Method Comparison Studies with R

Kevin O'Brien

Method Comparison Studies

- The problem of assessing the **agreement** between two or more methods of measurement is ubiquitous in scientific research, particularly with clinical sciences, and is commonly referred to as a ‘method comparison study’.
  - “Do two methods of measurement agree statistically?".
  - “Can the two methods be used interchangeably?"
- Published examples of method comparison studies can be found in disciplines as diverse as Pharmacology ludbrook97, Anaesthesia Myles, and cardiac imaging methods Krumm

Accuracy and Precision

- A method of measurement should ideally be both accurate and precise.
- Barnhart describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown ‘true value’.
- The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value.
- A precise and accurate method will yield results consistently close to the true value.

Accuracy and Precision

- Of course a method may be accurate, but not precise, if the average of its measurements is close to the true value, but those measurements are highly dispersed.
- Conversely a method that is not accurate may be quite precise, as it consistently indicates the same level of inaccuracy.
- The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias.
- The smaller the systematic bias, the greater the accuracy of the method.

Accuracy and Precision

- The FDA define precision as the closeness of agreement (degree of scatter) between a series of measurements obtained from multiple sampling of the same homogeneous sample under prescribed conditions.
- Barnhart describes precision as being further subdivided as
  1. within-run, intra-batch precision or repeatability (which assesses precision during a single analytical run),
  2. between-run, inter-batch precision or repeatability (which measures precision over time)

Inter-Method Bias

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

Inter-Method Bias

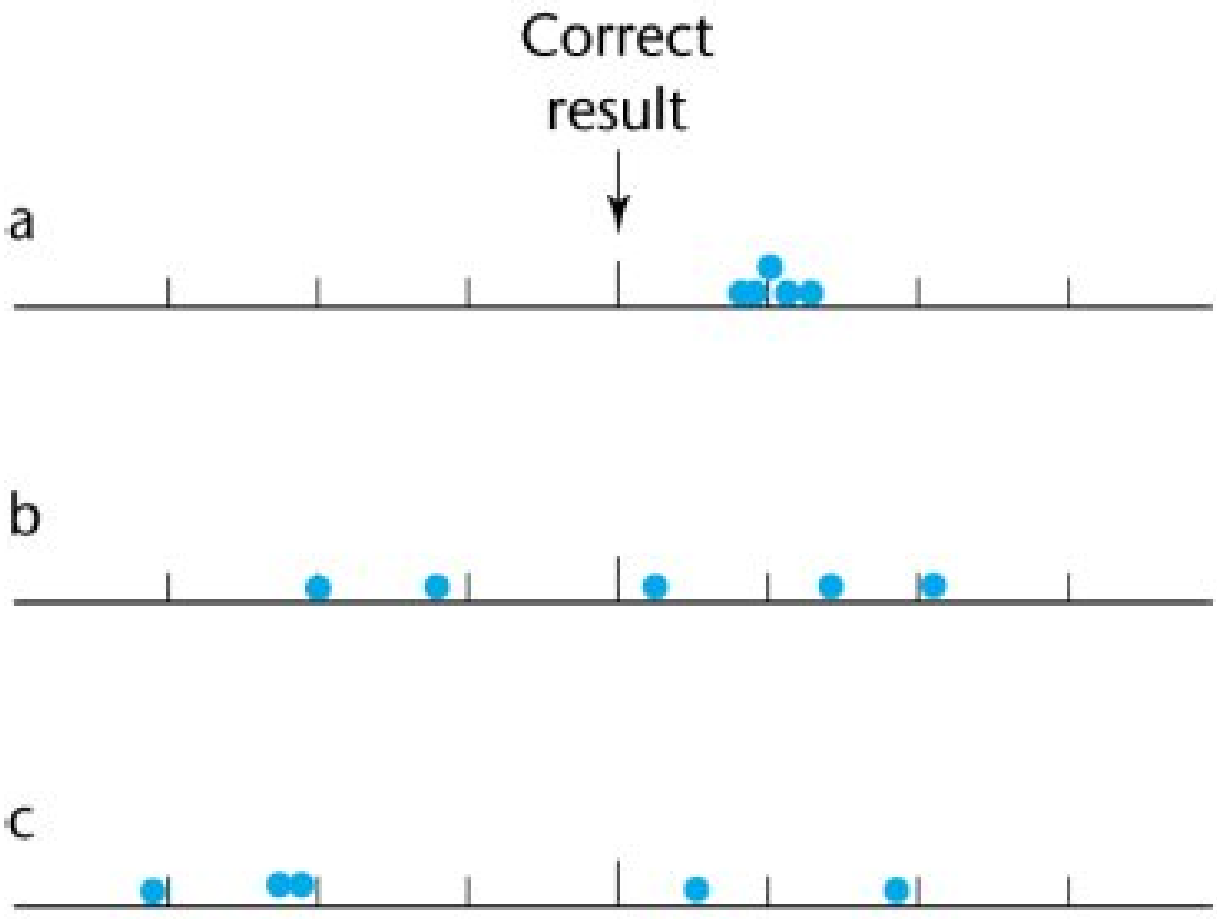
- A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements.
- The data in Table 1.2 (*next slide*) are a good example of possible inter-method bias; the ‘Fotobalk’ consistently recording smaller velocities than the ‘Counter’ method.
- Consequently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

Experimental scientists make a fundamental distinction between ***random***, and ***systematic*** errors. To distinguish between random and systematic errors let us consider a real experiment.
Four students (A-D) each perform an analysis in which exactly 10.00 *ml* of exactly 0.1 M sodium hydroxide is titrated with exactly 0.1 NI hydrochloric acid. Each student performs five replicate titrations, with the results shown in Table 1.1.

Student	Results	(ml)				Comment
A	10.08	10.11	10.09	10.10	10.12	Precise, biased
B	9.88	10.14	10.02	9.80	10.21	Imprecise unbiased
C	10.19	9.79	9.69	10.05	9.78	Imprecise, biased
D	10.04	9.98	10.02	9.97	10.04	Precise, unbiased

Graphical illustration

The results of experiment represented by dot-plots. (The true value is 10.00).



Recall the average for each student $10.0950, 9.9600, 9.9300, 10.0025$ respectively.

Systematic error and bias

Systematic error is a deviation of all measurements in one direction from the true value. It is well represented by the difference between the average value of the determined values and the true value of the measured quantity. This difference is called the bias of measurements.

Random error and precision

Random error is a deviation of a measurement from the average of measured values. It is well represented by the standard deviation of measurements. This value is often called precision of measurements.

Combined error vs. accuracy

Accuracy is in inverse relation to the total deviation of a single measurement from the true value.

Three Conditions

For two methods of measurement to be considered interchangeable, the following conditions must apply [1]:

- No significant inter-method bias
- No difference in the between-subject variabilities of the two methods
- No difference in the within-subject variabilities of the two methods (repeatability)

- To illustrate the characteristics of a typical method comparison study consider the data in Table I Grubbs73.
- In each of twelve experimental trials a single round of ammunition was fired from a 155mm gun, and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels '***Fotobalk***', '***Counter***' and '***Terma***'.

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

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

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

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: Difference between Fotobalk and Counter measurements.

