

## Formal Tests

The approach proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of case-wise differences and means ( $\rho_{ad}$ ). According to the authors, this test is equivalent to the ‘Pitman Morgan Test’. For the Grubbs data, the correlation coefficient estimate ( $r_{ad}$ ) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers ‘r to z’ transformation (Cohen, Cohen, West, and Aiken, Cohen et al.). The null hypothesis ( $\rho_{ad}=0$ ) would fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has been no further mention of this particular test in Bland and Altman (1986), although Bland and Altman (1999) refers to Spearman’s rank correlation coefficient. Bland and Altman (1999) comments ‘we do not see a place for methods of analysis based on hypothesis testing’. Bland and Altman (1999) also states that consider structural equation models to be inappropriate.

### 0.0.1 The Problem of Identifiability

Dunn (2002) highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example  $\alpha$  may take the value of the inter-method bias estimate from Bland-Altman methodology. Another assumption is that the precision ratio  $\lambda = \frac{\sigma_\epsilon^2}{\sigma_\delta^2}$  may be known.

Dunn (2002) considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that is very difficult

to get replicate observations when the measurement method requires invasive medical procedure.

Dunn (2002) recommends the following approach for analyzing method comparison data. Firstly he recommends conventional Bland-Altman methodology; plotting the scatterplot and the Bland-Altman plot, complemented by estimate for the limits of agreement and the correlation coefficient between the difference and the mean. Additionally boxplots may be useful in considering the marginal distributions of the observations. The second step is the calculations of summary statistics; the means and variances of each set of measurements, and the covariances.

When both methods measure in the same scale (i.e.  $\beta = 1$ ), Dunn (2002) recommends the use of Grubbs estimators to estimate error variances, and to test for their equality. A test of whether the intercept  $\alpha$  may be also be appropriate.

## 0.1 Formal Models and Tests

The Bland-Altman plot is a simple tool for inspection of data, and Kinsella (1986) comments on the lack of formal testing offered by that methodology. Kinsella (1986) formulates a model for single measurement observations for a method comparison study as a linear mixed effects model, i.e. model that additively combine fixed effects and random effects.

$$Y_{ij} = \mu + \beta_j + u_i + \epsilon_{ij} \quad i = 1, \dots, n \quad j = 1, 2$$

The true value of the measurement is represented by  $\mu$  while the fixed effect due to method  $j$  is  $\beta_j$ . For simplicity these terms can be combined into single terms;  $\mu_1 = \mu + \beta_1$  and  $\mu_2 = \mu + \beta_2$ . The inter-method bias is the difference of the two fixed effect terms,  $\beta_1 - \beta_2$ . Each of the  $i$  individuals are assumed to give rise to random error, represented by  $u_i$ . This random effects terms is assumed to have mean zero and be normally distributed with variance  $\sigma^2$ . There is assumed to be an attendant error for each measurement on each individual, denoted  $\epsilon_{ij}$ . This is also assumed to have mean zero. The variance of measurement error for both methods are not assumed to be identical for both methods variance, hence it is denoted  $\sigma_j^2$ . The set of observations  $(x_i, y_i)$  by methods  $X$  and  $Y$  are assumed to follow the bivariate normal distribution with expected values  $E(x_i) = \mu_i$  and  $E(y_i) = \mu_i$  respectively. The variance covariance of the observations  $\Sigma$  is given by

$$\Sigma = \begin{bmatrix} \sigma^2 + \sigma_1^2 & \sigma^2 \\ \sigma^2 & \sigma^2 + \sigma_2^2 \end{bmatrix}$$

The inter-method bias is the difference of the two fixed effect terms,  $\beta_1 - \beta_2$ .

Kinsella (1986) demonstrates the estimation of the variance terms and relative precisions relevant to a method comparison study, with attendant confidence intervals for

both. The measurement model introduced by Grubbs (1948, 1973) provides a formal procedure for estimate the variances  $\sigma^2, \sigma_1^2$  and  $\sigma_2^2$  devices. Grubbs (1948) offers estimates, commonly known as Grubbs estimators, for the various variance components. These estimates are maximum likelihood estimates, a statistical concept that shall be revisited in due course.

