Help for Dichotomous Agreement Program v.3.2a

Paul Barrett
Advanced Projects R&D Ltd.
Auckland
New Zealand

email: paul@pbarrett.net
Web: www.pbarrett.net

August 6th, 2014

Table of Contents

1 Measures of Dichotomous Agreement

The program uses a 2x2 crosstabulation table to compute two Chi-Square based hypothesis tests of independence between 2 variables: the Pearson Chi-Square and the Maximum Likelihood Chi-Square. These two tests, both based upon 1 degree of freedom, indicate whether there is any evidence to suggest that the observed frequencies of occurrence for the two variables could be due to chance only. The probabilities associated with the two Chi Square values give the estimate of the likelihood of observing the calculated values, given a Null Hypothesis of independence (which states that there is no difference between expected frequencies of chance occurrence and the observed frequencies). As the probability gets smaller, you can interpret this as indicating that the Null Hypothesis is becoming more and more unlikely to be true.

Specific Measures of Agreement
Pearson r/Phi
Phi/Phi-Max
Yule Q (Gamma)
Jaccard
G-Index (Hamman coefficient) ... point symmetry adjusted Phi
Bennett B-Index ... marginal symmetry adjusted Phi
Cohen's Kappa … for interrater agreement
Weighted Kappa

Medical/Decision Table Parameters
Sensitivity
Quality of Sensitivity
Specificity
Quality of Specificity

PPP (ppv, PVP-Positive Predictive Power)
NPP (npv, PVN-Negative Predictive Power)
PE (Predictive Efficiency)
RIOC (Relative Improvement Over Chance)

Base Rate (BR) or Prevalence
Level (of a test)

False Positive Rate (False Alarm)
False Negative Rate

Odds of Outcome Given Treatment or Predicted
Odds of Outcome Not Given Treatment (or not Predicted)
Odds Ratio
Relative Risk

Cohen's d` Effect Size
Estimated r (from d')

2 The Pearson Chi-Square

This is calculated as:

ChiSq = [(**A**-EfA)^2/EfA] +[(**B**-EfB)^2/EfB] + [(**C**-EfC)^2/EfC] + [(**D**-EfD)^2/EfD]

where:
The degrees of freedom = 1
A, B, C, D = Observed frequencies for Cells **A, B, C, D** respectively
EfA, EfB, EfC, EfD = the Expected frequencies under a Null Hypothesis of probabilistic independence between the two variables.

^2 = squared e.g. (D-EfD)^2 = (D-EfD)*(D-EfD)

Be careful in interpreting any Chi-Square that is computed on a table where at least one expected frequency is less than 3. The Yates correction is not used in this implementation - see Howell(1992) p. 135-136 for reasons. Also, the Pearson Chi-Square is sensitive to small sample size (< about 30). If you are using sample sizes less than this number, it is recommended that you use the Maximum Likelihood Chi-Square. This is less sensitive to small sample size bias.

Reference
Howell, D.C. (1992) Statistical Methods for Psychology 3rd. Edition. Duxbury Press

3 The Maximum Likelihood Chi-Square

This is calculated as:

ChiSq = 2*{[**A***Ln(**A**/EfA)] + [**B***Ln(B/EfB)] + [**C***Ln(C/EfC)] + [**D***Ln(D/EfD)] }

where:
The degrees of freedom = 1
A, B, C, D = Observed frequencies for Cells A, B, C, D respectively
EfA, EfB, EfC, EfD = the Expected frequencies under a Null Hypothesis of probabilistic independence between the two variables.
Ln = Log to the base e

If you are using sample sizes less than about 30, it is recommended that you use the Maximum Likelihood Chi-Square. This is less sensitive to small sample size bias. See pages 144-145 in Howell (1992).

Reference
Howell, D.C. (1992) Statistical Methods for Psychology 3rd. Edition. Duxbury Press

4 The Pearson r/Phi Coefficient

This is calculated here as:

Phi = (A*D-B*C)/SQRT((A+B)*(C+D)*(A+C)*(B+D))

Alternative Formula #2 ... Phi = SQRT(ChiSq/N)

where:
SQRT = square root
ChiSq = the calculated Chi Square value
N = Total Number of Observations (A+B+C+D)

It varies in value between -1.0 and + 1.0 *[Note: usually Phi is noted as varying between* **0.0 and 1.0** *- this is because the assignation of sign is quite arbitrary, depending upon how you code the data (1, 0 or 0, 1). Further, the second formula given above does not permit the calculation of negative values. The reason for using the first formula is to permit comparison with* Yule's Q *- which is also a signed coefficient that varies between -1.0 and +1.0]*

Phi is equivalent to the calculation of a conventional Pearson r correlation computed over two arrays of data, that have only two values, a 1 or 0. It is very sensitive (as is the Pearson) to skewed data (unequal numbers of 1s and 0s in each data array). That is, its value is attenuated by the amount of skew in each of the variables - as the proportions move away from a 50/50, 1 or 0 split, so its value is reduced (see Phi/Phi Max). Hence, the recommended use of the G-Indexor B-Index. These coefficients correct for the bias caused by skew in two different ways - see the helpfile sections for more details.

Further, phi is related to the kappa coefficient and weighted kappa. Take a look at these definitions for an explanation of their similarity.

5 Yule's Q Index

This is calculated as:

Q = (A*D - B*C)/(A*D + B*C)

It varies in value between **-1.0 and + 1.0** (0.0 = no agreement)

The Yule Q coefficient is a special case (in 2x2 tables) of the **gamma** coefficient of ordinal relationship, being based upon the ratio of cross products of the matrix. Essentially, the coefficient is a normed odds-ratio. The formula above for Q or gamma can be expressed as:

$$Q = (OR - 1)/(OR + 1)$$

where OR = odds ratio

WARNING ... Where any cell count is equal to 0, then Q will automatically be -1.0 or +1.0, regardless of the values in the other cells. In the program, the value of Q is set to "Invalid" under these conditions.

References
Ott, R.L., Larson, R., Rexroat, C., Mendenhall, W. (1992*) Statistics: A Tool for the Social Sciences. 5th Edition*. PWS-Kent. ISBN: 0-534-92931-1

Kraemer, H., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., and Kupfer, D.J. (1999) Measuring the Potency of Risk Factors for Clinical or Policy Significance. *Psychological Methods*, 4, 3, 257-271

6 The Jaccard Index

This is calculated as:

$$J = \mathbf{A}/(\mathbf{A}+\mathbf{B}+\mathbf{C})$$

This index varies between **0.0** (no agreement) and **1.0** (maximum agreement).

It represents the probability of a pair of variables exhibiting both of a pair of attributes when only those cases exhibiting one or the other variable are considered. i.e. it only uses cases that have an observation on either one, or both variables. It excludes cases where neither variable is "scored". For example, in a symptom checklist, only cases that respond YES or PRESENT on either symptom, or both, are used. Cases that respond NO or ABSENT on both symptoms are excluded. The rationale for this coefficient is that if no indication is present on either symptom (variable), then really there is no information that can be realistically used here in computing a measure of agreement between two variables or symptoms. Simply because a person does not have either symptom may not be suficient reason to conclude that the absence of one symptom is related to the absence of another symptom. A more direct measure is to exclude these "null" events and to base your measure of agreement on the ratio of agreements (A) to disagreements (B and C). This coefficient is highly recommended for work in the clinical and typological areas - where checklists of symptoms or objects is prevalent. However, you must justify your rationale appropriately.

