Variance Estimation of the Survey-Weighted Kappa Measure of Agreement

Moshe Feder, RTI International
E-mail address: mfeder@rti.org

Abstract

The standard formula due to Fleiss *et al.* (1969) for estimating the variance of the estimated Cohen's kappa, may be severely biased when using complex survey data, underestimating the variance. A procedure based on Taylor linearization is presented. The proposed procedure reduces to the Fleiss formula under a simple random sample design. Results from a small simulation study demonstrate the bias of the standard formula when clustering is present, and the good performance of the proposed procedure.

Keywords: Complex Sample, Taylor Linearization, Cohen's kappa, Re-interview.

1 Introduction

The National Survey on Drug Use and Health (NSDUH) is an annual survey of the civilian, noninstitutionalized, population of the United States, aged 12 years or older. It is the nation's primary source of statistical information on the use of illicit drugs. The survey employs a multistage area probability sample to produce population estimates of the prevalence of substance use and other health-related issues. Since 1999, the design was modified to allow state-level estimates with samples large enough for direct estimation of key substance abuse measures in 8 states and smaller samples requiring small area estimation procedures to produce state estimates in the other 42 states and the District of Columbia. The NSDUH is sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA). RTI has conducted the NSDUH since 1988. Prior to 2002, the survey was known as the National Household Survey on Drug Abuse (NHSDA).

NSDUH is currently conducting a study to assess the reliability of respondents' responses. An interview/reinterview method is employed where individuals are interviewed on two occasions, T1 and T2. The reliability of the responses is assessed by comparing the T1 and T2 responses. The anticipated sample size for this study is approximately 3,100.

To measure the reliability of categorical responses, Cohen's kappa (κ) index of inter-rater reliability is used (Cohen, 1960). This measure, κ, is the statistic most-often used to assess inter-rater reliability of categorical variables.

The common variance estimation approach is to use Fleiss *et al.* (1969) asymptotic variance formula (see also Agresti, 2002). It assumes an independent sample, with equal probabilities of inclusion. The NSDUH sample design is complex, involving clustering and unequal weighting to account for variable probabilities of inclusion and non-response adjustments. This may have significant effect on the point estimates of κ and the estimation of its variance. While correcting the point estimates of κ for the design is straightforward, the variance estimate is more involved. We present a Taylor linearization (TL) derivation, along with simulation results of the TL method and of the Fleiss *et al.* (1969) formula, and their assessment. The results show that failure to account for the clustering may result in negatively-biased variances (and and hence too-small standard errors).

2 The Kappa Parameter and Its Survey-Weighted Estimate

Let Y_{T1} and Y_{T2} be the responses given to a categorical question which have two levels, 0 and 1. The four possible combinations of T1 and T2 responses divide the population into four groups whose proportions are given in the following table.

	$Y_{T1}=0$	$Y_{T1}=1$	
$Y_{T2}=0$	p_{00}	p_{01}	p_{0+}
$Y_{T2}=1$	p_{10}	p_{11}	p_{1+}
	p_{+0}	p_{+1}	$p_{++}=1$

The probability of agreement is $p_e = p_{00}+p_{11}$. This parameter is a rough measure of agreement between the T1 and T2 responses. Note, however, that agreement could also occur by chance alone. Assuming independence of T1 and T2 responses, conditional on the marginal probabilities p_{0+},p_{1+},p_{+0},p_{+1}, the probability of agreement p_c (referred to as the 'probability of chance agreement') is given by $p_c = p_{0+}\cdot p_{+0}+p_{1+}\cdot p_{+1}$. To correct for a chance agreement, Cohen (1960) introduced the agreement measure κ, defined by the formula below.

$$\begin{aligned}\kappa &= \frac{\sum_{i=0,1} p_{ii} - \sum_{i=0,1} p_{i+}p_{+i}}{1-\sum_{i=0,1} p_{i+}p_{+i}} \\ &= \frac{(p_{00}+p_{11}) - (p_{0+}\cdot p_{+0} + p_{1+}\cdot p_{+1})}{1-(p_{0+}\cdot p_{+0} + p_{1+}\cdot p_{+1})}.\end{aligned}$$

Note that $-1 \le \kappa \le 1$. When $p_{00}+p_{11}=1$, a perfect agreement between T1 and T2, $\kappa = 1$. In the case of a complete disagreement $p_{01}+p_{10}=1$, $\kappa \le 0$. (Note, however, that the general case of complete disagreement—$p_{00}+p_{11}=0$ generally does not imply $\kappa=-1$. In fact,when $p_{01}+p_{10}=1$ and either p_{01} or p_{10} is small $\kappa \approx 0$.)

