SCRATCH

Kevin O'Brien

December 19, 2016

Contents

0.1 Bland Altman Methodology

0.1.1 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* [3] and propose estimating its value by determining the mean of the differences. The variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

0.2 The Bland Altman Plot

In 1986 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes. It has proved highly popular ever since. This is a simple, and widely used , plot of the differences of each data pair, and the corresponding average value. An important requirement is that the two measurement methods use the same scale of measurement.

Variations of the Bland Altman plot is the use of ratios, in the place of differences.

$$D_i = X_i - Y_i \tag{1}$$

Altman and Bland suggest plotting the within subject differences $D = X_1 - X_2$ on the ordinate versus the average of x_1 and x_2 on the abscissa.

0.2.1 Criticism of Bland Altman Plot

Unfortunately the Bland-Altman plot has a fatal flaw: it indicates incorrectly that there are systematic differences or bias in the relationship between two measures, when one has been calibrated against the other. (Hopkins)

0.2.2 Treatment of Outliers

Bland and Altman attend to the issue of outliers in their 1986 paper, wherein they present a data set with an extreme outlier

0.3 Paired T tests

This method can be applied to test for statisitcally significant deviations in bias. This method can be potentially misused for method comparison studies.

It is a poor measure of agreement when the rater's measurements are perpendicular to

the line of equality[Hutson et al]. In this context, an average difference of zero between the two raters, yet the scatter plot displays strong negative correlation.

Components in assessing agreement

1. The degree of linear relationship between the two sets
2. The amount of bias as represented by the difference in the means
3. The Differences in the two variances.

0.4 Methods of assessing agreement

1. Pearson's Correlation Coefficient
2. Intraclass correlation coefficient
3. Bland Altman Plot
4. Bartko's Ellipse (1994)
5. Blackwood Bradley Test
6. Lin's Reproducibility Index
7. Luiz Step function

Bland and Altman attend to the issue of repeated measures in 1996.
Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual. Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation ,and the correlation coefficient.

The above plot incorporates both the conventional limits of agreement (the inner pair of dashed lines), the 't' limits of agreement (the outer pair of dashed lines) centred around the inter-method bias (indicated by the full line). This plot is intended for expository purposes only, as the sample size is small.

0.4.1 Equivalence and Interchangeability

Limits of agreement are intended to analyse equivalence. How this is assessed is the considered judgement of the practitioner. In Bland and Altman (1986) an example of good agreement is cited. For two methods of measuring 'oxygen saturation', the limits of agreement are calculated as (-2.0,2.8).A practitioner would ostensibly find this to be sufficiently narrow.

If the limits of agreement are not clinically important, which is to say that the differences tend not to be substantial, the two methods may be used interchangeably. Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

0.5 Bland Altman Plots In Literature

Mantha et al. (2000) contains a study the use of Bland Altman plots of 44 articles in several named journals over a two year period. 42 articles used Bland Altman's limits of agreement, wit the other two used correlation and regression analyses. Mantha et al. (2000) remarks that 3 papers, from 42 mention predefined maximum width for limits of agreement which would not impair medical care.

The conclusion of Mantha et al. (2000) is that there are several inadequacies and inconsistencies in the reporting of results ,and that more standardization in the use of Bland Altman plots is required. The authors recommend the prior determination

of limits of agreement before the study is carried out. This contention is endorsed by Lin et al. (1991), which makes a similar recommendation for the sample size, noting that*sample sizes required either was not mentioned or no rationale for its choice was given.*

> In order to avoid the appearance of "data dredging", both the sample size and the (limits of agreement) should be specified and justified before the actual conduct of the trial. (Lin et al., 1991)

Dewitte et al. (2002) remarks that the limits of agreement should be compared to a clinically acceptable difference in measurements.

0.5.1 Gold Standard

This is considered to be the most accurate measurement of a particular parameter.

0.6 Discussion on Method Comparison Studies

The need to compare the results of two different measurement techniques is common in medical statistics.

