SCRATCH

Kevin O'Brien

February 11, 2017

Contents

Chapter 1

Introduction

1.1 Methods of assessing agreement

1.1.1 Agreement

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

Bland and Altman (1986)expressed this in the terms *we want to know by how much the new method is likely to differ from the old; if this is not enough to cause problems in clinical interpretation we can replace the old method by the new or use the two interchangeably. 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 comparisonand to choose the sample size .*

1.1.2 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* and propose estimating its value by determining the mean of the differences. The

variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measures.

Bland and Altman attend to the issue of repeated measures in 1996.

Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual.

Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation ,and the correlation coefficient.

The above plot incorporates both the conventional limits of agreement (the inner pair of dashed lines), the 't' limits of agreement (the outer pair of dashed lines) centred around the inter-method bias (indicated by the full line). This plot is intended for expository purposes only, as the sample size is small.

1.1.3 Equivalence and Interchangeability

Limits of agreement are intended to analyse equivalence. How this is assessed is the considered judgement of the practitioner. In Bland and Altman (1986) an example of good agreement is cited. For two methods of measuring 'oxygen saturation', the limits of agreement are calculated as (-2.0,2.8).A practitioner would ostensibly find this to be sufficiently narrow.

If the limits of agreement are not clinically important, which is to say that the differences tend not to be substantial, the two methods may be used interchangeably. Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

1.2 Introduction

The problem of assessing the agreement between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a 'method comparison study'. Published examples of method comparison studies can be found in disciplines

as diverse as pharmacology (Ludbrook, 1997), anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000).

To illustrate the characteristics of a typical method comparison study consider the data in Table I (Grubbs, 1973). In each of twelve experimental trials, a single round of ammunition was fired from a 155mm gun and its velocity was measured simultaneously (and independently) by three chronographs devices, identified here by the labels 'Fotobalk', 'Counter' and 'Terma'.

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

Table 1.1: 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.

A method of measurement should ideally be both accurate and precise. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown 'true value'. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise, if the average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely a method that is not accurate may be quite precise, as it consistently indicates the same level of inaccuracy. The tendency of a method of measurement to consistently give results above or below the true value is a source of systematic bias. The smaller the systematic bias, the greater the accuracy of the method.

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

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

to compare the precision of both methods of measurement.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.2: Difference between Fotobalk and Counter measurements.

1.3 Method Comparison Studies

Agreement between two methods of clinical measurement can be quantified using the differences between observations made using the two methods on the same subjects. The 95% limits of agreement, estimated by mean difference +/- 1.96 standard deviation of the differences, provide an interval within which 95% of differences between measurements by the two methods are expected to lie.

1.4 Discussion on Method Comparison Studies

The need to compare the results of two different measurement techniques is common in medical statistics.

In particular, in medicine, new methods or devices that are cheaper, easier to use, or less invasive, are routinely developed. Agreement between a new method and a traditional reference or gold standard must be evaluated before the new one is put into practice. Various methodologies have been proposed for this purpose in recent years.

Indications on how to deal with outliers in Bland Altman plots

We wish to determine how outliers should be treated in a Bland Altman Plot

In their 1983 paper they merely state that the plot can be used to 'spot outliers'.

In their 1986 paper, Bland and Altman give an example of an outlier. They state that it could be omitted in practice, but make no further comments on the matter.

In Bland and Altmans 1999 paper, we get the clearest indication of what Bland and Altman suggest on how to react to the presence of outliers. Their recommendation is to recalculate the limits without them, in order to test the difference with the calculation where outliers are retained.

The span has reduced from 77 to 59 mmHg, a noticeable but not particularly large

reduction.

However, they do not recommend removing outliers. Furthermore, they say:

We usually find that this method of analysis is not too sensitive to one or two large outlying differences.

We ask if this would be so in all cases. Given that the limits of agreement may or may not be disregarded, depending on their perceived suitability, we examine whether it would possible that the deletion of an outlier may lead to a calculation of limits of agreement that are usable in all cases?

Should an Outlying Observation be omitted from a data set? In general, this is not considered prudent.

Also, it may be required that the outliers are worthy of particular attention themselves.

Classifying outliers and recalculating We opted to examine this matter in more detail.

