Measures of Agreement

An interesting application is to measure how closely two individuals agree on a series of assessments.

A common application for this is to compare the consistency of judgments of experts. Some examples include

- compare psychiastrists' diagnosis for a patient
- comparing physicians' diagnosis of pneumonia versus bronchitis versus allergy
- comparing physicians' predictions of whether a tumor is malignant or not based on a biopsy, or based on medical images
- comparing two or more reviewers on Amazon have similar reviews of books
- comparing movie reviews from different websites

Measures of Agreement

A common approach to quantifying agreement is called the **kappa statistic**. One thing to note is that this measure only examines how well two reviewers agree with each other, regardless of whether their diagnoses are correct or not.

The kappa statistic puts the measure of agreement on a scale where 1 represents perfect agreement. A kappa of 0 indicates agreement being no better than chance. A difficulty is that there is not usually a clear interpretation of what a number like 0.4 means. Instead, a kappa of 0.5 indicates slightly more agreement than a kappa of 0.4, but there isn't an easy interpretation of a single kappa value. Negative values indicate worse than chance agreement.

Kappa

A typical guideline is

kappa value	interpretation
< 0	Less than chance agreement
0–0.2	Slight agreement
0.2–0.4	Fair agreement
0.4–0.6	Moderate agreement
0.6–0.8	Substantial agreement
0.8–1.0	Almost perfect agreement

Kappa

Part of the hope for people using the kappa statistic is to argue that diagnostic criteria can be used consistently, and that they therefore measure something "real". The easiest case is if there are two raters. Suppose two psychiatrists are diagnosing patients as either psychotic, neurotic, or other. Then we can represent the proportions of diagnoses in a table:

	Rater B			
Rater A	psychotic	neurotic	other	total
psychotic	0.75	0.01	0.04	0.80
neurotic	0.05	0.04	0.01	0.10
other	0	0	0.10	0.10
total	0.80	0.05	0.15	1.00

Kappa

To interept the table, 75% of patients were diagnosed as psychotic by both psychiatrists (this is not a random sample from the general population...). The two raters judged 89% of cases the same way, so by one measure there is 89% agreement.

You might notice however, that the raters have slightly different rates for some diagnoses. Rater A diagnosed 90% of patients as either psychotic or neurotic while Rater B diagnosed 85% as either psychotic or neurotic. The idea of the kappa statistic is to compare the observed amount of agreement divided by the expected amount of agreement based on the fact that the different raters assign different proportions of cases to different categories.

Kappa

The observed agreement can be written as

$$p_o = \sum_{i=1}^{k} p_{ii}$$

The expected amount of agreement due to chance is based on the probability that rater A assigns category i overall and that rater A assigns category i overall. This overall expected agreement due to chance is

$$p_e = \sum_{i=1}^{k} p_{i.} p_{.i}$$

where

$$p_{i.} = \sum_{j=1}^{k} p_{ij}, \quad p_{.i} = \sum_{j=1}^{k} p_{ji}$$

are the marginal proportions. For example $p_{1.}$ is the probability that rater A assigns a patient to category 1.

Kappa

Part of the idea of adjusting for chance agreement is that if some categories are naturally more likely than others, then there might be a lot of agreement due to chance even if the psychiatrists are making judgments independently of one another. It would be like rolling two independent but unfair dice that are both weighted for 6. These two dice will be more likely to agree than two unbiased dice.

Kappa

The overall kappa statistic is then

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

To see the limits on the range of κ, notice that the statistic can be rewritten as

$$1 - \frac{1 - p_o}{1 - p_e}$$

Because the fraction $(1 - p_o)/(1 - p_e)$ is positive, the kappa statistic is clearly less than or equal to 1. A κ value of less than or equal to 0 implies that

$$p_o - p_e \leq 0 \iff p_o \leq p_e$$

I've seen sources online claim that $-1 \leq \kappa \leq 1$. This is probably almost always the case in practice but I'm not sure if it is theoretically guaranteed for all cases.

