## 0.1 Carstensen's Model

Bendix Carstensen et al. proposed the use of LME models to allow for a more statistically rigourous approach to computing Limits of Agreement. The respective papers also discuss several shortcoming for techniques for dealing with replicate measurements, as proposed by Bland-Altman 1999.

Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement $y_{mi}$ by method $m$ on individual $i$ is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{1}$$

Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

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

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

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{3}$$

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

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

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{4}$$

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{5}$$

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

Let $y_{mir}$ denote the $r$th replicate measurement on the $i$th item by the $m$th method, where $m = 1, 2$ ; $i = 1, \ldots, N$; and $r = 1, \ldots, n_i$. When the design is balanced and there is no ambiguity we can set $n_i = n$. The LME model underpinning Roy's approach can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \tag{6}$$

Here $\beta_0$ and $\beta_m$ are fixed-effect terms representing, respectively, a model intercept and an overall effect for method $m$. The model can be reparameterized by gathering the $\beta$ terms together into (fixed effect) intercept terms $\alpha_m = \beta_0 + \beta_m$. The $b_{1i}$ and $b_{2i}$ terms are correlated random effect parameters having $\mathrm{E}(b_{mi}) = 0$ with $\mathrm{Var}(b_{mi}) = g_m^2$ and $\mathrm{Cov}(b_{1i}, b_{2i}) = g_{12}$. The random error term for each response is denoted $\epsilon_{mir}$ having $\mathrm{E}(\epsilon_{mir}) = 0$, $\mathrm{Var}(\epsilon_{mir}) = \sigma_m^2$, $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir}) = \sigma_{12}$, $\mathrm{Cov}(\epsilon_{mir}, \epsilon_{mir'}) = 0$ and $\mathrm{Cov}(\epsilon_{1ir}, \epsilon_{2ir'}) = 0$. Additionally these parameter are assumed to have Gaussian distribution. Two methods of measurement are in complete agreement if the null hypotheses $\mathrm{H}_1\colon \alpha_1 = \alpha_2$ and $\mathrm{H}_2\colon \sigma_1^2 = \sigma_2^2$ and $\mathrm{H}_3\colon g_1^2 = g_2^2$ hold simultaneously. Roy (2009) uses a Bonferroni correction to control the familywise error rate for tests of $\{\mathrm{H}_1, \mathrm{H}_2, \mathrm{H}_3\}$ and account for difficulties arising due to multiple testing. Additionally, Roy combines $\mathrm{H}_2$ and $\mathrm{H}_3$ into a single testable hypothesis $\mathrm{H}_4\colon \omega_1^2 = \omega_2^2$, where $\omega_m^2 = \sigma_m^2 + g_m^2$ represent the overall variability of method $m$.

Carstensen et al. (2008) also use a LME model for the purpose of comparing two methods of measurement where replicate measurements are available on each item. Their interest lies in generalizing the popular limits-of-agreement (LOA) methodology advocated by Bland and Altman (1986) to take proper cognizance of the replicate measurements. Carstensen et al. (2008) demonstrate statistical flaws with two approaches proposed by Bland and Altman (1999) for the purpose of calculating the variance of the inter-method bias when replicate measurements are available. Instead, they recommend a fitted mixed effects model to obtain appropriate estimates for the variance of the inter-method bias. As their interest mainly lies in extending the Bland-Altman methodology, other formal tests are not considered.

Carstensen et al. (2008) develop their model from a standard two-way analysis of variance model, reformulated for the case of replicate measurements, with random effects terms specified as appropriate. Their model can be written as

$$y_{mir} = \alpha_m + \mu_i + a_{ir} + c_{mi} + \varepsilon_{mir}. \tag{7}$$

The fixed effects $\alpha_m$ and $\mu_i$ represent the intercept for method $m$ and the 'true value' for item $i$ respectively. The random-effect terms comprise an item-by-replicate interaction term $a_{ir} \sim \mathcal{N}(0, \varsigma^2)$, a method-by-item interaction term $c_{mi} \sim \mathcal{N}(0, \tau_m^2)$, and model error terms $\varepsilon_{mir} \sim \mathcal{N}(0, \varphi_m^2)$. All random-effect terms are assumed to be independent. For the case when replicate measurements are assumed to be exchangeable for item $i$, $a_{ir}$ can be removed. The model expressed in (2) describes measurements by $m$ methods, where $m = \{1, 2, 3 \ldots\}$. Based on the model expressed in (2), Carstensen et al. (2008) compute the limits of agreement as

