

How do I analyse observer variation studies?

This is based on work done for a long term project by Doug Altman and Martin Bland. I recommend that you first read our three Statistics Notes on measurement error (Bland and Altman 1996, 1996b, 1996c), available online via my website.

1. Sources of variation

First we consider different sources of variation. Figure 1 shows three histograms of Peak Expiratory Flow Rate (PEFR) in male medical students: a set of single measurements of PEFR obtained from 58 different students and two sets of 20 repeated measurements of PEFR on single students A and B. The variability for 58 different students is much greater than that shown for students A or B, which are similar. There are two different kinds of variation here: variation within individuals because repeated measurements are not all the same, and variation between individuals because some people can blow harder than others.

We often find it useful to separate the error into different components, which might relate to observers, instruments, etc. For the data of Figure 1, we could model the data as the sum of two variables, $X_i + E_{ij}$, where X_i is the mean for subject i and E_{ij} is the deviation from that mean, or measurement error, of measurement j for that subject. If the variance of the measurement error E_{ij} is the same for all subjects, say σ^2 , the total variance is given by

$$\sigma^2 = \sigma_b^2 + \sigma_w^2$$

where σ_w^2 is the variance of the E_{ij} , the **within-subject variance**, and σ_b^2 is the variance of the X_i , the **between-subjects variance**. This is called a **components of variance** model.

2. Repeatability and measurement error

We first consider the problem of estimating the variation between repeated measurements for the same subject. Essentially, we want to know how far from the true value a single measurement is likely to be. This estimation will be simplest if we assume that the error is the same for everybody, irrespective of the value of the quantity being measured. This will not always be the case, and the error may depend on the magnitude of the quantity, for example being proportional to it.

Measurement error is assumed to be the same for everyone. This is a simple model, and it may be that some subjects will show more individual variation than others. If the measurement error varies from subject to subject, independently of magnitude so that it cannot be predicted, then we have to estimate its average value. We estimate the within-subject variability as if it were the same for all subjects.

Figure 1. Distribution of PEFR for 58 male medical students, with 20 repeated measurements for two students

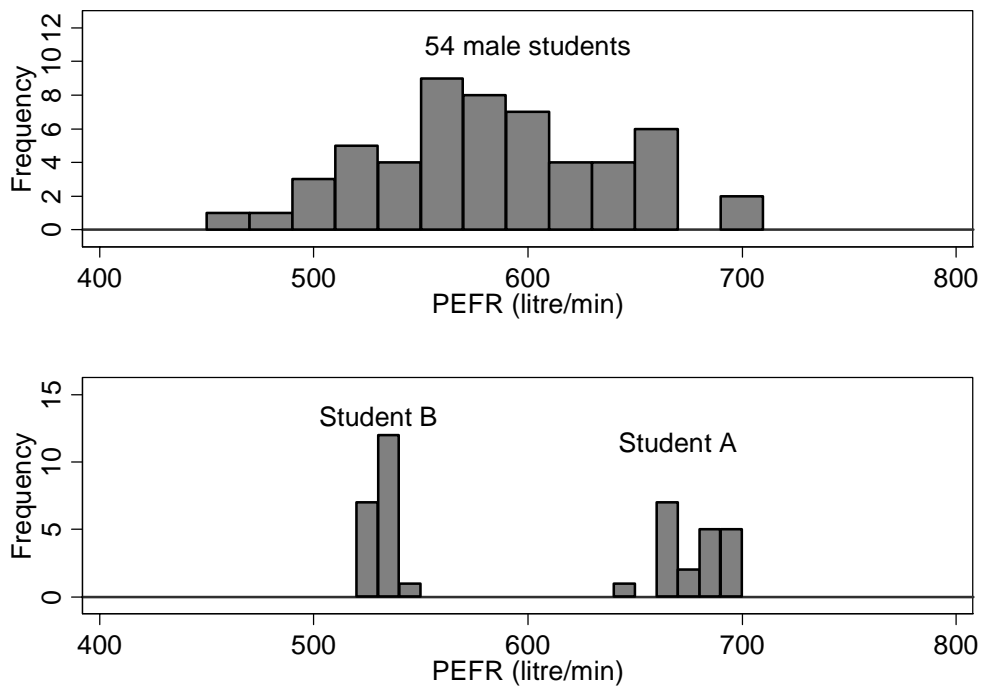


Table 1. One way analysis of variance for the data of Figure 1, Students A and B

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability
Total	39	215484.38			
Students	1	210975.63	210975.63	1778.1	<0.0001
Residual	38	4508.75	118.65		

