0.1 Bland-Altman Approach

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 paired sample $t-$test, correlation coefficients or simple linear regression. Simple linear regression is unsuitable for method comparison studies because of the required assumption that one variable is measured without error. In comparing two methods, both methods are assume to have attendant random error.

Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyzes and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be 'explained to non-statisticians'.

Notwithstanding previous remarks about linear 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 should also be shown, as it is necessary to give the correct interpretation of how both methods compare. In the case of good agreement, the observations would be distributed closely along the line of equality. A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that there is an inter-method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

Dewitte et al. (2002) notes that scatter plots were very seldom presented in the Annals of Clinical Biochemistry. This apparently results from the fact that the 'In-

structions for Authors' dissuade the use of regression analysis, which conventionally is accompanied by a scatter plot.

0.1.1 Bland-Altman plots

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, \ldots, n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \ldots, n$).

Altman and Bland (1983) proposes a scatterplot of the case-wise averages and differences of two methods of measurement. This scatterplot has since become widely known as the Bland-Altman plot. 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."

The case wise-averages capture several aspects of the data, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. Bland and Altman (1986) cautions that it would be the difference against either measurement value instead of their average, as the difference relates to both value. This methodology has proved very popular, and the Bland-Altman plots is widely regarded as powerful graphical methodology for making a visual assessment

of the data.

The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$. This inter-method bias is represented with a line on the Bland-Altman plot. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are also particularly relevant. The variances around this bias is estimated by the standard deviation of these differences S_d.

0.1.2 Bland-Altman plots for the Grubbs data

In the case of the Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. 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.

The Bland-Altman plot for comparing the 'Fotobalk' and 'Counter' methods, which shall henceforth be referred to as the 'F vs C' comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.8	793.2	0.6	793.5
2	793.1	793.3	-0.2	793.2
3	792.4	792.6	-0.2	792.5
4	794.0	793.8	0.2	793.9
5	791.4	791.6	-0.2	791.5
6	792.4	791.6	0.8	792.0
7	791.7	791.6	0.1	791.6
8	792.3	792.4	-0.1	792.3
9	789.6	788.5	1.1	789.0
10	794.4	794.7	-0.3	794.5
11	790.9	791.3	-0.4	791.1
12	793.5	793.5	0.0	793.5

Table 2: Fotobalk and Terma methods: differences and averages.

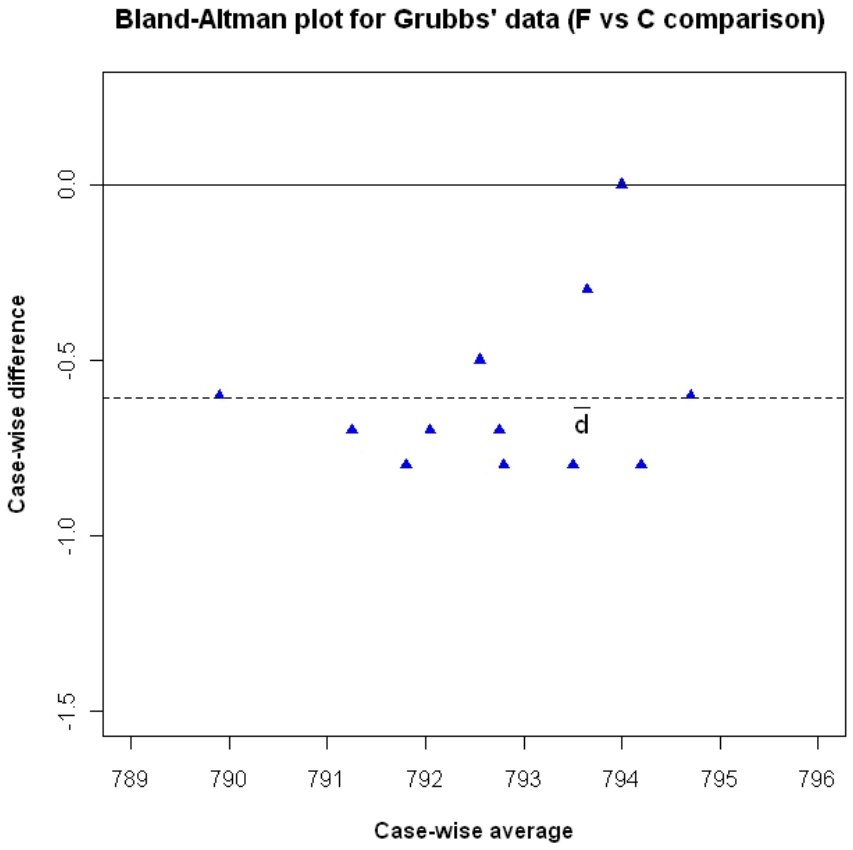

Figure 1: Bland-Altman plot For Fotobalk and Counter methods.

In Figure 1.3 Bland-Altman plots for the 'F vs C' and 'F vs T' comparisons are shown, where 'F vs T' refers to the comparison of the 'Fotobalk' and 'Terma' methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons. By inspection, there exists a larger inter-method bias in the 'F vs C' comparison than in the 'F vs T' comparison. Conversely there appears to be less precision in 'F vs T' comparison, as indicated by the greater dispersion of covariates.

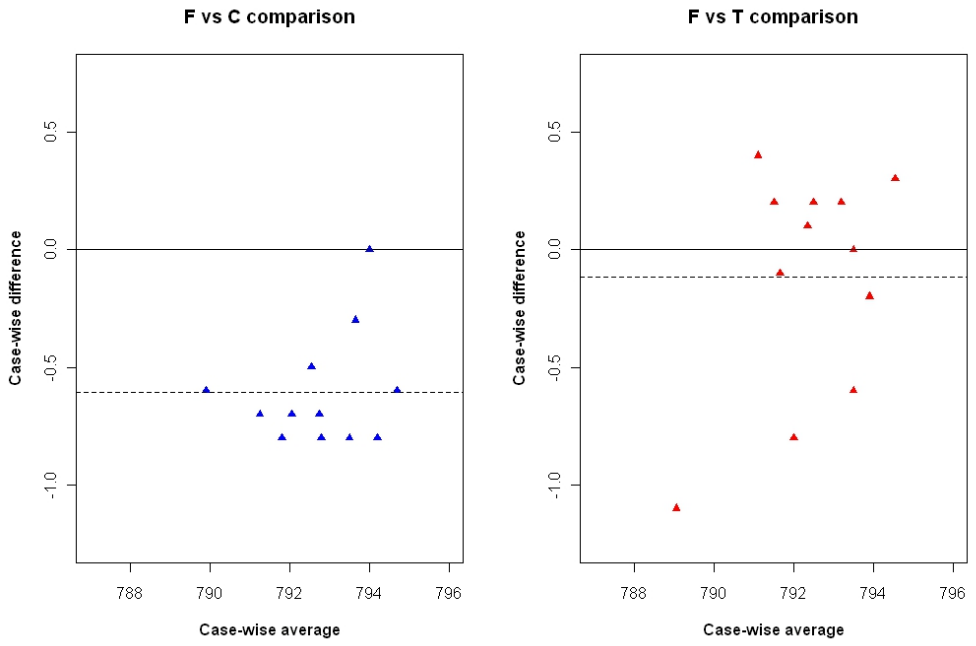

Figure 2: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

0.1.3 Prevalence of the Bland-Altman plot

Bland and Altman (1986), which further develops the Bland-Altman approach, was found to be the sixth most cited paper of all time by Ryan and Woodall (2005). Dewitte et al. (2002) describes the rate at which prevalence of the Bland-Altman plot has developed in scientific literature, by examining all articles in the journal 'Clinical Chemistry' between 1995 and 2001. This study concluded that use of the BlandAltman plot increased over the years, from 8% in 1995 to 14% in 1996, and 3136% in 2002.

