Chapter XX
More advanced approaches to the analysis of survey data

Gad Nathan
Hebrew University
Jerusalem, Israel

Abstract

In the present chapter, we consider the effects of complex sample design used in practice in most sample surveys on the analysis of the survey data. The cases in which the design may or may not influence analysis are specified and the basic concepts involved are defined. Once a model for analysis has been set up, we consider the possible relationships between the model and the sample design. When the design may have an effect on the analysis and additional explanatory variable related to the design cannot be added to the analytical model, two basic methodologies may be used: classical analysis, which could be modified to take the design into account; or a new analytical tool, which could be developed for each design. Different approaches are illustrated with real-data applications to linear regression, linear models, and categorical data analysis.

Key terms: complex sample design, analysis of survey data, linear regression, linear models, categorical data analysis, model-based analysis.

.

A. Introduction

1. Sample design and data analysis

1. The primary purpose of the vast majority of sample surveys, both in developed and in developing countries, is a descriptive one, namely, to provide point and interval estimates of descriptive measures of a finite population, such as means, medians, frequency distributions and cross-tabulations of qualitative variables. Nevertheless, as demonstrated in chapters XV-XIX and as will be demonstrated in chapter XXI, there is increasing interest in making inferences about the relationships among the variables investigated, as opposed to simply describing phenomena.

2. In the present chapter, we shall try to assess the effects of commonly used complex sample designs on the analysis of survey data. We shall attempt to identify cases where the design can influence the analysis. Usually, the sample design has no effect on the analysis when the variables on which the sample design is based are included in the analytical model. Frequently, however, some design variables are not included in the model, either owing to mis-specification or to a lack of interest in those design variables as explanatory factors. This can result in serious biases.

3. There are two basic methodologies we shall discuss for handling data from a complex sample when additional design-related variables are not added to the analysis. The first modifies a classical analytical tool developed for handling data from a simple random sample. The second develops a new analytical tool for the specific complex design.

4. In what follows, we present some examples of the possible effects of sample design on analysis, define a few basic concepts, and discuss the role of design effects in the analysis of complex sample data. Section B describes the two basic approaches to the analysis of complex sample data. In sections C and D, we discuss examples relating to the analysis of continuous and categorical data, respectively. The final section contains a summary and some conclusions. Formal definitions and technical results are given in the annex.

2. Examples of effects (and of non-effect) of sample design on analysis

5. In order to demonstrate the potential effects of sample design on the analysis, we consider the following simple, but illuminating, example (for details, see Nathan and Smith, 1989). Let Y be the variable of interest and X be an auxiliary variable. Assume that the linear regression model $Y_i = \alpha + \beta X_i + \varepsilon_i$, with $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}\left(0, \sigma_i^2\right)$, holds for the population. The model holds as well for any simple random sample selected from the population. Sometimes the assumption of independence across the $\varepsilon_i \mid X_i$ is better suited for the simple random sample than the population from which it was drawn. In a human population, for example, ε_i values may be correlated for different members of the same household, while in a simple random sample of individuals, with a very small probability that more than a single individual is selected per household, the correlation would be negligible.

6. Under simple random sampling the standard estimate of the regression coefficient is unbiased. The display in figure XX.1 plots *Y* against *X* for the total population; the display would look the same for a simple random sample. In the five displays of figures XX.2-XX.6, samples are selected from the population using methods very different from simple random sampling. Consider sample selection based entirely on the value of *X*, for instance, by truncating data points with *X* values beyond (or within) fixed limits as in figures XX.2 and XX.3. It is clear from these figures that the selection has no effect on the estimation of the intercept (α) and slope (β) parameters of the regression (although it may effect the variance of the estimators).

7. Now consider sample selection based on the values of the target variable, *Y*, for instance, by truncating those data points with *Y* values beyond (or within) fixed limits as in figures XX.4-XX.6.

Figure XX.1
No selection

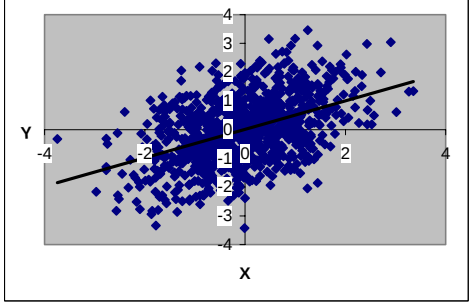


Figure XX.2
Selection on X: XL<X<XU

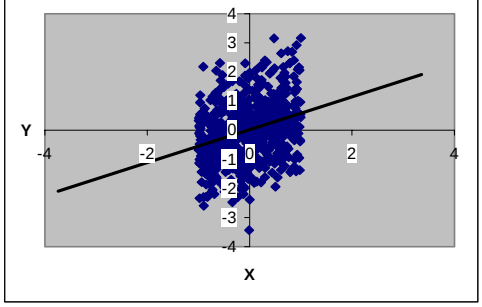


Figure XX.3
Selection on X: X<XL; X>XU

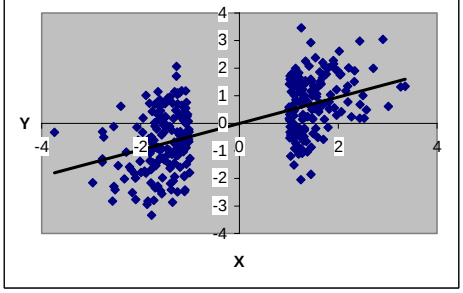


Figure XX.4
Selection on Y: YL<Y<YU

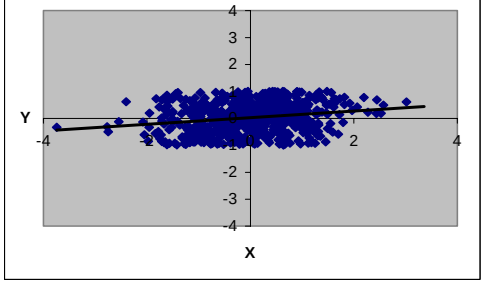


Figure XX.5
Selection on Y: Y<YL;Y>YU

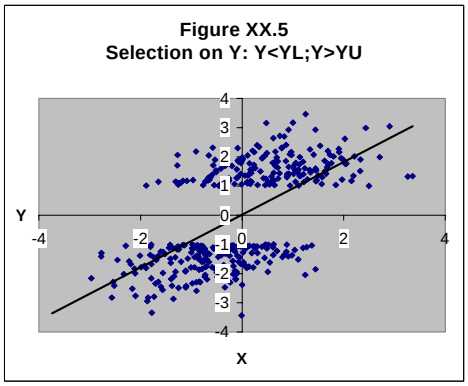
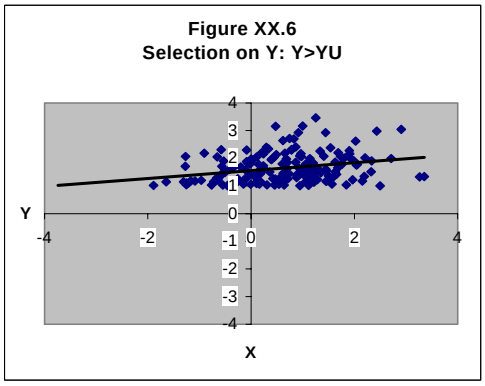


Figure XX.6
Selection on Y: Y>YU



In these cases, it is clear that the estimates of the slopes of the regression become biased. In the last case (figure XX.6), truncation is not symmetric, and the estimate of the intercept is also biased. These examples are extreme, since selection based on the truncation of the dependent variable is rare in sample surveys. It is quite common, however, in experimental or observational studies, such as case control studies in epidemiology or choice-based studies in economics [see, for example, Scott and Wild (1986) and Manski and Lerman (1977)]. Nevertheless, in many cases, samples are selected on the basis of design variables that may be closely related to the dependent variable. Thus, a common sampling procedure, widely used in surveys of establishments and of farms, is to select units with probability proportional to size. The size measure, say, the previous year's production, will obviously be related to the variable of interest when that variable is the current year's production. Standard estimates of model parameters, such as regression coefficients, can be biased when the sample design is ignored.

8. The examples above illustrate the dangers of carrying out an analysis based on complex sample data as if they came from simple random sampling. The examples reveal a need for identifying when it is likely that the design affects the analysis and for taking the design into account when it does.