The survey-weighted estimate of κ is

$$\hat\kappa = \frac{(\hat p_{00}+\hat p_{11}) - (\hat p_{0+}\cdot \hat p_{+0} + \hat p_{1+}\cdot \hat p_{+1})}{1-(\hat p_{0+}\cdot \hat p_{+0} + \hat p_{1+}\cdot \hat p_{+1})}$$

where $\hat p_{ij}$ is the survey-weighted estimate of p_{ij}. The estimate $\hat\kappa$ is a ratio estimate. Thus it is consistent, albeit biased.

3 Taylor Linearization Variance Estimation

To simplify notation let us denote $a = \hat p_{00}$, $b = \hat p_{01}$ and $c = \hat p_{10}$. Note that $\hat p_{11} = 1-a-b-c$. Also, denote

$$A = 1-(b+c)$$

(A is the estimated probability of agreement) and

$$B = (a+b)(a+c) + (1-a-b)(1-a-c),$$

the estimated probability of chance agreement.

Note that $B = 1 + 2(a+b)(a+c) - (2a+b+c)$.

Now,

$$\hat\kappa = 1-\frac{1-A}{1-B} = 1-\frac{b+c}{(2a+b+c)-2(a+b)(a+c)} = 1-U$$

where $U = (b+c)/[(2a+b+c)-2(a+b)(a+c)]$.

Clearly, $\mathrm{Var}(\hat\kappa) = \mathrm{Var}(U) = \mathrm{Var}(C/D)$ where $C = b+c$, $D = (2a+b+c)-2(a+b)(a+c)$ and $U = C/D$.

The first-order Taylor approximation of U is

$$\Delta U \approx \frac{\partial U}{\partial a}\Delta a + \frac{\partial U}{\partial b}\Delta b + \frac{\partial U}{\partial c}\Delta c,$$

where the partial derivatives are calculated at $\mathbf{g} = (E(a),E(b),E(c))' \approx (p_{00},p_{01},p_{10})'$, where $\Delta a = a - E[a]$, $\Delta b = b - E[b]$, $\Delta c = c - E[c]$, $\Delta U = U - U_0$ and where U_0 is the value of U at $\mathbf{g}$:

$$\begin{aligned}U_0 &= \frac{E[b]+E[c]}{(2E[a]+E[b]+E[c])-2(E[a]+E[b])(E[a]+E[c])} \\ &\approx E[U].\end{aligned}$$

Denote the partial derivatives of U with respect to a, b and c (at $\mathbf{g}$) by F, G and H respectively. Then,

$$F = \frac{\partial U}{\partial a} = \frac{\partial U}{\partial C}\cdot\frac{\partial C}{\partial a} + \frac{\partial U}{\partial D}\cdot\frac{\partial D}{\partial a} = -\frac{2C}{D^2}\left[1-(2a+b+c)\right]$$

$$G = \frac{\partial U}{\partial b} = \frac{\partial U}{\partial C}\cdot\frac{\partial C}{\partial b} + \frac{\partial U}{\partial D}\cdot\frac{\partial D}{\partial b} = \frac{1}{D} - \frac{C}{D^2}\left[1-2(a+c)\right]$$

$$H = \frac{\partial U}{\partial c} = \frac{\partial U}{\partial C}\cdot\frac{\partial C}{\partial c} + \frac{\partial U}{\partial D}\cdot\frac{\partial D}{\partial c} = \frac{1}{D} - \frac{C}{D^2}\left[1-2(a+b)\right].$$

We have

$$\Delta U \approx F\Delta a + G\Delta b + H\Delta c. \tag{1}$$

Define a new variable x_i for every individual i in the population U as follows:

$$x_i = F \cdot I_{[T1=0,T2=0]} + G \cdot I_{[T1=0,T2=1]} + H \cdot I_{[T1=1,T2=0]}. \tag{2}$$

Denote by the population by $\mathcal{U}$, its size by N, and the population mean of x by $\bar{X} = N^{-1}\sum_{i\in\mathcal{U}} x_i$. Then

$$\bar{X} = N(Fp_{00} + Gp_{01} + Hp_{10}).$$

An estimate of $\bar{X}$ is

$$\begin{aligned}\hat{\bar{X}} &= \hat{F}\hat{p}_{00} + \hat{G}\hat{p}_{01} + \hat{H}\hat{p}_{10}\\ &= F\hat{p}_{00} + G\hat{p}_{01} + H\hat{p}_{10} +\\ &\left\{(\hat{F}-F)\hat{p}_{00} + (\hat{G}-G)\hat{p}_{01} + (\hat{H}-H)\hat{p}_{10}\right\}.\end{aligned}$$

Thus,

$$\mathrm{Var}(\hat{\kappa}) \approx \mathrm{Var}(\hat{\bar{X}}). \tag{3}$$

Note that the expression in $\hat{\bar{X}}$ in $\{\cdot\}$ is asymptotically zero and thus will be ignored.

