International Journal of Methods in Psychiatric Research, Volume 7, Number 1

Complex sample design effects and inference for mental health survey data

STEVEN G. HEERINGA, Division of Surveys and Technologies, Institute for Social Research, University of Michigan, Ann Arbor USA

JINYUN LIU, Survey Design and Analysis Unit, Institute for Social Research, University of Michigan, Ann Arbor, USA.

ABSTRACT Mental health researchers world-wide are using large-scale sample survey methods to study mental health epidemiology and services utilization in general, non-clinical populations (Alegria et al. in press). This article reviews important statistical methods and software that apply to descriptive and multivariate analysis of data collected in sample surveys. A comparative analysis of mental health surveys in international locations is used to illustrate analysis procedures and 'design effects' for survey estimates of population statistics, model parameters and test statistics.

This article addresses the following questions. How should a research analyst approach the analysis of sample survey data? Are there software tools available to perform this analysis? Is the use of 'correct' survey analysis methods important to interpretation of survey data? It addresses the question of approaches to the analysis of complex sample survey data. The latest developments in software tools for the analysis of complex sample survey data are covered, and empirical examples are presented that illustrate the impact of survey sample design features on the interpretation of confidence intervals and test statistics for univariate and multivariate analyses.

Key words: population surveys, sample design, stratification, clustering, weighting, sampling variance, design effect, Wald statistics

Design effects on variance estimates and test statistics

Most surveys used in the study of psychiatric epidemiology and related fields are based on stratified multistage probability samples of household populations. For example, the multi-national archive of research projects that has been created by the International Consortium on Psychiatric Epidemiology (ICPE) currently includes 12 survey data sets. The sample designs for each of these data sets are distinct and adapted to the local, regional, and national population of interest; but they all share design and estimation features that must be taken into account when deriving population estimates and making statistical inferences to the corresponding survey populations. The survey literature labels the samples for these studies as 'complex designs', a term denoting the fact that the sample incorporates special design features such as stratification, clustering and weighted estimation that do not conform to the distributional assumptions of standard statistical analysis packages.

Standard analysis programs included in statistical software systems such as SAS (SAS Institute Inc. 1990), SPSS (SPSS Inc. 1993) or Stata (StataCorp 1997) assume simple random sampling (SRS) designs were used to collect the survey data. The SRS sampling assumption is synonymous with the assumption that the survey data are independent and identically distributed (IID). Standard errors of estimates and test statistics computed by standard analysis programs are only valid if the sample data meet the IID criteria (or closely approximate it). The assumption of independence of the sample observations does not hold for most data collected using complex sample designs and applying the SRS/IID assumption to complex sample survey data generally results in underestimation for standard errors of survey estimates of descriptive statistics and model parameters. Confidence intervals and test statistics based on computed standard errors that assume independence of observations will be biased.

Some useful terminology and notation is needed before turning to actual procedures and software for the analysis of sample survey data. Leslie Kish (1965) first defined the term 'design effect' as the ratio of the true design-based sampling variance of an estimated statistic to the simple random sampling variance for the same estimated statistic. As the confidence intervals for a statistic, t, is a linear function of the square root of its sampling variance, Kish defines DEFT(t) as the ratio of the true design-based standard error to the SRS standard error of t:

$$DEFT(t) = \frac{SE(t)_{des}}{SE(t)_{srs}} \tag{1}$$

This ratio measures the relative efficiency of a statistic estimated from complex sample survey data compared with a sample of equivalent size selected by an SRS design. The value of DEFT for an estimated statistic quantifies the relative increase/decrease of the true sampling error compared to the sampling error that would be reported by a standard statistical analysis program. Note that the exact form of the statistic is not specified in the expression for DEFT(t). The statistic, t, might correspond to any estimator function based on the survey data – an estimated mean, a proportion or rate, a pairwise or multiple correlation, a linear or logistic regression coefficient, or even a test statistic.

