RESEARCH Open Access

Background

Like many large-scale health surveys, the Australian Longitudinal Study on Male Health (Ten to Men) used a complex sampling scheme. This choice was made because sampling the target population using a simple random sample was not feasible. Sampling theory therefore plays an important role in our study design because it provides a framework for efficiency gains [1]. In Ten to Men, the key elements of the sample design were the use of stratification, multi-stage sampling and cluster sampling to select prospective participants and invite them to take part in the study. This design has implications for the analysis of data from Ten to Men for both inferences about population means or prevalences, and for quantifying the magnitude of associations between exposures and outcomes. Such analysis implications are, however, often poorly understood. At the extreme, views differ on whether to *al ay* adjust for aspects of the study design and sampling scheme at the analysis stage (including accounting for unequal sampling fractions using inverse-probability-of-selection sampling weights) or to *e e* adjust. Korn and Graubard [2] give an excellent example of this controversy using US National Health and Nutrition Examination Surveys (NHANES). At the heart of this debate is a trade-off between miti-

the SA2 as the PSU), adjustment for stratification (no adjustment, adjustment using the stratification variable as a covariate, adjustment using the survey command), and use of sample weights (yes or no). We also examine the association between self-rated health and smoking status using logistic regression, where the effect size of interest is an odds ratio. We again omit the results from analyses that use a multi-level logistic model for the same reasons discussed in the previous section.

In an analysis that makes no adjustment for the multi-stage design or for stratification or weighting (Table 2, row A), the mean difference between the two groups is −5.1 kg (95 % CI −5.8 to −4.5 kg). That is, those who describe themselves as having very good or excellent health report are, on average, 5.1 kg lighter than those who have good, fair or poor health. Adjusting for stratification by using a series of indicator variables for remoteness to enter it into the model as a categorical variable (row B) also gives a mean difference of −5.1 kg with 95 % CI −5.7 to −4.4 kg. Repeating the analysis in row A but with the use of sample weights to adjust for bias gives a smaller difference of −4.4 kg, but with a wider confidence interval than observed previously (95 % CI −5.6 to −3.3). Adjustment for stratification makes only a small difference to this result (row D).

Repeating the analysis to account for all stages of sampling using a multilevel model (rows E and F) gives a mean difference of −4.9 kg (95 % CI −5.5 to −4.2), with further adjustment for stratification giving a difference of −4.8 kg (95 % CI −5.5 to −4.2). As with estimating population prevalences using multi-level models, it is not possible to easily account for the sample weighting in this context.

The final four rows in Table 2 show results obtained using the survey commands to estimate the population mean difference. When SA1s are defined as the PSU and sample weights are used (row G), the mean difference between the two groups is −4.4 kg (95 % CI −5.5 to −3.2). When no weights are used, the difference is −5.1 kg (95 % CI −5.8 to −

effectively up-weighting data from SA1s with low participation fractions and thus poor participation.

Y Me a d Ad t

Following the calculation in Equation (1) above we get

$$\begin{aligned} P^{YM}_k &\propto C^B_k \times F^{YM}_k / T^{YM}_k \times U^{YM}_k / F^{YM}_k \\ &\propto C^B_k \times U^{YM}_k / T^{YM}_k \end{aligned}$$

$$\begin{aligned} W^{YM}_k &\propto 1/C^B_k \times T^{YM}_k / U^{YM}_k \\ &\propto T^{YM}_k / C^B_k \times 1/U^{YM}_k \end{aligned}$$

which, when we replace T^{YM}_k with C^{YM}_k, becomes

$$W^{YM}_k \propto C^{YM}_k / C^B_k \times 1/U^{YM}_k$$

Similarly for adults we get

$$W^A_k \propto C^A_k / C^B_k \times 1/U^A_k$$

I e a d O te Re a St ata B

For the inner and outer regional strata SA1s have equal probability of selection, so the term Pr(SA1$_k$ selected) does not vary within a remoteness stratum and can therefore be absorbed in the constant of proportionality.

To illustrate, for a boy in SA1$_k$ the probability of selection P^B_k is

$$\begin{aligned} P^B_k &\propto \Pr(\text{Boy found } | \text{SA1}_k \text{selected}) \\ &\quad \times \Pr(\text{Boy provides usable data} \mid \text{Boy found}) \\ &\propto F^B_k / T^B_k \times U^B_k / F^B_k \\ &\propto U^B_k / T^B_k \end{aligned} \tag{2}$$

Replacing T^B_k with C^B_k based on the assumption above, gives

$$P^B_k \propto U^B_k / C^B_k$$

and therefore

$$W^B_k \propto C^B_k / U^B_k$$

Y Me a d Ad t

Similarly for young men and adults we get $W^{YM}_k \propto C^{YM}_k / U^{YM}_k$ and $W^A_k \propto C^A_k / U^A_k$

Appendix 2

Stata code for incorporating baseline survey characteristics

In Stata, the survey characteristics of the study must be declared prior to undertaking any analysis that acknowledges the sampling design. The command that brings the stratification, multistage design (at the PSU level)

Global Health, The University of Melbourne, Melbourne 3010, Australia. [3]Murdoch Childrens Research Institute, Royal Children's Hospital, Parkville 3052, Australia. [4]Department of Paediatrics, Melbourne Medical School, The University of Melbourne, Melbourne 3010, Australia. [5]Statistical Consulting Centre, School of Mathematics and Statistics, The University of Melbourne, Melbourne 3010, Australia.

Published: 31 October 2016

References

1. Cochran WG. Sampling Techniques. New York: John Wiley & Sons; 2007.
2. Korn EL, Graubard BI. Epidemiologic studies utilizing surveys: accounting for the sampling design. Am J Public Health. 1991;81:1166–73.
3. StataCorp. Stata: Release 14.1. College Station: StataCorp LP; 2015.
4. Korn EL, Graubard BI. Analysis of Health Surveys. New York: John Wiley & Sons; 1999.
5. Australian Bureau of Statistics. Australian Statistical Geography Standard (ASGS): Volume 5 - Remoteness Structure. Canberra: Australian Bureau of Statistics; 2013.
6. Fitzmaurice GM, Laird NM, Ward M. Applied Longitudinal Analysis (Wiley Series in Probability and Statistics). 2004.
7. Scott AJ, Holt D. The Effect of Two-Stage Sampling on Ordinary Least Squares Methods. JASA. 1982;(7):848–54.
8. Neuhaus JM, Segal MR. Design effects for binary regression models fitted to dependent data. Stat Med. 1993;12:1259–68.
9. Lumley T. Complex surveys: a guide to analysis using R. New York: John Wiley & Sons; 2010
10. Thomas L, Peterson ED. The value of statistical analysis plans in observational research: defining high-quality research from the start. JAMA. 2012;308:773–4.
11. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. Stat Med. 2007;26:20–36.