relationship or association between them. The causing variable (or source variable) is called the independent variable and the effected variable the dependent variable. The normal convention is to use the columns to denote the independent variable (gender) and the rows to denote the dependent variable (sexism), but sometimes thisconvention is reversed (just to confuse you, but in any case the table won’t fit sideways on the page!).

```
             V348
            Count  :
            Col %  :Boys     Girls               Row
                   :                            Total
                   :     1  :     2  :    -1  :
SEXISM     --------:--------:--------:--------:
                0  :     0  :     3  :     1M :     3
                   :   0.0  :   6.3  :   0.0  :   3.1
                  -:--------:--------:--------:
                1  :     1  :    19  :     3M :    20
                   :   2.1  :  39.6  :   0.0  :  20.8
                  -:--------:--------:--------:
                2  :     6  :    10  :     2M :    16
                   :  12.5  :  20.8  :   0.0  :  16.7
                  -:--------:--------:--------:
                3  :     8  :     7  :     1M :    15
                   :  16.7  :  14.6  :   0.0  :  15.6
                  -:--------:--------:--------:
                4  :    11  :     5  :     3M :    16
                   :  22.9  :  10.4  :   0.0  :  16.7
                  -:--------:--------:--------:
                5  :     3  :     2  :     1M :     5
                   :   6.3  :   4.2  :   0.0  :   5.2
                  -:--------:--------:--------:
                6  :     3  :     0  :     2M :     3
                   :   6.3  :   0.0  :   0.0  :   3.1
                  -:--------:--------:--------:
                7  :     7  :     2  :     2M :     9
                   :  14.6  :   4.2  :   0.0  :   9.4
                  -:--------:--------:--------:
                8  :     6  :     0  :     0M :     6
                   :  12.5  :   0.0  :   0.0  :   6.3
                  -:--------:--------:--------:
                9  :     3  :     0  :     0M :     3
                   :   6.3  :   0.0  :   0.0  :   3.1
                  -:--------:--------:--------:
               -1  :    8M :    11M :    12M :    31M
                   :   0.0  :   0.0  :   0.0  :   0.0
                  -:--------:--------:--------:
           Column       48       48      27M      96
            Total     50.0     50.0      0.0   100.0

Number of missing observations =    46
```

Figure 6.1a: Sexism by Gender (Fifth Form survey)

(output from old version of SPSS (M = missing). For same output from SPSS11 for Windows 11 see Appendix 2)

Things are much clearer if we get rid of the missing cases on one or both variables, and we need to do this anyway before any statistics can be calculated.

6.3 CHARACTERISTICS OF ASSOCIATION There are four characteristics of association which together describe the relationship between two variables:

1. Existence of an association: i.e. whether, in fact an association exists. If there were no association, the table in fig. 1 would have column percentages the same in each row. We can define a characteristic of any pair of cells in the samerow by taking the difference between the cell percentages.These values are known by the general concept called `epsilon'.

2. Degree or strength of association: when the epsilon figuresare large, we speak of a strong degree of association between the two variables. Later we will discuss the problem ofdevising useful measures for this.

3. Direction of association. When we are dealing with ordinal ornominal variables, we can talk about the direction of association: if most cases appear in the diagonal from top left to bottom right we have a positive association. The other diagonal would indicate a negative association. For example, the top left and bottom right cells of our table are both zero,and a clear negative relationship between sexism and girls exists (i.e. girls are less sexist than boys).

4. Finally, nature of association is a more descriptive characteristic, defining the general pattern of an association:e.g. in our example, we could say that (1) the vast majority of girls have scores under 2 (78%) compared to only 35% ofboys; (2) that a large minority of boys have scores over 4 (43%) while only 4% of girls are in this range; and (3) that few boys or girls have intermediate scores (22% of boys and 18%of girls).

6.4 INDEPENDENCE - THE `NO ASSOCIATION MODEL'

We can define perfect independence as a situation in which allthe cell epsilon values are exactly zero. This is the no association model. In practice, we can never expect to findthis even when there is, in fact, no association between the underlying concepts. This leads us to the problem: how different do the percentages need to be before we can deduce that a relationship exists? The answer, which we can never really be sure about, depends on two criteria:

1. Theoretical considerations: i.e. how large a difference would be meaningful from the point of view of underlying sociological theory. It may be, for example, that a very small difference in percentages has a potentially large effect (e.g. election polls). Alternatively, we may think that even a very large difference does not mean much.

2. Statistical considerations concern the size and representativeness of the sample of cases. A small or unrepresentative sample would require substantial differences before we can declare an association `statistically significant'. However, these statistical considerations are to do with inferental statistics, and will be investigated later in the course.

6.5 EPSILON AND DELTA

We end this session with two useful measures of association (n.b. a more complete description will be discussed next session). First, the statistic **epsilon** mentioned earlier. This is defined as the difference between any two cell column percentages in the same row. Thus, the epsilon value for the difference between boys and girls for the sexism value 3 is 14.6 minus 16.7 or -2.1 **percentage points**. Ignore the minus sign for now, treating the difference in absolute terms: there is a 2.1 percentage point difference in the code 3 response between girls and boys. As a general (and very crude) rule, differences of over 10 percentage points are usually meaningful, and anything lower may be attributed to chance factors causing the difference. In this case, the value of 2.1 percentage points indicates a very low chance that there is any difference between boys and girls for this value of sexism. Now, the sign in front of the epsilon value can be used as an indicator of the direction of the difference. A positive sign indicates that there are more cases in the first column, a negative sign that there are more in the second. Thus, if we take the difference between boys and girls for sexism code 1, we get an epsilon of -37.5 percentage points, indicating that girls are much more likely to score 1 on the scale than boys. Another useful, but more complicated, measure of association uses the `no association table'. This table is produced from themarginals, or original univariate distributions, using the assumption that there is no relationship between the two variables (i.e. all the epsilon values are zero, or equivalently, the two sets of column percentages are matched row for row). This is then subtracted from the actual cell frequency to give us the delta value for each cell:

$$\text{Delta} = \text{no. in cell} - \frac{\text{row total X column total}}{\text{overall table total}}$$

The delta values for the table in figure 1 are given below in figure 2. The lower the value, the smaller the difference between `observed' and `expected' cell frequencies. A value of zero would indicate that the frequencies are the same, a value of 4 - say - would indicate that they differ by 4 cases, and so on.

```
                  Delta   :Boys     Girls      Row
                          :                    Total
                          :     1   :     2  :
         SEXISM   --------:---------:--------:
                       0  :  -1.5   :   1.5  :     3
                         -:---------:--------:
                       1  :  -9.0   :   9.0  :    20
                         -:---------:--------:
                       2  :  -2.0   :   2.0  :    16
                         -:---------:--------:
                       3  :   0.5   :  -0.5  :    15
                         -:---------:--------:
                       4  :   3.0   :  -3.0  :    16
                         -:---------:--------:
                       5  :   0.5   :  -0.5  :     5
                         -:---------:--------:
                       6  :   1.5   :  -1.5  :     3
                         -:---------:--------:
                       7  :   2.5   :  -2.5  :     9
                         -:---------:--------:
                       8  :   3.0   :  -3.0  :     6
                         -:---------:--------:
                       9  :   1.5   :  -1.5  :     3
                         -:---------:--------:
                  Total         48        48      96
```

Figure 6.2: Delta values for Sexism by Gender

Neither of these values (epsilon and delta) can be obtained directly from the SPSS CROSSTABS procedure at the moment, so need to be calculated from the printout. However, both are quite useful in interpreting the nature of the association, as distinct from the degree of association.

Section 7 - Measures of Association

References

1. Rowntree, **Statistics without Tears**: pages 155 to 164;
2. Loether & McTavish, **Descriptive Statistics for Sociologists**: Chapter 7;
3. Blalock, **Social Statistics**: Sections 15.4 & 18.4.

7.1 DEGREE OF ASSOCIATION

Last session we covered the four characteristics of association in general and examined two of them in some detail (i.e. Existence & Nature). This time we look at Degree (or strength) and Direction. The former gives a single measure of the strength of the relationship between the two variables, while the latter specifies the direction of association (if appropriate). We begin by looking at the possible combinations of the two levels of measurement (see Figure 1). Each variable can be either nominal, ordinal or interval - giving a total of 9 possible combinations.

		Independent Variable		
		Nominal	Ordinal	Interval
Dependent Variable	Nominal	Chi-square Cramer's V	←	←
	Ordinal	↑	Gamma Somers' d	←
	Interval	Eta	↑	Pearson's R

Figure 7.1: Choosing the Appropriate Measure of association

Not all possibilities can, however be dealt with by SPSS. We shall consider each possibility in detail according to dependent variable, giving a brief explanation of the measure to be used in each case.

7.2 NOMINAL DEPENDENT VARIABLES

Recall that we used Delta as an indication of the difference between the actual table and the table for`no association' for each cell. The measure, chi-square, uses the same idea, but for the table as a whole. It is by far the most commonly used (and abused) measure in social statistics. Chi-square is obtained from the cell delta values by taking their squares, dividing by the expected cell frequency (for the no association) model and summing over all cells:

$$\text{Chi-square} = \text{Sum of} \left(\frac{\text{Delta X Delta}}{\text{Expected cell frequency}} \right)$$

Chi-square can range from 0 (perfect match with the no-association model) to well over the number of cases in the table, so it is not always easy to tell really just how strong the association is from the chi-square value itself.

Standardisation

This brings us on to one of the recurring features of descriptive statistics, that of standardisation (or normalisation). In order to locate any measure in a meaningful range of values, it is standard practice to choose the value 0 to indicate "no" and the value 1 to mean "yes" or "perfect". In the case of chi-square, a measure called Cramer's V has been derived from the original measure in order to standardise it. A value of 0 for Cramer's V then indicates no association, while a value of 1 would indicate perfect or complete prediction (i.e. once the value of the independent variable is given, the dependent variable can be predicted with absolute certainty).