Consider the data of Figure 1. Calculating the standard deviations in the usual way, we get standard deviations $s_1 = 14.3178$ and $s_2 = 5.6835$ for the two students. We can get a combined estimate as in a two sample t test, which gives us

$$s_w^2 = \frac{(m_1-1)s_1^2 + (m_2-1)s_2^2}{(m_1-1) + (m_2-1)} = \frac{(20-1) \times 14.3178^2 + (20-1) \times 5.6835^2}{(20-1) + (20-1)} = 118.6509$$

where m_1 and m_2 are the numbers of measurements for subjects A and B respectively. The square root of this gives us the within-subjects standard deviation, $s_w = 10.8927$. Rounding we get $s_w = 10.9$ litre/min.

Table 2. Repeated PEFR measurements for 28 school children

Child number	PEFR (litre/min)					Last four readings	
						mean	s.d.
1	180	190	220	200	200	202.50	12.58
2	240	220	200	240	230	222.50	17.08
3	190	240	230	215	210	223.75	13.77
4	260	260	260	240	280	260.00	16.33
5	305	210	300	280	265	263.75	38.60
6	260	260	260	280	270	267.50	9.57
7	270	270	265	280	270	271.25	6.29
8	260	275	270	275	275	273.75	2.50
9	290	280	280	270	275	276.25	4.79
10	270	260	280	280	300	280.00	16.33
11	225	245	290	290	295	280.00	23.45
12	245	275	275	275	305	282.50	15.00
13	250	280	290	300	290	290.00	8.16
14	260	320	290	300	290	300.00	14.14
15	295	300	300	310	300	302.50	5.00
16	250	270	250	330	370	305.00	55.08
17	310	300	310	310	305	306.25	4.79
18	290	300	300	340	315	313.75	18.87
19	270	315	325	330	295	316.25	15.48
20	270	320	330	330	330	327.50	5.00
21	295	335	320	335	375	341.25	23.58
22	255	350	320	340	365	343.75	18.87
23	340	360	320	350	345	343.75	17.02
24	360	330	340	380	390	360.00	29.44
25	380	335	385	360	370	362.50	21.02
26	380	400	400	420	395	403.75	11.09
27	395	400	420	425	420	416.25	11.09
28	385	430	460	480	470	460.00	21.60

Table 3. One way analysis of variance for the data of Table 2

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability
Total	111	397973	3585		
Children	27	365604	13541	35.1	<0.001
Residual	84	32369	385		

In this way we obtain the standard deviation, s_w , of repeated measurements from the same subject, called the **within-subject standard deviation**. We would expect that about two thirds of observations would fall within one standard deviation of the subject's true value and about 95% within two standard deviations. If errors (differences between the observations and the true value) follow a Normal distribution, then we can formalise this by saying that we expect 68% of observations to lie within one standard deviation of the true value and 95% within 1.96 standard deviations. (I discuss the assumption of a Normal distribution in Sections 5 and 6).

This is the same as the residual mean square in one-way analysis of variance (ANOVA) and programs for one-way ANOVA may be used for the calculation, the subjects being the 'groups'. Table 1 shows the ANOVA table for students A and B. The estimate of within-subjects variance is the residual mean square, 118.65. The square root of this is the within-subjects standard deviation, $s_w = 10.89$ litre/min, as before. Another example is shown in Tables 2 and 3. We get $s_w = \sqrt{385} = 19.63$ litre/min.

3. Observer variation studies

The study of observer variation and observer agreement is straightforward in principle, but in practice is one of the most difficult areas in the study of clinical measurement. In principle, what we want to know is whether measurements taken on the same subject by different observers vary more than measurements taken by the same observer, and if so by how much. All we need to do is to ask a sample of observers, representative of the observers whose variation we wish to study, to make repeated observations on each of a sample of subjects, the order in which observers make their measurements being randomized. We then ask by how much the variation between measurements on the subject is increased when different observers make these measurements.

