Chapter 9

Turning Data into Findings

Introduction

In this chapter you will learn about:

- Things to think about at the outset that will help the analysing of quantitative data.
- How to analyse and interpret quantitative data.
- Different types of quantitative data including ordinal and categorical data.
- An introduction to the analysis of qualitative data.

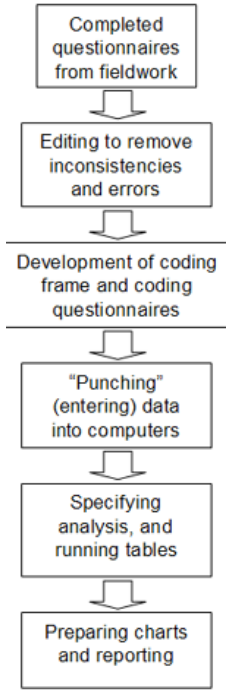
Planning for the analysis of quantitative data

Fieldwork generates hundreds and sometimes thousands of questionnaires. Except in the case of depth interviews, individual questionnaires are of little value or interest. What is required is to generalise from the aggregated data of either the whole sample or of some grouping of respondents out of the whole sample (a sub-sample). Data analysis is the process of aggregating the individual responses or "raw" data.

Very effective analysis can be carried out by using general purpose software such as spreadsheets and databases. Anyone carrying out the work regularly will need specialist software for survey analysis of which there are a dozen or so packages on the market offering various levels of sophistication at different price levels.

There are no paper questionnaires to deal with in a CATI or CAPI survey as the data is entered directly into computers. Where the output of the fieldwork is paper questionnaires, they require editing, coding and "punching" into the computer ready for analysis. The steps involved in data analysis are shown in Figure 9.1.

Figure 9.1 **Steps In The Data Analysis Process**



Editing ensures that each questionnaire is correctly completed, that all routing has been followed and that responses make sense. These checks are either carried out by a fieldwork supervisor or the data is punched into computers and cleaned up through logic checks. Where problems are identified in editing, the questionnaire is abandoned (if nothing can be done to resolve the problems) or corrected automatically by the computer (which follows pre-defined rules to ensure the answers match a logical response).

Key point

Open ended questions should only be included in a survey where it is vital to obtain a verbatim response and where there are less than around 200 interviews. Coding open ended questions requires skill and is costly

Open-ended questions must be coded. In order to code the questionnaires, a coding frame must first be developed by looking over the responses to open ended questions from a sample of questionnaires (enough to ensure all the themes have been captured – which probably will require 100 questionnaires to be looked at). The coding frame is a list of the themes, each with a numeric code which is written, as appropriate, next to the verbatim answers on the questionnaires. Typically, a frame will have no more than a dozen codes, this being sufficient to reduce the verbatims to a manageable listing but providing enough granularity to enable the analyst pick out the different issues.

The person who develops the coding frame must fully understand the objectives of the study as it is important to pull out the right issues. The coding frames are then used by the coding team (a number of coders is likely to be required in a survey of any size if the work is to be completed in a reasonable time).

The edited questionnaires, now have all questions assigned a numeric code. The closed questions have numbers circled next to the answers and the open questions have been closed down and coded. The codes for both types of question are then entered, questionnaire by questionnaire (each forming a record), into the analysis package.

There is always the possibility that the manual task of entering these data codes could be miss-keyed. Quality checks are carried out in which a sample of questionnaires are re-entered by another member of staff. Any inconsistencies that are found in the entries would need to be resolved or would trigger a check of the batch from which the verification sample was taken.

At this point the paper questionnaires, and those that have been entered directly from the keyboard in a CAPI or CATI survey, have reached the same point in the analysis process. The data in the computer must now be programmed to produce an output that helps the researcher analyse the results. The researcher decides what tables are needed (specified in terms of filters and the cross analysis of questions) and gives this specification to data programmers who write the data run programmes to produce the required output.

Some user-friendly analysis packages have menus that allow the researchers to do this themselves.

The cross analysis tables are then run and checked to ensure that they conform to the specification with appropriate filtering, the right cross analysis, suitable labelling and titling. These tables are the main data used for reporting the findings and drawing conclusions.

Think about

A survey about household appliances included the following question:

Why would you not consider buying the appliance in the next two years?

The question was open ended and the following verbatim replies were received to the first 9 questionnaires.

Respondent	Response
1	Too big to go in my kitchen.
2	I cannot to afford to buy one.
3	They look so ugly.
4	I don't like the colours and they cost too much.
5	I hear they are unreliable.
6	With only two of us at home we have no need of one.
7	I expect the prices will come down. I will wait until then.
8	I think they are complicated to use.
9	I don't know really.

