**University of York Department of Health Sciences**

# Measurement in Health and Disease

# Measurement error

## Accuracy and precision

In this lecture we shall consider the problem of the precision and repeatability of measurements which are numerical variables such as blood pressure and forced expiratory volume (FEV). We shall look at how good a measurement is from the clinical point of view, for giving us information about the individual subject or patient. We also look at the repeatability of measurement methods from the point of view of the researcher, that is how good a method is at telling us something about the population.

We shall have a lot to say about 'error', a word which comes from a Latin root meaning 'to wander'. In statistics we use the term **error** to mean the variation of observations or estimates about some central value. If we make several measurements of FEV on subject, they will not all be the same, because the subject cannot blow in exactly the same way each time. This variation is called error. It is not the same as a mistake, and does not imply any fault on the part of the observer. A measurement mistake might be if we transpose digits in recording the FEV, writing 9.4 litres instead of 4.9.

We will first distinguish precision and accuracy. A measurement is **precise** if repeated observations of the same quantity are close together. It is **accurate** if observations are close to the true value of the quantity. Thus a measurement can be precise without being accurate, but cannot be accurate without being precise. In this lecture I shall be concerned with precision.

## 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. The upper histogram shows a sample of single measurements of PEFR obtained from 54 different students, whereas the lower histograms each show 20 repeated measurements of PEFR on a single student Table 1. The variability between students shown in the upper histogram is much greater than that shown within the same student shown in the lower histograms. 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 measure PEFR for several reasons: for example, to compare a patient's PEFR to a reference interval for diagnostic purposes, to monitor changes in lung function over time, or to compare two groups of subjects as in a clinical trial or epidemiological study. In each case, we want to be sure that the variation between measurements, the within-subject variation, does not swamp the difference for which we looking. Because PEFR is known to have high variation between measurements, it is customary to make several observations to achieve this, and use their mean or maximum. The latter is used because of the special nature of this measurement, the maximum rate of flow which the subject can achieve.

**Figure 1. Distribution of PEFR for 54 male medical students, with 20 repeated measurements for two students**

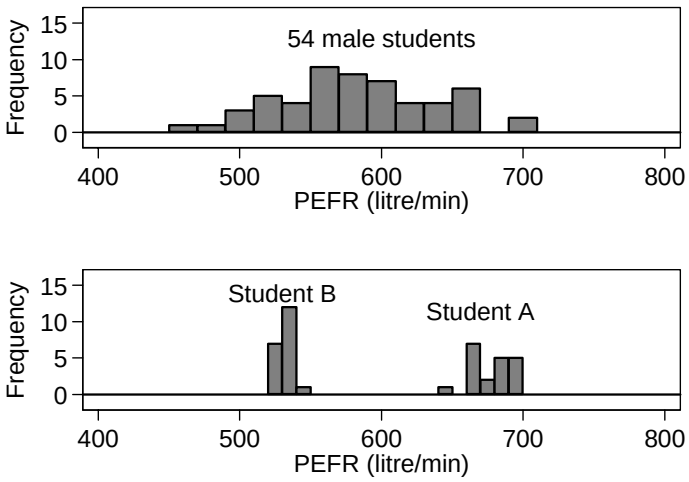



**Table 1. Repeated PEFR (litre/min) measurements for two male medical students**

| Student A | | | | | Student B | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 685 | 695 | 660 | 660 | 690 | 530 | 535 | 530 | 535 | 525 |
| 690 | 665 | 665 | 685 | 680 | 530 | 520 | 530 | 525 | 520 |
| 675 | 660 | 660 | 670 | 690 | 525 | 535 | 520 | 535 | 535 |
| 685 | 645 | 660 | 690 | 680 | 530 | 525 | 530 | 540 | 530 |

If we suppose that a subject has a true PEFR, which would be the mean of all possible measurements, then the difference between an individual measurement and the true value is its error. Many factors could influence this error. We would expect that a series of PEFR measurements made on a subject by different observers at different times spread over six months would vary more than a series over one morning by one observer. We might be interested in different types of variability for different purposes. Monitoring short term changes in blood pressure in a single patient requires one type of error, interpreting random blood pressure in a screening clinic another. In the first case, we are detecting shifts in mean blood pressure over a short period of time, in the second we are determining from one or two measurements whether the subject's mean blood pressure is above some cut-off point such as 90mm Hg diastolic. Thus we need to define what we mean by measurement error rather carefully. The British Standards Institution (1979) considered this question for laboratory measurements, and made the distinction between **repeatability**, incorporating variability between measurements made by the same operator in the same laboratory, and **reproducibility**, incorporating variability between measurements made by different operators working in different laboratories. The same considerations arise when we have complex measurements such as assays, where we might have the error estimated separately for different stages in the measurement, giving an intra-assay or within-assay error and an inter-assay or between-assay error. For the first we would take repeated readings from the same assay and estimate their error, and for the second we would take repeated assays on the same subject.

Sometimes we are able to separate the effects of the different sources of variation and sometimes not. In this lecture we describe techniques for estimating the variability between methods which work whether the measurements are all made by one observer on the same

occasion, or made by different observers on different occasions, or made repeatedly by the subjects themselves. We discuss studies where the same group of observers are used to measure several subjects in the next lecture.

