Measurement, Measurement Errors, and Measuring Measurement Errors in Clinical Medicine

1. Background

In many circumstances clinical decisions will be guided by measurements of some kind. Common examples include blood chemistry, such as serum sodium or serum potassium, haematological variables such as haematocrit or haemoglobin concentration, measures of physiological function such as glomerular filtration rate, or peak expiratory flow. Anthropological variables such as height, height velocity and skinfold thickness are of importance in paediatrics and sometimes in nutritional assessments.

All of the above are examples of quantitative, or continuous variables. That is they can take any value within their range. Binary variables, such as whether or not a patient is infected are also of importance but will not be considered here.

In all these cases, the process of obtaining the measurement will not be entirely satisfactory, that is, the *observed* measurement will be subject to measurement error. This can arise for many reasons, such as the degree of skill of the measurer (e.g. for height measurement), due to slight differences in the way the patient hold the peak flow meter when assessing expiratory flow, or even because of intrinsic randomness, as when measurements are based on counts of radioactive markers, as with some methods of assessing glomerular filtration rate or fat-free mass. However they arise, it is important that the clinician knows how precisely the *observed* value reflects the underlying true value, as this can influence the uses of and trust in the measurement.

2. Model for Measurement Errors

The idea is to postulate the existence of a "true" underlying quantity, such as the "true" height, peak flow, serum sodium, etc. and that the observed value is the true value perturbed by the error of measurement i.e.:

$$\textit{observed value} = \text{true value} + \textit{error}$$

The distribution of this combination can be illustrated by the distributions in figure 1, based on the heights of seven-year-old boys: the spread of the true value is measured by a standard deviation σ_B and the spread of the error distribution is σ, shown here as a perturbation around the mean of the true distribution. It follows from this that an observed value is likely (with 95% confidence to be strict) to be within $\pm 2\sigma$ of the true value.

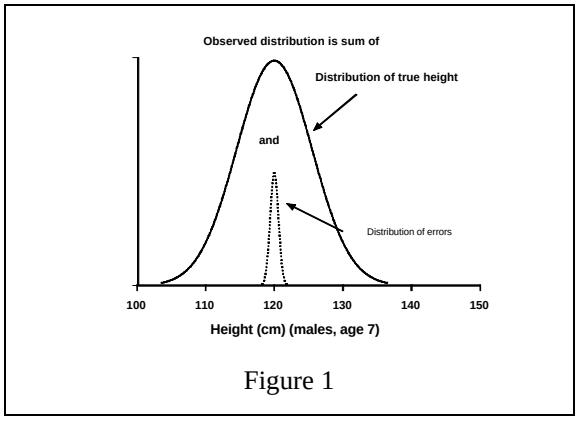

Figure 1

Effect of measurement error on the mean.

If the error is symmetrically distributed about 0, then the measurement is said to be *unbiassed* or *accurate*. In this case the mean of a large number of observations on the same individual will be close to the true value.

Effect of measurement error on the spread.

Clearly measurement error adds an extra source of variability and means that the standard deviation of the observed values will inevitably be larger than that of the true values. In fact, it can be shown that the standard deviation of the observed values is $\sqrt{\sigma_B^2 + \sigma^2}$, which is clearly larger than σ_B.

So, if a collection of measurements is made and the standard deviation is calculated then it is $\sqrt{\sigma_B^2 + \sigma^2}$ which is estimated. In many circumstances it would not be appropriate to try to estimate the 'true' standard deviation σ_B because it is the spread of the *observed* values that is important and allowance must be made for the contribution of measurement error. This would be done simply by calculating the standard deviation of the observed values. However, in some circumstances it may be of importance to know how much of the spread in the observed values is due to measurement error. If most of the variability in a measurement is due to measurement error, then this may limit the uses to which the variable can be put or the reliance that a clinician would place on the observation. As such it is useful to be able to estimate σ.

{*It must be admitted that the existence of a "true" value needs deeper consideration than it has received here, insofar as it is necessarily unobservable. The approach here is heuristic and is a model adopted to allow the problem of measurement errors to be analysed.*}

3. Measuring Measurement Errors

If the model "*observed value* = true value + *error"* is adopted then it is clear that from a single value it is impossible to disentangle how much is error and how much is "true value". If two independent measurements on the same individual are available then progress can be made* because both measurements contain the same "true value" but different measurement errors, i.e.

$$\textit{observed value}_1 = \text{true value} + \textit{error}_1 \qquad \textit{observed value}_2 = \text{true value} + \textit{error}_2$$

so that the difference *d* in the pair of measurements depends only on error. Consequently the standard deviation of a set of these differences *d* cannot depend on the spread of the true values σ_B and must depend only on σ. It turns out that the standard deviation of the difference between two independent errors, each with standard deviation σ, is $\sqrt{\sigma^2 + \sigma^2} = \sigma\sqrt{2}$, so σ can be estimated by computing the standard deviation of a series of differences of replicate measurements and then dividing this by $\sqrt{2} \approx 1.42$.

