Statistical Methods in Biomedical Research and Measurement Science

Júlia Volaufová[1]

Biostatistics Program, LSUHSC School of Public Health
1600 Canal Street, New Orleans, LA 70112, USA
E-mail: jvolau@lsuhsc.edu

Abstract. *In this paper we concentrate on few statistical methods that are widely used in the analysis of clinical trials and clinical studies and are also used in theory of measurement science.*

Keywords: *Mixed linear model; two-way mixed models; bias of a measurement, uncertainty of a measurement, methods comparison*

1. Introduction – Sources of Variability

I am often presented with a question what it was like after I left the Institute of Measurement Science more than 10 years ago and became a biostatistician, whether it was like switching jobs. Here I try to present ideas and illustrate that both in biostatistics and in some areas of measurement science the same statistical methods are used, and hence developing fundamentals for statistical methodology is important and useful for many seemingly different areas of application. I don't have any ambitions to present here a list of all statistical methods that may find their applications in both areas, in biomedical as well as in technical settings. Just to mention some – in medical setting and in actuarial science we talk about survival analysis or time-to-event models and in technical areas exactly the same statistical models are used in reliability theory; in developing diagnostic tests we talk about sensitivity and specificity and the same probabilistic approach is used in, say, laboratory settings when examining assays. In this paper we try to focus only on mixed linear models, and show their already wide use in metrology or theory of measurement.

Generally, statistics plays an essential role as soon as we start to deal with observed values - measurements - data - and we want to make reasonable inference from them, draw a more or less general conclusion based on observations. Hence, statistics is essential in measurement science particularly in metrology and theory of measurement, including data and error analysis, standards and calibration. These include also problems that are related to identifying, estimating, and combining uncertainties, combining and/or comparing results of measurements from different sources (measurement equipments/systems, laboratories, etc.), as well as developing methods for designing quality control procedures and establishing the reliability of measuring systems.

Biostatistics is essentially statistics. It concerns the development of statistical methods and plays a key role in design and analysis of studies in public health and biomedical research. In analyzing biological experiments we face several independent sources of variability. Usually the study (experiment) is designed so that the effectiveness of a certain intervention, say treatment, is to be established (i.e., Phase II Clinical Trials). For that, a set of markers (variables) is chosen that would be measured. Often the

[1] On long term leave from the Institute of Measurement Science, Slovak Academy of Sciences, Bratislava

subjects are followed for a certain period of time and the same quantity (variable) is measured repeatedly at predetermined time intervals, say, every month for 2 years. Obviously the measurement is carried out with some uncertainty (variability) that characterizes the measurement error of the measurement system. The efficacy of the treatment needs to be tested and conclusions are supposed to be made for a certain well-specified population. For that, a sample from the population under investigation is taken. In order to minimize the sampling bias, randomization is carried out, i.e., each individual from the sample is randomly assigned into, say, two groups – the intervention and the control group. The control group serves then as a standard, a reference for comparisons and for establishing efficacy of treatment (over time). When handling each individual measurement separately we can see that in this setting there are at least three potential sources of error (variability) that we may observe - the biological "between" subject effect, the "within" subject effect that characterizes the variability of an individual when observed over time, and the "measurement" error that characterizes the behavior of the equipment that is used in order to obtain the observed datum.

In technical areas, say when observing the unit under test (UUT), the three potential sources of error (the between system (unit) effect, the within system (unit) effect, and the measurement error) are also present, although under reasonable assumptions the between system error is negligibly small compared to the measurement error, and the within system error is essentially nonexistent since often the system test is performed at a given real time. On the other hand, in biomedical settings the biological variabilities, the between and within subject effects, are much higher than the measurement error, which in most cases is negligibly small.

