

Sample-Size Calculations for Cohen's Kappa

Alan B. Cantor

H. Lee Moffitt Cancer Center and Research Institute

In recent years, researchers in the psychosocial and biomedical sciences have become increasingly aware of the importance of sample-size calculations in the design of research projects. Such considerations are, however, rarely applied for studies involving agreement of raters. Published results on this topic are limited and generally provide rather complex formulas. In addition, they generally make the assumption that the raters have the same set of frequencies for the possible ratings. In this article I show that for the case of 2 raters and 2 possible ratings the assumptions of equal frequencies can be dropped. Tables that allow for almost immediate sample-size determination for a variety of common study designs are given.

Since its introduction by Cohen in 1960, variants of the parameter kappa have been used to address the issue of interrater agreement. Kappa takes the form

$$\kappa = \frac{\pi_o - \pi_c}{1 - \pi_c}, \quad (1)$$

where π_o is the proportion of rater pairs exhibiting agreement and π_c is the proportion expected to exhibit agreement by chance alone. Thus "perfect agreement" would be indicated by $\kappa = 1$, and no agreement (other than that expected by chance) means that $\kappa = 0$.

There have been several extensions of the original statistic. Weighted kappa allows different types of disagreement to have differing weights (Cohen, 1968). This might be appropriate if some types of disagreements were considered more critical than others. Extensions have also been made to allow for more than two raters (Posner, Sampson, Caplan, Ward, & Chenly, 1990), ordinal data (Fleiss, 1978), and continuous data (Rae, 1988). In this last case, kappa has been shown to be equivalent to the intraclass correlation coefficient (Rae, 1988).

When designing a study to estimate kappa in a

particular context or to perform one- or two-sample hypothesis tests concerning kappa, consideration should be given to the sample size needed to produce a desired precision for the estimate of power for the test. The report by Fleiss, Cohen, and Everitt (1969) giving the asymptotic variance of the estimate of kappa makes such discussions feasible, and Flack, Afifi, and Lachenbruch (1988) presented a method of sample-size determination for two raters and K possible ratings where $K \geq 2$. Their method requires the assumption that the raters' true marginal rating frequencies are the same. Donner and Eliasziw (1992) reported on a goodness-of-fit approach to this problem. They also made the assumption of equal marginals. In this article I show that for $K = 2$ this assumption can be dropped. In this case, with the aid of tables that are provided, the necessary sample-size calculation can be found almost immediately.

Thus, I restrict my attention to the situation of two raters, denoted Raters 1 and 2. They each rate N items with two possible ratings, denoted Ratings 1 and 2. The basic ideas presented below apply to more complex situations (weighted kappa and more ratings) as well, although the results will not be as simple.

Let π_{ij} represent the proportion of the population given rating i by Rater 1 and j by Rater 2. Let $\pi_{.j} = \pi_{1j} + \pi_{2j}$ and $\pi_{i.} = \pi_{i1} + \pi_{i2}$ be the proportion rate j by Rater 2 and the proportion rated i by Rater 1, respectively. Then $\pi_c = \pi_{1.}\pi_{.1} + \pi_{2.}\pi_{.2}$ and $\pi_o = \pi_{11} + \pi_{22}$.

Sample estimates of the above parameters are

Correspondence concerning this article should be addressed to Alan B. Cantor, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, Florida 36122-9497.

denoted by p_{ij} , p_j , p_i , p_e , and p_o . Then κ is estimated by

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}. \quad (2)$$

When designing a study to produce an estimate $\hat{\kappa}$ of kappa, the sample size should be chosen so that the standard error of $\hat{\kappa}$ will not exceed a preassigned value. Fleiss et al. (1969) showed that the asymptotic variance of $\hat{\kappa}$ can be written in the form Q/N , where

$$Q = (1 - \pi_e)^{-4} \left\{ \sum_i \pi_{ii} [(1 - \pi_e) - (\pi_i + \pi_i)(1 - \pi_o)]^2 + (1 - \pi_o)^2 \sum_{i \neq j} \pi_{ij} (\pi_i + \pi_j)^2 - (\pi_o \pi_e - 2\pi_e + \pi_o)^2 \right\}. \quad (3)$$

