Article

Educational and Psychological
Measurement
1–10
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0013164420973080
journals.sagepub.com/home/epm
SAGE

Large-Sample Variance of Fleiss Generalized Kappa

Kilem L. Gwet[1]

Abstract

Cohen's kappa coefficient was originally proposed for two raters only, and it later extended to an arbitrarily large number of raters to become what is known as Fleiss' generalized kappa. Fleiss' generalized kappa and its large-sample variance are still widely used by researchers and were implemented in several software packages, including, among others, SPSS and the R package "rel." The purpose of this article is to show that the large-sample variance of Fleiss' generalized kappa is systematically being misused, is invalid as a precision measure for kappa, and cannot be used for constructing confidence intervals. A general-purpose variance expression is proposed, which can be used in any statistical inference procedure. A Monte-Carlo experiment is presented, showing the validity of the new variance estimation procedure.

Keywords

Fleiss kappa, Cohen kappa, interrater reliability, Gwet AC1

Introduction

Cohen (1960) introduced the kappa coefficient as a measure of the extent of agreement among two raters, which adjusts for the possibility of agreement by pure chance. This measure became popular among researchers and underwent several extensions. Fleiss (1971), Light (1971), Krippendorff (1970), and Conger (1980), among others, proposed various extensions of Cohen's kappa to multiple raters. For several decades, researchers used Cohen's kappa and its generalizations despite its many well-documented deficiencies. Feinstein and Cicchetti (1990) as well as Cicchetti and Feinstein (1990) discussed several situations where kappa produces an unduly low agreement coefficient when the raters are in an almost perfect agreement.

[1] AgreeStat Analytics, Gaithersburg, MD, USA

Corresponding Author:
Kilem L. Gwet, AgreeStat Analystics, P.O. Box 2696, Gaithersburg, MD 20886, USA.
Email: gwet@agreestat.com

To avoid these kappa paradoxes, alternative agreement coefficients were developed. Among others, Gwet (2008) proposed the AC_1 coefficient, which is gaining in popularity and has been implemented in some major statistical packages such as SAS.

All agreement coefficients mentioned in the previous paragraph are summary measures that represent one possible approach to analyzing agreement data. Alternative approaches often used by researchers are based on statistical models that include latent class models and quasi-symmetric log-linear models. Log-linear models are extensively discussed by von Eye and Mun (2005), while latent class models are discussed by Schuster and Smith (2002) and Raykov et al. (2013). A key advantage of statistical models is their ability to describe the structure of the joint distribution of ratings and to test specific hypotheses that cannot be investigated with summary measures. Schuster (2002) also discussed a different type of models known as mixture models, which provide a hybrid approach between traditional summary statistics and log-linear models.

This article focuses on Fleiss' generalized kappa and its variance estimation. Although Fleiss (1971) labeled his coefficient as the generalized kappa, it does not reduce to Cohen's kappa when the number of raters is 2. Instead, it reduces to the pi coefficient proposed earlier by Scott (1955). In that sense, Fleiss' generalized kappa is strictly speaking an extension of Scott's pi coefficient. The standard errors provided by Fleiss (1971) were later deemed to be incorrect by Fleiss et al. (1979). The revised standard error provided by Fleiss et al. (1979), which is still widely being used today, is not meant to be used for quantifying the precision of Fleiss' kappa. Instead, it should be used solely for testing the hypothesis of zero agreement among raters. Equation (12) of Fleiss et al. (1979) is valid only under the assumption that there is no agreement among raters. If this assumption of no agreement is not satisfied, this equation becomes irrelevant.

On the very first page of their paper, Fleiss et al. (1979) indicated the following:

> In this article, formulas for the standard error of kappa in the case of different sets of equal numbers of raters that are valid when the number of subjects is large and the null hypothesis is true are derived.