The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (Hollis, 1996). Furthermore O'Brien et al. (1990) recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

0.1.4 Adverse features

Estimates for inter-method bias and variance of differences are only meaningful if there is uniform inter-bias and variability throughout the range of measurements. Fulfilment of these assumptions can be checked by visual inspection of the plot.The prototype Bland-Altman plots depicted in Figures 1.4, 1.5 and 1.6 are derived from simulated data, for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias, and is defined by Ludbrook (1997) as meaning that 'one method gives values that are higher (or lower) than those from the other by an amount that is proportional to the level of the measured variable'. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (Benjamini and Hochberg, 1995) test, should be also be used.

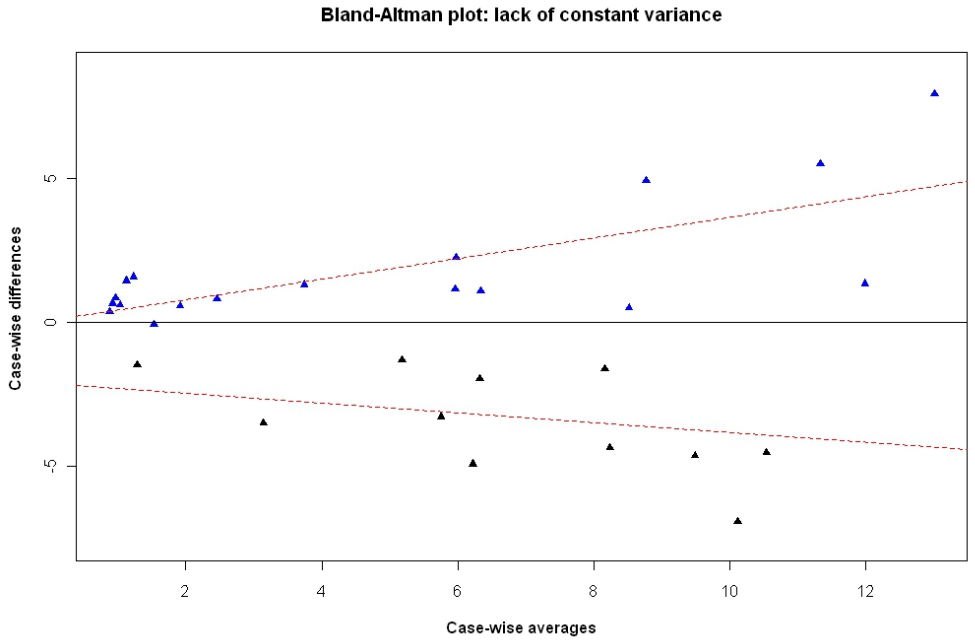

Figure 3: Bland-Altman plot demonstrating the increase of variance over the range.

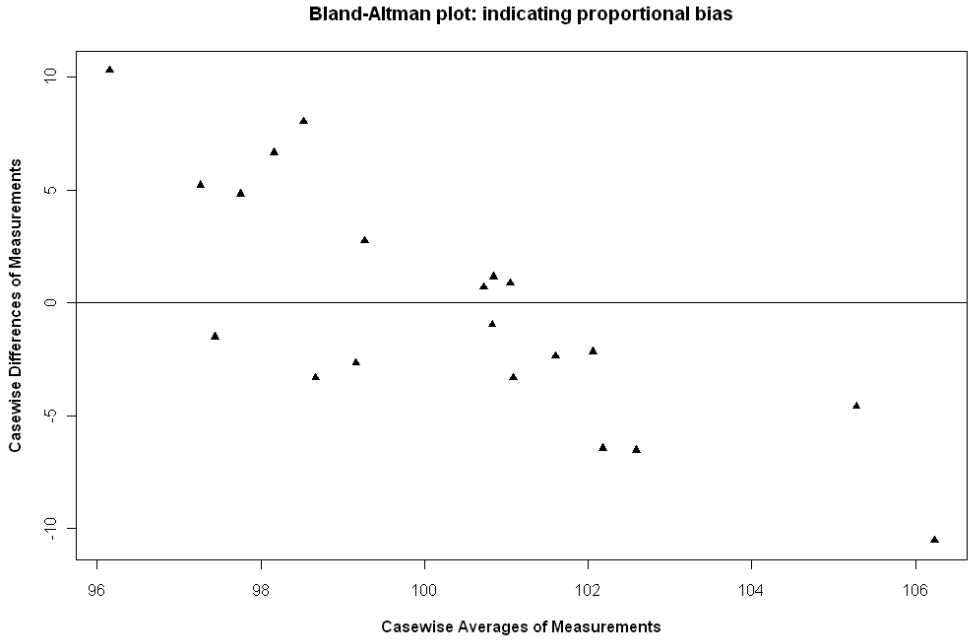

Figure 4: Bland-Altman plot indicating the presence of proportional bias.

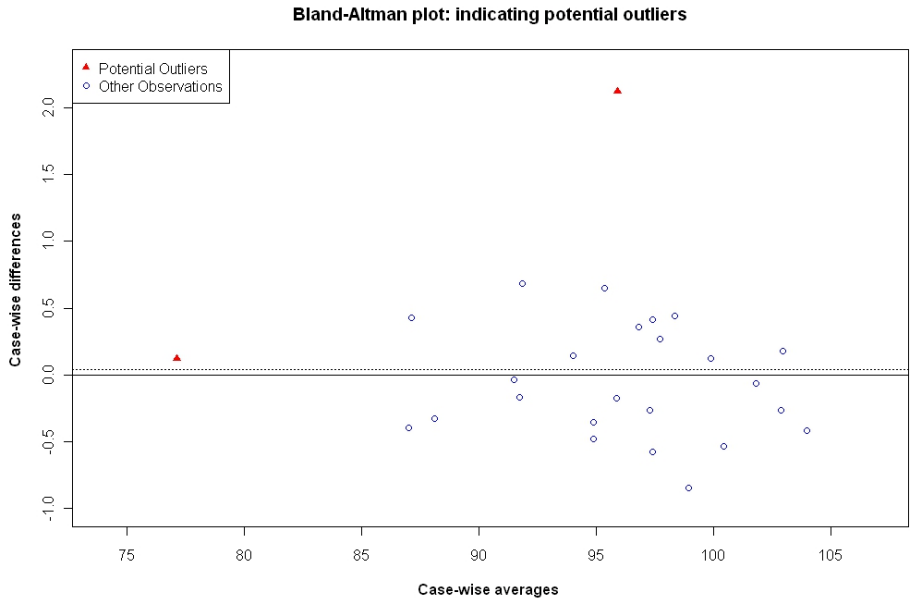

Figure 5: Bland-Altman plot indicating the presence of potential outliers.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. Bland and Altman (1999) do not recommend excluding outliers from analyzes, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that 'we usually find that this method of analysis is not too sensitive to one or two large outlying differences'. Figure 1.6 demonstrates how the Bland-Altman plot can be used to visually inspect the presence of potential outliers.

