RESEARCH METHODS & REPORTING

STATISTICS NOTES

Inverse probability weighting

Mohammad Ali Mansournia *assistant professor of epidemiology*[1], Douglas G Altman *professor of statistics in medicine*[2]

[1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran; [2]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford OX3 7LD, UK

Statistical analysis usually treats all observations as equally important. In some circumstances, however, it is appropriate to vary the weight given to different observations. Well known examples are in meta-analysis, where the inverse variance (precision) weight given to each contributing study varies, and in the analysis of clustered data.[1]

Differential weighting is also used when different parts of the population are sampled with unequal probabilities of selection. Two examples of intentional unbalanced sampling are:

1. *Surveys with unequal probabilities of selection*—In a national survey of hypertension prevalence, certain groups with relatively rare characteristics (such as people aged ≥65 years) were oversampled to improve the precision of estimates for those groups.[2]

2. *Two-phase prevalence studies*—In the first phase of a two-phase prevalence study of mental health status, the sampled patients completed a short screening questionnaire. In the second phase, a subsample was selected for a definitive diagnostic test with oversampling of the screen-positive cases to ensure precise estimates for diagnostic prevalence.[3]

In such cases the ordinary unweighted sample quantities, such as means or proportions, are likely to be biased estimates of their corresponding population quantities. This "selection bias" can be eliminated by performing a weighted estimation, giving each individual's data a weight inversely proportional to their probability of selection. Intuitively, the weighting is used to deflate the weight for those individuals who are oversampled. The weighted analysis can be thought of as creating a study with no differential selection.

Inverse probability weighting can also be used when individuals vary in their probability of having missing information. Two contexts where there may be unintentional unbalanced selection are:

3. *Studies with missing outcome data*—In surveys such as that mentioned in example 1, the response rates will be affected by availability or willingness to participate. Likewise in a cohort study of the effect of obesity on hypertension, some individuals are censored due to loss to follow-up (such as emigration) or competing risks (such as death from other causes).[4] In each case the amount of missing information will vary across subgroups.

4. *Randomised trials with crossing over from one arm to the other*—In a randomised trial 8010 postmenopausal women with early breast cancer were assigned to tamoxifen (n=2459) or letrozole (n=2463) for five years or to sequential treatment with two years of one of these agents followed by three years of the other. There was a selective crossover to letrozole of 619 patients in the tamoxifen arm after significant benefit was reported for letrozole compared with tamoxifen during the study. These 619 women may be artificially censored at the time they crossed over for analysis.[5]

In these situations, missing outcomes are unlikely to happen at random so that estimates will be biased. While the selection probabilities in examples 1 and 2 are known, the response or non-censoring probabilities in examples 3 and 4 are unknown. Inverse probability weighting can be used with weights estimated from a logistic regression model for predicting non-response or censoring. As in the first scenario, this application of the method aims to remove bias, but it is more controversial. Its validity relies on a correctly specified model including all prognostic variables associated with non-response or censoring, which cannot be assured.

In the breast cancer trial (example 4), although the intention-to-treat hazard ratio for overall survival (which ignores selective crossover) was 0.87 (95% confidence interval 0.77 to 1.00) in favour of letrozole, the adjusted hazard ratio using inverse probability of selection weights was 0.79 (0.69 to 0.90), suggesting that the true effect is greater than the intention-to-treat estimate.[5]

Correspondence to: M A Mansournia mansournia_ma@yahoo.com

In observational studies, the probability of exposure can depend on external factors (called confounders) that also affect the outcome. The causal effect of interest is then confused with the effects of confounders. Such confounding can be thought as a type of selection bias, because confounding essentially means that some causes of the outcome also influence selection for the exposure. A particularly important context is:

5 *Non-randomised studies comparing different treatments*—In a cohort study 12 552 warfarin-naive patients with atrial fibrillation admitted to hospital for ischaemic stroke and treated with warfarin were compared with patients who received no oral anticoagulant at discharge.[6]

Outside randomised trials the choice of treatment is likely to be influenced by predictors of outcome, so called "confounding by indication".[7] Various strategies are used to try to remove the bias in non-randomised treatment comparisons. The conventional approach is to use multivariable regression, but a recent alternative is inverse probability of treatment weighting. Here the weights are based on each individual's probability of receiving a specific treatment given the confounders, which is known as the propensity score (PS). The weights are $1/PS$ for the treated participants and $1/(1-PS)$ for the untreated participants.[8] The weights can be estimated from a logistic regression model for predicting treatment. Key assumptions are that all confounders have been measured and properly modelled in this treatment model. In the warfarin study (example 5) the unadjusted hazard ratio for cardiac events was 0.73 (99% confidence interval 0.67 to 0.80) in favour of warfarin, whereas the adjusted estimate using inverse probability of treatment weighting was 0.87 (0.78 to 0.98), about half the effect size.[6] If the cohort is also affected by censoring (see example 3 above), one can adjust simultaneously for confounding and selection bias due to censoring.[4 8]

Although helpful for bias reduction, estimates weighted by design weights (examples 1 and 2) tend to be less precisely estimated than the unweighted estimates, which is not necessarily true for examples 3-5. The ordinary 95% confidence interval for inverse probability weighted estimates may not provide the correct coverage and should be avoided. Instead, robust "sandwich" variance estimators or non-parametric bootstrapping should be used to provide valid confidence intervals.[8] Deeper discussion of inverse probability weighting methods can be found elsewhere.[8 9]

1 Kerry SM, Bland JM.. Analysis of a trial randomised in clusters. *BMJ* 1998;316: 54. 9451271
2 Korn EL, Graubard BI.. Epidemiologic studies utilizing surveys: accounting for the sampling design. *Am J Public Health* 1991;81: 1166-73. 1951829
3 Vázquez-Barquero JL, García J, Simón JA et al. Mental health in primary care. An epidemiological study of morbidity and use of health resources. *Br J Psychiatry* 1997;170: 529-35. 9330019
4 Alonso A, Seguí-Gómez M, de Irala J, Sánchez-Villegas A, Beunza JJ, Martínez-Gonzalez MA.. Predictors of follow-up and assessment of selection bias from dropouts using inverse probability weighting in a cohort of university graduates. *Eur J Epidemiol* 2006;21: 351-8. 16736275
5 Regan MM, Neven P, Giobbie-Hurder A et al. BIG 1-98 Collaborative Group International Breast Cancer Study Group (IBCSG). Assessment of letrozole and tamoxifen alone and in sequence for postmenopausal women with steroid hormone receptor-positive breast cancer: the BIG 1-98 randomised clinical trial at 8·1 years median follow-up. *Lancet Oncol* 2011;12: 1101-8. 22018631
6 Xian Y, Wu J, O'Brien EC et al. Real world effectiveness of warfarin among ischemic stroke patients with atrial fibrillation: observational analysis from Patient-Centered Research into Outcomes Stroke Patients Prefer and Effectiveness Research (PROSPER) study. *BMJ* 2015;351: h3786. 26232340
7 Freemantle N, Marston L, Walters K, Wood J, Reynolds MR, Petersen I.. Making inferences on treatment effects from real world data: propensity scores, confounding by indication, and other perils for the unwary in observational research. *BMJ* 2013;347: f6409. 24217206
8 Robins JM, Hernán MA, Brumback B.. Marginal structural models and causal inference in epidemiology. *Epidemiology* 2000;11: 550-60. 10955408
9 Seaman SR, White IR.. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2013;22: 278-95. 21220355