End of Award Report:

Practical Exemplars of the Analysis of Surveys

Background

The ESRC, through its support of the Economic and Social Data Service (ESDS) makes large amounts of data from UK surveys available for secondary analysis by academic researchers. Much of this data is collected by government to inform policy developments and researchers in government are also heavy users of these data sets. Most of these secondary analysts have relatively little background in survey design and analysis.

Survey organisations (including ONS) use sophisticated methods to design surveys and generally provide considerable detail on how this has been done. Other procedures such as post-stratification to match population totals are commonly carried out before the data are made available to secondary analysts. Appropriate analyses need to make use of information on the survey design and on the post-fieldwork procedures.

Methods of making inferences from survey data fall broadly into two main classes, those that are design-based and those that are model-based (Binder and Roberts, 2003). Design-based methods make inferences for population quantities defined in terms of population totals, ratios or other quantities. These methods are by no means new and much of the theory on which they were based was developed in the 1940s 50s and 60s (Cochrane, 1977) although methodological developments continue until today. They make no assumptions about the distribution the data, such as assuming common variances within subgroups.

Model-based methods, on the other hand, make assumptions about the model that has generated the data. To use a model-based method appropriately for a survey with a complex design the model needs to include all the features that are part of the design. So, for example, the model should include the stratification factors and make allowance for clustering of cases if this is part of the design. Even when this is attempted the model will usually involve some further assumptions. Survey statisticians have provided ample evidence of the robustness of their methods compared to model-based methods (e.g. Holt Smith and Winter, 1980).

Despite their obvious advantages for analysing complex surveys, design-based methods have been used very little by academic researchers in the UK, with a few very notable exceptions. Academics are more likely to be interested in model-based conclusions than in the estimates of finite population quantities so design-based methods may seem inappropriate. But unless model-based methods include all the design features they may give the wrong results. Design-based methods can be used to make inferences to super-populations that may have generated the data without any adjustments to their results (see Sarndal, Swenson and Wretman, 1992, Chapter 13, for a discussion of this).

Survey design features are often used for the convenience and economy of planning the survey rather than for any intrinsic interest in what they represent. This is very different from medical studies such as clinical trials or epidemiological studies where design features are generally of substantive interest. However, it is interesting that a recent paper on covariance adjustment in clinical trials and observational studies (Rosenbaum,

2002) has recently proposed a methodology that is essentially based on the same principles as design-based survey inference.

In recent years falling survey response rates have been a concern for all survey organisations both in the UK (Crockett, 2004) and elsewhere (Atrostic et al, 2001). This has led to the development of methods that correct for non-response. Within survey organisations the concern has chiefly been with unit non-response as households have become more difficult to contact and response rates for postal surveys have fallen. Within survey organisations this is most frequently addressed by post-stratifying the survey responses to match them with known population proportions on one or more sets of categories. New survey weights are then provided to incorporate this adjustment. This process also has the potential to improve precision.

Surveys that are post-stratified may make pure design-based inference more difficult. The approach loses its attraction of being free of modelling assumptions. On the practical side post-stratification gives problems if a survey is clustered, as are most government sponsored household surveys, because post-strata will cut across cluster boundaries and thus the survey cannot be analysed as though it was a simple stratified survey.

This difficulty with post-stratification was one of the motivations for the use of replication methods such as jackknives and bootstraps in survey analysis. These methods are implemented by providing a set of replicate weights that can be used to run the analysis many times. The difference between the replicates is then used to estimate the variability of the estimates. Each set of weights can be post-stratified. With this system it is claimed that the secondary analyst will not need to know anything about the survey design, but simply needs to use the replicate weights. It has also been claimed that using replicate weights will get over disclosure problems where individual primary sampling units might be identified (Yeo et al, 1999), though it is not clear how this would be any better than giving clusters non-disclosive identifiers.

