Contents

Chapter 1

Method Comparison Studies

1.1 Introduction

The problem of assessing the agreement between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a 'method comparison study'. Published examples of method comparison studies can be found in disciplines as diverse as pharmacology (Ludbrook, 1997), anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000).

To illustrate the characteristics of a typical method comparison study consider the data in Table I (Grubbs, 1973). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm gun and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels 'Fotobalk', 'Counter' and 'Terma'.

Round	Fotobalk [F]	Counter [C]	Terma [T]
1	793.8	794.6	793.2
2	793.1	793.9	793.3
3	792.4	793.2	792.6
4	794.0	794.0	793.8
5	791.4	792.2	791.6
6	792.4	793.1	791.6
7	791.7	792.4	791.6
8	792.3	792.8	792.4
9	789.6	790.2	788.5
10	794.4	795.0	794.7
11	790.9	791.6	791.3
12	793.5	793.8	793.5

Table 1.1.1: Velocity measurement from the three chronographs (Grubbs 1973).

An important aspect of the these data is that all three methods of measurement are assumed to have an attended measurement error, and the velocities reported in Table 1.1 can not be assumed to be 'true values' in any absolute sense.

A method of measurement should ideally be both accurate and precise. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown 'true value'. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise, if the average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely a method that is not accurate may be quite precise, as it consistently indicates the same level of inaccuracy. The tendency of a method of measurement to consistently give

results above or below the true value is a source of systematic bias. The smaller the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero.

1.1.1 Statement of a Model

Carstensen (2010) presents a useful formulation for comparing two methods X and Y, in their measurement of item i, where the unknown 'true value' is τ_i. Other authors, such as Kinsella (1986), present similar formulations of the same model, as well as modified models to account for multiple measurements by each methods on each item, known as replicate measurements.

$$X_i = \tau_i + \delta_i, \qquad \delta_i \sim \mathcal{N}(0, \sigma_\delta^2) \tag{1.1}$$

$$Y_i = \alpha + \beta\tau_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{1.2}$$

In some types of analysis, such as the conversion problems described by Lewis et al. (1991), an estimate for the scaling factor β may also be sought. For the time being, we will restrict ourselves to problems where β is assumed to be 1.

$$X_i = \tau_i + \delta_i, \qquad \delta_i \sim \mathcal{N}(0, \sigma_\delta^2) \tag{1.3}$$

$$Y_i = \alpha + \beta\tau_i + \epsilon_i, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2) \tag{1.4}$$

In this formulation, α represents the inter-method bias, and can be estimated as $E(X - Y)$. That is to say, a simple estimate of the inter-method bias is given by the differences between pairs of measurements. Table 1.1.2 is a good example of possible

inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. A cursory inspection of the table will indicate a systematic tendency for the Counter method to result in higher measurements than the Fotobalk method.

The absence of inter-method bias is, by itself, not sufficient to establish that two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than those of the other, they can not be considered to be in agreement. Hence, method comparison studies are required to take account of both inter-method bias and difference in precision of measurements.

Round	Fotobalk (F)	Counter (C)	Difference (F-C)
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.1.2: Difference between Fotobalk and Counter measurements.

Even without computing the mean difference, a cursory examination of the table will indicate that one method consistently provides a measurement less than the other.

1.2 Purpose of Method Comparison Studies

Carstensen (2010) provides a review of many descriptions of the purpose of Method Comparison studies, several of which are reproduced here.

> "The question being answered is not always clear, but is usually epxressed as an attempt to quantify the agreement between two methods" (Bland and Altman, 1995).
>
> "Some lack of agreement between different methods of measurement is inevitable. What matters is the amount by which they disagree. We want to know by how much the new method is likely to differ from the old, so that it is not enough to cause problems in the mathematical interpretation we can preplace the old method by the new, or even use the two interchangeably" (Bland and Altman, 1999).
>
> "It often happens that the same physical and chemical property can be measured in different ways. For example, one can determine For example, one can determine sodium in serum by flame atomic emission spectroscopy or by isotope dilution mass spectroscopy. The question arises as to which method is better" (Mandel, 1991).
>
> "In areas of inter-laboratory quality control, method comparisons, assay validations and individual bio-equivalence, etc, the agreement between observations and target (reference) values is of interest" (Lin et al., 2002).
>
> "The purpose of comparing two methods of measurement of a continuous biological variable is to uncover systematic differences, not to point to similarities'" (Ludbrook, 1997).
>
> "In the pharmaceutical industry, measurement methods that measure the quantity of prdocuts are regulated. The FDA (U.S. Food and Drug Administration) requires that the manufacturer show equivalency prior to approving the new or alternatice method in quality control" (Tan & Inglewicz,

1999).

While several major commonalities are present in each definitions, there is a different emphasis for each, which will inevitably give rise to confusion. Carstensen (2010) seems to endorse a simple phrasing of the research question that is proposed by Altman and Bland (1983), i.e. "*do the two methods of measurement agree sufficiently closely?*" with Carstensen (2010) expressing the view that other considerations (for example, the "equivalence" of two methods) to be treated as separate research questions. As such, we will revert to other research questions, such as "equivalence of methods" later, focussing on agreement and repeatability of methods.

1.3 Repeatability

Repeatability is the ability of a measurement method to give consistent results for a particular subject, i.e. a measurement will agree with prior and subsequent measurements of the same subject. Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study, a view endorsed by Carstensen et al. (2008). Before there can be good agreement between two methods, a method must have good agreement with itself. If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (Roy, 2009b). Barnhart et al. (2007) remarks that it is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors, while further remarking '*curiously replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked.* Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. However Roy (2009b) notes the lack of convenience in such calculations. 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.

A measurement is said to be repeatable when this variation is smaller than some pre-specified limit. 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.

The British Standards Institute (1979) defines 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%.

1.3.1 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 Dunn (2002), 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.

Dunn (2002) recognizes 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 $\bar{X}$. Advisably the sample size should 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 λ with $0.1\% \leq \lambda < 1\%$. Such thresholds are solely for expository purposes.

Chapter 2

Review of Current Methodologies

2.1 Bland-Altman Approach

The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically, comparison of two methods of measurement was carried out by use of paired sample $t-$test, correlation coefficients or simple linear regression. However, simple linear regression is unsuitable for method comparison studies due to the assumption that one variable is measured without error. In comparing two methods, both methods are assume to have attendant random error.

Altman and Bland (1983) highlighted the inadequacies of these approaches for comparing two methods of measurement, and proposed methodologies with this specific application in mind. Although the authors also acknowledge the opportunity to apply other, more complex, approaches, but argue that simpler approaches is preferable, especially when the results must be 'explained to non-statisticians'.

Notwithstanding previous remarks about linear regression, the first step recommended, which the authors argue should be mandatory, is the construction of a scatter plot of the data. Scatterplots can facilitate an initial judgement and helping to identify potential outliers, with the addition of the line of equality. In the case of good agree-

ment, the observations would be distributed closely along this line. However, they are not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation.

A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that inter-method bias is present, i.e. the Fotobalk device has a tendency to record a lower velocity.

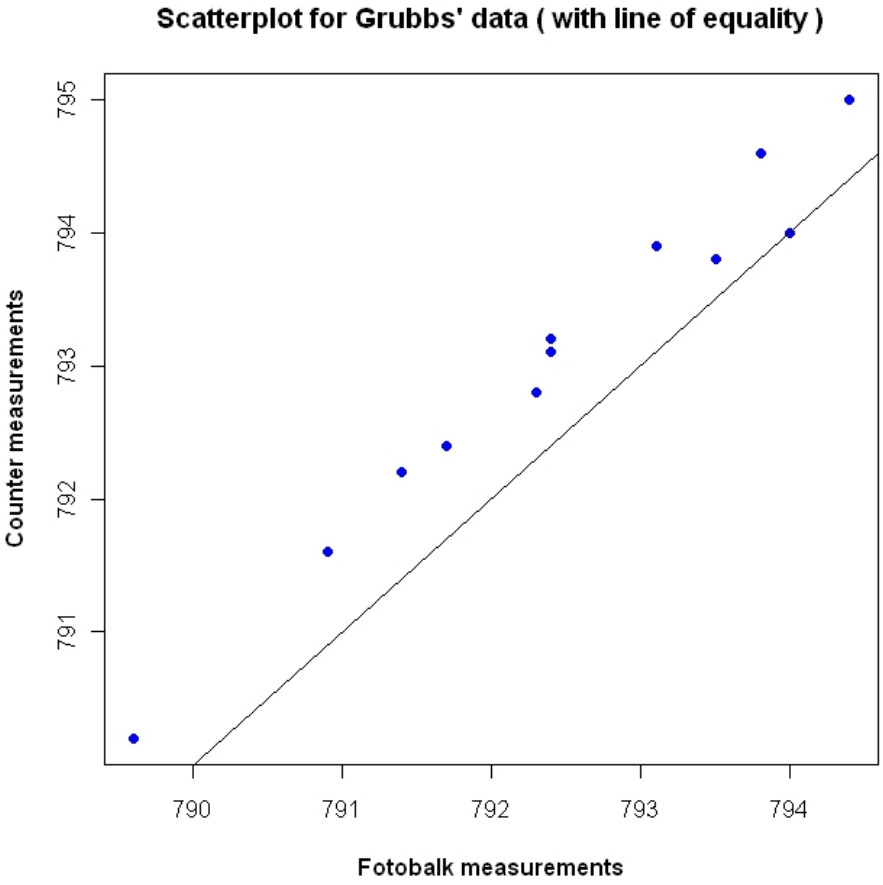

Figure 2.1.1: Scatter plot for Fotobalk and Counter methods.

Dewitte et al. (2002) notes that scatter plots were very seldom presented in the Annals of Clinical Biochemistry. This apparently results from the fact that the 'Instructions for Authors' dissuade the use of regression analysis, which conventionally is

accompanied by a scatter plot.

2.1.1 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 $i = 1, 2, \ldots, n$, on the same subject should be calculated, and then the average of those measurements, $(a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, \ldots, n$.

Following a technique known as the Tukey mean-difference plot, as noted by Kozak and Wnuk (2014) Altman and Bland (1983) proposed that a_i should be plotted against d_i, a plot now widely known as the Bland-Altman plot, and motivated this plot as follows:

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

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

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

relevant. The variances around this bias is estimated by the standard deviation of these differences S_d.

Rendering a Bland-Altma plot

Construction of a Bland-Altman plot can be implemented easily with `R` packages such as Bendix Carstensen's `MethComp` package, which is designed to *provide computational tools to manipulate, display and analyze data from method comparison studies* (Carstensen, 2010).

2.1.2 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 2.1.1: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.8	793.2	0.6	793.5
2	793.1	793.3	-0.2	793.2
3	792.4	792.6	-0.2	792.5
4	794.0	793.8	0.2	793.9
5	791.4	791.6	-0.2	791.5
6	792.4	791.6	0.8	792.0
7	791.7	791.6	0.1	791.6
8	792.3	792.4	-0.1	792.3
9	789.6	788.5	1.1	789.0
10	794.4	794.7	-0.3	794.5
11	790.9	791.3	-0.4	791.1
12	793.5	793.5	0.0	793.5

Table 2.1.2: Fotobalk and Terma methods: differences and averages.

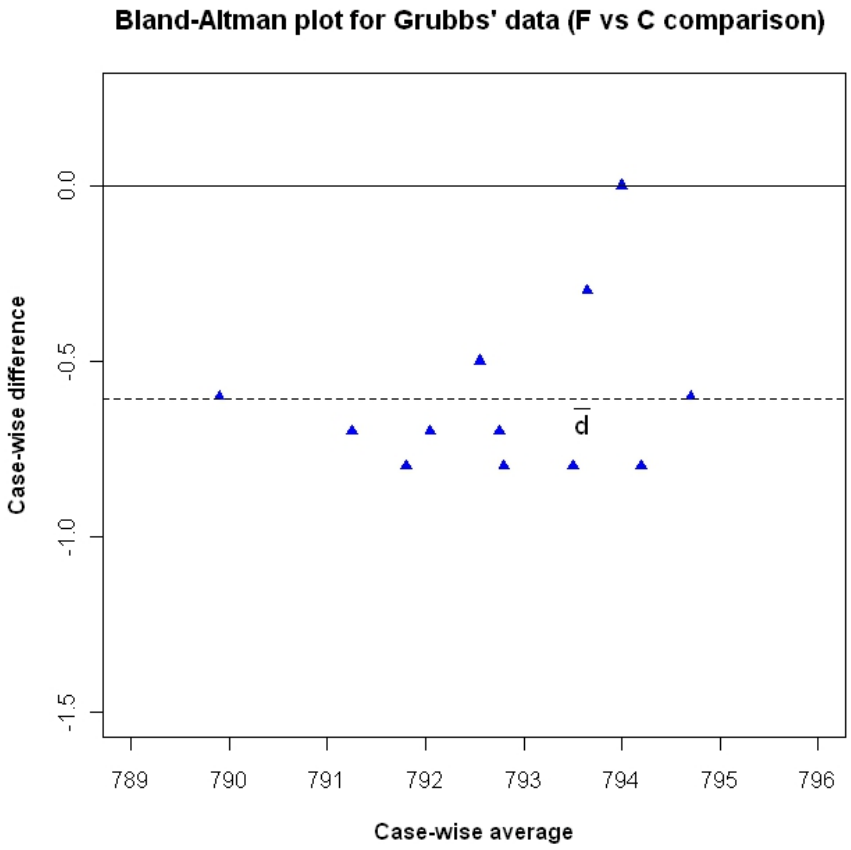

Figure 2.1.2: Bland-Altman plot For Fotobalk and Counter methods.

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.

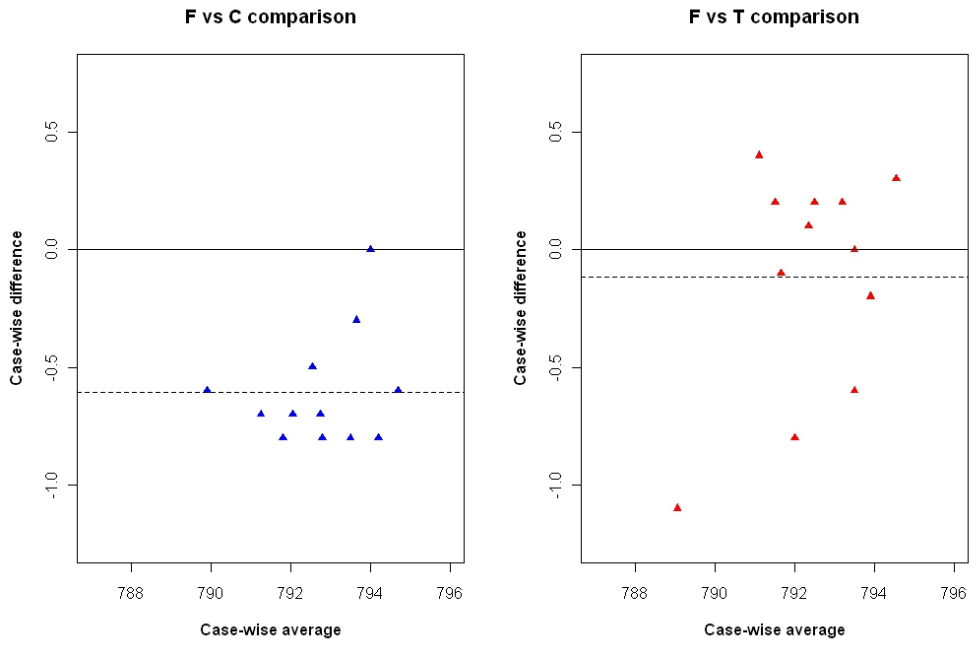

Figure 2.1.3: Bland-Altman plots for Grubbs' F vs C and F vs T comparisons.

2.1.3 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 approach.

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

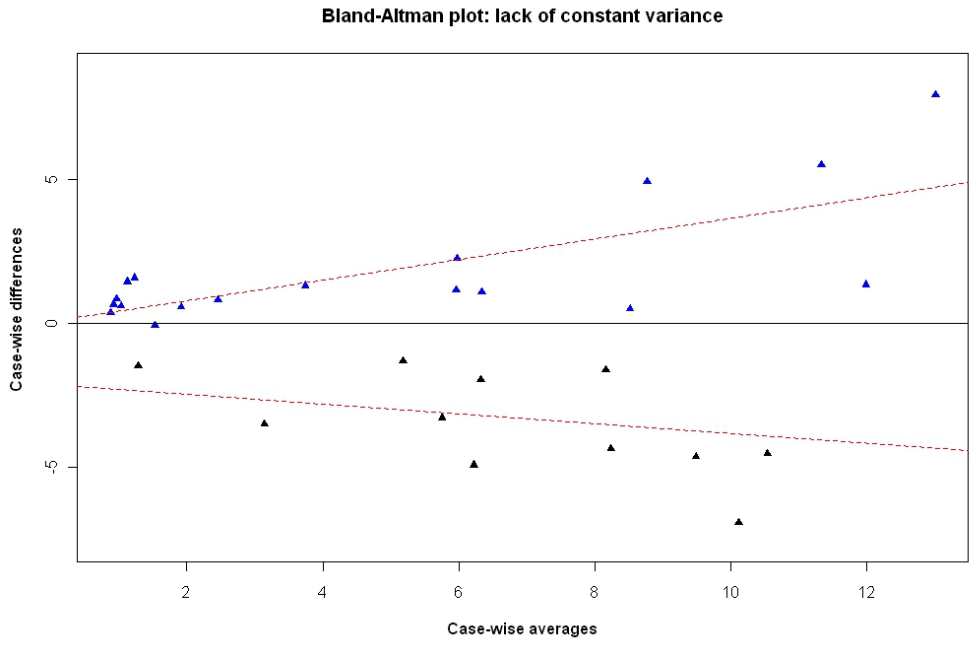

Figure 2.1.4: Bland-Altman plot demonstrating the increase of variance over the range.

