

# Method Comparison Studies

Kevin O'Brien

February 6, 2017

# Contents

# Chapter 1

## Introduction to Method

## Comparison Studies

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: Difference between Fotobalk and Counter measurements.

## 1.1 Bland-Altman methodology

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

Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which they are unsuitable for comparing two methods of measurement (?). Furthermore they proposed their simple methodology specifically constructed for method comparison studies. They acknowledge the opportunity to apply other valid, but complex, methodologies, but argue that a simple approach is preferable, especially when the results must be ‘explained to non-statisticians’.

Notwithstanding previous remarks about linear regression, the first step recommended, which the authors argue should be mandatory, is construction of a simple scatter plot of the data. The line of equality should also be shown, as it is necessary to give the correct interpretation of how both methods compare. In the case of good agreement, the observations would be distributed closely along the line of equality. A scatter plot of the Grubbs data is shown in Figure 1.1. Visual inspection confirms the previous conclusion that there is an inter-method bias present, i.e. Fotobalk device has a tendency to record a lower velocity.

? notes that scatter plots were very seldom presented in the *Annals of Clinical Biochemistry*. This apparently results from the fact that the ‘Instructions for Authors’

dissuade the use of regression analysis, which conventionally is accompanied by a scatter plot.

### 1.1.1 Bland-Altman plots

In light of shortcomings associated with scatterplots, ? recommend a further analysis of the data. Firstly case-wise differences of measurements of two methods  $d_i = y_{1i} - y_{2i}$  for  $i = 1, 2, \dots, n$  on the same subject should be calculated, and then the average of those measurements ( $a_i = (y_{1i} + y_{2i})/2$  for  $i = 1, 2, \dots, n$ ).

? proposes a scatterplot of the case-wise averages and differences of two methods of measurement. This scatterplot has since become widely known as the Bland-Altman plot. ? express the motivation for this plot thusly:

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

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

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

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

### 1.1.2 Bland-Altman plots for the Grubbs data

In the case of the Grubbs data the inter-method bias is  $-0.61$  metres per second, and is indicated by the dashed line on Figure 1.2. By inspection of the plot, it is also possible to compare the precision of each method. Noticeably the differences tend to increase as the averages increase.

The Bland-Altman plot for comparing the ‘Fotobalk’ and ‘Counter’ methods, which shall henceforth be referred to as the ‘F vs C’ comparison, is depicted in Figure 1.2, using data from Table 1.3. The presence and magnitude of the inter-method bias is indicated by the dashed line.



Round	Fotobalk [F]	Counter [C]	Differences [F-C]	Averages [(F+C)/2]
1	793.8	794.6	-0.8	794.2
2	793.1	793.9	-0.8	793.5
3	792.4	793.2	-0.8	792.8
4	794.0	794.0	0.0	794.0
5	791.4	792.2	-0.8	791.8
6	792.4	793.1	-0.7	792.8
7	791.7	792.4	-0.7	792.0
8	792.3	792.8	-0.5	792.5
9	789.6	790.2	-0.6	789.9
10	794.4	795.0	-0.6	794.7
11	790.9	791.6	-0.7	791.2
12	793.5	793.8	-0.3	793.6

Table 1.2: Fotobalk and Counter methods: differences and averages.

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

Table 1.3: Fotobalk and Terma methods: differences and averages.

In Figure 1.3 Bland-Altman plots for the ‘F vs C’ and ‘F vs T’ comparisons are shown, where ‘F vs T’ refers to the comparison of the ‘Fotobalk’ and ‘Terma’ methods. Usage of the Bland-Altman plot can be demonstrate in the contrast between these comparisons. By inspection, there exists a larger inter-method bias in the ‘F vs C’ comparison than in the ‘F vs T’ comparison. Conversely there appears to be less precision in ‘F vs T’ comparison, as indicated by the greater dispersion of covariates.

### 1.1.3 Prevalence of the Bland-Altman plot

?, which further develops the Bland-Altman methodology, was found to be the sixth most cited paper of all time by the ?. ? describes the rate at which prevalence of the Bland-Altman plot has developed in scientific literature. ? reviewed the use of Bland-Altman plots by examining all articles in the journal ‘Clinical Chemistry’ between 1995 and 2001. This study concluded that use of the BlandAltman plot increased over the years, from 8% in 1995 to 14% in 1996, and 3136% in 2002.

The Bland-Altman Plot has since become expected, and often obligatory, approach for presenting method comparison studies in many scientific journals (?). Furthermore ? recommend its use in papers pertaining to method comparison studies for the journal of the British Hypertension Society.

