A practical approach to Bland-Altman plots and variation coefficients for log transformed variables

A.M. Euser
F.W. Dekker
S. le Cessie

J Clin Epidemiol 2008; 61: 978-982
Chapter 5

Abstract

Objective
Indicators of reproducibility for log-transformed variables can often not be calculated straightforwardly and are subsequently incorrectly interpreted.

Methods and Results
We discuss meaningful Coefficients of Variation (CV) for log-transformed variables, which can be derived directly from the standard error of the log-transformed measurements. To provide easy interpretable Bland and Altman plots, we calculated limits of inter and intraobserver agreement (LA) for log-transformed variables and transform them back to the original scale. These LAs for agreement are subsequently plotted on the original scale in a conventional Bland and Altman plot. Both approaches were illustrated in a clinical example on the reproducibility of skinfold thickness measurements.

Conclusion
In reproducibility, it is important to calculate meaningful CVs, LAs, and Bland–Altman plots for log-transformed variables. We provide a practical approach in which existing statistical methods are applied in the field of reproducibility, thus leading to parameters of reproducibility which can be interpreted on the original scale.
Introduction

Reproducibility can be described as the repeatability of measurements in time or by different observers.1,2 Several indicators of reproducibility are applied in literature, with Intraclass Correlation Coefficients (ICC), Coefficients of Variation (CV), and limits of agreement (LA) being most frequently used. An ICC is a relative measurement of reliability, in which variation due to measurement error is compared with the variation between subjects. In a CV, reliability is expressed as the variation between measurements in relation to the mean value of all measurements. In contrast, LAs provide direct information about the absolute measurement error, which is plotted against the mean of the two measurements in a Bland and Altman Plot. This agreement forms an important measurement property by itself.3

Before reproducibility can be determined, data are frequently log transformed to approximate normality because of a skewed distribution of errors. Although ICCs after log transformation can still be calculated straight forward by estimating variance components on the log transformed data, a problem arises with the calculation and interpretation of other indicators of reproducibility, both for reliability and agreement measurements. CVs calculated in the conventional way have no natural interpretation anymore when estimated on a log-transformed scale without an actual zero. Bland and Altman describe the calculation of limits of agreement on log-transformed data.3 However, the advantage of the Bland and Altman plot as an easy interpretable indicator of reproducibility expressed in the absolute units of measurement used in the clinical situation doesn’t apply anymore.

In this article we discuss methods to calculate meaningful and interpretable CVs and LAs on log-transformed data, applied in an example from a clinical study on the reproducibility of skinfold thickness measurements.

Methods

Design and notations
In studies on reproducibility, one usually has observers (or instruments) measuring subjects. For simplicity, we start by assuming that each observer measures each subject once and focus on interobserver variability. Later on, we also consider the situation when more measurements are taken, which can be used to assess intraobserver variability. We denote the clinical measurement of interest by $X$ and write $X_{ij}$ for the measurement of the $j$th observer made on the $i$th subject ($i=1,\ldots,I$, $j=1,\ldots,J$).
In this article, we consider the situation that the distribution of $X$ is skewed and that a log transformation is performed to obtain an approximately normal distribution. The log-transformed measurement is denoted by $Z$. Although natural logarithms are mathematically more convenient, we consider here the 10-log transformation, because this is the transformation most frequently used in the applied field: $Z = 10\log (X)$.

**Random effect models**

Linear random effect models are often used to analyze this kind of data. Here we assume that the log-transformed variable $Z$ follows the following linear random effect model:

$$Z_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \ [\text{Model a}]$$

where $\mu$ is some fixed parameter, and where $\alpha_i$, the subject effect, $\beta_j$, the observer effect, and $\varepsilon_{ij}$ are independent random effects, normally distributed with mean 0 and with between-subject variance $\sigma_\alpha^2$, interobserver variance $\sigma_\beta^2$, and error variance $\sigma_\varepsilon^2$, respectively. In studying interobserver reproducibility, the interobserver variance $\sigma_\beta^2$ and error variance $\sigma_\varepsilon^2$ are expressed in relationship to the between-subject variance $\sigma_\alpha^2$.

**Coefficient of variation**

The coefficient of variation expresses the standard deviation as a percentage of the mean. It is a relative, unit-free measure, but it has only a useful interpretation if the measurement scale is positive, with value 0 the minimum value. For example, a CV for height or weight has a clear interpretation, but a CV for temperature measured in degrees Celsius or Fahrenheit not, because temperatures can be negative and the value 0 is not an absolute minimum.