The following points have to be considered

how to suitably identify an outlier (in a generalized sense)

Would a recalculation of the limits of agreement generally results in a compacted range between the upper and lower limits of agreement?

1.4.1 Agreement

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

1.4.2 Lack Of Agreement

1. Constant Bias

2. Proportional Bias

Constant Bias

This is a form of systematic deviations estimated as the average difference between the test and the reference method

Proportional Bias

Two methods may agree on average, but they may exhibit differences over a range of

1.5 Methods of assessing agreement

1. Pearson's Correlation Coefficient
2. Intraclass correlation coefficient
3. Bland Altman Plot
4. Bartko's Ellipse (1994)
5. Blackwood Bradley Test
6. Lin's Reproducibility Index
7. Luiz Step function

Bland and Altman attend to the issue of repeated measures in 1996. Repeated measurements on several subjects can be used to quantify measurement error, the variation between measurements of the same quantity on the same individual. Bland and Altman discuss two metrics for measurement error; the within-subject standard deviation ,and the correlation coefficient.

The above plot incorporates both the conventional limits of agreement (the inner pair of dashed lines), the 't' limits of agreement (the outer pair of dashed lines)

centred around the inter-method bias (indicated by the full line). This plot is intended for expository purposes only, as the sample size is small.

1.5.1 Equivalence and Interchangeability

Limits of agreement are intended to analyse equivalence. How this is assessed is the considered judgement of the practitioner. In Bland and Altman (1986) an example of good agreement is cited. For two methods of measuring 'oxygen saturation', the limits of agreement are calculated as (-2.0,2.8).A practitioner would ostensibly find this to be sufficiently narrow.

If the limits of agreement are not clinically important, which is to say that the differences tend not to be substantial, the two methods may be used interchangeably. Dunn (2002) takes issue with the notion of 'equivalence', remarking that while agreement indicated equivalence, equivalence does not reflect agreement.

1.6 Introductory Definitions

To illustrate the characteristics of a typical method comparison study consider the data in Table I, taken from Grubbs (1973). In each of twelve experimental trials a single round of ammunition was fired from a 155mm gun, and its velocity was measured simultaneously (and independently) by three chronographs devices, referred to here as 'Fotobalk', 'Counter' and 'Terma'.

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

Table 1.3: Measurement of 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 I can not be assumed to be 'true values' in any absolute sense. For expository purposes only the first two methods 'Fotobalk' and 'Counter' will enter in the immediate discussion.

While lack of agreement between two methods is inevitable, the question , as posed by Altman and Bland (1983), is 'do the two methods of measurement agree sufficiently closely?'

A method of measurement should ideally be both accurate and precise.An accurate measurement methods shall give a result close to the 'true value'. Precision of a

method is indicated by how tightly clustered its measurements are around their mean measurement value.

A precise and accurate method should yield results consistently close to the true value. However a method may be accurate, but not precise. The average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely an inaccurate method 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 lesser the systematic bias, the greater the accuracy of the method.

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

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

Round	Fotobalk (F)	Counter (C)	F-C
1	793.80	794.60	-0.80
2	793.10	793.90	-0.80
3	792.40	793.20	-0.80
4	794.00	794.00	0.00
5	791.40	792.20	-0.80
6	792.40	793.10	-0.70
7	791.70	792.40	-0.70
8	792.30	792.80	-0.50
9	789.60	790.20	-0.60
10	794.40	795.00	-0.60
11	790.90	791.60	-0.70
12	793.50	793.80	-0.30

Table 1.4: Difference between Fotobalk and Counter measurements

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree or not. These 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.

Therefore a methodology must be introduced that allows an analyst to estimate the inter-method bias, and to compare the precision of both methods of measurement.

1.6.1 Agreement Criteria

Roy's method considers two methods to be in agreement if three: no significant bias, i.e. the difference between the two mean readings is not "statistically significant", high

overall correlation coefficient, the agreement between the two methods by testing their repeatability coefficients. Two methods of measurement can be said to be in agreement if there is no significant difference between in three key respects.

Roy additionally uses the overall correlation coefficient to provide extra information about the comparison, with a minimum of 0.82 being required. Firstly, there is no inter-method bias between the two methods, i.e. there is no persistent tendency for one method to give higher values than the other. Secondly, both methods of measurement have the same within-subject variability. In such a case the variance of the replicate measurements would consistent for both methods. Lastly, the methods have equal between-subject variability. Put simply, for the mean measurements for each case, the variances of the mean measurements from both methods are equal. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009) allows for a formal test of each.

