Contents

Chapter 1

Current Methodologies: Bland-Altman Analysis

1.1 Bland-Altman Methodology

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. Statisticians Martin Bland and Douglas Altman recognized the inadequacies of several analyzes and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983), while instead recommending the use of graphical techniques to assess agreement.

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 1.2.1). These differences and averages are then plotted (Figure 1.2.2).

In 1986 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes. Principally their method is calculating, for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference d_i and mean a_i :

case-wise differences of measurements of two methods $d_i = x_i - y_i,$ for $i = 1, 2, \ldots, n$, on the same subject should be calculated, and then the average of those measurements, $a_i = (x_i + y_i)/2$ for $i = 1, 2, \ldots, n$. An important requirement is that the two measurement methods use the same scale of measurement. Following a technique known as the Tukey mean-difference plot, as noted by Kozak and Wnuk (2014), Altman and Bland (1983) proposed that a_i should be plotted against d_i, a plot now widely known as the Bland-Altman plot, and motivated this plot as follows:

> "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 approach has proved very popular, and the Bland-Altman plots is widely regarded as powerful graphical tool 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, the individual case-wise differences are also particularly relevant. The variances around this bias is estimated by the standard deviation of these differences S_d. 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 (1986) 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.

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* (Altman and Bland, 1983).

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. However it is worth mentioning, as it is a simple, powerful and elegant technique that is often overlooked in method comparison studies. The identity plot is a simple scatter-plot approach of measurements for both methods on either axis, with the line of equality (the $X = Y$ line, i.e. the 45 degree line through the origin).

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. In the case of good agreement, the observations would be distributed closely along this line. However, they are not useful for a thorough examination of the data. This plot can gives the analyst a cursory examination of how well the measurement methods agree. In the case of good agreement, the covariates of the plot accord closely with the line of equality.

O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation. A scatter plot of the Grubbs data is shown in Figure 1.1.1. Visual inspection confirms the previous conclusion that inter-method bias is present, i.e. the Fotobalk device has a tendency to record a lower velocity.

1.2 Bland Altman Plots

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 1.2.2,

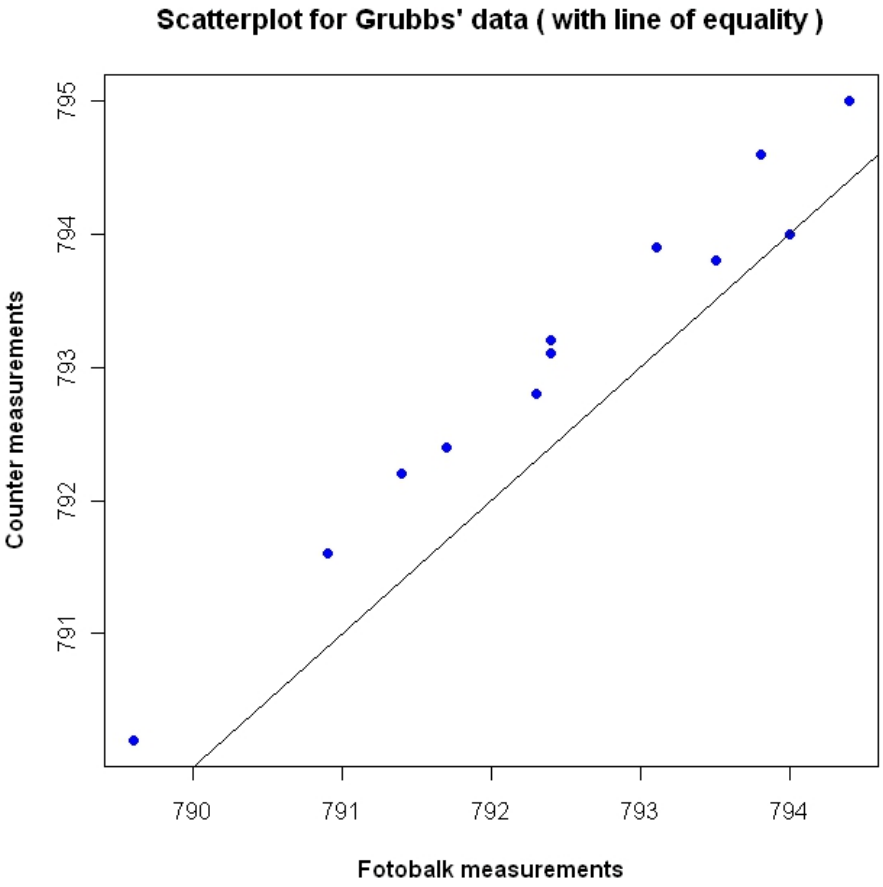

Figure 1.1.1: Scatter plot for Fotobalk and Counter methods.

using data from Table 1.2.1. The dashed line in Figure 1.2.2 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.

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.

1.2.1 Bland-Altman plots for the Grubbs data

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 1.2.1: 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 1.2.2: Fotobalk and Terma methods: Differences and Averages.