Bland-Altman Methods

Section 2 - Bland-Altman Methods

- Bland-Altman's Methods
- Limits of Agreement
- Implementation with R
- Other R Packages

The Bland-Altman Plot

- The Bland-Altman plot [2, 3] is a very simple graphical method to compare two measurements techniques.
- In this approach the case-wise differences between the two methods are plotted against the corresponding case-wise averages of the two methods.
- A horizontal lines is drawn at the mean difference(the inter-method bias) , and at the limits of agreement, which are defined as the inter-method bias plus and minus 2 times the standard deviation of the differences.

Bland-Altman Plots

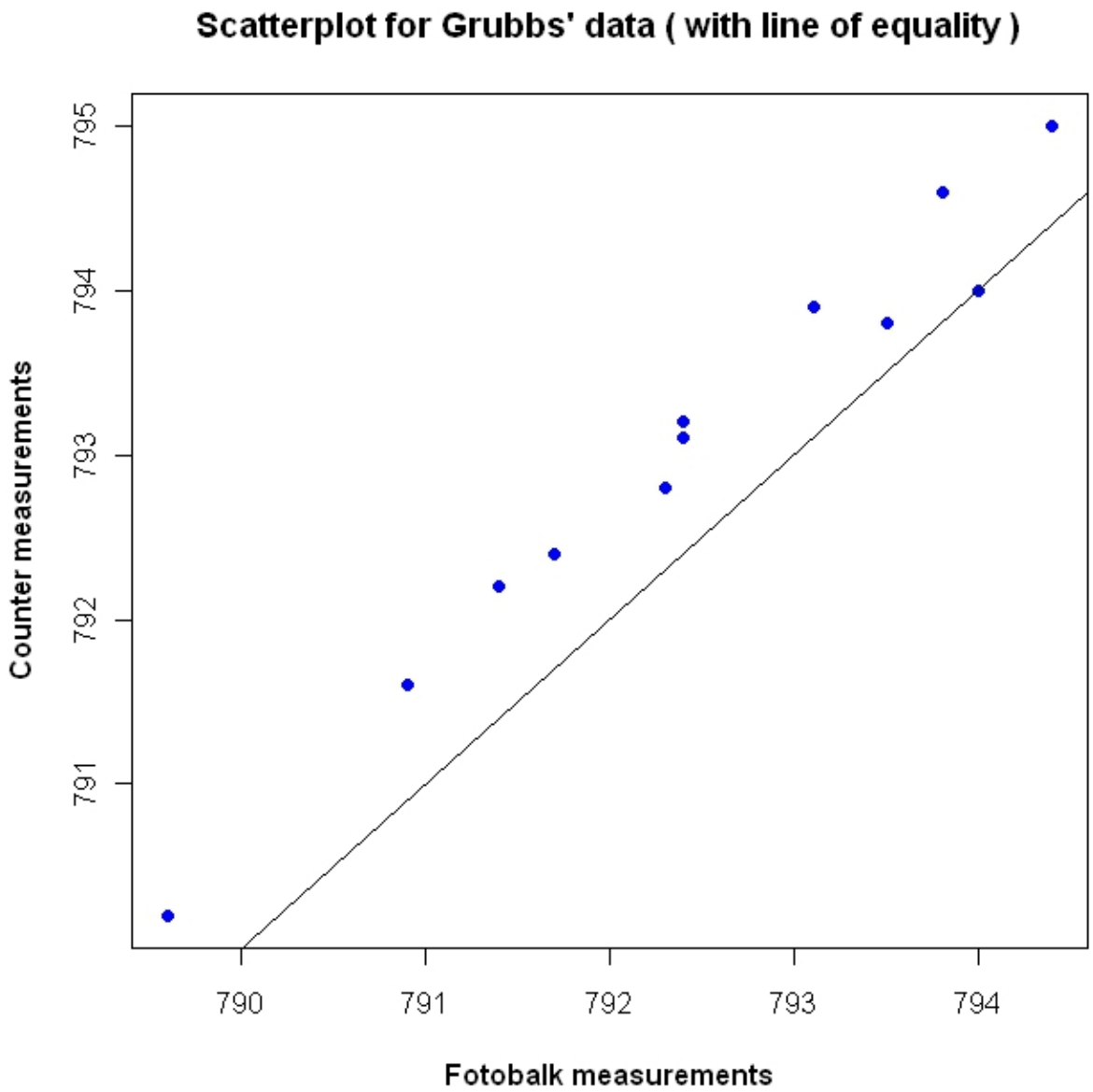
- The issue of whether two measurement methods comparable to the extent that they can be used interchangeably with sufficient accuracy is encountered frequently in scientific research.
- Historically comparison of two methods of measurement was carried out by use of paired sample t-test, correlation coefficients or simple linear regression.
- Statisticians Martin Bland and Douglas Altman recognized the inadequacies of these analyses and articulated quite thoroughly the basis on which of which they are unsuitable for comparing two methods of measurement BA83.

Bland-Altman Plots

- 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*'.

Bland-Altman Plots

- Notwithstanding previous remarks about 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*** must also be shown, as it is necessary to give the correct interpretation of how both methods compare.
- 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.


Scatterplot for Grubbs' data (with line of equality)
Counter measurements
791
792
793
794
795
790
791
792
793
794
Fotobalk measurements

The Bland-Altman Difference Plot

- In light of shortcomings associated with scatterplots, BA83 recommend a further analysis of the data.
- Firstly case-wise differences of measurements of two methods $d_i = y_{1i} - y_{2i}$ for $i = 1, 2, ..n$ on the same subject should be calculated, and then the average of those measurements ($a_i = (y_{1i} + y_{2i})/2$ for $i = 1, 2, ..n$).
- These differences and averages are then plotted.

Limits of Agreement

- Computing limits of agreement features prominently in many method comparison studies, further to Bland And Altman's Work
- Bland Altman 1999 addresses the issue of computing LoAs in the presence of replicate measurements, suggesting several computationally simple approaches. When repeated measures data are available, it is desirable to use all the data to compare the two methods.
- However, the original Bland-Altman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data. However, as a naive analysis, it may be used to explore the data because of the simplicity of the method.
- Carstensen et al computes the limits of agreement to the case with replicate measurements by using LME models.

- The magnitude of the inter-method bias between the two methods is simply the average of the differences $\bar{d}$.
- The variances around this bias is estimated by the standard deviation of the differences $S(d)$.
- This inter-method bias is represented with a line on the Bland-Altman plot.
- These estimates are only meaningful if there is uniform inter-bias and variability throughout the range of measurements, which can be checked by visual inspection of the plot.

- In the case of 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.

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: Fotobalk and Counter methods: differences and averages.

Round	Fotobalk [F]	Terma [T]	Differences [F-T]	Averages [(F+T)/2]
1	793.80	793.20	0.60	793.50
2	793.10	793.30	-0.20	793.20
3	792.40	792.60	-0.20	792.50
4	794.00	793.80	0.20	793.90
5	791.40	791.60	-0.20	791.50
6	792.40	791.60	0.80	792.00
7	791.70	791.60	0.10	791.65
8	792.30	792.40	-0.10	792.35
9	789.60	788.50	1.10	789.05
10	794.40	794.70	-0.30	794.55
11	790.90	791.30	-0.40	791.10
12	793.50	793.50	0.00	793.50

Table: Fotobalk and Terma methods: differences and averages.

- However, the original Bland-Altman method was developed for two sets of measurements done on one occasion (i.e. independent data), and so this approach is not suitable for replicate measures data.
- However, as a naive analysis, it may be used to explore the data because of the simplicity of the method.