7 The G-Index

This is calculated as:

$$G = ((A+D)-(B+C))/N$$

This index varies between **-1.0** (maximum negative agreement) through **0.0** (no agreement) to +**1.0** (maximum positive agreement).

Also known as the Hamman (1961) coefficient, and Holley and Guilford's G-index (1964). This is a point symmetry adjusted Phi coefficient [**point symmetry** = in a 2x2 table of profiles, it is realised if complementary patterns of 0s and 1s have similar frequencies]. A complementary pattern to 1,0 is 0,1 .. 1,1, and 0,0. If complementary profiles have different frequencies of occurrence, they may be averaged to produce equality, thus correcting the response skew that can so badly affect the phi coefficient. The main assumption with the G-Index is that it assumes a median response split at the population level (a 50/50

response split on each variable). Bennett's B-Index, in contrast, only assumes that each variable has a similar population distribution.

The use of either the G or B-Index is recommended where response skew may be a problem. However, little work has been done with this coefficient - use of either the G or B index requires some thought and consideration of the consequences of symmetry adjustment.

References

Hamman, U. (1961) Merkmalsbestand und Verwandtschaftsbeziehungen der Farinose. Ein Beitrag zum System der Monokotyledonen. *Wildenowia*, 2, 639-768.

Hammond, S. and Lienert, G.A. (1992) Point Symmetry Adjustment of phi coefficients in the factor analysis of psychometric test items. *Personality and Individual Differences*, 13, 2, 211-219.

Holley, J.W. and Guilford, J.P. (1964) A note on the G-index of agreement. *Educational and Psychological Measurement*, 24, 749-753.

8 Bennett B-Index

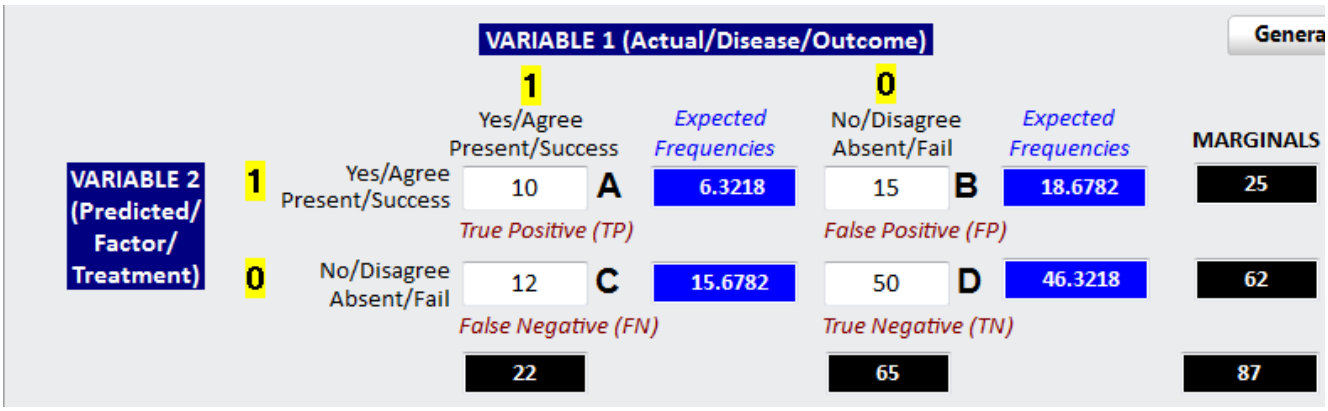
The Bennett index is a marginally adjusted phi coefficient. The effect of this transformation of the marginal totals of the 2x2 table is to correct the phi coefficient for the effects of a skewed response profile on each variable. The G-index makes a point symmetry adjustment that assumes the response splits are 50/50 (median split). The Bennett index does not make this assumption - rather it assumes instead that the two variables/items simply have equal skew in the population- whatever value this might be. Thus, the marginals are equated, holding cell counts A and D as fixed, and averaging cells B and C only. The Bennett coefficient is probably the one to be preferred to the G-Index; the median split assumption does appear slightly more restrictive. However, very little is known about the effects/disparity of using either coefficient.

This index varies between **-1.0** (maximum negative agreement) through **0.0** (no agreement) to +**1.0** (maximum positive agreement).

and the Harms and Ihm(1981) adjustment is made (to guard against A or D frequencies = 0) ....
A = A+1
B= B+1
C= C+1,
D=D+1

I've prepared an example showing how the Bennett is calculated, starting from a simple phi coefficient, through point-symmetry, to marginal symmetry.

The data ...



Phi coefficient

$a := 10$ $b := 15$ $c := 12$ $d := 50$ $n := 87$ Cell frequencies

Following the marginal frequency notation in Table 1, p. 227: Lienert, G., & Hammond, S. (1995). Modifed phi correlation coefficients for the multivariate analysis of ordinally scaled variables. *Educational and Psychological Measurement*, 55, 2, 225-236.

$A := a + b = 25$ $B := c + d = 62$ $C := a + c = 22$ $D := b + d = 65$ Marginals

$$Phi := \frac{(a \cdot d - b \cdot c)}{\sqrt{A \cdot B \cdot C \cdot D}} = 0.2149$$

Point-symmetry adjusted phi/G-Index

$a := \frac{(a+d)}{2} = 30$ $d := a = 30$ $b := \frac{(b+c)}{2} = 13.5$ $c := b = 13.5$ Cell frequencies

$A := \left(\frac{(a+d)}{2}\right) + \left(\frac{(b+c)}{2}\right) = 43.5$ $B := \left(\frac{(a+d)}{2}\right) + \left(\frac{(b+c)}{2}\right) = 43.5$ Marginals

$C := \frac{(b+c)}{2} + \left(\frac{(a+d)}{2}\right) = 43.5$ $D := \left(\frac{(a+d)}{2}\right) + \left(\frac{(b+c)}{2}\right) = 43.5$ Marginals

$$PS := \frac{((a \cdot d) - (b \cdot c))}{\sqrt{A \cdot B \cdot C \cdot D}} = 0.3793 \quad or \quad PS := \frac{((a+d) - (b+c))}{n} = 0.3793$$

Marginal symmetry Corrected Phi

Bennett Coefficient (without Harms and Ihm (1981) frequencies adjustment)

$a := 10 \quad b := 15 \quad c := 12 \quad d := 50$ Cell frequencies

$$A := a + \left\lfloor \frac{(b+c)}{2} \right\rfloor = 23.5 \qquad B := \left\lfloor \frac{(b+c)}{2} \right\rfloor + d = 63.5$$ Marginals

$$C := a + \left\lfloor \frac{(b+c)}{2} \right\rfloor = 23.5 \qquad D := \left\lfloor \frac{(b+c)}{2} \right\rfloor + d = 63.5$$ Marginals

$$BN := \frac{\left\lfloor a \cdot d - \left\lfloor \frac{(b+c)}{2} \right\rfloor^2 \right\rfloor}{\left\lfloor a + \left\lfloor \frac{(b+c)}{2} \right\rfloor \right\rfloor \cdot \left\lfloor d + \left\lfloor \frac{(b+c)}{2} \right\rfloor \right\rfloor} = 0.213$$ or $$BN := \frac{\left\lfloor a \cdot d - \left\lfloor \frac{(b+c)}{2} \right\rfloor^2 \right\rfloor}{\sqrt{A \cdot B \cdot C \cdot D}} = 0.213$$

Bennett Coefficient (with Harms and Ihm (1981) frequencies adjustment)

Harms,V. & Ihm. P. (1981). A rapid estimation method for the parameters in a contingency table with structural zeros. EDV in Med. u. Biol. 12, 76-80.

