0.1 Introductory Definitions

0.2 Introduction

The problem of assessing the agreement between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a 'method comparison study'. Published examples of method comparison studies can be found in disciplines as diverse as Pharmacology (**?**), Anaesthesia (**?**), and cardiac imaging methods (**?**). Method Comparison Studies is a branch of statistics used to compare the results of two different method of measurement, measuring the same subject samples. Consider a set of n samples. Measurements are taken on each of the n samples using both methods. This will enable comparison of the method used. In many cases the purpose of the study is to calibrate a new method of measurement against a Gold Standard method. A Gold Standard method is the known method that is considered most precise in its measurement. It should not be assumed that there is no error present in its measurements. The Gold Standard may not be financially feasible for general use, and therefore more economical methods, of suitable levels of precisions, must be devised. Method Comparison studies is used to ascertain the levels of precision of such methods.

To illustrate the characteristics of a typical method comparison study consider the data in Table I, taken from **?**. In each of twelve experimental trials a single round of ammunition was fired from a 155mm gun, and its velocity was measured simultaneously (and independently) by three chronographs devices, referred to here as 'Fotobalk', 'Counter' and 'Terma'.

Round	Fotobalk [F]	Counter [C]	Terma [T]
1	793.8	794.6	793.2
2	793.1	793.9	793.3
3	792.4	793.2	792.6
4	794.0	794.0	793.8
5	791.4	792.2	791.6
6	792.4	793.1	791.6
7	791.7	792.4	791.6
8	792.3	792.8	792.4
9	789.6	790.2	788.5
10	794.4	795.0	794.7
11	790.9	791.6	791.3
12	793.5	793.8	793.5

Table 1: Measurement of the three chronographs (Grubbs 1973)

An important aspect of the these data is that all three methods of measurement are assumed to have an attended measurement error, and the velocities reported in Table I can not be assumed to be ‘true values’ in any absolute sense. For expository purposes only the first two methods ‘Fotobalk’ and ‘Counter’ will enter in the immediate discussion.

A method of measurement should ideally be both accurate and precise.An accurate measurement methods shall give a result close to the ‘true value’. Precision of a method is indicated by how tightly clustered its measurements are around their mean measurement value.

A precise and accurate method should yield results consistently close to the true value. However a method may be accurate, but not precise. The average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely an inaccurate method may be quite precise , as it consistently indicates the same level of inaccuracy.

The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias. The lesser the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero.

A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently there is lack of agreement between the two methods.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.80	794.60	-0.80
2	793.10	793.90	-0.80
3	792.40	793.20	-0.80
4	794.00	794.00	0.00
5	791.40	792.20	-0.80
6	792.40	793.10	-0.70
7	791.70	792.40	-0.70
8	792.30	792.80	-0.50
9	789.60	790.20	-0.60
10	794.40	795.00	-0.60
11	790.90	791.60	-0.70
12	793.50	793.80	-0.30

Table 2: Difference between Fotobalk and Counter measurements

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree or not. These methods must also have equivalent levels of precision. Should one method yield results considerably more variable than that of the other, they can not be considered to be in agreement.

Therefore a methodology must be introduced that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

0.3 Bland Altman Plots

The issue of whether two measurement methods 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 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 (**?**).

Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge that there are other valid, but complex, methodologies, and argue that a simple approach is preferable to this complex approaches, *especially when the results must be explained to non-statisticians* (**?**).

Notwithstanding previous remarks about regression, 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, **?** 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 2.1). These differences and averages are then plotted (Figure 2.2).

The dashed line in Figure 2.2 alludes to the inter method bias between the two methods, as mentioned previously. Bland and Altman recommend the estimation of inter method bias by calculating the average of the differences. In the case of Grubbs data the inter method bias is -0.6083 metres per second.

GrubbsScatter.jpeg

Figure 1: Scatter plot For Fotobalk and Counter Methods

By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

0.3.1 Inspecting the Data

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. **?** 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

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 3: Fotobalk and Counter Methods: Differences and Averages

> values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study."

Figures 1.3 1.4 and 1.5 are three Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of trends that would adversely affect use of the recommended methodology. Figure 1.3 demonstrates how the Bland Altman plot would indicate increasing variance of differences over the measurement range. Figure 1.4 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias (**?**).

