Clinical and Experimental Pharmacology and Physiology (2002) **29**, 527–536

SPECIAL ARTICLE

STATISTICAL TECHNIQUES FOR COMPARING MEASURERS AND METHODS OF MEASUREMENT: A CRITICAL REVIEW

John Ludbrook

The University of Melbourne, Parkville, Victoria, Australia

SUMMARY

1. Clinical and experimental pharmacologists and physiologists often wish to compare two methods of measurement, or two measurers.

2. Biostatisticians insist that what should be sought is not agreement between methods or measurers, but disagreement or bias.

3. If measurements have been made on a continuous scale, the main choice is between the Altman–Bland method of differences and least products regression analysis. It is argued that although the former is relatively simple to execute, it does not distinguish adequately between fixed and proportional bias. Least products regression analysis, although more difficult to execute, does achieve this goal. There is almost universal agreement among biostatisticians that the Pearson product–moment correlation coefficient (*r*) is valueless as a test for bias.

4. If measurements have been made on a categorical scale, unordered or ordered, the most popular method of analysis is to use the kappa statistic. If the categories are unordered, the unweighted kappa statistic (K) is appropriate. If the categories are ordered, as they are in most rating scales in clinical, psychological and epidemiological research, the weighted kappa statistic (K_w) is preferable. But K_w corresponds to the intraclass correlation coefficient, which, like *r* for continuous variables, is incapable of detecting bias. Simple techniques for detecting bias in the case of ordered categorical variables are described and commended to investigators.

Key words: categorical variables, continuous variables, correlation, fixed bias, kappa statistic, least products regression analysis, limits of agreement, log-linear modelling, McNemar test, method of differences, proportional bias.

INTRODUCTION

Biomedical investigators often wish to compare two methods of measurement, usually to compare a new method with an established one. It is an important prerequisite for such studies that the same individual must make the measurements or ratings. Alternatively, investigators may wish to compare the performances of two measurers or raters who are using the same method of measurement, or to evaluate the repeatability of measurements made by the same observer.

In clinical and laboratory biomedical science, measurements of a variable are usually made on a continuous scale. Examples of this are measurements of blood pressure, blood gases, lung function and plasma concentrations of a variety of endogenous or exogenous substances. In contrast, categorical scales are used to describe or score attributes by epidemiologists, social scientists, clinicians and, occasionally, laboratory scientists. These categorical scales may be unordered, as in the description of eye colour or taste, or they may be ordered, as in rating scales for cardiovascular functional status, operative (anaesthetic) risk, severity of stroke and any number of *ad hoc* scales. Somewhere in between are quasi-continuous scales, such as indices of disability, quality of life and so forth, when the range of the scales can be between 20 and 42.

Whatever the scale of measurement, there is an unusual consensus among biostatisticians that the goal of making these comparisons should not be to demonstrate agreement, but to detect disagreement or bias. However, biostatisticians disagree, sometimes sharply, on how best to achieve this goal. The purpose of the present review is to describe some of the statistical techniques that can be used to detect bias and to evaluate them critically.

CONTINUOUS VARIABLES

This heading is shorthand for variables that are measured on a continuous, or interval, scale. These measurements include distance, weight, concentration, pressure, velocity, temperature, age and so forth. This section deals with three techniques for comparing methods of measurement: (i) regression analysis; (ii) the method of differences; and (iii) correlation.

Detecting bias by regression analysis

As a rule, the values obtained by one method of measurement are linearly related to those obtained by another. It seems to follow logically, therefore, that linear regression analysis would be a useful tool for comparing methods of measurement. But what sort of linear regression analysis?

The familiar form is least squares of *y* regression analysis, commonly known as ordinary least squares (OLS) regression. This is what is provided by most computer statistical programs. But, for comparing two methods of measurement, this is the wrong model on two counts. First, under statistical theory, it is an assumption of

Correspondence: Dr John Ludbrook, 563 Canning Street, Carlton North, Victoria 3054, Australia. Email: johnludbrook@bigpond.com

Received 15 December 2001; revision 18 February 2002; accepted 20 February 2002.

OLS regression that whereas the values of the *y* variable (predicted) are attended by random error, the values of the *x* variable (predictor) are fixed in advance and without random error. It is obvious that this assumption is rarely, if ever, fulfilled in method comparison studies. Second, by convention, in OLS regression the values of the *y* variable are regarded as 'true', the 'gold standard' or the 'benchmark'. Yet, when two methods are compared, it is usual that neither can be regarded as a benchmark. In this context, it is a property of OLS regression analysis that the line resulting from minimizing the sums of the squares of the deviations of the *y* values from the line and that resulting from minimizing the sums of the squares of the deviations of the *x* values from the line, are distinctly and sometimes markedly different.[1]

There is an alternative to the OLS regression model. It is ordinary least products (OLP) regression analysis, whose properties and method of execution have been described in detail elsewhere.[1] Least products regression analysis allows for both the *y* and the *x* values to be attended by random error. It depends on minimizing the sum of the products of the deviations of both *x* and *y* values from the estimated regression line. It requires no judgement that the *y* variable or the *x* variable provides 'true' or 'benchmark' values. Instead, it provides a technique for interchanging two meth-