[NB: On the SPSS print-out, a measure called phi is sometimes produced instead of Cramer's V. Don't worry! This is just a special name for Cramer's V with 2x2 tables.**]**

Ordinal or Interval Independent Variables

SPSS does not produce specific measures when the dependent variable is nominal and the independent variable is not. There are two ways of getting round this problem: (i) for 2x2 tables the dependent variable can normally be assumed to have the same level as the independent variable, and one of the measures described below can be used; (ii) otherwise the independent variable must be treated as nominal (losing some information about the relationship between values) and chi-square or Cramer's V can be used.

Direction and Nominal Measures

Before moving on to ordinal measures, it is important to note that direction of association has no meaning for nominal variables (no category can be higher or lower than another), although it does apply when nominal variables are not used. Thus, all standardised measures for nominal variables range from 0 to 1, whereas standardised measures for other levels normally range from -1 through 0 to +1, with the sign indication the direction of the association.

7.3 ORDINAL DEPENDENT VARIABLES

We consider first the case where both dependent and independent variables are ordinal. Nearly all measures of association for ordinal variables make use of the comparison of all possible pairs of cases in the sample. Pairs which preserve the same order over both variables are called concordant, while pairs which have the opposite ordering are called discordant. All other pairs are tied in some way. Consider Figure 2 below as an artificial illustration. Cases A and C form a concordant pair: satisfaction decreases from past to present in both cases. Conversely, cases A and B form a discordant pair, while case C `ties' with A on satisfaction now and with B on satisfaction past.

	Case A	Case B	Case C
Satisfaction past	very	fairly	fairly
Satisfaction now	not very	very	not very

Figure 7.2: Illustration of Ordinal Prediction

Restrictive Meaning of Perfect Association

The specific measure of association used for ordinal variables depends largely on what we mean by `perfect association'. In the more restricted sense, we consider only those relationships when we can predict with certainty to be perfect association relationships. These should always be square tables (with the same number of rows as columns), and with empty cells everywhere except on one (and only one) of the diagonals. Figure 3 illustrates the form of a restrictive perfect association.

```
             Low     High               Low     High
            :---:---:---:              :---:---:---:
      Low   : X :   :   :        Low   :   :   : X :
            :---:---:---:              :---:---:---:
            :   : X :   :              :   : X :   :
            :---:---:---:              :---:---:---:
      High  :   :   : X :        High  : X :   :   :
            :---:---:---:              :---:---:---:

         Perfect Positive          Perfect Negative
```

Figure 7.3: Illustration of Restrictive Perfect Association

Somers' d

The basic measure of association for ordinal variables which uses this more restrictive meaning is Somers' d. It ignores all pairs which are tied on the independent variable, and compares concordant with discordant pairs. The direction of association depends on the sign of d. If there are more discordant than concordant pairs, Somers' d is negative, indicating that the dependent variable tends to get lower as the independent variable gets higher. You will notice that SPSS gives two kinds of Somers' d values: asymmetric, with each variable in turn defined as the dependent variable, and symmetric where only one value is given. It is the asymmetric version that we mean here - the symmetric version is merely a combination of the two asymmetric measures. Such a combination is sometimes useful, for example, when one is trying to get an idea of a general association, rather than assuming an inherent causal relationship. If so, then it is betterto use another `combination' of asymmetric Somers' d called Kendall's Tau b. This is somewhat analagous to Cramer's V (for 2x2 tables the results are the same). However, Kendall's Tau b assumes a rectangular table. If you haven't got one, you should use Kendall's tau c instead.

Less Restrictive Meaning - Gamma

Quite often, especially in sociological applications, we are not interested in the rigorous restrictive definition of `perfect association'. We only want to know whether there are few who are not where they should be, rather than most where they should be. With ordinal variables, the more restrictive meaning of perfect association demands that all pairs should either be concordant (for a positive relationship) or be tied on both variables. The less restrictive version only demands that there should be no discordant pairs, and ignores the presence of ties of any kind. The usual measure for this less restrictive model is Gamma. It is always at least as high as any corresponding more restrictive measure (such as Somers' d), and often quite a lot higher. With 2x2 tables, Gamma becomes equivalent to another common measure, called Yule's Q.

Association with variables at other levels

Firstly, if we want to measure an association with a nominal independent variable, we should use nominal measures of association (ignoring the ordering inherent in the dependent variable), unless the independent variable is binary (only two categories) when the ordinal measures can be used. Secondly, if the independent variable is interval, we should still use an order-based measure.(Note: frequently, when the level is at least at an ordinal level, purely interval measures are used. This is not strictly applicable, but sometimes very useful as an initial summary of a relationship.)

7.4 INTERVAL DEPENDENT VARIABLES

The possibilities for analysing interval dependent variables are much greater than for the preceding cases. This becomes more so when we move on to three or more variables. But for now, we will just consider some of the most basic measures, restricting ourselves to the STATISTICS available in the CROSSTABS procedure.

(a) Nominal Independent Variable

There is a special measure, called eta, which has been developed specifically for this case (i.e. interval dependent variable and nominal independent variable). SPSS produces two values for eta, depending on which variable is dependent (row or column). Make sure you use the right one!

(b) Ordinal Independent Variable

There is no specific measure for this combination, and the eta measure defined above is the most appropriate. Sometimes the ordinal variable is assumed to be interval, but (as mentioned before) this is not strictly permissible from a statistical point of view.

(c) Interval Independent Variable

The final case we consider is when both variables are interval, when the appropriate measure is Pearson's r (often just called the correlation coefficient). A more thorough approach to the relationship between interval variables uses the method of regression which will be covered later in the course (but see Rowntree pages 176 to 184 if you are interested).

Part 8 - Measures of Association - Interval Variables

8.1 NOMINAL OR ORDINAL INDEPENDENT VARIABLES

We begin with the association between an interval dependent variable and a nominal or ordinal independent variable. Taking age in years against sex as an example, we can represent the association in terms of a number of conditional distributions (i.e. male or female), together with the marginal distribution for age as a whole (see fig 1)

[Where fig.1? It's disappeared!].[5]

These conditional distributions are commonly summarised in the form of a breakdown table (see fig. 2) containing the sum, mean, standard deviation and variance of the dependent variable for each conditional distribution and for the marginal distribution. The table also normally contains a count of the number of cases in each category of the independent variable. In addition, the SPSS procedure BREAKDOWN (now changed to MEANS) optionally produces (using STATISTICS 1, but now changed to STATISTICS ONEWAY) an 'analysis of variance' table (see fig. 3) which includes the sum-of-squares (short for the sum of square deviations of each case about the mean).

[Where figs. 1 and 2? They've disappeared as well!].

8.2 ETA SQUARED

The usual PRE measure for interval dependent variables and nominal or ordinal independent variables is eta-squared, which is derived as follows:

Guess 1: Let the value of the dependent variable for each case be the mean for the marginal distribution.

Guess 2: Let the value of the dependent variable be the mean for the conditional distribution corresponding to the category of the independent variable.

The error due to guess 1 is the sum-of squares for the marginal distribution, and for guess 2 the sum of the sum-of-squares for each conditional distribution (called the 'within-group sum-of-squares'). The formula is given by:

```
               total sum-of-squares - within-group sum-of-squares
Eta-square  =  --------------------------------------------------
                               total sum-of-squares
```

8.3 INTERVAL INDEPENDENT VARIABLES

When both variables are interval, we can look at the distribution in terms of a scatter-diagram (see fig. 4), in this case of sympathy with coloured immigrants by sympathy with coloured people born round here. The left hand [vertical] axis measures sympathy with coloured immigrants (the dependent variable(c) while the bottom [horizontal] axis measures sympathy with coloured people born round here (the independent variable).

[Where fig.4? Which data set?]

[5] Various references to figures in section 8, but none in text: not clear which data are used (Fifth form survey, SSRC Quality of Loife surveys) can they be reproduced? Must try to find an original or recreate something.

Each case is represented by a star on the diagram which corresponds to the particular pair of values for the case, and where two or more cases occupy the same point, the star is replaced by the appropriate number.

8.4 EXAMINING THE SCATTER-DIAGRAM

First, we need to obtain the statistics for the marginal distributions of each variable. We can then visualise a measure of association using the scatter-diagram as follows:

1. Plot the point corresponding to the joint means of both variables (point M in fig. 5). This is roughly central within the plot of points.

2. Divide the diagram into four quadrants about the joint means. Make the top right and bottom left quadrants positive and the other two negative.

3. Consider a particular case at point A in the diagram. We can construct a rectangle with the points A and M at opposite corners. We then note the area of this rectangle, giving it either a positive or a negative sign depending on the quadrant in which it is situated.

4. Finally, we add all these areas (including the appropriate sign(c) over all cases in the sample. Dividing by the number of cases minus one, we get the covariance for the two variables.

The covariance is given by the following formula:

$$\text{Covariance} = \frac{\text{Sum over cases } [(X - \text{mean of } X) \times (Y - \text{mean of } Y)]}{\text{Number of cases} - 1}$$

where X and Y symbolise the dependent variable and the independent variable respectively.

The more cases that are situated in the positive quadrants, the higher the covariance and similarly if most cases are in the negative quadrants, we get a high negative covariance.

Another measure of association, related to the covariance, is Pearson's product-moment correlation (or simply Pearson's correlation). It is given by the covariance divided by the standard deviation of both variables. It 'normalises' the covariance, giving a value between -1 and +1, with zero indicating no correlation.

8.5 LINEARITY AND PRE MEASURES

When we use the guessing rules for two interval variables, we must consider the prediction in the form of an equation linking the value of the dependent variable to a formula containing the value of the independent variable (i.e. a function). The usual formula is in the form of a linear equation, as follows:

$$Y = b \times X + c$$

where Y represents the value of the dependent variable, X the value of the independent variable, and b and c are two 'parameters' which represent fixed, but as yet unknown, values.

