Fabián Tibaldi [1], Liesbeth Bruckers [1], Herman Van Oyen [2], Johan Van der Heyden [2], Geert Molenberghs [1]

[1] *Limburgs Universitair Centrum, Center for Statistics, Diepenbeek*

[2] *Operational Public Health Research, Unit of Epidemiology, Scientific Institute of Public Health, Brussels*

Statistical software for calculating properly weighted estimates from Health Interview Survey Data

This Hinks & Kinks paper focuses on the interplay between the sampling design of a health interview survey and the choice of an appropriate statistical analysis. Data from the Belgian Health Interview Survey 1997 (HIS) are used as an illustration. The shortcomings of a simple (unweighted) analysis are discussed. The results from a weighted analysis are contrasted with those from a simple analysis. Finally, a weighted analysis incorporating stratification and cluster effects is conducted.

A main interest of surveys is the estimation of population parameters using the sample, often selected by complex schemes such as stratified multi-stage cluster sampling (Cochran 1977). Statistical methods for estimating population parameters and their associated variances are based on assumptions about the characteristics of the underlying distribution of the observations. Among these are that the observations were selected independently and have the same probability of being selected. The HIS violates both assumptions. For logistical reasons the selected households are clustered geographically, and within the household a subsample is taken.

We briefly outline the main aspects of the final sampling scheme for the selection of the households and respondents in the HIS.

The sampling of the households and respondents was a combination of different sampling techniques such as stratification and multistage sampling. Stratification was performed at the regional level (Flemish, Walloon and Brussels regions) and at the provincial level. At the regional level, unequal sampling rates were taken to guarantee sufficient precision of the results. Within a region, sampling was taken proportional to population size in each province. An extra refinement was needed for the German community, which was considered a proper entity on its own and was oversampled. Regional and provincial stratification aim at achieving a geographical spread of the interviews. The quota of interviews were also evenly distributed over quarters of the study year to obtain reasonable spread over time.

Within each stratum (province), a sample of individuals was obtained in three stages. At the first stage, municipalities (primary sampling units) were drawn by a systematic sampling procedure with probability proportional to their size. Each time a municipality was selected, a group of 50 individuals had to be successfully contacted. The next stage of random selection operated at household level (secondary sampling units). Finally, individuals (tertiary sampling units) were selected within households in such a way that at most four persons were interviewed in each household and the reference person and his or her partner were automatically selected. In addition to that extra households with similar characteristics (age of the reference person, gender, household size, and quarter) were selected to be used in case of non-response. We consider also these factors when calculating weights to perform the estimations. In Quataert et al. (1997) extra information about all these concepts can be found.

In this situation, sample units are not selected independently, nor are their responses likely to be independently distributed. Additionally, we are dealing with unequal selection probabilities because of regional stratification and oversampling in some provinces. Correct estimates can be obtained by re-weighting the data, inversely proportional to the selection probability (Levy & Lemeshow 1999). The principle behind estimation in a probability sample is that each person in the sample represents an entire slice of the population. The weight for each individual in the HIS is the product of the reciprocal of the selection probability within a household and a post-stratification factor for each province according to age, gender, household size distribution in each province, and quarter of the year in which the interview was done.

The proportion of persons reporting having had a steady general practitioner is used to illustrate the effect of design

Soz.- Präventivmed. 48 (2003) 269–271
0303-8408/03/040269–03
DOI 10.1007/s00038-003-3017-3

Table 1 Unweighted estimates

Region	N	p	S.E.
Flemish	4067	0.959	0.003103
Walloon	4889	0.955	0.002952
Brussels	2971	0.825	0.006962
Global	11927	0.924	0.002421

aspects on the estimated proportions and variances at regional and federal levels. Table 1 presents the results when ignoring all sampling design aspects of the HIS. Any statistical software package will give identical results. These estimated proportions and variances are *not* valid. They are only shown for the purpose of contrasting them with those obtained by incorporating unequal selection probabilities in the estimation process.

Table 2 shows the weighted estimated proportions of persons who had a steady general practitioner (Van Oyen et al. 1997). The results were obtained with SPSS for Windows software (Release 10.0.5) and STATA (Intercooled STAT 6.0 for Windows 98/95/NT). The weighted regional proportions are very close to the "uncorrected" ones based on a simple analysis. Since the selection probabilities within a region are more or less equal, the sample weights within a region are about constant. The weighted federal estimate for the proportion is closer to the proportions in the Walloon and Flemish regions than in the uncorrected analysis. This was to be expected since the sample weights are constructed in such a way that they correct for oversampling in the Brussels region. Note that the weighted variances obtained with SPSS are much smaller that the ones in the uncorrected analysis. SPSS employs weights representing *replicates*. Essentially, the data file is "inflated" so as to have a total sample size about equal to the population size of Belgium (i.e., the sum of the weights is approximately equal to the size of the Belgian population).