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'

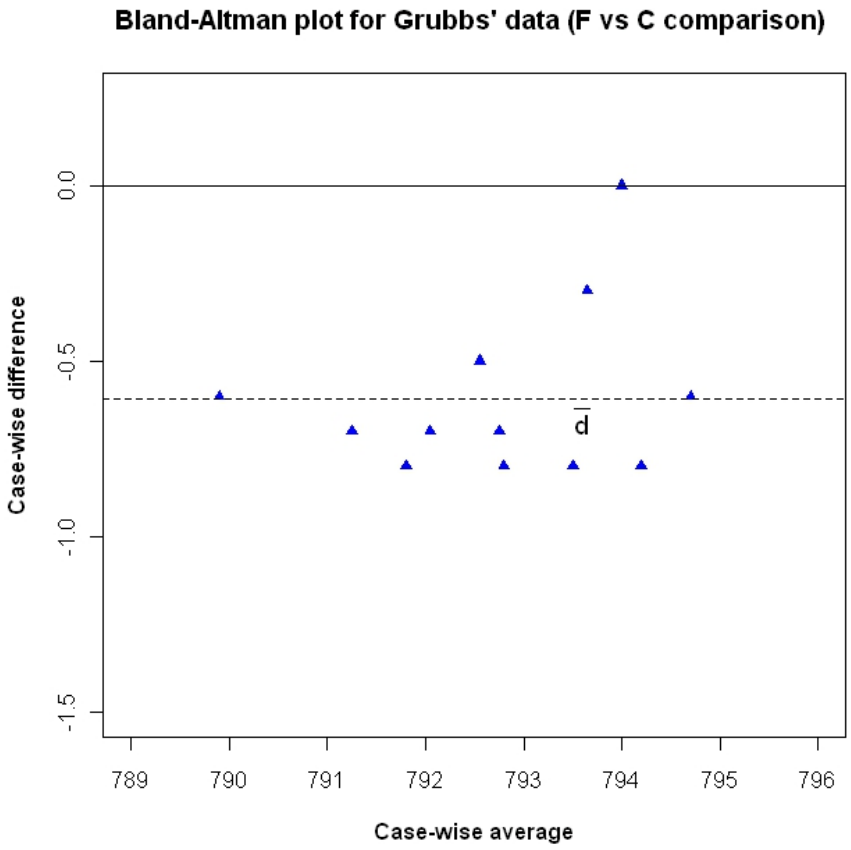

Figure 1.2.2: Bland-Altman plot For Fotobalk and Counter methods.

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.

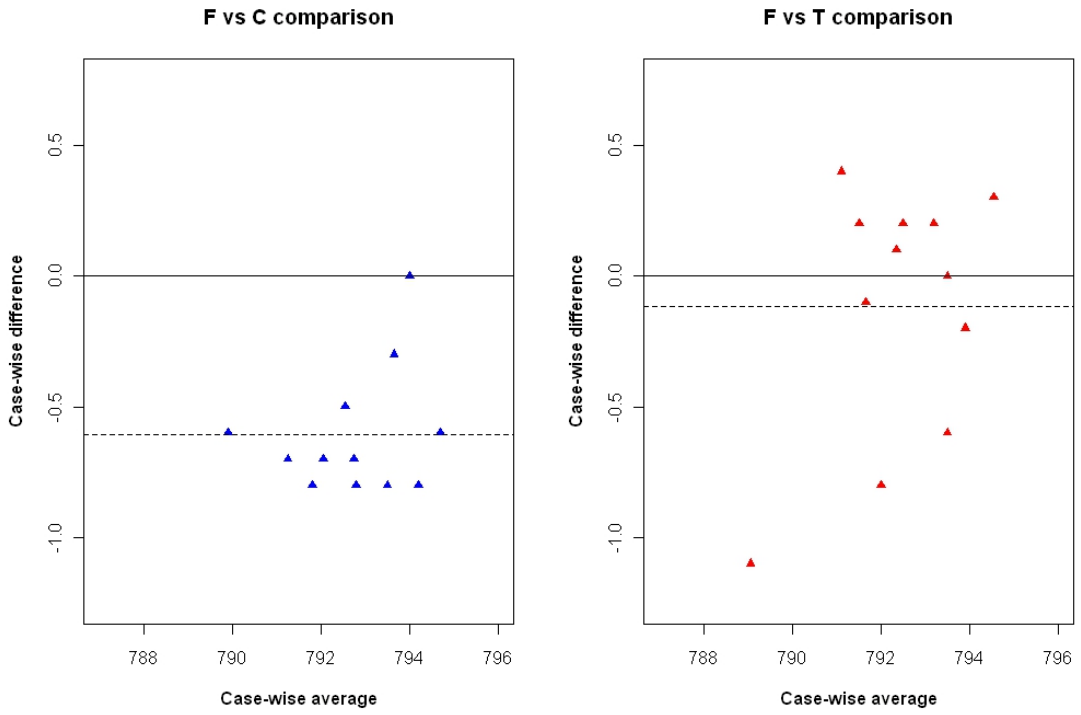

Figure 1.2.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

1.2.2 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."

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.

Figures 1.2.4, 1.2.5 and 1.2.6 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 1.2.4 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. 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 could be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (Benjamini and Hochberg, 1995) test, are advisable.

Figure 1.2.5 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'.

Both of these cases violate the assumptions necessary for further analysis using limits of agreement, which shall be discussed later.