## 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.

## The within-subject standard deviation, $s_w$

We start with the case where the 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.

Consider the data of Table 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 averaged over the two students. We actually average the variances, the squares of the standard deviations, allowing for possibly different samples sizes. It is the same method as used in a two sample t test. We get

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

where $m_1$ and $m_2$ are the numbers of measurements for subjects 1 and 2 respectively. The square root of this gives us the within-subject standard deviation, $s_w = 10.8927$. Rounding we get $s_w = 10.9$ litre/min.

In this way we obtain the standard deviation, $s_w$, of repeated measurements from the same subject, called the **within-subject standard deviation**. As for any standard deviation, we expect that about two thirds of observations will fall within one standard deviation of the mean, 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.

**Table 2. Repeated PEFR measurements for 28 school children**

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

For more than two subjects we could calculate the within-subject standard deviation, $s_w$, by extending this formula:

$$s_w^2 = \frac{(m_1 - 1)s_1^2 + (m_2 - 1)s_2^2 + (m_3 - 1)s_3^2 + \cdots + (m_n - 1)s_n^2}{(m_1 - 1) + (m_2 - 1) + (m_3 - 1) + \cdots + (m_n - 1)}$$

where $n$ is the number of subjects. We do not do this in practice, but use a method called one way analysis of variance, described below. For our purposes, it is the concept of the within-subject standard deviation which is important, not the mechanics of it.

Table 2 shows data taken from a larger study of lung function in schoolchildren. Five PEFR readings were made for each child. The first reading was treated as practice blow and ignored. Table 2 shows the second, third, fourth and fifth readings, which we shall use to estimate the repeatability of PEFR in 12 year old schoolchildren. Table 2 also shows the mean and standard deviation of the last four readings for each subject. For the common within-subject standard deviation, we have $s_w$ = 19.6 litre/min. This large variability in PEFR is well known and so individual PEFR readings are seldom used. In this study the variable used for analysis was the mean of the last three readings.

**Table 3. Analysis of variance table for the data of Table 2**

| Source | Sum of squares | Degrees of freedom | Mean Square | F ratio | P |
|---|---|---|---|---|---|
| Subject | 365604.24 | 27 | 13540.90 | 35.14 | 0.0000 |
| Residual | 32368.75 | 84 | 385.34 | | |
| Total | 397972.99 | 111 | 3585.342 | | |

## Analysis of variance

As its name suggests, analysis of variance (or 'anova') is a technique for estimating variances. It has many other uses, but in the study of measurement error it is used for its original function. We calculate a sum of squares for the repeated observations for each subject. This is the sum of squares about the subject mean. We add them together to get the combine sum of squares about the subject mean , or within-subject sum of squares. From this we get an estimate of the variance within the subjects, dividing the sum of squares by its degrees of freedom. We can also calculate a sum of squares and hence a variance for the subject means. These sums of squares are set out in an analysis of variance table (Table 3). Here the 'Subject' row is the variation between subjects and the 'Residual' row represents the variation within the subjects.

There are several things we can note about this table. The sum of squares add up, i.e. the subject and residual rows add up to give the total row. The degrees of freedom add up in the same way. We had 28 subjects, 4 observations on each. The degrees of freedom for subjects are given by 27 = 28 – 1. The degrees of freedom for the residual, i.e. within subjects, are given by 84 = 28 × (4 – 1). Each of the 28 subjects contributes 3 = 4 – 1 degrees of freedom within the subject. For the total, there are 112 observations, which gives 111 = 28 × 4 – 1 degrees of freedom. The mean squares are the sums of squares divided by the degrees of freedom: 13540.90 = 365604.24/27 and 385.34 = 32368.75/84. These are estimates of variance.

We do not need the F ratio. The F ratio or variance ratio is the ratio of the mean squares, in this case the mean square between subjects divided by the mean square within subjects: 35.14 = 13540.90/385.34. If the subjects are all the same, these two mean squares should both be estimates of the within-subject variance. Their ratio would be expected to be 1.00. Provided in the population the measurements themselves would follow a Normal distribution with uniform variance across subjects (as for a two sample t test), the ratio would be an observation from an F distribution if the null hypothesis that the subjects were all the same is true. We do not need the P value, we know the subjects are different.

We need the mean squares. The residual mean square is also called the within subjects mean square. It is the variance within the subject = the within-subject standard deviation squared: $\sqrt{385.34} = 19.63 = s_w$.

The subject mean square is also called the between subjects mean square. From it we can estimate the standard deviation and variance between the subjects: $13540.90 = 4s_b^2 + s_w^2$ è $s_b = 57.35$. The 4 comes from the 4 observations per subject. This is the standard deviation of the subjects true PEFR (i.e. average of many measurements).

## Reporting the measurement error

The within-subject standard deviation can be presented and used in several ways. We can report $s_w$ as it stands. There are other possibilities, which may or may not aid in interpretation of the statistic.

We can report the maximum difference which is likely to occur between the observation and the true mean, which is $1.96s_w$. For the children's PEFR data (Table 2) this is $1.96s_w = 1.96 \times 19.63 = 38.5$ litre/min. For 95% of measurements, the subject's true mean PEFR will be within 38.5 litre/min of that observed.