More recently there have been further developments in survey analysis that come under the broad heading of 'model-assisted survey sampling'. A number of different variants of these methods exist, but they share the characteristics of developing from design-based inference but relaxing some of the formal requirements, e.g. that the analyst should condition only on quantities that are fixed by the design. Their goal is to make inferences to the super-populations from which the population may be considered one realisation. The general term G-calibration has been applied to these methods and a review of them has been carried out for Statistics Belgium (Vanderhoeft , 2001)). In a comparison to replication methods for post-stratified data these methods performed very well compared to replication methods (Valliant, 1993). Once a program is available to do the computations they are less trouble than replication methods which can take up a lot of computing time and space.

Finally, recent years have seen considerable development of imputation methods for handling both unit and item non-response. These come in a variety of forms including both empirical and model-based procedures (Rubin, 1987). The use of model-based procedures has been encouraged by software implementations that have become available in recent years (e.g. Schafer, 1997).

This background overview is intended to outline the theoretical work on which the practical applications exemplified on the PEAS web site are based. Our focus in evaluating the software will not be on the theoretical properties of the algorithms that are implemented, but on the equally important matter of its usability.

Objectives

In our application we stated four aims.

1. *To provide a web-based resource that will contain practical examples of using recent methods for the analysis of social surveys. This resource will be developed by the two main applicants in collaboration with practitioners who analyse surveys from the academic community, central and local government.*
2. *To hold three workshops where material to be presented at the workshops will be developed along with users. The first of these will be on methods for complex surveys, the second on imputation methods for cross-sectional data and the third on imputation methods for cross sectional data*
3. *To provide resources for survey analysis, specific to the example surveys used as exemplars, that would be made available to researchers. Examples would be sets of bootstrap weights and data sets with missing values imputed*
4. *To disseminate new methods of survey analysis to UK researchers and to empower them to make contacts with the international community who develop tools for survey analysis.*

1. The web-based resource is now fully developed and is accessible at http://www.napier.ac.uk/depts/fhls/peas/. There are a total of 19 main web pages, each of which links to a large number of further resources. These consist of 9 pages on survey design and theory, 6 web pages on exemplars, 4 on software packages as well as individual pages on links and resources, FAQs and various index pages. Each of the 6 exemplar pages allows the user to download data in a choice of 4 different formats, to access code to run analyses and to view both code and output from each of 4 software packages, with comments, in a browser window. There are also a large number of other secondary pages linked from the main ones. There are extensive links between the theory sections and the exemplars.
2. We held three workshops, as we had planned, but the topics were not as we had originally envisaged. All three workshops were devoted to methods for analysing data from complex surveys and tested out the material for exemplars 1 to 4. The reason for this is explained in detail below. But, briefly, we found that there was an even greater need for researchers (including us) to understand fully how best to use this software. Also two new software packages for complex surveys (R survey and SPSS complex surveys) became available after we wrote our original proposal. The workshops participants were drawn from the groups we stated in our aims. A full report on the feedback from the workshops is in the evaluation report submitted with this report.
3. We envisaged having to make these resources like those mentioned in aim 3 available via the ESRC archive. Examples would be 1) new post-stratified weights 2) imputed values 3) replication weights. In putting data on the web we had to

anonymise it so it would no longer be exactly the same as that on the archive and thus outputs that would not be quite right would be of little interest to researchers. The methods we have used for most of these analyses are now available on standard packages so it will be easy for any researcher to repeat them for any survey using one of the exemplars as a model. One possible exception is the data that have been imputed from exemplar 6 (the Edinburgh Study of Youth Transitions and Crime). We are continuing to finalise the work on this exemplar, in collaboration with the ESYTC team and we anticipate submitting the final imputed longitudinal data set to the archive, but this will need to be subject to their agreement and to our deciding on the final best imputation model to present there.

4. Our success in meeting this objective is more difficult to assess, and in retrospect may have been rather too ambitious within our time scale. We hope that it will develop as more people use the site. In developing the PEAS web site we ourselves have had contact with researchers developing survey methodology and survey software, mainly in the USA and in Germany.

