Measurement System Analysis Book Six Sigma with R : Statistical Engineering for Process Improvement

Cano, Moguerza, Redchuk

The Table 5.1 of SixSigma book is noticeable in the data can paralleled with carstensens' formulation

- Voltmeter : Method
- Item : battery
- Replicate : Run
- y : voltage

```
http://onlinelibrary.wiley.com/doi/10.1002/uog.5256/pdf
```

```
http://link.springer.com/article/10.1023/A:1009982611386
```

Preiss (2000) : Limitations of BA

Correlation is a measure of association, but not agreement. However It still persists in literature.

The threshold for exchangability is a decision for tha analyst

Preiss and Fisher explore the limitations of the BA approach.

Summary : the outcome of BA analysis is dependent on the range of measurements such that narrow ranges will necessarily produce good results. This pitfall can be avoided by determining the probability thatvsuch observed agreements occur in pairing of unassociated measurements.

Standard BA analysis can then proceed without concerns about finding good agreement from unassociated measurements.

Chinn (1990) : Repeatability

0.1 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. **?** emphasizes the importance of repeatability as part of an overall method comparison study, a view endorsed by **?**. 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 (**?**). **?** 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.* **?** strongly recommends the simultaneous estimation of repeatability and agreement be collecting replicated data. However **?** notes the lack of convenience in such calculations. 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.

A measurement is said to be repeatable when this variation is smaller than some prespecified 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.

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%.

0.1.1 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.

? recognizes 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 $\bar{X}$. Advisably the sample size should 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.

Passing & Bablok (1983)

Passing & Bablok (1983) have described a linear regression procedure with no special assumptions regarding the distribution of the samples and the measurement errors. The result does not depend on the assignment of the methods (or instruments) to X and Y. The slope B and intercept A are calculated with their 95% confidence interval. These confidence intervals are used to determine whether there is only a chance difference between B and 1 and between A and 0.

Notes

- The Passing-Bablok procedure should only be used on variables that have a linear relationship and are highly correlated.

- Since it is a non-parametric procedure, Passing-Bablok regression is not influenced by the presence of one or relative few outliers.

- We advise to supplement the results of the Passing-Bablok procedure with a Bland-Altman plot. Literature

- Passing H, Bablok W (1983) A new biometrical procedure for testing the equality of measurements from two different analytical methods. Application of linear regression procedures for method comparison studies in Clinical Chemistry, Part I. J. Clin. Chem. Clin. Biochem. 21:709-720. [Abstract]

The very wide uptake of the limits of agreement approach has naturally been very pleasing. We have been aware, however, that sometimes the method has not been adopted with full understanding. For example, we have seen it suggested that two methods agree well because most of the observations lie within the 95% limits of agreement. The limits are calculated so that this will always be the case. (Bland & Altman 2002)

The 95% limits of agreement method has been widely cited and widely used, though many who cite it do not appear to have read the paper. (Bland & Altman 2003)

0.2 References

Problems with the Limits of Agreement method of comparing measurement systems Nathaniel T. Stevens, Stefan H. Steiner, R. Jock MacKay Business and Industrial Statistics Research Group Dept. of Statistics & Actuarial Science University of Waterloo

How do we compare measurement systems?

Standard plan: n subjects are measured once each by both measurement systems for a total of N = 2n measurements **Statistical model:**

$Y_{ij} = \alpha_j + \beta_j S_i + M_{ij}, i = 1, 2, \ldots, n, j = 1, 2$

$S_i \sin N(\mu, \sigma_s^2), M_{ij} \sim N(0, \sigma_j^2)$ independently

$-infty < \alpha_j < infty$ and $\beta_j > 0$ quantify the bias of measurement system j.

The goal Interchangeability $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$ and $\sigma_2^2\ leq \sigma_1^2$

Chinn

Repeatability and reference ranges for change

- A measurement that is totally unrepeatable clearly has no validity. Repeatability, however, or reproducibility, is an ambiguous concept without precise definition. To a clinical chemist it may mean reproducibility of results from an autoanalyser for a single blood sample. A technician may spend hours perfecting a non-automatic technique. But if two blood samples taken from the same subject within hours give very different results laboratory repeatability may be relatively unimportant.