The British Standards Institution (1979) recommended the **repeatability coefficient**, $r$, the maximum difference likely to occur between two successive measurements. This defined as $r = 2\sqrt{2}s_w = 2.83s_w$. This is because the variance of the difference between two measurements is the sum of the error variances of each measurement, i.e. $2\ s_w^2$, the standard deviation of the difference is the square root of this, and 95% of differences will be within 2 (or more precisely 1.96) standard deviations of the mean difference. The mean difference is of course zero. To correspond to a probability of 95%, $r = 1.96\sqrt{2}s_w = 2.77s_w$ would be better, but the difference is numerically unimportant. For the children's PEFR we have repeatability $r = 2.83s_w = 2.83 \times 19.63 = 55.6$ litre/min. This tells us that two measurements on the same subject are unlikely to be more than 55.6 litres apart.

We use the symbol "$r$" to mean both "repeatability coefficient" and "correlation coefficient". This should not be confusing, as it is usually clear from the context what is intended.

We can use the **coefficient of variation** (**CV** or **cv**), defined as the ratio of the standard deviation to the mean. It is not really appropriate to use the coefficient of variation when the error is independent of the mean, although such usage is widespread. For the PEFR data, for example, we would have $cv = s_w / \bar{x} = 19.63/307.0 = 0.064$, or 6.4%. The CV is usually quoted as a percentage. The implication is that the error is proportional to the magnitude of the measurement. This is often the case, but then the calculation of $s_w$ assuming a constant error, as described above, is incorrect. We discuss the appropriate circumstances for the use of the coefficient of variation and its calculation below.

## Assumptions in the calculation of the within-subject standard deviation

Two assumptions are required for the calculation of $s_w$: that the measurement error does not depend on the magnitude of the measurement, and that the measurement errors for each subject follow a Normal distribution.

Independence of the magnitude is essential if we are to have one estimate of standard deviation. If measurement error depends on the magnitude of the measurement, any estimate $s_w$ will be correct at only one particular point on the scale. The assumption that measurement errors are Normally distributed is not necessary for the calculation of $s_w$, and about 95% of observations will be within $2s_w$ of the subject mean whether the errors follow a Normal distribution or not.

**Figure 2. Histogram of the within-subject residuals for the data of Table 2**

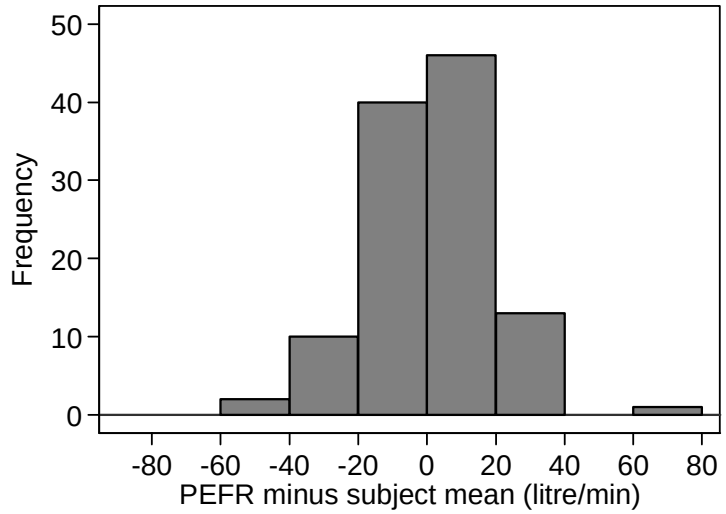
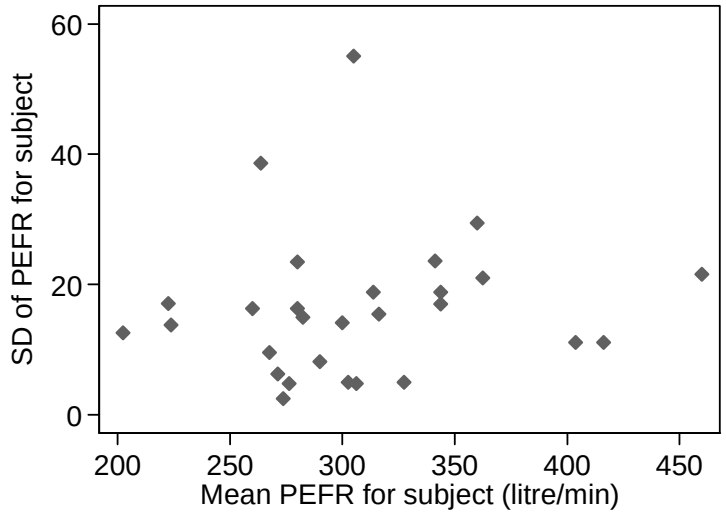



**Figure 3. Subject standard deviation against subject mean for the data of Table 2**




The Normal assumption is reasonable and checkable, for example by using histograms like in Figure 1. We need to take the difference from the subject mean, because it is the distribution of the errors in which we are interested (Figure 2). A more important assumption is that the within-subject standard deviation is independent of the subject mean, in other words, that the measurement error is constant over the range of measurement. We assume this so that we can calculate a common $s_w$ for all subjects. This assumption can be checked by plotting subject standard deviation against subject mean. For the schoolchild PEFR data (Table 2) we have Figure 3. Inspection suggests that there is no tendency for the standard deviation to increase as the mean increases.

