Comparison of measurement systems

Drs. Jan Willem Bikker PDEng

Introduction of the comparison problem

In DfSS and six sigma, the gage R&R study is a commonly known investigation of a measurement system. It studies how different the outcomes are when you take multiple measurements of the same thing, split over immediate repetitions and operator inferences.

Here, we deal with another situation around measurement systems that is not covered by the gage r&R study: the comparisons of measurement systems. Examples are:

- Does the new measurement device do the same thing as the existing one?
- Does a quick-and-dirty protocol give very different results from the standard one?
- Can an entirely different measurement principle be used instead of the standard method?
- Is the prototype new measurement system comparable to the widely accepted standard?

In the medical disciplines, the paper of Bland and Altman from 1986 is widely known [1]. It describes the analysis of a test, and some pitfalls. The points presented here can be found largely in that article, which is readily found in internet search engines using the search terms “bland altman pdf”. Also see the Wikipedia page [3].

Although the Bland-Altman article was written with biostatistics in mind, its lessons are applicable in any technical environment.

Design of a test

We describe the simplest case. A number of objects (products, patients, cases) are measured in pairs by the new measurement system X and the reference Y. In the simplest case, the cases should be independent. For instance, when you have 5 cases per patient over time, and several patients, the data come in groups and are not independent.

You should take care that the sample of cases cover the range of situations of interest. For instance, in medical settings, your sample should include people from all target ages for which you wish to compare the measurement systems. And when it measures the heart rate, should it be only when sitting or lying, or also when running?

More elaborate versions of the test may include short-term repeats of the paired measurements, which enables to estimate the repeatability of systems X and Y; this is described in [1].

Analysis

As always, a good analysis uses graphs of the data. A basic plot shows the Y vs X. Figure 1 shows a few examples.

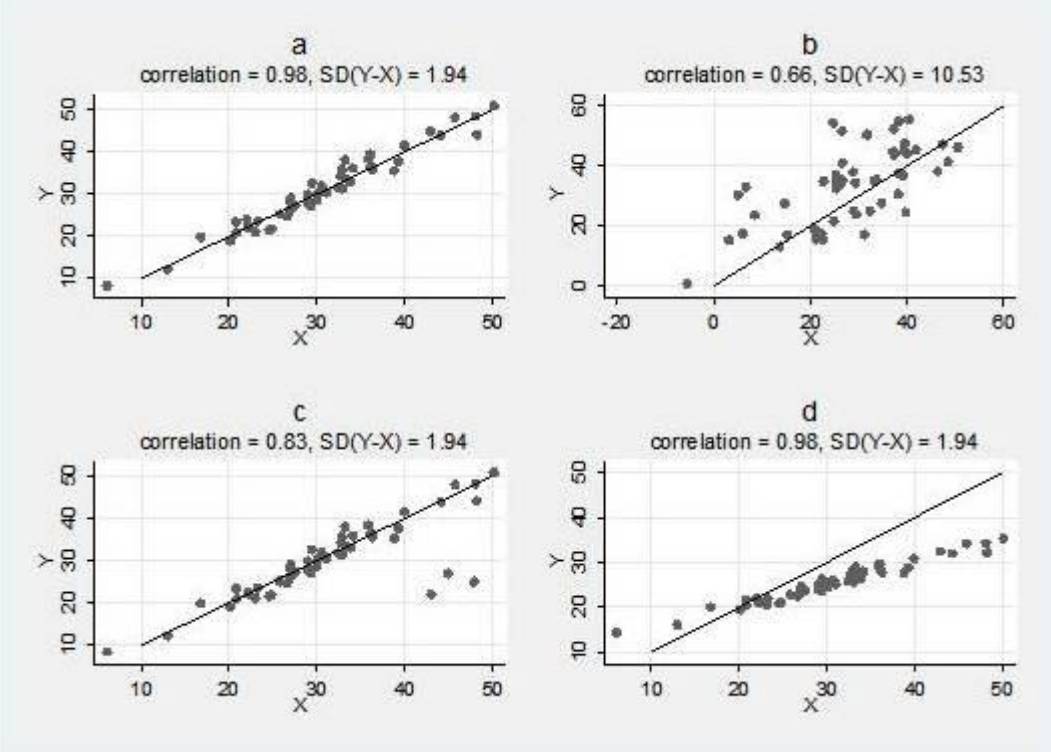

Figure 1 - Y vs X. a: good agreement, b: bad agreement, c: some outliers calling for the question whether they are due to system X or Y, and d: high correlation between X and Y but poor agreement.

As the article from Bland and Altman [1] argue, it is tempting to use the correlation coefficient to describe the degree of agreement of measurement systems X and Y, but this is wrong. The points should be near the diagonal line “Y=X”, and you can have a high correlation even if this is not the case. For instance, the cloud of points could be shifted or tilted, as in Figure 1d. The correlation coefficient may accompany the graph as a piece of extra information though.

