Contents

Chapter 1

Method Comparison Studies

1.1 What is a method comparison study?

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'. Ludbrook (1997) states that the purpose of comparing two measurements "of a continuous biological variable" is to uncover systematic differences, not to point to similarities". The need to compare the results of two different measurement techniques is common in medical statistics. 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'.

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.

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).

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. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

The absence of inter-method bias by itself is not sufficient to establish whether 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. With this in mind a methodology is required that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

Round	Fotobalk (F)	Counter (C)	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.

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. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

1.2 Agreement

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

To carry their idea a step further, we define a specific numerical measure of agreement as twice the expected squared perpendicular distance of the pair of random variables (X_1, X_2) to the line of equality or agreement in the (X_1, X_2)-plane, that is, $E(X_1 - X_2)2$, where X1 and X2 denote the continuous measurements of rater 1 and rater 2, respectively.

Obviously, other L_p norms may be considered for the purpose of numerically measuring agreement and warrant future consideration. Note that we will use the term rater and measuring device interchangeably throughout this article.

Agreement is the extent to which the measure of the variable of interest, under a constant set of experimental conditions, yields the same result on repeated trials (Sanchez et al). The more consistent the results, the more reliable the measuring procedure.

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 either 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.

1.3 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. Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

1.4 Gold Standard

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

1.5 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.

1.6 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.

Two methods may agree on average, but they may exhibit differences over a range of measurements. Proportional Bias is a difference in the two measures which is proportional to the scale of the measurement.

Using a naive estimation of bias, such as the mean of differences, it may incorrectly indicate absence of bias, by yielding a mean difference close to zero. This would be caused by positive differences in the measurements at one end of the range of measurements being canceled out by negative differences at the other end of the scale.

1.7 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* and propose estimating its value by determining the mean of the differences. The variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

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.

1.8 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.

1.9 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.

Historically comparison of two methods of measurement was carried out by use of paired sample $t-$test, correlation coefficients or simple linear regression. Simple linear regression is unsuitable for method comparison studies because of the required assumption that one variable is measured without error. In comparing two methods, both methods are assume to have attendant random error.

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, taken from 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, referred to here as 'Fotobalk', 'Counter' and 'Terma'.

1.10 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.

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.

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

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

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 **??** 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.

Chapter 2

Improper MCS Techniques

2.1 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.

2.2 Inappropriate Methodologies

Use of the Pearson Correlation Coefficient , although seemingly intuitive, is not appropriate approach to assessing agreement of two methods. Arguments against its usage have been made repeatedly in the relevant literature. It is possible for two analytical methods to be highly correlated, yet have a poor level of agreement.

2.3 Paired T tests

This method can be applied to test for statistically significant deviations in bias. This method can be potentially misused for method comparison studies.
It is a poor measure of agreement when the rater's measurements are perpendicular to the line of equality[Hutson et al]. In this context, an average difference of zero between the two raters, yet the scatter plot displays strong negative correlation.

2.4 Inappropriate use of the Correlation Coefficient

It is well known that Pearson's correlation coefficient is a measure of the linear association between two variables, not the agreement between two variables (e.g., see Bland and Altman 1986).

This is a well known as a measure of linear association between two variables. Nonetheless this is not necessarily the same as Agreement. This method is considered wholly inadequate to assess agreement because it only evaluates only the association of two sets of observations.

It is intuitive when dealing with two sets of related data, i.e the results of the two raters, to calculate the correlation coefficient (r). Bland and Altman attend to this in

their 1999 paper.

They present a data set from two sets of meters, and an accompanying scatterplot. An hypothesis test on the data set leads us to conclude that there is a relationship between both sets of meter measurements. The correlation coeffiecient is determined to be $r = 0.94$. However, this high correlation does not mean that the two methods agree. It is possible to determine from the scatterplot that the intercept is not zero, a requirement for stating both methods have high agreement. Essentially, should two methods have highly correlated results, it does not follow that they have high agreement.

Intra-class correlation coefficient

- 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.