This equation can be represented in the scatter-diagram as a straight line (see fig. 6), where the parameter c is indicated by the point on the vertical Y axis where the line crosses it, and the parameter b is represented by the 'slope' of the line (i.e. the distance measured vertically when a point on the line is shifted by one unit on the horizontal axis).

We can predict the value of Y for a given value of X (assuming that the parameters b and c are known) as follows:

1. Mark off the value of the independent variable on the horizontal axis.

2. Extend a line vertically from this point to a point on the line representing the relationship (point B).

3. Then extend a line horizontally (from point B) to the vertical axis.

4. Read off the value on the vertical axis as the predicted value of the dependent variable.

The position of this line (i.e. the parameters b and c) is determined (1) by ensuring that it passes through the point representing the joint means (i.e. point M) and (2) by making the sum-of-squares difference between the predicted value (on the line) and the actual value of the dependent variable as small as possible.

Finally, the PRE measure for two interval variables is determined as follows:

Guess 1: Choose the mean of the dependent variable in every case.

Guess 2: Choose the linear prediction as above for the value of the dependent variable.

It turns out that this PRE measure (called R-squared) is the square of the Pearson correlation described earlier. The error for guess 1 is called the total sum-of-squares and the error for guess 2 is called the unexplained sum-of-squares, so that:

$$R^2 = \frac{\text{Total sum-of-squares - Unexplained sum-of-squares}}{\text{Total sum-of-squares}}$$

9 Elaboration

Having discussed the relationship between two variables, we move on to investigate the effect of other variables on this relationship. This is the elaboration process.

References:

Loether & McTavish, Ch 8

Bowen & Weisberg, Ch 8

Blalock, Sect. 15.4 and Ch 20

Rosenberg **The Logic of Survey Analysis**

Moser & Kalton, Ch 17, Sect 4

9.1 The Control Variable

The additional variables used to investigate the original relationship are called control variables or test variables. The effect of a test variable T on the relationship between an independent variable X and a dependent variable Y can be represented diagrammatically as follows:

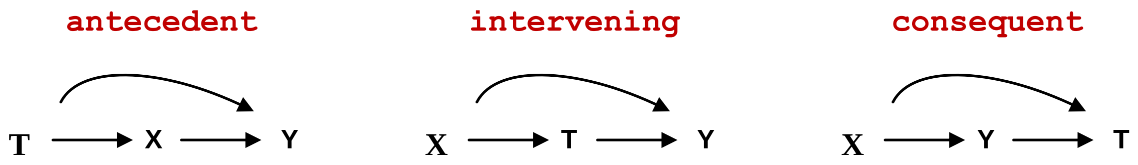


Fig 9.1 Three models for one control variable

An antecedent model has the control variable as a causal factor for both X and Y; an intervening model has the control variable as a causative factor for the dependent variable Y, but is itself affected by the independent variable X; finally a consequent model has the control variable as an effect of both X and Y. The choice of model - antecedent, intervening or consequent - depends entirely on the nature of the underlying theory and the specific hypotheses you want to test. Statistical analysis cannot by itself determine which model is correct in any particular instance: it can only be used to investigate the form of the associations (or lack of associations) between variables once the model has been defined.

9.2 Conditional Tables

We obtain conditional tables by dividing the sample into two or more groups according to the value of the control variable. Each conditional table thus produces a description of the relationship between X and Y for each value of T. These three-variable tables are known as first order tables (one control variable) as distinct from the original two-variable tables which are known as zero-order tables (no control variable). If we introduce a second or even a third control variable, this produces second- or third-order tables and so on for as many test variables as are included in the model.

In addition to these conditional tables, it is also useful to examine the original (zero-order) associations between the test variable T and each of the original variables X and Y. We then have a more or less complete picture of the whole model. This enables us to produce three kinds of measures of association (or PRE measures) as follows:

1 Original total association: between X and Y

2 Conditional associations: between X and Y for each value of T

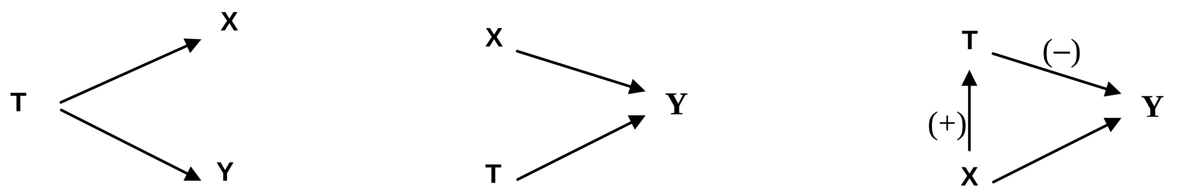
3 Total associations with between X and T and the control variable: between Y and T

By comparing these different measures of association, and bearing in mind the model of the relationships between variables, we can then investigate in detail the effect of the test variable. For example, three possible outcomes might be:

1 *Spurious relationships*: in an antecedent model when the original total association is strong, but the conditional associations are both weak. The total associations with the control variable will also have been strong.

2 *Independent causation*: in an intervening or an antecedent model, when the original total association is weak, the conditional associations are strong, and the total associations with the control variable are strong.

3 *Suppressor control* variable: in an intervening model, when the original total association is as strong as each of the conditional associations (i.e. they are all about the same), and the total associations with the control variable are weak for the independent variable X, but strong for the dependent variable Y.



spurious relationship

independent causation

suppressor control variable

Fig 9.2 Examples of possible relationships

9.3 Partial Measures of Association

In many of the above examples, we are interested only in the summary of the conditional associations - i.e. we want an "average" of the conditional measures of association. This average is most often achieved by means of a partial coefficient of association or partial correlation coefficient. If the original conditional associations were PRE measures, then the partial coefficients are also usually PRE measures. As with conditional associations, we can also define zero-, first- etc. order partial correlations. For example SPSS produces partial gamma coefficients.

More often partial coefficients are used for interval dependent variables, especially when either or both of the independent and control variables are also interval. This type of coefficient, normally referred to as partial correlation, is particularly appropriate to regression analysis and analysis of variance.

10 - Inferential Statistics

10.1 THE BASIS OF STATISTICAL INFERENCE

Many research problems in Social Science concern the generalization from the observation of a particular, relatively small quantity of information. Statistical Inference provides a method for this generalization process by means of which inferences about the population under investigation are made on the basis of information obtained from a sample taken from this population.

References:

1. Rowntree, "Statistics Without Tears", Chapter 5.

2. Loether & McTavish, "Inferential Statistics", Chapters 3, 4.

3. Bowen & Weisberg, "Introduction to Data Analysis", Chapter 10.

4. Blalock, "Social Statistics", Chapters 8, 10 and 11.

10.2 FOUR STEPS IN STATISTICAL INFERENCE

The procedure for statistical inference normally involves four steps, even though some of these steps may only be 'assumed':

1. Th*e Statistical Model*: making assumptions about the expected behaviour of the population as a whole.

2. *The Sampling Distribution*: making assumptions about the expected behaviour of a particular sample.

3. *Estimation*: forming an estimate of a population value for the statistical model, on the basis of the observations from the sample.

4. *Decision*: using the estimate to make a decision about the population as a whole.

We shall be covering the first two topics this session, and the other two next session, continuing with specific examples thereafter.

10.3 THE STATISTICAL MODEL

The first step in the inferential process is to produce a Statistical Model which defines the expected behaviour of the population as a whole. The statistical model usually contains the following components:

1. The distribution of each variable of interest, or the joint distribution of all the variables under interest. This is typically in the form of a standard theoretical distribution, such as the Normal distribution.

2. A formula expressing the assumed relationships between variables, linking dependent variables to independent variables. Such a formula defines both the causality of relationships and their mathematical form.

3. Either of these two components may contain a number of fixed but unknown values - parameters - whose values are to be estimated during the inferential process.

EXAMPLE

We wish to examine the differing attitudes towards women amongst adolescents. We begin by assuming a model that links these attitudes to sex, and we therefore take the following two variables:

1. An attitude to women index consisting of the number of positive statements concerning attitudes to women from a list of nine possible statements.

2. Self-reported sex

We then assume that the joint distribution of the two variables takes the form of two distinct Normal distributions of the attitude to women index, one for boys and one for girls.

This gives us a total of five possible parameters:

1. The proportion (or percentage(c) of the population who are girls.

2. The mean of the index for girls.

3. The mean of the index for boys.

4. The standard deviation of the index for girls.

5. The standard deviation of the index for boys.

For this example, we do not need a specific formula linking the two variables, since the information can already be obtained from the parameters. But if we wanted to specify such a formula, it would take the form:

Expected index value = mean value for girls, if a girl, or mean value for boys, if a boy.

10.4 THE SAMPLING DISTRIBUTION

The second step involves the concept of estimators. These are statistical formulae which produce a given statistic from a particular sample. This statistic is such that it provides a best estimate for a given population parameter.

This process requires a great deal of statistical theory in order to explain it in any detail, and so we will only describe the essential features:

1. We consider all the possible samples which could have been collected using the same method as for the actual sample. This gives us our 'population' of samples.

2. Any particular estimator for a parameter would produce an estimate for each of these possible samples. The estimate can be thought of as a derived variable, and the estimator as the formula for deriving the variable.

3. We could then envisage the distribution of the estimates over all possible samples as a theoretical statistical distribution. This is called the Sampling Distribution of the estimate. Its mathematical form can be calculated on the basis of (1) the statistical model and (2) the method of collecting the sample, once a formula for the estimator is given.

4. The problem then becomes one of finding the 'best' estimator for the parameter in question. This is done first by considering the mean of the sampling distribution. We look at all possible estimators which have the mean equal to the population parameter itself - these are the unbiassed estimators.

