Confidence Intervals for Population Attributable Fractions Using Complex Survey Data

Akhil K. Vaish,[1] Olga Khavjou[2]
[1,2]RTI International, 3040 East Cornwallis Road, RTP, NC 27709

Abstract

This paper describes an easy-to-implement method for producing confidence intervals (CIs) for population attributable fraction (PAF) statistics using survey data. The PAF is used by epidemiologists and policymakers to assess how much of the disease burden in a population could be reduced if the exposure to certain risk factors were eliminated. The PAF is defined as $p(rr-1)/rr$, where p denotes the proportion of cases exposed to a risk factor and rr denotes the model-based relative risk comparing the proportion of cases among the exposed group with the proportion of cases among the unexposed group. The rr is obtained by modeling the log of the prevalence of the disease as a linear function of covariates where exposure to the risk factor is included as one of the model covariates. The proposed methodology is based on the Taylor series linearization and properly accounts for survey design features in estimating the variances and covariances of the estimated quantities. The methodology is implemented using the VARGEN procedure of SUDAAN® version 11.0 software on 2013 Behavioral Risk Factor Surveillance System data to produce state-by-age group PAF estimates and CIs of hypertension with diabetes as the risk factor.

Key Words: SUDAAN, VARGEN procedure, population attributable fraction (PAF), Taylor series linearization, Delta method, survey variance estimation, confidence intervals (CIs)

1. Introduction

How much of the disease burden in a population could be eliminated if the effects of certain risk factors were eliminated from the population? To address this question, epidemiologists calculate the population attributable fraction (PAF). In the presence of confounding, the PAF is defined by Kleinbaum, Kupper, and Morgenstern (1982) as $p(rr-1)/rr$, where p denotes the proportion of cases exposed to a risk factor and rr denotes the relative risk comparing the proportion of cases among the exposed group with the proportion of cases among the unexposed group. The rr can be adjusted for the covariates, such as age, race, and gender; in that case, the rr is obtained by modeling the log of the prevalence of a disease as a linear function of covariates where exposure to the risk factor is included as one of the model covariates. Although estimating the PAF is easy because it is a nonlinear function, estimating its standard error (SE) is not as straightforward, especially when complex survey data are being analyzed.

Most mental and physical health-related data along with information about a broad range of other topics are obtained via surveys that employ a complex, multistage sample selection process involving stratification and clustering. For example, the Behavioral Risk Factor

Surveillance System[1] (BRFSS) collects data about U.S. residents via a telephone survey regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. BRFSS completes more than 400,000 adult interviews each year in all 50 states as well as the District of Columbia and three U.S. territories, making it the largest continuously conducted health survey system in the world. Other prominent national surveys include the National Health Interview Survey[2], the National Health and Nutrition Examination Survey[3], and the National Survey on Drug Use and Health[4]. To obtain valid estimates of target population characteristics and associated SEs from selected survey data, statistical analysis methods using survey weights, stratification, clustering, and information about sample selection in the estimation process are used to analyze the survey data (Cochran, 1977; Lohr, 2010).

The objective of this paper is to obtain $100(1-\alpha)$ percent CIs for the PAFs for the 50 states and the District of Columbia by age groups, where p is the conditional probability and rr is the model-based relative risk estimated from complex survey data. We considered 4 age groups (18-44, 45-64, 65-74, and 75+), which entailed fitting 204 models (51 states × 4 age groups) to estimate model-based rr. In this case, estimating the variance of the estimated PAF by resampling methods (e.g., the jackknife method) is cumbersome. For example, to draw 20 resamples and run 204 models for each of the resamples to estimate model-based rr, which requires checking the model convergence for 4,080 models (204 models × 20 resamples), and to implement a customized solution for each of the models that did not converge would be highly time-consuming and impractical.

