Defining parameters of interest in longitudinal studies and some implications for design

T.M.Fred Smith
University of Southampton, Department of Mathematics,
Southampton, SO17 1BJ, U.K.
tmfs@maths.soton.ac.uk

1. Introduction

A sampling strategy involves the selection of the pair ($\hat{\theta}_s$, p(s)), where $\hat{\theta}_s$ is an estimator of a target parameter, θ, and p(s) is the design for selecting the sample. However, before a strategy can be evaluated the statistician must first specify the target parameter, θ. This obvious point is frequently neglected by survey statisticians, but for optimum survey design, or redesign, there should be a measure of accuracy, such as total survey error, requiring both variance and bias and hence a target parameter. Three examples from the U.K. where the survey documentation does not include the definition of target parameters are:

1. the Labour Force Survey, where the target measure of unemployment is not defined;
2. the Average Earnings Index, where the lack of a clear target led to the continued use of an out-dated design and to misinterpretation of the results of the survey, and to serious problems when the design and estimator were changed;
3. all time series seasonally adjusted by the X11 programme.

The absence of a target parameter leads to confusion in methodological studies involving variance calculations, which in the case of the Average Earnings Index led to a crisis within the CSO and wider.

There are two major sources of data in official statistics; complete records based on administrative systems and sample surveys. For complete records the statistical problem is to provide adequate summaries of the information in the data. For cross-section data this is usually provided by cross-tabulations which disaggregate the data into sub-groups of interest. A major issue is where to stop this process. For dynamic analysis the cross-section summaries form a time series which adds another dimension to the analysis. Should change be measured by a comparative analysis of totals or by the flows that determine the change? Most would argue that the flows are needed to explain change but flows are severely affected by measurement and matching errors and are rarely available in sufficient detail from administrative records. For samples the problems of measurement and matching are now compounded by the addition of sampling error. For disaggregated data we run into the small domain (area) problem, similarly for gross flows. For time series analysis the observed covariance structure of the series is complicated by the superposition of the sampling errors.

How should target parameters be defined? The comparison above of complete records and samples has given us one mechanism for defining target parameters for samples. The target parameter is the value that would be calculated if there were a complete record (census) with no errors. It is this exercise which is frequently missing from survey documentation. Instead parameters are defined implicitly as that which is estimated by an estimator. This allows variances to be calculated but is inadequate for consideration of total survey error. It is also a potential source of confusion for users. In the UK some users of the Average Earnings Index thought that it was a measure of the average earnings of those in employment, namely total earnings in some time period divided by total number of employees, while others thought that it was a Laspeyres-type index of the change in earnings of a given labour force. In practice it was neither of these. What it should be is still not clear.

2. Micro-analysis

The choice of parameter depends on the objective of the analysis. Explanation of a phenomenon starts with the behaviour of the individual units in the population, with a micro-analysis. It is a bottom-up approach, for behaviour varies with context, and context puts the social into social science. Building context into behavioural models leads to multi-level (hierarchical) models creating interesting design issues.

Dynamic properties depend also on the treatment of the time axis. Should unemployment be defined continuously, daily, weekly, or for some other period? Administrative records may change daily while the ILO definition is based on a specified week. Is either correct? A complete record of continuous data can be represented on a Lexis diagram. But how can the information in such a diagram be summarised? How can context be incorporated into the diagram? How much information is lost by sampling the time axis? See Smith and Holt (1989) for a more detailed discussion. Practical considerations seem frequently to over-ride these theoretical issues.

For data recorded in continuous time one method of analysis of the time between events is survival analysis. Survival models can depend on variables that are fixed, such as sex, that change over time, education and age, or are contextual at different levels and may also change with time. An efficient design for a complex survival model must sample data from all these sources in such a way that parameters can be estimated efficiently. If the outcomes depend strongly on the contextual variables then there must be adequate samples within each selected context. This suggests a highly clustered design. Efficient estimation of contextual effects may also mean that the contexts should be chosen to include extremes. Random selections may concentrate too much on common contexts at the centre of the data set which give little information on the extreme conditions that may contain the most useful information. My conclusion is that design for micro-analysis should be based on the principles of experimental design rather than those of overall representative sampling. However, designs for monitoring population predictions should be based on representative samples.