Limits of Agreement

- bxc2008 computes the limits of agreement to the case with replicate measurements by using LME models.
- Roy formulates a very powerful method of assessing whether two methods of measurement, with replicate measurements, also using LME models. Roy's approach is based on the construction of variance-covariance matrices.
- Importantly, Roy's approach does not directly address the issue of limits of agreement (although another related analysis , the *Coefficient of Repeatability*, is mentioned).

Limits of Agreement

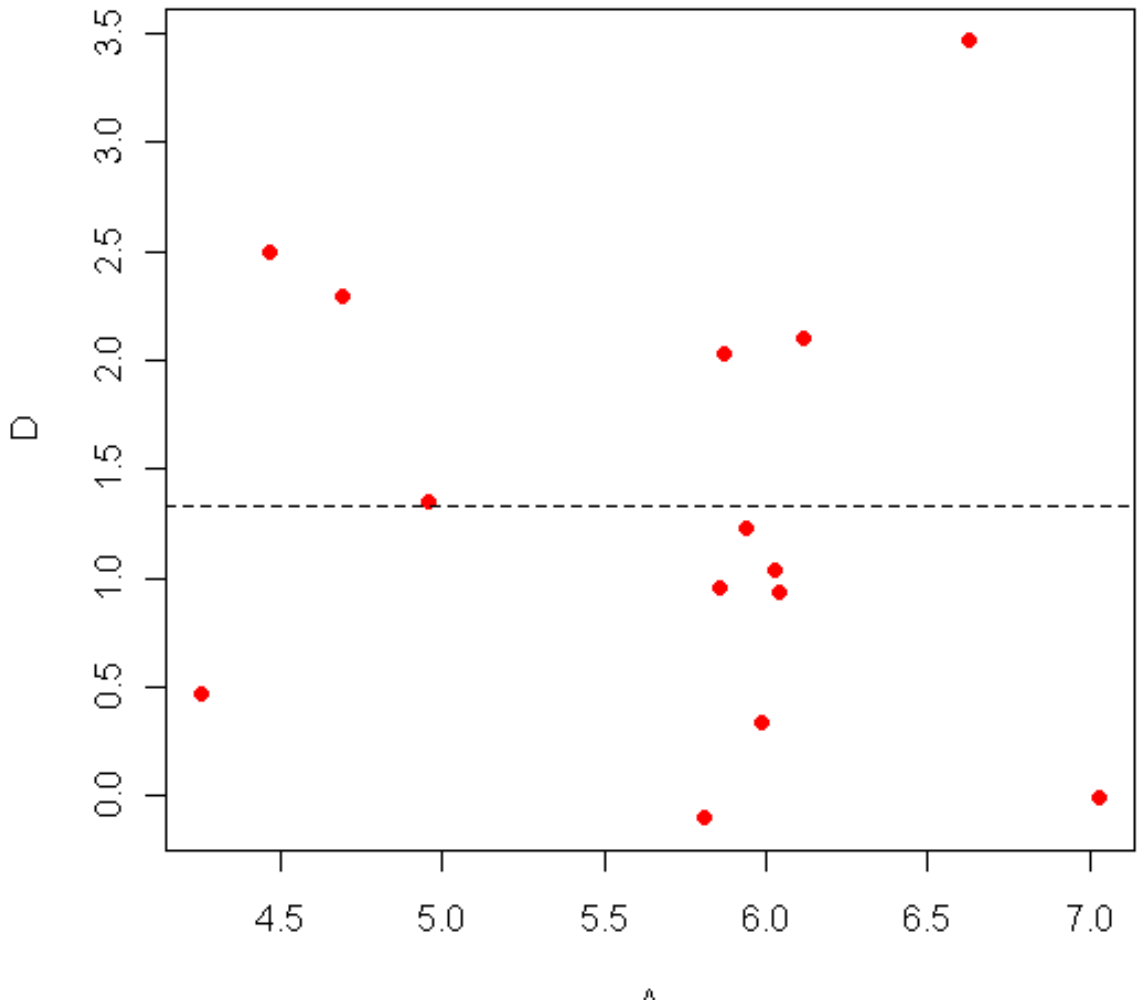
- This paper seeks to use Roy's approach to estimate the limits of agreement. These estimates will be compared to estimates computed under Carstensen's formulation.
- In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required.
- As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus is solely on the standard deviation.

Bland-Altman Plot

```
>X = rnorm(14,6,1);Y = rnorm(14,5.3,1.1)
>
>A=(X+Y)/2 #case-wise averages
>D=X-Y #case-wise differences
>
>Dbar=mean(D) #inter-method bias
>SdD=sd(D) #standard deviation of the differences
>
>plot(A,D,pch=16,col="red", ylim=c(-3,3))
>
>abline(h=Dbar,lty=2)
>abline(h=(Dbar-2*SdD),lty=2)
>abline(h=(Dbar+2*SdD),lty=2)
```

Simple Bland-Altman Plot

Inter-method



Using Bland-Altman Plots

Bland-Altman plots are a powerful graphical methodology for making a visual assessment of the data. BA83 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 Bland-Altman plot is simply a scatterplot of the case-wise averages and differences of two methods of measurement.
- 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 particularly.
- Later it will be shown that case-wise differences are the sole component of the next part of the methodology, the limits of agreement.

- For creating plots, the case wise-averages fulfil several functions, 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.
- BA86 cautions that it would be the difference against either measurement value instead of their average , as the difference relates to both value.

Next Slide

- 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.