$a1 := 10 + 1 \quad b1 := 15 + 1 \quad c1 := 12 + 1 \quad d1 := 50 + 1$ Adjusted Cell frequencies

$$BN := \frac{\left\lfloor a1 \cdot d1 - \left\lfloor \frac{(b1+c1)}{2} \right\rfloor^2 \right\rfloor}{\left\lfloor a1 + \left\lfloor \frac{(b1+c1)}{2} \right\rfloor \right\rfloor \cdot \left\lfloor d1 + \left\lfloor \frac{(b1+c1)}{2} \right\rfloor \right\rfloor} = 0.21$$

References

Hammond, S. M., & Lienert, G. A. (1992). Point symmetry adjustment of phi-coefficients in the factor analysis of psychometric test items. *Personality and Individual Differences*, 13, 2, 211-219.

Harms,V. & Ihm. P. (1981). A rapid estimation method for the parameters in a contingency table with structural zeros. EDV in Med. u. Biol. 12, 76-80.

Lienert, G., & Hammond, S. (1995). Modifed phi correlation coefficients for the multivariate analysis of ordinally scaled variables. *Educational and Psychological Measurement*, 55, 2, 225-236.

9 Sensitivity

This is the probability that an actual (outcome) observed event is predicted correctly. For example, the probability that a patient who commits a violent act being predicted to do so. It is calculated by dividing the probability of a True Positive by the sum of the probabilities of obtaining a True Positive and False Negative result. A **False negative** is defined as the probability of predicting **no outcome**, when in fact one actually occurred. In terms of a formula, it is calculated as

$$SE = A/(A+C)$$

where A = True Positive
and C = False Negative

The coefficient varies between 0 and +1.0

10 Phi/Phi Max

Cureton's (1959) Phi over Phi-Max is the correction to the standard Phi Coefficient for restriction of range resulting from unequal marginal probabilities for the two variables. That is, phi can only ever reach its maximum value of 1.0 when both variables have an equal probability of occurrence. Two other coefficients also address this problem, although in a different way – see the G-Index or B-Index sections. The Phi/Phi Max coefficient adjusts the Phi coefficient for inequality between these two marginal probabilities using the following formula:

If Phi = or is > 0.0 then
PhiMax = (Pit-Pi*Pt)/(P'-Pi*Pt)

If Phi < 0.0 and Pi < or = Qt then
PhiMax = (Pit-Pi*Pt)/(Pi*Pt)

If Phi < 0.0 and Pi > Qt then
PhiMax = (Pit-Pi*Pt)/(Pi*Pt – (Pi-Qt))

Where:
Pit = probability of joint occurrence of both variables = (cell A)
Pi = marginal probability for row 1 of the table = (A+B)/N
Pt = marginal probability for column 1 of the table = (A+C)/N
Qt = marginal probability for column 2 of the table = (B+D)/N
P' = is the smaller of Pi and Pt.

Phi/Phi Max varies between -1 and + 1.0 as does phi. I do not recommend use of this coefficient as the size discrepancy between the observed vs corrected coefficient can sometimes be twice the size or more of the original coefficient. It is better to use the Bennett B-Index under conditions of marginal probability inequality, rather than rely upon the Phi/Phi Max correction. Unlike the Phi/Phi Max correction, the Bennett coefficient does not try to compute what the phi coefficient would be IF we had equal marginal probabilities, but rather, computes a coefficient that simply assumes that the skewed marginals are equal -

so preserving the joint probability of success and failure, and averaging the number of false positives and negatives (or their probabilities). I hope you are getting the message that with all these coefficients, their usefulness (or bias) or otherwise is a function of what assumptions you wish to make about the nature of agreement between your two variables!

References

Cureton, E.E. (1959) Note on phi/phi max. Psychometrika, 24, 89-91.

Glass, G.V. and Hopkins, K.D. (1996) Statistical Methods in Education and Psychology 3rd. Edition. Allyn and Bacon. ISBN: 0-205-14212-5

Shannon, G.A. and Cliver, B.A. (1987) An application of item response theory in the comparison of four conventional item discrimination indices for criterion referenced tests. Journal of Educational Measurement, 24, 4, 347-356.

11 Quality of the Sensitivity Coefficient

This coefficient is derived and presented in Kraemer (1992). It is also known as a weighted kappa. It expresses the sensitivity coefficient as a function of the Level of the test (the probability associated with making a prediction of an actual outcome or event). The Level of a test is the sum of the True Positive and False Positive probabilities; essentially, the proportion of cases that are predicted to result in an outcome event. In effect, this coefficient re-expresses sensitivity as a function of the prediction rate (the number of predictions made). For example, the sensitivity of a test may be 99.9%, but if the Level is also 99.9%, the quality coefficient is 0.0. It is calculated as:

$$\mathbf{K(1) = (Sensitivity - Q)/(1-Q)}$$

Where:
$K(1)$ = Quality coefficient
$Q = TP + FP = (A+B)/N$

The coefficient varies between 0 and +1.0

Essentially, what is happening here is that the investigator can adjust the kappa coefficient where differential consideration of false negatives or false positives is required. That is, if our concern is to treat false negatives as of major importance to us (in violence risk prediction, this is where we decide that where we have made a prediction of “no risk”, but an individual goes on to commit a violent offence [the false negative]), then we weight these cases more than the false positives in our calculations. The converse is the case if we decide that making a prediction of violent outcome, but observing no outcome (the false-positive) is more important to us than the false negative. Obviously, if we can attain high values for our weighted kappa, for both sensitivity and specificity (the quality indices), then we can be assured that we have a very good test indeed.

It is useful to look at how we might re-express this formula, showing the explicit weighting function, and so clearly seeing how it weights the sensitivity coefficient. As Kraemer et al (1999) have shown, the conventional kappa coefficient is weighted equally for false-positive and false-negatives.

The formula above may be re-expressed as:

$$K(r) = \frac{(p_A \cdot p_D - p_B \cdot p_C)}{P \cdot Q' \cdot r + P' \cdot Q \cdot r'}$$

where *pA* = probability of occurrence of observations in cell A; likewise for B, C, and D.
P = (A+C)/N (the Base Rate or Prevalence of a test)
Q = (A+B)/N (the Level of a test
r = the weight to be applied (varies between 1 and 0)
P', Q', and r' = (1-P), (1-Q), and (1-r) respectively.

We can further re-express the formulae in terms of cell frequencies as:

$$K(r) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot r\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (1-r)\right\}\right]}$$

which can be re-expressed as :

$$K(r) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{P \cdot Q' \cdot r + P' \cdot Q \cdot r'}$$

In order to show how the divisor is composed in detail, where we use an ***r* weight value of 1.0**, to provide maximum weight for the **false negatives**

$$K(1) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot 1\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (0)\right\}\right]}$$

$$K(1) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right)\right]}$$

As epidemiologists may observe , this particular coefficient is known as attributable risk in their domain. The attributable risk is defined as the proportion of cases in the total population that are attributable to a risk factor.

