Approaches to the Analysis of Survey Data

March 2001

The University of Reading
Statistical Services Centre

Biometrics Advisory and
Support Service to DFID

Contents

1. Preparing for the Analysis

1.1 Introduction

This guide is concerned with some fundamental ideas of analysis of data from surveys. The discussion is at a statistically simple level; other more sophisticated statistical approaches are outlined in our guide *Modern Methods of Analysis*. Our aim here is to clarify the ideas that successful data analysts usually need to consider to complete a survey analysis task purposefully.

An ill-thought-out analysis process can produce incompatible outputs and many results that never get discussed or used. It can overlook key findings and fail to pull out the subsets of the sample where clear findings are evident. Our brief discussion is intended to assist the research team in working systematically; it is no substitute for clear-sighted and thorough work by researchers. We do not aim to show a totally naïve analyst exactly how to tackle a particular set of survey data. However, we believe that where readers can undertake basic survey analysis, our recommendations will help and encourage them to do so better.

Chapter 1 outlines a series of themes, after an introductory example. Different data types are distinguished in section 1.2. Section 1.3 looks at data structures; simple if there is one type of sampling unit involved, and hierarchical with e.g. communities, households and individuals. In section 1.4 we separate out three stages of survey data handling – exploration, analysis and archiving – which help to define expectations and procedures for different parts of the overall process. We contrast the research objectives of description or estimation (section 1.5), and of comparison (section 1.6) and what these imply for analysis. Section 1.7 considers when results should be weighted to represent the population – depending on the extent to which a numerical value is or is not central to the interpretation of survey results. In section 1.8 we outline the coding of non-numerical responses. The use of ranked data is discussed in brief in section 1.9.

In Chapter 2 we look at the ways in which researchers usually analyse survey data. We focus primarily on tabular methods, for reasons explained in section 2.1. Simple one-way tables are often useful as explained in section 2.2. Cross-tabulations (section 2.3) can take many forms and we need to think which are appropriate. Section 2.4 discusses issues about ‘accuracy’ in relation to two- and multi-way tables. In section 2.5 we briefly discuss what to do when several responses can be selected in response to one question.

Cross-tabulations can look at many respondents, but only at a small number of questions, and we discuss profiling in section 2.6, cluster analysis in section 2.7, and indicators in sections 2.8 and 2.9.

1.2 Data Types

Introductory Example: On a nominal scale the categories recorded, usually counted, are described verbally. The ‘scale’ has no numerical characteristics. If a single one-way table resulting from simple summarisation of nominal (also called categorical) scale data contains frequencies:-

Christian	Hindu	Muslim	Sikh	Other
29	243	117	86	25

there is little that can be done to present exactly the same information in other forms. We could report highest frequency first as opposed to alphabetic order, or reduce the information in some way e.g. if one distinction is of key importance compared to the others:-

Hindu	Non-Hindu
243	257

On the other hand, where there are *ordered* categories, the sequence makes sense only in one, or in exactly the opposite, order:-

Excellent	Good	Moderate	Poor	Very Bad
29	243	117	86	25

We could reduce the information by combining categories as above, but also we can summarise, somewhat numerically, in various ways. For example, accepting a degree of arbitrariness, we might give scores to the categories:-

Excellent	Good	Moderate	Poor	Very Bad
5	4	3	2	1

and then produce an ‘average score’ – a numerical indicator – for the sample of:-

$$\frac{29 \times 5 + 243 \times 4 + 117 \times 3 + 86 \times 2 + 25 \times 1}{29 + 243 + 117 + 86 + 25} = 3.33$$

This is an analogue of the arithmetical calculation we would do if the categories really were numbers e.g. family sizes.

The same average score of 3.33 could arise from differently patterned data e.g. from rather more extreme results:-

Excellent	Good	Moderate	Poor	Very Bad
79	193	117	36	75

Hence, as with any other indicator, this ‘average’ only represents one feature of the data and several summaries will sometimes be needed.

A major distinction in statistical methods is between quantitative data and the other categories exemplified above. With quantitative data, the difference between the values from two respondents has a clearly defined and incontrovertible meaning e.g. “It is 5C° hotter now than it was at dawn” or “You have two more children than your sister”. Commonplace statistical methods provide many well-known approaches to such data, and are taught in most courses, so we give them only passing attention here.

In this guide we focus primarily on the other types of data, coded in number form but with less clear-cut numerical meaning, as follows. Binary – e.g. yes/no data – can be coded in 1/0 form; while purely categorical or nominal data – e.g. caste or ethnicity – may be coded 1, 2, 3… using numbers that are just arbitrary labels and cannot be added or subtracted. It is also common to have ordered categorical data, where items may be rated *Excellent, Good, Poor, Useless,* or responses to attitude statements may be *Strongly agree, Agree, Neither agree nor disagree, Disagree, Strongly disagree*. With ordered categorical data the number labels should form a rational sequence, because they have some numerical meaning e.g. scores of 4, 3, 2, 1 for *Excellent* through to *Useless*. Such data supports limited quantitative analysis, and is often referred to by statisticians as ‘qualitative’ – this usage does not imply that the elicitation procedure must satisfy a purist’s restrictive perception of what constitutes qualitative research methodology.