Kappa

	Rater B			
Rater A	psychotic	neurotic	other	total
psychotic	0.75	0.01	0.04	0.80
neurotic	0.05	0.04	0.01	0.10
other	0	0	0.10	0.10
total	0.80	0.05	0.15	1.00

$$p_o = 0.89$$

$$p_e = (0.80)(0.80) + (0.10)(0.05) + (0.10)(0.15) = 0.66$$

$$\kappa = \frac{.89 - .66}{1 - .66} \approx 0.68$$

This indicates substantial agreement based on typical guidelines.

Kappa

The null hypothesis is that the two raters are independent of one another. If this occurs, then $p_0 \approx p_e$. A way to test this is to convert this to a z-score using the z-score (Fleiss, Cohen, Everitt; 1969):

$$se(\kappa) = \frac{1}{(1-p_e)\sqrt{n}} \sqrt{p_e + p_e^2 - \sum_{i=1}^{k} p_{i.} p_{.i} (p_{i.} + p_{.i})}$$

The z score is then

$$z = \frac{\kappa}{se(\kappa)}$$

For this example with $n = 100$,

$$se(\kappa) = \frac{1}{(1-.66)\sqrt{100}} \sqrt{.66 + .66^2 - 1.0285} = 0.076$$

$$z = 0.68/0.076 = 8.95$$

Suggesting that 0.68 is highly different from 0.

Kappa

The z-score approach is based on a large-sample approximation. It is also possible to simulate the distribution of κ conditional on the marginal proportions. In this case, for $n = 100$, each patient can be thought of as two multinomials, one for each rater, where the category probabilities are the marginal proportions for the two raters.

However, we are conditioning on the number of counts in the marginals, so it makes sense to permute the assignments for the individual patients. This way each simulated κ value has the same p_e value but a random p_o value.

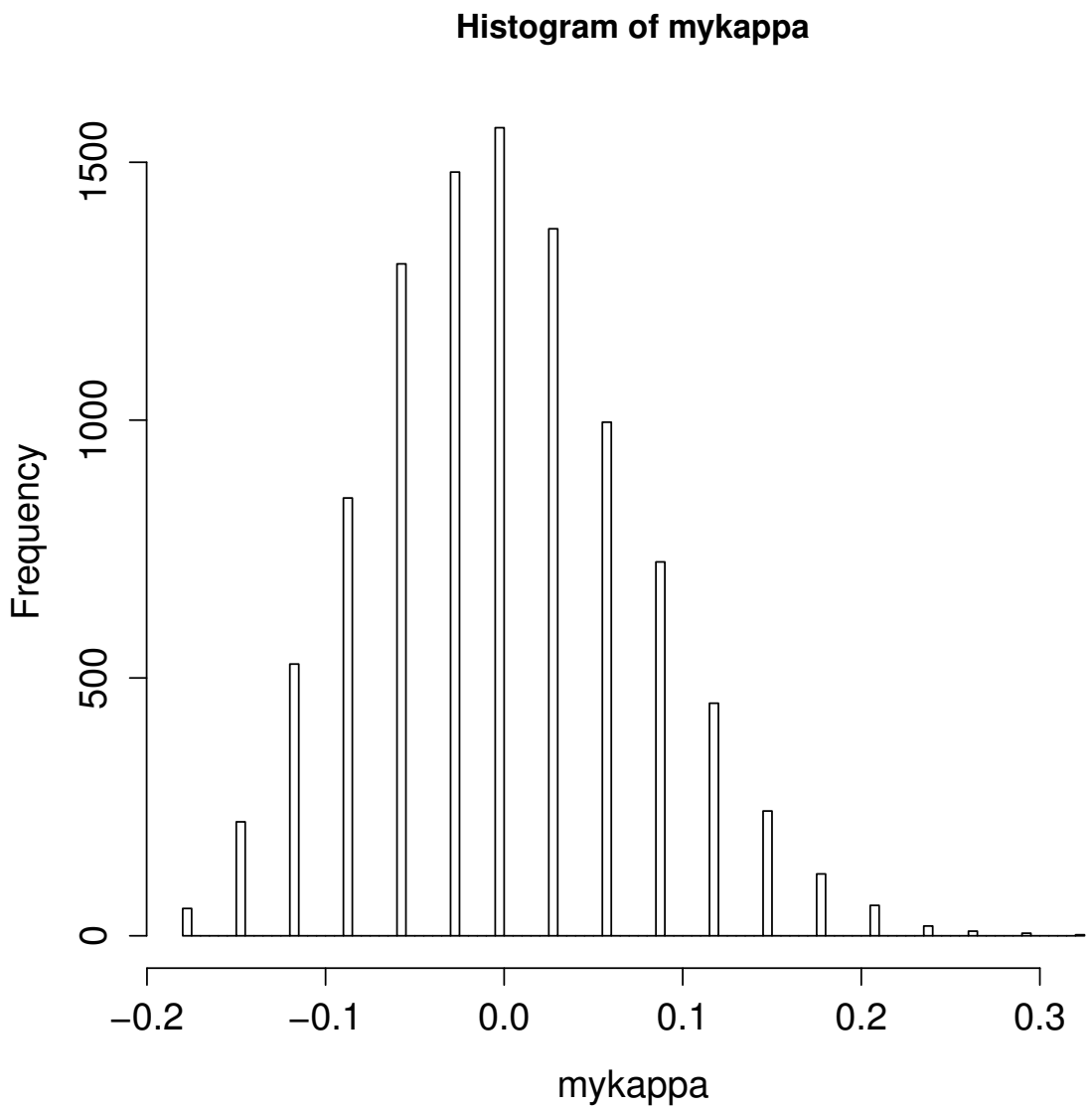
Kappa

```
n=100
raterA <- c(rep(1,80),rep(2,10),rep(3,10))
raterB <- c(rep(1,80),rep(2,5),rep(3,15))

#these are for the n=20 case
#raterA <- c(rep(1,16),rep(2,2),rep(3,2))
#raterB <- c(rep(1,16),rep(2,1),rep(3,3))

pe <- .8^2 + .1*.05 + .1*.15
I <- 10000
n <- 100
mykappa <- 1:I
for(i in 1:I) {
  raterBtemp <- sample(raterB)
  po <- sum(raterA==raterBtemp)/n
  mykappa[i] <- (po-pe)/(1-pe)
}
```