As a complement to the Bland-Altman plot, Bartko (1994) proposes the use of a bivariate confidence ellipse, constructed for a predetermined level. Altman (1978) provides the relevant calculations for the ellipse. This ellipse is intended as a visual guidelines for the scatter plot, for detecting outliers and to assess the within- and between-subject variances.

The minor axis relates to the between subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other. Consequently Bartko's ellipse provides a visual aid to determining the relationship between variances. If var(a) is greater than var(d), the orientation of the ellipse is horizontal. Conversely if var(a) is less than var(d), the orientation of the ellipse is vertical.

The Bland-Altman plot for the Grubbs data, complemented by Bartko's ellipse, is depicted in Figure 1.7. The fourth observation is shown to be outside the bounds of the ellipse, indicating that it is a potential outlier.

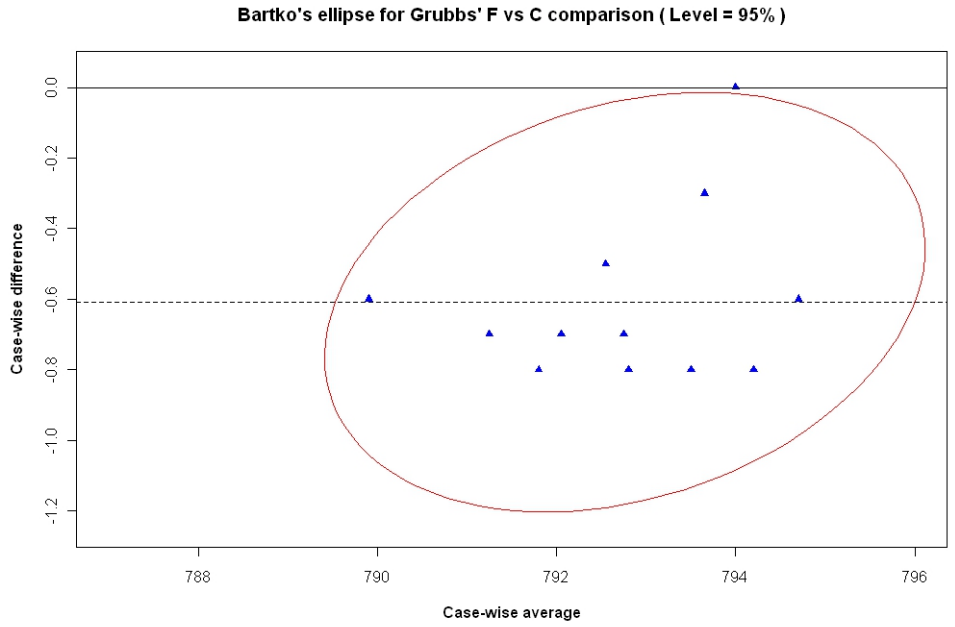

Figure 6: Bartko's Ellipse For Grubbs' Data.

The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko's ellipse. A covariate is added to the 'F vs C' comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Table 1.8 depicts a 95% confidence ellipse for this manipulated

data set. By inspection of the confidence interval, a conclusion would be reached that this extra covariate is an outlier, in spite of the fact that this observation is wholly consistent with the conclusion of the Bland-Altman plot.

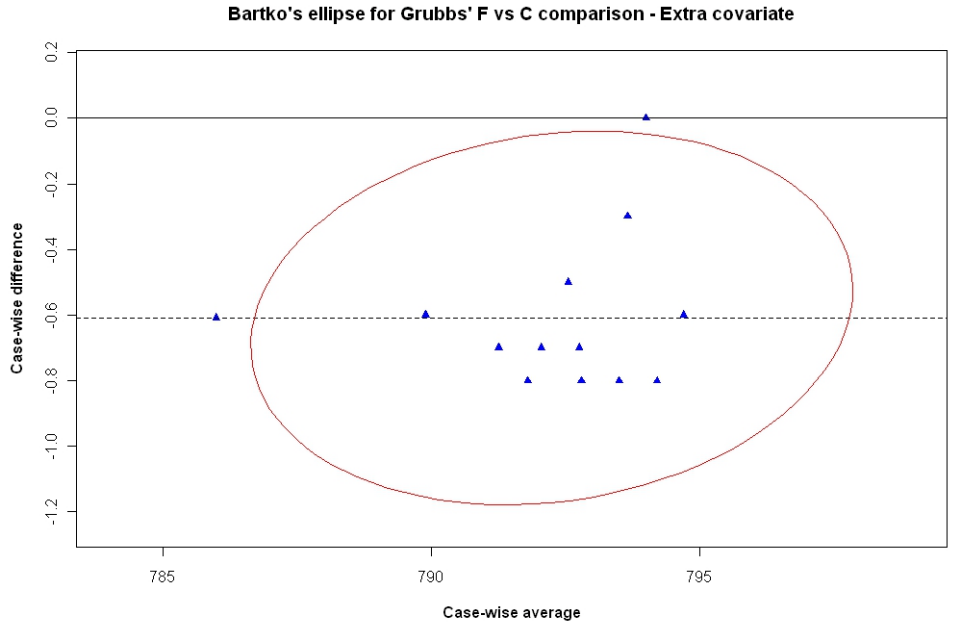

Figure 7: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

Importantly, outlier classification must be informed by the logic of the data's formulation. In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Any observation should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra covariate. Conversely, the fourth observation, from the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

In classifying whether a observation from a univariate data set is an outlier, many formal tests are available, such as the Grubbs test for outliers. In assessing whether a covariate in a Bland-Altman plot is an outlier, this test is useful when applied to the case-wise difference values treated as a univariate data set. The null hypothesis of the

Grubbs test procedure is the absence of any outliers in the data set. Conversely, the alternative hypotheses is that there is at least one outlier present.

The test statistic for the Grubbs test (G) is the largest absolute deviation from the sample mean divided by the standard deviation of the differences,

$$G = \max_{i=1,\ldots,n} \frac{\left|d_i - \bar{d}\right|}{S_d}.$$

For the 'F vs C' comparison it is the fourth observation gives rise to the test statistic, $G = 3.64$. The critical value is calculated using Student's t distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}.$$

For this test $U = 0.75$. The conclusion of this test is that the fourth observation in the 'F vs C' comparison is an outlier, with $p-$value $= 0.003$, according with the previous result using Bartko's ellipse.

0.2 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.

0.2.1 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.

0.2.2 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.

0.3 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.

