1.1.2 Introduction to Survey Data

[Updated 23 August 2010, from 1992 original.]

1: Introduction

These notes are based on the very first session of the practice-oriented **Survey Analysis Workshop** (postgraduate, hands-on, part-time, evening) which I designed and taught from 1976[1] until 1992 at the then Polytechnic of North London (now part of London Metropolitan University). The course was one of two courses intended as a replacement for the Summer Schools in Survey Methods given by the SSRC Survey Unit from 1970 to 1976. The survey processing and analysis components were covered in **Survey Analysis Workshop**. Our sister course **Survey Research Practice**, taught entirely by senior practitioners[2] from outside the Polytechnic, covered all other aspects (questionnaire design, sampling, attitude measurement, qualitative methods, interviewing, costing and fieldwork).

On arrival the first evening, students were given a short pre-course questionnaire to complete, asking about their previous experience of computing and statistics, and a few items replicated from the British Social Attitudes survey: they also received an information pack containing a set of course booklets and a facsimile questionnaire for the latest BSA survey. We would introduce ourselves briefly with name, institutional affiliation (if any), previous qualifications and experience, and reasons for coming on the course. An all too frequent lament was "I've got a degree in Sociology and I want a job!", but even the ones with jobs had often received little or no training in statistical or technical skills in their undergraduate or postgraduate courses (much of which was in any case inadequate). Some of these wanted a better job and/or had been sent by their employers (mainly central and local government or the voluntary sector). Some were MPhil or PhD students registered elsewhere.

This was followed by a session in the computer lab in which students familiarised themselves with the terminals and the line-printers, learned to log on to the Vax, copy a short pre-prepared SPSS job into their area, run it with a specially written front-end program[3] to make it easier to use SPSS on the Vax, print out the results and return to class with their printout for a brief explanation and discussion. No student ever left empty-handed from this or later sessions, (but the spare copies of output always came in handy for one or two of them!) This greatly assisted student motivation and learning.

The session ended with nibbles and drinks, non-alcoholic for drivers, Bulgarian red or white wine (from the very first Majestic Wine Warehouse, newly opened in Tottenham) for everyone else.

---

[1] In 1976 the statistics were taught by John Utting and the computing by myself. There were no computing facilities on our campus: all data and programs had to be written on coding sheets which Maureen Ashman (Senior Programmer with responsiblity for SPSS) collected and arranged to be punched on 80-column cards and run by Computer Services (PNLCS). Results were returned by courier: errors would not be detected until the following week! When PNLCS supplied an ICL card punch, I could correct errors, but still had to send cards back by courier to get results the next day. Jim Ring taught the statistics from 1981 to 1988 when I took over the whole course. When SPSS went interactive and we got a computer lab with 16 terminals and 4 fast servers, the course was transformed. From 1989 I used the pre-course questionnaire and information pack and had Debbie Youdell and Katie Featherstone, two of my sophomore students, as demonstrators. I also sold the Norusis book at cost. There'll never be another course like it.

[2] Visiting lecturers included: Cathie Marsh (SPS, Cambridge Univ) Nick Moon, John O'Brien, (NOP) Alan Marsh (OPCS/PSI) Colin Airey, Martin Collins, Barry Hedges, Roger Jowell, Jean Morton-Williams, Jane Ritchie, Bridget Taylor, Roger Thomas, Sharon Witherspoon (SCPR/Natcen) Gordon Heald (Gallup) Malcolm Brighton (Document Reading Services) and Wendy Sykes (Independent researcher)

[3] The program was written by Jim Ring while he was Senior Research Officer in the Survey Research Unit at PNL. It limited SPSS output files to two editions (to avoid users running out of disk space) and had excellent error trapping. If errors occurred it returned users to the point in their syntax file where they had left off, although SPSS didn't always precisely identify the type of error. It greatly assisted students and researchers with a series of prompts in editing SPSS syntax files, correcting and running of SPSS jobs and local printing of results and enabled a great deal of work to be completed in a very short time..