**Figure 4. Individual observations against subject mean for the data of Table 2**

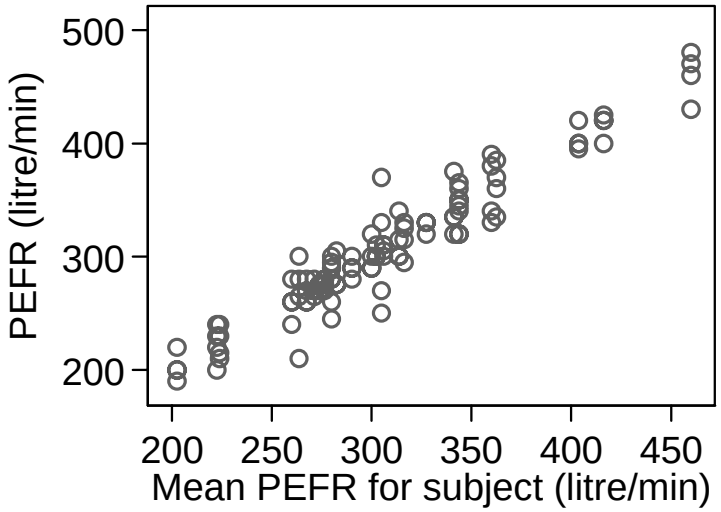
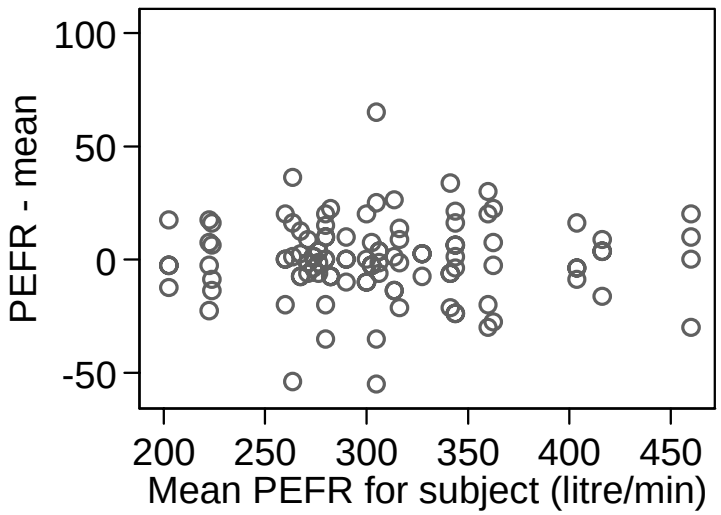



**Figure 5. Within-subject residuals against subject mean for the data of Table 2**




We can also plot the observations against the mean for the subject, as in Figure 4. If we subtract the subject mean from each observation, to get the within-subject residuals, we can plot residuals against subject mean to see whether they get more variable as the subject mean increases (Figure 5).

When we have only two observations per subject, as in Table 4, the subject standard deviation is equal to $\sqrt{2}$ times the absolute value of the difference. Thus we can plot the absolute difference against the subject mean to show the relationship between mean and standard deviation. Figure 6 shows this for the FEV data. There is little evidence of any relationship between mean and standard deviation. The assumption of independence looks very reasonable. This is not so for the data of Table 5, which shows cotinine measured in the same children (Figure 7), where the difference increases as the mean increases.

**Table 4. Pairs of measurements of FEV1 (litres) a few weeks apart, from 164 Scottish schoolchildren (D. Strachan, personal communication)**

