Chapter 19

Agreement and the kappa statistic

19. Agreement

Besides the 2×2 contingency table for unmatched data and the 2×2 table for matched data, there is a third common occurrence of data appearing summarised as a 2×2 table. When two judges in a pie-baking contest are asked to rate a series of home-made pies as acceptable or not, or two psychiatrists are asked to rate a set of patients as neurotic or not, or two tests of learning disability are tried on a set of children (such that each test indicates the presence or absence of the disability), one is interested in how much the two measurers agree, the idea being that

1. more agreement is better, because the raters are finding the same quality
2. the more agreement, the more likely we are to rely, in the future, on the opinion of one judge, or one psychiatrist, or one test (probably the shorter one).

There are a number of measures of agreement which have been proposed, such as the overall proportion of agreement. These are discussed quite completely by Fleiss et al. (2003, Chapter 18). However, in this course we shall concentrate on the *kappa* statistic, denoted by κ, whose estimator $\hat{\kappa}$ is defined as

$$(I_o - I_e)/(1 - I_e), \tag{19.1}$$

where I_o is the observed value of any commonly used index (such as the overall proportion of agreement), and I_e is the expected value of this index on the basis of chance alone. The statistic $\hat{\kappa}$ has the properties that

1. if there is complete agreement, then $\hat{\kappa} = 1$,
2. if there is no agreement other than by chance, then $\hat{\kappa} = 0$,
3. if the observed agreement is greater than chance agreement, then $\hat{\kappa} > 0$
4. if the observed agreement is less than chance agreement, then $\hat{\kappa} < 0$

19.1. Two raters

If the frequency of the evaluations of the two raters, of two psychiatrists, or two tests, is given by the following table:

Table 19.1: table for calculation of interrater agreement for 2 raters			
	rater 1		
rater 2	Yes	No	Total
Yes	p_{11}	p_{12}	r_1
No	p_{21}	p_{22}	r_2
Total	c_1	c_2	T

where the p's refer to the **proportion** in each category, and the c's and r's are the column and row **proportions**, respectively, and T is the total **number** of subjects. The value of $\hat{\kappa}$ given by **any** common index of agreement is

$$2(p_{11}p_{22} - p_{12}p_{21})/(c_1r_2 + c_2r_1) \tag{19.2}$$

For example, in the case of two tests of a particular learning disability, we might have

Table 19.2: example of table of frequencies			
	test 1		
test 2	Present	Absent	Total
Present	40	15	55
Absent	10	35	45
Total	50	50	100

which yields the following table of proportions:

Table 19.3: example of table of proportions			
	test 1		
test 2	Present	Absent	Total
Present	0.40	0.15	0.55
Absent	0.10	0.35	0.45
Total	0.50	0.50	100

so that the overall proportion of agreement, denoted by p_o, is $0.40 + 0.35 = 0.75$. The expected agreement by chance, denoted by p_e, is

$$r_1c_1 + r_2c_2 = 0.50(0.55) + 0.50(0.45) = 0.50$$

Using these values and definition (10.1) of $\hat{\kappa}$, we have

$$\hat{\kappa} = (0.75 - 0.50)/(1 - 0.50) = 0.25/0.50 = 0.50$$

Using the second definition (10.2), we have

$$\hat{\kappa} = 2(0.40(0.35) - 0.15(0.10))/(0.50(0.45) + 0.50(0.55))$$

$$= 2(0.14 - 0.015)/0.50 = 0.25/0.50 = 0.50$$

19.2. More than 2 raters

For the case of k categories and two raters, we have the following table:

Table 19.4: table for calculation of interrater agreement for k raters					
	rater 1				
rater 2	1	2	...	k	Total
1	p_{11}	p_{12}	...	p_{1k}	r_1
2	p_{21}	p_{22}	...	p_{2k}	r_2
.					
.					
.					
k	p_{k1}	p_{k2}	...	p_{kk}	r_k
Total	c_1	c_2	...	c_k	T