Reference
Kraemer, H.C.(1992) Evaluating Medical Tests. Sage. ISBN: 0-8039-4612-0

Kraemer, H., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., and Kupfer, D.J. (1999) Measuring the Potency of Risk Factors for Clinical or Policy Significance. Psychological Methods, 4, 3, 257-271

12 Specificity

This is the probability that the actual **non-occurrence** of an event is predicted correctly. For example, the probability of a patient who does **NOT** commit a violent act being predicted correctly. It is calculated by dividing the probability of a True Negative by the sum of the probabilities of obtaining a True Negative and False Positive result. A **False positive** is defined as the probability of predicting **an outcome**, when in fact one did not actually occur. In terms of a formula, it is calculated as

$$SE = D/(B+D)$$

where D = True Negative
and B = False Positive

The coefficient varies between 0 and +1.0

13 Quality of the Specificity Coefficient

This coefficient is derived and presented in Kraemer (1992). It is also known as a weighted kappa. It adjusts the specificity coefficient (the probability of correctly predicting *no outcome* as a function of all no-outcome occurrences.. D/(B+D)) for the *inverse* Level (the probability associated with making a prediction of NO actual outcome or event – the sum of all negative predictions made divided by N) and Level (Q) of the test. The Level of a test is the sum of the True Positive and False Positive probabilities – essentially, the proportion of cases that are predicted to result in an outcome event (the number of all positive predictions made divided by N). The Inverse Level is simply (1-Q). For example, the specificity of a test may be 99%, but if the Level is 0.01%, then the quality coefficient is 0.0. It is calculated as:

$$K(0) = (Specificity - (1-Q))/Q$$

Where:
K(0) = Quality coefficient
Q = TP + FP = (A+B)/N

The coefficient varies between 0 and +1.0

Essentially, what is happening here is that the investigator can adjust the kappa coefficient where differential consideration of false negatives or false positives is required. That is, if our concern is to treat false negatives as of major importance to us (in violence risk prediction, this is where we decide that where we have made a prediction of "no risk", but an individual goes on to commit a violent offence [the false negative]), then we weight these cases more than the false positives in our calculations. The converse

is the case if we decide that making a prediction of violent outcome, but observing no outcome (the false-positive) is more important to us than the false negative. Obviously, if we can attain high values for our weighted kappa, for both sensitivity and specificity (the quality indices), then we can be assured that we have a very good test indeed.

It is useful to look at how we might re-express this formula, showing the explicit weighting function, and so clearly seeing how it weights the sensitivity coefficient. As Kraemer et al (1999) have shown, the conventional kappa coefficient is weighted equally for false-positive and false-negatives.

The formula above may be re-expressed as:

$$K(r) = \frac{(p_A \cdot p_D - p_B \cdot p_C)}{P \cdot Q' \cdot r + P' \cdot Q \cdot r'}$$

where pA = probability of occurrence of observations in cell A; likewise for B, C, and D.
P = (A+C)/N (the Base Rate or Prevalence of a test)
Q = (A+B)/N (the Level of a test
r = the weight to be applied (varies between 1 and 0)
P', Q', and r' = (1-P), (1-Q), and (1-r) respectively.

We can further re-express the formulae in terms of cell frequencies as:

$$K(r) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot r\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (1-r)\right\}\right]}$$

which can be re-expressed as :

$$K(r) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{P \cdot Q' \cdot r + P' \cdot Q \cdot r'}$$

In order to show how the divisor is composed in detail, where we use an ***r* weight value of 0.0**, to provide maximum weight for the **false positives**

$$K(0) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot 0\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (1)\right\}\right]}$$

$$K(0) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right)\right\}}$$

Reference

Kraemer, H.C.(1992) *Evaluating Medical Tests*. Sage. ISBN: 0-8039-4612-0

14 PPP (Positive Power of Prediction)

The probability that a prediction correctly predicted the occurrence of an outcome event. E.g. the probability that a patient predicted to be violent was actually violent. It is calculated by expressing the number of True Positives (successful predictions) as a function of the total number of positive predictions (positive as in Yes, Event Occurring etc.). The initials **PPP** = **P**ositive **P**redictive **P**ower. The initials **ppv** = **P**ositive **P**redictive **V**alue. Kraemer uses **PVP** = the **P**redictive **V**alue of a **P**ositive test.

PPP = A/(A+B)

Where
A = True Positive
B = False Positive

The coefficient varies between 0 and +1.0

References

Kraemer, H.C.(1992) *Evaluating Medical Tests*. Sage. ISBN: 0-8039-4612-0

Kraemer, H., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., and Kupfer, D.J. (1999) Measuring the Potency of Risk Factors for Clinical or Policy Significance. *Psychological Methods*, 4, 3, 257-271

15 NPP (Negative Power of Prediction)

The probability that a prediction correctly predicted the **non-occurrence** of an outcome event. E.g. the

probability that a patient predicted not to be violent was actually not violent. It is calculated by expressing the number of True Negatives (successful predictions of non-occurrence) as a function of the total number of negative predictions (negative as in No, No Event Occurring etc.). The initials **NPP** = **N**egative **P**redictive **P**ower. The initials **npv** = **N**egative **P**redictive **V**alue. Kraemer uses **PVN** = the **P**redictive **V**alue of a **P**ositive test.

$$\mathbf{NPP = D/(C+D)}$$

Where
C = False Negative
D = True Negative

The coefficient varies between 0 and +1.0

References

Kraemer, H.C.(1992) *Evaluating Medical Tests*. Sage. ISBN: 0-8039-4612-0

Kraemer, H., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., and Kupfer, D.J. (1999) Measuring the Potency of Risk Factors for Clinical or Policy Significance. *Psychological Methods*, 4, 3, 257-271

16 PE (Predictive Efficiency)

Otherwise known as **Classification Accuracy**, Efficiency, or **Overall Prediction Accuracy**; this is the overall probability of prediction success, summing both True Positives AND True Negatives. It is calculated by summing the probabilities in cells A and D (that for the true positive (A) and that for the true negative (D)). Of course, this coefficient tells you nothing about the errors made, but only how far your test was able to correctly predict both positive and negative outcomes.

The formula in terms of cell frequencies =

$$\mathbf{PE = (A+D)/N}$$

17 IOC (Improvement Over Chance)

The Improvement Over Chance (IOC) coefficient literally computes the improvement in prediction using a test, over chance prediction.

The coefficient can be expressed as a % (and is done so in the program) , and can be negative or positive (as you can make predictions which are worse than expected by chance alone). Its maximum value is 100%.

The formula is:

$$chance := \left\{\frac{(A+C)}{N} \cdot \frac{(A+B)}{N}\right\} + \left\{\frac{(B+D)}{N} \cdot \frac{(C+D)}{N}\right\}$$

$$IOC := \left\{\frac{(A+D)}{N}\right\} - chance$$

Alternatively, it may be expressed as:

$$PE := \frac{(A+D)}{N} \quad \text{predictive efficiency}$$

$$BR := \frac{(A+C)}{N} \quad \text{Base Rate}$$

$$SR := \frac{(A+B)}{N} \quad \text{Selection Ratio}$$

$$IOC := PE - ((BR \cdot SR) + (1-BR) \cdot (1-SR))$$

Two worked examples are given in the RIOC help topic.

18 RIOC (Relative Improvement Over Chance)

Loeber and Dishion's statistic indexes the improvement of prediction over chance, relative to the Base Rate, using your test. This is an extremely valuable statistic that gives a clear indication of just how good your test really is in terms of predictive accuracy. Whereas the IOC indexes the basic improvement over chance, the IOC index is sensitive to both Base Rate and Selection Ratio. By expressing the IOC relative to the maximum possible accuracy (the Base Rate) given the lowest possible accuracy (chance), the RIOC provides a universal measure of effect which is much less dependent upon sample characteristics than the IOC.

