Matched Pairs

Rater Agreement

Demetris Athienitis

Rater agreement

Sometimes we have matched data where each matched pair consists of ratings by two separate individuals.

- Each of the two individuals rate the same quantity and we are interested in understanding how good their agreement is
- This comes up frequently when there are subjective tests
  - Multiple reviewers are used to improve robustness
  - Interested in understanding how often they agree

Rater agreement

Interested in hypotheses regarding the existence or not of an association but agreement and association are not the same thing. Agreement requires association, but association does not require agreement. E.g.

- Two people can strongly disagree
- One person can consistently review higher than the other

Example (Movie reviews)

Two movie reviewers give their opinion on 160 movies

	Reviewer 2			
Reviewer 1	Con	Mixed	Pro	Total
Con	24	8	13	45
Mixed	8	13	11	32
Pro	10	9	64	83
Total	42	30	88	160

Cohen's Kappa (unweighted)

Let $\pi_{ij} = P(R1 = i, R2 = j)$,

$$P(agree) = \sum_i \pi_{ii} \overset{ind.}{=} \sum_i \pi_{i+}\pi_{+i}$$

$$\kappa = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}$$

- $\kappa = 0$ if agreement only equals that expected under independence
- $\kappa = 1$ if perfect agreement
- Denominator = maximum difference for numerator, attained if agreement is perfect, since perfect agreement implies $\sum_i \pi_{ii} = 1$
- Possible -ve value for κ, implies no effective agreement between the two raters or the agreement is worse than random

Cohen's Kappa (unweighted)

Asymptotic normality can be established

$$\hat{\kappa} \overset{H_0}{\sim} N(0, V(\hat{\kappa}))$$

and calculation of standard error is left to software.

R

There are multiple packages such as `irr`, `psych`, `concord` that have their own functions and their own *weight* scheme. We will use `cohen.kappa{psych}`.

Example (Movie reviews continued)

```
> print(movie)
      Con Mixed Pro
Con    24     8  13
Mixed   8    13  11
Pro    10     9  64
> library(psych)
> cohen.kappa(movie)
Cohen Kappa and Weighted Kappa correlation coefficients
and confidence boundaries
                 lower estimate upper
unweighted kappa  0.27     0.39  0.51
weighted kappa    0.32     0.46  0.60

> sqrt(cohen.kappa(movie)$var.kappa)
[1] 0.05979313
```

Cohen's Kappa (weighted)

Weighted kappa lets you count disagreements differently and is especially useful when codes are ordered. Three matrices are involved:

- matrix of observed scores, n_{ij}
- matrix of expected scores based on independence, $m_{ij} = n_{i+}n_{+j}$,
- weight matrix w_{ij}

Derivations of weighted kappa are sometimes expressed in terms of similarities, and sometimes in terms of dissimilarities. In the latter case, the weights on the diagonal are 1 and off the diagonal are less than 1.

Example (Movie reviews continued)

```
> cohen.kappa(movie)
Cohen Kappa and Weighted Kappa correlation coefficients
and confidence boundaries
                 lower estimate upper
unweighted kappa  0.27     0.39  0.51
weighted kappa    0.32     0.46  0.60

> cohen.kappa(movie)$weight
       Con Mixed  Pro
Con   1.00  0.75 0.00
Mixed 0.75  1.00 0.75
Pro   0.00  0.75 1.00
```

Notice that cells with 0.75 although they represent disagreement it is not as severe as disagreements with 0 weight.

Cohen's Kappa (weighted)

Remark

In `cohen.kappa{psych}` you can also create your own custom weights as an argument to the function.

Repeat the previous example but use 0.5 instead on 0.75 in the weight matrix.

We learned

- Metric for special case of matched pairs: rater agreement
- Cohen's kappa, unweighted and weighted for ordered levels