1.3 Data Structure

SIMPLE SURVEY DATA STRUCTURE: the data from a single-round survey, analysed with limited reference to other information, can often be thought of as a ‘flat’ rectangular file of numbers, whether the numbers are counts/measurements, or codes, or a mixture. In a structured survey with numbered questions, the flat file has a column for each question, and a row for each respondent, a convention common to almost all standard statistical packages. If the data form a perfect rectangular grid with a number in every cell, analysis is made relatively easy, but there are many reasons why this will not always be the case and flat file data will be incomplete or irregular. Most importantly:-

- Surveys often involve ‘skip’ questions where sections are missed out if irrelevant e.g. details of spouse’s employment do not exist for the unmarried. These arise legitimately, but imply different subsets of people respond to different questions. ‘Contingent questions’, where not everyone ‘qualifies’ to answer, often lead to inconsistent-seeming results for this reason. If the overall sample size is just adequate, the subset who ‘qualify’ for a particular set of contingent questions may be too small to analyse in the detail required.
- If some respondents fail to respond to some questions (item non-response) there will be holes in the rectangle. Non-informative non-response occurs if the data is missing for a reason unrelated to the true answers e.g. the interviewer turned over two pages instead of one! Informative non-response means that the absence of an answer itself tells you something, e.g. you are almost sure that the missing income value will be one of the highest in the community. A little potentially informative non-response may be ignorable, if there is plenty of data. If data are sparse or if informative non-response is frequent, the analysis should take account of what can be inferred from knowing that there are informative missing values.

HIERARCHICAL DATA STRUCTURE: another complexity of survey data structure arises if the data are hierarchical. A common type of hierarchy is where a series of questions is repeated say for each child in the household, and combined with a household questionnaire, and maybe data collected at community level.

For analysis, we can create a rectangular flat file, at the ‘child level’, by repeating relevant household information in separate rows for each child. Similarly, we can summarise information for the children in a household, to create a ‘household level’ analysis file. The number of children in the household is usually a desirable part of the summary; this “post-stratification” variable can be used to produce sub-group analyses at household level separating out households with different numbers of child members. The way the sampling was done can have an effect on interpretation or analysis of a hierarchical study. For example if children were chosen at random, households with more children would have a greater chance of inclusion and a simple average of the household sizes would be biased upwards: it should be corrected for selection probabilities.

Hierarchical structure becomes important, and harder to handle, if there are many levels where data are collected e.g. government guidance and allocations of resource, District Development Committee interpretations of the guidance, Village Task Force selections of safety net beneficiaries, then households and individuals whose vulnerabilities and opportunities are affected by targeting decisions taken at higher levels in the hierarchy. In such cases, a relational database reflecting the hierarchical structure is a much more desirable way than a spreadsheet to define and retain the inter-relationships between levels, and to create many analysis files at different levels. Such issues are described in the guide *The Role of a Database Package for Research Projects*. Any one of the analysis files may be used as we discuss below, but any such study will be looking at one facet of the structure, and several analyses will have to be brought together for an overall interpretation. A more sophisticated approach using multi-level modelling, described in our guide on *Modern Methods of Analysis*, provides a way to look at several levels together.

1.4 Stages of Analysis

It is often worth distinguishing the three stages of exploratory analysis, deriving the main findings, and archiving.

EXPLORATORY DATA ANALYSIS (EDA) means looking at the data files, maybe even before all the data has been collected and entered, to get an idea of what is there. It can lead to additional data collection if this is seen to be needed, or savings by stopping collecting data when a conclusion is already clear, or existing results prove worthless. It is *not* assumed that results from EDA are ready for release as study findings.

- EDA usually overlaps with data cleaning; it is the stage where anomalies become evident e.g. individually plausible values may lead to a way-out point when combined with other variables on a scatterplot. In an ideal situation, EDA would end with confidence that one has a clean dataset, so that a single version of the main datafiles can be finalised and ‘locked’ and all published analyses derived from a single consistent form of ‘the data’. In practice later stages of analysis often produce additional queries about data values.
- Such exploratory analysis will also show up limitations in contingent questions e.g. we might find we don’t have enough currently married women to analyse their income sources separately by district. EDA should include the final reconciliation of analysis ambitions with data limitations.
- This phase can allow the form of analysis to be tried out and agreed, developing analysis plans and program code in parallel with the final data collection, data entry and checking. Purposeful EDA allows the subsequent stage of deriving the main findings to be relatively quick, uncontroversial, and well organised.

DERIVING THE MAIN FINDINGS: the second stage will ideally begin with a clear-cut clean version of the data, so that analysis files are consistent with one another, and any inconsistencies, e.g. in numbers included, can be clearly explained. This is the stage we amplify upon, later in this guide. It should generate the summary findings, relationships, models, interpretations and narratives, and first recommendations that research users will need to *begin* utilising the results.

Of course one needs to allow time for ‘extra’ but usually inevitable tasks such as:-

- follow-up work to produce further more detailed findings, e.g. elucidating unexpected results from the pre-planned work.
- a change made to the data, each time a previously unsuspected recording or data entry error comes to light. Then it is important to correct the database and all analysis files already created that involve the value to be corrected. This will mean repeating analyses that have already been done using, but not revealing, the erroneous value. If that analysis was done “by mouse clicking” and with no record of the steps, this can be very tedious. This stage of work is best undertaken using software that can keep a log: it records the analyses in the form of program instructions that can readily and accurately be re-run.

ARCHIVING means that data collectors keep, perhaps on CD, all the non-ephemeral material relating to their efforts to acquire information. Obvious components of such a record include:-

(i) data collection instruments, (ii) raw data, (iii) metadata recording the what, where, when, and other identifiers of all variables, (iv) variable names and their interpretations, and labels corresponding to values of categorical variables, (v) query programs used to extract analysis files from the database, (vi) log files defining the analyses, and (vii) reports. Often georeferencing information, digital photographs of sites and scans of documentary material are also useful. Participatory village maps, for example, can be kept for reference as digital photographs.