Note that all of the values needed are uniquely determined by $\pi_{1\cdot}$, $\pi_{\cdot 1}$, and κ . Specifically,

$$\begin{aligned} \pi_{2\cdot} &= 1 - \pi_{1\cdot} \\ \pi_{\cdot 2} &= 1 - \pi_{\cdot 1} \\ \pi_e &= \pi_{1\cdot} \pi_{\cdot 1} + \pi_{2\cdot} \pi_{\cdot 2} \\ \pi_o &= \kappa(1 - \pi_e) + \pi_e \\ \pi_{22} &= (\pi_o - \pi_{1\cdot} + \pi_{2\cdot})/2 \\ \pi_{11} &= \pi_o - \pi_{22} \\ \pi_{12} &= \pi_{1\cdot} - \pi_{11} \\ \pi_{21} &= \pi_{\cdot 1} - \pi_{11}. \end{aligned} \quad (4)$$

Table 1 gives values for Q for values of $\pi_{1\cdot}$, $\pi_{\cdot 1}$, and κ . For $\pi_{1\cdot} \neq \pi_{\cdot 1}$, kappa has an upper bound less than one. Thus the table has no values of Q for values of kappa that are not permissible. To use this table, one should specify the proportion expected to get Rating 1 from each rater and a value of kappa. The value of N that will yield a standard error C , for $\hat{\kappa}$, is found by dividing the corresponding value of Q by C^2 . If one wants to allow for a range of values for $\pi_{1\cdot}$, $\pi_{\cdot 1}$, and κ , the largest value of Q associated with these values should be chosen. Table 2 gives, for each pair $(\pi_{1\cdot}, \pi_{\cdot 1})$, the largest possible value of Q . It can be used

when one is reluctant to make any prior assumption concerning κ .

As an example, suppose two observers are asked to observe a group of subjects and to decide whether or not each exhibits a particular behavior. One would like to estimate kappa with an 80% confidence interval of the form of $\hat{\kappa} \pm d$, where d does not exceed 0.1. Suppose that one expects each to observe the behavior about 30% of the time. From a table of the normal distribution this requires a standard error not exceeding $0.1/1.28 = 0.078$. If no assumption is made about the value of kappa, one would use the maximum value of Q for $\pi_{1\cdot} = \pi_{\cdot 1} = 0.3$, which is 1.0. Thus one requires $N = 1.0/0.078^2 \approx 165$ subjects. Now suppose the study is done with the following results: $p_{1\cdot} = 0.4$, $p_{\cdot 1} = 0.3$, $\hat{\kappa} = 0.3$. Then one finds from Table 1, $Q = 0.929$ so that the estimated standard error is $\sqrt{0.929/165} = 0.075$ and an 80% confidence interval for kappa is 0.3 ± 0.096 .

If one wants to test a null hypothesis of the form $H_0: \kappa = \kappa_0$ with significance level α and power $1 - \beta$ if $\kappa = \kappa_1$, elementary calculations show that

$$N = \left[\frac{Z_\alpha \sqrt{Q_0} + Z_\beta \sqrt{Q_1}}{\kappa_1 - \kappa_0} \right]^2. \quad (5)$$

Here $Z_\alpha = \Phi^{-1}(1 - \alpha)$, where $\Phi(\cdot)$ is the standard normal distribution function. Q_0 and Q_1 are the values from Table 1 for the null hypothesis and alternative, respectively. Note that $\kappa_0 = 0$ is not excluded from the above discussion.

For an example one turns to the previous scenario in which two observers are determining whether or not a behavior is observed. Suppose one wishes to test the null hypothesis of $H_0: \kappa = 0.3$ against the one-sided alternative $H_A: \kappa > 0.3$ with significance level 0.05 and power 0.80 for $\kappa = 0.5$. Then $Z_\alpha = 1.645$ and $Z_\beta = 0.842$. If one expects both observers to see the behavior in about half the subjects, one has $Q_0 = 0.910$ and $Q_1 = 0.750$. From Equation 5 one gets $N = 131$.