Design effects in survey data are caused by three features of the sample design and estimation process:

- stratification of the survey population prior to selection;
- clustering or grouping of elements in the process of sample selection; and
- differential weighting of sample units in estimation and analysis.

Stratification is intended to improve the efficiency of the sample relative to simple random samples. Therefore, effective stratification of the sample design tends to reduce the design effect for estimates computed from the survey sample data (Cochran 1977). Clustering or grouping of sample elements in survey designs is generally required to reduce data collection costs to acceptable levels. Multistage probability sample designs used for most in-person population surveys involve selecting clusters of households and individuals. Sampling units selected as clusters are not independent sample selections. Intra-class correlations among the responses for clustered subjects reduce the efficiency of the sample design and therefore tend to increase the design effect for sample estimates (Kish 1965). Weighting serves several purposes in the analysis of survey data. Sample selection weight factors and non-response weighting corrections compensate for unequal probabilities of selecting and observing survey subjects. Sample-based estimates of survey statistics may not be unbiased if these weighting adjustments are ignored. Survey estimates may also include post-stratification weight factors that adjust weighted sample marginals for key demographic variables (for example age, sex, geographic region) to population distributions measured in censuses or other administrative data collections. Weighted estimation can increase or decrease the design effect for survey estimates depending on the correlation of the weight values with the standard deviations of the variable(s) used in the estimation of the survey statistic t (Kish 1965).

The complex effects and interactions of stratification, clustering and estimation weighting that produce design effects in survey estimates are difficult if not impossible to model mathematically. Simple models of design effects and variance increase due to estimation weighting have been proposed for means and proportions (Kish 1965) and simple linear regression coefficients (Skinner, Holt and Smith 1989). The lack of formal analytical models for studying design effects is primarily a problem for the designers of samples who must be guided in their design work by these simple models and observations of estimated design effects from previous studies of similar design. Design effects are essentially an empirical problem for survey analysts since there are well-established methods and software that allow these analysts to perform correct estimation and inference from the sample survey data. These methods and software are covered in the following section.

Analysis methods and programs

Since the late 1930s, advances in survey sampling theory have guided the development of a number of methods for correctly estimating variances from complex sample survey data sets, but for many years analysts who were interested in applying these methods in the analysis of their survey data were limited to a number of specialized stand-alone programs. By today's standards most of these programs were limited in the scope of statistical procedures they included, were poorly documented, and not 'user friendly'. The past few years have brought important improvements in the scope and usability of software that is specifically designed for the analysis of survey data.

What is special about software that is intended for the analysis of survey data? The software must correctly handle case-specific weights in estimation and inference. Weighted computations of the sample statistics are required for unbiased estimation of population values. Weighting effects must be incorporated in the estimation of standard errors of these unbiased estimates in order to develop correct confidence intervals or test statistics required for inference about population values. Standard analysis procedures found in the major statistical software packages enable weighted estimation but fail to account for the effects of weighting in the computation of standard errors and test statistics. To prevent biased inferences, the survey data analysis program must account for the design effects due to stratification, clustering, and weighting of sample observations.

Researchers (Goldstein 1987) have proposed parametric, model-based approaches to the analysis of data from stratified and clustered sample designs, but the vast majority of work in this area has focused on non-parametric, design-based approaches that provide reliable estimates of standard errors and robust inferences from large-sample survey data sets (Rust 1985). The focus in this paper will be on the design-based methods and software for survey estimation and inference (Wolter 1985).

The two most common approaches to the estimation of standard errors for complex sample survey statistics are the Taylor series linearization method and resampling variance estimation procedures such as balanced repeated replication (BRR) or jack-knife repeated replication (JRR). Bootstrap methods for variance estimation can also be included among the resampling approaches (Rao and Wu 1988).

Taylor series linearization method

When survey data are collected using a complex sample design with unequal size clusters, most statistics of interest will not be simple linear functions of the observed data. The linearization approach applies Taylor's method to derive an approximate form of the estimator that is linear in those statistics for which variances and covariances can be directly and easily estimated (Woodruff 1971). SUDAAN and Stata are two commercially available statistical software packages that include procedures that apply the Taylor series method to estimation and inference for complex sample data.