- The repeatability of most respiratory measurements is necessarily of the ”time to time” type with the interval between measurements measured at least in minutes but possibly in hours or days. Differences in the results obtained depend on the time gap, the variation increasing-that is, repeatability decreasing-with the length of the gap. Neild 7t al ’ measured forced expiratory volume in one second (FEV,), peak expiratory flow (PEF), respiratory resistance, and specific airway conductance in 25 non-asthmatic subjects three times on each of three consecutive days, and reported the ”within subject within day” variance of each, and also the estimate of additional variance due to variation within subjects but between days.

- Variances were used in the paper by Neild et al’ because components of variance may be added. The usual measure of repeatability when only within day or only between day variation is studied is the within subject standard deviation. If more than two repeat measurements are carried out for some or all subjects then repeatability is most easily calculated from the results of a one way anaiysis of variance, with ”subjects” as the ”group” variable.

- The within subject standard deviation is the square root of the pooled within subject sum of squares divided by its degrees of freedom (that is, for those used to analysis of variance terminology, of the residual ”mean square”).

- Frequently repeatability studies are carried out with just two repeat measurements per subject; indeed, for most purposes this is the most efficient design. ’Fhen it is natural to take the differences between the first and second measurements for cach subject.

convert a standard deviation of differences, which has double the variance of a single FEV,, to a within subject standard deviation of a single FEV, we divide 0-42 by , 2 to get 0-30.

- Strictly speaking, the fact that the mean difference is not exactly zero but 0-01 should be taken into account.

  The within subject standard deviation is actually

  /I06(0 42)2 + 107(0.01)2

  107 x 2

  but the result is the same to 2 decimal places.

- The fact that either the standard deviation of the differences or the within subject standard deviation may be reported as a measure of repeatability may lead to confusion.

- 'I'he within subject standard deviation is referred to as the single determination standard deviation, and the 951991, p391) derived from it as the single determination 95% range.

- This will indicate the limits around a single measurement that must be regarded as possible values for the true measurement-that is, how much reliance, sav for diagnosis, can be placed on that reading.

If a patient is being monitored then we are interested in the change in values. 'lThe inherent variability, as measured by the single determination standard deviation or range, enters into both the initial and the subsequent measurement, and so the standard deviation and 95% range for change, are greater, by a factor of the square root of 2, than the single dletermination values. 'Fhe 95change can be calculated directly from the standard deviation of differences between the repeat measurements, or from the within subject standard deviation provided that the extra factor of the square root of 2 is remembered. It is recommended that the single determination standard deviation or 95for measuring repeatability as such and the 95% range for change for use in assessment of patients. it is, however, of prime importance that whichever is used is clearly stated, so that results will not be misinterpreted and can be converted to the aiternative form when this is needed.

Not all calculations are carried out on the original scale of measurement. Tihe third article in this series will explain now to choose the scale. Many mcasurements for example, PD,O in bronchiai challenge testing-are analysed on a log scale. All calculations should be carried out on the log values, but an arithmetic mean on the log scale can be antilogged to give the geometric mean value and the 95% range, expressed as + k, can be antilogged to x / antilog(k). Thus Chinn et al 2 reported a within subject standard deviation of log10 PD20 histamine as 0 27. The single determination 95or PD20 x / . 3-47 jimol. Such a range is usually expressed, however, in units of doubling doses, obtained by dividing 054 by logl0(2) (= 0.301), to give + 1-79 doubling doses. The possibility of an overall shift in mean value between the first and the second test can be investigated by calculating the standard error and 95% confidence interval for the mean difference. For the post-saline FEV, mentioned above the standard error of the mean difference was 0A42/,/107 = 0 04. The 95% confidence interval for the mean difference was thus from -0-07 to + 0-09 1. This tells us only the likely size of the bias between the first and the second occasion postsaline FEVy, and little about repeatability. A true confidence interval is not a measure of repeatability.