3.1 Variance Estimation Under a Simple Random Sample Design

Under a simple random sample design (SRS),

$$\begin{aligned}\mathrm{var}(\hat{\bar{X}}) &= \mathrm{var}(Fa + Gb + Hc)\\ &= \frac{1}{n}\left[F^2a(1-a) + G^2b(1-b) + H^2c(1-c)\right.\\ &\left.-2FGab - 2FHac - 2GHbc\right]\end{aligned} \tag{4}$$

using the variance and covariance formulae for the components of a multinomial random variable.

Note: If X is a random variable, and $\mathrm{Var}(X)$ is its variance, $\mathrm{var}(X)$ (with lower-case v) will denote an estimate of $\mathrm{Var}(X)$.

Comment: The Fleiss *el al.* (1969) formula (see also Agresti, 2002, pp. 434–435) is

$$\begin{aligned}\mathrm{var}(\hat{\kappa}) = \frac{1}{n}&\left\{\frac{A(1-A)}{(1-B)^2}\right.\\ &+ \frac{2(1-A)\left[2AB - \sum_{i=0}^{1} p_{ii}(p_{i+} + p_{+i})\right]}{(1-B)^3}\\ &\left.+ \frac{(1-A)^2\left[\sum_{i=0}^{1}\sum_{j=0}^{1} p_{ij}(p_{j+} + p_{+i})^2 - 4B^2\right]}{(1-B)^4}\right\}.\end{aligned} \tag{5}$$

Equation (4) agrees with (5). Thus, the procedure we present generalizes Fleiss *el al.* (1969).

3.2 Variance Estimation Under Complex Designs

Assume for the moment that F, G and H are constants (population values). Let w_i be the survey weights, then

$$\hat{\bar{X}} = \frac{1}{N}\sum w_i x_i = F\hat{p}_{00} + G\hat{p}_{01} + H\hat{p}_{10}.$$

Using (2) and (3), we arrive at the procedure below.

3.3 Procedure

1. Calculate a, b, c and then F, G and H.
2. Calculate a new variable x_i for every $i \in s$ as in (2).
3. Calculate $\mathrm{var}(\hat{\bar{X}})$, the estimated variance of $\hat{\bar{X}}$ accounting for the sample's complex design (e.g., using PROC DESCRIPT of SUDAAN® [Research Triangle Institute (2004)]).

Then $\mathrm{var}(\hat{\kappa}) = \mathrm{var}(\hat{\bar{X}})$.

Comment:: When $D = (2a+b+c) - 2(a+b)(a+c) \approx 0$ the method may be unstable.

3.4 The Bias of the Standard Formula Under Complex Design

Let $\hat{\theta}$ be an estimate of a population parameter θ, $\mathrm{Var}(\hat{\theta})$ and $\mathrm{Var}_{\mathrm{SRS}}(\hat{\theta})$, respectively its variance under the complex design, and under a simple random design. Then the design effect is defined by $DEFF = \text{Var}(\hat{\theta})/\text{Var}_{\text{SRS}}(\hat{\theta})$. In the case where θ is a sum or a mean of a variable y, the design effect of a clustered sample is related to the intraclass correlation of y, ρ, by $DEFF \approx 1 + (\bar{b} - 1)\rho$, where $\bar{b}$ is the average number of elements drawn from each cluster (see Kish, 1965). Thus,

$$\frac{\text{Var}_{\text{SRS}}(\hat{\theta}) - \text{Var}(\hat{\theta})}{\text{Var}(\hat{\theta})} \approx \frac{(\bar{b} - 1)\rho}{DEFF}$$

is the relative bias when failing to account for the design.