The goal that Fleiss et al. (1979) set for their paper was clearly stated on the first page. Moreover, when these authors started the standard error derivation on page 2, they specifically made the following assumption:

> Consider the hypothesis that the ratings are purely random in the sense that for each subject, the frequencies $n_{i1}, n_{i2}, \ldots, n_{ik}$ are a set of multinomial frequencies with parameters n and $(P_1, P_2, \ldots, P_k)$, where $\sum P_j = 1$..

Since the ratings are assumed to be purely random, no agreement should be expected to occur beyond chance. This is not a problem, as long as the researcher is solely concerned about testing the null hypothesis of no agreement. But if precision measures are to be reported, or confidence intervals constructed, then the standard error proposed by Fleiss et al. (1979) should be avoided. For unknown reasons, researchers in various fields of research have been using this standard error as a measure of precision. The implementation of this standard error in a major statistical software such as SPSS® or in the R package ‘‘rel’’ must have contributed to its widespread misuse.

Notation

Let us consider an interrater reliability experiment, which involves n subjects, r raters, and q categories into which each of the r raters is expected to classify all n subjects (there could be missing ratings in case some raters do not rate all subjects, but we will ignore these practical considerations for now). A total of r_{ik} out of r raters have classified subject i into category k. Let π_k be the probability for a random rater to classify a random subject into category k. The exact value of π_k will generally be unknown. However, once rating data are collected, one would typically replace the unknown π_k with its estimated value $\widehat{\pi}_k$ defined as follows:

$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} r_{ik}/r. \tag{1}$$

Note that π_k represents the theoretical value to which $\widehat{\pi}_k$ converges (in probability) as the number of subjects n increases. The distinction between these two quantities will become essential later in this paper, when deriving the large-sample variance of Fleiss’ coefficient. For simplicity, $\pi_k^\star$ will denote the complement of π_k, given by $1-\pi_k$. Likewise, the complement of $\widehat{\pi}_k$ would labeled as $\widehat{\pi}_k^\star$.

Fleiss’ Kappa and Its Variance

Fleiss’ generalized kappa coefficient is defined as follows:

$$\widehat{\kappa} = \frac{p_a - p_e}{1-p_e} \text{ where } \begin{cases} p_a = \frac{1}{n}\sum\limits_{i=1}^{n}\sum\limits_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)}, \\ p_e = \sum\limits_{k=1}^{q}\widehat{\pi}_k^2. \end{cases} \tag{2}$$

In Equation 2, p_a and p_e represent the percent agreement and the percent chance agreement, respectively. Fleiss’ generalized kappa is denoted by $\widehat{\kappa}$ in Equation 2 and will converge (in probability) to a fixed quantity labeled as κ.

The variance estimate that Fleiss et al. (1979) proposed is given by

$$var_{\mathrm{F}}(\widehat{\kappa}) = \frac{2}{nr(r-1)(\sum\limits_{k=1}^{q}\widehat{\pi}_k\widehat{\pi}_k^\star)^2}[(\sum_{k=1}^{q}\widehat{\pi}_k\widehat{\pi}_k^\star)^2 - \sum_{k=1}^{q}\widehat{\pi}_k\widehat{\pi}_k^\star(\widehat{\pi}_k^\star - \widehat{\pi}_k)] \tag{3}$$

Equation 3 is the expression many researchers currently use for estimating the variance of Fleiss' generalized kappa coefficient. It is a direct function of the classification probabilities $\widehat{\pi}_k$ and is independent of the kappa coefficient itself. This is due to the assumption of no agreement among raters, which underlies the derivation of this variance, making it suitable for hypothesis testing only.

The next section is devoted to the derivation of a general-purpose variance estimator that can be used as a precision measure as well as for hypothesis testing or any other inferential procedure involving Fleiss' generalized kappa coefficient.

