Contents

Chapter 1

Techniques for Method Comparison

1.1 The Bland-Altman Approach to Method Comparison

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

In 1983 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes (Altman and Bland, 1983). Bland-Altman plots are a powerful graphical technique for making a visual assessment of the data. Altman and Bland (1983) express the motivation for this plot:

> *"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."*

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.

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.

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 magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$, and is represented with a line on the Bland-Altman plot. 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) do, however, state that the 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 propose their simple approach specifically constructed for method comparison studies. They acknowledge that there are other valid, but complex approaches, but argue that a simple approach is preferable, *"especially when the results must be explained to non-statisticians"* (Altman and Bland, 1983).

1.1.1 Inspecting Method Comparison Data

The first step recommended, which the authors argue should be mandatory, is construction of an identity plot, which is a simple scatter-plot approach of measurements for both methods on either axis. The line of equality (the $X = Y$ line, i.e. the 45 degree line through the origin) should also be shown, as it is necessary to give the correct interpretation of how both methods compare.

This plot can gives the analyst a cursory examination of how well the measurement methods agree. 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 on the left 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.

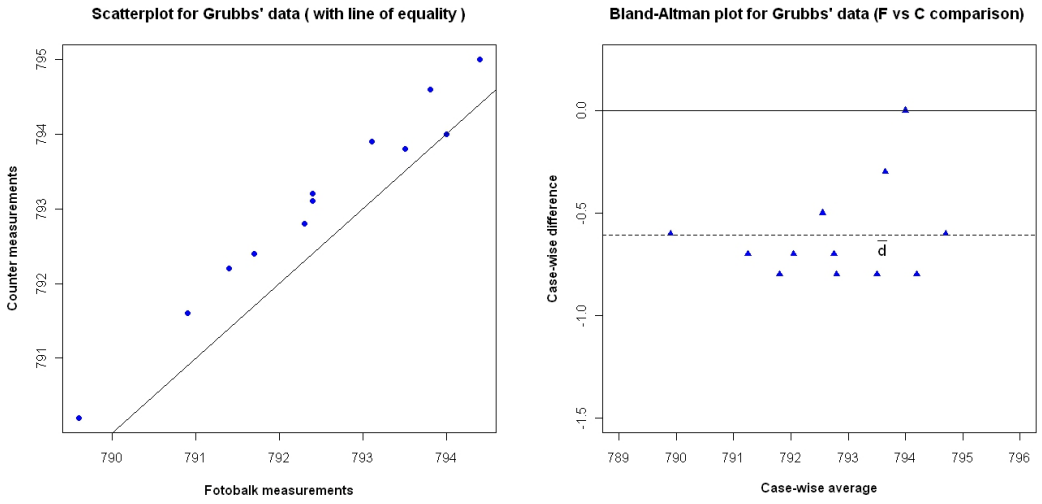

Figure 1.1.1: Identity Plot and Bland-Altman Plot For Fotobalk and Counter methods.

However scatter-plots, such as these, are not sufficient 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 plot for comparing the 'Fotobalk' and 'Counter' methods, which shall henceforth be referred to as the 'F vs C' comparison, is depicted on the right in

Figure 1.1.1, using data from Table 1.1.1. The dashed line in the Bland-Altman plot alludes to the inter-method bias between the two methods, estimated 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, one would notice that the differences tend to increase as the averages increase.