1.2.3 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 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

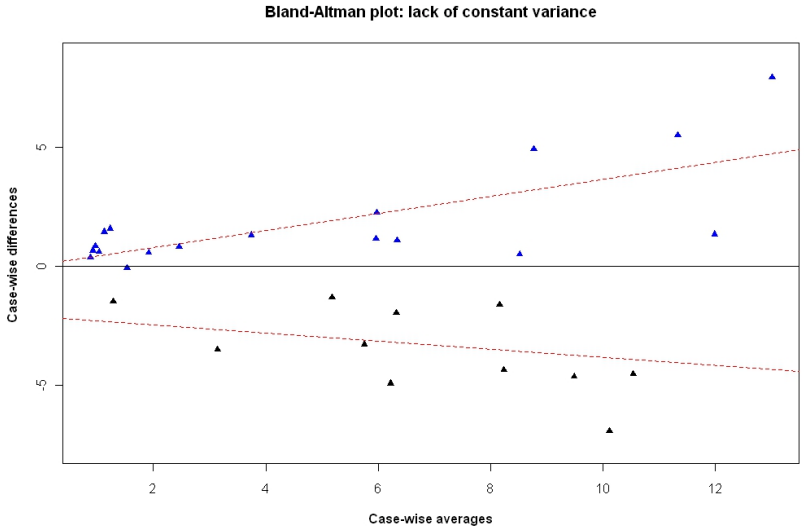

Figure 1.2.4: Bland-Altman Plot demonstrating the increase of variance over the range

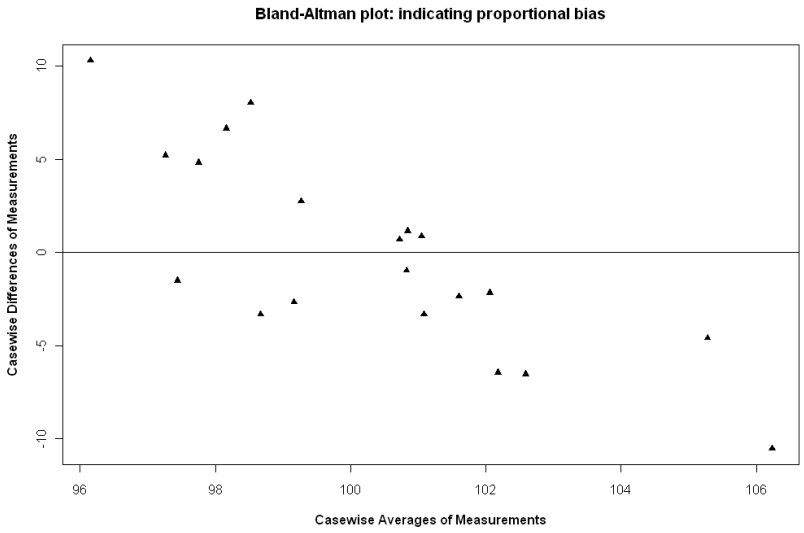

Figure 1.2.5: Bland-Altman Plot indicating the presence of proportional bias

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.

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 1.2.6 is a Bland Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively.

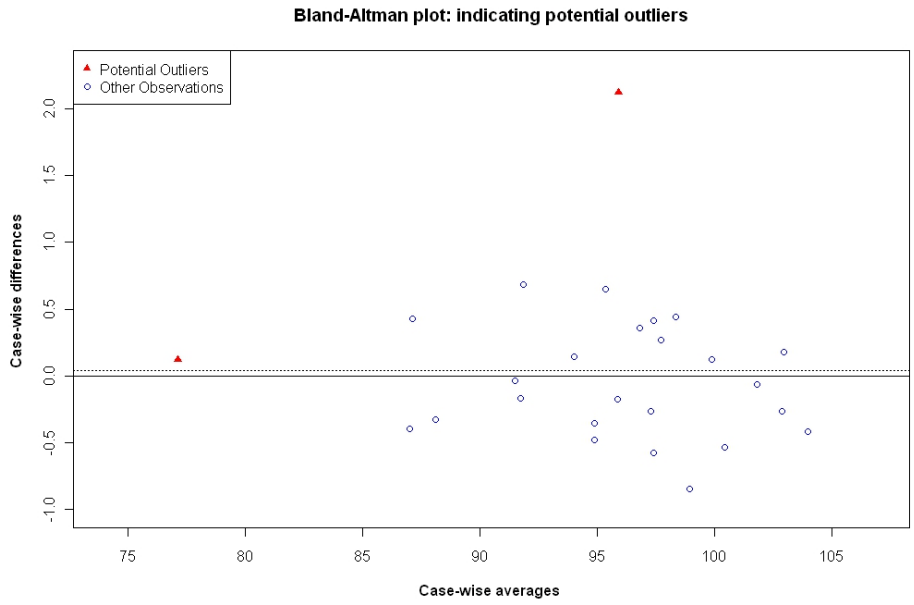

Figure 1.2.6: Bland-Altman Plot indicating the presence of Outliers

Importantly, outlier classification must be informed by the logic of the mechanism that produces the data.

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 1.2.6, 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. 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.

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."*

1.2.4 Detection of Outliers

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.

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?

1.2.5 Bartko's Ellipse

Bartko (1994) offers a graphical complement to the Bland-Altman plot, in the form of a bivariate confidence ellipse, constructed for a predetermined level.

As an enhancement on the Bland Altman Plot, Bartko (1994) has expounded a confidence ellipse for the covariates. Bartko (1994) proposes a bivariate confidence ellipse as a boundary for dispersion. The stated purpose is to 'amplify dispersion', which presumably is for the purposes of outlier detection. The orientation of the the ellipse is key to interpreting the results.

