Statistical validation criteria for drinking-water microbiological methods

NIWA Client Report: HAM2003-012
February 2003

NIWA Project: MOH03201

Statistical validation criteria for drinking-water microbiological methods

Graham McBride

Prepared for

Ministry of Health

NIWA Client Report: HAM2003-012
February 2003

NIWA Project: MOH03201

National Institute of Water & Atmospheric Research Ltd
Gate 10, Silverdale Road, Hamilton
P O Box 11115, Hamilton, New Zealand
Phone +64-7-856 7026, Fax +64-7-856 0151
www.niwa.co.nz

Contents

Reviewed by: *Approved for release by:*

James Sukias David Ray

Formatting checked

……

Executive Summary

Methods for establishing equivalence between a "gold standard method" and an alternative candidate method for enumeration of micro-organisms in drinking-water are discussed. The focus is on presence/absence data, because "Maximum Acceptable Values" in current Drinking-Water Standards for New Zealand (2000) define a breach of standard as the presence of *at least one* micro-organism in a given tested volume.

A simple test is proposed for presence/absence data for all micro-organisms in Table 14.1 of the Drinking-Water Standards. First one calculates the lower one-sided 95% confidence limit for Cohen's kappa (being a measure of chance-corrected agreement between the methods). If that limit exceeds 0.6 it can be inferred that the alternative method is equivalent to the standard method. (In more formal technical language this is a one-sided precautionary 5% level test of the hypothesis that the true value of Cohen's kappa statistic is less than 0.6).

It is also concluded that the minimum number of samples to be used in equivalence testing is 50, with the proviso that if the alternative candidate method is borderline for equivalence with that many samples, one should increase this to 150 samples and assess equivalence again.

Some discussion of appropriate methodology for enumeration data is given. The concordance correlation coefficient (first proposed in 1989) is recommended for further investigation (ensuring that recent – 2000 – corrections are included).

1. Introduction

NIWA has been contracted by the Ministry of Health to advise on desirable approaches for establishing validation criteria for microbiological methods that may be proposed as alternatives to the referee methods specified in Section 11 of the "Drinking-Water Standards of New Zealand, 2000" (signed agreement dated 20 December 2002).

We are required to:

- Establish equivalence criteria for membrane filtration *E. coli* enumeration against a referee MPN method (Colilert®).
- Demonstrate the statistical techniques that may be used to assess such equivalence.
- Recommend an optimal number (or a range of numbers) of split-sample analyses for the equivalence assessment calculations.

The agreement notes that the most common scenario will be the need to validate a membrane-filtration *E. coli* method against the referee method (the IDEXX Colilert QuantiTray system), in which case the main questions are:

- How similar should the results be for the method to be considered valid?
- What is the optimal number of split-sample analyses that should be tested for the validation to be considered reliable?

A draft report was required in the contract by 24 January 2003 (it was delivered on 19 January) and a final report by 21 February 2003 (subsequently extended by one week – email from Dr A Kouzminov, 21 February 2003).

This investigation is mostly concerned with presence/absence data, rather than enumeration data (email correspondence with Dr M.E.U. Taylor Ministry of Health on 10 January 2003 refers). This is because the "Maximum Acceptable Values" in the Ministry of Health's Drinking-Water Standards 2000 (Table 14.1) are all stated as being less than 1 organism per unit volume (i.e., MAV for *E. coli*, pathogenic bacteria, viruses, protozoa, helminths and Cyanobacteria).

Accordingly, the main part of this report (section 2) focuses on presence/absence data. However, issues arising for enumeration data are also discussed (in section 3). Recommendations addressing the contract's brief are given in section 4. Some technical detail is placed in footnotes, to aid readability of the main text.

2. Equivalence criteria for presence/absence data

Commonly this situation is addressed using McNemar's test (Fleiss 1981; Zar 1996). Essentially this test examines the effectiveness of alternative treatments (such as a medicine, where there is an effective/ineffective outcome). In such cases the interest lies in testing *association* between treatments and outcomes. Furthermore, in its use there has been a focus on testing a barren "null" hypothesis (that treatments are not merely equivalent, but *exactly* equal in their effect).