Kappa

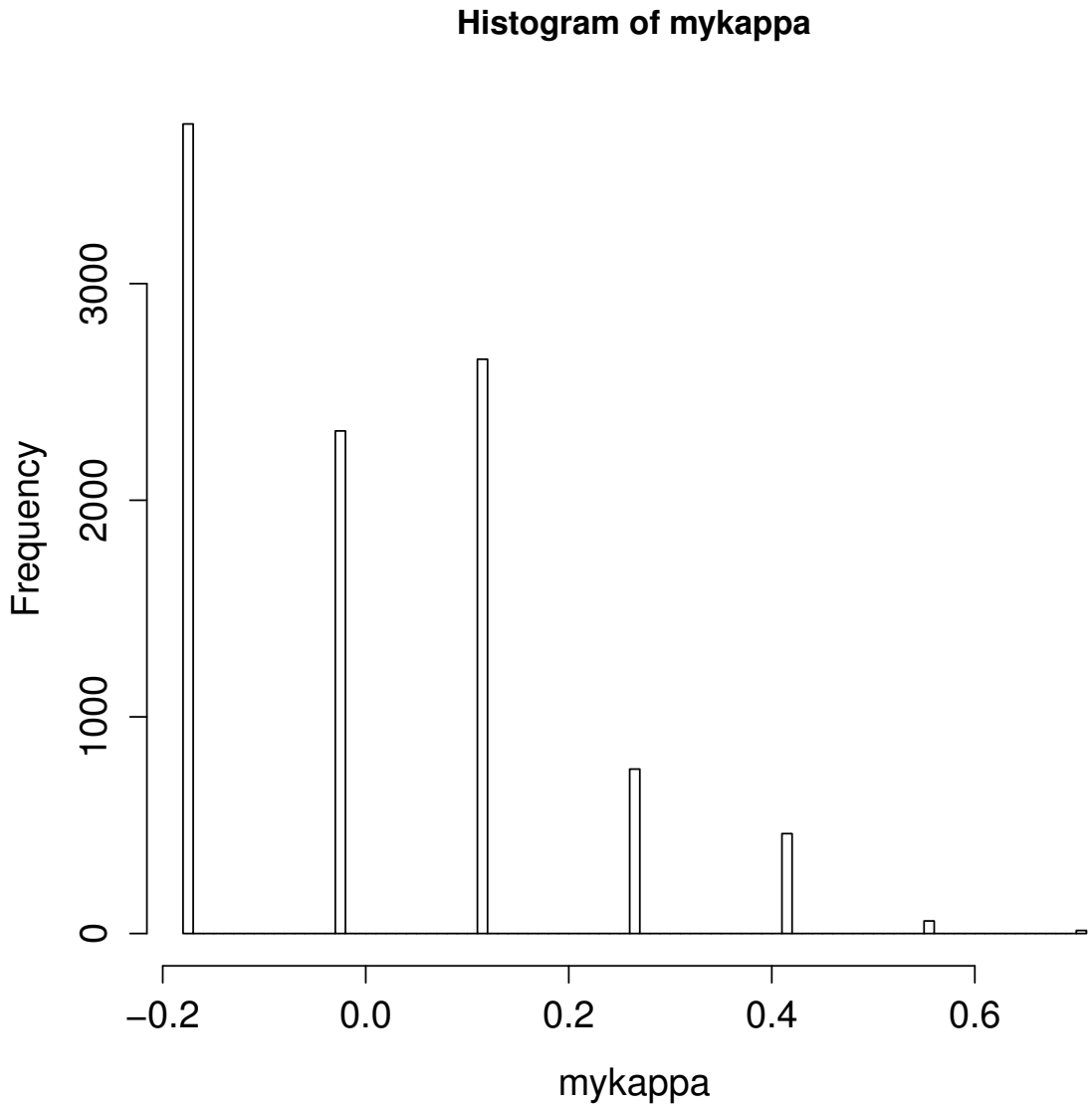

Histogram of mykappa
Frequency
0
500
1000
1500
−0.2
−0.1
0.0
0.1
0.2
0.3
mykappa

Kappa

If the sampe size was smaller (say, $n = 20$), we get a more moderate p-value of 0.017.

```
> source(‘‘kappa.r’’)
> sum(mykappa>.68)/I
[1] 0.0014
```

Kappa


Histogram of mykappa
Frequency
0
1000
2000
3000
−0.2
0.0
0.2
0.4
0.6
mykappa

Kappa

The kappa statistic forms a kind of weird discrete distribution. There are finitely many possible values, but they are typically not simple fractions.

```
> table(mykappa)
mykappa
 -0.176470 -0.029411   0.117647   0.264705
      3737      2320       2651        759
  0.411764   0.558823   0.705882
       461         58         14
```

Kappa

The z score approach gives the following

$$se(\kappa) = \frac{1}{(1-.66)\sqrt{10}}\sqrt{.66 + .66^2 - 1.0285} = 0.170$$

$$z = 0.68/0.170 = 4.0$$

which suggests a one-sided p-value of 3.2×10^{-5} This suggests much stronger evidence against the null than using the simulated distribution. I would trust the simulated distribution for κ more. For smaller sample sizes (say, $n = 10$), the distribution of kappa can be more noticeably skewed (especially left-skewed).

