Paper 323-2012

A Bootstrapped Kappa Statistic for a Multiple-Rater Multiple-Category Problem

Yubo Gao, University of Iowa, Iowa City, IA

ABSTRACT

In many research fields, kappa coefficient is a popular tool for measuring the degree of agreement between two or more raters where the raters rate the subjects on a categorical scale. SAS® procedures and macros exist for calculating kappa and its confidence interval. But this is one time calculation. Sometimes there is a need to estimate the precision of this statistic through resampling technique such as bootstrap so as to get a full understanding about its confidence interval. There is no literature so far about calculating kappa confidence intervals using bootstrap resampling methodology in a multiple-rater multiple-category situation. This article just wants to fill this gap using SAS procedure PROC SURVEYSELECT and a macro.

INTRODUCTION

It is a common practice for researchers to measure the interrater agreement when the subjects are judged on a categorical scale by different raters. The degree of agreement among the various raters gives some indication as to the consistency of the values. If agreement is high, we feel more confident in the ratings that reflect the actual circumstance. If agreement among the raters is low, we are less confident in the results. It is well known that kappa is one of the most popular indicators of interrater agreement for categorical data. When the number of raters is two, this is easily accomplished by using SAS procedure PROC FREQ with the AGREE option to obtain the kappa statistic derived from a technique developed by Fleiss (1981). In multiple-rater (more than two) and multiple-category (more than two) contexts, based on the methodology proposed by Fleiss (1981), Green (1997) suggested an implementation in SAS, but whole codes were not available. King (2004) used Excel to implement the formula. A SAS macro called MAGREE provided by SAS Institute and a SAS macro called Inter_Rater developed by Gwet (2002) can be used to compute the kappa estimates, including the tests and confidence interval. Remember that this is only one time calculation. In order to get better understanding of the estimate there is a need to use simulation technique to validate that estimate. The bootstrap methodology pioneered by Efron and Tibshirani (1993) is perfectly fitted in such a computationally intensive statistical environment. For two-rater case, Vierkant (1997) presented a SAS macro for calculating confidence intervals about kappa or weighted kappa using bootstrap resampling methodology. There is no literature so far about calculating kappa confidence intervals using bootstrap resampling methodology in a multiple-rater multiple-category situation. The purpose of this article is to show a SAS way to confirm/evaluate the accuracy of the kappa statistic for a multiple-rater multiple-category problem using SAS procedure PROC SURVEYSELECT and a macro.

BOOTSTRAPPING KAPPA STATISTIC WITH SAS

The bootstrap is such a computationally intensive resampling statistical technique that allows the researcher to make inferences from data without making strong distributional assumptions about the data or the statistic being calculated.

The bootstrap method is a resampling technique for estimating the precision of a statistic. A bootstrap test does not make any assumptions about the distribution of the test statistic. Depending on the statistic of interest, there are a variety of possible bootstrap sampling schemes. For simplicity, we use non-parametric (that is, case resampling) bootstrapping here. Since bootstrapping is a computer-intensive technique, we need seek a fast and efficient approach to generate the samples we need in the simulation process. In SAS, Cassell (2007) suggested that PROC SURVEYSELECT do a good job. First we need to generate bootstrap samples (replicates).

```
proc surveyselect data=cvm out=outboot          /* 1 */
        seed=30459584                           /* 2 */
        method=urs                              /* 3 */
        samprate=1                              /* 4 */
        outhits                                 /* 5 */
        rep=2000;                               /* 6 */
run;
```

PROC SURVEYSELECT, which appeared in SAS 8.2, allows one to generate random samples of many kinds from an input data set. In line [1], we invoke PROC SURVEYSELECT and tell it the input and output data set names, through the SAS options DATA= and OUT=.