When assessing the interobserver reproducibility, one would like to relate the mean of the observations to the spread of the measurements from different observers on the same subject. This measurement error between observations equals $\sqrt{\frac{\sigma_\beta^2 + \sigma_\varepsilon^2}{Z^2}}$, and is sometimes called the agreement standard error of the measurement: $\text{SEM}_{\text{agreement}}$. In the linear random effects model, the interobserver CV would be calculated by $100\% \times \sqrt{\frac{\sigma_\beta^2 + \sigma_\varepsilon^2}{Z^2}}$, with $Z$ the sample mean of measurements of $Z$ and $\sigma$ indicates the estimate of $\sigma$. However, it makes no sense to calculate this CV for $Z$, since on the log scale, 0 is no absolute minimum. Values of $X$ smaller than 1 correspond to values of $Z$ smaller than 0, and it is well possible that $Z$ is 0 or even negative.

Therefore, CV should be defined on the original scale. It can easily be shown, using Taylor expansion, that the standard deviation of a naturally log-transformed variable is approximately equal to the CV on the original scale. Therefore the $\text{SEM}_{\text{agreement}}$ of the natural-log transformed measurement is quite commonly used as interobserver CV on the original scale. Here, we
use as interobserver coefficient of variation, \( CV_{\text{inter}} = 100\% \times \ln(10) \sqrt{\left(\frac{\sigma_0^2}{M} + \sigma_1^2\right)} \), where \( \sqrt{\left(\frac{\sigma_0^2}{M} + \sigma_1^2\right)} \) is the spread of the log-transformed measurements from different observers on the same subject. The value \( \ln(10) \) is needed since we consider 10-log transformations. Bland and Altman\(^4\) suggest a different CV for log-transformed variables. There are no strong arguments in favor of their way of calculating CVs and the two approaches will yield very similar results when the CV is not large.

**Limits of agreement and Bland–Altman plots**

Assessing agreement between two observers or measurement methods can be done by using Bland and Altman plots and calculating limits of agreement. In a Bland and Altman plot, the difference between the two measurements per subject is plotted against the mean of the two measurements. In our situation, we have random observers, and assume that the mean difference between two arbitrary observers is 0. The limits of agreement are then defined as \(-1.96 \times s\) and \(+1.96 \times s\), with \( s \) the observed standard deviation of the difference between the two measurements per subject. If the spread of the differences increases with increasing mean of the observations, the Bland Altman plot and limits of agreement should be calculated on a log scale. This is straightforward to do, but it is difficult to interpret log-transformed variables in clinical practice.

We transformed these limits of agreement back to the original scale by taking anti-logs. This yields an interval for the ratio between two measurements. If the limits of agreement for \( Z = \log_{10}(X) \) are between \(-a\) and \(a\), with \( a = 1.96 \times s\), this implies that the ratio between two measures on the original scale \( X_i/X_j \) is between \(10^{-a}\) and \(10^{a}\). Then, for a given value for \( X \), it can be shown that \(X_i - X_j\) is between \(-2X(10^a - 1)/(10^a + 1)\) and \(2X(10^a - 1)/(10^a + 1)\). Although a ratio of measures is still difficult to conceptualize, these LAs on the original scale can be plotted in a conventional Bland and Altman plot of \( X \) to clearly visualize the reproducibility of the measurement for each different value of \( X \).

So far, we considered only two observers. Rousson et al. extended the definition of limits of agreements to several observers by: \( LA_{\text{inter}} = 0 \pm 1.96 \times \sqrt{2(\sigma_0^2 + \sigma_1^2)} \). This upper limit of \( 1.96 \times \sqrt{2(\sigma_0^2 + \sigma_1^2)} \) is also called the smallest detectable change,\(^2\) that is, the smallest change in measurement, which is unlikely to occur by differences between observers. In the same way as described previously, the limits of agreement can be calculated for the log transformed variable \( Z \) and transformed back to the original scale. A Bland and Altman plot on the original scale of \( X \) can then be made by drawing these back transformed limits of agreement as function of the mean of \( X \). An impression of the distribution of the individual data can be obtained by considering all possible pairs of observers and plotting the difference between the measurements per observer pair on the same subject versus the mean of the measurements of an observer pair on this subject.
Intraobserver reproducibility

To assess the intraobserver reproducibility, an observer has to measure a subject more than one time. Let \( Z_{ijk} \) be the \( k \)th measurement of observer \( i \) on subject \( j \). Model \([a]\) as described above can be extended to:

\[
Z_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \quad \text{[Model b].}
\]

The extra random term \( \gamma_{ij} \) models interaction between observer and subject and is assumed to be normally distributed, with mean 0 and variance \( \sigma^2_{\gamma} \). The residual error term \( \epsilon_{ijk} \) with variance \( \sigma^2_{\epsilon} \) indicates the random error occurring within measurements made by one observer on one subject. For good intraobserver reproducibility this within-subject-observer variation should be as small as possible.

Following the same reasoning as described for the interreproducibility measures, for the IntraCoefficient of Variation of \( X \), the \( CV_{intra} = 100\% \times \ln(10) \hat{\Delta}_{in} \) can be used.

