Contents

0.1 Conventional Approaches to Unreplicated Design . . . 3

0.1.1 The Pitman-Morgan test . . . 4

0.1.2 The Bradley-Blackwood test . . . 4

0.2 Conventional Approaches to Replicated Designs . . . 6

0.3 Outliers . . . 6

0.3.1 Bland-Altman's . . . 6

0.3.2 Bartko . . . 6

0.3.3 Hawkins . . . 6

0.3.4 Outliers in the LME Model framework . . . 6

The issue of whether two methods of measurements are comparable to the extent that they can be used interchangeably with sufficient accuracy and measurement precision is encountered frequency in scientfic research. Published examples of method comparison studies can be found in disciplines as diverse as pharmacology (Ludbrook, 1997), anaesthesia (Myles, 2007), and cardiac imaging methods (Krummenauer et al., 2000)(references).

In the most basic design, items (such as people in medical studies) are measures once only by each of two measurement methods. If the recorded measurements by the two instruments differ systematically, a problem of inter-method bias exists. Oftentimes this bias can be mitigated by some technical adjustment or recalibration of the readings. However, if the method variances differ, no comparable adjustment is possible, and a more serious problem exists.

This problem has received significant attention in statistical literature over many decades. Statistical tests for equality of measurements precisions were devised by Pitman (1939) and Morgan (1939). Grubbs (1948, 1973) formulate a model testing framework for comparing multiple devices.

A graphical tool advocated by Altman and Bland (1983); Bland and Altman (1986) shifted the analysis from concerns over statistical hypothesis testing to concerns of statistical equivalence. Known as the Bland-Altman plot, this technique has become the most popular (and in some cases, obligatory) method of presenting method comparison studies in journals. Dunn (2002) prefers an approach based in measurement error models. Carstensen et al. (2008) extend the technique to replicated deisign using LME frameworks to replicated designs using an LME framework, and supports this work with an `R` package. Broemeling (2009) lays out a Bayesian strategy.

With some few exceptions, e.g. Hawkins (1978) and Bartko (1994), the issue of outliers and anomalous values have not featured prominently in method comparison literature.

This chapter is organized as follows; firstly a review of the tools used in the analysis of unreplicated deisgns. We then consder ther extension to replicated designs. The

LME framework advanced by Carstensen et al. (2008) and Roy (2009) are given special attention, and we conclude with some remarks about outliers, a topic that we will particular attention to in a later chapter.

0.1 Conventional Approaches to Unreplicated Design

Let the random variables Y_1 and Y_2 be distributed bivariate normal with $\mathrm{E}(Y_1) = \mu_1,\ \mathrm{E}(Y_2) = \mu_2,\ \mathrm{var}(Y_1) = \sigma_1^2,\ \mathrm{var}(Y_2) = \sigma_2^2$, and correlation coefficient $-1 < \rho < 1$. Of particular interest are tests of the unconditional marginal hypotheses $\mathrm{H}'\colon\ \mu_1 = \mu_2$ and $\mathrm{H}''\colon\ \sigma_1^2 = \sigma_2^2$, and tests of the joint hypothesis $\mathrm{H}^{\mathrm{J}}\colon\ \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$. The random variables $D = Y_1 - Y_2$ and $S = Y_1 + Y_2$ are bivariate normal with expectations $\mathrm{E}(D) = \mu_D = \mu_1 - \mu_2$ and $\mathrm{E}(S) = \mu_S = \mu_1 + \mu_2$, variances $\mathrm{var}(D) = \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$ and $\mathrm{var}(S) = \sigma_S^2 = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2$, and covariance $\mathrm{cov}(D, S) = \sigma_1^2 - \sigma_2^2$. The conditional distribution of D given S is normal with expectation $\mu_{D|S=s} = \mu_D + [(\sigma_1^2 - \sigma_2^2)/\sigma_S^2](s - \mu_S)$ and variance $\sigma_{D|S}^2 = \sigma_D^2 - (\sigma_1^2 - \sigma_2^2)^2/\sigma_S^2$. These differences and sums are the building blocks of the test procedures: of H', due to Gossett (1908); of H'', devised concurrently by Pitman (1939) and Morgan (1939); and of H^{J}, proposed by Bradley and Blackwood (1989). Notably, the classic test procedure of H' due to Gossett (1908) makes no assumptions about the equality, or otherwise, of the variance parameters σ_1^2 and σ_2^2.

