

# Contents

<b>1</b>	<b>Bland-Altman Methodology</b>	<b>4</b>
1.1	Bland Altman Plots In Literature . . . . .	4
1.2	The Bland Altman Plot . . . . .	4
1.3	Criticism of Bland Altman Plot . . . . .	5
1.4	Bland Altman Plots . . . . .	5
1.4.1	Inspecting the Data . . . . .	8
1.5	Limits of Agreement . . . . .	12
1.5.1	Variations of the Bland Altman Plot . . . . .	12
1.6	Bland Altman Plots . . . . .	12
1.7	Bland Altman plot and the Treatment of Outliers . . . . .	17
1.7.1	Criticism of Bland Altman Plot . . . . .	20
1.7.2	Treatment of Outliers . . . . .	22
1.8	Bland Altman plot and the Treatment of Outliers . . . . .	22
1.8.1	Motivation for Bland-Altman plots . . . . .	23
1.8.2	Prevalence of the Bland-Altman plot . . . . .	27
1.8.3	Adverse features . . . . .	28
1.8.4	Inferences on Bland-Altman estimates . . . . .	34
1.9	Limits of Agreement . . . . .	38
1.9.1	Formal definition of limits of agreement . . . . .	39
1.9.2	Replicate Measurements . . . . .	40
1.9.3	Bland-Altman correlation test . . . . .	41

1.9.4	Identifiability . . . . .	42
1.10	Regression Methods . . . . .	43
1.11	Limits of Agreement . . . . .	44
1.12	Outline of Thesis . . . . .	45
1.12.1	Prevalence of the Bland-Altman plot . . . . .	49
1.13	Variations of the Bland-Altman Plot . . . . .	49
1.14	Variations of the Bland Altman Plot . . . . .	50
1.15	Multivariate . . . . .	52
1.15.1	Mahalanobis Distance . . . . .	52
1.16	Bartko's Ellipse . . . . .	52
1.16.1	Relevance of Repeatability . . . . .	54
1.17	Repeatability . . . . .	54
1.17.1	Repeatability and gold standards . . . . .	54
<b>2</b>	<b>Coefficient of Repeatability</b>	<b>55</b>
2.1	Add Ins . . . . .	55
2.2	Bland and Altman . . . . .	55
2.3	Carstensen . . . . .	56
2.3.1	Repeatability . . . . .	57
2.4	Reproducibility . . . . .	57
2.4.1	2 The Coefficient of Repeatability . . . . .	57
2.5	Repeatability . . . . .	58
2.6	Repeatability . . . . .	58
2.6.1	Repeatability . . . . .	59
2.6.2	Bland and Altman 1999 . . . . .	60
2.6.3	Notes from BXC Book (chapter 9) . . . . .	60
2.7	Coefficient of Repeatability . . . . .	64
2.7.1	Repeatability . . . . .	64
2.7.2	Coefficient of Repeatability . . . . .	64

2.7.3	Note 1: Coefficient of Repeatability . . . . .	64
2.7.4	Repeatability coefficient . . . . .	65
2.7.5	Repeatability . . . . .	66
2.8	The Identity plot . . . . .	66
2.9	Variations of the BA plot . . . . .	66
2.10	Survival Agreement Plot (Luiz et al) . . . . .	66
2.11	Eksborg's Plot . . . . .	67
2.12	Bartko's Ellipse . . . . .	68
2.13	dewitte et al . . . . .	69
2.14	Mountain Plot . . . . .	70
2.14.1	Survival Plots . . . . .	75
2.14.2	Deming Regression . . . . .	76

# Chapter 1

## Bland-Altman Methodology

### 1.1 Bland Altman Plots In Literature

Mantha et al. (2000) contains a study the use of Bland Altman plots of 44 articles in several named journals over a two year period. 42 articles used Bland Altman's limits of agreement, wit the other two used correlation and regression analyses. Mantha et al. (2000) remarks that 3 papers, from 42 mention predefined maximum width for limits of agreement which would not impair medical care.

The conclusion of Mantha et al. (2000) is that there are several inadequacies and inconsistencies in the reporting of results ,and that more standardization in the use of Bland Altman plots is required. The authors recommend the prior determination of limits of agreement before the study is carried out. This contention is endorsed by Lin et al. (1991), which makes a similar recommendation for the sample size, noting *thatsample sizes required either was not mentioned or no rationale for its choice was given.*

### 1.2 The Bland Altman Plot

In 1986 Bland and Altman published a paper in the Lancet proposing the difference plot for use for method comparison purposes. It has proved highly popular ever since.

This is a simple, and widely used, plot of the differences of each data pair, and the corresponding average value. An important requirement is that the two measurement methods use the same scale of measurement.

Variations of the Bland Altman plot is the use of ratios, in the place of differences.

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

Altman and Bland suggest plotting the within subject differences  $D = X_1 - X_2$  on the ordinate versus the average of  $x_1$  and  $x_2$  on the abscissa.

### 1.3 Criticism of Bland Altman Plot

Unfortunately the Bland-Altman plot has a fatal flaw: it indicates incorrectly that there are systematic differences or bias in the relationship between two measures, when one has been calibrated against the other. (Hopkins)

### 1.4 Bland Altman Plots

The issue of whether two measurement methods are comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of matched pairs correlation coefficients or simple linear regression. Bland and Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983).

As an alternative they proposed a simple statistical methodology specifically appropriate for method comparison studies. They acknowledge that there are other valid methodologies, but argue that a simple approach is preferable to complex approaches, *“especially when the results must be explained to non-statisticians”* (Altman and Bland, 1983).

The first step recommended, which the authors argue should be mandatory, is construction of a simple scatter plot of the data. The line of equality ( $X = Y$ ) should also be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 1.1). The averages of the two measurements is considered by Bland and Altman to be the best estimate for the unknown true value. Importantly both methods must measure with the same units. These results are then plotted, with differences on the ordinate and averages on the abscissa (figure 1.2).

The dashed line in figure 1.2 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  $-0.61$  metres per second.

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 1.4.1: Fotobalk and Counter Methods: Differences and Averages

From a visual inspection of Bland-Altman plot, it is also possible to compare the precision of each method in addition to the inter-method bias. Evidently the data points in the figure 1.2 tend to cluster near the bias line, particularly at the lower end of the range of measurements. The variances of the differences seem to increase along the range. In case of small data sets, any decision on the level of precision is subjective.

### 1.4.1 Inspecting the Data

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. Altman and Bland (1983) express the motivation for this plot thusly:

”From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

Figures XX YY and ZZ 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 XX demonstrates how the Bland Altman plot would indicate increasing variance of differences over the measurement range.

Figure ZZ is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias (Ludbrook, 1997).



