Contents

Chapter 1

Appendices

1.1 Appendix 1 : Improper Method Comparison Techniques

The issue of whether two measurement methods are comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research. Historically, comparison of two methods of measurement was carried out by use of paired sample $t-$test, simple linear regression, or correlation coefficients.

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 as a t random variable with $n - 1$ degrees of freedom and is calculated as follows,

$$t^* = \frac{\bar{d}}{s_d/\sqrt{n}} \tag{1.1}$$

where $\bar{d}$ and s_d is the average of the differences of the n observations. This method can be potentially misused for method comparison studies. Paired $t-$tests test only whether the mean responses are the same, and so provides a useful test for inter-method bias. However, no insight can be obtained about the variability of the case-wise differences

by the paired $t-$test, critically undermining it as a stand-alone procedure. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

The Correlation Coefficient

Correlation is inadequate to assess agreement because it only evaluates only the linear association of two sets of observations. Nonetheless linear association is not the same as agreement. It is possible for two methods to be highly correlated, yet have poor agreement due to any combination of constant and proportional bias. Arguments against its usage have been made repeatedly in the relevant literature, with Altman and Bland (1983), Bland and Altman (1986), **?** and **?** as examples.

Regression Methods

On account of the fact that one set of measurements are linearly related to another, one could surmise that simple linear Regression is the most suitable approach to analyzing comparisons. However simple linear regression is considered by many authors to be wholly unsuitable for method comparison studies (Altman and Bland, 1983; Cornbleet and Cochrane, 1979; Ludbrook, 1997). Simple linear regression is defined as such with the name 'Model I regression' by Cornbleet and Cochrane (1979), in contrast to 'Model II regression' models, which shall be discussed later on.

A key assumptions of simple linear regression is that the independent variable values are without random error. For method comparison studies, both sets of measurement must be assumed to be measured with imprecision and neither case can be taken to be a reference method. Arbitrarily selecting either method as the reference (i.e. the independent variable) will yield conflicting outcomes: a regression of X on Y would yield an entirely different model from fitting Y on X.

Further criticisms of linear regression exist. Firstly regression methods are uninformative about the variability of the differences. Secondly regression models are unduly

influenced by outliers. Lastly, regression models can not be used to effectively analyze repeated measurements.

The Identity Plot

Altman and Bland (1983) states that regression analysis can offer useful insights, and recommending an 'Identity Plot', a simple graphical approach that yields a cursory examination of how well the measurement methods agree. In the case of good agreement, the co-variates of the Identity plot accord closely with the $X = Y$ line. This plot is not useful for a thorough examination of the data. O'Brien et al. (1990) notes that data points will tend to cluster around the line of equality, obscuring interpretation. An identity plot shall complement demonstrations of commonly used approaches in the next chapter.

Decomposition of Inter-Method Bias

Regression approaches are useful for a making a detailed examination of the biases across the range of measurements, allowing inter-method bias to be decomposed into constant bias and proportional bias. Regression methods can determine the presence of inter-method bias, and the levels of constant bias and proportional bias thereof Ludbrook (1997, 2002).

Constant bias describes the case where one method gives values that are consistently different to the other across the whole range. 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. Proportional Bias exists when two methods agree on average, but exhibit differences over a range of measurements, i.e. the differences are proportional to the scale of the measurement. A measurement method may be subject to any combination of fixed bias or proportional bias, or both (Ludbrook, 2002).

Constant or proportional bias using linear regression can be detected by an individual test on the intercept or the slope of the line regressed from the results of the two methods to be compared. If there is no constant bias, the intercept is equal to zero and, similarly, if there is no proportional bias, the slope is equal to one. Thus, carrying out hypothesis tests on these coefficients (where the null hypotheses are $\beta_0 = 0$ and $\beta_1 = 1$) allow us to test for the presence of both types of bias.

If the basic assumptions underlying linear regression are not met, the regression equation, and consequently the estimations of bias are undermined.