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

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

There is a substantial difference in the number of fixed parameters used by the respective models; the model in (6) requires two fixed effect parameters, i.e. the means of the two methods, for any number of items $N$, whereas the model in (7) requires $N + 2$ fixed effects.

Allocating fixed effects to each item $i$ by (7) accords with earlier work on comparing methods of measurement, such as Grubbs (1948). However allocation of fixed effects in ANOVA models suggests that the group of items is itself of particular interest, rather than as a representative sample used of the overall population. However this approach seems contrary to the purpose of LOAs as a prediction interval for a population of items. Conversely, Roy (2009) uses a more intuitive approach, treating the observations as a random sample population, and allocating random effects accordingly.

## 0.2 The Research of Carstensen et al

### 0.2.1 Bendix Carstensen's data sets

Carstensen et al. (2008)describes the sampling method when discussing of 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 notes that every machine was calibrated every day to the manufacturers guidelines.

### 0.2.2 Limits of agreement for Carstensen's data

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 $(-0.23, 0.32)$

Carstensen 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 $(-12.18, 17.12)$.

### 0.2.3 Using LME Models for Method Comparison

**?** formulates an LME model, both in the absence and the presence of an interaction term.**?** uses both to demonstrate the importance of using an interaction term. Failure to take the replication structure into account results in over-estimation of the limits of agreement. For the Carstensen estimates below, an interaction term was included when computed.

### 0.2.4 Computing LoAs with LMEs

Carstensen presents a model where the variation between items for method $m$ is captured by $\sigma_m$ and the within item variation by $\tau_m$.

Further to his model, Carstensen computes the limits of agreement as

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

## 0.3 Carstensen's Model

Carstensen et al. (2008) proposes an approach for comparing two or more methods of measurement based on linear mixed effects models. This approach extends the well established Bland-Altman methodology for the case of replicate measurements on each item. Carstensen considers the matter of computing an appropriate estimate for the standard deviation of case-wise differences, so as to determine the limits of agreement. As the interest lies in extending the Bland-Altman methodology, other formal tests are not described.

Carstensen (2004) proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The following model ( in the authors own notation) is formulated as follows, where $y_{mir}$ is the $r$th replicate measurement on subject $i$ with method $m$.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{8}$$

The intercept term $\alpha$ and the $\beta_m \mu_i$ term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value $\mu_i$. $c_{mi}$ is a interaction term to account for replicate, and $e_{mir}$ is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that mobservations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).

Using Carstensen's notation, a measurement $y_{mi}$ by method $m$ on individual $i$ the measurement $y_{mir}$ is the $r$th replicate measurement on the $i$th item by the $m$th method, where $m = 1, 2, \ldots, M$ $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + a_{ir} + \epsilon_{mir}, \qquad c_{mi} \sim \mathcal{N}(0, \tau_m^2), a_{ir} \sim \mathcal{N}(0, \varsigma^2), \varepsilon_{mi} \sim \mathcal{N}(0, \varphi_m^2). \tag{9}$$

Here the terms $\alpha_m$ and $\mu_i$ represent the fixed effect for method $m$ and a true value for item $i$ respectively. The random effect terms comprise an interaction term $c_{mi}$ and the residuals $\varepsilon_{mir}$. The $c_{mi}$ term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau_m^2$.

The random error term for each response is denoted $\varepsilon_{mir}$ 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.

When only two methods are to be compared, separate estimates of $\tau_m^2$ can not be obtained. Instead the average value $\tau^2$ is obtained and used.

Carstensen's approach is that of a standard two-way mixed effects ANOVA with replicate measurements. With regards to the specification of the variance terms, Carstensen remarks that using his approach is common, remarking that *The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods* (**?**).

