Repeatability

Kevin O'Brien

September 9, 2015

1 Repeatability

A measurement method can be said to have a good level of repeatability if there is consistency in repeated measurements on the same subject using that method. Conversely, a method has poor repeatability if there is considerable variation in repeated measurements.

This is relevant to method comparison studies because the 'repeatabilities' of the two methods of measurement affects the level of agreement of those methods. Poor repeatability in one method would result in poor agreement. More so if there is poor repeatability in both methods.

The quality of repeatability is the ability of a measurement method to give consistent results for a particular subject. That is to say that a measurement will agree with prior and subsequent measurements of the same subject.

Repeatability is defined by the **?** as 'the closeness of agreement between independent results obtained with the same method on identical test material, under the same conditions (same operator, same apparatus, same laboratory and after short intervals of time)' and is determined by taking multiple measurements on a series of subjects.

Repeatability is important in the context of method comparison because the repeatability of two methods influence the amount of agreement which is possible between those methods. If one method have poor repeatability, then agreement with that method and another will necessarily be poor also. **?** and **?** highlight the importance of reporting repeatability in method comparison, because it measures the purest random error not influenced by any external factors. Statistical procedures on within-subject variances of two methods are equivalent to tests on their respective repeatability coefficients. A formal test is introduced by **?**, which will discussed in due course.

1.1 Relevance of Repeatability

Repeatability of two method limit the amount of agreement that is possible.If one method has poor repeatability, the agreement is bound to be poor. If both methods have poor repeatability, agreement is even worse.

The British standards Insitute [1979] define a coefficient of repeatability as *the value below which the difference between two single test results....may be expected*

to lie within a specified probability. Unless otherwise instructed, the probability is assumed to be 95%.

The Bland Altman method offers the analyst a measurement on the repeatability of the methods. The *Coefficient of Repeatability* (CR) can be calculated as 1.96 (or 2) times the standard deviations of the differences between the two measurements (d2 and d1).

1.2 Bland-Altman recommendations

? strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. **?** notes the lack of convenience in such calculations.

If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (**?**).

It is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors (**?**).

1.2.1 Add Ins

importance of repeatability 'cursiously replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked.

lack of repaeatability can interfere with the comparsion of two methods because if one methods has poor repeatability, in the sense that there is considerable variation in repeated measurements on the same subject, the agreement between two methods is bound to be porr.

? strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. **?** notes the lack of convenience in such calculations.

If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (**?**).

It is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors (**?**).

1.3 Repeatability and gold standards

Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to **?**, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer, which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability. Some authors, such as [cite] and [cite] have recognized this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a 'bronze standard'. Again, no formal definition of a 'bronze standard' exists.

The coefficient of repeatability may provide the basis of formulation a formal definition of a 'gold standard'. For example, by determining the ratio of CR to the sample mean $\bar{X}$. Further to [Lin], it is preferable to have a sample size specified in advance. A gold standard may be defined as the method with the lowest value

of $\lambda = CR/\bar{X}$ with $\lambda < 0.1\%$. Similarly, a silver standard may be defined as the method with the lowest value of λ with $0.1\% \leq \lambda < 1\%$. Such thresholds are solely for expository purposes.

1.4 Repeatability coefficient

The British standards Insitute[1979] define a coefficient of repeatability as *the value below which the difference between two single test results..may be expected to lie within a specified probability.*In the absence of other indications, the probability is 95%.

? introduces the repeatability coefficient for a method, which is defined as the upper limits of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances (**?**).

σ_x^2 is the within-subject variance of method x. The repeatability coefficient is $2.77\sigma_x$ (i.e. $1.96 \times \sqrt{2}\sigma_x$). For 95% of subjects, two replicated measurement by the same method will be within this repeatability coefficient.

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

2 Repeatability in Bland-Altman Blood Data Analysis

- Two readings by the same method will be within $1.96\sqrt{2}\sigma_w$ or $2.77\sigma_w$ for 95% of subjects. Thisvalue is called the repeatability coefficient.
- For observer J using the sphygmomanometer $\sigma_w = \sqrt{37.408} = 6.116$ and so the repeatability coefficient is $2:77 \times 6.116 = 16:95$ mmHg.
- For the machine S,$\sigma_w = \sqrt{83.141} = 9.118$ and the repeatability coefficient is $2:77 \times 9.118 = 25.27$ mmHg.
- Thus, the repeatability of the machine is 50% greater than that of the observer.

