Paper 155-30

A Macro to Calculate Kappa Statistics for Categorizations by Multiple Raters

Bin Chen, Westat, Rockville, MD
Dennis Zaebst, National Institute of Occupational and Safety Health, Cincinnati, OH
Lynn Seel, Westat, Rockville, MD

ABSTRACT

It is often necessary to assess the agreement on multi-category ratings by multiple raters in various studies in many fields. Kappa is one of the most popular indicators of interrater agreement for categorical data. SAS® procedure PROC FREQ can provide kappa statistic for two raters. A SAS® macro MAGREE computes kappa for multiple raters with multi-categorical ratings. Both have limited applications. The author wrote a macro which implements the Fleiss (1981) methodology measuring the agreement when both the number of raters and the number of categories of the rating are greater than two. This macro can handle missing data as well as data that are not square, which are the two major limits of PROC FREQ and macro MAGREE. It is very easy to use. Anyone who has minimum SAS® programming skill and basic understanding of the kappa statistic can use the macro with ease. All the users need to do is to start the macro and input some parameters. SAS® products involved are Base SAS® and SAS® MACRO. This macro is tested on Windows SAS®8 and above.

Keyword: SAS® MACRO, kappa statistic, categorical data.

INTRODUCTION

Assessing the agreement on multi-category ratings by multiple raters is often necessary in various studies in many fields. Kappa is one of the most popular indicators of interrater agreement for categorical data. The SAS® procedure PROC FREQ can provide the kappa statistic for two raters and multiple categories, provided that the data are square, which will be explained in a later section. A SAS® macro MAGREE can compute the estimates and tests of agreement among multiple raters and multiple categories. Unfortunately, the MAGREE macro was not designed to handle missing data. A user-friendly procedure that can handle missing and/or non-square data is needed. The macro described here concentrates on the measure of agreement when both the number of raters and the number of categories of the rating are at least two based on Fleiss (1981) methodology. This macro can handle missing data as well as data that are not square.

KAPPA STATISTICS

The kappa statistic was first proposed by Cohen (1960). Some extensions were developed by others, including Cohen (1968), Everitt (1968), Fleiss (1971), and Barlow et al (1991). This paper implements the methodology proposed by Fleiss (1981), which is a generalization of the Cohen kappa statistic to the measurement of agreement among multiple raters. Each of the n target subjects is rated by m (≥2) raters independently into one of k (≥2) mutually exclusive and exhaustive categories. Let x_{ij} be the number of ratings on subject i (i=1,…,n) into category j (j=1,…,k) by the m raters, $\bar{p}_j$ denote the overall proportion of ratings in category j, and $\hat{k}_j$ the value of kappa for category j, then we have:

$$\bar{p}_j = \frac{\sum_{i=1}^{n} x_{ij}}{nm},$$

$$\hat{k}_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}(m - x_{ij})}{nm(m-1)\bar{p}_j(1-\bar{p}_j)}.$$

The overall kappa $\hat{\bar{k}}$ is weighted average of $\hat{k}_j$:

$$\hat{\bar{k}} = \frac{\sum_{j=1}^{k} \bar{p}_j (1-\bar{p}_j)\hat{k}_j}{\sum_{j=1}^{k} \bar{p}_j (1-\bar{p}_j)},$$

which is equivalent to:

$$\hat{\bar{k}} = 1 - \frac{nm^2 - \sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}^{\ 2}}{nm(m-1)\sum_{j=1}^{k} \bar{p}_j (1-\bar{p}_j)}.$$

The standard error of $\hat{\bar{k}}$ is:

$$s.e.(\hat{\bar{k}}) = \frac{\sqrt{2}}{\sum_{j=1}^{k} \bar{p}_j (1-\bar{p}_j)\sqrt{nm(m-1)}} \sqrt{(\sum_{j=1}^{k} \bar{p}_j (1-\bar{p}_j))^2 - \sum_{j=1}^{k} \bar{p}_j (1-\bar{p}_j)(1-2\bar{p}_j)}$$

AVAILABLE PROCEDURES FROM SAS®

SAS® procedure PROC FREQ with the AGREE option can provide the kappa statistic when there are only two raters. PROC FREQ only works with complete or square data, i.e., each rater uses each possible choice on the rating scale at least once. For example, two raters (1 and 2) rated n subjects into three categories (1, 2 and 3). Rater 1 did not give "1" to any of the n subjects. The frequency table would look like:

		Rater 1		
		2	3	
Rater2	1	b	c	r1
	2	e	f	r2
	3	h	i	r3
		c2	c3	

instead of:

		Rater 1			
		1	2	3	
Rater2	1	a	b	c	r1
	2	d	e	f	r2
	3	g	h	i	r3
		c1	c2	c3	

The first table is not square while the second one is. In the case of first table, PROC FREQ will not provide a kappa or will provide a wrong kappa. In real world studies, it is not uncommon to see a non-square frequency table.