However, the situation we are faced with concerns questions of *reasonable agreement* between alternative methods. This invokes the notion of equivalence, not of equality. Accordingly, it seems most appropriate to use a measure of "inter-rater agreement",[1] for which the appropriate candidate is "Cohen's kappa" (Cohen 1960, Bishop *et al*. 1975, Fleiss 1981). Many of the issues to do with this statistic have been presented earlier (McBride 1997), but are repeated here – where appropriate.

2.1 Cohen's kappa

This statistic is "chance-corrected". That is, one may expect a certain amount of agreement by chance alone, and kappa accounts for that possibility. Its basis is as follows. Consider any index that assumes the value 1 when there is complete agreement. Let p_0 denote the observed value of the index and let p_e denote its value expected on the basis of chance alone. Then the observed *excess beyond chance* is $p_0 - p_e$, whereas the maximum possible excess is $1 - p_e$. The ratio of these differences defines kappa (κ), i.e.,

$$\hat{\kappa} = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

[1] This concerns the degree of agreement between two assessors (i.e., raters) when examining the same statistical population.

where the circumflex over κ makes it clear that it is the observed value of κ, being our estimate of its true value based on the data at hand. In this report "kappa" will be used to denote this estimated value, rather than its true value.

Let us now consider the case of two microbiological methods, with one being the gold standard method (e.g., Colilert for *E. coli*) and the other being the alternative method (e.g., a membrane-filtration method for *E. coli*). We can summarise the presence/absence data as in the following table:

Table 1: Presence/absence table.

Alternative	**Standard**	
	Presence	Absence
Presence	n_{pp}	n_{pa}
Absence	n_{ap}	n_{aa}

where the total number of data is $n = n_{pp} + n_{pa} + n_{ap} + n_{aa}$.

These data are then conveniently "normalised" by dividing each item in the data table by their total number, so that we arrive at the following frequency table.

Table 2: Frequency table.

Alternative	**Standard**		
	Presence	Absence	Total
Presence	a	b	p_1
Absence	c	d	q_1
Total	p_2	q_2	1

where $a = n_{pp}/n$, $b = n_{pa}/n$, $c = n_{ap}/n$, $d = n_{aa}/n$, and we have defined marginal frequencies as row and column sums (e.g., $p_1 = a + b$). We can use these marginal frequencies to construct a chance-expected table, as follows

Table 3: Chance-expected table.[2]

Alternative	Standard		
	Presence	Absence	Total
Presence	p_1p_2	p_1q_2	p_1
Absence	q_1p_2	q_1q_2	q_1
Total	p_2	q_2	1

Now the frequency table (Table 2) shows us that the overall proportion of agreement is

$$p_0 = a + d \tag{2}$$

and the chance-expected table (Table 3) shows that the overall proportion of chance-expected agreement is

$$p_e = p_1p_2 + q_1q_2 \tag{3}$$

Cohen's kappa is simply calculated by inserting equations 2 and 3 into equation 1.

2.2 Equivalence criterion

Popular software (SPSS Inc. 1996) advises that a value of kappa beyond 0.75 signifies "excellent agreement", and this has sometimes been used as a criterion for equivalence of microbiological methods (e.g., Whyte & Finlay 1995). This in turn is based on the view expressed in an influential book (Fleiss 1981, page 218): "For most purposes, values greater than .75 *or so* may be taken to represent excellent agreement beyond chance..." (italics added). However, the reasoning for "0.75 or so" is not clear, because

[2] The table is constructed as follows. Consider the case where the number of "present" results obtained by the standard method is independent of the number obtained by the proposed alternative method (and so, similarly, for "absent" results). In that case row totals in the frequency table (Table 2) are independent of the column totals. Now, even if that were the case one can still obtain chance-agreement in individual trials (i.e., "present"/"present" or "absent"/"absent"). The probabilities of this happening, to be inserted into the north-west and south-east cells of the chance-expected table (Table 3), follow from the definition of statistical independence, i.e., Prob(P and A) = Prob(P)Prob(A). Therefore the probabilities in those cells – the frequencies of chance-agreement in a long run of trials – are simply the product of the appropriate row (p_1 and p_2) and column (q_1 and q_2) totals, i.e., p_1p_2 and q_1q_2. A similar argument applies for chance-disagreement, in which case the probabilities to be inserted into the north-east and south-west cells of the table are p_1q_2 and q_1p_2. Note that this distribution of chance-expected probabilities still preserves the row and column totals of Table 2, e.g., in the first row of Table 3, $p_1p_2 + p_1q_2 = p_1(p_2 + q_2) = p_1$.