5. Finally we choose the unbiassed estimator which has the smallest variance of the sampling distribution. This gives the minimum variance unbiassed estimator. The standard deviation of the sampling distribution for this estimator gives us the standard error of the estimate. Statistical theory also demands that the estimator be 'sufficient' for the computation of the parameter concerned - i.e. it does not contain any unknown elements.

EXAMPLE

1. Take the sample of 142 fifth-formers and treat it as our 'population'.

2. We take a number of 'samples' of 10 selected randomly from this population, using the SPSS facilities SEED and SAMPLE.

3. We estimate the percentage of girls for each sample using the estimator:

$$\text{Estimator} = \frac{\text{number of girls in sample}}{\text{number who gave their sex}} \times 100$$

This can be obtained by running the FREQUENCIES procedure on the selected sample.

4. The list of estimates can give us an idea of the form of the sampling distribution for percentage of girls.

5. In particular, the mean of the distribution should be close to the percentage for the whole 'population'.

6. Finally, the standard deviation of the distribution should be close to the average 'standard error' for each sample.

10.5 REPRESENTATIVE SAMPLES AND WEIGHTING

In practice, most computer packages - including SPSS - work on the assumption that the sample is representative. This would be true if each case in the sample were collected randomly, or in such a way that it could be assumed to be random (e.g. systematic sampling).

It is not always possible, or practible, to be able to use a truly representative sample. Then each case must be weighted, in inverse proportion to its chance of being chosen for the sample. If, for example, we were sampling households in both urban and rural areas, and urban households had twice as much chance of being selected as rural, then we would give a weight for rural households twice that for urban households. In addition, it is useful in statistical theory if the average weight over the whole sample is exactly one. Thus, for our households, we would give each rural household a weight of 4/3 and each urban household a weight of 2/3.

QUESTION

Suppose we discovered that the population of fifth-formers from which the survey was taken had exactly 50% boys and 50% girls. What weights would we apply to each case, given that there are 56 boys and 59 girls, with 27 not answered?

11 - Estimation

References

1. Rowntree, 'Statistics without Tears': Chapter 5, pages 155 to 164;

2. Loether & McTavish, 'Inferential Statistics for Sociologists': Chapter 4;

3. Blalock, 'Social Statistics': Chapter 12.

11.1 OVERVIEW

Last session we examined the general principles of statistical inference. We have a statistical model of the population and a sample selected from that population. Probability theory enables us to make decisions about the population on the basis of the data collected from the sample. Such decisions normally take two forms:

(i) the estimation of fixed, but unknown, quantities (parameters) which form part of the statistical model, and

(ii) the testing of a statistical hypothesis about the model.

This latter type of decision (hypothesis testing) involves the former (estimation) as a preliminary stage. So this session's lecture will concentrate on estimation.

We shall look at estimates of four kinds of parameters: proportions (or percentages)" means" standard deviations and Pearson correlations. For each, we go through three of the four steps outlined last week: the statistical model, the sampling distribution and the estimate itself. But first, we examine briefly each of these three steps, specifying what is required from them.

11.2 Formulating The Statistical Model

1. How many variables are involved? [One; two; three or more]

2. What level of measurement for each variable? [Nominal; ordinal; interval]

3. What is the expected or assumed distribution for the variables? [Binomial; normal; Poisson; multinomial; multivariate normal or unknown]

4. For two or more variables: which variables are dependent and which independent?

5. How many fixed but unknown quantities are included in the model (i.e. how many parameters)?

There are many other components of models where three or more variables are involved, but we shall not be dealing with these cases.

11.3 Deriving The Sampling Distribution

1. What is the size of the sample?

2. What is the selection procedure for defining the sample? [Random; stratified random etc.]

3. What is the formula for estimation of each parameter (i.e. the estimator) given the statistical model and the sampling procedure? [You will have to look in a text-book for this one!]

4. What is the theoretical distribution of the result of this formula (i.e. the sampling distribution of the estimator)? [Each estimator is usually associated with a particular sampling distribution: see the text-book again!]

One useful result of probability theory comes in here: the Central Limit Theorem tells us that if we have a large enough sample (usually more than 50) then the sampling distribution can be assumed to be normal.

11.4 Examining The Estimate

1. The actual value of the estimate is given by the result of the estimator formula for a particular sample.

2. The standard error of the estimate - defined as the estimate of the standard deviation of the sampling distribution - is obtained from a formula akin to that for the estimator itself. For large samples, the standard error usually has a standard form:

```
                             Sample standard deviation
Standard error   =   ------------------------------------
                     Square-root of number of valid cases
```

3. To make a useful conclusion about the possible range of values for the estimate, we can derive a confidence interval for the estimate. We first choose a level of confidence suitable for the problem at hand: for example, we want to be confident that 95% of all possible samples will produce confidence intervals which actually include the parameter. Then only once in twenty are we going to get a sample where the parameter is outside this range. For normal sampling distributions (i.e. samples with 50 or more valid cases(c) the 95% confidence level gives a range of approximately twice the standard error about the estimate" and the 99% level gives a range of about 2.5 times the standard error.

11.5 PROPORTIONS

We use the sex of the respondent as a working example:

Sampling Distribution: The normal distribution (the binomial distribution is more appropriate if valid cases X expected proportion is less than 5), assuming a random sample.

Estimator:

```
                             Number of girls in sample
expected proportion   =   -----------------------------
                               Number of valid cases
```

Statistical Model: A two-valued (i.e. binary) variable, with a binomial distribution for the population" one parameter (i.e. proportion girls), fixed but unknown.

Estimate:

```
                        59
Estimated proportion =  ---   =  0.51  (or 51%)
                        115
```

```
      Standard error:

                              (Number of girls X Number of boys)
                       Sq.  root of (--------------------------------)
                              (      number of valid cases      )
  Standard Error = ---------------------------------------------
                                      number of valid cases


                                (59 X 56)
                         Sq.  root of (-------)
                                (  115  )   5.36
                       = --------------------- = ---- = 0.05  (or 5%)
                                    115              115
```

11.6 MEANS

We use the derived score 'attitude to women' to illustrate both the mean and the standard deviation.

Statistical model: An interval variable, with a normal distribution for the population, two parameters, mean and standard deviation, both fixed but unknown.

```
    Estimator:

                             Sum of valid scores
         Expected mean   =  ---------------------  (= sample mean)
                            Number of valid cases
```

Sampling distribution: The normal distribution, given a random sample (n.b. if the number of valid cases is less than 100, we would use the Student's-t distribution).

```
    Estimate:

                                   668
                Estimated mean = --- = 5.9
                                   113


    Standard Error:

                                   Sample standard deviation
                Standard error = --------------------------
                                   Square root of valid cases

                                   2.055
                                 = ----- = 0.19
                                   10.63
```

11.7 STANDARD DEVIATION

The statistical model for our interval variable - attitude to women - had a second unknown parameter, the standard deviation. In order to estimate it, we must first assume a value for the population mean, as estimated above.

Statistical Model: Interval variable with a normal distribution for the population" two parameters, the mean assumed to be equal to the estimated value 5.9, and unknown standard deviation.

Sampling Distribution: The chi-square distribution, defined as the distribution of a sum of squares of variables each with a normal distribution.

```
           Estimator:

                                                  2
             (Sum of squares of score - (mean  X valid cases))
 Sq.  root of (-----------------------------------------------)
             (            valid cases - 1                     )

         N.b.  this is the same as the sample standard deviation.

           Estimate:

                            2
               (4422 - (5.9  X 113))
 = Sq.  root of (-------------------) = Sq.  root of 4.2   = 2.1
               (       112         )


           Standard Error:

                              Standard Deviation
           Standard Error = ------------------------------ = 2.1
                            Sq.  root of (2 X valid cases)
```

12 - Hypothesis Testing

The final stage of statistical inference - decision making - concerns hypothesis testing. We use the sample to test a hypothesis on the population, making a decision - whether to accept or reject the hypothesis - on the basis of the results of the test. We first look at the problems associated with the formulation of the hypothesis, and then we go through the four stages of statistical inference with the aid of a particular example.

12.1 THE FORMULATION OF THE HYPOTHESIS

We begin with a theoretical general hypothesis about the population under investigation. The general hypothesis makes general statements about theoretical concepts - for example that girls have a more positive attitude to women than boys. This must then be 'translated' into one or more statistical hypotheses which make statements about a statistical model of the population, typically taking the form of a set of equations or inequalities relating the various parameters of the model. In the above example, we could use the statistical hypothesis that - for a given 9-item index measuring 'attitude to women' - the population of girls has a higher mean than the population of boys. Such a formulation, however, presents a number of problems which need to be discussed.

1: The fallacy of affirming the consequent:

Since we are testing a number of derived statistical hypotheses rather than the initial general hypothesis, we must be careful about the conclusions we draw from them. If a statistical hypothesis is confirmed by a test on a given sample, it does not necessarily follow that the original general hypothesis is true, but we can say that if the statistical hypothesis is rejected, then this indicates that the general hypothesis should also be rejected. Thus, suppose our test led us to comfirm that the attitude to women index has a higher mean for girls than for boys - then we cannot necessarily infer that girls have a more positive attitude to women. There may be other equally valid reasons for this result - that there is some other factor influencing the relationship (i.e. that girls turn out to be more liberal than boys, and that liberalism indicates a more positive attitude to women). On the other hand, if we rejected the statistical hypothesis, and found that there appeared to be no difference between the two means, then we could infer that girls do not have a more positive attitude to women.

The solution to this problem is to devise a statistical test that, if confirmed, refutes the general hypothesis. Such a test can be derived form a statistical hypothesis that necessarily denies the general hypothesis - the null hypothesis. Thus, if we confirm a null hypothesis that the means are the same, then we must reject the original general hypothesis that girls are more positive to women.

