0.1 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.

0.2 The Conversion Problem

In this section, we will reconsider the conversion problem, where by the methods of measurements are denominated in different units. Conversion problems arise when the comparison is between two approximate methods of measurement each of which measures the quantity in different units.

This situation can arise when the methods in question proceed by measuring different proxies for the underlying quantity of interest. (lewis 1991)

For the single measurement case, the author can not foresee any scope for insights that are not already offered by using a structural relation model, as proposed by lewis et 1991, or error-in-variables regression. In the case of orthonormal regression, it is not reasonable to assume that both methods have equal measurement variance, when they are denominated in different units. The analyst may attempt to mitigate the problem by scaling the variance of one method, but even still problems remain. Similarly for Deming regression, no further insights on how to properly estimate the variance ratio can be offered.

0.3 Other Types of Studies

Lewis et al. (1991) categorize method comparison studies into three different types. The

key difference between the first two is whether or not a 'gold standard' method is used. In situations where one instrument or method is known to be 'accurate and precise', it is considered as the'gold standard' (Lewis et al., 1991). A method that is not considered to be a gold standard is referred to as an 'approximate method'. In calibration studies they are referred to a criterion methods and test methods respectively.

1. Calibration problems. The purpose is to establish a relationship between methods, one of which is an approximate method, the other a gold standard. The results of the approximate method can be mapped to a known probability distribution of the results of the gold standard (Lewis et al., 1991). (In such studies, the gold standard method and corresponding approximate method are generally referred to a criterion method and test method respectively.) Altman and Bland (1983) make clear that their methodology is not intended for calibration problems.

2. Comparison problems. When two approximate methods, that use the same units of measurement, are to be compared. This is the case which the Bland-Altman methodology is specfically intended for, and therefore it is the most relevant of the three.

3. Conversion problems. When two approximate methods, that use different units of measurement, are to be compared. This situation would arise when the measurement methods use 'different proxies', i.e different mechanisms of measurement. Lewis et al. (1991) deals specifically with this issue. In the context of this study, it is the least relevant of the three.

Dunn (2002, p.47) cautions that 'gold standards' should not be assumed to be error free. 'It is of necessity a subjective decision when we come to decide that a particular method or instrument can be treated as if it was a gold standard'. The clinician gold standard , the sphygmomanometer, is used as an example thereof. The sphyg-

momanometer 'leaves considerable room for improvement' (Dunn, 2002). **?** similarly addresses the issue of glod standards, 'well-established gold standard may itself be imprecise or even unreliable'.

The NIST F1 Caesium fountain atomic clock is considered to be the gold standard when measuring time, and is the primary time and frequency standard for the United States. The NIST F1 is accurate to within one second per 60 million years (NIST, 2009).

Measurements of the interior of the human body are, by definition, invasive medical procedures. The design of method must balance the need for accuracy of measurement with the well-being of the patient. This will inevitably lead to the measurement error as described by Dunn (2002). The magnetic resonance angiogram, used to measure internal anatomy, is considered to the gold standard for measuring aortic dissection. Medical test based upon the angiogram is reported to have a false positive reporting rate of 5% and a false negative reporting rate of 8%. This is reported as sensitivity of 95% and a specificity of 92% (ACR, 2008).

In literature they are, perhaps more accurately, referred to as 'fuzzy gold standards' (**?**). Consequently when one of the methods is essentially a fuzzy gold standard, as opposed to a 'true' gold standard, the comparison of the criterion and test methods should be consider in the context of a comparison study, as well as of a calibration study.

0.4 Lewis Conversion

While regarding a comparison of two pump meters under operational conditions

..It is suspected that the various assumptions made by each method are weak under operational conditions Lewis listed several sources of variation that relate to the

practical aspects of each measurement method.

There is little reasons to believe that the laboratory conditions of the devise provide a suitable basis for the conversion of data gathered under operational conditions.

0.4.1 Latent Variables

Latent variables are variables that are not measured (i.e. not observed) but whose values is observed from other observed variables. One advantage of using latent variables is that it reduces the dimensionality of data. A large number of observable variables can be aggregated in a model to represent an underlying concept, making it easier for humans to understand the data. [wikipedia]

0.4.2 Using LMEs for Conversion Problems

For the case of conversion problem with replicate measurements, a framework that incorporates the ideas offered by Roy (2009) can be proposed. Estimates for between-subject and within-subject variances may be sought. However Roy's tests on variability are no longer applicable, as one would not expect the method to have similar estimates. An estimate for the scaling factor β may be sought, where $Y_i \approx \beta X$.

$$X_i = \tau_i + \delta_i$$

$$Y_i = \alpha + \beta X \tau_i + \epsilon_i$$

We will simulate a data set based in lewis conversion problems, provide three replicates values for both measurements. To acheive this we add "jitter noise" to three copies of each original measurement.

Bibliography

ACR (2008). Acute Chest Pain (suspected aortic dissection) - American College of Radiology Expert Group Report.

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.

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

Lewis, P., P. Jones, J. Polak, and H. Tillitson (1991). The problem of conversion in method comparison studies. *Applied Statistics 40*, 105–112.

NIST (2009). Cesium fountain atomic clock : The primary time and frequency standard for the United States. http://tf.nist.gov/timefreq/cesium/fountain.htm.