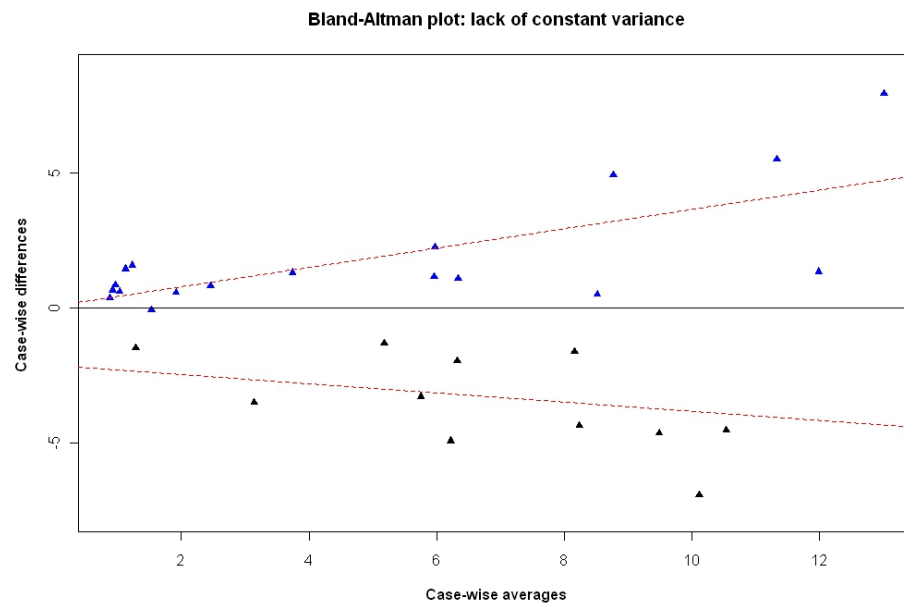


Figure 1.4.1: Bland-Altman Plot demonstrating the increase of variance over the range

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

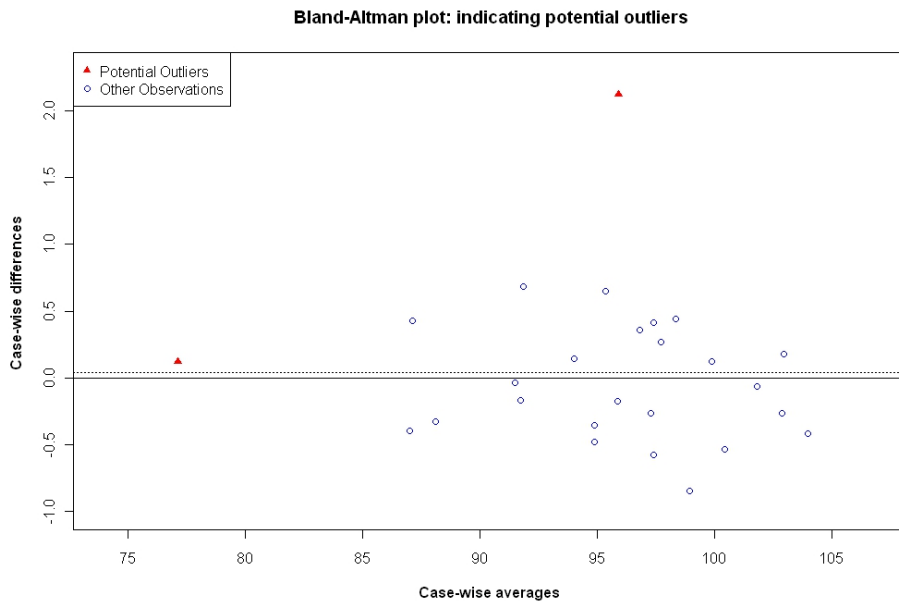


Figure 1.4.2: Bland-Altman Plot indicating the presence of Outliers

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 YY, should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster. The one on the extreme left should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

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

## 1.5 Limits of Agreement

Bland and Altman (1986) introduces an elaboration of the plot, adding to the plot ‘limits of agreement’ to the plot. These limits are based upon the standard deviation of the differences. The discussion shall be reverted to these limits of agreement in due course.

### 1.5.1 Variations of the Bland Altman Plot

Bland and Altman (1999) remarks that it is possible to ignore the issue altogether, but the limits of agreement would wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. Bland and Altman (1999) acknowledge that this is not easy to interpret, and that it is not 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 casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases.

## 1.6 Bland Altman Plots

The issue of whether two measurement methods are comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically comparison of two methods of measurement was carried out by use of matched pairs correlation coefficients or simple linear regression. Bland and Altman recognized the inadequacies of these analyses and articulated quite

thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement (Altman and Bland, 1983).

As an alternative they proposed a simple statistical methodology specifically appropriate for method comparison studies. They acknowledge that there are other valid methodologies, but argue that a simple approach is preferable to complex approaches, *”especially when the results must be explained to non-statisticians”* (Altman and Bland, 1983).

The first step recommended which the authors argue should be mandatory is construction of a simple scatter plot of the data. The line of equality ( $X = Y$ ) should also be shown, as it is necessary to give the correct interpretation of how both methods compare. A scatter plot of the Grubbs data is shown in figure 2.1. A visual inspection thereof confirms the previous conclusion that there is an inter method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 1.1). The averages of the two measurements is considered by Bland and Altman to be the best estimate for the unknown true value. Importantly both methods must measure with the same units. These results are then plotted, with differences on the ordinate and averages on the abscissa (figure 1.2). Altman and Bland (1983) express the motivation for this plot thusly:

”From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 1.6.2: Fotobalk and Counter Methods: Differences and Averages

The dashed line in figure 1.2 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  $-0.61$  metres per second.

From a visual inspection of Bland-Altman plot, it is also possible to compare the precision of each method in addition to the inter-method bias. Evidently the data points in the figure HH tend to cluster near the bias line, particularly at the lower end of the range of measurements. The variances of the differences seem to increase along the range. In case of small data sets, Any decision on the level of precision is subjective.

Figures XX and ZZ are Bland-Altman plots of data simulated for expository purposes. Figure XX demonstrates how the Bland Altman plot would indicate increasing variance of differences over the measurement range. Figure ZZ is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias (Ludbrook, 1997).

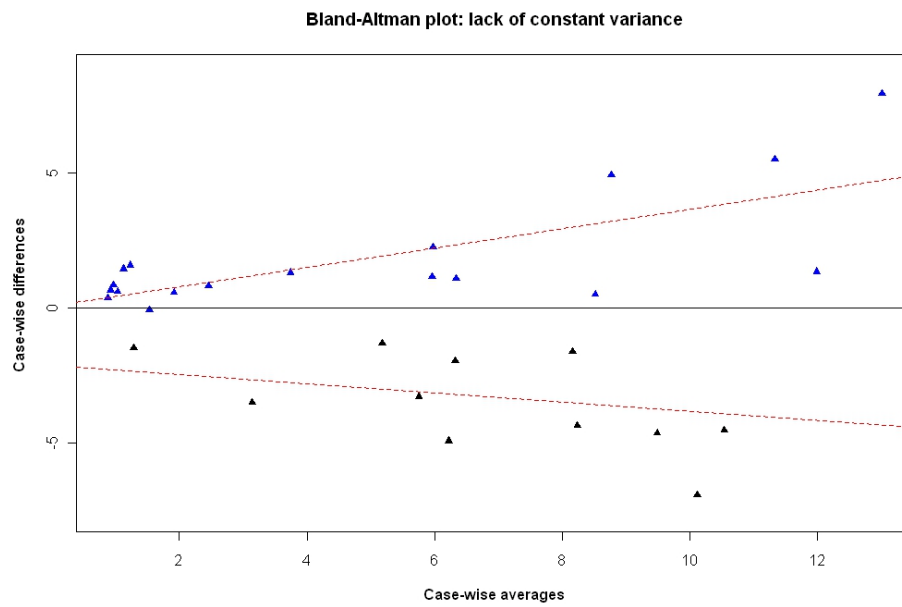


Figure 1.6.3: Bland-Altman Plot demonstrating the increase of variance over the range



The fourth pair of measurements from table 1.1 show both methods recording the same value, hence the difference is zero. In assessing the impact of the corresponding data point, two conflicting conclusions can be drawn.

One conclusion is that it is an outlier, and the precision of the differences is consistent along the range of measurements. The other conclusion is that it is not an outlier, and the precision increases proportionately along the range of measurements.

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 ( $\rho_{AD}$ ). 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 ( $\rho_{AD} = 0$ ) would fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected.

There has been no further mention of this particular test in the subsequent article published by Bland and Altman, although Bland and Altman (1999) refers to Spearman’s rank correlation coefficient.

The second variation is a plot of casewise ratios as percentage of averages.

## 1.7 Bland Altman plot and the Treatment of Outliers

We wish to determine how outliers should be treated in a Bland Altman Plot. In their 1983 paper Bland and Altman merely state that the plot can be used to ‘spot outliers’. However they pay more attention to the issue of outliers in their 1986 paper, wherein

they present a data set with an extreme outlier.