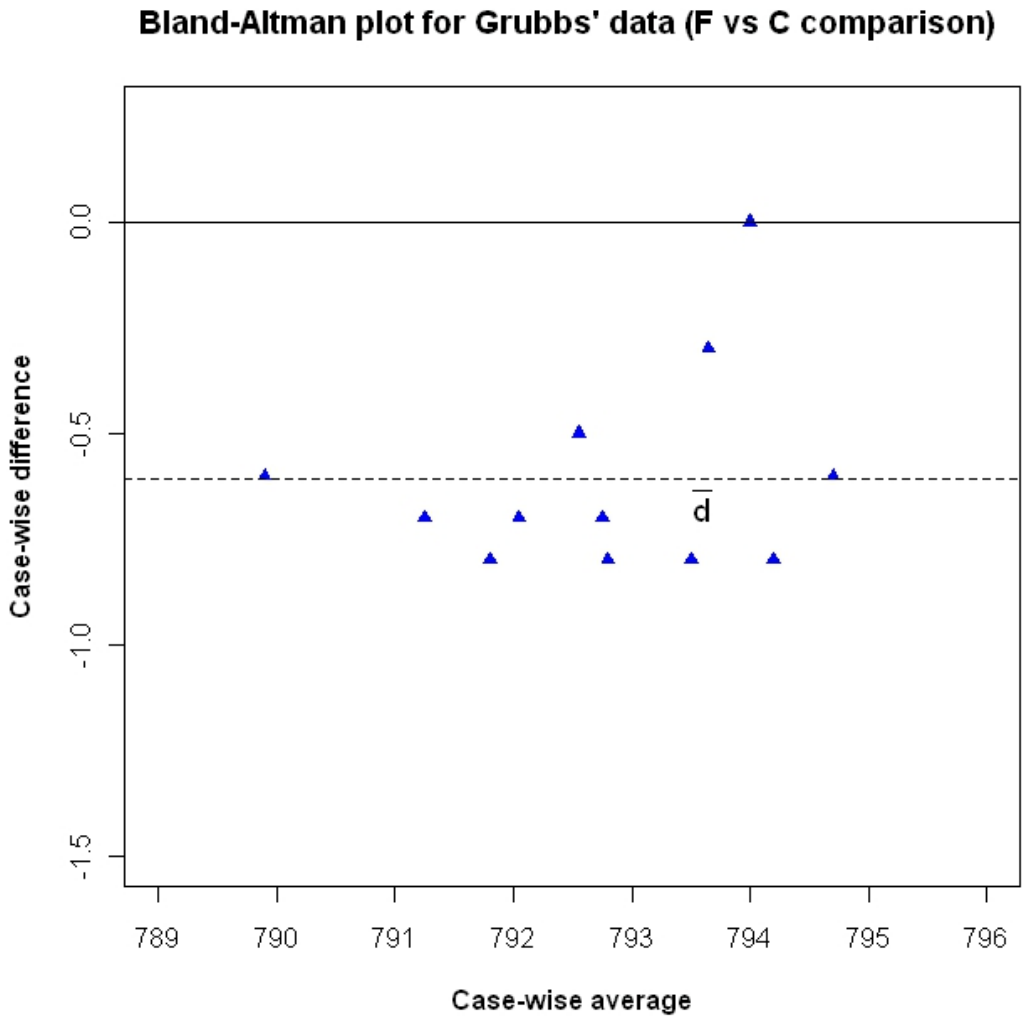
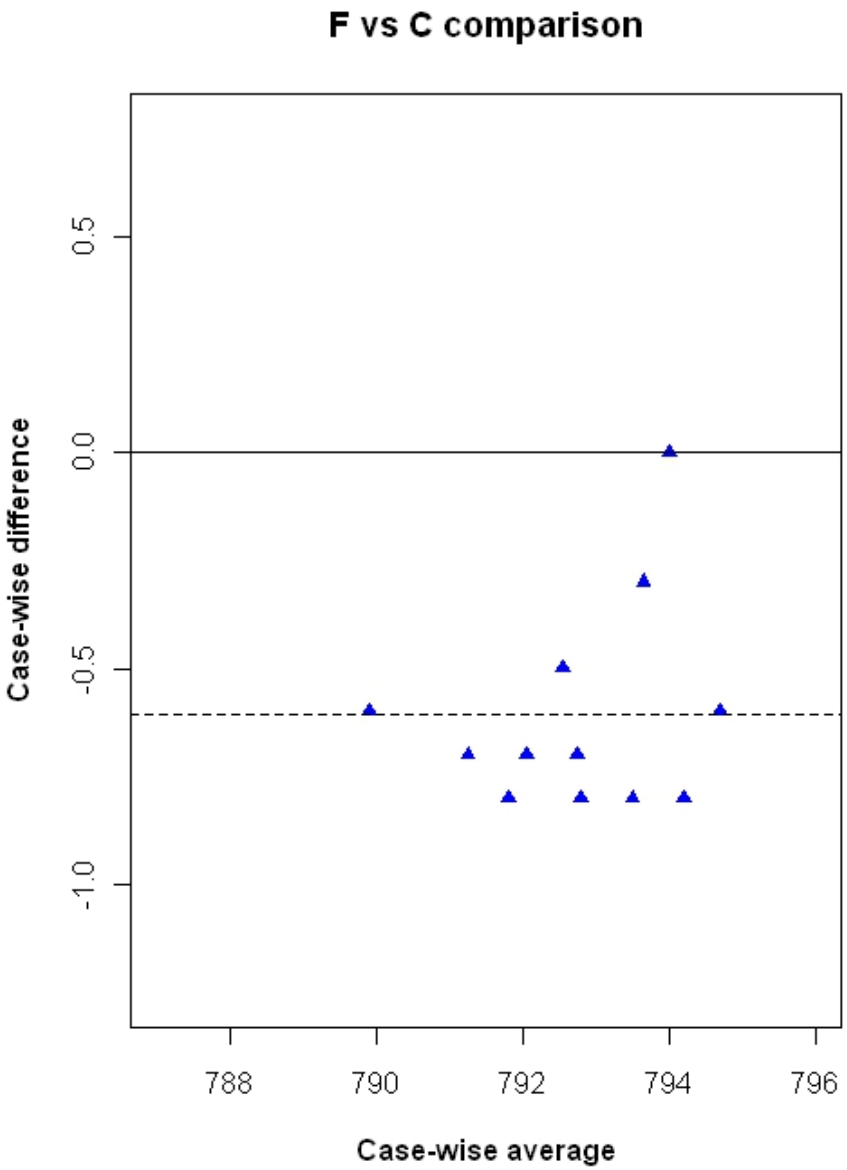
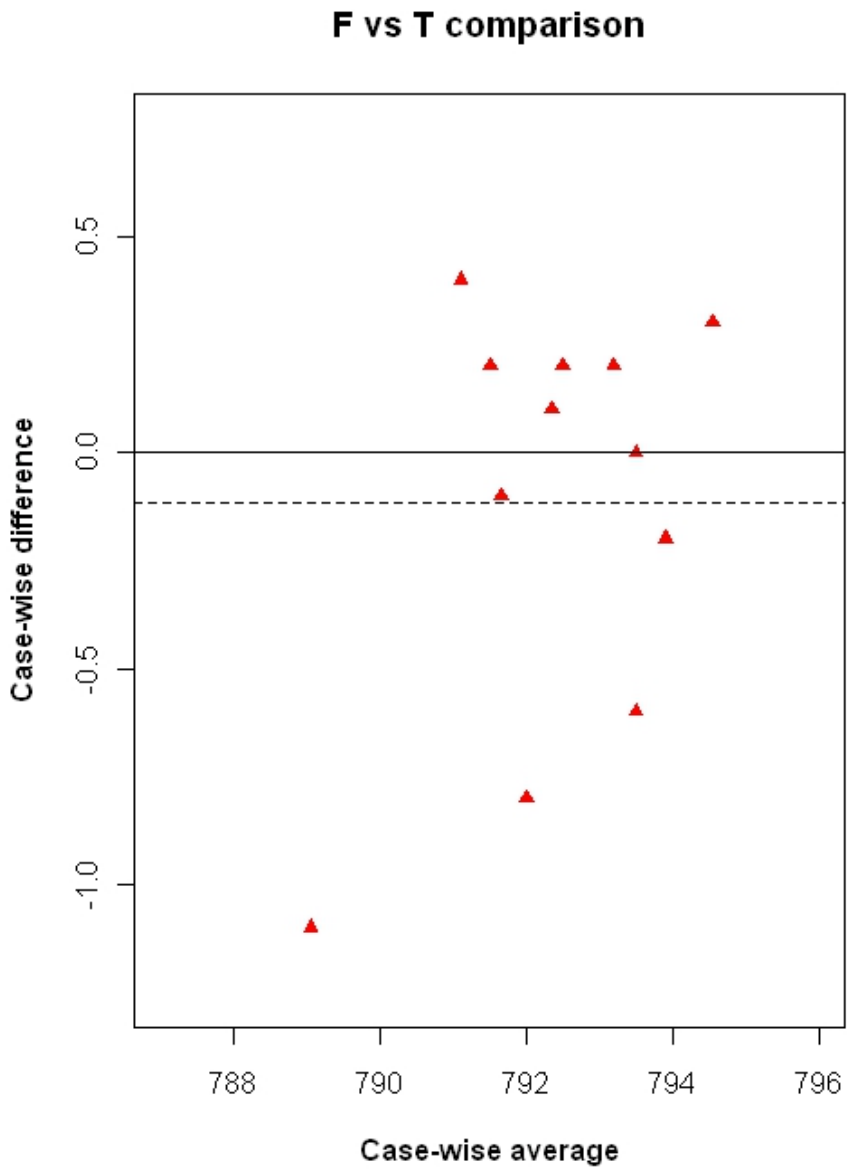


Figure: Bland-Altman plot For Fotobalk and Counter methods.

