Centre for Mental Health Research
The Australian National University

Psychiatric Epidemiology and Social Issues (PEaSI)
Working Paper Series

Calculating sample weights for the *PATH through Life survey*

Peter Butterworth, Sarah Olesen, Liana Leach and Trish Jacomb

PEaSI WP: 01
May 2010

Calculating sample weights for the PATH through Life survey

Peter Butterworth [1], Sarah Olesen [1], Liana Leach [1] and Trish Jacomb [1]

PEaSI WP: 01

Author contact details:

1. Centre for Mental Health Research, The Australian National University, ACT 0200

Acknowledgements

The PATH Through Life Project was supported by NHMRC Program Grant 179805 and NHMRC Project Grant 157125. PB is funded by NHMRC Population Health Career Development Award Fellowship 525410.

We thank Anthony Jorm, Helen Christensen, Bryan Rodgers, Trish Jacomb, Karen Maxwell and the PATH interviewing team for their contribution to the PATH Through Life Project.

Introduction

The aim of this brief technical report is to describe the development of sample weights for Wave 1 of the Personality and Total Health Through Life (PATH) dataset. The PATH Project is a longitudinal, community survey assessing the health and well-being of the residents of Canberra and Queanbeyan (NSW) in Australia. Using a prospective cohort design, the survey follows three cohorts of participants, born in the years 1975-1979, 1957-1960 and 1937-1941, interviewing them once every four years over a planned 20 year period.

The PATH project used a simple random sample, with participants drawn from the Australian electoral roll. The recruitment process targeted people in three age brackets: 20-24, 40-44 and 60-64. Potential participants in the two younger age brackets were initially identified from a larger population with a 10-year age range (i.e., 20-29 and 40-49), as this was the minimum age range released for research purposes by the Australian Electoral Commission at this time. A modification of these laws provided a more targeted 5-year age range for the 60-64 year olds. To contact participants aged 20-24, an introductory letter explaining the study was sent to 12,414 people listed as 20-29 years old on the electoral role. To contact participants aged 40-44, a letter was sent to 9,033 people listed as 40-49 years old. A more targeted group of 4,831 people listed as 60-64 years old was also sent the introductory letter. Participation rates for those who were in the correct age range and could be located were: 20-24 - 58.6%, 40-44 – 64.6%, 60-64 – 58.3%. The final sample (n= 7485) for Wave 1 of PATH was: 1163 males and 1241 females aged 20-24, 1192 males and 1338 females aged 40-44, and 1319 males and 1232 females aged 60-64.

Although the absence of stratification and clustered sampling avoids potential survey design effects, subgroups of the population showed a differential response rate: that is, some people were more or less likely to respond to the invitation to participate. Literature suggests non-response is more likely

amongst men, young people, and people who are socio-economically disadvantaged (Lundberg et al., 2005) This may mean that the PATH sample does not match or represent the population from which it was drawn (i.e., Canberra and Queanbeyan). Earlier descriptive analysis (Leach et al. (submitted); Windsor & Rodgers (unpublished analysis)) showed that the sample was similar to the target population on a range of known characteristics, but somewhat different on others. For example, whilst only 10.9% of women in the youngest cohort of PATH were not in the labour force, the 2001 Australian Census for the Canberra/Queanbeyan area found that this rate was 16.2%.

In the context of the primary aims of the PATH Through Life Project, response bias amongst these population subgroups may affect the accuracy of population estimates for the prevalence of mental illness and its risk/protective factors. That is, if there is a response bias the estimates drawn from the PATH sample may not reflect the true prevalence of these factors in the relevant population (Canberra and Queanbeyan). Therefore, our aim in this working paper was to develop sample weights so that PATH baseline data more accurately reflects the general population of Canberra and Queanbeyan. Sample weights adjust for the bias in the probability that certain groups of individuals in the reference population (in the case of PATH, Canberra or Queanbeyan) were selected for and responded to the original PATH survey. Each sample member is assigned a sample weight, which is a value that represents the inverse probability of their inclusion in the sample on the basis of a range of personal and demographic characteristics. These weights are 'applied to' (i.e., multiplied by) participants' responses so that the responses of people who share characteristics with those less likely to respond are given 'more weight' than others.

We did not base weights on cross-classification of a number of covariates (e.g., gender × education × marital status × employment status) because it is difficult to obtain such detailed data on the general

population. Moreover, such cross-tabulations would likely include several very small (or empty) cell sizes. Rather, we used the process of "*raking"* as implemented in STATA svywgt procedure. This produces weights based on marginal results from multiple variables. The process draws on information from several simpler cross-tabulations (age × gender × critical factor) from the Census (e.g., education, marital status, employment status). It uses an iterative process: utilising one variable at a time, calculating weights that adjust the sample to the specific population characteristics; then using this weighted data, adjusts these weights to match the population on the next variable in the sequence. The process is iterative in that, once all variables have been used in the calculation of weights, the sequence is then repeated until the weights "converge" or no longer change substantially across sequences. For details, see Battaglia et al. (2004), Izrael et al., (2009) and Johnson (2008).