In light of shortcomings associated with scatterplots, Altman and Bland (1983) recommend a further analysis of the data. Firstly differences of measurements of two methods on the same subject should be calculated, and then the average of those measurements (Table 1.1). The averages of the two measurements is considered by Bland and Altman to be the best estimate for the unknown true value. Importantly both methods must measure with the same units. These results are then plotted, with differences on the ordinate and averages on the abscissa (figure 1.2). Altman and Bland (1983) express the motivation for this plot thusly:

”From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.80	794.60	-0.80	794.20
2	793.10	793.90	-0.80	793.50
3	792.40	793.20	-0.80	792.80
4	794.00	794.00	0.00	794.00
5	791.40	792.20	-0.80	791.80
6	792.40	793.10	-0.70	792.80
7	791.70	792.40	-0.70	792.00
8	792.30	792.80	-0.50	792.50
9	789.60	790.20	-0.60	789.90
10	794.40	795.00	-0.60	794.70
11	790.90	791.60	-0.70	791.20
12	793.50	793.80	-0.30	793.60

Table 1.7.3: Fotobalk and Counter Methods: Differences and Averages

The dashed line in figure 1.2 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  $-0.61$  metres per second.

From a visual inspection of Bland-Altman plot, it is also possible to compare the precision of each method in addition to the inter-method bias. Evidently the data points in the figure HH tend to cluster near the bias line, particularly at the lower end of the range of measurements. The variances of the differences seem to increase along the range. In case of small data sets, Any decision on the level of precision is subjective.

Figures XX and ZZ are Bland-Altman plots of data simulated for expository purposes. Figure XX demonstrates how the Bland Altman plot would indicate increasing variance of differences over the measurement range. Figure ZZ is an example of cases

where the inter-method bias changes over the measurement range. This is known as proportional bias (Ludbrook, 1997).

The fourth pair of measurements from table 1.1 show both methods recording the same value, hence the difference is zero. In assessing the impact of the corresponding data point, two conflicting conclusions can be drawn.

One conclusion is that it is an outlier, and the precision of the differences is consistent along the range of measurements. The other conclusion is that is not an outlier, and the precision increases proportionately along the range of measurements.

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 ( $\rho_{AD}$ ). 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 ( $\rho_{AD} = 0$ ) 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.

### 1.7.1 Criticism of Bland Altman Plot

Hopkins[8] argues that the plot indicates incorrectly that there are systematic differences or bias in the relationship between two measures, when one has been calibrated against the other.

An Evaluation of the correlation between the difference and means complement the

analysis.

Bland and Altman caution that the calculations are based on the assumption that the data is normally distributed. This can be verified by using a histogram. If Data is not normally distributed, it can be transformed.

### 1.7.2 Treatment of Outliers

Bland and Altman attend to the issue of outliers in their 1986 paper, wherein they present a data set with an extreme outlier

## 1.8 Bland Altman plot and the Treatment of Outliers

We wish to determine how outliers should be treated in a Bland Altman Plot. In their 1983 paper Bland and Altman merely state that the plot can be used to 'spot outliers'. However they pay more attention to the issue of outliers in their 1986 paper, wherein they present a data set with an extreme outlier.

In Bland and Altman's 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.*

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.

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 be 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.

### 1.8.1 Motivation for Bland-Altman plots

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

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

“From this type of plot it is much easier to assess the magnitude of disagreement (both error and bias), spot outliers, and see whether there is any trend, for example an increase in (difference) for high values. This way of plotting the data is a very powerful way of displaying the results of a method comparison study.”

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

The magnitude of the inter-method bias between the two methods is simply the average of the differences  $\bar{d}$ . This inter-method bias is represented with a line on the Bland-Altman plot. As the objective of the Bland-Altman plot is to advise on the

agreement of two methods, it is the case-wise differences that are also particularly relevant. The variances around this bias is estimated by the standard deviation of these differences  $S_d$ .

### **Bland-Altman plots for the Grubbs data**

In the case of the Grubbs data the inter-method bias is  $-0.61$  metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

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



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

Table 1.8.4: 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 1.8.5: Fotobalk and Terma methods: differences and averages.

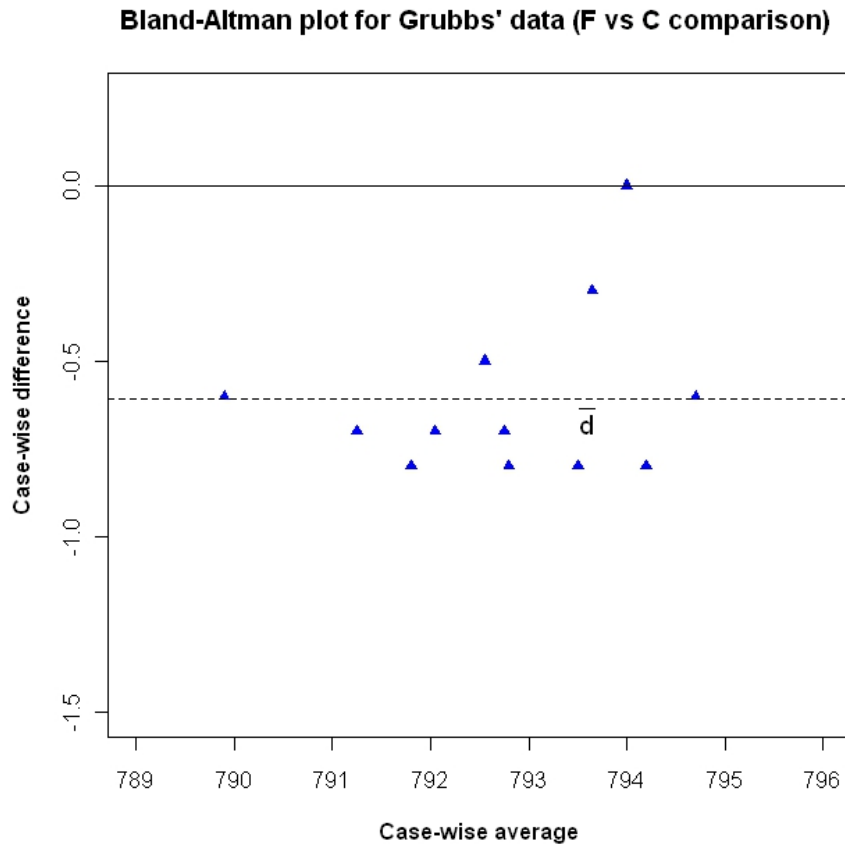


Figure 1.8.4: Bland-Altman plot For Fotobalk and Counter methods.

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

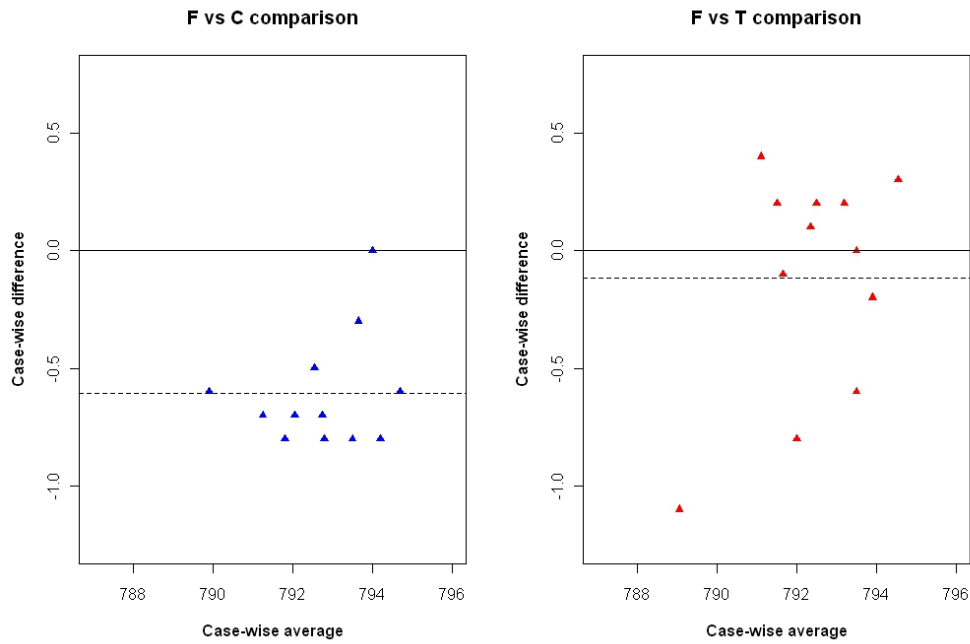


Figure 1.8.5: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

## 1.8.2 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 BlandAltman plot increased over the years, from 8% in 1995 to 14% in 1996, and 3136% in 2002.