What are "independent measurements" and why are they needed?

The pairs of replicates needed above were stated to be *independent*. This is because it is essential that each member of the pair be affected by error in exactly the same way as if each were just a single measurement being made *de novo*. Problems can arise if the way the replicate measurements were obtained means that the result of the first measurement influences the value of second. In this way the two measurement become artificially close and the value of σ is underestimated.

In first two practicals the replicate height measurements were made close in time and knowledge of the first value may well influence the value of the reported second measurement (e.g. if the first attempt at a second measurement is 'very different' from the first measurement then the second measurement may be repeated). In the second two practicals, the minimetre was moved between the replicates, so that

* more than two measurements helps but this is ignored to avoid technicalities

a) the time interval between making the measurements was increased and b) the second measurement would differ by the (unknown) offset in the two attachment points, so it was no longer known what value on the second reading would reproduce the first reading.

An example: errors in height measurement.

Errors in Height Measurement.

In the appendix are data on replicate height measurements from the two practicals. While it is clear that male and females have different mean heights (males 180.57 cm, females 167.10 cm), the difference between replicate measurements should not differ between males and females. This is because the difference between heights measures the measurement error on height, and this is more likely to be an attribute of the measurer not the measured. Even if there is a difference between males and females in their ability to measure height (and there is no reason to suppose that there is) we would of course had to have recorded the sex of the measurer not the measured.

Consequently, the differences between replicate height measurements from practicals 1 and 2 for males and females can be pooled. The standard deviation (SD) of these differences is 0.258 cm, and dividing by $\sqrt{2}$ gives an SD for the errors of 0.18cm. We can also pool male and female data from practicals 3 and 4 and these differences have an SD of 0.566, giving an SD for the errors of 0.40cm. Thus the error in the second two practicals, where the second measurement could not be adjusted in the light of the first measurement, is over twice as large as the value obtained when such adjustments were possible.

The method of data collection in practicals 1 and 2 suggests that the replicates obtained will not be independent, and hence that the estimate of error obtained would be an underestimate. This is certainly compatible with the results obtained, suggesting that it is the estimate from the last two practicals that gives a fairer representation of the error SD.

Effect of Error on Measurement of Height

Growth charts tell us that in the UK seven-year-old boys have a mean height of 121 cm, with 94% of children between 110 cm and 131 cm. If we assume that the above error would also apply when measuring children (and this is a conservative assumption, it is likely to be larger for children), then a seven-year-old boy of mean height could have height between $121 \pm 2 \times 0.40 \cong 120.2, 121.8$, which means the child would be between the 44th and 56th centile, rather than on the 50th centile. This kind of discrepancy hardly matters, so it seems that even inexperienced observers do not add an important component of error to measurements of height.

Effect of Error on Height Velocity

The monitoring of children referred to specialist clinics because of short stature (height) depends to some extent on evaluating and interpreting their *height velocity*. The height velocity is simply a measure of the rate at which the child is growing and is calculated from two height measurements H_1 and H_2 made at different times, t_1 and t_2 respectively. The velocity v is then defined as

$$\frac{H_2 - H_1}{t_2 - t_1}$$

and as with height, centile charts exist to permit the interpretation of this quantity. This quantity is usually expressed in cm/yr. For a boy measured at ages 6.5 and 7.5 years, the mean growth in this time (the one year growth velocity) is 5.8cm, with 94% of the population falling between 4.2 cm and 7.4 cm.

This velocity can be written as :

$$\begin{gathered}(\text{true height}_2 + \text{error}_2) - (\text{true height}_1 + \text{error}_1) \\ = \\ (\text{true height}_2 - \text{true height}_1) + \text{error}_2 - \text{error}_1 \\ = \\ \text{true velocity} + \text{error in velocity}\end{gathered}$$

so the error in velocity is simply the difference in the errors of the constituent heights. It can be shown that if each error in height has standard deviation σ, then the difference of two such errors has standard deviation $\sigma\sqrt{2}$. If the errors are as in the above then this is 0.57cm, so a calculated velocity of 5.8 cm/yr may be in error by $\pm 2\times$ 0.57, that is between 4.7 and 7.0 cm (approximately), that is between the 10th and 90th centile! So, when the errors are applied to a velocity, they are very substantial, and imply that the result cannot reliably place a child on the velocity chart. Even in experienced hands the error in the velocity rarely gets below 0.15 cm. Although this reduces the uncertainty in the above centile range to between the 30th and 70th, it is clear that errors of measurement have a much more substantial impact on the assessment of height velocity than they do on the assessment of height itself.