Surveys are often complicated endeavours where analysis covers only a fraction of what could be done. Reasons for developing a good management system, of which the archive is part, include:-

- keeping the research process organised as it progresses;
- satisfying the sponsor’s (e.g. DFID’s) contractual requirement that data should be available if required by the funder or by legitimate successor researchers;
- permitting a detailed re-analysis to authenticate the findings if they are questioned;
- allowing a different breakdown of results e.g. when administrative boundaries are redefined;
- linking several studies together, for instance in longer-term analyses carrying baseline data through to impact assessment.

1.5 Population Description as the Major Objective

In the next section we look at the objective of comparing results from sub-groups, but a more basic aim is to estimate a characteristic like the absolute number in a category of proposed beneficiaries, or a relative number such as the prevalence of HIV sero-positives. The estimate may be needed to describe a whole population or sections of it.

In the basic analyses discussed below, we need to bear in mind both the planned and the achieved sampling structure.

Example: Suppose 'before' and 'after' surveys were each planned to have a 50:50 split of urban and rural respondents. Even if we achieved 50:50 splits, these would need some manipulation if we wanted to generalise the results to represent an actual population split of 70:30 urban:rural. Say we wanted to assess the change from 'before' to 'after' and the achieved samples were in fact split 55:45 and 45:55. We would have to correct the results carefully to get a meaningful estimate of change.

Samples are often stratified i.e. structured to capture and represent particular segments of the target population. This may be much more sophisticated than the urban/rural split in the previous paragraph. Within-stratum summaries serve to describe and characterise each of these parts individually. If required by the objectives, overall summaries, which put together the strata, need to describe and characterise the whole population.

It may be fine to treat the sample as a whole and produce simple, unweighted summaries if (i) we have set out to sample the strata proportionately, (ii) we have achieved this, and (iii) there are no problems due to hierarchical structure. Non-proportionality arises from various quite distinct sources, in particular:-

- Case A: often sampling is disproportionate across strata by design, e.g. the urban situation is more novel, complex, interesting or accessible, and gets greater coverage than the fraction of the population classed as rural.
- Case B : sometimes particular strata are bedevilled with high levels of non-response, so that the data are not proportionate to stratum sizes, even when the original plan was that they should be.

If we ignore non-proportionality, a simple-minded summary over all cases is *not* a proper representation of the population in these instances.

The 'mechanistic' response to 'correct' both the above cases is (1) to produce within-stratum results (tables or whatever), (2) to scale the numbers in them to represent the true population fraction that each stratum comprises, and then (3) to combine the results.

There is often a problem with doing this in case B, where non-response is an important part of the disproportionality: the reasons why data are missing from particular strata often correspond to real differences in the behaviour of respondents, especially those omitted or under-sampled, e.g. “We had very good response rates everywhere except in the north. There a high proportion of the population are nomadic, and we largely failed to find them.” Just scaling up data from settled northerners does not take account of the different lifestyle and livelihood of the missing nomads. If you have largely missed a complete category, it is honest to report partial results making it clear which categories are not covered and why.

One common ‘sampling’ problem arises when a substantial part of the target population is unwilling or unable to cooperate, so that the results in effect only represent a limited subset – those who volunteer or agree to take part. Of course the results are biased towards e.g. those who command sufficient resources to afford the time, or e.g. those who habitually take it upon themselves to represent others. We would be suspicious of any study which appeared to have relied on volunteers, but did not look carefully at the limits this imposed on the generalisability of the conclusions.

If you have a low response rate from one stratum, but are still prepared to argue that the data are somewhat representative, the situation is at the very least uncomfortable. Where you have disproportionately few responses, the multipliers used in scaling up to ‘represent’ the stratum will be very high, so your limited data will be heavily weighted in the final overall summary. If there is any possible argument that these results are untypical, it is worthwhile to think carefully before giving them extra prominence in this way.

1.6 Comparison as the Major Objective

One sound reason for disproportionate sampling is that the main objective is a comparison of subgroups in the population. Even if one of two groups to be compared is very small, say 10% of the total number in the population, we now want roughly equally many observations from each subgroup, to describe both groups roughly equally accurately. There is no point in comparing a very accurate set of results from one group with a very vague, ill-defined description of the other; the comparison is at least as vague as the worse description. The same broad principle applies whether the comparison is a wholly quantitative one looking at the difference in means of a numerical measure between groups, or a much looser verbal comparison e.g. an assessment of differences in pattern across a range of cross-tabulations.

If for a subsidiary objective we produce an overall summary giving ‘the general picture’ of which both groups are part, 50:50 sampling may need to be re-weighted 90:10 to produce a quantitative overall picture of the sampled population.

The great difference between true experimental approaches and surveys is that experiments usually involve a relatively specific comparison as the major objective, while surveys much more often do not. Many surveys have multiple objectives, frequently ill defined, often contradictory, and usually not formally prioritised. Along with the likelihood of some non-response, this tends to mean there is no sampling scheme which is best for all parts of the analysis, so various different weighting schemes may be needed in the analysis of a single survey.

1.7 When Weighting Matters

Several times in the above we have discussed issues about how survey results may need to be scaled or weighted to allow for, or ‘correct for’, inequalities in how the sample represents the population. Sometimes this is of great importance, sometimes not. A fair evaluation of survey work ought to consider whether an appropriate trade-off has been achieved between the need for accuracy and the benefits of simplicity.