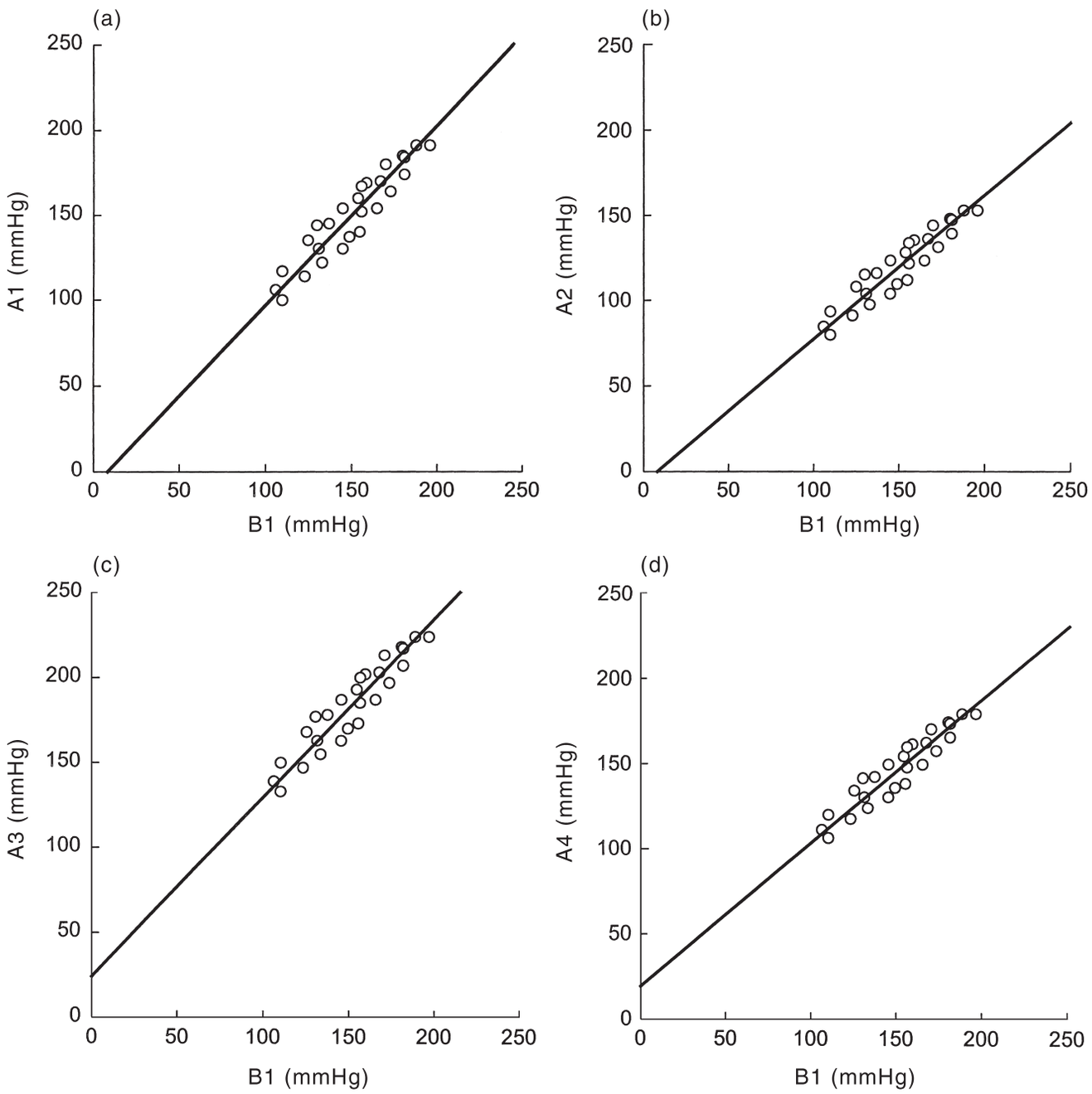

Fig. 1 Data points and ordinary least products regression lines for four hypothetical comparisons of Methods A and B for measuring systolic blood presssure. (a) No bias; (b) proportional bias; (c) fixed bias; (d) proportional and fixed bias. Data points are given in Appendix 1. Details are given in Table 1. Regression model: $E(A) = a + b(B)$. Coefficients for the four regression lines are given in Table 1. Proportional bias: 95% confidence interval (CI) for slope (b) does not include 1. Fixed bias: 95% CI for intercept (a) does not include 0. In all four cases, Pearson's $r = 0.939$.

ods, or calibrating one against the other. It detects proportional and fixed bias (or both) between two methods.

There is a drawback to OLP regression analysis: whereas OLS regression analysis can be executed by a handheld calculator or even with a pencil and paper, OLP regression analysis is an iterative process that requires a computer program that can execute non-linear regression analysis by way of a loss function.[1] One assumption that underlies OLP (and OLS) regression analysis is that the scatter of values around the regression line is constant over the whole range of y (and x) values. In biological work, it is common that the scatter increases with the level of y (and x). In this circumstance, weighted least products (WLP) regression analysis can be used.[1] But this, too, is catered for by most computer statistics programs.

Detecting bias by the method of differences

The notion of examining the differences between the results of two methods of measurement was conceived by Altman and Bland[2] and has recently been elaborated by the same authors.[3] Their proposal is that the differences should be plotted against the means. If the OLS regression line fitted to the plot has a slope (b) that differs 'significantly' from 0, then it is argued that proportional bias exists. If the mean value for the difference ($\bar{d}$) differs 'significantly' from 0 on the basis of a one-sample t-test, then it is argued that there is fixed (or, in their words,[2,3] relative) bias. Altman has been a strong and persistent advocate for making inferences in medical research by using estimation by means of confidence intervals (CI) rather than using P values resulting from hypothesis testing.[4] The CI technique can be applied to the analysis of differences.[3]

The slope of the regression of differences on means is a satisfactory method for detecting proportional bias. Although OLS regression analysis is used by Altman and Bland to estimate slope, this can be defended by the argument that differences must always be the dependent (predicted) variable, means never. There is, however, a serious flaw in detecting fixed bias by the method of differences.[1] It is that if there is proportional bias ($b \neq 0$), then the mean difference ($\bar{d}$) will almost inevitably deviate from zero. Thus, there is the risk that fixed bias will be overdiagnosed. The exception to this is if proportional bias is in one direction (e.g. $b > 0$) and fixed bias in the opposite direction. Then $\bar{d}$ may be close to 0 and fixed bias will be underdiagnosed.

