

### 0.0.1 Introduction to Limits of Agreement

- Comparing two methods of measurement is normally done by computing limits of agreement (LoA), i.e. prediction limits for a future difference between measurements with the two methods. When the difference is not constant it is not clear what this means, since the difference between the methods depends on the average; hence, unlike the case where the difference is constant, LoA cannot directly be translated into a prediction interval for a measurement by one method given that of another.
- The main point in the paper by Bland and Altman [1] is however different from the outlook in this paper; Bland and Altman mainly discuss whether two methods of measurement can be used interchangeably and how to assess this with the help of proper statistical methods to derive LoA, i.e. prediction limits for differences between two methods. This paper takes as starting point that the classical LoA can be converted to a prediction interval for one method given a measurement by the other (details in the next section). This sort of relationship can be shown in a plot as a line with slope 1 and prediction limits as lines also with slope 1; applicable for the prediction both from method 1 to method 2 and vice versa. In the case of non-constant difference it would be desirable to be able to produce a similar plot, usable both ways. Thus, the aim of this paper is to produce a conversion from one method to another that also applies in the case where the difference between methods is not constant.
- In this paper, I set up a proper model for data for method comparison studies which in the case of constant difference between methods leads to the classical LoA, and in the case of linear bias gives a simple formula for the prediction. The paper only addresses the situation where only one measurement by each

method is available, although replicate measurements by each method are desirable whenever possible [2]. Moreover, the situation with non-constant variance over the range of measurements is not covered either.

## 0.0.2 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line.

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis

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

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

between measurements by the two methods will lie.”

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

$$LoA = \bar{d} \pm 1.96S(d) \quad (1)$$

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. (However, in some literature, 2 standard deviations are used instead for simplicity.) 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.

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. As Bland and Altman (1986) point out this may not be the case. Bland and Altman advises on how to calculate of confidence intervals for the inter-method bias and the limits of agreement. Importantly the authors recommend prior determination of what would and would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion.

‘How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size.’(Bland and Altman, 1986)

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 limits, the mean and standard deviation, are not based on any sample used for an analysis, but on the process's historical values, a key difference with Bland-Altman limits of agreement.

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

$$\bar{d} \pm t_{(0.975, n-1)} S_d \sqrt{1 + \frac{1}{n}} \quad (2)$$

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

Luiz et al. (2003) describes limits of agreement as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence.

### 0.0.3 Problems with Limits of Agreement

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence a confidence interval, they are not constructed a such. They are designed for future values. The formulation is also heavily influenced by outliers. An Example in Altman and Bland (1983) demonstrates the effect of recalculating without a particular outlier.

Referring to the VCF data set in the same paper, there is more than one outlier.

#### **0.0.4 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.

#### **0.0.5 Limits Of Agreement**

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line.

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis.

The bias is computed as the average of the difference of paired assays.

If one method is sometimes higher, and sometimes the other method is higher, the average of the differences will be close to zero. If it is not close to zero, this indicates that the two assay methods are producing different results systematically.

## **Precision of Limits of Agreement**

The limits of agreement are estimates derived from the sample studied, and will differ from values relevant to the whole population. A different sample would give different limits of agreement. Bland and Altman (1986) advance a formulation for confidence intervals of the inter-method bias and the limits of agreement. These calculations employ quantiles of the 't' distribution with  $n - 1$  degrees of freedom.

### **0.0.6 Appropriate Use of Limits of Agreement**

Importantly Bland and Altman (1999) makes the following point:

These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

Carstensen attends to the issue of repeated data, using the expression replicate to express a repeated measurement on a subject by the same methods. Carstensen formulates the data as follows Repeated measurement - Arrangement of data into groups, based on the series of results of each subject.

### **0.0.7 Appropriate Use of Limits of Agreement**

Importantly Bland and Altman (1999) makes the following point:

These estimates are meaningful only if we can assume bias and variability are uniform throughout the range of measurement, assumptions which can be checked graphically.

The import of this statement is that , should the Bland Altman plot indicate that these assumptions are not met, then their entire methodology, as posited thus far, is inappropriate for use in a method comparison study. Again, in the context of potential outlier in the Grubbs data (figure 1.2), this raises the question on how to correctly continue.

measurements

## 0.1 Limits of Agreement

A third element of the Bland-Altman approach, an interval known as ‘limits of agreement’ is introduced in Bland and Altman (1986) (sometimes referred to in literature as 95% limits of agreement). 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 question to which limits of agreement are intended as the answer to must be established clearly. Bland and Altman (1995) comment that the limits of agreement show ‘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 (sometimes rounded to 2) is the 95% quantile for the standard normal distribution. 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 constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion. However Mantha et al. (2000) highlight inadequacies in the correct application of limits of agreement, resulting in contradictory estimates of limits of agreement in various papers.

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.

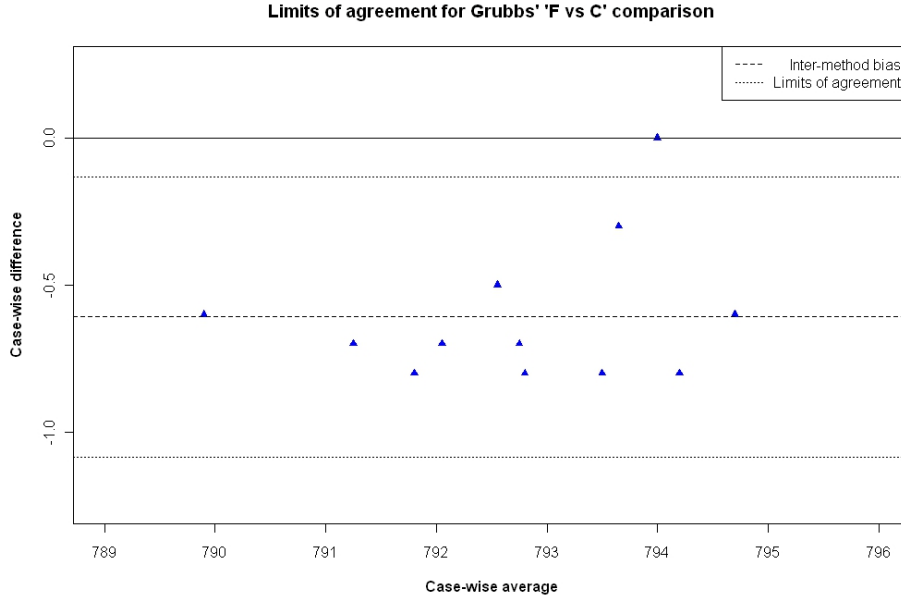


Figure 1: Bland-Altman plot with limits of agreement



### 0.1.1 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:  $\bar{d} \pm t_{(\alpha/2, n-1)} S_d / \sqrt{n}$ . The confidence intervals and standard error for the limits of agreement follow from the variance of the limits of agreement, which is shown to be

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

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.

### 0.1.2 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.

### 0.1.3 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).

#### 0.1.4 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 they were Shewhart control limits.

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

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

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 the 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).

# Bibliography

- 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.
- Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* i, 307–310.
- Bland, J. and D. Altman (1995). Comparing methods of measurement - why plotting difference against standard method is misleading. *The Lancet* 346, 1085–87.
- Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160.
- Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 4(1).
- 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.