To estimate the increase in variation when different observers are used, we use analysis of variance. We assume that the effects of subject, observer and measurement error are added. The observer may have a bias, a fixed effect where one observer consistently measures high or low. There may also be a random effect, which we will call the heterogeneity, where the observer measures higher than others for some subjects and lower for others. The final measurement is made up of the overall mean, the difference from the mean for that particular subject, the difference from the mean for the observer, and the measurement error. The statistical model for measurements by different observers can be written:

$$Y_{ijk} = \mu + X_i + O_j + H_{ij} + E_{ijk}$$

where Y_{ijk} is the observed value, μ is the mean of all possible measurements, X_i is difference between μ and the mean value for subject i , O_j is the difference between μ and the mean value of observations by observer j (the observer's bias), H_{ij} is the heterogeneity, i.e. the extra variation in subject i due to observer j , the subject times observer interaction, and E_{ijk} is the error in measurements by a single observer on a single subject.

The meaning of heterogeneity may be obscure, and a thought experiment may make it clearer. In the film *10* (Edwards 1979), Dudley Moore scores sexual attractiveness out of ten. Suppose we wish to estimate the observer variation of this highly subjective measurement. We persuade several observers to rate several subjects, and repeat the rating the several times. Now there will be an overall mean rating, for all subjects by all observers on all occasions. This mean is μ . Some subjects will receive higher mean scores than others. The difference of the subject mean from the overall mean is X_i . If we get the same observer to rate the same subject several times, the ratings will vary. The difference between the individual measurement and the mean for that observer's measurement of that subject is the measurement error, E_{ijk} . Some observers will be more generous in their ratings than others. The difference of the observer mean from the overall mean is O_j . For a given observer, this is the bias, the tendency to rate high or low. What about the heterogeneity? It is well known that people tend to be attracted to partners who look like them. Tall, thin women marry tall, thin men, and short, fat men marry short, fat women, for example. (Take a good look at your friends if you don't believe us.) Thus Bland, who is short, may give higher ratings to short women than to tall ones, and Altman, who is tall, may give higher ratings to tall women than to short, *even though their overall mean ratings may be the same*. This is the heterogeneity, or observer times subject interaction, and it may be just as important as the observer bias. H_{ij} is the difference between the mean rating for subject i by observer j and the mean we would expect given the mean rating

over all observers for subject i and the mean rating over all subjects by observer j . Physical measurements can behave in the same way. Measured blood pressure is said to be higher when subject and observer are of opposite sex than when they are the same sex. If both observers and subjects include both sexes, this will contribute to heterogeneity. In general, there may be unknown observer and subject factors which contribute to heterogeneity and our method of analysis must allow for their presence.

The variables X_i , O_j , H_{ij} , and E_{ijk} all represent deviations from a central value and so have mean zero. We shall assume that they follow Normal distributions, are independent of one another, and denote their variances by σ_b^2 , σ_o^2 , σ_h^2 , and σ_w^2 respectively. Thus σ_b^2 is the variance between subjects, i.e. between the true values for subjects, σ_o^2 is the variance between observers, σ_h^2 is the variance between different observers on different subjects, over and above the variance between the average values of the observers and of the subjects, and σ_w^2 is the variance of observations by one observer on one subject. The assumptions that these variables are Normal, independent and have uniform variances are quite strong, particularly that the measurement error variance σ_w^2 is the same for all observers, but as Healy (1989) notes, it is very difficult to proceed without them.

Under the assumption of independent errors, the variance of single observed measurements by different observers on different subjects is

$$\sigma^2 = \sigma_b^2 + \sigma_o^2 + \sigma_h^2 + \sigma_w^2$$

We can estimate the components of variance by analysis of variance, which is straightforward provided we have repeated measurements by each observer on each subject.

4. Repeated observations by each observer

Table 4 shows measurements of abdominal circumference measured by fetal ultrasound. Four observers each made three measurements on three patients. This is an extract from a larger data set, given for illustration. We have measurements by each observer on each subject, and we can carry out a simple two-way analysis of variance as shown in Table 5.

Each of the numbers in the ‘mean square’ column of Table 5 is an estimate of variance. Analysis of variance is usually introduced as a method of comparing the means of more than two groups, and the name ‘analysis of variance’ may seem rather inappropriate. The application to the comparison of means is called the **fixed effects model**, because we are comparing groups defined by some fixed attribute, such as a particular treatment. For estimation of variances we use the **random effects model**, where the subjects are regarded as a sample of subjects and the observers as a sample of observers. We are interested in estimating the variance of the population of observers, not in comparing the means for these particular observers.