For this table we do not have a short form of the calculation of $\hat{\kappa}$ such as that given by (10.2). However, the formula (10.1) may be written as

$$\hat{\kappa} = (p_o - p_e)/(1 - p_e)$$

where

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

and

$$p_e = \sum_{i=1}^{k} c_i r_i$$

For example, consider the results of two tests for learning disability where the results are presence, uncertain and absence: sample data is given below:

Table 19.5: Proportions for 3 classes				
	test 1			
test 2	Present	Uncertain	Absent	Total
Present	0.40	0.05	0.05	0.50
Uncertain	0.05	0.10	0.05	0.20
Absent	0.05	0.05	0.20	0.30
Total	0.50	0.20	0.30	1.00

For this data, we have

$$p_o = 0.40 + 0.10 + 0.20 = 0.70,$$

$$p_e = 0.50(0.50) + 0.20(0.20) + 0.30(0.30) = 0.25 + 0.04 + 0.09 = 0.38$$

so that

$$\hat{\kappa} = (0.70 - 0.38)/(1.0 - 0.38) = (0.32)/0.62 = 0.516$$

It has been suggested (by Fleiss et al, 2003, p604) that values of kappa

1. greater than or equal to 0.75 show excellent agreement,
2. between 0.40 and 0.75 represent fair to good agreement,
3. less than or equal to 0.40 show poor agreement.
   These were originally suggested by Landis and Koch(1977).

19.3. Inferences about agreement

For the purpose of testing the hypothesis that $\kappa = 0$, that is, that there is no agreement, except by chance, we may use the following formula for the standard error of $\hat{\kappa}$

$$\frac{\sqrt{[p_e + p_e^2 - \sum r_i c_i (r_i + c_i)]/T}}{(1 - p_e)} \tag{10.3}$$

However, for the more interesting cases of

1. Testing the hypothesis of $H_o\colon \kappa \le 0.40$ versus $H_A\colon \kappa > 0.40$,
2. constructing a confidence interval for κ after having rejected the hypothesis $H_o\colon \kappa = 0$, an alternative, but more complicated, formula must be used. It allows for the fact that κ is non-zero. It is written

   $$\sqrt{[(A + B - C)/T]}/(1 - p_e) \tag{10.4}$$

   where

   $$A = \sum p_{ii}[1 - (r_i + c_i)(1 - \tilde{\kappa})]^2,$$

   $$B = (1 - \tilde{\kappa})^2 \sum_{i=1}^{k} \sum_{j \ne i}^{k} p_{ij}(r_j + c_i)^2,$$

   and

   $$C = [\tilde{\kappa} - p_e(1 - \tilde{\kappa})]^2.$$

   This formula was developed by Fleiss, Cohenn and Everitt (1969).

To test the hypothesis that $\kappa = \kappa_o$, Fleiss, Cohen and Everitt say to use the statistic

$$\frac{|\hat{\kappa} - \kappa_o|}{se(\hat{\kappa})},$$

To construct a confidence interval for κ, use

$$\hat{\kappa} \pm z_{\alpha/2}\, se(\hat{\kappa}).$$

where $se(\hat{\kappa})$ is (10.4) evaluated at $\hat{\kappa}$.

Example 19.1:

Consider the data given in table 19.3 for which we have $\hat{\kappa} = 0.50$:

1. to test $\kappa = 0$, we require

$$\sum r_i c_i (r_i + c_i) = 0.55(0.50)(1.05) + 0.45(0.50)(0.95) = 0.5025$$

so that the standard error of $\hat{\kappa}$ is

$$\sqrt{(0.50 + 0.25 - 0.5025)/100}/0.50 = \sqrt{0.002475}/0.5 = 0.04975/0.50 = 0.0995$$

and the test statistic is $0.50/0.0995 = 5.03$, which is highly significant, (that is, $p < 0.0001$, even for a two-sided alternative),

