Annals of Epidemiology 25 (2015) 174–178

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Attributable fraction estimation from complex sample survey data

Steven G. Heeringa PhD [a,*], Patricia A. Berglund MBA [a], Brady T. West PhD [a], Edmundo R. Mellipilán MS [a], Kenneth Portier PhD [b]

[a] Survey Methodology Program, Institute for Social Research, University of Michigan, Ann Arbor, MI
[b] Statistics and Evaluation Division, American Cancer Society, Inc., Atlanta, GA

ARTICLE INFO

Article history:
Received 7 November 2014
Accepted 9 November 2014
Available online 15 November 2014

Keywords:
Attributable fraction
Population attributable risk
Complex sample design
Weighting
Stratified sampling
Cluster sampling
Jackknife repeated replication
Bootstrap

ABSTRACT

Purpose: A review of methods for the estimation of attributable fraction (AF) statistics from case-control, cross-sectional, or cohort data collected under a complex sample design. Provide guidance on practical methods of complex sample AF estimation and inference using contemporary software tools.
Methods: Statistical literature on AF estimation from complex samples for the period 1980 to 2014 is reviewed. A general approach based on weighted sum estimators of the AF and application of Jackknife repeated replication and Bootstrap resampling methods for estimating the variance of AF estimates is outlined and applied to an example analysis of risk factors for alcohol dependency.
Results: The literature lays the theoretical foundation to address the problem of AF estimation and inference from complex samples. To date, major statistical software packages do not provide a complete program but the approach is easily implemented using the modeling software and macro/function language capabilities available in major statistical analysis packages. In an example application, weighted sum estimation and inference for the population AF showed stable and consistent results under both Jackknife repeated replication and Bootstrap methods of variance estimation.
Conclusions: Future work on AF estimation for complex samples should focus on simulation studies and empirical testing to investigate the properties of the resampling variance estimation methods across a range of complex study design features and populations.

© 2015 Elsevier Inc. All rights reserved.

Introduction

Contemporary epidemiologic research on the association between disease or other health outcomes and putative exposure risk factors makes extensive use of observational data from large sample survey programs such as the National Health and Nutrition Examination Survey (NHANES). Because most surveys are designed to be representative of defined target populations, it is natural that they are used not only to analyze risk but also to estimate the attributable fraction (AF) of disease or other measurable population outcomes that can be associated with a given level of exposure to a risk factor.

Survey data collections are used in the full range of epidemiologic study designs [1–11]. Cross-sectional sample data from the Behavioral Risk Factor Surveillance Survey have been used to study the AF of prevalent cancer diagnoses associated with adverse childhood experiences [12]. Prospective cohort studies of epidemiologic risk and AFs often use baseline survey data. The NHANES III Mortality Follow-up Study has been used to estimate hazards ratios and AFs for all-cause mortality among diabetic adults [13]. The AFs of breast cancer deaths associated with age at first birth and family history of breast cancer have been estimated from the NHANES I Epidemiological Follow-up Survey data [14]. Although a less common use in practice, sample survey designs are also used to conduct case-control investigations of epidemiologic risk factors [15].

Complex sample designs and design effects

The NHANES, Behavioral Risk Factor Surveillance Survey, and other epidemiologic survey data collections recruit subjects from complex probability sample designs that feature stratification of the target population, multistage cluster sampling, and disproportionate sample selection (and thus compensatory weighting in estimation) [16].

The term "complex sample" originates in the specialized features of these sample designs relative to simple random sampling (SRS) [17].