Round	Fotobalk [F]	Counter [C]	Terma [T]	[F-C]	[(F+C)/2]	[F-T]	[(F+T)/2]
1	793.8	794.6	793.2	-0.8	794.2	0.6	793.5
2	793.1	793.9	793.3	-0.8	793.5	-0.2	793.2
3	792.4	793.2	792.6	-0.8	792.8	-0.2	792.5
4	794.0	794.0	793.8	0.0	794.0	0.2	793.9
5	791.4	792.2	791.6	-0.8	791.8	-0.2	791.5
6	792.4	793.1	791.6	-0.7	792.8	0.8	792.0
7	791.7	792.4	791.6	-0.7	792.0	0.1	791.6
8	792.3	792.8	792.4	-0.5	792.5	-0.1	792.3
9	789.6	790.2	788.5	-0.6	789.9	1.1	789.0
10	794.4	795.0	794.7	-0.6	794.7	-0.3	794.5
11	790.9	791.6	791.3	-0.7	791.2	-0.4	791.1
12	793.5	793.8	793.5	-0.3	793.6	0.0	793.5

Table 1.1.1: Fotobalk : Differences and Averages with Counter and Terma.

In Figure 1.1.2 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.

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 Bland-Altman plot.

Figure 1.1.3 and Figure 1.1.4 show two Bland-Altman plots derived from simulated

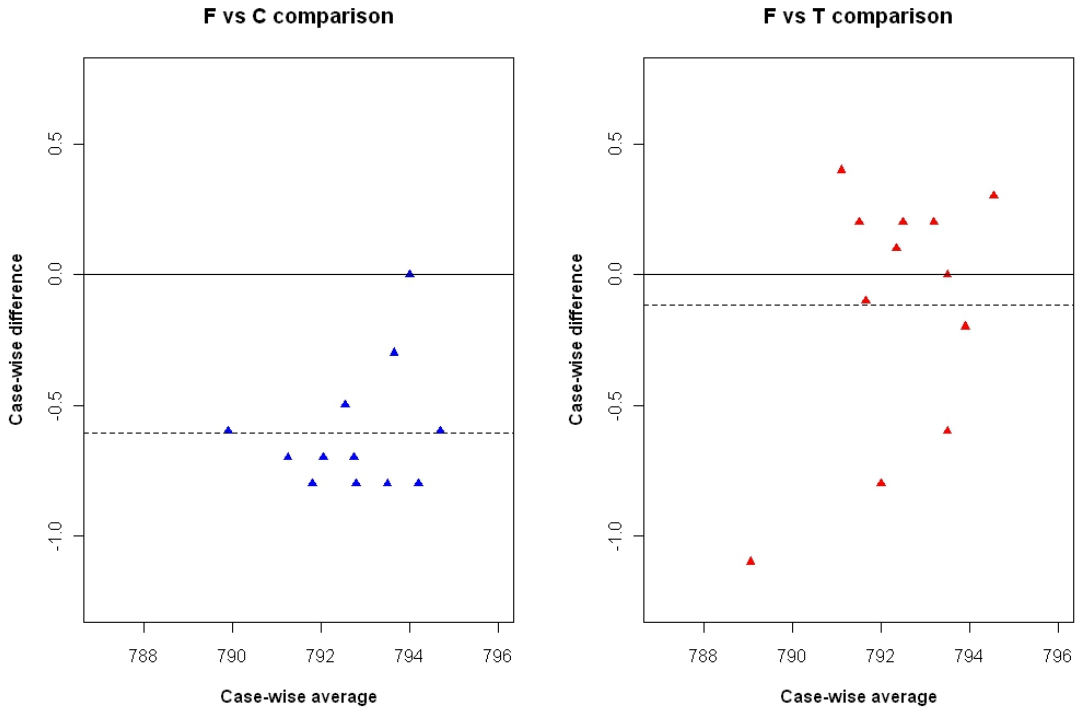

Figure 1.1.2: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

data, each for the purpose of demonstrating how the plot would inform an analyst of trends that would adversely affect use of the recommended approach. Additionally the procedure is not properly constructed to deal with outliers, which shall be reverted to later.

Figure 1.1.3 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, have 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 are advisable, for example, the Benjamini-

Hochberg Test (Benjamini and Hochberg, 1995).

Figure 1.1.4 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.