Quantifying disagreement between methods of measurement

Bland and Altman make the important point that the mere absence of bias often does not provide sufficient information to allow a judgement that one method can, or cannot, be substituted for the other.[3] The ability to make this judgement may be very important in clinical practice. They suggest that a useful aid to making this judgement is to calculate what they call the 95% limits of agreement. Their formula for this is:

$$\bar{d} \pm z_{2\alpha} s_d \qquad [1]$$

where $\bar{d}$ is the mean difference between methods, s_d is the standard deviation of the difference between methods and $z_{2\alpha}$ is the standardized normal deviate corresponding to two-sided $P = 0.05$ (1.960).

A better, somewhat more conservative, formulation is safer in the case of small samples, for instance $n < 100$.[5] It is:

$$\bar{d} \pm t_{n-1,2\alpha} s_d \sqrt{(1 + 1/n)} \qquad [2]$$

where $\bar{d}$ is the mean difference between methods, s_d is the standard deviation of the difference between methods, $t_{n-1,2\alpha}$ is the value of t corresponding to two-sided $P = 0.05$ for d.f. $= n - 1$ and $\sqrt{(1 + 1/n)}$ is an adjustment for small sample size.

Equations 1 and 2 are more commonly known as 95% tolerance limits for the population.[5] That is, it is predicted that 95% of values in the parent population of differences will fall within these limits.

Hypothetical example of a continuous variable

Imagine that four different studies have been conducted in which one indirect method for measuring systolic blood pressure has been compared with another. One method (B) is common to all studies. The other four methods are coded A1, A2, A3 and A4. The data resulting from these studies, each in 26 subjects, are provided in Appendix 1.

These four studies were analysed first by ordinary least products (OLP) regression analysis.[1] The outcomes are given in Fig. 1 and Table 1. They are described as indicating no bias, proportional bias, fixed bias or both proportional and fixed bias on the basis of the 95% CI attached to the OLP regression coefficients (Fig. 1; Table 1).

The four hypothetical studies were then re-analysed by the Altman–Bland method of differences.[2,3] The results of these analyses are presented in Fig. 2 and Tables 2,3. Proportional bias is indicated if the slope ($b_{A.B}$) of the OLS regression of differences on means differs 'significantly' from 0 ($P \leq 0.05$) or, equivalently, if the 95% CI for $b_{A.B}$ does not include 0. Fixed bias is indicated if the difference ($\bar{d}$) differs 'significantly' from 0 ($P \leq 0.05$) by one-sample t-test or, equivalently, if the 95% CI for $\bar{d}$ does not include 0. In Fig. 1b and Table 1b, OLP regression analysis indicates that there is proportional, but not fixed, bias. However, in Fig. 2b

Table 1 Outcome of analyses by ordinary least products regression

Proportional	r	a	95% CI	b	95% CI	Proportional bias	Fixed bias
(a) A1B	0.939	−8.8	−32.6, 15.0	1.056	0.900, 1.211	No	No
(b) A2B	0.939	−7.0	−26.0, 12.0	0.844	0.720, 0.969	Yes	No
(c) A3B	0.939	24.2	0.4, 48.0	1.056	0.900, 1.211	No	Yes
(d) A4B	0.939	19.4	0.4, 38.4	0.844	0.720, 0.969	Yes	Yes

For data used in regressions, see Appendix 1.

r, product–moment correlation coefficient. a, b, coefficients in ordinary least products regression model E(A) = $a + b$(B); a, A (y axis) intercept; b, slope; proportional bias, if 95% confidence interval (CI) for b does not include 1; fixed bias, if 95% CI for a does not include 0.

See Fig. 1 for graphical display.

and Tables 2b,3b, the method of differences declares that there is fixed as well as proportional bias.

The outcomes of OLP regression versus the method of differences, with reference to the panels in Figs 1,2, can be summarized as shown in Table 4.

In short, if the method of analysing differences is used, fixed bias can be confounded by proportional bias. If the method of OLP regression analysis is used, this confounding effect does not occur.

As for the 95% level of agreement (95% tolerance limits for the population of differences), this is reliable only when there is no proportional bias (Tables 5a and 5c).

Correlation

Every investigator knows of the product–moment correlation coefficient (r), described by Karl Pearson over 100 years ago.[6] But, how many know how to interpret r? It is a common misapprehension that correlation is synonymous with cause-and-effect. That is simply