If the objective is formal estimation, e.g. of total population size from a census of a sample of communities, we are concerned to produce a *strictly numerical answer*, which we would like to be as accurate as circumstances allow. We should then correct as best we can for a distorted representation of the population in the sample. If groups being formally compared run across several population strata, we should try to ensure the comparison is fair by similar corrections, so that the groups are compared on the basis of consistent samples. In these cases we have to face up to problems such as unusually large weights attached to poorly-responding strata, and we may need to investigate the extent to which the final answer is dubious because of sensitivity to results from such subsamples.

Survey findings are often used in *‘less numerical’ ways*, where it may not be so important to achieve accurate weighting e.g. “whatever varieties they grow for sale, a large majority of farm households in Sri Lanka prefer traditional red rice varieties for home consumption because they prefer their flavour”. If this is a clear-cut finding which accords with other information, if it is to be used for a simple decision process, or if it is an interim finding which will prompt further investigation, there is a lot to be said for keeping the analysis simple. Of course it saves time and money. It makes the process of interpretation of the findings more accessible to those not very involved in the study. Also, weighting schemes depend on good information to create the weighting factors and this may be hard to pin down.

Where we have *worryingly large weights*, attaching to small amounts of doubtful information, it is natural to want to put limits on, or ‘cap’, the high weights, even at the expense of introducing some bias, i.e. to prevent any part of the data having too much impact on the result.

The ultimate form of capping is to express doubts about all the data, and to give equal weight to every observation. The rationale, not usually clearly stated, even if analysts are aware they have done this, is to minimise the maximum weight given to any data item. This lends some support to the common practice of analysing survey data as if they were a simple random sample from an unstructured population. For ‘less numerical’ usages, this may not be particularly problematic as far as simple description is concerned. Of course it is wrong – and may be very misleading – to follow this up by calculating standard deviations and making claims of accuracy about the results which their derivation will not sustain!

1.8 Coding

We recognise that purely qualitative researchers may prefer to use qualitative analysis methods and software, but where open-form and other verbal responses occur alongside numerical data it is often sensible to use a quantitative tool. From the statistical viewpoint, basic coding implies that we have material, which can be put into nominal-level categories. Usually this is recorded in verbal or pictorial form, maybe on audio- or videotape, or written down by interviewers or self-reported. We would advocate computerising the raw data, so it is archived. The following refers to extracting codes, usually describing the routine comments, rather than unique individual ones which can be used for subsequent qualitative analysis.

By scanning the set of responses, themes are developed which reflect the items noted in the material. These should reflect the objectives of the activity. It is not necessary to code rare, irrelevant or uninteresting material.

In the code development phase, a large enough range of the responses is scanned to be reasonably sure that commonly occurring themes have been noted. If previous literature, or theory, suggests other themes, these are noted too. Ideally, each theme is broken down into unambiguous, mutually exclusive and exhaustive, categories so that any response segment can be assigned to just one, and assigned the corresponding code value. A ‘codebook’ is then prepared where the categories are listed and codes assigned to them. Codes do not have to be consecutive numbers. It is common to think of codes as presence/absence markers, but there is no intrinsic reason why they should not be graded as ordered categorical variables if appropriate, e.g. on a scale such as *fervent, positive, uninterested/no opinion, negative*.

The entire body of material is then reviewed and codes are recorded. This may be in relevant places on questionnaires or transcripts. Especially when looking at 'new' material not used in code development, extra items may arise and need to be added to the codebook. This may mean another pass through material already reviewed, to add new codes e.g. because a particular response is turning up more than expected.

From the point of view of analysis, no particular significance attaches to particular numbers used as codes, but it is worth bearing in mind that statistical packages are usually excellent at sorting, selecting or flagging, for example, 'numbers between 10 and 19' and other arithmetically defined sets. If these all referred to a theme such as 'forest exploitation activities of male farmers' they could easily be bundled together. It is of course impossible to separate out items given the same code, so deciding the right level of coding detail is essential at an early stage in the process.

When codes are analysed, they can be treated like other nominal or ordered categorical data. The frequencies of different types of response can be counted or cross-tabulated. Since they often derive from text passages and the like, they are often particularly well-adapted for use in sorting listings of verbal comments – into relevant bundles for detailed non-quantitative analysis.

1.9 Ranking & Scoring

A common means of eliciting data is to ask individuals or groups to rank a set of options. The researchers' decision to use ranks in the first place means that results are less informative than scoring, especially if respondents are forced to choose between some nearly-equal alternatives and some very different ones. A British 8-year-old offered baked beans on toast, or fish and chips, or chicken burger, or sushi with hot radish might rank these 1, 2, 3, 4 but score them 9, 8.5, 8, and 0.5 on a zero to ten scale!

Ranking is an easy task where the set of ranks is not required to contain more than about four or five choices. It is common to ask respondents to rank, say, their best four from a list of ten, with 1 = best, etc. Accepting a degree of arbitrariness, we would usually replace ranks 1, 2, 3, 4, and a string of blanks by pseudo-scores 4, 3, 2, 1, and a string of zeros, which gives a complete array of numbers we can summarise – rather than a sparse array where we don't know how to handle the blanks. A project output paper[†] available on the SSC website explores this in more detail.

[†] *Converting Ranks to Scores for an ad hoc Assessment of Methods of Communication Available to Farmers* by Savitri Abeyasekera, Julie Lawson-Macdowell & Ian Wilson. This is an output from DFID-funded work under the Farming Systems Integrated Pest Management Project, Malawi and DFID NRSP project R7033, Methodological Framework for Combining Qualitative and Quantitative Survey Methods.

