Psychological Bulletin
1971, Vol. 76, No. 5, 378-382

MEASURING NOMINAL SCALE AGREEMENT AMONG MANY RATERS [1]

JOSEPH L. FLEISS [2]

Biometrics Research, New York State Department of Mental Hygiene

The statistic kappa was introduced to measure nominal scale agreement between a fixed pair of raters. In this paper kappa is generalized to the case where each of a sample of subjects is rated on a nominal scale by the same number of raters, but where the raters rating one subject are not necessarily the same as those rating another. Large sample standard errors are derived, and a numerical example is given.

Cohen introduced the statistics kappa (Cohen, 1960) and weighted kappa (Cohen, 1968) to measure the degree of agreement between two raters who rate each of a sample of subjects on a nominal scale. Both kappa and weighted kappa incorporate a correction for the extent of agreement expected by chance alone. Kappa is useful when all disagreements may be considered equally serious, and weighted kappa is useful when the relative seriousness of the different kinds of disagreement can be specified. Properties of these two statistics have been studied by Everitt (1968) and by Fleiss, Cohen, and Everitt (1969).

The use of kappa and weighted kappa is restricted to the case both where the number of raters is two and where the same two raters rate each subject. Generalizations are therefore necessary for the case of more than two raters and for the case where the raters judging one subject are not necessarily the same as those judging another.

In this paper we consider only the generalization of unweighted kappa to the measurement of agreement among any constant number of raters where there is no connection between the raters judging the various subjects. An important special case is when subjects are rated by different *pairs* of raters. We do not consider the case of varying numbers of raters, or the case where each subject is rated by the same group of more than two raters, or the case where weights can be assigned to the various disagreements.

NOTATION

Let N represent the total number of subjects, n the number of ratings per subject, and k the number of categories into which assignments are made. Let the subscript i, where $i = 1, \ldots, N$, represent the subjects, and the subscript j, where $j = 1, \ldots, k$, represent the categories of the scale.

Define n_{ij} to be the number of raters who assigned the ith subject to the jth category, and define

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}. \qquad [1]$$

The quantity p_j is the proportion of all assignments which were to the jth category. Since $\Sigma_j n_{ij} = n$, therefore $\Sigma_j p_j = 1$.

Table 1 presents data on which of $k = 5$ diagnoses were given each of $N = 30$ patients by $n = 6$ psychiatrists. Thus, $n_{14} = 6$ and $n_{1j} = 0$ for $j \neq 4$; $n_{22} = 3$, $n_{25} = 3$, and $n_{2j} = 0$ for $j \neq 2$ and $j \neq 5$; etc. The overall proportions are $p_1 = .144$, $p_2 = .144$, $p_3 = .167$, $p_4 = .306$, and $p_5 = .239$. A total of 43 psychiatrists provided diagnoses. In the actual study (Sandifer, Hordern, Timbury, & Green, 1968), between 6 and 10 psychiatrists from the pool of 43 were unsystematically selected to diagnose a subject. For the illustrative purposes of this paper, randomly selected diagnoses were dropped to bring the number of assignments per patient down to a constant of six.

[1] This work was supported in part by Grant MH 08534 and in part by Grant MH 09191, both from the National Institute of Mental Health. Myron G. Sandifer, Jr., Associate Dean for Academic Affairs of the University of Kentucky College of Medicine, kindly provided the data on which Table 1 is based.

[2] Also at Columbia University. Requests for reprints should be sent to Joseph L. Fleiss, Biometrics Research, State of New York Department of Mental Hygiene, 722 West 168 Street, New York, N. Y. 10032.