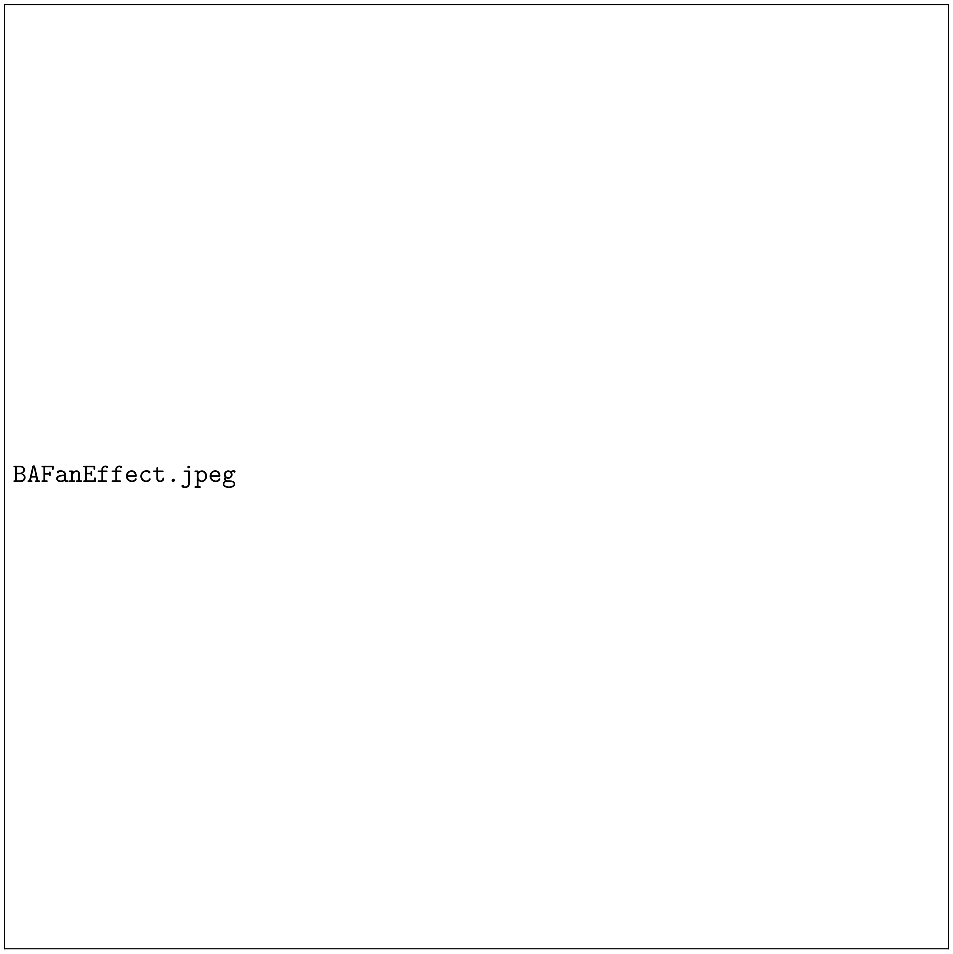

Figure 2: Bland-Altman Plot demonstrating the increase of variance over the range

Figure 1.4 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias (Ludbrook, 1997). Both of these cases violate the assumptions necessary for further analysis using limits of agreement ,which shall be discussed later. The plot also can be used to identify outliers. An outlier is an observation that is numerically distant from the rest of the data. Classification thereof is a subjective decision in any analysis, but must be informed by the logic of the formulation. Figure 1.5 is a Bland Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively.

PropBias.jpeg

Figure 3: Bland-Altman Plot indicating the presence of proportional bias

In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Hence any observation , such as the one on the extreme right of figure 1.5, should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster. The one on the extreme left should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

? do not recommend excluding outliers from analyses. However recalculation of the inter-method bias estimate , and further calculations based upon that estimate, are useful for assessing the influence of outliers.(**?**) states that *"We usually find that this method of analysis is not too sensitive to one or two large outlying differences."*

BAOutliers.jpeg

Figure 4: Bland-Altman Plot indicating the presence of Outliers

0.3.2 Limits of Agreement

? introduces an elaboration of the plot, adding to the plot ‘limits of agreement’ to the plot. These limits are based upon the standard deviation of the differences. The discussion shall be reverted to these limits of agreement in due course.

