UniS

Sociology at Surrey

University of Surrey

social research UPDATE

- Survey analysts routinely ignore complex design factors such as clustering, stratification and weighting.
- This results in biased estimates of standard errors and increased likelihood of Type I errors.
- A substantive example is used to illustrate the problem and appropriate software applications are briefly reviewed.

Analysing Complex Survey Data: Clustering, Stratification and Weights

Patrick Sturgis

The vast majority of surveys analysed by the UK social research community employ complex sample designs and weighting adjustments, yet are often treated as un-weighted simple random samples by analysts. This is unfortunate because failing to take these factors into account is likely to result in biased point and variance estimation. In this short article, I use data from the 2000 UK Time Use Survey (UKTUS) to show how these factors should be incorporated into the estimates produced from complex surveys and to illustrate the threats to accurate inference if they are ignored. This is not intended as a detailed statistical treatment of these issues but as a general discussion aimed at substantive and policy-oriented users of large scale social surveys. Readers in search of more detailed treatments are directed toward Kish (1962) and Groves et al (2004).

Clustering, Stratification and Weighting

Clustering – or multi-stage selection of sample units – is almost always used on national, face-to-face interview surveys, as non-clustered designs are both impractical from the perspective of data collection agencies and prohibitively expensive for funders of research.

For a fixed cost, clustering produces more precise population estimates than a simple random design would achieve. However, for a fixed sample size, clustered designs are subject to larger standard errors. This is because there tend to be greater similarities, on many attributes, between members of the same geographical sub-unit than between independently selected members of the total population. For instance, size of garden, number of bedrooms and household income are all variables that are intuitively likely to be more similar within than they are between postcode sectors. Clustering, therefore, underestimates true population variance and this is reflected in standard errors that are larger, if correctly estimated, than those that would have been obtained from a simple random sample of the same size.

National probability surveys also standardly employ stratification in the selection of sample units. Stratification divides the sample up into separate sub-groups and then selects random samples from within each group. These sub-samples are then combined to form the complete issued sample. Strata are created through the cross-classification of variables contained on the sampling frame, which are known or believed to correlate with key survey variables. So long as the latter assumption holds true, stratification will reduce sampling error, relative to an un-stratified sample design of the same size.

Sampling within strata can either be proportionate or disproportionate to population totals. In addition to obtaining increases in statistical efficiency, disproportionate stratification is often used to ensure that robust estimates can be

Patrick Sturgis is a lecturer in the department of sociology at the University of Surrey. He completed his PhD at the London School of Economics after working for several years in the Survey Methods Centre at the National Centre for Social Research. He is currently an investigator on three ESRC funded projects, examining public attitudes to genomics and changing social and political attitudes in the UK.

ISSUE 43

made within substantively important strata. For instance, surveys of the GB population might disproportionately sample within strata formed by the three countries of Great Britain. 'Over-sampling' within the Wales stratum would enable separate estimates to be produced for people living in Wales, where sample size might be too small under a proportionate stratification. To produce estimates representative of the GB population from such a disproportionate allocation, however, sample units from Wales would need to be down-weighted to their correct population proportion.

A third complex design factor employed by most national probability samples is the use of post-survey weighting. Weighting is generally applied to correct for unequal selection probabilities and nonresponse. The main purpose of this weighting is to reduce bias in population estimates by up-weighting population sub-groups that are under-represented and down-weighting those that are over-represented in the sample. A less desirable by-product of weighting however is that it can, when the variance of the weights is large, result in standard errors that are larger than they would be for un-weighted estimates.

Complex Designs and Variance Estimation

The net effect of clustering, stratification and weighting, therefore, is that the standard errors of these 'complex' sample designs tend to be different (smaller or larger, but usually larger) than those of a simple random sample. The difference in the precision of the estimates produced by a complex design relative to a simple random sample is known as the design effect (*deff*). The design effect is the ratio of the actual variance, under the sampling method used, to the variance calculated under the assumption of simple random sampling. This number will vary for different variables in the survey – some may be heavily influenced by design effects and others less so.

For cluster samples, the main components of *deff* are the intraclass correlation or *rho*, and the number of units within each cluster. *Rho* is a statistical estimate of within cluster homogeneity. It represents the probability that two units drawn randomly from the same cluster will have the same value on the variable in question, relative to two units drawn at random from the population as a whole. Thus, a *rho* of 0.10 indicates that two units randomly selected from within the same cluster are 10% more likely to have the same value than are two randomly selected units in the population as a whole. The design effect is calculated as follows:

$$deff = 1 + rho\ (n - 1),$$

where:

- *deff* is the design effect,
- *Rho* is the intra-class correlation for the variable in question,
- and n is the size of the cluster.