The method use to calculate sample weights for PATH data

Weights were calculated independently for men and women in each of the three cohorts (6 separate calculations in all). The final weights were designed to reweight to the original sample size in each cohort. The comparative data were from special tables requested from the ABS and drawn from the 2001 census for the Canberra and Queanbeyan region (ABS, 2001). The tables reported data for Australian citizens who usually resided in Canberra and Queanbeyan so as to approximate the population on the Australian electoral roll. The weighted estimate of the number of men and women within each cohort was also based on this 2001 ABS census data. Census data was available in five-year age bands corresponding to the PATH sampling frame. Variables used in the raking process were: employment status (employed, unemployed and not in the labour force), marital status (legally married, defacto, and not married; separated/divorced/widowed was also included for middle and older age cohorts), high school education (not complete senior certificate or equivalent), occupational classification (identifying those employed in labouring or elementary clerical/sales from ABS ASCO

coding), and having completed a tertiary qualification (at Bachelor level or above). For the weighting process, census data on the prevalence of these population characteristics in the Canberra and Queanbeyan area were expressed as a proportion of the sample size for each sex by cohort group; Because the raking procedure requires complete data, it was necessary to impute missing data on the PATH variables used to construct weights. Missing data was minimal (maximum missingness of 0.43% for ASCO occupational variable). As all data is categorical and as we seek to adopt the most conservative approach, we replaced missing data with the most common category.

Results

The ranking procedure converged for all subgroups within the default parameters of the svywght procedure. Table 1 presents details of the weights for each gender within each cohort.

Table 1. Descriptive statistics on sample weights by cohort and sex

		Characteristics of sample weights					
Cohort	**Sex**	**Mean**	**Median**	**25th percentile**	**75th percentile**	**Minimum**	**Maximum**
1975-1979	M	1.040	1.047	0.657	1.047	0.239	3.336
	F	0.963	0.844	0.700	1.123	0.304	2.882
1956-1960	M	1.019	0.928	0.748	1.194	0.733	3.223
	F	0.983	0.973	0.688	1.052	0.471	2.664
1937-1941	M	0.967	1.012	0.688	1.299	0.310	2.752
	F	1.035	1.286	0.710	1.309	0.198	1.811

The sample weights for each of the groups are (on average) close to 1. This was expected given that the aim was to reweight to the same overall sample size. Further, the inter-quartile range and full range of weights is relatively narrow. Thus, there were no exceptionally large (or small) weights.

Table 2. Comparison of population parameters with estimates based on weighted and unweighted PATH data.

		ABS population characteristics			Unweighted PATH			Weights PATH		
Cohort	Sex	Unemployed (%)	Married (%)	<12 yrs school (%)	Unemployed (%)	Married (%)	<12 yrs school (%)	Unemployed (%)	Married (%)	<12 yrs school (%)
1975-1979	M	9.15	4.55	17.19	6.74	6.14	8.99	8.64	**4.58**	**17.32**
	F	5.37	9.20	13.95	4.79	11.44	8.69	4.49	**9.28**	**14.09**
1956-1960	M	3.19	67.46	33.95	2.01	73.99	25.50	3.21	**67.41**	**33.92**
	F	2.50	65.44	36.08	2.62	68.71	31.59	2.44	**65.43**	**36.09**
1937-1941	M	2.50	79.81	43.79	1.29	82.72	38.58	2.19	**79.74**	**43.90**
	F	0.68	66.56	54.69	0.57	66.75	53.65	0.56	66.59	54.90

Bold = unweighted estimate not included in weighted 95% CI.

We anticipated the need to trim extreme weights to prevent a small number of cases having a disproportionate influence on any analyses and/or to reduce the impact of large weights on calculations of variance. However, there were no weights outside the truncation criteria (set apriori as the median weight plus six times the interquartile range; Izrael et al., 2009). The weights were left as derived by the raking process and, therefore, approximate the characteristics of the general Canberra/Queanbeyan population.

To investigate the effect of the application of the sample weights on the population estimates derived from the PATH survey, Table 2 presents data for three characteristics (unemployment, legal marriage, and having not completed 12 years of high school) for each sex within each cohort. The table initially presents data from the 2001 Census and then unweighted and weights results from the PATH survey. The table also indicates (in bold) instances where the 95% confidence interval of the weighted PATH estimates did not include the unweighted estimate. This shows where the inclusion of sample weights makes a significant difference to the calculation of the weighted and unweighted estimates.