### 1.1.4 Adverse features

Estimates for inter-method bias and variance of differences are only meaningful if there is uniform inter-bias and variability throughout the range of measurements. Fulfilment of these assumptions can be checked by visual inspection of the plot. The prototype Bland-Altman plots depicted in Figures 1.4, 1.5 and 1.6 are derived from simulated data, for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Figure 1.4 demonstrates how the Bland-Altman plot would indicate increasing variance of differences over the measurement range. Fitted regression lines, for both the upper and lower half of the plot, has been added to indicate the trend. Figure 1.5 is an example of cases where the inter-method bias changes over the measurement range. This is known as proportional bias, and is defined by ? as meaning that ‘one method gives values that are higher (or lower) than those from the other by an amount that

is proportional to the level of the measured variable'. In both Figures 1.4 and 1.5, the assumptions necessary for further analysis using the limits of agreement are violated.

Application of regression techniques to the Bland-Altman plot, and subsequent formal testing for the constant variability of differences is informative. The data set may be divided into two subsets, containing the observations wherein the difference values are less than and greater than the inter-method bias respectively. For both of these fits, hypothesis tests for the respective slopes can be performed. While both tests can be considered separately, multiple comparison procedures, such as the Benjamini-Hochberg (?) test, should be also be used.



Figure 1.1: Bland-Altman plot demonstrating the increase of variance over the range.

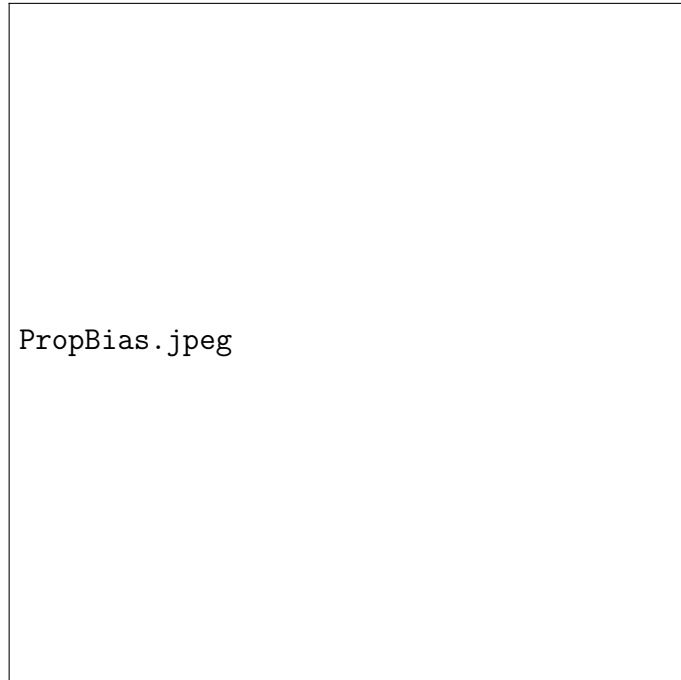


Figure 1.2: Bland-Altman plot indicating the presence of proportional bias.

The Bland-Altman plot also can be used to identify outliers. An outlier is an observation that is conspicuously different from the rest of the data that it arouses suspicion that it occurs due to a mechanism, or conditions, different to that of the rest of the observations. ? do not recommend excluding outliers from analyzes, but remark that recalculation of the inter-method bias estimate, and further calculations based upon that estimate, are useful for assessing the influence of outliers. The authors remark that ‘we usually find that this method of analysis is not too sensitive to one or two large outlying differences’. Figure 1.6 demonstrates how the Bland-Altman plot can be used to visually inspect the presence of potential outliers.