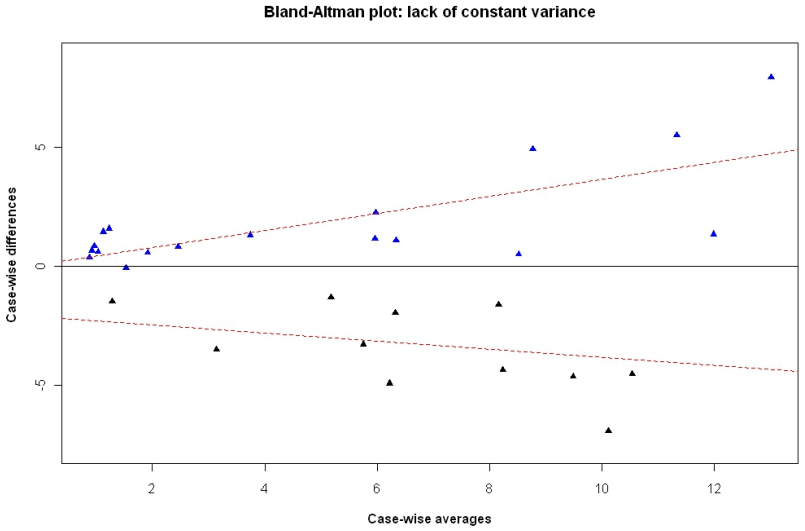

Figure 1.1.3: Bland-Altman Plot demonstrating the increase of variance over the range

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

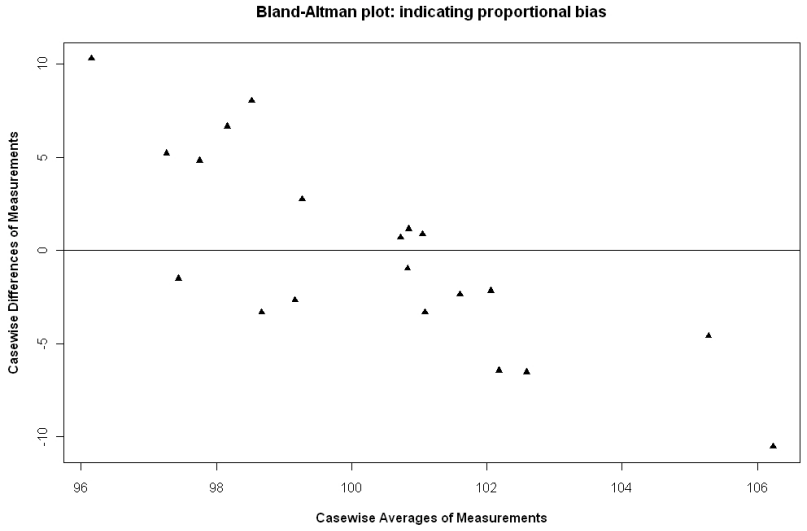

Figure 1.1.4: Bland-Altman Plot indicating the presence of proportional bias

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 be 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, removing 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*'.

1.2 Limits of Agreement

A third element of the Bland-Altman approach, an interval known as limits of agreement is introduced in Bland and Altman (1986) (sometimes referred to in literature as 95% limits of agreement). These limits centre on the average difference line, and are computed as $LOA = \bar{d} \pm 1.96 s_d$ with $\bar{d}$ as the estimate of the inter method bias and s_d the standard deviation of the differences. Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably, by demonstrating the range in which 95% of the sample data should lie. The limits of agreement requires an assumption of a constant level of bias throughout the range of measurements.

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.*

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.

Calculation of the limits of agreement relies on the assumption that the case-wise differences are normally distributed, although the measurements themselves are not assumed to follow any distribution. This assumption is justified because variation

between subjects has been removed, leaving measurement error, which is likely to be normally distributed (Bland and Altman, 1986). Bland and Altman (1999) remark that this assumption is easy to check using commonly used methods, i.e. a normal probability plot.

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 1.2.5 shows the resultant Bland-Altman plot, with the limits of agreement shown in dotted lines.