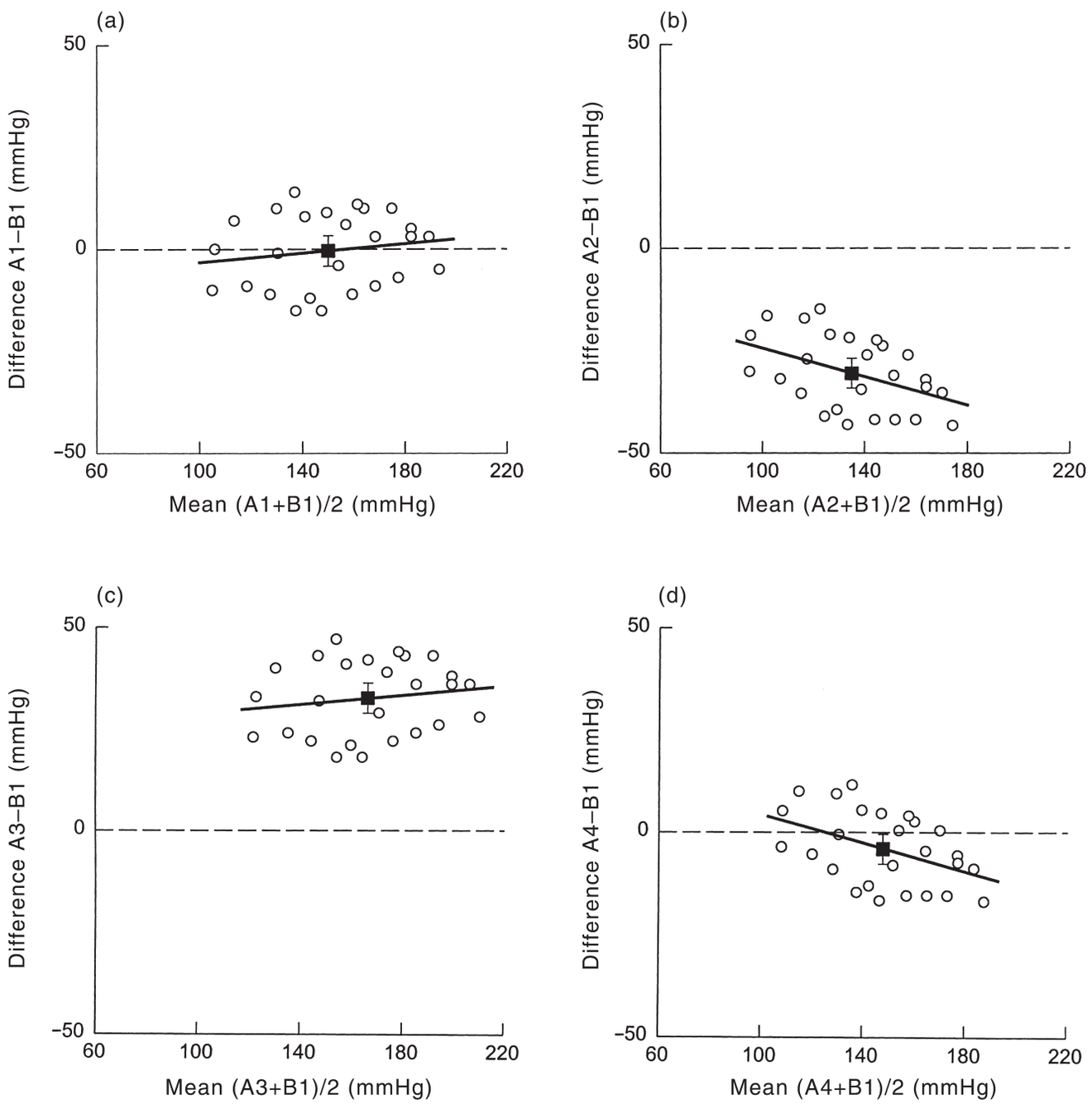

Fig. 2 Data points for plot of differences against means for four hypothetical comparisons of Methods A and B for measuring systolic blood pressure. (a) No bias; (b) proportional and fixed bias; (c) fixed bias; (d) proportional and fixed bias. Data points are given in Appendix 1. Details are given in Tables 2,3. Solid lines, ordinary least squares regression of differences on means. (■), mean value of difference with 95% confidence interval (CI) bars. Proportional bias: slope (b) differs from 0 at $P \leq 0.05$. Fixed bias: mean difference ($\bar{d}$) differs from 0 at $P \leq 0.05$ (and 95% CI does not include 0).

Table 2 Outcomes of analyses of differences by ordinary least squares regression

Regression	*r*	*a*	*b*	*P*	Proportional bias
(a) (A1 – B)/(meanA1B)	0.155	–8.8	0.056	0.450	No
(b) (A2 – B)/(meanA2B)	0.442	–6.9	–0.174	0.024	Yes
(c) (A3 – B)/(meanA3B)	0.155	23.3	0.056	0.450	No
(d) (A4 – B)/(meanA4B)	0.442	21.8	–0.174	0.024	Yes

For data used in regressions, see Appendix 1.

r, product–moment correlation coefficient. *a*, *b*, coefficients in ordinary least squares regression model E(A – B) = *a* + *b*(mean AB); proportional bias, if *b* differs significantly from 0 (i.e. $P \leq 0.05$).

See Fig. 2 for graphical display.

Table 3 Outcomes of analysis of differences by one-sample *t*-test

Difference	Mean difference (±SEM)	95% CI for mean difference	*t*	*P*	Fixed bias
(a) (A1 – B)	–0.39 ± 1.79	–4.07, 3.30	–0.215	0.832	No
(b) (A2 – B)	–30.50 ± 1.76	–34.12, –26.88	–17.345	< 0.0001	Yes
(c) (A3 – B)	32.62 ± 1.79	28.93, 36.30	18.214	< 0.0001	Yes
(d) (A4 – B)	–4.10 ± 1.76	–7.72, –0.48	–2.332	0.028	Yes

For data used in differences, see Appendix 1.

SEM, standard error of the mean. 95% CI, 95% confidence interval for mean difference; *t*, one-sample *t* statistic at d.f. = 25; *P*, two-sided *P* value from *t*-test; fixed bias, if $P \leq 0.05$ or 95% CI does not include 0.

See Fig. 2 for graphical display.

Table 4 Outcomes of ordinary least products regression versus the method of differences

Panel	OLP regression	Method of differences
(a)	No bias	No bias
(b)	Proportional bias	Proportional and fixed bias
(c)	Fixed bias	Fixed bias
(d)	Proportional and fixed bias	Proportional and fixed (just) bias

OLP, ordinary least products.

Table 5 Outcomes of analysis of differences in terms of 95% limits of agreement (95% tolerance limits for population of differences)

Difference	$\bar{d}$ ± 95% tolerance limits for population	95% tolerance limits for population
(a) (A1 – B)	–0.38 ± 19.16	–19.55–18.78
(b) (A2 – B)	–30.50 ± 18.82	–49.32 to –11.68
(c) (A3 – B)	32.62 ± 19.16	13.45–51.78
(d) (A4 – B)	–4.10 ± 18.82	–22.92–14.72

For raw data, see Appendix 1.

$\bar{d}$, mean difference. The formula for 95% tolerance limits for population is: $\bar{d} - t_{n-1,2\alpha} s_d \sqrt{(1 + 1/n)}$, where $t_{n-1,2\alpha}$ is the two-sided value of the *t* statistic, *s* is the standard deviation of $\bar{d}$ and *n* is the number of observations ($n = 26$).

not so. In a vague way, *r* is a measure of association. It is better interpreted as an index of goodness-of-fit of a linear regression model to the observed values. The formula for calculating *r* can be found in most elementary statistical texts. It incorporates the notion of deviations of both *y* and *x* values from a regression line.