In contrast to Roy's model, Carstensen's model requires that commonly used assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off-diagonal elements are also zero. Also, implementation requires that the between-item variances are estimated as the same value: $\tau_1^2 = \tau_2^2 = \tau^2$. Also, implementation requires that the between-item variances are estimated as the same value: $g_1^2 = g_2^2 = g^2$. As a consequence, Carstensen's method does not allow for a formal test of the between-item variability.

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

The presence of the true value term $\mu_i$ gives rise to an important difference between Carstensen's and Roy's models. The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. In other words, Roy considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

## 0.4 Carstensen's Mixed Models

Carstensen (2004) proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The following model ( in the authors own notation) is formulated as follows, where $y_{mir}$ is the $r$th replicate measurement on subject $i$ with method $m$.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{10}$$

The intercept term $\alpha$ and the $\beta_m \mu_i$ term follow from Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value $\mu_i$. $c_{mi}$ is a interaction term to account for replicate, and $e_{mir}$ is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis. There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. (Exchangeability means that future samples from a population behaves like earlier samples).

Carstensen (2004) uses the above formula to predict observations for a specific individual $i$ by method $m$;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m} \mu_i + c_{mi} \tag{11}$$

. Under the assumption that the $\mu$s are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates. The assumption of exchangeability can be unrealistic in certain situations. Carstensen (2004) provides an amended formulation which includes an extra interaction term $(d_{mr} \sim N(0, \omega_m^2)$to account for this.

Carstensen et al. (2008) sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{12}$$

Carstensen et al. (2008) proposes a methodology to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman methodology in this regard. It is not possible to estimate the interaction variance components $\tau_1^2$ and $\tau_2^2$ separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{13}$$

### 0.4.1 Bendix Carstensen's data sets

**?**describes the sampling method when discussing of 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 notes that every machine was calibrated every day to the manufacturers guidelines.

### 0.4.2 Carstensen Methods

Components

```
Section 5.3 Models for replicate measurements
Section 5 Replicate measurements.

Carstensen page 56
%-----------------------------------------------------------------%
air extra random effect that does not depend on method.
It is treated as an extension of i.
The variance of air represents the variation between replication condition (common for all methods), wi
```

$$ymir = m + i + cmi + emir$$

$$cmi = N(0, m2)$$

$$emir = N(0, m2)$$

```
Carstensen page 58

var(y10-y20) =12+22+12+22

1-2222+12+22

Roy further to Carstensen

ymir=m+i+cmi+emir
```

Section 7 A general model for method comparisons.
Carstensen discusses the model and its use as if all parameter estimates are available.
In this model, intermethod bias is assumed to be constant at all measurement levels.
i : True value for item i
The parameter i can be thought of as the underlying, but unobtainable, true measurement for item i.

m: Fixed effect for method m

**Carstensen et al - Mixed Models**

Carstensen et al [4] also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value.

The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that *inter-method bias* is the only difference between the two methods. A measurement $y_{mi}$ by method $m$ on individual $i$ is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{14}$$

Carstensen et al [5] sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{15}$$

Carstensen *et al* Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement $y_{mi}$ by method $m$ on individual $i$ is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{16}$$

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \tag{17}$$

The following model (in the authors own notation) is formulated as follows, where $y_{mir}$ is the $r$th replicate measurement on subject $i$ with method $m$.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{18}$$

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \tag{19}$$

$$e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)$$

The intercept term $\alpha$ and the $\beta_m \mu_i$ term follow from *Dunn* Dunn (2002), expressing constant and proportional bias respectively , in the presence of a real value $\mu_i$. $c_{mi}$ is a interaction term to account for

replicate, and $e_{mir}$ is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

The above formulation doesn't require the data set to be balanced. However, it does require a sufficient large number of replicates and measurements to overcome the problem of identifiability. The import of which is that more than two methods of measurement may be required to carry out the analysis.

There is also the assumptions that observations of measurements by particular methods are exchangeable within subjects. ***Exchangeability*** means that future samples from a population behaves like earlier samples).