The RIOC expresses the improvement over chance (IOC) as a function of the difference between the random correct (RC) and maximum correct (MC) values in a given study. Thus, the percent improvement over chance in a given study always falls between the random correct value and the maximum correct value.

From Loeber and Dishion, pp. 72-73..
"A major problem in the evaluation of predictive efficiency is that it depends to a great extent on how well the selection ratio matches the base rate in a given study. A discrepancy between the selection ratio and the base rate, which is very common in delinquency studies, influences the magnitude of the maximum correct value. The present use of the RIOC measure, which is partly based on the maximum correct value,

largely eliminates this problem. To test this, the IOC and RIOC indices were correlated with both the selection ratio and the base rate for each of the studies reviewed. It was thought that the best index of predictive efficiency would be the least correlated with either of the selection parameters. As expected, the IOC correlated .54 and .38 and the RIOC correlated. 13 and .22 with the base rates and selection ratios. We concluded that the RIOC measure is more independent of varying base rates and selection ratios and therefore superior as an overall evaluative index of predictive efficiency."

The coefficient can be expressed as a % (and is done so in the program) , and can be negative or positive (as you can make predictions which are worse than expected by chance alone). Its maximum value is 100%.

The formula is:

$$chance := \left\{\frac{(A+C)}{N} \cdot \frac{(A+B)}{N}\right\} + \left\{\frac{(B+D)}{N} \cdot \frac{(C+D)}{N}\right\}$$

$$PE := \frac{(A+D)}{N} \quad \text{predictive efficiency}$$

$$BR := \frac{(A+C)}{N} \quad \text{Base Rate}$$

$$SR := \frac{(A+B)}{N} \quad \text{Selection Ratio}$$

$$IOC := PE - ((BR \cdot SR) + (1-BR) \cdot (1-SR)) \quad \text{or} \quad IOC := \left\{\frac{(A+D)}{N}\right\} - chance$$

$$MC := 1 - |SR - BR| \quad \text{or} \quad MC := \frac{((BR \cdot N) + (B+D) - |(A+B-(BR \cdot N))|)}{N}$$

$$RIOC := \frac{IOC}{MC - chance}$$

From the example in Loeber and Dishion (1983) - Figure 1 (p. 70):

$A := 46 \qquad B := 102$

$C := 30 \qquad D := 118 \qquad N := A + B + C + D = 296$

$$chance := \left(\frac{(A+C)}{N} \cdot \frac{(A+B)}{N}\right) + \left(\frac{(B+D)}{N} \cdot \frac{(C+D)}{N}\right) = 0.5$$

$$PE := \frac{(A+D)}{N} = 0.5541 \quad \text{predictive efficiency}$$

$$BR := \frac{(A+C)}{N} = 0.2568 \quad \text{Base Rate}$$

$$SR := \frac{(A+B)}{N} = 0.5 \quad \text{Selection Ratio}$$

$$IOC := PE - ((BR \cdot SR) + (1-BR) \cdot (1-SR)) = 0.0541 \quad \text{or} \quad IOC := \left(\frac{(A+D)}{N}\right) - chance = 0.0541$$

$$MC := 1 - |SR - BR| = 0.7568 \quad \text{or} \quad MC := \frac{((BR \cdot N) + (B+D) - |(A+B-(BR \cdot N))|)}{N} = 0.7568$$

$$RIOC := \frac{IOC}{MC - chance} = 0.2105$$

and where we have perfect prediction ..

$A := 100 \qquad B := 0$

$C := 0 \qquad D := 100 \qquad N := A + B + C + D = 200$

$$chance := \left(\frac{(A+C)}{N} \cdot \frac{(A+B)}{N}\right) + \left(\frac{(B+D)}{N} \cdot \frac{(C+D)}{N}\right) = 0.5$$

$$PE := \frac{(A+D)}{N} = 1 \qquad \text{predictive efficiency}$$

$$BR := \frac{(A+C)}{N} = 0.5 \qquad \text{Base Rate}$$

$$SR := \frac{(A+B)}{N} = 0.5 \qquad \text{Selection Ratio}$$

$$IOC := PE - ((BR \cdot SR) + (1 - BR) \cdot (1 - SR)) = 0.5 \qquad \text{or} \qquad IOC := \left(\frac{(A+D)}{N}\right) - chance = 0.5$$

$$MC := 1 - |SR - BR| = 1 \qquad \text{or} \qquad MC := \frac{((BR \cdot N) + (B+D) - |(A+B-(BR \cdot N))|)}{N} = 1$$

$$RIOC := \frac{IOC}{MC - chance} = 1$$

or where our test is actually working less well than chance (usually where the base-rate for success is very high):

$$A := 815 \quad B := 109$$

$$C := 209 \quad D := 12 \quad N := A + B + C + D = 1145$$

$$chance := \left\{\frac{(A+C)}{N} \cdot \frac{(A+B)}{N}\right\} + \left\{\frac{(B+D)}{N} \cdot \frac{(C+D)}{N}\right\} = 0.7421$$

$$PE := \frac{(A+D)}{N} = 0.7223 \quad \text{predictive efficiency}$$

$$BR := \frac{(A+C)}{N} = 0.8943 \quad \text{Base Rate}$$

$$SR := \frac{(A+B)}{N} = 0.807 \quad \text{Selection Ratio}$$

$$IOC := PE - ((BR \cdot SR) + (1-BR) \cdot (1-SR)) = -0.0198 \quad \text{or} \quad IOC := \left\{\frac{(A+D)}{N}\right\} - chance = -0.0198$$

$$MC := 1 - |SR - BR| = 0.9127 \quad \text{or} \quad MC := \frac{((BR \cdot N) + (B+D) - |(A+B-(BR \cdot N))|)}{N} = 0.9127$$

$$RIOC := \frac{IOC}{MC - chance} = -0.1163$$

This example shows just how important is the IOC and especially the RIOC .. because if we just reported the Predictive Efficiency (**PE** = Overall classification accuracy) in isolation of the other important statistics, 72% classification accuracy looks 'acceptable'. But, when you see the RIOC is **-11.63%** *(your test is actually performing less well than if we made predictions by tossing a coin)*, it adds a certain cold reality to any claim of 'predictive accuracy'!

References

Loeber, R. and Dishion, T. (1983) Early Predictors of male delinquency: a review. *Psychological Bulletin*, 94, 68-99

Mossman, D. (1994) Assessing Predictions of Violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, 62, 4, 783-792.

19 BR (Base Rate)

Otherwise known as **Prevalence**, this statistic indexes the probability of an occurrence, usually expressed as a %. Base rates are defined for specific populations of interest and are restricted to them. A base rate is equivalent to a proportion. It is calculated by summing the number of occurrences and dividing these by

the total number of observations made. E.g. Assume we are trying to find out the base rate of any violent offence being committed by a patient within a hospital ward. We know there are 100 patients on the ward during the period of interest; we note that 30 of the patients committed at least one violent offence over the period of interest. This gives us a base rate of 30/100 = 0.30 or 30%. In a 2x2 classification table, it is calculated as the sum of the True Positives (A) and False Negatives (C), divided by the total number of observations.