$$\begin{aligned}\hat{\sigma}^2 &= \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{n-1} = S_{xy} \\ \hat{\sigma}_1^2 &= \sum \frac{(x_i - \bar{x})^2}{n-1} = S^2_x - S_{xy} \\ \hat{\sigma}_2^2 &= \sum \frac{(y_i - \bar{y})^2}{n-1} = S^2_y - S_{xy}\end{aligned}$$

Thompson (1963) defines  $\Delta_j$  to be a measure of the relative precision of the measurement methods, with  $\Delta_j = \sigma^2/\sigma_j^2$ . Thompson also demonstrates how to make statistical inferences about  $\Delta_j$ . Based on the following identities,

$$\begin{aligned}C_x &= (n-1)S_x^2, \\ C_{xy} &= (n-1)S_{xy}, \\ C_y &= (n-1)S_y^2, \\ |A| &= C_x \times C_y - (C_{xy})^2,\end{aligned}$$

the confidence interval limits of  $\Delta_1$  are

$$\begin{aligned}\Delta_1 &> \frac{C_{xy} - t\left(\frac{|A|}{n-2}\right)^{\frac{1}{2}}}{C_x - C_{xy} + t\left(\frac{|A|}{n-2}\right)^{\frac{1}{2}}} \\ \Delta_1 &> \frac{C_{xy} + t\left(\frac{|A|}{n-2}\right)^{\frac{1}{2}}}{C_x - C_{xy} - t\left(\frac{|A|}{n-1}\right)^{\frac{1}{2}}}\end{aligned}\tag{1}$$

The value  $t$  is the  $100(1 - \alpha/2)\%$  upper quantile of Student's  $t$  distribution with  $n - 2$  degrees of freedom (Kinsella, 1986). The confidence limits for  $\Delta_2$  are found by substituting  $C_y$  for  $C_x$  in (1.3). Negative lower limits are replaced by the value 0.

The case-wise differences and means are calculated as  $d_i = x_i - y_i$  and  $a_i = (x_i + y_i)/2$  respectively. Both  $d_i$  and  $a_i$  are assumed to follow a bivariate normal distribution with  $E(d_i) = \mu_d = \mu_1 - \mu_2$  and  $E(a_i) = \mu_a = (\mu_1 + \mu_2)/2$ . The variance matrix  $\Sigma_{(a,d)}$  is

$$\Sigma_{(a,d)} = \begin{bmatrix} \sigma_1^2 + \sigma_2^2 & \frac{1}{2}(\sigma_1^2 - \sigma_2^2) \\ \frac{1}{2}(\sigma_1^2 - \sigma_2^2) & \sigma^2 + \frac{1}{4}(\sigma_1^2 + \sigma_2^2) \end{bmatrix}. \quad (2)$$

### 0.1.1 Morgan Pitman Testing

An early contribution to formal testing in method comparison was made by both Morgan (1939) and Pitman (1939), in separate contributions. The basis of this approach is that if the distribution of the original measurements is bivariate normal. Morgan and Pitman noted that the correlation coefficient depends upon the difference  $\sigma_1^2 - \sigma_2^2$ , being zero if and only if  $\sigma_1^2 = \sigma_2^2$ .

The classical Pitman-Morgan test is a hypothesis test for equality of the variance of two data sets;  $\sigma_1^2 = \sigma_2^2$ , based on the correlation value  $\rho_{a,d}$ , and is evaluated as follows;

$$\rho(a, d) = \frac{\sigma_1^2 - \sigma_2^2}{\sqrt{(\sigma_1^2 + \sigma_2^2)(4\sigma_s^2 + \sigma_1^2 + \sigma_2^2)}} \quad (3)$$

The correlation constant takes the value zero if, and only if, the two variances are equal. Therefore a test of the hypothesis  $H : \sigma_1^2 = \sigma_2^2$  is equivalent to a test of the hypothesis  $H : \rho(D, A) = 0$ . This corresponds to the well-known  $t$  test for a correlation coefficient with  $n - 2$  degrees of freedom. Bartko (1994) describes the Morgan-Pitman test as identical to the test of the slope equal to zero in the regression of  $Y_{i1}$  on  $Y_{i2}$ , a result that can be derived using straightforward algebra.