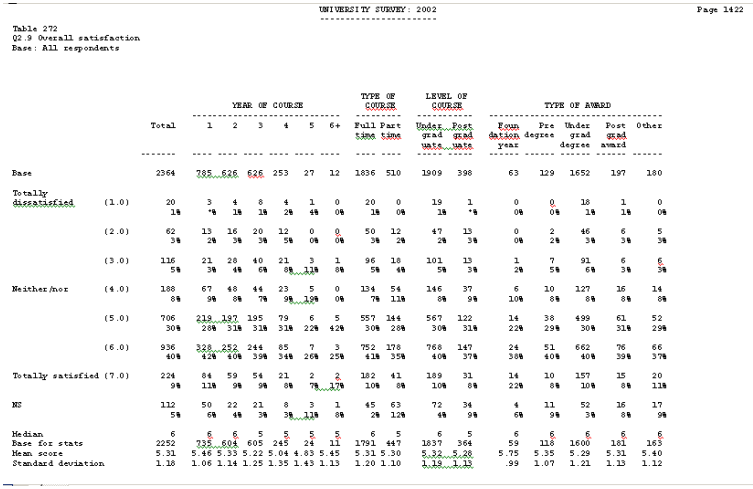
Look through the responses and see how you would group them into a smaller number of themes that could be the basis of a coding frame. There are no right or wrong answers to this and my suggestion is given at the end of this chapter.

An introduction to the analysis of quantitative data

The researcher must decide, out of the surfeit of tables, which is the data that are relevant to the objectives and the survey. This is the vital task of data reduction leading to interpretation. In a survey of students carried out for a university, questions were asked about the

courses that were being studied and the satisfaction with those courses. Well over a thousand pages of tables were generated by cross analyzing each question against a long list of classification questions. Using these cross analyses, the analyst looks for interesting differences between groups of respondents. In the case of the student survey the classification questions included the age of the student, their gender, their religion, where they lived and so on. Figure 9.2 shows an example of a page from this survey .

Figure 9.2 **Table From A University Survey**



UNIVERSITY SURVEY: 2002 — Page 1422

Table 272
Q2.9 Overall satisfaction
Base: All respondents

			YEAR OF COURSE						TYPE OF COURSE		LEVEL OF COURSE		TYPE OF AWARD				
		Total	1	2	3	4	5	6+	Full time	Part time	Under grad uate	Post grad uate	Foun dation year	Pre degree	Under grad degree	Post grad award	Other
Base		2364	785	626	626	253	27	12	1836	510	1909	398	63	129	1652	197	180
Totally dissatisfied	(1.0)	20	3	4	8	4	1	0	20	0	19	1	0	0	18	1	0
		1%	*%	1%	1%	2%	4%	0%	1%	0%	1%	*%	0%	0%	1%	1%	0%
	(2.0)	62	13	16	20	12	0	0	50	12	47	13	0	2	46	6	5
		3%	2%	3%	3%	5%	0%	0%	3%	2%	2%	3%	0%	2%	3%	3%	3%
	(3.0)	116	21	28	40	21	3	1	96	18	101	13	1	7	91	6	6
		5%	3%	4%	6%	8%	11%	8%	5%	4%	5%	3%	2%	5%	6%	3%	3%
Neither/nor	(4.0)	188	67	48	44	23	5	0	134	54	146	37	6	10	127	16	14
		8%	9%	8%	7%	9%	19%	0%	7%	11%	8%	9%	10%	8%	8%	8%	8%
	(5.0)	706	219	197	195	79	6	5	557	144	567	122	14	38	499	61	52
		30%	28%	31%	31%	31%	22%	42%	30%	28%	30%	31%	22%	29%	30%	31%	29%
	(6.0)	936	328	252	244	85	7	3	752	178	768	147	24	51	662	76	66
		40%	42%	40%	39%	34%	26%	25%	41%	35%	40%	37%	38%	40%	40%	39%	37%
Totally satisfied	(7.0)	224	84	59	54	21	2	2	182	41	189	31	14	10	157	15	20
		9%	11%	9%	9%	8%	7%	17%	10%	8%	10%	8%	22%	8%	10%	8%	11%
NS		112	50	22	21	8	3	1	45	63	72	34	4	11	52	16	17
		5%	6%	4%	3%	3%	11%	8%	2%	12%	4%	9%	6%	9%	3%	8%	9%
Median		6	6	6	5	5	5	5	6	5	6	5	6	6	6	6	6
Base for stats		2252	735	604	605	245	24	11	1791	447	1837	364	59	118	1600	181	163
Mean score		5.31	5.46	5.33	5.22	5.04	4.83	5.45	5.31	5.30	5.32	5.26	5.75	5.35	5.29	5.31	5.40
Standard deviation		1.18	1.06	1.14	1.25	1.35	1.43	1.13	1.20	1.10	1.18	1.15	.99	1.07	1.21	1.13	1.12