SAS® also provides a macro MAGREE that can calculate the kappa statistic and the corresponding standard error and p-value for cases with multiple raters and multiple categories of ratings. However MAGREE only functions when each rater rates each subject exactly once. It won't function if any of the following happens:

- Missing values appear in the subject, rater, or rating variables,
- A rater does not rate some subject(s),
- A subject is not rated by some rater(s),
- All subjects do not have an equal number of ratings,
- More than one rating of a subject by a rater.

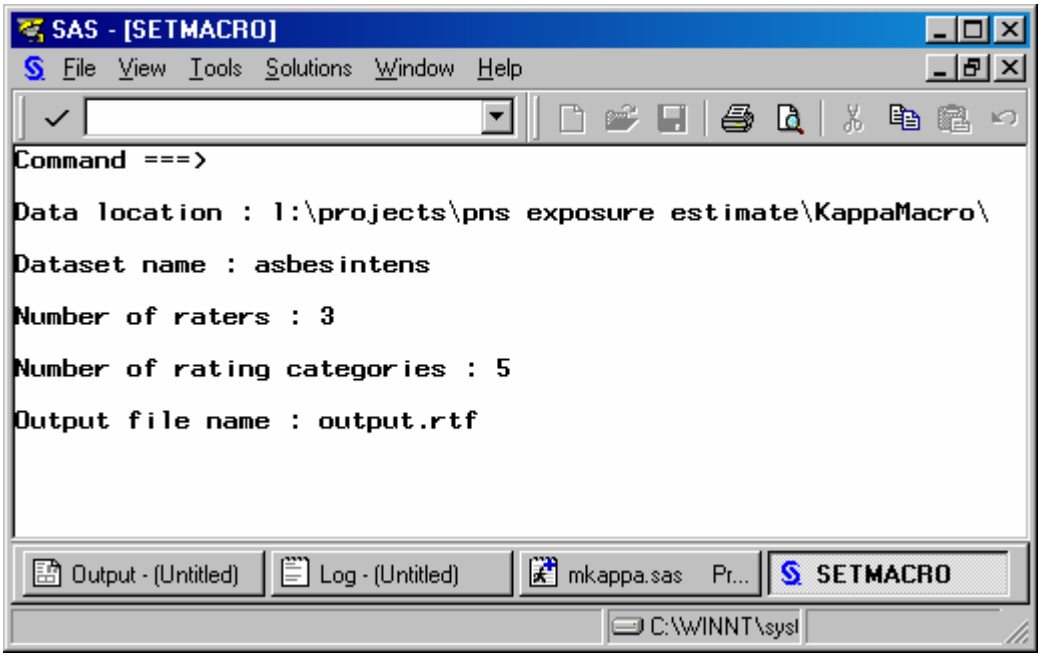
In practice, it is very common to have some of the values missing in the data. For example, one rater might not rate one subject or might give the subject an invalid value due to various reasons. Many times, there is no opportunity to correct these data mistakes, either due to the unavailability of the rater or the restrictions by the study design. In these cases, MAGREE will not calculate the kappa even though all other raters provided valid ratings.

Besides the macro MAGREE, some other SAS® programs have been developed to compute kappa for multiple raters using categorical classifications (Green, 1997). However, they are not very user-friendly and it is hard for someone who does not understand Fleiss's (1981) methodology to use the programs.

MKAPPA MACRO

While conducting a study which needs a multi-categorical kappa from multiple raters with missing data, the author developed the macro MKAPPA. The MKAPPA macro provides kappa, the standard error and p-value of the kappa statistic. The macro is tested on Windows SAS®8 and above.

When the user invokes the program, a new window will pop up asking the user to input the required information: the input SAS® dataset name and location, the number of raters, the number of rating categories, and the output file name. The output file will be put in the same folder as the input dataset.



The proper form of the input dataset is shown in Table 1. The values of the rater variables are the categories into which the rater placed the subject. The categories need be numeric.

Table 1. Dataset format for MKAPPA macro

Rater1	Rater2	Rater3	Rater4	Rater5
1	1	2	1	2
2	2	2	2	1
4	5	5	5	4
3	4	4	3	3
3	5	4	5	5
…	…	…	…	…

The output of the macro is a table in rich text format. All of the output statistics, including the kappa, the standard error, and the p-value of the kappa, are also available in a SAS® data set called koutput in the same location of the input SAS® data set.

APPLICATION EXAMPLE

The MKAPPA macro was applied to evaluate the retrospective assessments of asbestos exposures at a shipyard by three raters. For each shop, job title, and time period combination, the three raters were asked to rate the asbestos exposure level independently on a 5-category scale. Table 2 shows part of the data. The distribution of each rater's rating is listed in Table 3. We have missing data from rater1. Using macro MKAPPA, we obtained the kappa statistics listed in Table 4.

