1 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 (**?**).

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.

2 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.

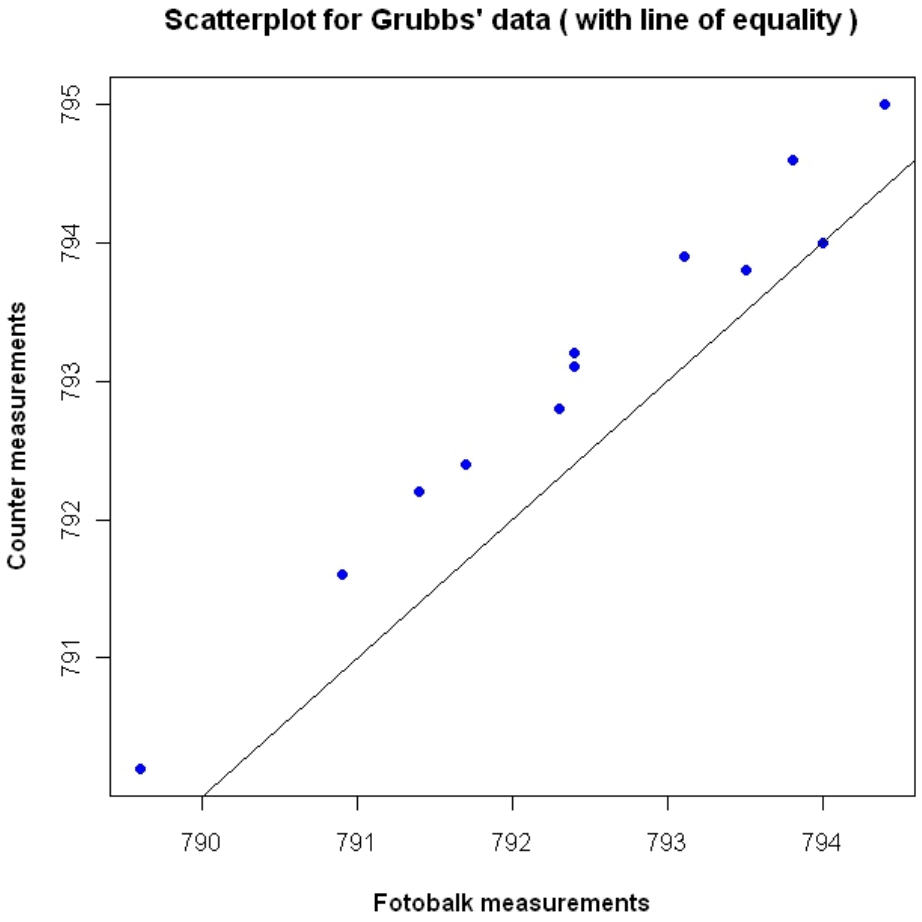

Figure 1: Scatter plot For Fotobalk and Counter Methods

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.

2.1 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 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 3: Fotobalk and Counter Methods: Differences and Averages

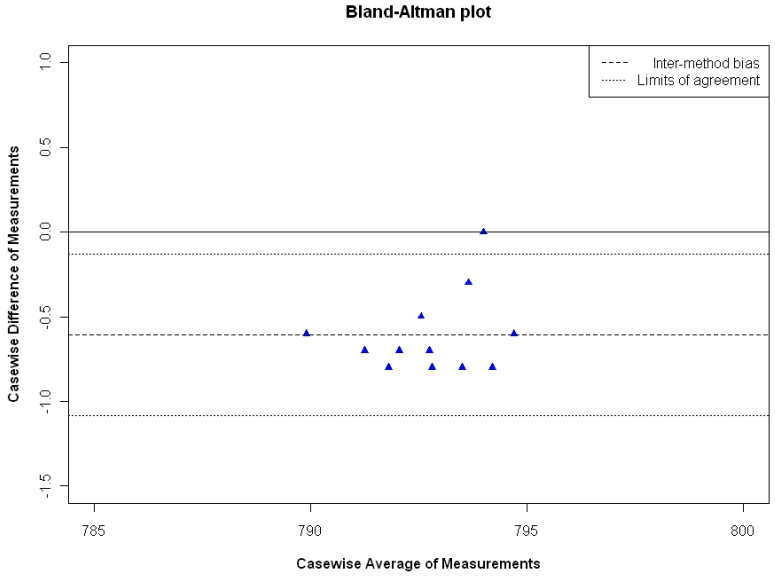

Figure 2: Bland Altman Plot For Fotobalk and Counter Methods

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 demon-

strates 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 (**?**).