The table has rows and columns of data showing the satisfaction of students with the University. In this survey, satisfaction was measured using a numeric score from 1 to 7 where 1 is totally dissatisfied and 7 is totally satisfied. The first column of data is the overall or total figure for all students at the University. At the bottom of the "total" column are some key statistics. The mean score shows the arithmetic average and for the University as a whole, this is 5.31 out of 7. (Note: the 112 students who didn't give a score – "not specified" or NS – were not included in this computation). Most of the courses at the university are undergraduate degrees which take three or four years, in the latter case including a year out on a placement.

Casting your eye over the mean scores it becomes apparent that there is an interesting trend in the data. The satisfaction with the

University appears to decline as the students move nearer to the completion of their course. Looking further across the row of mean scores, we see that in the "Type of award" column that there appears to be lower satisfaction levels amongst undergraduates taking a degree than other types of undergraduates or post graduates.

Before we move on to consider the implications of these results it is worth noting the other information on the lower rows of data:

- the median shows the value where half the results larger than this figure and a half are smaller
- the base for the statistics shows the number of people who have answered the question. This is useful to establish the robustness of the result. For example, in the table there are very few respondents that have been studying 5 years or more so we would have to be very careful in the interpretation of this group. In all the other groups there are a few hundred respondents and this gives us more confidence in the result (as long as it has been answered by a representative group).
- The standard deviation shows the spread or dispersion of the results around the mean. A low standard deviation indicates that there is a clustering of responses around the mean.

We have been concentrating on the summary statistics at the bottom of the table. The researcher could be just as interested in the percentage of people who gave a score of 6 or 7 (we call this the "top box" result) as this indicates how many are really satisfied. Equally, there will be interest in the percentage who give very low scores because addressing their problems could be one route to improving overall satisfaction.

The researcher must not just report on the data, but attempt to interpret what it could mean. What are the possible causes of a lower level of satisfaction amongst students who have been studying at the University for longer? We do not know for sure but we can hypothesize that it could be the result of one or more factors:

- Students arrive with high expectations which gradually get disillusioned the longer they stay at the University
- The courses get harder nearer to graduation and this takes the enjoyment and satisfaction out of studying at the University

- Students move towards the end of their studies, they become more concerned about the ability to get a job with a degree from this University
- Teaching deteriorates as the courses progress into the third and fourth years.

Key point

Cross tabulations are the main output from a quantitative survey and are used to pick out different responses between different groups of people.

There may be other factors as well influencing the decline in satisfaction over time spent at the University. It is important to establish what the reason is for the decline because only then can it be addressed and rectified. Other data, elsewhere in the survey may help establish the reason (or reasons) or further research may be needed which concentrates on this question alone.

Think about

Consider the "top box" results in figure 9.2 (ie those respondents giving a score of 6 or 7 as a satisfaction score). What patterns of response can you spot across the different groups of respondents? Which result do you prefer to use – the mean score or the top box? Why?

Types of quantitative data

The rating of satisfaction in the University survey used a scale from 1 to 7 – 1 being not satisfied and 7 being totally satisfied. This is an *ordinal* scale as opposed to an interval scale (such as degrees centigrade). In an ordinal scale, the distance between each number is not necessarily uniform. For example, the half way point in a satisfaction scale of this type is not necessarily 4. It is quite easy to obtain satisfaction sores of around 5 but getting scores of 6 or 7 is more difficult. The effort required to boost the average satisfaction score just one or two fractions would be high when the levels of satisfaction are over 5 (still using the scale from 1 to 7).

Other data market researcher analyse is categorical – in other words it measures how many respondents are in each category. We may be interested in which categories of information are used to find out about the University. The result of this type of analysis can be pre-

sented as a table such as Figure 9.3. In this case the table shows the responses from 2,364 students. Note that the column does not total 100%; this is because of multi-response – some students mentioned more than one source of information.

