Agreement, reliability and repeatability studies

Categorical variables

Jonathan Bartlett

London School of Hygiene and Tropical Medicine

15th November 2012

Agreement for categorical data

- Today we consider agreement (‘reliability’) for categorical variables.

Outline

Agreement with truth - detecting disease

- Suppose we are interested in how well a diagnostic test detects disease ($D = 1$) from non-disease ($D = 0$).
- For each subject the test either gives a positive ($X = 1$, indicative of disease) or negative ($X = 0$, indicative of no disease) result.
- For a sample of subjects we obtain their true disease status D and their test result X:

	$X = 0$	$X = 1$	Total
$D = 0$	a	b	$a + b$
$D = 1$	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Agreement with truth - detecting disease

- Sensitivity: $P(X = 1|D = 1)$, estimated by $d/(c + d)$
- Specificity: $P(X = 0|D = 0)$, estimated by $a/(a + b)$
- Positive predictive value: $P(D = 1|X = 1)$, estimated by $d/(b + d)$ (provided prevalence in sample is representitive of population of interest!)

	$X = 0$	$X = 1$	Total
$D = 0$	a	b	$a + b$
$D = 1$	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Outline

Inter-rater agreement

- Comparing 'ratings/scores/assessments' from two raters.

	$R_2 = 0$	$R_2 = 1$	Total
$R_1 = 0$	a	b	$a + b$
$R_1 = 1$	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Percentage agreement

- The most obvious way to summarize agreement is by % agreement.
- We can estimate this by $(a + d)/n$.

	$R_2 = 0$	$R_2 = 1$	Total
$R_1 = 0$	a	b	$a + b$
$R_1 = 1$	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Chance agreement

- But even if the two raters rated at random, we would expect some agreement by chance.
- This idea was the motivation behind Cohen's kappa.
- If the two raters rate randomly, their ratings on a given subject are independent.
- Based on the observed margins, we estimate $P(R_1 = 0)$ by $(a + b)/n$ and $P(R_2 = 0)$ by $(a + c)/n$.

	$R_2 = 0$	$R_2 = 1$	Total
$R_1 = 0$	a	b	$a + b$
$R_1 = 1$	c	d	$c + d$
	$a + c$	$b + d$	$n = a + b + c + d$

Cohen's kappa

- Then 'by chance' we would expect the two to agree with 0 with prob. $P(R_1 = 0) \times P(R_2 = 0)$
- and to agree with 1 with prob. $P(R_1 = 1) \times P(R_2 = 1)$.
- Overall chance agreement (CA) is then $P(CA) = P(R_1 = 0) \times P(R_2 = 0) + P(R_1 = 1) \times P(R_2 = 1)$
- Let $P(OA)$ denote the overall agreement.
- Cohen's kappa is then defined as

$$\kappa = \frac{P(OA) - P(CA)}{1 - P(CA)}$$

Example

```
. tab simpleA simpleB
Radiologis |     Radiologist B´s
    t A´s  |       assessment
assessment |    Normal  Not norma |     Total
-----------+----------------------+----------
    Normal |        21         12 |        33
Not normal |         7         45 |        52
-----------+----------------------+----------
     Total |        28         57 |        85
```

- Overall agreement: $(21+45)/85 = 0.776$
- Agreement expected under independence: $(33/85)\times(28/85)+(52/85)\times(57/85) = 0.538$
- Kappa:

$$\frac{0.776-0.538}{1-0.538} = 0.515$$

The `kap` command

```
. kap simpleA simpleB
                 Expected
Agreement       Agreement      Kappa    Std. Err.          Z      Prob>Z
--------------------------------------------------------------------------
  77.65%         53.81%       0.5160     0.1076          4.80      0.0000
```

Dependence on marginal probabilities / prevalence

- There has been lots of discussion/analysis (over decades) about kappa and its suitability for quantifying agreement.
- Some of this focuses on its dependence on the marginal probabilities / prevalence (e.g. Feinstein 1990).
- Under certain assumptions you can show that kappa varies with prevalence, even when rater's sensitivity/specificity is constant.
- Whether this constitutes a weakness of kappa is a measure is debateable (Vach 2005).
- Indeed, the ICC/reliablity coefficient for continuous measures differs for populations with different levels of heterogeneity.

Extension to more than two categories

- Kappa can also be defined/estimated when we have more than two categories.

```
. use http://www.stata-press.com/data/r12/rate2, clear
(Altman p. 403)

. tabulate rada radb
```

Radiologist A´s assessment	Radiologist B´s assessment Normal	benign	suspect	cancer	Total
Normal	21	12	0	0	33
benign	4	17	1	0	22
suspect	3	9	15	2	29
cancer	0	0	0	1	1
Total	28	38	16	3	85

```
. kap rada radb
```

Agreement	Expected Agreement	Kappa	Std. Err.	Z	Prob>Z
63.53%	30.82%	0.4728	0.0694	6.81	0.0000

Outline

Ordinal/ordered variables

- Sometimes the variable in question is ordinal or has a natural ordering to its levels, e.g. an integer score from 0-4.
- Cohen later proposed a modified version of kappa, in which partial 'credit' is given for 'small' disagreements.
- A number of different weighting schemes have been proposed.

```
. kap rada radb, wgt(w)

Ratings weighted by:
   1.0000   0.6667   0.3333   0.0000
   0.6667   1.0000   0.6667   0.3333
   0.3333   0.6667   1.0000   0.6667
   0.0000   0.3333   0.6667   1.0000

             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  86.67%      69.11%     0.5684     0.0788       7.22      0.0000
```

Weighting schemes

- The choice of weighting, and whether it is appropriate, is important to think about...

```
. kap rada radb, wgt(w2)

Ratings weighted by:
   1.0000    0.8889    0.5556    0.0000
   0.8889    1.0000    0.8889    0.5556
   0.5556    0.8889    1.0000    0.8889
   0.0000    0.5556    0.8889    1.0000

             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  94.77%      84.09%     0.6714     0.1079       6.22      0.0000
```

More than two raters

- Extensions have also been made to the case of more than two raters.
- These are implemented in Stata's `kappa` (different from the `kap` command) command.

Conclusions

- Quantifying agreement with categorical data is more difficult than the continuous case (I think!).
- Arguable whether a single index/parameter can ever sufficiently well summarize agreement in this setting.
- The ideal (?) is to present the contingency table itself, although this becomes tricky with more than two raters.

References

- J. Cohen (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20 (1): 3746
- A. R. Feinstein (1990). High agreement but low kappa: 1. The problems of two paradoxes. Journal of Clinical Epidemiology 43 (6):543-549
- W. Vach (2005). The dependence of Cohen's kappa on the prevalence does not matter. Journal of Clinical Epidemiology 58 (7): 655-661