The explicit formula is: P = (A+C)/N

20 Level or Selection Ratio (of a test)

The level (Q) of a test (alternatively known as the **Selection Ratio**) is defined as the sum of the probabilities of True Positives (A) and False Positives (B) for a test. Essentially, the sum of the predicted occurrences of an event (to be contrasted with **the Inverse Level** – which is the sum of the predicted probabilities of non-occurrences of an event (1-Q), i.e. probabilities of the prediction of no violent outcome, but outcome occurs (false negative, C) + prediction of no outcome and no violent offense occurs (true negative, D)). This coefficient is basically an indicator of the probability (or proportion) of **positive** predictions made (or the proportion of cases with a positive factor) out of the total numbers of predictions (both positive and negative).

In terms of cell frequencies …

Q = (A+B)/N

The Inverse Level is:

Q` = (1-Q) or (C+D)/N

21 Cohen's Kappa

This coefficient is implemented here in its dichotomous form – however, it can be used for more than two rating categories. Kappa was designed specifically as a measure of agreement between 2 judges, where ratings are categorical, and where a correction for *chance agreement* is made. This coefficient thus differs from the percent agreement approach adopted by some, because this simple calculation does not take into account what the chance-level agreement between judges would be alone, assuming they both guessed randomly. The formula for kappa is:

$$\kappa = \frac{\sum f_o - \sum f_e}{N - \sum f_e}$$

where $\sum f_o$ = observed frequencies in the diagonal

$\sum f_e$ = expected frequencies in the diagonal

N = Number of Patients

The expected frequencies are the same as those calculated for the Pearson Chi-Square calculation, except we use just the diagonal values (A and D) for both observed and expected frequencies.

In contrast to this formula, we might consider use of the Jaccard coefficient , which is another measure of interrater agreement, but one that excludes joint-negatives from its calculation.

A useful point is that both kappa and the Jaccard coefficients can be interpreted as % values. Kappa can be interpreted as the % agreement after correcting for chance. The Jaccard coefficient can be interpreted as the % agreement after excluding joint negative pairs. Both coefficients vary between **0** and **1** (or **0** to **100%**)

On a mathematical note, we can use Kraemer's simplified formula to show that Cohen's kappa is equivalent to a generalised weighted kappa formula, where false positives and false negatives are given equal weight. See also the Quality Coefficients for Sensitivity and Specificity; these use the same formula as below, except that for the Quality of Sensitivity, r = 1 (our concern is for false negatives), and for quality of Specificity, we set r = 0 (our concern is now entirely for false positives)

Reference

Cohen, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 10, 37-46.

22 False Positive Rate (False Alarms)

The proportion of cases where a prediction for a positive outcome is made, but no outcome is observed. Alternatively, it can be expressed as the proportion of cases subjected to a factor, but where no outcome effect is observed.

It is calculated as:

Fpr=B/(B+D)

23 False Negative Rate

The proportion of cases where a prediction for a negative outcome is made, but an outcome is observed. Alternatively, it can be expressed as the proportion of cases not subjected to a factor, but where an outcome effect is observed.

It is calculated as:

Fpr=C/(A+C)

24 Odds of Outcome Given Treatment (or Predicted)

These are simply the ratio of the probability of a **positive outcome given a positive factor or prediction** to the probability of **a negative outcome**, given a positive factor or prediction.. It is calculated as:

Odds of an occurrence given Positive Factor/Prediction **-=A/B**

25 Odds of Outcome NOT Given Treatment (or NOT Predicted)

These are simply the ratio of the probability of a **positive outcome given a negative factor or prediction** to the probability of **a negative outcome**, given a negative factor or prediction.. It is calculated as:

Odds of an occurrence given Negative Factor/Prediction **-=C/D**

26 Odds Ratio

The ratio of odds under two conditions. The best way of understanding an odds-ratio is in terms of relative risk. Relative Risk is the ratio of the probability of a positive outcome given a positive factor vs the probability of positive outcome given a negative factor. The Odds ratio is virtually equivalent to relative risk under certain conditions, and hence tends to be used as an easily computable index of relative risk.

First, let us look at the formula for relative risk …

$$RR = \frac{\left(\frac{p_A}{(p_A + p_B)}\right)}{\left(\frac{p_C}{(p_C + p_D)}\right)}$$

where

p_A = probability of A cell occurrence.

p_B = probability of B cell occurrence.

p_C = probability of C cell occurrence.

p_C = probability of D cell occurrence.

If the probability of a positive outcome relative to all outcomes for a positive factor/prediction (*pA* relative

to *pB*) is small, and likewise for the probability of a positive outcome relative to all outcomes for a negative factor/prediction (*pC* relative to *pD*), then the relative risk equation is approximated by:

$$RR = \frac{\left(\frac{p_A}{p_B}\right)}{\left(\frac{p_C}{p_D}\right)} \text{ which can be written as : } \frac{(A \cdot D)}{(B \cdot C)}$$

where

p_A = probability of A cell occurrence.

p_B = probability of B cell occurrence.

p_C = probability of C cell occurrence.

p_C = probability of D cell occurrence.

A,B,C,D are cell frequencies

Note here: the formula should look familiar in terms of odds. If you look at the formula for the Odds of a Positive Outcome given treatment and the Odds of a Positive Outcome Given no Treatment, then you'll see that the formula above is simply the ratio of these two odds.

For example:
from a real study quoted in Samuels and Witmer (1999), p.436, Table 10.27, we have the cell layout as:

		Low Birthweight	Normal
Birthweight			
Risk Factor SMOKING	-Smoker	237 (A)	3489 (B)
	Non-Smoker	197 (C)	5870 (D)

The relative risk is **1.9589.** That is, the relative risk or conditional probability of having a low birthweight baby is about twice as great for smokers as for non-smokers. The odds ratio is **2.024**. Very close. If we make the probabilities associated with A and C cells even smaller, you'll see how close the estimates become …

		Low Birthweight	Normal
Birthweight			
Risk Factor SMOKING	-Smoker	237 (A)	13489 (B)
	Non-Smoker	197 (C)	15870 (D)

The relative risk is **1.408**. That is, the relative risk or conditional probability of having a low birthweight baby is about twice as great for smokers as for non-smokers. The odds ratio is **1.415**. Even closer. Try adding 1000 to cells A and C, with the original values in B and D – and see the disparity increase.

Finally, it is very easy to make sweeping statements about odds-ratios. Douglas, Cox, and Webster (1999) are highly misleading in their statement on p. 160, lines17-18 "A correlation of 0.50 translates roughly into an odds ratio of 9.0". This is not necessarily the case at all. For example, the dataset below …

	Actual Violence	No Violence
Predicted Risk – Positive Prediction	25 (A)	120 (B)

```
                Negative Prediction          140 (C)      7000 (D)
```

The Relative Risk is: **8.7931**. The Odds Ratio is **10.4167**. Estimated **r correlation is 0.1610** and phi is **0.1434**. Obviously, this has a lot to do with the asymmetry of the marginal frequencies in the table, and the manner of computing a measure of agreement.

References

Armitage, P. and Berry, G. (1994) *Statistical Methods in Medical Research 3rd. Edit*. Blackwell Science. ISBN: 0-632-03695-8

Douglas, K.S., Cox, D.N., and Webster, C.D. (1999) Violence Risk Assessment: Science and Practice. *Legal and Criminological Psychology*, 4, 149-184.