2. to test $\kappa = 0.40$, we require $se(\kappa = 0.50)$, whose components are

$$A = 0.40[1 - (1.05)(0.50)]^2 + 0.35[1 - (0.95)(0.50)]^2$$

$$= 0.40(0.475)^2 + 0.35(0.525)^2 = 0.09025 + 0.09647 = 0.18672$$

$$B = (0.50)^2\left[0.15(0.50 + 0.45)^2 + 0.10(0.50 + 0.55)^2\right] = 0.25(0.135375 + 0.11025)$$

which is 0.06141, and

$$C = [0.50 - 0.50(0.50)]^2 = 0.25^2 = 0.0625$$

so that the standard error is

$$\sqrt{(0.18671 + 0.06141 - 0.0625)/100}/0.50 = 0.08617$$

and the test statistic is

$$|0.50 - 0.40|/0.08617 = 1.160,$$

so that $p = 0.1230$ (for a one-sided alternative), which is not significant (even at $\alpha = 0.10$),

3. to construct a 95% confidence interval, we use the same standard error, that is 0.08617, and the 95% confidence interval is

$$[0.50 - 1.96(0.08617), 0.50 + 1.96(0.08617)],$$

that is, (0.331,0.669).

Example 19.2:

The data in table 19.5 has 3 categories and we have already shown that $\hat{\kappa} = 0.516$.

1. to test $\kappa = 0$, we require

$$\sum r_i c_i (r_i + c_i) = 0.50(0.50)(1.00) + 0.20(0.20)(0.40) + 0.30(0.30)(0.60)$$

$$= 0.25 + 0.016 + 0.054 = 0.320$$

so that the standard error of $\hat{\kappa}$ is

$$\sqrt{(0.38 + 0.38^2 - 0.320)/100}/0.62 = 0.04521/0.62 = 0.0729$$

and the test statistic is

$$0.52/0.0729 = 7.08,$$

which is highly significant, (that is, $p < 0.0001$),

2. to test $\kappa = 0.40$,we require

$$A = 0.40[1 - (1.00)(0.484)]^2 + 0.10[1 - (0.40)(0.484)]^2 + 0.20[1 - (0.60)(0.484)]^2$$

$$= 0.40(0.516)^2 + 0.10(0.8064)^2 + 0.20(0.7096)^2 = 0.10650 + 0.06503 + 0.10071$$

which is 0.27224,

$$B = (0.484)^2\Big[0.05(0.50 + 0.20)^2 + 0.05(0.50 + 0.30)^2 + 0.05(0.20 + 0.30)^2$$

$$+0.05(0.20 + 0.50)^2 + 0.05(0.30 + 0.50)^2 + 0.05(0.30 + 0.20)^2\Big]$$

$$= 0.36(0.135375 + 0.11025)$$

$$= 0.23426(0.05)(0.49 + 0.64 + 0.25)2 = 0.0323,$$

and

$$C = [0.516 - 0.38(0.484)]^2 = 0.33226^2 = 0.110396$$

so that the standard error is

$$\sqrt{(0.27224 + 0.0323 - 0.110396)/100}/0.62 = 0.07107$$

and the test statistic is

$$|0.516 - 0.40|/0.07107 = 1.632,$$

with $p = 0.0514$ (for a one-sided alternative), so that the result is not significant at $\alpha = 0.05$.

3. to construct a 95% confidence interval, we use the same standard error, 0.07107 so that the 95% confidence interval is

$$[0.516 - 1.96(0.07107), 0.516 + 1.96(0.07107)],$$

that is, (0.377,0.655).

empty page

19.4. Weighted kappa

In a rating scale with more that two levels, the usual kappa statistic considers all the disagreements equally bad, that is, being in cell (1,3) is just as bad as being in cell (1,2). For a nominal scale variable, for example, favourite colour, red, green or blue, this is fine. However for an ordinal scale variable, for disease activity, with levels, none, moderate, much, this would be saying that having two raters saying (none, much) is as bad as having them saying (none, moderate).

One way around this problem when the scales are ordinal is to weight the discrepancies so that those ratings more discordant are considered worse. In computation of the kappa statistic, this is done so that the more discordant are given less weight in computing agreement.