2.5 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 $i = 1, 2, ...n$.

$$X_i = \tau_i + \delta_i \tag{2.1}$$

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

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 $\beta = 1$.

$$E(X_i) = \tau_i \tag{2.2}$$

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

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

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

$$z_0 = d = 0 \tag{2.3}$$

$$z_{n+1} = z_n^2 + c \tag{2.4}$$

Chapter 3

Formal Testing Procedures

3.0.1 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.

3.1 Model Formulation and Formal Testing

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

$$\begin{aligned} Y_{ij} = \quad & \mu_j + S_i + \epsilon_{ij} \quad i = 1, 2...n \quad j = 1, 2 \\ & S \sim N(0, \sigma_s^2) \qquad \epsilon_{ij} \sim N(0, \sigma_j^2) \end{aligned} \tag{3.1}$$

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

$$\begin{aligned} var(Y_{ij}) = \quad & \sigma_s^2 + \sigma_j^2 \\ cov(Y_{i1}, Y_{i2}) = \quad & \sigma_s^2 \end{aligned} \tag{3.2}$$

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

$$\begin{aligned} \hat{\sigma_s^2} \quad &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad = Sxy \\ \hat{\sigma_1^2} \quad &= \sum \frac{(x_i - \bar{x})^2}{n-1} \quad = S^2x - Sxy \\ \hat{\sigma_2^2} \quad &= \sum \frac{(y_i - \bar{y})^2}{n-1} \quad = S^2y - Sxy \end{aligned} \tag{3.3}$$

The standard error of these variance estimates are:

$$\begin{aligned} var(\sigma_1^2) = \quad & \frac{2\sigma_1^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \\ var(\sigma_2^2) = \quad & \frac{2\sigma_2^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \end{aligned} \tag{3.4}$$

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.

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}} \tag{3.5}$$

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.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned} \tag{3.6}$$

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

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_S^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix} \tag{3.7}$$

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

$$\begin{aligned} \hat{\sigma_1^2} &= \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\ \hat{\sigma_2^2} &= \sum (y_{i2} - \bar{y_2})(D_i - \bar{D}) \end{aligned} \tag{3.8}$$

3.1.1 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;

$$t^* = \bar{D} / \frac{S_D}{\sqrt{n}} \tag{3.9}$$

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

3.1.2 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;

$$\rho(D, A) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}} \tag{3.10}$$

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.

The Pitman-Morgan test for equal variances is based on the correlation of D with S. The correlation coefficient is zero if, and only if, the variances are equal. The test statistic is the familiar t-test with n-2 degree of freedom. Bradley and Blackwood (1989) construct the conditional expectation of D given S as linear model. They used this result to propose a test of the joint hypothesis of the mean difference and equal variances. If the intercept and slope estimates are zero, the two methods have the same mean and variance. The Pitman-Morgan test is equivalent to the marginal test of the slope estimate in Bradley-Blackwoods model.

3.1.3 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.

3.2 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 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 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.

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

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

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 $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$. The variance matrix $\Sigma_{(a,d)}$ is

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \tag{3.11}$$

3.3 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.

$$\Delta_1 > \frac{C_{xy} - t(\frac{|A|}{n-1}))^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-1}))^{\frac{1}{2}}}, \tag{3.12}$$

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.

Thompson (1963) presents three relations that hold simultaneously with probability $1-2\alpha$ where $2\alpha = 0.01$ or 0.05. Thompson (1963) contains tables for K and M.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned} \tag{3.13}$$

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

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_S^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix} \tag{3.14}$$

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

$$\begin{aligned} \hat{\sigma_1^2} &= \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\ \hat{\sigma_2^2} &= \sum (y_{i2} - \bar{y_2})(D_i - \bar{D}) \end{aligned} \tag{3.15}$$

3.4 Model Formulation and Formal Testing

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

$$\begin{aligned} Y_{ij} = \quad \mu_j + S_i + \epsilon_{ij} \quad i = 1, 2...n \quad j = 1, 2 \\ S \sim N(0, \sigma_s^2) \qquad \epsilon_{ij} \sim N(0, \sigma_j^2) \end{aligned} \tag{3.16}$$

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

$$\begin{aligned} var(Y_{ij}) = &\quad \sigma_s^2 + \sigma_j^2 \\ cov(Y_{i1}, Y_{i2}) = &\quad \sigma_s^2 \end{aligned} \tag{3.17}$$