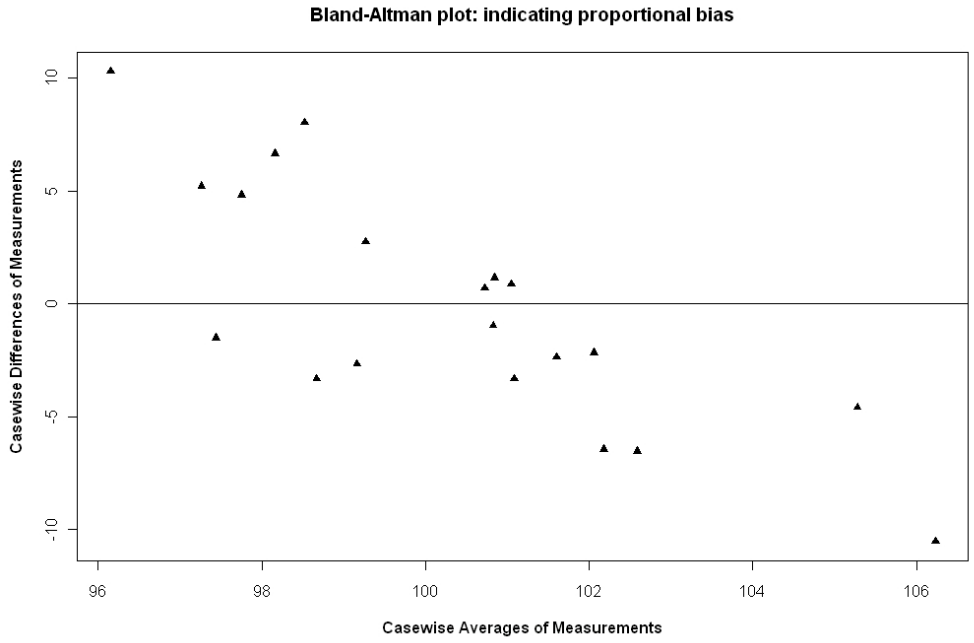

Figure 2.1.5: Bland-Altman plot indicating the presence of proportional bias.

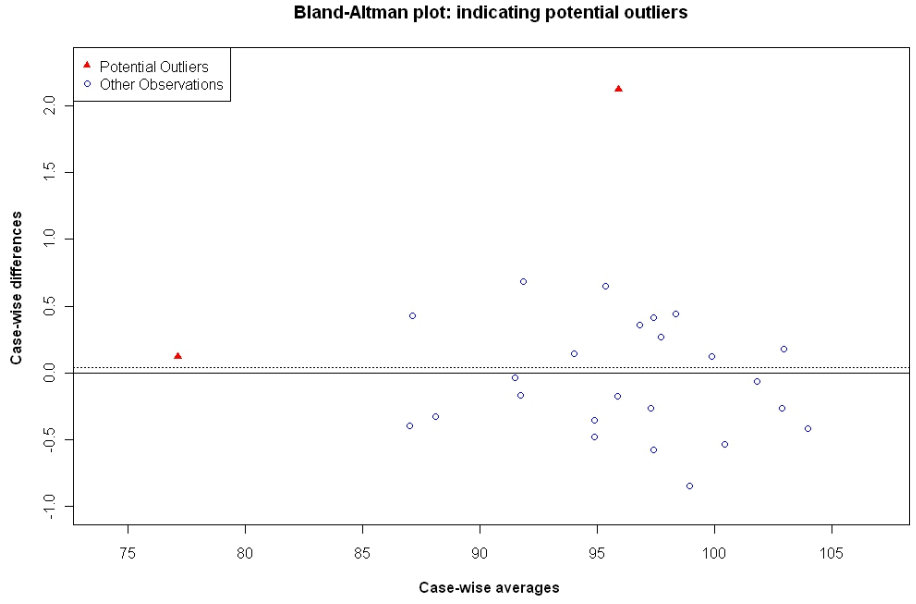

Figure 2.1.6: Bland-Altman plot indicating the presence of potential outliers.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. Bland and Altman (1999) do not recommend excluding outliers from analyses, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that 'we usually find that this method of analysis is not too 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.

The minor axis relates to the between subject variability, whereas the major axis relates to the error mean square, with the ellipse depicting the size of both relative to each other. Consequently Bartko's ellipse provides a visual aid to determining the relationship between variances. If var(a) is greater than var(d), the orientation of the ellipse is horizontal. Conversely if var(a) is less than var(d), the orientation of the ellipse is vertical.

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

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

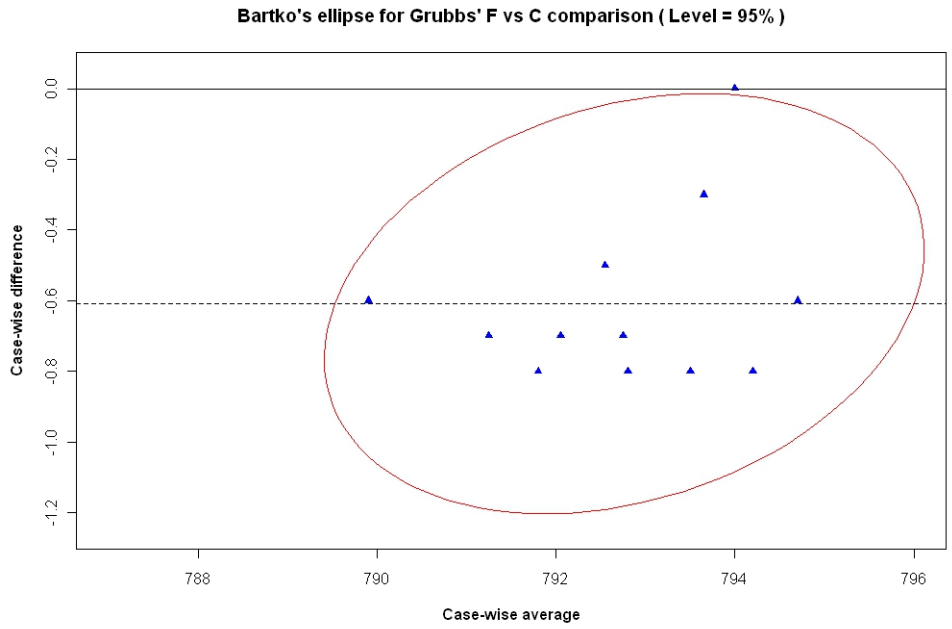

Figure 2.1.7: Bartko's Ellipse for Grubbs' data.

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

Importantly, outlier classification must be informed by the logic of the mechanism that produces the data. In the Bland-Altman plot, the horizontal displacement (i.e. the average) of any observation is supported by two separate 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.

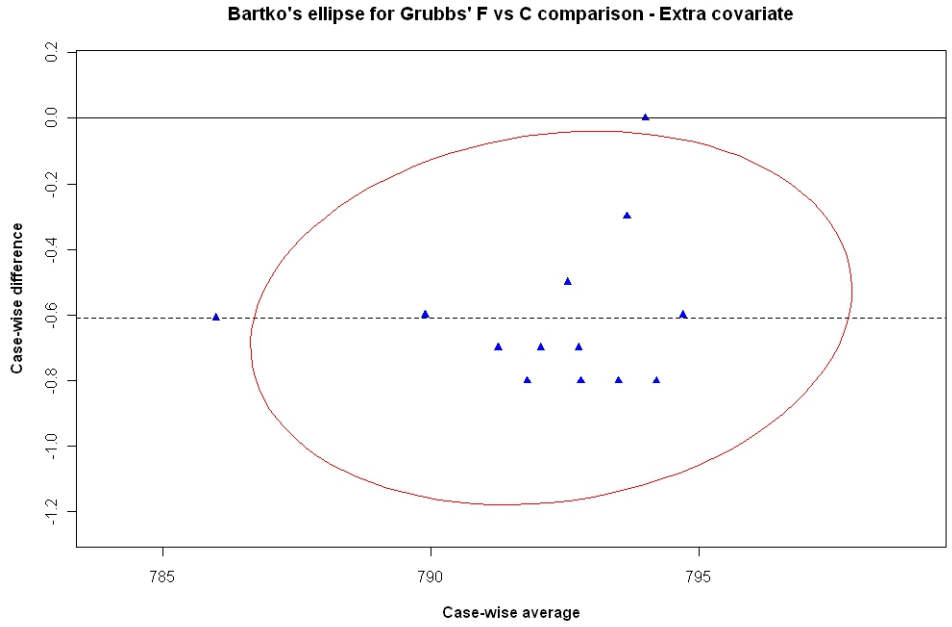

Figure 2.1.8: Bartko's Ellipse for Grubbs' data, with an extra covariate.

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

$$G = \max_{i=1,\ldots,n} \frac{|d_i - \bar{d}|}{S_d}. \tag{2.1}$$

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

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

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

2.2 Limits of Agreement

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

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

with $\bar{d}$ as the estimate of the inter method bias, s_d as the standard deviation of the differences and 1.96 (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.

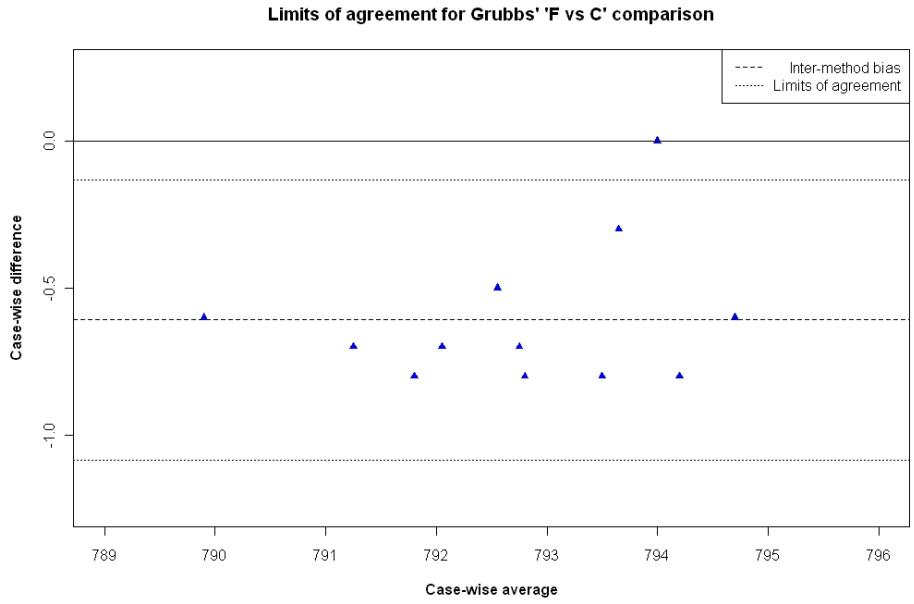

Figure 2.2.9: Bland-Altman plot with limits of agreement

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

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

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

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

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.

2.2.2 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).

2.2.3 Alternative Agreement Indices

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

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

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

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

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

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

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

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

Table 2.2.3: Example data set

To illustrate the use of EAD, consider table 2.2.3. The inter-method bias is 0.03, which is quite close to zero, which is desirable in the context of agreement. However, an identity plot would indicate very poor agreement, as the points are noticeably distant from the line of equality.

The limits of agreement are $[-9.61, 9.68]$, a wide interval for this data. As with the identity plot, this would indicate lack of agreement. As with inter-method bias, an EAD value close to zero is desirable. However, from table 2.2.3, the EAD can be computed as 3.71. The Bland-Altman plot remains a useful part of the analysis. In 2.2.11, it is clear there is a systematic decrease in differences across the range of measurements.

Barnhart et al. (2007) remarks that a comparison of EAD and MSD , using simulation studies, would be interesting, while further adding that '*It will be of interest to investigate the benefits of these possible new unscaled agreement indices*'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously

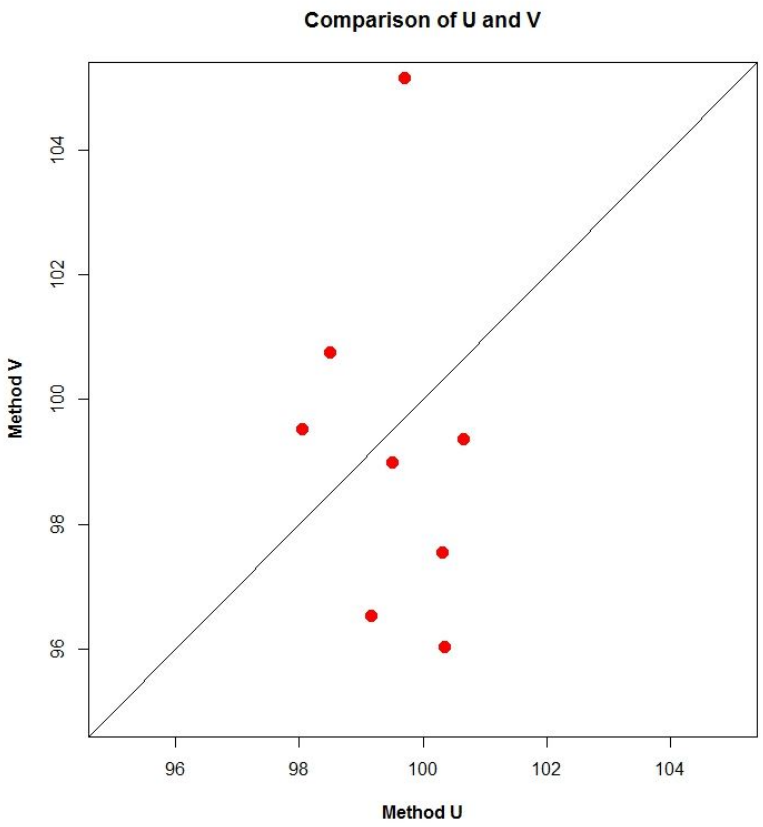

Figure 2.2.10: Identity Plot for example data

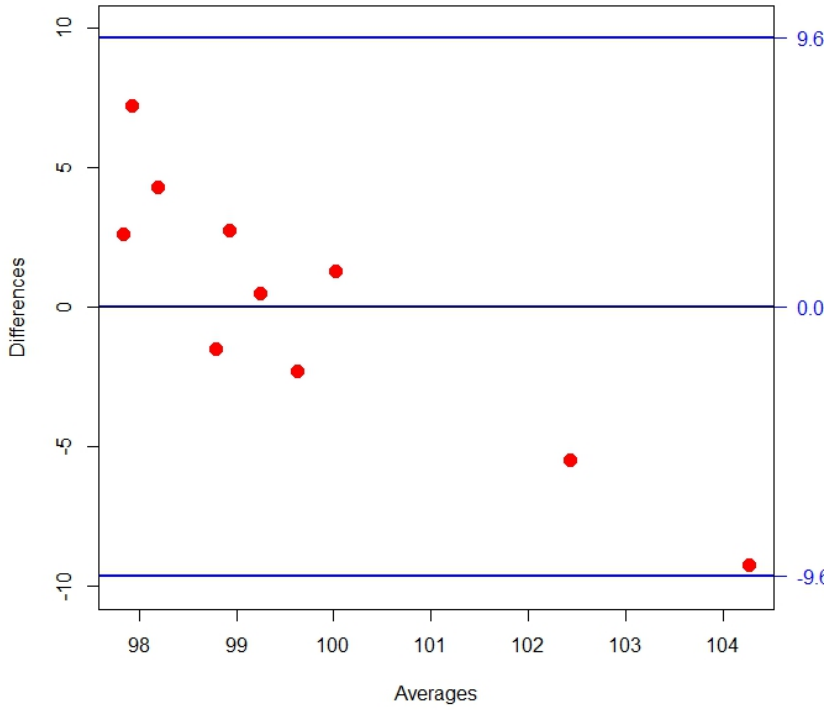

Figure 2.2.11: Bland-Altman Plot for UV comparison

on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. Hence the EAD values for both comparisons are much closer.

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

Table 2.2.4: Agreement indices for Grubbs' data comparisons.

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

$$CP = P(|x_i - y_i| \leq d_0) \tag{2.2}$$

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

2.2.4 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.

Chapter 3

Improper MCS Techniques

3.0.5 Paired sample T-test

Bartko (1994) discusses the use of the well known paired sample t test to test for inter-method bias; $H : \mu_D = 0$. The test statistic is distributed a t random variable with $n - 1$ degrees of freedom and is calculated as follows;

$$t^* = \bar{D} / \frac{S_D}{\sqrt{n}} \tag{3.1}$$

where $\bar{D}$ and S_D is the average of the differences of the n observations.

- Paired t tests test only whether the mean responses are the same. Certainly, we want the means to be the same, but this is only a small part of the story. The means can be equal while the (random) differences between measurements can be huge.

- The correlation coefficient measures linear agreement–whether the measurements go up-and-down together. Certainly, we want the measures to go up-and-down together, but the correlation coefficient itself is deficient in at least three ways as a measure of agreement. The correlation coefficient can be close to 1 (or equal to 1!) even when there is considerable bias between the two methods. For example, if one method gives measurements that are always 10 units higher than the other method, the correlation will be 1 exactly, but the measurements will always be 10 units apart.

- The magnitude of the correlation coefficient is affected by the range of subject-s/units studied.

- The correlation coefficient can be made smaller by measuring samples that are similar to each other and larger by measuring samples that are very different from each other. The magnitude of the correlation says nothing about the magnitude of the differences between the paired measurements which, when you get right down to it, is all that really matters.