Table 2. Asbestos Intensity Estimate by the Three Raters

Subject	Rater 1	Rater 2	Rater 3
1	1	2	2
2	1	2	3
3	1	2	2
4	1	2	1
5	1	2	2
6	1	1	1
7	1	1	2
8	1	1	1
9	1	1	1
…	…	…	…

Table 3. Distribution of the Asbestos Intensity Estimate by Each Rater

Rate	Rater1	Rater2	Rater3
1	2246	2706	3247
2	894	798	48
3	267	9	205
4	46	10	23
5	66	0	0
missing	4	0	0

Table 4. Kappa statistics

	Kappa	SE	p-value
Overall	0.593	0.0078	<0.001

CONCLUSION

MKAPPA is a user-friendly SAS® macro which can calculate the kappa statistics for multiple raters with multi-categorical ratings. It will calculate the kappa statistics when there are missing data. Anyone who has minimum SAS® programming skills and a basic understanding of the kappa statistic can use this macro with ease.

REFERENCES

Barlow W, Lai NY, and Azen SP (1991). A comparison of methods for calculating a stratified kappa. Statist. Med., 10, 1465-1472.

Cohen, J (1960). A coefficient of agreement for nominal scales. Educ. Psychol. Meas., 20, 37-46.

Cohen, J (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psych. Bull., 70, 213-220.

Everitt BS (1968). Moments of the statistics kappa and the weighted kappa. British J. Math. Statist. Psych., 21, 97-103.

Fleiss JL (1971). Measuring nominal scale agreement among many raters. Psych. Bull., 76, 378-382.

Fleiss JL (1981). Statistical Methods for Rates and Proportions. John Wiley & Sons, Inc., New York.

Green AM (1997). Kappa statistics for multiple raters using categorical classifications. Proceedings of the 22nd annual SAS User Group International conference, 1110-1115.

TRADEMARK CITATION

DISCLAIMER

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Please contact the authors at:

Bin Chen
WESTAT Inc.
5555 Ridge Ave
MS-R44
Cincinnati, OH 45213
Phone: 513-841-4316
Fax: 513-841-4470
Email: bchen2@cdc.gov

APPENDIX

The MKAPPA macro:

```
%window setmacro rows = 18
  #2  "Data location : " dloc 80
  #4  "Dataset name : " dset 40
  #6  "Number of raters : " Nrater 8
  #8  "Number of rating categories : " Ncat 8
  #10 "Output file name : " outfile
  ;

%let dloc = c:\documents and settings\bin chen\my documents\projects\KappaMacro\ ;
%let dset = testno1 ;
%let Nrater  = 3  ;
%let Ncat  = 5  ;
%let outfile = output.rtf ;

%display setmacro ;

%macro mkappa;
libname kap "&dloc";

data temp; set kap.&dset;
subj=_n_;
data temp; set temp;
```

```
%do i=1 %to &nrater;
rate=rater&i; rater=&i; output; %end;
run;
proc freq data=temp noprint; tables subj*rate / out=xij ;
run;
data subj; set xij; keep subj;
proc sort data=subj nodupkey out=subj; by subj; run;
proc sql noprint; select count(*) into :tsub from subj;

data temp2; do subj=1 to &tsub; do rate=1 to &ncat; output; end; end; run;
data xij; set xij; drop percent;
proc sort data=xij; by subj rate;
proc sort data=temp2; by subj rate;
data xij; merge xij temp2(in=a); by subj rate; if a;
data xij; set xij; if count=. then count=0; run ;
data xij; set xij; xmx=count*(&nrater-count); run;

proc sort data=xij; by rate;
data xij; set xij; by rate;
retain x xmx2;
if first.rate then do; x=0; xmx2=0; end;
x=x+count; xmx2=xmx2+xmx;
if last.rate then output;
run;
data xij; set xij; keep rate x xmx2 ; run;

data xij; set xij;
p=x/(&tsub*&nrater);
pq=p*(1-P);
pqp=p*(1-p)*(1-2*p);
kj=abs(1 - xmx2/(&tsub*&nrater*(&nrater-1)*pq));
numj=kj*pq;
run;
data xij; set xij; if p=0 then do;
kj=0; numj=0; end; run;

data final; set xij end=end; retain num den pqpsum;
if _n_=1 then do; num=0; den=0; pqpsum=0; end;
num=num+numj;
den=den+pq;
pqpsum=pqpsum+pqp;
if end then output;
run;

data kap.koutput; set final; keep Kappa SE pvalue;
Kappa=num/den;
SE=sqrt(2*(den*den-pqpsum))/(den*sqrt(&tsub*&nrater*(&nrater-1)));
pvalue=1-probnorm(kappa/se);
run;

ods rtf file="&dloc&outfile"   ;
options nonumber ;
goptions dev=png target=png htitle=1.5 htext=1.2 ftext=swissb ftitle=swissb
         xmax=6 ymax=5 ;
proc print data=kap.koutput noobs label;
  label kappa="Kappa" SE="Standard Error"  pvalue="Prob>Z";
  title "Kappa statistics for &Ncat-category ratings by &Nrater raters";
run;
ods rtf close;

%mend mkappa;
```