Our target users for the PEAS web site are secondary analysts, not methodologists. Thus we aim to present practical and usable methods that could be used safely by anyone, including those without detailed knowledge of the underlying statistical theory. Using the analogy of driving a modern car, we would aim to help people to be good and careful drivers who would know what their vehicle was capable of, how to use the controls and how to respond to warning signals. Out target audience would not be the car designers (methodologists in the analogy) or even the mechanics (software authors), although we might expect that our practical conclusions would be relevant to both of these, especially the latter.

Methods

In this section we will describe the process by which the material that is now on the PEAS web site was developed.

In the first months of the project Susan Purdon and Gillian Raab started work on developing the material to go on the site. This involved exchanges of ideas about how a survey organization tackled survey design, compared with approaches in an academic environment. The capabilities of current statistical software for surveys was explored and experiences compared and possible structures for the site were discussed.

With the appointment of Iona Waterston as web-designer these initial ideas had to be transformed into a web resource that would help others to learn about survey design and analysis. An overall design structure was mapped out, in collaboration with Kathy Buckner. There were to be two main sections of the site that, for short, we termed ‘theory’ and ‘exemplars’, although the section on theory also covered the practical side of survey design and analysis. Susan Purdon took primary responsibility for much of the theory section of the site, ensuring it would be well grounded in her practical experience of designing and analysing large social surveys. Gillian Raab took the leading role in preparing the exemplars and testing out the software, with assistance from staff at the National Centre as well as considerable help, at later stages, from workshop participants and others. A software section was added to the site and links between different sections were established. Figure 1 illustrates the basic structure.