2: Accepting the null hypothesis:

A further logical problem occurs if the null hypothesis is confirmed. As in the previous problem, a positive result for the test does not necessarily indicate that the null hypothesis is true, although a negative result implies that the null hypothesis should be rejected. Thus, if our test on the two means for the attitude to women index showed that we could not reject the null hypothesis that there is no difference between the two means, then we cannot then confirm that there was a difference.

Thus the logic of hypothesis testing proceeds by finding a number of possible null hypotheses that, if rejected, tend to confirm (but not prove conclusively(c) the original general hypothesis. The more null hypotheses that we fail to reject, the more likely it is that the general hypothesis is false.

3. Alternative hypotheses:

With statistical tests, we must also define the possible alternatives to the null hypothesis. For example, the alternative to the null hypothesis that the two means are the same would be that the mean for girls is higher than that for boys, ignoring the possibility that the mean for boys is higher than that for girls. This single alternative requires a one-tailed test, whereas both alternatives would need a two-tailed .

4. Significance:

The final problem connected with hypothesis formulation concerns the significance level as a criterion for deciding whether to reject the null hypothesis. The test takes the form of a probability that our sample comes from a population for which the null hypothesis is true. The significance level is a specific probability which we choose to distinguish tests which indicate rejection of the null hypothesis (and therefore have a lower probability than the significance level). Thus if the null hypothesis that two means are the same turned out to give a probability of 2% (i.e. a 2% chance of having selected a sample with at least the same difference in means as the sample actually selected), then we can say that the null hypothesis is significant at the 2% level.

12.2 THE PROCESS OF HYPOTHESIS TESTING

We use the example of attitude to women to illustrate the process step by step. First, our general hypothesis takes the form: 'Girls have, on average, a more positive attitude to women than boys'. Next, we devise a null hypothesis, 'There is no difference between the means of the attitude to women index for girls and that for boys'. Then we define an alternative hypothesis, 'That the mean for girls is higher than that for boys'. Finally, we fix an acceptable significance level, 5%, which gives us the maximum acceptable chance of choosing a sample which would lead to the rejection of the null hypothesis. We are now ready to follow through the four stages of statistical inference:

12.2.1 The Statistical Model

We assume that we have two independent normal distributions of the attitude to women index, one for girls and one for boys. In addition we assume the null hypothesis that both means are the same.

12.2.2 The Sampling Distribution

We use an estimator, called the test_statistic, which gives us a known distribution for determining the differences between the means. In this case, the statistic is given by:

$$\text{Test statistic} = \frac{\text{Mean for girls - Mean for boys}}{\text{Pooled standard error of the Means}}$$

Where Pooled standard error of the means is given by

$$\text{Sq. root of} \quad \left(\frac{\text{Variance for girls}}{\text{number of girls}} + \frac{\text{Variance for boys}}{\text{number of boys}}\right)$$

This gives us a sampling distribution in the form of Student's_t, which has a known distribution (rather like the normal distribution but more peaked).

12.2.3 Estimation

We get an estimate for the Student's t statistic of -6.99. In addition - for this statistic - we must provide one more piece of information, the number of degrees of freedom, given by the number of valid cases less 2 - or 96.

12.2.4 Decision Making

We observe from statistical tables that the value of Student's t which has 5% chance of being exceeded is +1.66, given approximately 100 degrees of freedom and a one-tailed test. Since this is certainly lower than the absolute value of the estimate obtaimed (i.e. 6.99), we conclude that the null hypothesis can certainly be rejected, and the test tends to conform our original general hypothesis.

13: More on hypothesis testing

In this, the last formal lecture of the course, we examine the standard Tests of Hypothesis used in Statistical Inference. These include tests of central tendency (i.e. means, proportions), tests of dispersion (i.e. variance(c) and tests of association (e.g. correlation, two-sample tests). All of these tests take the form of a null hypothesis which, if rejected, adds more evidence to the confirmation of the original general hypothesis. N.B. if a null hypothesis is not rejected, then the test is said to be inconclusive or that there is insufficient_evidence to reject the null hypothesis.

13.1 TESTS FOR CENTRAL TENDENCY

13.1.1 Proportions

General Hypothesis: The proportion of the population in a given category is higher (or lower) than a specified figure. E.g. There are more girls than boys in fifth forms.

Statistical model: Nominal or ordinal variables which, when grouped into two categories (i.e. within or outside the range specified by the general hypothesis), form a binomial distribution in the population. One parameter (proportion in specified category), fixed but unknown. E.g. Self-reported sex with parameter 'proportion girls'.

Null Hypothesis: Proportion in specified category is equal to a specified figure. E.g. Proportion girls is equal to 0.5.

Alternative Hypothesis: Proportion is greater than (or less than) the specified figure (i.e. a one-tailed test). E.g. Proportion girls is greater than 0.5.

Sampling distribution: Under the null hypothesis, the sample proportion will have a normal sampling distribution, with mean 1/2 specified proportion and standard deviation 1/2 standard error of specified proportion, provided the number of valid cases is sufficiently large (a good cut-off is valid cases X specified proportion greater than 5).

Test statistic: The most useful test statistic is the z-score, which is defined as:

```
          Sample proportion - Specified proportion
z-score = ---------------------------------------
           Standard error of specified proportion
```

Decision: Choose a significance level (i.e. probability of picking a sample from the population assuming the null hypothesis(c) and compare the z-score with the corresponding value tabulated in the z-score tables. E.g. we choose a 5% significance level, giving a z-score for a one-tailed test of 1.65. This is clearly higher than our result of 0.408, and hence our result is not within the range required to reject the hypothesis. We must therefore decide not to reject the null hypothesis and hence the sample does not give us cause to reject the original general hypothesis.

(N.B. Instead of choosing a significance level first, we could have looked up the probability of achieving a z-score at least as high as our result, and made a decision on the basis of this figure. E.g. the result z-score of .408 gives a probability of 59% of obtaining a sample with at least this value, given our null hypothesis. This is clearly insufficient evidence to reject the null hypothesis.)

(N.B. Since SPSS does not provide a specific difference of proportion test, the best procedure to use is NPAR TESTS with the CHI-SQUARE method. For example, the following procedure will test if proportion girls was 0.5:

```
NPAR TESTS CHI-SQUARE = V348 (1, 2) /EXPECTED=EQUAL .
```

13.1.2 Means

General hypothesis: The mean score of a variable for the population is higher (or lower) than a specified value.

Statistical model: Interval variables with a normal distribution in the population; two fixed but unknown parameters: mean and standard deviation.

Null hypothesis: Mean value is equal to specified figure.

Alternative hypothesis: Mean value is higher (or lower or not equal to(c) specified value; one-tailed test except for 'not equal to' option, when a two-talied test is required.

Sampling Distribution: Student's t Distribution with degrees of freedom given by (number of valid cases less one).

Test Statistic: Student's t score given by the formula:

```
                                    sample mean - specified mean
t = Sq root (valid cases - 1) X     ----------------------------
                                     sample standard deviation
```

Decision: Choose a significance level and look up the t-score corresponding to this level (given the degrees of freedom and whether one- or two-tail test). Compare this value with the test statistic value and decide whether to reject the null hypothesis accordingly. Alternatively, look up the significance level corresponding to the test statistic value and decide on the basis of this figure.

13.2 TESTS FOR DISPERSION

13.2.1 Standard Deviation

General hypothesis: The standard deviation of a variable for the population is at least as high as a specified figure.

Sampling distribution: Interval variable with Normal Distribution for the population; one parameter, mean, estimated and one parameter, variance, unknown.

Null Hypothesis: The standard deviation is equal to the specified value.

Alternative Hypothesis: The standard deviation is greater than the specified value.

Sampling Distribution: The F distribution with degrees of freedom given by 1 and (valid cases - 1). N.B. The F distribution has two kinds of degrees of freedom associated with it.

Test statistic: F-score given by the formula:

```
           (Sample sum-of-squares) / (Valid cases - 1)
F-score =  -------------------------------------------
            Square of (specified standard deviation)
```

Decision: As for t-score above, but using the tables for the F-statistic.

13.3 TESTS FOR ASSOCIATION

13.3.1 Chi-square

General hypothesis:

Some relationship exists between two nominal (or ordinal(c) variables in the population.

Statistical Model:

Two nominal variables.

Null Hypothesis:

Independent variables - i.e. no relationship.

Alternative Hypothesis:

Some relationship

Statistical Distribution:

Chi-square, with number of degrees of freedom equal to (rows less one) X (columns less one).

Test Statistic:

Chi-square given by the formula:

```
                (Square of (observed freq.  - expected freq.))
Chi-square = Sum(-------------------------------------------)
                (          expected frequency               )
```

Decision:

The SPSS CROSSTABS procedure gives the significance level associated with the chi-square result. A significance of 0.05 or below indicates that the null hypothesis should not be rejected, and contributes to the confirmation of the assumption that a relationship exists.

13.3.2 Gamma, Kendal's Tau Etc.

The tests for the various measures of strength of association, such as gamma and Kendal's tau are given in the SPSS CROSSTABS procedure, together with their associated significance levels.

The null hypothesis is usually of zero association, and the decision is made in the same way as for the chi-square test above.

13.3.3 Pearson's Correlation

General hypothesis: A linear or near-linear relationship exists between two interval scale variables.

Sampling Distribution: Bi-variate normal, with unknown means, standard deviations and covariance.

Null Hypothesis: Zero covariance.

Alternative hypothesis: Non-zero covariance.

Test statistic: F-statistic, with one and N-2 degrees of freedom. Use the SPSS procedure PEARSON CORR to give the significance level of the F-statistic.

Decision: as for chi-square.

13.3.4 Two-sample Tests

The tests for comparing parameters from two populations follow closely the above tests for central tendency and dispersion. The difference being that, in general, the specific parameter value is replaced by the sample estimate for the second sample. Also, in the case of the difference of means, we must assume that the standard deviations for each sample are equal (this is the same as the T-TEST example used in the last lecture). This is the simplest case of an association between an interval dependent variable and a nominal (i.e. binary(c) independent variable.