In particular, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and a traditional reference or gold standard must be evaluated before the new one is put into practice. Various methodologies have been proposed for this purpose in recent years.

Indications on how to deal with outliers in Bland Altman plots

We wish to determine how outliers should be treated in a Bland Altman Plot

In their 1983 paper they merely state that the plot can be used to 'spot outliers'.

In their 1986 paper, Bland and Altman give an example of an outlier. They state that it could be omitted in practice, but make no further comments on the matter.
In Bland and Altmans 1999 paper, we get the clearest indication of what Bland and Altman suggest on how to react to the presence of outliers. Their recommendation is to recalculate the limits without them, in order to test the difference with the calculation where outliers are retained.

The span has reduced from 77 to 59 mmHg, a noticeable but not particularly large reduction.
However, they do not recommend removing outliers. Furthermore, they say:
We usually find that this method of analysis is not too sensitive to one or two large outlying differences.
We ask if this would be so in all cases. Given that the limits of agreement may or may not be disregarded, depending on their perceived suitability, we examine whether it would possible that the deletion of an outlier may lead to a calculation of limits of agreement that are usable in all cases?
Should an Outlying Observation be omitted from a data set? In general, this is not considered prudent.
Also, it may be required that the outliers are worthy of particular attention themselves.
Classifying outliers and recalculating We opted to examine this matter in more detail.
The following points have to be considered
how to suitably identify an outlier (in a generalized sense)
Would a recalculation of the limits of agreement generally results in a compacted range between the upper and lower limits of agreement?

0.6.1 Agreement

Bland and Altman (1986) define Perfect agreement as 'The case where all of the pairs of rater data lie along the line of equality'. The Line of Equality is defined as the 45 degree line passing through the origin, or X=Y on a XY plane.

0.6.2 Lack Of Agreement

1. Constant Bias

2. Proportional Bias

Constant Bias

This is a form of systematic deviations estimated as the average difference between the test and the reference method

Proportional Bias

Two methods may agree on average, but they may exhibit differences over a range of measurements

0.7 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.
Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

0.7.1 Bland Altman plots using 'Gold Standard' raters

According to Bland and Altman, one should use the methodology previous outlined, even when one of the raters is a Gold Standard.

0.7.2 Bias Detection

further to this method, the presence of constant bias may be indicated if the average value differences is not equal to zero. Bland and Altman does, however, indicate the indication of absence of bias does not provide sufficient information to allow a judgement as to whether or not one method can be substituted for another.

Chapter 1

The Bland Altman Plot

1.1 Bland Altman Plots

The issue of whether two measurement methods are comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of matched pairs correlation coefficients or simple linear regression. Bland and Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983).

As an alternative they proposed a simple statistical methodology specifically appropriate for method comparison studies. They acknowledge that there are other valid methodologies, but argue that a simple approach is preferable to complex approaches, *"especially when the results must be explained to non-statisticians"* (Altman and Bland, 1983).

The first step recommended which the authors argue should be mandatory is construction of a simple scatter plot of the data. The line of equality ($X = Y$) should also

be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 1.1). The averages of the two measurements is considered by Bland and Altman to the best estimate for the unknown true value. Importantly both methods must measure with the same units. These results are then plotted, with differences on the ordinate and averages on the abscissa (figure 1.2). Altman and Bland (1983)express the motivation for this plot thusly:

> "From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study."

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 1.1: Fotobalk and Counter Methods: Differences and Averages

1.1.1 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods.

The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this.

Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

1.1.2 Criticism of Bland Altman Plot

Hopkins[8] argues that the plot indicates incorrectly that there are systematic differences or bias in the relationship between two measures, when one has been calibrated against the other.

An Evaluation of the correlation between the difference and means complement the analysis.

Bland and Altman caution that the calculations are based on the assumption that the data is normally distributed. This can be verified by using a histogram. If Data is not normally distributed, it can be transformed.