A better way to express the agreement between the measurement systems is uses the bandwidth of the cloud of points and how far it is away from the line Y=X.

The Bland-Altman plot

In the article, the plot of the difference vs the average of X and Y is proposed. The average may be regarded as a version that is probably close to the “true” value of the measured quantity. Ideally,

- The mean of the difference is close to 0
- The spread is small
- The differences are not related to the average.[1]

Consider the case where X and Y measure the same object with the same measurement spread (no trends). The BA article points out that the difference Y-X will have a correlation to X (and Y), which is the reason the difference is not plotted against X or Y, but the average. However, although the article does not mention it, there will even be a slight correlation between Y-X and (X+Y)/2 in case the measurement spreads are different; see [2]. Also, in case the reference can be regarded as having very small measurement spread, it would be safe to plot the difference vs the reference.

1

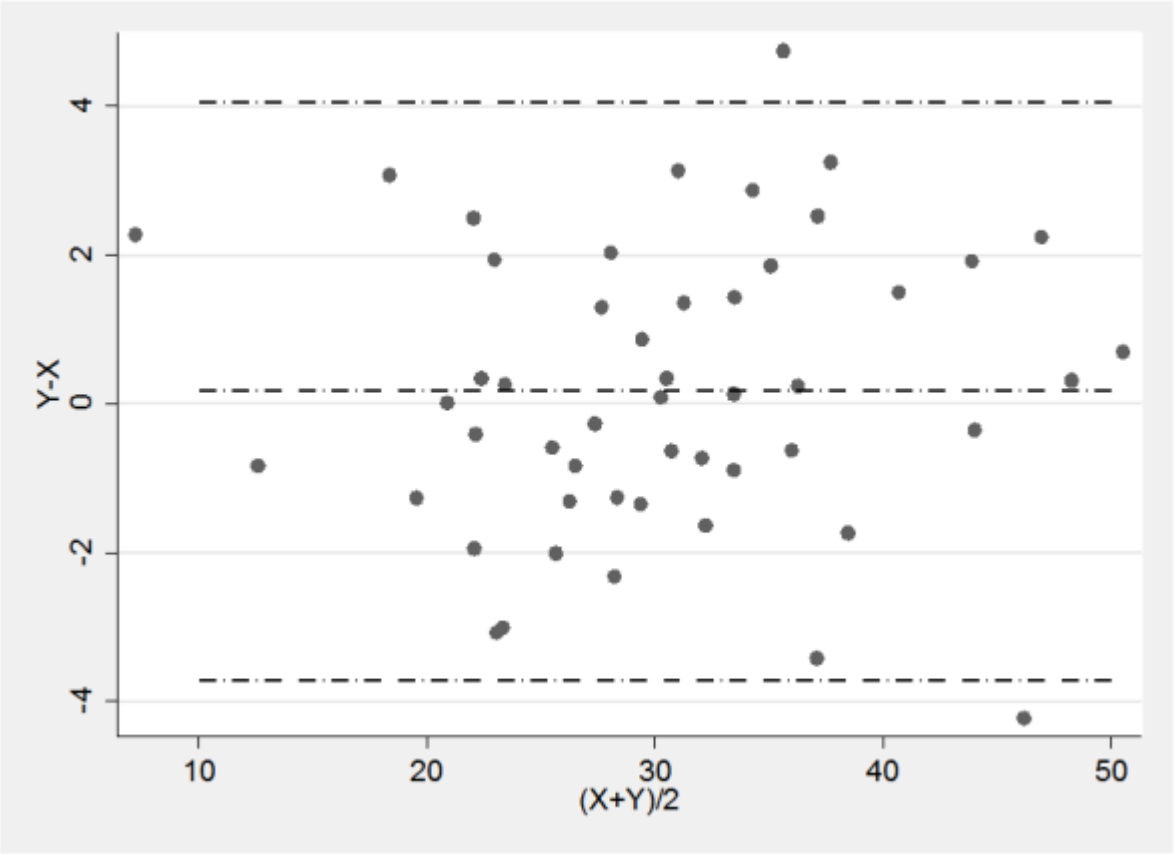

Figure 2 - Bland Altman plot for Figure 1a, with lines for the mean difference and limits of agreement

In the plot, three horizontal lines are given: the mean, and the two limits of agreement, which are equal to the mean +- 2 standard deviations. We expect that about 95% of the differences of all future cases would be between these limits of agreements. For that reason, the limits of agreement are a good measure to express the comparison of the measurement systems.

In this example, we have 50 points, the mean of the difference is 0.17, the standard deviation 1.94, and the limits of agreement are at -3.71 and 4.05.