13.3.5 One-Way ANOVA

One-way analysis of variance is used to investigate the relationship between an interval dependent variable and a nominal (or ordinal) independent variable.

General Hypothesis: Some relationship.

Statistical Model: Interval dependent variable, nominal independent variable; parameters consist of a list of means and standard deviations of the dependent variable for each category of the independent variable, plus the proportion of cases in each category. Also all standard deviations are assumed to be equal.

Null hypothesis: All means are equal.

Sampling distribution: The F-distribution with (categories - 1) and (valid cases - categories) degrees of freedom.

Test Statistic: Use the SPSS procedure ANOVA.

Appendix 1

References:

Statistics:

BLALOCK
Social Statistics

LOETHER & MCTAVISH
Descriptive Statistics for Sociologists

MUELLER et al
Statistical Reasoning in Sociology

ROSENBERG
The Logic of Survey Analysis

ROWNTREE
Statistics without Tears

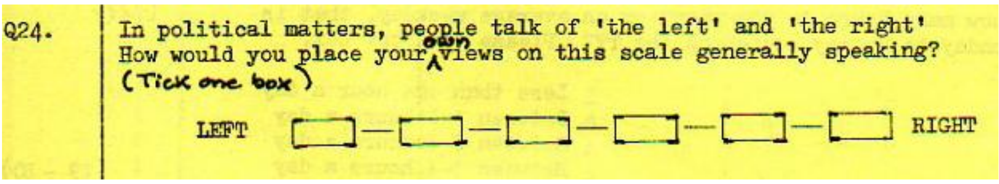
Fifth Form survey:

Details and downloadable resources for this survey can be be found on Playground to Politics.

Paul Ahmed, Harriet Cain and Alan Cook
Playground to Politics: a study of values and attitudes among fifth formers in a North London comprehensive school
Report on 2nd year project for BA Applied Social Studies (Social Research) Polytechnic of North London 1982

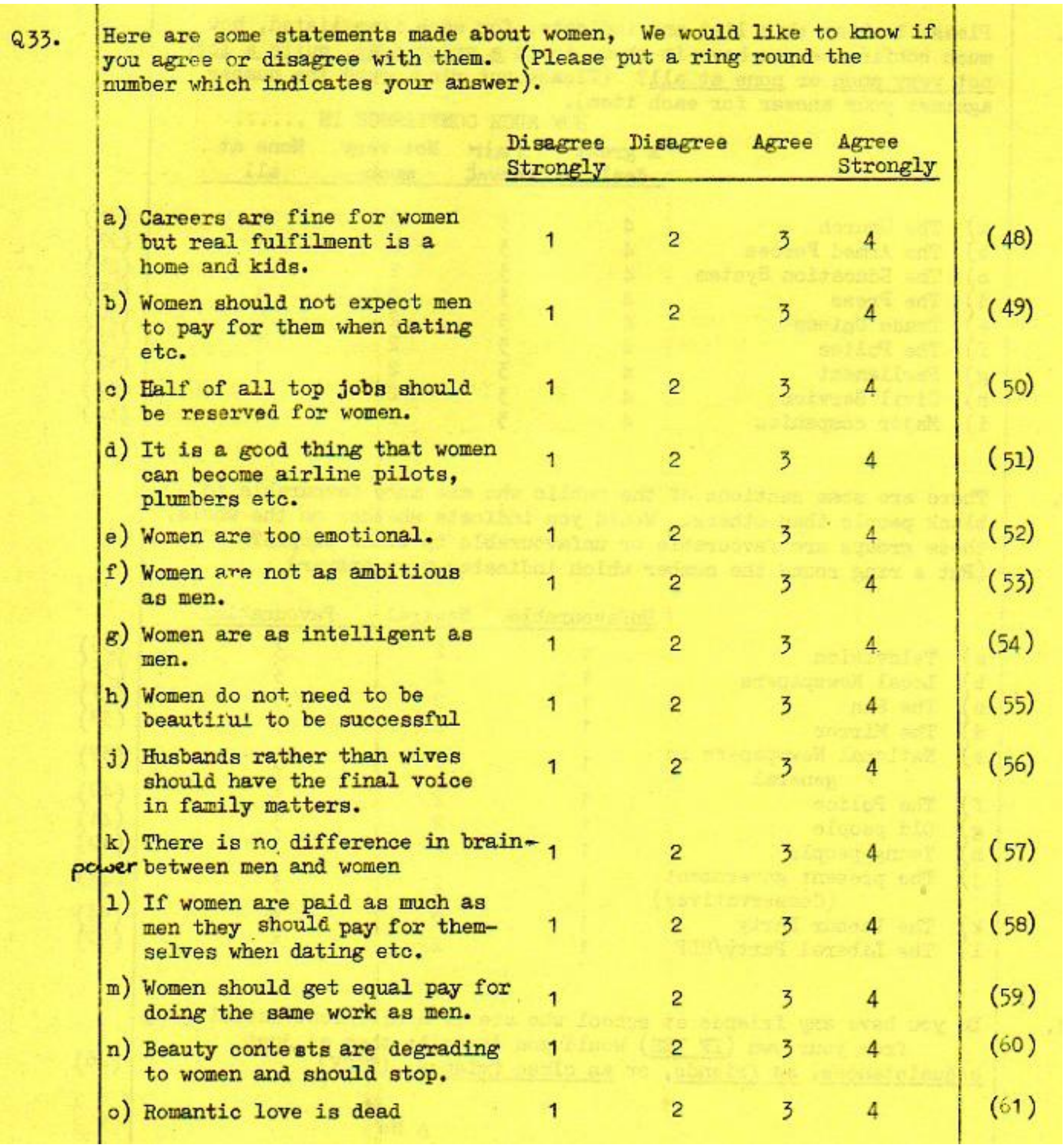
John Hall and Alison Walker,
User manual for Playground to Politics: a study of values and attitudes among fifth formers in a North London comprehensive school
Survey Research Unit, Polytechnic of North London 1982 (mimeo 40 pp – codebook, questionnaire, coding notes)

Responses coded 1 to 6 (left to right)



Q24. In political matters, people talk of 'the left' and 'the right'
How would you place your own views on this scale generally speaking?
(Tick one box)

LEFT □—□—□—□—□—□ RIGHT



Q33. Here are some statements made about women, We would like to know if you agree or disagree with them. (Please put a ring round the number which indicates your answer).

	Disagree Strongly	Disagree	Agree	Agree Strongly	
a) Careers are fine for women but real fulfilment is a home and kids.	1	2	3	4	(48)
b) Women should not expect men to pay for them when dating etc.	1	2	3	4	(49)
c) Half of all top jobs should be reserved for women.	1	2	3	4	(50)
d) It is a good thing that women can become airline pilots, plumbers etc.	1	2	3	4	(51)
e) Women are too emotional.	1	2	3	4	(52)
f) Women are not as ambitious as men.	1	2	3	4	(53)
g) Women are as intelligent as men.	1	2	3	4	(54)
h) Women do not need to be beautiful to be successful	1	2	3	4	(55)
j) Husbands rather than wives should have the final voice in family matters.	1	2	3	4	(56)
k) There is no difference in brain-power between men and women	1	2	3	4	(57)
l) If women are paid as much as men they should pay for themselves when dating etc.	1	2	3	4	(58)
m) Women should get equal pay for doing the same work as men.	1	2	3	4	(59)
n) Beauty contests are degrading to women and should stop.	1	2	3	4	(60)
o) Romantic love is dead	1	2	3	4	(61)

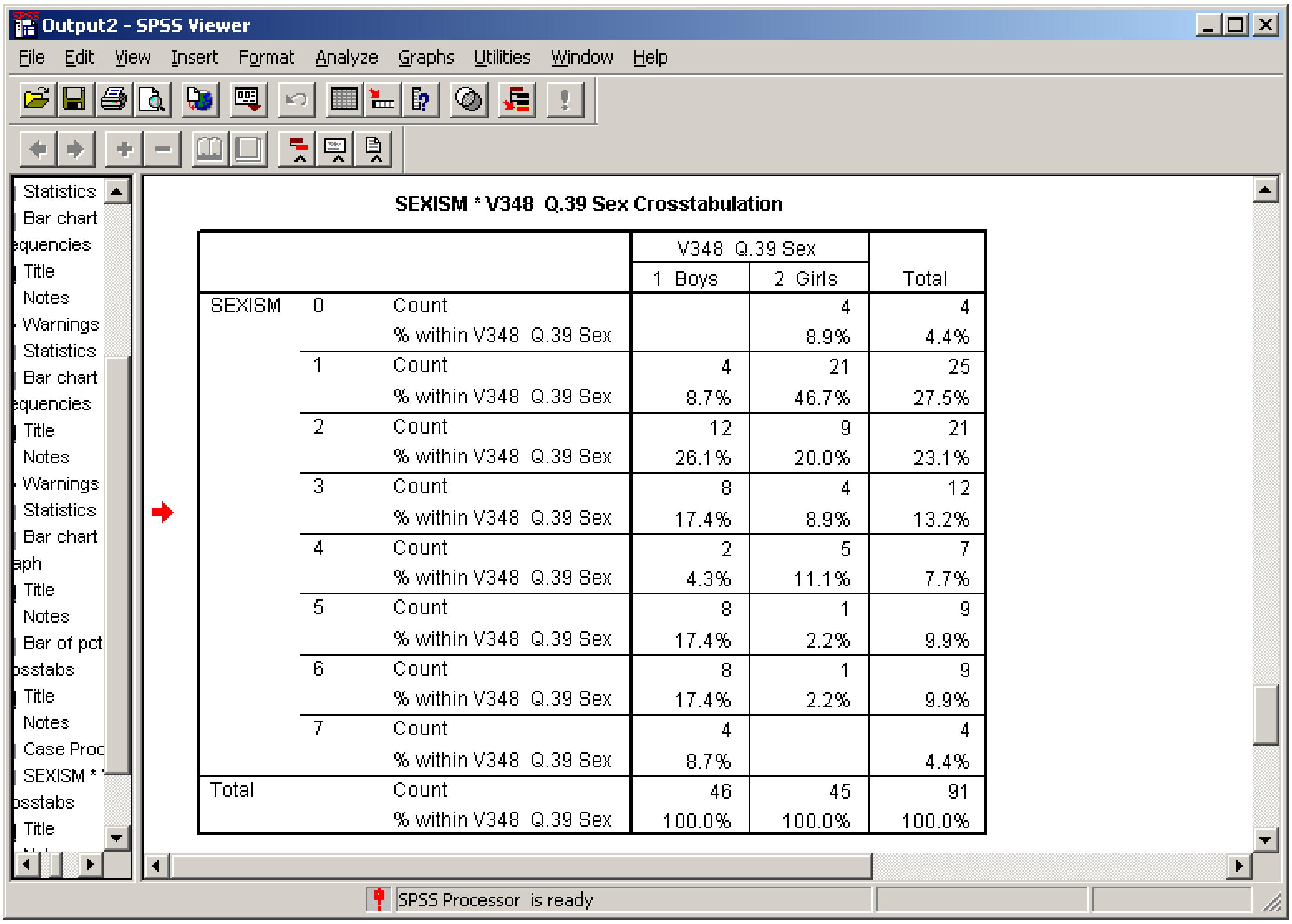
[Various references to figures in section 8, but none in text: not clear which data are used (?SSRC QoL surveys) can they be reproduced?]