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

### 1.8.3 Adverse features

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

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

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

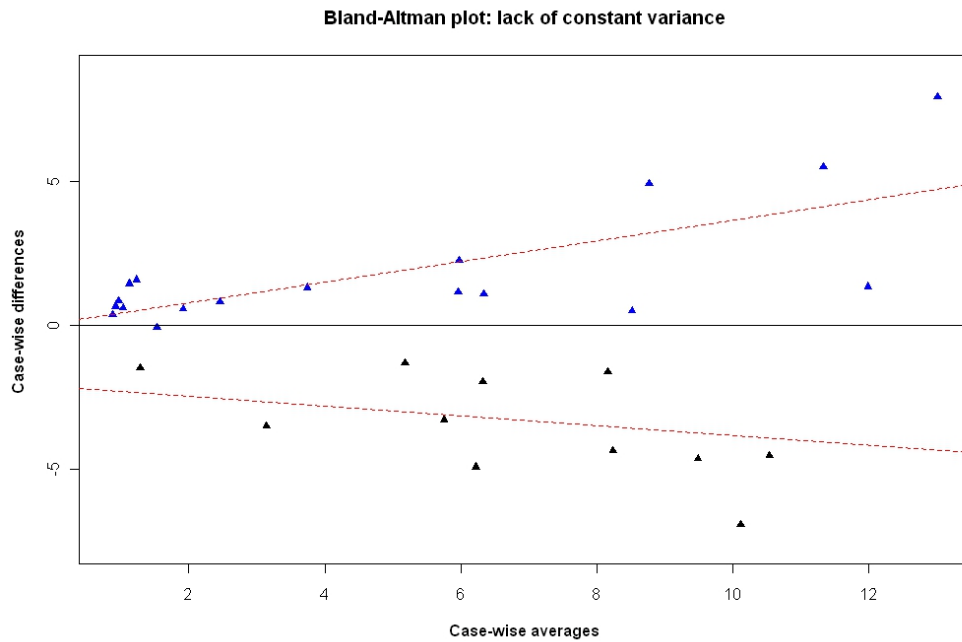


Figure 1.8.6: Bland-Altman plot demonstrating the increase of variance over the range.

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

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

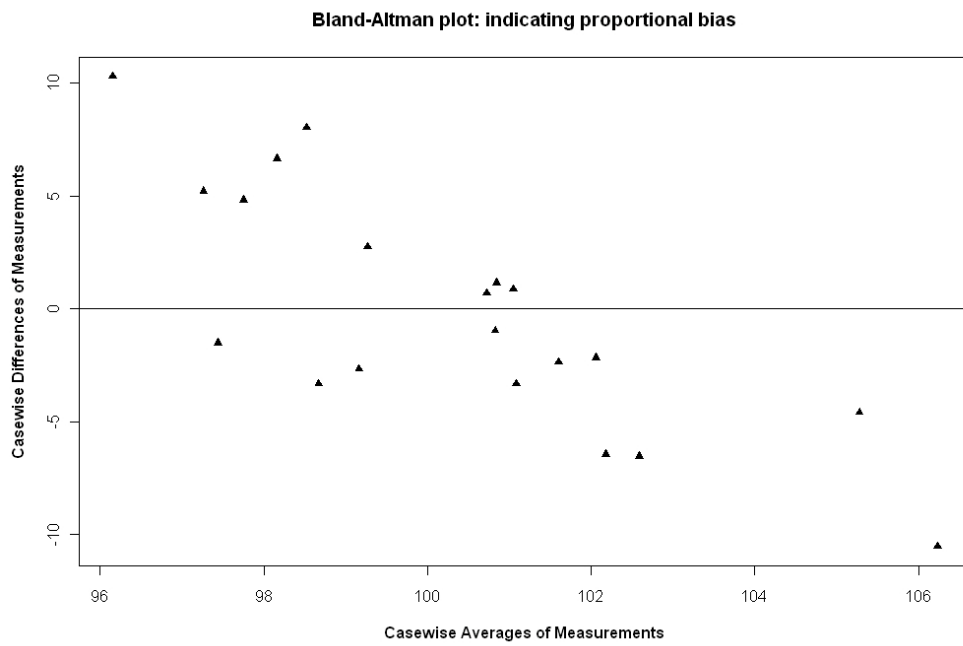


Figure 1.8.7: Bland-Altman plot indicating the presence of proportional bias.

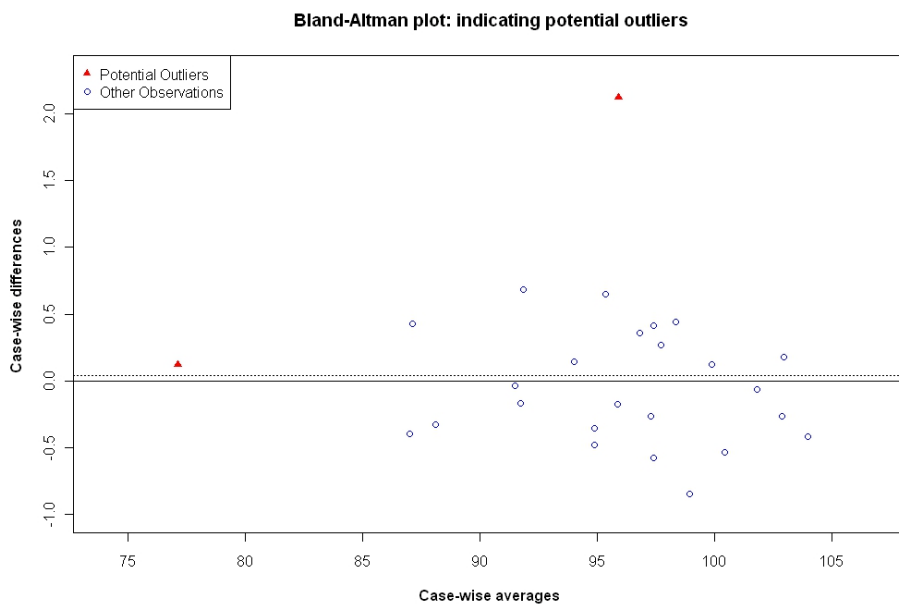


Figure 1.8.8: Bland-Altman plot indicating the presence of potential outliers.

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  $\text{var}(a)$  is greater than  $\text{var}(d)$ , the orientation of the ellipse is horizontal. Conversely if  $\text{var}(a)$  is less than  $\text{var}(d)$ , the orientation of the ellipse is vertical.