Chapter 1

The Bland Altman Plot

1.1 Bland Altman Plots

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, 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 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

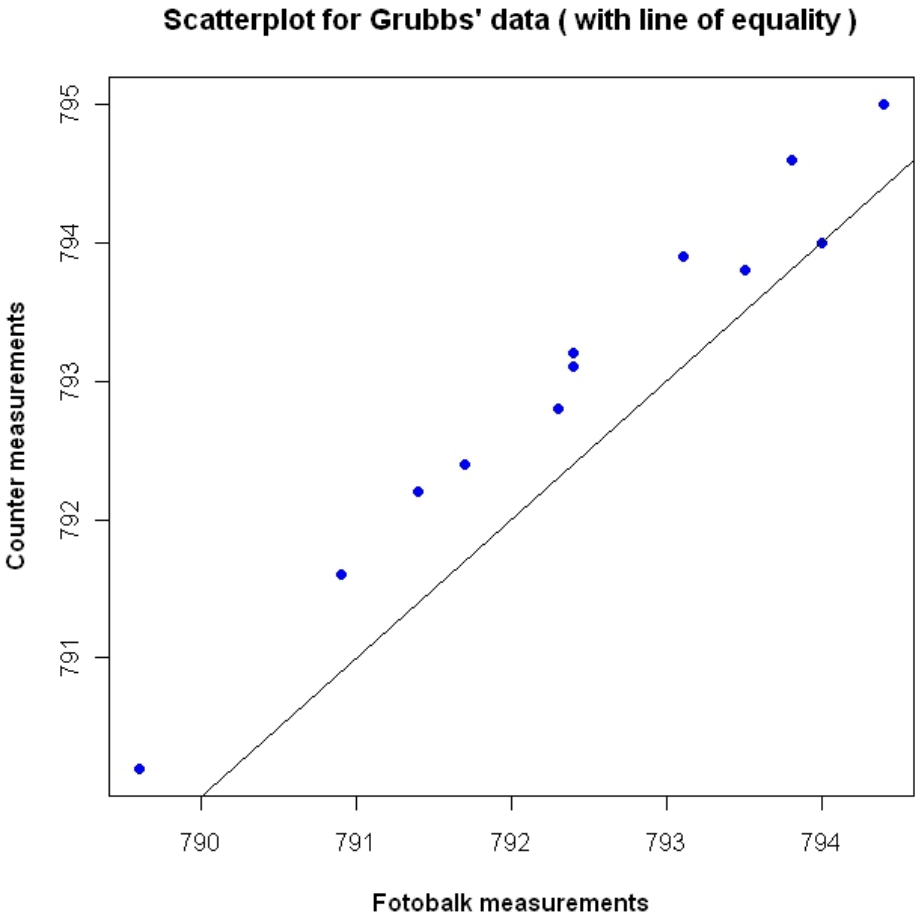

Figure 1.1: Scatter plot For Fotobalk and Counter Methods

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.

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

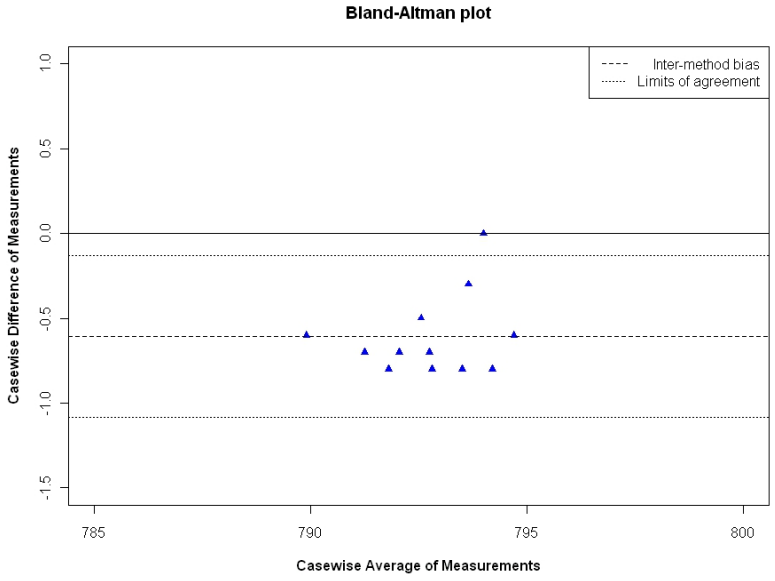

Figure 1.2: Bland Altman Plot For Fotobalk and Counter Methods

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.

1.1.1 Inspecting the Data

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

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 (Ludbrook, 1997).

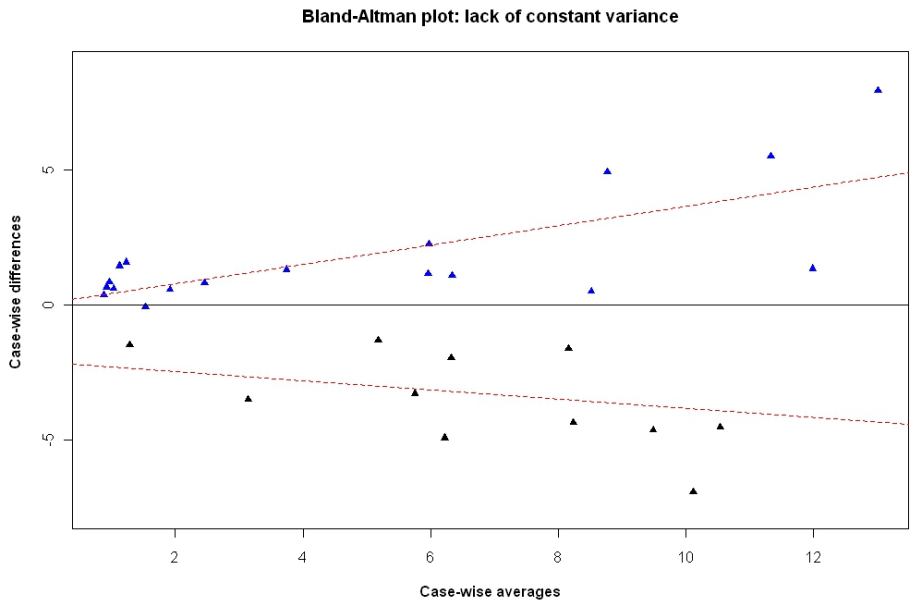

Figure 1.3: Bland-Altman Plot demonstrating the increase of variance over the range

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.

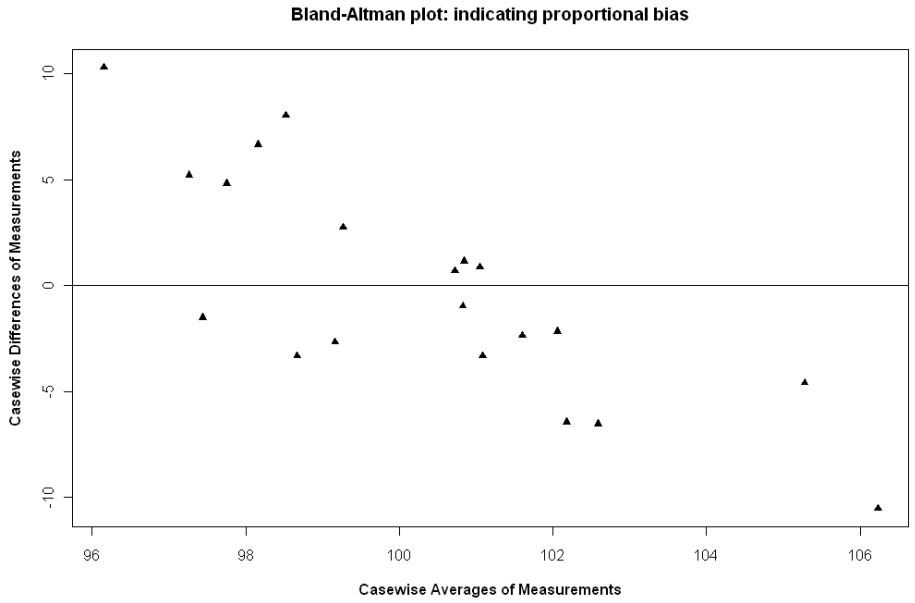

Figure 1.4: Bland-Altman Plot indicating the presence of proportional bias

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.