3. Basic concepts

9. Most sample surveys are designed primarily for descriptive (or enumerative) purposes. They aim at estimating the values of finite population parameters, such as the median household income or the proportion of all adults with AIDS. These are statistics that, in principle, can be measured exactly if the whole population were included in the survey, that is to say, if a census of the population was enumerated rather than a random sample of the population. The standard theory of survey sampling ensures that data from a random sample can be used to provide unbiased estimates of the finite population parameters and of their sampling errors no matter how complex the sample design. This assumes that the sample design is a probability sample design, that is to say, each unit in the population has a known positive probability of being sampled. These classical methods of estimation of finite population parameters are known as design-based (or randomization-based) methods, since all inference is based on the properties of the sample design, via the sample probability distribution. It should be noted, however, that the efficiency of different estimation strategies (a sample design coupled with an estimation formula) can usually be evaluated only when extensive information about the population is available. This is usually not the case in practice. Thus, even classical sampling texts (for example, Cochran, 1977) often rely on models to justify specific methods of sampling or estimation. If the population values follow a simple regression model, for example, then the ratio estimator will be more efficient than the simple expansion estimator, under certain assumptions. Design-based methods are thus often model-assisted, but they are not model-based (or model-dependent): with model-assisted methods of sample design and estimation of descriptive statistics, model assumptions are not required for the nearly unbiased estimation of finite population parameters.

10. The model-based (or prediction theory) approach to sample design and estimation assumes that the finite population values are, in fact, realizations of a super-population distribution based on a hypothetical model with super-population (model) parameters. For further detail and discussion, see Brewer and Mellor (1973), Hansen, Madow and Tepping (1983),

Särndal, Swensson and Wretman (1992) and Valliant, Dorfman and Royall (2000). In contrast with model-assisted estimation methods, the increased efficiency attained by model-based estimation methods relies on the validity of the assumed model. Thus, if there is any doubt about the validity of the model assumptions, the apparent reduction in mean square may not justify the use of purely model-based analysis. A good example of this is demonstrated in Hansen, Madow and Tepping (1983), where assuming no intercept in a regression model, even when the intercept is in fact very close to zero, results in invalid model-based inference.

11. Increasingly, surveys are being used for analytical purposes as well as descriptive ones. Often surveys are designed with their analytical uses in mind. This is because decision makers and researchers are interested in the processes underlying the raw data, in modelling the relationships among the variables investigated. Such analyses obviously require assumptions about models. The aim of an analysis is to confirm the validity of an assumed model and to estimate the model's parameters rather than the parameters of the finite population. Thus analysis is inherently model-based. Inference about model parameters must, practically by definition, be based on the models of interest.

12. It should be pointed out, however, that when the population is very large and the hypothesized model indeed holds, there is in practice very little difference between the model parameters and their finite population counterparts. For instance, if the standard linear regression model $Y_i = \alpha + \beta X_i + \varepsilon_i$, with $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}\left(0, \sigma_i^2\right)$ holds, and the population size is very large, then the value of the standard population regression coefficient, *B* (see annex), will be very close to the value of the model parameter, β, because of to the Central Limit Theorem. Thus, although we shall focus on the estimation of model parameters in what follows, these parameters will sometimes be replaced by their finite-population counterparts. For the sake of simplicity in presentation, most of the examples in what follows will be formulated in terms of univariate distributions (that is to say, a single dependent variable and a single explanatory variable). The extension of the results to the multivariate case is usually straightforward.

13. To summarize, hypothetical models are an integral part of statistical analysis. In order to analyse data from sample surveys, the choice of a good model to fit the data is a critical part of the analysis. Researchers and analysts must have a good understanding of the models underlying the processes they wish to study before applying analytical methods. As we shall see in the following sections, the application to data from complex sample designs requires both an understanding of the underlying model and of the way in which the analysis can be affected by the complex design.

4. Design effects and their role in the analysis of complex sample data

14. The topics of design effects and their estimation have been treated extensively in chapters VI and VII, primarily in relation to their role in the design and estimation for enumerative surveys. We shall see in this chapter that they also play an important role in the analysis of data from complex sample surveys. The underlying idea is based on the fact that, assuming that the model assumptions hold, unbiased estimates of the model parameters and estimates of the variances of these estimates are readily available under simple random sampling. These estimates and variance estimates form the basis for testing hypotheses relating to the model

parameters. For example, under simple random sampling and assuming the simple regression model $Y_i = \alpha + \beta X_i + \varepsilon_i$, with $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}(0, \sigma_i^2)$, the ordinary least squares sample estimator, b, is an unbiased estimator of β, and an unbiased estimator of its variance, $v(b)$, is available (see annex). The standard test of the null hypothesis that β=0 is then based on the test statistic $b/\sqrt{v(b)}$, by invoking the Central Limit Theorem. When the sample design is a complex one, for example, a stratified cluster sample, the estimator b remains model-unbiased, if the regression model holds and the sample design does not depend on the values of Y_i (for example, in constrast to the situation in figures XX.4-XX.6). By this we mean that under the specified regression model, the expected value of b is β, where the expectation is with respect to the super-population distribution of the values of Y_i. As we shall see in section C.1, this may no longer be true if the model does not hold. However, even if the model assumptions hold, $v(b)$ is no longer a valid estimator of the model variance of b and must be modified. Often, a direct estimator of the correct model variance can be computed, for example, by using one of the software packages described in chapter XXI, and can be used to replace $v(b)$. If a direct estimator is not available, often a good estimator of the design effect, denoted by $d^2(b)$, can be obtained. This can be used for the modification of the test statistic to replace $v(b)$ by $d^2(b)$ x $v(b)$. We shall present further specific uses of the design effects to modify standard test statistics for other applications below.

B. Basic approaches to the analysis of complex sample data

1. Model specification as the basis of analysis

15. Correct specification of the underlying model is a fundamental step in any analysis. The consequences of model mis-specification - both of exclusion of relevant explanatory variables (or the inclusion of superfluous ones) and of using a wrong functional form (for example, linear instead of quadratic) - are well known and documented in standard texts. They can take the form of biases in the estimation of the model parameters (primarily the exclusion of relevant variables), losses of efficiency (mostly connected with erroneous inclusion of explanatory variables) and altered sizes and power in tests of hypotheses. These effects may be exacerbated when the mis-specification relates directly to sample design variables or to variables correlated with design variables. Nevertheless, it is important to realize that the survey design variables may not be relevant to the research objectives. Moreover, there may be no subject-matter justification for their inclusion in the analytical model.

16. There are two basic approaches to incorporating survey design variables in the model. The aggregated approach considers the model of interest to be at the population level and conceptually independent of the sample design employed to obtain the data. Under this approach, design variables would be included in the model only when they are relevant to the subject-matter analysis. For instance, say we wish to explain the binary variable employed/unemployed by the explanatory variable years of education, irrespective of geographical location. The sample is stratified, say, by geographical regions, for which different models may be relevant. This would be the case even had simple random sampling been used. As a result, the stratification can be included in the model (see the disaggregated approach discussed below) to reflect regional variations in the relationships among the model variables. If, by contrast, the stratification and

the sample allocation to strata were carried out simply for operational reasons (convenience or cost), the sample weights would likely not be relevant to the population model. The incorporation of sampling weights into an analysis otherwise free of stratum effects will lead to some loss of efficiency. Nevertheless, it is susceptible to the easy interpretation of a model free from stratum effects, while being robust to model failure if some of the ignored stratum effects really exist.

17. The disaggregated approach extends the analyst's model to include not only the survey variables of interest but also variables used in the survey design and those relating to the structure of the population reflected in the design. Design variables relating to the stratification and clustering are included in the model to reflect the complex structure of the population. For instance, in the previous example, the model would contain a different set of coefficients (both an intercept and a slope) for each geographical stratum. Inference under the disaggregated approach takes the sample design fully into account, assuming that all design variables are correctly included in the model. The large number of parameters that need to be estimated under this approach may cause difficulties and lead to less accurate estimates when compared with those of more parsimonious aggregated models. The disaggregated approach is appropriate only when the analyst believes that the hypothesized model is relevant for his purposes.

18. The appropriate approach to be used - the aggregated or disaggregated approach - will depend on the analyst's aims. The aggregated approach is more suitable for studying factors affecting the population as a whole and, as such, may be more useful for evaluating national-policy actions. The disaggregated approach is more suitable for studying micro-effects and the effects of local and sector-specific decision-making. For further examples and discussion, see Skinner, Holt and Smith (1989) and Chambers and Skinner (2003).

2. Possible relationships between the model and sample design: informative and uninformative designs

19. It is important to draw a distinction between informative and non-informative sample designs when analysing complex survey data. Once a model has been hypothesized, the analyst must consider whether, after conditioning on the model covariates, the sample selection probabilities are related to the values of the response variable. A sampling process is informative if the joint conditional model distribution of the observations for the sample, given the values of the covariates in the model, differs from their conditional distribution in the population. Only when these distributions are identical is the sample design non-informative (or ignorable), in which case standard analytical methods can be employed as if the observations came from a simple random sample. When the sample design is informative, the model holding for the sample data is different from the population model. Ignoring the sampling process in such a case may yield biased point estimators and distort the analysis just as when variables are excluded from the model in a conventional analysis. Note that the correct inclusion of design-related variables in the model will ensure that the design is non-informative.