Samuels, M.L. and Witmer, J.A. (1999) *Statistics for the Life Sciences, 2nd Edition*. Prentice Hall. ISBN: 0-13-649211-8

27 Relative Risk

Relative Risk is the ratio of the probability of a positive outcome given a positive factor vs the probability of positive outcome given a negative factor

The formula is:

$$RR = \frac{\left(\frac{p_A}{(p_A + p_B)}\right)}{\left(\frac{p_C}{(p_C + p_D)}\right)}$$

where

p_A = probability of A cell occurrence.

p_B = probability of B cell occurrence.

p_C = probability of C cell occurrence.

p_C = probability of D cell occurrence.

If the probability of a positive outcome relative to all outcomes for a positive factor/prediction (*pA* relative to *pB*) is small, and likewise for the probability of a positive outcome relative to all outcomes for a negative factor/prediction (*pC* relative to *pD*), then the relative risk equation is approximated by the odds ratio:

$$RR = \frac{\left(\frac{p_A}{p_B}\right)}{\left(\frac{p_C}{p_D}\right)} \text{ which can be written as : } \frac{(A \cdot D)}{(B \cdot C)}$$

where

p_A = probability of A cell occurrence.

p_B = probability of B cell occurrence.

p_C = probability of C cell occurrence.

p_C = probability of D cell occurrence.

A,B,C,D are cell frequencies

Note here: the formula should look familiar in terms of odds. If you look at the formula for the Odds of a Positive Outcome given treatment and the Odds of a Positive Outcome Given no Treatment, then you'll see that the formula above is simply the ratio of these two odds.

For example:
from a real study quoted in Samuels and Witmer (1999), p.436, Table 10.27, we have the cell layout as:

		Low Birthweight	Normal Birthweight
Risk Factor SMOKING	**Smoker**	237 (A)	3489 (B)
	Non-Smoker	197 (C)	5870 (D)

The relative risk is **1.9589.** That is, the relative risk or conditional probability of having a low birthweight baby is about twice as great for smokers as for non-smokers. The odds ratio is **2.024**. Very close.

We can take another example, this time from the VRAG results of Webster et al (1994)…

		Actual Violence	No Violence
Predicted Risk –	**Positive Prediction**	115 (A)	94 (B)
	Negative Prediction	76 (C)	333 (D)

Relative Risk is: **2.9612**. The Odds Ratio is **5.3604**. The interpretation of the relative risk is that an individual who commits a violent offence is about 3 times more likely to have been predicted as such in comparison with an individual who committed a recidivist violent offence who was not predicted to do so. The danger of not understanding the relationship between relative risk and the odds ratio (and the approximation of relative risk by the odds ratio) is that we would over-estimate the relativity by using the odds ratio (saying we are 5 times more likely to correctly predict the positive outcome). The odds ratio interpretation is after all, **a ratio between odds of occurrence, not between probabilities of occurrence**. If I change cell B to 55 instead of 94, **the relative risk is 3.6404, whilst the odds ratio is 9.1615**.

Reference

Webster, C.D., Harris, G.T., Rice, M.E., Cormier, C., Quinsey, V.L. (1994) *The Violence Prediction Scheme: Assessing Dangerousness in High Risk Men*. University of Toronto, Centre of Criminology. ISBN: 0-919584-74-8

28 Cohen's d`

The is the index, attributed to Cohen, that is best known as providing a measure of effect size within statistical power analysis. Within decision table analysis, or more generally, discrimination judgement analysis, Cohen's d is referred to as d` (d-prime). However, although Cohen's d and d` sound as though they are different indexes, they are not. The same index is called two different names, as it had been developed in two different domains.

Cohen's d – from now on I'm going to refer to it as d` - reflects the standardized distance between two distributions. Generally, it is the distance between the mean of a null and alternative hypothesis distribution. Alternatively, in ROC analysis (Receiver Operating Characteristic analysis) it is the distance between the means of the **signal+noise** distribution and **no-signal+noise** distribution. In a prediction environment, it is the distance between the "judgement distributions" of actual outcomes and no outcomes.

More formally:

$$d' = \frac{\mu_1 - \mu_0}{\sigma}$$

where μ_1 = the mean of the alternative hypothesis distribution
or the signal + noise distribution
μ_0 = the mean of the null hypothesis distribution
or the no-signal + noise distribution
σ = the common standard deviation of both distributions

In most applications, and certainly in all forensic applications I have seen, binormal distributions are assumed. That is, it is assumed both sampling distributions are normal, with sigma (population standard deviation of 1.0). The calculations here assume normal distributions.

In order to compute d` here, I use the approach outlined in Swets (1996), pp.21-22. I first compute

FPR=B/(B+D) …which is the probability of a false positive, or false positive rate.

Given this probability value, which indexes (as area under a normal curve) the position of the decision criterion being used **relative to the Null Hypothesis distribution mean**, we can convert this area/probability into standard deviation units by computing the inverse normal value (the z-score, deviation-value) for the area. Here our area is computed as that between the mean of the null hypothesis distribution and the SD z-value – which extends toward the right of the null distribution mean (into the alternative hypothesis distribution area).

Then, we compute the Sensitivity coefficient as:

Sensitivity = A/(A+C)

Given this probability value, which indexes (as area under a normal curve) the position of the decision

criterion being used **relative to the Alternative Hypothesis distribution mean**, we can convert this area/ probability into standard deviation units by computing the inverse normal value (the z-score, deviation-value) for the area. Here our area is computed as that between the mean of the alternative hypothesis distribution and the SD z-value – which extends toward the left of the alternative distribution mean (into the null hypothesis distribution area).

It's time for a graph, and an example – so that you can see what is going on here!

Using the example from Swets (1994), we have the following decision table…

		stimulus	**no-stimulus**
Response Type -	**Positive Response**	80 (A)	10 (B)
	Negative Response	20 (C)	90 (D)

The problem to be solved is what is the value of the internal subjective criterion an individual is using that causes the response pattern as observed to the stimuli. **The False Positive Rate is 0.10.** This value tells us that 10% of all responses identified as False Positives (indicate positive response when no stimulus exists) lie beyond our unknown cutoff value. Of course, knowing that 10% of responses lie beyond this point, given the known distribution function (normal), we immediately know where the cutoff point is in relation to the Null distribution (no stimulus) discriminations. This is found using the inverse normal distribution – which gives us the z-value (critical value xc) that excludes the upper 10% of the area of this Null distribution. **The sensitivity is 0.80**. This value tells us that 20% (1-0.80) of the Alternative Hypothesis distribution is excluded by the critical value (xc). Knowing this now enables us to position the Alternative distribution (stimulus distribution) precisely in relation to the cutoff point, which also enables us to compute d`. How? Well, if we know that 20% of the total area in the Alternative distribution lies to the left of its mean, then we can compute the Inverse function as before – which gives us the corresponding z-value (standard deviation units). **Given both z-values are in the same metric, it is now a simple matter of addition to compute d` as the sum of the two z-values of the FPR and Sensitivity**.

Note, when we start with our 2x2 table, we do not know where the decision criterion is, or what distance might lie between the distributions. By following the procedures above, you can see how we find out these values, step-by-step.

The graph (in a separate .jpg image file that comes with the software) shows how these parameters are defined – and the various regions of interest. The metric of the graph is exactly as per example… you can count the SD units on the bottom axis – each tick = 0.10 SD units. You can also see graphically the relation of the parameter values to measurements on the graph.