Now consider a two-sample test of $H_0: \kappa_1 = \kappa_2$ versus $H_A: \kappa_1 \neq \kappa_2$, where κ_1 and κ_2 are estimated from independent samples, each of size N . Let Q_{01} and Q_{02} be the values of Q expected under the null hypothesis and Q_{A1} and Q_{A2} be the values of Q expected under an alternative. Elementary calculations show that

Table 1
Values of Q

		Kappa									
π_1	π_2	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.1	1.000	1.598	1.984	2.179	2.205	2.083	1.835	1.481	1.043	0.542
0.2	0.1	0.852	1.159	1.350	1.434	1.425	1.331	1.166			
0.2	0.2	1.000	1.182	1.284	1.312	1.272	1.172	1.018	0.817	0.576	0.301
0.3	0.1	0.654	0.808	0.899	0.931	0.911					
0.3	0.2	0.931	1.029	1.075	1.072	1.024	0.935	0.808	0.647		
0.3	0.3	1.000	1.055	1.070	1.046	0.986	0.893	0.768	0.614	0.433	0.228
0.4	0.1	0.490	0.550	0.580							
0.4	0.2	0.793	0.832	0.840	0.819	0.771	0.697				
0.4	0.3	0.953	0.973	0.965	0.929	0.867	0.780	0.668	0.533		
0.4	0.4	1.000	1.004	0.984	0.940	0.872	0.781	0.668	0.533	0.376	0.198
0.5	0.1	0.360	0.356								
0.5	0.2	0.640	0.634	0.614	0.582						
0.5	0.3	0.840	0.832	0.806	0.764	0.706	0.630	0.538			
0.5	0.4	0.960	0.950	0.922	0.874	0.806	0.720	0.614	0.490	0.346	
0.5	0.5	1.000	0.990	0.960	0.910	0.840	0.750	0.640	0.510	0.360	0.190
0.6	0.1	0.257	0.207								
0.6	0.2	0.490	0.448	0.408							
0.6	0.3	0.691	0.659	0.621	0.576	0.524					
0.6	0.4	0.852	0.830	0.796	0.748	0.686	0.610	0.519			
0.6	0.5	0.960	0.950	0.922	0.874	0.806	0.720	0.614	0.490	0.346	
0.7	0.1	0.174									
0.7	0.2	0.350	0.280								
0.7	0.3	0.524	0.472	0.424	0.379						
0.7	0.4	0.691	0.659	0.621	0.576	0.524					
0.7	0.5	0.840	0.832	0.806	0.764	0.706	0.630	0.538			
0.8	0.1	0.105									
0.8	0.2	0.221	0.129								
0.8	0.3	0.350	0.280								
0.8	0.4	0.490	0.448	0.408							
0.8	0.5	0.640	0.634	0.614	0.582						
0.9	0.1	0.048									
0.9	0.2	0.105									
0.9	0.3	0.174									
0.9	0.4	0.257	0.207								
0.9	0.5	0.360	0.356								

$$N = \left[\frac{A_\alpha \sqrt{Q_{01} + Q_{02}} + Z_\beta \sqrt{Q_{A1} + Q_{A2}}}{\kappa_1 - \kappa_2} \right]^2 \quad (6)$$

As an example, consider a questionnaire designed to measure how well a patient is coping emotionally and psychologically with a serious chronic illness. For simplicity, and in order to fit the current discussion, assume the result is dichotomous: satisfactory or unsatisfactory. One measure of such a questionnaire's utility is internal validity. This is measured by agreement of a patient's results on

two separate administrations. As part of a study to compare two such questionnaires, patients with a serious chronic illness are randomized to one of the two questionnaires. In each case the questionnaire is administered to the patient twice, at study entry and 1 month later. The estimates of kappa are to be compared. Specifically, one tests $H_0: \kappa_1 = \kappa_2$ against $H_A: \kappa_1 \neq \kappa_2$ with $\alpha = 0.05$. Suppose that the previous work with one of the questionnaires causes one to expect it to have $\kappa_1 \approx 0.7$ and that about half the patients will be judged to be coping satisfactorily. One would like a power of

