Program NM-CLASS

For Normal Approximation Classification Consistency and Accuracy

August 2019

Stella Y. Kim
University of North Carolina at Charlotte
Email: stella-kim@uncc.edu

NM-CLASS is the R code that computes classification consistency and accuracy indices using the normal approximation (NM) procedure discussed in Peng and Subkoviak (1980). NM-CLASS is developed to be used to estimate two classification consistency indices, the agreement index *P* and the kappa coefficient, and three classification accuracy indices including an accuracy index γ, the false-positive error rate, and the false-negative error rate discussed in Kim and Lee (in press).

Normal Approximation Procedure

The NM procedure, as its name implies, is an approximation of Huynh's (1976) strong true score procedure, relying on the assumption that a joint distribution of scores from two parallel forms is a bivariate normal distribution. The bivariate normal distribution is defined to have a mean and standard deviation of the observed scores from a single test form and a correlation equal to test reliability, r. Note that two parallel forms are assumed to have an identical mean and standard deviation. In the case of binary decisions (e.g., pass/fail), the proportions of the bivariate distribution that are either below or above a specified cut score on both forms are regarded as the proportions of examinees who are consistently assigned into the same category on two parallel forms of a test (i.e., agreement index *P*).

Although the original paper (Peng & Subkoviak, 1980) did not discuss explicitly, a similar approach can be employed for estimating classification accuracy. In this case, the true scores and observed scores are assumed to follow a bivariate normal distribution with a correlation as the square root of reliability, $\sqrt{r}$ (see Kim & Lee, 2019).

The NM procedure can deal with any types of scores such as scores from polytomous items, weighted composite scores, and scale scores, as long as the normality assumption is tenable. The program can also be used to compute classification consistency and accuracy indices for multiple category classifications with multiple cut scores.

Program Operation

In order to execute the program, the user needs R software installed on their computer (https://www.r-project.org). Additionally, the pbivnorm package needs to be downloaded (https://cran.r-project.org/web/packages/pbivnorm/index.html). To download the pbivnorm package, the user types in "install.packages("pbivnorm") in the R console.

The user loads the pbivnorm package by typing in "library(pbivnorm)." The "pbivnorm" function in the pbivnorm package is used to find the cumulative bivariate frequency.

Index

`nm`	*Find estimates of classification consistency and accuracy using the normal approximation procedure.*

Usage

```
nm(data, rel, cut)
```

Arguments

The following arguments need to be specified.

`data`	A matrix of raw data with a single column
`rel`	reliability estimate for the data
`cut`	A vector of cut score(s)

Note

Running NM-CLASS requires a reliability estimate as an input. Several candidates exist as a reliability estimate. However, as the primary goal, in the context of classification, often is to identify an examinee's absolute performance level with respect to the cut score(s) regardless of the performance of other examinees, a reliability coefficient that involves absolute error variance is suggested. Such reliability coefficients include those from the compound multinomial model (Lee, 2007) or phi coefficient in generalizability theory (Brennan, 2001).

Example

The format of data file is explained through a hypothetical example. The example data file is distributed along with the program.

Data File

```
25
27
22
17
27
7
11
11
23
⋮
```

The example data file includes 100 examinees' summed scores with a possible score range of 0 - 30. Note that each line is associated with a single examinee. Note again that the type of scores that can be used is not limited to summed scores.

Execution of R Code

```
setwd("Desktop")      #set working directory to the folder
                      where the data file is located
data <- read.table("example.dat")    #read the data file
library(pbivnorm)          #load the pbivnorm package
nm(data,.8,c(20,25))       #specify the reliability as .8
                            and two cut scores (20 and 25)
```

In this example, a reliability estimate of .8 is specified. Also, two cut scores (20 and 25) are used. If multiple cut scores are identified, results for simultaneous classification (i.e., when all cut scores are applied simultaneously) will be printed along with results for binary classifications with each cut score being applied one at a time.

Output

The output includes estimates of five classification indices: phi, kappa, gamma, false-positive error rate, and false-negative error rate. Under `$`Binary Classifications`,` each row represents the results for binary classification(s) based on its corresponding cut score. As mentioned previously, the specification of multiple cut scores invokes analysis for simultaneous classification. As such, `$`Simultaneous Classification`` is produced with the number of possible performance categories (i.e., the number of cut scores plus 1).

```
> nm(data, .8, c(20,25))
$`Binary Classifications`
                   PHI     KAPPA     GAMMA FALSE_POSITIVE FALSE_NEGATIVE
cut score 1 0.7983122 0.5888838 0.8547743     0.07661441     0.06861129
cut score 2 0.8907881 0.5330344 0.9232532     0.05150827     0.02523856

$`Simultaneous Classification`
                    PHI     KAPPA    GAMMA FALSE_POSITIVE FALSE_NEGATIVE
3 categories 0.7023259 0.4783902 0.779905      0.1267672     0.09332775
```

References

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer-Verlag.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*, 253-264.

Kim, S. Y., & Lee, W. (2019). Classification consistency and accuracy for mixed-format tests. *Applied Measurement in Education, 32* (2), 97-115.

Kim, S. Y., & Lee, W. (in press). Classification consistency and accuracy with atypical score distributions. *Journal of Educational Measurement.*

Lee, W. (2007). Multinomial and compound multinomial error models for tests with complex item scoring. *Applied Psychological Measurement, 31*, 255-274.

Peng, C. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17*, 359-368.