Another commonly used method is to employ Taylor series linearization (the "delta" method) to approximate the estimated PAF as a linear function of estimated rr and p. The variance of the linearized version of the estimated PAF involves determining the correlation between estimated rr and p, which is not available because estimated rr is a model-based estimate and estimated p is a weighted mean of a binary variable. To circumvent this problem, Natarajan, Lipsitz, and Rimm (2007) used 97.5 percent CIs for rr and p, then combined both the CIs using the Bonferroni inequality to obtain a 95 percent CI for the PAF. Natarajan et al. (2007) also noted that the Bonferroni-based CI for the PAF was approximately 18 to 20 percent wider than the delta method and that the jackknife method sometimes produced wider CIs than the Bonferroni method. Although the Bonferroni method is fast (compared with the jackknife method) and easy to implement using survey data analysis software, such as SUDAAN® (RTI International, 2012), it does not produce the variance of the estimated PAF, which is needed for the testing of hypotheses about the PAF. In this paper, we propose an approximation of the correlation between an estimated rr and p and use the delta method to produce the variance of the estimated PAF.

2. Methodology

2.1 Notations

[1] https://www.cdc.gov/brfss/

[2] https://www.cdc.gov/nchs/nhis/

[3] https://www.cdc.gov/nchs/nhanes/index.htm

[4] https://www.samhsa.gov/data/population-data-nsduh

The methodology is described via an example where the disease is hypertension and the risk factor is diabetes. BRFSS data from 2013 were used to demonstrate the methodology. The following notations are used. Let

d : diabetes ($d=1$ denotes having diabetes and $d=0$ denotes not having diabetes),
h : hypertension ($h=1$ denotes having hypertension and $h=0$ denotes not having hypertension),
$P(d=1)$: population probability of having diabetes,
$P(h=1)$: population probability of having hypertension,
$P(d=1|h=1)$: probability of having diabetes among the population members who have hypertension,
$P(h=1|d=1)$: probability of having hypertension among the population members who have diabetes,
$P(h=1|d=0)$: probability of having hypertension among the population members who do not have diabetes, and
$P(d=1, h=1)$: population probability of having diabetes and hypertension.

2.2 Estimation of the PAF and Its Associated Standard Error

Using the above notations, the unadjusted relative risk (*urr*) of hypertension associated with diabetes is defined as $urr = \dfrac{P(h=1|d=1)}{P(h=1|d=0)}$. The model-based (or adjusted for covariates) relative risk of hypertension associated with diabetes is defined as $rr = e^{\beta}$, where β is estimated by fitting the following log-linear model with *x* as the covariates, which are *d (diabetes), male, age (in years), black, hisp (Hispanic),* and *othr (other race)*:

$$\begin{aligned}\log[P(h=1|x)] = \beta_0 + \beta \times d + \beta_1 \times male + \beta_2 \times age + \\ \beta_3 \times black + \beta_4 \times hisp + \beta_5 \times othr.\end{aligned} \tag{1}$$

The model-based population fraction of hypertension attributable to diabetes (adjusted for the covariates mentioned above) is defined as

$$PAF = P(d=1|h=1)\left[\frac{rr-1}{rr}\right] = P(d=1|h=1)(1-e^{-\beta}), \tag{2}$$

where β is estimated from model (1). From equation (2), we note that an estimate of the PAF is a product of two random quantities: $P\hat{A}F = XY$, where $X = \hat{P}(d=1|h=1)$ and $Y=(1-e^{-\hat{\beta}})$ are estimates of $P(d=1|h=1)$ and $(1-e^{-\beta})$, respectively, and $\hat{\beta}$ is an estimate of β from model (1). The variance of $P\hat{A}F$ is approximated using the delta method and is given as follows:

$$\text{var}(P\hat{A}F) = \text{var}(XY) \approx y^2\,\text{var}(X) + x^2\,\text{var}(Y) + 2xy\,\text{cov}(X,Y). \tag{3}$$