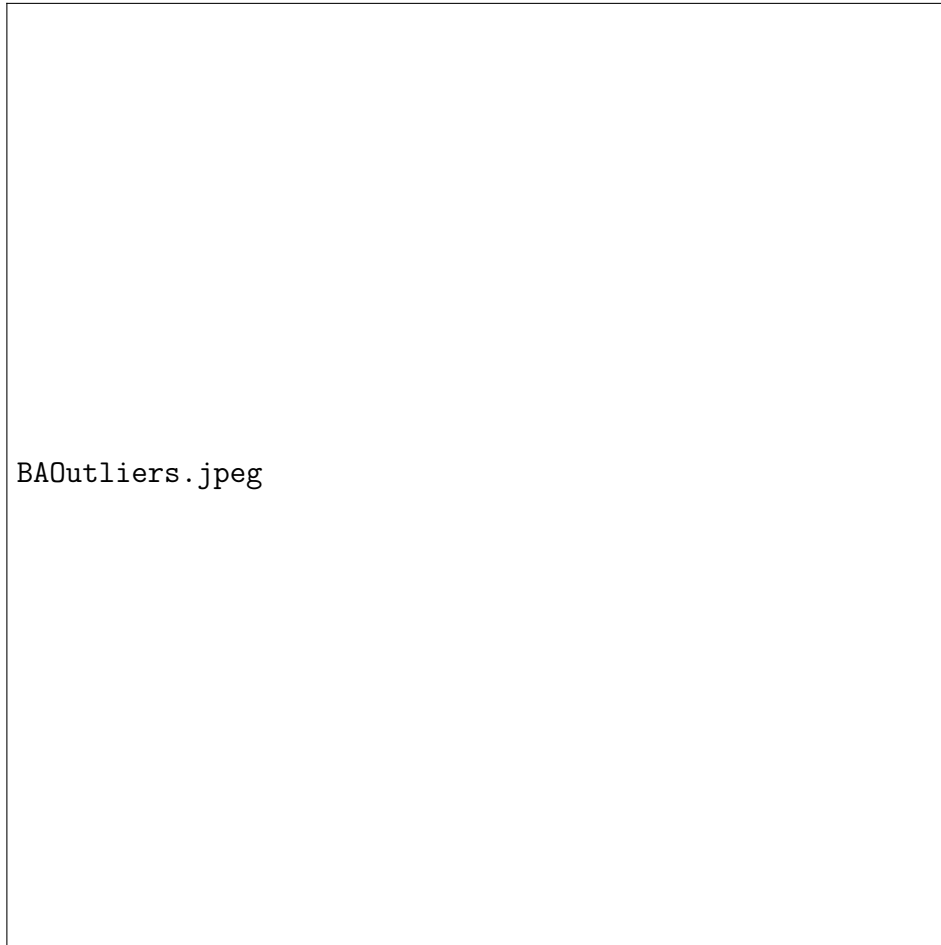


Figure 1.3: Bland-Altman plot indicating the presence of potential outliers.

## 1.2 Limits of Agreement

A third element of the Bland-Altman methodology, an interval known as ‘limits of agreement’ is introduced in ? (sometimes referred to in literature as 95% limits of agreement). Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably. ? refer to this as the ‘equivalence’ of two measurement methods. The specific purpose of the limits of agreement must be established clearly. ? comment that the limits of agreement ‘how far apart measurements by the two methods were likely to be for most individuals’, a definition echoed in their 1999 paper:

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

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

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

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

The limits of agreement methodology assumes a constant level of bias throughout the range of measurements. Importantly the authors recommend prior determination of what would and would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion. However ? highlights inadequacies in the correct application of limits of agreement, resulting in contradictory estimates limits of agreement in various papers.

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

For the Grubbs ‘F vs C’ comparison, these limits of agreement are calculated as -0.132 for the upper bound, and -1.08 for the lower bound. Figure 1.9 shows the resultant Bland-Altman plot, with the limits of agreement shown in dashed lines.

### 1.2.1 Inferences on Bland-Altman estimates

?advise on how to calculate confidence intervals for the inter-method bias and limits of agreement. For the inter-method bias, the confidence interval is simply that of a mean:  $\bar{d} \pm t_{(0.5\alpha, n-1)} S_d / \sqrt{n}$ . The confidence intervals and standard error for the limits of agreement follow from the variance of the limits of agreement, which is shown to be

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

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

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

Consequently the standard errors of both limits can be approximated as 1.71 times the standard error of the differences.

A 95% confidence interval can be determined, by means of the  $t$  distribution with  $n-1$  degrees of freedom. However ? comment that such calculations may be ‘somewhat optimistic’ on account of the associated assumptions not being realized.

### 1.2.2 Formal definition of limits of agreement

? note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as ‘being like a reference interval’.

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the parameters used to determine the Shewhart limits are not based on any sample used for an analysis, but on the process’s historical values, a key difference with Bland-Altman limits of agreement.



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

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

where  $n$  is the number of subjects. Carstensen is careful to consider the effect of the sample size on the interval width, adding that only for 61 or more subjects is there a quantile less than 2.

? offers an alternative description of limits of agreement, this time as tolerance limits. A tolerance interval for a measured quantity is the interval in which a specified fraction of the population's values lie, with a specified level of confidence. ? describes them as a probability interval, and offers a clear description of how they should be used; 'if the absolute limit is less than an acceptable difference  $d_0$ , then the agreement between the two methods is deemed satisfactory'.

The prevalence of contradictory definitions of what limits of agreement strictly are will inevitably attenuate the poor standard of reporting using limits of agreement, as mentioned by ?.