*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.*

### 0.4.3 Using LME models to create Prediction Intervals

Carstensen et al. (2008) proposes a methodology to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman methodology in this regard. It is not possible to estimate the interaction variance components $\tau_1^2$ and $\tau_2^2$ separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{20}$$

### 0.4.4 Computation

Modern software packages can be used to fit models accordingly. The best linear unbiased predictor (BLUP) for a specific subject $i$ measured with method $m$ has the form $BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi}$, under the assumption that the $\mu$s are the true item values.

**?** generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This methodology proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements.

Maximum likelihood estimation is used to estimate the parameters. The REML estimation is not considered since it does not lead to a joint distribution of the estimates of fixed effects and random effects parameters, upon which the assessment of agreement is based.

### 0.4.5 Carstensen's Mixed Models

Carstensen *et al*[4] presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland-Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

Carstensen *et al*[4] proposes linear mixed effects models for deriving conversion calculations similar to Deming's regression, and for estimating variance components for measurements by different methods. The following model ( in the authors own notation) is formulated as follows, where $y_{mir}$ is the $r$th replicate measurement on subject $i$ with method $m$.

$$y_{mir} = \alpha_m + \beta_m\mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{21}$$

The intercept term $\alpha$ and the $\beta_m \mu_i$ term follow from Dunn[7], expressing constant and proportional bias respectively , in the presence of a real value $\mu_i$. $c_{mi}$ is a interaction term to account for replicate, and $e_{mir}$ is the residual associated with each observation. Since variances are specific to each method, this model can be fitted separately for each method.

This model includes a method by item interaction term.

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen.

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \tag{22}$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$.

For the replicate case, an interaction term $c$ is added to the model, with an associated variance component.

### 0.4.6 Computing LoAs from LME models

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.

Carstensen presents a model where the variation between items for method $m$ is captured by $\sigma_m$ and the within item variation by $\tau_m$.

Further to his model, Carstensen computes the limits of agreement as

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

The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.

Carstensen et al. (2008) formulates an LME model, both in the absence and the presence of an interaction term.Carstensen et al. (2008) uses both to demonstrate the importance of using an interaction term. Failure to take the replication structure into account results in over-estimation of the limits of agreement. For the Carstensen estimates below, an interaction term was included when computed.

Carstensen et al. (2008) uses an approach based on linear mixed effects (LME) models for the purpose of computing the limits of agreement for two methods of measurement, where replicate measurements are taken on items. As the emphasis of this methodology lies on the inter-method bias and the limits of agreement, the two key elements of the Bland-Altman methodology, other formal tests are not described.

Using Carstensen's notation, a measurement $y_{mi}$ by method $m$ on individual $i$ the measurement $y_{mir}$ is the $r$th replicate measurement on the $i$th item by the $m$th method, where $m = 1, 2$, $i = 1, \ldots, N$, and $r = 1, \ldots, n_i$ is formulated as follows;

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + \epsilon_{mir}, \qquad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \tag{23}$$

Here the terms $\alpha_m$ and $\mu_i$ represent the fixed effect for method $m$ and a true value for item $i$ respectively. The random effect terms comprise an interaction term $c_{mi}$ and the residuals $\epsilon_{mir}$. The $c_{mi}$ term represent random effect parameters corresponding to the two methods, having $\mathrm{E}(c_{mi}) = 0$ with $\mathrm{Var}(c_{mi}) = \tau_m^2$. Carstensen specifies the variance of the interaction terms as being univariate normally distributed. As such, $\mathrm{Cov}(c_{mi}, c_{m'i}) = 0$. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

With regards to specifying the variance terms, Carstensen remarks that using his approach is common, remarking that *The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods* (**?**).

The presence of the true value term $\mu_i$ gives rise to an important difference between Carstensen's and Roy's models. The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. In other words, Roy considers the group of items being measured as a sample taken from a population. Therefore a distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

## 0.5 Carstensen's Limits of agreement

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required. The method of computation is the same as Roy's model, but with the covariance estimates set to zero.

