

CALCULATING COHEN'S KAPPA

A MEASURE OF INTER-RATER RELIABILITY

FOR QUALITATIVE RESEARCH INVOLVING NOMINAL CODING

WHAT IS COHEN'S KAPPA?

- COHEN'S KAPPA IS A STATISTICAL MEASURE CREATED BY JACOB COHEN IN 1960 TO BE A MORE ACCURATE MEASURE OF RELIABILITY BETWEEN TWO RATERS MAKING DECISIONS ABOUT HOW A PARTICULAR UNIT OF ANALYSIS SHOULD BE CATEGORIZED.
- KAPPA MEASURES NOT ONLY THE % OF AGREEMENT BETWEEN TWO RATERS, IT ALSO CALCULATES THE DEGREE TO WHICH AGREEMENT CAN BE ATTRIBUTED TO CHANCE.

JACOB COHEN, A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES, *EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT* 20: 37-46, 1960.

THE EQUATION FOR K

$$\underline{K = Pr(a) - Pr(e)}$$

$$N - Pr(e)$$

N = TOTAL NUMBER
OF RATED ITEMS,
ALSO CALLED
"CASES"

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

THE EQUATION FOR K

THE FANCY "K" STANDS FOR KAPPA

$$\underline{K = Pr(a) - Pr(e)}$$

$$N - Pr(e)$$

N = TOTAL NUMBER
OF RATED ITEMS,
ALSO CALLED
"CASES"

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

THE EQUATION FOR K

THE FANCY "K" STANDS FOR KAPPA

→ $K = \text{Pr}(a) - \text{Pr}(e)$
 $N - \text{Pr}(e)$

N = TOTAL NUMBER
OF RATED ITEMS,
ALSO CALLED
"CASES"

$\text{Pr}(A)$ = SIMPLE
AGREEMENT AMONG
RATERS

$\text{Pr}(E)$ = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

CALCULATING K BY HAND USING A CONTINGENCY TABLE

RATER 1 →

		A	B	C
RATER 2 ↓	A			
	B			
	C			

THE SIZE OF THE TABLE IS DETERMINED BY HOW MANY CODING CATEGORIES YOU HAVE

THIS EXAMPLE ASSUMES THAT YOUR UNITS CAN BE SORTED INTO THREE CATEGORIES, HENCE A 3X3 GRID

CALCULATING K BY HAND USING A CONTINGENCY TABLE

RATER 1 \longrightarrow

		A	B	C
RATER 2 \downarrow	A	# of agreements on A	disagreement	disagreement
	B	disagreement	# of agreements on B	disagreement
	C	disagreement	disagreement	# of agreements on C

THE DIAGONAL HIGHLIGHTED HERE REPRESENTS AGREEMENT (WHERE THE TWO RATERS BOTH MARK THE SAME THING)

DATA: RATING BLOG COMMENTS

- USING A RANDOM NUMBER TABLE, I PULLED COMMENTS FROM ENGLISH LANGUAGE BLOGS ON BLOGGER.COM UNTIL I HAD A SAMPLE OF 10 COMMENTS
- I ASKED R&W COLLEAGUES TO RATE EACH COMMENT: "PLEASE CATEGORIZE EACH USING THE FOLLOWING CHOICES: RELEVANT, SPAM, OR OTHER."
- WE CAN NOW CALCULATE AGREEMENT BETWEEN ANY TWO RATERS

CALCULATING K FOR RATERS 1 & 2

RATER 1 →

		R	S	O	
R A T E R 2 ↓	R	6 (Item #2,3, 4-8)	0	0	6
	S	1 (Item #1)	1 (Item #10)	0	2
	O	2 (Item #4 & 9)	0	0	2
		9	1	0	10

ADD ROWS & COLUMNS
SINCE WE HAVE 10 ITEMS, THE TOTALS SHOULD ADD UP TO 10 FOR EACH

CALCULATING K

COMPUTING SIMPLE AGREEMENT

RATER 1 →

		R	S	O
R A T E R 2 ↓	R	6 (Item #2,3, 4-8)	0	0
	S	1 (Item #1)	1 (Item #10)	0
	O	2 (Item #4 & 9)	0	0

ADD VALUES
OF DIAGONAL
CELLS &
DIVIDE BY
TOTAL
NUMBER OF
CASES TO
COMPUTE
SIMPLE
AGREEMENT
OR
"PR(A)"

$$(6+1)/10$$

THE EQUATION FOR K: RATERS 1 & 2

$$K = 7 - \text{Pr}(e)$$

$$10 - \text{Pr}(e)$$

WE ALSO SUBSTITUTE 10 AS
THE VALUE OF N

RATERS 1 & 2 AGREED ON 70%
OF THE CASES. BUT HOW
MUCH OF THAT AGREEMENT
WAS BY CHANCE?

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

THE EQUATION FOR K: RATERS 1 & 2

WE CAN NOW ENTER THE VALUE OF PR(A)

$$K = 7 - \text{Pr}(e)$$

$$10 - \text{Pr}(e)$$

WE ALSO SUBSTITUTE 10 AS
THE VALUE OF N

RATERS 1 & 2 AGREED ON 70%
OF THE CASES. BUT HOW
MUCH OF THAT AGREEMENT
WAS BY CHANCE?