4 Simulation Study

4.1 Simulation Set-up

As mentioned in the introduction, failure to account for the clustering in complex survey data may result in negatively biased variance estimates. Subjects within clusters (primary sampling units) tend to be more alike than ones in different clusters. In order to generate data with clustering, we first assumed the following model for a continuous response variable $x_{h,i,j}$, for subject i in primary sampling unit (PSU) i, in stratum h.

$$x_{h,i,j} = \frac{\phi}{\phi^2 + 1}\varepsilon_{h,i} + \frac{1}{\phi^2 + 1}\varepsilon_{h,i,j},$$

$\varepsilon_{h,i} \overset{iid}{\sim} N(0,1)$, $\varepsilon_{h,i,j} \overset{iid}{\sim} N(0,1)$, and where the $\varepsilon_{h,i,j}$s are independent of the $\varepsilon_{h',i'}$s. The parameter ϕ determines the clustering in the data. Clearly, when $\phi = 0$, no clustering is present—the intra cluster correlation (ICC) is zero. When ϕ grows, so does the ICC. Next, we created two discrete variables Y_{T1} and Y_{T2} from x by defining

$$(Y_{T1}, Y_{T2}) = \begin{cases} (0,0) & \text{if } x < \Phi^{-1}(0.4) \\ (0,1) & \text{if } \Phi^{-1}(0.4) \leq x < \Phi^{-1}(0.5) = 0 \\ (1,0) & \text{if } 0 \leq x < \Phi^{-1}(0.7) \\ (1,1) & \text{if } \Phi^{-1}(0.7) \leq x, \end{cases} \tag{5}$$

where Φ is the cumulative function of the standard normal distribution.

Clearly, as $|\phi|$ increases, so do the ICC of the dicretized variables Y_{T1} and Y_{T2}.

We drew our observations from the stratified and clustered infinite population given in (5), using a with-replacement (WR) at the first stage and a simple random sample (SRS) in the second stage. The common assumptions made in the analysis of the NSDUH data are a WR selection in the first stage, and an SRS selection in the second stage[1]. In the simulation results below, the number of strata was $L = 2$, with 8 PSUs drawn from each, and 10 individuals drawn from each PSU.

4.2 Simulation Results—$\phi = 1$

Figures 1 and 2 show histograms of the variance estimates from the proposed method (under the title of "TL") and those from the standard (Fleiss *et al.*, 1969) formula (under the title of ASE[2]) when $\phi = 1$. The empirical variance (defined as the variance of the 500 simulations: $(500-1)^{-1}\sum_{m=1}^{500}\left(\hat{\kappa}_m - \bar{\hat{\kappa}}\right)^2$) is also shown by a red line. While the estimates from the proposed method are distributed around the empirical mean, those from the standard formula show a clear negative bias.

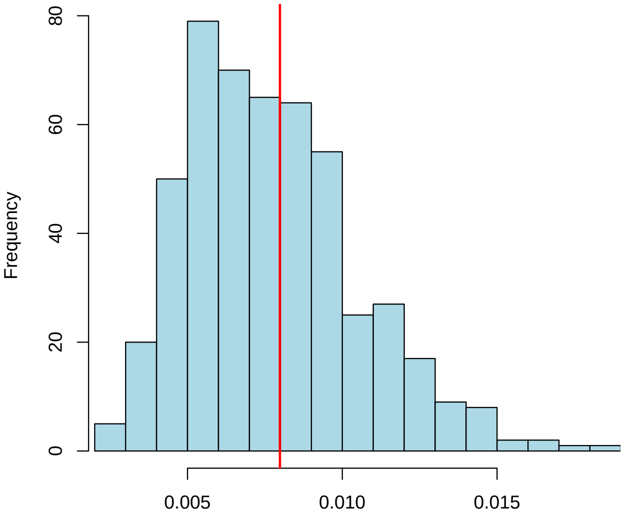

Figure 1: Histograms of Taylor Linearization Variance Estimates—Clustering Effect Present

[1] This is not the actual design, of course.

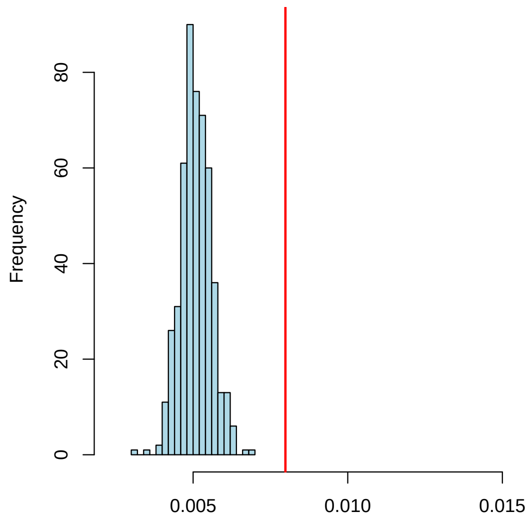

Figure 2: Histograms of Fleiss et al. (1969) Variance Estimates—Clustering Effect Present

4.3 Simulation Results—$\phi = 0$

Figures 3 and 4 show histograms of the variance estimates from the two methods when $\phi = 1$, along with the empirical variance. Both sets of estimates are distributed around the empirical mean. Also evident is that the estimates from the standard formula are less dispersed. Thus, when no clustering is present, one can, and in fact should, ignore the design (as the variance estimation would then be more efficient).