Bland and Altman (1999) 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.(Bland and Altman, 1999) states that *"We usually find that this method of analysis is not too sensitive to one or two large outlying differences."*

1.1.2 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 .*

1.1.3 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.

1.1.4 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.

1.1.5 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.

1.1.6 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.

1.2 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.

$$D_i = X_i - Y_i \tag{1.1}$$

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.

1.3 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 paired sample t-test, correlation coefficients or simple linear regression. Statisticians Martin Bland and Douglas 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). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, 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 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 1.1. Visual inspection 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 case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, ..n$ on the same subject should be calculated,

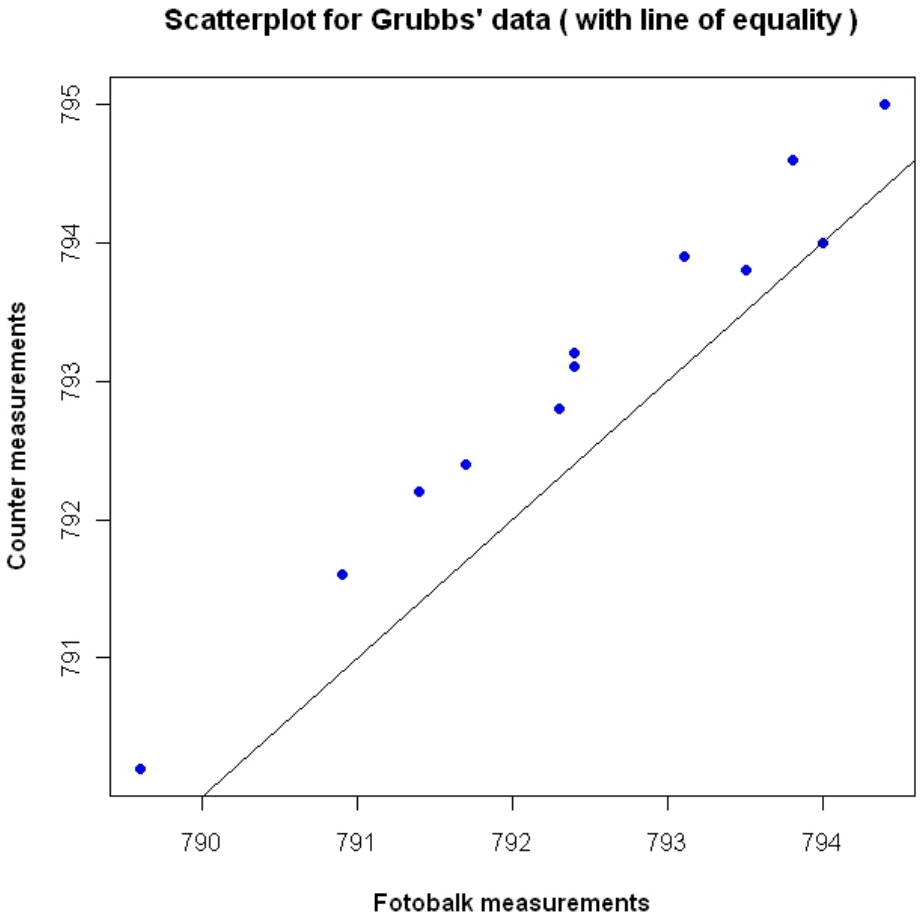

Figure 1.5: Scatter plot For Fotobalk and Counter Methods.

and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, ..n$). These differences and averages are then plotted. This methodology, now commonly known as the 'Bland-Altman Plot', has proved very successful. Bland and Altman (1986), which further develops the methodology, was found to be the sixth most cited paper of all time by the Ryan and Woodall (2005). Dewitte et al. (2002) also commented on the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (Hollis, 1996). Furthermore O'Brien et al. (1990) recommend its use in papers pertaining to

method comparison studies for the journal of the British Hypertension Society.

The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$. The variances around this bias is estimated by the standard deviation of the differences $S(d)$. This inter-method bias is represented with a line on the Bland-Altman plot. These estimates are only meaningful if there is uniform inter-bias and variability throughout the range of measurements, which can be checked by visual inspection of the plot. In the case of Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. 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.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.2: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.80	793.20	0.60	793.50
2	793.10	793.30	-0.20	793.20
3	792.40	792.60	-0.20	792.50
4	794.00	793.80	0.20	793.90
5	791.40	791.60	-0.20	791.50
6	792.40	791.60	0.80	792.00
7	791.70	791.60	0.10	791.65
8	792.30	792.40	-0.10	792.35
9	789.60	788.50	1.10	789.05
10	794.40	794.70	-0.30	794.55
11	790.90	791.30	-0.40	791.10
12	793.50	793.50	0.00	793.50

Table 1.3: Fotobalk and Terma methods: differences and averages.

1.3.1 Using Bland-Altman Plots

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

The Bland-Altman plot is simply a scatterplot of the case-wise averages and differences of two methods of measurement. As the objective of the Bland-Altman plot is to advise on the agreement of two methods, it is the case-wise differences that are particularly. Later it will be shown that case-wise differences are the sole component of the next part of the methodology, the limits of agreement.

For creating plots, the case wise-averages fulfil several functions, such as expressing the range over which the values were taken, and assessing whether the assumptions of constant variance holds. Case-wise averages also allow the case-wise differences to be presented on a two-dimensional plot, with better data visualization qualities than a one dimensional plot. Bland and Altman (1986) cautions that it would be the difference against either measurement value instead of their average , as the difference relates to both value.

The Bland-Altman plot for comparing the 'Fotobalk' and 'Counter' methods, which shall henceforth be referred to as the 'F vs C' comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.

In Figure 1.3 Bland-Altman plots for the 'F vs C' and 'F vs T' comparisons are shown, where 'F vs T' refers to the comparison of the 'Fotobalk' and 'Terma' methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons.

By inspection, there exists a larger inter-method bias in the 'F vs C' comparison than in the 'F vs T' comparison. Conversely there appears to be less precision in'F vs T' comparison, as indicated by the greater dispersion of co-variates.

Figures 1.4, 1.5 and 1.6 are three prototype Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias. In both Figures 1.4 and 1.5, the assumptions