From this formula, we can see that the design effect increases as the cluster size (in most instances the number of addresses sampled within a postcode sector) increases, and as rho (within cluster homogeneity) increases.

A somewhat more readily interpretable derivation of the design effect is the design factor or '*deft*', which is the square root of *deff*. *Deft* gives us an inflation factor for the standard errors obtained using a complex survey design. For example, a *deft* value of 2, indicates that the standard errors are twice as large as they would have been had the design been a simple random sample. Deft can also be used to obtain the effective sample size, *neff*, which gives, for a complex survey design, the sample size that would have been required to obtain the same level of precision in a simple random sample.

In order to correctly estimate variance when analyzing survey data with a complex design, two main statistical approaches are available: Taylor Series approximation and Balanced Repeated Replication (BRR)[1]. An extended discussion of the properties of these estimators is beyond the scope of this article but see Groves et al (2004) for a detailed treatment. For the substantive analyst, however, the important thing to note is that many popular statistical software packages (such as SPSS and SAS) do not implement these procedures as standard. This means that, for a great many statistics, these packages produce standard

error estimates as if they were taken from a simple random sample, ignoring any complex design factors. If there is significant within cluster homogeneity on particular survey variables, if stratification has been used, or if any form of weighting has been applied during estimation, standard errors will, therefore, be biased. In the next section of this article, I illustrate the effect of complex design factors on standard survey estimates using a recently collected, publicly available data set.

Example: the UK 2000 Time Use Survey

The UKTUS 2000 collects information about how people spend their time, using 'own words' daily diaries to record detailed information about the activities people participate in during a particular day (see Deacon 2003). It uses a multi-stage sample design involving the stratified selection of a sample of postcode sectors. Addresses selected for the survey are taken only from these sectors and are thus 'clustered'. The data set contains weight variables which correct for unequal selection probabilities and differential nonresponse.

Table 1 shows standard errors and design effects for mean time spent per day sleeping, broken down by age and sex. The point estimate of the mean for each sub-group appears in column 1. The second column shows the estimated true standard error, that is the standard error taking into account the effects of clustering, stratification and weighting. The third column shows the 95% confidence interval around the point estimate using the true standard error and the fourth column shows the design factor, *deft* (the estimated ratio of the true standard error to the standard error of a simple random sample of the same size). Column six presents the size of the sample (or sub-sample) on which the estimate is based, while the final column shows *neff*, the sample size that would be required using a simple random sample to obtain the same level of precision. The standard errors in Table 1 were calculated using the software package Stata 7.0, which employs the Taylor Series approximation method for standard error estimation.

Table 1 Standard Errors for Mean Number of Minutes per day Sleeping

Base		Estimate	True s.e.	[95% Conf.	Interval]	Deft	n	neff
male	16-24	544.6	6.5	531.9	557.3	1.63	1090	412
	25-34	506.3	3.9	498.7	513.8	1.30	1406	827
	35-44	486.0	3.4	479.3	492.8	1.30	1590	937
	45-54	477.9	3.3	471.5	484.4	1.23	1569	1036
	55-64	491.6	3.9	484.0	499.3	1.39	1101	571
	65-74	505.8	4.4	497.1	514.4	1.44	874	419
	75+	522.8	6.0	511.0	534.5	1.50	486	216
female	16-24	545.7	4.2	537.3	554.0	1.14	1371	1058
	25-34	517.2	3.1	511.1	523.2	1.18	1804	1299
	35-44	501.0	2.9	495.4	506.6	1.18	1877	1353
	45-54	493.8	3.4	487.0	500.5	1.25	1746	1125
	55-64	492.5	3.6	485.3	499.7	1.37	1208	646
	65-74	503.1	3.7	495.9	510.3	1.27	1022	638
	75+	525.7	5.3	515.3	536.2	1.54	728	306
	Total	**518.1**	**1.2**	**515.7**	**520.6**	**1.57**	**20976**	**8533**

The design effects in Table 1 are, relative to those commonly found on similar surveys, large. The majority of *deft* values are above 1.2, a value commonly taken to indicate sizeable variance inflation. Quite a number are above 1.5, indicating an effective loss of more than 50% of the sample relative to a simple random sample.

Looking next at each of the design factors in isolation, the majority of this loss of precision results from the weights rather than the clustering or stratification[2]. Table 2 illustrates this by showing the design effects estimated for each of the three design factors on their own (estimates shown for men only). From Table 2 we can see that the effect of stratification on variance estimates is almost non-existent. This is probably because the three stratification variables used[3] are derived from the census, aggregated to the postcode sector level. This level of aggregation makes such variables, at best, only weakly predictive of individual level measurements.