In line [2], we specify a random seed. Without the seed, we cannot reproduce our results. In line [3], the METHOD= option lets us specify the type of random sampling. For a bootstrap, we need a simple random sample with replacement, and we need the sample to be of the same size as the original data set. Remember that simple random sampling with replacement is also called Unrestricted Random Sampling, which is abbreviated as URS in the METHOD= option. In order to get a sample of the same size as our original data set, in line [4], we use the SAMPRATE= option to get that without having to figure out the data set size first. SAMPRATE= option accepts either whole numbers (as percents) or proportions up to one. Either SAMPRATE=1 or SAMPRATE=100 will give us that 100% sample. In line [5], we use the keyword OUTHITS. This is for use when we ask for samples which could return a record more than once – like URS samples. OUTHITS makes sure that the procedure generates an output record every time it hits a given record, rather than only the first time. This gives us the bootstrap sample that we need in the next step. In line [6], we specify the number of bootstrap samples that we want to generate. This automatically generates a variable called REPLICATE, which keeps track of the replicate number for the samples and will be used later. Here, we have generated 2000 samples (replicates) contained in data set outboot. According to Efron and Tibshirani (1993), 2000 samples should be enough to produce useful results.

For each sample identified by variable replicate, we used the macro Inter_Rater (v.1.0) to compute its overall kappa coefficient. Specifically, macro myboot below does those things.

```
%global i;
%macro myboot;
        %do i=1 %to 2000;
                data b&i(drop=NumberHits Replicate);                    /* 7 */
                    set outboot(where=(Replicate=&i));
                run;

                %Inter_Rater(InputData=b&i,                             /* 8 */
                             DataType=c,
                             VarianceType=c,
                             CategoryFile=CatFile,
                             Outfile=a&i);

                data d&i;                                               /* 9 */
                    set a&i(keep=KappaStat q);
                    replicate=&i;
                    if q not in (1,2,3);
                run;

                proc append base=big data=d&i;                          /* 10 */
                run;

        %end;

        proc univariate data=big;                                       /* 11 */
                var KappaStat;
                output out=final pctlpts=2.5,97.5 pctlpre=ci;
        run;

        proc print data=final;                                          /* 12 */
        run;

%mend myboot;

%myboot;                                                                /* 13 */
```

First, define global macro variable i to identify sample i. Within the macro myboot, there are 2000 do-end iterations that correspond to the 2000 samples generated by PROC SURVEYSELECT. Each iteration uses a different sample distinguished by variable Replicate. Line [7] picks up Replicate (sample) &i.

Line [8] uses the macro Inter_Rater to computes the kappa statistic for Replicate (sample) &i. The input parameters in macro Inter_Rater are adapted for this case. InputData is the data set used, which is equal to the Replicate

(sample) &i here. DataType= identifies which of the 2 forms of input file is used, here DataType=c corresponds to the input file format in data set cvm. VarianceType=c is related to another statistic AC1 in Inter_Rater that is not used now. CategoryFile= contains a single variable file about the categories information, and here it has three records, N, I, and S. Outfile= designates the output file for kappa statistic, and here it is data set a&i. For detailed parameter information, see Gwet (2002).

Line [9] just pulls out the overall kappa statistic for Replicate (sample) &i from a&i and saves it to d&i. Line [10] appends d&i to data set big. From the 2000 samples, we will have 2000 kappa coefficients in data set big. Next we need find a way to get the 95 percent interval for kappa. One way to estimate the 100(1-alpha)% confidence intervals from the 2000 kappa coefficients is to take the (alpha/2) and (1 – alpha/2) quantiles of the estimated values. These are called bootstrap percentile intervals. For example, a bootstrap 95% percentile interval would be the interval from the 2.5th percentile to the 97.5th percentile. We can get both these values from PROC UNIVARIATE. Line [11] generates the 2.5% and 97.5% percentile kappa statistics from data set big and saves the 95% confidence interval to data set final. Line [12] prints out final results.

AN EXAMPLE

The data used in this paper come from the Example 3.12 “CVM Data” on page 108 of Shoukri and Pause (1998), which is reproduced here.

Four clinicians are asked to classify 20 x-rays to detect spinal cord damage in young foals believed to have Cervical Vertebral Malformation (CVM). The results are given in the following Table.