SUDAAN

SUDAAN (Shah et al. 1996) is a commercially available software system developed and marketed by the Research Triangle Institute of Research Triangle Park, North Carolina (USA). It was developed as a stand-alone software system with capabilities for the more important methods for descriptive and multivariate analysis of survey data, including: estimation and inference for means, proportions and rates (PROC DESCRIPT and PROC RATIO); contingency table analysis (PROC CROSSTAB); linear regression (PROC REGRESS); logistic regression (PROC LOGISTIC); log-linear models (PROC CATAN); and survival analysis (PROC SURVIVAL). SUDAAN V7.0 and earlier versions were designed to read directly from ASCII and SAS system data sets. The latest versions of SUDAAN permit procedures to be called directly from the SAS system. Information on SUDAAN is available at the following website address: http://www.rti.org.

Stata

Stata (StataCorp 1997) is a more recent commercial entry to the available software for analysis of complex sample survey data and has a growing body of research users. Stata includes special versions of its standard analysis routines that are designed for the analysis of complex sample survey data. A special survey analysis program is available for descriptive estimation of means, ratios, proportions and population totals (SVYMEAN). Stata programs for multivariate analysis of survey data currently include linear regression (SVYREG), logistic regression (SVYLOGIT) and probit regression (SVYPROBT). Information on the Stata analysis software system can be found on the Web at: http://www.stata.com.

Resampling methods

Balanced repeated replication (BRR), JRR and the bootstrap comprise a second class of non-parametric methods for conducting estimation and inference from complex sample data. As suggested by the generic label for this class of methods, BRR, JRR and the bootstrap use replicated subsampling of the sample database to develop sampling variance estimates for linear and non-linear statistics. WesVar PC (Brick et al. 1996) is a publicly available software system for personal computers that employs replicated variance estimation methods to conduct the more common types of statistical analysis of complex sample survey data. WesVar PC was developed by Westat Inc. and is distributed along with documentation free of charge to researchers from Westat's Web site: http://www.westat.com/wesvarpc/. WesVar PC includes a Windows-based application generator that enables the analyst to select the form of data input (SAS data file, SPSS for Windows database, dBase file, ASCII data set) and the computation method (BRR or JRR methods). Analysis programs contained in WesVar PC provide the capability for basic descriptive (means, proportions, totals, crosstabulations) and regression (linear, logistic) analysis of complex sample survey data.

These new and updated software packages include an expanded set of user-friendly, well-documented analysis procedures. Difficulties with sample design specification, data preparation, and data input in the earlier generations of survey analysis software created a barrier to use by analysts who were not survey design specialists. The new software enables the user to input data and output results in a variety of common formats, and the latest versions accommodate direct input of data files from the major analysis software systems. Readers who are interested in a more detailed comparison of these and other survey analysis software alternatives are referred to Cohen (1997).

Design effects for estimates of means, proportions and rates

Empirical research (Kish, Groves, Krotki 1975) involving sampling error analysis of large numbers of survey data sets has demonstrated that the greatest design effects occur for simple population estimates of univariate statistics, including estimates of means, proportions, and ratios. In most complex sample surveys the design effects for these statistics will be greater than 1.0. Consequently, the true confidence intervals for point estimates of these univariate statistics can be much wider than the output from standard analysis programs would suggest.

To illustrate the nature of complex sample design effects for simple univariate statistics, estimates, standard errors, and design effects for two percentages were computed from four mental health survey data sets that are included in the ICPE archive:

- the 1990 National Comorbidity Survey (NCS) of United States' household residents;
- the 1990 Mental Health Supplement to the Ontario Health Survey (MHS OHS) of Ontario, Canada's household population;
- the 1996 Netherlands Mental Health Survey and Incidence Study (NEMESIS) of the Dutch household population; and
- the 1992 Puerto Rico Mental Health Care Utilization (MHCU) Project survey of households in lower income regions of Puerto Rico.