1.2 Appendix 2 : Variations of the Bland-Altman Plot

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. As Bland and Altman (1986) point out this may not be the case. Importantly Bland and Altman (1999) makes the following point:

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

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

Due to limitations of the conventional difference plot, a series of alternative formulations for the Bland-Altman approach have been proposed. Referring to the assumption that bias and variability are constant across the range of measurements, Bland and Altman (1999) address the case where there is an increase in variability as the magnitude increases. They remark that it is possible to ignore the issue altogether, but the limits

of agreement would be wider apart than necessary when just lower magnitude measurements are considered. Conversely the limits would be too narrow should only higher magnitude measurements be used.

To address the issue, they propose the logarithmic transformation of the data. The plot is then formulated as the difference of paired log values against their mean. Bland and Altman acknowledge that this is not easy to interpret, and may not be suitable in all cases.

Bland and Altman (1999) offers two variations of the Bland-Altman plot that are intended to overcome potential problems that the conventional plot would be inappropriate for. The first variation is a plot of case-wise differences as percentage of averages, and is appropriate when there is an increase in variability of the differences as the magnitude increases.

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

1.3 Appendix 3 : The Coefficient of Repeatability

Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement by collecting replicated data. Roy (2009b) notes the lack of convenience in such calculations. The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999).

As mentioned previously, Barnhart et al. (2007) emphasize 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). **?** 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.* Bland and Altman (1999) defines the repeatability coefficient 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.

Once the within-item variability for both methods has been estimated, the relevant calculations for the coefficients of repeatability are straightforward. The coefficient is calculated from the within-item variability σ_m^2 as $1.96 \times \sqrt{2} \times \sigma_m = 2.83\sigma_m$. For 95% of subjects, two replicated measurement by the same method will be within this repeatability coefficient.

The coefficient of repeatability may provide the basis for the formulation a formal definition of a 'gold standard'. For example, by determining the ratio of the repeatability coefficient (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.

1.3.1 Carstensen Coefficient of Repeatability

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.

Under the model for linked replicates, 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 $2.8\sigma_m$.

1.4 Appendix 4 : Other Types of Studies

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

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard (Lewis et al., 1991). In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively. Altman and Bland (1983) make clear that their framework is not intended for calibration problems.

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

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

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). In even extreme cases, there must be an assumption of inaccuracy with gold standard systems.

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 tests based upon the angiogram are reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8% (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.

? discusses the importance of gold standards in the context of method comparison studies. Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to Dunn (2002), various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer (i.e. a blood pressure measurement cuff), which is prone to measurement error. Consequently it can be said that a measurement method can be

the 'gold standard', yet have poor repeatability. Dunn (2002) recognizes this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a 'bronze standard'. Again, no formal definition of a bronze standard exists.

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free and that 'it is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The clinician gold standard, the sphygmomanometer, is used as an example thereof. The sphygmomanometer 'leaves considerable room for improvement'. Pizzi (1999) similarly addresses the issue of gold standards, 'well-established gold standard may itself be imprecise or even unreliable'.

In literature gold standards are, perhaps more accurately, can be 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 both in the context of a comparison study and a calibration study.

According to Bland and Altman, one should use the approach previous outlined, even when one of the methods is a gold standard.

1.4.1 Similar Problems

Lewis et al. (1991) categorize method comparison studies into three different types, namely: calibration, comparison and conversion. 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 as criterion methods and test methods respectively.

1. Calibration problems. The purpose is to establish a relationship between meth-

ods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard (Lewis et al., 1991). In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively. Altman and Bland (1983) make clear that their framework 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 for which Bland and Altman's Methodology is intended, and therefore it is the most relevant of the three for this thesis.

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

? discusses the importance of gold Standards in the context of method comparison studies. Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to Dunn (2002), various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer (i.e. a blood pressure measurement cuff), which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability. Dunn (2002) recognizes this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a 'bronze standard'. Again, no formal definition of a bronze standard exists.

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free and that 'it is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The

clinician gold standard, the sphygmomanometer, is used as an example thereof. The sphygmomanometer 'leaves considerable room for improvement'. Pizzi (1999) similarly addresses the issue of gold 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 tests based upon the angiogram are reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8% (ACR, 2008).