References

Swets, J.A. (1994) *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics*. Lawrence Erlbaum. ISBN:0-8058-1834-0

Swets, J.A. (1998) Separating Discrimination and Decision in Detection, Recognition, and Matters of Life and Death. In Scarborough, D. and Sternberg, S.(eds). *An Invitation to Cognitive Science: Methods, Models, and Conceptual Issues*. 2nd Edition.Vol. 4.MIT Press. ISBN:0-262-65046-0

29 Estimated r from d`

From Cohen (1988), he gives the equation for estimating the product-moment (pearson) correlation from a value of d`. This is useful where we do not have the correlation between the raw data for outcomes and factors/predictions. Although we can obtain 2x2 table measures of agreement, these are estimates based upon the table of data at hand, using a specific criterion value to define success and failure, and thus false positives and false negatives. However, given a d` value, we have a measure (assuming both sampling distributions are normal) that can be related directly to the expected correlation in the total sample between our judgements and outcomes, irrespective of any decision criterion we might later use.

There are two formulas normally used:
Equation 1

$$r = \sqrt{\frac{d^2}{(d^2+4)}}$$

Equation 2

$$r = \sqrt{\frac{d^2}{\left(d^2+\frac{1}{PQ}\right)}} = \frac{d}{\sqrt{\left(d^2+\frac{1}{PQ}\right)}}$$

where P = the proportion of individuals (or stimuli, or positive outcomes) in the total population.
Q = the proportion (1-P)

For example, in the case of violence prediction, it is likely that the two proportions of cases (offenders vs non-offenders will not be a-priori equal in the population. Therefore, it would be incorrect to use Equation 1.**I have used Equation 2 in this program – as most work in the forensic area assumes unequal proportions of potential outcomes (disease fatalities, injuries, judgments of risk etc.)**. I use the observed marginal frequencies (A+C and B+D) as my estimates of population proportion. To see the effects of being this conservative, let us take a d` of 0.9, with P = 0.3, and Q = 0.7….

Equation 1 = 0.4103, Equation 2 = 0.3813 As P and Q approach 0.5, so does the inverse of their multiplication approach 4.0. The greater the inequality, the greater the constant value above 4.

References
Cooper, H., Hedges, L.V.(Eds.). (1997) *Handbook of Research Synthesis*. Russell Sage Foundation. ISBN: 0-87154-226-9 (especially chapter 16 by Robert Rosenthal)

30 Weighted Kappa

The kappa coefficient can be adjusted for differential consideration of false negatives or false positives.It is already taking into account the base rate or prevalence. That is, if our concern is to treat false negatives as of major importance to us (in violence risk prediction, this is where we decide that where we have made a

prediction of "no risk", but an individual goes on to commit a violent offence [the false negative]), then we weight these cases more than the false positives in our calculations. The converse is the case if we decide that making a prediction of violent outcome, but observing no outcome (the false-positive) is more important to us than the false negative. Obviously, if we can attain high values for our weighted kappa, for both sensitivity and specificity (the quality indices), then we can be assured that we have a very good test indeed. We might equally be dealing with rater reliabilities, and thus be placing more importance on false-positive vs false negative errors. However, in general for these rater reliability issues, we mostly use an equal-weighted kappa. **However, it is a matter of choice, not of necessity.**

It is useful to look at how we might re-express this formula, showing the explicit weighting function, and so clearly seeing how it weights the sensitivity coefficient. As Kraemer et al (1999) have shown, the conventional kappa coefficient is weighted equally for false-positive and false-negatives.

The formula above may be re-expressed as:

$$K(r) = \frac{(p_A \cdot p_D - p_B \cdot p_C)}{P \cdot Q' \cdot r + P' \cdot Q \cdot r'}$$

where pA = probability of occurrence of observations in cell A; likewise for B, C, and D.
P = (A+C)/N (the Base Rate or Prevalence of a test)
Q = (A+B)/N (the Level of a test
***r* = the weight to be applied (varies between 1 and 0)**
P', Q', and r' = (1-P), (1-Q), and (1-r) respectively.

We can further re-express the formulae in terms of cell frequencies as:

$$K(r) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot r\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (1-r)\right\}\right]}$$

which can be re-expressed as :

$$K(r) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{P \cdot Q' \cdot r + P' \cdot Q \cdot r'}$$

In order to show how the divisor is composed in detail, where we use an ***r* weight value of 1.0**, to provide maximum weight for the **false negatives**

$$K(1) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot 1\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (0)\right\}\right]}$$

$$K(1) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right)\right]}$$

Obviously, for an equally weighted kappa, the formula is:

$$K(1) = \frac{\left(\frac{A \cdot D}{N^2}\right) - \left(\frac{B \cdot C}{N^2}\right)}{\left[\left\{\left(\frac{(A+C)}{N}\right) \cdot \left(1 - \left(\frac{(A+B)}{N}\right)\right) \cdot 0.5\right\} + \left\{\left(1 - \left(\frac{(A+C)}{N}\right)\right) \cdot \left(\frac{(A+B)}{N}\right) \cdot (0.5)\right\}\right]}$$

The phi coefficient is the geometric mean of Kraemer's 1, and 0 weighted kappas via:

$$\phi = \sqrt{k(0) \cdot k(1)}$$

Reference

Kraemer, H.C.(1992) *Evaluating Medical Tests*. Sage. ISBN: 0-8039-4612-0

Kraemer, H., Kazdin, A.E., Offord, D.R., Kessler, R.C., Jensen, P.S., and Kupfer, D.J. (1999) Measuring the Potency of Risk Factors for Clinical or Policy Significance. *Psychological Methods*, 4, 3, 257-271

31 Attributable Risk

Attributable risk is defined as the proportion of cases in the total population that are attributable to a risk factor. Generally it is used in circumstances in which it is considered justifiable to infer causation from an observed association. Then, it may be used as a measure of importance of eliminating a factor as part of a prevention strategy. The formula is as that for the Quality of Sensitivity coefficient.

We can re-express the formula as the ratio of the difference between the Base Rate (probability of a positive outcome in the population) and the probability of being outcome positive although not being subjected to the risk factor, divided by the base rate..:

AR = (P-[C/(C+D)])/P …….where P = the Base Rate

In essence, we are expressing the excess incidence attributable to the proposed risk factor. Alternatively, it can be expressed as:

AR = (A/(A+C))*((rr-1)/rr) ……where rr = relative risk.

In risk prediction work, attributable risk cannot really be used – as predictions do not cause outcomes! Although we can see the meaning of the quality coefficient in this case (for the specificity), the prediction of "no-outcome" is not a causal variable.

If however, we were looking at say smoking and low birthweight, then we might conceive of a smoking as a causal variable for low birthweight. Taking the example from a real study quoted in Samuels and Witmer (1999), p.436, Table 10.27, we have the cell layout as:

		Low Birthweight	**Normal Birthweight**
Risk Factor SMOKING	**Smoker**	237 (A)	3489 (B)
	Non-Smoker	197 (C)	5870 (D)

The attributable risk is 0.2673 (26.73% of all low-birthweight cases (indexed by the base rate) within the total population of all cases (N) may be attributable to smoking). Put another way, 26.73% of the 4.43% of low-birthweight babies can be attributed to the effects of smoking.The relative risk is 1.9589. That is, the relative risk or conditional probability of having a low birthweight baby is about twice as great for smokers as for non-smokers.

References

Armitage, P. and Berry, G. (1994) *Statistical Methods in Medical Research 3rd. Edit*. Blackwell Science. ISBN: 0-632-03695-8

Samuels, M.L. and Witmer, J.A. (1999) *Statistics for the Life Sciences, 2nd Edition*. Prentice Hall. ISBN: 0-13-649211-8

Index

- Y -

Yule Q 6