Where the instructions were to rank as many as you wish from a fixed, long list, we would tend to replace the variable length lists of ranks with scores. One might develop these as if respondents each had a fixed amount, e.g. 100 beans, to allocate as they saw fit. If four were chosen these might be scored 40, 30, 20, 10, or with five chosen 30, 25, 20, 15, 10, with zeros again for unranked items. These scores are arbitrary e.g. 40, 30, 20, 10 could instead be any number of choices e.g. 34, 28, 22, 16 or 40, 25, 20, 15; this reflects the rather uninformative nature of rankings, and the difficulty of *post hoc* construction of information that was not elicited effectively in the first place.

Having reflected and having replaced ranks by scores we would usually treat these like any other numerical data, with one change of emphasis. Where results might be sensitive to the actual values attributed to ranks, we would stress sensitivity analysis more than with other types of numerical data, e.g. re-running analyses with (4, 3, 2, 1, 0, 0, …) pseudo-scores replaced by (6, 4, 2, 1, 0, 0 , …). If the interpretations of results are insensitive to such changes, the choice of scores is not critical.

2. Doing the Analysis

2.1 Approaches

Data listings are readily produced by database and many statistical packages. They are generally on a case-by-case basis, so are particularly suitable in EDA as a means of tracking down odd values, or patterns, to be explored. For example, if material is in verbal form, such a listing can give exactly what every respondent was recorded as saying. Sorting these records – according to who collected them, say – may show up great differences in field workers' aptitude, awareness or approach. Data listings can be an adjunct to tabulation: in Excel, for example, the Drill Down feature allows one to look at the data from individuals who appear together in a single cell.

There is a place for the use of graphical methods, especially for presentational purposes, where simple messages need to be given in easily understood, and attention-grabbing form. Packages offer many ways of making results bright and colourful, without necessarily conveying more information or a more accurate understanding. A few basic points are covered in the guide on *Informative Presentation of Tables, Graphs and Statistics*.

Where the data are at all voluminous, it is a good idea selectively to tabulate most 'qualitative' but numerically coded data i.e. the binary, nominal or ordered categorical types mentioned above. Tables can be very effective in presentations if stripped down to focus on key findings, crisply presented. In longer reports, a carefully crafted, well documented, set of cross-tabulations is usually an essential component of summary and comparative analysis, because of the limitations of approaches which avoid tabulation:-

- Large numbers of charts and pictures can become expensive, but also repetitive, confusing and difficult to use as a source of detailed information.
- With substantial data, a purely narrative full description will be so long-winded and repetitive that readers will have great difficulty getting a clear picture of what the results have to say. With a briefer verbal description, it is difficult not to be overly selective. Then the reader has to question why a great deal went into collecting data that merits little description, and should question the impartiality of the reporting.
- At the other extreme, some analysts will skip or skimp the tabulation stage and move rapidly to complex statistical modelling. Their findings are just as much to be distrusted! The models may be based on preconceptions rather than evidence, they may fit badly and conceal important variations in the underlying patterns.

- In terms of producing final outputs, data listings seldom get more than a place in an appendix. They are usually too extensive to be assimilated by the busy reader, and are unsuitable for presentation purposes.

2.2 One-Way Tables

The most straightforward form of analysis, and one that often supplies much of the basic information need, is to tabulate results, question by question, as ‘one-way tables’. Sometimes this can be done using an original questionnaire and writing on it the frequency or number of people who ‘ticked each box’. Of course this does not identify which respondents produced particular combinations of responses, but this is often a first step where a quick and/or simple summary is required.

2.3 Cross-Tabulation: Two-Way & Higher-Way Tables

At the most basic level, cross-tabulations break down the sample into two-way tables showing the response categories of one question as row headings, those of another question as column headings. If for example each question has five possible answers the table breaks the total sample down into 25 subgroups.

If the answers are subdivided e.g. by sex of respondent, there will be one three-way table, 5x5x2, probably shown on the page as separate two-way tables for males and for females. The total sample size is now split over 50 categories and the degree to which the data can sensibly be disaggregated will be constrained by the total number of respondents represented.

There are usually many possible two-way tables, and even more three-way tables. The main analysis needs to involve careful thought as to which ones are necessary, and how much detail is needed.

Even after deciding that we want some cross-tabulation with categories of ‘question J’ as rows and ‘question K’ as columns, there are several other decisions to be made:

- The number in the cells of the table may be just the frequency i.e. the number of respondents who gave that combination of answers. This may be rephrased as a proportion or a percentage of the total. Alternatively, percentages can be scaled so they total 100% across each row or down each column, so as to make particular comparisons clearer.
- The contents of a cell can equally well be a statistic derived from one or more other questions e.g. the proportion of the respondents falling in that cell who were economically-active women. Often such a table has an associated frequency table to show how many responses went in to each cell. If the cell frequencies represent

small subsamples the results can vary wildly, just by chance, and should not be over-interpreted.

- Where interest focuses mainly on one ‘area’ of a two-way table it may be possible to combine rows and columns that we don’t need to separate out, e.g. ruling party supporters vs. supporters of all other parties. This simplifies interpretation and presentation, as well as reducing the impact of chance variations where there are very small cell counts.
- Frequently we don’t just want the cross-tabulation for ‘all respondents’. We may want to have the same table separately for each region of the country – described as segmentation – or for a particular group on whom we wish to focus such as ‘AIDS orphans’ – described as selection.
- Because of varying levels of success in covering a population, the response set may end up being very uneven in its coverage of the target population. Then simply combining over the respondents can mis-represent the intended population. It may be necessary to show the patterns in tables, sub-group by sub-group to convey the whole picture. An alternative, discussed in Part 1, is to weight up the results from the sub-groups to give a fair representation of the whole.