Next Slide

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.

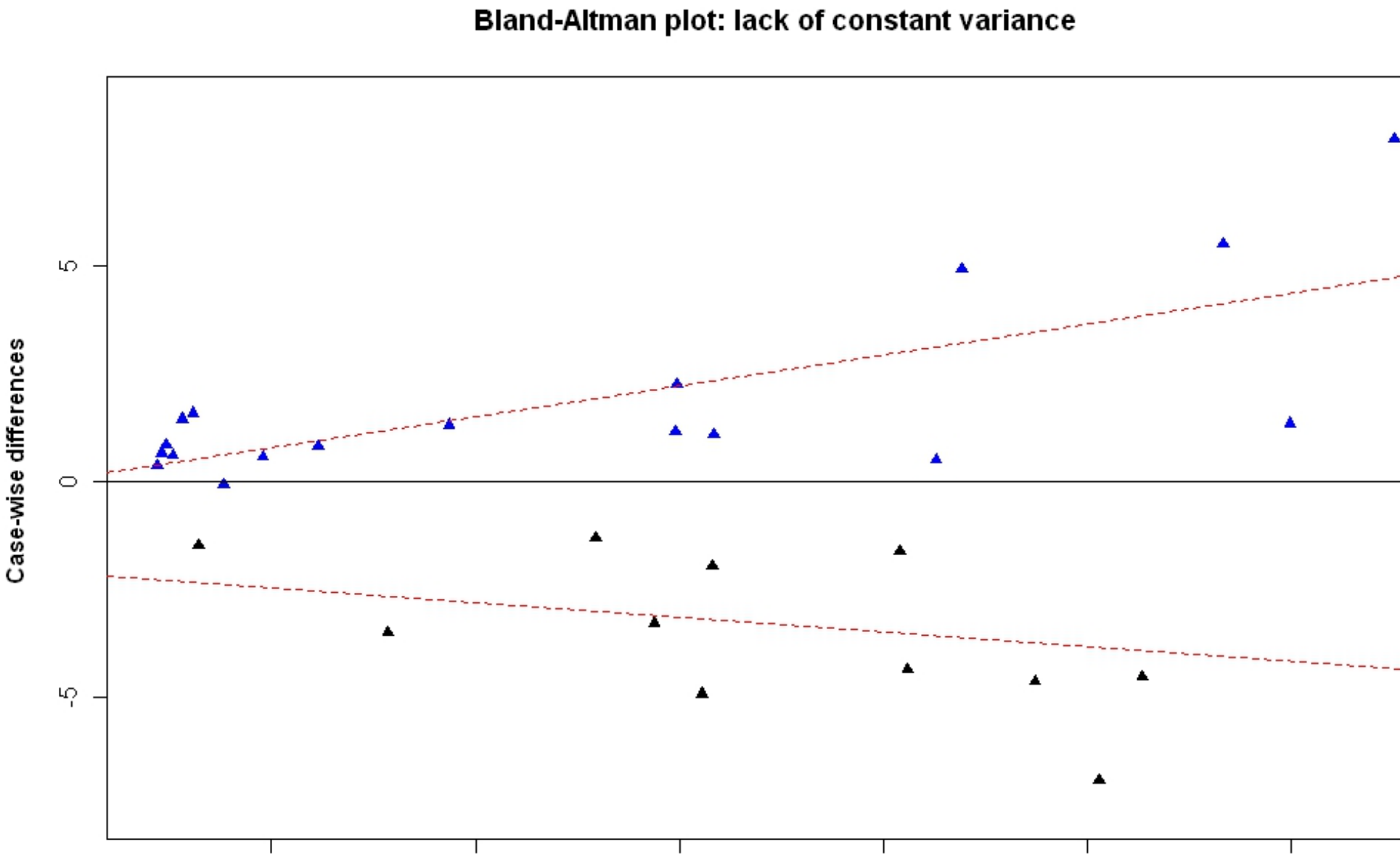

F vs C comparison
Case-wise difference
0.5
0.0
-0.5
-1.0
788
790
792
794
796
Case-wise average