Fleiss cites a paper by Landis & Koch (1977) in which the following agreement criteria are suggested:

Table 4: Strength of agreement criteria (Landis & Koch 1977).

Kappa statistic	Strength of agreement
< 0.00	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost Perfect

In my view the appropriate criteria should be based on having strong confidence that the *true* kappa value is at least 0.60, in which case the strength of agreement would be at least "Substantial". Of course, that does not mean that the *observed* kappa value merely has to be above 0.6. One must take account of "statistical sampling error" when making inferences from an observation about true values. For example, if our observation is kappa = 0.71 (substantial agreement), it may be that by chance we were lucky and obtained a high value when the true value was in fact 0.58 (moderate agreement). The alternative is also possible of course – obtaining an observed value of 0.58 when the true value was in fact 0.71.

In the context of public health one should take a precautionary approach in handling statistical sampling error. That is, assume the worst situation (e.g., true kappa value is less than 0.60) and then see if the data obtained lead to the conclusion that this assumption should be rejected. This is achieved by testing the following hypothesis:

H: the true value of kappa is *less than* 0.60

A one-sided hypothesis test should then be performed, at the usual significance level of 5%. If *H* is rejected, one has strong grounds for inferring that the true kappa value is *greater* than 0.60, and so agreement is at least "substantial". For such rejection to occur, the observed kappa will have to be somewhat *larger* than 0.60 (precisely because a precautionary approach is being taken).[3]

[3] For this reason, testing a true kappa value of 0.8 ("Almost Perfect") is unrealistic.

The mechanics of the test are as follows. Reject H if the lower 95% one-sided confidence limit on kappa is greater than 0.6, where

$$\text{lower one-sided 95\% confidence limit for } \kappa = \hat{\kappa} - 1.6449 S_{\hat{\kappa}} \quad (4)$$

where "1.6449" is the 95%ile of the unit normal distribution, and "$S_{\hat{\kappa}}$" is the standard error of estimate for kappa (for which formulae are given in Appendix A).

2.2.1 Bayesian interpretation of the test result

Because the Drinking-Water Standards 2000 include the use of Bayesian statistics (as stated in section 1.1 of the Standards), it may be instructive to consider how the test proposed above can be interpreted in a Bayesian way. Fortunately, this is a rather straightforward procedure for a one-sided problem (De Groot 1973, Casella & Berger 1987, Lee 1997).

To elaborate, the test procedure given above derives from the usual "frequentist" paradigm in which probability statements are made about data, assuming a hypothesis to be true. In particular, the hypothesis (H) is rejected if there is a "small" probability of obtaining a kappa value at least as high as was obtained if H is true.[4] This probability is the "p value" and the test's decision rule is to reject H when p is less than an *a priori* significance level (α). Common practice is to take $\alpha = 0.05$. This (i.e., $p < \alpha$) happens when the lower one-sided 95% confidence limit exceeds 0.6. That is, the final statement made about the validity of the tested hypothesis follows from a probability statement made *about data*.

The Bayesian approach enables us to make direct probability statements *about the hypothesis*. This is achieved using Bayes' rule to update prior information or belief in the light of the new data. Assuming that this prior information is "vague"[5] the p-value is the probability of H being true. So if $p < 0.05$ the hypothesis is very likely to be false and we may say that agreement is at least "Substantial" (i.e., that the true value of kappa exceeds 0.6). That is, if the lower one-sided 95% confidence limit exceeds 0.6 we may say that we have at least 95% confidence that the agreement is "Substantial".

[4] It is taken to be *only just* true, i.e., $\kappa = 0.6$.

[5] Sometimes this is referred to as an "uninformative" prior, though no statement is truly uninformative. Technically, one must assume that the prior probability distribution is unimodal and has half its total probability on each side of $\kappa = 0.6$ (Casella & Berger 1987).