Classification is based on scores where:

1-5 =N (slight or no damage)
6-11 =I (intermediate damage)
>=12 =S (severe damage)

CVM Classification by 4 Clinicians

X-rays	Clinicians			
	1	2	3	4
1	S	I	S	I
2	I	I	S	I
3	N	N	I	N
4	N	N	N	N
5	S	S	S	I
6	N	N	N	I
7	N	N	I	N
8	N	N	N	N
9	S	S	I	S
10	S	S	S	S
11	I	I	S	I
12	I	S	S	S
13	S	I	I	I
14	N	I	I	N
15	I	I	N	N
16	I	N	N	I
17	S	S	S	S
18	S	I	S	S
19	I	S	S	S
20	N	N	I	N

We call it data set cvm that will be used in the simulation as our input data. First, run PROC SURVEYSELECT to get 2000 samples (of course, we may put PROC SURVEYSELECT part at the beginning of macro myboot). Then invoke macro myboot in Line [13].

For this example, the overall kappa coefficient is 0.34959 with Standard Error=0.06469. From Landis and Koch (1977), the strength of agreement among raters is fair. The 95% confidence interval is 0.3496±1.96*0.065, i.e., (0.2222, 0.4770).

From the simulation, the 95% confidence interval of the kappa coefficient is as follows.

Obs	ci2_5	ci97_5
1	0.17830	0.48460

So, we expect that the bootstrapped 95% confidence interval of the kappa coefficient for this problem is (0.1783, 0.4826). Compared with (0.2222, 0.4770) obtained before from one-time calculation, this interval is little bigger. Here, we only presented the bootstrap percentile intervals here. If interested, with some coding efforts, you can get other refined bootstrap confidence intervals, such as BCa and Bootstrap-T intervals.

CONCLUSION

The kappa statistic has become a wide accepted tool for measuring the amount of agreement between raters. Sometimes there is a need to estimate the kappa precision through resampling technique such as bootstrap so as to get a full understanding about its confidence interval. In a multiple-rater multiple-category situation, this paper illustrated a process of getting kappa confidence intervals using bootstrap resampling methodology through the PROC SURVEYSELECT and a macro.

REFERENCES

Cassell, David L. (2007), “Don't Be Loopy: Re-Sampling and Simulation the SAS® Way”. Proceedings of the SAS Global Forum. Orlando, Florida.
Efron, B., & Tibshirani, RJ. (1993), An Introduction to the Bootstrap. New York: Chapman & Hall/CRC.
Fleiss, JL.(1981), Statistical Methods for Rates and Proportions. 2nd Edition. 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. San Diego, California March 16-19, 1997.
Gwet, Kilem (2002), Inter_Rater macro, http://www.stataxis.com/files/sas/Inter_Rater.txt --accessed June 19, 2008.
King, Jason E. (2004), Software Solutions for Obtaining a Kappa-Type Statistic for Use with Multiple Raters, the annual meeting of the Southwest Educational Research Association, Dallas, Texas.
Landis, JR, and Koch, GG. (1977), The measurement of observer agreement for categorical data. Biometrics 33:159-174.
SAS Institute, MAGREE macro for Computing estimates and tests of agreement among multiple raters, http://support.sas.com/kb/25/006.html --accessed May 8,2008.
Shoukri, M.M., and Pause, C.A. (1998), Statistical Methods for Health Sciences, 2nd Ed. CRC Press.
Vierkant R.A.(1997), A SAS® Macro for Calculating Bootstrapped Confidence Intervals About a Kappa Coefficient. Proceedings of the 22nd Annual SAS User Group International conference. San Diego, California March 16-19, 1997.

CONTACT INFORMATION

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

Yubo Gao
University of Iowa Hospitals and Clinics (UIHC)
Orthopaedic Surgery, 01066 JPP
200 Hawkins Dr.
Iowa City, IA 52242
Phone: 319-356-1674
Email: yubo-gao@uiowa.edu