Table 2
Maximum Values of Q

π_1	π_2	Q_{MAX}	κ
0.1	0.1	2.21417	0.366
0.2	0.1	1.44128	0.339
0.2	0.2	1.31200	0.289
0.3	0.1	0.93136	0.310
0.3	0.2	1.07992	0.243
0.3	0.3	1.07003	0.187
0.4	0.1	0.58424	0.256
0.4	0.2	0.84096	0.177
0.4	0.3	0.97355	0.121
0.4	0.4	1.00558	0.067
0.5	0.1	0.36000	0.000
0.5	0.2	0.64000	0.000
0.5	0.3	0.84000	0.000
0.5	0.4	0.96000	0.000
0.5	0.5	1.00000	0.000
0.6	0.1	0.25684	0.000
0.6	0.2	0.48980	0.000
0.6	0.3	0.69136	0.000
0.6	0.4	0.85207	0.000
0.6	0.5	0.96000	0.000
0.6	0.5	0.96000	0.000
0.7	0.1	0.17355	0.000
0.7	0.2	0.34964	0.000
0.7	0.3	0.52438	0.000
0.7	0.4	0.69136	0.000
0.7	0.5	0.84000	0.000
0.7	0.5	0.84000	0.000
0.8	0.2	0.22145	0.000
0.8	0.3	0.34964	0.000
0.8	0.4	0.48980	0.000
0.8	0.5	0.64000	0.000
0.8	0.5	0.64000	0.000
0.8	0.5	0.64000	0.000
0.8	0.5	0.64000	0.000
0.9	0.1	0.04819	0.000
0.9	0.2	0.10519	0.000
0.9	0.3	0.17355	0.000
0.9	0.4	0.25684	0.000
0.9	0.5	0.36000	0.000
0.9	0.5	0.36000	0.000
0.9	0.5	0.36000	0.000
0.9	0.5	0.36000	0.000
0.9	0.5	0.36000	0.000

at least 80% if $\kappa_2 = 0.5$ or 0.9 . Using the notation above, one has $Z_\alpha = 1.96$, $Z_\beta = 0.841$, $Q_{01} = Q_{02} = 0.510$, $Q_{A1} = 0.510$, and $Q_{A2} = 0.750$. From Fleiss et al. (1969), $N = 214$.

Several authors have stressed the importance of

adequate power for studies designed to test the efficacy of interventions (A'Kern, 1995; Cohen, 1977; Lachin, 1981), and their arguments are not reprised here. The same considerations hold for studies of interrater agreement. Studies should be designed with the precision of estimates and the power of statistical tests taken into consideration. This article should facilitate this.

References

A'Kern, R. P. (1995). Statistical power: A measure of the quality of a study. *British Journal of Urology*, 75, 5-8.

Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provisions for scaled disagreement or partial credit. *Psychological Bulletin*, 70, 213-220.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. Orlando, FL: Academic Press.

Donner, A., & Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: Confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine*, 11, 1511-1519.

Flack, V. F., Afifi, A. A., & Lachenbruch, P. A. (1988). Sample size determinations for the two rater kappa statistic. *Psychometrika*, 53, 321-325.

Fleiss, J. L. (1978). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

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

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, 2, 93-113.

Posner, K. L., Sampson, P. D., Caplan, R. A., Ward, R. J., & Chenly, F. W. (1990). Measuring interrater reliability among multiple raters: An example of methods for nominal data. *Statistics in Medicine*, 9, 1103-1116.

Rae, G. (1988). The equivalence of multirater kappa statistics and intraclass correlation coefficients. *Educational and Psychological Measurement*, 48, 921-933.

Received July 15, 1995

Revision received December 8, 1995

Accepted December 18, 1995 ■