

# SCRATCH

Kevin O'Brien

February 3, 2017

# Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Introduction . . . . .	2
1.2	Introduction . . . . .	5
1.3	Method Comparison Stduies with R . . . . .	9
1.3.1	Accuracy and Precision . . . . .	9
1.3.2	What is Agreement . . . . .	9
1.4	Bland Altman Methodology . . . . .	10
1.4.1	Bias . . . . .	10
1.5	Discussion on Method Comparison Studies . . . . .	10
1.5.1	Agreement . . . . .	11
1.5.2	Lack Of Agreement . . . . .	12
1.6	Bland Altman Plot . . . . .	12
1.6.1	Bland Altman plots using 'Gold Standard' raters . . . . .	12
1.6.2	Bias Detection . . . . .	13

# Chapter 1

## Introduction

### 1.1 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 these data is that all three methods of measurement are assumed to have an attendant 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 1.2: Difference between Fotobalk and Counter measurements.

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

An important aspect of these data is that all three methods of measurement are assumed to have an attendant 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 1.4: Difference between Fotobalk and Counter measurements.

## 1.3 Method Comparison Studies with R

### 1.3.1 Accuracy and Precision

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

### 1.3.2 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)

## 1.4 Bland Altman Methodology

### 1.4.1 Bias

Bland and Altman define bias as *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 measures.

## 1.5 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 be 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 result in a compacted range between the upper and lower limits of agreement?

### **1.5.1 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.

## 1.5.2 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 measurements

## 1.6 Bland Altman Plot

Bland Altman have recommended the use of graphical techniques to assess agreement. Principally their method is calculating , for each pair of corresponding two methods of measurement of some underlying quantity, with no replicate measurements, the difference and mean. Differences are then plotted against the mean.

Hopkins argued that the bias in a subsequent Bland-Altman plot was due, in part, to using least-squares regression at the calibration phase.

### 1.6.1 Bland Altman plots using 'Gold Standard' raters

According to Bland and Altman, one should use the methodology previous outlined, even when one of the raters is a Gold Standard.

### **1.6.2 Bias Detection**

further to this method, the presence of constant bias may be indicated if the average value differences is not equal to zero. Bland and Altman does, however, indicate the indication of absence of bias does not provide sufficient information to allow a judgement as to whether or not one method can be substituted for another.

# Bibliography

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