Proposed Variance

To compute the variance of Fleiss' generalized kappa of Equation 2, the following equation is recommended:

$$var_{\mathrm{G}}(\widehat{\kappa}) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\kappa_i^{\star} - \widehat{\kappa})^2, \tag{4}$$

where

$$\kappa_i^{\star} = \kappa_i - 2(1-\widehat{\kappa})\frac{p_{e|i} - p_e}{1-p_e}, \tag{5}$$

with $\kappa_i = (p_{a|i} - p_e)/(1-p_e)$. Moreover, $p_{a|i}$ and $p_{e|i}$ representing the percent agreement and percent chance agreement evaluated on subject i alone, are, respectively, given by

$$p_{a|i} = \sum_{k=1}^{q}\frac{r_{ik}(r_{ik}-1)}{r(r-1)} \quad \text{and} \quad p_{e|i} = \sum_{k=1}^{q}\widehat{\pi}_k r_{ik}/r. \tag{6}$$

The f factor of Equation 4 represents the sampling fraction (i.e., $f = n/N$) to be used if the n subjects were selected from a larger universe of N subjects. In the finite population sampling literature, $1-f$ is often referred to as the finite population correction. This correction may be useful if the n sample subjects are selected from a universe of N subjects of small to moderate size. Otherwise, one can safely set $f = 0$.

Large Sample Approximation

This section outlines the general steps leading to the formulation of the variance given by Equation 4. A rigorous mathematical proof of the validity of this equation would be out of the scope of this paper. Instead, the derivations are carried out heuristically and a Monte-Carlo simulation study is presented to provide an empirical validation.

The general strategy consists of using the linearization method to demonstrate that for large subject samples, $\widehat{\kappa}$ has the same probability distribution as a linear expression, which is a simple average of subject-level values and whose variance is simple to derive. That common limit probability distribution is in fact Normal.

Dealing With the Denominator of Fleiss' Kappa

One reason calculating the variance of Fleiss' kappa appears challenging is the term $1 - p_e$ that appears in its denominator as seen in Equation 2. The first thing to do is to show that the large-sample distribution of kappa is the same as the large-sample distribution of a statistic that does not involve any sample dependent term in the denominator. This is accomplished by realizing that if $\widehat{\pi}_k$ converges in probability to π_k for each category k (i.e., $\widehat{\pi}_k \overset{P}{\rightarrow} \pi_k$) then it follows from the Continuous Mapping Theorem[1] that p_e will converge (in probability) to P_e defined by

$$P_e = \sum_{k=1}^{q} \pi_k^2$$

Convergence in probability will often be referred to mathematically as $p_e \overset{P}{\rightarrow} P_e$. It follows that

$$1/(1 - p_e) = \frac{1}{(1 - P_e)(1 - \varepsilon_n)}, \tag{7}$$

where $\varepsilon_n = (p_e - P_e)/(1 - P_e)$. Since $\varepsilon_n \overset{P}{\rightarrow} 0$, it follows from Taylor's theorem[2] that for large samples $1/(1 - \varepsilon_n) = 1 + \varepsilon_n + \mathit{Remainder}$ where the remainder goes to 0 (in probability) at a faster rate than ε_n. In large sample theory, this equation is often written as $1/(1 - \varepsilon_n) = 1 + \varepsilon_n + o_p(\varepsilon_n)$ to indicate that the remainder goes to 0 faster than ε_n. It follows from Slutsky's theorem[3] that the large-sample probability distribution of $1/(1 - p_e)$ is the same as the large-sample distribution of L_e defined by

$$L_e = (1 + \varepsilon_n)/(1 - P_e).$$

Consequently, Fleiss' kappa has the same large-sample probability distribution as the quantity κ_0, given by,

$$\kappa_0 = (p_a - p_e)L_e. \tag{8}$$

Since κ_0 involves unknown terms such as P_e, it cannot be evaluated. But this should not be a problem, since it is its probability distribution that is of interest to us for the time being. Unknown quantities will be dealt with at the time of estimating the probability distribution variance after it has been clearly formulated.