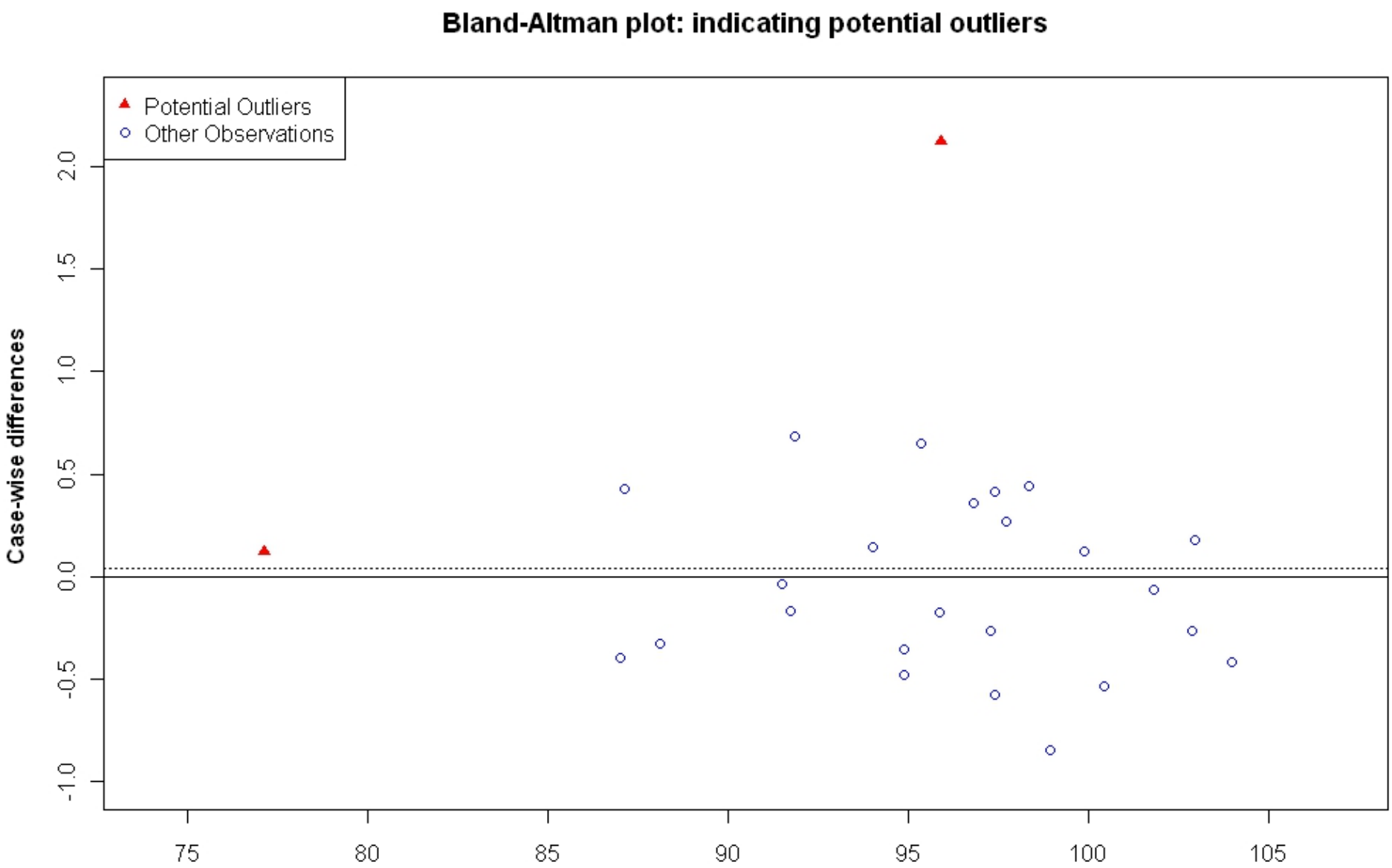

F vs T comparison
Case-wise difference
0.5
0.0
-0.5
-1.0
788
790
792
794
796
Case-wise average

- 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 co-variates.
- Figures 1.4, 1.5 and 1.6 are three prototype Bland-Altman plots derived from simulated data, each for the purpose of demonstrating how the plot would inform an analyst of features that would adversely affect use of the recommended methodology.

Replicate Measurements

- Bland and Altman's approach originally devised for a single measurement on each item by each of the methods.
- Their 1999 paper [3] extended their approach to replicate measurements:
  By replicates we mean two or more measurements on the same individual taken in identical conditions.
  In general this requirement means that the measurements are taken in quick succession.
- Emphasis put on "repeatability".


Bland-Altman plot: lack of constant variance
Case-wise differences
5
0
-5


Bland-Altman plot: indicating potential outliers
Potential Outliers
Other Observations
Case-wise differences
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
75
80
85
90
95
100
105

Thus far, the formulation for comparison of two measurement methods is one where one measurement by each method is taken on each subject. Should there be two or more measurements by each methods, these measurement are known as ‘replicate measurements’. BXC2008 recommends the use of replicate measurements, but acknowledges that additional computational complexity.

BA86 address this problem by offering two different approaches. The premise of the first approach is that replicate measurements can be treated as independent measurements. The second approach is based upon using the mean of the each group of replicates as a representative value of that group. Using either of these approaches will allow an analyst to estimate the inter method bias.
However, because of the removal of the effects of the replicate measurements error, this would cause the estimation of the standard deviation of the differences to be unduly small. BA86 propose a correction for this.

BXC2008 takes issue with the limits of agreement based on mean values, 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. BXC2008 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.

REgression Techniques for MCS

Regression Based Techniques

- Conventional regression models are estimated using the ordinary least squares (OLS) technique, and are referred to as ‘Model I regression’CornCoch,ludbrook97.
- 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 BA83,ludbrook97, 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. CornCoch,ludbrook97, 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.

Deming Regression

- Both Variables are assumed to have attended measurement error.
- Orthonormal Regression (Variance Ratio's are assumed to be equal)
- Deming Regression (Variance Ration is specifed, with default setting of 1)
- Dunn 2002 advises caution with model
- Model Diagnostics?

Editing Comments

-
-
-
-

- The most commonly known Model II methodology is known as Deming's Regression, (also known an Ordinary Least Product regression). Deming regression is recommended by CornCoch as the preferred Model II regression for use in method comparison studies.
- As previously noted, the Bland Altman Plot is uninformative about the comparative influence of proportional bias and fixed bias. Deming's regression provides independent tests for both types of bias.
- For a given λ, Kummel derived the following estimate for the Demimg regression slope parameter. (α is simply estimated by using the identity $\bar{Y} - \hat{\beta}\bar{X}$.)

$$\hat{\beta} = \frac{S_{YY} - \lambda S_{XX} + [(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2]^{1/2}}{2S_{XY}} \quad (1)$$

Deming Regression

As with conventional regression methodologies, Deming's 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.

Deming Regression

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 Demings regression. Measurements of transmitral volumetric flow (MF) by doppler echocardiography, and left ventricular stroke volume (SV) by cross sectional echocardiography in 21 patients withour aortic valve disease are tabulated in zhang. This data set features in the discussion of method comparison studies in [p.398]AltmanBook .

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

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

Deming Regression

Deming's Regression suffers from some crucial drawback. Firstly it is computationally complex, and it requires specific software packages to perform calculations.Secondly it is uninformative about the comparative precision of two methods of measurement. Most importantly CarollRupert states that Deming's regression is acceptable only when the precision ratio (λ, in their paper as η) is correctly specified ,but in practice this is often not the case, with the λ being underestimated.

Using TDI and CP

- Total Deviation Index
- Coverage Probability
- Mountain Plots (Krouwer and Monti)

Using LME Models

Section 7 - Using LME Models in Method comparison

- Carstensen et al
- Roy 2009

BXC2008 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.

BXC2004 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.

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

Carstensen's Mixed Models

- Carstensen *et al* [1] 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 rth replicate measurement on subject i with method m.

$$y_{mir} = \alpha_m + \beta_m \mu_i + c_{mi} + e_{mir} \tag{4}$$

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

Carstensen's Mixed Models

- The intercept term α and the $\beta_m \mu_i$ term follow from *Dunn* [3], expressing constant and proportional bias respectively , in the presence of a real value μ_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.

Carstensen's Mixed Models

- 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.

Carstensen's Mixed Models

- 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).

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 *et al* [1] 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;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad (e_{mi} \sim N(0, \sigma_m^2)) \qquad (5)$$

- 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.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir} \tag{6}$$

- Carstensen *et al* [2] 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 τ_1^2 and τ_2^2 separately. Therefore it must be assumed that they are equal. The variance of the difference can be estimated as follows:

$$var(y_{1j} - y_{2j}) \tag{7}$$

Section 8 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.

Roy's method

- Roy proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects (alertRoy).
- Using Roy's method, four candidate models are constructed, each differing by constraints applied to the variance covariance matrices.
- In addition to computing the inter-method bias, three significance tests are carried out on the

Method Comparison Studies with R

- The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree.
- The two methods must also have equivalent levels of precision.
- Should one method yield results considerably more variable than that of the other, they can not be considered to be in agreement.
- With this in mind a methodology is required that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

BXC2004 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 rth replicate measurement on subject i with method m.

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

The intercept term α and the $\beta_m \mu_i$ term follow from DunnSEME, expressing constant and proportional bias respectively , in the presence of a real value μ_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).

BXC2004 uses the above formula to predict observations for a specific individual *i* by method *m*;

$$BLUP_{mir} = \hat{\alpha_m} + \hat{\beta_m}\mu_i + c_{mi} \quad (9)$$

. Under the assumption that the μs 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. BXC2004 provides an amended formulation which includes an extra interaction term ($d_{mr} \sim N(0, \omega_m^2)$to account for this.

Editing Notes - Roys 2009 paper

-
-

LME models

- In a linear mixed-effects model, responses from a subject are due to both fixed and random effects. A random effect is an effect associated with a sampling procedure.
- Replicate measurements would require use of random effect terms in model.
- Can have differing number of replicate measurements for different subjects.