2: The language of survey data (and SPSS)

I would start by writing, across the top of the (double-width) chalk-board, a few things typically measured by questionnaires (some solicited from the class):

Sex	Age last birthday	Marital status	Social class	How voted	Height in metres	How happy are you?	No. of children

I then divided the board into columns defined by the items listed and filled in the responses of imaginary respondents (with a running commentary on the kind of thing they might mutter before giving a response to a questionnaire-clutching stranger knocking at the door just as they've settled down to watch their favourite TV programme).

Sex	Age last birthday	Marital status	Social Class	How voted	Height in metres	How happy are you?	No. of children
Male	26	Single	C1	Lib-Dem	1.82	Fairly	None
Female	35	Married	C2	Never vote	1.63	Very	2
Female	56	Married	D	Tory	1.55	Very	3
Male	42	Divorced	AB	Labour	1.74	Not very	1
Female	83	Widowed	E	[Refused]	[Don't know]	[Not aswered]	4
Male	16	Single	Still at school	[Too young to vote]	1.65	Fairly	[Does not apply]

It is also typical of social surveys that for some cases the value of a particular variable is **missing**. There are many reasons for this, but common ones are:

a) the question was not answered because it is **inapplicable** (e.g. the 16-year old who is too young to vote)
b) because the respondent **refused** to answer,
c) could not decide or **did not know** (eg the lady who didn't know her height) or
d) because the question was **missed** altogether (see empty cell above for the lady with no reply for, *"How happy are you?"*).

Such **missing values** need to be taken into account, but the way they are treated in analysis will depend on a variety of factors.

At this point I would write **VARIABLES** across the top, **CASES** down the side, note that the entries inside the grid were **VALUES** and that we had just generated a **DATA MATRIX**.

I would then shorten the column headings and number the rows to yield something like the following data matrix.

V A R I A B L E S → → →

	SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
CASES ↓								
1	Male	26	Single	C1	LibDem	1.82	Fairly	None
2	Female	35	Married	C2	Never vote	1.63	Very	2
3	Female	56	Married	D	Tory	1.55	Very	3
4	Male	42	Divorced	AB	Labour	1.74	Not very	1
5	Female	83	Widowed	E	Refused	Don't know		4
6	Male	16	Single	Still at school	Too young to vote	1.65	Fairly	Does not apply

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

VALUES etc. etc. until all cases are entered

This constituted the first introduction to formal terminology (and to **keywords** in the SPSS language).

Spreadsheets such as Excel and statistical analysis packages such as SPSS operate with just such a matrix of data values (actually just the cells in the **body** of the table above).

3: What's a data matrix?

We have just seen a simple data matrix containing information on imaginary respondents' answers to precoded or post-coded open-ended questions, in which the **columns** represent **VARIABLES**

V A R I A B L E S → → →

		SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
CASES ↓									
1		Male	26	Single	C1	**LibDem**	1.82	Fairly	None
2		Female	35	Married	C2	**Never vote**	1.63	Very	2
3		Female	56	Married	D	**Tory**	1.55	Very	3
4		Male	42	Divorced	AB	**Labour**	1.74	Not very	1
5		Female	83	Widowed	E	**Refused**	Don't know		4
6		Male	16	Single	Still at school	**Too young to vote**	1.65	Fairly	Does not apply

the **rows** represent individual **CASES**, e.g.

V A R I A B L E S → → →

	SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
CASES ↓								
1	Male	26	Single	C1	LibDem	1.82	Fairly	None
2	Female	35	Married	C2	Never vote	1.63	Very	2
3	**Female**	**56**	**Married**	**D**	**Tory**	**1.55**	**Very**	**3**
4	Male	42	Divorced	AB	Labour	1.74	Not very	1
5	Female	83	Widowed	E	Refused	Don't know		4
6	Male	16	Single	Still at school	Too young to vote	1.65	Fairly	Does not apply

… and the **cells** contain the initial **VALUES** of each variable for each case e.g.