2. Mixed Linear Model – Two-Way Heteroscedastic Mixed Model

The effects and possible sources of variability are usually modeled such that the sampling design is taken closely into consideration. Here we assume that the observations on different sampling units (subjects, etc.) are independent. As soon as the model is set up, it dictates the dependencies between observations, the covariances (correlations) between each two. Formalizing the above described setting, we may observe the following. Any single measurement is a function of a subject (unit), treatment, time, and possibly replication. The simplest model that takes all that into consideration is expressed as:

$$y_{ijkl} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + \epsilon_{ijkl} \, . \tag{1}$$

Here y_{ijkl} is the l-th observation (measurement) at the k-th time point on the j-th subject (entity, concentration, etc.) in the i-th treatment group (by measurement system, laboratory, etc.). μ represents the common unknown population mean and α_i is the fixed effect, say the treatment effect that we would like to test. All the other terms in the model represent random effects. β_{ij} is the random between subject effect with mean zero and population variance σ_b^2; γ_{ijk} is the random within subject effect again with zero mean and with variance σ_g^2, and finally ϵ_{ijkl} denotes the measurement error that has zero mean, too, and variance σ_e^2. In the special case that we have only one observation at each time point, within-subject error becomes $\gamma_{ijk} + \epsilon_{ijkl}$, and the two terms are not separately identifiable. All random variables in the model representing random effects are mutually independent, that's how the model is set up. Hence the variance of a single observation Y_{ijkl}, denoted by σ_y^2, is then

$$\text{var}\,(Y_{ijkl}) = \sigma_y^2 = \sigma_b^2 + \sigma_g^2 + \sigma_e^2 \;\; \text{for all } i, j, k, l. \tag{2}$$

The covariances between different replications on the same subject at the same time point are

$$\text{cov}\,(Y_{ijkl}, Y_{ijkl'}) = \sigma_b^2 + \sigma_g^2 \;\; \text{for all } i, j, k, \text{and } l \neq l'. \tag{3}$$

Similarly, the covariances between observations on the same sampling unit (subject) at different time points are

$$\text{cov}\,(Y_{ijkl}, Y_{ijk'l'}) = \sigma_b^2 \;\text{ for all } i, j, l, l',\; k \neq k', \tag{4}$$

and the covariances between any two observations on different sampling units are 0. If we arrange the observations in lexicographic order (the indices from the right move the fastest) then we end up with a covariance matrix that has a block diagonal form, and the resulting covariance matrix for each sampling unit can be expressed as a linear combination of known matrices and unknown parameters known as variance components. Our model, since it contains fixed (nonrandom) and random effects, is an example of a *mixed linear model* and particularly model (1) represents a *two-way mixed linear model*.

There are obviously possible variations of model (1). For example, it is often reasonable to assume that the replications of measurements taken at the same time point on the same subject are independent and they are carried out in order to increase the precision of each measurement entering the final analysis, i.e., to decrease the variance of the measurement error. Then after averaging over replications the model takes the form

$$\bar{y}_{ijk.} = \mu + \alpha_i + \beta_{ij} + \gamma_{ijk} + \bar{\epsilon}_{ijk.}. \tag{5}$$

Depending on what level of complexity the model has to achieve, we take the investigation further. If, e.g., the number of replications at each time point for all subjects (units) is the same then if we combine the last two terms in expression (5) into one, say ϵ^*_{ijk}, and reparametrize its variance we get

$$\text{var}\,(\epsilon^*_{ijk}) = \sigma_g^2 + \frac{\sigma_e^2}{n} = \sigma_{e^*}^2, \tag{6}$$

where n is the number of replicates. If it is not the case and the number of replications is, say n_{jk}, then we end up with different variance components for different sets of measurements. Then we talk about *heteroscedasticity*.

The theory of mixed linear models is pretty well developed and has been around for more than 50 years. The origins and motivation for development of these models dates back to late thirties and early forties. Methods for estimation of fixed effects, for prediction of random effects as well as for estimation of variance components are implemented in many statistical software packages such as SAS, SPSS, Stata or S+. In applications, though, there are unaccountable situations and special cases for which the general approach gives unsatisfactory results due to poor properties of approximations or restrictions on the model. As an example we can mention an application in medical research that involves, e.g., comparison of several measurement/calculation methods of certain quantity such as percent body fat, blood pressure, etc.. Similarly, if combining experiments from different trials or, in other words, applying a meta-analysis approach to several sets of data, there are questions that have to be addressed very specifically (see, e.g., [5] or [7]). If we change a setting and talk about theory of measurement then the same can be said about comparing or combining measurements from, say, different laboratories or measurement systems (see, e.g., [14], [20], or [19]).

