Chapter 1

Limits of Agreement

1.1 Limits of Agreement

A third element of the Bland-Altman methodology, an interval known as 'limits of agreement' is introduced in **?** (sometimes referred to in literature as 95% limits of agreement). Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference line and are 1.96 times the standard deviation above and below the average difference line. Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably. **?** refer to this as the 'equivalence' of two measurement methods. The specific purpose of the limits of agreement must be established clearly. **?** comment that the limits of agreement '*how far apart measurements by the two methods were likely to be for most individuals*', a definition echoed in their 1999 paper:

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

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

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

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

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

> "How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size (**?**)".

However **?** highlights inadequacies in the correct application of limits of agreement, resulting in contradictory estimates for limits of agreement in various papers.

For the Grubbs 'F vs C' comparison, these limits of agreement are calculated as -0.132 for the upper bound, and -1.08 for the lower bound. Figure 1.1.1 shows the resultant Bland-Altman plot, with the limits of agreement shown in dashed lines.

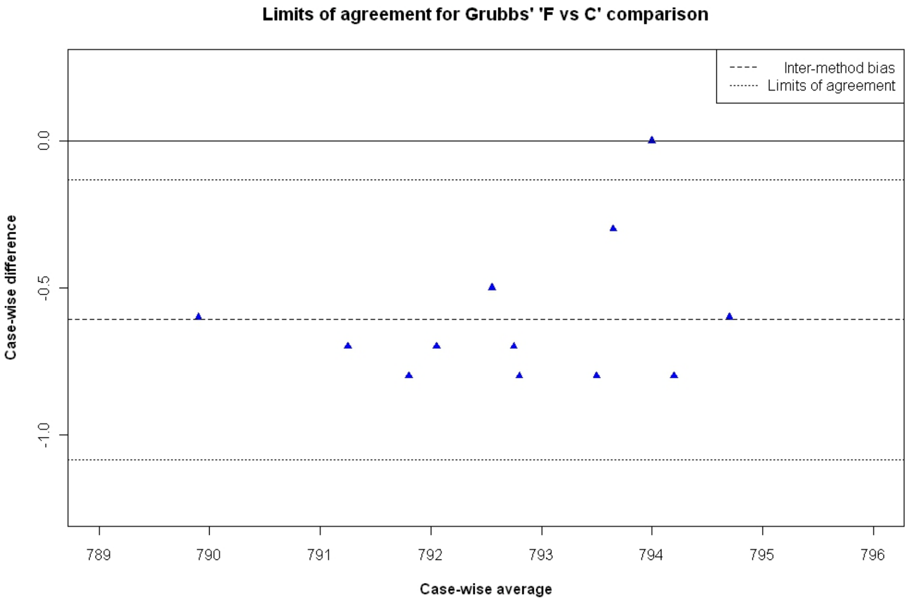

Figure 1.1.1: Bland-Altman plot with limits of agreement

1.2 Interpretation of Limits Of Agreement

How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable.

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

1.2.1 Formal Definition of Limits of Agreement

? 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.

? 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. **?** 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.

Various other interpretations as to how limits of agreement should properly be defined. **?** 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. **?** 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 discussed by **?**.

1.3 Limits of Agreement for Replicate Measurements

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

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. ? advance a formulation for confidence intervals of the inter-method bias and the limits of agreement. These calculations employ quantiles of the 't' distribution with $n - 1$ degrees of freedom.

1.3.1 Appropriate Use of Limits of Agreement

Importantly ? makes the following point:

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

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

1.4 Coefficient of Repeatability

1.4.1 Repeatability

As mentioned previously, **?** emphasizes the importance of repeatability as part of an overall method comparison study. The coefficient of repeatability was proposed by **?**, and is referenced in subsequent papers, such as **?**. The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (**?**). The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (**?**). Once the the standard deviations of the differences between the two measurements (in some texts called the residual standard deviation or within-item variability) $sigma_m$ is determined, the computation of the coefficients of repeatability for both methods is straightforward. The coefficient is calculated from the (in some texts called the residual standard deviation) as $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$).

1.4.2 Repeatability coefficient

? 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 (**?**).

σ^2_x 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.