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

A method of measurement should ideally be both accurate and precise. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown 'true value'. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise. If the average of its measurements is close to the true value, but those measurements are highly dispersed. Conversely a method that is not accurate may be quite precise, as it consistently indicates the same

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

Table 1.5: Velocity measurement from the three chronographs (Grubbs 1973).

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.

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 et al. (2007) describes precision as being further subdivided as either within-run, intra-batch precision or repeatability (which assesses precision during a single analytical run), or between-run, inter-batch precision or repeatability (which measures precision over time)

In the context of the agreement of two methods, there is also a tendency of one

measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than 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.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.6: Difference between Fotobalk and Counter measurements.

1.7 Introduction

The problem of assessing the agreement between two or more methods of measurement is ubiquitous in scientific research, and is commonly referred to as a 'method comparison study'. Published examples of method comparison studies can be found in disciplines as diverse as Pharmacology (Ludbrook, 1997), Anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000).

To illustrate the characteristics of a typical method comparison study consider the data in Table I (Grubbs, 1973). In each of twelve experimental trials a single round of ammunition was fired from a 155mm gun, and its velocity was measured

simultaneously (and independently) by three chronographs devices, identified here by the labels ‘Fotobalk’, ‘Counter’ and ‘Terma’.

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

Table 1.7: 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.

A method of measurement should ideally be both accurate and precise. Barnhart et al. (2007) describes agreement as being a broader term that contains both of those qualities. An accurate measurement method will give results close to the unknown 'true value'. The precision of a method is indicated by how tightly measurements obtained under identical conditions are distributed around their mean measurement value. A precise and accurate method will yield results consistently close to the true value. Of course a method may be accurate, but not precise. If the average of its measurements

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

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 et al. (2007) describes precision as being further subdivided as either within-run, intra-batch precision or repeatability (which assesses precision during a single analytical run), or between-run, inter-batch precision or repeatability (which measures precision over time)

In the context of the agreement of two methods, there is also a tendency of one measurement method to consistently give results above or below the other method. Lack of agreement is a consequence of the existence of 'inter-method bias'. For two methods to be considered in good agreement, the inter-method bias should be in the region of zero. A simple estimation of the inter-method bias can be calculated using the differences of the paired measurements. The data in Table 1.2 are a good example of possible inter-method bias; the 'Fotobalk' consistently recording smaller velocities than the 'Counter' method. Consequently one would conclude that there is lack of agreement between the two methods.

The absence of inter-method bias by itself is not sufficient to establish whether two measurement methods agree. The two methods must also have equivalent levels of precision. Should one method yield results considerably more variable than 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.

Round	Fotobalk (F)	Counter (C)	F-C
1	793.8	794.6	-0.8
2	793.1	793.9	-0.8
3	792.4	793.2	-0.8
4	794.0	794.0	0.0
5	791.4	792.2	-0.8
6	792.4	793.1	-0.7
7	791.7	792.4	-0.7
8	792.3	792.8	-0.5
9	789.6	790.2	-0.6
10	794.4	795.0	-0.6
11	790.9	791.6	-0.7
12	793.5	793.8	-0.3

Table 1.8: Difference between Fotobalk and Counter measurements.

1.8 Gold Standard

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

The Gold Standard may not be financially feasible for general use, and therefore more economical methods, of suitable levels of precisions, must be devised. Method Comparison studies is used to ascertain the levels of precision of such methods.

1.9 Method Comparison Stduies with R

1.9.1 Accuracy and Precision

An important consideration in discussing methods of measurement are the issues of accuracy and precision.

1.9.2 What is Agreement

Agreement between two methods of clinical measurement can be quantified using the differences between observations made using the two methods on the same subjects. (Bland and Altman 1999)

1.9.3 Bias

Bland and Altman define bias a *a consistent tendency for one method to exceed the other* [3] and propose estimating its value by determining the mean of the differences. The variation about this mean shall be estimated by the standard deviation of the differences. Bland and Altman remark that these estimates are based on the assumption that bias and variability are constant throughout the range of measurements.

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.

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i*, 307–310.

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

Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics 15*(1), 53–66.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

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

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

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