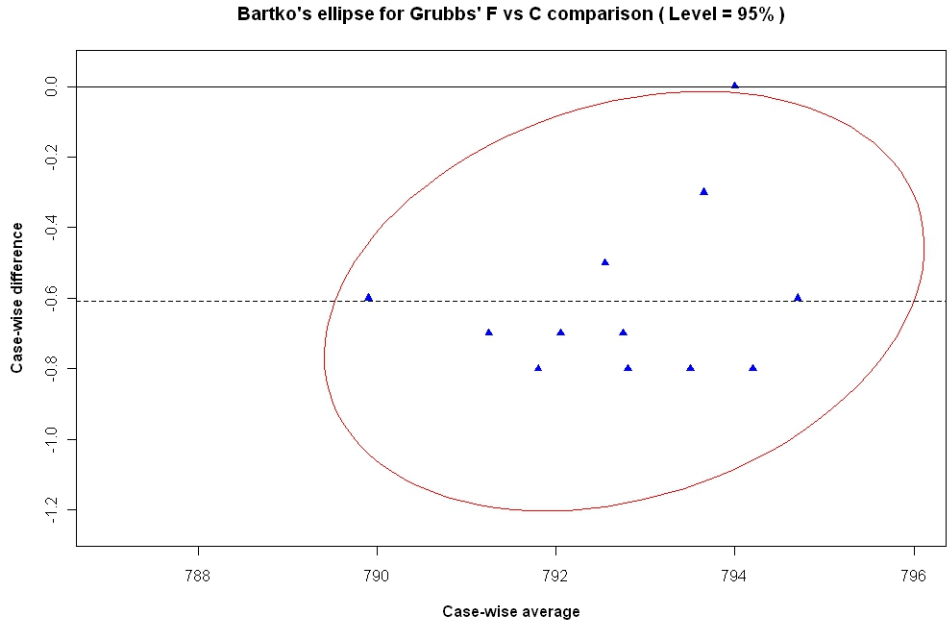

Figure 1.2.7: Bartko's Ellipse For Grubbs Data

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.

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, we would conclude that this extra covariate is an outlier, in spite of the fact that this observation is very close to the inter-method bias as determined by this approach.

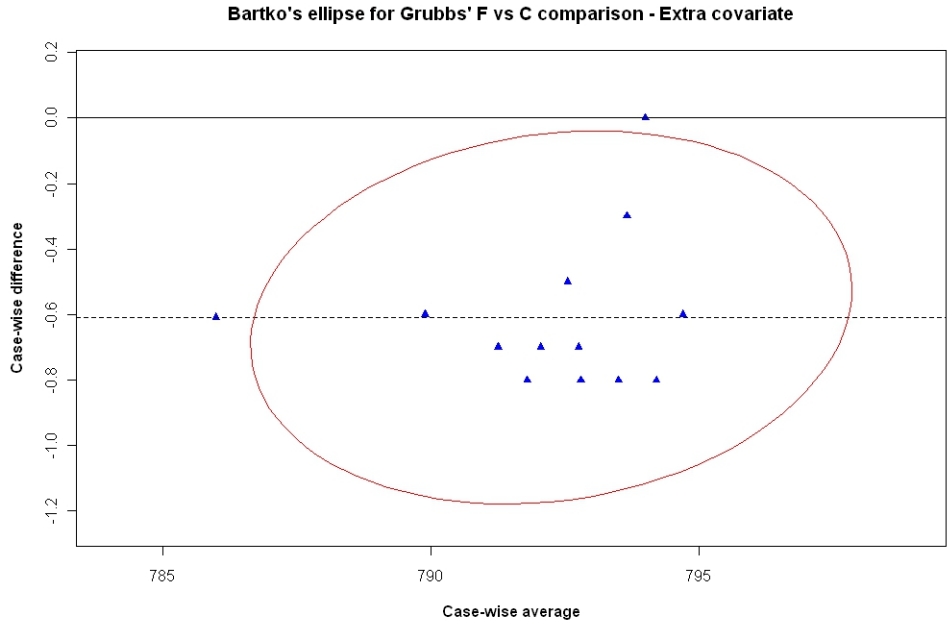

Figure 1.2.8: Bartko's Ellipse for Grubbs' data, with an extra covariate.

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

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.

In the Bland-Altman plot, the horizontal displacement of any point on the plot is supported by two independent measurements. Any point 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 point, from the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

Additionally Bartko's ellipse provides a visual aid to determining the relationship between variances. Furthermore, the ellipse provides a visual aid to determining the relationship between the variance of the means $Var(a_i)$ and the variance of the differences $Var(d_i)$. 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 more horizontal the ellipse, the greater the degree of agreement between the two methods being tested.

1.2.6 Grubbs' Test for Outliers

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.

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,

$$G = \max_{i=1,\ldots,n} \frac{\left|d_i - \bar{d}\right|}{S_d}. \tag{1.1}$$

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 $= 0.003$, in accordance with the previous result of Bartko's ellipse.

1.3 Limits of Agreement

As Bland and Altman (1986) point out this may not be the case. Bland and Altman advises on how to calculate of confidence intervals for the inter-method bias and the limits of agreement.

Bland and Altman (1999) note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as "being like a reference interval."

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process con-

trol. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine the limits, the mean and standard deviation, are not based on any sample used for an analysis, but on the process's historical values, a key difference with Bland-Altman limits of agreement.

Carstensen et al. (2008) regards the limits of agreement as a prediction interval for the difference between future measurements with the two methods on a new individual, but states that it does not fit the formal definition of a prediction interval, since the definition does not consider the errors in estimation of the parameters. Prediction intervals, which are often used in regression analysis, are estimates of an interval in which future observations will fall, with a certain probability, given what has already been observed. Carstensen et al. (2008) offers an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.975,n-1)} S_d \sqrt{1 + \frac{1}{n}} \tag{1.2}$$

where n is the number of subjects. Only for 61 or more subjects is there a quantile less than 2.

Luiz et al. (2003) describes limits of agreement as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence. Luiz et al. (2003) offers an alternative description of limits of agreement, this time as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence. Barnhart et al. (2007) describes them as a probability interval, and offers a clear description of how they should be used; '*if the absolute limit is less than an acceptable difference* d_0*, then the agreement between the two methods is deemed satisfactory*'.