3 Carstensen

- The limits of agreement are not always the only issue of interest the assessment of method specific repeatability and reproducibility are of interest in their own right.
- Repeatability can only be assessed when replicate measurements by each method are available.
- The repeatability coefficient for a method is defined as the upper limits of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances.

- If the standard deviation of a measurement is σ the repeatability coefficient is $2 \times \sqrt{2}\sigma = 2.83 \times \sigma \approx 2.8\sigma$.
- The repeatability of measurement methods is calculated differently under the two models
- Under the model assuming exchangeable replicates (1), the repeatability is based only on the residual standard deviation, i.e. $2.8\sigma_m$
- Under the model for linked replicates (2) there are two possibilities depending on the circumstances.
- If the variation between replicates within item can be considered a part of the repeatability it will be $2.8\sqrt{\omega^2 + \sigma_m^2}$.
- However, if replicates are taken under substantially different circumstances, the variance component ω^2 may be considered irrelevant in the repeatability and one would therefore base the repeatability on the measurement errors alone, i.e. use $2.8\sigma_m$.

3.1 Repeatability

? strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. **?** notes the lack of convenience in such calculations.

If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (**?**).

It is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors (**?**).

4 Reproducibility

It is advisable to be able to distinguish between Repeatability and a similar concept Reproducibility. Reproducibility is

4.1 Repeatability

Repeatability (or *test-retest reliability*) describes the variation in measurements taken by a single method of measurement on the same item and under the same conditions. A less-than-perfect test-retest reliability causes test-retest variability. Such variability can be caused by, for example, intra-individual variability and intra-observer variability. A measurement may be said to be repeatable when this variation is smaller than some agreed limit.

Test-retest variability is practically used, for example, in medical monitoring of conditions. In these situations, there is often a predetermined "critical difference", and for differences in monitored values that are smaller than this critical difference, the possibility of pre-test variability as a sole cause of the difference may be considered in addition to, for examples, changes in diseases or treatments.

According to the Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, the following conditions need to be fulfilled in the establishment of repeatability:

- the same measurement procedure
- the same observer
- the same measuring instrument, used under the same conditions
- the same location
- repetition over a short period of time.

4.2 Bland and Altman 1999

As noted by Bland and Altman 1999, the repeatability of two methods of measurement can potentially limit Repeatability (using Bland-Altman plot) The Bland-Altman plot may also be used to assess a methods repeatability by comparing repeated measurements using one single measurement method on a sample of items. The plot can then also be used to check whether the variability or precision of a method is related to the size of the characteristic being measured. Since for the repeated measurements the same method is used, the mean difference should be zero. Therefore the Coefficient of Repeatability (CR) can be calculated as 1.96 (often rounded to 2) times the standard deviation of the case-wise differences.

4.3 Notes from BXC Book (chapter 9)

The assessment of method-specific repeatability and reproducibility is of interest in its own right. Repeatability and reproducibility can only be assessed when replicate measurements by each method are available. If replicate measurements by a method are available, it is simple to estimate the measurement error for a method, using a model with fixed effects for item, then taking the residual standard deviation as measurement error standard deviation. However, if replicates are linked, this may produce an estimate that biased upwards. The repeatability coefficient (or simply repeatability) for a method is defined as the upper limit of a prediction interval for the absolute difference between two measurements by the same method on the same item under identical circumstances (see above conditions)

$$y_{mir} = \alpha_m + \beta_m(\mu_i + a_{ir} + c_{mi}) + e_{mir}$$

The variation between measurements under identical circumstances.

5 Repeatability

5.1 Repeatability and gold standards

Currently the phrase 'gold standard' describes the most accurate method of measurement available. No other criteria are set out. Further to **?**, various gold standards have a varying levels of repeatability. Dunn cites the example of the sphygmomanometer, which is prone to measurement error. Consequently it can be said that a measurement method can be the 'gold standard', yet have poor repeatability. Some authors, such as [cite] and [cite] have recognized this problem. Hence, if the most accurate method is considered to have poor repeatability, it is referred to as a 'bronze standard'. Again, no formal definition of a 'bronze standard' exists.

The coefficient of repeatability may provide the basis of formulation a formal definition of a 'gold standard'. For example, by determining the ratio of CR to the sample mean $\bar{X}$. Further to [Lin], it is preferable to have a sample size specified in advance. A gold standard may be defined as the method with the lowest value of $\lambda = CR/\bar{X}$ with $\lambda < 0.1\%$. Similarly, a silver standard may be defined as the method with the lowest value of λ with $0.1\% \leq \lambda < 1\%$. Such thresholds are solely for expository purposes.