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

$$\begin{aligned} \hat{\sigma_s^2} &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} &= Sxy \\ \hat{\sigma_1^2} &= \sum \frac{(x_i - \bar{x})^2}{n-1} &= S^2x - Sxy \\ \hat{\sigma_2^2} &= \sum \frac{(y_i - \bar{y})^2}{n-1} &= S^2y - Sxy \end{aligned} \tag{3.18}$$

The standard error of these variance estimates are:

$$\begin{aligned} var(\sigma_1^2) = &\quad \frac{2\sigma_1^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \\ var(\sigma_2^2) = &\quad \frac{2\sigma_2^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \end{aligned} \tag{3.19}$$

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.

$$\Delta_1 > \quad \frac{C_{xy} - t(|A|/n-2))^{\frac{1}{2}}}{C_x - C_{xy} + t(|A|/n-2))^{\frac{1}{2}}} \tag{3.20}$$

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.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned} \tag{3.21}$$

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

$$\Sigma = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma_S^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix} \tag{3.22}$$

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

$$\begin{aligned} \hat{\sigma_1^2} &= \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\ \hat{\sigma_2^2} &= \sum (y_{i2} - \bar{y_2})(D_i - \bar{D}) \end{aligned} \tag{3.23}$$

3.5 Identifiability

Dunn (2002) 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). Dunn (2002) 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 3.5.1: 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 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.

3.6 Morgan Pitman Testing

An early contribution to formal testing in method comparison was made by both **?** and **?**, in separate contributions.

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;

$$\rho(a,d) = \quad \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}} \tag{3.24}$$

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;

$$\rho(D,A) = \quad \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_S^2 + \sigma_1^2 + \sigma_2^2)}}. \tag{3.25}$$

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.

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 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.

3.7 Morgan Pitman

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.

3.8 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,

$$t^* = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \tag{3.26}$$

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.

- 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!

Structural Equation Modelling

Authors, such as a Lewis et al. (1991), **?** 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.

Chapter 4

Regression Procedures

4.1 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).

4.2 Regression Methods

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

The use of regression models that assumes the presence of error in both variables X

and Y have been proposed for use instead (Cornbleet and Cochrane, 1979; Ludbrook, 1997). These methodologies are collectively known as 'Model II regression'. They differ in the method used to estimate the parameters of the regression.

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

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

4.2.1 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.

$$D = (X_1 - X_2) \tag{4.1}$$

$$M = (X_1 + X_2)/2 \tag{4.2}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{4.3}$$

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.

4.3 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 $\hat{\beta}_1$).

$$D = (X_1 - X_2) \tag{4.4}$$

$$M = (X_1 + X_2)/2 \tag{4.5}$$

The Bradley Blackwood Procedure fits D on M as follows:

$$D = \beta_0 + \beta_1 M \tag{4.6}$$

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 $\beta_0 = \beta_1 = 0$)

- 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.

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:

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

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

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.

4.4 Bartko's BB

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:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg} \tag{4.7}$$

	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 4.4.2: 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 4.10. 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.

	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 4.4.3: 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.

4.5 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.

4.6 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 $\bar{Y} - \hat{\beta}\bar{X}$;

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + [(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2]^{1/2}}{2S_{xy}} \tag{4.8}$$

, 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 4.6.4: 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.