2.2.2 Worked example

Say we have 120 pairs of data representing present/absent results from a gold standard method (e.g., Colilert ®) and an alternative method, of which 24 are "present/present", 8 are "present/absent", 5 are "absent/present" and 83 are "absent/absent". Then applying the formula above we obtain:

$p_0 = (24 + 83)/120 = 0.89167$

$p_1 = (24 + 8)/120 = 0.26667$

$p_2 = (24 + 5)/120 = 0.24167$

$q_1 = (5 + 83)/120 = 0.73333$

$q_2 = (8 + 83)/120 = 0.75833$

and so

$p_e = 0.26667 \times 0.24167 + 0.73333 \times 0.75833 = 0.62055$

and Cohen's kappa becomes

$$\hat{\kappa} = \frac{p_0 - p_e}{1 - p_e} = \frac{0.89167 - 0.62055}{1 - 0.62055} = 0.71451$$

Application of the standard error formula in Appendix A results in the estimated standard error $S_{\hat{\kappa}} = 0.07382$. Inserting these values (of $\hat{\kappa}$ and $S_{\hat{\kappa}}$) into equation 4 we obtain a lower one-sided 95% confidence limit of 0.71451 – 1.6449×0.07382 = 0.59308. That is, because kappa is less than 0.6, we do not have sufficient evidence to infer substantial agreement.

2.3 Recommended number of samples

There is never a simple and unequivocal answer to the question "how many samples do I need?" The answer depends particularly on how borderline the agreement might be, and on the cost of obtaining and analysing samples.

However, we can use a fundamental property of the standard error – and hence of confidence limits – to establish workable ranges of desired number of samples. That property is that the standard error varies with the reciprocal square root of the number of samples.[6]

Figure 1 indicates this pictorially. It was constructed using the above calculation procedure (via a computer program) for a range of numbers of samples (from 5 to 200), assuming that the frequencies remain unaltered from that obtaining for 120 samples as the number of samples is changed.[7] Two cases are shown, with a marginal value of kappa (0.71) and a more satisfactory value (0.81).

In each case we see the upward curvature of the lower confidence limit tapering off with increasing number of samples, as may be expected from the above-noted reciprocal square-root dependency. Two important features should be noted:

- an ever diminishing increase in precision (i.e., decrease in width between the kappa estimate and its confidence limit) as one increases the number of samples, with rapidly diminishing returns (given the effort to be spent in collecting and analysing samples) in the region of 40–60 samples.

- the larger the value of kappa, the earlier the lower 95% one-sided confidence limit crosses the critical $\kappa = 0.6$ line (in which case the two methods could be declared to be in substantial agreement, and therefore equivalent).

These observations lead to the inference that a minimum number of samples to be used in equivalence studies is on the order of 50, with the proviso that if a method is borderline for equivalence with that many samples, one could increase this to 150 samples and judge equivalence again.[8]

[6] To see this, note that the identical standard error formulae in the Appendix (eqs A.1 and A.5) both have $\sqrt{n}$ in their denominators. All other terms in those equations are in terms of frequencies, and are therefore independent of n.

[7] Of course, these curves are somewhat artifactual as the value of kappa itself will change as the number of data changes. However, for simplicity – and without loss of generality – we have ignored this. (A more rigorous procedure would require extensive Monte Carlo simulation modelling, beyond the scope of this project.)

[8] Strictly this contravenes the "stopping rule" (Berger & Berry 1998). Certainly it would be *ultra vires* if one kept on increasing the number of samples and stopping to look each time to see if the confidence limit passes over the 0.6 value and, when that happens, inferring equivalence and stopping.