2.4 Tabulation & the Assessment of Accuracy

Tabulation is usually purely descriptive, with limited effort made to assess the ‘accuracy’ of the numbers tabulated. We caution that confidence intervals are sometimes very wide when survey samples have been disaggregated into various subgroups: if crucial decisions hang on a few numbers it may well be worth putting extra effort into assessing – and discussing – how reliable these are. If the uses intended for various tables are not very numerical or not very crucial, it is likely to cause unjustifiable delay and frustration to attempt to put formal measures of precision on the results.

Usually, the most important considerations in assessing the ‘quality’ or ‘value’ or ‘accuracy’ of results are *not* those relating to ‘statistical sampling variation’, but those which appraise the following factors and their effects:-

- evenness of coverage of the target (intended) population
- suitability of the sampling scheme reviewed in the light of field experience and findings
- sophistication and uniformity of response elicitation and accuracy of field recording
- efficacy of measures to prevent, compensate for, and understand non-response
- quality of data entry, cleaning and metadata recording
- selection of appropriate subgroups in analysis

If any of the above factors raises important concerns, it is necessary to think hard about the interpretation of ‘statistical’ measures of precision such as standard errors. A factor that has uneven effects will introduce biases, whose size and detectability ought to be dispassionately appraised and reported with the conclusions.

Inferential statistical procedures can be used to guide generalisations from the sample to the population, where a survey is not badly affected by any of the above. Inference addresses issues such as whether apparent patterns in the results have come about by chance or can reasonably be taken to reflect real features of the population. Basic ideas are reviewed in *Understanding Significance: the Basic Ideas of Inferential Statistics*. More advanced approaches are described in *Modern Methods of Analysis*.

Inference is particularly valuable, for instance, in determining the appropriate form of presentation of survey results. Consider an adoption study, which examined socio-economic factors affecting adoption of a new technology. Households are classified as male or female headed, and the level of education and access to credit of the head is recorded. At its most complicated the total number of households in the sample would be classified by adoption, gender of household head, level of education and access to credit resulting in a 4-way table.

Now suppose, from chi-square tests we find no evidence of any relationship between adoption and education or access to credit. In this case the results of the simple two-way table of adoption by gender of household head would probably be appropriate. If on the other hand, access to credit were the main criterion affecting the chance of adoption and if this association varied according to the gender of the household head, the simple two-way table of adoption by gender would no longer be appropriate and a three-way table would be necessary. Inferential procedures thus help in deciding whether presentation of results should be in terms of one-way, two-way or higher dimensional tables.

Chi-square tests are limited to examining association in two-way tables, so have to be used in a piecemeal fashion for more complicated situations like that above. A more general way to examine tabulated data is to use log-linear models described in *Modern Methods of Analysis.*

2.5 Multiple Response Data

Surveys often contain questions where respondents can choose a number of relevant responses, e.g.

If you are not using an improved fallow on any of your land, please tick from the list below, any reasons that apply to you:-

(i) Don't have any land of my own

(ii) Do not have any suitable crop for an improved fallow

(iii) Can not afford to buy the seed or plants

(iv) Do not have the time/labour

There are three ways of computerising these data. The simplest is to provide as many columns as there are alternatives. This is called a "multiple dichotomy”, because there is a yes/no (or 1/0) response in each case indicating that the respondent ticked/did not tick each item in the list.

The second way is to find the maximum number of ticks from anyone and then have this number of columns, entering the codes for ticked responses, one per column. This is known as “multiple response” data. This is a useful method if the question asks respondents to put the alternatives in order of importance, because the first column can give the most important reason, and so on.

A third method is to have a separate table for the data, with just 2 columns. The first identifies the person and the second gives their responses. There are as many rows of data as there are reasons. There is no entry for a person who gives no reasons. Thus, in this third method the length of the columns is equal to the number of responses rather than the number of respondents.

If there are "follow-up" questions about each reason, the third method above is the obvious way to organise the data, and readers may identify the general concept as being that of data at another "level", i.e. the reason level. More information on organising this type of data is provided in the guide *The Role of a Database Package for Research Projects*.

Essentially such data are analysed by building up counts of the numbers of mentions of each response. Apart from SPSS, few standard statistics packages have any special facilities for processing multiple response and multiple dichotomy data. Almost any package can be used with a little ingenuity, but working from first principles is a time-consuming business. On our web site we describe how Excel may be used.

2.6 Profiles

Usually the questions as put to respondents in a survey need to represent ‘atomic’ facets of an issue, expressed in concrete terms and simplified as much as possible, so that there is no ambiguity and so they will be consistently interpreted by respondents.

Basic cross-tabulations are based on reporting responses to such individual questions and are therefore narrowly issue-specific. A rather different approach is needed if the researchers' ambitions include taking an overall view of individual, or small groups', responses as to their livelihood, say. Cross-tabulations of individual questions are not a sensible approach to 'people-centred' or 'holistic' summary of results.

Usually, even when tackling issues a great deal less complicated than livelihoods, the more important research outputs are 'complex molecules' which bring together responses from numerous questions to produce higher-level conclusions described in more abstract terms. For example several questions may each enquire whether the respondent follows a particular recommendation, whereas the output may be concerned with overall 'compliance' – the abstract concept behind the questioning. A profile is a description synthesising responses to a range of questions, perhaps in terms of a set of abstract nouns like compliance. It may describe an individual, cluster of respondents or an entire population.