These analyses were restricted to survey respondents between the ages of 18–54 at the time of interview.

Table 1 presents estimates, standard errors and design effects for survey estimates of the percentage of adults in each survey population who sought treatment from a formal mental health service provider within the twelve month period prior to interview. Table 2 presents these same statistics for the survey estimates of the percentage of respondents who reported at least one episode of major depression within the 12-month period preceding the survey interview. Results for each example statistic are presented for the total sample and separately for male and female respondents.

The weighted estimates and standard errors presented in Table 1 and Table 2 were computed using SUDAAN V7.0. Standard errors reported in Tables 1 and Table 2 therefore include the effects of stratification, clustering, and weighting.

The results of this simple empirical test demonstrate several points. First, the estimated values of DEFT range from 0.91 to 1.63 and are greater than 1.0 for all but two of the estimated percentages. Second, design effects for the estimated percentages vary substantially across the independent studies and across gender subclasses within the studies. We will use the NCS estimate of the percentage of women reporting a major depression episode in the prior 12 months to illustrate the effect that correct estimates of sampling error have on the interpretation of these simple statistics. Uncorrected for sample design effects, the weighted NCS estimate of the percent of US women age 18–54 who have experienced major depression in the past 12 months is 12.5% with an estimated standard error of

Table 1: Design effects for survey estimates of percentage of population (aged 18–54) with 12-month use of formal mental health services

	Total population			Males			Females		
Survey	p(%)	se(p)	deft(p)	p(%)	se(p)	deft(p)	p(%)	se(p)	deft(p)
NCS	8.13%	0.49%	1.31	5.94%	0.60%	1.32	10.32%	0.74%	1.26
MHS OHS	6.56%	0.34%	1.37	4.75%	0.55%	1.44	8.33%	0.64%	1.30
NEMESIS	13.45%	0.53%	1.18	10.27%	0.53%	0.95	16.75%	0.89%	1.27
MHCU	8.26%	0.55%	1.02	7.44%	0.84%	1.14	9.00%	0.70%	0.91

Table 2: Design effects for survey estimates of percentage of population (18–54) with 12-month diagnosis of major depression

	Total population			Males			Females		
Survey	p(%)	se(p)	deft(p)	p(%)	se(p)	deft(p)	p(%)	se(p)	deft(p)
NCS	9.86%	0.57%	1.40	7.22%	0.67%	1.35	12.50%	0.86%	1.35
MHS OHS	4.48%	0.43%	1.63	3.06%	0.48%	1.54	5.88%	0.60%	1.41
NEMESIS	6.14%	0.40%	1.26	4.25%	0.46%	1.23	8.11%	0.63%	1.23
MHCU	6.17%	0.59%	1.27	4.29%	0.71%	1.24	7.86%	0.86%	1.19

0.63% (not shown). When the design-based procedures of SUDAAN V7.0 are applied, the weighted estimate of the 12-month depression rate remains unchanged at 12.5% but the corrected estimate of the standard error rises to 0.86%. The value of DEFT for the estimated percentage is therefore 0.86%/0.63% = 1.35. Researchers who ignore the design effect correction in analysis and publication would report a 95% confidence interval for the estimated 12-month percentages as 12.50% +/– 1.96 * 0.63% = [11.26%, 13.74%]. The corrected confidence interval that reflects the complex sample design effects on the standard error should be 12.50% +/– 1.96 * 0.86% = [10.81%, 14.19%].