Further, note that $\text{cov}(X,Y) = corr(X,Y)\sqrt{\text{var}(X)\,\text{var}(Y)} = \rho\sqrt{\text{var}(X)\,\text{var}(Y)}$, where $\rho = corr(X,Y)$ is the correlation between $\hat{P}(d=1|h=1)$ and $(1-e^{-\hat{\beta}})$. As mentioned earlier, estimating ρ by resampling methods (e.g., the jackknife method) is cumbersome and time-consuming. The correlation ρ, which involves an estimated adjusted relative risk $(e^{\hat{\beta}})$, is approximated by the correlation based on the estimated unadjusted relative risk, $u\hat{r}r = \dfrac{\hat{P}(h=1|d=1)}{\hat{P}(h=1|d=0)}$, i.e.,

$$\begin{aligned}
&\rho[\hat{P}(d=1|h=1),(1-e^{-\hat{\beta}})]\\
&\quad = -\rho[\hat{P}(d=1|h=1),e^{-\hat{\beta}}]\\
&\quad \approx -\rho[\hat{P}(d=1|h=1),\frac{\hat{P}(h=1|d=0)}{\hat{P}\left(h=1|d=1\right)}] = -\rho[X,r],
\end{aligned}$$

where $r = \dfrac{\hat{P}(h=1|d=0)}{\hat{P}\left(h=1|d=1\right)} = \dfrac{1}{u\hat{r}r}$ and

$$\begin{aligned}
\rho[X,r] &= \frac{\text{cov}(X,r)}{\sqrt{\text{var}(X)\,\text{var}(r)}}\\
&= \frac{[\text{var}(X)+\text{var}(r)-\text{var}(X-r)]/2}{\sqrt{\text{var}(X)\,\text{var}(r)}}.
\end{aligned}$$

Note that the $P\hat{A}F$ and $\text{var}(P\hat{A}F)$ depends on $\hat{P}(d=1|h=1)$, $\hat{P}(h=1|d=0)$, $\hat{P}\left(h=1|d=1\right)$, $\hat{\beta}$, $\text{var}(\hat{\beta})$, $\text{var}(X)$, $\text{var}(r)$, and $\text{var}(X-r)$. These quantities can easily be obtained by using the VARGEN and LOGLINK procedures in SUDAAN 11.0, which properly takes into account the sampling design features (e.g., weights, stratification, and clustering) in the estimation process. To use the VARGEN procedure efficiently, define

$$\begin{aligned}
x1 &= 1 \text{ if } (h=1 \text{ and } d=1)\\
&= 0 \text{ otherwise}\\
x2 &= 1 \text{ if } (h=1 \text{ and } d=0)\\
&= 0 \text{ otherwise,}
\end{aligned}$$

then $\hat{P}(d=1|h=1) = \dfrac{\hat{P}(d=1,h=1)}{\hat{P}(h=1)} = \dfrac{\text{weighted mean of } x1}{\text{weighted mean of } h}$,

$\hat{P}(h=1|d=0) = \dfrac{\hat{P}(h=1,d=0)}{\hat{P}(d=0)} = \dfrac{\text{weighted mean of } x2}{(1-\text{weighted mean of } d)}$, and

$\hat{P}\left(h=1|d=1\right) = \dfrac{\hat{P}(d=1,h=1)}{\hat{P}(d=1)} = \dfrac{\text{weighted mean of } x1}{\text{weighted mean of } d}$.

The VARGEN procedure also allows us to obtain r, $\text{var}(X)$, $\text{var}(r)$, and $\text{var}(X-r)$ as described in the SUDAAN 11.0 program (see Section 3.1). The LOGLINK procedure in

SUDAAN 11.0 is used to estimate β and the associated variance. Using $P\hat{A}F$ and $\text{var}(P\hat{A}F)$, a 95 percent normal CI for the PAF is given by $P\hat{A}F \pm 1.96\sqrt{\text{var}(P\hat{A}F)}$. If the $P\hat{A}F > 0$, then the CI can be constructed on the logit scale, i.e., $(\frac{e^L}{1+e^L}, \frac{e^U}{1+e^U})$, where $L = \ln\frac{P\hat{A}F}{1-P\hat{A}F} - 1.96\frac{\sqrt{\text{var}(P\hat{A}F)}}{P\hat{A}F(1-P\hat{A}F)}$, $U = \ln\frac{P\hat{A}F}{1-P\hat{A}F} + 1.96\frac{\sqrt{\text{var}(P\hat{A}F)}}{P\hat{A}F(1-P\hat{A}F)}$, and ln is the natural logarithmic function.