- The usual significance test involving a correlation coefficient– whether the population value is 0–is irrelevant to the comparability problem. What is important is not merely that the correlation coefficient be different from 0. Rather, it should be close to (ideally, equal to) 1!

intra-class correlation coefficient

- The intra-class correlation coefficient has a name guaranteed to cause the eyes to glaze over and shut the mouth of anyone who isn't an analyst. The ICC, which takes on values between 0 and 1, is based on analysis of variance techniques. It is close to 1 when the differences between paired measurements is very small compared to the differences between subjects. Of these three procedures–t test, correlation coefficient, intra-class correlation coefficient–the ICC is best because it can be large only if there is no bias and the paired measurements are in good agreement, but it suffers from the same faults ii and iii as ordinary correlation coefficients. The magnitude of the ICC can be manipulated by the choice of samples to split and says nothing about the magnitude of the paired differences.

Regression Methods

- Regression analysis is typically misused by regressing one measurement on the other and declare them equivalent if and only if the confidence interval for the regression coefficient includes 1. Some simple mathematics shows that if the measurements are comparable, the population value of the regression coefficient will be equal to the correlation coefficient between the two methods. The population correlation coefficient may be close to 1, but is never 1 in practice. Thus, the only things that can be indicated by the presence of 1 in the confidence interval for the regression coefficient is (1) that the measurements are comparable but there weren't enough observations to distinguish between 1 and the population regression coefficient, or (2) the population regression coefficient is 1 and therefore, the measurements aren't comparable.

- There is a line whose slope will be 1 if the measurements are comparable. It is known as a structural equation and is the method advanced by Kelly (1985). Altman and Bland (1987) criticize it for a reason that should come as no surprise: Knowing the data are consistent with a structural equation with a slope of 1 says something about the absence of bias but *nothing* about the variability about Y = X (the difference between the measurements), which, as has already been stated, is all that really matters.

3.1 The Technology Acceptance Model

Davis (1989) proposes the TAM model, which suggests an hypothesis as to why users may adopt particular technologies, and not others. According to this theory, when users are presented with a new technology, two important factors will influence their decision about how and when they will adopt it.

Perceived usefulness (PU) - This was defined by Fred Davis as "the degree to which a person believes that using a particular system would enhance his or her

job performance".

Perceived ease-of-use (PEOU) - Davis defined this as "the degree to which a person believes that using a particular system would be free from effort"

Davis's explanations of these term can be rephrased for application to statistical analysis. Perceived Use could refer to the degree to which an user would deem a particular statistical method would properly establish the results of an analaysis. In the case of method comparison studies, proper indication of agreement, or lack thereof.

Perceived ease-of-use requires only applying the context of a satistical problem. A very modest statistical skill set is the only prerequistive for constructing a Bland-Altman plot, and computing limits of agreement. The main building blocks are simple descriptive, statistics and a knowledge of the normal distribution. These are topics that feature in almost every undergraduate statistics courses. Furthermore **?** recommends including the Bland-Altman method itself in undergraduate teaching.

In short, the user perceives the Bland-Altman methodology to be an easy-to-implement technique, that will properly address the question of agreement.

Conversely the Survival plot is a derivative of the Kaplan-Meier Curve, a non-parametric graphical technique that features in Survival Analysis. This subject area is a well known domain of statistics, but would be encountered on curriculums of specialist courses.

The Mountain Plot is formally called the empirical folder cumulative distribution plot. While not particularly hard to render, the procedure is not straight-forward for the casual user. Currently there is only one software implementation, ***medcalc.be*** toolkit.

3.2 Variations and Alternative Graphical Methods

In this section, we will look at some variations and enhancements of the Bland-Altman plot, as well as some alternative graphcial techniques. Strictly speaking, the Identity

Plot is advised by Bland and Altman as a prior analysis to the Bland-Alman plot, and therefore is neither a variant nor an alternative approach. However it is worth mentioning, as it is a simple, powerful and elegant technique that is often overlooked in method comparison studies. The identity plot is a simple scatter-plot approach of measurements for both methods on either axis, with the line of equality (the $X = Y$ line, i.e. the 45 degree line through the origin). This plot can gives the analyst a cursory examination of how well the measurement methods agree. In the case of good agreement, the covariates of the plot accord closely with the line of equality.

3.2.1 Variants of the Bland-Altman Plot

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

Referring to the assumption that bias and variability are constant across the range of measurements, Bland and Altman (1999) address the case where there is an increase in variability as the magnitude increases. They remark that it is possible to ignore the issue altogether, but the limits of agreement would be wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used. To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. Bland and Altman acknowledge that this is not easy to interpret, and may not be suitable in all cases.

Bland and Altman's Percentage and Ratio Plots

Bland and Altman (1999) offer two variations of the Bland-Altman plot intended to overcome situations where the conventional plot is inappropriate. The first variation is a plot of casewise differences as percentage of averages, and is appropriate when the variability of the differences increase as the magnitude increases.

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

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

Bartko's Ellipse

As an enhancement on the Bland Altman Plot, Bartko (1994) has expounded a confidence ellipse for the covariates. Bartko (1994) proposes a bivariate confidence ellipse as a boundary for dispersion. The stated purpose is to 'amplify dispersion', which presumably is for the purposes of outlier detection. The orientation of the the ellipse is key to interpreting the results. The minor axis is related to the between-item variability whereas the major axis is related to the mean squared error (referred to here as Error Mean Square).The ellipse illustrates the size of both relative to each other.

Consequently Bartko's ellipse provides a visual aid to determining the relationship between variances. Furthermore, the ellipse provides a visual aid to determining the relationship between the variance of the means $Var(a_i)$ and the variance of the differences $Var(d_i)$. If var(a) is greater than var(d), the orientation of the ellipse is horizontal. Conversely if var(a) is less than var(d), the orientation of the ellipse is vertical. The more horizontal the ellipse, the greater the degree of agreement between the two methods being tested.

Bartko states that the ellipse can, inter alia, be used to detect the presence of outliers (furthermore Bartko (1994) proposes formal testing procedures, that shall be discussed in due course). The Bland-Altman plot for the Grubbs data, complemented by Bartko's ellipse, is depicted in Figure 3.2.1. 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

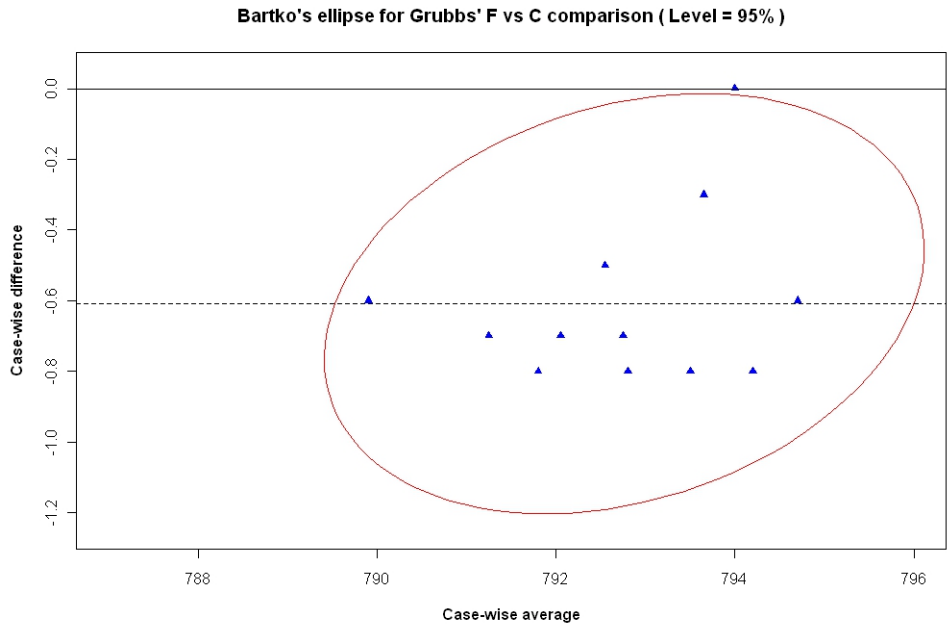

Figure 3.2.1: Bartko's Ellipse For Grubbs' Data.

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.

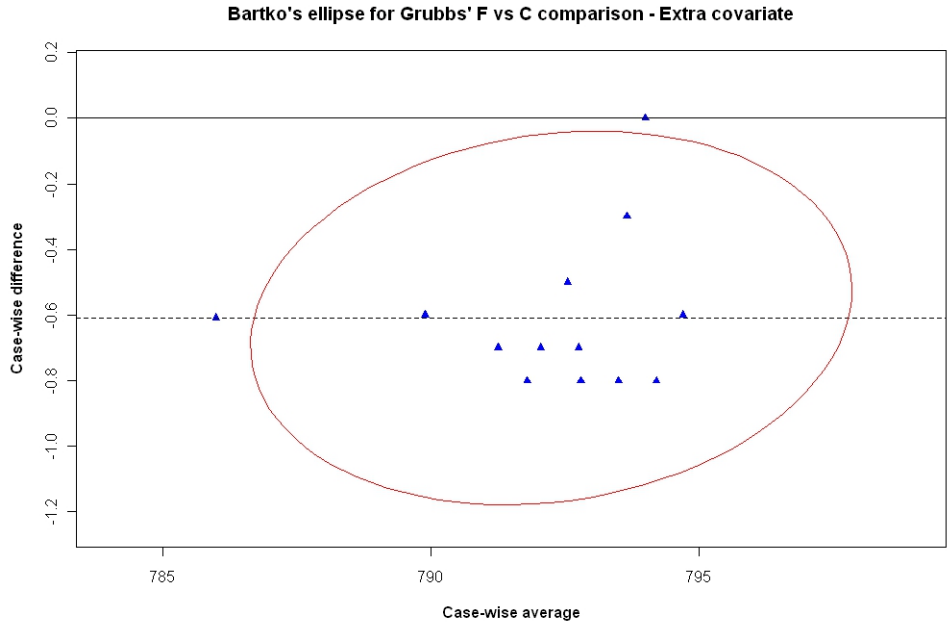

Figure 3.2.2: Bartko's Ellipse For Grubbs' Data, with an extra covariate.

In the Bland-Altman plot, the horizontal displacement of any point on the plot is supported by two independent measurements. Any point should not be considered an outlier on the basis of a noticeable horizontal displacement from the main cluster, as in the case with the extra co-variate. Conversely, the fourth point, from the original data set, should be considered an outlier, as it has a noticeable vertical displacement from the rest of the observations.

Survival-Agreement Plot

A graphical technique for method comparison studies, that is entirely different to the Bland-Altman plot, was proposed by Luiz et al. (2003). This approach, known as the survival-agreement plot, is used to determine the degree of agreement using the Kaplan-Meier method, a well known graphical technique in the area of Survival Analysis. Furthermore Luiz et al. (2003) propose that commonly used survival analysis techniques should complement this method, *providing a new analytical insight for agreement.* Two survival?agreement plots are used to detect the bias between to measurements of the same variable. The presence of inter-method bias is tested with the log-rank test, and its magnitude with Cox regression.

The degree of agreement (or disagreement) of a measure is expressed as a function of several limits of tolerance, using the Kaplan-Meier method, where the failures occur exactly at absolute values of the differences between the two methods of measurement.

According to Luiz et al, the survival-agreement plot is a step function of a typical survival analysis without censored data, where the Y axis represents the proportion of discordant cases. This is equivalent to a step function where the X axis represents the absolute observed differences and the Y axis is the proportion of the cases with at least the observed difference (x_i).

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: percentile = 100 - 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.

3.2.2 Replicate Measurements

Thus far, the formulation for comparison of two measurement methods is based on one measurement by each method per subject. Should there be two or more measurements by each method, these measurements are known as 'replicate measurements'. Carstensen et al. (2008) recommends the use of replicate measurements, but acknowledges the additional computational complexity.

Bland and Altman (1986) address this situation via 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 one single representative value.

Although either approach may be used to estimate the inter-method bias, removal of the effects of replicate measurements error leads to the underestimation of the standard deviation of the differences. Bland and Altman (1986) propose a correction for this.

Carstensen et al. (2008) take issue with the limits of agreement based on mean values of replicate measurements, since these must be interpreted as prediction limits for the difference between means of repeated measurements by both methods, rather than the difference of individual measurements. 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.

3.3 Formal Models and Tests

While the Bland-Altman plot is useful for inspection of data, **?** notes the lack of formal testing offered by this methodology. Furthermore, **?** formulates a model for single measurement observations as a linear mixed effects model, i.e. a model that additively combines fixed effects and random effects:

$$Y_{ij} = \quad \mu + \beta_j + u_i + \epsilon_{ij} \qquad i = 1, \ldots, n \qquad j = 1, 2$$

The true value of the measurement is represented by μ while the fixed effect due to method j is β_j. For simplicity these terms can be combined into single terms; $\mu_1 = \mu + \beta_1$ and $\mu_2 = \mu + \beta_2$. The inter-method bias is the difference of the two fixed effect terms, $\beta_1 - \beta_2$. Each individual is assumed to give rise to a random error, represented by u_i. This random effects term is assumed to have mean zero and be normally distributed with variance σ^2. There is assumed to be an attendant error for each measurement on each individual, denoted ϵ_{ij}. This is also assumed to have mean zero. The variance of measurement error for both methods are not assumed to be identical for both methods variance, hence it is denoted σ_j^2. The set of observations (x_i, y_i) by methods X and Y are assumed to follow a bivariate normal distribution with expected values $E(x_i) = \mu_i$ and $E(y_i) = \tau_i$ respectively. The variance covariance of the observations $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \sigma_2^2 \end{bmatrix}$$

? demonstrates the estimation of the variance terms and relative precisions relevant to a method comparison study, with attendant confidence intervals for both. The measurement model introduced by Grubbs (1948, 1973) provides a formal procedure for estimating the variances σ^2, σ_1^2 and σ_2^2. Grubbs (1948) offers estimates, commonly known as Grubbs estimators, for the various variance components. These estimates

are maximum likelihood estimates, which shall be revisited in due course.

$$\hat{\sigma^2} = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} = Sxy$$
$$\hat{\sigma_1^2} = \sum \frac{(x_i - \bar{x})^2}{n-1} = S^2x - Sxy$$
$$\hat{\sigma_2^2} = \sum \frac{(y_i - \bar{y})^2}{n-1} = S^2y - Sxy$$

Thompson (1963) defines $\Delta_j = \sigma^2/\sigma_j^2, j = 1, 2$, to be a measure of the relative precision of the measurement methods, and demonstrates how to make statistical inferences about Δ_j. Based on the following identities,

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ |A| &= C_x \times C_y - (C_{xy})^2, \end{aligned}$$

the confidence interval limits of Δ_1 are

$$\frac{C_{xy} - t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}} < \Delta_1 < \frac{C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}}$$

The value t is the $100(1-\alpha/2)\%$ upper quantile of Student's t distribution with $n-2$ degrees of freedom (**?**). The confidence limits for Δ_2 are found by substituting C_y for C_x in (1.2). Negative lower limits are replaced by the value 0.

The case-wise differences and means are calculated as $d_i = x_i - y_i$ and $a_i = (x_i + y_i)/2$ respectively. Both d_i and a_i are assumed to follow a bivariate normal distribution with $E(d_i) = \mu_d = \mu_1 - \mu_2$ and $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$, and the variance matrix $\Sigma_{(a,d)}$ is

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \tag{3.2}$$

3.4 Measurement Error Models

DunnSEME proposes a measurement error model for use in method comparison studies. Consider n pairs of measurements X_i and Y_i for $i = 1, 2, ...n$.

$$X_i = \tau_i + \delta_i \tag{3.3}$$

$$Y_i = \alpha + \beta\tau_i + \epsilon_i$$

In the above formulation is in the form of a linear structural relationship, with τ_i and $\beta\tau_i$ as the true values , and δ_i and ϵ_i as the corresponding measurement errors. In the case where the units of measurement are the same, then $\beta = 1$.

$$E(X_i) = \tau_i \tag{3.4}$$

$$E(Y_i) = \alpha + \beta\tau_i$$

$$E(\delta_i) = E(\epsilon_i) = 0$$

The value α is the inter-method bias between the two methods.

$$z_0 = d = 0 \tag{3.5}$$

$$z_{n+1} = z_n^2 + c \tag{3.6}$$

3.5 Model Formulation and Formal Testing

? formulates a model for un-replicated observations for a method comparison study as a mixed model.

$$Y_{ij} = \mu_j + S_i + \epsilon_{ij} \quad i = 1, 2...n \quad j = 1, 2 \tag{3.7}$$

$$S \sim N(0, \sigma_s^2) \qquad \epsilon_{ij} \sim N(0, \sigma_j^2)$$

As with all mixed models, the variance of each observation is the sum of all the associated variance components.

$$\begin{aligned} var(Y_{ij}) = &\quad \sigma_s^2 + \sigma_j^2 \\ cov(Y_{i1}, Y_{i2}) = &\quad \sigma_s^2 \end{aligned} \tag{3.8}$$

Grubbs (1948) offers maximum likelihood estimators, commonly known as Grubbs estimators, for the various variance components:

$$\begin{aligned} \hat{\sigma_s^2} &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} &= Sxy \\ \hat{\sigma_1^2} &= \sum \frac{(x_i - \bar{x})^2}{n-1} &= S^2x - Sxy \\ \hat{\sigma_2^2} &= \sum \frac{(y_i - \bar{y})^2}{n-1} &= S^2y - Sxy \end{aligned} \tag{3.9}$$

The standard error of these variance estimates are:

$$\begin{aligned} var(\sigma_1^2) = &\quad \frac{2\sigma_1^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \\ var(\sigma_2^2) = &\quad \frac{2\sigma_2^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \end{aligned} \tag{3.10}$$

Thompson (1963)presents confidence intervals for the relative precisions of the measurement methods, $\Delta_j = \sigma_S^2/\sigma_j^2$ (where $j = 1, 2$), as well as the variances σ_S^2, σ_1^2 and σ_2^2.