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using Carstensen's model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ. In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen, when covariance is present.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

## 0.6 Repeated Measurements

In cases where there are repeated measurements by each of the two methods on the same subjects , Bland Altman suggest calculating the mean for each method on each subject and use these pairs of means to compare the two methods. The estimate of bias will be unaffected using this approach, but the estimate of the standard deviation of the differences will be too small, because of the reduction of the effect of repeated measurement error. Bland Altman propose a correction for this. Carstensen attends to this issue also, adding that another approach would be to treat each repeated measurement separately.

## 0.7 Carstensen 2004 's Mixed Models

Carstensen (2004) uses the above formula to predict observations for a specific individual $i$ by method $m$;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \tag{24}$$

. Under the assumption that the $\mu$s are the true item values, this would be sufficient to estimate parameters. When that assumption doesn't hold, regression techniques (known as updating techniques) can be used additionally to determine the estimates. The assumption of exchangeability can be unrealistic in certain situations. Carstensen (2004) provides an amended formulation which includes an extra interaction term ($d_{mr}d_{mr} \sim N(0, \omega_m^2)$to account for this.

Carstensen et al. (2008) sets out a methodology of computing the limits of agreement based upon variance component estimates derived using linear mixed effects models. Measures of repeatability, a characteristic of individual methods of measurements, are also derived using this method.

## 0.8 LME

Consistent with the conventions of mixed models, **?** formulates the measurement $y_{ij}$from method $i$ on individual $j$ as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{25}$$

The design matrix $P_{ij}$ , with its associated column vector $\theta$, specifies the fixed effects common to both methods. The fixed effect specific to the $j$th method is articulated by the design matrix $W_{ij}$ and its column vector $v_i$. The random effects common to both methods is specified in the design matrix $X_{ij}$, with vector $b_j$ whereas the random effects specific to the $i$th subject by the $j$th method is expressed by $Z_{ij}$, and vector $u_j$. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{26}$$

These vectors are assumed to be independent for different $i$s, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{27}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate $x$ is the key step in how this methodology assesses agreement.

# Chapter 1

# Limits of Agreement

## 1.1 Using LME models to create Prediction Intervals

Carstensen (2004) also advocates the use of linear mixed models in the study of method comparisons. The model is constructed to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods. A measurement $y_{mi}$ by method $m$ on individual $i$ is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \tag{1.1}$$

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

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \qquad (e_{mi} \sim N(0, \sigma_m^2), c_{mi} \sim N(0, \tau_m^2)) \tag{1.2}$$

Carstensen et al. (2008) proposes a methodology to calculate prediction intervals in the presence of replicate measurements, overcoming problems associated with Bland-Altman methodology in this regard. It is not possible to estimate the interaction variance components $\tau_1^2$ and $\tau_2^2$ separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{1.3}$$

### 1.1.1 Computation

Modern software packages can be used to fit models accordingly. The best linear unbiased predictor (BLUP) for a specific subject $i$ measured with method $m$ has the form $BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi}$, under the assumption that the $\mu$s are the true item values.

### 1.1.2 Limits Of Agreement

Bland and Altman proposed a pair of Limits of agreement. These limits are intended to demonstrate the range in which 95% of the sample data should lie. The Limits of agreement centre on the average difference

line and are 1.96 times the standard deviation above and below the average difference line.
How this relates the overall population is unclear. It seems that it depends on an expert to decide whether or not the range of differences is acceptable. In a study A Bland-Altman plots compare two assay methods. It plots the difference between the two measurements on the Y axis, and the average of the two measurements on the X axis

### 1.1.3 Appropriate Use of Limits of Agreement

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

## 1.2 Intervals

### 1.2.1 Precision of Limits of Agreement

The limits of agreement are estimates derived from the sample studied, and will differ from values relevant to the whole population. A different sample would give different limits of agreement. Bland and Altman (1986) advance a formulation for confidence intervals of the inter-method bias and the limits of agreement. These calculations employ quantiles of the 't' distribution with $n - 1$ degrees of freedom.

### 1.2.2 Purpose of Limits of Agreement

It must be established clearly the specific purpose of the limits of agreement. Bland and Altman (1995) 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(Bland and Altman, 1999).