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

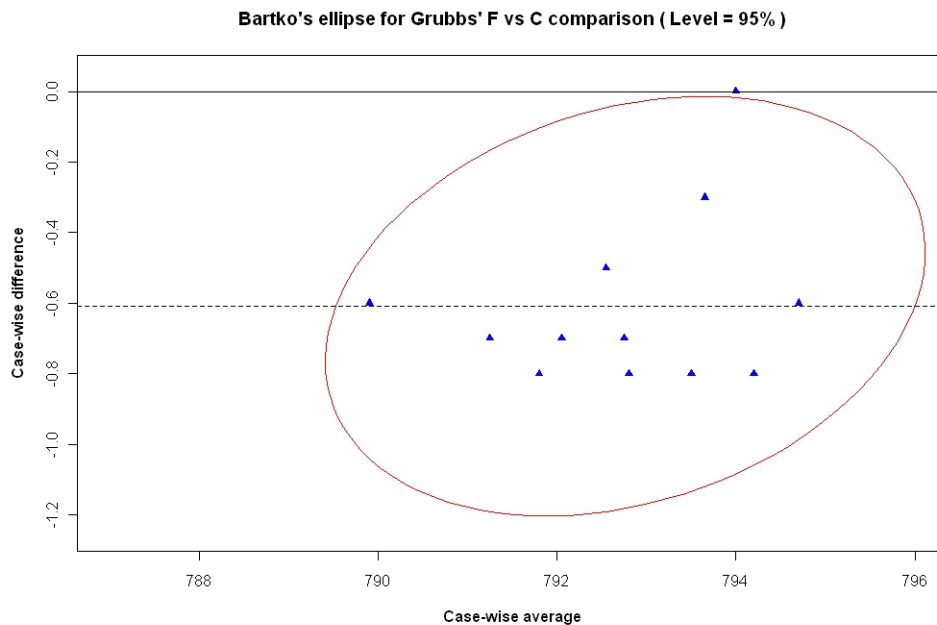


Figure 1.8.9: Bartko's Ellipse For Grubbs' Data.

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

this extra covariate is an outlier, in spite of the fact that this observation is wholly consistent with the conclusion of the Bland-Altman plot.

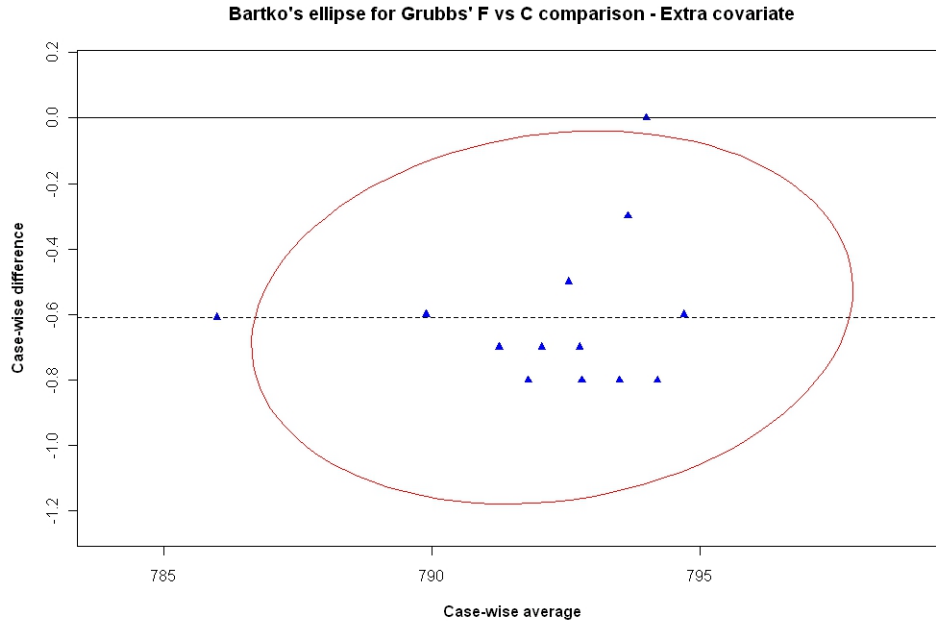


Figure 1.8.10: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

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

In classifying whether a observation from a univariate data set is an outlier, many formal tests are available, such as the Grubbs test for outliers. In assessing whether a covariate in a Bland-Altman plot is an outlier, this test is useful when applied to the case-wise difference values treated as a univariate data set. The null hypothesis of the Grubbs test procedure is the absence of any outliers in the data set. Conversely, the alternative hypotheses is that there is at least one outlier present.



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

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

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

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

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

### 1.8.4 Inferences on Bland-Altman estimates

Bland and Altman (1999) advises on how to calculate confidence intervals for the inter-method bias and limits of agreement. For the inter-method bias, the confidence interval is simply that of a mean:  $\bar{d} \pm t_{(0.5\alpha, n-1)} S_d / \sqrt{n}$ . 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

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

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

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

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

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

# 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.

- Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician* 43(4), 234–235.
- 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.
- Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry* 24(2), 342–345.
- 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.
- IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. <http://goldbook.iupac.org/R05293.html>.
- 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.

- Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry* 44, 1024–1031.
- 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 significance 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.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics* 32(5), 461 – 474.

## 1.9 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). 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:

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

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 accounts of Bland-Altman plots use a multiple of 2 standard deviations instead for simplicity.)

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

“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)”.

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

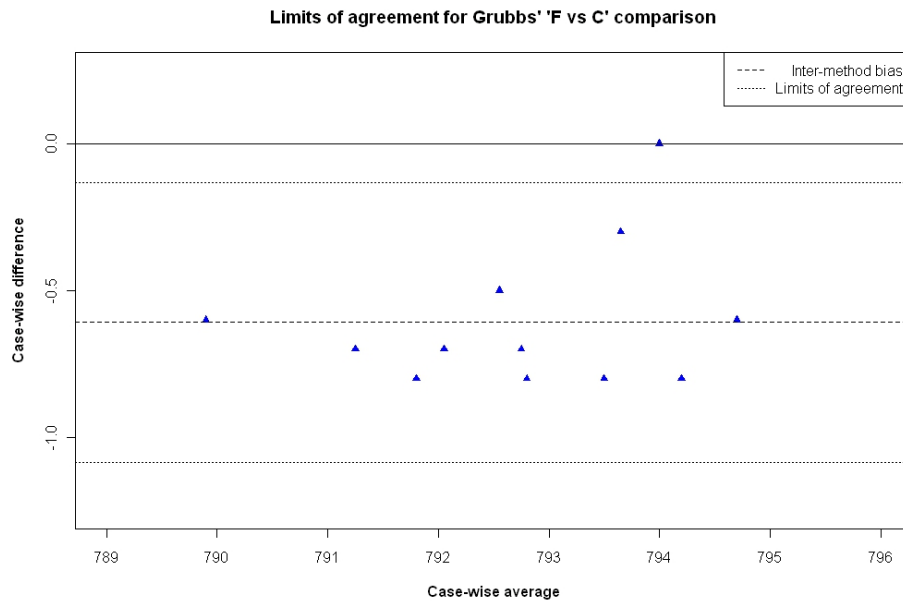


Figure 1.9.11: Bland-Altman plot with limits of agreement

### 1.9.1 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’.

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine

the Shewhart limits 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

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

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 there a quantile less than 2.

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’.

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 mentioned by Mantha et al. (2000).

## 1.9.2 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as ‘replicate mea-



surements’. Carstensen et al. (2008) recommends the use of replicate measurements, but acknowledges the additional computational complexity.

Bland and Altman (1986) address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.

However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. Bland and Altman (1986) propose a correction for this.

Carstensen et al. (2008) takes issue with the limits of agreement based on mean values of replicate measurements, in that they can only be interpreted as prediction limits for difference between means of repeated measurements by both methods, as opposed to the difference of all measurements. Incorrect conclusions would be caused by such a misinterpretation. Carstensen et al. (2008) demonstrates how the limits of agreement calculated using the mean of replicates are ‘much too narrow as prediction limits for differences between future single measurements’. This paper also comments that, while treating the replicate measurements as independent will cause a downward bias on the limits of agreement calculation, this method is preferable to the ‘mean of replicates’ approach.

