Adjustment for Non-Response in the Minnesota Nurses Study

Steven J. Mongin

1 May 2001

This section addresses differences in the probabilities of reporting a physical assault between selected subgroups as well as the probability of such a report overall. At least one form of the survey was sent to 6298 randomly selected individuals from the Minnesota Nurses license database (MNLDB). Of these 6298, 3989 responded (63.34%) to the question on physical assault with a "Yes" or a "No".

Basic demographic data from the MNLDB was available including age, sex, home location, and nursing license type. These covariates were used to define strata according to the following subgroups:

Age : 21-30, 31-40, 41-50, 51-60, 61 and over.

Sex : Female, male.

Home Location : Metro vs non-metro.

License Type : LPN only vs RN (with or without LPN).

The probability of response appears to vary among these strata. For example, Figure (**??**) shows the observed response rates by age group. Estimates of event probabilities in the population of nurses which gave us our sample could be biased if they are not adjusted for such differential response rates.

1 Horvitz-Thompson Adjustment

To account for the differential response rates, a Horvitz-Thompson (1952) type adjustment was applied under the following assumption:

A1 : The event probabilities are the same for all nurses in a given stratum, regardless of survey response status.

This adjustment is referred to as "H-T" in what follows. Under this assumption, the subset of responders in each stratum provide an unbiased estimate of the event probability of interest. However, estimates which pertain to a combination of strata are based on data weighted according to H-T.

For example, over the 40 strata determined by the four covariates presented above, we seek an estimate of the overall probability of physical assault which pertains to the population of eligible nurses in the MNLDB. Let us define the following:

Eligible : A nurse identified in the MNLDB who worked as a nurse at any time in the 12 months preceding the time during which surveys were completed.

Survey : The long or short form of the study survey.

$\mathcal{S}_k$: The set of subscripts i associated with all subjects in the k^{th} stratum; $k = 1, \ldots, K$.

ν : The probability of an Eligible nurse reporting a physical assault in the Survey.

w_i : The weight given to the i^{th} subject; $i = 1, \ldots, n$.

ρ_k : The probability that an Eligible nurse in the k^{th} stratum will respond to the physical assault question.

r_i : For the i^{th} study subject, the indicator of an Eligible nurse responding with a "Yes" or a "No" to the physical assault question. For any $i \in \mathcal{S}_k$, $E[r_i] = \rho_k$.

v_i : For the i^{th} study subject, the indicator of responding with a "Yes" to the physical assault question. If "Yes", $v_i = 1$; if "No", $v_i = 0$. This value is missing for 6298 - 3989 = 2309 nurses. For a randomly chosen i, $E[v_i] = \nu$.

v_{k1} : The sum of the v_i, hence the number of reported physical assaults, in the k^{th} stratum.

n_{k0} : The number of Eligible nurses from the k^{th} stratum who were sent a Survey.

n_{k1} : The number of Eligible nurses from the k^{th} stratum who were sent a Survey and responded to the physical assault question.

Applying H-T to this case, we have $n = 6298$ and $K = 40$. For each k, we can estimate ρ_k by $\hat{\rho}_k = n_{k1}/n_{k0}$. The estimate of ν is then given by:

$$\begin{aligned} \hat{\nu} &= \frac{\sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} w_i v_i}{\sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} w_i r_i} && (1) \\ &= \frac{\sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} (1/\hat{\rho}_k) v_i}{\sum_{k=1}^{K} \sum_{i \in \mathcal{S}_k} (1/\hat{\rho}_k) r_i} \\ &= \frac{\sum_{k=1}^{K} v_{k1} \,/\, \hat{\rho}_k}{\sum_{k=1}^{K} n_{k1} \,/\, \hat{\rho}_k}. && (2) \end{aligned}$$

Taking v_{k1}/n_{k1} as an estimate of the incidence rate in the k^{th} stratum, (2) is mathematically equivalent to producing a standardized incidence rate by *direct adjustment* (Szklo and Nieto, 2000; pp 265+), where the "standard population" is defined by the entire sample of 6298 nurses and their distribution across the K strata.

2 Eligibility Weights

The target population is restricted to Eligible nurses. Of those 4920 nurses who indicated their Eligibility, 921 (18.72%) were not Eligible. The estimates of the stratum-specific response probabilities as well as all other estimates pertaining to the target population should reflect this restriction. Let e_i denote the indicator of a sufficient response from the i_{th} subject to determine their Eligibility, and ϵ the probability that a nurse selected for this study is indeed Eligible. Each nurse can be cross-classified according to the observed values of e_i and r_i. Let the following refer to subjects within the k^{th} stratum, and suppress the k subscript. Then the totals of each subcategory may be denoted a, b, c, and d as shown at (3).

$$\begin{aligned} a &= \sum \mathbf{1}_{\{i\,:\,e_i=0,\,r_i=0\}} \quad b = \sum \mathbf{1}_{\{i\,:\,e_i=0,\,r_i=1\}} \\ c &= \sum \mathbf{1}_{\{i\,:\,e_i=1,\,r_i=0\}} \quad d = \sum \mathbf{1}_{\{i\,:\,e_i=1,\,r_i=1\}} \end{aligned} \tag{3}$$

Each of the counts represented by a, b, c, and d may be further subscripted with a 1 or a 0 to convey how many are actually Eligible or not. For example, a_1 are Eligible, a_0 are not, and $a_1 + a_0 = a$.