### 0.1.2 Paired sample $t$ test

Bartko (1994) discusses the use of the well known paired sample  $t$  test to test for inter-

method bias;  $H : \mu_d = 0$ . The test statistic is distributed a  $t$  random variable with  $n - 1$  degrees of freedom and is calculated as follows,

$$t^* = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}} \quad (4)$$

where  $\bar{d}$  and  $s_d$  is the average of the differences of the  $n$  observations. Only if the two methods show comparable precision then the paired sample student t-test is appropriate for assessing the magnitude of the bias.

$$t^* = \frac{\bar{d}}{s_d/\sqrt{n}} \quad (5)$$

### 0.1.3 Bland-Altman correlation test

The approach proposed by Altman and Bland (1983) is a formal test on the Pearson correlation coefficient of case-wise differences and means ( $\rho_{AD}$ ). According to the authors, this test is equivalent to the ‘Pitman Morgan Test’. For the Grubbs data, the correlation coefficient estimate ( $r_{AD}$ ) is 0.2625, with a 95% confidence interval of (-0.366, 0.726) estimated by Fishers ‘ $r$  to  $z$ ’ transformation (Cohen, Cohen, West, and Aiken, Cohen et al.). The null hypothesis ( $\rho_{AD} = 0$ ) fail to be rejected. Consequently the null hypothesis of equal variances of each method would also fail to be rejected. There has no been no further mention of this particular test in Bland and Altman (1986), although Bland and Altman (1999) refers to Spearman’s rank correlation coefficient. Bland and Altman (1999) state that they ‘do not see a place for methods of analysis based on hypothesis testing’. Bland and Altman (1999) also states that consider structural equation models to be inappropriate.

## 0.2 Measurement Error Models

Dunn (2002) proposes a measurement error model for use in method comparison studies. Consider  $n$  pairs of measurements  $X_i$  and  $Y_i$  for  $i = 1, 2, \dots, n$ .

$$X_i = \tau_i + \delta_i \quad (6)$$

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

In the above formulation is in the form of a linear structural relationship, with  $\tau_i$  and  $\beta\tau_i$  as the true values, and  $\delta_i$  and  $\epsilon_i$  as the corresponding measurement errors. In the case where the units of measurement are the same, then  $\beta = 1$ .

$$E(X_i) = \tau_i \quad (7)$$

$$E(Y_i) = \alpha + \beta\tau_i$$

$$E(\delta_i) = E(\epsilon_i) = 0$$

The value  $\alpha$  is the inter-method bias between the two methods.

$$z_0 = d = 0 \quad (8)$$

$$z_{n+1} = z_n^2 + c \quad (9)$$

## 0.3 Measurement Error Models

Dunn (2002) proposes a measurement error model for use in method comparison studies. Consider  $n$  pairs of measurements  $X_i$  and  $Y_i$  for  $i = 1, 2, \dots, n$ .

$$X_i = \tau_i + \delta_i \quad (10)$$

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

In the above formulation is in the form of a linear structural relationship, with  $\tau_i$  and  $\beta\tau_i$  as the true values, and  $\delta_i$  and  $\epsilon_i$  as the corresponding measurement errors. In the case where the units of measurement are the same, then  $\beta = 1$ .

$$E(X_i) = \tau_i \quad (11)$$

$$E(Y_i) = \alpha + \beta\tau_i$$

$$E(\delta_i) = E(\epsilon_i) = 0$$

The value  $\alpha$  is the inter-method bias between the two methods.

$$z_0 = d = 0 \quad (12)$$

$$z_{n+1} = z_n^2 + c \quad (13)$$

### 0.3.1 Identifiability

In many models, naive assumptions are required to overcome issues of identifiability. Precision is defined by the reciprocal of the variance of the random errors. Also it is assumed that the error variance is independent of the amount of material being measured. However, in practice, this is often not the case. Variability increases over the



scale of measurements over many cases. Estimators of scale parameters are estimable only if the analyst is prepared to make naive, if not unacceptable, assumptions.

Equation 4  $\psi$  and  $\varepsilon$  are statistically independent of each other. Contamination effect that arises from non-specificity / specimen specific bias. Random error is measured by . Homogeneity of variances is assumed. If there are no replicate measures, both variances are completely confounded, and there is no way of telling them apart. Scaling of new measurements is measured by .

? 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 methodology 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.

Maximum likelihood estimation is used to estimate the parameters. The REML estimation is not considered since it does not lead to a joint distribution of the estimates of fixed effects and random effects parameters, upon which the assessment of agreement is based.

## Pitman's Test on Correlated variances

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Pitman's test is identical to the slope equal to zero in the regression of  $y$  on  $x$ .