Precision of the measurement system

So far, we described a simple analysis, which does not express the uncertainty due to how small or large the number of cases really is. We use the standard way of expressing such uncertainty by 95% confidence intervals.

Consider the vertical coordinates in the Bland Altman plot (**Error! Reference source not found.**), the differences D=Y-X. The mean of D, $\overline{D}$, has as standard error $\sqrt{S^2/n}$ where S is the estimated standard deviation and n the number of cases. A 95% confidence interval of the mean is given by its estimate plus or minus two times the standard error: $\overline{D} \pm 2\sqrt{S^2/n}$.

The limits of agreement, $\overline{D} + 2S$ and $\overline{D} - 2S$, have larger uncertainty. In fact, the approximation is easy to interpret: $\overline{D} + 2S$ has three elements of uncertainty, one $\overline{D}$ and two S-s. The standard error is $\sqrt{3}$ times as large, and the upper limit of agreement has 95% confidence interval:

$$\overline{D} + 2S \pm 2\sqrt{3S^2/n}$$
and
$$\overline{D} - 2S \pm 2\sqrt{3S^2/n}.$$

In this example, $\overline{D} = 0.17,\ n = 50, S = 1.94$ so the half-width of the confidence intervals evaluate to $\pm 2\sqrt{3 \cdot 1.94^2/50} = \pm 2\sqrt{0.2256} = \pm 0.95$.

The limits of agreement have confidence intervals of -3.71 ± 0.95 and 4.05 ± 0.95.

Tolerance intervals, k-statistics, prediction intervals

Prediction intervals are often given in regression problems. The simplest form for a normal distribution with known parameters is that 95% of future observations will lie within the interval $\mu \pm 1.96 \cdot \sigma$. In the usual case where the parameters are estimated, the 1.96 needs to be larger.

Tolerance intervals are intervals that contain at least e.g. 99% of the population with a given level of confidence, e.g. 95%. In a sense, the outer ends of the confidence intervals of the limits of agreement (i.e., $\overline{D} - 2S - 2\sqrt{3S^2/n}$ to $\overline{D} + 2S + 2\sqrt{3S^2/n}$), intend to express the same idea. For instance, for n=30, with 95% confidence, 95% of future observations will lie within the interval $m \pm 2.555 \cdot S$.

Where tolerance intervals contain a fraction of the population *with a certain confidence*, prediction intervals do the same *by point estimate.* In fact, a prediction interval is a tolerance interval with 50% confidence (the actual proportion is expected to be the required one, but with 50% chance too low and 50% chance too high).

This uses the method for normal distributions; there is also a calculation variant based on percentiles of the sample. The last is explained in a dfss.nl article “Sample size done differently” [4].

The so-called k-factor is the k in $m \pm k \cdot S$; so $k = 2.555$ in the example above. The k-factor is sometimes used in process capability studies (of manufacturing processes) to incorporate the sample uncertainty into the conclusions.

Designing a test for claiming similarity

How can you use these formulas? Suppose the goal in a project is to prove that the prototype Y and reference X will give very similar measurement values, at least within 30 units. The true agreement should be considerable better than ± 30. If the entire 95% confidence intervals of the upper and lower limits of agreement fit within ± 30, we have proof that the measurement systems are similar.

If you have an a-priori idea of how large the standard deviation of the differences will be (the S in the formulas), the you can choose the size of the test (n cases) from the width of the confidence interval, $\pm 2\sqrt{3S^2/n}$. For instance, if you expect $S = 2$, and wish to know the limits up to ± 0.8, you need to solve for $0.8 = 2\sqrt{3 \cdot 2^2/n}$ which gives $n = 75$. This corresponds to the earlier example, which had fewer points ($n = 50$), a similar S, and a somewhat larger uncertainty (± 0.95).

Conclusion

A comparison of measurement systems may came up in a wide range of settings and disciplines. The text above describes the basics from the Bland Altman article teaches, which is a classic in certain medical disciplines.

Two type of plots are proposed. The correlation may be tempting to use as a measure of agreement, but it is better to focus on the difference in outcomes when you measure the same thing using the two systems. Here, the mean plus or minus two standard deviations of the differences are taken as definition for the “limits of agreement”.

These limits have relatively simple expressions for their uncertainty, using 95% confidence intervals.

References

[1] Bland, J. M., & Altman, D. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. The lancet, 327(8476), 307-310.

[2] Tu, Y. K., & Gilthorpe, M. S. (2007). Revisiting the relation between change and initial value: a review and evaluation. Statistics in medicine, 26(2), 443.

[3] https://en.wikipedia.org/wiki/Bland%E2%80%93Altman_plot

[4] http://www.dfss.nl/sample-size-done-differently/

For more information, please contact CQM
T +31 40 750 23 23
info@cqm.nl