### **1.9.3 Bland-Altman correlation test**

The approach proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of case-wise differences and means ( $\rho_{AD}$ ). According to the authors, this test is equivalent to the ‘Pitman Morgan Test’. 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 ( $\rho_{AD} = 0$ ) fail to be rejected. Consequently

the null hypothesis of equal variances of each method would also fail to be rejected. There has been no further mention of this particular test in Bland and Altman (1986), although Bland and Altman (1999) refers to Spearman's 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.

### 1.9.4 Identifiability

Dunn (2002) highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example in literature the variance ratio  $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$  must often be assumed to be equal to 1 (Linnet, 1998). Dunn (2002) considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

Bradley and Blackwood (1989) offers a formal simultaneous hypothesis test for the mean and variance of two paired data sets. Using simple linear regression of the differences of each pair against the sums, a line is fitted to the model, with estimates for intercept and slope ( $\hat{\beta}_0$  and  $\hat{\beta}_1$ ). The null hypothesis of this test is that the mean ( $\mu$ ) and variance ( $\sigma^2$ ) of both data sets are equal if the slope and intercept estimates are equal to zero (i.e  $\sigma_1^2 = \sigma_2^2$  and  $\mu_1 = \mu_2$  if and only if  $\beta_0 = \beta_1 = 0$  )

A test statistic is then calculated from the regression analysis of variance values (Bradley and Blackwood, 1989) and is distributed as ' $F$ ' random variable. The degrees of freedom are  $\nu_1 = 2$  and  $\nu_2 = n - 2$  (where  $n$  is the number of pairs). The critical

value is chosen for  $\alpha\%$  significance with those same degrees of freedom. Bartko (1994) amends this methodology for use in method comparison studies, using the averages of the pairs, as opposed to the sums, and their differences. This approach can facilitate simultaneous usage of test with the Bland-Altman methodology. Bartko's test statistic take the form:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Averages	1	0.04	0.04	0.74	0.4097
Residuals	10	0.60	0.06		

Table 1.9.6: Regression ANOVA of case-wise differences and averages for Grubbs Data

For the Grubbs data,  $\Sigma d^2 = 5.09$ ,  $SSReg = 0.60$  and  $MSreg = 0.06$  Therefore the test statistic is 37.42, with a critical value of 4.10. Hence the means and variance of the Fotobalk and Counter chronometers are assumed to be simultaneously equal.

Importantly, this methodology determines whether there is both inter-method bias and precision present, or alternatively if there is neither present. It has previously been demonstrated that there is a inter-method bias present, but as this procedure does not allow for separate testing, no conclusion can be drawn on the comparative precision of both methods.

## 1.10 Regression Methods

Conventional regression models are estimated using the ordinary least squares (OLS) technique, and are referred to as 'Model I regression' (Cornbleet and Cochrane, 1979; Ludbrook, 1997). A key feature of Model I models is that the independent variable is assumed to be measured without error. As often pointed out in several papers (Altman and Bland, 1983; Ludbrook, 1997), this assumption invalidates simple linear

regression for use in method comparison studies, as both methods must be assumed to be measured with error.

The use of regression models that assumes the presence of error in both variables  $X$  and  $Y$  have been proposed for use instead (Cornbleet and Cochrane, 1979; Ludbrook, 1997). These methodologies are collectively known as ‘Model II regression’. They differ in the method used to estimate the parameters of the regression.

Regression estimates depend on formulation of the model. A formulation with one method considered as the  $X$  variable will yield different estimates for a formulation where it is the  $Y$  variable. With Model I regression, the models fitted in both cases will entirely different and inconsistent. However with Model II regression, they will be consistent and complementary.

Regression approaches are useful for a making a detailed examination of the biases across the range of measurements, allowing bias to be decomposed into fixed bias and proportional bias. Fixed bias describes the case where one method gives values that are consistently different to the other across the whole range. Proportional bias describes the difference in measurements getting progressively greater, or smaller, across the range of measurements. A measurement method may have either an attendant fixed bias or proportional bias, or both. (?). Determination of these biases shall be discussed in due course.

## 1.11 Limits of Agreement

Computing limits of agreement features prominently in many method comparison studies, further to Bland and Altman (1986, 1999). 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 (i.e. independent data), 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. ? computes the limits of agreement to the case with replicate measurements by using LME models.

## 1.12 Outline of Thesis

Thus the study of method comparison is introduced. The intention of this thesis is to progress the study of method comparison studies, using a statistical method known as Linear mixed effects models. Chapter two shall describe linear mixed effects models, and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented, using the R programming language.

Model diagnostics are an integral component of a complete statistical analysis. In chapter three model diagnostics shall be described in depth, with particular emphasis on linear mixed effects models, further to chapter two.

For the fourth chapter, important linear mixed effects model diagnostic methods shall be extended to method comparison studies, and proposed methods shall be demonstrated on data sets that have become well known in literature on method comparison. The purpose is to both calibrate these methods and to demonstrate applications for them. The last chapter shall focus on robust measures of important parameters such as agreement.

# 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.

- Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician* 43(4), 234–235.
- 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.
- Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry* 24(2), 342–345.
- 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.
- IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. <http://goldbook.iupac.org/R05293.html>.
- 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.

- Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry* 44, 1024–1031.
- 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 significance 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.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics* 32(5), 461 – 474.



### 1.12.1 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.

## 1.13 Variations of the Bland-Altman Plot

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 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 *log* 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’.

## 1.14 Variations of the Bland Altman Plot

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 casewise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases.

The second variation is a plot of casewise ratios as percentage of averages.

## **Classifying outliers and recalculating**

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?

## 1.15 Multivariate

### 1.15.1 Mahalanobis Distance

The Mahalanobis Distance is a descriptive statistic that provides a relative measure of a data point's distance (residual) from a common point. It is a unitless measure introduced by P. C. Mahalanobis in 1936.[1] The Mahalanobis distance is used to identify and gauge similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant. In other words, it has a multivariate effect size.

## 1.16 Bartko's Ellipse

As a complement to the Bland-Altman plot, *Bartko* proposes the use of a bivariate confidence ellipse, constructed for a predetermined level. *AltmanEllipse* 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  $\text{var}(a)$  is greater than  $\text{var}(d)$ , the orientation of the ellipse is horizontal. Conversely if  $\text{var}(a)$  is less than  $\text{var}(d)$ , 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, 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.

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

In classifying whether a observation from a univariate data set is an outlier, many formal tests are available, such as the Grubbs test for outliers. In assessing whether a covariate in a Bland-Altman plot is an outlier, this test is useful when applied to the case-wise difference values treated as a univariate data set. The null hypothesis of the Grubbs test procedure is the absence of any outliers in the data set. Conversely, the alternative hypotheses is that there is at least one outlier present.

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

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

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

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

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

### 1.16.1 Relevance of Repeatability

Repeatability of two method limit the amount of agreement that is possible.

If one method has poor repeatability, the agreement is bound to be poor. If both methods have poor repeatability, agreement is even worse.

## 1.17 Repeatability

### 1.17.1 Repeatability and gold standards

Currently the phrase ‘gold standard’ describes the most accurate method of measurement available. No other criteria are set out. Further to ?, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer, which is prone to measurement error. Consequently it can be said that a measurement method can be the ‘gold standard’, yet have poor repeatability. Some authors, such as [cite] and [cite] have recognized this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a ‘bronze standard’. Again, no formal definition of a ‘bronze standard’ exists.