Hmalett

- Hamlett re-analyses the data of **Lam** to generalize their model to cover other settings not covered by the Lam method.
- In many cases, repeated observation are collected from each subject in sequence and/or longitudinally.

$$y_i = \alpha + \mu_i + \epsilon$$

Hmalett

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

$$y_{mi} + \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim \backslash(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.
Like wise the separate α can not be estimated, only theiir difference can be estimated as $\bar{D}$

Roy's Approach

- Roy proposes an LME model with Kronecker product covariance structure in a doubly multivariate setup.

- Response for ith subject can be written as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + b_{1i} z_{i1} + b_{2i} z_{i2} + \epsilon_i$$

- β_1 and β_2 are fixed effects corresponding to both methods. (β_0 is the intercept.)

- b_{1i} and b_{2i} are random effects corresponding to

Roy's LME model

- Let $\boldsymbol{y}_i$ be the set of responses for subject i (in matrix form).
- $\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \epsilon_i$
- $\boldsymbol{b}_i \sim N_m(0, \boldsymbol{D})$ (m: number of methods)
- $\epsilon_i \sim N_{n_i}(0, \boldsymbol{R})$ (n_i: number of measurements on subject i)

Variance-covariance matrix

- Overall variance covariance matrix for response vector $\boldsymbol{y}_i$

$$\text{Cov}(\boldsymbol{y}_i) = \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i' + \boldsymbol{R}_i$$

- can be re-expressed as follows:

$$\boldsymbol{Z}_i \begin{bmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{bmatrix} \boldsymbol{Z}_i' + \left(V \otimes \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right)$$

- Overall variability between the two methods is sum of between-subject and within-subject variability,

$$\text{Block } \boldsymbol{\Omega}_i = \begin{bmatrix} d_1^2 & d_{12} \\ d_{12} & d_2^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Roy's method

formulates a very powerful method of assessing whether two methods of measurement, with replicate measurements, also using LME models. Roy's approach is based on the construction of variance-covariance matrices. Importantly, Roy's approach does not address the issue of limits of agreement (though another related analysis , the coefficient of repeatability, is mentioned).

Roy's method

This paper seeks to use Roy's approach to estimate the limits of agreement. These estimates will be compared to estimates computed under Carstensen's formulation.
In computing limits of agreement, it is first necessary to have an estimate for the standard deviations of the differences. When the agreement of two methods is analyzed using LME models, a clear method of how to compute the standard deviation is required. As the estimate for inter-method bias and the quantile would be the same for both methodologies, the focus is solely on the standard deviation.

Roy's method

Roy proposes a novel method using the LME model with Kronecker product covariance structure in a doubly multivariate set-up to assess the agreement between a new method and an established method with unbalanced data and with unequal replications for different subjects Roy.

Using Roy's method, four candidate models are constructed, each differing by constraints applied to the variance covariance matrices. In addition to computing the inter-method bias, three significance tests are carried out on the respective formulations to make a judgement on whether or not two methods

Implementation with R]Implementation

Variance-Covariance Structures

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

- Symmetric structure specifies that σ_1^2 may differ from σ_2^2.
- Compound symmetric structure specifies that $\sigma_1^2 = \sigma_2^2$.
- In both cases, σ_{12} may take value other than 0.

The `nlme` Package

- LME models can be implemented in `R` using the `nlme` package, one of the core packages.
- Authors: Jose Pinheiro, Douglas Bates (up to 2007), Saikat DebRoy (up to 2002), Deepayan Sarkar (up to 2005), the `R` Core team (source: `nlme` package manual)
- “Mixed-Effects Models in S and S-PLUS" by JC Pinheiro and DM Bates (Springer,2000)

The Reference Model

```
REF = lme(y ~ meth,
    data = dat,
    random = list(item=pdSymm(~
meth-1)),
    weights=varIdent(form=~1|meth),
    correlation = corSymm(form=~1 |
item/repl),
    method="ML")
```

- LME model that specifies a symmetric matrix structure for both between-subject and within-subject variances.

The Nested Model 1

```
NMB = lme(y ~ meth,
    data = dat,
    random = list(item=pdCompSymm(~
meth-1)),
    weights=varIdent(form=~1|meth),
    correlation = corSymm(form=~1 |
item/repl),
    method="ML")
```

- LME model that specifies a compound symmetric matrix structure for between-subject and symmetric structure within-subject variances.

The Nested Model 2

```
NMW = lme(y ~ meth,
    data = dat,
    random = list(item=pdSymm(~
meth-1)),
    #weights=varIdent(form=~1|meth),
    correlation = corCompSymm(form=~1 |
item/repl),
    method="ML")
```

- LME model that specifies a symmetric matrix structure for between-subject and compound symmetric structure within-subject variances.

The Nested Model 3

```
NMO = lme(y ~ meth,
    data = dat,
    random = list(item=pdCompSymm(~
meth-1)),
    #weights=varIdent(form=~1|meth),
    correlation = corCompSymm(form=~1 |
/repl),
    method="ML")
```

- LME model that specifies a compound symmetric matrix structure for both between-subject and within-subject variances.

Example]Example

Example: Blood Data

- Used in Bland and Altman's 1999 paper [3]. Data was supplied by Dr E O'Brien.
- Simultaneous measurements of systolic blood pressure each made by two experienced observers, J and R, using a sphygmometer.
- Measurements also made by a semi-automatic blood pressure monitor, denoted S.
- On 85 patients, 3 measurement made in quick succession by each of the three observers (765

Example: Blood Data

Inter-method Bias between J and S: 15.62 mmHg

```
>summary(REF)
.....
Fixed effects: y ~ meth
               Value Std.Error  DF t-value
(Intercept) 127.41     3.3257 424  38.310
methS        15.62     2.0456 424   7.636
.....
```

Between-subject variance covariance matrix

```
..
Random effects:
 Formula: ~method - 1 | subject
 Structure: General positive-definite
            StdDev    Corr
methodJ  30.396975 methdJ
methodS  31.165565 0.829
Residual  6.116251
..
```

$$\hat{\boldsymbol{D}} = \begin{pmatrix} 923.97 & 785.34 \\ 785.34 & 971.29 \end{pmatrix}$$

Within-subject variance covariance matrix

```
Correlation Structure: General
 Formula: ~1 | subject/obs
 Parameter estimate(s):
 Correlation:
  1
2 0.288
Variance function:
 Structure: Different standard deviations
 Formula: ~1 | method
 Parameter estimates:
       J        S
1.00000 1.49806
```

Overall variance covariance matrix

- Overall variance

$$\text{Block } \hat{\Omega} = \hat{\boldsymbol{D}} + \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix}$$

- Standard deviation of the differences can be computed accordingly : 20.32 mmHg.

- Furthermore, limits of agreement can be computed: $[15.62 \pm (2 \times 20.32)]$ (mmHg).

Some useful R commands

- `intervals` :

  This command obtains the estimate and confidence intervals on the parameters associated with the model. This is particularly useful in writing some code to extract estimates for inter-method bias and variances, and hence estimates for the limits of agreement.

- `anova` :

  When a reference model and nested model are specified as arguments, this command performs a likelihood ratio test.