Notice from the table that there were 14 cases out of 10000 simulations where $\kappa \geq 0.68$; hence the simulated p-value is 0.0014.

Kappa when there are 2 raters and 2 categories

What happens with the kappa statistics when there are only two categories?

In this case there is a 2x2 table (just when you thought you knew everything about 2x2 tables!). Let's see what happens.

Kappa when there are 2 raters and 2 categories

In this case let the entries of the table be p_{ij}, $i = 1, 2$, $j = 1, 2$. Let n be the total sample size, and let n_{ij} be the number of individuals in each cell, and $n_{i.}$ and $n_{.}$ the marginal totals. Then the diagonals are

$$p_{ii} = n_{ii}/n$$

Thus

$$p_0 = \frac{n_{11} + n_{22}}{n}$$

The expected cell counts work just like for a χ^2 test; they are the product of the marginal totals divided by the overall total. Thus

$$p_{i.} = n_{i.}/n = (n_{i1} + n_{i2})/n$$
$$p_{.i} = n_{.i}/n = (n_{1i} + n_{2i})/n$$
$$p_e = p_{1.}p_{.1} + p_{2.}p_{.2} = \frac{n_{11} + n_{12}}{n}\frac{n_{11} + n_{21}}{n} + \frac{n_{21} + n_{22}}{n}\frac{n_{12} + n_{22}}{n}$$