Altman and Bland were among the first to point out that *r* is useless for detecting bias in method comparison studies,[2,3] an opinion that is wholeheartedly supported by the author.[1] If some doubt this, they should inspect Table 1 and Fig. 1. These show clearly that a large and statistically 'highly significant' value for *r* (0.939) can coexist with gross bias. In short, the correlation coefficent has no place in this review. The same is true of the intraclass correlation coefficient.[7]

Commentary

It appears that the Altman–Bland method of differences is not always safe as a technique for detecting fixed bias. Does this matter?

It probably does. Proportional bias is not uncommon when methods of measuring variables such as cardiac output or blood pressure are compared. This may not matter too much if the main clinical interest is in percentage change from baseline. However, fixed bias is a more serious phenomenon. It means that the starting points (or end-points) of the two methods are different. This, in turn, implies that there is a serious and irremediable difference between the two methods. For this reason, it is important that the method for detecting fixed bias be accurate.

As regards the so-called 95% limits of agreement (Table 5),[3] it is clear from the foregoing that only when there is no proportional bias can this technique be used safely to decide whether one method corresponds well enough to the other so that either may be used in clinical practice, as in Tables 5a,5b. In Table 5a, when there is no bias (Fig. 1, Table 1), the difference between the two methods for measuring systolic blood pressure could be as great as $19.16 \times 2 = 36.32$ mmHg. The same is the case if there is fixed bias only. It is unlikely that this would be acceptable to hypertensionologists.

There are two reasons for these unacceptably wide limits of agreement in the hypothetical example: (i) the group (sample) size is very small ($n = 26$); and (ii) systolic blood pressure is a difficult variable to investigate in this way. Both the minimal value (say, 50 mmHg) and the maximal value (say, 250 mmHg) are far removed from zero. This means that the coefficients in OLP regression analysis are almost inevitably attended by wide confidence limits.[1] These could be narrowed only by studying a much larger group (sample).

CATEGORICAL VARIABLES

What exactly is a categorical variable? It is when the outcome of an investigation is measured by assigning it to two or more categories. The categorical scales may be unordered (as in eye colour or taste sensation) or ordered. A good example of the latter is the categorization of the severity of heart failure according to the prescription of the New York Heart Association (NYHA) or the Canadian Cardiovascular Society (CCS), both of which use five-point, ordered rating scales.[8] Is there a sharp separation between categorical and contnuous scales? In theory, there is. Obviously, blood pressure is measured on a continuous scale. Equally obviously, the presence or absence of a feature such as alive or dead is expressed on a two-category scale. However, there is a grey area in between these extremes. What about rating scales in which there are eight, 16 or 24 ordered categories? If there are eight categories, the scale is best regarded as categorical. If there are 24 or more,[9,10] most would regard this as a continuous scale (although the distribution of values is often far from normal). But, what if the maximum range of categories is 12, as in the composite Glasgow Coma Scale?[11] I have no clear answer to this dilemma, other than to collapse the number of categories to three, four or six if the number of observations is less than, say, 200.

The main interest of experimental and, especially, clinical pharmacologists and physiologists is in ordered categorical variables, measured on scales in which there are more than two categories. This will be the focus of the exposition that follows.

The kappa statistic

This was described by Cohen in 1960 as a method for comparing raters when the rating scale is unordered.[12] Cohen later extended this to ordered scales.[13] There are monographs on how to calculate and evaluate the kappa statistics[14–16] and an excellent review article by Kramer and Feinstein.[17] Calculation of both the unweighted (K) and weighted (K_w) kappa statistics is designed to take into account the effects of chance. Originally, the kappa statistic was of interest mainly to social scientists, but, increasingly, it has been embraced by clinical scientists. A search of PubMed (National Library of Medicine, Washington DC; http://www.ncbi.nlm.nih.gov/PubMed/) for the term kappa + statistic shows that from less than five citations annually 20 years ago, these reached approximately 80 annually in 1996–2000 (Fig. 3). This increase seems to be because clinical scientists have been slow to discover the kappa statistic but, having done so, have embraced it enthusiastically, although not always wisely.

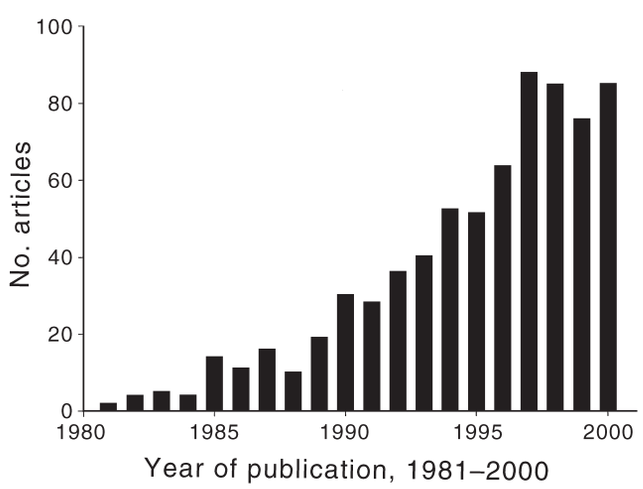

Fig. 3 Articles in the biomedical literature in which the kappa statistic was used, 1981–2000, by year. Data from PubMed (National Library of Medicine, Washington DC; http://www.ncbi.nlm.nih.gov/PubMed/).

Cohen's original intention was that the kappa statistic(s) could be used to compare two raters who use the same categorical scale, rather than to compare methods of rating.[12,13] It is vital to recognize that for the kappa technique to have any value in analysing a study, the two raters must always be the same individuals. It has no place in, for instance, evaluating the performance of peer reviewers of manuscripts or grant applications, when the pairs inevitably consist of different individuals on each occasion[18] (although Fleiss[14] seems to condone this practice). It is also doubtful whether it has a place in comparing methods of categorical measurement. This is because the measurements (ratings) are almost always subjective, in the sense that they are value judgements made by human raters. In a rapidly changing clinical setting, it is impossible to believe that an individual rater could apply two different methods of rating to the same subject within a short space of time without bias. It is more credible that a rater could re-evaluate subjects at, say, an interval of a week or month. That is, it may have a place in evaluating the reproducibility of a single rater's scores.