0.3.3 Variations of the Bland Altman Plot

? remarks that it is possible to ignore the issue altogether, but the limits of agreement would wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. **?** acknowledge that this is not easy to interpret, and that it is not suitable in all cases.

? offers two variations of the Bland -Altman plot that are intended to overcome potential problems that the conventional plot would inappropriate for.

The first variation is a plot of casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases. The second variation is a plot of casewise ratios as percentage of averages.

0.3.4 Agreement

Bland and Altman (1986) defined perfect agreement as the case where all of the pairs of rater data lie along the line of equality, where the line of equality is defined as the 45 degree line passing through the origin(i.e. the $X = Y$ line).

Bland and Altman (1986)expressed this in the terms *we want to know by how much the new method is likely to differ from the old; if this is not enough to cause problems in clinical interpretation we can replace the old method by the new or use the two interchangeably. How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparisonand to choose the sample size .*

0.3.5 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* 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.

0.3.6 Inappropriate assessment of Agreement

Paired T tests

This method can be applied to test for statistically 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.

Inappropriate Methodologies

Use of the Pearson Correlation Coefficient , although seemingly intuitive, is not appropriate approach to assessing agreement of two methods. Arguments against its usage have been made repeatedly in the relevant literature. It is possible for two analytical methods to be highly correlated, yet have a poor level of agreement.

Pearson's Correlation Coefficient

It is well known that Pearson's correlation coefficient is a measure of the linear association between two variables, not the agreement between two variables (e.g., see Bland and Altman 1986)..This is a well known as a measure of linear association between two variables.Nonetheless this is not necessarily the same as Agreement. This method is considered wholly inadequate to assess agreement because it only evaluates only the association of two sets of observations.

0.3.7 Inappropriate use of the Correlation Coefficient

It is intuitive when dealing with two sets of related data, i.e the results of the two raters, to calculate the correlation coefficient (r). Bland and Altman attend to this in their 1999 paper.

They present a data set from two sets of meters, and an accompanying scatterplot. An hypothesis test on the data set leads us to conclude that there is a relationship between both sets of meter measurements. The correlation coeffiecient is determined to be r =0.94.However, this high correlation does not mean that the two methods agree. It is possible to determine from the scatterplot that the intercept is not zero, a requirement for stating both methods have high agreement. Essentially, should two methods have highly correlated results, it does not follow that they have high agreement.

0.3.8 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.

0.3.9 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.

scatter plots

The authors advise the use of scatter plots to identify outliers, and to determine if there is curvilinearity present. In the region of linearity ,simple linear regression may yield results of interest.

0.3.10 Effect of Outliers

Another argument against the use of model I regression is based on outliers. Outliers can adversely influence the fitting of a regression model. Cornbleet and Cochrane compare a regression model influenced by an outlier with a model for the same data set, with the outlier excluded from the data set. A demonstration of the effect of outliers was made in Bland Altman's 1986 paper. However they discourage the exclusion of outliers.

0.3.11 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line.

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis.

The bias is computed as the average of the difference of paired assays.

If one method is sometimes higher, and sometimes the other method is higher, the average of the differences will be close to zero. If it is not close to zero, this indicates that the two assay methods are producing different results systematically.

Precision of Limits of Agreement

The limits of agreement are estimates derived from the sample studied, and will differ from values relevant to the whole population. A different sample would give different limits of agreement. **?** advance a formulation for confidence intervals of the inter-method bias and the limits of agreement. These calculations employ quantiles of the 't' distribution with $n - 1$ degrees of freedom.

0.3.12 Appropriate Use of Limits of Agreement

Importantly **?** makes the following point:

> These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

Carstensen attends to the issue of repeated data, using the expression replicate to express a repeated measurement on a subject by the same methods. Carstensen formulates the data as follows Repeated measurement - Arrangement of data into groups, based on the series of results of each subject.

0.3.13 The Bland Altman Plot - Variations

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.

0.3.14 Pitman & Morgan Test

This test assess tthe equaltiy of population vairances. Pitman's test tests for zero corrleation between the sums and products.

Correlation between differences and means is a test statistics for the null hypothesis of equal variances given bivariate normality.

0.3.15 Lin's Reproducibility Index

Lin proposes the use of a reproducibility index, called the Concordance Correlation Coefficent (CCC).While it is not strictly a measure of agreement as such, it can form part of an overall method comparision methodology.