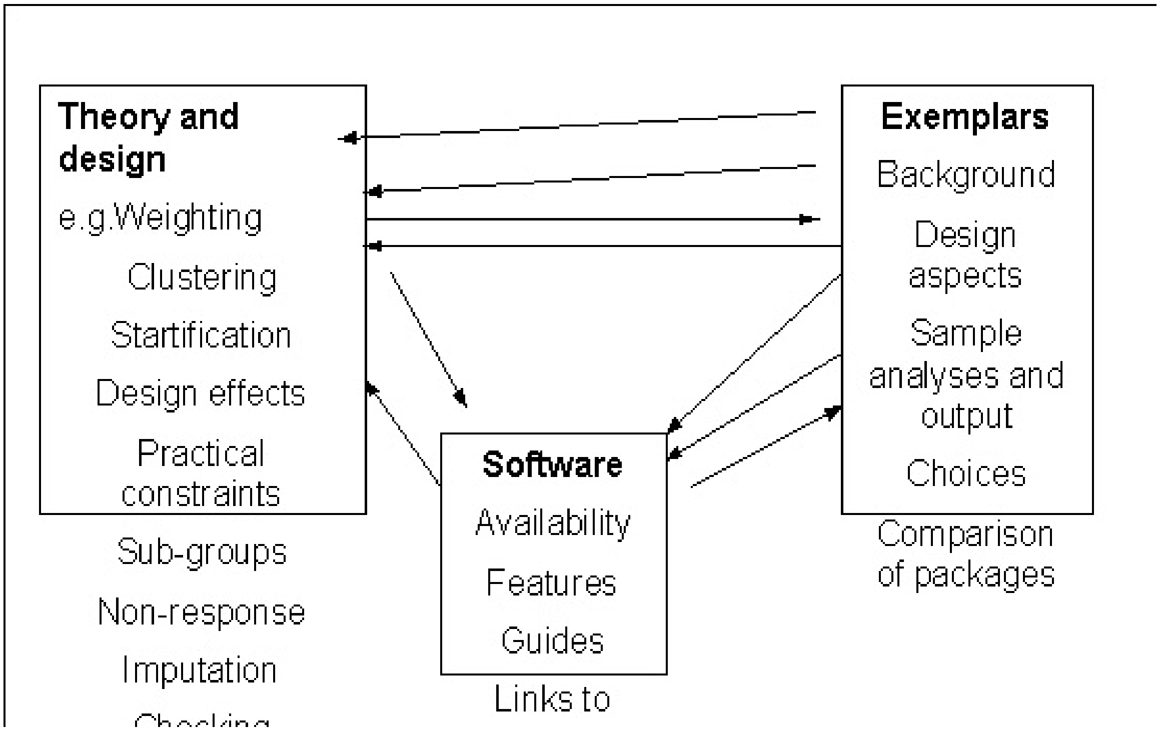


Figure 1: Illustration of structure of PEAS web site

Iona Waterston was responsible for designing the look of the pages and translated our material into web pages that were initially checked by the research team. There were some false starts in terms of the material provided to go on the web, particularly on the presentation of exemplars. The initial exemplars were too complicated and were difficult to follow. They had to be simplified and rewritten. It was decided that each one should only attempt to show one or two analyses and be restricted to a relatively small number of variables.

At the same time the capabilities of software for analysing surveys was investigated. We had originally thought that we might have to have recourse to specialist software (such as WesVar for replication methods) in order to cover the full range of methods of analysis that might be required. But with the development of survey methods for standard packages there was no need to do go further than four packages all of which are main-stream statistical analysis software (SPSS, SAS, R and STATA).

By the end of the summer we had enough material to start running our series of workshops. These took place in September, October and November. They consisted of introductory sessions (the slides from these can be found on the PEAS web site) followed by practical sessions and by a variety of feedback activities. The first two workshops were by invitation and, in particular, were attended by people associated with some of the surveys on which the exemplars were based, as well as some other local researchers we knew to be keen to learn about survey methods. The final workshop was advertised round the ScotStat (http://www.scotland.gov.uk/stats/scotstats.asp) mailing list for users of Scottish statistics and was over-subscribed. It attracted a wider range of users with, on average, less experience of survey analysis than had been the case for the other two workshops. Further detail of the workshops can be found in the evaluation report. The workshops evaluated general features of the web site and were based on the first four exemplars, all of which were based on government sponsored surveys available via ESDS.

Negotiations with data providers to make the data from the surveys available on the web site were on-going throughout this time. Some of the surveys requested that we carry out more extensive data anonymisation procedures than we had originally planned and this entailed re-doing some of the analyses on the anonymised data. In conjunction with the data providers we arrived at a set of principles that would guide us in making data available (Appendix 1).

After the last workshop we had the task of finalising the web site. A fifth exemplar on re-weighting for non-response was developed with few problems. This was based on a postal survey and provided an exemplar with a simpler deign, typical of this type of survey. We had not, as yet, tackled imputation procedures, which were part of our original specification. There was less interest in carrying out imputation from researchers than we had originally envisaged and it did not prove possible to use the data set we had originally planned to use for this (SHARE longitudinal data) because there were no plans to make it available via the data archive. Fortunately, we were contacted by researchers from the Edinburgh Study of Youth Transitions and Crime (Supported by ESRC Research grants R000237157 and R000239150) who wanted help with dealing with missing data problems in this longitudinal data set. This enabled the final exemplar to be developed, in collaboration with the research team in time for it to be placed on the web site.

Work on imputation was facilitated by Gillian Raab's attendance at a workshop organised by another Research Methods project (James Carpenter and Mike Kenward, 'Missing data in multilevel models' award reference H333 25 0047). The types of analysis being discussed here were very different from those we have been working on for the PEAS site, being wholly model-based rather than design-based. But some use of modelling techniques, even implicitly by making assumptions about conditional distributions, is inevitable when data are missing. The imputation exemplar used simpler methods than they have proposed and is limited to those that are part of packages used for survey analysis.

Three things held up the final release of the web site after the funding from the ESRC stopped (March 05), described under 'difficulties' above. Identifying the problems with results from the imputation exemplar (6) was the biggest task. Some of the methods involved have been criticised on theoretical grounds on Carpenter and Kenward's web site (http://www.missingdata.org.uk) and by a recent draft review (Durant, 2005). Our difficulties were of a different type and were mainly, but not entirely, caused by the size of the imputation problems we were attempting. This is an important message for survey analysts who may attempt to use these model-based methods on large data sets. The PEAS site now contains examples of how imputed data can be examined to check if there is any evidence of problems and suggestions for where problems can arise. But we still some residual worries that undetected problems may still be lurking in what has been done. Empirical techniques of the 'hot-deck' variety that are carried out close to the original data collection may be more robust.