4.7 Other Types of Studies / gold Standards

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

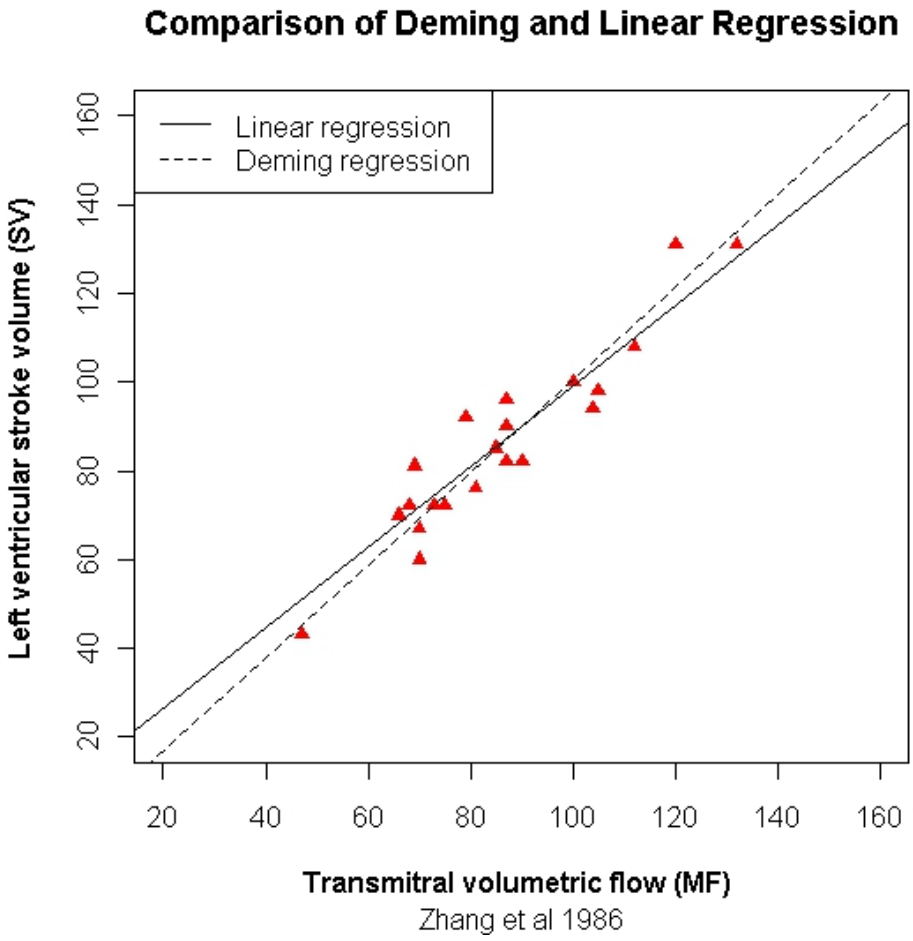

Figure 4.6.1: Deming Regression For Zhang's Data

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 mea-

surement 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.

Dunn (2002, 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' (Dunn, 2002). 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 Dunn (2002). 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.

4.8 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.

Chapter 5

Repeated Measurements and Repeatability

5.1 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.

5.2 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.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \text{ where } \boldsymbol{\mu} = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \tag{5.1}$$

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

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \tag{5.2}$$

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.

5.3 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 $\mathrm{Var}(e_{mir}) = \sigma^2_{Wm}$

5.4 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.

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.

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.

5.5 Linkage

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.

5.6 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);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{5.3}$$

(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 $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

5.7 Exchangeable and Linked measurements

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.

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

5.8 Replicate Measurements

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

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

However, because of the removal of the effects of the replicate measurements error,

this would cause the estimation of the standard deviation of the differences to be unduly small. Bland and Altman (1986) propose a correction for this.

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

5.9 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).

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).

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

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 prespecified 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%.

5.10 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.

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.

5.11 Coefficient of Repeatability - Good

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).

5.12 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.

5.13 Repeatability and Gold Standards

Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to **?**, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer,

which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability.

? 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.

5.14 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.

5.15 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.

5.16 Repeatability coefficient from LME Models-Chapter 2

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.

5.17 Carstensen Move to Chapter 2

- 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 $2.8\sigma_m$.

5.18 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.

5.19 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)

5.20 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.

5.21 Lin's Reproducibility Index

Lin proposes the use of a reproducibility index, called the Concordance Correlation Coefficent (CCC).While it is not strictly a measure of agreement as such, it can form

part of an overall method comparision methodology.

5.22 Outline of Thesis

In the first chapter the study of method comparison is introduced, while the second chapter provides a review of current methodologies. The intention of this thesis is to progress the study of method comparison studies, using a statistical method known as Linear mixed effects models. Chapter three shall describes linear mixed effects models, and how the use of the linear mixed effects models have so far extended to method comparison studies. Implementations of important existing work shall be presented, using the `R` programming language.

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

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

Bibliography

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. 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.

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.

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.

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

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.

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.

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.

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.

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., 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.

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.

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.

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

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.

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.

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

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.

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.