### 0.3.2 Identifiability

Dunn (2002) highlights an important issue regarding using models such as these, the identifiability problem. This comes as a result of there being too many parameters to be estimated. Therefore assumptions about some parameters, or estimators used, must be made so that others can be estimated. For example in literature the variance ratio  $\lambda = \frac{\sigma_1^2}{\sigma_2^2}$  must often be assumed to be equal to 1 (Linnet, 1998). Dunn (2002) considers methodologies based on two methods with single measurements on each subject as inadequate for a serious study on the measurement characteristics of the methods. This is because there would not be enough data to allow for a meaningful analysis. There is, however, a contrary argument that in many practical settings it is very difficult to get replicate observations when the measurement method requires invasive medical procedure.

Bradley and Blackwood (1989) offers a formal simultaneous hypothesis test for the mean and variance of two paired data sets. Using simple linear regression of the differences of each pair against the sums, a line is fitted to the model, with estimates for intercept and slope ( $\hat{\beta}_0$  and  $\hat{\beta}_1$ ). The null hypothesis of this test is that the mean ( $\mu$ ) and variance ( $\sigma^2$ ) of both data sets are equal if the slope and intercept estimates are equal to zero (i.e  $\sigma_1^2 = \sigma_2^2$  and  $\mu_1 = \mu_2$  if and only if  $\beta_0 = \beta_1 = 0$  )

A test statistic is then calculated from the regression analysis of variance values (Bradley and Blackwood, 1989) and is distributed as ‘ $F$ ’ random variable. The degrees of freedom are  $\nu_1 = 2$  and  $\nu_2 = n - 2$  (where  $n$  is the number of pairs). The critical value is chosen for  $\alpha\%$  significance with those same degrees of freedom. Bartko (1994) amends this methodology for use in method comparison studies, using the averages of the pairs, as opposed to the sums, and their differences. This approach can facilitate simultaneous usage of test with the Bland-Altman methodology. Bartko’s test statistic

take the form:

$$F.test = \frac{(\Sigma d^2) - SSReg}{2MSReg}$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Averages	1	0.04	0.04	0.74	0.4097
Residuals	10	0.60	0.06		

Table 1: Regression ANOVA of case-wise differences and averages for Grubbs Data

For the Grubbs data,  $\Sigma d^2 = 5.09$ ,  $SSReg = 0.60$  and  $MSreg = 0.06$  Therefore the test statistic is 37.42, with a critical value of 4.10. Hence the means and variance of the Fotobalk and Counter chronometers are assumed to be simultaneously equal.

Importantly, this methodology determines whether there is both inter-method bias and precision present, or alternatively if there is neither present. It has previously been demonstrated that there is a inter-method bias present, but as this procedure does not allow for separate testing, no conclusion can be drawn on the comparative precision of both methods.

# Bibliography

- 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.
- Bartko, J. (1994). Measures of agreement: A single procedure. *Statistics in Medicine* 13, 737–745.
- Bland, J. and D. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* i, 307–310.
- Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160.
- Bradley, E. L. and L. G. Blackwood (1989). Comparing paired data: A simultaneous test for means and variances. *The American Statistician* 43(4), 234–235.
- Cohen, J., P. Cohen, S. West, and L. Aiken. *Applied multiple regression / correlation analysis for the behavioral sciences* (Third ed.). Laurence Erlbaum Associates.
- Dunn, G. (2002). *Statistical Evaluation of Measurement Error* (Second ed.). Stanford: American Mathematical Society and Digital Press.
- Grubbs, F. (1948). On estimating precision of measuring instruments and product variability. *Journal of the American Statistical Association* 43, 243–264.

- Grubbs, F. (1973). Errors of measurement, precision, accuracy and the statistical comparison of measuring instruments. *Technometrics* 15(1), 53–66.
- Kinsella, A. (1986). Estimating method precision. *The Statistician* 35, 421–427.
- 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.
- Linnet, K. (1998). Performance of deming regression analysis in case of misspecified analytical error ratio in method comparison studies. *Clinical Chemistry* 44, 1024–1031.
- Morgan, W. A. (1939). A test for the significance of the difference between two variances in a sample from a normal bivariate population. *Biometrika* 31, 13–19.
- Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika* 31, 9–12.
- Thompson, W. (1963). Precision of simultaneous measurement procedures. *Journal of American Statistical Association* 58, 474–479.