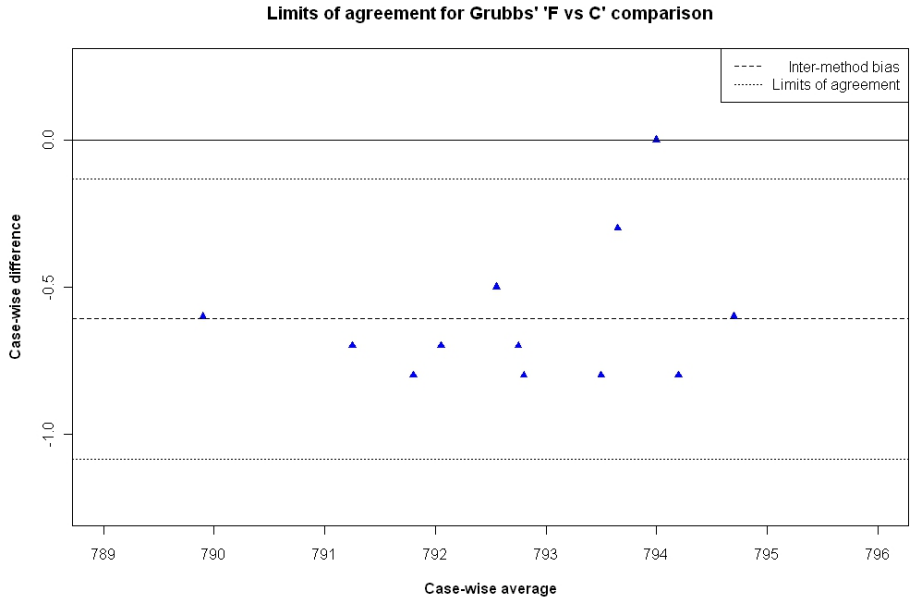

Figure 1.2.5: Bland-Altman plot with limits of agreement

1.2.1 Interpretation 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.

The Shewhart chart is a well known graphical technique used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as

if they were Shewhart control limits. Importantly the parameters used to determine the limits, the mean and standard deviation, are not based on any randomly ordered sample used for an analysis, but on a statistical process's time ordered 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) offer an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.025,n-1)} S_d \sqrt{1 + \frac{1}{n}} \tag{1.1}$$

where n is the number of subjects. With consideration of the effect of the sample size on the interval width, Carstensen et al. (2008) remarks that only for 61 or more subjects is the 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.

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 will inevitably attenuate the poor standard of reporting using limits of agreement, as discussed by Mantha et al. (2000).

1.2.2 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. Bland and Altman (1986) advance a formulation for confidence intervals of the inter-method bias and the limits of agreement, arguing that it is possible to make such estimates if it is assumed that the case-wise differences approximately follow a normal distribution. However Bland and Altman (1999) caution that such calculations may be 'somewhat optimistic' if the associated assumptions are not valid. A 95% confidence interval can be determined, by means of the t distribution with $n-1$ degrees of freedom. For the inter-method bias, the confidence interval is 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

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

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

$$\mathrm{Var}(LOA) \approx 1.71^2 \frac{s_d^2}{n},$$

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

1.3 Detection of Outliers in the Bland-Altman Framework

The Bland-Altman plot can be used to identify outliers. Here we use a simple definition of an outlier as 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. 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 Altman (1999), we get the clearest indication of what they 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. 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."* Classification thereof is a subjective decision in any analysis, but must be informed by the logic of the mechanism that produces the data.. Figure 1.3.6 is a Bland-Altman plot with two conspicuous observations, at the extreme left and right of the plot respectively. In the Bland-Altman plot depicted in Figure 1.3.6, consider the