20. There are two major problems with including all design-related variables in a model. First, exactly which variables were used in the design may not be known or, if known, their values may not be available. Even when the design variables are identified and measured, the

analyst may not know the exact form of the relationship (for example, linear or exponential) between them and the variable of interest. For instance, if the design is a stratified one, then the possibility that a regression relationship has different slopes and intercepts for the different strata will need to be checked.

21. Second, when design variables are correctly included in the model, resulting estimates may be of little value to the analyst, since the variables added are not of intrinsic subject-matter interest (recall the discussion on aggregated and disaggregated analysis). This implies that the effect of a complex sample design on analysis cannot always be dealt with solely by modifying the underlying model. In what follows, we shall consider both how to modify standard analytical methods to take a complex design into account and how to construct special design-specific methods of estimation and analysis.

3. Problems in the use of standard software analysis packages for analysis of complex samples

22. The nearly universal use of standard software for statistical analysis has led to widespread abuse of sound statistical practice. This abuse is frequently exacerbated when analysing complex sample survey data.

23. The advantages of statistical software in facilitating analysis unfortunately come with the possibilities for performing analysis without any basic understanding of the underlying principles involved. This has become a serious problem in quantitative work, especially in the social sciences. This problem is compounded by the fact that most commonly available software treats data as if they resulted from simple random sampling. As pointed out previously, this can lead to seriously biased inference when the design is informative. Nevertheless, with due care, standardized software can often be adapted to approximately capture or account for the effect of a complex design. In particular the *SURVEYREG* procedure in the latest versions of SAS® (versions 8 and 9) features regression analysis that takes the sample design into account in ways similar to those described below [see An and Watts (2001)].

24. For instance, consider the linear (heteroscedastic) regression model defined by: $Y_i = \alpha + \beta X_i + \varepsilon_i$, with $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}\left(0, \sigma_i^2\right)$. Standard computer programmes ordinarily compute b, the ordinary least squares (OLS) estimator of β, or the generalized least squares (GLS) estimator, b_G, where sums and products are weighted by the reciprocals of σ_i^2, whose values (or relative values) are assumed to be known (see annex). Both of these are unbiased estimators for the parameter β, if the model holds, although b_G is a more efficient estimator in the heteroscedastic case. The standard programmes also provide estimators of the variances of the OLS estimator, $v(b)$, and of the GLS estimator, $v(b_G)$, which are each model-unbiased under the appropriate model (the homoscedastic model in the case of $v(b)$).

25. In many cases, there may be doubt about the validity of the model, so that instead of estimating β, it may be more appropriate to estimate the finite population counterpart of β, which we denote as B (see annex). Although b (the OLS estimator) is a model-unbiased estimator for $\beta \approx B$, it is not in general design-unbiased. The sample-weighted (Horvitz-Thompson) estimator,

b_W, with cross-products and squares weighted by the reciprocals of the inclusion probabilities, is both design-consistent and model-unbiased, under appropriate conditions. Furthermore b_W can be obtained from the weighted regression options of many standard programmes by using the values of w_i as the weights. Alternatively b_W can be obtained by unweighted regression of the transformed variables $Y_i/\sqrt{\pi_i}$, $X_i/\sqrt{\pi_i}$ with the intercept replaced by $1/\sqrt{\pi_i}$. It must be emphasized, however, that under both these alternatives, the estimates of the variance-covariance matrix reported by most standard programmes are incorrect -- both as estimators for design mean squared error and as estimators for model variance -- except in unusual circumstances.

26. In summary, the use of standard software programmes that do not take into account complex survey design should be avoided unless it can be determined that the complex design does not have a serious effect on estimation. This can often be achieved through the suitable use of standard software. See the example in section C.2. The use of software packages specifically dealing with complex sample designs is recommended (see chap. XXI).

C. Regression analysis and linear models

1. Effect of design variables not in the model and weighted regression estimators

27. Regression analysis and linear modelling are very common applications where standard models developed for simple random samples are routinely applied to data from complex sample surveys. As already pointed out, this can often lead to erroneous analyses and conclusions. A key source of protection against error is the identification of variables determining or influencing the sample design, so that they can be included in the model. As we have seen, however, even when these variables are identified, their inclusion in the model may not be warranted from the subject-matter point of view. In the present subsection, we will study the effects on traditional estimators of not including design variables in the model and investigate the possibilities for modifying these estimators so as to take the complex design into account. For ease of exposition, we consider the case of a single dependent variable, Y (denoted in this section for technical reasons by X_1), where the model of interest has a single explanatory variable (X_2), and there is a single design variable (X_3). The model of interest is therefore $\mathrm{E}(X_1) = \mu_1 + \beta_{12}(X_2 - \mu_2)$, where β_{12} is the parameter of interest, rather than the full model, which includes the design variable, X_3. See the annex for the formulae used in this subsection.

28. Under fairly general conditions specified in Nathan and Holt (1980), the standard OLS estimator for β_{12}, $b_{12} = s_{12}/s_2^2$, can be both (model-) biased conditional on X_3 and on the sample, S**,** and biased unconditionally. Expressions for the conditional model expectation and its unconditional (joint model and design) expectation show that, in general, b_{12} is asymptotically biased, unless $\rho_{23,}$ the correlation between X_2 and $X_{3,}$ is zero or if the simple sample variance of X_3 is an unbiased estimator for its true variance. It can be shown that this second condition does hold asymptotically for a large number of equal probability (epsem) sample designs, but rarely for unequal probability designs (for example, non-proportional stratified sample designs).

29. A corrected, asymptotically unbiased estimator based on the maximum likelihood estimator under normality, $\hat{\beta}_{12}$, can be used instead of the OLS estimator. Expressions for the variances of b_{12} and of $\hat{\beta}_{12}$ are given in Nathan and Holt (1980). It should be noted that the usual estimator for the variance of b_{12}, $v(b_{12})$, may not be nearly unbiased, even when b_{12} is a consistent estimator for β_{12}. This can happen when the ε_i's are not independent and identically distributed among the observations in the sample.

30. Neither the estimator b_{12} nor $\hat{\beta}_{12}$ depends on the sample design, though their properties do. Information on the sample design may be useful for improving these estimators, either when information on values of the design variable X_3 is not available for the whole population (so that S_3^2 cannot be used for estimation) or when the analysts wishes to ensure robustness to departures from the model. This can be done by using sample-weighted estimators based on Horvitz-Thompson estimation for each of the components of the unweighted estimators. One can replace the unweighted sample moments by their weighted versions in the expressions for b_{12} and for $\hat{\beta}_{12}$ to obtain the weighed estimators, b_{12}^* and $\hat{\beta}_{12}^*$.

31. Note that b_{12}^* can be used when the population variance of X_3, S_3^2, is unknown, but that $\hat{\beta}_{12}^*$ cannot be used in that situation. It is easily seen that, under fairly general conditions, both of these estimators are design-consistent estimators of the finite population parameter, B_{12}.

32. Empirical comparisons of the performance of these four estimators were made by Nathan and Holt (1980) for a population of N = 3,850 farms, about which data on crop land (X_1), total acreage (X_2), and total value of produce in previous year (X_3), were available. Farms were stratified on the basis of the X_3 values resulting in six strata of sizes 563, 584, 854, 998, 696 and 155. The following six sample designs were used to select samples of size n = 400 (see table XX.1):

(A) Simple random sampling;

(B) Proportional stratified simple random sampling;

(C) Fixed size stratified simple random sampling;

(D) Stratified simple random sampling with higher-than- proportional allocation to strata with high X_3 values (25, 30, 60, 80, 130, 75);

(E) Stratified simple random sampling with U-shaped allocation (100, 80, 20, 20, 80, 100).

.

Table XX.1. Bias and Mean square of ordinary least squares estimator and variances of unbiased estimators for population of 3,850 farms using various survey designs

Survey design	$E(b_{12})-\beta_{12}$	MSE(b_{12})	$V(\hat{\beta}_{12})$	$V(b_{12}^{*})$	$V(\hat{\beta}_{12}^{*})$
A	0-000	0-000214	0-000197	0-000226	0-000197
B	0-000	0-000200	0-000198	0-000222	0-000196
C	0-031	0-001102	0-000160	0-000222	0-000196
D	0-027	0-000879	0-000163	0-000220	0-000195
E	0-042	0-001877	0-000152	0-000225	0-000196

Source: Nathan and Holt (1980); table 1.