Although there are many ways of calculating weights, the following rules are used by SAS

1 $w_{ii} = 1$, $w_{ij} = w_{ji}$ and $0 \le w_{ij} \ge 1$

2 (following Fleiss, Cohen and Everitt, 1969)

$$p_{o(w)} = \sum_i \sum_j w_{ij} p_{ij}$$

$$p_{e(w)} = \sum_i \sum_j w_{ij} r_i c_j$$

$$\hat{\kappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

3 There are two types of weights used by SAS:

3.1 Cicchetti and Allison (1971) weights:

$$w_{ij} = 1 - \frac{|c_i - c_j|}{c_c - c_1}$$

where c_i is the score for column i (see next section) and c is the number of columns
For a default column score of 1, 2, 3, this yields
$w_{ii} = 1$
$w_{12} = w_{21} = 0.5$
and
$w_{22} = 0$

3.2 Fleiss and Cohen (1973) weights:

$$w_{ij} = 1 - \frac{(c_i - c_j)^2}{(c_c - c_1)_2}$$

For a default column score of 1, 2, 3, this yield $w_{ii} = 1$
$w_{12} = w_{21} = 0.75$
and
$w_{22} = 0$

4 The column scores can vary:

4.1 Table scores are the default. These are the column (and row) numbers; see preceding for example

4.2 input values, for example, 1, 2, 4, rather than the default 1, 2, 3;

4.3 others such as Rank, Ridit and modified Ridit (see SAS documentation);

5 The resulting variance is

$$Var(\hat{\kappa}_w) = \frac{\sum_i \sum_j p_{ij}\left[w_{ij} - (w_{i.} + w_{.j})(1-\hat{\kappa}_w)\right]^2 - \left[\hat{\kappa}_w - P_{e(w)}(1-\hat{\kappa}_w)\right]^2}{(1-P_{e(w)})^2 n}$$

where

$$w_{i.} = \sum_j p_{.j} w_{ij}$$

and

$$w_{.j} = \sum_i p_{i.} w_{ij}$$

6 For hypothesis testing, under the null hypothesis, this variance reduces to

$$Var(\hat{\kappa}_w) = \frac{\sum_i \sum_j p_{i.} p_{.j}\left[w_{ij} - (w_{i.} + w_{.j})\right]^2 - P_{e(w)}^2}{(1-P_{e(w)})^2 n}$$

19.5. Some other estimators

1 Scott's kappa:

The kappa statistic discussed so far, usually referred to as Cohen's kappa (1960), assumes that the rates (proportions) may be different for both raters, so that the

$$\hat{p}_{11} = r_1 c_1$$

so that in the first example, we have

$$\hat{p}_{11} = 0.55 * 0.50 = 0.275$$

However, Scott assumed that the rates (proportions) are the same for both raters so that one has to average the proportion over row and column, as in

$$\frac{r_1 + c_1}{2}$$

so that

$$\hat{p}_{11} = \frac{(r_1 + c_1)(r_1 + c_1)}{4}$$

so that in the first example, we have

$$\hat{p}_{11} = \frac{(0.55 + 0.50)^2}{4} = 0.275625$$

Moreover,

$$p_e = \hat{p}_{11} + \hat{p}_{22}$$

$$= \left(\frac{r_1 + c_1}{2}\right)^2 = \left(\frac{r_2 + c_2}{2}\right)^2$$

In this example, this becomes

$$\frac{(0.55 + 0.50)^2 + (0.45 + .50)^2}{4}$$

$$= \frac{1.1025 + 0.9025}{4} = 0.50125$$

Recall that Cohen's kappa was 0.50.

2 a maximum likelihood estimator:

One defines the common correlation probability model:

$$Pr(X_1 = 1, X_2 = 1) = \pi^2 + \pi(1 - \pi)\kappa$$

$$Pr(X_1 = 1, X_2 = 0) = \pi(1 - \pi)(1 - \kappa)$$

$$Pr(X_1 = 0, X_2 = 1) = \pi(1 - \pi)(1 - \kappa)$$

$$Pr(X_1 = 0, X_2 = 0) = (1 - \pi)^2 + \pi(1 - \pi)\kappa$$

Under this model one has a maximum likelihood estimator and its asymptotic variance. In fact, the maxiumum likelihood estimator is Scott's estimator.