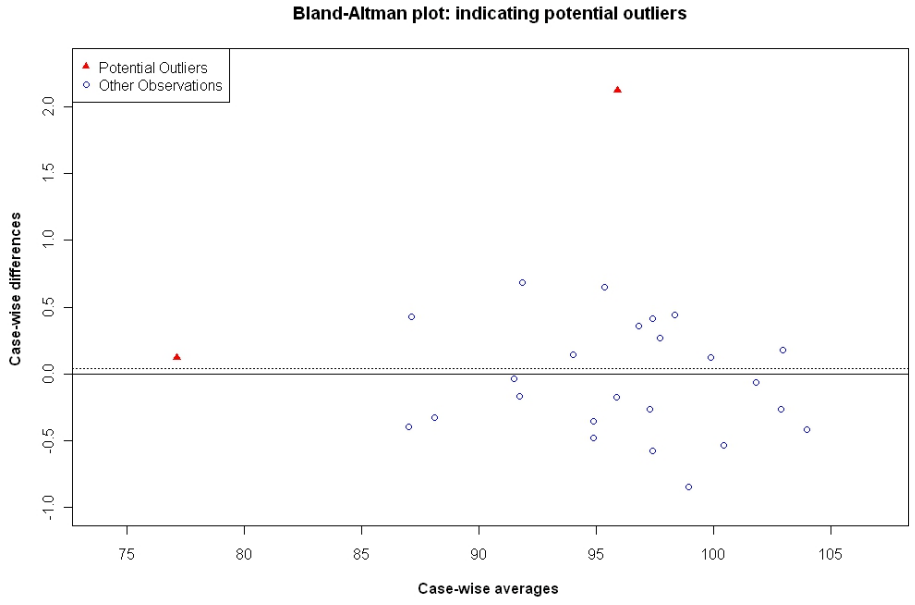

Figure 1.3.6: Bland-Altman plot indicating the presence of outliers

covariate located on the extreme left of the plot. Ordinarily we would conclude that this point due to it's horizontal displacement from the main cluster of points. However

this horizontal displacement is supported by two independent measurements and is very close to the inter-method bias, i.e. very close to its expected value. Therefore that observation, should not be considered an outlier at all.

Conversely the observation located at the top of the plot, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations. There are no mitigating factors.

1.3.1 Bartko's Ellipse

As an enhancement on the Bland-Altman Plot, Bartko (1994) has expounded a confidence ellipse for the covariates. Bartko (1994) offers a graphical complement to the Bland-Altman plot in the form of a bivariate confidence ellipse as a boundary for dispersion, with Altman (1978) providing the relevant calculations. This ellipse is intended as a visual guideline for the scatter plot, for detecting outliers and to assess the within- and between-subject variability. 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. Additionally Bartko's ellipse provides a visual aid to determining the relationship between 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.

Furthermore, the ellipse provides a visual aid to determining the relationship between the variance of the means $\mathrm{Var}(a)$ and the variance of the differences $\mathrm{Var}(d)$. If $\mathrm{Var}(a)$ is greater than $\mathrm{Var}(d)$, the orientation of the ellipse is horizontal. Conversely if $\mathrm{Var}(a)$ is less than $\mathrm{Var}(d)$, the orientation of the ellipse is vertical. The more horizontal the ellipse, the greater the degree of agreement between the two methods being tested.

Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers. The limitations of using bivariate approaches to outlier detection in the Bland-