The number of attributable cases in the population is defined as $NPAF = N\,P(h=1)PAF$, where N is the population size. Note that

$$\begin{aligned} NPAF &= N\ P(h=1)\ PAF \\ &= N\ P(h=1)\left[P(d=1\,|\,h=1)(1-e^{-\beta})\right] \\ &= N\ [P(d=1,h=1)(1-e^{-\beta})] \\ &= N\ PAF2. \end{aligned}$$

Hence, a 95 percent normal CI for NPAF is given by

$$\left[\hat{N}\left(P\hat{A}F2 - 1.96\sqrt{\text{var}(P\hat{A}F2)}\right), \hat{N}\left(P\hat{A}F2 + 1.96\sqrt{\text{var}(P\hat{A}F2)}\right)\right],$$

where $\hat{N}$ is the sum of sample weights and $P\hat{A}F2$ and $\text{var}(P\hat{A}F2)$ can easily be obtained by using the methodology described above for estimating $P\hat{A}F$ and $\text{var}(P\hat{A}F)$.

3. Data Analysis

3.1 SUDAAN Code

To demonstrate the methodology, we used 2013 BRFSS data which included 483,865 respondents, BRFSS sample design variables (e.g., sampling weight, stratum, primary sampling unit [PSU]), covariates (e.g. age and gender), and indicator variables for hypertension and diabetes. To estimate the model-based relative risk for each of the state × age group combinations, we used the LOGLINK procedure in SUDAAN 11.0. For some of the state × age group combinations, the model did not converge, so a simpler model without a variable for "black" was fitted. To estimate the PAF and its associated SE, we used the VARGEN procedure in SUDAAN 11.0. To our knowledge, VARGEN is the only survey data analysis procedure that can produce means and variances of complicated statistics such as the PAF without explicitly deriving Taylor deviates (which are needed for computing valid survey variances for complex statistics; see, e.g., Graubard & Fears, 2005) and programming them using SAS or R. The following SUDAAN code explains the VARGEN procedure. SUDAAN's keywords are in boldface letters, user inputs are italicized for ease of understanding, and the text within /* */ denotes the estimated quantities. For a detailed discussion about SUDAAN's keywords, refer to the SUDAAN 11.0 user manuals (e.g., RTI International, 2012). The VARGEN procedure can also be used to obtain an estimate of unadjusted PAF which is defined as

$$\begin{aligned} U\hat{P}AF &= \hat{P}(d=1|h=1)(1-\frac{\hat{P}(h=1|d=0)}{\hat{P}\left(h=1|d=1\right)}) \\ &= \hat{P}(d=1|h=1)-\hat{P}(h=1|d=0)\frac{\hat{P}(d=1)}{\hat{P}\left(h=1\right)}. \end{aligned}$$

```
proc vargen data=temp1 design=wr;
class state age_group;
nest ststr psu/missunit;           /*BRFSS stratum and PSU variables*/
weight finalwt;                    /*BRFSS sampling weight*/
xmean x1mean: x1;                  /*1: $\hat{P}(d=1,h=1)$ */
xmean hmean: h;                    /*2: $\hat{P}(h=1)$ */
xmean x2mean: x2;                  /*3: $\hat{P}(h=1,d=0)$ */
xmean dmean: d;                    /*4: $\hat{P}(d=1)$ */
parameter r1: x1mean/hmean;        /*5: $\hat{P}(d=1|h=1)$ */
parameter r2: dmean/hmean;         /*6: $\hat{P}(d=1)/\hat{P}(h=1)$ */
parameter r3: x2mean/(1-dmean);          /*7: $\hat{P}(h=1|d=0)$ */
parameter r4: x1mean/dmean;              /* 8: $\hat{P}\left(h=1|d=1\right)$ */
parameter r5: r1-r2*r3;            /*9: unadjusted PAF */
parameter r6: r3/r4;               /*10: $\hat{P}(h=1|d=0)/\hat{P}\left(h=1|d=1\right)$=1/ $u\hat{r}r$ */
parameter r7: r1-r6;               /*11: $\hat{P}(d=1|h=1)-[\hat{P}(h=1|d=0)/\hat{P}\left(h=1|d=1\right)]$*/
tables state*age_group; setenv colwidth=20 decwidth=8;
output estim seestim/estimfmt=f11.8 seestimfmt=f11.8 filename=serr filetype=sas
replace;
run;
```