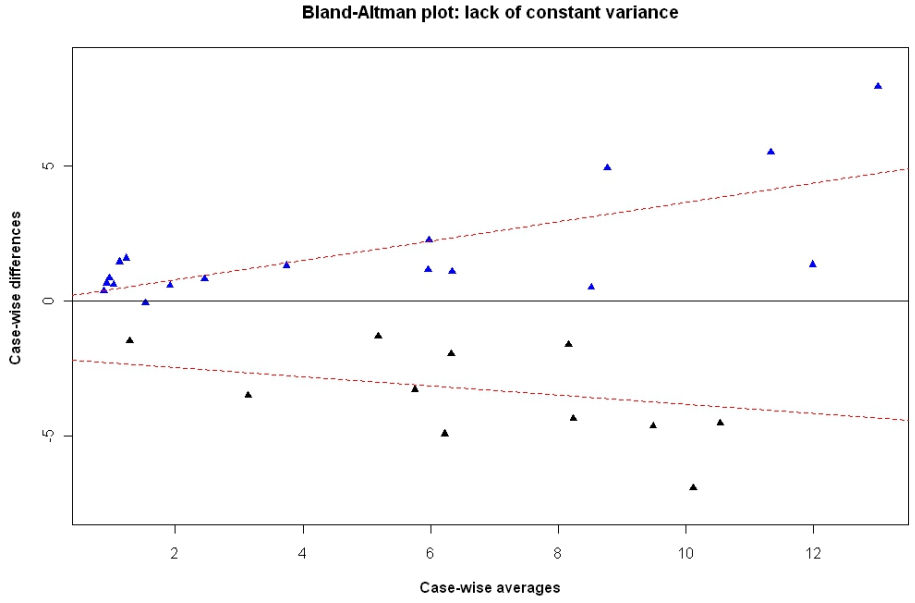

Figure 3: Bland-Altman Plot demonstrating the increase of variance over the range

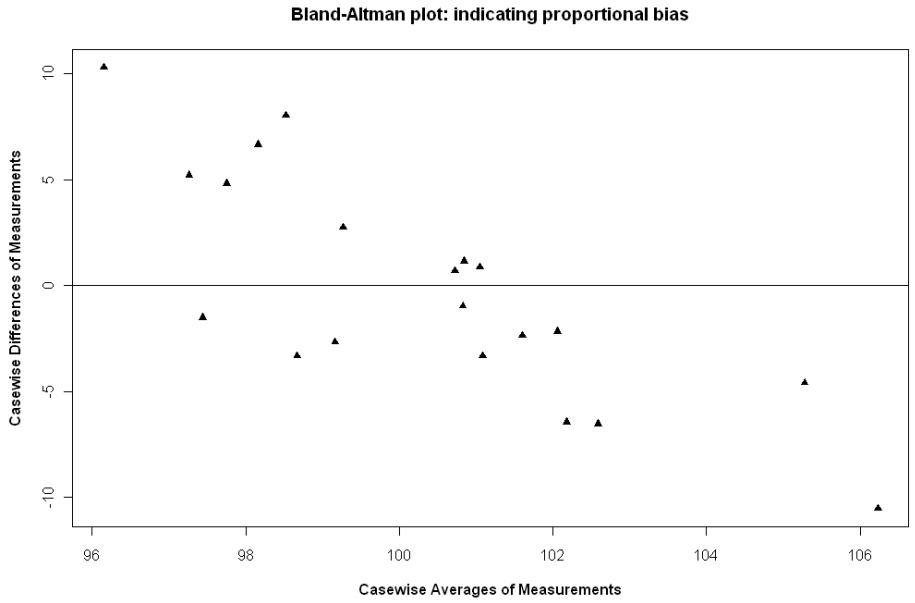

Figure 4: Bland-Altman Plot indicating the presence of proportional bias

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.

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."*

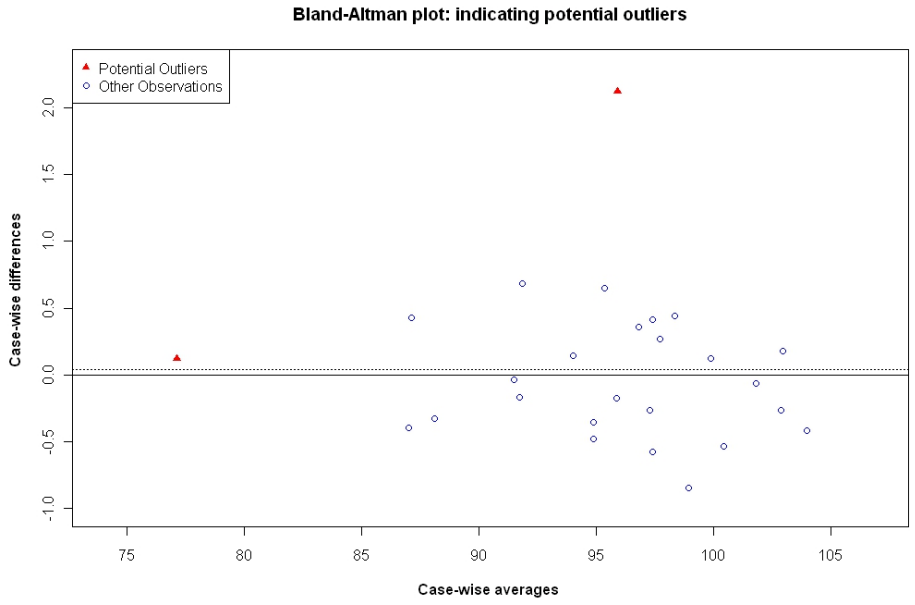

Figure 5: Bland-Altman Plot indicating the presence of Outliers

2.2 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.

2.3 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.