Table 2 Weighted estimates using SPSS with unequal weights

Region	Weights			Normed weights	
	N	p	S.E	p	S.E.
Flemish	4067	0.967	0.00008	0.967	0.00297
Walloon	4889	0.946	0.00014	0.946	0.00324
Brussels	2971	0.833	0.00038	0.833	0.00840
Global	11927	0.945	0.00007	0.945	0.00208

The weighted variances shown in Table 2 underestimate the true variances by a factor of 100 to 1000 at the regional level, obtained by dividing the sample size by the one corresponding to the region in each case. STATA uses a Taylor series linearisation method (see STATA Manual) to estimate the weighted variances (Tab. 3). The precision of the STATA estimates is of the same order of magnitude as the uncorrected variances.

It is possible to obtain reasonable estimates for the variances with SPSS by "norming" the weights. The normed weights are calculated by dividing each original weight by a factor such that the sum of the normed weights equals the actual number of persons considered in the analysis. As a result the variances are of the correct order of magnitude (Tab. 2). Actually, obtaining the normed weights involved performing the latter process four times, once for each region and once for the federal level. This re-scaling of the weights has no effect on the estimated proportions.

Although the estimated proportions resulting from a weighted analysis are valid, the precision on the other hand is not yet fully correct. It is well known that in general stratification has the effect of increasing the precision, while clustering leads to a loss in precision. SPSS does not handle both stratification and clustering in the estimation of weighted proportions. STATA has "SURV" commands that allow specification of these characteristics for the survey data. Table 3 shows the results of a STATA analysis taking into account different selection probabilities as well as stratification at the provincial level.

Notice that stratification has no effect on the proportions, but that the variances are – as expected – slightly smaller than the ones obtained when ignoring it. The more homogenous the persons within a stratum are with respect to the

Table 3 Weighted estimates using STATA and accounting for unequal weights, stratification, and clustering effects

Region	Weights			Weights and stratification		Weights, stratification and clustering	
	N	p	S.E.	p	S.E.	p	S.E.
Flemish	4067	0.967	0.003352	0.967	0.0033419	0.967	0.004099
Walloon	4889	0.946	0.006963	0.946	0.0069565	0.946	0.008304
Brussels	2971	0.833	0.007575	0.833	0.0075762	0.833	0.010894
Global	11927	0.945	0.002958	0.945	0.0029401	0.945	0.003628

variable of interest, the larger the effect on the variance will be (Levy & Lemeshow 1999)

Table 3 finally shows the result of a *complete* STATA analysis. Different selection probabilities, stratification, and clustering at the household level were taken into account. The SURV STATA commands can only handle one level of clustering. In the HIS there is clustering at the household level and at the municipality (a group of 50 subjects) level. Clustering at the household level is expected to be strongest. The analysis therefore corrects for the dependence of family members. As might be expected, the weighted estimates for the variances are inflated by the clustering.

Conclusions

In this paper the importance of the appropriate use of the sampling design aspects in producing valid estimates for survey data is emphasised. Even from a simple example it is clear that a normed-weighted analysis results in correct point estimates and reasonable estimates for the standard error. Valid estimates for the precision are obtained when stratification and clustering are taken in account.

References

Cochran WG (1977). Sampling techniques. New York: Wiley.

Levy P, Lemeshow S (1999). Sampling of populations. New York: Wiley.

Quataert P, Van Oyen H, Tafforeau J, et al. (1997). The Health Interview Survey 1997: protocol for the selection of the households and the respondents. Brussels: Scientific Institute of Public Health. (Episerie; 12).

Van Oyen H, Tafforeau J, Hermans H, et al. (1997). The Belgian Health Interview Survey. Arch Public Health *55*: 79–82.

Address for correspondence

Fabián Tibaldi
Limburgs Universitair Centrum
Center for Statistics
Universitaire Campus, Building D
B-3590 Diepenbeek
Tel.: +32-11-26 82 83
Fax: +32-11-26 82 99
e-mail: fabian.tibaldi@luc.ac.be