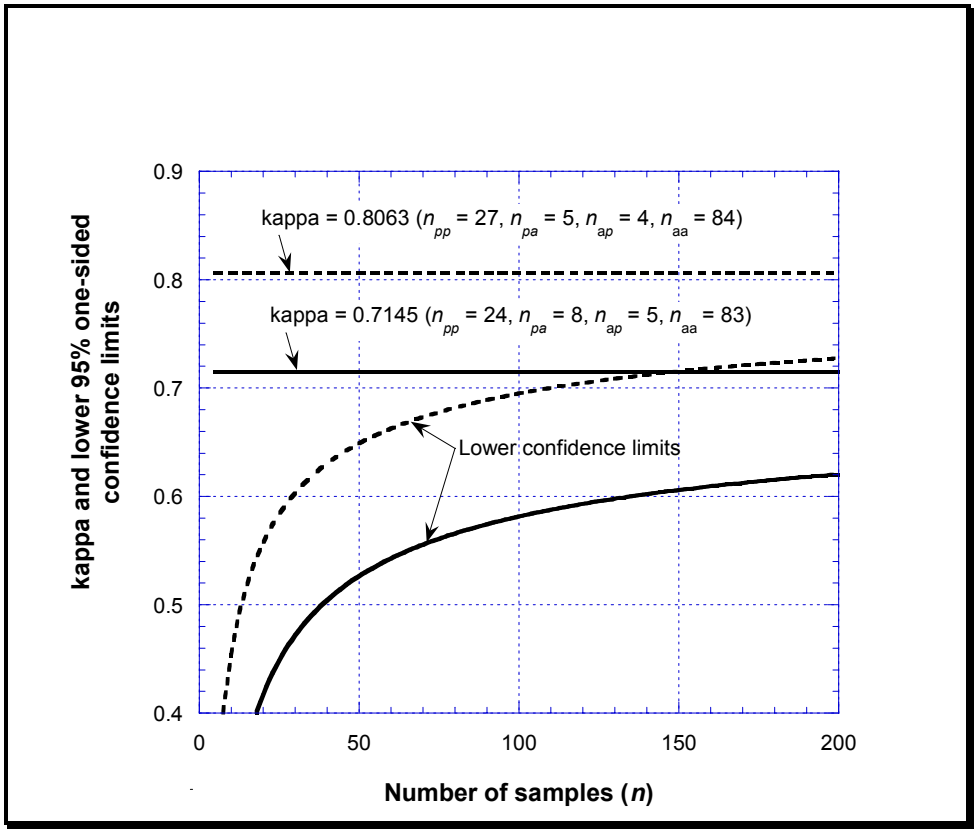

Figure 1: Variation of the lower confidence limit with number of samples for two cases (marginal and acceptable).

3. Equivalence criteria for continuous data

Particularly for environmental samples, there is some literature on establishing equivalence between methods on the basis of their ability to enumerate samples (e.g., Rippey *et al*. 1987, Abbott *et al*. 1998). Various measures have been used therein, all with some shortcomings in this (more complicated) case of enumeration data.

The sample concordance correlation coefficient (denoted as $\hat{\rho}_c$) was first proposed by Lin (1989) for assessment of concordance in continuous data. It represents a breakthrough in assessing concordance between alternative methods where the data are continuous (i.e., enumerative, not discrete),[9] in that it appears to avoid *all* the shortcomings associated with the panoply of usual procedures (Pearson correlation coefficient *r*, paired *t*-tests, least squares analysis for slope and intercept, coefficient of variation, intraclass correlation coefficient). It also appears to be superior to the previously proposed limits-of-agreement procedure (Altman & Bland 1983, Bland & Altman 1986), as discussed by Steichem & Cox (2002). The full equations are listed in Appendix B.

The concordance correlation coefficient can range from –1 to +1, as does Pearsons *r*, but it cannot exceed *r* in absolute value. It is robust on as few as 10 pairs of data (Lin 1989). It appears, with a worked example, in a recent edition of a popular biostatistical text (Zar 1996).[10] It is important to note that there are typographical errors in Lin's original paper (and therefore repeated in Zar's book). Corrections have recently been published (Lin 2000), along with the claim that they have negligible effect. However, it has been shown by Steichen & Cox (2002) that the errors can be problematic when the assessed relationship approaches strong concordance – highly relevant in studies of equivalence.

It would appear that this measure could be examined for equivalence criteria, possibly using Monte Carlo techniques.[11]

[9] Enumerative data may at first be thought to be discrete, but the use of dilutions and reporting to standard volumes (e.g., 100 mL in the case of *E. coli*) effectively makes the data continuous.
[10] Zar denotes the coefficient as r_c.
[11] This too is beyond the scope of the current project.

4. Recommendations

For presence/absence data use a one-sided precautionary 5% level test of the hypothesis that the true value of Cohen's kappa is less than 0.6 (Cohen's kappa being a measure of chance-corrected agreement). If this hypothesis is rejected the conclusion can be made that the agreement between the two microbiological methods is at least "substantial", in which case equivalence between them may be inferred. This procedure is equally applicable to all micro-organisms listed in Table 14.1 of the Drinking-Water Standards for New Zealand (2000).