The intra-observer limits of agreement on the log scale are: \( LA_{intra} = 0 \pm 1.96 \cdot \sqrt{2 \sigma^2_{\epsilon}} \), and can be transformed back to limits of agreement for the difference of two measurements made by the same observer on the same subject \( X_{ij1} - X_{ij2} \) being equal to \(-2\bar{X}(10^a - 1)/(10^a + 1)\) and \(2\bar{X}(10^a - 1)/(10^a + 1)\), where \( a = 1.96 \cdot \sqrt{2 \sigma^2_{\epsilon}} \).

Clinical example

To demonstrate the advantages of the methods described above, especially for the Bland and Altman plots, we will show data from a clinical study on the reproducibility of skinfold thickness measurements in young adults. In this study, skinfold thickness measurements at four locations (triceps, biceps, subscapular, and iliacal) were taken in duplicate on four subjects by 13 observers. Every subject was measured in duplicate at the four skinfold locations by all 13 observers. In the estimation of interobserver reproducibility, the mean of the two measurements by one observer was taken for every skinfold location. The objectives and methods of this study are described in detail elsewhere.6,7 In this example, we take the data from the biceps skinfold measurement. Indicators of reproducibility and variance components and of the biceps skinfold measurement are displayed in Table 1.

Coefficient of variation

At first glance, one should be tempted to apply the normal formula for calculating an interobserver CV on the log-transformed data and thus divide \( \sqrt{\sigma^2_{\alpha} + \sigma^2_{\epsilon}} = \sqrt{(12.9 \cdot 10^{-1} + 7.142 \cdot 10^{-1})} = 0.142 \) by the mean log-transformed biceps skinfold measurement, which is 1.14 (see Table 1).
Table 1. Indicators of reproducibility of the biceps skinfold measurement

### Biceps skinfold

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Value (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean and range of all measurements on original scale (mm)</td>
<td>14.5 (range 7.3–29.0)</td>
</tr>
<tr>
<td>Mean and range of all measurements of 10log transformed variable</td>
<td>1.14 (range 0.86–1.46)</td>
</tr>
</tbody>
</table>

### Variance Components of 10log transformed variable

<table>
<thead>
<tr>
<th>Component</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersubject variance $\sigma_i^2$</td>
<td>2.904$^{-3}$</td>
</tr>
<tr>
<td>Interobserver variance $\sigma_{ij}^2$</td>
<td>12.90$^{-3}$</td>
</tr>
<tr>
<td>Error-variance* $\sigma_i^2$</td>
<td>7.142$^{-3}$</td>
</tr>
<tr>
<td>Observer-subject variance $\sigma_{os}^2$</td>
<td>6.592$^{-3}$</td>
</tr>
<tr>
<td>Residual-error variance $\sigma_{ir}^2$</td>
<td>1.099$^{-3}$</td>
</tr>
</tbody>
</table>

### Limits of agreement

<table>
<thead>
<tr>
<th>Type</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intraobserver LA of 10log biceps</td>
<td>-0.092 to 0.092</td>
</tr>
<tr>
<td>Intraobserver LA of ratio of two biceps measurements</td>
<td>0.809 to 1.235</td>
</tr>
<tr>
<td>Intraobserver LA of difference of two biceps measurements as function of the mean $\bar{X}$</td>
<td>-0.21$\bar{X}$ to 0.21$\bar{X}$</td>
</tr>
<tr>
<td>Interobserver LA of 10log biceps</td>
<td>-0.392 to 0.392</td>
</tr>
<tr>
<td>Interobserver LA of ratio of two biceps measurements</td>
<td>0.400 to 2.499</td>
</tr>
<tr>
<td>Interobserver LA of difference of two biceps measurements as function of the mean $\bar{X}$</td>
<td>-0.85$\bar{X}$ to 0.85$\bar{X}$</td>
</tr>
</tbody>
</table>

### Coefficients of variation

<table>
<thead>
<tr>
<th>Type</th>
<th>Value (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intraobserver CV</td>
<td>7.6%</td>
</tr>
<tr>
<td>Interobserver CV</td>
<td>33.1%</td>
</tr>
</tbody>
</table>

* $a^{-3}$ was written as $a^3$.
* In this example $\sigma_L = \sigma_{os}^2 + \sigma_{ir}^2/2$, since each observer measured a subject twice.
This yields an interobserver CV of 12.5%. The quantitative value of this CV might look attractive, but as explained above it is a completely meaningless value. Therefore, one should apply the formula for log-transformed data, which yields a CV of $100\% \ln(10) \times 0.142 = 33.1\%$. This is a true, meaningful value and indicates that the interobserver reproducibility of this skinfold is not that good compared with other literature on this topic.8