Appendix 2

Sample output from SPSS 11 for Windows

A: Crosstabulations: Sexism score (Fifth Form survey)

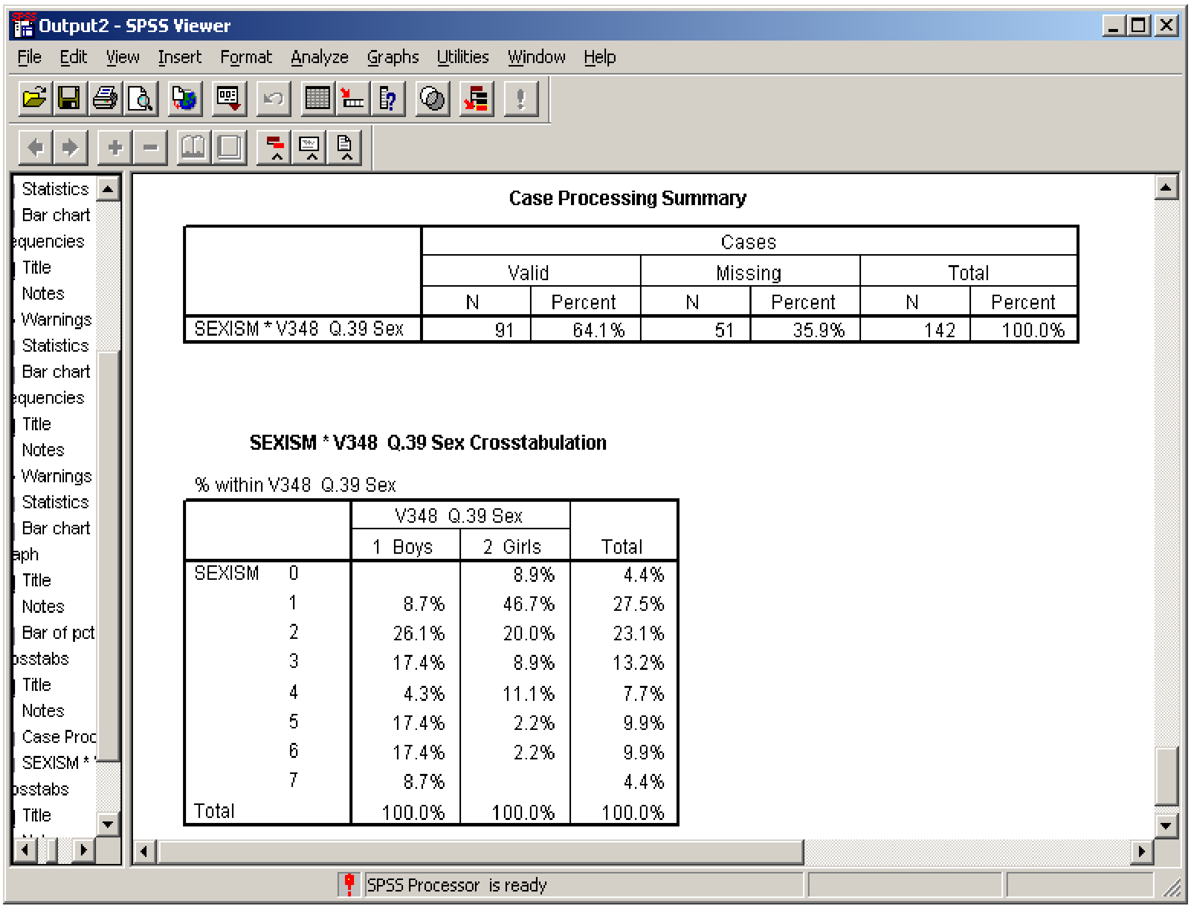
Screen dump of SPSS output:



SEXISM * V348 Q.39 Sex Crosstabulation

			V348 Q.39 Sex		Total
			1 Boys	2 Girls	
SEXISM	0	Count		4	4
		% within V348 Q.39 Sex		8.9%	4.4%
	1	Count	4	21	25
		% within V348 Q.39 Sex	8.7%	46.7%	27.5%
	2	Count	12	9	21
		% within V348 Q.39 Sex	26.1%	20.0%	23.1%
	3	Count	8	4	12
		% within V348 Q.39 Sex	17.4%	8.9%	13.2%
	4	Count	2	5	7
		% within V348 Q.39 Sex	4.3%	11.1%	7.7%
	5	Count	8	1	9
		% within V348 Q.39 Sex	17.4%	2.2%	9.9%
	6	Count	8	1	9
		% within V348 Q.39 Sex	17.4%	2.2%	9.9%
	7	Count	4		4
		% within V348 Q.39 Sex	8.7%		4.4%
Total		Count	46	45	91
		% within V348 Q.39 Sex	100.0%	100.0%	100.0%

This looks bad enough, but it’s even worse when copied to Word document (the lines are stripped out automatically).   You can improve it by just asking for the column percentages, but then SPSS loses the column counts so you don’t know what the base is unless you look in the case processing summary above it.   You really need them in the same table, but this requires some nifty editing.



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEXISM * V348 Q.39 Sex	91	64.1%	51	35.9%	142	100.0%

SEXISM * V348 Q.39 Sex Crosstabulation

% within V348 Q.39 Sex

		V348 Q.39 Sex		Total
		1 Boys	2 Girls	
SEXISM	0		8.9%	4.4%
	1	8.7%	46.7%	27.5%
	2	26.1%	20.0%	23.1%
	3	17.4%	8.9%	13.2%
	4	4.3%	11.1%	7.7%
	5	17.4%	2.2%	9.9%
	6	17.4%	2.2%	9.9%
	7	8.7%		4.4%
Total		100.0%	100.0%	100.0%

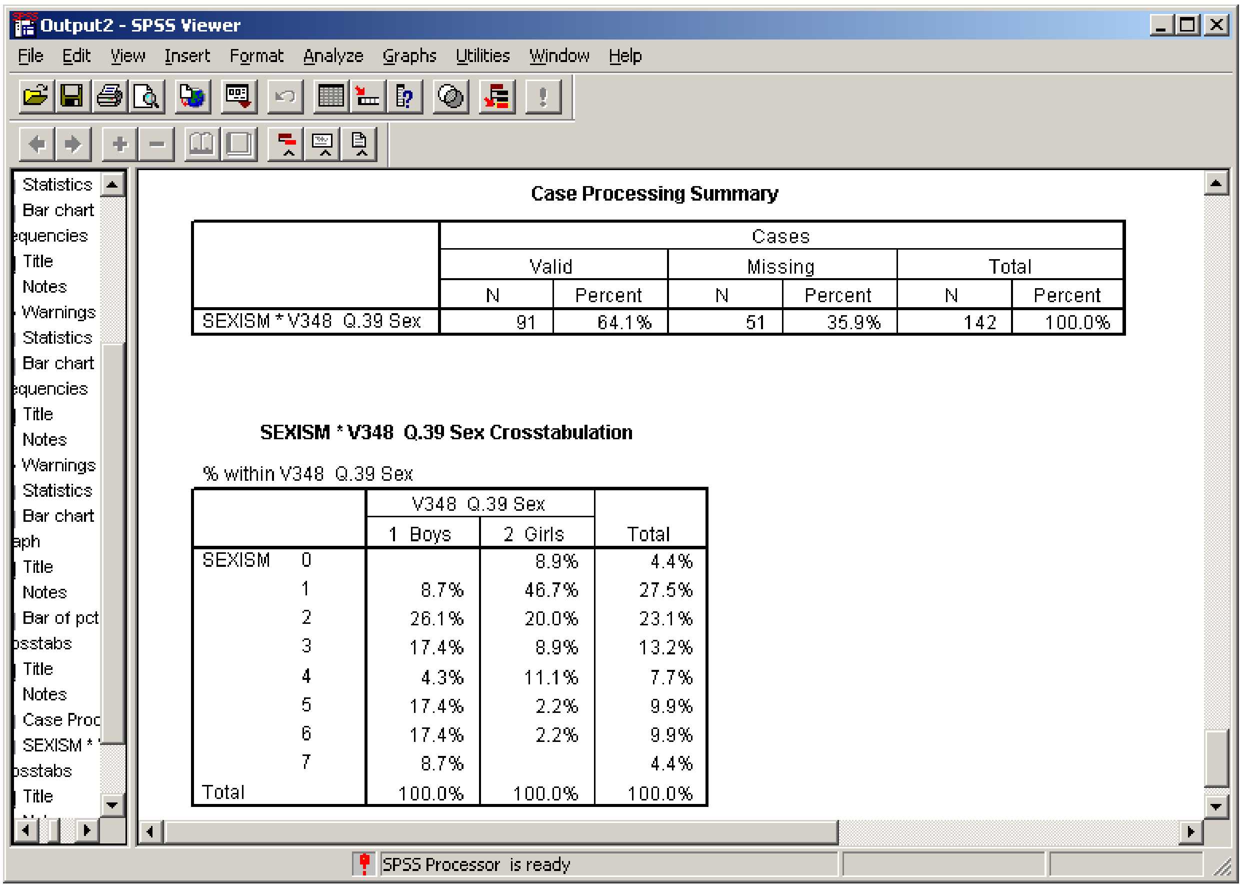
SPSS Processor is ready

(First table above as copied into Word document, then converted from tables to text after adjusting column width)