Design effects in multivariate estimation and hypothesis testing

To illustrate how design effects may influence multivariate analysis of complex sample survey data, we consider a logistic regression model of the probability that 18–54-year-old US residents used a formal mental health service in the 12-month period prior to interview. The dependent variable in this analysis is the dichotomous indicator: 0 – no formal mental health service used; 1 – more than one visit to a formal mental health provider in the past 12 months. The independent variables included in the model are age (four categories), education level (four categories), marital status (three categories), and gender. The data for this exercise are from the 1990 National Comorbidity Survey (NCS). Table 3 presents three series of estimated coefficients and standard errors for the fitted model. Column (1) contains estimated parameters and standard errors when the model is fitted by SAS PROC LOGISTIC, and both the sample weights and other design effects are ignored. Refitting the model using SAS PROC LOGISTIC and including the NCS analysis weight produces the estimates and standard error shown in Column (2) of Table 3. The fitted model described in Column (3) was estimated using SUDAAN V7.0 PROC LOGISTIC and includes the effects of the NCS weights and sample design features in the estimation of the model coefficients and their standard errors. Comparison of the estimated logistic regression coefficients in columns (1) and (2) points out the effect of the analysis weights on the point estimates of the logistic model parameters. Note two general results from the comparison of the fitted models represented in Columns (2) and (3). First, since both model fits incorporate the analysis weights, the point estimates of the logistic regression coefficients are identical. Second, the comparison illustrates the increase in standard errors of parameters estimates when the model estimation correctly accounts for the design effects of weights, stratification and clustering. Taking as an example the standard error of the estimated coefficient for gender, DEFT(β_{gen}) = 0.129/0.104 = 1.240. Confidence intervals for β_{gen} are therefore 24% wider when the design effects are correctly included in the estimation of the standard error.

Table 4 extends the empirical exercise based on the logistic model of formal mental health service use to three additional data sets from the ICPE archive. Point estimates of the gender coefficient and standard errors are again reported for the three estimation approaches. The patterns of differences between unweighted and weighted parameter estimates and the observed design effects when standard errors are correctly estimated follow those shown in Table 3. Note again that the unweighted and weighted estimates of the logistic regression coefficients differ for each survey data set. Comparing results for models fitted with and without design effect corrections for standard errors, the values of DEFT(β_{gen}) are: NCS (1.38); MHS OHS (1.22); NEMESIS (1.08); MHCU (1.03). The smaller DEFT (β_{gen}) values for the Netherlands and Puerto Rico model coefficients can be attributed to reduced clustering and weighting effects relative to those for the US and Ontario samples.

The process of building parsimonious multivariate models requires the analyst to test the significance of the contribution of main effects and their interactions to the overall fit of the model. For logistic and other generalized linear models (GLMs), tests of significance of effects are most often based on the likelihood ratio test statistic (McCullagh and Nelder 1989). The logis-

Table 3: Effect of weighting and sample design on model estimation – logistic model of 12-month use of formal services, data from the 1990 NCS

Independent variables and effects	(1) SAS V6.12 No weights		(2) SAS V6.12 Weighted		(3) SUDAAN V7.0 Weighted, design		
	$\hat{\beta}$	$se\hat{\beta}$	$\hat{\beta}$	$se\hat{\beta}$	$\hat{\beta}$	$se\hat{\beta}$	$DEFT(\hat{\beta})$
Intercept	–1.931	0.169	–2.573	0.206	–2.573	1.243	6.03
Age 1	–1.090	0.180	–0.686	0.207	–0.686	0.242	1.17
2	–0.253	0.130	–0.221	0.143	–0.221	0.227	1.59
3	0.0162	0.127	0.047	0.138	0.047	0.210	1.52
4	0.000	–	0.000	–	0.000	–	–
Educ 1	–0.015	0.149	0.112	0.168	0.112	0.227	1.35
2	–0.410	0.124	–0.142	0.137	–0.142	0.167	1.22
3	–0.077	0.121	0.148	0.143	0.148	0.169	1.18
4	0.000	–	0.000	–	0.000	–	–
Marital status							
1	–0.365	0.123	–0.111	0.153	–0.111	0.165	1.08
2	–0.372	0.140	0.250	0.189	0.250	0.196	1.04
3	0.000	–	0.000	–	0.000	–	–
Gender	0.575	0.094	0.597	0.104	0.597	0.129	1.24

