University of York Department of Health Sciences

Measurement in Health and Disease

Cohen's Kappa

Martin Bland

http://martinbland.co.uk/

Percentage agreement: a misleading approach

Answers to the question: 'Have you ever smoked a cigarette?', by Derbyshire school children

		Interview Yes	Interview No	Total
Self-administered questionnaire	Yes	61	2	63
	No	6	25	31
Total		67	27	94

How closely do the children's answers agree?

Percentage agreement = 100×(61+25)/94 = 91.5%.

Can be misleading because it does not take into account the agreement which we would expect even if the two observations were unrelated.

Artificial tabulation of observations by three observers

Obsvr A	Obsvr B Yes	No	Total
Yes	10	10	20
No	10	70	80
Total	20	80	100

Obsvr A	Obsvr C Yes	No	Total
Yes	0	20	20
No	0	80	80
Total	0	100	100

Percentage agreement:

100×(10+70)/100 = 80%   100×(0+80)/100 = 80%

Observer C always chooses 'No'.

Artificial tabulation of observations by two observers

Observer A	Obsvr D Yes	No	Total
Yes	4	16	20
No	16	64	80
Total	20	80	100

Percentage agreement = 68%.

Frequencies equal to those expected under the null hypothesis of independence (chi^2=0.0).

No more agreement than would be expected by chance.

Another example:

Obsvr X	Obsvr Y Yes	No	Total
Yes	1	9	10
No	9	81	90
Total	10	90	100

This time percentage agreement = 82%, best yet.

The frequencies are equal to the expected values, chi^2 = 0.0, and the two "observer's" assessments are unrelated.

Percentage agreement is widely used, but may be highly misleading.

Example, Barrett *et al.* (1990) reviewed the appropriateness of caesarian section in a group of cases, all of whom had had a section due to fetal distress.

Quoted the percentage agreement between each pair of observers in their panel: between 60% and 82.5%.

Barrett, J.F.R., Jarvis, G.J., Macdonald, H.N., Buchan, P.C., Tyrrell S.N., and Lilford, R.J. (1990) Inconsistencies in clinical decision in obstetrics. *Lancet* **336,** 549-551.

Barrett *et al.* (1990): the percentage agreement between each pair of observers in their panel: between 60% and 82.5%.

If they made decisions at random, with an equal probability for ‘appropriate’ and ‘inappropriate’, the expected agreement would be 50%.

If they tended to rate a greater proportion as ‘appropriate’ this would be higher, e.g. if they rated 80% ‘appropriate’ the agreement expected by chance would be 68% ($0.8\times0.8 + 0.2\times0.2 = 0.68$).

In the absence of the percentage classified as ‘appropriate’ we cannot tell whether their ratings had any validity at all.

Esmail, A. and Bland, M. (1990) Caesarian section for fetal distress. *Lancet* **336,** 819.

The proportion of subjects for which there is agreement tells us nothing at all.

To look at the extent to which there is agreement other than that expected by chance, we need a different method of analysis: Cohen's kappa.

p = proportion of units where there is agreement,
p_e = proportion of units which would be expected to agree, by chance.

Cohen’s kappa (κ) is then defined by

$$\kappa = \frac{p - p_e}{1 - p_e}$$

$$\kappa = \frac{p - p_e}{1 - p_e}$$

Kappa = amount by which agreement exceeds chance, divided by maximum possible amount by which agreement could exceed chance.

Answers to the question: 'Have you ever smoked a cigarette?', by Derbyshire school children

		Interview		
		Yes	No	Total
Self-administered	Yes	61	2	63
questionnaire	No	6	25	31
Total		67	27	94

$$p = (61 + 25)/94 = 0.915$$

$$p_e = \frac{(63\times 67)/94 + (31\times 27)/94}{94} = 0.572$$

$$\kappa = \frac{0.915 - 0.572}{1 - 0.572} = 0.801$$

Artificial tabulation of observations by three observers

Obsvr A	Obsvr B Yes	No	Total
Yes	10	10	20
No	10	70	80
Total	20	80	100

Obsvr A	Obsvr C Yes	No	Total
Yes	0	20	20
No	0	80	80
Total	0	100	100

	Obsvr B	Obsvr C
Percentage agreement:	80%	80%
Kappa:	0.37	0.00

Observer A	Obsvr D Yes	No	Total
Yes	4	16	20
No	16	64	80
Total	20	80	100

Percentage agreement: 68%
Kappa: 0.00

$$\kappa = \frac{p - p_e}{1 - p_e}$$

Perfect agreement when all agree so $p = 1$, $\kappa = 1$.