Dealing With the Percent Agreement

The percent agreement p_a is defined in Equation 2 and can be seen as a sample mean obtained by averaging the $p_{a|i}$ values defined by Equation 6. It follows from the Law of Large Numbers[4] that p_a converges in probability to a fixed quantity denoted by P_a. That is, $p_a \xrightarrow{P} P_a$. It follows that κ_0 of Equation 8 can be rewritten as follows:

$$\kappa_0 = \frac{p_a - P_e}{1 - P_e} - (1 - \kappa)\frac{p_e - P_e}{1 - P_e}, \tag{9}$$

where $\kappa = (P_a - P_e)/(1 - P_e)$. In Equation 9, only p_a and p_e are sample dependent. The other terms are fixed and are not subject to any variation. While the percent agreement p_a is a regular sample mean of subject-level values, this is not the case for the percent chance agreement p_e (see Equation 2). Therefore, the percent chance agreement must be further processed to linearize it.

Linearization of the Percent Chance Agreement

As previously indicated, the estimated propensity $\widehat{\pi}_k$ for classification into category k converges in probability toward the fixed quantity π_k. It follows from the Taylor's theorem that in the neighborhood of π_k, $\widehat{\pi}_k$ can be expressed as follows: $\widehat{\pi}_k^2 = \pi_k^2 + 2\pi_k(\widehat{\pi}_k - \pi_k) + \mathit{Remainder}$, where the remainder is term that converges (in probability) toward 0, faster than the difference $\widehat{\pi}_k - \pi_k$ as the number of subjects increases. Consequently, it follows from another application of Slutsky's theorem that the large-sample distribution of the difference $p_e - P_e$ of Equation 9 is the same as that of $2(p_{e|0} - P_e)$ where $p_{e|0}$ is given by

$$p_{e|0} = \sum_{k=1}^{q} \pi_k \widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} p_{e|i}, \text{ where } p_{e|i} = \sum_{k=1}^{q} \pi_k r_{ik}/r. \tag{10}$$

Therefore, the large-sample distribution of κ_0 of Equation 9 is the same as the distribution of κ_1 given by

$$\kappa_1 = \frac{p_a - P_e}{1 - P_e} - 2(1 - \kappa)\frac{p_{e|0} - P_e}{1 - P_e} = \frac{1}{n}\sum_{i=1}^{n} \kappa_i^\star, \tag{11}$$

where $\kappa_i^\star$ is defined by

$$\kappa_i^\star = \kappa_i - 2(1 - \kappa)\frac{p_{e|i} - P_e}{1 - P_e}. \tag{12}$$

Equation 11 is the linear expression that was needed. It follows from the Central Limit Theorem that the large-sample probability distribution of κ_1 is Normal with mean κ and a variance that can be estimated with Equation 4.

Monte Carlo Simulation

This article establishes that the large-sample probability distribution of Fleiss' generalized kappa is Normal with a mean at its expected value κ and a variance that can be estimated by Equation 4. The way to verify the accuracy of this result is conduct a simulation. A universe of subjects of a certain size N must first be created along with the ratings as if the raters rated that entire universe. A Fleiss' generalized kappa will be calculated from the universe data to obtain the κ value (i.e., the fixed population parameter that will have to be estimated from smaller samples). For each given sample size n, a large number of samples of n subjects will selected from the universe. Using the corresponding ratings, the sample-based Fleiss' kappa $\widehat{\kappa}$ will be computed along with the associated 95% confidence interval as follows:

$$[\widehat{\kappa} - 1.96\sqrt{v(\widehat{\kappa})}\ ;\ \widehat{\kappa} + 1.96\sqrt{v(\widehat{\kappa})}] \tag{13}$$

Alternatively, for a better coverage rate, one may use the Student's critical value $t_{0.05}(n-1)$, which decreases with the sample size n, as opposed to the fixed value 1.96. Both critical values will get closer and closer as the number of subjects n in the sample grows. The confidence interval based on the Student's distribution is calculated as follows:

$$[\widehat{\kappa} - t_{0.05}(n-1)\sqrt{v(\widehat{\kappa})}\ ;\ \widehat{\kappa} + t_{0.05}(n-1)\sqrt{v(\widehat{\kappa})}] \tag{14}$$

Confidence intervals were evaluated in this Monte Carlo experiment based on Equation 14.