What if there are more than two raters on the same subject? Fleiss describes a rather complex way of using a modified kappa statistic to evaluate the association between several raters[14] but, in my view, it is best to stick with just two.

The main focus of this review will be on the ordered kappa statistic (K_w) and alternatives to it, although it will be necessary to consider first the original, unweighted kappa statistic (K). It should be said, in advance, that the argument will be that the kappa statistic provides no useful information about the presence or absence of bias.

First, how the results of a study to compare two raters are set out must be described: it is as observed frequencies of occurrence of the categories according to rater, in an $r \times c$ table in which $r = c$ (Table 6). It should be noted that the diagonal of cells from top left to bottom right is described as the agreement diagonal, because only on this diagonal do the two raters agree completely. In addition, the expected frequencies of occurrence for each cell in the table are calculated by a formula that will be familiar to those who use the Chi-squared statistic and which is given in Appendix 2. Because

Table 6 Stereotype 3 × 3 table for kappa test of association: observed cell frequencies

	Rater A			
	1	2	3	Row total
Rater B				
1	*a*	*b*	*c*	*a* + *b* + *c*
2	*d*	*e*	*f*	*d* + *e* + *f*
3	*g*	*h*	*i*	*g* + *h* + *i*
Column total	*a* + *d* + *g*	*b* + *e* + *h*	*c* + *f* + *i*	*N*

Expected cell frequency for a given cell is calculated as ((Row total)/(Column total)/*n*).

Agreement diagonal, *a*, *e*, *i*.

the simplest formulations of the kappa statistics make use of the observed and expected proportions for each cell (p_o and p_e), the formulae for these are also given in Appendix 2.

The unweighted kappa statistic (K)

This is designed to be used only in cases in which the categorical variables are unordered. Its formula is based on p_o and p_e, where p_o is the observed proportion and p_e is the expected proportion:

$$K = (\Sigma p_o - \Sigma p_e)/(1 - \Sigma p_e) \quad [3]$$

where the summing of p_o and p_e (Σp_o and Σp_e, respectively) is done only for the cells on the agreement diagonal.

The values of K can range from – 1 to + 1, although, in practice, they range from 0 (when agreement is no better than chance) to + 1 (perfect agreement). A value of – 1 can occur only under very special conditions.[16]

The weighted kappa statistic (K_w)

This is designed to be used in cases in which the categories are ordered (although it has sometimes been used for unordered catgories). It introduces the notion of a weighting factor, *w*. This is usually intended to exact a greater 'penalty' for greater degrees of disagreement. The formula is:

$$K_w = 1 - ((\Sigma(w)(p_o)\text{across all cells})/(\Sigma(w)(p_e)\text{across all cells})) \quad [4]$$

There are several systems of weighting. Cohen originally described an arbitrary weighting system, in which the investigator decides arbitrarily which disagreements should attract the worst 'penalty'.[13] It is rarely used nowadays. Another is the linear weighting system, in which the weights for cell proportions from the agreement diagonal outwards in each direction progress linearly: 0, 1, 2, 3, 4 and so forth. Paradoxically, Fleiss described an inverse weighting system,[14] under which cells on the agreement diagonal are assigned the maximal weight, while disagreements are assigned less than maximal weightings. It is hard to see merit in this. The most popular system is the quadratic weighting system, in which the weights for cell proportions from the agreement diagonal outwards progress geometrically: 0^2, 1^2, 2^2, 3^2, 4^2 and so forth. This quadratic system provides the greatest leverage to entries in cells that are remote from the agreement diagonal. In all the weighting systems (except Fleiss' inverse system), the weight for cells on the agreement diagonal is $w = 0$. Thus, the value of K_w depends only on the off-diagonal entries. The greater the value of $(\Sigma(w)(p_o))/(\Sigma(w)(p_e))$ in eqn 4, the smaller the value of K_w.

As in the case of K, the values of K_w can range from – 1 to + 1. This is reminiscent of the product–moment correlation coefficient. In fact, an important property of K_w (quadratic weighting) is that it corresponds to one form of the intraclass correlation coefficient.[19]

Table 7 Hypothetical 3 × 3 table of frequencies for ordered categories from comparison of two raters

	Rater A			Row
	I	II	III	total
Rater B				
I	5	2	1	8
II	2	5	3	10
III	1	2	5	8
Column total	8	9	9	26

Value of weighted kappa statistic (quadratic weighting), $K_w = 0.485$ (exact two-sided $P = 0.020$).

Bias by method of modified McNemar or single binomial tests, exact two-sided $P > 0.999$ (no bias).

Evaluating the kappa statistics

There are three different approaches to this matter: (i) applying classical statistical theory; (ii) an empirical approach; and (iii) the use of permutation.

Asymptotic CI and P *values*

This approach depends on the assumption that K and K_w are approximately normally distributed. It provides reasonably accurate values for 95% CI and *P* provided the number of observations is sufficiently large (for instance, > 100). The asymptotic approach depends on estimating asymptotic standard errors (ASE). For estimating CI, the formulation is (ASE_1). A simple method of doing this for K was described by Cohen,[12] a better method by Fleiss *et al.*[14,20] However, for testing the null hypothesis K = 0, a different formulation is required (ASE_0). Cohen's original formulation of ASE_0[12] has since been improved on by others.[14,16,20]

Different formulations of ASE are needed for evaluating K_w. Again, a simple method was described by Cohen.[13] Subsequently, more complex but theoretically more accurate methods were described by Fleiss and others.[14,16,20]

An empirical approach

A very simple method for evaluating agreement for the unweighted kappa statistic was proposed by Landis and Koch.[21] It is as follows: K < 0.00, poor; K = 0.00–0.20, slight; K = 0.21–0.40, fair; K = 0.41–0.60, moderate; K = 0.61–0.80, substantial; K = 0.81–1.00, almost perfect.