33. The results demonstrate the bias of b_{12} for the non-epsem designs (C,D,E), whereas the other estimators are either design-consistent or model-consistent (or both). They also demonstrate the advantage of $\hat{\beta}_{12}$ over the weighted estimators for all the designs considered. This holds even though the full model assumptions appear not to hold for the population. When S_3^2 is unknown, however, the less efficient, but still consistent, b_{12}^{*}, is a reasonable estimator.

34. To summarize, when data are based on unequal probability designs, it is worthwhile to consider both weighted and unweighted maximum likelihood estimators, rather than the simple OLS estimators. The unweighted estimator seems to be more efficient. In many applications, however, the analyst will not have the information needed to compute maximum likelihood estimators; and less efficient, but consistent, sample-weighted estimators are appropriate and, indeed, are routinely used [see Korn and Graubard (1999)].

2. Testing for the effect of the design on regression analysis

35. Many analysts prefer to use simple weighted or unweighted estimators of regression coefficients, which can be obtained from standard packages, rather than the modified estimators proposed in section C.1. We have seen that the simple OLS estimator is consistent when the design is non-informative or the effect of the design is negligible, but that its weighted counterpart is preferable when that is not the case. DuMouchel and Duncan (1983) proposed a simple test, based on standard software packages, for deciding whether weights should be used when based on data from a non-clustered sample. Consider the univariate case with a single explanatory variable. The extension to the multivariate case is straightforward. Letting $\hat{\Delta} = b_W - b$, the goal is to test the hypothesis: $\Delta = \mathrm{E}\left(\hat{\Delta}\right) = 0$. DuMouchel and Duncan showed that the test for $\Delta = 0$ is the same as the test for $\gamma = 0$, under the model $Y_i = \alpha + \beta X_i + \gamma Z_i + \varepsilon_i$, where $Z_i = w_i X_i$ and $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}\left(0, \sigma^2\right)$. The authors gave a numerical example for the multivariate case involving a subset of data from the University of Michigan Survey Research Center's Panel Study of Income Dynamics. The sample of 658 individuals was selected with varying probabilities, resulting in weights ranging from 1 to 83. The final model used to explain educational attainment included a constant and 17 explanatory variables, such as parents'

education, income, age, race employment and interactions. The following analysis of variance (ANOVA) table is obtained:

Table XX.2. ANOVA table comparing weighted and unweighted regressions

Source	df	Sum of squares	Mean square	*F*	Significance
Regression	17	730.6	43.0	17.35	<.0001
Weights	18	43.3	2.5	.97	.494
Error	622	1542.2	2.5		
Total	657	2315.9			

36. Taken together, the 18 variables corresponding to Z_i (the 17 explanatory variable and the constant, each multiplied by w_i) have an F value of .97 and a significance level of only .494. Thus an unweighted regression is justified, even though it may entail some loss of power.

37. In general, analysts may be equally concerned about accepting the null hypothesis that $\Delta = \mathrm{E}(\hat{\Delta}) = 0$, when it is false, as about rejecting it when it is true. As a result, they may decide to conduct a weighted analysis (with the appropriate software) or develop a less parsimonious model when the significance level is considerable larger that the standard .05. In the example above, the significance level is very close to 0.5, which suggests that the weights can be ignored. In an earlier version of the model, however, the significance level for the Z_i was .056, at which point, DuMouchel and Duncan added some interaction terms. The final results are those displayed in the table.

38. The DuMouchel-Duncan test described above assumes that the ε_i's are independent and identically distributed. Often survey data come from multistage sample designs. When the ε_i values of observations from the same sample cluster are correlated or when the observations have an unknown heteroscedasticity regardless of the design, this test in inappropriate. Nevertheless, an analyst may feel that using sample weights adds unneeded variance to the resulting estimates. A Wald test along the lines proposed by Fuller (1984) can be employed. In practice this involves using software like SAS/SURVEYREG and entering each data point twice, once with the sample weight set to 1 and once with the sample weight set to the actual weight.

39. Pfeffermann and Sverchkov (1999) proposed an alternative set of weights for use when the linear model is correct and the errors are independent and identically distributed but the sample design is informative. The test described above can be used to assess their weights relative to the sample weights. For further discussion of the role of sampling weights when modelling survey data, see Pfeffermann (1993) and Korn and Graubard (1999).

3. Multilevel models under informative sample design

40. Recently, there has been increased use of multilevel models for the analysis of data from populations with complex hierarchic structures. For instance, in most household surveys, individuals nested within households are the units of investigation, and there is interest in both

relationships among the individuals and among the households. Similar hierarchic structures exist for surveys of pupils within schools and employees within establishments.

41. The usual single-level linear models can easily be extended to take a hierarchy into account by using mixed (random and fixed effects) models with an error structure that reflects the hierarchic configuration. For example, what is known as the random intercept model can be formulated (for a single explanatory variable) as follows:

$$y_{ij} = \beta_{oi} + \beta x_{ij} + \varepsilon_{ij}; \quad \varepsilon_{ij} \mid x_{ij} \sim \mathrm{N}\left(0, \sigma_{\varepsilon}^2\right), \quad (i = 1, \ldots, N; j = 1, \ldots, M_i)$$

where y_{ij} is the outcome variable for first-level unit j (say, individual) within second-level unit i (say, household), x_{ij} is a known explanatory variable and β an unknown parameter. The intercept, β_{oi}, is here a random variable, which is further modelled as

$$\beta_{oi} = \alpha + \gamma z_i + u_i; \quad u_i \mid z_i \sim \mathrm{N}\left(0, \sigma_u^2\right), \quad (i = 1, \ldots, N)$$

where z_i is a known second-level unit explanatory variable and α and γ are unknown parameters.

42. Under simple random sampling, models of this type can be analysed using straightforward extensions of single-level linear model theory. Unfortunately, closed forms for the estimates of the model parameters ($\alpha, \beta, \gamma, \sigma_{\varepsilon}^2, \sigma_u^2$, for the model above) are not available. Instead, an iterative procedure, Iterated Generalized Least Squares (IGLS), is used. It produces estimates that converge to maximum-likelihood solutions. Thus, the closed-form methods of adapting weighted least squares to take sample design into account cannot be employed in this case. A sample-weighted version of IGLS (PWIGLS), which weights the first- and second-level estimating equations by appropriate weights based on the selection probabilities, has been developed to obtain consistent estimators of the parameters [see Pfeffermann and others (1998) for the details].

43. More recently, Pfeffermann, Moura and Silva (2001) have proposed a model-dependent (purely model-based) approach for multilevel analysis that accounts for informative sampling. The idea behind the proposed approach is to extract the hierarchic model holding for the sample data as a function of the population model and the first-order sample inclusion probabilities, and then fit the sample model using classical estimation techniques. The selection probabilities become additional outcome variables to be modelled and to thereby strengthen the performance of the estimators. Further detail is beyond the scope of this chapter but can be found in Pfeffermann, Moura and Silva (2001). A simulation experiment that follows closely the design of the Rio de Janeiro Basic Education Evaluation study of 1996 indicates that the results of applying the proposed method are promising.

D. Categorical data analysis

1. Modifications to chi-square tests for tests of goodness of fit and of independence

44. Initial attempts to assess the effects of complex sample design on the analysis of categorical data (data such that each point falls into one of a finite number of categories or cells) concerned modifications to the chi-squared tests that are commonly employed either to assess the goodness of fit between the distribution of a single categorical variable and a hypothesized distribution or to test for independence between two categorical variables. Although several modified chi-squared tests have been proposed in the literature for data from proportionate stratified simple random sampling, the effect of that design is usually very small in practice. Thus, in a study of modified chi-square statistics in eight data sets from proportionate stratified samples in Israel, presented in table XX.3 [from Kish and Frankel (1974)], none of the final iteration statistics differed by more than 4 per cent from those that would have been obtained under simple random sampling (SRS) assumptions, and most differed by less than 1 per cent.

Table XX.3. Ratios of three iterated chi-squared tests to SRS tests [a/]

Data set	*No. of strata*	*Row x columns*	*Sample size*	*Nathan's three tests*					
					First iteration			*Last iteration*	
				X^2	χ_1^2	*G*	X^2	χ_1^2	*G*
1	4	3x3	845	1·028	0·992	1·017	1·004	1·004	1·005
2	4	3x3	821	1·088	0·963	1·043	0·999	1·003	1·001
3	4	3x3	491	1·740	0·707	1·406	1·011	1·001	1·009
4	4	3x3	2 528	1·095	0·959	1·049	1·003	1·005	1·003
5	6	2x4	500	1·079	0·967	1·040	1·004	1·003	1·003
6	3	2x2	120	1·013	0·967	1·009	1·008	0·969	1·007
7	5	2x2	269	1·076	0·989	1·043	1·011	1·015	1·011
8	2	2x4	81	1·368	0·889	1·186	1·029	1·037	1·029

Source: Adapted from data in Nathan (1972).

a/ Eight contingency tables based on proportionate stratified samples from Israel: Nos. 1-4 of savings, No. 5 of attitudes, No. 6 of hospital data, No. 7 of poultry medicament and No. 8 of perception experiments.