In September 05 all the permissions to place data on the web had finally been received, the imputation exemplar was complete and most of the software upgrades that had taken place during the life of the project had been incorporated. The site has been advertised over various mailing lists and links with other sites (e.g. SOSIG) are established.

Results

We obtained two kinds of results in developing the PEAS material. We learned about the capabilities of the different statistical software packages to carry out survey analysis procedures. The analyses of the exemplars also provided some interesting results about the relative merits of different designs, or different weighting or imputation procedures for the surveys we analyzed. Space does not allow us to include this here, but details are in the exemplar pages of the PEAS site.

Survey analysis software

A package for analysing surveys with a complex design needs to do the following:-

- Allow the user to specify a variety of designs
- Check that the data and the design agree and take action if not
- Implement different methods of survey analysis document each method fully
- Carry out a variety of analyses with adjustment for survey design
- Present the results of analyses

Software included

We have used the four packages R (survey package), SAS , SPSS (Complex surveys module) and STATA. The most recent versions of the packages are currently R survey package 3.3, SAS version 9, SPSS version 13 with complex surveys and STATA version 9. The comments are a result of using the latest versions of all of these packages, except for STATA, where only version 8 has been used but published information on the new version has been incorporated into the comments. It is planned to update the site once the new version of STATA has been tested.

What designs can be accommodated?

All four programmes allow the specification of the standard features that are part of most survey designs, e.g. weighting, clustering and stratification. It is possible, in all cases to specify the design in terms of just a single level of clustering and then make use of methods using the variation between cluster totals (Hansen, Hurwitz and Madow, 1953) to compute standard errors. In SPSS it is necessary to select the 'with replacement' option in this case and finite population corrections cannot be used . Three of the four packages (not SAS) have capabilities to allow multi-stage designs to be specified, but we do not think they would ever be needed by secondary survey analysts, so these methods are not featured on the PEAS site (see section 2.1 of the section of the PEAS site on clustering for a discussion of this).

Post-stratification and raking to match survey totals can be handled by R and by STATA. The SAS macro CALMAR available from INSEE (French National Statistics Agency) can also be used and is the most comprehensive implementation of this method. Its documentation is only in French and the version on the web site is not the most recent. More recent versions allow post-stratification at different levels (e.g individual and household) to be carried out simultaneously. We illustrate the earlier version with Exemplar 1. Olivier Sautory, the author of Calmar, has informed us that the more up-to-date version with documentation in English as described in Sautory (2003) will be made available in autumn 2005, and the site will be updated with this information if and when this happens.

How well do the packages check the adequacy of the design?

This is the aspect that varies most between the packages. The sort of problem that can happen is that a survey that is supposed to be clustered within strata may have clusters with members that are in more than one stratum. Strata with a single primary sampling unit (PSU) can also cause problems. At one extreme SPSS complex surveys will produce an answer even when the design is quite wrong, though with a caution that the design has not been checked. SAS and STATA produce mixed results, with STATA (**at least in version 8 – SUSAN can you check 9 if you have it**) refusing to provide any output when any stratum has a single PSU, even as a result of sub-setting the survey. The best choice comes from R survey which gives a warning and a list of options as to how to deal with this. If only a few cases are affected then one of the approximate solutions can be chosen. If many units are affected then the survey may have been wrongly described.

What methods are implemented?

The latest versions of all four packages now include methods of survey inference by Taylor linearization along with adjusted chi-squared statistics for tests of association in tables from complex surveys.

STATA and R include replication methods (e.g. jackknife methods) and procedures for producing replicate weights. R has the widest range and the most efficient way of storing the weights. Some SAS macros for replication methods have been published, but these are clumsy to use. All of the methods we have tested gave results that agreed well with Taylor series methods when both could be used.