A natural estimate of ρ, the probability of response to the physical assault question by an Eligible nurse, is then given by $(b_1+d_1)/(a_1+b_1+c_1+d_1)$.

However, we cannot observe a_1 and b_1. Consider these simplifying assumptions:

A2 The probability of Eligibility, ϵ, does not depend on whether or not Eligibility could be determined.

A3 Among those in whom Eligibility could not be determined, ϵ does not depend on whether or not they responded to the physical assault question.

In terms of classification into a, b, c, or d at (3), these assumptions mean that information on actual Eligibility available from the $c + d$ nurses can be used to estimate Eligibility for those who did not provide it. Then, an unbiased estimator of ϵ is $\hat{\epsilon} = (c_1 + d_1)/(c + d)$. It follows that a_1 may be estimated by $\hat{a_1} = \hat{\epsilon}a$. Similarly, $\hat{b_1} = \hat{\epsilon}b$. The estimated probability of response is then:

$$\hat{\rho} \quad = \quad \frac{\hat{b_1} + d_1}{\hat{a_1} + \hat{b_1} + c_1 + d_1}, \tag{4}$$

Based on counts a, b, c, c_1, d, and d_1 specific to each stratum, we then have $\hat{\rho_k}$, $k = 1, \ldots, K$ as stratum-specific estimates of the probability of response to the physical assault question among the Eligible nurses.

In order to use the data corresponding to b at (3) in any statistical estimates, an additional factor for their weight terms w_i as in (1) may be considered. This is to account for the lack of response regarding Eligibility among subjects in this subcategory. Using assumptions A2 and A3 for ϵ, the appropriate case weight factor to incorporate in w_i for these subjects is $\hat{\epsilon}$. The effect of this increment is to downweight these subjects by the estimated probability of not being Eligible, that is, $1 - \epsilon$ for their respective stratum.

The case weights $\{w_i\}_{i=1}^n$ form the cornerstone of a procedure of these steps:

1. Set all w_i to 1.
2. For i such that $e_i = 0$ (i.e., subjects in the a or b counts), multiply w_i by $\hat{\epsilon}$. This is trivial when all w_i equal 1, but less so for the bootstrap component of the procedure applied later.
3. Create the stratum-specific nonresponse weights $\{1/\hat{\rho_k}\}_{k=1}^K$ according to (4), weighting all counts as dictated by the w_i resulting from step 2.
4. Estimate ν by (1).

3 Inference Based on the BC_a Bootstrap

Due to the use of the data in the creation of the weights, and then a subset of the data for the estimation of the probability parameters of interest, the distribution of an estimator such as $\hat{\nu}$ is difficult to specify analytically. A nonparametric bootstrap is well-suited for this situation. For confidence limits of higher accuracy than provided by the standard percentile bootstrap, a "BC_a" (bias-corrected and accelerated) bootstrap was used (Efron and Tibshirani, 1993).

The case weights $\{w_i\}$ provide a convenient vehicle for computing bootstrap confidence intervals. The first part of this procedure is the same as with a percentile bootstrap. A sample of size n with replacement is drawn from the original data. This resampling is communicated as a new set of weights $\{w_i\}_{i=1}^n$ such that $\sum_{i=1}^n w_i = n$, but each w_i may take a value from 0 to n. This bootstrap weight vector is used in step 1 of the procedure outlined in the previous section, in place of the vector of ones.

This procedure is repeated B times (with B typically at least 2000), saving $\hat{\nu}_b$, the estimate of ν based on the b^{th} bootstrap sample, for $b \in \{1, \ldots, B\}$. A percentile bootstrap, 95% confidence interval would then use the $q_{lo} = 0.025$ and $q_{hi} = 0.975$ quantiles of the distribution of $\{\hat{\nu}_b\}_{b=1}^B$. The supplemental BC_a procedure adjusts q_{lo} and q_{hi} to remove bias in the original estimator detected by the bootstrap, and to adjust for apparent lack of optimality in the current metric of the parameter. This subprocedure also involves resampling, but in a jackknife-type algorithm over the n cases. With each iteration, the $\{w_i\}_{i=1}^n$ are initialized with all ones except for the case to be left out, which is assigned $w_i = 0$. The resulting BC_a quantiles, q_{lo}^* and q_{hi}^*, are then used to identify 95% confidence limits from the original bootstrap distribution of $\{\hat{\nu}_b\}_{b=1}^B$.

[More can be added at this point regarding inference about differences in probabilities between selected subgroups.]

4 References

Efron, B., and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap,* New York: Chapman and Hall.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement from a Finite Universe," *Journal of the American Statistical Association,* 47, 663-685.

Szklo, M., and Nieto, F. J. (2000), *Epidemiology: Beyond the Basics,* Gaithersburg, Maryland, USA: Aspen.