Measuring Overall Agreement

The extent of agreement among the n raters for the ith subject may be indexed by the proportion of agreeing pairs out of all the $n(n-1)$ possible pairs of assignments. This proportion is, say,

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij} - 1)$$

$$= \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right). \quad [2]$$

Thus, $P_1 = 1$; $P_2 = .40$; etc. The overall extent of agreement may then be measured by the mean of the P_is, say,

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i$$

$$= \frac{1}{Nn(n-1)} \left(\sum_{i=1}^{N} \sum_{j=1}^{k} n_{ij}^2 - Nn \right). \quad [3]$$

For the data of Table 1,

$$\bar{P} = \frac{1}{(30)(6)(5)}(680 - 30 \times 6) = .5556. \quad [4]$$

The value $\bar{P} = .5556$ may be interpreted as follows. Let a subject be selected at random and diagnosed by a randomly selected psychiatrist. If the subject were also diagnosed by a second randomly selected psychiatrist, the second diagnosis would agree with the first over 55% of the time. Measures like $\bar{P}$ have been used to measure diagnostic agreement by Sandifer, Pettus, and Quade (1964), by Sandifer, Hordern, Timbury, and Green (1968) and by Sandifer, Fleiss, and Green (1968).

It is clear, however, that some degree of agreement is to be expected solely on the basis of chance. In fact, if the raters made their assignments purely at random, one would expect the mean proportion of agreement to be

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2. \quad [5]$$

For the data of Table 1,

$$\bar{P}_e = .144^2 + .144^2 + \cdots + .239^2 = .2201. \quad [6]$$

TABLE 1

Diagnoses on 30 Subjects by Six Raters per Subject

Subject	Category				
	Depression ($j=1$)	Personality disorder ($j=2$)	Schizophrenia ($j=3$)	Neurosis ($j=4$)	Other ($j=5$)
1				6	
2		3			3
3		1	4		1
4					6
5		3		3	
6	2		4		
7			4		2
8	2		3	1	
9	2			4	
10					6
11	1			5	
12	1	1		4	
13		3	3		
14	1			5	
15		2		3	1
16			5		1
17	3			1	2
18	5	1			
19		2		4	
20	1		2		3
21					6
22		1		5	
23		2		1	3
24	2			4	
25	1			4	1
26		5		1	
27	4				2
28		2		4	
29	1		5		
30					6
Total	26	26	30	55	43
p_j	.144	.144	.167	.306	.239

The quantity $1 - \bar{P}_e$ measures the degree of agreement attainable over and above what would be predicted by chance. The degree of agreement actually attained in excess of chance is $\bar{P} - \bar{P}_e$, so that a normalized measure of overall agreement, corrected for the amount expected by chance, is

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}. \quad [7]$$

For the data of Table 1,

$$\kappa = \frac{.5556 - .2201}{1 - .2201} = .430. \quad [8]$$

An equivalent expression for κ is

$$\kappa = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^2 - Nn[1 + (n-1)\sum_{j=1}^{k} p_j^2]}{Nn(n-1)(1 - \sum_{j=1}^{k} p_j^2)} \quad [9]$$

Standard Error of κ

It seems reasonable to interpret the absence of real agreement among the n raters as their inability to distinguish one subject from another. Such an inability implies that the raters apply the overall rates of assignment, $\{p_j\}$, to each and every subject. This in turn implies that under the hypothesis of no agreement, the n assignments on each subject are multinomial variables with probabilities $p_1, \ldots, p_k$.

If N, the number of subjects, is large, we may assume that the proportions $p_1, \ldots, p_k$ given by Equation 1 are constants so that by Equation 9, κ is essentially a function only of the random variable $\Sigma_i \Sigma_j n_{ij}^2$. We assume further that even though $p_1, \ldots, p_k$ are held constant, N is large enough so that the sum of the squared numbers of assignments for the ith subject, $\Sigma_j n_{ij}^2$, is independent of that for the i'th subject, $\Sigma_j n_{i'j}^2$. Thus, if N is large, the variance of κ under the hypothesis of no agreement beyond chance is approximately equal to

$$\mathrm{Var}(\kappa) = \frac{\sum_{i=1}^{N} \mathrm{Var}(\sum_{j=1}^{k} n_{ij}^2)}{N^2 n^2 (n-1)^2 (1 - \sum_{j=1}^{k} p_j^2)^2} \quad [10]$$

Now,

$$\mathrm{Var}(\sum_j n_{ij}^2) = E(\sum_j n_{ij}^2)^2 - [E(\sum_j n_{ij}^2)]^2$$
$$= E(\sum_j n_{ij}^4) + E(\sum\sum_{j \neq m} n_{ij}^2 n_{im}^2) - [E(\sum_j n_{ij}^2)]^2. \quad [11]$$