For each sample size n, a long series of such confidence intervals will be constructed, and coverage of the population kappa κ by each of them will be checked. If the variance formula is correct, then the coverage rate is expected to be close to its nominal value of 95% for each value of the sample size n. As the value of n increases, the interval coverage rate is expected to get closer and closer to 95%.

Population Data

Initially, a dataset of 3,500 subjects and seven raters was created. Each rater has to classify all 3,500 subjects into one of five possible categories. The process of creating this initial dataset was set up so that each subject would be classified into one randomly chosen category with probability 0.8 and will be classified into each of the remaining four categories with the same probability of 0.05. These 3,500 represent our population of subjects from which small subject samples will be selected~~,~~ ~~and the~~ population value Fleiss' kappa is $\kappa = 0.5612$.

Results

This investigation was conducted for different values of n varying from 10 to 50 with a 5 increment. For each of these sample size values, 100,000 samples were selected from the universe of 3,500 subjects. For each of the 100,000 samples 2 confidence intervals were calculated. One interval was based on the new variance Equation 4 also referred to as Var2020 in this section. The second interval was based on the

Table 1. Results of the Monte Carlo Experiment.

n	Mean kappa	Mean Var1979	Mean Var2020	True variance	Coverage Var1979	Coverage Var2020
10	0.532	0.00144	0.00883	0.00920	60.1%	93.0%
15	0.542	0.00090	0.00579	0.00587	58.0%	93.8%
20	0.547	0.00065	0.00430	0.00436	56.9%	94.1%
25	0.550	0.00051	0.00342	0.00347	56.2%	94.2%
30	0.552	0.00042	0.00284	0.00284	56.1%	94.5%
35	0.553	0.00036	0.00243	0.00244	55.7%	94.5%
40	0.555	0.00031	0.00212	0.00210	55.7%	94.7%
45	0.555	0.00028	0.00188	0.00186	55.6%	94.8%
50	0.556	0.00025	0.00169	0.00166	55.8%	94.9%

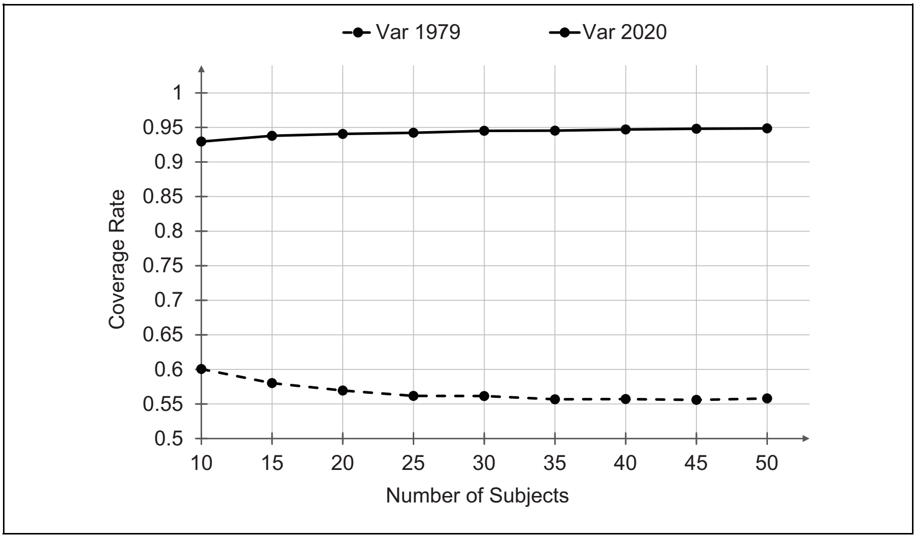

Figure 1. Monte Carlo simulation coverage rates.

variance proposed by Fleiss et al. (1979), also referred to as Var1979 in this section. The results of this experiment are shown in Table 1 and depicted in Figure 1