The test procedure for H^{J} advanced by Bradley and Blackwood (1989) additively decomposes into independent tests of H'' and the conditional marginal hypothesis $\mathrm{H}^{\dagger}\colon\ \mu_1 = \mu_2$, assuming the additional restriction $\sigma_1^2 = \sigma_2^2$. The former test in this decomposition is the Pitman-Morgan procedure referred to above. The latter test in the decomposition is based on the F-ratio with $(1, n-2)$ degrees-of-freedom, denoted below by F_0^*. Conveniently, all three test procedures can be calculated from the fitted simple linear regression of observed differences on observed sums.

0.1.1 The Pitman-Morgan test

The test of the hypothesis that the variances σ_1^2 and σ_2^2 are equal, which was devised concurrently by **?** and **?**, is based on the correlation of D with S, the coefficient being $\rho_{DS} = (\sigma_1^2 - \sigma_2^2)/(\sigma_D \sigma_S)$, which is zero if, and only if, $\sigma_1^2 = \sigma_2^2$. Consequently a test of $\mathrm{H}''\colon\ \sigma_1^2 = \sigma_2^2$ is equivalent to a test of $\mathrm{H}\colon\ \rho_{DS} = 0$ and the test statistic is the familiar t-test for a correlation coefficient with $(n-2)$ degrees-of-freedom:

$$T^*_{\mathrm{PM}} = R\sqrt{\frac{n-2}{1-R^2}},$$

where $R = \sum(D_i - \bar{D})(S_i - \bar{S})/[\sum(D_i - \bar{D})^2 \sum(S_i - \bar{S})^2]^{\frac{1}{2}}$ is the sample correlation coefficient of the n case-wise differences $D_i = Y_{i1} - Y_{i2}$ and sums $S_i = Y_{i1} + Y_{i2}$. Throughout this paper the summation $\sum$ is taken to imply $\sum_{i=1}^{n}$. The procedure is to reject the hypothesis H'' in favour of $\sigma_1^2 \neq \sigma_2^2$ if $|T^*_{\mathrm{PM}}| > t_{\alpha/2,(n-2)\mathrm{df}}$.

0.1.2 The Bradley-Blackwood test

Bradley and Blackwood (1989) write $\mu_{D|S=s} = \mu_D + [(\sigma_1^2 - \sigma_2^2)/\sigma_S^2](s - \mu_S) = \beta_0 + \beta_1 s$ where $\beta_0 = \mu_D - [(\sigma_1^2 - \sigma_2^2)/\sigma_S^2]\mu_S$ and $\beta_1 = (\sigma_1^2 - \sigma_2^2)/\sigma_S^2$. They use this result to propose a test of the joint hypothesis H^{J}, which is true if, and only if, $\beta_0 = \beta_1 = 0$. Their test procedure follows directly from the theory of linear models (Hogg and Craig, 1995, for example) and is based on the F-ratio

$$F^* = (\frac{n-2}{2})(\frac{\sum D_i^2 - \mathrm{SSE}}{\mathrm{SSE}}) \sim F_{(2,n-2)\mathrm{df}}, \tag{1}$$

where SSE is the residual error sum-of-squares from the fitted regression $\hat{D}_i = \hat{\beta}_0 + \hat{\beta}_1 s_i$ of the case-wise differences on the case-wise sums. The procedure is to reject the hypothesis H^{J} in favour of $\mu_1 \neq \mu_2$ and (or) $\sigma_1^2 \neq \sigma_2^2$ if $F^* > F_{\alpha,(2,n-2)\mathrm{df}}$. The F distribution in (1) is valid conditional on S, and since the distribution does not depend on S it is also the unconditional distribution of the test statistic F^*.