One approach to discussing a larger concept is to produce numerous cross-tabulations reflecting actual questions and to synthesise their information content verbally. This tends to lose sight of the 'profiling' element: if particular groups of respondents tend to reply to a range of questions in a similar way, this overall grouping will often come out only weakly. If you try to follow the group of individuals who appear together in one corner cell of the first cross-tab, you can't easily track whether they stay together in a cross-tab of other variables.

Another type of approach may be more constructive: to derive synthetic variables – indicators – which bring together inputs from a range of questions, say into a measure of 'compliance', and to analyse those, by cross-tabulation or other methods. See section 2.8 below. If we have an analysis dataset with a row for each respondent and a column for each question, the derivation of a synthetic variable just corresponds to adding an extra column to the dataset. This is then used in analysis just like any other column. A profile for an individual will often comprise a set of values of a suite of indicators.

2.7 Looking for Respondent Groups

Profiling is often concerned with acknowledging that respondents are not just a homogeneous mass, and distinguishing between different groups of respondents.

Cluster analysis is a data-driven statistical technique that can draw out – and thence characterise – groups of respondents whose response profiles are similar to one another. The response profiles may serve to differentiate one group from another if they are somewhat distinct. This might be needed if the aim were, say, to define target groups for distinct safety net interventions. The analysis could help clarify the distinguishing features of the groups, their sizes, their distinctness or otherwise, and so on. Unfortunately there is no guarantee that groupings derived from data alone will make good sense in terms of profiling respondents. Cluster analysis does not characterise the groupings; you have to study each cluster to see what they have in common. Nor does it prove that they constitute suitable target groups for meaningful development interventions

Cluster analysis is thus an exploratory technique, which may help to screen a large mass of data, and prompt more thoughtful analysis by raising questions such as:-

- Is there any sign that the respondents *do* fall into clear-cut sub-groups?
- How *many* groups do there seem to be, and how important are their separations?
- If there *are* distinct groups, what sorts of responses do "typical" group members give?

2.8 Indicators

Indicators are summary measures. Magazines provide many examples, e.g. an assessment of personal computers may give a score in numerical form like 7 out of 10 or a pictorial form of quality rating, e.g.

Very good	Good	Moderate	Poor	Very Poor
★	☆	○	◍	●

This review of computers may give scores – indicators – for each of several characteristics, where the maximum score for each characteristic reflects its importance e.g. for one model:- build quality (7/10), screen quality (8/20), processor speed (18/30), hard disk capacity (17/20) and software provided (10/20). The maximum score over all characteristics in the summary indicator is in this case (10 + 20 + 30 + 20 + 20) = 100, so the total score for each computer is a percentage e.g. above (7 + 8 + 18 + 17 + 10) = 60%. The popularity of such summaries demonstrates that readers find them accessible, convenient and to a degree useful. This is either because there is little time to absorb detailed information, or because the indicators provide a baseline from which to weigh up the finer points.

Many disciplines of course are awash with suggested indicators from simple averages to housing quality measures, social capital assessment tools, or quality-adjusted years of life. Of course new indicators should be developed only if others do nor exist or are unsatisfactory. Well-understood, well-validated indicators, relevant to the situation in hand are quicker and more cost-effective to use. Defining an economical set of meaningful indicators before data collection ought ideally to imply that at analysis, their calculation follows a pre-defined path, and the values are readily interpreted and used.

Is it legitimate to create new indicators after data collection and during analysis? This is to be expected in genuine ‘research’ where fieldwork approaches allow new ideas to come forward e.g. if new lines of questioning have been used, or if survey findings take the researchers into areas not well covered by existing indicators. A study relatively early on in a research cycle, e.g. a baseline survey, can fall into this category. Usually this means the available time and data are not quite what one would desire in order to ensure well-understood, well-validated indicators emerge in final form from the analysis.

Since the problem does arise, how does the analyst best face up to it?

It is important not to create unnecessary confusion. An indicator should synthesise information and serve to represent a reasonable measure of some issue or concept. The concept should have an agreed name so that users can discuss it meaningfully e.g. ‘compliance’ or ‘vulnerability to flooding’. A specific meaning is attached to the name, so it is important to realise that the jargon thus created needs careful explanation to ‘outsiders’. Consultation or brainstorming leading to a consensus is often desirable when new indicators are created. Indicators created ‘on the fly’ by analysts as the work is rushed to a conclusion are prone to suffer from their hasty introduction, then to lead to misinterpretation, often over-interpretation, by enthusiast would-be users. It is all too easy for a little information about a small part of the issue to be taken as ‘the’ answer to ‘the problem’!

As far as possible, creating indicators during analysis should follow the same lines as when the process is done *a priori* i.e. (i) deciding on the facets which need to be included to give a good feel for the concept, (ii) tying these to the questions or observations needed to measure these facets, (iii) ensuring balanced coverage, so that the right input comes from each facet, (iv) working out how to combine the information gathered into a synthesis which everyone agrees is sensible. These are all parts of ensuring face (or content) validity as in the next section. Usually this should be done in a simple enough way that the user community are all comfortable with the definitions of what is measured.

There is some advantage in creating indicators when datasets are already available. You can look at how well the indicators serve to describe the relevant issues and groups, and select the most effective ones. Some analysts rely too much on data reduction techniques such as factor analysis or cluster analysis as a substitute for thinking hard about the issues. We argue that an intellectual process of indicator development should build on, or dispense with, more data-driven approaches.