Kappa when there are 2 raters and 2 categories

To look at $p_0 - p_e$, you can expand

$$p_0 = p_{11} + p_{22} = \frac{n_{11} + n_{22}}{n} = \frac{(n_{11} + n_{22})(n_{11} + n_{12} + n_{21} + n_{22})}{n^2}$$

After some algebra, $p_0 - p_e$ simplifies to

$$2[p_{11}p_{22} - p_{12}p_{21}]$$

This is the numerator of the kappa statistic. The denominator is still complicated. However the numerator (except for the two) looks similar to the square root of an expression for the $\chi 2$ statistic in 2x2 tables

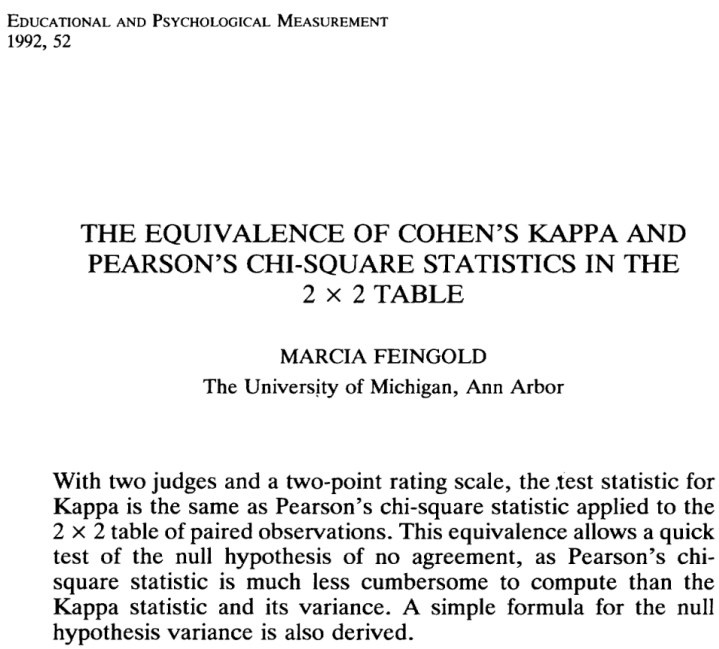
$$\chi^2 = (n(ad - bc)^2)/[(a + b)(a + c)(b + d)(c + d)]$$

where $a = n_{11}$, $b = n_{12}$, $c = n_{21}$ and $d = n_{22}$. The denominator for the kappa statistic is $1 - p_e$, which you can try to manipulate using

$$1 = \frac{n^2}{n^2} = \frac{(n_{11} + n_{12} + n_{21} + n_{22})^2}{n^2}$$

but I don't know if this gives you anything useful. I was hoping to find a simple formula for the kappa statistic in the 2x2 table case.

Kappa



EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1992, 52

THE EQUIVALENCE OF COHEN'S KAPPA AND PEARSON'S CHI-SQUARE STATISTICS IN THE 2 × 2 TABLE

MARCIA FEINGOLD
The University of Michigan, Ann Arbor

With two judges and a two-point rating scale, the test statistic for Kappa is the same as Pearson's chi-square statistic applied to the 2 × 2 table of paired observations. This equivalence allows a quick test of the null hypothesis of no agreement, as Pearson's chi-square statistic is much less cumbersome to compute than the Kappa statistic and its variance. A simple formula for the null hypothesis variance is also derived.

Figure : Equivalence of kappa and chi-square in 2x2 case.

