Paper SA02-2012

Using SAS to Determine the Sample Size on the Cohen's Positive Kappa Coefficient Problem

Yubo Gao, University of Iowa, Iowa City, IA

ABSTRACT

The determination of sample size is a very important early step when conducting study. This paper considers the Cohen's Kappa coefficient _based sample size determination in epidemiology. Historically, the Kappa coefficient is used to express the degree of agreement between two raters when the same two raters rate each of a sample of *n* subjects independently, with the ratings being on a categorical scale consisting of 2 or more categories. In the context of epidemiology, the Kappa coefficient is similarly applied to indicate the degree of the agreement between two diagnostic tests for detecting a certain disease when neither of the two tests can be regarded as a gold standard. This paper addresses the minimum required sample size in such case when the Kappa coefficient is desired to be no less than a certain positive proportion or number, and presents the SAS code for the calculation, based on three parameters, namely, the co-positive proportion (p_{11}) of, and the positive proportions of, the two diagnostic tests ($p_{1.}$ and $p_{.1}$).

INTRODUCTION

Conventionally, the Cohen's kappa coefficient (Cohen 1960) is used to express the degree of agreement between two raters when the same two raters rate each of a sample of n subjects independently, with the ratings being on a categorical scale consisting of 2 or more categories (Fleiss (1981)). In the context of epidemiology or medical diagnoses, the kappa coefficient is similarly applied to indicate the degree of the agreement between two diagnostic tests for detecting a certain disease when neither of the two tests can be regarded as a gold standard. A simple example is given in Table 1, where two diagnostic tests, Test 1 and Test 2, are employed to screen out a sample of susceptible individuals. In Table 1, p_{ij} denote the true population proportion in the ith row category and the jth column category, whereas a, b, c, and d denote the observed numbers from the selected sample.

Table 1 A 2 by 2 table for the joint distribution of two diagnostic tests.

	Test 2		
Test 1	Positive (+)	Negative (-)	
Positive (+)	p_{11} (a)	p_{12} (b)	$p_{1.}$
Negative (-)	p_{21} (c)	p_{22} (d)	$p_{2.}$.
	$p_{.1}$	$p_{.2}$	

Note that the parameters in Table 1 satisfy

$$p_{11} + p_{21} = p_{.1}, \quad (1.1)$$
$$p_{12} + p_{22} = p_{.2} = 1 - p_{.1}, \quad (1.2)$$
$$p_{11} + p_{12} = p_{1.}, \quad (1.3)$$
$$p_{21} + p_{22} = p_{2.} = 1 - p_{1.} \quad (1.4)$$

The Cohen's kappa coefficient (κ), which is a measurement of agreement between two tests, is defined by (Fleiss (1981)

$$\kappa = (p_0 - p_e)/(1 - p_e) \quad (1.5)$$

Where p_o and p_e are given, respectively, by

$$p_o = p_{11} + p_{22}, \quad (1.6)$$

and

$$p_e = p_{1.}\, p_{.1} + p_{2.} p_{.2} \quad (1.7)$$

Now we want to decide under the condition that neither of the two tests are regarded as gold standards the minimum required sample size if the kappa coefficient is desired to be kept at no less than a certain positive fraction number?

SAMPLE SIZE FORMULA

In order to decide the minimum required sample size, a pair of hypotheses is set up as follows (Lee (2002)):

H_0:$\kappa \leq \kappa_0$,

versus (2.1)

H_1: $\kappa > \kappa_0$,

where $\kappa_0 > 0$ is the desired level.

A point estimator of κ, denoted by $\kappa 1$, is given by

$$\kappa 1 = 2(ad-bc)/[(a+b)(a+c)+(b+d)(c+d)] \quad (2.2)$$

with its variance given by (Fleiss, et al (1969))

$$var(\kappa 1) = (E+F-G)/[n(1-p_e)^4] \quad (2.3)$$

where p_o and p_e are given, respectively, by Eqs. (1.6) and (1.7), *E, F*, and *G* are given, respectively, by

$$E = p_{11}[(1-p_e)-(p_{.1}+p_{1.})(1-p_0)]^2 + p_{22}[(1-p_e)-(p_{.2}+p_{2.})(1-p_0)]^2 \quad (2.4)$$

$$F = (1-p_0)^2[p_{12}(p_{.1}+p_{2.})^2 + p_{21}(p_{.2}+p_{1.})]^2 \quad (2.5)$$

and

$$G = [p_0(1+p_e)-2p_e)]^2 \quad (2.6)$$

Note that it has been shown that $\kappa 1$ of Eq. (2.2) is an unbiased estimator of κ (Everitt (1968)).

Let α and β denote, respectively, the Type I and Type II errors, and z_p $(0 < p < 1)$ be the upper p^{th} percentile of a standard normal distribution. To maintain the power of the test to be at lease $1 - \beta$, the sample size (n) has to satisfy the following equation (Bickel and Doksum (1977)):

$$n = (z_\alpha + z_\beta)^2(E+F-G)/[(1-p_e)^2(\kappa 1 - \kappa_0)] , \quad (2.7)$$

where *E, F,* and *G* are given, respectively, by Eqs. (2.4)-(2.6), while p_e is given by Eq. (1.7).