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}} \tag{3.11}$$

where

$$\begin{aligned} C_x = &\quad (n-1)S_x^2 \\ C_{xy} = &\quad (n-1)S_{xy} \\ C_y = &\quad (n-1)S_y^2 \\ A = &\quad C_x \times C_y - (C_{xy})^2 \end{aligned}$$

t is the $100(1-\alpha/2)\%$ quantile of Student's t distribution with $n-2$ degrees of freedom. Δ_2 can be found by changing C_y for C_x. A lower confidence limit can be found by calculating the square root. This inequality may also be used for hypothesis testing.

For the interval estimates for the variance components, Thompson (1963) presents three relations that hold simultaneously with probability $1-2\alpha$ where $2\alpha = 0.01$ or 0.05.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned} \tag{3.12}$$

The case-wise differences and means are $D_i = Y_{i1} - Y_{i2}$ and $A_i = (Y_{i1} + Y_{i2})/2$ respectively. Both D_i and A_i follow a bivariate normal distribution with $E(D_i) = \mu_D = \mu_1 - \mu_2$ and $E(A_i) = \mu_A = (\mu_1 + \mu_2)/2$. The variance matrix Σ is

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_S^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix} \tag{3.13}$$

? demonstrates how the Grubbs estimators for the error variances can be calculated using the difference values, providing a worked example on a data set.

$$\begin{aligned} \hat{\sigma_1^2} &= \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\ \hat{\sigma_2^2} &= \sum (y_{i2} - \bar{y_2})(D_i - \bar{D}) \end{aligned} \tag{3.14}$$

3.5.1 Morgan Pitman

The test of the hypothesis that the variance of both methods are equal is based on the correlation value $\rho_{D,A}$ which is evaluated as follows;

$$\rho(D, A) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}} \tag{3.15}$$

The correlation constant takes the value zero if, and only if, the two variances are equal. Therefore a test of the hypothesis $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a test of the hypothesis $H : \rho(D, A) = 0$. The corresponds to the well-known t test for a correlation coefficient with $n - 2$ degrees of freedom.

Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of Y_{i1} on Y_{12}, adding that this result can be shown using straightforward algebra.

3.5.2 Bartko's Bradley-Blackwood Test

This is a regression based approach that performs a simulataneous test for the equivalence of means and variances of the respective methods.

$$D = (X_1 - X_2) \tag{3.16}$$

$$M = (X_1 + X_2)/2 \tag{3.17}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{3.18}$$

Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.
We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from the results of a regression of D on M.
We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.
Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept.

subsectiont-test

3.5.3 Blackwood Bradley Model

This is a regression based approach that performs a simultaneous test for the equivalence of means and variances of the respective methods.
We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.

$$D = (X_1 - X_2) \tag{3.19}$$

$$M = (X_1 + X_2)/2 \tag{3.20}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{3.21}$$

Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.
We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from the results of a regression of D on M.
Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept. Bradley and Blackwood have developed a regression based approach assessing the agreement.
The Bradley Blackwood test is a simultaneous test for bias and precision. They propose a regression approach which fits D on M, where D is the difference and average of a pair of results.

3.5.4 Pitman & Morgan Test

This test assess the equality of population variances. Pitman's test tests for zero correlation between the sums and products.

Correlation between differences and means is a test statistics for the null hypothesis of equal variances given bivariate normality.

3.6 Thompson 1963

Thompson (1963) defines Δ_j to be a measure of the relative precision of the measurement methods, with $\Delta_j = \sigma_S^2/\sigma_j^2$(where $j = 1, 2$). Confidence intervals for Δ_j are also presented.

$$\Delta_1 > \frac{C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-1}))^{\frac{1}{2}}}, \tag{3.22}$$

where

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ A &= C_x \times C_y - (C_{xy})^2. \end{aligned}$$

The value t is the $100(1-\alpha/2)\%$ quantile of Student's t distribution with $n-2$ degrees of freedom. The ratio Δ_2 can be found by interchanging C_y and C_x. A lower confidence limit can be found by calculating the square root. The inequality in equation 1.10 may also be used for hypothesis testing.

For the interval estimates for the variance components, Thompson (1963) presents three relations that hold simultaneously with probability $1-2\alpha$ where $2\alpha = 0.01$ or 0.05.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned}$$

Thompson (1963) contains tables for K and M.

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 (μ) and variance (σ^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 thereof are $\nu_1 = 2$ and $\nu_1 = n - 2$ (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:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg} \tag{3.23}$$

	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 3.6.1: 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.

3.6.1 Formal Testing

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

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

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

3.7 Bartko's Regression and Ellipse

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 (β_0 and β_1). The null hypothesis of this test is that the mean (μ) and variance (σ^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 thereof are $\nu_1 = 2$ and $\nu_1 = n - 2$ (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 metholodogy for calculation using the from the averages of the pairs, as opposed to the sums, and their differences. This would facilitate simultaneous usage of test with the Bland Altman methodology. Bartko's test statistic take the form:

$$F.test = \frac{(\Sigma D^2) - SSReg}{2MSReg} \tag{3.24}$$

	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 3.7.2: 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 4.102821 (calculate using r code $qf(0.95, 2, 10)$). 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 seperate testing, no conclusion can be drawn on the comparative precision of both methods.

3.8 Formal Models and Tests

The Bland-Altman plot is a simple tool for inspection of data, and **?** comments on the lack of formal testing offered by that methodology. **?** formulates a model for single measurement observations for a method comparison study as a linear mixed effects model, i.e. model that additively combine fixed effects and random effects.

$$Y_{ij} = \quad \mu + \beta_j + u_i + \epsilon_{ij} \qquad i = 1, \ldots, n \qquad j = 1, 2$$

The true value of the measurement is represented by μ while the fixed effect due to method j is β_j. For simplicity these terms can be combined into single terms; $\mu_1 = \mu + \beta_1$ and $\mu_2 = \mu + \beta_2$. The inter-method bias is the difference of the two fixed effect terms, $\beta_1 - \beta_2$. Each of the i individuals are assumed to give rise to random error, represented by u_i. This random effects terms is assumed to have mean zero and be normally distributed with variance σ^2. There is assumed to be an attendant error for each measurement on each individual, denoted ϵ_{ij}. This is also assumed to have mean zero. The variance of measurement error for both methods are not assumed to be identical for both methods variance, hence it is denoted σ_j^2. The set of observations (x_i, y_i) by methods X and Y are assumed to follow the bivariate normal distribution with expected values $E(x_i) = \mu_i$ and $E(x_i) = \mu_i$ respectively. The variance covariance of the observations $\mathbf{\Sigma}$ is given by

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \sigma_2^2 \end{bmatrix}$$

The inter-method bias is the difference of the two fixed effect terms, $\beta_1 - \beta_2$.

? demonstrates the estimation of the variance terms and relative precisions relevant to a method comparison study, with attendant confidence intervals for both. The measurement model introduced by Grubbs (1948, 1973) provides a formal procedure for estimate the variances σ^2,σ_1^2 and σ_2^2 devices. Grubbs (1948) offers estimates, commonly known as Grubbs estimators, for the various variance components. These estimates are maximum likelihood estimates, a statistical concept that shall be revisited in due

course.

$$\hat{\sigma^2} = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} = Sxy$$
$$\hat{\sigma_1^2} = \sum \frac{(x_i - \bar{x})^2}{n-1} = S^2x - Sxy$$
$$\hat{\sigma_2^2} = \sum \frac{(y_i - \bar{y})^2}{n-1} = S^2y - Sxy$$

Thompson (1963) defines Δ_j to be a measure of the relative precision of the measurement methods, with $\Delta_j = \sigma^2/\sigma_j^2$. Thompson also demonstrates how to make statistical inferences about Δ_j. Based on the following identities,

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ |A| &= C_x \times C_y - (C_{xy})^2, \end{aligned}$$

the confidence interval limits of Δ_1 are

$$\begin{aligned} \Delta_1 &> \frac{C_{xy} - t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}} \\ \Delta_1 &> \frac{C_{xy} + t(\frac{|A|}{n-2}))^{\frac{1}{2}}}{C_x - C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}} \end{aligned} \tag{3.25}$$

The value t is the $100(1-\alpha/2)\%$ upper quantile of Student's t distribution with $n-2$ degrees of freedom (**?**). The confidence limits for Δ_2 are found by substituting C_y for C_x in (1.3). Negative lower limits are replaced by the value 0.

The case-wise differences and means are calculated as $d_i = x_i - y_i$ and $a_i = (x_i + y_i)/2$ respectively. Both d_i and a_i are assumed to follow a bivariate normal distribution with $E(d_i) = \mu_d = \mu_1 - \mu_2$ and $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$. The variance matrix $\Sigma_{(a,d)}$ is

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \tag{3.26}$$

3.8.1 Morgan-Pitman Testing

An early contribution to formal testing in method comparison was made by both **?** and **?**, in separate contributions. The basis of this approach is that if the distribution of the original measurements is bivariate normal. Morgan and Pitman noted that the correlation coefficient depends upon the difference $\sigma_1^2 - \sigma_2^2$, being zero if and only if $\sigma_1^2 = \sigma_2^2$.

The classical Pitman-Morgan test is a hypothesis test for equality of the variance of two data sets; $\sigma_1^2 = \sigma_2^2$, based on the correlation value $\rho_{a,d}$,and is evaluated as follows;

$$\rho(a,d) = \quad \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}} \tag{3.27}$$

The correlation constant takes the value zero if, and only if, the two variances are equal. Therefore a test of the hypothesis $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a test of the hypothesis $H : \rho(D, A) = 0$. The corresponds to the well-known t test for a correlation coefficient with $n - 2$ degrees of freedom. Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of Y_{i1} on Y_{12}, a result that can be derived using straightforward algebra.

3.9 Model Formulation and Formal Testing

? formulates a model for un-replicated observations for a method comparison study as a mixed model.

$$\begin{aligned} Y_{ij} = \quad & \mu_j + S_i + \epsilon_{ij} \quad i = 1, 2...n \quad j = 1, 2 \\ & S \sim N(0, \sigma_s^2) \qquad \epsilon_{ij} \sim N(0, \sigma_j^2) \end{aligned} \tag{3.28}$$

As with all mixed models, the variance of each observation is the sum of all the associated variance components.

$$\begin{aligned} var(Y_{ij}) = \quad & \sigma_s^2 + \sigma_j^2 \\ cov(Y_{i1}, Y_{i2}) = \quad & \sigma_s^2 \end{aligned} \tag{3.29}$$

Grubbs (1948) offers maximum likelihood estimators, commonly known as Grubbs estimators, for the various variance components:

$$\begin{aligned} \hat{\sigma_s^2} &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} &= Sxy \\ \hat{\sigma_1^2} &= \sum \frac{(x_i - \bar{x})^2}{n-1} &= S^2x - Sxy \\ \hat{\sigma_2^2} &= \sum \frac{(y_i - \bar{y})^2}{n-1} &= S^2y - Sxy \end{aligned} \tag{3.30}$$

The standard error of these variance estimates are:

$$\begin{aligned} var(\sigma_1^2) = &\quad \frac{2\sigma_1^4}{n-1} + \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \\ var(\sigma_2^2) = &\quad \frac{2\sigma_2^4}{n-1} + \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \end{aligned} \tag{3.31}$$

Thompson (1963)presents confidence intervals for the relative precisions of the measurement methods, $\Delta_j = \sigma_S^2/\sigma_j^2$ (where $j = 1, 2$), as well as the variances σ_S^2, σ_1^2 and σ_2^2.

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}} \tag{3.32}$$

where

$$\begin{aligned} C_x = &\quad (n-1)S_x^2 \\ C_{xy} = &\quad (n-1)S_{xy} \\ C_y = &\quad (n-1)S_y^2 \\ A = &\quad C_x \times C_y - (C_{xy})^2 \end{aligned}$$

t is the $100(1-\alpha/2)\%$ quantile of Student's t distribution with $n-2$ degrees of freedom. Δ_2 can be found by changing C_y for C_x. A lower confidence limit can be found by calculating the square root. This inequality may also be used for hypothesis testing.

For the interval estimates for the variance components, Thompson (1963) presents three relations that hold simultaneously with probability $1 - 2\alpha$ where $2\alpha = 0.01$ or 0.05.

$$
\begin{aligned}
|\sigma^2 - C_{xy}K| &\leqslant M(C_x C_y)^{\frac{1}{2}} \\
|\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\
|\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}}
\end{aligned}
\tag{3.33}
$$

The case-wise differences and means are $D_i = Y_{i1} - Y_{i2}$ and $A_i = (Y_{i1} + Y_{i2})/2$ respectively. Both D_i and A_i follow a bivariate normal distribution with $E(D_i) = \mu_D = \mu_1 - \mu_2$ and $E(A_i) = \mu_A = (\mu_1 + \mu_2)/2$. The variance matrix Σ is

$$
\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_S^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}
\tag{3.34}
$$

? demonstrates how the Grubbs estimators for the error variances can be calculated using the difference values, providing a worked example on a data set.

$$
\begin{aligned}
\hat{\sigma_1^2} &= \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\
\hat{\sigma_2^2} &= \sum (y_{i2} - \bar{y_2})(D_i - \bar{D})
\end{aligned}
\tag{3.35}
$$

3.9.1 Morgan Pitman

The test of the hypothesis that the variance of both methods are equal is based on the correlation value $\rho_{D,A}$ which is evaluated as follows;

$$
\rho(D, A) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}}
\tag{3.36}
$$

The correlation constant takes the value zero if, and only if, the two variances are equal. Therefore a test of the hypothesis $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a test of the hypothesis $H : \rho(D, A) = 0$. The corresponds to the well-known t test for a correlation coefficient with $n - 2$ degrees of freedom.

Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of Y_{i1} on Y_{12}, adding that this result can be shown using straightforward algebra.

3.9.2 Morgan Pitman

The test of the hypothesis that the variance of both methods are equal is based on the correlation value $\rho_{D,A}$ which is evaluated as follows;

$$\rho(D, A) = \quad \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}} \tag{3.37}$$

The correlation constant takes the value zero if, and only if, the two variances are equal. Therefore a test of the hypothesis $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a test of the hypothesis $H : \rho(D, A) = 0$. The corresponds to the well-known t test for a correlation coefficient with $n - 2$ degrees of freedom.

Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of Y_{i1} on Y_{12}, adding that this result can be shown using straightforward algebra.

3.10 Model Formulation and Formal Testing

? formulates a model for un-replicated observations for a method comparison study as a mixed model.

$$\begin{aligned} Y_{ij} = \quad & \mu_j + S_i + \epsilon_{ij} \quad i = 1, 2...n \quad j = 1, 2 \\ & S \sim N(0, \sigma_s^2) \qquad \epsilon_{ij} \sim N(0, \sigma_j^2) \end{aligned} \tag{3.38}$$

As with all mixed models, the variance of each observation is the sum of all the associated variance components.

$$\begin{aligned} var(Y_{ij}) = \quad & \sigma_s^2 + \sigma_j^2 \\ cov(Y_{i1}, Y_{i2}) = \quad & \sigma_s^2 \end{aligned} \tag{3.39}$$

Grubbs (1948) offers maximum likelihood estimators, commonly known as Grubbs estimators, for the various variance components:

$$\begin{aligned} \hat{\sigma_s^2} &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} &= Sxy \\ \hat{\sigma_1^2} &= \sum \frac{(x_i - \bar{x})^2}{n-1} &= S^2x - Sxy \\ \hat{\sigma_2^2} &= \sum \frac{(y_i - \bar{y})^2}{n-1} &= S^2y - Sxy \end{aligned} \tag{3.40}$$

The standard error of these variance estimates are:

$$\begin{aligned} var(\sigma_1^2) = &\quad \frac{2\sigma_1^4}{n-1} + \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \\ var(\sigma_2^2) = &\quad \frac{2\sigma_2^4}{n-1} + \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \end{aligned} \tag{3.41}$$

Thompson (1963)presents confidence intervals for the relative precisions of the measurement methods, $\Delta_j = \sigma_S^2/\sigma_j^2$ (where $j = 1, 2$), as well as the variances σ_S^2, σ_1^2 and σ_2^2.

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}} \tag{3.42}$$

where

$$\begin{aligned} C_x = &\quad (n-1)S_x^2 \\ C_{xy} = &\quad (n-1)S_{xy} \\ C_y = &\quad (n-1)S_y^2 \\ A = &\quad C_x \times C_y - (C_{xy})^2 \end{aligned}$$

t is the $100(1-\alpha/2)\%$ quantile of Student's t distribution with $n-2$ degrees of freedom. Δ_2 can be found by changing C_y for C_x. A lower confidence limit can be found by calculating the square root. This inequality may also be used for hypothesis testing.

For the interval estimates for the variance components, Thompson (1963) presents three relations that hold simultaneously with probability $1 - 2\alpha$ where $2\alpha = 0.01$ or 0.05.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_x C_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned} \tag{3.43}$$

The case-wise differences and means are $D_i = Y_{i1} - Y_{i2}$ and $A_i = (Y_{i1} + Y_{i2})/2$ respectively. Both D_i and A_i follow a bivariate normal distribution with $E(D_i) = \mu_D = \mu_1 - \mu_2$ and $E(A_i) = \mu_A = (\mu_1 + \mu_2)/2$. The variance matrix Σ is

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_S^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix} \tag{3.44}$$

? demonstrates how the Grubbs estimators for the error variances can be calculated using the difference values, providing a worked example on a data set.

$$\begin{aligned} \hat{\sigma_1^2} &= \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\ \hat{\sigma_2^2} &= \sum (y_{i2} - \bar{y_2})(D_i - \bar{D}) \end{aligned} \tag{3.45}$$

3.10.1 Paired sample T-test

Bartko (1994) discusses the use of the well known paired sample t test to test for inter-method bias; $H : \mu_D = 0$. The test statistic is distributed a t random variable with $n - 1$ degrees of freedom and is calculated as follows;

$$t^* = \bar{D} / \frac{S_D}{\sqrt{n}} \tag{3.46}$$

where $\bar{D}$ and S_D is the average of the differences of the n observations.