The coefficient of repeatability may provide the basis of formulation a formal definition of a ‘gold standard’. For example, by determining the ratio of  $CR$  to the sample mean  $\bar{X}$ . Further to [Lin], it is preferable to have a sample size specified in advance. A gold standard may be defined as the method with the lowest value of  $\lambda = CR/\bar{X}$  with  $\lambda < 0.1\%$ . Similarly, a silver standard may be defined as the method with the lowest value of  $\lambda$  with  $0.1\% \leq \lambda < 1\%$ . Such thresholds are solely for expository purposes.

# Chapter 2

## Coefficient of Repeatability

### 2.1 Add Ins

importance of repeatability 'curiously replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked.

lack of repeatability can interfere with the comparison of two methods because if one method has poor repeatability, in the sense that there is considerable variation in repeated measurements on the same subject, the agreement between two methods is bound to be poor.

### 2.2 Bland and Altman

- Two readings by the same method will be within  $1.96\sqrt{2}\sigma_w$  or  $2.77\sigma_w$  for 95% of subjects. This value is called the repeatability coefficient.
- For observer J using the sphygmomanometer  $\sigma_w = \sqrt{37.408} = 6.116$  and so the repeatability coefficient is  $2 : 77 \times 6.116 = 16 : 95$  mmHg.
- For the machine S,  $\sigma_w = \sqrt{83.141} = 9.118$  and the repeatability coefficient is  $2 : 77 \times 9.118 = 25.27$  mmHg.

- Thus, the repeatability of the machine is 50% greater than that of the observer.

## 2.3 Carstensen

- The limits of agreement are not always the only issue of interest the assessment of method specific repeatability and reproducibility are of interest in their own right.
- Repeatability can only be assessed when replicate measurements by each method are available.
- The repeatability coefficient for a method 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.
- If the standard deviation of a measurement is  $\sigma$  the repeatability coefficient is  $2 \times \sqrt{2}\sigma = 2.83 \times \sigma \approx 2.8\sigma$ .
- The repeatability of measurement methods is calculated differently under the two models
- Under the model assuming exchangeable replicates (1), the repeatability is based only on the residual standard deviation, i.e.  $2.8\sigma_m$
- Under the model for linked replicates (2) there are two possibilities depending on the circumstances.
- If the variation between replicates within item can be considered a part of the repeatability it will be  $2.8\sqrt{\omega^2 + \sigma_m^2}$ .
- However, if replicates are taken under substantially different circumstances, the variance component  $\omega^2$  may be considered irrelevant in the repeatability and one would therefore base the repeatability on the measurement errors alone, i.e. use  $2.8\sigma_m$ .



### 2.3.1 Repeatability

Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement by collecting replicated data. Roy (2009b) notes the lack of convenience in such calculations.

If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (Roy, 2009b).

It is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors (Barnhart et al., 2007).

## 2.4 Reproducibility

It is advisable to be able to distinguish between Repeatability and a similar concept Reproducibility. Reproducibility is

### 2.4.1 2 The Coefficient of Repeatability

Since for the repeated measurements the same method is used, the mean difference should be zero.

Therefore the Coefficient of Repeatability (CR) can be calculated as 1.96 (or 2) times the standard deviations of the differences between the two measurements ( $d_2$  and  $d_1$ ): WRONG

## 2.5 Repeatability

The quality of repeatability is the ability of a measurement method to give consistent results for a particular subject. That is to say that a measurement will agree with prior and subsequent measurements of the same subject.

Repeatability is defined by the IUPAC (2009) as ‘the closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time)’ and is determined by taking multiple measurements on a series of subjects.

Repeatability is important in the context of method comparison because the repeatability of two methods influence the amount of agreement which is possible between those methods. If one method have poor repeatability, then agreement with that method and another will necessarily be poor also. ? and Roy (2009a) highlight the importance of reporting repeatability in method comparison, because it measures the purest random error not influenced by any external factors. Statistical procedures on within-subject variances of two methods are equivalent to tests on their respective repeatability coefficients. A formal test is introduced by Roy (2009a), which will discussed in due course.

## 2.6 Repeatability

A measurement method can be said to have a good level of repeatability if there is consistency in repeated measurements on the same subject using that method. Conversely, a method has poor repeatability if there is considerable variation in repeated measurements.

This is relevant to method comparison studies because the ‘repeatabilities’ of the two methods of measurement affects the level of agreement of those methods. Poor repeatability in one method would result in poor agreement. More so if there is poor

repeatability in both methods.

The British standards Insitute[1979] define a coefficient of repeatability as *the value below which the difference between two single test results..may be expected to lie within a specified probability*.In the absence of other indications, the probability is 95%.

### 2.6.1 Repeatability

Repeatability (or *test-retest reliability*) describes the variation in measurements taken by a single method of measurement on the same item and under the same conditions. A less-than-perfect test-retest reliability causes test-retest variability. Such variability can be caused by, for example, intra-individual variability and intra-observer variability. A measurement may be said to be repeatable when this variation is smaller than some agreed limit. Test-retest variability is practically used, for example, in medical monitoring of conditions. In these situations, there is often a predetermined "critical difference", and for differences in monitored values that are smaller than this critical difference, the possibility of pre-test variability as a sole cause of the difference may be considered in addition to, for examples, changes in diseases or treatments.

According to the Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, the following conditions need to be fulfilled in the establishment of repeatability:

- the same measurement procedure
- the same observer
- the same measuring instrument, used under the same conditions
- the same location
- repetition over a short period of time.

## 2.6.2 Bland and Altman 1999

As noted by Bland and Altman 1999, the repeatability of two methods of measurement can potentially limit Repeatability (using Bland-Altman plot) The Bland-Altman plot may also be used to assess a methods repeatability by comparing repeated measurements using one single measurement method on a sample of items. The plot can then also be used to check whether the variability or precision of a method is related to the size of the characteristic being measured. Since for the repeated measurements the same method is used, the mean difference should be zero. Therefore the Coefficient of Repeatability (CR) can be calculated as 1.96 (often rounded to 2) times the standard deviation of the case-wise differences.

## 2.6.3 Notes from BXC Book (chapter 9)

The assessment of method-specific repeatability and reproducibility is of interest in its own right. Repeatability and reproducibility can only be assessed when replicate measurements by each method are available. If replicate measurements by a method are available, it is simple to estimate the measurement error for a method, using a model with fixed effects for item, then taking the residual standard deviation as measurement error standard deviation. However, if replicates are linked, this may produce an estimate that biased upwards. The repeatability coefficient (or simply repeatability) for a method is defined as the upper limit of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances (see above conditions)

$$y_{mir} = \alpha_m + \beta_m(\mu_i + a_{ir} + c_{mi}) + e_{mir}$$

The variation between measurements under identical circumstances.

# 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.

- Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician* 43(4), 234–235.
- 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.
- Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry* 24(2), 342–345.
- 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.
- IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. <http://goldbook.iupac.org/R05293.html>.
- 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.

- Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry* 44, 1024–1031.
- 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 significance 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.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics* 32(5), 461 – 474.

## 2.7 Coefficient of Repeatability

### 2.7.1 Repeatability

Barnhart emphasizes the importance of repeatability as part of an overall method comparison study. Before there can be good agreement between two methods, a method must have good agreement with itself. The coefficient of repeatability, as proposed by Bland and Altman (1999) is an important feature of both Carstensen's and Roy's methodologies. The coefficient is calculated from the residual standard deviation (i.e.  $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$ ).

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 within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.

### 2.7.2 Coefficient of Repeatability

The Bland Altman Method offers the analyst a measurement on the repeatability of the methods.

The Coefficient of Repeatability (CR) can be calculated as 1.96 (or 2) times the standard deviations of the differences between the two measurements (d2 and d1).

### 2.7.3 Note 1: Coefficient of Repeatability

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 within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.



The Bland Altman method offers the analyst a measurement on the repeatability of the methods.

The Coefficient of Repeatability (CR) can be calculated as 1.96 (or 2) times the standard deviations of the differences between the two measurements.

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 within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.