USING SAS TO DETERMINE SAMPLE SIZES

Now we use SAS to calculate sample size needed to satisfy kappa requirement. The sample size formula for n in Eq. (2.7) depends seemingly on several parameters including p_{ij}, i, j = 1, 2, p_o, p_e, and κ. But, since all of p_o, p_e, and κ depend on p_{ij}, i, j =1, 2, Eq. (2.7) actually depends only on p_{ij}, i, j =1, 2. However, due to the constraint that the sum of total probability must be one, among four parameters of p_{ij}, i, j = 1, 2, only three of them are considered to be independent parameters.

From simple manipulations, Eq. (2.2) for $\kappa 1$ can be expressed in terms of these three parameters as

$$\kappa 1 = 2(p_{11}-p_{1.}p_{.1})/(p_{1.}+p_{.1}-2p_{1.}p_{.1}) \quad (3.1)$$

Generally, there is no uniform guideline to help in deciding which three should be chosen to serve as independent parameters. Here we choose p_{11}, the co-positive proportion between two tests, $p_{1.}$, the positive proportion of Test 1, and $p_{.1}$, the positive proportion of Test 2 as three independent parameters for Eq. (2.7). Since there are three independent parameters, there exist too many different possible scenarios, which cannot be completely exhausted. We only calculate limited combinations among those parameters as follows.

Suppose desired κ_0=0.5, α=0.05, β is from 0.05 to 0.2, and the ranges for other three can be seen in the code, where $p_{1s}=p_{1.}$, $p_{s1}=p_{.1}$, and so on.

```
/*sample size for kappa*/
options nocenter formdlim=' ';

data kp;
        alpha=0.05;
        k0=0.5;
        do beta=0.05 to 0.2 by 0.05;;
          do p11=0.4 to 0.7 by 0.005;
            do p1s=p11+0.1;
              do ps1=p11+0.05;
                p12=p1s-p11;
                p21=ps1-p11;
                p22=1-p11-p12-p21;
                k1=2*(p11-p1s*ps1)/(p1s+ps1-2*p1s*ps1);

                p0=p11+p22;
                p2s=p21+p22;
                ps2=p12+p22;
                pe=p1s*ps1+p2s*ps2;

                e=p11*((1-pe)-(ps1+p1s)*(1-p0))**2+p22*((1-pe)-(ps2+p2s)*(1-p0))**2;
                f=(1-p0)**2*(p12*(ps1+p2s)**2+p21*(ps2+p1s)**2);
                g=(p0*(1+pe)-2*pe)**2;

                n=floor((probit(alpha)+probit(beta))**2*(e+f-g)/((1-pe)**2*(k1-k0))**2)+1;
                output;
              end;
            end;
          end;
        end;
run;
```

From the calculations, holding everything else are the same, it can be easily seen that the higher the power of the test, or equivalently, the smaller the probability of type II error β, the larger the sample size. Also, the bigger the difference ($\kappa 1$- κ_0) between the true unknown kappa coefficient ($\kappa 1$) and its desired level (κ_0) is, the smaller the sample size. In addition, the closer the other two marginal probabilities ($p_{1.}$ and $p_{.1}$) to the first diagonal joint probability (p_{11}), the smaller the sample size.

CONCLUSIONS

This paper discussed the minimum required sample size if the kappa coefficient is desired to be kept at no less than a certain positive proportion under the condition that neither of the two tests are regarded as gold standards in diagnosing a certain disease in epidemiology. A SAS solution is provided through the derived sample size formula. We chose the co-positive proportion of, and the positive proportions of, the two diagnostic tests as the three parameters needed in the formula. Though in theory there are too many different possible combinations for three independent parameters, in reality it is easy to decide the sample size based on specific question requirements.

REFERENCES

- Cohen, J. A. (1960). Coefficient of agreement for nominal scales. *Educational and Psychological Measurement,20,* 37-46.
- Fleiss, J. L. (1981), Statistical Methods for Rates and Proportions, second edition. *John Wiley & Sons, Inc*., New York.
- Tze-San Lee(2002), On Determination of Sample Size for the Positive Kappa Coefficient. Accessed on June 20, 2012 at http://www.amstat.org
- Fleiss, J. L., Cohen, J., and Everitt, B. S. (1969), Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327.
- Everitt, B. S. (1968), Moments of the statistics kappa and weighted kappa. *British Journal of*

Mathematical and Statistical Psychology, 21, 97-103.
- Bickel, P.J. and Doksum, K.A. (1977), Mathematical Statistics, *Holden-Day, Inc*., Oakland, California.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Yubo Gao
Department of Orthopaedics and Rehabilitation
University of Iowa Hospitals and Clinics
200 Hawkins Dr.
Iowa City, Iowa 52242
yubo-gao@uiowa.edu
(319)356-1674

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.