0.1 Alternative agreement indices

As an alternative to limits of agreement, Lin et al. (2002) proposes the use of the mean square deviation is assessing agreement. The mean square deviation is defined as the expectation of the squared differences of two readings.

The MSD is usually used for the case of two measurement methods X and Y , each making one measurement for the same subject, and is given by

$$MSDxy = E[(x-y)^2] = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2 + 2\sigma_x\sigma_y(1-\rho_{xy}).$$

Barnhart et al. (2007) advises the use of a predetermined upper limit for the MSD value, MSD_{ul}, to define satisfactory agreement. However, a satisfactory upper limit may not be properly determinable, thus creating a drawback to this methodology.

Barnhart et al. (2007) proposes both the use of the square root of the MSD or the expected absolute difference (EAD) as an alternative agreement indices. Both of these indices can be interpreted intuitively, being denominated in the same units of measurements as the original measurements. Also they can be compare to the maximum acceptable absolute difference between two methods of measurement d_0.

$$EAD = E(|x-y|) = \frac{\sum |x_i - y_i|}{n}$$

The EAD can be used to supplement the inter-method bias in an initial comparison study, as the EAD is informative as a measure of dispersion, is easy to calculate and requires no distributional assumptions.

Barnhart et al. (2007) remarks that a comparison of EAD and MSD , using simulation studies, would be interesting, while further adding that '*It will be of interest to investigate the benefits of these possible new unscaled agreement indices*'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding

Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the EAD penalizes the comparison for having a greater variance of differences. Hence the EAD values for both comparisons are much closer.

	F vs C	F vs T
Inter-method bias	-0.61	0.12 3
Difference variances	0.06	0.22
Limits of agreement	(-1.08, -0.13)	(-0.81,1.04)
EAD	0.61	0.35

Table 1: Agreement indices for Grubbs' data comparisons.

Further to Lin (2000) and Lin et al. (2002), individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If d_0 is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than d_0 can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{1}$$

If π_0 is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is π_0 may be determined. This boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

0.2 Alternative Agreement Indices

Alternative indices, proposed by Barnhart et al. (2007), are the square root of the MSD and the expected absolute difference (EAD).

$$EAD = E(|x - y|) = \frac{\sum |x_i - y_i|}{n}$$

The EAD can be used to supplement the inter-method bias in an initial comparison study, as the EAD is informative as a measure of dispersion, is easy to calculate and requires no distributional assumptions. A consequence of using absolute differences is that high variances would result in a higher EAD value.

	X	Y	U	V
1	101.83	102.52	98.05	99.53
2	101.68	102.69	99.17	96.53
3	97.89	99.01	100.31	97.55
4	98.15	99.57	100.35	96.03
5	99.94	100.85	99.51	99.00
6	98.85	98.86	98.50	100.76
7	99.86	97.85	100.66	99.37
8	101.57	100.21	99.66	108.87
9	100.12	99.85	99.70	105.16
10	99.49	98.77	101.55	94.31

```
Differences  2.5% limit 97.5% limit    SD(diff)
-0.08078844 -2.39471014  2.23313327  1.15696085
```

	X	Y	$X - Y$	$\|X - Y\|$
1	98.05	99.53	-1.49	1.49
2	99.17	96.53	2.64	2.64
3	100.31	97.55	2.75	2.75
4	100.35	96.03	4.32	4.32
5	99.51	99.00	0.51	0.51
6	98.50	100.76	-2.26	2.26
7	100.66	99.37	1.29	1.29
8	99.66	108.87	-9.21	9.21
9	99.70	105.16	-5.45	5.45
10	101.55	94.31	7.24	7.24

Table 2: Example data set

To illustrate the use of EAD, consider table 2. The inter-method bias is 0.03, which is quite close to zero, and conducive to agreement between methods. However, an identity plot would indicate very poor agreement, as the points are noticeably distant from the line of equality.

The limits of agreement are $[-9.61, 9.68]$, a wide interval for this data. As with the identity plot, this would indicate lack of agreement. The EAD is 3.71.