Strata are nonoverlapping groupings of all population elements or clusters of elements formed by the study designer before the selection of the probability sample. Sample elements or clusters are sampled independently within strata, thereby eliminating any

* Corresponding author. Institute for Social Research, University of Michigan, 426 Thompson, Ann Arbor, MI 48109. Tel.: +1 734-647-4621; fax: +1 734-764-8263.
E-mail address: sheering@umich.edu (S.G. Heeringa).

http://dx.doi.org/10.1016/j.annepidem.2014.11.007
1047-2797/© 2015 Elsevier Inc. All rights reserved.

between-strata component of the sampling error and reducing the total variance of estimates. In practice, stratification also facilitates disproportionate allocation of the sample to strata that define subpopulations of interest (e.g., stratified sampling of exposed and unexposed persons in a prospective cohort study of disease) [18].

Single stage or multistage cluster sampling is used by survey designers primarily to reduce interviewing costs by amortizing travel and related expenditures over a group of observations. In almost all cases, sampling plans that incorporate cluster sampling result in standard errors for survey estimates that are greater than those from an SRS of equal size. The general increase in variances of sample estimates due to clustered sampling is caused by correlations (nonindependence) of observations within sample clusters [19].

Weighting of the survey data in analysis is required to "map" the sample back to an unbiased representation of the target population. Generally, the final analysis weights in survey data sets are the product of the sample selection weight (w_{sel}), a survey nonresponse adjustment factor (w_{nr}) and a poststratification adjustment factor (w_{ps}):

$$w_{final} = w_{sel} \times w_{nr} \times w_{ps}$$

The objective in applying weights in survey estimation is to compensate for varying sample inclusion probabilities and to attenuate potential bias due to differential nonresponse in the selected probability sample. The statistical price paid for bias reduction using w_{sel} and w_{nr} is increased standard errors for weighted estimates. In contrast, poststratification weighting (w_{ps}) to external population controls can lead to reduced standard errors of sample estimates or may attenuate sampling biases due to sample frame noncoverage [20].

In the context of finite population sampling, standard statistical approaches to estimating relative risk and its variance assume that the observed data are obtained from a SRS. This closely approximates the probability model assumption that, conditional on exposure levels and other covariates, the observations of disease outcomes are independent, random draws from a probability distribution (e.g., binomial for the logistic model).

The need to apply complex sample population weights changes the approach to estimation of population statistics or model parameters [21]. As noted previously, also relative to SRS designs, stratification, cluster sampling, and weighting all influence the sizes of standard errors and the associated confidence intervals (CIs) for survey estimates. Even for simple statistics such as population proportions or means, the net influence of these design effects on the sampling variance of complex sample estimates is difficult to model analytically. Empirically, experience shows that for multistage-stratified clustered sample designs such as the NHANES the true variance for the complex sample estimates is greater than that for data from an SRS of equal size [16].

Review and recommended approaches

Estimators of AFs take a number of different forms depending on (1) the epidemiologic study design (cross-sectional, cohort, and case control); (2) whether estimates of risk ratios are adjusted for covariates other than the exposure factor; and (3) the source of prevalence information for exposure/covariate patterns (sample estimates, population censuses, or registers) for the population of inference [22–26].

Following Bruzzi et al. [22], the adjusted population AF for exposures at $m = 0,\ldots, M-1$ levels and $c = 1,\ldots, C$ unique covariate patterns $X = \{X_1,\ldots X_p\}$ can be expressed as follows

$$AF = 1 - \sum_{c=1}^{C} \sum_{m=0}^{M-1} f_{m,c} \Big/ RR_{m0|c}; \quad (1)$$

where $f_{m,c}$ is the proportion of the diseased population with exposure level m and covariate pattern c and $\mathrm{RR}_{m0|c}$ is the relative risk of disease at exposure level m (relative to the baseline, $m = 0$) for covariate pattern c.

The adjusted AF statistic is thus a function of both the vector of population prevalences, $f = \{f_{m,c}\}$, and corresponding relative risks, $\mathrm{RR} = \{\mathrm{RR}_{m0|c}\}$. In the general case where f and RR must be estimated from sample data, the sampling variance of the estimator of AF is a complex nonlinear function of the individual estimates' variances and covariances. Further complicating estimation and inference for AF, the prevalences, f, may be estimated from a different source than the associated risk ratios or drawn from population data sources [27–29]. This article focuses on covariate-adjusted estimates of population AFs in which both the relative risk and exposure/covariate prevalence components must be estimated from a single complex sample data source, although the general method of estimation and inference we discuss applies equally to cases where estimates are sourced from multiple surveys and/or determined from registries or census sources.