V A R I A B L E S → → →

CASES ↓		SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
1		Male	26	Separated	C1	LibDem	1.82	Fairly	None
2		Female	35	Married	C2	Never vote	1.63	Very	2
3		Female	56	Married	D	Tory	1.55	Very	3
4		Male	42	Divorced	AB	Labour	1.74	Not very	2
5		Female	83	Widowed	E	Refused	Don't know		4
6		Male	16	Single	Still at school	Too young to vote	1.65	Fairly	Does not apply

4: The nature of survey data

A sample survey studies **part** of a group (the **sample**) in order to make inferences about the **whole** group (the **population**) from which the sample is drawn. We usually try to ensure that the sample is representative of the population; but a sample which is known to be biased can sometimes be useful if sufficient is known about the bias to be able to make an allowance for it, or if it can be used to set an upper or lower limit to some population value. From a sample we obtain a **statistic** which is an **estimate**, within certain limits, of the **true population value**, known as a **parameter**.

Social surveys study social entities (persons, families, business firms, political parties, clubs etc.). The survey data consist of information about each entity in a sample selected from the whole population of such entities. Each entity in the sample is known as a **CASE**.

Data are obtained by measuring, observing, asking questions about, various characteristics (height, colour of eyes, voting behaviour, number of children) of the cases in the sample. The characteristics studied are called **VARIABLES** and the descriptions of the characteristics for each case (e.g. **height in metres** - 1.63m; **colour of eyes** - blue; **party voted for at last election** - Labour; **number of children** -2) are called their **VALUES**.

Thus survey data typically consist of the **value** of each **variable** for each **case**, and can be represented in a rectangular **data matrix** (just like a spreadsheet) as in the examples above.

Although computers can handle alphabetic data obtained from questionnaire surveys (and it is sometimes easier for anxious beginners) **alphabetic** values are normally changed to **numeric** values before or immediately after data entry. Numeric data can be processed much more quickly and efficiently, and so the non-numeric **alphabetic codes** needed to be coded using **numeric codes** representing the original responses.

Leaving the original responses on the board, using a different colour chalk (yes, chalk!) and referring to an imaginary **coding frame** (perhaps pausing to explain what a coding frame was) I would write in a numeric code alongside each alphabetic response.

V A R I A B L E S → → →

CASES ↓	SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
1	Male 1	26	Single 1	C1 2	LibDem 2	1.82	Fairly 2	None 0
2	Female 2	35	Married 2	C2 3	Never vote 8	1.63	Very 3	2
3	Female 2	56	Married 2	D 4	Tory 3	1.55	Very 3	3
4	Male 1	42	Divorced 3	AB 1	Labour 1	1.74	Not very 1	1
5	Female 2	83	Widowed 4	E 5	Refused 7	Don't know 9		4
6	Male 1	16	Single 1	Still at school 8	Too young to vote 0	1.65	Fairly 2	Does not apply 8

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

VALUES

I then cleaned off the original alphabetic codes, leaving only the numeric codes. The matrix then looked something like this:

V A R I A B L E S → → →

CASES ↓	SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
1	1	26	1	2	2	1.82	2	0
2	2	35	2	3	8	1.63	3	2
3	2	56	2	4	3	1.55	3	3
4	1	42	3	1	1	1.74	1	1
5	2	83	5	5	7	9	.	4
6	1	16	1	8	0	1.65	2	88

[NB: I have replaced the blank cell with a **full stop (period).** I'll explain it later, but this is because it is the **missing value** automatically assigned by SPSS when it finds blanks or alphabetic characters in fields expected to contain numeric data.

This data matrix is now in a format ready to be fed into SPSS ready for editing and analysis. These days it will usually already be in a computer readable format generated by a research agency or after being entered by the researcher directly from the questionnaire.

To help you understand how numbers are used to represent numeric. alphabetic or missing data. I have modified the matrix to leave the original **numbers** unchanged, the numeric codes used to replace alphabetic responses indicated in **dark red** and those for responses to be treated as missing in **blue**.