Various other interpretations as to how limits of agreement should properly be defined.

The prevalence of contradictory definitions of what limits of agreement strictly are will inevitably attenuate the poor standard of reporting using limits of agreement, as

discussed by Mantha et al. (2000).

1.4 Limits of Agreement

A third element of the Bland-Altman methodology, an interval known as 'limits of agreement' is introduced in Bland and Altman (1986) (sometimes referred to in literature as 95% 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.

The Limits of agreement are intended to demonstrate the range in which 95% of the sample data should lie. Following basic principles of the normal probability disribution, the Limits of Agreement are centred on the average difference line (which indicates the inter-method bias) and are 1.96 times the standard deviation above and below the average difference line.

Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably. Bland and Altman (1986) refer to this as the 'equivalence' of two measurement methods. The specific purpose of the limits of agreement must be established clearly. Bland and Altman (1995) comment that the limits of agreement '*how far apart measurements by the two methods were likely to be for most individuals*', a definition echoed in their 1999 paper:

> "We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% limits of agreement. These values define the range within which most differences between measurements by the two methods will lie."

The limits of agreement (LoA) are computed by the following formula:

$$LoA = \bar{d} \pm 1.96 s_d$$

with $\bar{d}$ as the estimate of the inter method bias, s_d as the standard deviation of the differences and 1.96 is the 95% quantile for the standard normal distribution. (Some descriptions of the Bland-Altman plot use 2 standard deviations instead for simplicity.)

Importantly the authors recommend prior determination of what would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion. However Mantha et al. (2000) highlight inadequacies in the correct application of limits of agreement, resulting in contradictory estimates of limits of agreement in various papers.

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. Importantly the authors recommend prior determination of what would and would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion.

> "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 comparison and to choose the sample size (Bland and Altman, 1986)".

However Mantha et al. (2000) highlights inadequacies in the correct application of limits of agreement, resulting in contradictory estimates for limits of agreement in various papers.

For the Grubbs 'F vs C' comparison, these limits of agreement are calculated as -0.132 for the upper bound, and -1.08 for the lower bound. Figure **??** shows the resultant Bland-Altman plot, with the limits of agreement shown in dashed lines.

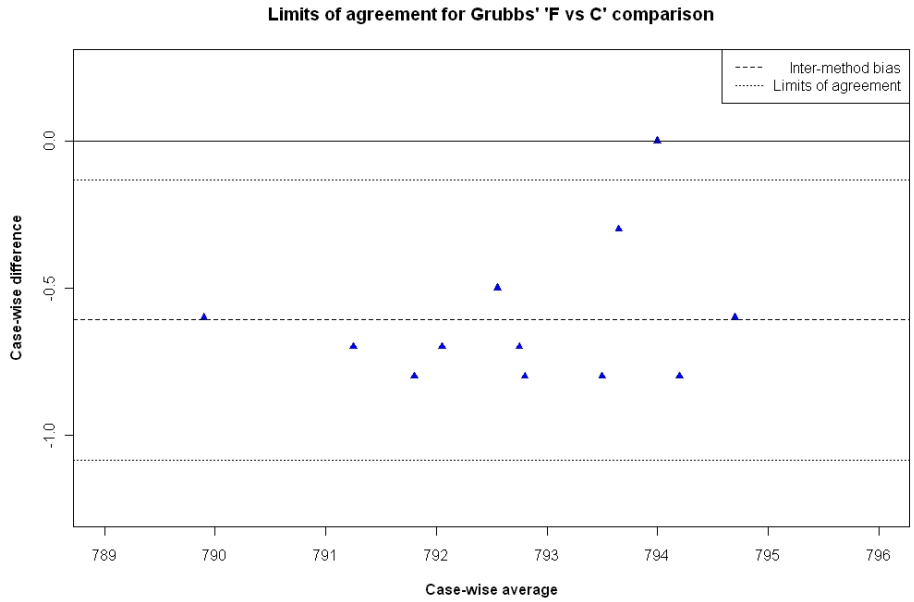

Figure 1.4.9: Bland-Altman plot with limits of agreement

1.5 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.

Calculation of the limits of agreement relies on the assumption that the case-wise differences are normally distributed. The calculation removes a lot of the variation between subjects, leaving measurement error, which is likely to be normally distributed. (Bland and Altman (1999) remark that this assumption is easy to check using a normal

plot.)

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

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.

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.

The difference are assumed to be normally distributed, although the measurements themselves are not assumed to follow any distribution. Therefore the authors argue that the 95% of differences are expected to lie within these limits. This assumption is justified because variation between subjects has been removed, leaving only measurement error (Bland and Altman, 1986). There are formal methodologies to test whether this assumption holds.

1.6 Interpretation of Limits Of Agreement

The purpose specifically intended for the limits of agreement must be established clearly. Bland and Altman (1995) comment that the limits of agreement *how far apart measurements by the two methods were likely to be for most individuals.*, a definition echoed in their 1999 paper:

> We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% limits of agreement. These values define the range within which most differences between measurements by the two methods will lie(Bland and Altman, 1999).

Carstensen et al. (2008) offers an alternative, more specific, definition of the limits of agreement *"a prediction interval for the difference between future measurements with the two methods on a new individual."* Luiz et al. (2003) describes them as tolerance limits.

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence a confidence interval, they are not constructed a such. They are designed for future values , hence Carstensen et al. (2008) refering to them as Prediction Intervals.

As with the Bland-Altman plot, the formulation of the Limits of Agreement is also heavily influenced by outliers. An example in Altman and Bland (1983) demonstrates the effect of recalculating without a particular outlier. Refering to the VCF data set in the same paper, there is more than one outlier.