### 1.2.3 Alternative agreement indices

As an alternative to limits of agreement, ? 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

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

? 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 properly determinable, thus creating a drawback to this methodology.

? proposes both the use of the square root of the MSD or the expected absolute difference (EAD) as an alternative agreement indices. Both of these indices can be interpreted intuitively, being denominated in the same units of measurements as the original measurements. Also they can be compare to the maximum acceptable absolute difference between two methods of measurement  $d_0$ .

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

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.

? remarks that a comparison of EAD and MSD, using simulation studies, would be interesting, while further adding that ‘It will be of interest to investigate the benefits of these possible new unscaled agreement indices’. For the Grubbs’ ‘F vs C’ and ‘F vs T’ comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for ‘F vs C’ and ‘F vs T’ comparisons were depicted previously on Figure 1.3. While the inter-method bias for the ‘F vs T’ comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. Hence the EAD values for both comparisons are much closer.

	F vs C	F vs T
Inter-method bias	-0.61	0.12 3
Difference variances	0.06	0.22
Limits of agreement	(-1.08, -0.13)	(-0.81,1.04)
EAD	0.61	0.35

Table 1.4: Agreement indices for Grubbs’ data comparisons.

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

$$CP = P(|x_i - y_i| \leq d_0) \quad (1.1)$$

If  $\pi_0$  is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is  $\pi_0$  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.

### 1.3 Variations of the Bland-Altman Plot

Referring to the assumption that bias and variability are constant across the range of measurements, ? 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 wider apart than necessary when just lower magni-

tude 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.

? 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. This will remove the need for *log* transformation. This approach is useful when there is an increase in variability of the differences as the magnitude of the measurement increases. ? proposed such a ratio plot, independently of Bland and Altman. ? commented on the reception of this article by saying ‘Strange to say, this report has been overlooked’.

### 1.3.1 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 method, these measurements are known as ‘replicate measurements’. ? recommends the use of replicate measurements, but acknowledges the additional computational complexity.

? 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 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. ? propose a correction for this.

? 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. ? 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.

## 1.4 Formal Models and Tests

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

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

The true value of the measurement is represented by  $\mu$  while the fixed effect due to method  $j$  is  $\beta_j$ . For simplicity these terms can be combined into single terms;  $\mu_1 = \mu + \beta_1$  and  $\mu_2 = \mu + \beta_2$ . The inter-method bias is the difference of the two fixed effect terms,  $\beta_1 - \beta_2$ . Each of the  $i$  individuals are assumed to give rise to random error, represented by  $u_i$ . This random effects terms is assumed to have mean zero and be normally distributed with variance  $\sigma^2$ . There is assumed to be an attendant error for each measurement on each individual, denoted  $\epsilon_{ij}$ . 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  $\sigma_j^2$ . The set of observations  $(x_i, y_i)$  by methods  $X$  and  $Y$  are assumed to follow the bivariate normal distribution with expected values  $E(x_i) = \mu_i$  and  $E(y_i) = \mu_i$  respectively. The variance covariance of the observations  $\Sigma$  is given by

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

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

? demonstrates the estimation of the variance terms and relative precisions relevant to a method comparison study, with attendant confidence intervals for both. The measurement model introduced by ?? provides a formal procedure for estimate the

variances  $\sigma^2, \sigma_1^2$  and  $\sigma_2^2$  devices. ? offers estimates, commonly known as Grubbs estimators, for the various variance components. These estimates are maximum likelihood estimates, a statistical concept that shall be revisited in due course.

$$\begin{aligned}\hat{\sigma}^2 &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} = S_{xy} \\ \hat{\sigma}_1^2 &= \sum \frac{(x_i - \bar{x})^2}{n-1} = S^2_x - S_{xy} \\ \hat{\sigma}_2^2 &= \sum \frac{(y_i - \bar{y})^2}{n-1} = S^2_y - S_{xy}\end{aligned}$$

? defines  $\Delta_j$  to be a measure of the relative precision of the measurement methods, with  $\Delta_j = \sigma^2/\sigma_j^2$ . Thompson also demonstrates how to make statistical inferences about  $\Delta_j$ . 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  $\Delta_1$  are

$$\begin{aligned}\Delta_1 &> \frac{C_{xy} - t(\frac{|A|}{n-2})^{\frac{1}{2}}}{C_x - C_{xy} + t(\frac{|A|}{n-2})^{\frac{1}{2}}} \\ \Delta_1 &> \frac{C_{xy} + t(\frac{|A|}{n-2})^{\frac{1}{2}}}{C_x - C_{xy} - t(\frac{|A|}{n-1})^{\frac{1}{2}}}\end{aligned}\tag{1.2}$$

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  $\Delta_2$  are found by substituting  $C_y$  for  $C_x$  in (1.3). Negative lower limits are replaced by the value 0.