The first column of Table 1 shows the different sample sizes used. The second column labeled as "Mean Kappa" represents for each sample size *n*, the average of all 100,000 kappa values generated. Columns "Mean Var1979" and "Mean Var2020" represent the mean values of all 100,000 Var1979 (see Equation 3) and Var2020 (see Equation 4) calculated for a given sample size. The "True Variance" column is the Monte-Carlo variance calculated as follows:

$$\text{True Variance} = \frac{1}{100,000}\sum_{s=1}^{100,000}(\widehat{\kappa}_s - \overline{\widehat{\kappa}})^2,$$

where $\overline{\widehat{\kappa}}$ is the average of all 100,000 replicate $\widehat{\kappa}_s$ values. The last two columns represent the two coverage rates of the confidence intervals based of Var1979 and Var2020, respectively.

It follows from Table 1 and Figure 1 that even for sample sizes as small as $n = 10$ the coverage rate~~s~~ of ~~the~~ confidence intervals based on the recommended variance of Equation 4 is very close to its nominal value of 95%. ~~These~~ rate~~s~~ ~~get~~ closer and closer to this nominal value as the sample size increases. However, the coverage rate~~s~~ of ~~the~~ confidence intervals based of the variance proposed by Fleiss et al. (1979) ~~are~~ dramatically low. Figure 1 even suggests that this coverage rate tends to decrease as the sample size increases. This is an indication that this variance formula is invalid for calculating confidence intervals. Table 1 also shows that Mean2020 is consistently close to the ''True Variance'' for all values of the sample size.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Kilem L. Gwet https://orcid.org/0000-0001-7968-1432

Notes

1. The Continuous Mapping Theorem stipulates that any continuous function preserves the stochastic convergence of a sequence of random variables.
2. Taylor's theorem approximates any differentiable function by a linear function in the neighborhood of a given point.
3. Slutsky's theorem is well known in probability theory. It stipulates that if a sequence of random variables X_n converges in probability to a constant value c and the large-sample distribution of another sequence Y_n is the same as the distribution of a random variable Y, then the large-sample distribution of any continuous function $g(X_n, Y_n)$ is the same as the distribution of $g(c, Y)$.
4. The Law of Large Numbers, in its ''weak'' version due to Aleksandr Khinchin (1894-1959) stipulates that the sample average converges in probability toward its expected value.

References

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551-558. https://doi.org/10.1016/0895-4356(90)90159-M

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Conger, A. J. (1980). Integration and generalization of kappas for multiple raters. *Psychological Bulletin*, *88*(2), 322-328. https://doi.org/10.1037/0033-2909.88.2.322

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543-549. https://doi.org/10.1016/0895-4356(90)90158-L

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Fleiss, J. L., Nee, J. C. M., & Landis, J. R. (1979). The large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, *86*(5), 974-977. https://doi.org/10.1037/0033-2909.86.5.974

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29-48. https://doi.org/10.1348/000711006X126600

Krippendorff, K. (1970). Estimating the reliability, systematic error, and random error of interval data. *Educational and Psychological Measurement*, *30*(1), 61-70. https://doi.org/10.1177/001316447003000105

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin*, *76*(5), 365-377. https://doi.org/10.1037/h0031643

Raykov, T., Dimitrov, D. M., von Eye, A., & Marcoulides, G. A. (2013). Interrater agreement evaluation: A latent variable modeling approach. *Educational and Psychological Measurement*, *73*(3), 512-531. https://doi.org/10.1177/0013164412449016

Schuster, C. (2002). A mixture model approach to indexing rater agreement. *British Journal of Mathematical and Statistical Psychology*, *55*(2), 289-303. https://doi.org/10.1348/000711002760554598

Schuster, C., & Smith, D. (2002). Indexing systematic rater agreement with a latent-class model. *Psychological Methods*, *7*(3), 384-395. https://doi.org/10.1037/1082-989X.7.3.384

Scott, W. A. (1955). Reliability of content analysis: the case of nominal scale coding. *Public Opinion Quarterly*, *19*(3), 321-325. https://doi.org/10.1086/266577

von Eye, A., & Mun, E. Y. (2005). *Analyzing rater agreement: Manifest variable methods*. Erlbaum.