In literature gold standards are, perhaps more accurately, can be 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 both in the context of a comparison study and a calibration study.

According to Bland and Altman, one should use the methodology previous outlined, even when one of the methods is a gold standard.

1.5 Appendix 5 : Indices and Graphical Techniques

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

same subject, and is given by:

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

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

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

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

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

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

To illustrate the use of EAD, consider Table 1.5.1. The inter-method bias of 0.03, which is desirably close to zero in the context of agreement. However, an identity plot would indicate very poor agreement, as the points are noticeably distant from the line of equality.

The limits of agreement are $[-9.61, 9.68]$, which is a wide interval for this data. As with the identity plot, this would indicate lack of agreement. As with inter-method bias, an EAD value close to zero is desirable. However, from Table 1.5.1, the EAD can be computed as 3.71. The Bland-Altman plot remains a useful part of the analysis. In Figure 1.5.2, it is clear there is a systematic decrease in differences across the range of measurements.

Barnhart et al. (2007) remarks that a comparison of EAD and MSD, using simulation studies, would be interesting, while further adding that '*It will be of interest*

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

Table 1.5.1: Example data set

to investigate the benefits of these possible new unscaled agreement indices'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. The EAD values for both comparisons are therefore much closer.

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

Table 1.5.2: Agreement indices for Grubbs' data comparisons.

Further to Lin (2000) and Lin et al. (2002), individual agreement between two

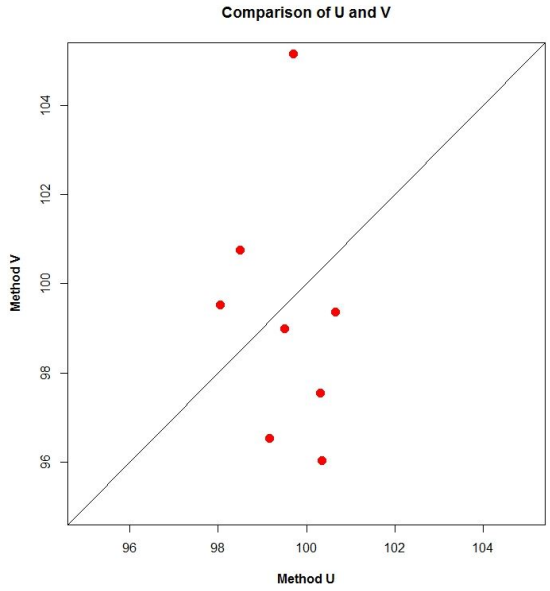

Figure 1.5.1: Identity Plot for example data

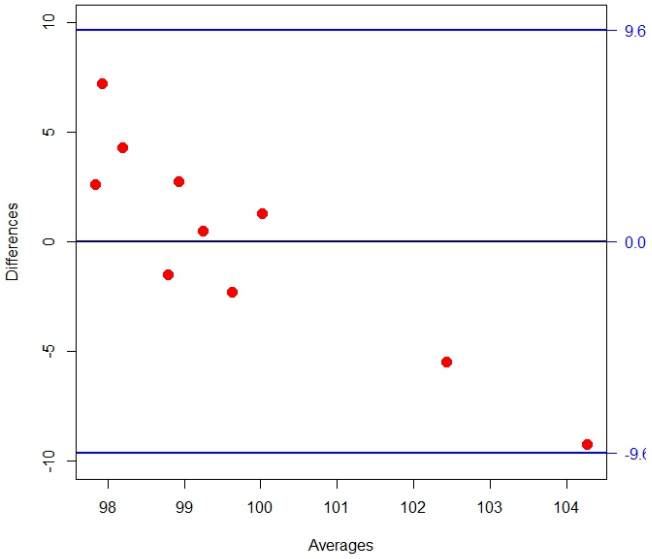

Figure 1.5.2: Bland-Altman Plot for UV comparison

measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If d_0 is predetermined as the maximum acceptable

absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{1.2}$$

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

1.6 Appendix 6 : Measurement Error Models

Dunn (2002) proposes a measurement error model for use in method comparison studies. Consider n pairs of measurements X_i and Y_i for $i = 1, 2, ...n$.

