IRR

Brandon Shields

11/30/2021

#The following Inter-rater reliability analysis will comapre reliability between a human and AI grader. This is for 50 students whose total score could be a value between 1 and 6.

human\_grader <- read.csv(file = "C:/Users/bshield6/OneDrive - Kent State University/Desktop/Copy of Pilot\_Grading 10-28-21.csv")  
  
human\_grader <- human\_grader[1:51,]  
  
#Change Column names so when merged, We can tell the difference between graders  
  
human\_col\_names <- colnames(human\_grader)  
  
colnames(human\_grader) <- paste("Human", human\_col\_names, sep="\_")  
  
human\_grader$User\_Name <- human\_grader$Human\_ï..User.Name  
  
human\_grader <- human\_grader[,-c(1)]  
  
  
#Scores for AI grader  
  
ai\_grader <- read.csv(file = "C:/Users/bshield6/OneDrive - Kent State University/Desktop/Copy of Copy of KSU-Assignment1-10-5-21.csv")  
  
ai\_grader <- ai\_grader[,c(1,3,5:9)]  
  
ai\_grader$Total\_Score <- ai\_grader$Score  
  
ai\_col\_names <- colnames(ai\_grader)  
  
colnames(ai\_grader) <- paste("AI", ai\_col\_names, sep="\_")  
  
ai\_grader$User\_Name <- ai\_grader$AI\_ï..User.Name  
  
ai\_grader <- ai\_grader[,-c(1:2)]

irr\_data\_set <- inner\_join(human\_grader,ai\_grader, by = "User\_Name")  
  
#creating data sets for analysis  
irr\_total\_score <- select(irr\_data\_set,c("Human\_Final\_Score", "AI\_Total\_Score"))  
irr\_focus <- select(irr\_data\_set,c("Human\_Focus", "AI\_Focus"))  
irr\_content <- select(irr\_data\_set,c("Human\_Content", "AI\_Content"))  
irr\_organization <- select(irr\_data\_set,c("Human\_Organization", "AI\_Organization"))  
irr\_language <- select(irr\_data\_set,c("Human\_Language", "AI\_Language"))  
irr\_mechanics <- select(irr\_data\_set,c("Human\_Mechanics", "AI\_Mechanics"))

#Level of agreement  
agree\_total <- irr::agree(irr\_total\_score)  
agree\_focus <- irr::agree(irr\_focus)  
agree\_content <- irr::agree(irr\_content)  
agree\_organization <- irr::agree(irr\_organization)  
agree\_language <- irr::agree(irr\_language)  
agree\_mechanics <- irr::agree(irr\_mechanics)  
  
#level of agreement with a tolerance of one  
  
agree\_total1 <- irr::agree(irr\_total\_score,1)  
agree\_focus1 <- irr::agree(irr\_focus,1)  
agree\_content1 <- irr::agree(irr\_content,1)  
agree\_organization1 <- irr::agree(irr\_organization,1)  
agree\_language1 <- irr::agree(irr\_language,)  
agree\_mechanics1 <- irr::agree(irr\_mechanics,1)  
  
#Limitation with agreement is that it does not consider agreement due to random chance

#Cohen’s kappa is a measure of the agreement between two raters who have recorded a categorical outcome for a number of individuals. Cohen’s kappa factors out agreement due to chance and the two raters either agree or disagree on the category that each subject is assigned to (the level of agreement is not weighted). If you have more than two raters you need to consider an alternative approach which is detailed below.  
  
#To calculate a Cohen’s Kappa the following assumptions need to be met:  
#1. The response being measured by the two raters is categorical (either a nominal or ordinal variable).   
#2. The responses are paired observations of the same measure and the two raters individually assess the measure for each individual.  
#3. The two raters remain fixed i.e. it is the same two raters assessing each individual.  
#4. The two raters are independent of each other.  
#5. For each assessment each rater makes an assessment based on the exact same number and definition of categories.  
  
#https://www.sheffield.ac.uk/polopoly\_fs/1.885169!/file/88\_Kappa.pdf  
  
cohen\_total <- irr::kappa2(irr\_total\_score)  
cohen\_focus <- irr::kappa2(irr\_focus)  
cohen\_content <- irr::kappa2(irr\_content)  
cohen\_organization <- irr::kappa2(irr\_organization)  
cohen\_language <- irr::kappa2(irr\_language)  
cohen\_mechanics <- irr::kappa2(irr\_mechanics)  
  
#The calculation of Cohen’s kappa accounts for disagreement between two raters but does not take into account the level of the disagreement which is an important consideration when the ratings are in ordered categories.  
  
#Any chosen weights can be used; common weights are calculated as either linear or quadratic sets, where k denotes the total number of categories (5 in this example).   
  