Figure 5 shows the bias as a function of ϕ.

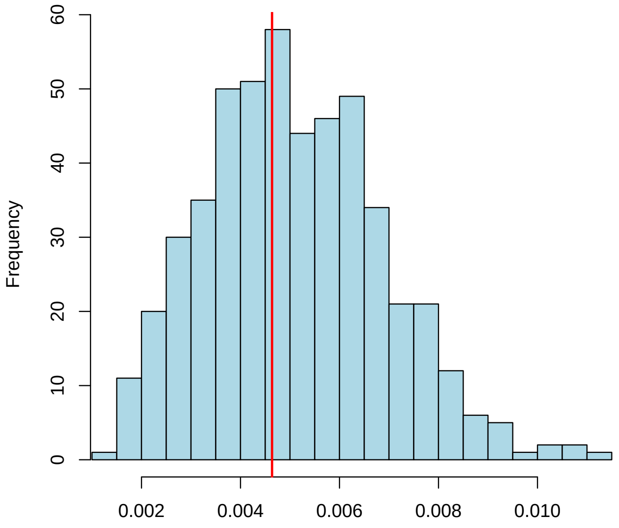

Figure 3: Histograms of Taylor Linearization Variance Estimates—No Clustering Effect

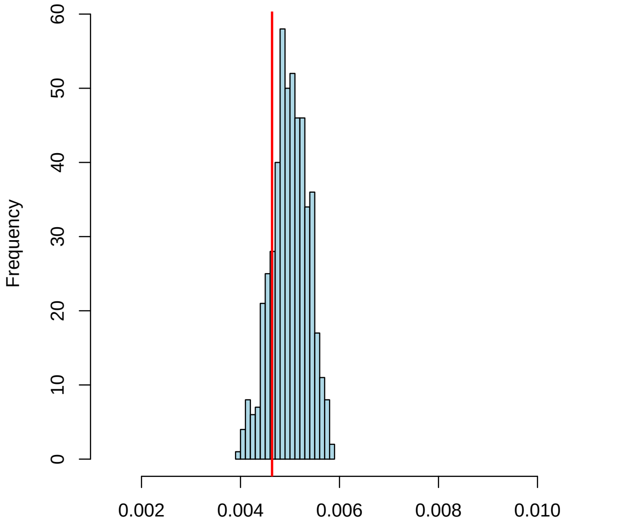

Figure 4: Histograms of Fleiss et al. (1969) Variance Estimates—No Clustering Effect

Acknowledgments

This research was developed with the support of the Substance Abuse and Mental Health Services Administration, Office of Applied Studies, under Contract no. 283-2004-00022.

The views expressed in this presentation do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention of trade names, commercial practices, or organizations imply endorsement by the U.S. Government

The simulations were carried out in the statistical software package R (www.R-project.org).

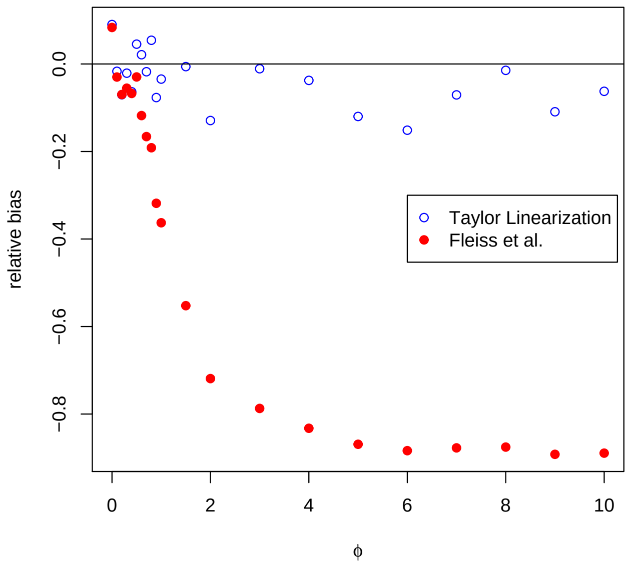

Figure 5: Bias of the Standard Formula as a Function of ϕ

References

Agresti, A. (2002), Categorical Data Analysis. Wiley.

Cohen J. (1960), A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 37–46.

Fleiss, J.L., Cohen, J.; Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72**(5), 323–327.

Kish, L. (1965). Survey Sampling. New York: Wiley.

Research Triangle Institute (2004). SUDAAN Language Manual, Release 9.0. Research Triangle Park, NC: Research Triangle Institute.