3. Macro-analysis

Macro-analysis is a top-down approach. The outputs from longitudinal surveys form a time series of estimates which track the dynamics of a process. The target parameters should be the corresponding time series of population values which would be the ideal input into a dynamic model. There are many problems to consider from the perspective of design. One is the level of aggregation at which the estimates are produced. If the target parameter at time t is the proportion of people of working age who are unemployed how is that to be interpreted? If it is to be treated as a probability then the individual units should be exchangeable, there should be no grouping that changes the proportions in an informative manner. Forming homogeneous subgroups is a basic objective of much survey analysis but is plagued with difficulties such as those of Simpson’s paradox. Group sizes can become very small and in the limit may contain only one unit. Designing to allow many levels of disaggregation involves very large sample sizes and cost constraints will often preclude this option.

Macro-analysis is essentially descriptive and here the sampling of the time axis may be less important. It is the general path of a series that matters and the time interval can be determined by the uses to which the data are put. Series which affect policy decisions should be sampled at a frequency that relates to the policy levers and the time scale for effects to take place. If the policy lever is to change legislation in order to affect benefit payments then the time scale is long and quarterly or even annual data may suffice. If the policy lever is to change interest rates then the time scale for an effect may be less than a year and monthly data is required. Again there is little discussion in the statistics literature about the sampling of time and its relation to policy decisions.

Given a point in time there is little dispute about design principles. The target is a population total or mean and representative samples are needed. Efficiency is determined by the variation of the data and the cost of alternative methods of data collection. Sampling over time is more problematical since the target parameters are not only the cross-section totals at each time but also the components of change in these totals, see Holt and Skinner (1989). The design issues for totals and changes in totals are well covered in the literature, see, for example, Cochran (1977, Ch.12), and the review papers by Duncan and Kalton (1987) and Binder and Dick (1989). For measuring change the samples should overlap as much as possible, given practical constraints, while for totals the samples should be independent.

4. Time series analysis

How do you describe a time series of population values? There are several possibilities. One is to plot the series, which is a description, and to predict future values, which is an application. A prediction requires that the series be modelled in some way and this necessitates the estimation of the time series covariances. A second possibility is to simplify the series by extracting certain components and plotting the residuals. This is the approach of seasonal adjustment. A third is to model the structure of the series and to estimate the components of this structure. Once estimated the series can be deseasonalised or future values can be predicted. Possible models include the ARIMA class and the Basic Structural Model (BSM) of trend, seasonal and irregular components. See Harvey (1989) for details of these models which are descriptive rather than explanatory. Macro-econometric models are attempts at explanation. The approach selected determines the target parameters.

When a time series of population values is estimated from a survey the sampling errors complicate the analysis. Complex rotation patterns give rise to complex covariance structures which are superimposed on the covariance structure of the time series and should be taken account of in any analysis. Composite estimation further complicates the covariance structure Since the whole series is being estimated a second order sufficient statistic is the vector of cross-section estimates together with the sample covariance matrix. These should form the inputs to the estimation of the target parameters which may well take the form of composite estimators. The design issue is to choose a rotation pattern appropriate to the selected method of analysis. A design for trend estimation will be different from that for seasonal adjustment. Steel (1999) discusses designs for trend estimation.

Seasonal adjustment methods rarely involve target parameters. Algorithms, such as X11, do not specify any target parameters, they are judged solely by the properties of the outputs. Let Y_t, t=1,...,T, be a time series of population values. If $B(Y_t) = Y_{t-1}$, then the output of the X11 algorithm can be approximated by the filtered series $X(B)Y_t = X_t$, say. This combination of population values becomes the target for inference. If y_t, t = 1,...,T, is a vector of unbiased estimates of $Y_{t,}$ and X(B) is known then $x_t = X(B)y_t$ is an unbiased estimate of X_t. Is $X(B)Y_t$ a relevant target parameter? In practice the choice of X(B) is data based and is affected by the covariance structure of the sampling errors, and so the actual filter is sub-optimal, Hausman and Watson (1985).

