Agreement Amongst Movie Reviewers

Greg Johnson

Introduction

Gene Siskel and Roger Ebert are both household names, known for their famous tv show where they each review new and upcoming movies. Agresti & Winner's 1997 article *Evaluating Agreement and Disagreement among Movie Reviewers* aims to explore their rivalry by looking at their agreement on movie ratings and comparing them with another pair of movie reviewers, Michael Medved and Jeffrey Lyons. Data were the ratings of movies from April 1995 to September 1996; the sample size was 160. Reviews were classified into three ordered categories: con, mixed, and pro. Siskel and Ebert were each treated as a separate variable of ordered categories, forming a 3 by 3 two-way contingency table.

```
SEtab <- matrix(c(24, 8, 10, 8, 13, 9, 13, 11, 64), 3, dimnames = list(Siskel = c("Con",
    "Mixed", "Pro"), Ebert = c("Con", "Mixed", "Pro")))
SEtab
```

```
##        Ebert
## Siskel  Con Mixed Pro
##   Con    24     8  13
##   Mixed   8    13  11
##   Pro    10     9  64
```

Agreement between Siskel and Ebert

Agreement was the sum of the elements of the main diagonal of contingency table: the number of reviews that were rated by both as con, or as mixed, or as pro. Siskel & Ebert agreed 63% of the time. Perfect agreement would require all off-diagonal elements to be 0. This would require marginal homogeneity - that row marginal percentages match column marginal percentages (or that Siskel & Ebert have the same distribution of ratings - neither has a tendency to give out more cons than pros or vice versa). Note that marginal homogeneity follows from perfect agreement but the opposite is not necessarily true. As with any statistical analysis, it's important to remember that random chance can produce the percentage of agreement. In fact, if Siskel & Ebert rated independently of each other, their agreement rate would be expected to be 39.6%!

Having shown that there is an agreement, Agresti & Winner explore methods of investigating the strength of the agreement. Cohen's Kappa for categorical scales is introduced. This statistic takes the difference between number of observed agreements and number of expected agreements under independence and standardizes it by the maximum possible difference. For the Siskel & Ebert data, this standardized difference is 0.389. The discrepancy between observed and expected agreements is only 39% of the largest discrepancy possible. When Weighted Kappa is used to incorporate the ordinal nature of the ratings (weighting the differences by severity of disagreement), the discrepancy is still only 43% of the maximum. Not very impressive.

Continuing to explore the nature of the agreement/disagreement of Siskel & Ebert's ratings, Agresti & Winner next look at the structure of the disagreements. Suprisingly, the structure is symmetric! For example, the probability of a Siskel "Con" and an Ebert "Mixed" is the same as the probability of a Siskel "Mixed" and an Ebert "Con." This was tested with a modified Pearson chi-squared test statistic for the six disagreement cells on three degrees of freedom (because the six cells are hypothesized to really just be 3 disagreement pairs). Since the ratings structure was symmetric, it's also marginally homogeneous. However these conditions don't guarantee strong agreement or statistical dependence.

Siskel and Ebert vs. Medved and Lyons

Another way of examining the strength of of Siskel & Ebert's agreement is to compare it to another movie reviewer pair - Michael Medved and Jeffrey Lyons form the comparison group.

```
MLtab <- matrix(c(22, 5, 21, 7, 7, 18, 8, 7, 28), 3, dimnames = list(Lyons = c("Con",
    "Mixed", "Pro"), Medved = c("Con", "Mixed", "Pro")))
MLtab
```

```
##        Medved
## Lyons    Con Mixed Pro
##   Con     22     7   8
##   Mixed    5     7   7
##   Pro     21    18  28
```

The first thing noted is that the contingency table of Medved & Lyons does not fulfill marginal homogeneity - the overall distribution of ratings of the two is very different. Medved tended to have more Con reviews and Lyons tended to have more Pro reviews. This difference is reflected in the low weighted kappa of 0.204, about two standard errors lower than the weighted kappa of Siskel and Ebert. Disagreement between Medved & Lyons is so severe that a quasi-model of independence applied to all cells except one (Medved is Con and Lyons is Pro) fits very well. The pairwise weighted kappas between Siskel & Ebert and Medved & Lyons are all very low (below 0.30). So it looks like relative to Medved & Lyons, Siskel & Ebert are very much in agreement.

So while movie reviewers in general appear to be in weak agreement, perhaps the real takeaway is that each of us should individually find the movie reviewer that most matches our own reviews and listen to them more closely than the other reviewers.

Replicating Weighted Kappa

Agreement on movie reviews is essentially strength of association between two ordinal variables. Therefore in order to compare the agreement of Siskel & Ebert and Medved & Lyons, we simply need to compare their two contingency tables using some sort of measure of strength of association, specifically a measure that incorporates the ordinal nature of the ratings. Agresti & Winner used Weighted Kappa for both pairs which we will replicate here.

```
SE.wk<-Kappa(SEtab,weights="Equal-Spacing")
SE.wk[["Weighted"]]
```

```
##      value        ASE
## 0.42687402 0.06349523
```

```
ML.wk<-Kappa(MLtab,weights="Equal-Spacing")
ML.wk[["Weighted"]]
```

```
##      value        ASE
## 0.20353077 0.07240848
```

Just as Agresti & Winner found, the weighted kappa for Siskel & Ebert was 0.427 and for Medved and Lyons, 0.204. Clearly, Siskel & Ebert have higher agreement, as measured by weighted kappa, than do Medved and Lyons. However we cannot rule out the difference as random error. One quick method to check if the two are statistically close is to construct 95% confidence intervals for both and see if they overlap. For Siskel & Ebert, their true weighted kappa will be in the 95% confidence approximately 95% of the time. For this sample, that confidence interval is (0.30,0.55). For Medved & Lyons, their 95% confidence interval is (0.06,0.35). The intervals do overlap so while Siskel & Ebert have stronger agreement in movie ratings, we cannot rule out the possibility that this is just random chance (though it is quite small).

Measuring Concordance

Another measure of ordinal association is gamma which is, conditional on being untied, the probability of a pair being concordant minus the probability of a pair being discordant. In the context of the movie reviewing data, concordant pairs would be pairs of ratings in which one review was higher rated for both Siskel & Ebert in comparison to the other.

```r
# we're going to use a built-in R function so that we may get standard
# errors and confidence intervals
GKgamma(SEtab)
```

```
## gamma        : 0.625
## std. error   : 0.079
## CI           : 0.471 0.779
```

```r
GKgamma(MLtab)
```

```
## gamma        : 0.327
## std. error   : 0.122
## CI           : 0.088 0.566
```

The proportion of concordant ratings for Siskel & Ebert is 63% larger than the proportion of discordant ratings. This is much higher than Medved & Lyons' 33%. However the 95% confidence interval for the former (0.471,0.779) intersects the confidence interval for the latter (0.088,0.566) so we cannot rule out random chance.

Conclusion

In summary, by two different measures of ordinal association, Siskel & Ebert do agree more than Medved & Lyons but it's hard to rule out random chance as the explanation for the difference.