No agreement in the sense of no relationship, $p = p_e$, $\kappa = 0$.

No agreement when there is an inverse relationship, e.g. if children who said no the first time said yes the second and vice versa.

We have $p < p_e$ and so $\kappa < 0$.

The lowest possible value for κ is $-p_e/(1-p_e)$, so depending on p_e, κ may take any negative value.

Thus κ is not like a correlation coefficient, lying between −1 and +1.

Only values between 0 and 1 have any useful meaning.

Kappa is always less than the proportion agreeing, p.

We can see this mathematically because:

$$
\begin{aligned}
p - \kappa &= p - \frac{p - p_e}{1 - p_e} \\
&= \frac{p(1 - p_e) - (p - p_e)}{1 - p_e} \\
&= \frac{p - pp_e - p + p_e}{1 - p_e} \\
&= \frac{p_e - pp_e}{1 - p_e} \\
&= \frac{p_e(1 - p)}{1 - p_e}
\end{aligned}
$$

and this must be greater than 0 because p_e, $1-p$, and $1-p_e$ are all greater than 0.

Hence p must be greater than κ.

Several categories

Answers to a question about cough during day or at night during past two weeks

		Interview			
		Yes	No	Don't know	Total
Self-	Yes	12	4	2	18
administered	No	12	56	0	68
questionnaire	Don't Know	3	4	1	7
Total		27	64	3	94

p = 0.73, p_e = 0.55, κ = 0.41.

Combining the 'No' and 'Don't know' categories

		Interview		
		Yes	No/DK	Total
Self-administered	Yes	12	6	18
questionnaire	No/DK	15	61	76
	Total	27	67	94

p = 0.78, p_e = 0.63, κ = 0.39.

κ does not necessarily increase because p increases.

Physical health of 366 subjects as judged by a health visitor and the subject's general practitioner, expected frequencies in parentheses (data from Lea MacDonald)

General	Health Visitor				
Practitioner	Poor	Fair	Good	Excellent	Total
Poor	2 (1.1)	12 (5.5)	8 (11.4)	0 (4.1)	22
Fair	9 (4.1)	35 (23.4)	43 (48.8)	7 (17.7)	94
Good	4 (8.0)	36 (45.5)	103 (95.0)	40 (34.5)	183
Excellent	1 (2.9)	8 (16.7)	36 (36.8)	22 (12.6)	67
Total	16	91	190	69	366

p = 0.443, p_e = 0.361, κ = 0.13

When categories are ordered, so that incorrect judgments tend to be in the categories on either side of the truth, and adjacent categories are combined, kappa tends to increase.

Physical health of 366 subjects as judged by a health visitor and the subject's general practitioner, expected frequencies in parentheses (data from Lea MacDonald)

General Practitioner	Health Visitor: Poor	Fair	Good	Excellent	Total
Poor	2 (1.1)	12 (5.5)	8 (11.4)	0 (4.1)	22
Fair	9 (4.1)	35 (23.4)	43 (48.8)	7 (17.7)	94
Good	4 (8.0)	36 (45.5)	103 (95.0)	40 (34.5)	183
Excellent	1 (2.9)	8 (16.7)	36 (36.8)	22 (12.6)	67
Total	16	91	190	69	366

$p = 0.443$, $p_e = 0.361$, $\kappa = 0.13$

If we combine the categories 'poor' and 'fair' we get $\kappa = 0.19$. If we then combine categories 'good' and 'excellent' we get $\kappa = 0.31$.

Kappa increases as we combine adjoining categories.

Data with ordered categories are better analysed using weighted kappa.

Example of the use of kappa:

Kappa statistics for a series of questions asked self-administered and at interview

Morning cough, two weeks	0.62
Day or night cough, two weeks	0.41
Morning cough, since Christmas	0.24
Day or night cough, since Christmas	0.10
Ever smoked	0.80
Smokes now	0.82

How large should kappa be to indicate good agreement?