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

THE EQUATION FOR K: RATERS 1 & 2

WE CAN NOW ENTER THE VALUE OF PR(A)

→ $K = 7 - \text{Pr}(e)$
 $10 - \text{Pr}(e)$

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

WE ALSO SUBSTITUTE 10 AS
THE VALUE OF N

RATERS 1 & 2 AGREED ON 70%
OF THE CASES. BUT HOW
MUCH OF THAT AGREEMENT
WAS BY CHANCE?

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

CALCULATING K

EXPECTED FREQUENCY OF CHANCE AGREEMENT

RATER 1 →

RATER 2 ↓

	R	S	O
R	6 (5.4)	0	0
S	1 (Item #1)	1 (.2)	0
O	2 (Item #4 & 9)	0	0 (0)

FOR EACH
DIAGONAL
CELL, WE
COMPUTE
EXPECTED
FREQUENCY OF
CHANCE (EF)

$$EF = \frac{\text{ROW TOTAL} \times \text{COL TOTAL}}{\text{TOTAL \# OF CASES}}$$

$$EF \text{ FOR "RELEVANT"} = (6 * 9) / 10 = 5.4$$

CALCULATING K

EXPECTED FREQUENCY OF CHANCE AGREEMENT

RATER 1 \longrightarrow

		R	S	O
R A T E R 2 \downarrow	R	6 (5.4)	0	0
	S	1 (Item #1)	1 (.2)	0
	O	2 (Item #4 & 9)	0	0 (0)

ADD ALL
VALUES OF (EF)

TO GET
"PR(E)"

$PR(E) =$

$$5.4 + .2 + 0 =$$

5.6

THE EQUATION FOR K: RATERS 1 & 2

$$K = \frac{7 - 5.6}{10 - 5.6}$$

$$K = .3182$$

THIS IS FAR BELOW THE ACCEPTABLE
LEVEL OF AGREEMENT, WHICH SHOULD
BE AT LEAST .70

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

THE EQUATION FOR K: RATERS 1 & 2

WE CAN NOW ENTER THE VALUE OF PR(E)
& COMPUTE KAPPA

$$K = \frac{7 - 5.6}{10 - 5.6}$$

$$K = .3182$$

THIS IS FAR BELOW THE ACCEPTABLE
LEVEL OF AGREEMENT, WHICH SHOULD
BE AT LEAST .70

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

THE EQUATION FOR K: RATERS 1 & 2

WE CAN NOW ENTER THE VALUE OF PR(E)
& COMPUTE KAPPA

$$\rightarrow K = \frac{7 - 5.6}{10 - 5.6}$$

$$K = .3182$$

THIS IS FAR BELOW THE ACCEPTABLE
LEVEL OF AGREEMENT, WHICH SHOULD
BE AT LEAST .70

PR(A) = SIMPLE
AGREEMENT AMONG
RATERS

PR(E) = LIKLIHOOD
THAT AGREEMENT IS
ATTRIBUTABLE TO
CHANCE

DATA: RATERS 1 & 2

$K = .3182$ How can we improve?

Item #	1	2	3	4	5	6	7	8	9	10
Rater 1	R	R	R	R	R	R	R	R	R	S
Rater 2	S	R	R	O	R	R	R	R	O	S

- LOOK FOR THE PATTERN IN DISAGREEMENTS CAN SOMETHING ABOUT THE CODING SCHEME BE CLARIFIED?
- TOTAL # OF CASES IS LOW, COULD BE ALLOWING A FEW STICKY CASES TO DISPROPORTIONALLY INFLUENCE AGREEMENT

DATA: RATERS 1-5

Item #	1	2	3	4	5	6	7	8	9	10
Rater 1	R	R	R	R	R	R	R	R	R	S
Rater 2	S	R	R	O	R	R	R	R	O	S
Rater 3	R	R	R	O	R	R	O	O	R	S
Rater 4	R	R	R	R	R	R	R	R	R	S
Rater 5	S	R	R	O	R	O	O	R	R	S

CASE 1 SHOWS A PATTERN OF DISAGREEMENT BETWEEN "SPAM" & "RELEVANT," WHILE CASE 4 SHOWS A PATTERN OF DISAGREEMENT BETWEEN RELEVANT & OTHER

EXERCISES & QUESTIONS

Item #	1	2	3	4	5	6	7	8	9	10
Rater 1	R	R	R	R	R	R	R	R	R	S
Rater 2	S	R	R	O	R	R	R	R	O	S
Rater 3	R	R	R	O	R	R	O	O	R	S
Rater 4	R	R	R	R	R	R	R	R	R	S
Rater 5	S	R	R	O	R	O	O	R	R	S

1. COMPUTE COHEN'S K FOR RATERS 3 & 5
2. REVISE THE CODING PROMPT TO ADDRESS PROBLEMS YOU DETECT; GIVE YOUR NEW CODING SCHEME TO TWO RATERS AND COMPUTE K TO SEE IF YOUR REVISIONS WORKED; BE PREPARED TO TALK ABOUT WHAT CHANGES YOU MADE
3. COHEN'S KAPPA IS SAID TO BE A VERY CONSERVATIVE MEASURE OF INTER-RATER RELIABILITY...CAN YOU EXPLAIN WHY? WHAT ARE ITS LIMITATIONS AS YOU SEE THEM?

DO I HAVE TO DO THIS BY HAND?

NO, YOU COULD GO HERE:

<http://faculty.vassar.edu/lowry/kappa.html>