The required moments are found to be

$$E(\sum_j n_{ij}^4) = n + n(n-1) \times [7\sum_j p_j^2 + 6(n-2)\sum_j p_j^3 + (n-2)(n-3)\sum_j p_j^4], \quad [12]$$

$$E(\sum\sum_{j \neq m} n_{ij}^2 n_{im}^2) = n(n-1) + n(n-1)[(2n-5)\sum_j p_j^2 + (n-2)(n-3)(\sum_j p_j^2)^2 - 2(n-2)\sum_j p_j^3 - (n-2)(n-3)\sum_j p_j^4], \quad [13]$$

and

$$[E(\sum_j n_{ij})]^2 = n^2 + n(n-1) \times [2n\sum_j p_j^2 + n(n-1)(\sum_j p_j^2)^2]. \quad [14]$$

Thus,

$$\mathrm{Var}(\sum_j n_{ij}^2) = 2n(n-1)[\sum_j p_j^2 - (2n-3)(\sum_j p_j^2)^2 + 2(n-2)\sum_j p_j^3], \quad [15]$$

and, because $\Sigma_i \mathrm{Var}(\Sigma_j n_{ij}^2) = N \mathrm{Var}(\Sigma_j n_{ij}^2)$, therefore

$$\mathrm{Var}(\kappa) = \frac{2}{Nn(n-1)} \times \frac{\sum_j p_j^2 - (2n-3)(\sum_j p_j^2)^2 + 2(n-2)\sum_j p_j^3}{(1 - \sum_j p_j^2)^2}. \quad [16]$$

For the data of Table 1, we have already found that $\Sigma_j p_j^2 = .2201$. Further, $(\Sigma_j p_j^2)^2 = .0484$ and $\Sigma_j p_j^3 = .0529$. Thus, the variance of κ is approximately

$$\mathrm{Var}(\kappa) = \frac{2}{(30)(6)(5)} \times \frac{.2201 - (9)(.0484) + (2)(4)(.0529)}{(1 - .2201)^2} = .000759, \quad [17]$$

and the standard error (SE) of κ is approximately

$$\mathrm{SE}(\kappa) = .028. \quad [18]$$

Under the hypothesis of no agreement beyond chance, $\kappa/\mathrm{SE}(\kappa)$ will, by the central limit theorem, be approximately distributed as a

standard normal variate. In our case,

$$\frac{\kappa}{\mathrm{SE}(\kappa)} = \frac{.430}{.028} = 15.4, \qquad [19]$$

so we would infer that the overall agreement in assignment to the k categories is significantly greater than chance.

Agreement on a Particular Category

We consider now the extent of agreement in assigning a subject to category j. To motivate the choice of the basic index of agreement, we suppose that each subject is assigned to one of the k categories by a randomly selected rater. Let only those subjects assigned to category j be rated again, now by another randomly selected rater. The index is the conditional probability that the second assignment is to category j, given that the first was to category j. It is calculated as

$$\bar{P}_j = \frac{\sum_{i=1}^{N} n_{ij}(n_{ij} - 1)}{\sum_{i=1}^{N} n_{ij}(n - 1)} = \frac{\sum_{i=1}^{N} n_{ij}^2 - Nnp_j}{Nn(n-1)p_j}, \qquad [20]$$

where p_j is defined by Equation 1.

Under the hypothesis of no agreement beyond chance, we would expect this probability to equal the unconditional probability of an assignment to category j, namely p_j. A measure of the extent of agreement beyond chance in assignment to category j is then

$$\kappa_j = \frac{\bar{P}_j - p_j}{1 - p_j} = \frac{\sum_{i=1}^{N} n_{ij}^2 - Nnp_j[1 + (n-1)p_j]}{Nn(n-1)p_jq_j}, \qquad [21]$$

where $q_j = 1 - p_j$. Note that κ, the measure of overall agreement (Equation 9), is a weighted average of the κ_js. In fact,

$$\kappa = \Sigma_j p_j q_j \kappa_j / \Sigma_j p_j q_j.$$

If N is large, we can assume that p_j is a constant and, even with this restriction, that n_{ij}^2 is independent of $n_{i'j}^2$ for $i \neq i'$. Under the hypothesis of no more than chance agreement in assignment to category j, the quantity n_{ij} has a binomial distribution with parameters n and p_j.