Table 4: Effect of weighting and sample design on model estimation – logistic model of 12-month use of formal services, estimates of gender coefficient and standard errors. Data from four ICPE survey data sets

Survey	(1) SAS V6.12 No weights		(2) SAS V6.12 Weighted		(3) SUDAAN V7.0 Weighted, design		
	$\hat{\beta}$	$se\hat{\beta}$	$\hat{\beta}$	$se\hat{\beta}$	$\hat{\beta}$	$se\hat{\beta}$	$DEFT(\hat{\beta})$
NCS (United States)	0.575	0.094	0.597	0.104	0.597	0.129	1.38
MHS OHS (Ontario)	0.694	0.113	0.571	0.108	0.571	0.138	1.22
NEMESIS (Netherlands)	0.505	0.079	0.589	0.080	0.589	0.085	1.08
MHCU (Puerto Rico)	0.141	0.145	0.126	0.147	0.126	0.150	1.03

tic regression model likelihood assumes a binomial distribution for the dichotomous dependent variable, y ~ B(n,p). The logistic link function defines the regression relationship between the binomial parameter, p (the expected value of y) and the effects of interest. If the specified likelihood for the data is correct, the LRT statistic is distributed as a chi-square random variable with q degrees of freedom. Application of the LRT to models based on complex sample survey data is complicated by the fact that the design effects (stratification, clustering, and weighting) alter the data likelihood. In simpler cases, it may be possible to respecify the likelihood function to directly incorporate the design effects and apply the LRT based on the revised correct data likelihood. However, in most analyses of survey data, the only practical choice is to use a Wald test (Skinner, Holt and Smith 1989) to evaluate significance of model effects and interactions.

Table 5 summarizes the results from an empirical comparison of LRT and Wald tests of the significance of the four main effects considered in the logistic regression model for the probability that eligible adults received formal mental health services in the prior twelve-month period. The significance of the four main effects (age, education, marital status, and gender) is tested individually (the remaining three main effects are included in the reduced model). SAS V6.12 PROC LOGISTIC was used to obtain the LRT statistic that assumes a binomial likelihood and independence of observations. The corresponding Wald statistsics were computed by SUDAAN 7.0 and correctly incorporate design effects in the estimation of the variance/covariance matrix for the model. The LRT and Wald test statistics are both referred to the Chi-square distribution with the indicated degrees of freedom. For reference, critical values (α = 0.05) have been provided along with the values of the test statistics in Table 5.

The results presented in Table 5 demonstrate that inclusion of sample design effects in the calculation of test statistics can in fact change the analyst's interpretation of the significance of effects in the model. The uncorrected LRT statistics lead us as analysts to conclude that age, education, marital status and gender are significant predictors (α = 0.5) of 12-month use of formal treatment services. However, the apparent significance of the marital effect for NCS disappears when the test is based on the Wald statistic that correctly incorporates the design effect on the variances and covariances of the model parameters. In contrast, the LRT and Wald statistics to test the significance of age, education, and gender are numerically different but both lead to the consistent conclusion that these are significant predictors of 12-month use of treatment services.

To further illustrate the need for sample design

Table 5: Effect of weighting and sample design on test statistics. Logistical model of 12-month use of formal mental health services. Data from four ICPE survey data sets

			NCS	MHS OHS	NEMESIS	MHCU
Effect	d.f.	Test	(United States)	(Ontario)	(Netherlands)	Puerto Rico)
Age	3	LRT	16.543	28.313	23.896	30.777
		WALD	10.372	21.704	16.079	22.100
		$\chi^2_{3.95}$	7.815	7.815	7.815	7.815
Education	3	LRT	5.851	5.814	6.267	1.635
		WALD	3.147	2.378	5.141	1.615
		$\chi^2_{3.95}$	7.815	7.815	7.815	7.815
Marital status	2	LRT	6.665	37.443	78.285	10.612
		WALD	5.544	29.096	135.851	7.227
		$\chi^2_{2.95}$	5.991	5.991	5.991	5.991
Gender	1	LRT	33.878	29.086	55.621	0.732
		WALD	21.456	18.137	48.880	0.680
		$\chi^2_{1.95}$	3.841	3.841	3.841	3.841