Figure 9.3 **The Importance Of Different Sources Of Information On The University**

Sources Of Information On The University	**%**
University prospectus	65
University open day	43
A friend or relation told you about it	33
University web site	32
A teacher at school or college told you about it	24
Visited a friend at the university	15
The city web site	9
Someone from the University visiting your school or college	7
Base:	2,364

Students filled in a self-completion questionnaire which had a pre-defined list of sources from which they were asked to select those that they used to obtain information on the University. The results in Figure 9.3 have been presented in declining order of frequency of mention to help the reader focus on the most important. Tables showing this type of response can, of course, also include cross analyses and not just show the response for the total sample.

By looking at data that has been cross analysed, the researcher can pick out a relationship between two variables – as we have seen, the relationship between student satisfaction and the number of years they have been studying at the University. The relationship between three dimensions can (if with more difficulty) also be examined in a table. We could for example take the age of the student as the third dimension. The relationship between the three variables can also be represented in some sort of three dimensional table with a vertical axis although to do so and read it would be no easy task. And, why stop at three variables? The investigation of relationships between any number of variables may be worthwhile and produce a model which offers useful insights into how a market works and, therefore, provides guidance to effective marketing. The relationship between more than two or three variables is the outcome of *multivariate analysis* which is increasingly used in market research –

particularly for handling product attribute and attitude data. In part the uptake of these techniques is because the mechanics of carrying out complex statistical operations has been made so much easier through widely available software run on PCs.

Marketing planning is now very much based on segmentation. The age of mass markets is waning and increasingly, strategies are aimed at influencing specific market segments or niches. Segments can be defined in terms of "objective" variables such as demographics – an approach that has been widely used for many years. Conventional cross analysis of data is usually sufficient to segment markets in these terms. However, another approach is to focus on more subjective factors and especially consumers' attitudes. Using appropriate scalar questions any number of such attitude variables can be obtained. The question then arises as to how these can be used to group consumers into homogeneous segments, each with people in them that have a bundle of common attitudes. Two multivariate techniques are used for such segmentation – *factor analysis* and *cluster analysis*.

Factor analysis focuses on the attitude attributes themselves and where a lot of attitude questions have been asked, it reduces them to a smaller number of *component factors* or groupings of attitudes which on the basis of responses appear to be empirically linked.

Key point

Using market research to develop a needs based segmentation can give a company a substantial competitive advantage.

Cluster analysis on the other hand focuses on respondents themselves. As the term suggests it groups or clusters the data into relatively homogeneous groups on the basis of their attitudes to the product. These clusters may represent people with particular needs such as low prices or convenience. Factor analysis and cluster analysis are often carried out together with clusters defined in terms of component factors from preliminary factor analysis.

Products and brands can be analysed in terms of any number of attributes, limited only by what is included in the questionnaire. Not all attributes are, however, equally important; they almost certainly fall into some sort of hierarchy; in some markets value for money may lead the ranking followed by the product quality, followed by availability, ease of doing business etc. This hierarchy can be established by direct questioning (eg please rank the following in terms of their importance....) but this simple approach may not produce a realistic model of the consumer choice process – consumers just do not think in this way when making an

actual purchase decision. An alternative is to link preferences for products (which can be bundles of attributes) to the price they are prepared to pay for them. The importance of the attributes is then derived from the two sets of data at the analysis stage.

Key point

Using conjoint analysis, researchers can work out the value attached to the components of an offer without directly asking how much these individual components are valued.

A widely used multi-variate technique to achieve this is *conjoint analysis* which calculates "utility values" for attributes. Conjoint analysis is a sophisticated technique and there are technical issues that need to be considered. In particular, the design of attributes is a crucial step in a conjoint project as choices between poorly defined levels can make the exercise meaningless. Also there are different types of conjoint analysis each suited to a particular application – the trade off approach (adaptive conjoint analysis) is the most common, but there is also full-profile conjoint analysis. In the trade-off approach respondents are asked to rank all combinations of attribute levels taking two attributes at a time. Conversely in the full profile approach respondents are requested to rank alternatives described in terms of all associated attributes. The trade-off approach is much easier to use over the telephone where pairs of attributes can be read out to people and they can choose between them whereas the full conjoint requires people to see the whole concept laid out, often with a picture, a description of all its features and attributes and its price.

If you want to read up further visit www.sawtooth.com. Sawtooth provide most of the conjoint interviewing software around the world and have both simple and detailed explanations on their web site.

An introductionn to the recording and analysis of qualitative data

In qualitative research, the samples are smaller than in quantitative surveys and there are likely to be 50 or less questionnaires to process. However, the data may (or should) be more subtle and complex. It is likely, for example, that questions will be mainly open-ended and the interviewer will have prompted for full responses. Also the interview or discussion may be unstructured with the sequence of questions varying between different respon-

dents. In coding open ended responses there will be a loss of detail but this is often necessary to obtain a feel for the magnitude of the response.