$$X_i = \tau_i + \delta_i \tag{1.3}$$

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

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

$$E(X_i) = \tau_i \tag{1.4}$$

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

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

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

$$z_0 = d = 0 \tag{1.5}$$

$$z_{n+1} = z_n^2 + c \tag{1.6}$$

1.6.1 Thompson 1963: Model Formulation and Formal Testing

Kinsella (1986) formulates a model for un-replicated observations for a method comparison study as a mixed model.

$$\begin{aligned} Y_{ij} = \quad & \mu_j + S_i + \epsilon_{ij} \quad i = 1, 2...n \quad j = 1, 2 \\ & S \sim N(0, \sigma_s^2) \qquad \epsilon_{ij} \sim N(0, \sigma_j^2) \end{aligned} \tag{1.7}$$

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

$$\begin{aligned} \text{var}(Y_{ij}) = \quad & \sigma_s^2 + \sigma_j^2 \\ \text{cov}(Y_{i1}, Y_{i2}) = \quad & \sigma_s^2 \end{aligned} \tag{1.8}$$

The standard error of these variance estimates are:

$$\begin{aligned} \text{var}(\sigma_1^2) = \quad & \frac{2\sigma_1^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \\ \text{var}(\sigma_2^2) = \quad & \frac{2\sigma_2^4}{n-1} + \quad \frac{\sigma_S^2\sigma_1^2 + \sigma_S^2\sigma_2^2 + \sigma_1^2\sigma_2^2}{n-1} \end{aligned} \tag{1.9}$$

Kinsella (1986) 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.

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

$$\begin{aligned} \hat{\sigma_1^2} \quad & = \sum (y_{i1} - \bar{y1})(D_i - \bar{D}) \\ \hat{\sigma_2^2} \quad & = \sum (y_{i2} - \bar{y_2})(D_i - \bar{D}) \end{aligned} \tag{1.10}$$

The value t is the $100(1 - \alpha/2)\%$ upper quantile ofStudent's t distribution with $n - 2$ degrees of freedom (Kinsella, 1986). 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

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.

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.

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.

$$\begin{aligned} |\sigma^2 - C_{xy}K| &\leqslant M(C_xC_y)^{\frac{1}{2}} \\ |\sigma_1^2 - (C_x - C_{xy})K| &\leqslant M(C_x(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \\ |\sigma_2^2 - (C_y - C_{xy})K| &\leqslant M(C_y(C_x + C_y - 2C_{xy}))^{\frac{1}{2}} \end{aligned} \tag{1.11}$$

Estimating the Variance Ratio for Deming Regression

$$\begin{aligned} x_i &= \mu + \beta_0 + \epsilon_{xi} \\ y_i &= \mu + \beta_1 + \epsilon_{yi} \end{aligned}$$

The inter-method bias is the difference of these biases. In order to determine an estimate for the residual variances, one of the method biases must be assumed to be zero, i.e. $\beta_0 = 0$. The inter-method bias is now represented by β_1.

$$
\begin{aligned}
x_i &= \mu + \epsilon_{xi} \\
y_i &= \mu + \beta_1 + \epsilon_{yi}
\end{aligned}
$$

The residuals can be expressed as

$$
\begin{aligned}
\epsilon_{xi} &= x_i - \mu \\
\epsilon_{yi} &= y_i - (\mu + \beta_1)
\end{aligned}
$$

The variance of the residuals are equivalent to the variance of the corresponding observations, $\sigma^2_{\epsilon x} = \sigma^2_x$ and $\sigma^2_{\epsilon y} = \sigma^2_y$.

$$
\lambda = \frac{\sigma^2_{yx}}{\sigma^2_y}. \tag{1.12}
$$

Assuming constant standard deviations, and given duplicate measurements, the analytical standard deviations are given by

$$
\begin{aligned}
SD^2_{ax} &= \frac{1}{2n}\sum (x_{2i} - x_{1i})^2 \\
SD^2_{ay} &= \frac{1}{2n}\sum (y_{2i} - y_{1i})^2
\end{aligned}
$$

Using duplicate measurements, one can estimate the analyticalstandard deviations and compute their ratio. This ratio is then used for computing the slope by the Deming method (Linnet, 1998).