Interpretation of kappa, after Landis and Koch (1977)

Value of kappa	Strength of agreement
<0.20	Poor
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Good
0.81–1.00	Very good

Landis, J.R. and Koch, G.G. (1977) The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-74.

Standard error and confidence interval for κ

The standard error of κ is given by

$$\mathrm{SE}(\kappa) = \sqrt{\frac{p(1-p)}{n(1-p_e)^2}}$$

where n is the number of subjects. The 95% confidence interval for κ is κ-1.96×SE(κ) to κ+1.96×SE(κ) as κ is approximately Normally Distributed, provided np and $n(1$-$p)$ are large enough, say greater than five.

Answers to the question: 'Have you ever smoked a cigarette?', by Derbyshire school children

		Interview		
		Yes	No	Total
Self-administered	Yes	61	2	63
questionnaire	No	6	25	31
Total		67	27	94

p = 0.915, p_e = 0.572, κ = 0.801.

$$\mathrm{SE}(\kappa) = \sqrt{\frac{p(1-p)}{n(1-p_e)^2}} = \sqrt{\frac{0.915\times(1-0.915)}{94\times(1-0.572)^2}} = 0.067$$

95% confidence interval: 0.801-1.96×0.067 to 0.801+1.96×0.067 = 0.67 to 0.93.

Significance test of the null hypothesis of no agreement.

$$\mathrm{SE}(\kappa) = \sqrt{\frac{p(1-p)}{n(1-p_e)^2}} = \sqrt{\frac{p_e(1-p_e)}{n(1-p_e)^2}} = \sqrt{\frac{p_e}{n(1-p_e)}}$$

For the example, SE(κ) = 0.119, κ/SE(κ) = 0.801/0.119 = 6.73, P < 0.0001. This test is one tailed, as zero and all negative values of κ mean no agreement.

Possible to get a significant difference when the confidence interval contains zero.

Problems with kappa

Kappa depends on the proportions of subjects who have true values in each category.

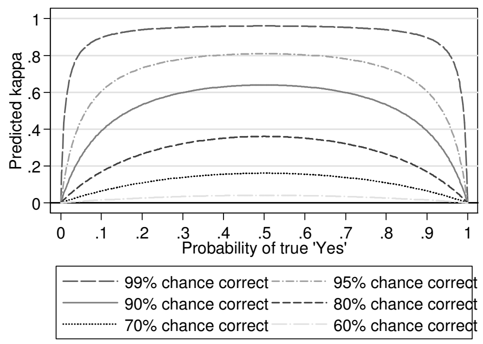
Suppose we have two categories, and the proportion in the first category is p_1, probability that an observer is correct is q, unrelated to the subject's true status.

Expected chance agreement will be

$$\kappa = \frac{p_1(1-p_1)}{\dfrac{q(1-q)}{(1-2q)^2} + p_1(1-p_1)}$$

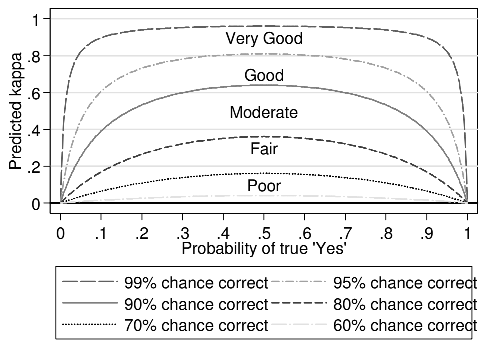
$$\kappa = \frac{p_1(1-p_1)}{\dfrac{q(1-q)}{(1-2q)^2} + p_1(1-p_1)}$$

Kappa depends on proportion of 'yes's.



$$\kappa = \frac{p_1(1-p_1)}{\dfrac{q(1-q)}{(1-2q)^2} + p_1(1-p_1)}$$

Landis and Koch criteria:



Kappa will be specific for a given population.

Like the intra-class correlation coefficient, to which kappa is related, and has the same implications for sampling.

If we choose a group of subjects to have a larger number in rare categories than does the population we are studying, kappa will be larger in the observer agreement sample than it would be in the population as a whole.

When one category is rare, kappa is almost always small.