| 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd |
|---|---|---|---|---|---|---|---|---|---|
| 0.92 | 0.94 | 1.37 | 1.39 | 1.49 | 1.51 | 1.60 | 1.63 | 1.75 | 1.87 |
| 1.04 | 1.72 | 1.37 | 1.52 | 1.49 | 1.60 | 1.60 | 1.66 | 1.76 | 1.62 |
| 1.05 | 1.18 | 1.38 | 1.16 | 1.50 | 1.45 | 1.60 | 1.68 | 1.76 | 1.82 |
| 1.08 | 1.28 | 1.38 | 1.29 | 1.50 | 1.47 | 1.60 | 1.75 | 1.77 | 1.78 |
| 1.10 | 1.11 | 1.38 | 1.37 | 1.50 | 1.58 | 1.61 | 1.44 | 1.77 | 1.85 |
| 1.17 | 1.24 | 1.38 | 1.39 | 1.51 | 1.51 | 1.61 | 1.53 | 1.78 | 1.72 |
| 1.19 | 1.25 | 1.38 | 1.40 | 1.51 | 1.54 | 1.61 | 1.55 | 1.78 | 1.76 |
| 1.19 | 1.26 | 1.38 | 1.43 | 1.51 | 1.73 | 1.61 | 1.61 | 1.80 | 1.72 |
| 1.19 | 1.37 | 1.39 | 1.44 | 1.52 | 1.53 | 1.61 | 1.61 | 1.80 | 1.76 |
| 1.20 | 1.24 | 1.40 | 1.38 | 1.53 | 1.46 | 1.62 | 1.57 | 1.80 | 1.79 |
| 1.21 | 1.19 | 1.40 | 1.42 | 1.53 | 1.48 | 1.62 | 1.68 | 1.80 | 1.82 |
| 1.22 | 1.26 | 1.40 | 1.57 | 1.53 | 1.48 | 1.63 | 1.70 | 1.80 | 1.82 |
| 1.22 | 1.38 | 1.42 | 1.45 | 1.53 | 1.51 | 1.64 | 1.61 | 1.82 | 1.88 |
| 1.23 | 1.28 | 1.42 | 1.46 | 1.53 | 1.56 | 1.64 | 1.72 | 1.85 | 1.73 |
| 1.23 | 1.54 | 1.42 | 1.83 | 1.53 | 2.01 | 1.65 | 1.43 | 1.85 | 1.81 |
| 1.27 | 1.31 | 1.43 | 1.38 | 1.54 | 1.56 | 1.65 | 1.60 | 1.85 | 1.89 |
| 1.28 | 1.27 | 1.43 | 1.38 | 1.54 | 1.57 | 1.65 | 2.05 | 1.86 | 1.90 |
| 1.28 | 1.29 | 1.43 | 1.41 | 1.55 | 0.69 | 1.66 | 1.64 | 1.87 | 1.88 |
| 1.28 | 1.38 | 1.43 | 1.51 | 1.55 | 1.56 | 1.67 | 1.50 | 1.88 | 1.82 |
| 1.29 | 1.23 | 1.43 | 1.54 | 1.55 | 1.60 | 1.67 | 1.63 | 1.89 | 1.90 |
| 1.29 | 1.28 | 1.43 | 1.65 | 1.56 | 1.60 | 1.69 | 1.67 | 1.89 | 2.00 |
| 1.32 | 1.37 | 1.45 | 1.29 | 1.57 | 1.57 | 1.69 | 1.69 | 1.92 | 2.00 |
| 1.33 | 1.32 | 1.45 | 1.42 | 1.57 | 1.60 | 1.69 | 1.79 | 1.92 | 2.10 |
| 1.33 | 1.35 | 1.45 | 1.48 | 1.58 | 1.36 | 1.70 | 1.82 | 1.94 | 1.43 |
| 1.33 | 1.42 | 1.46 | 1.47 | 1.58 | 1.49 | 1.72 | 1.69 | 1.94 | 2.10 |
| 1.34 | 1.39 | 1.46 | 1.49 | 1.58 | 1.60 | 1.72 | 1.73 | 1.95 | 2.27 |
| 1.34 | 1.44 | 1.47 | 1.19 | 1.58 | 1.60 | 1.72 | 1.74 | 1.97 | 2.03 |
| 1.35 | 1.40 | 1.47 | 1.44 | 1.58 | 1.65 | 1.73 | 1.73 | 2.10 | 2.20 |
| 1.35 | 1.40 | 1.47 | 1.53 | 1.58 | 1.67 | 1.74 | 1.71 | 2.10 | 2.21 |
| 1.35 | 1.40 | 1.47 | 1.65 | 1.59 | 1.41 | 1.74 | 1.79 | 2.11 | 2.13 |
| 1.35 | 1.59 | 1.48 | 1.35 | 1.59 | 1.60 | 1.74 | 1.80 | 2.15 | 2.07 |
| 1.36 | 1.25 | 1.48 | 1.48 | 1.59 | 1.71 | 1.75 | 1.61 | 2.21 | 2.02 |
| 1.36 | 1.32 | 1.49 | 1.47 | 1.60 | 1.58 | 1.75 | 1.84 | | |

**Figure 6. Absolute difference against mean for the data of Table 4**

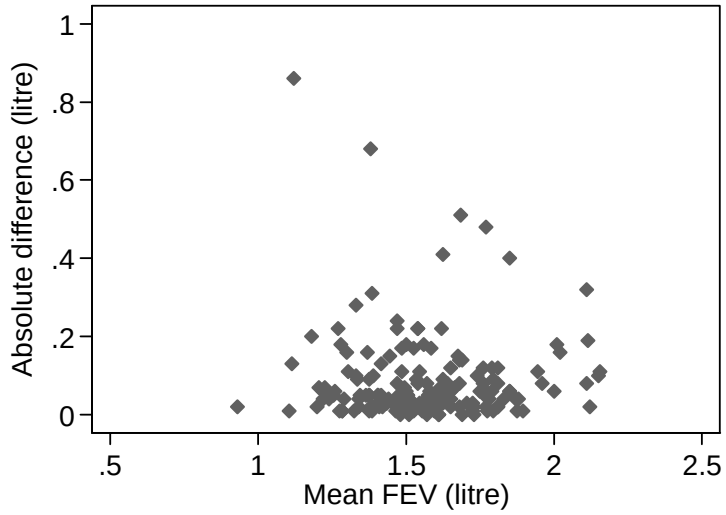
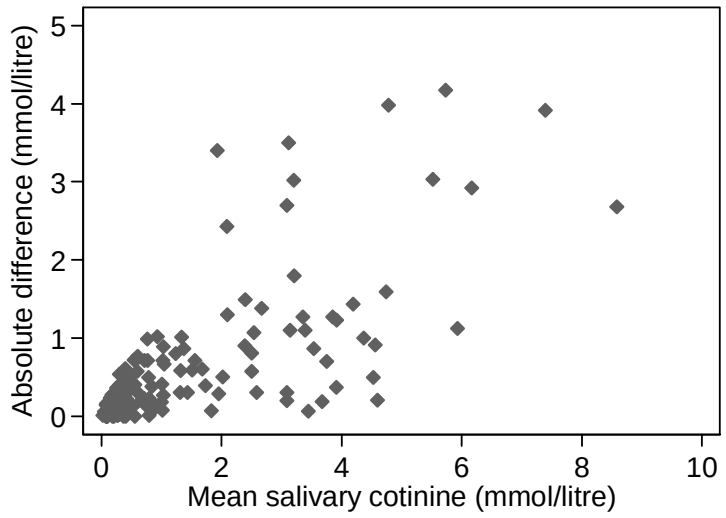