Method comparison

- One aspect in which methods can be compared is in their repeatability. For example, Oldham and Cole3 reported the repeatability of nine indices of FEV,, each calculated from five repeat blows.
- The most repeatable on the basis ofthe within subject standard deviation was the mean of all five, and the least repeatable the maximum of the first three.
- With most alternative methods there is a possibility that they do not, on average, give the same answer.

- Bland and Altman4 described in detail how to compare two methods on the same scale of measurement, illustrating their recommendations with PEF data.

- They give several reasons why the correlation coefficient should not be used, the most important of which are, firstly, that to agree perfectly the results from two measurements must lie on the line of identity, not just any straight line, and, secondly, that the correlation coefficient is influenced by the range ofvariation between the subjects, blood samples, or other units chosen to test the two methods.

- For a given level of agreement the correlation coefficient increases as the variance between the units increases. Bland and Altman4 recommended plotting the difference in the two results against the mean value from the two methods.

- Provided that there is no relation between differences and mean values, "limits of agreement" can be calculated as d + 2s, where d is the mean difference and s the standard deviation of the differences. If the sample size, n, is less than 100, 2 should be replaced by $t, 1, 0.05n + 1)/n$.

- Agreement, or more accurately lack of it, thus has two components, the relative bias as estimated by d and the random variation as estimated by s, which is at least as great as that predicted by the repeatability of each method.
- If the within subject (or between replicate if more appropriate) standard deviations are s, and s, then 2 s /S21+ S2 A 95can be calculated as for a paired t test-that is, d + tn - 1.o0.05o1;-
- As d estimates only the systematic component of lack of agreement, this is not a measure of agreement. As with repeatability, care must be taken to distinguish between confidence interval and the 95

- Measurements on different scales Bland and Altman4 dealt only with measurements on the same scale. We cannot compare repeatability as measured by standard deviations in different units, as may be required in comparing repeatability of different indices of histamine challenge tests6-for example, PC20 in log (mg/ml) and the slope of the FEV, dose-response curve in l(mg/ml)-'.

- The solution is to calculate a dimensionless statistic. Although dividing the standard deviation by the mean to give the coefficient of variation is still used, it is valid only in certain circumstances7 (to be described in the third article in this series).

- It is better to calculate the ratio of between subject to total variation, known as the intraclass correlation coefficient, as used by Dehaut et al.6

- The maximum value of the intraclass correlation coefficient is 1, achieved only when repeatability is perfect. A value of zero (or less) denotes repeatability that is no better (or worse) than would be expected by chance. To be useful a measurement should have an intraclass correlation coefficient of at least 0 6.

- Baseline FEV, measured on two occasions 1-14 days apart2 in 1 11 subjects had an intraclass correlation coefficient of 0-88. Repeated measurements of FEV, on the same day may give a value as high as 0Q99.3 In the simplest case of two components of variation, between subject (or other unit) and within subject, estimating the two components as described by Armitage and Berry8 is straightforward.

- A statistician should be consulted before data collection, however, if there are more than two components. The components of variance are of direct use-for example, the effect of averaging three or five measurements can be compared.'

- The use of the intraclass correlation coefficient implies that each component of variance has been estimated appropriately, from sufficient data (at least 25 degrees of freedom) and from a sample representing the population to which the results will be applied. When intraclass correlation coefficients are compared they should be obtained from data on the same sample of subjects, or from samples from the same population.

- With two methods on different scales there is no "line of identity" on which the data should lie for perfect agreement. Indeed, there is no

Mantha et al. (2000) and Dewitte et al. (2002)

- Repeatability not assessed, incorrectly calculated limits of agreement, wrong axes on difference plots, ignoring relationships between differences and averages, no CIs for estimates

- In less than 10% of the articles did the authors actually define a clinically acceptable difference and compare their limits of agreement to it

A literature review of anesthesia journals revealed several inadequacies and inconsistencies in statistical reports of results of comparison studies with regard to interchangeability of measure- ment methods. We encourage journal editors to evaluate submissions on this subject carefully to ensure that their readers can draw valid conclusions about the value of new technologies.