Table 4. Ultrasound abdominal circumference measurements (cm) by 4 observers, data of Chitty

Observer	Subject 1			Subject 2			Subject 3		
1	13.6	13.3	12.9	14.7	14.8	14.7	17.1	17.1	18.3
2	13.8	14.2	13.2	14.9	14.1	14.5	17.2	17.5	17.6
3	13.2	13.1	13.1	14.5	14.2	13.8	16.3	15.2	16.1
4	13.7	13.7	13.4	14.4	14.3	13.6	16.8	16.8	17.5

Table 5. Analysis of variance for fetal abdominal circumference, three subjects and four observers

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability
Total	35	90.4222	2.5835		
Subjects	2	79.9439	39.9719	250.26	<0.0001
Observers	3	3.9089	1.3030	8.16	0.0006
Subs × obs	6	2.7361	0.4560	2.86	0.03
Residual	24	3.8333	0.1597		

Table 6. Expected values of mean squares in a two-way analysis of variance table, random effects model, o observers each measuring n subjects m times

Source of variation	Degrees of freedom	Mean square
Total	$mno-1$	
Subjects	$n-1$	$m\sigma_b^2 + m\sigma_h^2 + \sigma_w^2$
Observers	$o-1$	$mn\sigma_o^2 + m\sigma_h^2 + \sigma_w^2$
Subjects × observers	$(n-1)(o-1)$	$m\sigma_h^2 + \sigma_w^2$
Residual error	$(m-1)no$	σ_w^2

The expected values of the sums of squares are shown in symbolic form in Table 6. We can use Tables 5 and 6 to estimate the components of variance for the data of Table 4. From the residual mean square, we have $\sigma_w^2 = 0.1597$. If we subtract this from the interaction mean square we get $3\sigma_h^2 = 0.4560 - 0.1597$, $\sigma_h^2 = 0.0988$. If we subtract the interaction row from the observers row we get $3 \times 3\sigma_o^2 = 1.3030 - 0.4560$, $\sigma_o^2 = 0.0941$. If we subtract the interaction row from the subjects row we get $4 \times 3\sigma_b^2 = 39.9719 - 0.4560$, $\sigma_b^2 = 3.2930$.

We can now calculate the variance of observations on the same subject by two different observers:

$$\sigma_o^2 + \sigma_h^2 + \sigma_w^2 = 0.0941 + 0.0988 + 0.1597 = 0.3526$$

We can interpret this as in Section 2. We can compare this to the variance of observations on the same subject by the same observer, $\sigma_w^2 = 0.1597$, and we see that the variance is more than doubled when different observers are used. The maximum difference likely between to measurements by the same observer (the repeatability) is estimated as $2.83 \times \sigma_w^2 = 2.83 \times \sqrt{0.1597} = 1.13$. If different observers were employed the corresponding figure is estimated to be $2.83 \times \sqrt{0.3526} = 1.68$.

To estimate the intra-class correlation coefficient for measurements by different observers, we estimate the total variance of measurements by different observers on different subjects: $\sigma^2 = \sigma_b^2 + \sigma_o^2 + \sigma_h^2 + \sigma_w^2 = 3.2930 + 0.3526 = 3.6456$. The ICC is then estimated by $ICC = \sigma_b^2 / \sigma^2 = 3.2930 / 3.6456 = 0.90$. We can also estimate the intra-observer ICC, for measurements by a single observer, to be $\sigma_b^2 / (\sigma_b^2 + \sigma_w^2) =$

$3.2930/(3.2930+0.1597) = 0.95$. This is greater than the inter-observer ICC because using different observers increases the variation.

One of the possible difficulties in this method of analysis is that we may get negative estimates of variance, which are, of course, impossible. We solve this problem by setting any negative estimates to zero.

It is possible to get confidence intervals for the variances and ICCs, but it can be quite complicated. You can ignore the P values in the ANOVA table.

5. Checking assumptions

Although I have continued the discussion of the components of variance above, the first step in the analysis should be to check whether the within-subjects standard deviation is independent of the mean and variation within the subject is approximately Normal. This is best done graphically, as illustrated in Section 6.

6. An example with only one measurement per subject and deviation from assumptions

Moertel and Hanley (1976) made model tumours from 12 solid spheres, arranged in random order on a soft mattress and covered with foam rubber. They then invited 16 experienced oncologists to measure the diameter of each sphere, each observer using the technique and equipment which they routinely used in clinical practice. The data are shown in Table 7.