3 Goodness of fit approach:

This is alternative approach to estimating κ in that it produces asymmetric confidence intervals, thus imitating the asymmetric distribution of any estimator of κ.
In this approach, Donner and Eliasziw (199?) suggested

1 Let $P_l = X_1 + X_2$
and let n_i be the number of times i positive scores are indicated;

2 look at the Goodness of fit statistic, which in this case may be written as

$$\sum_{l=1}^{3} \frac{(n_l - N\hat{P}_l)^2}{N\hat{P}_l}$$

which, under the null hypothesis, has a χ_1^2 distribution.

3 If we set this statistic to its critical 0.05-level of 3.84, we have an equation in κ.
This is a cubic equation with 3 possible roots, two of which may be imaginary.

4 If all roots are real, we select two of them to produce upper and lower values of a 95% confidence interval for κ.
In general, if all three roots are real, one of them is outside the interval (-1,1), and the other two are in that interval, and these latter two roots provide the values for the confidence interval.

19.6. Using SAS to calculate kappa statistic

Here is a SAS program for a simple 2x2 table:

```
title1 ’Chapter 19 - agreement ’;
title2 ’Kappa statistic’;
options ls=80 ps=60;
proc format;
     value rating 0 = ’absent’
              1 = ’presnt’;
data mary;
     input rat1 rat2 freq;
     label rat1 = ’rater 1’
     rat2 =’rater 2’;
     format rat1 rating.;
     format rat2 rating.;
datalines;
1 1 40
1 0 15
0 1 10
0 0 35
;
proc freq order=data;
     tables rat1*rat2/kappa;
     exact kappa;
     weight freq;
```

and here is the output:

```
Chapter 19 - agreement                                   1
           Kappa statistic
          The FREQ Procedure
        Table of rat1 by rat2
 rat1(rater 1)     rat2(rater 2)
 Frequency|
 Percent  |
 Row Pct  |
 Col Pct  |presnt  |absent  |  Total
 ---------+--------+--------+
 presnt   |     40 |     15 |     55
          |  40.00 |  15.00 |  55.00
          |  72.73 |  27.27 |
          |  80.00 |  30.00 |
 ---------+--------+--------+
 absent   |     10 |     35 |     45
          |  10.00 |  35.00 |  45.00
          |  22.22 |  77.78 |
          |  20.00 |  70.00 |
 ---------+--------+--------+
 Total          50       50      100
             50.00    50.00   100.00
 Statistics for Table of rat1 by rat2
           McNemar’s Test
       -----------------------
       Statistic (S)    1.0000
       DF                    1
       Pr > S           0.3173
       Simple Kappa Coefficient
   --------------------------------
   Kappa (K)                 0.5000
   ASE                       0.0862
   95% Lower Conf Limit      0.3311
   95% Upper Conf Limit      0.6689
        Test of H0: Kappa = 0
   ASE under H0              0.0995
   Z                         5.0252
   One-sided Pr >  Z         <.0001
   Two-sided Pr > |Z|        <.0001
   Exact Test
   One-sided Pr >=  K     4.178E-07
   Two-sided Pr >= |K|    8.356E-07
          Sample Size = 100
```

Here is the calculation of kappa for a second 2x2 table:

```
title1 ’Chapter 19 - agreement ’;
title2 ’Kappa statistic example 2’;
options ls=80 ps=60;
proc format;
     value rating 0 = ’absent’
              1 = ’presnt’;
data mary;
     input rat1 rat2 freq;
     label rat1 = ’rater 1’
     rat2 =’rater 2’;
     format rat1 rating.;
     format rat2 rating.;
datalines;
1 1 20
1 0 25
0 1 20
0 0 35
;
proc freq order=data;
     tables rat1*rat2/kappa;
     exact kappa;
     weight freq;
```