The case-wise differences and means are calculated as  $d_i = x_i - y_i$  and  $a_i = (x_i + y_i)/2$

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

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \quad (1.3)$$

### 1.4.1 Morgan Pitman Testing

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

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

$$\rho(a, d) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_s^2 + \sigma_1^2 + \sigma_2^2)}} \quad (1.4)$$

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. ? describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of  $Y_{i1}$  on  $Y_{i2}$ , a result that can be derived using straightforward algebra.

### 1.4.2 Paired sample $t$ test

? 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,

$$t^* = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \quad (1.5)$$

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.

$$t^* = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (1.6)$$

### 1.4.3 Bland-Altman correlation test

The approach proposed by ? is a formal test on the Pearson correlation coefficient of case-wise differences and means ( $\rho_{AD}$ ). According to the authors, this test is equivalent to the ‘Pitman Morgan Test’. For the Grubbs data, the correlation coefficient estimate ( $r_{AD}$ ) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers ‘ $r$  to  $z$ ’ transformation (?). The null hypothesis ( $\rho_{AD} = 0$ ) fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has no been no further mention of this particular test in ?, although ? refers to Spearman’s rank correlation coefficient. ? comments ‘we do not see a place for methods of analysis based on hypothesis testing’. ? also states that consider structural equation models to be inappropriate.

### 1.4.4 Identifiability

? highlights an important issue regarding using models such as these, 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 literature the variance ratio  $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$

must often be assumed to be equal to 1 (?).? considers methodologies 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 contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

? 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 ( $\mu$ ) and variance ( $\sigma^2$ ) 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 (?) 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). The critical value is chosen for  $\alpha\%$  significance with those same degrees of freedom. ? 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:

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

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

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

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.

## 1.5 Regression Methods

Conventional regression models are estimated using the ordinary least squares (OLS) technique, and are referred to as ‘Model I regression’ (??). 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 (??), 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 (??). 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 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.

### 1.5.1 Deming Regression

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

Deming regression is a regression fitting approach that assumes error in both variables. Deming regression is recommended by ? 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 angle specified by the ratio  $\lambda$  of the residual variance of both variables. When  $\lambda$  is one, the angle is 45 degrees. In ordinary linear regression, the distances are minimized in the vertical directions (?). In cases involving only single measurements by each method,  $\lambda$  may be unknown and is therefore assumed a value of one. While this will bias the estimates, it is 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  $\lambda$ , ? derived the following estimate that would later be used for the Deming regression slope parameter. The intercept estimate  $\alpha$  is simply estimated in the same way as in conventional linear regression, by using the identity  $\bar{Y} - \hat{\beta}\bar{X}$ ;

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + [(S_{yy} - \lambda S_{xx})^2 + 4\lambda S_{xy}^2]^{1/2}}{2S_{xy}} \quad (1.7)$$

, with  $\lambda$  as the variance ratio. As stated previously  $\lambda$  is often unknown, and therefore must be assumed to equal one. ? states that Deming regression is acceptable only when the precision ratio ( $\lambda$ , in their paper as  $\eta$ ) is correctly specified, but in practice this is often not the case, with the  $\lambda$  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 ?. This data set features in the discussion of method comparison studies in ?, 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 1.6: Transmitral volumetric flow(MF) and left ventricular stroke volume (SV) in 21 patients. (Zhang et al 1986)

? states that Deming's regression is acceptable only when the precision ratio ( $\lambda$ , in their paper as  $\eta$ ) is correctly specified, but in practice this is often not the case, with the  $\lambda$  being underestimated.



ZhangDeming.jpeg

Figure 1.4: Deming Regression For Zhang's Data

## 1.6 Other Types of Studies

? 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' (?). 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 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 (?). (In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively.) ? 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 specifically 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. ? deals specifically with this issue. In the context of this study, it is the least relevant of the three.

?, p.47 cautions that 'gold standards' should not be assumed to be error free. 'It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The clinician gold standard, the sphygmomanometer, is used as an example thereof. The sphygmomanometer 'leaves considerable room for improvement' (?). ? 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 (?).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement



with the well-being of the patient. This will inevitably lead to the measurement error as described by ?. The magnetic resonance angiogram, used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. This is reported as sensitivity of 95% and a specificity of 92% (?).

In literature they are, perhaps more accurately, referred to as ‘fuzzy gold standards’ (?). 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.