effect corrections in estimation and inference for multivariate models, we consider a discrete time logistic model example (Yamaguchi 1991). Discrete time logistic models are used in survival analysis to model the length of time to an event. The event that is modelled in this example is the age at which an individual reports the first use of an illicit drug. Individuals who report never having used an illicit drug are considered to be censored observations with censoring occurring at the age of interview. The data for this analysis are drawn from the 1990 Mental Health Supplement to the Ontario Health Survey. A dichotomous variable indicating first drug use (0 = no; 1 = yes) is modelled as a function of age (Age 0–10, Age11, . . ., Age 25, Age 26–30, Age 31–37), birth year cohort (Cohort 1, Cohort 2, Cohort 3) and gender. Terms for first and second order interaction between age categories, birth cohort and gender are also included in the model. Each model was first estimated using SAS PROC LOGISTIC and then re-estimated using SUDAAN V7.0 PROC LOGISTIC. Analysis weights were used in both cases but only the SUDAAN 7.0 analysis incorporates the design effects in the estimation of standard errors and test statistics.

Table 6 compares the estimated model coefficients and standard errors for the main effects in the model estimated from the MHS OHS (Ontario) data. Again, we observe that the design effects for model coefficients tend to be greater than 1. Table 7 compares results for LRT and Wald tests of hypotheses concerning the significance of the main effects and interaction terms in discrete time logistic models of age at first illicit drug use. As in the previous example, the numerical value of the corrected Wald statistic is always much smaller than the uncorrected LRT statistic. However, in this example the Wald test that reflects the sample design effects leads to the same decision concerning the significance of the model terms for main effects and interactions.

Summary

This article has described special methods and statistical software for analysis of survey data. The new generation of survey analysis software provides researchers with flexible and easy-to-use program tools for correctly analysing complex sample survey data. Examples drawn from analysis of mental health survey data have shown the importance of the complex sample design effects on survey-based estimation and inference for univariate and multivariate statistics.

Table 6: Effect of weighting and sample design on model estimation. Discrete time logistic model of age of first illicit drug use. Data from the 1990 *Mental Health Supplement*, Ontario Health Survey

	(1) (1)SAS V6.12 Weighted		(2) SUDAAN V7.0 Weighted, design		
Var Name	$\hat{\beta}$	$se\hat{\beta}$	$\hat{\beta}$	$se\hat{\beta}$	$DEFT(\hat{\beta})$
Age 0 – 10	–11.56	0.64	–11.56	0.81	1.27
Age 11	–8.92	0.56	–8.92	0.55	0.98
Age 12	–8.24	0.44	–8.24	0.71	1.61
Age 13	–7.32	0.34	–7.32	0.49	1.44
Age 14	–6.86	0.31	–6.86	0.36	1.16
Age 15	–6.36	0.30	–6.36	0.40	1.33
Age 16	–6.08	0.29	–6.08	0.41	1.41
Age 17	–5.66	0.28	–5.66	0.35	1.25
Age 18	–5.95	0.29	–5.95	0.41	1.41
Age 19	–6.82	0.32	–6.82	0.41	1.28
Age 20	–6.70	0.32	–6.70	0.36	1.13
Age 21	–6.51	0.31	–6.51	0.45	1.45
Age 22	–6.82	0.33	–6.82	0.60	1.82
Age 23	–8.03	0.47	–8.03	0.72	1.53
Age 24	–8.37	0.54	–8.37	0.55	1.02
Age 25	–6.54	0.32	–6.54	0.49	1.53
Age 26 – 30	–8.03	0.33	–8.03	0.36	1.09
Age 31 – 37	–8.70	0.36	–8.70	0.43	1.19
Gender	–0.76	0.09	–0.76	0.14	1.56
Cohort 1	2.71	0.28	2.71	0.36	1.29
Cohort 2	2.53	0.27	2.53	0.36	1.33
Cohort 3	2.09	0.27	2.09	0.35	1.30