Clustering serves to considerably reduce precision for all the estimates in table 2, although for five out of the seven estimates the relative contribution to the overall design effect is less than that from weighting.

Looking at the breakdown of design effects presented in table 2, we might be tempted

social research UPDATE is distributed without charge on request to social researchers in the United Kingdom by the Department of Sociology at the University of Surrey as part of its commitment to supporting social research training and development.

Contributions to social research UPDATE that review current issues in social research and methodology in about 2,500 words are welcome. All UPDATE articles are peer-reviewed.

Table 2 Relative Contributions of Design Factors to Variance Inflation for Mean Time Sleeping for Men by Age

Age	deft	deft due to stratification	deft due to clustering	deft due to weighting
16-24	1.62	0.99	1.27	1.29
25-34	1.30	1.00	1.02	1.20
35-44	1.30	0.99	1.10	1.19
45-54	1.23	1.00	1.09	1.11
55-64	1.39	1.00	1.17	1.19
65-74	1.44	0.99	1.28	1.18
75+	1.50	0.99	1.28	1.23

to conclude that the weights should not be used, due to the large loss in precision that their application clearly entails. This, however, would be a mistake as the reduction in bias from the application of the weights more than outweighs any loss of precision. This can be demonstrated by estimating the Mean Square Error (MSE) of these estimates, which is the sum of the variance and the square of the bias. It gives us the mean, or expected, difference between the true population figure we are attempting to estimate and the actual survey estimate (Groves 1989).

Table 3 shows MSE estimates for mean time sleeping for men by age, firstly taking into account clustering and stratification but not weighting and secondly taking into account all three design factors. Note that table 3 makes the simplifying assumption that the weighted estimates are unbiased, although we have no way of knowing the true population values. The substantially higher MSE estimates for the un-weighted data in Table 3 clearly indicate that weighting produces more accurate estimates, despite the loss in precision that this can sometimes produce.

The *de facto*, usually implicit, assumption of many substantive social researchers is that surveys with complex sample designs can be analysed as if they were simple random samples. This is largely due to a lack of awareness of the variance estimation problems caused by clustering, stratification and weighting but also because people simply trust the software they use to provide correct estimates of these parameters.

Software options for correct variance estimation are, of course, constantly and rapidly evolving. SPSS, in the most recent version (SPSS 12.0), can now provide correct variance estimates for means, proportions, crosstabulations and ratios under complex sample designs. A wider range of statistics, including a variety of regression estimators, is available in the most recent version of Stata (Stata 8.0). SUDAAN, produced by the Research Triangle Institute (http://www.rti.org) can handle a variety of descriptive and multivariate estimators. All these options are commercially licensed. Free software (CENVAR and VPLX) for a range of estimators is available from the US Census Bureau (http://www.census.gov).

Table 3 Mean Square Error Estimates for Mean Time Sleeping for Men by Age

	Un-weighted				Weighted			
Age	Estimate	S.E.	bias	MSE	Estimate	S.E.	bias	MSE
16-24	563.4	5.3	18.8	379.5	544.6	6.5	0.0	41.8
25-34	521.9	3.5	15.6	255.3	506.3	3.9	0.0	14.9
35-44	496.1	3.2	10.0	110.8	486.0	3.4	0.0	11.9
45-54	489.3	3.0	11.4	137.8	477.9	3.3	0.0	10.8
55-64	497.4	3.4	5.8	45.0	491.6	3.9	0.0	15.1
65-74	508.4	3.8	2.6	21.5	505.8	4.4	0.0	19.2
75+	521.1	5.5	-1.7	33.5	522.8	6.0	0.0	35.6

References

Deacon, K. (2003) The UK 2000 *Time Use Survey: Technical Report*, GSS.

Groves, R. (1989). *Survey Errors and Survey Costs*. New York: John Wiley & Sons.

Groves, R. Fowler, F., Couper, M. Lepkowski, J, Singer, E. and Tourangeau, R. (2004) *Survey Methodology*, NY: Wiley.

Kish, L (1962) Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Skinner, C., Holt, D. and Smith, T. (1989) *Analysis of Complex Surveys*, NY: Wiley.

Endnotes

[1] An alternative to BRR, based on the jackknife, can also be used to take account of these complex design factors (see Skinner et al 1989).

[2] Note that the design effects estimated in isolation need not sum exactly to the total *deft*.

[3] % households with head of household in social economic group 1-5, population density, and % unemployed.

social research UPDATE

(ISSN: 1360-7898)

is published by

Department of Sociology
University of Surrey
Guildford GU2 7XH
United Kingdom.
Tel: 01483 300800
Fax: 01483 689551
Edited by Nigel Gilbert
(e-mail: **n.gilbert@soc.surrey.ac.uk**)