#If the difference between the first and second category is equally as important as a difference between the second and third category, etc., use linear weights. Use quadratic weights if the difference between the first and second category is less important than a difference between the second and third category, etc.  
  
  
cohen\_totalW <- irr::kappa2(irr\_total\_score, weight = "equal")  
cohen\_focusW <- irr::kappa2(irr\_focus, weight = "equal")  
cohen\_contentW <- irr::kappa2(irr\_content, weight = "equal")  
cohen\_organizationW <- irr::kappa2(irr\_organization, weight = "equal")  
cohen\_languageW <- irr::kappa2(irr\_language, weight = "equal")  
cohen\_mechanicsW <- irr::kappa2(irr\_mechanics, weight = "equal")  
  
  
#Value of K Strength of agreement  
#< 0.20 Poor  
#0.21 - 0.40 Fair  
#0.41 - 0.60 Moderate  
#0.61 - 0.80 Good  
#0.81 - 1.00 Very good

#Krippendorff’s alpha (also called Krippendorff’s Coefficient) is an alternative to Cohen’s Kappa for determining inter-rater reliability.  
  
#Krippendorff’s alpha:  
  
#Ignores missing data entirely.  
#Can handle various sample sizes, categories, and numbers of raters.  
#Applies to any measurement level (i.e. (nominal, ordinal, interval, ratio).  
  
#Commonly used in content analysis to quantify the extent of agreement between raters, it differs from most other measures of inter-rater reliability because it calculates disagreement (as opposed to agreement). This is one reason why the statistic is arguably more reliable, but some researchers report that in practice, the results from both alpha and kappa are similar  
  
  
  
kripp\_total <- irr::kripp.alpha(as.matrix(irr\_total\_score), method = c("ordinal"))  
kripp\_focus <- irr::kripp.alpha(as.matrix(irr\_focus), method = c("ordinal"))  
kripp\_content <- irr::kripp.alpha(as.matrix(irr\_content), method = c("ordinal"))  
kripp\_organization <- irr::kripp.alpha(as.matrix(irr\_organization), method = c("ordinal"))  
kripp\_language <- irr::kripp.alpha(as.matrix(irr\_language), method = c("ordinal"))  
kripp\_mechanics <- irr::kripp.alpha(as.matrix(irr\_mechanics), method = c("ordinal"))  
  
#Values range from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. Krippendorff suggests: “[I]t is customary to require α ≥ .800. Where tentative conclusions are still acceptable, α ≥ .667 is the lowest conceivable limit

Category <- c("Total Score", "Focus","Content", "Organization", "Language", "Mechanics" )  
Agreement <- c(agree\_total$value, agree\_focus$value, agree\_content$value,  
 agree\_organization$value, agree\_language$value, agree\_mechanics$value)  
  
  
Agreement\_Tolerance <- c(agree\_total1$value, agree\_focus1$value, agree\_content1$value,  
 agree\_organization1$value, agree\_language1$value, agree\_mechanics1$value)  
  
C\_Kappa <- c(cohen\_total$value, cohen\_focus$value, cohen\_content$value,  
 cohen\_organization$value, cohen\_language$value, cohen\_mechanics$value)  
  
WeightedC\_Kappa <- c(cohen\_totalW$value, cohen\_focusW$value, cohen\_contentW$value,  
 cohen\_organizationW$value, cohen\_languageW$value, cohen\_mechanicsW$value)  
  
Kripp\_Alpha <- c(kripp\_total$value, kripp\_focus$value, kripp\_content$value,  
 kripp\_organization$value, kripp\_language$value, cohen\_mechanics$value)  
  
  
IRR\_Summary <- data.frame(Category,Agreement,Agreement\_Tolerance,C\_Kappa,  
 WeightedC\_Kappa,Kripp\_Alpha)  
  
print(IRR\_Summary)

## Category Agreement Agreement\_Tolerance C\_Kappa WeightedC\_Kappa  
## 1 Total Score 45.09804 90.19608 0.24324324 0.3634361  
## 2 Focus 41.17647 90.19608 0.12371134 0.2091272  
## 3 Content 49.01961 92.15686 0.25379854 0.3667219  
## 4 Organization 45.09804 86.27451 0.22769064 0.3013699  
## 5 Language 35.29412 35.29412 0.09418730 0.1633110  
## 6 Mechanics 37.25490 82.35294 0.09634551 0.1404494  
## Kripp\_Alpha  
## 1 0.003928965  
## 2 -0.018907516  
## 3 -0.003533659  
## 4 0.094675356  
## 5 -0.003609754  
## 6 0.096345515

#https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4974794/