It is evident that the weighted estimate is closer to the actual population parameter in all cases in which there is substantial difference between the weighted and unweighted PATH estimates. This shows the weighted estimate better approximates the actual known prevalence of the characteristic in the Canberra/Queanbeyan population than the unweighted PATH data. The application of weights had little effect on estimates for women in the oldest cohort. The weights also did not substantially alter the estimates of unemployment, though most weighted estimates (particularly young males) were closer to the population parameter than the unweighted estimates.

Conclusions

In this brief report we have described the process of deriving sample weights for the PATH Survey. We used the "raking" process within STATA to calculate the weights based on a series of cross tabulations of data from the 2001 Census. The weights covered a modest range, with no extreme values. The application of sample weights resulted in a better approximation to population parameters than the unweighted data and may overcome the effects of response bias.

These sample weights will be available as a variable in the PATH through Life dataset. They should be used to ensure representativeness of results. This is particularly important when interpreting results as indicative of population characteristics. These sample weights do not address attrition between waves. There remains, therefore, a need to develop longitudinal weights: either specifically for each individual project conducted using the longitudinal data, or general longitudinal weights that can also be added to the dataset.

Key points for potential users

- **What are sample weights?** Sample weights adjust for biases in the probability that an individual in the reference population (in the case of PATH, Canberra or Queanbeyan) was selected for and responded to the original PATH survey. These biases can lead to a sample having different characteristics to the population it aims to represent. Each sample member is given a sample weight, which is a value that represents the probability of their inclusion in the sample on the basis of their personal/demographic characteristics. These weights are 'applied to' (i.e., multiplied by) participants' responses so that the responses of people who share characteristics with those less likely to respond are given 'more weight' than others.
- **Why use PATH sample weights?** Weighting the data means that your results will be representative of the general (Canberra) community, rather than the PATH sample specifically. Whilst PATH uses a random sample, like any survey, there were some biases in the response rate (i.e., certain population subgroups were more or less likely to respond than others).
- **When to use the weights?** Weighted data should be used in any analyses where you intend to relate your findings to the general community (i.e., generalise your results beyond PATH sample members / assume your findings are true of people in the general community). This will almost always be the case.
- **Where do I find the weights?** PATH sample weights are contained in a variable called "sample_weight" in all current PATH data files on the CMHR O: Drive.
- **How do I use the weights?** PATH sample weights can be applied to any analyses in SPSS by (do either of these commands *before* your analyses, and every time you open the dataset or re-run your analyses):
    - Using the drop-down menu
      Data > Weight Cases > (select) "sample_weight" > (select) Weight cases by > (click the arrow) > (click OK)
    - Using this syntax
      WEIGHT BY sample_weight.
      EXECUTE.
- **Will using the weights (or not) influence my results?** This will depend on the type of analyses you are doing. Prevalence rates may differ somewhat in weighted and unweighted data, however, analyses that test the association between variables are unlikely to be substantially affected.

References

Australian Bureau of Statistics. (2001). Tables requested from ABS 2001 Census: Australian Citizenship. ABS: Canberra.

Battaglia M.P, Izrael, D. Hoaglin D.C, and Frankel M.R. (2004). Tips and Tricks for Raking Survey Data (a.k.a. Sample Balancing) American Association for Public Opinion Research. Available at: http://www.amstat.org/sections/srms/proceedings/y2004/files/Jsm2004-000074.pdf

Battaglia M.P, Izrael, D. Hoaglin D.C, and Frankel M.R. Practical Considerations in Raking Survey Data. Survey Practice: Practical Information for Survey Researchers. Available at: http://www.abtassoc.us/presentations/raking_survey_data_2_JOS.pdf

Izrael, D., Battaglia, M.P. and Frankel, M.R. (2009). Extreme survey weight adjusting as a component of sample balancing (aka raking). SAS Global Forum 2009. Available at: http://support.sas.com/resources/papers/proceedings09/247-2009.pdf

Johnson, D.R. (2008). Using weights in the analysis of survey data. Available at: http://help.pop.psu.edu/help-by-statistical-method/weighting/Introduction%20to%20survey%20weights%20pri%20version.ppt.

Leach L.S, Butterworth P, Strazdins, L., Rodgers, B., Broom, D.H., and Olesen S.C. (submitted). The limitations of employment as a tool for social inclusion. BMC Public Health.

Lundberg I, Damström Thakker K, Hällström T, Forsell Y. (2005). Determinants of non-participation, and the effects of non-participation on potential cause-effect relationships, in the PART study on mental disorders. Soc Psychiatry Psychiatr Epidemiol. 2005 Jun;40(6):475-83.

Windsor, T.D. and Rodgers, B. (unpublished analysis). Comparison of PATH Wave 1 sample to Census data for Canberra and Queanbeyan ABS census 2001 (Usual residents profile for Australian Citizens).