The Bland-Altman plot remains a useful part of the analysis. In 2, it is clear there is a systematic decrease in differences across the range of measurements.

Coverage Probability and Total Deviation Index

As elaborated by Lin and colleagues (Lin, 2000; Lin et al., 2002), an intuitive measure of agreement is a measure that captures a large proportion of data within a boundary

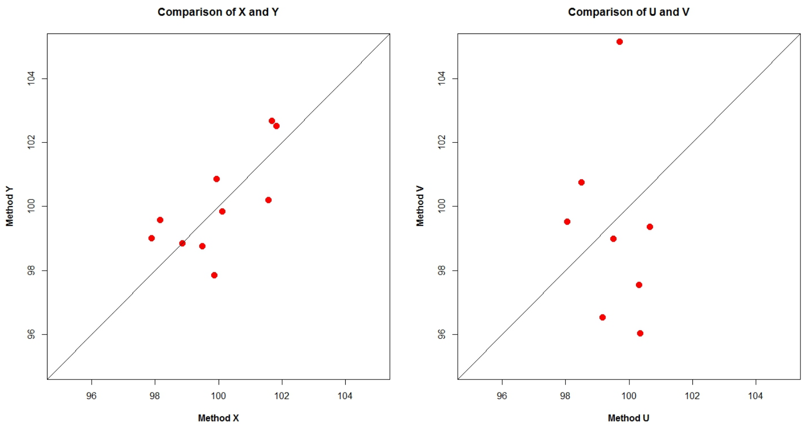

Figure 1: Identity Plot for example data

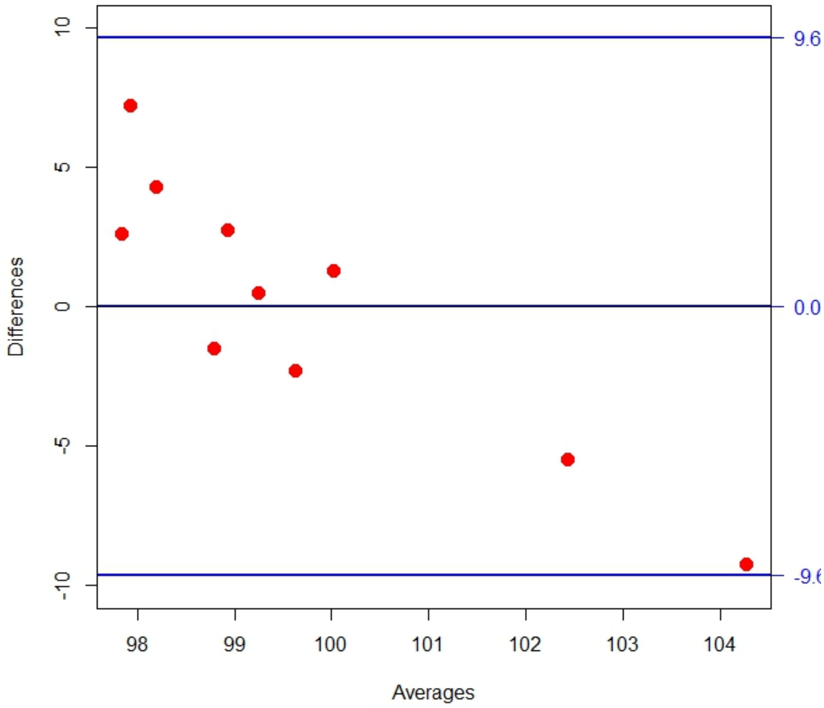

Figure 2: Expected Absolute Difference

for allowed observers differences. The proportion and boundary are two quantities that correspond to each other. If we set d0 as the predetermined boundary; i.e., the maximum acceptable absolute difference between two observers readings, we can compute the probability of absolute difference between any two observers readings less than d0.

