Interrater Reliability

2024-01-09

The first thing I needed to do for my dissertation was determine whether my research assistants were all in agreement when providing assessments of vaginal pH using litmus test strips. One way to do this is to calculate weighted Kappa values, which will indicate the degree to which different researchers agreed on the pH of the test strips.

First, we need to load in our two data sets- one data frame in which 4 different raters provided ratings for 88 different cases, and another data frame in which 3 different raters provided ratings for 166 cases.

```
three_raters <- read_csv("3 raters 1.9.24.csv")
```

```
## New names:
## Rows: 998 Columns: 25
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): Before/After dbl (5): ID, Session, Haylee, Tara, Gavin lgl (19): ...7,
## ...8, ...9, ...10, ...11, ...12, ...13, ...14, ...15, ...16,...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...24`
## * `` -> `...25`
```

```
four_raters <- read_csv("4 raters 1.9.24.csv")
```

```
## New names:
## Rows: 920 Columns: 26
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (1): Before/After dbl (6): ID, Session, Haylee, Maddie, Tara, Gavin lgl (19):
## ...8, ...9, ...10, ...11, ...12, ...13, ...14, ...15, ...16, ...17...
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
## * `` -> `...11`
## * `` -> `...12`
## * `` -> `...13`
## * `` -> `...14`
## * `` -> `...15`
## * `` -> `...16`
## * `` -> `...17`
## * `` -> `...18`
## * `` -> `...19`
## * `` -> `...20`
## * `` -> `...21`
## * `` -> `...22`
## * `` -> `...23`
## * `` -> `...24`
## * `` -> `...25`
## * `` -> `...26`
```

Then, we can use the kappam.fleiss function in the 'irr' package to calculate weighted kappa for 3 and 4 independent raters, respectively.

```r
three_rate_kappa <- kappam.fleiss(three_raters[, c("Haylee", "Tara", "Gavin")])
```

```r
four_rate_kappa <- kappam.fleiss(four_raters[, c("Haylee", "Maddie", "Tara", "Gavin")])
```

```r
three_rate_kappa
```

```
##  Fleiss' Kappa for m Raters
##
##  Subjects = 166
##    Raters = 3
##     Kappa = 0.531
##
##         z = 17.9
##   p-value = 0
```

```r
four_rate_kappa
```

```
##  Fleiss' Kappa for m Raters
##
##  Subjects = 88
##    Raters = 4
##     Kappa = 0.543
##
##         z = 18.7
##   p-value = 0
```

Weighted Kappa values in both the three-rater and four-rater data frames are not terrible, but they also aren't great. In other words, the data indicate that there was at least some discrepancy in how the research assistants (and myself) interpreted pH values as indicated by the test strips.