Figure 2 shows the measured value against the true value, showing that the variability increases with increasing size. (If we did not have the true value, as is usually the case, we could use the subject mean instead of the true value in this plot.) The increasing variation is associated with the marked digit preference in Table 7. The digit preference is very strong, the majority of terminal digits being '0' and none being '9', but is less marked for the small 'tumours' and strongest for the large, all of the terminal digits for 'tumour' 12 being '0'. As Figure 2 also shows, the relationship between variability and diameter disappears after log transformation, apart from a possible reduction in variability for the largest 'tumour'.

Sometimes it is impracticable to take repeated measurements by the same observer on a subject, because knowledge of the first reading would bias the second. The artificial tumour data of Table 7 are like this. We cannot then estimate the measurement error, but we can still get a reasonable estimate of the variability between measurements by different observers on the same subject. We calculate the analysis of variance as in Section 4, but this time we cannot include the observer times subject interaction, because it cannot be estimated separately from single observations. It is included in the residual. The expected mean squares are shown in Table 8.

Following the method of Section 4, we can use Tables 8 and 9 to estimate the components of variance in Table 7. There are 12 subjects (in this case dummy tumours) and 16 observers, so $n = 12$, $o = 16$. From the residual mean square, we have $\sigma_h^2 + \sigma_w^2 = 0.0167$. Subtracting this from the observers mean square we get $12\sigma_o^2 = 0.2320 - 0.0167$, so $\sigma_o^2 = 0.0179$. Subtracting the residual mean square from the subjects mean square we get $16\sigma_b^2 = 6.2387 - 0.0167$, so $\sigma_b^2 = 0.3889$.

Table 7. Measurements of the diameter of 12 model tumours by 16 observers (Moertel and Hanley 1976)

		'Tumour' number and diameter (cm)}											
Observer	1	2	3	4	5	6	7	8	9	10	11	12	
1	1.8	2.3	2.7	3.5	4.3	4.4	5.3	5.5	6.4	7.5	8.6	14.5	
2	1.3	1.8	2.5	3.0	4.0	4.0	5.0	6.5	5.0	7.0	7.5	15.0	
3	1.5	2.0	3.0	4.0	4.5	3.0	5.0	4.5	5.0	6.5	7.5	13.0	
4	1.5	2.0	2.5	3.5	4.0	4.0	5.5	5.0	6.0	7.0	8.0	15.0	
5	1.5	1.8	2.5	3.3	4.0	3.8	4.7	4.5	5.0	6.4	7.0	13.0	
6	1.8	2.0	3.0	4.0	4.5	5.0	5.5	6.0	6.0	6.5	9.0	15.0	
7	1.8	2.3	3.0	4.0	4.5	5.0	5.5	5.5	6.0	7.5	7.5	15.0	
8	1.0	2.0	2.5	3.0	4.0	3.5	4.0	5.0	4.0	6.0	6.0	15.0	
9	1.5	2.0	3.0	4.0	5.0	3.5	7.0	5.5	6.5	6.5	4.0	19.0	
10	1.6	2.0	2.5	3.5	4.0	4.5	5.5	5.5	5.5	7.5	9.0	14.0	
11	1.2	1.5	2.0	3.0	4.0	5.4	4.5	6.0	6.0	7.0	7.5	15.0	
12	1.5	2.2	3.0	4.3	4.2	5.2	5.5	5.4	6.0	7.5	10.5	18.0	
13	1.5	1.5	2.0	3.0	4.5	3.0	6.0	4.0	4.5	4.5	8.5	14.0	
14	2.0	2.0	3.0	3.5	4.0	4.5	5.5	5.5	6.0	7.0	8.0	15.0	
15	1.6	2.2	2.8	3.8	4.5	5.6	5.3	5.4	6.2	6.8	8.1	14.0	
16	0.8	1.0	1.5	2.0	3.0	2.0	4.0	3.0	4.0	5.0	6.0	12.0	
16	1.7	1.8	2.2	3.5	3.8	3.8	4.8	4.6	5.0	6.0	6.0	13.0	

Figure 2. Measurements by 16 observers of the diameter of 12 model tumours, against true diameter