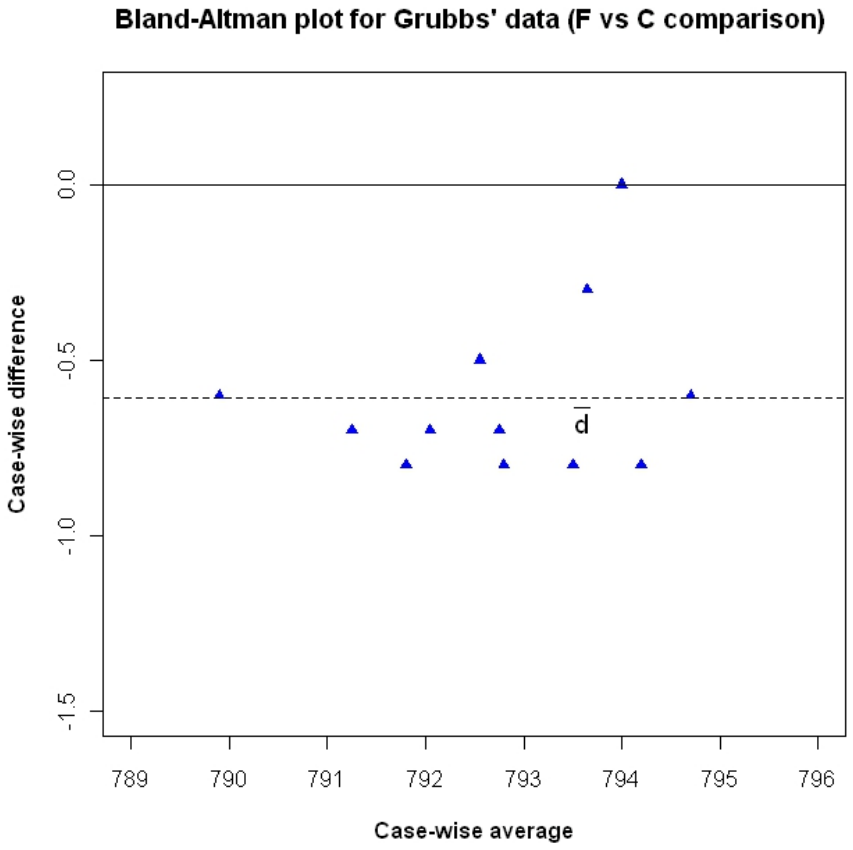

Figure 1.6: Bland-Altman plot For Fotobalk and Counter methods.

necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (Benjamini and Hochberg, 1995) test, should be also be used.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses

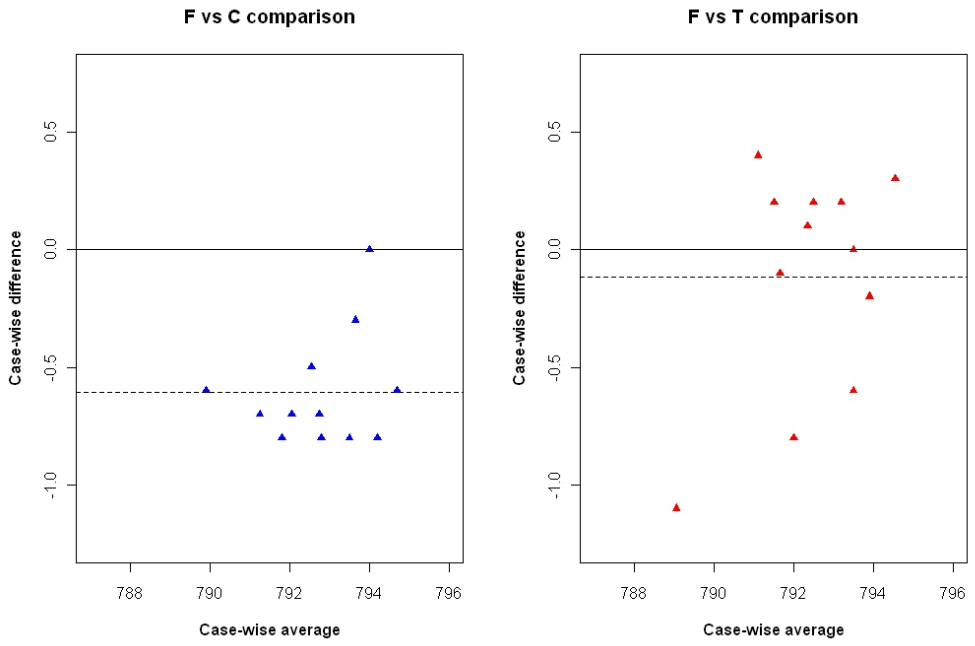

Figure 1.7: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

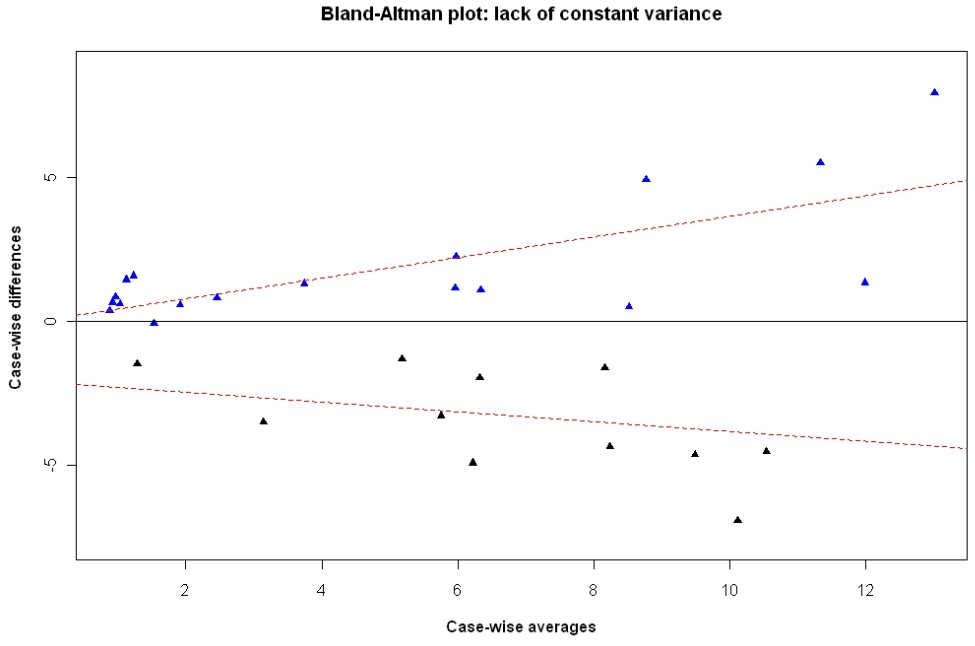

Figure 1.8: Bland-Altman plot demonstrating the increase of variance over the range.

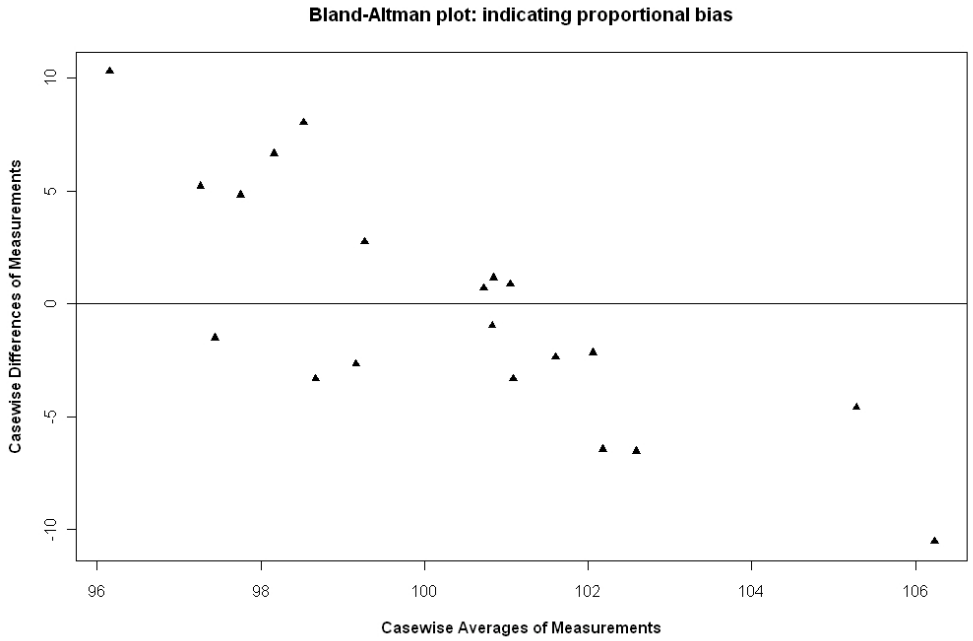

Figure 1.9: Bland-Altman plot indicating the presence of proportional bias.