Weighted kappa

General Practitioner	Health Visitor: Poor	Fair	Good	Excellent	Total
Poor	2 (1.1)	12 (5.5)	8 (11.4)	0 (4.1)	22
Fair	9 (4.1)	35 (23.4)	43 (48.8)	7 (17.7)	94
Good	4 (8.0)	36 (45.5)	103 (95.0)	40 (34.5)	183
Excellent	1 (2.9)	8 (16.7)	36 (36.8)	22 (12.6)	67
Total	16	91	190	69	366

$p = 0.443$, $p_e = 0.361$, $\kappa = 0.13$

Disagreement between 'good' and 'excellent' is not as great as between 'poor' and 'excellent'.

Weight the disagreement.

Weights for disagreement between ratings of physical health as judged by health visitor and general practitioner

General practitioner	Health visitor: Poor	Fair	Good	Excellent
Poor	0	1	2	3
Fair	1	0	1	2
Good	2	1	0	1
Excellent	3	2	1	0

Weight for cell i,j by w_{ij}, the proportion in cell i,j by p_{ij} and the expected proportion in i,j by $p_{e,ij}$, maximum weight, $w_{\max}$.

$$\kappa_w = \frac{p - p_e}{1 - p_e} = \frac{1 - \sum w_{ij} p_{ij} / w_{\max} - \left(1 - \sum w_{ij} p_{e,ij} / w_{\max}\right)}{1 - \left(1 - \sum w_{ij} p_{e,ij} / w_{\max}\right)} = 1 - \frac{\sum w_{ij} p_{ij}}{\sum w_{ij} p_{e,ij}}$$

If all the $w_{ij} = 1$ except on the main diagonal, $w_{ii} = 0$, we get the usual unweighted kappa.

General Practitioner	Health Visitor Poor	Fair	Good	Excellent	Total
Poor	2 (1.1)	12 (5.5)	8 (11.4)	0 (4.1)	22
Fair	9 (4.1)	35 (23.4)	43 (48.8)	7 (17.7)	94
Good	4 (8.0)	36 (45.5)	103 (95.0)	40 (34.5)	183
Excellent	1 (2.9)	8 (16.7)	36 (36.8)	22 (12.6)	67
Total	16	91	190	69	366

$p = 0.443$, $p_e = 0.361$, $\kappa = 0.13$

Weights for disagreement

General practitioner	Health visitor Poor	Fair	Good	Excellent
Poor	0	1	2	3
Fair	1	0	1	2
Good	2	1	0	1
Excellent	3	2	1	0

κ_w=0.23, larger than the unweighted value.

Unweighted $\kappa = 0.13$

Weights for disagreement

General practitioner	Health visitor Poor	Fair	Good	Excellent
Poor	0	1	2	3
Fair	1	0	1	2
Good	2	1	0	1
Excellent	3	2	1	0

κ_w=0.23, larger than the unweighted value.

Alternative weights

General practitioner	Health visitor Poor	Fair	Good	Excellent
Poor	0	1	4	9
Fair	1	0	1	4
Good	4	1	0	1
Excellent	9	4	1	0

κ_w = 0.35.

	Poor	Fair	Good	Excellent
Poor	0	1	2	3
Fair	1	0	1	2
Good	2	1	0	1
Excellent	3	2	1	0

These are sometimes called linear weights. Linear weights are proportional to number of categories apart.

	Poor	Fair	Good	Excellent
Poor	0	1	4	9
Fair	1	0	1	4
Good	4	1	0	1
Excellent	9	4	1	0

These are sometimes called quadratic weights. Quadratic weights are proportional to the square of the number of categories apart.

Weights for agreement

Some programs define weights for agreement instead of Cohen's original weights for disagreement.

Stata does this.

SPSS 16 does not do weighted kappa.