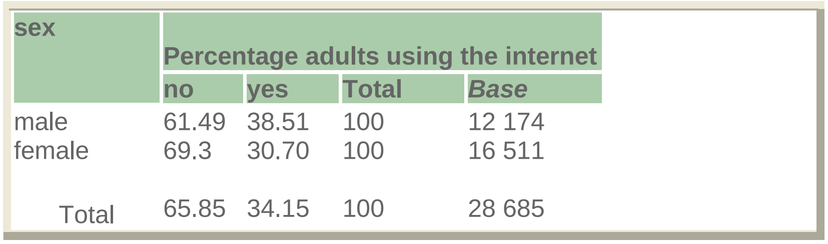
The main justification for using replication methods is to make proper allowance for post-stratification in clustered samples. But the alternative of using calibration methods would be a more up-to-date solution to this problem. Only R implements this at present

What analyses can be carried out?

All the packages can now handle basic descriptive statistics, adjusted chi-squared tests, linear and logistic regression. R and STATA each have many additional procedures such as ordered logistic regression and survival analysis. Their range of procedures is similar and not entirely overlapping.

Presentation of results

All the packages produce standard errors and confidence intervals and all, except SAS, can also print design effects and/or design factors. The presentation of tables in the standard form for survey reports, with the base numbers shown in a separate column, as illustrated in Table1 taken from exemplar 2, can only be produced automatically in SPSS and with the SAS version 9 SURVEYFREQ procedure.



sex	Percentage adults using the internet			
	no	yes	Total	*Base*
male	61.49	38.51	100	12 174
female	69.3	30.70	100	16 511
Total	65.85	34.15	100	28 685

Table 1: Sample table

Ease of use of the survey packages

Familiarity with the original packages is the most important aspect in how user-friendly the survey methods software appears. For experienced users none of the software provides any problems. SPSS and STATA are both easy to learn and widely used by UK researchers. SAS is slightly more difficult but widely used in government, business and industry. The R package, which has the most comprehensive range of survey procedures in terms of coverage and options, is much less used and largely confined to the academic community and some environmental research organisations. It is unlikely that, with its command-driven interface and its complex data structures, it would ever become a major tool in survey research organisations. But it deserves to be used more widely as a tool for some secondary analysts and for specialists in survey organisations to evaluate potential designs.

Software for imputation

The packages R, SAS and STATA provide software that can be used for imputation. Space does not permit a detailed description of this here but the imputation section of the web site provides all the details (Imputation section 6.6) http://www.napier.ac.uk/depts/fhls/peas/imputation.asp#for

The difficulties we encountered with this software were of two kinds. In the first case there were problems with it not running correctly. Many of these were ironed out with the software providers or are still being fixed. The others were the more difficult problem that the results were apparently not correct. We have included extensive notes on this on the web site.

There also appear to be problems with the maintenance of these routines on R and the porting of them to the latest version. Again, we hope that these may be sorted out soon and we are in touch with potential providers to try to make this happen.

Activities

The main activities we carried out were the workshops that are mentioned above and described in detail in the evaluation report. Both Gillian Raab and Susan Purdon have given presentations about survey research methods, incorporating information about PEAS at several formal and informal occasions

Outputs

The main output is the PEAS website described above, freely available over the web.

Impacts

Gillian Raab was invited to take part in the review teams of two major Scottish surveys (Scottish Household survey and Scottish Health Survey). The work carried out on these surveys (exemplars 2 and 3) for PEAS enabled her to provide input to these reviews and to discuss design issues with other members of the review groups.

Future Research Priorities

We have identified a need for more investigation of the best imputation methods to be used for complex surveys.

References

Atrostic K, Bates N, Burt G, and Silberstein A (2001) Nonresponse in U.S. Government Household Surveys: Consistent Measures, Recent Trends, and New Insights *Journal of Official Statistics*, Vol.17, No.2, 2001. pp. 209-226

Binder DA and Roberts GR Design-based and model-based methods for estimating model parameters in Chambers RL and Skinner CJ (2002) *Analysis of survey data*, Wiley, Chichester

Cochrane W. G. (1977, first published 1953) *Sampling Techniques***,** Wiley, New York.