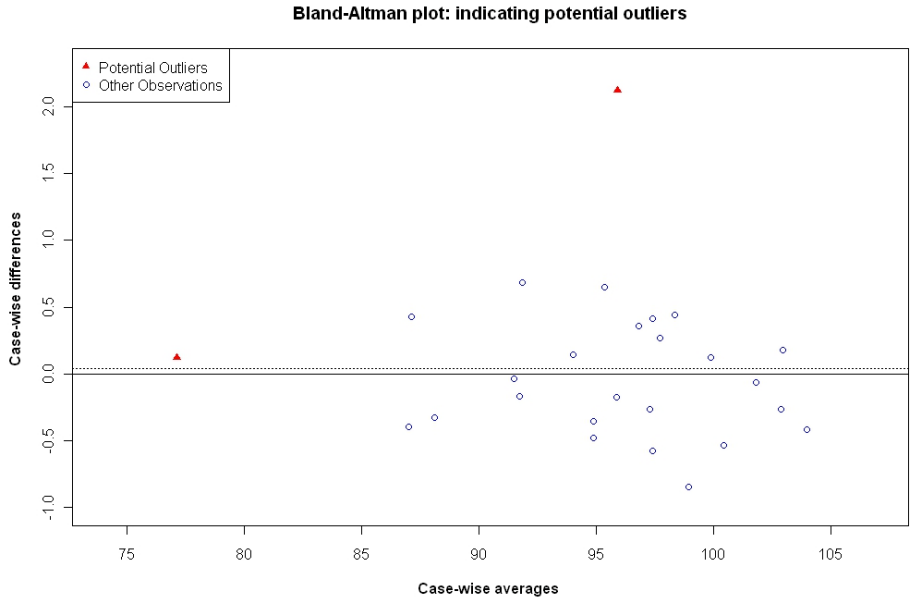

Figure 1.10: Bland-Altman plot indicating the presence of potential outliers.

suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. Classification of outliers can be determined with numerous established approaches, such as the Grubb's test, but always classification must be informed by the logic of the data's formulation. Figure 1.6 is a Bland-Altman plot with two potential outliers.

Bland and Altman (1999) do not recommend excluding outliers from analyses, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that 'we usually find that this method of analysis is not too sensitive to one or two large outlying differences'.

In classifying whether a observation from a univariate data set is an outlier, Grubbs' outlier test is widely used. In assessing whether a co-variate in a Bland-Altman plot is an outlier, this test is useful when applied to the difference values treated as a univariate data set. For Grubbs' data, this outlier test is carried out on the differences, yielding the following results.

The null and alternative hypotheses is the absence and presence of at least one outlier respectively. Grubbs' outlier test statistic G is the largest absolute deviation from the sample mean divided by the standard deviation of the differences. For the 'F vs C' comparison, $G = 3.6403$. The critical value is calculated using Student's t distribution and the sample size,

$$U = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2+t^2_{\alpha/(2n),n-2}}}. \tag{1.2}$$

For this test $U = 0.7501$. The conclusion of this test is that the fourth observation in the 'F vs C' comparison is an outlier, with $p - value = 0.002799$.

As a complement to the Bland-Altman plot, Bartko (1994) proposes the use of a bivariate confidence ellipse, constructed for a predetermined level.

The minor axis relates to the between subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other.Altman (1978) provides the relevant calculations for the ellipse. Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers (furthermore Bartko (1994) proposes formal testing procedures, that shall be discussed in due course). Inspection of Figure 1.7 shows that the fourth observation is outside the bounds of the ellipse, concurring with the conclusion that it is an outlier.

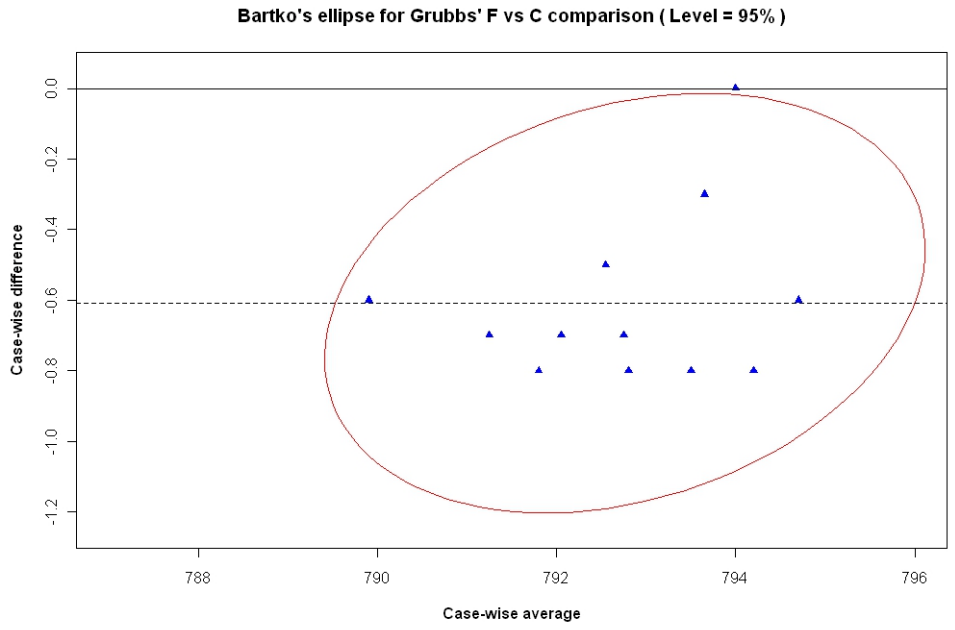

Figure 1.11: Bartko's Ellipse For Grubbs' Data.

The limitations of using bivariate approaches to outlier detection in the Bland-Altman plot can demonstrated using Bartko's ellipse. A co-variate is added to the 'F vs C' comparison that has a difference value equal to the inter-method bias, and an average value that markedly deviates from the rest of the average values in the comparison, i.e. 786. Table 1.8 depicts a 95% confidence ellipse for this enhanced data set. By inspection of the confidence interval, a conclusion would be reached that this extra co-variate is an outlier, in spite of the fact that this observation is consistent with

the intended conclusion of the Bland-Altman plot.