0.3 Total Deviation Index and Coverage Probability

Lin et al. (2002) proposes a measure called the 'Total Deviation Index'. This assumes that the differences of paired measurements are a random sample from a normal distribution, and consequently the approach is to construct a probability interval, known as a tolerance interval, for these differences. A tolerance interval is a statistical range within which a specified proportion of the population lies. Smaller values of q indicate better agreement. P_0 is specified by the practitioner.

? generalize this approach to account for situations where the distributions are not identical, which is commonly the case. The TDI is not consistent and may not preserve its asymptotic nominal level, and that the coverage probability approach of Lin et al. (2002) is overly conservative for moderate sample sizes. This approach proposed by **?** is a regression based approach that models the mean and the variance of differences as functions of observed values of the average of the paired measurements. These techniques have been adopted by Mayo Clinic (Research Section).

0.4 Coverage Probability and Tolerance Deviation Index

The CP is the most intuitively clear approach; it mirrors the information provided by the TDI. Both TDI and CP depend on the normality assumption and offer better power for inference than the CCC. The CP would have difficulty discriminating among instruments or assays that have excellent agreement, all because the CP values would be very close to 1. In this case, the TDI can be used to discriminate among these. When a meaningful clinical range is known and the study is conducted over that range, the CCC offers a meaningful geo- metric interpretation and is unit free. Furthermore, the accuracy and precision components of the CCC offer more insight.

Therefore, the CCC, accuracy, and precision remain very useful tools. Note that when Y and X are not linearly related, the CCC will capture the total deviation. However, it will treat the nonlinear deviation as imprecision rather than inaccuracy. The CCC, ICC, and Pearson correlation coefficient depend largely on the analytical range and the intrasample variation.

0.5 Coverage probability

This term refers to the probability that a procedure for constructing random regions will produce an interval containing, or covering, the true value. It is a property of the interval producing procedure, and is independent of the particular sample to which such a procedure is applied. We can think of this quantity as the chance that the interval constructed by such a procedure will contain the parameter of interest.

Coverage probability (CP)

Another user friendly measure of agreement which is related to the computation of the TDI is the so called coverage probability (CP) . The CP describes the proportion captured within a pre-specified boundary of the absolute paired-measurement differences from two devices, i.e., the value of pκ such that $P(|D| < \kappa) = p_\kappa$. Therefore one can find pκ for a specified boundary κ using standard methods for computing probability quantities under normal assumptions [11]:

(13) and to obtain a CP estimate, pκ can be computed by replacing μ_D and σ_D by their REML estimate counterparts derived from model (1).

As with the TDI, the CP criterion can also be translated into a hypothesis test specification. In this case the interest is to ensure that a specified boundary of the absolute paired-measurement differences captures at least a predetermined proportion, p0:

The proposed TI method for inference about the TDI can be utilized to perform inferences about the CP estimates. From the TI in (10) it follows that

(14) Now κ is a fixed known boundary, and our interest lies in finding a lower confidence bound for the CP estimate. Thus, one can find a lower confidence bound for a non-central Student-t proportion with confidence level 1 - α by searching the non-centrality parameter, that depends on and hence on pκ, that satisfies

(15) and once the non-centrality parameter is achieved, a lower bound about the proportion pκ is found using equation (5),

However, the non-centrality parameter cannot be found in a closed form, so one may use again a modified version of the binary search algorithm as follows:

1. begin with the interval [low = 0; high = 1], as pκ is bounded by the interval (0,1);

2. calculate the midpoint of the interval *mid = (low + high)/2* and compute the difference ;

3. if d is greater than 0 up to a tolerance bound δ (i.e.,), then recalculate the interval [low = mid + δ; high = 1]; if it is lower than 0 up to a tolerance bound δ (i.e.), then recalculate the interval [low = 0; high = mid - δ];

4. repeat steps 2-3 until convergence, i.e. until d satisfies .