Chapter 2

REGRESSION

2.1 Model II Regression

2.1.1 Simple Linear Regression

Simple Linear Regression is well known statistical technique , wherein estimates for slope and intercept of the line of best fit are derived according to the Ordinary Least Square (OLS) principle.This method is known to Cornbleet and Cochrane as Model I regression.

In Model I regression, the independent variable is assumed to be measured without error. For method comparison studies, both sets of measurement must be assumed to be measured with imprecision and neither case can be taken to be a reference method. Arbitrarily selecting either method as the reference will yield two conflicting outcomes. A fitting based on 'X on Y' will give inconsistent results with a fitting based on 'Y on X'. Consequently model I regression is inappropriate for such cases.

Conversely, Cornbleet Cochrane state that when the independent variable X is a pre-

cisely measured reference method, Model I regression may be considered suitable. They qualify this statement by referring the X as *the 'correct' value*, tacitly implying that there must still be some measurement error present. The validity of this approach has been disputed elsewhere.

2.1.2 Model II regression

Cochrane and Cornbleet argue for the use of methods that based on the assumption that both methods are imprecisely measured ,and that yield a fitting that is consistent with both 'X on Y' and 'Y on X' formulations. These methods uses alternatives to the OLS approach to determine the slope and intercept.

They describe three such alternative methods of regression; Deming , Mandel, and Bartlett regression. Collectively the authors refer to these approaches as Model II regression techniques.

2.1.3 Distribution of Maxima

It is possible to use Order Statistics theory to assess conditional probabilities. With two random variables T_0 and T_1, we define two variables Z and W such that they take the maximum and minimum values of the pair of T values.

2.1.4 Plot of the Maxima against the Minima

In Figure 1, The Maximas are plotted against their corresponding minima. The Critical values of the Maxima and Minima are displayed in the dotted lines.The Line of Equality depicts the obvious logical constraint of the each Maximum value being greater than its corresponding minimum value.

The scientific question at hand is the correct approach to assessing whether two

methods can be used interchangeably. Bland and Altman (1999) expresses this as follows:

> We want to know by how much (one) method is likely to differ from the (other), so that if it not enough to cause problems in the mathematical interpretation we can ... use the two interchangeably.

Consequently, of the categories of method comparison study, comparison studies, the second category, is of particular importance, and the following discussion shall concentrate upon it. Less emphasis shall be place on the other three categories.

Further to Bland and Altman (1986), 'equivalence' of two methods expresses that both can be used interchangeably. Dunn (2002, p.49) remarks that this is a very restrictive interpretation of equivalence, and that while agreement indicated equivalence, equivalence does not necessarily reflect agreement.

The main difference between Myers proposed method and the Bland Altman is that the random effects model is used to estimate the within-subject variance after adjusting for known and unknown variables. The Bland Altman approach uses one way analysis of variance to estimate the within subject variance. In general, the random effects model is an extension of the analysis of the ANOVA method and it can adjust for many more covariates than the ANOVA method

2.1.5 Criticism of Bland Altman Plots

An Evaluation of the correlation between the difference and means complement the analysis.

Bland and Altman caution that the calculations are based on the assumption that the data is normally distributed. This can be verified by using a histogram. If Data is not normally distributed, it can be transformed.

Luiz *et al* remarks that that Bland Altman Plot should be used only for small data sets, as the use of an index will be of little value to the analysis.

2.2 Conclusions about Existing Methodologies

Scatterplots are recommended by Altman and Bland (1983) for an initial examination of the data, facilitating an initial judgement and helping to identify potential outliers. They are not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation.

The Bland Altman methodology is well noted for its ease of use, and can be easily implemented with most software packages. Also it doesn't require the practitioner to have more than basic statistical training. The plot is quite informative about the variability of the differences over the range of measurements. For example, an inspection of the plot will indicate the 'fan effect'. They also can be used to detect the presence of an outlier.