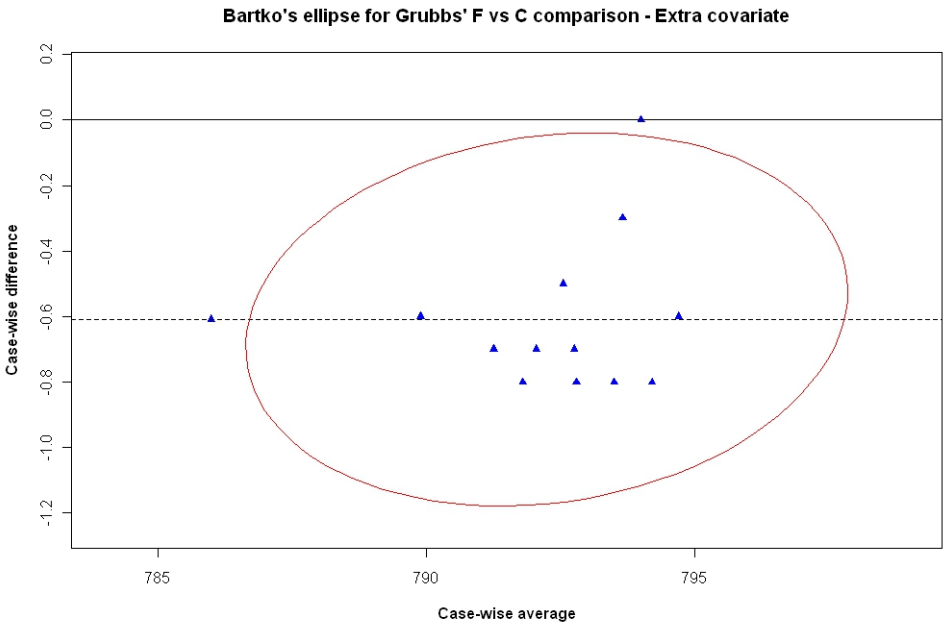

Figure 1.12: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

In the Bland-Altman plot, the horizontal displacement of any observation is supported by two independent measurements. Any observation should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra co-variate. Conversely, the fourth observation, from the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

Bartko's ellipse provides a visual aid to determining the relationship between variances. If $\text{var}(a_i)$ is greater than $\text{var}(d_i)$, the orientation of the ellipse is horizontal. Conversely if $\text{var}(a_i)$ is less than $\text{var}(d_i)$, the orientation of the ellipse is vertical.

1.4 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."

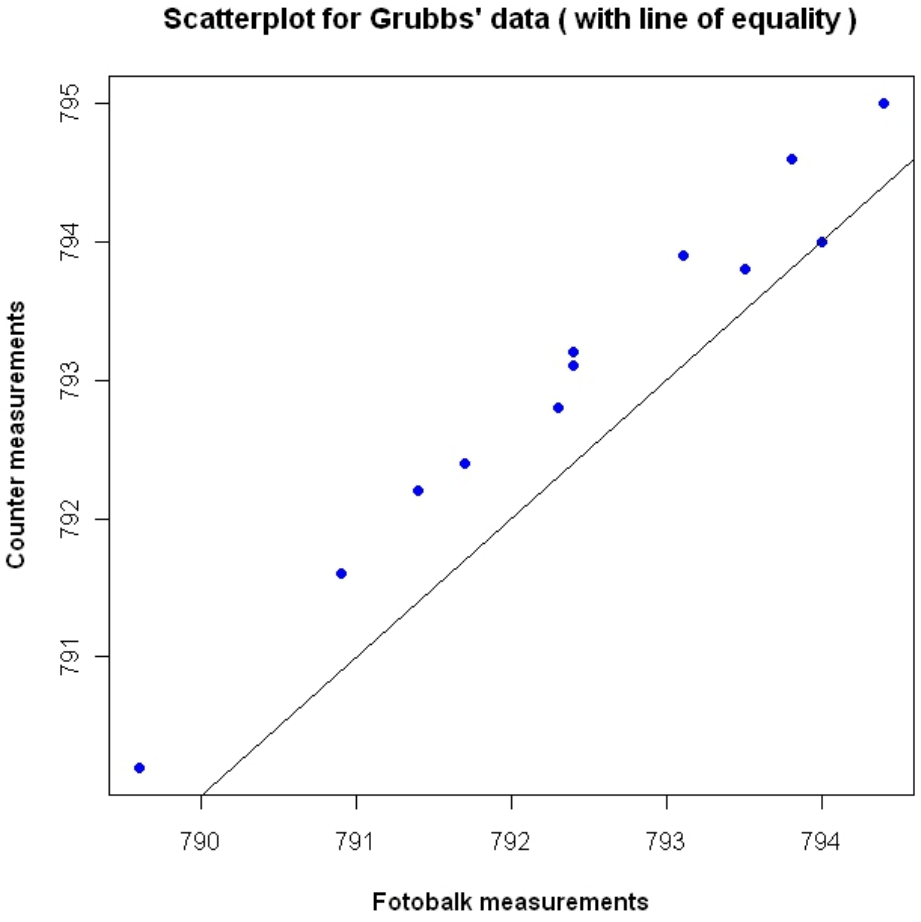

Figure 1.13: Scatter plot For Fotobalk and Counter Methods.

In light of shortcomings associated with scatterplots, Altman and Bland (1983)

recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, ..n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, ..n$). These differences and averages are then plotted. This methodology, now commonly known as the 'Bland-Altman Plot', has proved very successful. Bland and Altman (1986), which further develops the methodology, was found to be the sixth most cited paper of all time by the Ryan and Woodall (2005). Dewitte et al. (2002) also commented on the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (Hollis, 1996). Furthermore O'Brien et al. (1990) recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$. The variances around this bias is estimated by the standard deviation of the differences $S(d)$. This inter-method bias is represented with a line on the Bland-Altman plot. These estimates are only meaningful if there is uniform inter-bias and variability throughout the range of measurements, which can be checked by visual inspection of the plot. In the case of Grubbs data the inter-method bias is -0.61 metres per second, and is indicated by the dashed line on Figure 1.2. 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.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.4: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.80	793.20	0.60	793.50
2	793.10	793.30	-0.20	793.20
3	792.40	792.60	-0.20	792.50
4	794.00	793.80	0.20	793.90
5	791.40	791.60	-0.20	791.50
6	792.40	791.60	0.80	792.00
7	791.70	791.60	0.10	791.65
8	792.30	792.40	-0.10	792.35
9	789.60	788.50	1.10	789.05
10	794.40	794.70	-0.30	794.55
11	790.90	791.30	-0.40	791.10
12	793.50	793.50	0.00	793.50

Table 1.5: Fotobalk and Terma methods: differences and averages.

1.4.1 Discussion

I have here proposed a simple twist to the results from regression of the differences on the sums in the case of a linear relationship between two methods of measurement. It is

consistent with the obvious underlying model, and exploits the fact that although the parameters of the model cannot be estimated, those functions of the parameters that are needed for creating predictions can be estimated. The prediction limits provided have the attractive property that if the prediction line with limits is drawn in a coordinate system, the chart will apply in both ways; hence, both the line and the limits are symmetric. Precisely as the prediction intervals derived from the classical LoA are in the case where the difference between methods is constant. The drawback is that the regression of the differences on the means ignores that the averages are correlated with the residuals (i.e. the error terms), and therefore gives biased estimates if the slope linking the two methods is far from 1 or the residual variances are very different. However, both of these are rather uncommon in method comparison studies, so the method proposed here is widely applicable. When considering LoA, the only feasible transformation is the log-transform, which gives LoA for the ratio of measurements, which is immediately understandable. If, for example, the measurements are fractions where some are close to either 0 or 1 a logit transform may be adequate.

LoA would then be for (log) odds-ratios, not very easily understood. For other more arbitrarily chosen transformation the situation may be even worse. But if a plot with conversion lines and limits are constructed, then the plot is readily back-transformed to the original scale for practical use.

1.4.2 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.

1.4.3 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

1.5 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.

1.6 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.

1.7 Gold Standard

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

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.

Bibliography

Altman, D. (1978). Plotting probability ellipses. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 27*(3), 347–349.

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.

Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine 13*, 737–745.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B 57*(1), 289–300.

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.

Hollis, S. (1996). Analysis of method comparison studies. *Ann Clin Biochem 33*, 1–4.

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.

Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics 32*(5), 461 – 474.