**Table 5. Duplicate salivary cotinine measurements for a group of Scottish schoolchildren, ordered by magnitude (D. Strachan, personal communication)**

| 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd | 1st | 2nd |
|---|---|---|---|---|---|---|---|---|---|
| ND | ND | 0.2 | 0.6 | 0.4 | 0.3 | 0.9 | 0.2 | 2.7 | 2.4 |
| ND | ND | 0.3 | ND | 0.4 | 0.4 | 0.9 | 0.3 | 2.7 | 4.0 |
| ND | ND | 0.3 | ND | 0.4 | 0.4 | 0.9 | 0.7 | 2.8 | 2.2 |
| ND | ND | 0.3 | ND | 0.4 | 0.4 | 0.9 | 0.7 | 2.8 | 3.9 |
| ND | 0.1 | 0.3 | ND | 0.4 | 1.1 | 0.9 | 3.3 | 2.8 | 6.8 |
| ND | 0.1 | 0.3 | ND | 0.4 | 1.4 | 1.0 | 0.2 | 3.1 | 1.6 |
| ND | 0.1 | 0.3 | ND | 0.5 | 0.1 | 1.0 | 1.6 | 3.2 | 2.9 |
| ND | 0.2 | 0.3 | 0.1 | 0.5 | 0.1 | 1.1 | 0.4 | 3.2 | 3.0 |
| ND | 0.2 | 0.3 | 0.1 | 0.5 | 0.3 | 1.1 | 0.9 | 3.2 | 4.5 |
| ND | 0.2 | 0.3 | 0.1 | 0.5 | 0.3 | 1.1 | 1.0 | 3.3 | 4.5 |
| ND | 0.2 | 0.3 | 0.2 | 0.5 | 0.3 | 1.2 | 0.8 | 3.5 | 3.4 |
| ND | 0.6 | 0.3 | 0.2 | 0.5 | 0.4 | 1.2 | 0.9 | 3.5 | 4.9 |
| 0.1 | ND | 0.3 | 0.3 | 0.5 | 1.0 | 1.2 | 1.5 | 3.6 | 0.2 |
| 0.1 | 0.1 | 0.3 | 0.3 | 0.6 | ND | 1.2 | 1.8 | 3.7 | 2.6 |
| 0.1 | 0.1 | 0.3 | 0.3 | 0.6 | 0.3 | 1.3 | 0.3 | 3.8 | 3.6 |
| 0.1 | 0.2 | 0.3 | 0.4 | 0.6 | 0.5 | 1.4 | 0.7 | 3.9 | 5.5 |
| 0.1 | 0.2 | 0.3 | 0.4 | 0.6 | 0.6 | 1.5 | 0.6 | 4.0 | 3.1 |
| 0.1 | 0.4 | 0.3 | 0.4 | 0.6 | 0.8 | 1.6 | 0.8 | 4.1 | 3.4 |
| 0.1 | 0.5 | 0.3 | 0.4 | 0.6 | 0.8 | 1.6 | 1.3 | 4.1 | 3.7 |
| 0.2 | ND | 0.3 | 0.5 | 0.6 | 1.0 | 1.7 | 4.7 | 4.1 | 5.0 |
| 0.2 | ND | 0.3 | 0.6 | 0.7 | 0.1 | 1.8 | 0.9 | 4.4 | 1.7 |
| 0.2 | ND | 0.4 | ND | 0.7 | 0.2 | 1.8 | 1.9 | 4.7 | 4.5 |
| 0.2 | 0.1 | 0.4 | ND | 0.7 | 0.3 | 1.8 | 2.1 | 4.8 | 4.3 |
| 0.2 | 0.1 | 0.4 | 0.1 | 0.7 | 0.3 | 1.8 | 2.3 | 4.9 | 1.4 |
| 0.2 | 0.1 | 0.4 | 0.1 | 0.7 | 0.8 | 1.9 | 1.2 | 4.9 | 3.9 |
| 0.2 | 0.1 | 0.4 | 0.1 | 0.7 | 0.9 | 1.9 | 1.5 | 6.5 | 5.4 |
| 0.2 | 0.1 | 0.4 | 0.1 | 0.7 | 1.4 | 1.9 | 2.8 | 7.0 | 4.0 |
| 0.2 | 0.2 | 0.4 | 0.2 | 0.8 | 0.4 | 2.0 | 1.4 | 7.6 | 4.7 |
| 0.2 | 0.2 | 0.4 | 0.2 | 0.8 | 0.5 | 2.0 | 3.1 | 7.8 | 3.6 |
| 0.2 | 0.3 | 0.4 | 0.3 | 0.8 | 0.8 | 2.0 | 3.4 | 9.3 | 5.4 |
| 0.2 | 0.3 | 0.4 | 0.3 | 0.8 | 0.9 | 2.1 | 2.9 | 9.9 | 7.2 |
| 0.2 | 0.3 | 0.4 | 0.3 | 0.8 | 1.8 | 2.3 | 4.1 | | |
| 0.2 | 0.5 | 0.4 | 0.3 | 0.9 | 0.2 | 2.7 | 1.4 | | |

Figure 7. Absolute difference against mean for the data of Table 5.