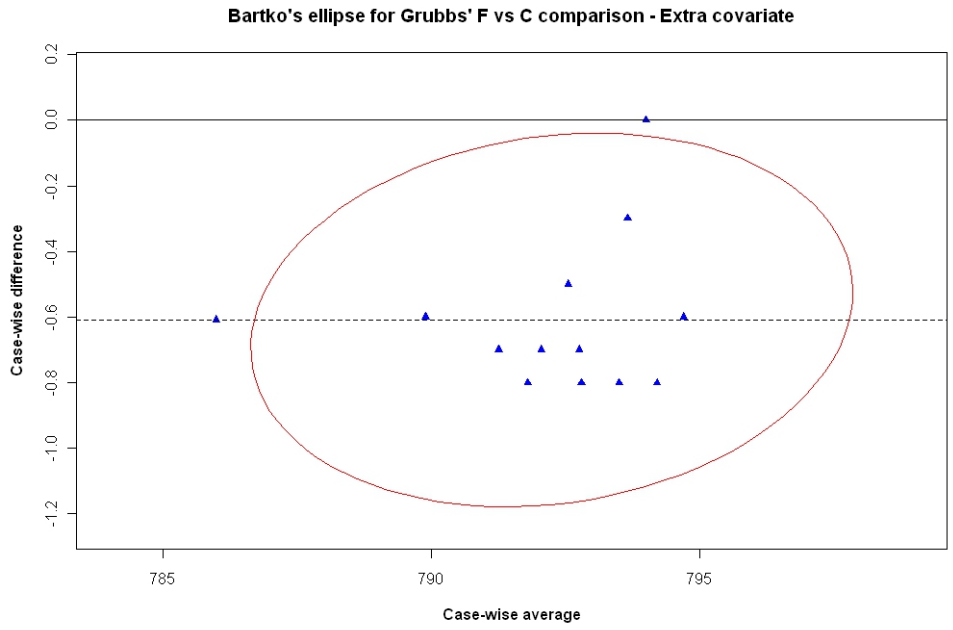

Figure 1.3.7: Bartko's ellipse for Grubbs data

Altman plot can demonstrated using Bartko's ellipse.

The Bland-Altman plot for the Grubbs data, complemented by Bartko's ellipse, is depicted in Figure 1.3.7. However, both observation that were previously considered as potential outliers (i.e. the extreme left and the uppermost) are shown to be outside the bounds of the ellipse, indicating both to be outliers.

1.3.2 Grubbs' Test for Outliers

In classifying whether an 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,

$$G = \max_{i=1,\ldots,n} \frac{\left|d_i - \bar{d}\right|}{S_d}. \tag{1.2}$$

For the 'F vs C' comparison it is the fourth observation that 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.

1.4 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) reviews the use of Bland-Altman plots by examining all articles in the journal 'Clinical Chemistry' between 1995 and 2001, describing the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. This study concludes 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 the expected, and often the 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 on 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, while the other two used correlation and regression analyses. Mantha et al. (2000) remark that 3 papers, from 42 mention predefined maximum width for limits of agreement that 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) remark that the limits of agreement should be compared to a clinically acceptable difference in measurements.

1.5 Criticism of Limits of Agreement

The Bland-Altman approach is well noted for its ease of use, and can be easily implemented with most software packages. Also it does not 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' or the presence of an outlier.

However the approach comes in for criticism in a number of respects. In the first instance, some caution must be given to the inter-method bias estimate. If one method is sometimes higher, or sometimes lower, the average of the differences will be close to zero. If the inter-method bias is close to zero, there be an indication that the two measurement methods are in agreement, when in fact they are producing different results systematically.

Several problems have been highlighted regarding limits of agreement. One is the somewhat arbitrary manner in which they are constructed. 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) percentage points. According to the authors, a knowledgeable practitioner in the field should 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. Furthermore Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

While in essence they are similar to confidence intervals, limits of agreement are not constructed as such; they are designed for future values. Lack of clarity in this regards can give rise to confusion, and incorrect interpretations.

Ludbrook (1997, 2002) criticizes Bland-Altman plots on the basis that they present 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. There is no guidance on how to deal with outliers. Bland and Altman recognize the effect they would have on the limits of agreement, but offer no guidance on how to correct for those effects. Finally the adaptation of the approach to deal with replicate measurements, as specified by Bland and Altman (1999), is flawed.

1.6 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. Contrary to Bland and Altman (1999), Carstensen et al. (2008) computes the limits of agreement to the case with replicate measurements by using LME models. This approach will be discussed in due course.

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., 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).

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, 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.

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.

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.