3.10.2 Paired sample T-test

Bartko (1994) discusses the use of the well known paired sample t test to test for inter-method bias; $H : \mu_D = 0$. The test statistic is distributed a t random variable with $n-1$ degrees of freedom and is calculated as follows;

$$t^* = \bar{D} / \frac{S_D}{\sqrt{n}} \tag{3.47}$$

where $\bar{D}$ and S_D is the average of the differences of the n observations.

3.11 Thompson 1963

Thompson (1963) defines Δ_j to be a measure of the relative precision of the measurement methods, with $\Delta_j = \sigma_S^2/\sigma_j^2$(where $j = 1, 2$). Confidence intervals for Δ_j are also presented.

$$\Delta_1 > \frac{C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-1}))^{\frac{1}{2}}}, \tag{3.48}$$

where

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ A &= C_x \times C_y - (C_{xy})^2. \end{aligned}$$

The value t is the $100(1-\alpha/2)\%$ quantile of Student's t distribution with $n-2$ degrees of freedom. The ratio Δ_2 can be found by interchanging C_y and C_x. A lower confidence limit can be found by calculating the square root. The inequality in equation 1.10 may also be used for hypothesis testing.

For the interval estimates for the variance components, Thompson (1963) presents three relations that hold simultaneously with probability $1 - 2\alpha$ where $2\alpha = 0.01$ or 0.05.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned}$$

Thompson (1963) contains tables for K and M.

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 (μ) and variance (σ^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 thereof are $\nu_1 = 2$ and $\nu_1 = n - 2$ (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:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg} \tag{3.49}$$

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.

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

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.

3.11.1 Formal Testing

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

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

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

3.12 Thompson 1963

Thompson (1963) defines Δ_j to be a measure of the relative precision of the measurement methods, with $\Delta_j = \sigma_S^2/\sigma_j^2$(where $j = 1, 2$). Confidence intervals for Δ_j are also presented.

$$\Delta_1 > \frac{C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-1}))^{\frac{1}{2}}}, \tag{3.50}$$

where

$$\begin{aligned} C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ A &= C_x \times C_y - (C_{xy})^2. \end{aligned}$$

The value t is the $100(1-\alpha/2)\%$ quantile of Student's t distribution with $n-2$ degrees of freedom. The ratio Δ_2 can be found by interchanging C_y and C_x. A lower confidence limit can be found by calculating the square root. The inequality in equation 1.10 may also be used for hypothesis testing.

For the interval estimates for the variance components, Thompson (1963) presents three relations that hold simultaneously with probability $1-2\alpha$ where $2\alpha = 0.01$ or 0.05.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned}$$

Thompson (1963) contains tables for K and M.

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 (μ) and variance (σ^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 thereof are $\nu_1 = 2$ and $\nu_1 = n - 2$ (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:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg} \tag{3.51}$$

	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 3.12.4: 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.

3.12.1 Formal Testing

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

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

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

3.13 Bartko's Regression and Ellipse

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 (β_0 and β_1). The null hypothesis of this test is that the mean (μ) and variance (σ^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 thereof are $\nu_1 = 2$ and $\nu_1 = n - 2$ (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 metholodogy for calculation using the from the averages of the pairs, as opposed to the sums, and their differences. This would facilitate simultaneous usage of test with the Bland Altman methodology. Bartko's test statistic take the form:

$$F.test = \frac{(\Sigma D^2) - SSReg}{2MSReg} \tag{3.52}$$

	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 3.13.5: 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 4.102821 (calculate using r code $qf(0.95, 2, 10)$). 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 seperate testing, no conclusion can be drawn on the comparative precision of both methods.

3.13.1 Bartko's Ellipse

Bartko (1994) offers a graphical complement to the Bland-Altman plot, in the form of a bivariate confidence ellipse. Altman (1978) provides the relevant calculations.

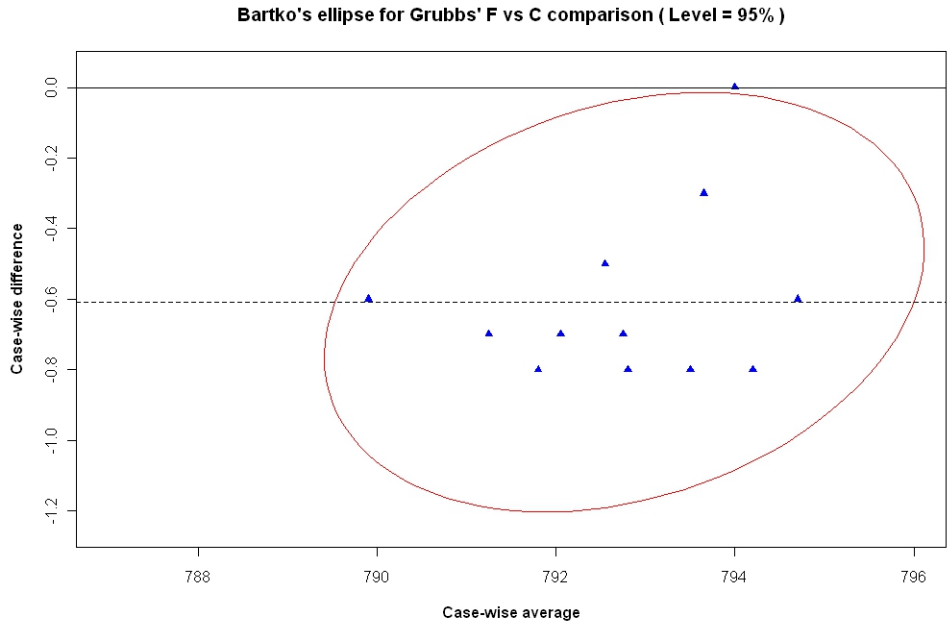

Figure 3.13.3: Bartko's Ellipse For Grubbs Data

3.13.2 Bartko's Bradley-Blackwood Test

This is a regression based approach that performs a simulataneous test for the equivalence of means and variances of the respective methods.

$$D = (X_1 - X_2) \tag{3.53}$$

$$M = (X_1 + X_2)/2 \tag{3.54}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{3.55}$$

Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.
We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from the results of a regression of D on M.
We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.
Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept.

subsectiont-test

3.13.3 Blackwood Bradley Model

This is a regression based approach that performs a simultaneous test for the equivalence of means and variances of the respective methods.
We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.

$$D = (X_1 - X_2) \tag{3.56}$$

$$M = (X_1 + X_2)/2 \tag{3.57}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{3.58}$$

Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.
We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from

the results of a regression of D on M.

Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept. Bradley and Blackwood have developed a regression based approach assessing the agreement.

The Bradley Blackwood test is a simultaneous test for bias and precision. They propose a regression approach which fits D on M, where D is the difference and average of a pair of results.

3.13.4 Pitman & Morgan Test

This test assess the equality of population variances. Pitman's test tests for zero correlation between the sums and products.

Correlation between differences and means is a test statistics for the null hypothesis of equal variances given bivariate normality.

3.14 Bartko's Regression and Ellipse

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 (β_0 and β_1). The null hypothesis of this test is that the mean (μ) and variance (σ^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 thereof are $\nu_1 = 2$ and $\nu_1 = n - 2$ (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 metholodogy for calculation using the from the averages of the pairs, as opposed to the sums, and their differences. This would facilitate simultaneous usage of test with the Bland Altman methodology. Bartko's test statistic take the

form:

$$F.test = \frac{(\Sigma D^2) - SSReg}{2MSReg} \tag{3.59}$$

	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 3.14.6: 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 4.102821 (calculate using r code $qf(0.95, 2, 10)$). 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 seperate testing, no conclusion can be drawn on the comparative precision of both methods.

3.14.1 Bartko's Ellipse

Bartko (1994) offers a graphical complement to the Bland-Altman plot, in the form of a bivariate confidence ellipse. Altman (1978) provides the relevant calculations.

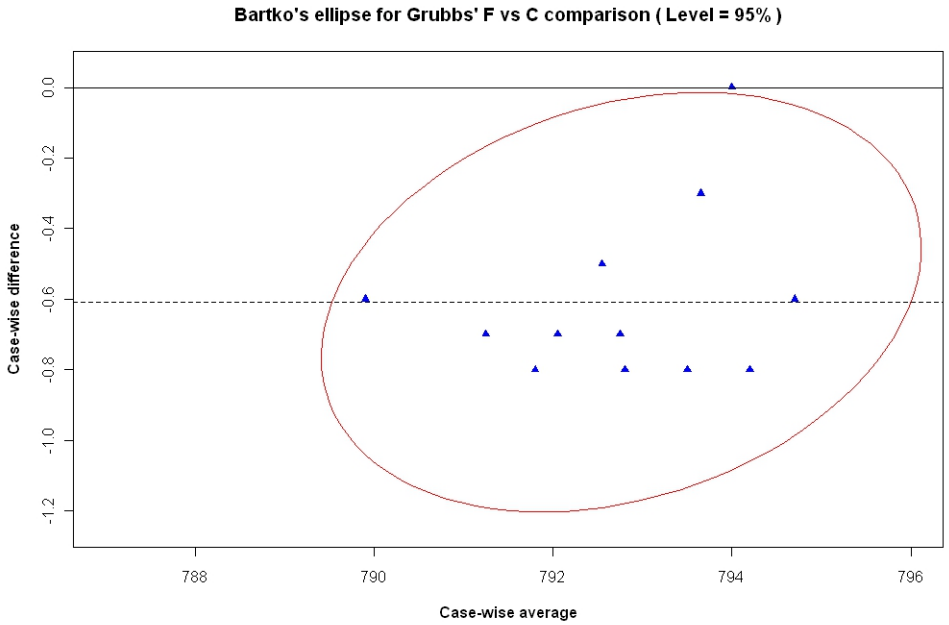

Figure 3.14.4: Bartko's Ellipse For Grubbs Data

3.14.2 Morgan Pitman Testing

An early contribution to formal testing in method comparison was made by both **?** and **?**, in separate contributions. The basis of this approach is that the distribution of the original measurements is bivariate normal. Morgan and Pitman noted that the correlation coefficient depends upon the difference $\sigma_1^2 - \sigma_2^2$, being zero if and only if $\sigma_1^2 = \sigma_2^2$.

The classical Pitman-Morgan test is a hypothesis test for equality of the variance of two data sets; $\sigma_1^2 = \sigma_2^2$, based on the correlation value $\rho_{a,d}$,and is evaluated as follows;

$$\rho(a,d) = \quad \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}} \tag{3.60}$$

The correlation constant takes the value zero if, and only if, the two variances are equal. Therefore a test of the hypothesis $H : \sigma_1^2 = \sigma_2^2$ is equivalent to a test of the hypothesis $H : \rho(D, A) = 0$. This corresponds to the well-known t test for a correlation coefficient with $n - 2$ degrees of freedom. Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of Y_{i1} on Y_{12}, a result that can be derived using straightforward algebra.

3.14.3 Paired sample *t*-test

Bartko (1994) discusses the use of the well known paired sample t test to test for inter-method bias; $H : \mu_d = 0$. The test statistic is distributed a t random variable with $n - 1$ degrees of freedom and is calculated as follows,

$$t^* = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{3.61}$$

where $\bar{d}$ and s_d is the average of the differences of the n observations. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

Structural Equation Modelling

Authors, such as a Lewis et al. (1991), Dunn (2002) and Voelkel and Siskowski (2005), strongly advocate the use of *Structural Equation Models* for the purposes of method comparison. Conversely Bland and Altman (1999) also states that consider structural equation models to be inappropriate.

3.15 Blackwood -Bradley Model

Bradley and Blackwood (1989) have developed a regression based procedure for assessing the agreement. This approach performs a simultaneous test for the equivalence of means and variances of the respective methods. 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 $\hat{\beta}_1$).

$$D = (X_1 - X_2) \tag{3.62}$$

$$M = (X_1 + X_2)/2 \tag{3.63}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{3.64}$$

This technique offers a formal simultaneous hypothesis test for the mean and variance of two paired data sets. The null hypothesis of this test is that the mean (μ) and variance (σ^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 $\nu_1 = n - 2$ (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 approach 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 approach. 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 3.15.7: 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 approach 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.

3.15.1 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 (ρ_{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 (ρ_{AD} =0) 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 Bland and Altman (1986), although Bland and Altman (1999) refers to Spearman's rank correlation coefficient. Bland and Altman (1999) state that they ' do not see a place for methods of analysis based on hypothesis testing'.

3.15.2 Identifiability

? highlights an important issue regarding using models such as structural equation modelling, which is 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 the literature, the variance ratio $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$ must often be assumed to be equal to 1 (Linnet, 1998). **?** considers approaches 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 counter-argument that in many practical settings it is very difficult to get replicate observations when, for example, the measurement method requires invasive medical procedure.

Bradley and Blackwood (1989) offer a formal simultaneous hypothesis test for the mean and variance of paired data sets. This approach is based upon regressing the differences of each pair on the sum of each pair, 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 (μ) and variance (σ^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). Bartko (1994) amends this approach 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 approach. 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 3.15.8: 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 3.742, 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.

sectionBartko's Bradley-Blackwood Test This is a regression based approach that performs a simultaneous test for the equivalence of means and variances of the respective methods.We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.

$$D = (X_1 - X_2) \tag{3.65}$$

$$M = (X_1 + X_2)/2 \tag{3.66}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{3.67}$$

- The Bradley Blackwood test is a simultaneous test for bias and precision. They propose a regression approach which fits D on M, where D is the difference and average of a pair of results.

- Both beta values, the intercept and slope, are derived from the respective means and standard deviations of their respective data sets.

- We determine if the respective means and variances are equal if both beta values are simultaneously equal to zero. The Test is conducted using an F test, calculated from the results of a regression of D on M.

- We have identified this approach to be examined to see if it can be used as a foundation for a test perform a test on means and variances individually.

- Russell et al have suggested this method be used in conjunction with a paired t-test , with estimates of slope and intercept.

3.15.3 Blackwood Bradley Model

Bradley and Blackwood have developed a regression based approach assessing the agreement.

3.16 Bradley-Blackwood Test (Kevin Hayes Talk)

This work considers the problem of testing $\mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$ using a random sample from a bivariate normal distribution with parameters $(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$.

The new contribution is a decomposition of the Bradley-Blackwood test statistic (*Bradley and Blackwood, 1989*)for the simultaneous test of $\mu_1 = \mu_2$; $\sigma_1^2 = \sigma_2^2$ as a sum of two statistics.

One is equivalent to the Pitman-Morgan (*Pitman, 1939; Morgan, 1939*) test statistic for $\sigma_1^2 = \sigma_2^2$ and the other one is a new alternative to the standard paired-t test of $\mu_D = \mu_1 = \mu_2 = 0$.

Surprisingly, the classic Student paired-t test makes no assumptions about the equality (or otherwise) of the variance parameters.

The power functions for these tests are quite easy to derive, and show that when $\sigma_1^2 = \sigma_2^2$, the paired t-test has a slight advantage over the new alternative in terms of power, but when $\sigma_1^2 \neq \sigma_2^2$, the new test has substantially higher power than the paired-t test.

While Bradley and Blackwood provide a test on the joint hypothesis of equal means and equal variances their regression based approach does not separate these two issues.

The rejection of the joint hypothesis may be due to two groups with unequal means and unequal variances; unequal means and equal variances, or equal means and unequal variances. We propose an approach for resolving this (model selection) problem in a manner controlling the magnitudes of the relevant type I error probabilities.

3.17 Regression Methods for Method Comparison

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. However this assumption invalidates simple

linear regression for use in method comparison studies, as both methods must be assumed to be measured with error (Altman and Bland, 1983; Ludbrook, 1997).

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. (Ludbrook, 1997). Determination of these biases shall be discussed in due course.

3.17.1 Deming Regression

As stated previously, the fundamental flaw of simple linear regression is that it allows for measurement error in one variable only. This causes a downward biased slope estimate.

Deming regression is a regression fitting approach that assumes error in both variables. Deming regression is recommended by Cornbleet and Cochrane (1979) as the preferred Model II regression for use in method comparison studies. The sum of squared distances from measured sets of values to the regression line is minimized at an angles

specified by the ratio λ of the residual variance of both variables. I When λ is one, the angle is 45 degrees. In ordinary linear regression, the distances are minimized in the vertical directions (Linnet, 1999). In cases involving only single measurements by each method, λ may be unknown and is therefore assumes a value of one. While this will produce biased estimates, they are less biased than ordinary linear regression.

The Bland Altman Plot is uninformative about the comparative influence of proportional bias and fixed bias. Model II approaches, such as Deming regression, can provide independent tests for both types of bias.

For a given λ, Kummel (1879) derived the following estimate that would later be used for the Deming regression slope parameter. The intercept estimate α is simply estimated in the same way as in conventional linear regression, by using the identity $\bar{Y} - \hat{\beta}\bar{X}$;

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + [(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2]^{1/2}}{2S_{xy}} \tag{3.68}$$

, with λ as the variance ratio. As stated previously λ is often unknown, and therefore must be assumed to equal one. Carroll and Ruppert (1996) states that Deming regression is acceptable only when the precision ratio (λ,in their paper as η) is correctly specified, but in practice this is often not the case, with the λ being underestimated. Several candidate models, with varying variance ratios may be fitted, and estimates of the slope and intercept are produced. However no model selection information is available to determine the best fitting model.

As with conventional regression methodologies, Deming regression calculates an estimate for both the slope and intercept for the fitted line, and standard errors thereof. Therefore there is sufficient information to carry out hypothesis tests on both estimates, that are informative about presence of fixed and proportional bias.