Following Scott, Smith and Jones (1977) we can write the survey estimates as

$$y_t = Y_t + e_t,$$

where e_t is the survey error at time t. Assuming, as is reasonable, that the survey error is independent of the population time series error structure, we see that if $X(B)e_t$ is small then

$$X(B)Y_t \cong X(B)y_t.$$

This leads to Hausman and Watson's conclusion that a good rotation design is one that makes the survey error pattern as seasonal as possible. In fact, since X(B) is a complex filter, any pattern which is eliminated by the filter will suffice. If e_t is not eliminated by the filter then the residual (adjusted) series now confounds survey error with any trend and irregular components Estimates of trend based on seasonally adjusted data will suffer from this problem.

An alternative approach is to model the covariance structure of the survey errors and to remove this from the overall covariance matrix. Using the independence assumption we have

$$V(\mathbf{y}) = V(\mathbf{Y}) + V(\mathbf{e}),$$

so that the covariance matrix of the population series can be estimated by subtracting the covariance matrix of the survey errors from that of the observed series. Scott et al (1977) used this approach to fit ARMA models and Pfeffermann (1991) used it to fit the BSM. Harvey (1989) shows how both ARIMA models and the BSM can be expressed as Kalman filters which simplifies the computation of estimates and predictions.

The BSM with added survey error can be written as

$$y_t = T_t + S_t + I_t + e_t \; .$$

Identification of the separate terms requires that they be defined in some way and that their covariances also have a known structure. The Kalman filter defines the terms in the BSM recursively and imposes the assumption that the residual terms are uncorrelated with one another and with the survey errors. Another approach is to model the trend and seasonal terms explicitly as some form of polynomial, but this lacks flexibility. The modelling approach does not appear to impose any constraints on the survey errors other than that the error covariance matrix should be estimated as precisely as possible. This suggests that a non-overlapping design will be efficient.

Given a model overall change can be estimated by the deviation from prediction, as proposed by Smith (1978). Overall change can be due to change in any or all of the components and to explain change it is necessary to estimate the change in each component. This is best done by modelling the series and estimating parameters adaptively using the Kalman filter. This avoids the problem of confounding if trends are estimated from seasonally adjusted series.

Specifying target parameters helps to avoid some of the ambiguities and difficulties which have arisen in practice and gives guidance on the choice of design.

REFERENCES

Cochran, W.G. (1977). Sampling Techniques, 3rd ed. Wiley.

Duncan, G.J. and Kalton, G. (1987). Issues of design and analysis of surveys across time. Int. Statist. Rev., 55, 97-117.

Harvey, A.C. (1989). Forecasting Structural Time Series Models and the Kalman Filter. C.U.P.

Hausman, J.A. and Watson, M.W. (1985). Errors in variables and seasonal adjustment procedures. J. Amer. Statist. Assoc. 80, 531-540.

Holt, D. and Skinner, C.J. (1989). Components of change in repeated surveys. Int. Statist. Rev., 52, 1-18.

Pfefferman, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. J. of Bus. and Econ. Stats. 9, 163-175.

Scott, A.J., Smith, T.M.F. and Jones R. G. (1977). The application of time series methods to the analysis of repeated surveys. Int. Statist. Rev., 45, 13-28.

Smith, T.M.F. (1978). Principles and problems in the analysis of repeated surveys, in Survey Sampling and Measurement, ed. N.K. Namboodiri, Academic Press. 201-216.

Smith, T.M.F. and Holt, D. (1989). Some inferential problems in the analysis of surveys over time. Proc. of 47th Session of the ISI, C.1-17.1, 405-424.

Steel, D. and McLaren, C.H. (1999). Choosing rotation patterns for trend estimation. Proc. 52nd Session of ISI.

RESUME

Avant de parler du choix du plan de sondage, il est nécessaire dedéfinir les paramètres cibles. Ceux-ci dépendent des objectifs denotre inférence. Nous discuterons comment ces paramètres influencent le choix du plan de sondage. En particulier, nous mettrons l'accenssur les méthodes d'ajustement saisonnier.