1.6.2 Using LME models to estimate the ratio (BXC)

$$
y_{mi} = \mu + \beta_m + b_i + \epsilon_{mi}
$$

with β_m is a fixed effect for the method m and b_i is a random effect associated with patient i, and ϵ_{mi} as the measurement error. This is a simple single level LME model.

Pinheiro and Bates (1994) provides for the implementation of fitting a model. The variance ratio of the residual variances is immediately determinable from the output.

1.7 Appendix 7: Model II regression

Cornbleet and Cochrane (1979) argue for the use of methods that based on the assumption that both methods are imprecisely measured ,and that yield a fitting that is consistent with both 'X on Y' and 'Y on X' formulations. These methods uses alternatives to the OLS approach to determine the slope and intercept.

They describe three such alternative methods of regression; Deming, Mandel, and Bartlett regression. Collectively the authors refer to these approaches as Model II regression techniques.

The authors make the distinction between model I and model II regression types.

Model II regression is the appropriate type when the predictor variable x is measured with imprecision.

Cornbleet and Cochrane (1979) remark that clinical laboratory measurements usually increase in absolute imprecision when larger values are measured.

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

Model II regression method also calculates a line of best fit for two sets of data. It differs from Model I regression in that it is derived in a way that factors in for error in the x-axis, as well as the y-axis. Cornbleet and Cochrane (1979) refer to it as 'Model II regression'.

1.7.1 Ordinary Least Product Regression

Ludbrook (1997) states that the grouping structure can be straightforward, but there are more complex data sets that have a hierarchical(nested) model.

Observations between groups are independent, but observations within each groups are dependent because they belong to the same subpopulation. Therefore there are two sources of variation: between-group and within-group variance.

1.7.2 Least Products Regression

Used as an alternative to Bland-Altman Analysis, this method is also known as 'Geometric Mean Regression' and 'Reduced Major Axis Regression'. This regression model minimizes the areas of the right triangles formed by the data points' vertical and horizontal deviations from the fitted line and the fitted line.

Model II regression analysis caters for cases in which random error is attached to both dependent and independent variables. Comparing methods of measurement is just such a case (Ludbrook, 1997, 2002).

Least products regression is the Ludbrookes preferred technique for analysing the Model II case. In this, the sum of the products of the vertical and horizontal deviations of the x,y values from the line is minimized.

Least products regression analysis is suitable for calibrating one method against another. It is also a sensitive technique for detecting and distinguishing fixed and proportional bias between methods.

Least-products regression can lead to inflated SEEs and estimates that do not tend to their true values an N approaches infinity (Draper and Smith, 1998)

1.7.3 Classical model for Single Measurements

1.8 Appendix 8: Carstensen's Model

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

This model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2 \ldots$. We use the term *item* to denote an individual, subject or sample, to be measured, being randomly sampled from a population

The classical model is based on measurements y_{mi} by method $m = 1, 2$ on item $i = 1, 2 \ldots$

$$y_{mi} = \alpha_m + \mu_i + e_{mi}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2). \tag{1.13}$$

Here α_m is the fixed effect associated with method m, μ_i is the true value for item i (fixed effect) and e_{mi} is a random effect term for errors.

The random error term for each response is denoted ε_{mir} having $\mathrm{E}(\varepsilon_{mir}) = 0$, $\mathrm{Var}(\varepsilon_{mir}) = \varphi_m^2$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

1.8.1 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 item. The number of repeated measurements for each variable may differ for each item. 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.

$$\begin{pmatrix} y_{Aij} \\ y_{Bij} \end{pmatrix} \sim \mathcal{N}(\mu, \Sigma) \text{ where } \mu = \begin{pmatrix} \mu_A \\ \mu_B \end{pmatrix} \tag{1.14}$$

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

$$\Sigma = \begin{pmatrix} \sigma_A^2 & \sigma_{AB} \\ \sigma_{AB} & \sigma_B^2 \end{pmatrix} \tag{1.15}$$

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 Σ does not depend on a particular time point, and is the same over all time points.