45. Although the impact of the design on categorical data analysis under proportionate stratified simple random sampling is usually small, this is often not the case under clustered sampling, as was demonstrated in a seminal paper by Rao and Scott (1981). When testing for goodness of fit, they showed that, under the null hypothesis, the usual chi-square statistic, X^2, is distributed asymptotically as a weighted sum of k-1 independent χ_1^2 (that is to say, squared normal) random variables. The weights are the eigenvalues of a matrix D (see annex). The matrix can be viewed as a natural multivariate extension of the design effect for univariate statistics (see chaps. VI and VII). Its eigenvalues, λ_{0i}^2, are termed generalized design effects and can be shown to be the design effects for certain linear combinations of the design effects, d_i^2, of $\hat{p}_i$ (= the estimated proportion of the population in category i). A modified chi-square statistic,

X_C^2, can be obtained by dividing the standard X^2 statistics by the average of estimates of these generalized design effects, denoted by $\hat{\lambda}^2$. This modification requires knowledge only about the design effects of the cell estimates. Although X_C^2 does not have an asymptotic χ_{k-1}^2 distribution under the null hypothesis, it has the same asymptotic expected value as χ_{k-1}^2 (that is to say, *k*-1), but with a larger variance. It turns out that X_C^2 can be used empirically to test goodness-of-fit by comparing the value of this statistic to the critical value of χ_{k-1}^2. This can be seen in table XX.4 [from Rao and Scott (1981)], which displays the true sizes of tests based on X^2 and on X_C^2, respectively, for six items of data from the 1971 General Household Survey of the United Kingdom of Great Britain and Northern Ireland. The survey had a stratified three-stage design.

Table XX.4. Estimated asymptotic sizes of tests based on X^2 and on X_C^2 for selected items from the 1971 General Household Survey of the United Kingdom of Great Britain and Northern Ireland; nominal size is .05

Variable	*k*	*m*	$\hat{\lambda}_{\cdot}^2$	*Size* (X^2)	*Size* (X_C^2)
G1: Age of building	3	33.1	3.42	.41	.05
G2: Ownership type	3	33.4	2.54	.37	.06
G3: Type of accommodation	4	27.7	2.17	.30	.06
G4: Number of rooms	10	34.6	1.19	.14	.06
G5: Household gross weekly income	6	26.6	1.14	.10	.06
G6: Age of head of household	3	34.6	1.26	.10	.05

The results show that the use of the standard chi-square statistic, X^2, can be very misleading, whereas the modified statistic, X_C^2, performs very well.

46. Similar results hold when testing for independence in a two-way contingency table. For a contingency table with *r* columns and *c* rows, the null hypothesis of interest is $H_0 : h_{ij} = p_{ij} - p_{i+}p_{+j} = 0 \quad (i = 1,\ldots,r; j = 1,\ldots,c)$, where p_{ij} is the population proportion in the (*i,j*)th cell and p_{i+}, p_{+j} are the marginal totals. The usual chi-square statistic for data from a simple random sample, X_I^2, is asymptotically distributed as chi-square with *b*=(*r*-1)(*c*-1) degrees of freedom under the null hypothesis. This need not be true when the sample design is complex. In fact, the asymptotic distribution of X_I^2 is a weighted sum of *b* independent χ_1^2 random variables, similar to the case when testing for goodness-of-fit.

47. A generalized Wald statistic can be constructed based on estimating the complete variance-covariance matrix of the estimates, $\hat{h}_{ij} = \hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j}$ [see details in Rao and Scott (1981)]. Fortunately, a first-order correction, which requires only estimates of the variances of $\hat{h}_{ij}$, $\hat{\mathrm{v}}(\hat{h}_{ij})$, seems to be an adequate approximation. The modified statistic is defined as $X_{I(C)}^2 = X_I^2 / \hat{\delta}_{\cdot}^2$, where $\hat{\delta}_{\cdot}^2$ is a weighted average of the estimated design effects of $\hat{h}_{ij}$. When

estimates of these design effects are unavailable, as often happens with secondary analysis of published data, an alternative modification can be obtained by replacing $\hat{\delta}_{\cdot}^{2}$ by $\hat{\lambda}^{2}$, a weighted average of the estimated design effects of the cell proportions, $\hat{d}_{ij}^{2}$. The adequacy of these approximations depends to a large extent on the relative variance of the design effects. A second-order correction is available for use when the relative variance is large.

48. Empirical results for 15 two-way contingency tables based on data from the General Household Survey of the United Kingdom of Great Britain and Northern Ireland are given in table XX.5 [from Rao and Scott (1981)]. They indicate again that: (a) the uncorrected chi-square statistic, X_{I}^{2}, performs very poorly in many cases; (b) the corrected statistic, $X_{I(C)}^{2}$, based on $\hat{\delta}_{\cdot}^{2}$ attains the nominal size almost exactly; and (3) the corrected statistic based on $\hat{\lambda}^{2}$ errs on the conservative side.

Table XX.5. Estimated asymptotic sizes of tests based on X_{I}^{2}, $X_{I}^{2}/\hat{\delta}_{\cdot}^{2}$, and on $X_{I}^{2}/\hat{\lambda}_{\cdot}^{2}$ for cross-classification of selected variables from the 1971 General Household Survey of the United Kingdom of Great Britain and Northern Ireland; nominal size is .05

Cross Classification	*r + c*	$\hat{\delta}_{\cdot}^{2}$	$\hat{\lambda}^{2}$	*Size* (X_{I}^{2})	*Size* $(X_{I}^{2}/\hat{\delta}_{\cdot}^{2})$	*Size* $(X_{I}^{2}/\hat{\lambda}_{\cdot}^{2})$
G1 X G2	2 X 2	1.99	3.18	.16	.05	.01
G1 X G3	2 X 3	1.97	2.36	.22	.05	.03
G1 X G4	2 X 3	1.24	1.98	.09	.05	.01
G1 X G5	2 X 6	.91	1.23	.04	.05	.02
G1 X G6	2 X 3	.97	1.75	.05	.05	.01
G2 X G3	2 X 3	1.94	2.49	.21	.05	.03
G2 X G4	2 X 3	1.41	1.86	.12	.05	.02
G2 X G5	2 X 6	1.02	1.18	.06	.05	.03
G2 X G6	2 X 3	1.13	1.61	.08	.05	.02
G3 X G4	3 X 3	1.26	1.72	.11	.05	.01
G3 X G5	3 X 6	.93	1.14	.03	.05	.02
G3 X G6	3 X 3	.96	1.51	.05	.05	.01
G4 X G5	3 X 6	.94	1.05	.05	.05	.03
G4 X G6	3 X 3	.93	1.21	.04	.05	.02
G5 X G6	6 X 3	.85	.94	.03	.05	.04

2. Generalizations for log-linear models

49. The results above for two-way tables have been generalized by Rao and Scott (1984) to the log-linear model used in analysing multi-way tables. Denote by $\boldsymbol{\pi}$ the *T*-vector of population cell proportions, π_{i}, in the multi-way table with $\sum_{1}^{T}\pi_{t}=1$ (for example, $T = 4$ for a 2 x 2 table). Denote the saturated log-linear model (that which includes all possible interactions)

as M_1. We consider testing of the hypothesis that a reduced nested sub-model, M_2, is sufficient. Let $\hat{\boldsymbol{\pi}}$ be the pseudo maximum likelihood estimator of $\boldsymbol{\pi}$ under M_1. This is defined as the solution of the sample estimate of the census likelihood equations (those that would have been obtained on the basis of the population data) and is based on a design-consistent estimator of $\boldsymbol{\pi}$ under the survey design (see annex). Similarly, let $\hat{\hat{\boldsymbol{\pi}}}$ be the pseudo maximum likelihood estimator of $\boldsymbol{\pi}$ under M_2. The standard Pearson chi-square statistic for testing H_0, based on $\hat{\boldsymbol{\pi}}$, and on $\hat{\hat{\boldsymbol{\pi}}}$, does not usually have an asymptotic chi-square distribution under the null hypothesis. This case is similar to that of the two-way table, inasmuch as the standard Pearson chi-square statistic's asymptotic distribution is a weighted sum of u independent χ_1^2 random variables with weights δ_i^2, which are the eigenvalues of a generalized design effect matrix (see annex for details).