A 95% confidence interval for the intercept estimate can be used to test the intercept, and hence fixed bias, is equal to zero. This hypothesis is accepted if the confidence interval for the estimate contains the value 0 in its range. Should this be, it can be concluded that fixed bias is not present. Conversely, if the hypothesis is rejected, then it is concluded that the intercept is non zero, and that fixed bias is present.

Testing for proportional bias is a very similar procedure. The 95% confidence interval for the slope estimate can be used to test the hypothesis that the slope is equal to 1. This hypothesis is accepted if the confidence interval for the estimate contains the value 1 in its range. If the hypothesis is rejected, then it is concluded that the slope is significant different from 1 and that a proportional bias exists.

For convenience, a new data set shall be introduced to demonstrate Deming regression. Measurements of transmitral volumetric flow (MF) by doppler echocardiography, and left ventricular stroke volume (SV) by cross sectional echocardiography in 21 patients with aortic valve disease are tabulated in Zhang et al. (1986). This data set features in the discussion of method comparison studies in Altman (1991, p.398) .

Patient	MF (cm^3)	SV (cm^3)	Patient	MF (cm^3)	SV (cm^3)	Patient	MF (cm^3)	SV (cm^3)
1	47	43	8	75	72	15	90	82
2	66	70	9	79	92	16	100	100
3	68	72	10	81	76	17	104	94
4	69	81	11	85	85	18	105	98
5	70	60	12	87	82	19	112	108
6	70	67	13	87	90	20	120	131
7	73	72	14	87	96	21	132	131

Table 3.17.9: Transmitral volumetric flow(MF) and left ventricular stroke volume (SV) in 21 patients. (Zhang et al 1986)

Carroll and Ruppert (1996) states that Deming's regression is acceptable only when the precision ratio (λ, in their paper as η) is correctly specified, but in practice this is often not the case, with the λ being underestimated.

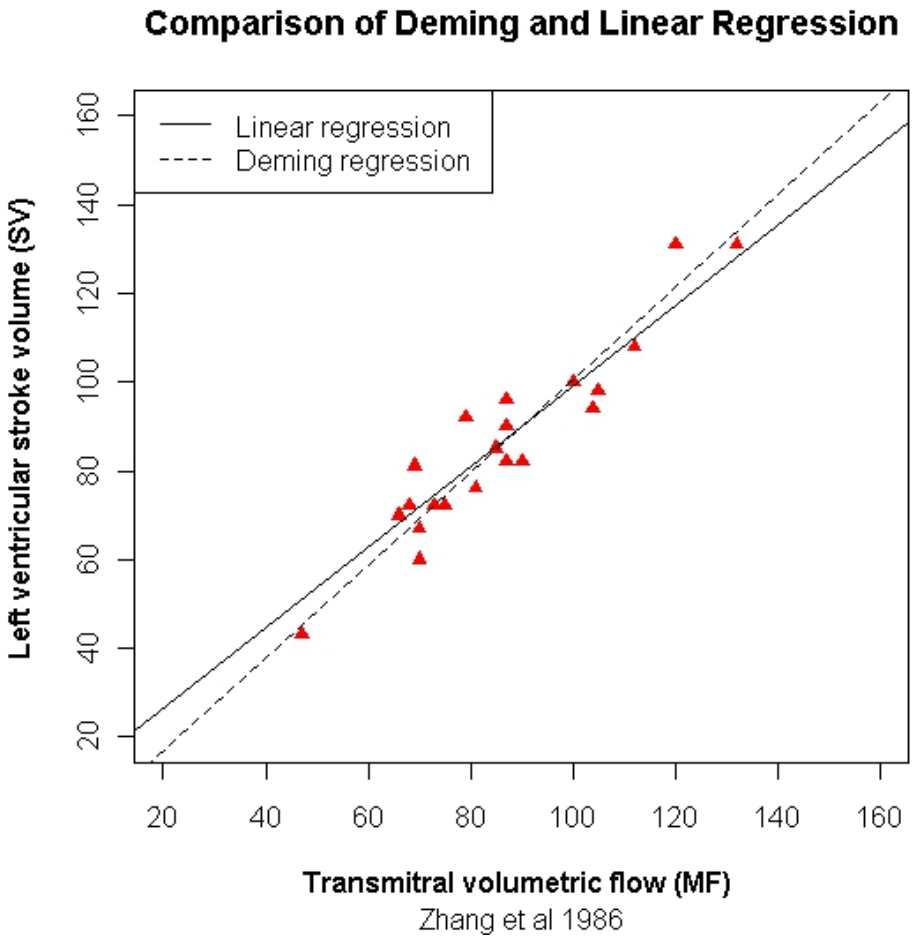

Figure 3.17.5: Deming Regression For Zhang's Data

3.18 Other Types of Studies

Lewis et al. (1991) categorize method comparison studies into three different types. The key difference between the first two is whether or not a 'gold standard' method is used. In situations where one instrument or method is known to be 'accurate and precise', it is considered as the'gold standard' (Lewis et al., 1991). A method that is not considered to be a gold standard is referred to as an 'approximate method'. In calibration studies they are referred to a criterion methods and test methods respectively.

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The

results of the approximate method can be mapped to a known probability distribution of the results of the gold standard (Lewis et al., 1991). (In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively.) Altman and Bland (1983) make clear that their methodology is not intended for calibration problems.

2. Comparison problems. When two approximate methods, that use the same units of measurement, are to be compared. This is the case which the Bland-Altman methodology is specfically intended for, and therefore it is the most relevant of the three.

3. Conversion problems. When two approximate methods, that use different units of measurement, are to be compared. This situation would arise when the measurement methods use 'different proxies', i.e different mechanisms of measurement. Lewis et al. (1991) deals specifically with this issue. In the context of this study, it is the least relevant of the three.

?, p.47 cautions that'gold standards' should not be assumed to be error free. 'It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphygmomanometer 'leaves considerable room for improvement' (**?**). Pizzi (1999) similarly addresses the issue of glod standards, 'well-established gold standard may itself be imprecise or even unreliable'.

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well-being of the patient. This will inevitably lead to the measurement error as

described by **?**. The magnetic resonance angiogram, used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. This is reported as sensitivity of 95% and a specificity of 92% (ACR, 2008).

In literature they are, perhaps more accurately, referred to as 'fuzzy gold standards' (Phelps and Hutson, 1995). Consequently when one of the methods is essentially a fuzzy gold standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

3.19 Methods of assessing agreement

1. Pearson's Correlation Coefficient
2. Intraclass correlation coefficient
3. Bland Altman Plot
4. Bartko's Ellipse (1994)
5. Blackwood Bradley Test
6. Lin's Reproducibility Index
7. Luiz Step function

Bland and Altman attend to the issue of repeated measures in 1996. Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual. Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation ,and the correlation coefficient.

The above plot incorporates both the conventional limits of agreement (the inner pair of dashed lines), the 't' limits of agreement (the outer pair of dashed lines) centred around the inter-method bias (indicated by the full line). This plot is intended for expository purposes only, as the sample size is small.

3.19.1 Equivalence and Interchangeability

Limits of agreement are intended to analyse equivalence. How this is assessed is the considered judgement of the practitioner. In Bland and Altman (1986) an example of good agreement is cited. For two methods of measuring 'oxygen saturation', the limits of agreement are calculated as (-2.0,2.8).A practitioner would ostensibly find this to be sufficiently narrow.

If the limits of agreement are not clinically important, which is to say that the differences tend not to be substantial, the two methods may be used interchangeably. **?** takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

3.20 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 that*sample sizes required either was not mentioned or no rationale for its choice was given.*

> In order to avoid the appearance of "data dredging", both the sample size and the (limits of agreement) should be specified and justified before the actual conduct of the trial. (Lin et al., 1991)

Dewitte et al. (2002) remarks that the limits of agreement should be compared to a clinically acceptable difference in measurements.

3.20.1 Gold Standard

This is considered to be the most accurate measurement of a particular parameter.

3.21 Discussion on Method Comparison Studies

The need to compare the results of two different measurement techniques is common in medical statistics.

In particular, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and a traditional reference or gold standard must be evaluated before the new one is put into practice. Various methodologies have been proposed for this purpose in recent years.

Indications on how to deal with outliers in Bland Altman plots
We wish to determine how outliers should be treated in a Bland Altman Plot
In their 1983 paper they merely state that the plot can be used to 'spot outliers'.
In their 1986 paper, Bland and Altman give an example of an outlier. They state that it could be omitted in practice, but make no further comments on the matter.
In Bland and Altmans 1999 paper, we get the clearest indication of what Bland and Altman suggest on how to react to the presence of outliers. Their recommendation is to recalculate the limits without them, in order to test the difference with the calculation where outliers are retained.

The span has reduced from 77 to 59 mmHg, a noticeable but not particularly large reduction.
However, they do not recommend removing outliers. Furthermore, they say:
We usually find that this method of analysis is not too sensitive to one or two large outlying differences.
We ask if this would be so in all cases. Given that the limits of agreement may or may not be disregarded, depending on their perceived suitability, we examine whether it would possible that the deletion of an outlier may lead to a calculation of limits of agreement that are usable in all cases?
Should an Outlying Observation be omitted from a data set? In general, this is not

considered prudent.

Also, it may be required that the outliers are worthy of particular attention themselves. 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?

3.21.1 Agreement

Bland and Altman (1986) define Perfect agreement as 'The case where all of the pairs of rater data lie along the line of equality'. The Line of Equality is defined as the 45 degree line passing through the origin, or X=Y on a XY plane.

3.21.2 Lack Of Agreement

1. Constant Bias
2. Proportional Bias

Constant Bias

This is a form of systematic deviations estimated as the average difference between the test and the reference method

Proportional Bias

Two methods may agree on average, but they may exhibit differences over a range of measurements

3.22 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.
Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

3.22.1 Bland Altman plots using 'Gold Standard' raters

According to Bland and Altman, one should use the methodology previous outlined, even when one of the raters is a Gold Standard.

3.22.2 Bias Detection

further to this method, the presence of constant bias may be indicated if the average value differences is not equal to zero. Bland and Altman does, however, indicate the indication of absence of bias does not provide sufficient information to allow a judgement as to whether or not one method can be substituted for another.

3.23 Coefficient of Repeatability

3.23.1 Repeatability

As mentioned previously, Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study. The coefficient of repeatability was proposed by Bland and Altman (1999), and is referenced in subsequent papers, such as Carstensen et al. (2008). The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999). The coefficient of repeatability is a measure of how well a measure-

ment method agrees with itself over replicate measurements (Bland and Altman, 1999). Once the the standard deviations of the differences between the two measurements (in some texts called the residual standard deviation or within-item variability) $sigma_m$ is determined, the computation of the coefficients of repeatability for both methods is straightforward. The coefficient is calculated from the (in some texts called the residual standard deviation) as $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$).

3.23.2 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.

3.23.3 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).

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

3.24 Repeatability

3.24.1 What is 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.

3.24.2 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.
- same objectives

Repeatability is defined by the **IUPAC** 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.

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.

3.25 Importance of Repeatability in MCS

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 & 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. $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$).

Barnhart et al. (2007) 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. $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$).

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

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

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

? 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.

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 method?s 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.

3.25.1 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 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.* Unless otherwise instructed, the probability is assumed to be 95%.

The Bland Altman method offers 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 (d_2 and d_1).

3.25.2 Repeatability coefficient from LME Models

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

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

3.25.3 Repeatability in Bland-Altman Blood Data Analysis

- Two readings by the same method will be within $1.96\sqrt{2}\sigma_w$ or $2.77\sigma_w$ for 95% of subjects. Thisvalue 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.

3.26 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 σ 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 ω^2 may be considered irrelevant in the repeatability and one would therefore base the repeatability on the measurement errors alone, i.e. use $2.8\sigma_m$.

3.26.1 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.

Chapter 4

Linear Mixed effects Models

4.1 Linear Mixed effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (random effects are also known as variance components). LME models are a generalization of the classical linear model, which contain fixed effects only. When the levels of factors are considered to be sampled from a population, and each level is not of particular interest, they are considered random quantities with associated variances. The effects of the levels, as described, are known as random effects. Random effects are represented by unobservable normally distributed random variables. Conversely fixed effects are considered non-random and the levels of each factor are of specific interest.

Fisher (1918) introduced variance components models for use in genetical studies. Whereas an estimate for variance must take an non-negative value, an individual variance component, i.e. a component of the overall variance, may be negative.

The framework has developed since, including contributions from Tippett (1931), who extend the use of variance components into linear models, and Eisenhart (1947), who introduced the 'mixed model' terminology and formally distinguished between mixed and random effects models. Henderson (1950) devised a framework for deriving

estimates for both the fixed effects and the random effects, using a set of equations that would become known as 'mixed model equations' or 'Henderson's equations'. LME methodology is further enhanced by Henderson's later works (Henderson, 1953; Henderson et al., 1959, 1963, 1973, 1984). The key features of Henderson's work provide the basis for the estimation techniques.

Hartley and Rao (1967) demonstrated that unique estimates of the variance components could be obtained using maximum likelihood methods. However these estimates are known to be biased 'downwards' (i.e. underestimated) , because of the assumption that the fixed estimates are known, rather than being estimated from the data. Patterson and Thompson (1971) produced an alternative set of estimates, known as the restricted maximum likelihood (REML) estimates, that do not require the fixed effects to be known. Thusly there is a distinction the REML estimates and the original estimates, now commonly referred to as ML estimates.

Laird and Ware (1982) provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. SAS Institute added PROC MIXED to its software suite in 1992 (Singer, 1998). Pinheiro and Bates (1994) described how to compute LME models in the `S-plus` environment.

Using Laird-Ware form, the LME model is commonly described in matrix form,

$$y = X\beta + Zb + \epsilon \tag{4.1}$$

where y is a vector of N observable random variables, β is a vector of p fixed effects, X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N, respectively, random effects such that $\mathrm{E}(b) = 0,\ \mathrm{E}(\epsilon) = 0$ and

$$\mathrm{var} \begin{pmatrix} b \\ \epsilon \end{pmatrix} = \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}$$

where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ. The variance-covariance matrix for the vector of

observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$. It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

4.1.1 Estimation

Estimation of LME models involve two complementary estimation issues'; estimating the vectors of the fixed and random effects estimates $\hat{\beta}$ and $\hat{b}$ and estimating the variance covariance matrices D and Σ. Inference about fixed effects have become known as 'estimates', while inferences about random effects have become known as 'predictions'. The most common approach to obtain estimators are Best Linear Unbiased Estimator (BLUE) and Best Linear Unbiased Predictor (BLUP). For an LME model given by (4.1), the BLUE of $\hat{\beta}$ is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

whereas the BLUP of $\hat{b}$ is given by

$$\hat{b} = DZ'V^{-1}(y - X\hat{\beta}).$$

Henderson's equations

Because of the dimensionality of V (i.e. $n \times n$) computing the inverse of V can be difficult. As a way around the this problem Henderson (1953); Henderson et al. (1959, 1963, 1973, 1984) offered a more simpler approach of jointly estimating $\hat{\beta}$ and $\hat{b}$. Henderson (1950) made the (ad-hoc) distributional assumptions $y|b \sim \mathrm{N}(X\beta + Zb, \Sigma)$ and $b \sim \mathrm{N}(0, D)$, and proceeded to maximize the joint density of y and b

$$\left| \begin{array}{cc} D & 0 \\ 0 & \Sigma \end{array} \right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix}' \begin{pmatrix} D & 0 \\ 0 & \Sigma \end{pmatrix}^{-1} \begin{pmatrix} b \\ y - X\beta - Zb \end{pmatrix} \right\}, \quad (4.2)$$

with respect to β and b, which ultimately requires minimizing the criterion

$$(y - X\beta - Zb)'\Sigma^{-1}(y - X\beta - Zb) + b'D^{-1}b. \quad (4.3)$$

This leads to the mixed model equations

$$\begin{pmatrix} X'\Sigma^{-1}X & X'\Sigma^{-1}Z \\ Z'\Sigma^{-1}X & X'\Sigma^{-1}X + D^{-1} \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X'\Sigma^{-1}y \\ Z'\Sigma^{-1}y \end{pmatrix}. \tag{4.4}$$

Using these equations, obtaining the estimates requires the inversion of a matrix of dimension $p+q \times p+q$, considerably smaller in size than V. **?** shows that these mixed model equations do not depend on normality and that $\hat{\beta}$ and $\hat{b}$ are the BLUE and BLUP under general conditions, provided D and Σ are known.

Robinson (1991) points out that although Henderson (1950) initially referred to the estimates $\hat{\beta}$ and $\hat{b}$ from (4.4) as "joint maximum likelihood estimates", Henderson (1973) later advised that these estimates should not be referred to as "maximum likelihood" as the function being maximized in (4.3) is a joint density rather than a likelihood function. Lee et al. (2006) remarks that it is clear that Henderson used joint estimation for computational purposes, without recognizing the theoretical implications.

Estimation of the fixed parameters

The vector y has marginal density $y \sim \mathrm{N}(X\beta, V)$, where $V = \Sigma + ZDZ'$ is specified through the variance component parameters θ. The log-likelihood of the fixed parameters (β, θ) is

$$\ell(\beta, \theta|y) = -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\beta)'V^{-1}(y - X\beta), \tag{4.5}$$

and for fixed θ the estimate $\hat{\beta}$ of β is obtained as the solution of

$$(X'V^{-1}X)\beta = X'V^{-1}y. \tag{4.6}$$

Substituting $\hat{\beta}$ from (4.6) into $\ell(\beta, \theta|y)$ from (4.5) returns the *profile* log-likelihood

$$\begin{aligned} \ell_P(\theta \mid y) &= \ell(\hat{\beta}, \theta \mid y) \\ &= -\frac{1}{2}\log|V| - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}) \end{aligned}$$

of the variance parameter θ. Estimates of the parameters θ specifying V can be found by maximizing $\ell_P(\theta \mid y)$ over θ. These are the ML estimates.