Weights for agreement

Subtract the disagreement weight from the maximum weight, then divide by the maximum:

	Poor	Fair	Good	Excellent
Poor	0	1	2	3
Fair	1	0	1	2
Good	2	1	0	1
Excellent	3	2	1	0

becomes

	Poor	Fair	Good	Excellent
Poor	1	2/3	1/3	0
Fair	2/3	1	2/3	1/3
Good	1/3	2/3	1	2/3
Excellent	0	1/3	2/3	1

Weights for agreement

Subtract the disagreement weight from the maximum weight, then divide by the maximum:

	Poor	Fair	Good	Excellent
Poor	0	1	2	3
Fair	1	0	1	2
Good	2	1	0	1
Excellent	3	2	1	0

becomes

	Poor	Fair	Good	Excellent
Poor	1.00	0.67	0.33	0.00
Fair	0.67	1.00	0.67	0.33
Good	0.33	0.67	1.00	0.67
Excellent	0.00	0.33	0.67	1.00

Weights for agreement

Subtract the disagreement weight from the maximum weight, then divide by the maximum:

	Poor	Fair	Good	Excellent
Poor	0	1	4	9
Fair	1	0	1	4
Good	4	1	0	1
Excellent	9	4	1	0

becomes

	Poor	Fair	Good	Excellent
Poor	1	8/9	5/9	0
Fair	8/9	1	8/9	5/9
Good	5/9	0.89	1.00	0.89
Excellent	0.00	0.55	0.89	1.00

Weights for agreement

Subtract the disagreement weight from the maximum weight, then divide by the maximum:

	Poor	Fair	Good	Excellent
Poor	0	1	4	9
Fair	1	0	1	4
Good	4	1	0	1
Excellent	9	4	1	0

becomes

	Poor	Fair	Good	Excellent
Poor	1.00	0.89	0.55	0.00
Fair	0.89	1.00	0.89	0.55
Good	0.55	0.89	1.00	0.89
Excellent	0.00	0.55	0.89	1.00

Choice of weights

Clearly, we should define these weights in advance rather than derive them from the data.

Cohen (1968) recommended that a committee of experts decide them, but in practice it seems unlikely that this happens.

When using weighted kappa we should state the weights used.

I suspect that in practice people use the default weights of the program.

If we combine categories, weighted kappa may still change, but it should do so to a lesser extent than unweighted kappa.

Agreement between many observers

Ratings of 40 statements as 'Adult', 'Parent' or 'Child by 10 transactional analysts, Falkowski *et al.* (1980)

Statement	Observer									
	A	B	C	D	E	F	G	H	I	J
1	C	C	C	C	C	C	C	C	C	C
2	P	C	C	C	C	P	C	C	C	C
3	A	C	C	C	C	P	P	C	C	C
4	P	A	A	A	P	A	C	C	C	C
5	A	A	A	A	P	A	A	A	A	P
6	C	C	C	C	C	C	C	C	C	C
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
38	C	C	C	C	C	C	C	C	C	P
39	A	C	C	C	C	C	C	C	C	C
40	A	P	C	A	A	A	A	A	A	A

Fleiss (1971) extended Cohen's kappa to the study of agreement between many observers.

Fleiss, J.L. (1971) Measuring nominal scale agreement among many raters. *Psychological Bulletin* **76**, 378-38

Agreement between many observers

Fleiss' method has a problem.

It does not use the identity of the observers.

It assumes that each observation is by a new observer.

Compare observer variation studies where the outcome variable is quantitative: we have two sources of variation, between observers (systematic) and heterogeneity (observer and subject interaction).

Agreement between many observers

Ratings of 40 statements as 'Adult', 'Parent' or 'Child by 10 transactional analysts, Falkowski *et al.* (1980)

Statement	Observer									
	A	B	C	D	E	F	G	H	I	J
1	C	C	C	C	C	C	C	C	C	C
2	P	C	C	C	C	P	C	C	C	C
3	A	C	C	C	C	P	P	C	C	C
4	P	A	A	A	P	A	C	C	C	C
5	A	A	A	A	P	A	A	A	A	P
6	C	C	C	C	C	C	C	C	C	C
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
38	C	C	C	C	C	C	C	C	C	P
39	A	C	C	C	C	C	C	C	C	C
40	A	P	C	A	A	A	A	A	A	A

$\kappa = 0.43$, $P < 0.001$.

There is some agreement, but only moderate.

Agreement between many observers

There is also a weighted version of Fleiss' method.

These methods are not much implemented in software.

Even Stata does not do weighted kappa for many observers.

Conclusions

- Kappa has problems as a measure of agreement.
- It is difficult to interpret, particularly when one category is small.
- Weighted kappa depends on the weights.
- Multi-observer kappas do not deal with the data structure properly.
- There is no other accepted method.