If there are only a small number of responses it may be sufficient for the researcher to read through the scripts. Common topics can be cut and pasted into an Excel spread sheet to sort into groups, each tagged, where possible, with classification data. In the report these verbatim comments make powerful illustrations of points as they are seen to come from the mouths of respondents and not the person giving the presentation.

Where interviews or group discussions have been taped or digitally recorded – a common practice in qualitative research – it is generally considered good practice to type them up and it is from these transcriptions that the researcher carries out the analysis. The verbatim transcriptions of these discussions may require some tidying up of the text so that it makes sense (but without any attempt to change the meaning). It will be obvious that whilst tape recording interviews is an efficient way of capturing accurately all that is said at an interview, it creates additional work later as it requires approximately as long to transcribe the interview as it does to conduct it. This is one reason why qualitative research is an expensive process.

Software packages offer some help to the qualitative researcher in the analysis of their material. In the main, these look for words or word strings and are useful for carrying out frequency counts of words to establish their importance in the discussion. However, nothing has replaced the value of the qualitative researcher themselves who, having carried out the interviews and soaked themselves in the output, are best placed to prepare and deliver the report.

In the analysis of qualitative market research data the researcher is seeking to do four things:

- Identify themes showing how and where they originate
- Clarifying the meaning of these themes in the context of the research project
- Identifying the frequency with which the themes crop up and therefore establishing so far as is possible, how important they are
- Noting down exceptions and unusual themes in order to see if they are the beginning of a trend or exceptional occurrences.

Key point

There are very few short cuts in the analysis of qualitative data. The researchers have to steep themselves in the data and pull out the important issues.

In qualitative research, much depends on the flair and interpretation put on the data by the practitioners involved. No two qualitative researchers are likely to produce identical outputs from their fieldwork and nor would they analyse the data in the same way. To this extent, the analysis of qualitative data is very different from quantitative data. It involves a small team of experienced researchers from beginning to end. Compare this with the large teams of interviewers, editors, coders and programmers that work on quantitative surveys.

Regardless of which method is used in the qualitative research (depth interviews, focus groups, observation), there are a number of recommended procedures:

- Carry out the analysis as soon as possible after data collection
- Look at the amount of time respondents spend on particular themes
- Look at how much data was given naturally and spontaneously, rather than prompted
- Identify the force of reactions to different issues (this is done by watching people's faces and behaviour as well as listening to what they say and how they say it)
- Differentiate between honest and socially acceptable answers
- Look for majority and minority opinions
- Look for consistencies and inconsistencies in answers and reactions (and probe if you don't understand something)

SCARY STORY

A customer satisfaction survey of buyers of PVC raw materials was nearing completion. Around 200 interviews had been carried out and the analyst was working through the cross tabulations, preparing the charts for the report. Time as always was short and there were only a few days to go to the presentation.

However, the researcher was uneasy. Bits of data did not stack up. There were inconsistencies in the products companies said they were making and the raw materials they were buying. Some suppliers of raw materials had higher market shares and some had lower market shares than was expected. A request to the bureau that produced the tables provided the assurance that the results were spot on.

This was not a very large survey and it was an easy job for the researcher to carry out a quick count on two or three questions using the hard copy questionnaires. This produced counts that did not tally with the tables.

The bureau producing the tables was asked to re-punch the data and produce new tables at breakneck speed. The new tables were quite different from the original ones, but at least they made sense.

A post mortem on the project showed that in the panic to get the job done on time, the tabulation bureau had taken on new labour to enter the data. The "punchers" are paid on a piece rate – so much per questionnaire – and one of them had entered only a handful of questionnaires, copying each several times. As a result, some questionnaires had multiple entries and this was skewing the data.

The story illustrates the difficulty of spotting errors in data analysis. Double punching a sample of questionnaires is a useful check to ensure that the data entry is carried out correctly, but there could still be errors in the programming.

The key learning from this story is not to fall into the temptation of force fitting data into the findings, just because you are working to a tight deadline and there is no time to carry out new work.

If something doesn't look right, it probably isn't and the best check of all is to get as near to the raw data as possible. This could involve looking at the original questionnaires and carrying out a count of responses to key questions or looking through the data files of each entry on a CATI or CAPI system. Tedious it may be but worth it because no researcher will rest easy unless they are confident that what they are presenting is as close to the truth as possible.

Suggested Coding Of Open-Ended Question on Page 153

Response Group	Respondents included
Design of the appliance	1,3,4,8
Cost factors	2,4,7
Unreliability	5
Have no need	6
Don't know	9