3. Measurement Methods Comparisons

Let's look closely at one particular application of mixed linear models that is developed in the biomedical area and is used also in metrology.

In several of their papers, J.M. Bland and D.G.Altman, see, e.g. [1], [2], and [3], propose a graphical method for comparison of two different methods of measurement of some quantity and assessment of an *agreement* of the two methods of measurements. Their paper in *Lancet* is as of March 2005 cited more

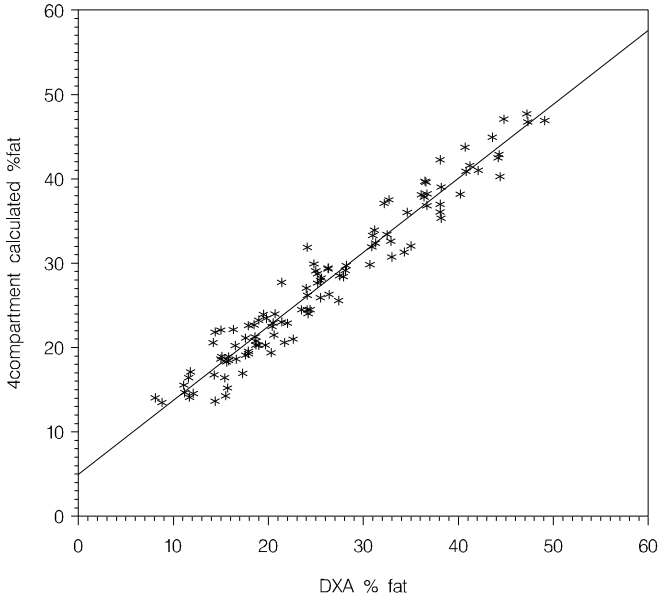

Figure 1: Graph of original data; four-compartment calculated vs. DXA % body fat with the regression line. The estimated regression equation is $4c = 5.076 + 0.872DXA$, the correlation coefficient is $\rho = 0.975$.

than 9000 times. Their graphical approach is extremely popular in medical applications because of its simplicity and transparency. Still this topic stirs up discussions and the results are often over-interpreted.

Let's illustrate the comparisons on an example. In the study, 108, 12-year-old children were measured by several different body composition measurement methods in order to establish their percent body fat. For detailed description of the study see [4]. One of the main interests was in comparison of an expensive method by dual X-ray energy absorptiometry (DXA) with a calculated percent body fat using four compartments (body density, body water, height, and weight). Each method of establishing percent body fat was applied exactly once on each child. As a first step we usually plot the data. Figure 1 shows the raw data, the four-compartment % body fat vs. DXA % body fat for all children in a study.

For modeling the observations we use the following model.

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij},\ i = 1, 2, \ldots, n,\ j = 1, 2. \tag{7}$$

Here we changed the notation since we do not have treatments. Under the present notation, α_i is a random between subject effect, assuming $\alpha_i \sim N(0, \sigma_a^2)$, $i = 1, 2, \ldots, n$; β_j is a fixed method effect where β_j represents the bias of each method, $j = 1, 2$; and ε_{ij} is the error term, $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, allowing for different variances for the two methods. All random variables in a model are assumed to be mutually independent. Here we have added an assumption of normality.

We say that two (or more) methods *agree* if they have the same bias and the same variance. The model for two measurement methods comparisons involves a common mean (% body fat) that is influenced by a random between-subject effect (each child has its own value of percent body fat) and two different within method variabilities (the actual observed values). In order to establish whether the methods agree we have to test for equality of biases, $\beta_1 = \beta_2$, and for equality of variances, $\sigma_1^2 = \sigma_2^2$.

