Factors affecting intercoder reliability: a Monte Carlo experiment

Feng GC (2013). Quality and Quantity; 47 (2959-2982)

Harvard Catalyst Biostatistics Program Journal Club

06/05/2019

Camden Bay, Ph.D.

Overview of Agreement (2 raters, 2 categories)

- Intercoder reliability = interrater reliability= interrater agreement = consensus = …
- Used in “communication, computational linguistics, marketing, management, psychology, sociology, education, medical science, ecology, and geography” (“Death to Kappa” in remote sensing)
- Cohen’s kappa is the dominant measure of agreement even though there are many viable alternatives
- Measures of agreement are typically “improvements” of basic percent agreement which does not account for chance agreement (i.e., if two raters truly flip a coin for each rating, percent agreement will be 50%)
- Numerous articles have been published proposing new methods of agreement and assessing Cohen’s kappa
- Cohen’s kappa, like its alternatives, can produce “paradoxical” results (rather, hard to interpret)

Overview of Article

- Feng provides an overview of different agreement measures, focusing on how they differ, how they are similar, and how they are affected by prevalence, rater bias, sensitivity, and specificity
  - Rater Bias: Differences in marginal distributions between raters
- Early agreement methods modified percent agreement by accounting for chance agreement in a simple manner (1/number of categories); later ones incorporated information about categories and raters
- To this day, new methods are continuing to be developed, including Gwet's AC1, but all with the same difficulties

Overview of Article

- Feng then discusses criticisms of agreement measures; these often stem from a dependence on prevalence and rater bias and agressive corrections for chance agreement
- Some potential (and controversial) solutions have included calculating agreement within categories and using measures such as PABAK that account for bias and prevalence in Cohen's kappa
- In the second half of the article, Feng devises and presents the results of a simulation study analyzed using response surfaces. This study shows how different agreement measures are affected (linearly or quadratically) by prevalence, sensitivity, and specificity

Example (note prevalence and rater bias)

Rater A/Rater B	I	II
I	1500	212
II	10	40

Agreement Measures
Percent agreement: 0.87 [considered very good…right?]
Cohen's kappa: 0.23 [poor]
Finn's r (modification of ICC): 0.75 [good]
ICC(1): 0.20 [poor]
ICC(2,agreement): 0.23 [poor]
Light's kappa: 0.23 [?]
Krippendorff's alpha: 0.04 [abysmal]
Maxwell's RE: 0.75 [?]
PABAK (bias and variance corrected kappa): 0.75 [good]

All calculated using the R packages irr (Gamer M, Lemon J, Fellows I, Singh P) and epiR (Stevenson M, et al.)

Questions for Discussion

- What do you use?
- Percent agreement is easy to interpret. Is Cohen's kappa? Are any of these agreement measures?
- Some argue for reporting **many** measures of agreement; why not just report the raw contingency table? Is raw percent agreement really so bad?
- What do you do when different methods produce very different results?
- Can we get anything out of this paper, or is agreement still a confusing mess?
- Feng provides recommendations for assessing agreement based on his response-surface analysis on page 2978. Do you think this is feasible?