SEXISM * V348 Q.39 Sex Crosstabulation

			V348 Q.39 Sex		Total
			1 Boys	2 Girls	
SEXISM	0	Count		4	4
		% within V348 Q.39 Sex		8.9%	4.4%
	1	Count	4	21	25
		% within V348 Q.39 Sex	8.7%	46.7%	27.5%
	2	Count	12	9	21
		% within V348 Q.39 Sex	26.1%	20.0%	23.1%
	3	Count	8	4	12
		% within V348 Q.39 Sex	17.4%	8.9%	13.2%
	4	Count	2	5	7
		% within V348 Q.39 Sex	4.3%	11.1%	7.7%
	5	Count	8	1	9
		% within V348 Q.39 Sex	17.4%	2.2%	9.9%
	6	Count	8	1	9
		% within V348 Q.39 Sex	17.4%	2.2%	9.9%
	7	Count	4		4
		% within V348 Q.39 Sex	8.7%		4.4%
Total		Count	46	45	91
		% within V348 Q.39 Sex	100.0%	100.0%	100.0%

Still a bit of a mess. It's a bit clearer like this:



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
SEXISM * V348 Q.39 Sex	91	64.1%	51	35.9%	142	100.0%

SEXISM * V348 Q.39 Sex Crosstabulation

% within V348 Q.39 Sex

		V348 Q.39 Sex		Total
		1 Boys	2 Girls	
SEXISM	0		8.9%	4.4%
	1	8.7%	46.7%	27.5%
	2	26.1%	20.0%	23.1%
	3	17.4%	8.9%	13.2%
	4	4.3%	11.1%	7.7%
	5	17.4%	2.2%	9.9%
	6	17.4%	2.2%	9.9%
	7	8.7%		4.4%
Total		100.0%	100.0%	100.0%

but it still needs editing to get the column totals in, and even then it's cluttered.

SEXISM * V348 Q.39 Sex Crosstabulation
% within V348 Q.39 Sex

		V348 Q.39 Sex		Total
		1 Boys	2 Girls	
SEXISM	0		8.9%	4.4%
	1	8.7%	46.7%	27.5%
	2	26.1%	20.0%	23.1%
	3	17.4%	8.9%	13.2%
	4	4.3%	11.1%	7.7%
	5	17.4%	2.2%	9.9%
	6	17.4%	2.2%	9.9%
	7	8.7%		4.4%
(N=100%)		46	45	91

Yet again SPSS has left blank cells, the decimal points need aligning and the % signs clutter things up.

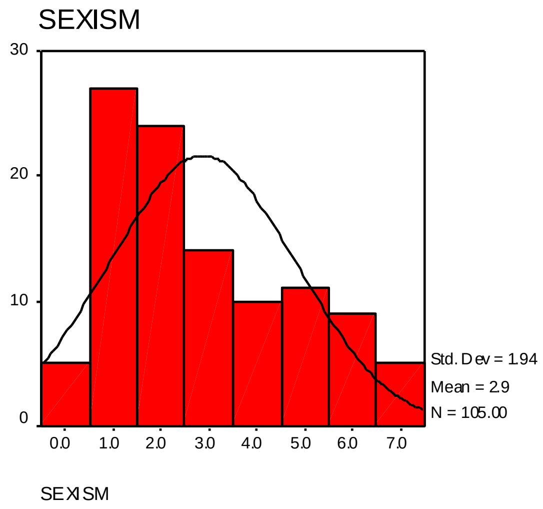
What you want is something like this, finicky, but much clearer:

SEXISM by Sex

				I
		Boys	**Girls**	**Total**
		%	**%**	**%**
SEXISM	0	0.0	8.9	4.4
	1	8.7	46.7	27.5
	2	26.1	20.0	23.1
	3	17.4	8.9	13.2
	4	4.3	11.1	7.7
	5	17.4	2.2	9.9
	6	17.4	2.2	9.9
	7	8.7	0.0	4.4
(N=100%)		46	45	91

B: **Histograms:** Sexism score (Fifth Form survey)

Figure 1: All pupils



NB: This chart has been overlaid with a normal distribution conforming to the mean and standard deviation calculated for the distribution of scores on the sexism scale. Note that the distribution is positively skewed (ie the tail has been pulled out towards the higher scores which have in turn affected the calculated mean. Perhaps the median would be a better measure?)

Figure 2: Boys only

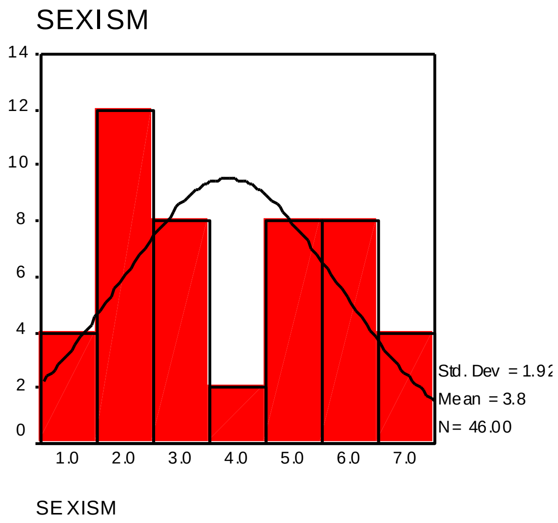
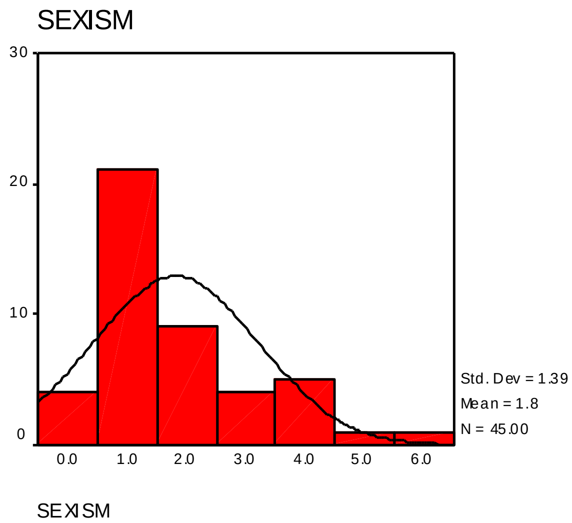
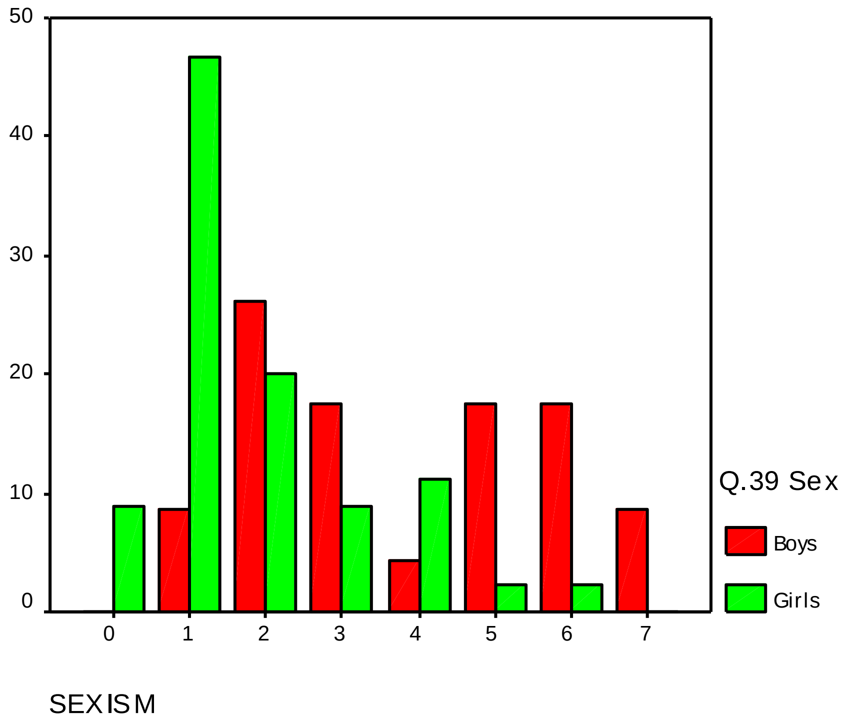


Figure 3: Girls only



This clearly shows a big difference between boys and girls, and also illustrates the dangers of taking initial results from only one variable! Notice SPSS has chopped the 0 bar from the boys' chart and the 7 bar from the girls'. Notice also the bi-modal distribution for the boys, a surefire indication that another variable is at work. The only way of forcing all values from 0 to 7 is to put boys and girls on the same chart as on Fig 4 below, but it could still do with a spot of red for the boys in the 0 column.

Figure 4: Boys and girls separately


SEXISM
Q.39 Sex
🟥 Boys
🟩 Girls