1.6.1 Formal definition of Limits of Agreement

Bland and Altman (1999) note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as '*being like a reference interval*', offering no elaboration.
where n is the number of subjects. Carstensen is careful to consider the effect of the sample size on the interval width, adding that only for 61 or more subjects is the quantile less than 2.

1.6.2 Purpose of Limits of Agreement

It must be established clearly the specific purpose of the limits of agreement. Bland and Altman (1995) comment that the limits of agreement *how far apart measurements by the two methods were likely to be for most individuals.*, a definition echoed in their 1999 paper:

> We can then say that nearly all pairs of measurements by the two methods will be closer together than these extreme values, which we call 95% limits

of agreement. These values define the range within which most differences between measurements by the two methods will lie(Bland and Altman, 1999).

? offers an alternative, more specific, definition of the limits of agreement *"a prediction interval for the difference between future measurements with the two methods on a new individual."* Luiz et al. (2003) describes them as tolerance limits.

Importantly they have the same construction as Shewhart Control limits.

1.7 Interpretation of Limits Of Agreement

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.

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.

1.8 Limits of Agreement Outliers

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence a confidence interval, they are not constructed a such. They are designed for future values.

The formulation is also heavily influenced by outliers. An example in Altman and Bland (1983) demonstrates the effect of recalculating without a particular outlier. Refering to the VCF data set in the same paper, there is more than one outlier.

1.9 Inferences on Bland-Altman estimates

The limits of agreement are estimates derived from the sample studied, and will differ from values relevant to the whole population. Bland and Altman (1986) advance a for-

mulation for confidence intervals of the inter-method bias and the limits of agreement.

1.9.1 Confidence Intervals and Standard Error

Bland and Altman (1999) argue that it is possible to estimate confidence intervals and standard error, if it is assumed that the differences approximately follow a normal distribution,

These calculations employ quantiles of the 't' distribution with $n-1$ degrees of freedom. For the inter-method bias, the confidence interval is a simply that of a mean: $\bar{d} \pm t_{(\alpha/2,n-1)} S_d/\sqrt{n}$. The confidence intervals and standard error for the limits of agreement follow from the variance of the limits of agreement, which is shown to be

$$\mathrm{Var}(LoA) = (\frac{1}{n} + \frac{1.96^2}{2(n-1)}) s_d^2.$$

If n is sufficiently large this can be following approximation can be used

$$\mathrm{Var}(LoA) \approx 1.71^2 \frac{s_d^2}{n}.$$

Consequently the standard errors of both limits can be approximated as 1.71 times the standard error of the differences.

If n is sufficiently large this can be following approximation can be used

$$\mathrm{Var}(LoA) \approx 1.71^2 \frac{S_d^2}{n}. \tag{1.3}$$

Consequently the standard errors of both limits can be approximated as $1.71 S.E.(\bar{d})$

A 95% confidence interval can be determined, by means of the t distribution with $n-1$ degrees of freedom. Bland and Altman (1999) comment that such calculations may be 'somewhat optimistic' on account of the associated assumptions not being realized.

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. Bland and Altman (1986) 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.

1.10 Variations of the Bland-Altman Plot

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. As Bland and Altman (1986) point out this may not be the case. Bland and Altman advises on how to calculate of confidence intervals for the inter-method bias and the limits of agreement. Importantly Bland and Altman (1999) 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. Due to limitations of the conventional difference plot, a series of alternative formulations for the Bland-Altman approach have been proposed. Referring to the assumption that bias and variability are constant across the range of measurements, Bland and Altman (1999) address the case where there is an increase in variability as the magnitude increases. They remark that it is possible to ignore the issue altogether, but the limits of agreement would be 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. Bland and Altman acknowledge that this is not easy to interpret, and may not be suitable in all cases.

Bland and Altman (1999) 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 case-wise 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 case-wise ratios as percentage of averages. This will remove the need for logarithmic transformation. This approach is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. Eksborg (1981) proposed such a ratio plot, independently of Bland and Altman. Dewitte et al. (2002) commented on the reception of this article by saying '*Strange to say,this report has been overlooked*'.

1.11 Limits of Agreement for Replicate Measurements

Computing limits of agreement features prominently in many method comparison studies since the publication of Bland and Altman (1986). Bland and Altman (1999) addresses the issue of computing LoAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the two methods. However, the original Bland-Altman method was developed for two sets of measurements done on one occasion, and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method. In addition to Bland and Altman (1999), Carstensen et al. (2008) computes the limits of agreement to the case with replicate measurements by using LME models, an approach that will be discussed in due course.

1.11.1 Appropriate Use of Limits of Agreement

Importantly Bland and Altman (1999) 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.

1.12 Prevalence of the Bland-Altman plot

Bland and Altman (1986), which further develops the Bland-Altman methodology, was found to be the sixth most cited paper of all time by the Ryan and Woodall (2005). Dewitte et al. (2002) describes the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. Dewitte et al. (2002) reviewed the use of Bland-Altman plots by examining all articles in the journal 'Clinical Chemistry' between 1995 and 2001. This study concluded that use of the Bland-Altman plot increased over the years, from 8% in 1995 to 14% in 1996, and 31-36% 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.

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.

As an alternative to limits of agreement, Lin et al. (2002) proposes the use of the mean square deviation in assessing agreement. The mean square deviation is defined as the expectation of the squared differences of two readings. The MSD is usually used for the case of two measurement methods X and Y, each making one measurement for the same subject, and is given by

$$MSDxy = E[(x-y)^2] = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2 + 2\sigma_x\sigma_y(1-\rho_{xy}).$$

Barnhart et al. (2007) advises the use of a predetermined upper limit for the MSD value, MSD_{ul}, to define satisfactory agreement. However, a satisfactory upper limit may not be easily determinable, thus creating a drawback to this methodology.