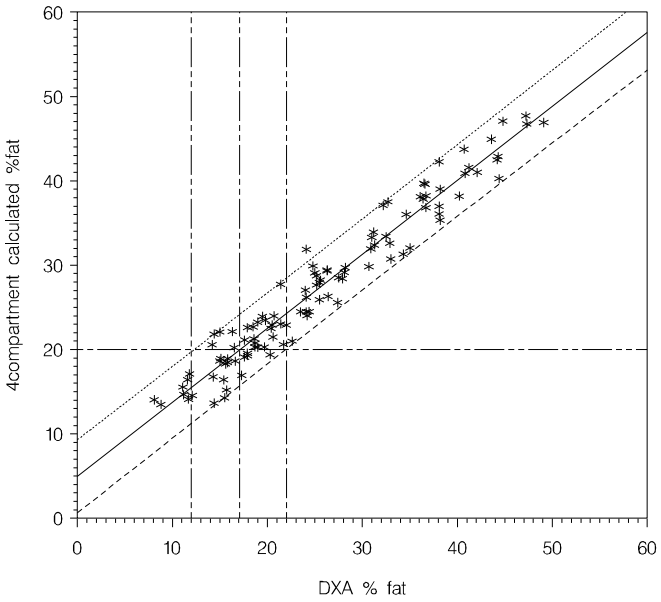

Figure 2: Four-compartment calculated vs. DXA % body fat with the regression line and prediction limits. If the observed value by the 4-compartment method is 20 then we obtain the % body fat by DXA be inverse regression. In this case the corresponding value is 17. The confidence interval is obtained by inverting the prediction interval at the observed value of 20, resulting in $[12, 22]$ in this particular case. The presented values are rounded.

Since we have only one measurement taken by each method, the measurement error variances are not estimable, but we can test for their equality. However, the idea of testing for equality of measurement error variances is not new and was already developed concurrently by Pitman in [12] and Morgan in [11] in 1939. Bland and Altman in their approach use the same idea. Subsequently the same was also emphasized in [9]. We use the fact that if we create the difference between the two random variables that represent the measurements by the two methods, $Y_1 - Y_2$, and their sum, $Y_1 + Y_2$, then the covariance between these two newly created quantities is

$$\text{cov } [(Y_1 - Y_2), (Y_1 + Y_2)] = \sigma_1^2 - \sigma_2^2. \tag{8}$$

Hence, if we regress the difference between the two measurements against their average and test the hypotheses that the slope of the regression line is equal to zero, it is equivalent to testing that the two measurement variances are equal.

Notice that model (7) covers also the case when one of the methods is considered a gold standard. In general, even in that case the bias or the variance of the gold standard should be compared to the newly investigated method, although often this fact is omitted. Some authors introduce simpler models. St. Laurent in [16] investigates comparison of a new method with a gold standard and considers a model

$$X_i = G_i + \epsilon_i, \ i = 1, 2, \ldots, n, \tag{9}$$

where X_i is the new or approximate measure with measurement error ε_i that has mean 0, variance σ_e^2, and is independent of the gold standard measurement; the variable G_i is the gold standard measurement

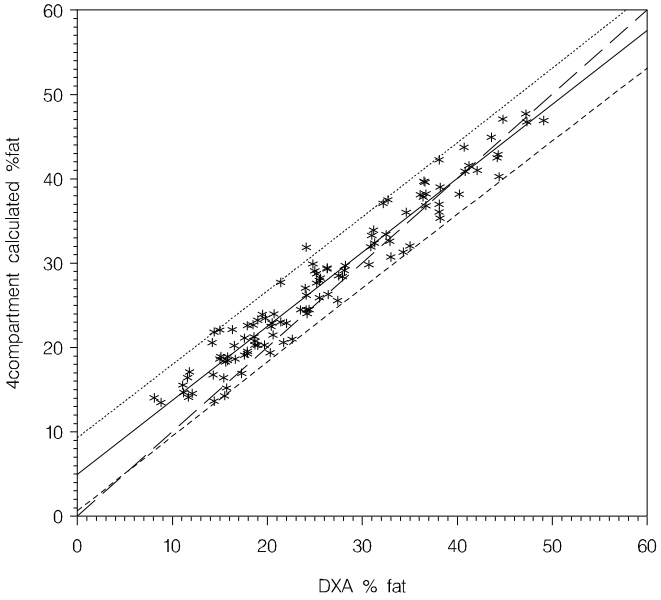

Figure 3: Four-compartment calculated vs. DXA % body fat with the regression line, prediction limits, and line of identity. In a regression the test of the simultaneous hypothesis $H_0 : \beta_0 = 0$ and $\beta_1 = 1$ results in the observed value $F(2, 107) = 59.12$ and p-value $p < 0.0001$, hence the hypothesis is rejected.