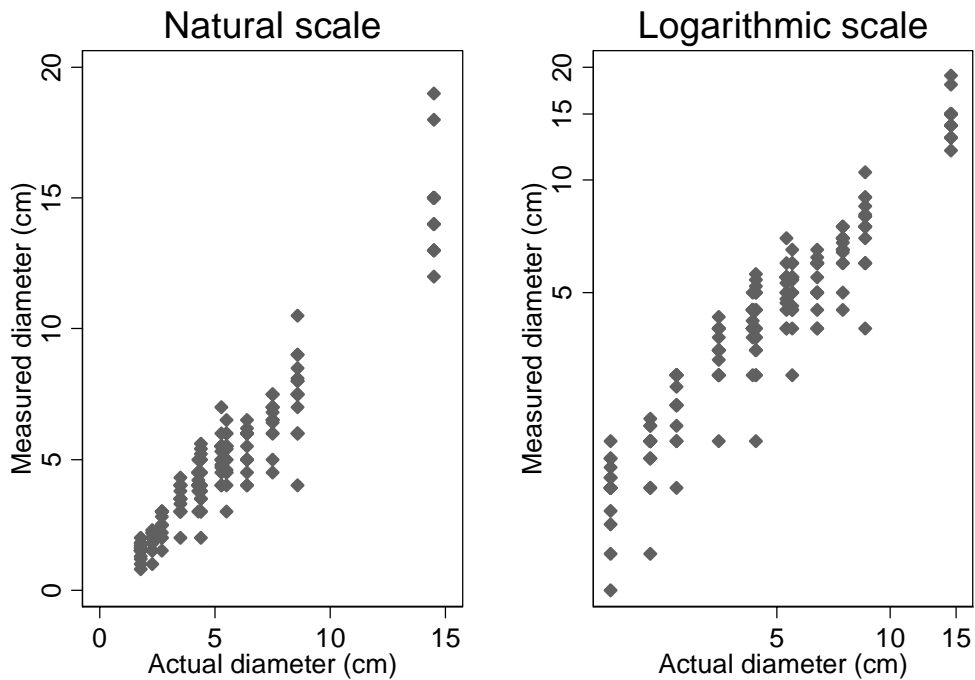


Table 8. Expected values of sums of squares in a two-way analysis of variance table without replicated measurements, random effects model

Source of variation	Degrees of freedom	Mean square
Total	$no-1$	
Subjects	$n-1$	$o\sigma_b^2 + \sigma_h^2 + \sigma_w^2$
Observers	$o-1$	$n\sigma_o^2 + \sigma_h^2 + \sigma_w^2$
Residual error	$(n-1)(o-1)$	$\sigma_h^2 + \sigma_w^2$

Table 9. Analysis of variance for log 'tumour' diameter

Source of variation	Degrees of freedom	Sum of squares	Mean square	Variance ratio (F)	Probability
Total	191	74.8586	0.3919		
Subjects	11	68.6254	6.2387	373.95	<0.0001
Observers	15	3.4806	0.2320	13.91	<0.0001
Residual	165	2.7527	0.0167		

We can now estimate the variance of observations on the same subject by two different observers:

$$\sigma_o^2 + \sigma_h^2 + \sigma_w^2 = 0.0179 + 0.0167 = 0.0346$$

We can interpret this variance estimated on the logarithmic scale as described by Bland and Altman (1996c). The standard deviation is $\sqrt{0.0346} = 0.186$. The antilog is $\exp(0.186) = 1.20$. The coefficient of variation is found by subtracting 1, giving $1.20 - 1.00 = 0.20$ or 20%. The inter-observer coefficient of variation is thus 20%.

The variance between subjects is estimated by $16\sigma_b^2 = 6.2387 - 0.0167$, $\sigma_b^2 = 0.3889$. The inter-observer ICC might be estimated by $\sigma_b^2 / (\sigma_b^2 + \sigma_o^2 + \sigma_h^2 + \sigma_w^2) = 0.3889 / (0.3889 + 0.0346) = 0.92$. However, as this is not a sample of subjects but a set of true values predetermined by the investigator, correlation is not appropriate and we should not quote it in this particular case.

We can estimate the between-observer variation from this design, but as we cannot estimate the within-observer measurement error, we cannot say by how much observer variation increases the error.

References

- Bland JM, Altman DG. (1996) Measurement error. *British Medical Journal*, **313**, 744.
- Bland JM, Altman DG. (1996b) Measurement error and correlation coefficients. *British Medical Journal*, **313**, 41-2.
- Bland JM, Altman DG. (1996c) Measurement error proportional to the mean. *British Medical Journal*, **313**, 106.
- Edwards B. (1979) *10*. Los Angeles, Orion Pictures Corporation.
- Healy MJR. (1989) Measuring measuring errors. *Statistics in Medicine* **8**, 893-906.
- Moertel CG and Hanley JA. (1976) The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer* **38**, 388-394.

J. M. Bland
May 2004