Alternative indices, proposed by Barnhart et al. (2007), are the square root of the MSD and the expected absolute difference (EAD).

$$EAD = E(|x-y|) = \frac{\sum |x_i - y_i|}{n}$$

Both of these indices can be interpreted intuitively, since their units are the same as that of the original measurements. Also they can be compared to the maximum acceptable absolute difference between two methods of measurement d_0. For the sake of brevity, the EAD will be considered solely.

The EAD can be used to supplement the inter-method bias in an initial comparison study, as the EAD is informative as a measure of dispersion, is easy to calculate and requires no distributional assumptions. A consequence of using absolute differences is that high variances would result in a higher EAD value.

	U	V	$U - V$	$\|U - V\|$
1	98.05	99.53	-1.49	1.49
2	99.17	96.53	2.64	2.64
3	100.31	97.55	2.75	2.75
4	100.35	96.03	4.32	4.32
5	99.51	99.00	0.51	0.51
6	98.50	100.76	-2.26	2.26
7	100.66	99.37	1.29	1.29
8	99.66	108.87	-9.21	9.21
9	99.70	105.16	-5.45	5.45
10	101.55	94.31	7.24	7.24

Table 1.12.3: Example data set

To illustrate the use of EAD, consider table 1.12.3. The inter-method bias is 0.03, which is quite close to zero, which is desirable in the context of agreement. However, an identity plot would indicate very poor agreement, as the points are noticeably distant from the line of equality.

The limits of agreement are $[-9.61, 9.68]$, a wide interval for this data. As with the identity plot, this would indicate lack of agreement. As with inter-method bias, an EAD value close to zero is desirable. However, from table 1.12.3, the EAD can be computed as 3.71. The Bland-Altman plot remains a useful part of the analysis. In 1.12.11, it is clear there is a systematic decrease in differences across the range of measurements.

Barnhart et al. (2007) remarks that a comparison of EAD and MSD , using simulation studies, would be interesting, while further adding that '*It will be of interest*