with zero mean μ and variance σ_g^2. For the measure of agreement between the two methods he considers the correlation coefficient between the two measurements, which, assuming model (9), is $\rho = \dfrac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}$. Nevertheless, as Bland and Altman pointed out correctly, for assessing agreement between two measurement methods even if one of them is a gold standard, using the correlation coefficient is misleading since it does not take into account the relative bias. We also advocate here against such practice. For comparisons as well as for calibration, regression methods seem to be the most appropriate. In our example, if we assume that the DXA method is the gold standard, then Figure 2 illustrates the use of regression for calibration. A simultaneous test for intercept being zero and slope being one in a regression of the new against the gold standard methods would serve the goal if the need is to compare and/or assess the agreement of the two methods. Notice that in our example the correlation coefficient between the two body fat measures is high, $\rho = 0.975$, and we see that the simultaneous test for intercept $\beta_0 = 0$ and slope $\beta_1 = 1$ is rejected. Figure 3 illustrates the observed regression line and the line of identity.

4. Testing for equality of biases and variances

It is easy to show that if the number of observations by each of the two methods is equal (balanced case), the estimates of the means $\mu+\beta_1$ and $\mu+\beta_2$ do not depend on estimates of the unknown variances and are equal to simple averages of measurements by each method over all units. The test of the hypothesis that $\beta_1 = \beta_2$ against a two-sided alternative based on the estimated difference $\widehat{\beta_1 - \beta_2} = \dfrac{1}{n}\sum_{i=1}^{n}(y_{i1} - y_{i2})$ is the well known paired t-test with $n - 1$ degrees of freedom irrespective of whether the variances σ_1^2 and σ_2^2 are equal. The test for the equality of variances is in this case the well known t-test that tests that the

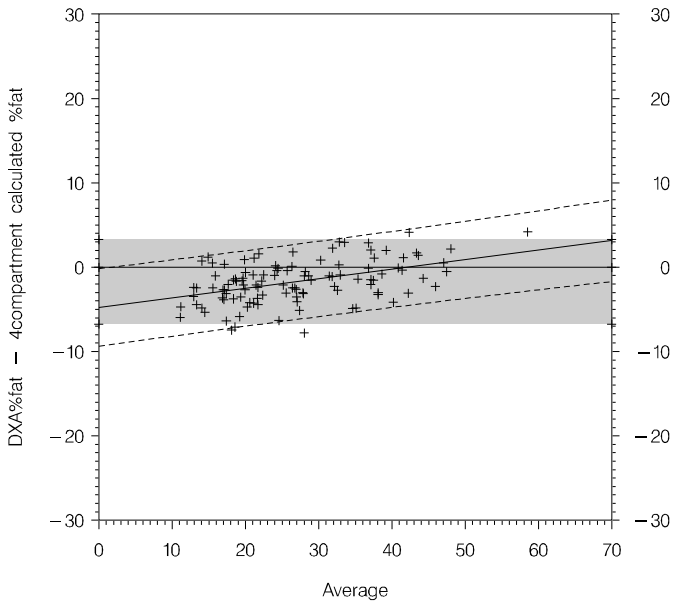

Figure 4: Four-compartment calculated vs. DXA % body fat using the Bland-Altman graph. The difference is plotted against the average of the two methods. The shaded area represents the area of agreement defined by Bland and Altman, the area of the mean of the difference $\pm$ $2\times$ the standard deviation of the difference. The prediction region for the difference is also displayed here.

slope of the regression line in regression of differences against the averages is zero. Figure 4 illustrates the procedure.