If there is a relationship between standard deviation and mean, we cannot use the within-subject standard deviation as a measure of repeatability, as it will not be the same through the range of measurement. Instead, we try to transform the data so that the relationship disappears.

**Data which go off the scale**

Many assays have some limit below which no measurement can be made, and the result is recorded as below the limit of detection. Table 5 shows pairs of salivary cotinine measurements made on a sample of schoolchildren. Many of the cotinine levels were so low as be undetectable. When such data are used as outcome or predictor variables in regression analyses, the undetectable observations can be set to an arbitrary low value, such as half the lowest possible detectable value. Provided there are not many such observations, the presence of these arbitrary values will not influence the analysis much. This will not work for the estimation of measurement error, because serious bias may be introduced. In particular, individuals for whom both measurements are recorded as not detectable will have differences of zero, which will not occur in the higher parts of the scale and violate the assumption that the measurement error is uniform throughout the scale of measurement.

*Provided the measurement error is uniform*, we can simply omit observations which are below the detectable range. Variables which have 'not detectable' observations are unlikely to meet this assumption, however, but usually have error increasing as the quantity being measured increases, as does salivary cotinine (Figure 7). (For the graphs I have set all the 'none detectable' readings to 0.05, which is half the lowest observable value, 0.1.) We can usually deal with this relationship between error and subject mean by transformation, as described below.

If we have two observations per case, as in Table 5, to omit an observation means that the subject will be omitted. If we have more than two observations per case, as in Table 2, omitting only observations below the limit of detection and keeping the rest will mean that subjects with some observations below and the limit and some above will have small individual standard deviations, as the range of their observations will be artificially narrowed. We should omit all such cases. It may also happen that the quantity being measured is too large and all we know is that it is above some value. We can deal with these in the same way, provided the assumption of uniform error is met.

## Repeatability dependent on the magnitude of the variable

When the within-subject standard deviation is related to the magnitude of the measurement, as in Figure 7, we cannot estimate $s_w$ as described above, because it is not constant. The simplest alternative model to consider is that the standard deviation is proportional to the mean. We then estimate the ratio of standard deviation to the mean, the coefficient of variation. I shall omit the details of this. If the standard deviation is proportional to the mean CV should be a constant. For the cotinine example, the coefficient of variation is 67%.

When the standard deviation is proportional to the mean we have a valid use and method of estimation of the coefficient of variation. From it, we can estimate the standard deviation of repeated measurements at any point within the range of measurement, by multiplying by the mean at that point.

**Figure 8. Second measurement against first for FEV (data of Table 4)**

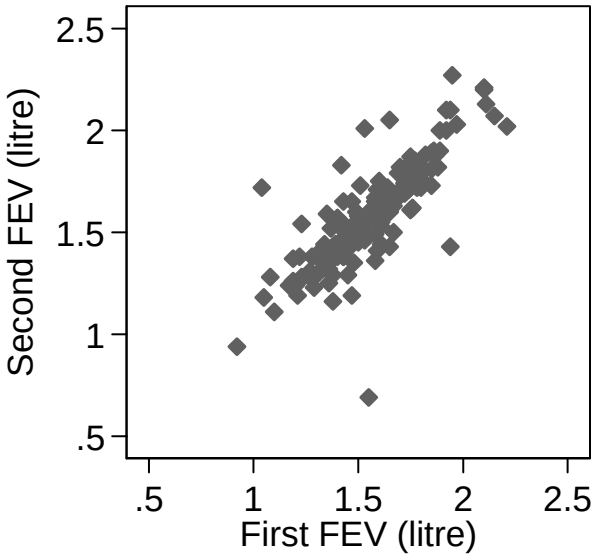
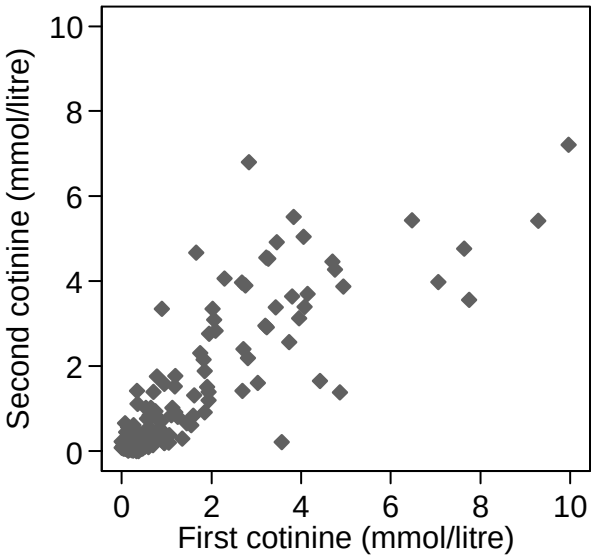



**Figure 9. Second against first measurements of plasma cotinine, data of Table 5.**




The within-subject variability for salivary cotinine seems very large, but the possible range of values, from these very lightly exposed children to heavy smokers, is very wide and salivary cotinine is sufficient precise to distinguish between many different levels of exposure. The precision of a measurement must be interpreted in the light of the use to which the method is to be put.