Total Deviation Index

Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. Lin LI. *Stat Med. 2000 Jan 30;19(2):255-70* *http://www.ncbi.nlm.nih.gov/pubmed/10641028*

In areas of inter-laboratory quality control, method comparisons, assay validation and individual bioequivalence, etc., the agreement between observations and target (reference) values is of interest. The mean of the squared difference between observations and target values (MSD) is a good measure of the total deviation. A new user-friendly statistic, the total deviation index (TDI(1-p)), is introduced that translates the MSD into an index that can be directly compared to a predetermined criterion.

The TDI(1-p) describes a boundary such that a majority, 100(1-p) per cent, of the observations are within the boundary (measurement unit and/or per cent) from their target values. Statistical inference using the sample counter part (estimate) is presented. A Monte Carlo experiment with 5000 runs was performed to confirm the estimate's validity. Applications in laboratory performance and validation, as well as individual bioequivalence, are presented.

Individual agreement between two measurement systems is determined using the total deviation index (TDI) or the coverage probability (CP) criteria as proposed by Lin (2000) and Lin et al. (2002). We used a variance component model as proposed by Choudhary (2007). Using the bootstrap approach, Choudhary (2007), and generalized confidence intervals, we construct bounds on TDI and CP. A simulation study was conducted to assess whether the bounds maintain the stated type I error probability of the test. We also present a computational example to demonstrate the statistical methods described in the paper.

- http://artax.karlin.mff.cuni.cz/r-help/library/MethComp/html/TDI.html

Total Deviation Index - TDI - Escaramis

The Total Deviation Index estimated by Tolerance Intervals to evaluate the concordance of measurement devices.

Gergia Escarams1, Carlos Ascaso1 and Josep L Carrasco1.

In an agreement assay, it is of interest to evaluate the degree of agreement between the different methods (devices, instruments or observers) used to measure the same characteristic. We propose in this study a technical simplification for inference about the total deviation index (TDI) estimate to assess agreement between two devices of normally-distributed measurements and describe its utility to evaluate inter- and intra-rater agreement if more than one reading per subject is available for each device.

We propose to estimate the TDI by constructing a probability interval of the difference in paired measurements between devices, and thereafter, we derive a tolerance interval (TI) procedure as a natural way to make inferences about probability limit estimates. We also describe how the proposed method can be used to compute bounds of the coverage probability.

The approach is illustrated in a real case example where the agreement between two instruments, a handle mercury sphygmomanometer device and an OMRON 711 automatic device, is assessed in a sample of 384 subjects where measures of systolic blood pressure were taken twice by each device. A simulation study procedure is implemented to evaluate and compare the accuracy of the approach to two already established methods, showing that the TI approximation produces accurate empirical confidence levels which are reasonably close to the nominal confidence level.

The method proposed is straightforward since the TDI estimate is derived directly from a probability interval of a normally-distributed variable in its original scale, without further transformations. Thereafter, a natural way of making inferences about this estimate is to derive the appropriate TI. Constructions of TI based on normal populations are implemented in most standard statistical packages, thus making it simpler for any practitioner to implement our proposal to assess agreement.

0.6 Unscaled Agreement Indices

- Summary agreement indices based on the absolute difference of readings by observers are grouped here as unscaled agreement indices.

- They are usually defined as the expectation of a function of the difference, or features of the distribution of the absolute difference.

- These indices include mean squared deviation, repeatability coefficient, repeatability variance, reproducibility variance (ISO), limits of agreement (Bland and Altman, 1999), coverage probability (CP) and total deviation index (TDI) (Lin et al., 2002 Choudhary and Nagaraja, 2007; Choudhary, 2007a).

0.7 Information Approach

PURPOSE: Disagreement on the interpretation of diagnostic tests and clinical decisions remains an important problem in medicine. As no strategy to assess agreement seems to be fail-safe to compare the degree of agreement, or disagreement,

0.7.1 Example: Systolic Blood Pressure

Bland and Altman (1999) present the example of measurements of systolic blood pressure of 85 individuals, by two observers (observer J and observer R) with sphygmomanometer, and one other measurement, by a semiautomatic device (device S).