Table 7: Effect of weighting and sample design on test statistics. Discrete time logistic model of age of first illicit drug use. Data from the 1990 MHS HIS

	MHS OHS (Ontario)			
Ho:Effect=0	$\chi^2_{df.95}$	df	LRT	Wald
Gender × Age × Cohort	12.60	6	8.76	4.35
Gender × Cohort	5.99	2	10.71	6.90
Cohort × Age	12.59	6	29.68	21.49
Gender × Cohort	7.82	3	4.49	2.85
Cohort	7.82	3	210.52	67.41
Gender	3.84	1	72.85	30.34

Acknowledgements

The research reported here was carried out in conjunction with the international Consortium in Psychiatric Epidemiology (ICPE). More information about ICPE can be obtained from http://www.hcp.med.harvard. edu/icpe

References

Alegria M, Kessler R, Bijl R, Lin E, Heeringa SG, Takeuchi DT, Kolody B (in press) To appear in The Unmet Need for Treatment. Proceedings of a Symposium of the World Psychiatric Association, Sydney, Australia, October, 1997.

Brick, JM, Broene P, James P, Severynse J. A User's Guide to WesVar PC. Rockville MD: Westat Inc, 1996.

Cochran WG. Sampling Techniques. New York: John Wiley & Sons, 1977.

Cohen SB. An evaluation of alternative PC-based software packages developed for the analysis of complex survey data. The American Statistician 1997; 51(3): 285–92.

Goldstein H. Multi-level Models in Educational and Social Research. London: Oxford University Press, 1987.

Kish L. Survey Sampling. New York: John Wiley & Sons Inc, 1965.

Kish L, Groves RM, Krotki KP. Sampling errors for fertility surveys. Occasional Paper No. 17. Voorburg, Netherlands: World Fertility Survey, International Statistical Institute, 1975.

McCullagh PM, Nelder JA. Generalized Linear Models (2 edn). London: Chapman & Hall, 1989.

Rao JNK, Wu CFJ. Resampling inference with complex sample data. Journal of the American Statistical Association 1988; 83: 231–9.

Rust K. Variance estimation for complex estimators in sample surveys. Journal of Official Statistics 1985; 1(4): 381–97.

SAS Institute Inc. SAS/STAT® User's Guide, Version 6 (4 edn) vol. 2. Cary NC: SAS Institute Inc, 1990.

Shah BV, Barnwell BG, Biegler GS. SUDAAN User's Manual: Software for Statistical Analysis of Correlated Data. Research Triangle Park, NC: Research Triangle Institute, 1996.

Skinner CJ, Holt D, Smith TMF. Analysis of Complex Surveys. New York: John Wiley & Sons, 1989.

SPSS Inc. SPSS® for Windows™. BASE System User's Guide, Release 6.0. Chicago Il: SPSS Inc, 1993.

Stata Corp. Stata Statistical Software: Release 5.0. College Station TX: Stata Corporation, 1997.

Wolter KM. Introduction to Variance Estimation. New York: Springer-Verlag, 1985.

Woodruff RS. A simple method for approximating the variance of a complicated estimate. Journal of the American Statistical Association 1971; 66: 411–14.

Yamageuchi K. Event History Analysis. Applied Social Research Methods Series, vol. 28. Newbury Park, CA/London: Sage Publications, 1991.

Correspondence to Steven G Heeringa, Institute for Social Research, University of Michigan, 426 Thompson Street,
Ann Arbor, MI 48104, USA. Tel. (+1)(734)-936-0991.
Fax (+1)(734)-763-3750. E-mail: sheering@isr.umich.edu.