1.9 Appendix 9: Carstensen's Examples

1.9.1 Diabetes (HB1Ac) data set (2008 paper)

Carstensen et al. (2008) describes the sampling method when discussing a motivating example. Diabetes patients attending an outpatient clinic in Denmark have their HbA_{1c} levels routinely measured at every visit. Venous and Capillary blood samples were obtained from all patients appearing at the clinic over two days. Samples were measured on four consecutive days on each machines, hence there are five analysis days.

Carstensen et al. (2008) notes that every machine was calibrated every day to the manufacturers guidelines. Measurements are classified by method, individual and replicate. In this case the replicates are clearly not exchangeable, neither within patients nor simulataneously for all patients.

1.9.2 the Oximetry Data

Carstensen et al. (2008) introduces a second data set; the oximetry study. This study done at the Royal Children's 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.

Carstensen et al. (2008) demonstrates the use of the interaction term when computing the limits of agreement for the 'Oximetry' data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

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);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{1.16}$$

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 2)

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 $\hat{\mathbf{\Lambda}}$. 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.

(N.B. To complement the blood pressure 'J vs S' analysis, the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

Limits of agreement are determined using Roys's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$.

1.9.3 The Fat Data Set

Carstensen et al. (2008) presents a data set 'fat', which is a comparison of measurements of subcutaneous fat by two observers at the Steno Diabetes Center, Copenhagen. Measurements are in millimeters (mm). Each person is measured three times by each observer. The observations are considered to be 'true' replicates. Carstensen et al. (2008) describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}.$$

For the 'Fat' data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

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

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{1.17}$$

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

$$\bar{y_A} - \bar{y_B} \pm 1.96 \times \sqrt{\sigma_A^2 + \sigma_B^2 - 2(\sigma_{AB})}. \tag{1.18}$$

Roy (2009b) has demonstrated a methodology whereby d_A^2 and d_B^2 can be estimated separately. Also covariance terms are present in both G and Λ. Using ARoy2009's methodology, the variance of the differences is

$$\text{var}(y_{iA} - y_{iB}) = d_A^2 + \lambda_B^2 + d_A^2 + \lambda_B^2 - 2(d_{AB} + \lambda_{AB}) \tag{1.19}$$

1.9.4 RV-IV

Remark: what paper is this from? For the the RV-IC comparison, $\hat{D}$ is given by

$$\hat{G} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \tag{1.20}$$

The estimate for the within-subject variance covariance matrix is given by

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \tag{1.21}$$

The estimated overall variance covariance matrix for the the 'RV vs IC' comparison is given by

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \tag{1.22}$$

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.

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 (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

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

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).

Cornbleet, P. J. and D. Cochrane (1979). Regression methods for assessing agreement between two methods of clinical measurement. *Journal of Clinical Chemistry 24*(2), 342–345.

Dewitte, K., C. Fierens, D. Stckl, and L. M. Thienpont (2002). Application of the Bland Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry 48*, 799–801.

Draper, N. R. and H. Smith (1998). Fitting a straight line by least squares. *Applied Regression Analysis, Third Edition*, 15–46.

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

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

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.

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

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

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

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

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

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

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

Ludbrook, J. (2002). Statistical techniques for comparing measurers and methods of measurement: a critcal review. *Clinical and Experimental Pharmacology and Physiology 29*, 527–536.

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

O'Brien, E., J. Petrie, W. Littler, M. de Swiet, P. L. Padfield, D. Altman, M. Bland, A. Coats, and N. Atkins (1990). The British Hypertension Society protocol for the evaluation of blood pressure measuring devices. *Journal of Hypertension 8*, 607–619.

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

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

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.