University of York Department of Health Sciences

Research Methods

Measurement in Healthcare Research

‘All science is measurement’ -- Helmholtz

‘All science is measurement but not all measurement is science.’ -- Kelvin

The role of measurement

This lecture provides an overview of some of the issues concerning measurement which will be covered in different parts of the M.Sc. programme.

Measurement has a central role both in clinical care and in healthcare research. In clinical work, diagnosis often depends on measurement. Some diagnoses depend on a measurement being above some predetermined level (e.g. hypertension or diabetes). We use measurements to monitor the progress of patients (e.g. serial lung function measurements in the management of asthma, or temperature measurements in fever). In research, most studies depend upon measurements of some sort. The quality of data depends upon the measurement techniques. Poor measurement techniques can introduce so much random variation into data that the research question cannot be answered.

Much research is concerned directly with the development and testing of methods of measurement. Other studies concern the interpretation of measurements, such as the evaluation of diagnostic tests.

In this lecture we shall take a wide definition of measurement. This includes:

- direct physical measures (e.g. height, weight, blood pressure),
- questionnaire based scales (e.g. anxiety, depression),
- subjective assessments (e.g. patient's condition as poor, fair, good or excellent),
- presence or absence of a sign.

The same issues of repeatability, variation between different observers, etc., arise with all of them.

Making measurements

The outcome of a measurement is influenced by several things:

- the true value of the quantity we want to measure,
- biological variation over time,
- the measurement instrument itself,
- the skill, experience and expectations of the observer,
- the relationship between observer and subject.

For example, consider a familiar measurement, blood pressure. This is measured for several different purposes, but we shall look at blood pressure measured as case-finding for hypertension. Here the true value of the quantity we want to measure is the subject’s long term average peak arterial pressure. The measurement we actually make is not this average, because there is biological variation over time, from heartbeat to heartbeat, over the day, seasonally over the year, and so on. The measurement instrument itself might have an effect.

Are we using a mercury sphygmomanometer and stethoscope or an automated sphygmomanometer which uses a microphone to detect the blood flow? Blood pressure also varies with the cuff size and the position in which the subject is placed, sitting or supine. It also depends on the skill, experience and expectations of the observer, how good they are at detecting the Korotkov sounds, for example, and how closely they adhere to a protocol as the points at which systolic and diastolic pressure should be recorded. Does the relationship between the observer and the subject have any effect? I have been told, but never tracked down a reference, that blood pressure is higher when measured by an observer of the opposite sex to the subject. The phenomenon of white coat hypertension is documented, where the prospect of having BP measured is enough to raise it.

Some of these factors are outside the control of the observer (e.g. variation within the subject). Some factors are not (e.g. position). It is important to standardize these, for example, such things as the accuracy with which we read scales and record the result. A survey of health professionals has shown important differences in the way blood pressure is measured, for example, some observers recording to the nearest 5 mm Hg, others to the nearest 10 mm Hg.

Accuracy and precision

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.

‘Error’, in the sense in which we use it in ‘measurement error’, is not the same as ‘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. Table 1 shows an example of a mistake. These systolic blood pressure measurements were recorded in a clinical trial comparing different types of CABG surgery. These observations have been arranged in ascending order in columns and it is easy to see that the first observation, 16 mm Hg, is a mistake. I do not believe that the observer made a measurement of 16 mm Hg. As this patient was also recorded as ‘hypertensive’, it is highly likely that the original reading was 160 mm Hg.

We usually distinguish between ‘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. We cannot judge the accuracy of many measurements, because we do not know what the truth is, but we can usually see how close together repeated measurements of the same thing are.

Table 1. 210 systolic blood pressure measurements from CABG patients

16	105	110	116	120	123	126	130	130	135	140	144	150	160
88	105	110	116	120	123	126	130	131	135	140	145	152	160
95	106	111	117	120	123	126	130	131	135	140	145	153	160
98	106	112	117	120	123	127	130	131	135	140	145	153	160
99	107	112	117	120	123	127	130	132	135	140	145	154	160
99	107	112	117	120	124	127	130	132	136	140	145	154	164
99	107	112	118	120	125	127	130	132	138	140	145	154	165
100	108	112	118	120	125	128	130	132	138	140	146	155	165
100	108	112	118	120	125	128	130	132	139	140	147	155	166
100	109	113	119	121	125	128	130	132	139	140	147	156	170
100	109	113	119	122	125	128	130	132	139	141	148	158	170
102	109	115	120	122	125	128	130	132	140	141	148	158	175
102	110	115	120	122	126	128	130	133	140	142	150	159	176
103	110	115	120	123	126	128	130	134	140	143	150	159	189
104	110	116	120	123	126	129	130	135	140	143	150	160	198