An optimal number of samples for performing such a test is 50, with the proviso that if a candidate method is borderline for equivalence with that many samples one could increase this to 150 samples and assess equivalence again.

For enumeration data use the recent (2000) amendments to Lin's concordance correlation procedure.

5. References

Abbott, S.; Caughley, B.; Scott, G. (1998). Evaluation of Enterolert® for the enumeration of enterococci in the marine environment. *New Zealand Journal of Marine and Freshwater Research 32*: 505–513.

Altman, D. G.; Bland, J. M. (1983). Measurement in medicine: the analysis of method comparison studies. *Statistician 32*: 307–317.

Berger, J. O.; Berry, D. A. (1988): Statistical analysis and the illusion of objectivity. *American Scientist 76*: 159–165.

Bland, J. M.; Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, February 8*: 307–310.

Casella, G.; Berger, R. L. (1987). Reconciling Bayesian and Frequentist evidence in the one-sided testing problem. *Journal of the American Statistical Association 82*: 106–111.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20(1)*: 37–46.

DeGroot, M. H. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association 68*: 966–969.

Fleiss, J. L. (1981). Statistical Methods for Rates and Proportions. Wiley, New York.

Landis, J. R.; Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics 33*: 159–174.

Lee, P. M. (1997). Bayesian Statistics: An Introduction. 2nd ed. Arnold, London.

Lin, L.I-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics 45*: 255–268.

Lin, L.I-K. (2000). A note on the concordance correlation coefficient. *Biometrics 56*: 324–325.

McBride, G. B. (1997). Evaluation of Colilert® procedures using Cohen's "kappa" statistic as the measure of agreement with a standard. NIWA Consultancy Report SCJ70201. Report to Environmental Diagnostics Ltd. 19 p.

Rippey, S. R.; Adams, W. N.; Watkins, W. D. (1987). Enumeration of fecal coliforms and *E. coli* in marine and estuarine waters an alternative to the APHA-MPN approach. *Journal of the Water Pollution Control Federation 59(8)*: 795–798.

SPSS Inc. (1996). Systat 6.0 for Windows. SPSS Inc. Illinois, Chicago.

Steichen, T. J.; Cox, N. J. (2002). A note on the concordance correlation coefficient. *Stata Journal 2(2)*: 183–189.

Whyte, R.; Finlay, R. (1995). Monitoring the microbiological quality of drinking-waters. *Water and Wastes in New Zealand (November)*: 43–60.

Zar, J. H. (1997). Biostatistical Analysis. 3rd ed. Prentice Hall, Upper Saddle River, NJ.

Appendix A – Equation for standard error of estimates for kappa

Two equivalent forms of the standard error for $\hat{\kappa}$ estimated from a 2x2 frequency table appear in the literature. In both cases we transform the nomenclature in that literature to be consistent with that used in this report.

First, the standard error formula presented by Fleiss (1981, p. 221) may be written as

$$S_{\hat{\kappa}} = \frac{\sqrt{A + B - C}}{(1 - p_e)\sqrt{n}}, \tag{A.1}$$

where n is the total number of paired data, p_e is the chance-expected agreement proportion as defined in the main text (eq. 3) and

$$A = a\left[1 - (p_1 + p_2)(1 - \hat{\kappa})\right]^2 + d\left[1 - (q_1 + q_2)(1 - \hat{\kappa})\right]^2, \tag{A.2}$$

$$B = (1 - \hat{\kappa})^2\left[b(p_2 + q_1)^2 + c(p_1 + q_2)^2\right], \text{ and} \tag{A.3}$$

$$C = \left[\hat{\kappa} - p_e(1 - \hat{\kappa})\right]^2. \tag{A.4}$$

where a, b, c and d are presence/absence frequencies (see Table 2).