TABLE 2

Statistics for Measuring Agreement on Each of the Five Categories of Table 1

Category	$\Sigma_i n_{ij}^2$	p_j	P_j	κ_j	Var (κ_j)	$\kappa_j/\mathrm{SE}(\kappa_j)$
1	72	.144	.356	.248	.0130	2.17
2	72	.144	.356	.248	.0130	2.17
3	120	.167	.598	.517	.0136	4.44
4	229	.306	.632	.470	.0195	3.36
5	187	.239	.669	.565	.0163	4.43

Under this hypothesis,

$$\mathrm{Var}(n_{ij}^2) = np_jq_j \times [(1 + 2(n-1)p_j)^2 + 2(n-1)p_jq_j]. \qquad [22]$$

Therefore, the approximate variance of κ_j is

$$\mathrm{Var}(\kappa_j) = \frac{[1 + 2(n-1)p_j]^2 + 2(n-1)p_jq_j}{Nn(n-1)^2p_jq_j}. \qquad [23]$$

The hypothesis may be tested by referring the quantity $\kappa_j/\mathrm{SE}(\kappa_j) = \kappa_j/\sqrt{\mathrm{Var}(\kappa_j)}$ to tables of the standard normal distribution.

Table 2 gives summary statistics for each of the five categories. The calculations for Category 1 (consisting of involutional melancholia, manic-depressive depression, and psychotic depression) are illustrated in detail. Since $\Sigma_i n_{i1}^2 = 72$, and $p_1 = .144$, the basic index is, by Equation 20,

$$\bar{P}_1 = \frac{72 - (30)(6)(.144)}{(30)(6)(5)(.144)} = .356, \qquad [24]$$

indicating that, given that one assignment of a subject is to Category 1, the chances of a second one being to the same category are over one in three.

By Equation 21,

$$\kappa_1 = \frac{.356 - .144}{1 - .144} = .248. \qquad [25]$$

The approximate variance of κ_1 is, by Equation 23,

$$\mathrm{Var}(\kappa_1) = \frac{[1 + (2)(5)(.144)]^2 + (2)(5)(.144)(.856)}{(30)(6)(5)^2(.144)(.856)} = .0130, \qquad [26]$$

so that the approximate standard error is .114. Thus,

$$\frac{\kappa_1}{\mathrm{SE}(\kappa_1)} = \frac{.248}{.114} = 2.17, \qquad [27]$$

indicating that the agreement in assignment to Category 1 is significantly better than chance at the .05 level.

REFERENCES

COHEN, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, **20**, 37–46.

COHEN, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, **70**, 213–220.

EVERITT, B. S. Moments of the statistics kappa and weighted kappa. *British Journal of Mathematical and Statistical Psychology*, 1968, **21**, 97–103.

FLEISS, J. L., COHEN, J., & EVERITT, B. S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, **72**, 323–327.

SANDIFER, M. G., FLEISS, J. L., & GREEN, L. M. Sample selection by diagnosis in clinical drug evaluations. *Psychopharmacologia*, 1968, **13**, 118–128.

SANDIFER, M. G., HORDERN, A., TIMBURY, G. C., & GREEN, L. M. Psychiatric diagnosis: A comparative study in North Carolina, London and Glasgow. *British Journal of Psychiatry*, 1968, **114**, 1–9.

SANDIFER, M. G., PETTUS, C., & QUADE, D. A study of psychiatric diagnosis. *Journal of Nervous and Mental Diseases*, 1964, **139**, 350–356.

(Received May 12, 1970)

ERRATUM

In the article "Current Research on the Frequency of Dream Recall" by David Cohen in the June 1970 issue, lines 26–30 of the left column of page 437 should read as follows: "Relatively high anxiety was related to significantly lower frequency of dream recall in groups marked by intellectualization and abnegation. . . . "