50. In order to take the complex design into account, modified chi-square statistics, $X^2/\hat{\delta}_{\cdot}^2$, $X^2/\hat{\lambda}_{\cdot}^2$ and $X^2/\hat{d}_{\cdot}^2$ are proposed. Here $\hat{\delta}_{\cdot}^2$ is the average of the estimated eigenvalues, $\lambda_{\cdot}^2$ is the average of the estimated design effects of $\mathbf{X}'\hat{\mathbf{p}}$, and $\hat{d}_{\cdot}^2$, is the average of the estimated cell design effects (see annex for details). It should be noted that $\lambda_{\cdot}^2$ and $\hat{d}_{\cdot}^2$ do not depend on the null hypothesis, H_0, whereas $\hat{\delta}_{\cdot}^2$ does. Furthermore, $\hat{d}_{\cdot}^2$ requires knowledge only of the cell design effects as does $\lambda_{\cdot}^2$ when M_1 is the saturated model.

51. In the important case of models admitting explicit solutions for $\hat{\boldsymbol{\pi}}$ and for $\hat{\hat{\boldsymbol{\pi}}}$, Rao and Scott (1984) show that $\hat{\delta}_{\cdot}^2$ can be computed knowing only the cell design effects and those of their marginals. For instance, for the hypothesis of complete independence in a three-way $I\times J\times K$ table, $H_0 : \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$, where $\pi_{i++}, \pi_{+j+}, \pi_{++k}$ are the three-way marginals, the value of $\hat{\delta}_{\cdot}^2$ can be calculated explicitly as a function of the estimates of the design effects of the three-way marginals and of the estimates of the cell design effects.

52. The relative performances of these modified statistics and of the unmodified one are given in table XX.6 [from Rao and Scott (1984)] based on a 2×5×4 table from the Canada Health Survey 1978-1979. The variables are gender (I=2), drug use (J=5) and age group (K=4). The hypotheses tested were: (a) complete independence (denoted by $\overline{1}\otimes\overline{2}\otimes\overline{3}$); (b) partial independence (for example, $\pi_{ijk} = \pi_{i++}\pi_{+jk} \Leftrightarrow \overline{1}\otimes\overline{2}\,\overline{3}$) and, similarly, $\left(\overline{2}\otimes\overline{1}\,\overline{3}\right)$ and $\left(\overline{3}\otimes\overline{1}\,\overline{2}\right)$; (c) conditional independence (for example, $\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k} \Leftrightarrow \overline{1}\otimes\overline{2}\,|\,\overline{3}$) and, similarly, $\left(\overline{2}\otimes\overline{1}\,|\,\overline{3}\right)$and $\left(\overline{3}\otimes\overline{1}\,|\,\overline{2}\right)$. The design was complex, involving stratification and multistage sampling. Moreover, post-stratification was used to improve the estimates.

Table XX.6. Estimated asymptotic significance levels (SL) of X^2 and the corrected statistics $X^2/\hat{\delta}_{\cdot}^2$, $X^2/\hat{\lambda}_{\cdot}^2$, $X^2/\hat{d}_{\cdot}^2$. : 2 x 5 x 4 table and nominal significance level α = 0.05

	Hypothesis						
	(a)		(b)		(c)		
	$\bar{1}\otimes\bar{2}\otimes\bar{3}$	$\bar{1}\otimes\bar{2}\bar{3}$	$\bar{2}\otimes\bar{1}\bar{3}$	$\bar{3}\otimes\bar{1}\bar{2}$	$\bar{1}\otimes\bar{2}\vert\bar{3}$	$\bar{1}\otimes\bar{3}\vert\bar{2}$	$\bar{2}\otimes\bar{3}\vert\bar{1}$
SL (X^2)	0.72	0.33	0.76	0.72	0.43	0.30	0.78
SL ($X^2/\hat{\delta}_{\cdot}^2$)	0.16	0.11	0.14	0.13	0.095	0.11	0.12
SL ($X^2/\hat{\lambda}_{\cdot}^2$)	0.34	0.056	0.39	0.32	0.098	0.06	0.39
SL ($X^2/\hat{d}_{\cdot}^2$)	0.34	0.054	0.39	0.32	0.097	0.06	0.39
$\hat{\delta}_{\cdot}$	2.09	1.40	2.25	2.09	1.63	1.39	2.31
C.V. ($\hat{\delta}_i$)	1.54	1.02	1.37	1.27	0.86	1.05	1.11

53. The comparisons relate the actual significance levels (SL) to the desired nominal level, α = 0.05. The results again show unacceptably high values of SL for the uncorrected statistic. The modified statistics $X^2/\hat{\lambda}_{\cdot}^2$ and $X^2/\hat{d}_{\cdot}^2$, which do not depend on the hypothesis, perform very similarly with values of SL ranging from 0.06 to 0.39, which are too high. The modification based on marginal and cell design effects, $X^2/\hat{\delta}_{\cdot}^2$, has a more stable performance, with SL values ranging from 0.095 to 0.16, all above the nominal level, probably owing to the large coefficient of variation (CV) of the $\hat{\delta}_i^2$'s .

54. To summarize, correction methods are available for standard chi-squared test statistics in categorical data analysis. These corrections are often necessary for valid analysis, given a clustered sample, and can be applied with relative ease using estimated marginal and cell design effects. Details on available software to deal with the effects of complex sample design on chi-squared tests and logistic regression can be found in chapter XXI.

E. Summary and conclusions

55. In this chapter, we have illustrated methods for assessing the effects of commonly used complex sample designs on the analysis of survey data. The material is intended primarily as an introductory exposition of the issues rather than a prescriptive one. The assessment and treatment of the effects of sample design on analysis can be difficult and are not amenable to the formulation of easily applicable "how-to-do" rules. As we have shown, different problems can have different (or several different) possible methods of resolution. These are highly dependent on the hypothesized model and the validity of its underlying assumptions, on various aspects of the sample design (for example, unequal selection probabilities, clustering, etc.) and on the type of analysis contemplated. Knowledge about the relationship between the model and the sample design variables is imperative. Unfortunately, this information is not always readily available, which means that assumptions and approximations may have to be used instead.

56. A first and fundamental step in any analysis is the correct specification of the underlying model. This is the responsibility of the subject-matter analyst, although the final identification of the model can and should be based on appropriate statistical techniques. The initial exploratory analysis needed to identify the appropriate model can be conducted using standard graphical and descriptive methods without taking into account the effects of the sample design.

57. Once an initial working model has been hypothesized, it is necessary to determine whether the design has a confounding influence on the analysis. This can be done, for example, with a test comparing the weighted and the unweighted estimates of linear regression coefficients (see sect. C.2). If the complex design needs to be incorporated into the analysis, then one must choose the appropriate method for doing that. The disaggregated approach simply adds variables to the model related to the sample design.

58. In many situations, however, the model cannot be modified to fully reflect the effects of sample design in a meaningful way. When this is the case and the aggregated approach is to be used, two basic methodologies have been proposed to deal with the potential impact of the sample design. One entails the modification of classical analytical tools to take the design into account. This is the method best suited for dealing with categorical data analysis, where standard chi-square statistics can be modified on the basis of generalized design effects. The second approach is the development of appropriately defined analytical tools especially for the design. Sample-weighted estimators and a large-sample Wald statistic have been proposed. A reliable estimator of the covariance matrix is needed before using the Wald statistic. This is not always available in practice.

59. Considerable research into the problems of handling the effects of complex sample design on analysis has produced practical methods, some of which have been described in this chapter. Further research is under way, and many of the existing methods have already been incorporated in new and existing software. Unfortunately, owing to the complexity of the problem, it is unlikely that any overall uniform method will be developed in the future. The available methods and software must be applied with extreme caution. Their application requires both basic knowledge of the underlying theory and thorough understanding and experience in practical model construction.

Annex
Formal definitions and technical results

Regression models (sects. B.2 and B.3)

- Standard linear regression model: $Y_i = \alpha + \beta X_i + \varepsilon_i$, with $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}\left(0, \sigma^2\right)$

- Standard **population** regression coefficient: $$B = \frac{\sum_{i=1}^{N} Y_i \left(X_i - \overline{X}\right)}{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}$$

- Ordinary least squares (OLS) estimator of β: $$b = \frac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}\right)}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

- Unbiased estimator of variance of b: $$\mathrm{v}(b) = \frac{s^2}{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2},$$
  where s^2 is the unbiased estimator of σ^2, based on the variance of the estimated regression residuals.