Second, the formula presented by Bishop *et al*. (1975, p. 396) may be written as

$$S_{\hat{\kappa}} = \frac{\sqrt{E + F + G}}{(1 - p_e)\sqrt{n}}, \tag{A.5}$$

where [noting that their $\theta_1 = p_0$, as defined in the main text (eq. 2), and that their $\theta_1 = p_e$]

$$E = p_0(1 - p_0), \tag{A.6}$$

$$F = 2(1 - p_0)\frac{2p_0p_e - \theta_3}{1 - p_e}, \text{ and} \tag{A.7}$$

$$G = (1 - p_0)^2\frac{\theta_4 - 4p_e^2}{(1 - p_e)^2}, \tag{A.8}$$

with

$$\theta_3 = a(p_1 + p_2) + d(q_1 + q_2), \text{ and} \tag{A.9}$$

$$\theta_4 = a(p_1 + p_2)^2 + b(p_2 + q_1)^2 + c(p_1 + q_2)^2 + d(q_1 + q_2)^2 . \tag{A.10}$$

Using simple (tedious) algebra we can show that $A + B - C = E + F + G$, and so the two standard error formulations (A.1 and A.5) are in fact identical. This has also been tested in a computer program for many specific cases.

Appendix B – Equations for the concordance correlation coefficient

The *true* value of the concordance correlation coefficient ρ_c is defined as

$$\rho_c = 1 - \frac{\delta}{\delta^*}, \tag{B.1}$$

where

$$\delta = \text{expected squared perpendicular deviation from the } 45^{\circ}_0 \text{ line}, \tag{B.2}$$

$$\delta^* = \delta \text{ for uncorrelated data,} \tag{B.3}$$

where 45°_0 is the 1:1 concordance line passing through the origin.

To proceed, let us denote our n pairs of data as (X_1, Y_1), (X_2, Y_2), …, (X_n, Y_n). Also we define sample means and variances in the following manner, i.e.,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \; S_X^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \tag{B.4}$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \; S_Y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2, \tag{B.5}$$

and the sample covariance

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right). \tag{B.6}$$[12]

Then the *sample* concordance correlation coefficient (see Lin 1989 for its derivation) is

$$\hat{\rho}_c = \frac{2S_{XY}}{S_X^2 + S_Y^2 + \left(\overline{X} - \overline{Y}\right)^2}. \tag{B.7}$$[13]

[12] Note that the divisor for these S terms is n, whereas it is usual to use n–1 to ensure that the estimates are unbiased (unbiasedness is not always the most desirable property).

A standard error can be calculated for this estimate. In doing so it has been shown that the "Fisher transformation" is desirable to better meet the normal approximations to be invoked when confidence intervals are calculated or hypothesis tests are performed (Lin 1989). This transformation is

$$\hat{Z} = \tanh^{-1}(\hat{\rho}_c) = \frac{1}{2}\log_e\left(\frac{1+\hat{\rho}_c}{1-\hat{\rho}_c}\right), \tag{B.8}$$

and we obtain the sample standard error of estimate (of $\hat{Z}$) as

$$S_{\hat{Z}} = \sqrt{\frac{\left(1-r^2\right)\hat{\rho}_c^2}{\left(1-\hat{\rho}_c^2\right)r^2} + \frac{2\hat{\rho}_c^3(1-\hat{\rho})u^2}{r\left(1-\hat{\rho}_c^2\right)^2} - \frac{\hat{\rho}_c^4 u^4}{2r^2\left(1-\hat{\rho}_c^2\right)^2}}{n-2}}, \tag{B.9}[14]$$

where r is Pearson's correlation coefficient defined in the usual way [$r = S_{XY}/(S_X S_Y)$] and u is the "location shift relative to the scale" parameter, defined by

$$u = \left(\frac{n-1}{n}\right)\frac{\left(\overline{X}-\overline{Y}\right)}{\sqrt{S_X S_Y}}. \tag{B.10}$$

Then the lower one-sided confidence 95% confidence interval for Z is

$$L_Z = \text{lower one - sided 95\% confidence limit for } Z = \hat{Z} - 1.6449 S_{\hat{Z}}, \tag{B.11}$$

and so, inverting the transformation in eq. (B.8) the 95% lower confidence limit for ρ_c is

$$L_{\rho_c} = \tanh(L_Z). \tag{B.12}$$

[13] The n–1 term in the denominator of Zar's version of this equation (eq. 18.76) is wrong: it should be just n.

[14] In Lin (1989) and in Zar (1996) the second group of terms under the radical had a coefficient of 4 on the numerator and the third group had a coefficient of 2 on the numerator. These were corrected by Lin (2000), as shown in eq. (B.8), i.e., the first coefficient is 2 on the numerator and the second becomes a 2 *on the denominator*.