This approach is mentioned only to condemn it. It has no sound theoretical basis and can be positively misleading to investigators. It is regrettable that it is reproduced by Fleiss[14] and by the authors of several general texts of statistics (unnamed).

Exact P *values*

A much safer approach, if only because it caters equally well for large and small samples, is to use permutation to arrive at 'exact' *P* values. According to this technique, two-sided *P* for the null hypothesis that K or $K_w = 0$ when all possible permutations of the cell entries are listed, with the proviso that the marginal totals remain the same as those oberved, is:

$$\frac{\text{(No. values for K} \geq \text{that observed, in either direction)}}{\text{(Total no. permutations)}}$$

The 95% 'exact' CI corresponds to the 2.5 and 97.5% percentiles of the listed total number of possible permutations. StatXact 5 (Cytel Software, Cambridge, MA, USA) will execute this for *P* values and CI and provides for all the weighting systems mentioned above.

Detecting bias

The emphasis of this review has been on detecting bias. The kappa statistic cannot do this. The correspondence of K_w to a correlation coefficient invites this conclusion. The examples of Tables 7,8 confirm it.

Bias between two raters in this context means that one gives consistently higher (or lower) ratings than the other. In Table 7, there is disagreement between raters, but this is distributed more or less evenly between Rater A and Rater B. Quadratic $K_w = 0.485$ (exact two-sided $P = 0.020$), indicating reasonable agreement. However, in Table 8, it is clear that Rater A consistently gives higher ratings than Rater B. Yet, $K_w = 0.519$ and exact two-sided $P = 0.002$. On the face of it, on the data of Table 8 the agreement between raters is better than in Table 7. In Table 9, it is clear that the more 'leverage' that is exerted by the weighting system, the greater is the value of K_w. This is absurd.

The existence of this paradox has been recognized for at least 10 years. The solution that has been most often proposed is to use log-linear modelling to analyse the outcome of rater comparison studies.[22–24] However, the outcome of log-linear modelling is difficult for biomedical investigators and readers of their papers to understand, because they are used to simple outcomes such as *P* values or CI. There are simpler methods that are easier to execute and more readily comprehensible.

Simple methods for detecting bias

The procedures described here are based on the premise that if there is bias between raters, it will be reflected in an inequality of observed frequencies in the off-agreement diagonal entries in the table of frequencies. Under a null hypothesis of no bias, it would be expected that the sum of the entries in the diagonals at from 1 to *n* removes from the agreement diagonal will be equal for the upper right and lower left diagonals.

This can be explained by two hypothetical examples in which the performance of two raters is compared, in 3×3 ordered tables (Tables 7,8). What method(s) can be used to detect bias?

Modified McNemar test

Kramer and Feinstein suggested a modification of the McNemar test.[17] Their formula for this was:

$$\chi^2 = (\Sigma UR - \Sigma LL)2/(\Sigma UR - \Sigma LL) \text{ at d.f.} = 1$$

where ΣUR is the sum of the entries in the upper right off-agreement diagonal and ΣLL is the sum of the entries in the lower left off-agreement diagonal.

The attached *P* value can be obtained by reference to the Chi-squared distribution. However, it is much safer to obtain it by exact permutation (StatXact 5; Cytel Software), from which $P > 0.999$ for Table 7 and $P = 0.012$ for Table 8 (see Table 8).

In view of the gross asymmetry (bias) in Table 8, $P = 0.012$ is a rather disppointing outcome. It could be argued that if the weighted kappa statistic K_w is used to test for association (see above), then it is proper to use the same weighting system in performing the modified McNemar test. Thus, if quadratic weighting is used, as it was for K_w in the example of Table 8, so that ΣUR_w becomes 12 and ΣLL_w becomes 1, exact *P* from the McNemar test is $P = 0.0003$ (see Table 9). However, critics have argued that if one were to use a sufficently steep weighting system, for instance 0^{10}, 1^{10}, 2^{10} and so forth, then even the smallest difference between ΣUR and ΣLL would reveal bias. This argument is irrefutable, so that the only safe course to follow is not to use weighted entries and to increase the power of the test by increasing group (sample) size.

Exact single binomial test

This had been my first thought as a method for detecting bias. It tests whether the ratio between ΣUR and ΣLL is 0.5.[14] In its exact form, it gives precisely the same exact *P* values as the modified McNemar test, so that it is neither better nor worse and is susceptible to the same criticisms about choice of weighting system.

Commentary

It is clear that even when the weighted kappa statistic is used appropriately, it is incapable of detecting bias between two raters. This is not surprising, in view of the fact that the quadratic K_w corresponds to the intraclass correlation coefficient and that when variables are continuous the product–moment correlation coefficient is also incapable of detecting bias.

Table 8 Hypothetical 3×3 table of frequencies for ordered categories from comparison of two raters

	Rater A			
	I	II	III	Row total
Rater B				
I	5	4	2	11
II	0	5	4	9
III	0	1	5	6
Column total	5	10	11	26

Value of weighted kappa statistic (quadratic weighting), $K_w = 0.519$ (exact two-sided $P = 0.002$).

Bias by method of modified McNemar or single binomial tests, exact two-sided $P = 0.012$ (gross bias).

Table 9 Values of kappa and *P* for bias for the data of Tables 7 and 8

Type of kappa	Value of kappa	*P* for kappa = 0	*P* for McNemar test for bias (asymptotic)	*P* for McNemar test for bias (exact)
Data of Table 7				
K (unweighted)	0.364	0.013	0.763	> 0.999
K_w (linear)	0.423	0.001	0.782	> 0.999
K_w (quadratic)	0.485	0.020	0.808	> 0.999
Data of Table 8				
K (unweighted)	0.385	0.0042	0.0067	0.012
K_w (linear)	0.451	0.0010	0.0023	0.0034
K_w (quadratic)	0.519	0.0019	0.0003	0.0003

Note that the values for kappa for the data of Table 7 (no bias) are consistently smaller than those for the data of Table 8 (gross bias). The corresponding *P* values are consistently lower for Table 8 compared with Table 7. Note also the discrepancies between the asymptotic and exact *P* values for the McNemar test for bias.