Formal Tests: Between-subject Variances

- Test the hypothesis that both methods have equal between-subject variances.
- Constructed an alternative model “Nested Model B" using ***compound symmetric*** form for between-subject variance (hence specifying equality of between-subject variances).
- Use a likelihood ratio test to compare models.

```
...
> anova(REF,NMB)
    Model df ...      logLik   Test   L.Rat
REF     1  8 ...  -2030.736
```

Formal Tests: Within-subject Variances

- Test the hypothesis that both methods have equal within-subject variances.
- Constructed an alternative model “Nested Model W" using compound symmetric form for within-subject variance (hence specifying equality of within-subject variances).
- Again, use a likelihood ratio test to compare models.

```
...
> anova(REF,NMW)
    Model df ...     logLik   Test  L.Rati
REF     1  2 ...  -2020.726
```

Formal Tests : Outcomes

- Inter-method bias: Significant difference in mean values detected.
- Between-subject variance: No significant difference in between-subject variances between the two methods detected.
- Within-subject variance: A significant difference in within-subject variances is detected.
- Can not recommend switching between the two methods.

Remarks

- Can perform a test for equality of overall variances.
- This can be done by specifying a compound symmetry structure for both between-subject and within-subject variances when constructing a nested model.
- Roy controls the family-wise error rate in paper, using Bonferroni correction procedure.

Carstensen presents a model where the variation between items for method *m* is captured by σ_m and the within item variation by τ_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 limits of agreement computed by Roy's method are derived from the variance covariance matrix for overall variability.
- This matrix is the sum of the between subject VC matrix and the within-subject VC matrix.
- The standard deviation of the differences of methods x and y is computed using values from the overall VC matrix.

$$\text{var}(x - y) = \text{var}(x) + \text{var}(y) - 2\text{cov}(x, y)$$

Carstensen's LOAs

- The respective estimates computed by both methods are tabulated as follows. Evidently there is close correspondence between both sets of estimates.
- bxc2008 formulates an LME model, both in the absence and the presence of an interaction term.bxc 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.

Roy2006 uses the “Blood" data set, which featured in BA99.

Likelihood Ratio Tests

- The relationship between the respective models presented by roy is known as “nesting". A model A to be nested in the reference model, model B, if Model A is a special case of Model B, or with some specific constraint applied.
- A general method for comparing models with a nesting relationship is the **likelihood ratio test (LRTs)**.
- LRTs are a family of tests used to compare the value of likelihood functions for two models, whose respective formulations define a hypothesis to be tested (i.e. the nested and reference model).

- When testing hypotheses around covariance parameters in an LME model, REML estimation for both models is recommended by West et al. REML estimation can be shown to reduce the bias inherent in ML estimates of covariance parameters west.
- Conversely, pb advises that testing hypotheses on fixed-effect parameters should be based on ML estimation, and that using REML would not be appropriate in this context.

Variance Covariance Matrices

- Under Roy's model, random effects are defined using a bivariate normal distribution. Consequently, the variance-covariance structures can be described using 2×2 matrices.
- A discussion of the various structures a variance-covariance matrix can be specified under is required before progressing.
- The following structures are relevant:
  1. the identity structure,
  2. the compound symmetric structure
  3. the symmetric structure.

Variance Covariance Matrices

- The **identity** structure is simply an abstraction of the identity matrix.
- The **compound symmetric** structure and **symmetric** structure can be described with reference to the following matrix (here in the context of the overall covariance Block-$\mathbf{\Omega}_i$, but equally applicable to the component variabilities $\boldsymbol{G}$ and $\mathbf{\Sigma}$);

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix}$$

Symmetric structure requires the equality of all the diagonal terms, hence $\omega_1^2 = \omega_2^2$. Conversely compound symmetry make no such constraint on the diagonal elements. Under the identity structure, $\omega_{12} = 0$. A comparison of a model fitted using symmetric structure with that of a model fitted using the compound symmetric structure is equivalent to a test of the equality of variance.

In the presented example, it is shown that Roy's LOAs are lower than those of (BXC-model), when covariance between methods is present.

Remarks on the Multivariate Normal Distribution

- Diligence is required when considering the models. Carstensen specifies his models in terms of the univariate normal distribution.
- Roy's model is specified using the bivariate normal distribution.
- This gives rises to a key difference between the two model, in that a bivariate model accounts for covariance between the variables of interest.

- Roys uses and LME model approach to provide a set of formal tests for method comparison studies.
- Four candidates models are fitted to the data.
- These models are similar to one another, but for the imposition of equality constraints.
- These tests are the pairwise comparison of candidate models, one formulated without constraints, the other with a constraint.
  Roy's model uses fixed effects $\beta_0 + \beta_1$ and $\beta_0 + \beta_1$ to specify the mean of all observationsby methods 1 and 2 respectively.

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.

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

Remarks on the Multivariate Normal Distribution

The multivariate normal distribution of a k-dimensional random vector $X = [X_1, X_2, \ldots, X_k]$ can be written in the following notation:

$$X \sim \mathcal{N}(\mu, \Sigma),$$

or to make it explicitly known that X is k-dimensional,

$$X \sim \mathcal{N}_k(\mu, \Sigma).$$

with k-dimensional mean vector

$$\mu = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]$$

and $k \times k$ covariance matrix

$$\Sigma = [\mathrm{Cov}[X_i, X_j]], \; i = 1, 2, \ldots, k; \; j = 1, 2, \ldots, k$$

1. Univariate Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2),$$

2. Bivariate Normal Distribution

(a)

$$X \sim \mathcal{N}_2(\mu, \Sigma),$$

(b)

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Note 1: Coefficient of Repeatability

- The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements BA99.
- Once the within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.

Carstensen model in the single measurement case

- BXC2004 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 model in the single measurement case

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \qquad e_{mi} \sim \mathcal{N}(0, \sigma^2_m) \tag{10}$$

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.

Note 3: Model terms

It is important to note the following characteristics of this model.

- Let the number of replicate measurements on each item i for both methods be n_i, hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p. An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.
- Later on $\boldsymbol{X}_i$ will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be

Note 3: Model terms (Continued)

- ϵ is the $2n_i \times 1$ vector of residuals for measurements on item i.
- $\boldsymbol{G}$ is the 2×2 covariance matrix for the random effects.
- $\boldsymbol{R}_i$ is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i.
- The expected value is given as $\mathrm{E}(\boldsymbol{y}_i) = \boldsymbol{X}_i\beta$. hamlett
- The variance of the response vector is given by $\mathrm{Var}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i' + \boldsymbol{R}_i$ ...hamlett.

References

- Carstensen, Gurrin (2004): *Generalized Linear Models*, Chapman and Hall/CRC.
- Carstensen, Gurrin (2008): *Generalized Linear Models*, Chapman and Hall/CRC.
- Dunn, P. and Nelder, J. (1989): *SEME*, Chapman and Hall/CRC.
- A Roy (2009): *An application of linear mixed effects model to assess the agreement between two methods with replicated observations* Journal of Biopharmaceutical Statistics

References]References

References

- A Roy (2009): *An application of linear mixed effects model to assess the agreement between two methods with replicated observations* Journal of Biopharmaceutical Statistics
- Bland JM, Altman DG (1986) *Statistical method fo assessing agreement between two methods of clinical measurement*.
- Bland JM, Altman DG (1999) *Measuring agreement in method comparison studies.* Statistical Methods in Medical Research
- Pinheiro JC, Bates DM (2000): *Mixed-effects*

Thanks

- Dr Kevin Hayes, University of Limerick
- Mr Kevin Burke, University of Limerick