and the output

```
                Chapter 19 - agreement                            1
                           Kappa statistic example 2
                               The FREQ Procedure
                             Table of rat1 by rat2
                      rat1(rater 1)     rat2(rater 2)
                      Frequency|
                      Percent  |
                      Row Pct  |
                      Col Pct  |presnt  |absent  |  Total
                      ---------+--------+--------+
                      presnt   |     20 |     25 |     45
                               |  20.00 |  25.00 |  45.00
                               |  44.44 |  55.56 |
                               |  50.00 |  41.67 |
                      ---------+--------+--------+
                      absent   |     20 |     35 |     55
                               |  20.00 |  35.00 |  55.00
                               |  36.36 |  63.64 |
                               |  50.00 |  58.33 |
                      ---------+--------+--------+
                      Total          40       60      100
                                  40.00    60.00   100.00
                      Statistics for Table of rat1 by rat2
                                McNemar's Test
                            -----------------------
                            Statistic (S)    0.5556
                            DF                    1
                            Pr > S           0.4561
                            Simple Kappa Coefficient
                        --------------------------------
                        Kappa (K)                 0.0816
                        ASE                       0.0994
                        95% Lower Conf Limit     -0.1133
                        95% Upper Conf Limit      0.2765
                             Test of H0: Kappa = 0
                        ASE under H0              0.0995
                        Z                         0.8206
                        One-sided Pr >  Z         0.2059
                        Two-sided Pr > |Z|        0.4119
                        Exact Test
                        One-sided Pr >=  K        0.2690
                        Two-sided Pr >= |K|       0.5385
                               Sample Size = 100
```

Here is a SAS program for the dataset 2 with 3 categories:

```
title1 ’Chapter 19 - agreement ’;
title2 ’Kappa statistic 3’;
options ls=80 ps=60;
proc format;
     value rating 0 = ’absent’
              1 = ’uncertain’
              2 = ’presnt’;
data mary;
     input rat1 rat2 freq;
     label rat1 = ’rater 1’
     rat2 =’rater 2’;
     format rat1 rating.;
     format rat2 rating.;
datalines;
2 2 40
2 1 5
2 0 5
1 2 5
1 1 10
1 0 5
0 2 5
0 1 5
0 0 20
;
proc freq order=data;
     tables rat1*rat2/agree (wt=fc) norow nocol;
     test agree;
     exact agree;
     weight freq;
```

Note the `(wt=fc)` term in the Proc FREQ; this indicates that the weighting is the Fleiss-Cohen give above; the alternative is `(wt=ca)` for the Cicchetti-Alliston weighting.

Here are the results of this analysis:

```
Chapter 19 - agreement
               Kappa statistic
               The FREQ Procedure
             Table of rat1 by rat2
 rat1(rater 1)     rat2(rater 2)
 Frequency |
 Percent   |presnt  |uncertain|absent  |  Total
           |        |         |        |
 ----------+--------+---------+--------+
 presnt    |     40 |       5 |      5 |     50
           |  40.00 |   5.00  |   5.00 |  50.00
 ----------+--------+---------+--------+
 uncertain |      5 |      10 |      5 |     20
           |   5.00 |  10.00  |   5.00 |  20.00
 ----------+--------+---------+--------+
 absent    |      5 |       5 |     20 |     30
           |   5.00 |   5.00  |  20.00 |  30.00
 ----------+--------+---------+--------+
 Total           50        20       30      100
              50.00     20.00    30.00   100.00
      Statistics for Table of rat1 by rat2
               Test of Symmetry
            ----------------------
            Statistic (S)    0.0000
            DF                    3
            Pr > S           1.0000
            Simple Kappa Coefficient
        --------------------------------
        Kappa (K)                     0.5161 this agrees with the notes
        ASE                           0.0711 this agrees with the notes
        95% Lower Conf Limit          0.3768 this agrees with the notes
        95% Upper Conf Limit          0.6555 this agrees with the notes
             Test of H0: Kappa = 0
        ASE under H0                  0.0729 this agrees with the notes
        Z                             7.0780 this agrees with the notes
        One-sided Pr >  Z             <.0001
        Two-sided Pr > |Z|             <.0001 this agrees with the notes
        Exact Test
        One-sided Pr >=  K     1.342E-11
        Two-sided Pr >= |K|     1.342E-11

      Statistics for Table of rat1 by rat2
           Weighted Kappa Coefficient
        --------------------------------
        Weighted Kappa (K)            0.6053
        ASE                           0.0790
        95% Lower Conf Limit          0.4504
        95% Upper Conf Limit          0.7601
         Test of H0: Weighted Kappa = 0
        ASE under H0                  0.1000
        Z                             6.0526
        One-sided Pr >  Z             <.0001
        Two-sided Pr > |Z|             <.0001
```

```
Exact Test
One-sided Pr >=  K     2.883E-10
Two-sided Pr >= |K|     3.268E-10
       Sample Size = 100
```