Crockett A, (2004) Weighting the Social Surveys, *ESDS Government*, updated 2005 Afkhami R, (http://www.esds.ac.uk/government/resources/statguides.asp accessed 9/05)

Durrant GB (2005) *Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review ,*ESRC National Centre for Research Methods and Southampton

National Centre for Research Methods Working Paper Series, can be accessed from http://www.napier.ac.uk/depts/fhls/peas/pdfs/durrantreview.pdf.

Holt D, Smith TMF, Winter PD (1980) Regression analysis of data from complex surveys, *J R Statist Soc, Series A*, 143, 474-87.

Rosenbaum, P. R. (2002). Covariance adjustment in randomized experiments and observational studies (with discussion). *Statistical Science* 17, 286**–**327.

Sarndal CE, Swenson B and Wretman J (1992) *Model assisted survey sampling*, Springer,New York

Sautory **,** O (2003) *CALMAR 2: A new version of the CALMAR calibration and adjustment program*. Proceedings of Statistics Canada's Symposium 2003

Challenges in Survey Taking for the Next Decade (http://www.statcan.ca/english/freepub/11-522-XIE/2003001/session13/sautory.pdf accessed June 05).

Schafer J *Analysis of Incomplete multivariate data***.** CRC press, 1997.

Valliant R (1993) Post-stratification and conditional variance estimation. JASA 88: 89-96

Vanderhoeft C (2001) Generalized Calibration at Statistics Belgium. Statistics Belgium Working Paper No 3.(http://www.statbel.fgov.be/studies/paper03_en.asp accessed 9/05)

Yeo, D., Mantel, H. and Liu, TP.(1999). *Bootstrap Variance Estimation for the NationalPopulation Health Survey*. 1999 Proceedings of the Survey Research Methods Section,American Statistical Association, pp. 778-783.

Appendix 1

Precautions taken to anonymise data from the surveys

>> Why are we concerned?
>> Where does the data on the P|E|A|S site originate and where have we obtained them from?
>> Who has access to the survey data and what aspects may be restricted ?
>> What special conditions apply to the P|E|A|S data ?
>> How have we taken steps to make the data on this site anonymous?

Why do we need to be concerned with anonymity?

Survey subjects, in almost every case, are informed that the data they provide will only be used for 'statistical purposes' and that nothing that is published from the survey that will make it possible to identify individuals. Clearly, we must ensure that we never breach this promise in making real survey data available over the world wide web. This principle is enshrined in the Statement of Principles for the Office of National Statistics Code of Practice

http://www.statistics.gov.uk/about/national_statistics/cop/downloads/StatementRD.pdf

> " The National Statistician will set standards for protecting confidentiality, including a guarantee that no statistics will be produced that are likely to identify an individual unless specifically agreed with them."

Data may disclose information about individuals or organisations, even when it is not indexed by a name or another identifier that could be used directly to trace the respondent. Disclosure can happen because an individual has a unique combination of characteristics. An individual responding to a survey might be the only doctor in a village. If the survey identifies the village as a geographic area and has details of occupations, then anyone with access to the anonymised survey data could identify this person's responses to other questions.

Where does the survey data on the P|E|A|S site originate and where have we obtained it from?

The survey data on this site has come from three sources. Most commonly it is from large surveys that have been commissioned by government departments and carried out either by the Office of National Statistics or by Survey organisations. We also include a survey carried out by an academic group and sponsored by the ESRC and a survey carried out by a health board and funded by the NHS. In all but the last case we have obtained the data from the ESRC UK Data Archive . This service gives bona-fide researchers access to a wide range of data, including government surveys that are made available via the Economic and Social Data Service.

Who has access to the survey data and what aspects may be restricted?

The analysts working for the organisations who have commissioned the survey generally have access to all the data collected. This would exclude individual names and addresses which it would be good practice to hold separately from the survey data.