#### **2.7.4 Repeatability coefficient**

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).

$\sigma_x^2$  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.

### 2.7.5 Repeatability

Barnhart emphasizes the importance of repeatability as part of an overall method comparison study. Before there can be good agreement between two methods, a method must have good agreement with itself. The coefficient of repeatability, as proposed by Bland and Altman (1999) is an important feature of both Carstensen's and Roy's methodologies. The coefficient is calculated from the residual standard deviation (i.e.  $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$ ).

## 2.8 The Identity plot

This is a simple regression based approach. It 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  $X = Y$  line.

## 2.9 Variations of the BA plot

In light of some potential pitfalls associated with the conventional BA plot, a series of alternative formulations for the Bland-Altman plot have been proposed.

## 2.10 Survival Agreement Plot (Luiz et al)

This approach is put forward by *Luiz*. It seeks to extend the agreement evaluation through a graphic approach using step functions' capable of expressing the degree of agreement (or disagreement) as a function of several limits of tolerance.

- It expresses agreement or disagreement as a function of several limits of tolerance.
- Y axis represents the proportion of discordant cases.
- X axis represents the observed differences.

## 2.11 Eksborg's Plot

*Eksborg* proposes a plot of the relative values found by the two Methods being compared (Method 1/Method 2) vs the mean of the Method values.

This approach was discussed as an alternative to the BA approach by

## 2.12 Bartko's Ellipse

As an enhancement on the Bland Altman Plot, *bartko* has expounded a confidence ellipse for the covariates. *bartko* 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.

- The Minor Axis is related to the Variance between-subjects
- The Major Axis is related to the Error Mean Square.

The ellipse illustrates 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  $Var(a_i)$  and the variance of the differences  $Var(d_i)$ .

- If  $Var(a_i)$  is greater than  $Var(d_i)$ , the orientation of the ellipse is horizontal.
- If  $Var(a_i)$  is less than  $Var(d_i)$ , the orientation of the ellipse is vertical.

The more horizontal the ellipse, the greater the ICC.

## 2.13 dewitte et al

Bland ALtman recommend the logarithmic y scale others prefer the precent y scale. generally there is not much difference(except when the data extends over several orders of magnitude) percent method is recommends becuase the numbers can be read directly from the plot without the need for back transfromation.

absolute - small range

percentage - medium range

log scale - large range

we observe increasing use of the bland altman plot over the years, from 8

## 2.14 Mountain Plot

Krouwer and Monti have proposed a folded empirical cumulative distribution plot, otherwise known as a Mountain plot.

They argue that it is suitable for detecting large, infrequent errors. This is a non-parametric method that can be used as a complement with the Bland Altman plot. Mountain plots are created by computing a percentile for each ranked difference between a new method and a reference method. (Folded plots are so called because of the following transformation is performed for all percentiles above 50:  $\text{percentile} = 100 - \text{percentile}$ .) These percentiles are then plotted against the differences between the two methods.

Krouwer and Monti argue that the mountain plot offers some following advantages. It is easier to find the central 95% of the data, even when the data are not normally distributed. Also, comparison on different distributions can be performed with ease.

*A mountain plot (or "folded empirical cumulative distribution plot") is created by computing a percentile for each ranked difference between a new method and a reference method. To get a folded plot, the following transformation is performed for all percentiles above 50:  $\text{percentile} = 100 - \text{percentile}$ . These percentiles are then plotted against the differences between the two methods (Krouwer and Monti, 1995).*

*The mountain plot is a useful complementary plot to the Bland and Altman plot. In particular, the mountain plot offers the following advantages: It is easier to find the central 95% of the data, even when the data are not Normally distributed. Different distributions can be compared more easily.*

The folded cumulative distribution function for a random variable can be easily obtained by folding down the upper half of the cumulative distribution function (CDF). It is a simple graphical method for summarising distributions, and has been used for the evaluation of laboratory assays, clinical trials and quality control (Monti, 1995; Krouwer and Monti, 1995).

A mountain plot (or “folded empirical cumulative distribution plot”) is created by computing a percentile for each ranked difference between a new method and a reference method.

To get a folded plot, the following transformation is performed for all percentiles above 50:  $\text{percentile} = 100 - \text{percentile}$ . These percentiles are then plotted against the differences between the two methods (Krouwer & Monti, 1995). The calculations and plots are simple enough to perform in a spreadsheet.

The mountain plot is a useful complementary plot to the Bland & Altman plot. In particular, the mountain plot offers the following advantages:

- It is easier to find the central 95% of the data, even when the data are not Normally distributed.
- Different distributions can be compared more easily.
- Unlike a histogram, the plot shape is not a function of the intervals.

Compared with the Bland-Altman difference plot, the folded CDF stresses more the median and tails of the difference. If the two assays are unbiased 98 with each other (Krouwer and Monti, 1995), the median would be close to zero.

Bland-Altman and mountain plots each provide complementary perspectives on the data, and the authors recommend both plots.

### **2.14.1 Survival Plots**

- Survivalagreement plots have been suggested as a new graphical approach to assess agreement in quantitative variables. We propose that survival analytical

techniques can complement this method, providing a new analytical insight for agreement.

- Two survivalagreement plots are used to detect the bias between to measurements of the same variable. The presence of bias is tested with log-rank test, and its magnitude with Cox regression.
- An example on C-reactive protein determinations shows how survival analytical methods would be interpreted in the context of assessing agreement.
- Log-rank test, Cox regression, or other analytical methods could be used to assess agreement in quantitative variables; correct interpretations require good clinical sense



# 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.

- Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician* 43(4), 234–235.
- 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.
- Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry* 24(2), 342–345.
- 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.
- IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. <http://goldbook.iupac.org/R05293.html>.
- 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.

- Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry* 44, 1024–1031.
- 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 significance 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.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics* 32(5), 461 – 474.

- Model I regression
- Model II regression

Model I regression [Criterion v Test] [Cornbleet Gochman 1979] define this analysis as the case in which the independent variable, X, is measured without error, with y as the dependent variable.

In method comparison studies, the X variable is a precisely measured reference method. In the [Cornbleet Gochman 1979] paper. It is argued that criterion may be regarded as the correct value. Other papers dispute this.

Model II regression [Test V Test] In this type of analysis, both of the measurement methods are test methods, with both expected to be subject to error. Deming regression is an approach to model II regression.

## 2.14.2 Deming Regression

- The 95% confidence interval for the Intercept can be used to test the hypothesis that  $A=0$ . This hypothesis is accepted if the confidence interval for A contains the value 0.
  - If the hypothesis is rejected, then it is concluded that A is significant different from 0 and both methods differ at least by a constant amount.
  - The 95% confidence interval for the Slope can be used to test the hypothesis that  $B=1$ . This hypothesis is accepted if the confidence interval for B contains the value 1.
  - If the hypothesis is rejected, then it is concluded that B is significant different from 1 and there is at least a proportional difference between the two methods.
- Cochrane Cornbleet
- The authors make the distinction between model I and model II regression types.

- Model II regression is the appropriate type when the predictor variable  $x$  is measured with imprecision. Cornbleet and Cochrane remark that clinical laboratory measurements usually increase in absolute imprecision when larger values are measured.

## Guidelines

Always plot the data. Suspected outliers may be identified from the scatter plot.  $S_{ex}$  represents the precision of a single  $x$  measurement near the mean value of  $X$

$$\lambda = \frac{S_{ex}^2}{S_{ey}^2}$$

# 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.

- Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician* 43(4), 234–235.
- 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.
- Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry* 24(2), 342–345.
- 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.
- IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. <http://goldbook.iupac.org/R05293.html>.
- 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.

- Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry* 44, 1024–1031.
- 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 significance 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.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Ryan, T. P. and W. H. Woodall (2005). The most-cited statistical papers. *Journal of Applied Statistics* 32(5), 461 – 474.