Consequently there is no need to make special allowance for the fact that the case-wise sums encountered here are random sums, and not fixed, error-free explanatory variables as regression theory demands. This is the same argument that is generally used to show that t-test for a correlation coefficient is valid, e.g., T^*_{PM} above (Hogg and Craig, 1995, page 499).

Bland-Altman Plots

Altman and Bland (1983) correctly criticised the use the paired differnce reegresion and correlation anlayss for use in method comparison. Their graphical procedure is based on pair-wise differences versus pair-wise averages. This plot is essentially a visual analogue of the quantities underpinning the tests presented in the previous subsections.

The plot of differences versus average can be obtained by rotating the points in the original scatterplot of X and Y by 45 degrees and rescaling accordingly. This plot, in essence, serves the purpose of a diagnostic plot. From a historical perspective, a similar graphical tool was devised by Tukey several decades earlier Kozak and Wnuk (2014).

We will illustrate the workings of a Bland-Altman plot through a simple example. The Data is table 1 shows (GRUBBS)

The values in the final two columns contain the pairwise differecnces $d_i = x_i - y_i$ and $a_i = \frac{x_i + y_i}{2}$.

A plot of this quantities is show in figure 1.

Also included is a horizontal grey line repesenting th mean of the differences $\bar{d}$.

The horizontal dotted lines refer to the limits of agreement and are placed two standard deviations above and below $bard$

The rationale for this plot is that methods showing good agreement would be expected to have values falling predominantly between the limits of agreement.

Bland and Altman (1986) suggested that exact LOAs can be obtained by placing 1.96 in place of the 2 as a multuplier.

$$\bar{d} \pm t_{n-1}$$

Altman and Bland (1983) supplement their graphical tool wth a test of the equality of variances, based n the Pitman-Morgan procedre. This test was omitted from their lances paper Bland and Altman (1986).

In Bland and Altman (1999), they argue that they don't see a role in hypothesis testing in establishing equivalence of measurement methods.

Much of this analysis is based on classical assumptions of normally distributed data.

Enhancements proposed by Bland and Altman (1999), such as the use of confidence interval estimates for Limits of agreement have been seldom used in practice.

0.2 Conventional Approaches to Replicated Designs

It is worth noting that there are two type of replicate measurement, linked and unlinked.

For method comparison, two levels of sophistication exist beyond the classical design. Roy (2009) considers two instruments with replicate measurements.

0.3 Outliers

0.3.1 Bland-Altman's

Their protocol for the treatment of outliers is unclear.

0.3.2 Bartko

0.3.3 Hawkins

0.3.4 Outliers in the LME Model framework

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.

Broemeling, L. D. (2009). *Bayesian methods for measures of agreement.* CRC Press.

Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics 4*(1).

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

Gossett, W. S. G. (1908). The probable error of a mean. *Biometrika 6*(1), 1–25.

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.

Hawkins, D. M. (1978). Analysis of three tests for one or two outliers. *Statistica Neerlandica 32*(3), 137–148.

Hogg, R. V. and A. T. Craig (1995). *Introduction to mathematical statistics.(5"" edition).* Upper Saddle River, New Jersey: Prentice Hall.

Kozak, M. and A. Wnuk (2014). Including the tukey mean-difference (bland–altman) plot in a statistics course. *Teaching Statistics 36*(3), 83–87.

Krummenauer, F., I. Genevriere, and U. Nixdorff (2000). The biometrical comparison of cardiac imaging methods. *Computer Methods and Programs in Biomedicine 62*, 21–34.

Ludbrook, J. (1997). Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology 24*, 193–203.

Morgan, W. A. (1939). A test for the signicance of the difference between two variances in a sample from a normal bivariate population. *Biometrika 31*, 13–19.

Myles, P. (2007). Using the Bland Altman method to measure agreement with repeated measures. *British Journal of Anaesthesia 99*(3), 309–311.

Pitman, E. J. G. (1939). A note on normal correlation. *Biometrika 31*, 9–12.

Roy, A. (2009). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.