

# 1 Repeatability

Repeatability is the ability of a measurement method to give consistent results for a particular subject, i.e. a measurement will agree with prior and subsequent measurements of the same subject. Barnhart et al. (2007) emphasizes the importance of repeatability as part of an overall method comparison study, a view endorsed by Carstensen et al. (2008). Before there can be good agreement between two methods, a method must have good agreement with itself. If one method has poor repeatability in the sense of considerable variability, then agreement between two methods is bound to be poor (Roy, 2009b). Barnhart et al. (2007) remarks that it is important to report repeatability when assessing measurement, because it measures the purest form of random error not influenced by other factors, while further remarking ‘*curiously replicate measurements are rarely made in method comparison studies, so that an important aspect of comparability is often overlooked*’. Bland and Altman (1999) strongly recommends the simultaneous estimation of repeatability and agreement by collecting replicated data. However Roy (2009b) notes the lack of convenience in such calculations.

Repeatability is defined by the IUPAC (2009) 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.

A measurement is said to be repeatable when this variation is smaller than some pre-specified limit.

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:

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.

Repeatability is important in the context of method comparison because the repeatability of both of the measurement methods will necessarily influence the amount of agreement which is possible between those methods. If one method have poor repeatability, then agreement with that method with other other will necessarily be poor.

Lack of repeatability 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 poor. If there is poor repeatability in both methods, the lack of agreement would be even worse.

## The Coefficient of Repeatability

The British Standards Institute (1979) defines 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%.

Barnhart et al. (2007) and Roy (2009a) 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 Roy (2009a), which will be discussed in due course.

### 1.1 Coefficient of Repeatability

The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999). Once the within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.

## 1.2 Bland and Altman 1999

Bland and Altman (1999) remarks that their proposed 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.

## 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 Dunn (2002), 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 , 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  $\lambda$  with  $0.1\% \leq \lambda < 1\%$ . Such thresholds are solely for expository purposes.

## References

- 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 (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160.
- Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 4(1).
- Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.
- IUPAC (2009). IUPAC Compendium of Chemical Terminology - the Gold Book. <http://goldbook.iupac.org/R05293.html>.
- Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.
- Roy, A. (2009b). 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.