Kappa in R

```
> library(psych)
> x <- c(.75,.01,.04,.05,.04,.01,0,0,.1)
> x <- matrix(x,byrow=T,ncol=3)
> x
     [,1] [,2] [,3]
[1,] 0.75 0.01 0.04
[2,] 0.05 0.04 0.01
[3,] 0.00 0.00 0.10
> a <- cohen.kappa(x,n.obs=20)
> a
                  lower estimate upper
unweighted kappa  0.29     0.68   1.1 # silly: upper limit>1
weighted kappa    0.38     0.76   1.1
> a <- cohen.kappa(x,n.obs=20)
> a$kappa/sqrt(a$var.kappa)
[1] 3.449 # different from what I got out of the Fleiss book
> 1-pnorm(3.449)
[1] 0.0002813
```

Kappa with more than two observers

An extended version of the kappa statistic with more than two observers is called Fleiss' kappa. For this version, let N be the number of subjects, n the number of raters, and k the number of categories. T

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}}$$

where

$$n_{ij} = \text{number of raters assigning subject } i \text{ to category } j$$

$$p_j = \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}, \quad \text{proportion of assignments to category } j$$

$$P_i = \frac{1}{\binom{n}{2}} \sum_{i=1}^{k} \binom{n_{ij}}{2} \quad \text{(agreement on } i\text{th subject)}$$

$$\overline{P} = \frac{1}{N} \sum_{i=1}^{N} P_i; \qquad \overline{P}_e = \sum_{i=1}^{k} p_j^2$$

Kappa with more than two observers

Example with 10 subjects, 14 raters, 5 categories

n_{ij}	1	2	3	4	5	P_i
1	0	0	0	0	14	1.000
2	0	2	6	4	2	0.253
3	0	0	3	5	6	0.308
4	0	3	9	2	0	0.440
5	2	2	8	1	1	0.330
6	7	7	0	0	0	0.462
7	3	2	6	3	0	0.242
8	2	5	3	2	2	0.176
9	6	5	2	1	0	0.286
10	0	2	2	3	7	0.286
Total	20	28	39	21	32	
p_j	0.143	0.200	0.279	0.150	0.229	

Kappa with more than two observers

For this example,

$$\kappa = \frac{0.378 - 0.213}{1 - 0.213} = 0.210$$

suggesting slight agreement.

Kappa with more than two observers

Another R package for computing κ that can handle more than two observers is `kappam.fless()` in the `irr` package.

The input is a matrix where the ij entry is the classification for subject i by rater j. Thus, the matrix is $n \times m$. The previous matrix looks like this:

Kappa with more than two observers

we only need the matrix without the ID column

```
#ID rater1 rater2... rater 14
1  5  5  5  5  5  5  5  5  5  5  5  5  5  5
2  2  2  6  6  6  6  6  6  4  4  4  4  5  5
3  3  3  3  4  4  4  4  4  5  5  5  5  5  5
4  2  2  2  3  3  3  3  3  3  3  3  3  4  4
5  1  1  2  2  3  3  3  3  3  3  3  3  4  5
6  1  1  1  1  1  1  1  2  2  2  2  2  2  2
7  1  1  1  2  2  3  3  3  3  3  3  4  4  4
8  1  1  2  2  2  2  2  3  3  3  4  4  5  5
9  1  1  1  1  1  1  2  2  2  2  2  3  3  4
10  2  2  3  3  4  4  4  5  5  5  5  5  5  5
```

Kappa with more than two observers

```
> library(irr) >
> x <- read.table("fleisskappa.txt") # file with the previous
> kappam.fleiss(x[,2:15])
 Fleiss' Kappa for m Raters

 Subjects = 10
   Raters = 14
    Kappa = 0.23

        z = 14.4
  p-value = 0
```

Kappa with more than two observers

Is the κ value for multiple raters equal to the mean pairwise kappa value? It doesn't appear to be.

```
> mykappa <- NULL
> for(i in 1:14) {
+ for(j in (i+1):15) {
+ m <- x[,c(i,j)] # extract columns for raters i,j
+ mykappa <- c(mykappa,kappam.fleiss(m)$value)
+ }
+ }
> mean(mykappa)
[1] 0.1538346
> kappam.fleiss(x)$value
[1] 0.1886991
```

Kappa with more than two observers

The R function notes in the help details that it implements a slightly different formula from Fleiss, and that Fleiss's formula for multiple raters doesn't reduce to the two-rater kappa statistic when there are only two raters.