Ludbrook (1997, 2002)criticizes these plots on the basis that they presents no information on effect of constant bias or proportional bias. These plots are only practicable when both methods measure in the same units. Hence they are totally unsuitable for conversion problems. The limits of agreement are somewhat arbitrarily constructed. They may or may not be suitable for the data in question. It has been found that the limits given are too wide to be acceptable. There is no guidance on how to deal with outliers. Bland and Altman recognize effect they would have on the limits of agreeement, but offer no guidance on how to correct for those effects.

There is no formal testing procedure provided. Rather, it is upon the practitioner opinion to judge the outcome of the methodology.

Chapter 3

Appendix

3.0.1 Contention

Several papers have commented that this approach is undermined when the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined. Outliers are a source of error in regression estimates.In method comparison studies, the X variable is a precisely measured reference method. Cornbleet Gochman (1979) argued that criterion may be regarded as the correct value. Other papers dispute this.

3.0.2 Least Product Regression

Least Product Regression , also known as 'Model II regression' caters for cases in which random error is attached to both dependent and independent variables. Ludbrook cites this methodology as being pertinent to Method comparison studies.

The sum of the products of the vertical and horizontal deviations of the x,y values from the line is minimized.

Least products regression analysis is considered suitable for calibrating one method against another.Ludbrook comments that it is also a sensitive technique for detecting and distinguishing fixed and proportional bias between methods.

Proposed as an alternative to Bland & Altman methodology ,this method is also known as 'Geometric Mean Regression' and 'Reduced Major Axis Regression'.

Difference with Least Squares Regression

Least-products regression can lead to inflated SEEs and estimates that do not tend to their true values an N approaches infinity (Draper and Smith, 1998).

3.0.3 Ordinary Least Product Regression

Ludbrook (1997) states that the grouping structure can be straightforward, but there are more complex data sets that have a hierarchical(nested) model.

Observations between groups are independent, but observations within each groups are dependent because they belong to the same subpopulation. Therefore there are two sources of variation: between-group and within-group variance. Mean correction

is a method of reducing bias.

3.0.4 A regression based approach based on Bland Altman Analysis

Lu et al used such a technique in their comparison of DXA scanners. They also used the Blackwood Bradley test. However it was shown that, for particular comparisons,

agreement between methods was indicated according to one test, but lack of agreement was indicated by the other.

3.1 Measurement Error Models

Dunn (2002) proposes a measurement error model for use in method comparison studies. Consider n pairs of measurements X_i and Y_i for $i = 1, 2, ...n$.

$$X_i = \tau_i + \delta_i \tag{3.1}$$

$$Y_i = \alpha + \beta\tau_i + \epsilon_i$$

In the above formulation is in the form of a linear structural relationship, with τ_i and $\beta\tau_i$ as the true values , and δ_i and ϵ_i as the corresponding measurement errors. In the case where the units of measurement are the same, then $\beta = 1$.

$$E(X_i) = \tau_i \tag{3.2}$$

$$E(Y_i) = \alpha + \beta\tau_i$$

$$E(\delta_i) = E(\epsilon_i) = 0$$

The value α is the inter-method bias between the two methods.

$$z_0 = d = 0 \tag{3.3}$$

$$z_{n+1} = z_n^2 + c \tag{3.4}$$

Bibliography

Altman, D. and J. Bland (1983). Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society. Series D (The Statistician) 32*(3), 307–317.

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

Dewitte, K., C. Fierens, D. Stckl, and L. M. Thienpont (2002). Application of the Bland Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry 48*, 799–801.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Lin, S. C., D. M. Whipple, and charles S Ho (1991). Evaluation of statistical equivalence using limits of agreement and associates sample size calculation. *Communications in Statistics - Theory and Methods 27*(6), 1419–1432.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

Mantha, S., M. F. Roizen, L. A. Fleisher, R. Thisted, and J. Foss (2000). Comparing methods of clinical measurement: Reporting standards for bland and altman analysis. *Anaesthesia and Analgesia 90*, 593–602.

O'Brien, E., J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, D. Altman, M. Bland, A. Coats, and N. Atkins (1990). The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *Journal of Hypertension 8*, 607–619.