V A R I A B L E S → → →

CASES ↓		SEX	AGE	MARITAL	CLASS	VOTE	HEIGHT	HAPPY	KIDS
1		1	26	1	2	2	1.82	2	0
2		2	35	2	3	8	1.63	3	2
3		2	56	2	4	3	1.55	3	3
4		1	42	3	1	1	1.74	1	1
5		2	83	5	5	7	9	.	4
6		1	16	1	8	0	1.65	2	88

At this point students were asked if they had any comments on, or noticed anything about, the way numbers had been used in the matrix, but this was usually met by blank expressions all round. (Remember, these were mostly UK graduates in sociology and related subjects!)

I then asked whether there was any difference in the way numbers were used between two variables such as vote and number of children, or between height in centimetres and number of children (using a joke about average families having 2.4 children as a hint) or anything that, say, two sets of numbers had, but another didn't. It took different amounts of time with different waves of students, but eventually someone would get the idea and by this "Socratic" process the class would arrive at the solution without the phrase **levels of measurement** having once being mentioned, thus proving that they weren't as innumerate as they thought!

4: Levels of measurement

For ease of computer processing, the values of variables are usually, but not always, coded as numbers, in survey research most often as integers (numbers with no decimals), but in medical research much data will also have decimal places. It is important to remember that these numeric codes frequently do not have all the properties of real numbers. This can affect the kind of statistical presentation and manipulation which is appropriate or permissible.

One important quality is the **level of measurement**.

The basic level is **nominal** (or categorical). All that is necessary is that the categories are properly defined (precise, mutually exclusive, exhaustive of all cases). Religious affiliation is such a variable: so are marital status and parliamentary constituency. Surveys usually ensure that categories are exhaustive by including a residual category 'Other'. Numeric codes are **arbitrarily** assigned to categories and can be jumbled up without losing any information.

Ordinal implies that, in addition, the categories can be **ranked**, i.e. placed in order from **high** to **low** on some defined criterion (e.g. Very happy, Not too happy, Very unhappy). Numeric codes cannot be arbitrarily assigned to categories, but they **can be reversed**.

Interval has all the characteristics of nominal and ordinal plus a **defined unit of measurement**. Thus, for instance, the distance from 2 to 4 is the same as that from 4 to 6. Examples include age, height, income in ££, number of children. Numerical codes are neither arbitrary nor reversible. If the scale has a **true zero point** it is a **ratio** scale (e.g. 4 is twice 2 for number of children, but not for temperature in degrees Fahrenheit or Celsius)

Note that in sociological discussion things like age or years of schooling are frequently used as indicators of something less precise such as "experience" or "level of education". It is somewhat dubious whether they really ought to be treated as interval variables in such a context.

When all the cases are grouped into only two categories, according to whether they do, or do not, have a particular characteristic, (e.g. Male - Female, Yes - No) they are known as **dichotomies**. These can always be treated as interval measurements.

The values of some interval variables are **continuous** (Height, age) and sometimes need to be rounded to the nearest integer value. It is important to remember how the rounding has been done: e.g. when height is measured to the nearest inch, 68 inches means from 67.5 inches up to, but not including, 68.5 inches. Someone aged 68 last birthday means from 68 up to, but not including, 69 years old. Thus average age calculated on age last birthday will therefore need to be adjusted by adding six months to the result.

The values of other interval variables are **discrete** (e.g. number of children) and can only increase in increments of 1. Sometimes it is important for statistical calculation to bear this in mind.

You are advised to read **1.1.3 Introduction to the use of computers in survey analysis** before moving on.

Next section: **1.2: Coding data from questionnaires**

Next sessions: **1.2.1 Transfer sheet for data from your questionnaire(s)** (Print up and complete)
1.2.2 Preliminary data exercise (Type data from transfer sheet to *.txt file)

[Back to Block 1 menu]