In a cross-sectional, cohort, or a case-control study that is based on a complex probability sample design, estimates of the AF for rare diseases or outcomes can be computed as a weighted sum function of estimated relative risks [30,31].

$$\widehat{\mathrm{AF}}_1 = 1 - \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n(\alpha)} w_{h\alpha i} \cdot \frac{y_{h\alpha i}}{\mathrm{rr}_{h\alpha i}}}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n(\alpha)} w_{h\alpha i} y_{h\alpha i}} \quad (2)$$

where: h, α, i indexes sample observation i within stratum ($h = 1, \ldots, H$) and cluster ($\alpha = 1, \ldots, \alpha_h$); $w_{h\alpha i}$ = the population weight for observation i; $y_{h\alpha i}$ = indicator of disease status for observation i (1 = yes, 0 = no); $\mathrm{rr}_{h\alpha i}$ = the estimated relative risk for observation i, $\exp[\widehat{\beta}'(x_{h\alpha i} - x^0_{h\alpha i})]$; $\widehat{\beta}$ = vector of estimated Poisson (log rate) or logistic regression coefficients (log odds).

If the disease outcome is not rare, the estimator takes the form of a ratio of weighted sums of predicted probabilities from a logistic regression model [30].

$$\widehat{\mathrm{AF}}_2 = 1 - \frac{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n(\alpha)} w_{h\alpha i} \cdot p\left(x^0_{h\alpha i}\right)}{\sum_{h=1}^{H} \sum_{\alpha=1}^{a_h} \sum_{i=1}^{n(\alpha)} w_{h\alpha i} p(x_{h\alpha i})} \quad (3)$$

where: $p(x_{h\alpha i}) = \exp(\widehat{\beta}' x_{h\alpha i})/[1 + \exp(\widehat{\beta}' x_{h\alpha i})]$ = inverse logit transform estimate of the probability of disease at exposure level m; $p(x^0_{h\alpha i}) = \exp(\widehat{\beta}' x^0_{h\alpha i})/[1 + \exp(\widehat{\beta}' x^0_{h\alpha i})]$ = estimated probability of disease under counterfactual where exposure is set to level 0.

In the series of weighted sum AF estimators, the individual case weights have two functions. The weights are required to implicitly and correctly represent the fraction of all disease cases in the population that have a given joint distribution of exposure level m and covariate pattern c. The survey weights can also be used to support pseudomaximum likelihood estimation of the population parameters of the logistic, Poisson, or proportional hazards model that is used to derive the relative risk value for each observation [21,30,32].

Pseudomaximum likelihood estimation estimates of the regression parameters, $\boldsymbol{B}$, are obtained by maximizing the following unbiased estimate of the population pseudo likelihood which is a weighted function of the observed sample data and the predicted probability values, $\widehat{\pi}(\boldsymbol{x}_i)$—illustrated here for a binomial likelihood and the logistic link [33].

$$PL(\boldsymbol{B}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^{n} \left\{ \widehat{\pi}(\boldsymbol{x}_i)^{y_i} \cdot \left[1 - \widehat{\pi}(\boldsymbol{x}_i)\right]^{1-y_i} \right\}^{w_i} \quad (4)$$

with: $\widehat{\pi}(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i \widehat{\boldsymbol{B}})/[1 + \exp(\boldsymbol{x}_i \widehat{\boldsymbol{B}})]$.

Download English Version:

https://daneshyari.com/en/article/6148097

Download Persian Version:

https://daneshyari.com/article/6148097

Daneshyari.com