Weighted kappa statistic

The weighted kappa statistic allows the user to choose weights to quantify the amount of disagreement between two raters. This can be used for ordered categories. Here you use

$$p_o = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{ij}$$

$$p_e = \sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} p_{i.} p_{.i}$$

where w_{ij} is the weight associated with raters A and B classifying subjects into categories i and j respectively. Typical weight functions are

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \qquad w_{ij} = 1 - \frac{|i-j|}{k-1}$$

The second weight function penalizes by the number of categories of difference.

Cronbach's alpha

Another measure of agreement is Cronbach's alpha, developed in 1951. Cronbach's alpha is supposed to be a measure of reliability, where *reliability* is the ability of an *instrument* (such as a questionnaire) to measure consistently (this is a similar goal as the kappa statistic).

Like the kappa statistic, Cronbach's alpha doesn't measure the ability of an instrument to measure accurately, or to measure what it is intended to measure (this is called *validity*).

Cronbach's alpha

Theoretically, Cronbach's alpha is between 0 and 1, but estimates can be negative when there is little agreement. Cronbach's alpha is often used for Likert scales and compare the variance. Here each row is a subject.

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	Q18
4	5	2	4	4	4	3	5	5	4	5	4	4	4	4	3	5	4
5	5	2	3	4	4	4	4	1	4	4	3	5	3	2	3	4	4
4	3	4	4	2	5	4	4	4	4	3	3	4	2	4	3	4	3
3	5	5	2	3	3	2	4	4	2	4	4	4	3	3	3	4	2
4	5	3	4	4	3	3	4	2	4	4	3	3	4	4	4	4	3
4	5	3	4	2	2	2	4	2	3	4	2	3	3	1	4	3	1
4	3	3	4	4	4	4	4	3	4	4	4	4	4	4	4	4	4
4	4	3	4	4	4	4	4	2	4	4	3	4	4	3	4	4	4
3	3	5	2	1	3	1	1	3	3	1	1	1	1	1	1	1	1
4	2	2	3	4	5	4	4	3	4	4	3	4	5	4	3	4	4

Cronbach's alpha

The formula is

$$\widehat{\alpha} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} S_i^2}{S_y^2}\right)$$

where S_y^2 is the variance of the row sums, and S_i^2 are the individual variances of the the columns. I believe the idea is that this is measuring the within variability, $\sum_{i=1}^{k} S_i^2$, compared to a measure of variability on all of the questions. If the within variability is small, then the ratio should be close to 0, and subtracting from 1 gives you a value close to 1.
For a questionnaire, this is typically applied to a subset of questions that are believed to be measuring the same thing. On a personality test, some questions might essentially test how introverted versus extroverted someone is, how much they depend on feeling versus analytical thinking, etc. As an example, suppose columns 8, 11, and 17 measure "Math Fear".

Cronbach's alpha

Assuming the above is in a data frame *x*, we can type

```
> library(psych)
> library(dplyr)
> mathFear <- select(dat,8,11,17)
> alpha(mathFear)

Reliability analysis
Call: alpha(x = mathFear)

  raw_alpha std.alpha G6(smc) average_r S/N  ase mean   sd
      0.77      0.78    0.78      0.54 3.5 0.29  3.3 0.76

 lower alpha upper     95% confidence boundaries
0.19 0.77 1.34 # another silly CI for the upper boundary
```

Cronbach's alpha

The previous CI is quite wide because there were only 10 observations.
Note that there were three "raters" and we had a quantitative measure.

Cronbach's alpha

Let's see how Cronbach's alpha and the kappa statistic might be related. Here, I'll make random 3x3 tables. In this case, I won't fix the margins.

```
library(psych)
library(irr)

I <- 10000

mykappa <- NULL
myalpha <- NULL
for(i in 1:I) {
x <- sample(1:3,9,replace=T)
x <- matrix(x,ncol=3)
var1 <- var(x[,1]) + var(x[,2])+var(x[,3])
var2 <- var(rowSums(x))
if(var2 !=0) {
myalpha <- c(myalpha,(3/2)*(1-var1/var2))
mykappa <- c(mykappa,kappam.fleiss(x)$value)
}}
```

Kappa versus alpha

Here I only plotted cases where both statistics were positive. The correlation is 0.49 (moderate). This was only 1644 observations out of 10000. I simulated random data, so there was usually little agreement.