**?** offers an alternative, more specific, definition of the limits of agreement *"a prediction interval for the difference between future measurements with the two methods on a new individual."* Luiz et al. (2003) describes them as tolerance limits.

Importantly they have the same construction as Shewhart Control limits.

### 1.2.3 Tolerance Intervals

Shewhart formaulted a tolerance limits to determine acceptable ranges for a production process.

## 1.3 Limits of Agreement

Several problems have been highlighted regarding Limits of Agreement. One is the somewhat arbitrary manner in which they are constructed. While in essence a confidence interval, they are not constructed a such. They are designed for future values.
The formulation is also heavily influenced by outliers. An Example in Altman and Bland (1983) demonstrates the effect of recalculating without a particular outlier. Refering to the VCF data set in the same paper, there is more than one outlier.

## 1.4 Limits of Agreement

A third element of the Bland-Altman methodology, an interval known as 'limits of agreement' is introduced in Bland and Altman (1986), (sometimes referred to in literature as 95% limits of agreement). Limits of agreement are used to assess whether the two methods of measurement can be used interchangeably. Bland and Altman (1986) refer to this as the 'equivalence' of two measurement methods. It must be established clearly the specific purpose of the limits of agreement. Bland and Altman (1995) comment that the limits of agreement "how far apart measurements by the two methods were likely to be for most individuals", a definition echoed in their 1999 paper:

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

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

$$LoA = \bar{d} \pm 1.96 S(d) \tag{1.4}$$

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. (However, in some literature, 2 standard deviations are used instead for simplicity.) 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.

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. Bland and Altman advises on how to calculate of confidence intervals for the inter-method bias and the limits of agreement. Importantly the authors recommend prior determination of what would and would constitute acceptable agreement, and that sample sizes should be predetermined to give an accurate conclusion.

> 'How far apart measurements can be without causing difficulties will be a question of judgment. Ideally, it should be defined in advance to help in the interpretation of the method comparison and to choose the sample size.'(Bland and Altman, 1986)

Bland and Altman (1999) note the similarity of limits of agreement to confidence intervals, but are clear that they are not the same thing. Interestingly, they describe the limits as "being like a reference interval."

Limits of agreement have very similar construction to Shewhart control limits. The Shewhart chart is a well known graphical methodology used in statistical process control. Consequently there is potential for misinterpreting the limits of agreement as if equivalent to Shewhart control limits. Importantly the

parameters used to determine the limits, the mean and standard deviation, 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.

Carstensen et al. (2008) regards the limits of agreement as a prediction interval for the difference between future measurements with the two methods on a new individual, but states that it does not fit the formal definition of a prediction interval, since the definition does not consider the errors in estimation of the parameters. Prediction intervals, which are often used in regression analysis, are estimates of an interval in which future observations will fall, with a certain probability, given what has already been observed. Carstensen et al. (2008) offers an alternative formulation, a 95% prediction interval for the difference

$$\bar{d} \pm t_{(0.975,n-1)} S_d \sqrt{1 + \frac{1}{n}} \tag{1.5}$$

where $n$ is the number of subjects. Only for 61 or more subjects is there a quantile less than 2.

Luiz et al. (2003) describes limits of agreement 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.

### 1.4.1 Confidence Intervals and Standard Error

Bland and Altman (1999) argue that it is possible to estimate confidence intervals and standard error, if it is assumed that the differences approximately follow a normal distribution,

$$\mathrm{Var}(LoA) = (\frac{1}{n} + \frac{1.96^2}{2(n-1)}) S_d^2. \tag{1.6}$$

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

$$\mathrm{Var}(LoA) \approx 1.71^2 \frac{S_d^2}{n}. \tag{1.7}$$

Consequently the standard errors of both limits can be approximated as $1.71 S.E.(\bar{d})$

A 95% confidence interval can be determined, by means of the $t$ distribution with $n-1$ degrees of freedom. Bland and Altman (1999) comment that such calculations may be 'somewhat optimistic' on account of the associated assumptions not being realized.

# Bibliography


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.

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.

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).

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.

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.

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

Roy, A. (2009). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.