- General linear (heteroscedastic) regression model: $Y_i = \alpha + \beta X_i + \varepsilon_i$, with $\varepsilon_i \mid X_i \underset{ind}{\sim} \mathrm{N}\left(0, \sigma_i^2\right)$

- Weighted **population** regression coefficient: $$B^* = \frac{\sum_{i=1}^{N} Y_i \left(X_i - \overline{X}_\sigma\right) / \sigma_i^2}{\sum_{i=1}^{N} \left(X_i - \overline{X}_\sigma\right)^2 / \sigma_i^2},$$
  where $$\overline{X}_\sigma = \frac{\sum_{i=1}^{N} X_i / \sigma_i^2}{\sum_{i=1}^{N} 1 / \sigma_i^2}$$

- Generalized least squares (GLS) estimator of β: $b_G = \dfrac{\sum_{i=1}^{n} y_i \left(x_i - \bar{x}_\sigma\right)/\sigma_i^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}_\sigma\right)^2/\sigma_i^2}$,

  where $\bar{x}_\sigma = \dfrac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}$

- Variance of the GLS estimator: $v(b_G) = \dfrac{1}{\sum_{i=1}^{n}\left(x_i - \bar{x}_\sigma\right)^2/\sigma_i^2}$

- Design-weighted (Horvitz-Thompson) estimator: $b^* = \dfrac{\sum_{i=1}^{n} w_i y_i \left(x_i - \bar{x}^*\right)}{\sum_{i=1}^{n} w_i \left(x_i - \bar{x}^*\right)^2}$, where $w_i = \dfrac{1/\pi_i}{\sum_{k=1}^{n} 1/\pi_k}$,

  π_i is the inclusion probability, and $\bar{x}^* = \sum_{i=1}^{n} w_i x_i$

Effect of exclusion of design variables (sect. C.1)

- Model of interest: $\mathrm{E}(X_1) = \mu_1 + \beta_{12}(X_2 - \mu_2)$, where β_{12} is the parameter of interest.

- Full model with design variable, X_3: $\mathrm{E}(X_1) = \mu_1 + \beta_{12\cdot 3}(X_2 - \mu_2) + \beta_{13\cdot 2}(X_3 - \mu_3)$

- Notation:

  - Usual notation for multivariate analysis, for example, $\beta_{12\cdot 3}$ denotes the conditional regression coefficient of X_1 on X_2, given X_3

- First and second population moments of X_i: $\overline{X}_i = \frac{1}{N}\sum_{j=1}^{N} X_{ij}$;

  $S_i^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(X_{ij} - \overline{X}_i\right)^2$;

- $S_{ik} = \frac{1}{N-1}\sum_{j=1}^{N}\left(X_{ij} - \overline{X}_i\right)\left(X_{kj} - \overline{X}_i\right)$;

- Sample moments: $\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$;

  $s_{ik} = \frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}_i\right)\left(x_{kj} - \bar{x}_k\right)$; $s_i^2 = s_{ii}$, where we assume a sample, S, of fixed size *n,* selected by any design, possibly dependent on X_3.

- Standard OLS estimator of β_{12}: $b_{12} = \frac{s_{12}}{s_2^2}$

- Asymptotic model conditional expectation of b_{12}:

$$\mathrm{E}_M\left(b_{12}\middle|X_3,\mathbf{S}\right) = \frac{\beta_{12} + \beta_{13}\beta_{23}\left(s_3^2/\sigma_3^2 - 1\right)}{1 + \rho_{23}^2\left(s_3^2/\sigma_3^2 - 1\right)} + \mathrm{O}\left(n^{-1}\right)$$

- Unconditional (joint model and design) expectation:

$$\mathrm{E}_M\left(b_{12}\right) = \beta_{12} + \frac{\sigma_1}{\sigma_2}\frac{\rho_{13\cdot 2}\rho_{23}\left[\left(1-\rho_{12}^2\right)\left(1-\rho_{23}^2\right)\right]^{\frac{1}{2}}(Q-1)}{1+\rho_{23}^2(Q-1)} + \mathrm{O}\left(n^{-1}\right), \text{ where } Q = \mathrm{E}\left(s_3^2\right)/\sigma_3^2$$

  - OLS estimator, $b_{12,}$ is asymptotically biased, even unconditionally, unless $\rho_{23} = 0$ or $\mathrm{E}\left(s_3^2\right) = \sigma_3^2$, that is to say, Q=1

  - Corrected asymptotically unbiased estimator (maximum likelihood estimator (MLE) under normality):

$$\hat{\beta}_{12} = \frac{s_{12} + \left(s_{13}s_{23}/s_3^2\right)\left(S_3^2/s_3^2 - 1\right)}{s_3^2 + \left(s_{23}^2/s_3^2\right)\left(S_3^2/s_3^2 - 1\right)}$$

  - Weighted estimators: $b_{12}^* = \frac{s_{12}^*}{s_2^{*2}}$; $\hat{\beta}_{12}^* = \frac{s_{12}^* + \left(s_{13}^* s_{23}^*/s_3^{*2}\right)\left(S_3^2/s_3^{*2} - 1\right)}{s_3^{*2} + \left(s_{23}^{*2}/s_3^{*2}\right)\left(S_3^2/s_3^{*2} - 1\right)}$, where

$$\bar{x}_i^* = \sum_{j=1}^{n} \frac{x_{ij}}{N\pi_j};\ s_{ik}^* = \sum_{j=1}^{n} \frac{x_{ij}x_{kj}}{N\pi_j} - \bar{x}_i^*\bar{x}_k^*;\ s_i^{*2} = s_{ii}^*, \text{ and}$$

$\pi_j = \mathrm{p}\left(j \in \mathbf{S} \middle| X_{3j}\right)$ are the sample inclusion probabilities. Note that $\sum_{j=1}^{n} \frac{1}{N\pi_j} = 1$ under stratified simple random sampling, which is the design we are assuming here. For more general designs, $N\pi_j$ can be replaced by $1/w_j$, where

$$w_j = \frac{1/\pi_j}{\sum_{k=1}^{n} 1/\pi_k}$$

- Result: $\mathrm{E}_P\left(b_{12}^*\right) = \mathrm{E}_P\left(\hat{\beta}_{12}^*\right) = B_{12} + \mathrm{O}\left(n^{-1}\right)$, where E_P denotes design expectation (that is to say, the expectation over repeated sample selection).

Categorical data analysis (sect. D)

- Testing goodness-of-fit:
  - Assume known multinomial distribution with probabilities $\mathbf{p_0} = \left(p_{0,1}, \ldots, p_{0,k-1}\right)$, where k is the number of categories and $\sum_1^k p_{0,i} = 1$.
  - Under H_0, chi-square statistic $X^2 = n\sum_{i=1}^{k} \frac{\left(\hat{p}_i - p_{0i}\right)^2}{p_{0i}}$ (where $\hat{p}_i$ are sample estimates of p_{0i}) is distributed asymptotically as: $X^2 = \sum_{i=1}^{k-1} \lambda_{0i}^2 Z_i^2;\quad Z_i \underset{\text{ind.}}{\sim} \mathrm{N}(0,1)$, where λ_{0i}^2 are the eigenvalues of $\mathbf{D} = \mathbf{P}_0^{-1}\mathbf{V_0}$, $\mathbf{P}_0$ is the variance matrix of the sample estimates, under the null hypothesis for SRS, and $\mathbf{V}_0$ is their true variance matrix under H_0.
  - Modified chi-square statistic: $X_C^2 = X^2/\hat{\lambda}^2;\ \hat{\lambda}^2 = \sum_{i=1}^{k-1}(1-\hat{p}_i)\hat{d}_i^2/(k-1)$, where $\hat{d}_i^2$ are estimates of the design effects, d_i^2, of $\hat{p}_i$.
- Test of independence in two-way contingency tables:
  - Hypothesis of interest: $H_0 : h_{ij} = p_{ij} - p_{i+}p_{+j} = 0 \quad (i = 1, \ldots, r; j = 1, \ldots, c)$, where p_{ij} is the population proportion in the (i,j)th cell and $p_{i+} = \sum_1^c p_{ij},\quad p_{+j} = \sum_1^r p_{ij}$ are the marginal totals.

- Usual chi-square statistic:

$$X_I^2 = n \sum_{j=1}^{c} \sum_{i=1}^{r} \frac{\left(\hat{p}_{ij} - p_{i+} p_{+j}\right)^2}{p_{i+} p_{+j}}$$

where $\hat{p}_{ij}$ denotes the sample estimator of p_{ij}.

- X_I^2 is asymptotically distributed as weighted sum of b independent χ_1^2 random variables.
- First order correction: $X_{I(C)}^2 = X_I^2 / \hat{\delta}_{\cdot}^2$, where:

$\hat{\delta}_{\cdot}^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (1 - \hat{p}_{i+})(1 - \hat{p}_{+j}) \hat{\delta}_{ji}^2 / b$, and $\hat{\delta}_{ij}^2 = n \frac{\hat{\mathrm{v}}\left(\hat{h}_{ij}\right)}{\hat{p}_{i+} \hat{p}_{i+j} (1 - \hat{p}_{i+})(1 - \hat{p}_{+j})}$, is the estimated design effects of $\hat{h}_{ij}$.