5. Generalization for more than two methods comparisons – testing for equality of biases

Testing for equality of biases in model (7) in a general case is not a straightforward procedure. The situation is much more complicated if we deal with more than two methods and more than one observation by each method. Even in a balanced case the test for equality of biases is not exact and there are several approximate F-tests that we may choose from. The model for such a situation takes the form

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, J,\ k = 1, 2, \ldots, K. \qquad (10)$$

Here we have an additional random effect γ_{ij} due to a potential unit-by-method interaction, which is a realistic assumption mainly in chemical laboratories. We assume that γ_{ij} is random, normally distributed with zero mean and variance σ_g^2. The number of measurement methods is J and the number of replications is assumed to be K by each method. We have to test the hypothesis $\beta_1 = \beta_2 = \ldots = \beta_J$. Now we have J measurement error variances, as many as there are methods. The empirical best unbiased estimators of contrasts on biases of, say $\beta_1 - \beta_j$, for all $j = 2, \ldots, J$, are still independent of the unknown variance components. It is interesting to point out that the covariance matrix of the vector of contrasts is independent of the between subject variance σ_a^2 and depends only on σ_g^2 and all the measurement errors variances σ_j^2. We may choose to estimate the variance components in the model by some established method, say restricted maximum likelihood method (REML) and then to use an approximate F-test

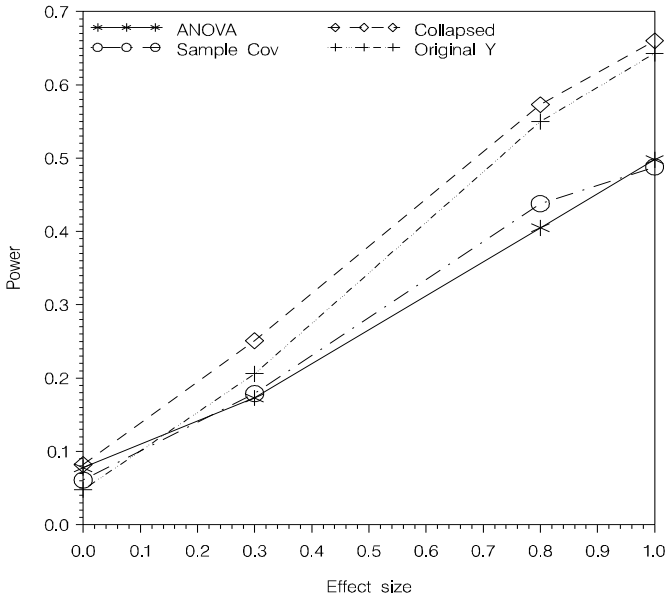

Figure 5: Size and power of four types of tests for equality of biases for four different levels of effect size. $\sigma_g^2 = 0$. For the sample size of 10, two replications by each of four methods, and the desired size of 0.05, the dominating power is for the test (+) that uses restricted maximum estimators (REML) of the variance components and the Fai-Cornelius approximation of the F-test.

based on the approach derived by Fai and Cornelius in [6] or based on Kenward-Roger approximate F-test, see [8]. Another way is to average the observations over replications and ignore the structure of dependencies and use the sample variance-covariance matrix estimator. The most intriguing is the approach that uses the analysis of variance (ANOVA) test as if the variances of all measurement errors were identical. In [18] a simulation method was used to compare several possible tests with respect to their size and power. Figures 5 and 6 below show the behavior of power for two different situations, with the subject-by-method variability being zero and high relative to measurement error variances.

For smaller sample sizes and zero variance $\sigma_g^2 = 0$ the effect of heteroscedasticity seems to be apparent. The test based on the full model with all variance components estimated by REML and then using Fai-Cornelius behaves the best with respect to size and power, but requires lots of computations. In spite of that, in medical but also in the most metrological applications this would be the most recommended approach. On the other hand, when σ_g^2, the subject-by-method variability increases, the effect of heteroscedasticity is washed out and in that case all tests behave very well with respect to their size and power. The recommended test is the ANOVA-based test since it is the simplest with respect to computational complexity. This may be used in some applications in chemical laboratories when the measurand and the method may interact and influence the response.

There is more work needed in the case of unbalanced designs where the numbers of replications by respective methods are unequal.

References

[1] D.G. Altman and J.M. Bland (1983). Measurement in Medicine: the Analysis of Method Comparison Studies. *The Statistician*, 32, 307–317.