Principal component analysis is data-driven, but readily provides weighted averages. These should be seen as no more than a foundation for useful forms of indicator.

2.9 Validity

The basic question behind the concept of validity is whether an indicator measures what we say or believe it does. This may be quite a basic question if the subject matter of the indicator is visible and readily understood, but the practicalities can be more complex in mundane, but sensitive, areas such as measurement of household income. Where we consider issues such as the value attached to indigenous knowledge the question can become very complex. Numerous variations on the validity theme are discussed extensively in social science research methodology literature.

Validity takes us into issues of what different people understand words to mean, during the development of the indicator and its use. It is good practice to try a variety of approaches with a wide range of relevant people, and carefully compare the interpretations, behaviours and attitudes revealed, to make sure there are no major discrepancies of understanding. The processes of comparison and reflection, then the redevelopment of definitions, approaches and research instruments, may all be encompassed in what is sometimes called triangulation – using the results of different approaches to synthesise robust, clear, and easily interpreted results. Survey instrument or indicator validity is a discussion topic, not a statistical measure, but two themes with which statistical survey analysts regularly need to engage are the following.

Content (or face) validity looks at the extent to which the questions in a survey, and the weights the results are given in a set of indicators, serve to cover in a balanced way the important facets of the notion the indicator is supposed to represent.

Criterion validity can look at how the observed values of the indicator tie up with something readily measurable that they should relate to. Its aim is to validate a new indicator by reference to something better established, e.g. to validate a prediction retrospectively against the actual outcome. If we measure an indicator of ‘intention to participate’ or ‘likelihood of participating’ beforehand, then for the same individuals later ascertain whether they did participate, we can check the accuracy of the stated intentions, and hence the degree of reliance that can in future be placed on the indicator.

As a statistical exercise, criterion validation has to be done through sensible analyses of good-quality data. If the reason for developing the indicator is that there is no satisfactory way of establishing a criterion measure, criterion validity is not a sensible approach.

2.10 Summary

In this guide we have outlined general features of survey analysis that have wide application to data collected from many sources and with a range of different objectives. Many readers of this guide should be able to use its suggestions unaided. We have pointed out ideas and methods which do not in any way depend on the analyst knowing modern or complicated statistical methods, or having access to specialised or expensive computing resources.

The emphasis has been on the importance of preparing the appropriate tables to summarise the information. This is not to belittle the importance of graphical display, but that is at the presentation stage, and the tables provide the information for the graphs. Often key tables will be in the text, with larger, less important tables in Appendices.

Often a pilot study will have indicated the most important tables to be produced initially. What then takes time is to decide on exactly the right tables. There are three main issues. The first is to decide on what is to be tabulated, and we have considered tables involving either individual questions or indicators. The second is the complexity of table that is required – one-way, two-way or higher. The final issue is the numbers that will be presented. Often they will be percentages, but deciding on the most informative base, i.e. what is 100% is also important.

2.11 Next Steps

We have mentioned the role of more sophisticated methods. Cluster analysis may be useful to indicate groups of respondents and principal components to identify data-driven indicators. Examples of both methods are in our *Modern Methods of Analysis* guide where we emphasise, as here, that their role is usually exploratory. When used, they should normally be at the start of the analysis, and are primarily to assist the researcher, rather than as presentations for the reader.

Inferential methods are also described in the *Modern Methods* guide. For surveys, they cannot be as simple as in most courses on statistics, because the data are usually at multiple levels and with unequal numbers at each subdivision of the data. The most important methods are log-linear and logistic models and the newer multilevel modelling. These methods can support the analysts' decisions on the complexity of tables to produce.

Both the more complex methods and those in this guide are equally applicable to cross-sectional surveys, such as baseline studies, and longitudinal surveys. The latter are often needed for impact assessment. Details of the design and analysis of baseline surveys and those specifically for impact assessment must await another guide!

The Statistical Services Centre is attached to the Department of Applied Statistics at The University of Reading, UK, and undertakes training and consultancy work on a non-profit-making basis for clients outside the University.

These statistical guides were originally written as part of a contract with DFID to give guidance to research and support staff working on DFID Natural Resources projects.

The available titles are listed below.

- *Statistical Guidelines for Natural Resources Projects*
- *On-Farm Trials – Some Biometric Guidelines*
- *Data Management Guidelines for Experimental Projects*
- *Guidelines for Planning Effective Surveys*
- *Project Data Archiving – Lessons from a Case Study*
- *Informative Presentation of Tables, Graphs and Statistics*
- *Concepts Underlying the Design of Experiments*
- *One Animal per Farm?*
- *Disciplined Use of Spreadsheets for Data Entry*
- *The Role of a Database Package for Research Projects*
- *Excel for Statistics: Tips and Warnings*
- *The Statistical Background to ANOVA*
- *Moving on from MSTAT (to Genstat)*
- *Some Basic Ideas of Sampling*
- *Modern Methods of Analysis*
- *Confidence & Significance: Key Concepts of Inferential Statistics*
- *Modern Approaches to the Analysis of Experimental Data*
- *Approaches to the Analysis of Survey Data*
- *Mixed Models and Multilevel Data Structures in Agriculture*

The guides are available in both printed and computer-readable form. For copies or for further information about the SSC, please use the contact details given below.

Statistical Services Centre, University of Reading
P.O. Box 240, Reading, RG6 6FN United Kingdom

tel: SSC Administration +44 118 378 8025
fax: +44 118 378 8458
e-mail: statistics@lists.reading.ac.uk
web: http://www.reading.ac.uk/ssc/