The simple methods for detecting bias that are described above are not based on sophisticated statistical theory but, rather, on simple logic. Yet, they seem to be effective, not only in detecting bias, but in doing so in a way that should be intelligible to biomedical investigators.

One of the important goals of comparing two methods of measurement of continuous variables is to be able to express the size of the differences in quantitative terms, so that clinical investigators and clinicians are in a position to judge whether the magnitude of the disagreement allows one method to be substituted for the other. Conceptually, it seems to be impossible to do this for the situation in which the performance of two raters is compared. All that can be concluded is that there is, or is not, bias between the two raters.

ACKNOWLEDGEMENTS

I am grateful to the many biostatisticians and non-statisticians who have helped me to develop this review and to the two reviewers of the manuscript. In particular, I thank the staff and higher-degree students of the National Stroke Research Institute of Australia (Heidelberg West, Victoria, Australia; Director Professor Geoffrey Donnan), who have listened patiently to my expositions of these ideas and who have offered constructive comments and criticisms.

REFERENCES

1. Ludbrook J. Comparing methods of measurement. *Clin. Exp. Pharmacol. Physiol.* 1997; **24**: 193–203.
2. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. *Statistician* 1983; **32**: 307–17.
3. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat. Meth. Med. Res.* 1999; **8**: 135–60.
4. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: Estimation rather than hypothesis testing. *BMJ* 1986; **292**: 746–50.
5. Lentner C (ed.). Introduction to statistics, statistical tables, mathematical formulae. In: *Geigy Scientific Tables*, Vol. 2. Ciba-Geigy, Basle. 1982; 205.
6. Pearson K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Phil. Trans. R. Soc. Lond. A* 1896; **187**: 253–318.
7. Bland JM, Altman DG. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. *Comput. Biol. Med.* 1990; **20**: 337–40.
8. Goldman L, Hashimoto B, Cook EF, Loscalzo A. Comparative reproducibility and validity of systems for assessing cardiovascular functional class: Advantages of a new specific activity scale. *Circulation* 1981; **64**: 1227–34.
9. Mahoney F, Barthel D. Functional evaluation: The Barthel Index. *Md State Med. J.* 1965; **14**: 61–5.
10. Brott T, Adams HP, Olinger CP *et al.* Measurements of acute cerebral infarction: A clinical examination scale. *Stroke* 1989; **20**: 864–70.
11. Teasdale G, Murray G, Parker L, Jennett B. Adding up the Glasgow Coma Score. *Acta Neurochir. Suppl.* 1979; **28**: 13–16.
12. Cohen J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 1960; **20**: 37–46.
13. Cohen J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 1968; **70**: 213–20.
14. Fleiss JL. *Statistical Methods for Rates and Proportions*, 2nd edn. John Wiley & Sons, New York. 1981.
15. Siegel S, Castellan NJ. *Nonparametric Statistics for the Behavioral Sciences*, 2nd edn. McGraw-Hill, New York. 1988.
16. Agresti A. *Categorical Data Analysis*. John Wiley & Sons, New York. 1990.
17. Kramer MS, Feinstein AR. Clinical biostatistics. LIV. The biostatistics of concordance. *Clin. Pharmacol. Ther.* 1981; **29**: 111–23.
18. Ludbrook J. Peer review of manuscripts. *J. Clin. Neurosci.* 2002; **9**: 105–8.
19. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* 1973; **33**: 613–19.
20. Fleiss JL, Cohen J, Everitt BS. Large sample standard errors of kappa and weighted kappa. *Psychol. Bull.* 1969; **72**: 323–7.
21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–74.
22. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J. Clin. Epidemiol.* 1993; **46**: 423–9.
23. Graham P, Jackson R. The analysis of ordinal agreement data: Beyond weighted kappa. *J. Clin. Epidemiol.* 1993; **46**: 1055–62.
24. May SM. Modelling observer agreement: An alternative to kappa. *J. Clin. Epidemiol.* 1994; **47**: 1315–24.

APPENDIX 1

Table A1 Dataset for example of two continuous variables. Methods A1, A2, A3 and A4 for indirect measurement of systolic blood pressure versus Method B, both in mmHg

A1	A2	A3	A4	B
106.0	84.8	139.0	111.2	106.0
100.0	80.0	133.0	106.4	110.0
117.0	93.6	150.0	120.0	110.0
114.0	91.2	147.0	117.6	123.0
135.0	108.0	168.0	134.4	125.0
130.0	104.0	163.0	130.4	131.0
144.0	115.2	177.0	141.6	130.0
122.0	97.6	155.0	124.0	133.0
145.0	116.0	178.0	142.4	137.0
130.0	104.0	163.0	130.4	145.0
137.0	109.6	170.0	136.0	149.0
154.0	123.2	187.0	149.6	145.0
140.0	112.0	173.0	138.4	155.0
152.0	121.6	185.0	148.0	156.0
160.0	128.0	193.0	154.4	154.0
169.0	135.2	202.0	161.6	159.0
154.0	123.2	187.0	149.6	165.0
170.0	136.0	203.0	162.4	167.0
164.0	131.2	197.0	157.6	173.0
180.0	144.0	213.0	170.4	170.0
167.0	133.6	200.0	160.0	156.0
174.0	139.2	207.0	165.6	181.0
185.0	148.0	218.0	174.4	180.0
191.0	152.8	224.0	179.2	188.0
184.0	147.2	217.0	173.6	181.0
191.0	152.8	224.0	179.2	196.0

APPENDIX 2

Table A2 Formulae for cells in $r \times c$ tables

	Formula
Expected frequency	((Observed row total) (Observed column total))/*N*
Observed proportion (p_o)	(Observed cell frequency)/*N*
Expected proportion (p_e)	(Expected cell frequency)/*N*