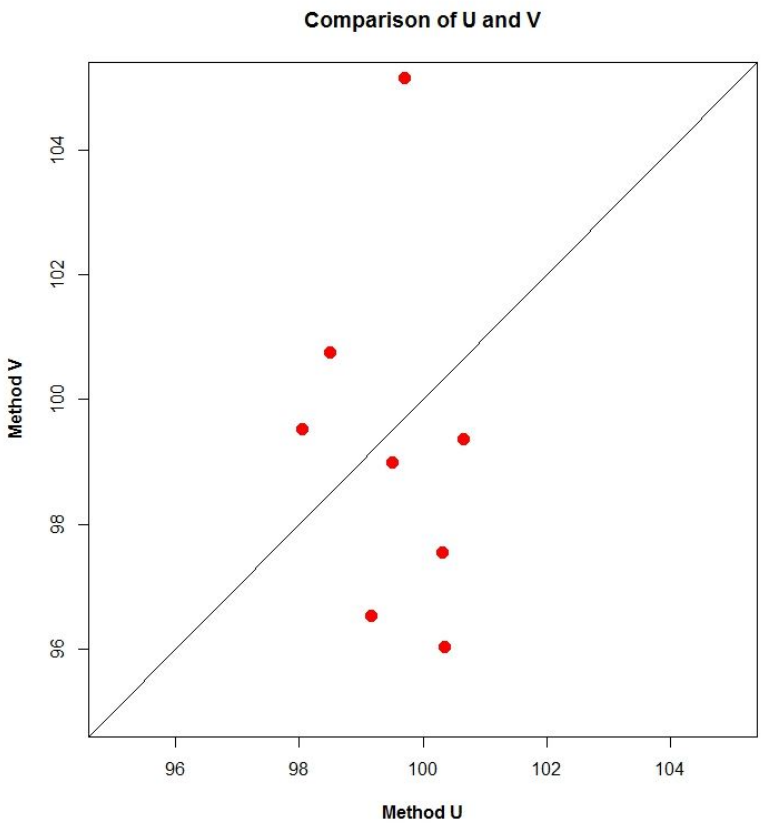

Figure 1.12.10: Identity Plot for example data

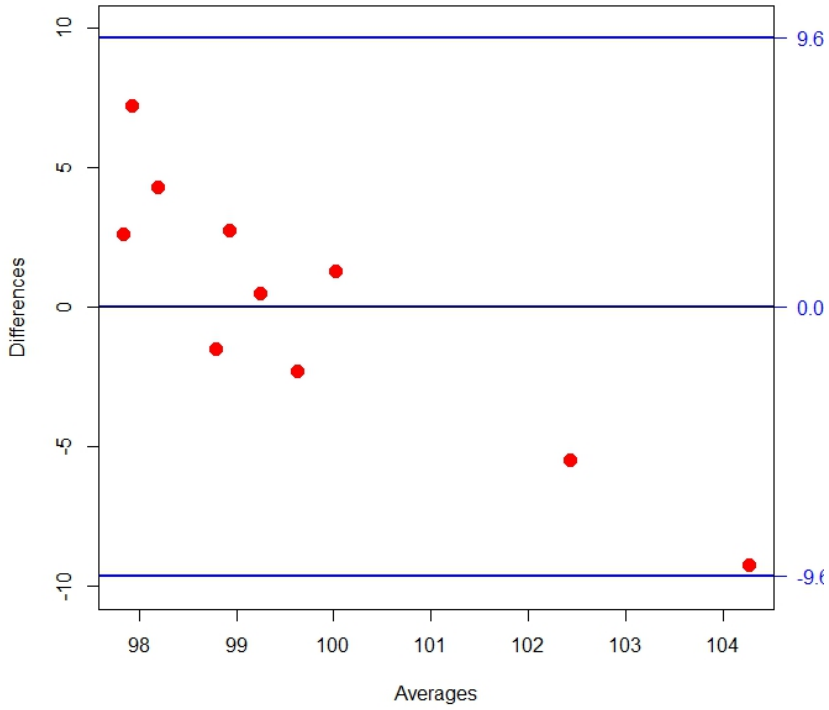

Figure 1.12.11: Bland-Altman Plot for UV comparison

to investigate the benefits of these possible new unscaled agreement indices'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. Hence the EAD values for both comparisons are much closer.

	F vs C	F vs T
Inter-method bias	-0.61	0.12
Difference variance	0.06	0.22
Limits of agreement	(-1.08, -0.13)	(-0.81,1.04)
EAD	0.61	0.35

Table 1.12.4: Agreement indices for Grubbs' data comparisons.

Further to Lin (2000) and Lin et al. (2002), individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{1.4}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. This boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

1.13 Coefficient of Repeatability

As mentioned previously, Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study. The coefficient of repeatability was proposed by Bland and Altman (1999), and is referenced in subsequent papers, such as Carstensen et al. (2008).

The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999). The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999).

Once the the standard deviations of the differences between the two measurements (in some texts called the residual standard deviation or within-item variability) $sigma_m$ is determined, the computation of the coefficients of repeatability for both methods is straightforward.

The coefficient is calculated from the (in some texts called the residual standard deviation) as $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$).

Bland and Altman (1999) introduces the repeatability coefficient for a method, which is defined as the upper limits of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances (Carstensen et al., 2008).

σ_x^2 is the within-subject variance of method x. The repeatability coefficient is $2.77\sigma_x$ (i.e. $1.96 \times \sqrt{2}\sigma_x$). For 95% of subjects, two replicated measurement by the same method will be within this repeatability coefficient.

1.14 Model Specification

The model underpinning the Bland-Altman approach can be presented as follows:

The case-wise differences $d_i = x_i - y_i$

$$\Sigma = \begin{pmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_b^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{pmatrix}$$

1.14.1 Formal Testing

The Bland Altman plot is a simple tool for inspection of the data, but in itself it offers no formal testing procedure in this regard. To this end, the approach proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of casewise differences and means (ρ_{AD}). According to the authors, this test is equivalent to a well established tests for equality of variances, known as the 'Pitman Morgan Test' (Pitman, 1939; Morgan, 1939).

For the Grubbs data, the correlation coefficient estimate (r_{AD}) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers "*r to z*" transformation (Cohen, Cohen, West, and Aiken, Cohen et al.). The null hypothesis (ρ_{AD} =0) would fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected.

There has no been no further mention of this particular test in the subsequent article published by Bland and Altman, although Bland and Altman (1999) refers to Spearmans' rank correlation coefficient. Bland and Altman (1999) comments 'we do not see a place for methods of analysis based on hypothesis testing'. Bland and Altman (1999) also states that consider structural equation models to be inappropriate.

1.14.2 Analysis of Variance

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{1.5}$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$. For the replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{1.6}$$

Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item.

1.14.3 Two Way ANOVA

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. Their model describing y_{mir}, again the rth replicate measurement on the ith item by the mth method ($m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n$), can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \tag{1.7}$$

The fixed effects α_m and μ_i represent the intercept for method m and the 'true value' for item i respectively. The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$, a method-by-item interaction term $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$, and model error terms $\varepsilon \sim \mathcal{N}(0, \varphi_m^2)$. All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item i, a_{ir} can be removed.

The model expressed in (2) describes measurements by m methods, where $m = \{1, 2, 3 \ldots\}$. Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

$$\alpha_1 - \alpha_2 \pm 2\sqrt{\tau_1^2 + \tau_2^2 + \varphi_1^2 + \varphi_2^2}$$

Carstensen et al. (2008) notes that, for $m = 2$, separate estimates of τ_m^2 can not be obtained. To overcome this, the assumption of equality, i.e. $\tau_1^2 = \tau_2^2$ is required.

1.14.4 Classical Model

The classical model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2 \ldots$

$$y_{mi} + \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim \backslash(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate α can not be estimated, only theiir difference can be estimated as $\bar{D}$

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.

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

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 (1995). Comparing methods of measurement - why plotting difference against standard method is misleading. *The Lancet 346*, 1085–87.

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

Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics 5*(3), 399–413.

Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics 4*(1).

Cohen, J., P. Cohen, S. West, and L. Aiken. *Applied multiple regression / correlation analysis for the behavioral sciences* (Third ed.). Laurence Erlbaum Associates.

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.

Eksborg, S. (1981). Evaluation of method-comparison data [letter]. *Clinical Chemistry 27*, 1311–1312.

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

Kozak, M. and A. Wnuk (2014). Including the tukey mean-difference (bland–altman) plot in a statistics course. *Teaching Statistics 36*(3), 83–87.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

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.

Luiz, R., A. Costa, P. Kale, and G. Werneck (2003). Assessment of agreement of a quantitative variable: a new graphical approach. *Journal of Clinical Epidemiology 56*, 963–967.

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.

Morgan, W. A. (1939). A test for the signicance of the difference between two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–19.

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.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika 31*, 9–12.

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