- Alternative modification obtained by replacing $\hat{\delta}_{\cdot}^2$ by $\hat{\lambda}_{\cdot}^2 = \frac{1}{rc - 1} \sum_{i=1}^{r} \sum_{j=1}^{c} \left(1 - \hat{p}_{ij}\right) \hat{d}_{ij}$

- Generalisations for log-linear models

  - Log-linear model: $\boldsymbol{\mu} = \tilde{u}(\boldsymbol{\theta}) \mathbf{1} + \mathbf{X}\boldsymbol{\theta}$, where $\boldsymbol{\pi}$ is the T-vector of population cell proportions, π_t, $\boldsymbol{\mu}$ is the T-vector of log probabilities $\mu_t = \ln \pi_t$, $\mathbf{X}$ is a known $T \times r$ matrix of full rank and $\mathbf{X}'\mathbf{1} = \mathbf{0}$, $\boldsymbol{\theta}$ is an r-vector of parameters and $\tilde{u}(\boldsymbol{\theta}) = \ln\{1/[\mathbf{1}'\exp(\mathbf{X}\boldsymbol{\theta})]\}$ is a normalizing factor.

  - Hypothesis of interest: H_0: $\boldsymbol{\theta}_2 = \mathbf{0}$, where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$, $\mathbf{X}_1$ is $T \times s$ and $\mathbf{X}_2$ is $T \times u$, $\boldsymbol{\theta}_1$ is $s \times 1$ and $\boldsymbol{\theta}_2$ is $u \times 1$.

  - Let $\hat{\boldsymbol{\pi}}$ be the pseudo maximum likelihood estimator of $\boldsymbol{\pi}$, under M_1, that is the solution of the pseudo-likelihood equations: $\mathbf{X}'\hat{\boldsymbol{\pi}} = \mathbf{X}'\hat{\mathbf{p}}$, where $\hat{\mathbf{p}}$ is a (design-) consistent estimator of $\boldsymbol{\pi}$, under the survey design. Similarly let $\hat{\hat{\boldsymbol{\pi}}}$ be the pseudo maximum likelihood estimator of $\boldsymbol{\pi}$, under M_2.

  - Standard Pearson chi-square statistic for testing H_0:

$$X^2 = n \sum_t \frac{\left(\hat{\pi}_t - \hat{\hat{\pi}}_t\right)^2}{\hat{\hat{\pi}}_t}.$$

  - Asymptotic distribution of X^2: $X^2 = \sum_{i=1}^{u} \delta_i^2 Z_i^2$; $Z_i \underset{\text{ind.}}{\sim} \mathrm{N}(0,1)$, where δ_i^2 are the eigenvalues of a generalised design effect matrix.

- Modified chi-square statistics: $X^2/\hat{\delta}_{\cdot}^2$, $X^2/\hat{\lambda}_{\cdot}^2$ and $X^2/\hat{d}_{\cdot}^2$ where:

- $\hat{\delta}_{\cdot}^2$ is the estimate of the average of the eigenvalues, $\delta_{\cdot}^2 = \frac{1}{u}\sum_i \delta_i^2$
- $\hat{\lambda}_{\cdot}^2$ is the estimate of the average of the design effects of $\mathbf{X}'\hat{\mathbf{p}}$
- $\hat{d}_{\cdot}^2 = \frac{1}{T}\sum_t \hat{d}_t^2$, where $\hat{d}_t^2 = n\frac{\hat{\mathrm{v}}(\hat{p}_t)}{\hat{\pi}_t(1-\hat{\pi}_t)}$ is the estimated design effect of cell *t*.

- Example: for the hypothesis of complete independence in a three-way $I\times J\times K$ table, $H_0 : \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$, where $\pi_{i++}, \pi_{+j+}, \pi_{++k}$ are the three-way marginals, the value of $\hat{\delta}_{\cdot}^2$ is given by:

$$\hat{\delta}_{\cdot}^2 = \frac{\sum_i\sum_j\sum_k(1-\hat{\pi}_{i++}\hat{\pi}_{+j+}\hat{\pi}_{++k})\hat{d}_{ijk}^2 - \sum_i(1-\hat{\pi}_{i++})\hat{d}_i^2(r) - \sum_j(1-\hat{\pi}_{+j+})\hat{d}_j^2(c) - \sum_k(1-\hat{\pi}_{++k})\hat{d}_k^2(l)}{IJK - I - J - K + 2}$$

where $\hat{d}_i^2(r)$, $\hat{d}_j^2(c)$, and $\hat{d}_k^2(l)$, are estimates of the design effects of the three-way marginals and $\hat{d}_{ijk}^2$ is the estimate of the cell design effect.

References

An, A., and D. Watts (2001). *New SAS Procedures for Analysis of Sample Survey Data*. SAS Users Group International (SUGI) paper, No. 23. Cary, North Carolina, SAS Institute, Inc. Available from http://www2.sas.com/proceedings/sugi23/stats/p247.pdf (accessed 2 July 2004).

Berthoud, R., and J. Gershuny, eds. (2000). *Seven Years in the Lives of British Families: Evidence on the Dynamics of Social Change from the British Household Panel Survey.* Bristol, United Kingdom: The Policy Press.

Brewer, K.R.W., and R.W. Mellor (1973). The effect of sample structure on analytical surveys. *Australian Journal of Statistics,* vol. 15, pp. 145-152.

Chambers, R. L., and C.J. Skinner, eds. (2003). *Analysis of Survey Data*. New York: Wiley and Sons, Inc.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley and Sons, Inc.

DuMouchel, W.H., and G.J. Duncan (1983). Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association,* vol. 78, pp. 535-543.

Duncan, G. J., and G. Kalton (1987). Issues of design and analysis of surveys across time. *International Statistical Review,* vol. 55, pp. 97-l17.

Feder, M., G. Nathan and D. Pfeffermann (2000). Time series multi-level modelling of complex survey longitudinal data with time varying random effects. *Survey Methodology,* vol. 26, pp. 53-65.

Fuller, W.A. (1984). Least squares and related analyses for complex survey designs. *Survey Methodology*, vol. 10, pp. 97-118.

Hansen, M.H., W.G. Madow and B.J. Tepping (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association,* vol. 78, pp. 776-793.

Kish, L., and M. Frankel (1974). Inference from complex samples. *Journal of the Royal Statistical Society B,* vol. 36, pp. 1-37.

Korn, E.L., and B.I. Graubard (1999). *Analysis of Health Surveys*. New York and Chichester, United Kingdom: Wiley and Sons, Inc.

Manski, C. F., and S.R. Lerman (1977). The estimation of choice probabilities from choice based samples. *Econometrica,* vol. 45, pp. 1977-1988.

Nathan, G. (1972). On the asymptotic power of tests of independence in contingency tables from stratified samples. *Journal of the American Statistical Association,* vol. 67, pp. 917- 920.

_____, and D. Holt (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society B,* vol. 43, pp. 377-386.

_____, and T.M.F. Smith (1989). The effect of selection on regression analysis. In *Analysis of Complex Surveys, C.J.* Skinner, D. Holt and T.M.F. Smith, eds. Chichester, United Kingdom: Wiley and Sons, Inc., pp. 227-250.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review,* vol. 61, pp. 317-337.

_____, and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya*, *Series B*, vol. 61, pt. 1, pp. 166-186.

Pfeffermann, D., F. Moura and N.S. Silva (2001). Multi-level modelling under informative probability sampling. Invited paper for the 53rd Session of the International Statistical Institute**,** Seoul.

Pfeffermann, D., and others (1998). Weighting for unequal selection probabilities in multi-level models. *Journal of the Royal Statistical Society B,* vol. 60, pp. 23-40.

Rao, J.N.K., and A.J. Scott (1981). The analysis of categorical data from complex sample surveys: hi-squared tests for of fit and independence in two-way tables. *Journal of the American Statistical Association,* vol. 76, pp. 221-230.

_____ (1984). On chi-squared tests for multi-way contingency tables with cell proportions estimated from survey data. *Annals of Statistics,* vol.12, pp. 46-60.

Särndal, C-E., B. Swensson and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Scott, A.J., and C.J. Wild (1986). Fitting logistic models under case control or choice-based sampling. *Journal of the Royal Statistical Society B,* vol. 48, pp. 170-182.

Skinner, C.J., D. Holt and T.M.F. Smith, eds. (1989). *Analysis of Complex Surveys*. Chichester, United Kingdom: Wiley and Sons, Inc.

Valliant, R., A.H. Dorfman and R.M. Royall (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Chichester, United Kingdom, and New York: Wiley and Sons, Inc.