## Correlation coefficients in the study of repeatability

When we have data like those of Tables 4 and 5, there is a great temptation to plot one measurement against the other. The resulting scatter diagram, Figures 8 and 9 for example, in turn tempts us to calculate a correlation coefficient between the first and second measurement. Such a correlation is also called a **reliability coefficient**, particularly in the social science literature. We usually specify the type of reliability, e.g. the **test-retest**

**reliability**, correlation between observations by the same observer on different occasions, or **inter-rater reliability**, the correlation between observations by different observers.

There are difficulties in interpreting the correlation coefficient as an index of repeatability. The correlation depends on the way the sample was chosen. The correlation obtained from a sample where all subjects are similar will be smaller than that obtained from a sample with large differences between subjects. Thus $r$ reflects both within and between subject variability.

For example, for the FEV data (Table 4) the correlation between repeated measurements is $r = 0.82$. Suppose we split the FEV sample into two sub-samples at 1.5 litres (close to the mean). The correlation for the first sub-sample (first FEV < 1.5) is $r = 0.54$ and for the second (first FEV ≥ 1.5) it is $r = 0.73$. For the full sample $r$ is bigger than for either sub-sample, because the variation between subjects is greater. This does not happen with the within-subject standard deviation. For the whole sample $s_w = 0.10$ litre, for subjects below 1.50 litres $s_w = 0.09$ litre, and for subjects above 1.5 litre $s_w = 0.11$ litre.

The correlation coefficient is thus dependent on the way the sample is chosen. It only has meaning for the population from which the study subjects can be regarded as a random sample. If we select subjects to give a wide range of the measurement, for example, this will inflate the correlation coefficient. The within-subject standard deviation is less susceptible to such problems and has a direct interpretation, so it may be preferred for describing the characteristics of methods of clinical measurement.

The correlation coefficient does have other uses in the study of repeatability. We can use it to test the null hypothesis that the first and second measurements are independent, i.e. that there is no repeatability at all. Thus it is useful in investigating the validity of measurement methods. It also enables us to compare the repeatability of different measurements collected on the same subjects. This might be useful if we are piloting a number of questionnaire scales to which best discriminates between individuals. We could make repeated measurements of all the scales on the same subjects and calculate correlations between the repeated measurements. The scales with the highest correlation between repeated measurements would discriminate best between subjects, in other words they would carry the most information.

## The intra-class correlation coefficient

There is another problem in the use of the correlation coefficient between the first and second measurements: there is no reason to suppose that the order is important. Indeed, if the order of measurement were important we would not have repeated observations of the same thing. We could reverse the order of any of the pairs and get a slightly different value of the correlation coefficient between repeated measurements. In fact for pairs of measurements on $n$ subjects, there are $2^n$ possible values of $r$. Most of these will be very similar, of course, and the best estimate of the population correlation coefficient will be in the middle.

The **intra-class correlation coefficient** or **ICC** avoids this problem. It estimates the average correlation between all possible pairs within the subject (the subject being the class). It also extends very easily to the case of several observations per subject, as for the PEFR data of Table 2. The intra-class correlation coefficient between repeated measurements is the correlation usually used for reliability statistics.

We shall omit the details of calculation. For the FEV data, ICC = 0.82. This is the same as the ordinary correlation coefficient found above. The effect of using the intra-class correlation rather than ordinary correlation coefficient is very small for so large a sample.

However, ICC has the great advantage that we can use it when there are more than two observations per subject. For the PEFR data of Table 2 it is 0.895.

The ICC is related to the variances within-subject and between subject as follows:

$$\text{ICC} = \frac{s_b^2}{s_b^2 + s_w^2}$$

For the data of Table 2, we have

$$\text{ICC} = \frac{57.35^2}{57.35^2 + 19.63^2} = 0.895$$

as before.

The intra-class correlation coefficient will be 1.00 when $s_w^2 = 0$, which happens when for each subject all measurements are identical. The correlation will be zero when there is no more difference between the subjects than would be expected by chance if the subjects were identical. Thus the intra-class correlation coefficient, like the ordinary correlation coefficient, depends on the range of the subject means.

For pairs of measurements, the intra-class correlation coefficient, ICC, and the ordinary product moment correlation coefficient, $r$, are estimates of the same thing. Unless the sample is small, they should be very similar, as for the FEV data of Table 4, for which ICC = 0.82 and $r = 0.82$. The main advantage of ICC is that it can be used when we have more than two observations per subject.

J. M. Bland

## Reference

British Standards Institution. (1979) *Precision of test methods 1: Guide for the determination and reproducibility for a standard test method (BS5497, part 1).* London: British Standards Institution.