However, if the velocity is assessed over only 3 months then 94% of the resulting growth increments will be between 1.0cm and 1.8cm. However the standard deviation of the error in these increments will remain at 0.25cm (at best), so over this short period error of measurement has a dominant role. In fact, it is inadvisable to attempt to assess a growth velocity based on measurements taken less than six months apart.

{further details on the assessment of growth can be found in Brook, CGD (1982), *Growth Assessment in Childhood and Adolescence*, Blackwell: more on the assessment of error can be found in Voss LD, *et al*., *Archives of Disease in Childhood*, 1990, volume 65, pages 1340-1344.}

What can be done about large measurement errors?

In the case of height and height velocity measurements the remedy to large measurements errors is to try to reduce σ. This can be done by training those who measure height so that their technique improves. Another approach is to change the measurement made to one with a lower σ but which carries a similar clinical message: in the case of height velocity, the technique of measuring very small changes in certain bones in the leg, knemometry, is under development.

In measurements of some variables it may be impossible to resort to such remedies. In these cases making several replicate measurements and using their mean is one alternative. It was pointed out above that the standard deviation of a single measurement is $\sqrt{\sigma_B^2 + \sigma^2}$, and if σ is a substantial proportion of σ_B then measurement error is making a substantial contribution to the total variability. If the mean of two *independent* measurements is used instead, then this has standard deviation $\sqrt{\sigma_B^2 + \frac{1}{2}\sigma^2}$; so the device of taking replicate measurements has reduced the contribution of measurement error to the total standard deviation. If *n* replicates were

taken then the standard deviation would have been $\sqrt{\sigma_B^2 + \frac{1}{n}\sigma^2}$. The number of replicates that is it sensible to use depends on the relative sizes of σ and σ_B: after a point taking more and more replicates becomes wasteful, as replication will never reduce the true variability of the measurement.

Appendix: raw data

Duplicate height readings from first two practicals (cm)

Male			Female		
1st reading	2nd reading	difference	1st reading	2nd reading	difference
170.9	171.1	0.2	170.6	170.8	0.2
168.8	168.5	-0.3	175.3	174.8	-0.5
180.5	180.5	0.0	162.9	162.9	0.0
180.5	180.6	0.1	162.2	162.2	0.0
177.2	177.7	0.5	168.2	167.5	-0.7
180.1	179.9	-0.2	169.1	169.0	-0.1
191.8	191.7	-0.1	168.2	168.0	-0.2
175.7	175.7	0.0	165.0	164.7	-0.3
174.9	174.8	-0.1	159.8	159.6	-0.2
187.5	187.7	0.2	162.9	163.0	0.1
180.0	180.0	0.0	168.2	167.9	-0.3
179.4	179.0	-0.4	167.0	167.0	0.0
185.9	186.0	0.1	169.9	170.0	0.1
177.5	177.6	0.1	177.1	177.1	0.0
178.9	178.7	-0.2	169.4	168.4	-1.0
175.5	175.9	0.4	166.1	166.0	-0.1
178.5	178.4	-0.1	175.2	175.1	-0.1
182.6	182.4	-0.2	168.0	168.2	0.2
170.2	170.3	0.1	172.6	172.8	0.2
174.9	174.8	-0.1	171.7	171.4	-0.3
182.0	182.0	0.0	170.4	170.2	-0.2
			166.9	166.5	-0.4
			162.6	162.6	0.0
			154.4	154.4	0.0
			168.2	168.4	0.2
			165.6	165.6	0.0
			157.2	157.1	-0.1
			164.0	164.0	0.0

Duplicate height readings from second two practicals (cm)

Male			Female		
1st reading	2nd reading	difference	1st reading	2nd reading	difference
183.3	183.2	-0.1	170.0	169.8	-0.2
179.2	178.6	-0.6	161.1	161.2	0.1
185.8	185.7	-0.1	173.2	173.4	0.2
200.0	200.0	0.0	167.6	168.7	1.1
186.1	186.3	0.2	166.0	166.3	0.3
172.2	172.4	0.2	170.4	170.2	-0.2
193.9	193.8	-0.1	159.0	159.2	0.2
180.4	180.6	0.2	167.5	166.4	-1.1
178.5	178.0	-0.5	171.2	171.6	0.4
184.5	184.5	0.0	163.2	163.8	0.6
185.5	186.4	0.9	179.2	179.9	0.7
173.9	174.3	0.4	174.2	174.0	-0.2
172.1	172.3	0.2	167.0	167.4	0.4
194.5	194.7	0.2	161.3	161.4	0.1
177.4	178.0	0.6	168.8	169.7	0.9
			168.4	168.3	-0.1
			172.0	172.2	0.2
			167.7	168.0	0.3
			167.5	167.4	-0.1
			158.7	159.0	0.3
			155.1	153.0	-2.1