For REML estimation the *restricted* log-likelihood is defined as

$$\ell_R(\theta \mid y) = \ell_P(\theta \mid y) - \frac{1}{2}\log|X'VX|.$$

The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (Dodge, 2003). Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in $\boldsymbol{\beta}$. Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood. The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Estimation of the random effects

The established approach for estimating the random effects is to use the best linear predictor of b from y, which for a given β equals $DZ'V^{-1}(y-X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (4.6) so that $\hat{b} = DZ'V^{-1}(y - X\hat{\beta})$. Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates $\hat{\beta}$ and $\hat{b}$ satisfy the equations in (4.4).

Algorithms for likelihood function optimization

Iterative numerical techniques are used to optimize the log-likelihood function and estimate the covariance parameters θ. The procedure is subject to the constraint that R and D are both positive definite. The most common iterative algorithms for optimizing the likelihood function are the Newton-Raphson method, which is the preferred method, the expectation maximization (EM) algorithm and the Fisher scoring methods.

The EM algorithm, introduced by Dempster et al. (1977), is an iterative technique for maximizing complicated likelihood functions. The algorithm alternates between performing an expectation (E) step and the maximization (M) step. The 'E' step computes the expectation of the log-likelihood evaluated using the current estimate for the variables. In the 'M' step, parameters that maximize the expected log-likelihood, found on the previous 'E' step, are computed. These parameter estimates are then used to determine the distribution of the variables in the next 'E' step. The algorithm alternatives between these two steps until convergence is reached.

The main drawback of the EM algorithm is its slow rate of convergence. Consequently the EM algorithm is rarely used entirely in LME estimation, instead providing an initial set of values that can be passed to other optimization techniques.

The Newton Raphson (NR) method is the most common, and recommended technique for ML and REML estimation. The NR algorithm minimizes an objective function defines as -2 times the log likelihood for the covariance parameters θ. At every iteration the NR algorithm requires the calculation of a vector of partial derivatives, known as the gradient, and the second derivative matrix with respect to the covariance parameters. This is known as the observed Hessian matrix. Due to the Hessian matrix, the NR algorithm is more time-consuming, but convergence is reached with fewer iterations compared to the EM algorithm. The Fisher scoring algorithm is an variant of the NR algorithm that is more numerically stable and likely to converge, but not recommended to obtain final estimates.

The extended likelihood

The desire to have an entirely likelihood-based justification for estimates of random effects, in contrast to Henderson's equation, has motivated Pawitan (2001, page 429) to define the *extended likelihood*. He remarks "In mixed effects modelling the extended likelihood has been called *h-likelihood* (for hierarchical likelihood) by Lee and Nelder (1996), while in smoothing literature it is known as the *penalized likelihood* (e.g. Green and Silverman 1994)." The extended likelihood can be written $L(\beta, \theta, b|y) =$

$p(y|b;\beta,\theta)p(b;\theta)$ and adopting the same distributional assumptions used by Henderson (1950) yields the log-likelihood function

$$\begin{aligned}\ell_h(\beta,\theta,b|y) &= -\frac{1}{2}\left\{\log|\Sigma| + (y-X\beta-Zb)'\Sigma^{-1}(y-X\beta-Zb)\right.\\ &\qquad\left.+\log|D| + b'D^{-1}b\right\}.\end{aligned}$$

Given θ, differentiating with respect to β and b returns Henderson's equations in (4.4).

The LME model as a general linear model

Henderson's equations in (4.4) can be rewritten $(T'W^{-1}T)\delta = T'W^{-1}y_a$ using

$$\delta = \begin{pmatrix}\beta\\ b\end{pmatrix},\ y_a = \begin{pmatrix}y\\ \psi\end{pmatrix},\ T = \begin{pmatrix}X & Z\\ 0 & I\end{pmatrix},\ \text{and } W = \begin{pmatrix}\Sigma & 0\\ 0 & D\end{pmatrix},$$

where Lee et al. (2006) describe $\psi = 0$ as quasi-data with mean $\mathrm{E}(\psi) = b$. Their formulation suggests that the joint estimation of the coefficients β and b of the linear mixed effects model can be derived via a classical augmented general linear model $y_a = T\delta + \varepsilon$ where $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{var}(\varepsilon) = W$, with *both* β and b appearing as fixed parameters. The usefulness of this reformulation of an LME as a general linear model will be revisited.

4.2 Repeated Measurements in LME models

In many statistical analyzes, the need to determine parameter estimates where multiple measurements are available on each of a set of variables often arises. Further to Lam et al. (1999), Hamlett et al. (2004) performs an analysis of the correlation of replicate measurements, for two variables of interest, using LME models.

Let y_{Aij} and y_{Bij} be the jth repeated observations of the variables of interest A and B taken on the ith subject. The number of repeated measurements for each variable may differ for each individual. Both variables are measured on each time points. Let n_i be the number of observations for each variable, hence $2 \times n_i$ observations in total.

It is assumed that the pair y_{Aij} and y_{Bij} follow a bivariate normal distribution.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix}$$

The matrix Σ represents the variance component matrix between response variables at a given time point j.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix}$$

σ_A^2 is the variance of variable A, σ_B^2 is the variance of variable B and σ_{AB} is the covariance of the two variable. It is assumed that $\boldsymbol{\Sigma}$ does not depend on a particular time point, and is the same over all time points.

4.2.1 Formulation of the Response Vector

Information of individual i is recorded in a response vector $\boldsymbol{y}_i$. The response vector is constructed by stacking the response of the 2 responses at the first instance, then the 2 responses at the second instance, and so on. Therefore the response vector is a $2n_i \times 1$ column vector. The covariance matrix of $\boldsymbol{y_i}$ is a $2n_i \times 2n_i$ positive definite matrix $\boldsymbol{\Omega}_i$.

Consider the case where three measurements are taken by both methods A and B,

$\boldsymbol{y}_i$ is a 6×1 random vector describing the ith subject.

$$\boldsymbol{y}_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})\boldsymbol{\prime}$$

The response vector $\boldsymbol{y_i}$ can be formulated as an LME model according to Laird-Ware form.

$$\begin{aligned} \boldsymbol{y_i} &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b_i} + \boldsymbol{\epsilon_i} \\ \boldsymbol{b_i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{D}) \\ \boldsymbol{\epsilon_i} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{R_i}) \end{aligned}$$

Information on the fixed effects are contained in a three dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)\boldsymbol{\prime}$. For computational purposes β_2 is conventionally set to zero. Consequently $\boldsymbol{\beta}$ is the solutions of the means of the two methods, i.e. $E(\boldsymbol{y}_i) = \boldsymbol{X}_i\boldsymbol{\beta}$. The variance covariance matrix $\boldsymbol{D}$ is a general 2×2 matrix, while $\boldsymbol{R}_i$ is a $2n_i \times 2n_i$ matrix.

4.2.2 Decomposition of the response covariance matrix

The variance covariance structure can be re-expressed in the following form,

$$\mathrm{Cov}(\mathrm{y}_i) = \boldsymbol{\Omega_i} = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R_i}.$$

$\boldsymbol{R_i}$ can be shown to be the Kronecker product of a correlation matrix $\boldsymbol{V}$ and $\boldsymbol{\Lambda}$. The correlation matrix $\boldsymbol{V}$ of the repeated measures on a given response variable is assumed to be the same for all response variables. Both Hamlett et al. (2004) and Lam et al. (1999) use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for $\boldsymbol{V}$. Roy (2009a) remarks that, with repeated measures, the response for each subject is correlated for each variable, and that such correlation must be taken into account in order to produce a valid inference on correlation estimates. Roy (2006) proposes various correlation structures may be assumed for repeated measure correlations, such as the compound symmetry and autoregressive structures, as alternative to the identity matrix.

However, for the purposes of method comparison studies, the necessary estimates are currently only determinable when the identity matrix is specified, and the results in Roy (2009a) indicate its use.

For the response vector described, Hamlett et al. (2004) presents a detailed covariance matrix. A brief summary shall be presented here only. The overall variance matrix is a 6×6 matrix composed of two types of 2×2 blocks. Each block represents one separate time of measurement.

$$\boldsymbol{\Omega}_i = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{D} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{\Sigma} & \boldsymbol{D} \\ \boldsymbol{D} & \boldsymbol{D} & \boldsymbol{\Sigma} \end{pmatrix}$$

The diagonal blocks are Σ, as described previously. The 2×2 block diagonal matrix in $\boldsymbol{\Omega}$ gives $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is the sum of the between-subject variability $\boldsymbol{D}$ and the within subject variability $\boldsymbol{\Lambda}$.

$\boldsymbol{\Omega_i}$ can be expressed as

$$\boldsymbol{\Omega_i} = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + (\boldsymbol{I}_{\boldsymbol{n_i}} \otimes \boldsymbol{\Lambda}).$$

The notation $\dim_{n_i}$ means an $n_i \times n_i$ diagonal block.

4.2.3 Correlation terms

Hamlett et al. (2004) demonstrated how the between-subject and within subject variabilities can be expressed in terms of correlation terms.

$$\boldsymbol{D} = \begin{pmatrix} \sigma_A^2 \rho_A & \sigma_A \sigma_b \rho_{AB} \delta \\ \sigma_A \sigma_b \rho_{AB} \delta & \sigma_B^2 \rho_B \end{pmatrix}$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \sigma_A^2 (1 - \rho_A) & \sigma_{AB} (1 - \delta) \\ \sigma_{AB} (1 - \delta) & \sigma_B^2 (1 - \rho_B) \end{pmatrix}.$$

ρ_A describe the correlations of measurements made by the method A at different times. Similarly ρ_B describe the correlation of measurements made by the method B

at different times. Correlations among repeated measures within the same method are known as intra-class correlation coefficients. ρ_{AB} describes the correlation of measurements taken at the same same time by both methods. The coefficient δ is added for when the measurements are taken at different times, and is a constant of less than 1 for linked replicates. This is based on the assumption that linked replicates measurements taken at the same time would have greater correlation than those taken at different times. For unlinked replicates δ is simply 1. Hamlett et al. (2004) provides a useful graphical depiction of the role of each correlation coefficients.

4.3 Using LME for method comparison

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. Rather than using the 'by hand' methods, estimates for required LME parameters can be read directly from program output. Furthermore, using computer approaches removes constraints associated with 'by-hand' approaches, such as the need for the design to be perfectly balanced.

4.3.1 Roy's Approach

For the purposes of comparing two methods of measurement, Roy (2009a) presents a framework that utilizes linear mixed effects model. This methodology provides for the formal testing of inter-method bias, between-subject variability and within-subject variability of two methods. The formulation contains a Kronecker product covariance structure in a doubly multivariate setup. By doubly multivariate set up, Roy means that the information on each patient or item is multivariate at two levels, the number of methods and number of replicated measurements. Further to Lam et al. (1999), it is assumed that the replicates are linked over time. However it is easy to modify to the unlinked case.

Roy (2009a) sets out three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

A formal test for inter-method bias can be implemented by examining the fixed ef-

fects of the model. This is common to well known classical linear model methodologies. The null hypotheses, that both methods have the same mean, which is tested against the alternative hypothesis, that both methods have different means. The inter-method bias and necessary $t-$value and $p-$value are presented in computer output. A decision on whether the first of Roy's criteria is fulfilled can be based on these values.

Importantly Roy (2009a) further proposes a series of three tests on the variance components of an LME model, which allow decisions on the second and third of Roy's criteria. For these tests, four candidate LME models are constructed. The differences in the models are specifically in how the the D and Λ matrices are constructed, using either an unstructured form or a compound symmetry form. To illustrate these differences, consider a generic matrix A,

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

A symmetric matrix allows the diagonal terms a_{11} and a_{22} to differ. The compound symmetry structure requires that both of these terms be equal, i.e $a_{11} = a_{22}$.

The first model acts as an alternative hypothesis to be compared against each of three other models, acting as null hypothesis models, successively. The models are compared using the likelihood ratio test. Likelihood ratio tests are a class of tests based on the comparison of the values of the likelihood functions of two candidate models.

4.3.2 Correlation

In addition to the variability tests, Roy advises that it is preferable that a correlation of greater than 0.82 exist for two methods to be considered interchangeable. However if two methods fulfil all the other conditions for agreement, failure to comply with this one can be overlooked. Indeed Roy demonstrates that placing undue importance to it can lead to incorrect conclusions. Roy (2009a) remarks that current computer implementations only gives overall correlation coefficients, but not their variances. Consequently

it is not possible to carry out inferences based on all overall correlation coefficients.

4.3.3 Variability test 1

The first test determines whether or not both methods A and B have the same between-subject variability, further to the second of Roy's criteria.

$$H_0 : \ d_A = d_B$$
$$H_A : \ d_A \neq d_B$$

This test is facilitated by constructing a model specifying a symmetric form for D (i.e. the alternative model) and comparing it with a model that has compound symmetric form for D (i.e. the null model). For this test $\hat{\mathbf{\Lambda}}$ has a symmetric form for both models, and will be the same for both.

4.3.4 Variability test 2

This test determines whether or not both methods A and B have the same within-subject variability, thus enabling a decision on the third of Roy's criteria.

$$H_0 : \ \lambda_A = \lambda_B$$
$$H_A : \ \lambda_A = \lambda_B$$

This model is performed in the same manner as the first test, only reversing the roles of $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. The null model is constructed a symmetric form for $\hat{\mathbf{\Lambda}}$ while the alternative model uses a compound symmetry form. This time $\hat{\boldsymbol{D}}$ has a symmetric form for both models, and will be the same for both.

As the within-subject variabilities are fundamental to the coefficient of repeatability, this variability test likelihood ratio test is equivalent to testing the equality of two coefficients of repeatability of two methods. In presenting the results of this test, Roy (2009a) includes the coefficients of repeatability for both methods.

4.3.5 Variability test 3

The last of the variability test examines whether or not methods A and B have the same overall variability. This enables the joint consideration of second and third criteria.

$$H_0 : \ \sigma_A = \sigma_B$$
$$H_A : \ \sigma_A = \sigma_B$$

The null model is constructed a symmetric form for both $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\Lambda}}$ while the alternative model uses a compound symmetry form for both.

4.3.6 Demonstration of Roy's testing

Roy provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used. The first two examples used are from the 'blood pressure' data set introduced by Bland and Altman (1999). The data set is a tabulation of simultaneous measurements of systolic blood pressure were made by each of two experienced observers (denoted 'J' and 'R') using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted 'S'). Three sets of readings were made in quick succession. Roy compares the 'J' and 'S' methods in the first of her examples.

The inter-method bias between the two method is found to be 15.62 , with a $t-$value of -7.64, with a $p-$value of less than 0.0001. Consequently there is a significant inter-method bias present between methods J and S, and the first of the Roy's three agreement criteria is unfulfilled.

Next, the first variability test is carried out, yielding maximum likelihood estimates of the between-subject variance covariance matrix, for both the null model, in compound symmetry (CS) form, and the alternative model in symmetric (symm) form. These matrices are determined to be as follows;

$$\hat{\boldsymbol{D}}_{CS} = \begin{pmatrix} 946.50 & 784.32 \\ 784.32 & 946.50 \end{pmatrix}, \qquad \hat{\boldsymbol{D}}_{Symm} = \begin{pmatrix} 923.98 & 785.24 \\ 785.24 & 971.30 \end{pmatrix}.$$

A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the null model is -2030.7, and for the alternative model -2030.8. The test statistic, presented with greater precision than the log-likelihoods, is 0.1592. The $p-$value is 0.6958. Consequently we fail to reject the null model, and by extension, conclude that the hypothesis that methods J and S have the same between-subject variability. Thus the second of the criteria is fulfilled.

The second variability test determines maximum likelihood estimates of the within-subject variance covariance matrix, for both the null model, in CS form, and the alternative model in symmetric form.

$$\hat{\mathbf{\Lambda}}_{\boldsymbol{CS}} = \begin{pmatrix} 60.27 & 16.06 \\ 16.06 & 60.27 \end{pmatrix}, \qquad \hat{\mathbf{\Lambda}}_{Symm} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix}.$$

Again, A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the alternative model model is -2045.0. As before, the null model has a log-likelihood of -2030.7. The test statistic is computed as 28.617, again presented with greater precision. The $p-$value is less than 0.0001. In this case we reject the null hypothesis of equal within-subject variability. Consequently the third of Roy's criteria is unfulfilled. The coefficient of repeatability for methods J and S are found to be 16.95 mmHg and 25.28 mmHg respectively.

The last of the three variability tests is carried out to compare the overall variabilities of both methods. With the null model the MLE of the within-subject variance covariance matrix is given below. The overall variabilities for the null and alternative models, respectively, are determined to be as follows;

$$\hat{\mathbf{\Sigma}}_{CS} = \begin{pmatrix} 1007.92 & 801.65 \\ 801.65 & 1007.92 \end{pmatrix}, \qquad \hat{\mathbf{\Sigma}}_{Symm} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix},$$

The log-likelihood of the alternative model model is -2045.2, and again, the null model has a log-likelihood of -2030.7. The test statistic is 28.884, and the $p-$value is less than 0.0001. The null hypothesis, that both methods have equal overall variability,

is rejected. Further to the second variability test, it is known that this difference is specifically due to the difference of within-subject variabilities.

Lastly, Roy considers the overall correlation coefficient. The diagonal blocks $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii}$ of the correlation matrix indicate an overall coefficient of 0.7959. This is less than the threshold of 0.82 that Roy recommends.

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

The off-diagonal blocks of the overall correlation matrix $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'}$ present the correlation coefficients further to Hamlett et al. (2004).

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'} = \begin{pmatrix} 0.9611 & 0.7799 \\ 0.7799 & 0.9212 \end{pmatrix}.$$

The overall conclusion of the procedure is that method J and S are not in agreement, specifically due to the within-subject variability, and the inter-method bias. The repeatability coefficients are substantially different, with the coefficient for method S being 49% larger than for method J. Additionally the overall correlation coefficient did not exceed the recommended threshold of 0.82.