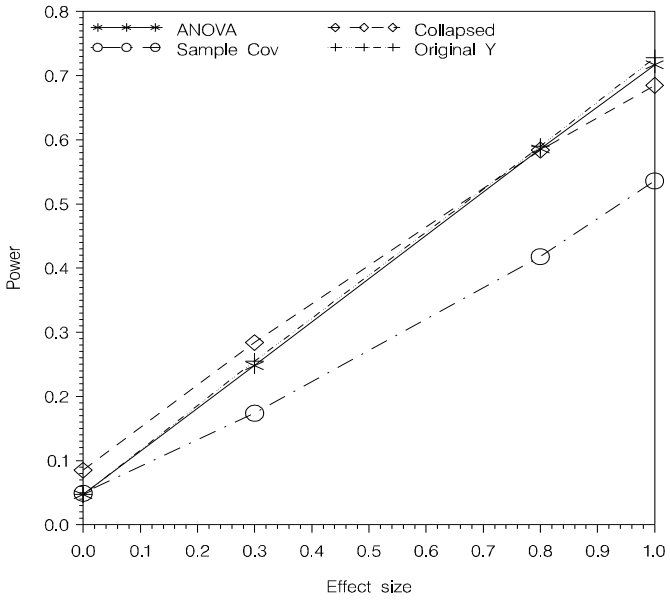

Figure 6: Size and power of four types of tests for equality of biases for four different levels of effect size. $\sigma_g^2 = 5$ is relatively high with respect to measurement error variances and hence the effect of their difference is washed out. For the sample size of 10, two replications by each of four methods, the ANOVA test (*) behaves as well as the computationally complex test that uses REML of variance components and the Fai-Cornelius approximation of the F-test (+).

[2] J.M. Bland and D.G. Altman (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, i, 307-310.

[3] J.M. Bland and D.G. Altman (1995). Comparing methods of measurement: why plotting difference against standard method is misleading. *Lancet*, 346, 1085-7.

[4] G.A. Bray, J.P. DeLany, J. Volaufova, D.W. Harsha, and C.C. Champagne (2002). Prediction of body fat in 12-y-old African American and white children: evaluation of methods. *American Journal of Clinical Nutrition*, 76(5): 980–990.

[5] S.E. Brockwell and I.R. Gordon (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, 20, 825–840.

[6] A.H.-T. Fai and P.L. Cornelius (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *J. Stat. Comput. Simul.*, 54, 363–378.

[7] J. Hartung, H.K. Makambi, and D. Arcac (2001). An extended ANOVA F-test with applications to the heterogeneity problem in meta-analysis. *Biometrical Journal*, 43(2), 135–146.

[8] M.G. Kenward and J.H. Roger (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, 53, 983–997.

[9] A. Kinsella (1986). Estimating Method Precisions. *The Statistician*, 35(4), 421–427.

[10] A.I. Khuri, T. Mathew, and B.K. Sinha (1998). *Statistical tests for Mixed Models*. John Wiley & Sons, Inc., New York.

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

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

[13] P.S.R.S. Rao, J. Kaplan, and W. Cochran (1981). Estimators for the One-Way Random Effects Model with Unequal Error Variances. *Journal of the American Statistical Association*, 76(373), 89 –97.

[14] A.L. Rukhin and M.G. Vangel (1998). Estimation of a Common Mean and Weighted Means Statistics. *Journal of the American Statistical Association*, 93(441), 303 – 308.

[15] S.R. Searle, G. Casella, and C.E. McCulloch (1992). *Variance Components*. John Wiley & Sons, Inc., New York.

[16] R.T. St.Laurent (1998). Evaluating Agreement with a Gold Standard in Method Comparison Studies. *Biometrics*, 54(2), 537–545.

[17] C.Y. Tan and B. Iglewicz (1999). Measurement-Methods Comparisons and Linear Statistical Relationship. *Technometrics*, 41(3),192 – 201.

[18] J. Volaufova and L.R. LaMotte (2003). On Estimation and Testing for Fixed Effects in Two-Way Heteroscedastic Mixed Models. Paper presented at international conference *Statistical Inference in Linear Models: STATLIN' 03*. Będlewo, Poland, August 21 – 27, 2003.

[19] V. Witkovský and G. Wimmer (2003). Consensus mean and interval estimators for the common mean. *Tatra Mountains Mathematical Publications*, 26, 183–194.

[20] P.L.H. Yu, Y. Sun and B.K. Sinha (1999). On Exact Confidence Intervals for the Common Mean of Several Normal Populations. *Journal of Statistical Planning and Inference*, 81, 263 – 277.