19.7. References

Cicchetti, DV, and Allison, T. (1971). A new procedure for assessing reliability of scoring EEG Sleep recordings. *American Journal of EEG Technology* **11**, 101-109.

Cohen, J (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37-46.

Donner, A. and Eliasziw, M. (1992). A goodness-of-fit approach to inference procedures for the kappa statistic: confidence interval construction, significance-testing and sample size estimation. *Statistics in Medicine* **11**, 1511-1519.

Fleiss, JL and Cohen, J (1973). The equivalence of weighted kappa and intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* **33**, 613-619.

Fleiss JL, Cohen J, and Everitt, BS (1969). Large-sample standard errors of kappa and weighted kappa. *Psychological Bulletin* **72**, 323-327.

Scott, WA. (1955). Reliability or content analysis: the case of nominal scale coding. *Public Opinion Quarterly* **19**, 321-325.

19.8. Exercises

1 Two tests (the Denver Development Screening Test(DDST) and the Early Screening Inventory(ESI)) of delayed development in children are used on a sample of 80 children with the following results:

	DDST		
ESI	delayed	not delayed	Total
delayed	40	10	50
not delayed	10	20	30
Total	50	30	80

1.1 Is there any agreement between the two tests.

1.2 Is the agreement fair to good?

1.3 Construct a 99% confidence interval for the coefficient of agreement.

2 Two tests (the Developmental Indicators for the Assessment of Learning (DIAL) and the Miller Assessment for Preschoolers(MAP)) of delayed development in children are used on a sample of 100 children with the following results:

	DIAL		
MAP	delayed	not delayed	Total
delayed	50	10	60
not delayed	10	30	40
Total	60	40	100

Construct a 95% confidence interval for a coefficient of agreement between the two tests.

3 Two tests, the Developmental Sentence Score (DSS) and the Miller Assessment for Preschoolers (MAP), of delayed development in children are used on a sample of 100 children with the following results:

	DSS		
MAP	delayed	not delayed	Total
delayed	50	10	60
not delayed	10	30	40
Total	60	40	100

At $\alpha = 0.01$, investigate the hypothesis that κ, the measure of agreement of the two tests, is less than or equal to 0.40.

4 Two tests of delayed development in children, Developmental Sentence Score (DSS) and Developmental Indicators for the Assessment of Learning (DIAL), were used on a sample of 100 children with the following results:

	DSS		
DIAL	delayed	not delayed	Total
delayed	40	10	50
not delayed	10	40	50
Total	50	50	100

4.1 Why should the measure of agreement be adjusted for "agreement by chance"

4.2 Construct a 95% confidence interval for a coefficient of agreement between the two tests.

5 Two tests (the Developmental Indicators for the Assessment of Learning (DIAL) and the Miller Assessment for Preschoolers(MAP)) of delayed development in children are used on a sample of 100 children with the following results:

	DIAL		
MAP	delayed	not delayed	Total
delayed	50	15	65
not delayed	5	30	35
Total	55	45	100

Construct a 99% confidence interval for a coefficient of agreement between the two tests.

6 Two raters were asked to indicate whether each of 100 children have delayed development or not. Their responses are summarised in the following table.

	rater 1		
rater 2	delayed	not delayed	Total
delayed	51	14	65
not delayed	6	29	35
Total	57	43	100

Construct a 99% confidence interval for a coefficient of agreement between the two raters (the interval may be one-sided or two-sided).