4.4 Limits of agreement in LME models

Limits of agreement are used extensively for assessing agreement, because they are intuitive and easy to use. Necessarily their prevalence in literature has meant that they are now the best known measurement for agreement, and therefore any newer methodology would benefit by making reference to them.

Carstensen et al. (2008) uses LME models to determine the limits of agreement. Between-subject variation for method m is given by d_m^2 and within-subject variation is given by λ_m^2. Carstensen et al. (2008) remarks that for two methods A and B, separate values of d_A^2 and d_B^2 cannot be estimated, only their average. Hence the assumption that $d_x = d_y = d$ is necessary. The between-subject variability $\boldsymbol{D}$ and within-subject variability $\boldsymbol{\Lambda}$ can be presented in matrix form,

$$\boldsymbol{D} = \begin{pmatrix} d_A^2 & 0 \\ 0 & d_B^2 \end{pmatrix} = \begin{pmatrix} d^2 & 0 \\ 0 & d^2 \end{pmatrix}, \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_A^2 & 0 \\ 0 & \lambda_B^2 \end{pmatrix}.$$

The variance for method m is $d_m^2 + \lambda_m^2$. Limits of agreement are determined using the standard deviation of the case-wise differences between the sets of measurements by two methods A and B, given by

$$\text{var}(y_A - y_B) = 2d^2 + \lambda_A^2 + \lambda_B^2. \tag{4.7}$$

Importantly the covariance terms in both variability matrices are zero, and no covariance component is present.

Carstensen et al. (2008) presents a data set 'fat', which is a comparison of measurements of subcutaneous fat by two observers at the Steno Diabetes Center, Copenhagen. Measurements are in millimeters (mm). Each person is measured three times by each observer. The observations are considered to be 'true' replicates.

A linear mixed effects model is formulated, and implementation through several software packages is demonstrated. All of the necessary terms are presented in the computer output. The limits of agreement are therefore,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{4.8}$$

Roy (2009a) has demonstrated a methodology whereby d_A^2 and d_B^2 can be estimated separately. Also covariance terms are present in both $\boldsymbol{D}$ and $\boldsymbol{\Lambda}$. Using Roy's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{4.9}$$

All of these terms are given or determinable in computer output. The limits of agreement can therefore be evaluated using

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{4.10}$$

For Carstensen's 'fat' data, the limits of agreement computed using Roy's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

4.4.1 Linked replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

Carstensen et al. (2008) introduces a second data set; the oximetry study. This study done at the Royal Childrens Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

In most cases, measurements were taken by both method at three different times. In some cases there are either one or two pairs of measurements, hence the data is unbalanced. Carstensen et al. (2008) describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

Carstensen et al. (2008) demonstrate the necessity of accounting for linked replicated by comparing the limits of agreement from the 'oximetry' data set using a model

with the additional term, and one without. When the interaction is accounted for the limits of agreement are (-9.62,14.56). When the interaction is not accounted for, the limits of agreement are (-11.88,16.83). It is shown that the failure to include this additional term results in an over-estimation of the standard deviations of differences.

Limits of agreement are determined using Roy's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. Roy's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of Roy's model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of Roy's model and the modified model (denoted 1 and 2 respectively);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{4.11}$$

(The variance of the additional random effect in model 2 is 3.01.)

Akaike (1974) introduces the Akaike information criterion (AIC), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.Two candidate models can said to be equally good if there is a difference of less than 2 in their AIC values.

The Akaike information criterion (AIC) for both models are $AIC_1 = 2304.226$ and $AIC_2 = 2306.226$, indicating little difference in models. The AIC values for the Carstensen 'unlinked' and 'linked' models are 1994.66 and 1955.48 respectively, indicating an improvement by adding the interaction term.

The $\hat{\mathbf{\Lambda}}$ matrices are informative as to the difference between Carstensen's unlinked and linked models. For the oximetry data, the covariance terms (given above as 11.67 and 2.6 respectively) are of similar magnitudes to the variance terms. Conversely for the 'fat' data the covariance term (-0.00032) is negligible. When the interaction term is added to the model, the covariance term remains negligible. (For the 'fat' data, the difference in AIC values is also approximately 2).

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\boldsymbol{\Lambda}}$. Therefore the test's proposed by Roy (2009a) can not be implemented. Conversely, accurate limits of agreement as determined by Carstensen's model may also be found using Roy's method. Addition of the interaction term erodes the capability of Roy's methodology to compare candidate models, and therefore shall not be adopted.

Finally, to complement the blood pressure (i.e.'J vs S') method comparison from the previous section (i.e.'J vs S'), the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

Chapter 5

Introduction

5.1 LME models in method comparison studies

Barnhart et al. (2007) describes the sources of disagreement in a method comparison study problem as differing population means, different between-subject variances, different within-subject variances between two methods and poor correlation between measurements of two methods. Further to this, Roy (2009b) states three criteria for two methods to be considered in agreement. Firstly that there be no significant bias. Second that there is no difference in the between-subject variabilities, and lastly that there is no significant difference in the within-subject variabilities. Roy further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

Roy (2009b) further proposes examination of the the overall variability by considering the second and third criteria be examined jointly. Should both the second and third criteria be fulfilled, then the overall variabilities of both methods would be equal.

The LME model approach has seen increased use as a framework for method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples)

Linear mixed effects (LME) models can facilitate greater understanding of the po-

tential causes of bias and differences in precision between two sets of measurement.

Lai and Shiao (2005) views the uses of linear mixed effects models as an expansion on the Bland-Altman methodology, rather than as a replacement.Lai and Shiao (2005) view the LME Models approach as an natural expansion to the Bland ? Altman method for comparing two measurement methods. Their focus is to explain lack of agreement by means of additional covariates outside the scope of the traditional method comparison problem. Lai and Shiao (2005) is interesting in that it extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable.

Lai and Shiao (2005) extends the usual method comparison study question. It correctly identifies LME models as a methodoloy that can used to make such questions tractable. The Data Set used in their examples is unavailable for independent use. Therefore, for the sake of consistency, a data set will be simulated based on the Blood Data that will allow for extra variables. Carstensen et al. (2008) remarks that modern statistical computation, such as that used for LME models, greatly improve the efficiency of calculation compared to previous 'by-hand' methods.

Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches, such as LME models, to method comparison studies. Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. Rather than using the 'by hand' methods, estimates for required LME parameters can be read directly from program output.

Carstensen et al. (2008) remarks upon 'by-hand' approaches advocated in Bland and Altman (1999) discouragingly, describing them as tedious, unnecessary and 'outdated'. Due to the prevalence of modern statistical software, Carstensen et al. (2008) advocates the adoption of computer based approaches to method comparison studies, allowing the use of LME models that would not have been feasible otherwise. Rather than using the 'by hand' methods, estimates for required parameters can be gotten directly from output code. Furthermore, using computer approaches removes constraints, such as

the need for the design to be perfectly balanced. In part this is due to the increased profile of LME models, and furthermore the availability of capable software.

Additionally a great understanding of residual analysis and influence analysis for LME models has been adchieved thanks to authors such as **?**, **?**, Cook (1986) West et al. (2007), amongst others. In this chapter various LME approaches to method comparison studies shall be examined.

Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

Roys uses and LME model approach to provide a set of formal tests for method comparison studies.

5.2 Introduction to LME Models, Fitting LME Models to MCS Data

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland and Altman (1999) suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods. The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be incorrect, (Carstensen, 2004). Carstensen (2004) recommends that replicate measurements for each method, but recognizes that resulting data are more difficult to analyze. To this end, Carstensen (2004) and Carstensen et al. (2008) recommend the use of LME models as a suitable framework for method comparison in the case of repeated measurements.

Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them.

This approach has seen increased use in method comparison studies in recent years (Lai & Shaio, Carstensen and Choudhary as examples). In part this is due to the increased profile of LME models, and furthermore the availability of capable software. Additionally LME based approaches may utilise the diagnostic and influence analysis techniques that have been developed in recent times.

In this section, we introduce the LME model, discusss how it can be applied to MCS problems, and how it is desirable in the case of replicate measurements, giving some examples from previous work (i.e. Carstensen et Al, Lai & Shaio, and Roy).

Further to that, there will be a demonstration on fitting various types LME models using freely available software.

While the MCS problem is conventionally poised in the context of two methods of measurements, LME models allow for a straightforward analysis whereby several methods of measurement can be measured simulataneously. However simple models only can only indicate agreement of lack thereof, and the presence of inter-method bias. To consider more complex questions, more complex LME models are required. Useful approaches will be introduced in a later section.

5.3 Definition of Replicate Measurements (Move to Chapter 1)

5.4 Definition of Replicate measurements

Further to Bland and Altman (1999), a formal definition is required of what exactly replicate measurements are

By replicates we mean two or more measurements on the same individual taken in identical conditions. In general this requirement means that the measurements are taken in quick succession.

Roy accords with Bland and Altman?s definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

5.5 Model for replicate measurements

We generalize the single measurement model for the replicate measurement case, by additionally specifying replicate values. Let y_{mir} be the $r-$th replicate measurement

for subject "i" made by method "m". Further to **?** fixed effect can be expressed with a single term α_{mi}, which incorporate the true value μ_i.

$$y_{mir} = \mu_i + \alpha_m + e_{mir}$$

Combining fixed effects (**?**), we write,

$$y_{mir} = \alpha_{mi} + e_{mir}.$$

The following assumptions are required

- e_{mir} is independent of the fixed effects with mean $\mathrm{E}(e_{mir}) = 0$.
- Further to **?** between-item and within-item variances $\mathrm{Var}(\alpha_{mi}) = \sigma^2_{Bm}$ and $\mathrm{Var}(e_{mir}) = \sigma^2_{Wm}$

5.6 Carstensen's Model

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

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m) \tag{5.1}$$

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m), \quad c_{mi} \sim \mathcal{N}(0, \tau^2_m). \tag{5.2}$$

Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

5.7 Two Way ANOVA

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

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \epsilon_{mir}. \tag{5.3}$$

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

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

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

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

5.8 Statistical Model For Replicate Measurements

Let y_{Aij} and y_{Bij} be the jth repeated observations of the variables of interest A and B taken on the ith subject. The number of repeated measurements for each variable may differ for each individual. Both variables are measured on each time points. Let n_i be the number of observations for each variable, hence $2 \times n_i$ observations in total.

It is assumed that the pair y_{Aij} and y_{Bij} follow a bivariate normal distribution.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \tag{5.4}$$

The matrix $\boldsymbol{\Sigma}$ represents the variance component matrix between response variables at a given time point j.

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \tag{5.5}$$

σ_A^2 is the variance of variable A, σ_B^2 is the variance of variable B and σ_{AB} is the covariance of the two variable. It is assumed that $\boldsymbol{\Sigma}$ does not depend on a particular time point, and is the same over all time points.

5.9 Exchangeable and Linked measurements

5.10 Sampling Scheme : Linked and Unlinked Replicates

Measurements taken in quick succession by the same observer using the same instrument on the same subject can be considered true replicates. Roy (2009b) notes that some measurements may not be 'true' replicates.

Roy's methodology assumes the use of 'true replicates'. However data may not be collected in this way. In such cases, the correlation matrix on the replicates may require a different structure, such as the autoregressive order one $AR(1)$ structure. However

determining MLEs with such a structure would be computational intense, if possible at all.

One important feature of replicate observations is that they should be independent of each other. In essence, this is achieved by ensuring that the observer makes each measurement independent of knowledge of the previous value(s). This may be difficult to achieve in practice. (Check who said this)

5.11 Replicate measurements

Roy (2009b) accords with Bland and Altman?s definition of a replicate, as being two or more measurements on the same individual under identical conditions. Roy allows the assumption that replicated measurements are equi-correlated. Roy allows unequal numbers of replicates.

Replicate measurements are linked over time. However the method can be easily extended to cover situations where they are not linked over time.

In this model , the variances of the random effects must depend on m, since the different methods do not necessarily measure on the same scale, and different methods naturally must be assumed to have different variances. Carstensen (2004) attends to the issue of comparative variances.

Bland and Altman (1999) also remark that an important feature of replicate observations is that they should be independent of each other. This issue is addressed by Carstensen (2010), in terms of exchangeability and linkage. Carstenen advises that repeated measurements come in two *substantially different* forms, depending on the circumstances of their measurement: exchangable and linked.

Repeated measurements are said to be exchangeable if no relationship exists between successive measurements across measurements. If the condition of exchangeability exists, a group of measurement of the same item determined by the same method can be re-arranged in any permutation without prejudice to proper analysis. There is no reason to believe that the true value of the underlying variable has changed over the course of the measurements.

Exchangeable repeated measurements can be treated as true replicates. For the purposes of method comparison studies the following remarks can be made. The $r-$th measurement made by method 1 has no special correspondence to the $r-$th measurement made by method 2, and consequently any pairing of repeated measurements are as good as each other.

Repeated measurements are said to be linked if a direct correspondence exists between successive measurements across measurements, i.e. pairing. Such measurements are commonly made with a time interval between them, but simultaneously for both methods. Paired measurements are exchangeable, but individual measurements are not.

If the paired measurements are taken in a short period of time so that no real systemic changes can take place on each item, they can be considered true replicates. Should enough time elapse for systemic changes, linked repeated measurements can not be treated as true replicates.

Bibliography

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

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.

Altman, D. G. (1991). *Practical Statistics for Medical Research*. Chapman and Hall.

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.

Carroll, R. and D. Ruppert (1996). The use and misuse of orthogonal regression in linear errors-in-variables models. *The American Statistician 50*(1), 1–6.

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

Carstensen, B. (2010). *Comparing Clinical Measurement Methods : A Practical Guide*. Wiley.

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.

Cook, R. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological) 48*(2), 133–169.

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.

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the em algorithme. *ournal of the Royal Statistical Society. Series B 39*(1), 1–38.

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.

Dodge, Y. (2003). *The Oxford Dictionary of Statistical Terms*. Oxford University Press.

Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics 3*(1), 1–21.

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

Fisher, R. (1918). The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh 2*, 399–433.

Green, P. J. and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*. Chapman & Hall Ltd.

Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association 43*, 243–264.

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

Hamlett, A., L. Ryan, and R. Wolfinger (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *Proceedings of the Statistics and Data Analysis Section, SAS Users Group International 198-229*, 1–7.

Hartley, H. and J. Rao (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika 54*(1/2), 93–108.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics 9*(2), 226–252.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1963). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1973). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C., O. Kempthorne, S. Searle, and C. von Krosigk (1984). The estimation of environmental and genetic trends from records subject to culling. *Biometrics 15*, 192–218.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics 21*, 309–310.

Henderson, C. R. (1973). Sire evaluation and genetic trends. In *Proceedings of the Animal Breeding and genetics Symposium in Honor of Dr Jay L. Lush*, pp. 10–41. Champaign, Illinois: Americian Society of Animal Science and American Dairy Science Association.

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`.

Kinsella, A. (1986). Estimating method precision. *The Statistician 35*, 421–427.

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

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

Kummel, C. (1879). Reduction of observation equations which contain more than one observed quantity. *The Analyst 6*, 97–105.

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

Laird, N. M. and J. H. Ware (1982). Random-effects models for longitudinal data. *Biometrics 38*(4), 963–974.

Lam, M., K. Webb, and D. O'Donnell (1999). Correlation between two variables in repeated measurements. *American Statistical Association, Proceedings of the Biometric Session*, 213–218.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models (Disc: P656-678). *Journal of the Royal Statistical Society, Series B: Methodological 58*, 619–656.

Lee, Y., J. A. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. Chapman & Hall Ltd.

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

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

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

Lin, S. C., D. M. Whipple, and charles S Ho (1991). Evaluation of statistical equivalence using limits of agreement and associates sample size calculation. *Communications in Statistics - Theory and Methods 27*(6), 1419–1432.

Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry 44*, 1024–1031.

Linnet, K. (1999). Necessary sample size for method comparison studies based on regression analysis. *Clinical Chemistry 45*(6), 882–894.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

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

Mantha, S., M. F. Roizen, L. A. Fleisher, R. Thisted, and J. Foss (2000). Comparing methods of clinical measurement: Reporting standards for bland and altman analysis. *Anaesthesia and Analgesia 90*, 593–602.

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

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.

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.

Patterson, H. and R. Thompson (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika 58*(3), 545–554.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Phelps, C. and A. Hutson (1995). Estimating diagnostic test accuracy using a fuzzy gold standard. *Medical decision making 15*, 144–57.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.

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

Pizzi, N. (1999). Fuzzy pre-processing of gold standards as applied to biomedical spectra classification. *Artificial Intelligence in Medicine 16*, 171–182.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (Disc: P32-51). *Statistical Science 6*, 15–32.

Roy, A. (2006). Estimating correlation coefficient between two variables with repeated observations using mixed effects models. *Biometric Journal 2*, 286–301.

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.

Singer, J. D. (1998). Using sas proc mixed to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics 24*(4), 323–355.

Thompson, W. (1963). Precision of simultaneous measurement procedures. *Journal of American Statistical Association 58*, 474–479.

Tippett, L. (1931). *The Methods of Statistics* (1st ed.). London: Williams and Norgate.

Voelkel, J. G. and B. E. Siskowski (2005). Center for quality and applied statistics kate gleason college of engineering rochester institute of technology technical report 2005–3.

West, B., K. Welch, and A. Galecki (2007). *Linear Mixed Models: a Practical Guide Using Statistical Software*. Chapman and Hall CRC.

Zhang, Y., S. Nitter-Hauge, H. Ihlen, K. Rootwelt, and E. Myhre (1986). Measurement of aortic regurgitation by doppler echocardiography. *British Heart Journal 55*, 32–38.