**Limits of agreement and Bland–Altman plots**

As can be clearly seen in the first conventional Bland and Altman plot, (Figure 1), the differences between the first and second measurement of the biceps skinfold by an observer are dependent of the skinfold thickness, with increasing intraobserver error with increasing

---

**Figure 1.** Conventional Bland and Altman plot. The differences between the first and second biceps skinfold measurement in relation to the mean of the two measurements of one observer on a subject. Lines are plotted indicating the limits of agreement (0 ± 1.96 S.D.).
Reliability and log transformations

thickness of the biceps skinfold. Therefore, the conventional LAs do not well represent any of the measurements. In the second Bland and Altman plot, (Figure 2), on log-transformed data, the spread of observations on the left hand side is comparable to the spread on the right hand side. The LAs plotted do fit better, although some skewness remains. However, these log-transformed values are difficult to interpret for use in clinical practice. The values on the x and y-axis can be anti-logged, yielding the same plot but with more interpretable axes: geometric means on the x-axis and the ratio of measurements on the y-axis. Still we prefer to study differences on the original scale and not ratios, because of their direct clinical interpretation. Therefore, in the third Bland and Altman plot, (Figure 3), we transformed the LAs (Table 1) back to the original scale using the methods described in

**Figure 2.** Bland and Altman plot of log-transformed data. The differences between the first and second 10 log biceps skinfold measurement in relation to the 10 log mean of the two measurements of one observer on a subject. Lines are plotted indicating the limits of agreement (0 ± 1.96 S.D.).
this article. We plotted these LAs into the conventional Bland–Altman plot on the original scale. This back transformation yields diagonal lines representing the intraobserver limits of agreement (formulas given in Table 1).

Note that the LAs for the differences are proportional to the mean. For example, a mean biceps value of 10 mm has limits of agreement between the measurements of two observers of −2.10 and 2.10 mm, whereas if the mean biceps value increases to 20 mm, the LAs increase to −4.21 and 4.21 mm (Figure 3).

**Figure 3.** Bland and Altman plot on the original scale with back transformed limits of agreement. The differences between the first and second biceps skinfold measurement in relation to the mean of the two measurements of one observer on a subject. Lines are plotted indicating the limits of agreement using the formulas in our paper.
Figure 4 shows the calculated interobserver LAs on the original scale for the difference between the measurements of two observers as function of the mean of the measurements of a pair of two observers on one subject. To illustrate the agreement between the 13 observers in our data set, we considered all possible pairs of observers and for each pair we plotted the differences between the measurements of biceps skinfold per subject versus the mean of the measurements. Whether the difference between two observers was positive or negative was decided arbitrarily, because there is no clear ordering of the observers. We should be careful not to overinterpret the observed patterns in this plot. Each observer here contributes
to 12 observer pairs. This explains the diagonal patterns of points in Figure 4 and results in downward trends for the smallest and largest mean values. Note that the downward trend for the observations with the largest mean values is caused by a small number of points and that the majority of the points is on the left side of the plot.

**Conclusion**

In conclusion, we have shown that correct and meaningful indicators of reproducibility can be estimated for log-transformed variables, which can be interpreted straightforwardly. As log transformations are frequently applied in reproducibility studies, it is important to use the correct formula in calculating a meaningful and interpretable CV and to provide easy interpretable Bland and Altman plots and LAs on the original scale to assess agreement. Apart from log transformations by which a scale without an absolute minimum arises, there are additional approaches to approximate normality in which an absolute minimum is preserved. For example, a square root transformation could normalize right skewed data while the null value remains zero. However, in clinical practice, log transformations are much more frequently applied. An important advantage of a log transformation is that differences on the logarithmic scale can be transformed back to ratios on the original scale, as shown in the calculated limits of agreement.

In Bland–Altman plots, it is also possible to express the difference between measurements as a percentage of the average of the measurements, as shown in an example by Dewitte et al. However, with this approach the advantage of a direct overview of the exact value of both the measurement error and the corresponding limits of agreement in one plot is lost. In this situation, the absolute measurement error must be calculated from the mentioned percentages and means.

The approach of Bland and Altman plots for log-transformed data with back transformed limits of agreement, we provide here has almost never shown in literature on reproducibility of clinical measurements, apart from Dewitte et al. who briefly mentioned this method to be used in clinical chemistry. Though the Bland and Altman plots obtained by this method might appear somewhat unconventional at first glance, they provide an easy and reliable tool to see the LAs for different values of the variable at once on a clinical relevant scale.

**Acknowledgments**

None of the authors had any financial or personal conflict of interest with respect to the material reported in this article.
References

6. Euser AM, Cessie le S, Finken MJJ, Wit JM, Dekker FW. Reliability studies can be designed more efficiently by using variance components estimates from different sources. J Clin Epidemiol 2007; 38:1010-1014.