The output SAS data *serr* contains est1-est11 variables corresponding to the 11 statistics defined above using **xmean** and **parameter** keywords, and the associated SEs are stored in se1-se11 variables, respectively. The VARGEN procedure took about 3 minutes to run, and the LOGLINK procedure took about 9 minutes to fit 204 log-linear models. Each fitted model was tested for convergence and refitted without the "black" covariate when the initial model failed to converge. Using SAS software, state × age group-level outputs from the LOGLINK and VARGEN procedures from SUDAAN 11.0 can easily be combined to produce model-based or adjusted PAF estimates (equation 2) and the associated variance (equation 3).

4. Selected Results

In this section, we present some selected results for illustration purposes only because the main focus of this paper is on the utility of the proposed methodology, not on the results.

The notations used in Table 1 are as follows: Hyptn (%): percent population with hypertension, Diab (%): percent population with diabetes, RR: relative risk, RRL: lower 95 percent limit for RR, RRU: upper 95 percent limit for RR, PAF: population attributable fraction, PAFL: lower 95 percent limit for PAF, and PAFU: upper 95 percent limit for PAF.

Table 1: PAF Estimates and Associated 95 Percent Confidence Intervals

State	Age Group	Hyptn (%)	Diab (%)	RR	(RRL	RRU)	PAF	(PAFL	PAFU)
Alabama	18-44	18.9	15.1	2.65	(1.97,	3.57)	0.09	(0.05,	0.14)
Alabama	45-64	52.1	27.9	1.61	(1.47,	1.75)	0.11	(0.08,	0.13)
Alabama	65-74	67.2	34.6	1.41	(1.30,	1.53)	0.10	(0.07,	0.13)
Alabama	75+	72.4	29.8	1.12	(1.01,	1.24)	0.03	(0.00,	0.06)

The notations used in Table 2 are as follows: NPOP: total number of persons in the population, N_Hyptn: number of persons with hypertension in the population, NPAF: number of hypertension cases in the population attributable to diabetes, NPAFL: lower 95 percent limit for NPAF, and NPAFU: upper 95 percent limit for NPAF.

Table 2: Number of Attributable Cases and Associated 95 Percent Confidence Intervals

State	Age Group	NPOP	N_Hyptn	NPAF	(NPAFL	NPAFU)
Alabama	18-44	1,707,032	323,331	30,442	(16,365	44,520)
Alabama	45-64	1,289,604	671,542	70,523	(55,410	85,635)
Alabama	65-74	424,963	285,393	28,759	(21,309	36,209)
Alabama	75+	305,261	221,119	6,884	(211	13,556)

5. Future Research

We plan to select a few state × age group combinations with low, medium, and high prevalence of hypertension and estimate the PAFs and their associated SEs by a resampling method (e.g., the jackknife method), then compare the widths of the resulting CIs with our method's results and the Bonferroni method's results.

References

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley & Sons, Inc.

Graubard, B. I., & Fears, T. R. (2005). Standard errors for attributable risk for simple and complex sample designs. *Biometrics*, *61*, 847-855. https://doi.org/10.1111/j.1541-0420.2005.00355.x

Kleinbaum, D. G., Kupper, L. L., & Morgenstern, H. (1982). *Epidemiologic research: Principles and quantitative methods*. New York, NY: Van Nostrand Reinhold Company, Inc. / John Wiley & Sons, Inc.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Pacific Grove, CA: Duxbury Press.

Natarajan, S., Lipsitz, S. R., & Rimm, E. (2007). A simple method of determining confidence intervals for population attributable risk from complex surveys. *Statistics in Medicine, 26*, 3229-3239. https://doi.org/10.1002/sim.2779

RTI International. (2012). *SUDAAN®, Release 11.0* [computer software]. Research Triangle Park, NC: Author.