Secondary analysts need to register with the Data Archive (see above) to obtain copies of the data and sign a guarantee of confidentiality. They must register the use they plan to make of the data in order to get access to specific resources. In some cases the more confidential aspects of data (e.g. geographic identifiers) may only be obtained with special permission from the organisation which has deposited the resource with the Archive. In other cases such confidential information may not be made available via the Archive at all.

Finally tabular data from the surveys appears in published reports and in internal documents that are available to the general public or to policy makers. In all of these cases the principle above needs to be adhered to by one or more of the following measures

- ensuring the data, either by itself or in conjunction with other sources, could never disclose an individual

- by making minor changes to the data (e.g. counts in tables) to prevent individuals being identified

- obtaining assurances from those with access to the data that no disclosure of individual information will take place

What special conditions apply to the P|E|A|S data ?

We want to make it easy for people learning about surveys to use real data, with all the problems this involves. It would have detracted from the usefulness of the P|E|A|S web site if it was necessary to apply to the Data Archive for permissions before trying out the methods. What is more, analysts require information on the survey deign to use methods for complex surveys. This may include variables that identify the PSU of a respondent for a clustered sample and/or the stratum for a stratified sample. These data are not always available for surveys deposited in the Data Archive, although they may sometimes be derived from the serial number of the case. Sometimes they are only available in restricted data sets.

We have made data sets of individual records available on the P|E|A|S with a subset of survey variables. These data can be accessed by anyone who holds the relevant software and finds the web site. Therefore the data (taken by itself) must comply with the same standards of confidentiality that would apply to tables that might be released in published reports. But other considerations apply because of the possibility of linking with data on the Data Archive. We can identify three different types of people who might access the data on the web page:-

1. People with no access to data from the Archive and thus have given no assurance of confidentiality
2. People with access to the data sets from the survey Archive who have obtained copies of the full data sets for those surveys from which we have extracted the data, but **without access to any special resources** for which depositor's permission is required.
3. People with access to the data sets as above and who have **obtained permission from the depositors** to access special resources.

For individuals of all three types we need to ensure that the data by itself is not disclosive of individual information. For those of types 2 and 3 we need to ensure that the combination of the data on the web site along with the data from the Archive is not more disclosive than the data from the Archive by itself. This might happen, for example, if the data on the web site provided data on the primary sampling units and this, along with the data on the Archive might enable the identification of individuals and access to other data about them.

How have we taken steps to make the data on this site anonymous?

There are various ways in which the data taken by itself can be disclosive:

1. If a small cell arises which contains someone who is unique in the population (e.g. the village doctor as described above).
2. If an individual is unique in the sample and could be identified by their individual characteristics
3. If a survey contains a small unit (cluster or stratum) with only a few respondents can be identified and this unit has a substantial sampling fraction.
4. As condition 3 but where the unit or stratum has a small sampling fraction.
5. Weights and sampling fractions can reveal the identity of clusters, as has been pointed out by de Waal and Willenborg, if the numbers of units in the population is known.

Items 1 and 3 above refer to uniqueness in the population, while 2 and 4 refer to uniqueness in the sample. Sample uniqueness will only be disclosive if it is known that a respondent took part in the survey.

We have used the following methods where they were required to prevent this type of disclosure:

- adding random noise to the data as a Poisson variable for counts or a normal random variable for continuous data
- providing some continuous variables as ranges only
- ensuring that categorical variables have no rare classes
- limiting the number of variables provided in the data sets to those required for our analyses (usually very few)
- removing any identifying information, such as the original numbering, that might identify small clusters or strata
- adding random noise to weights and sampling fractions where this might lead to their identification

For surveys where we are providing additional information on the web site that is either held as restricted files in the Data Archive, or is not available from the Data Archive at all, we have carried out the following steps to prevent the web data being merged with the files from the Archive:-

- changed the serial numbers that identify individual cases
- added random noise to continuous variables and to the weights, where they take a large number of unique values

For each exemplar we have checked to make sure that these procedures have worked, paying special attention to any population unique cases. We have also made sure that none of these procedures distorts the conclusions from the analysis of the exemplars compared to what would have been obtained from the original data.