Luiz et al. (16) re-analyze the data and also observe, with a graphical approach, a greater agreement between the two observers than between the observers and the semiautomatic device.

Using our information-based measure of disagreement; we also obtained a significantly more disagreement between each observer and the semiautomatic device than between the two observers (Table 1).

0.7.2 Discussion

- We can look at disagreement between observers as the distance between their ratings, so the metric properties are important. Moreover, the proposed measure of disagreement is scale-invariant, i.e., the degree of disagreement between two observers should be the same if the measurements are analyzed in kilograms or in grams, for example.

- Differential weighting is another property of the proposed information-based measure of disagreement: each comparison between two ratings is divided by a normalizing factor, depending on each pair of ratings alone, before summing. Therefore, the information-based measure of disagreement is appropriate for ratio scale measurements (with a natural 0) and it is not appropriate for interval scale measurements (without a natural 0).

- For example, outside air temperature in Celsius (or Fahrenheit) scale does not have a natural 0. The 0 is arbitrary and it does not make sense to say that 20 is twice as hot as 10. Outside air temperature in Celsius (or Fahrenheit) scale is an interval scale. On the other hand, height has a natural 0 meaning: the absence of height. Therefore, it makes sense to say that 80 inches is twice as large as 40 inches. Height is a ratio scale.

- Suppose the heights of a sample of subjects measured independently by two different observers. A difference between the two observers of 1 inch in a child subject represents a worse observers' error than a disagreement between observers of 1 inch in an adult subject.

- Due to differential weighting property of the information-based measure of disagreement, a difference between the observers of one inch in a child in fact weights less to the estimate of information-based measure of disagreement between observers than a difference between the observers of 1 inch in an adult.

- The usual approaches used to evaluate agreement have the limitation of the comparability of populations. In fact, ICC depends on the variance of the trait in the population; although this characteristic can be considered an advantage it does not permit one to compare the degree of agreement across different populations. Also the interpretation of the limits of agreement depends on what can be considered clinically relevant or not, which could be subjective and different from reader to reader.

- The comparison of the degree of agreement in different populations is not straightforward. Other approaches 16 and 17 to assess observer agreement have been proposed, however the comparability of populations is still not easy with these approaches.

- The proposed information-based measure of disagreement, used as a complement to current approaches for evaluating agreement, can be useful to compare the degree of disagreement among different populations with different characteristics, namely with different variances.

- Moreover, we believe that information theory can make an important contribution to the relevant problem of measuring agreement in medical research, providing not only better quantification but also better understanding of the complexity of the underlying problems related to the measurement of disagreement.

0.8 Probability-based Measures of Agreement

There are two measures of agreement based on the probability criteria. The first is the p_0-th percentile of jDj, say Q(p_0), where p_0 (¿ 0.5) is a specified large probability (usually 0:80). It was introduced by Lin (2000) who called it the total deviation index (TDI). Its small value indicates a good agreement between (X; Y). The TDI can be expressed as,

! EQUATION HERE -distribution with a single degree of freedom and non-centrality parameter ***Del***.

The second measure, introduced by Lin et al. (2002), is the **coverage probability** (CP) of the interval [0; 0], where a difference under 0 is considered practically equivalent to zero. There is no loss of generality in taking this interval to be symmetric around zero as it can be achieved by a location shift.

A high value of F(0) implies a good agreement between the methods.

0.9 LME - Pankaj Choudhury

Consistent with the conventions of mixed models, (**?**) formulates the measurement $y_{ij}$from method i on individual j as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{2}$$

The design matrix P_{ij} , with its associated column vector θ, specifies the fixed effects common to both methods. The fixed effect specific to the jth method is articulated by the design matrix W_{ij} and its column vector v_i. The random effects common to both methods is specified in the design matrix X_{ij}, with vector b_j whereas the random effects specific to the ith subject by the jth method is expressed by Z_{ij}, and vector u_j. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{3}$$

These vectors are assumed to be independent for different is, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{4}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate x is the key step in how this approach assesses agreement.

Bibliography

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.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.