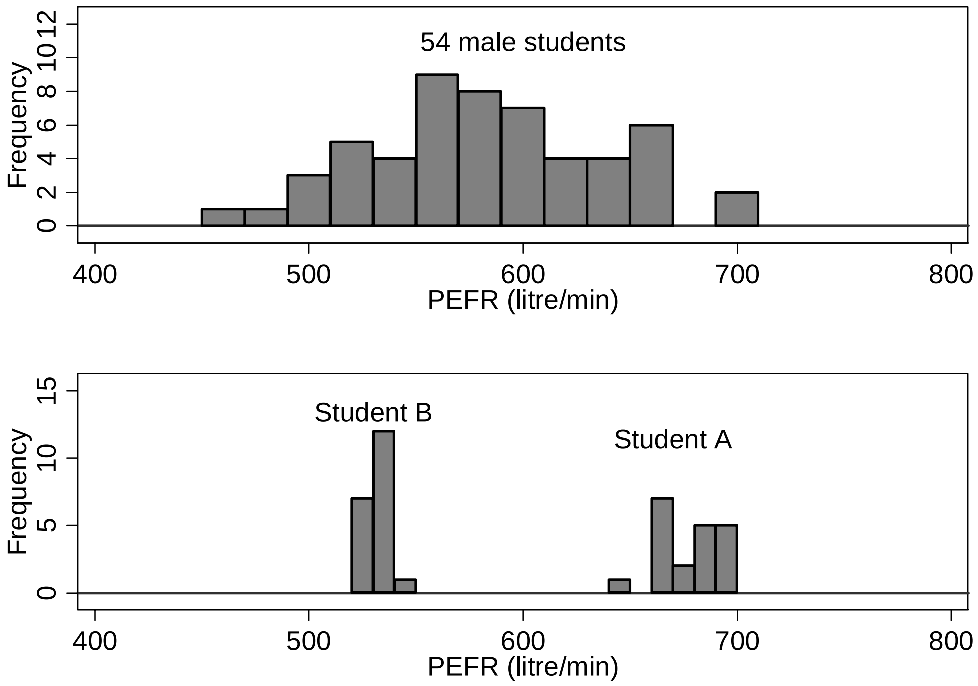
Sources of variation

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. 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.

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.

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



Repeatability and measurement error

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.

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.

For example, Table 2 shows repeated measurements of FEV on a sample of Scottish schoolchildren. We can have a look at the measurement error by plotting one measurement against the other (Figure 2). We can see that there is quite a lot of variation from one measurement to another. One observation looks like a mistake, with the first observation being 1.55 litres and the second 0.69 litres, considerably smaller than any other observation.

It looks quite plausible that the measurement error is the same for large and small FEV. Another way to check this is to plot the difference against the average, as in Figure 3. The average of the two repeat measurements is the best estimate we have of the magnitude of FEV for that child. There appears to be no obvious relationship between error and magnitude.

Table 2. 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		

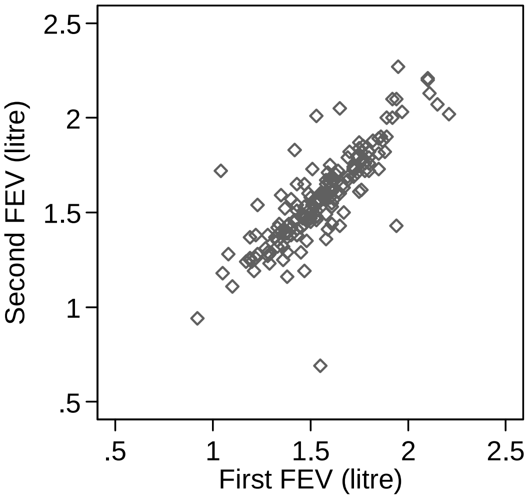


Figure 2. One FEV measurement against another on the same child

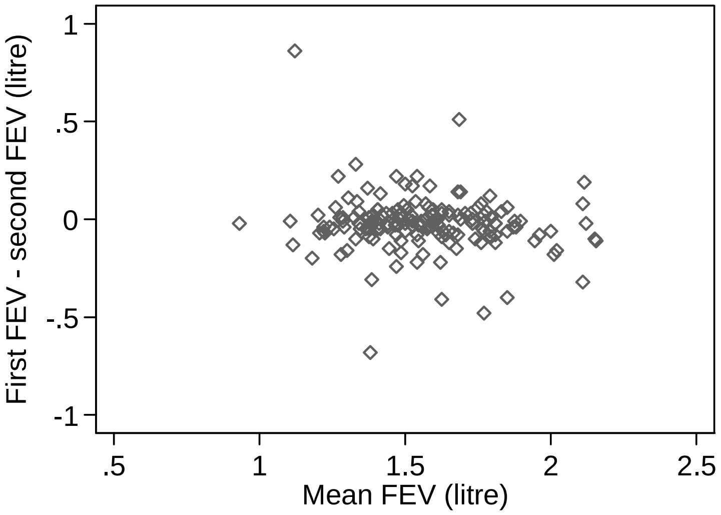


Figure 3. Difference in FEV against average of the two.

Table 3. 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		

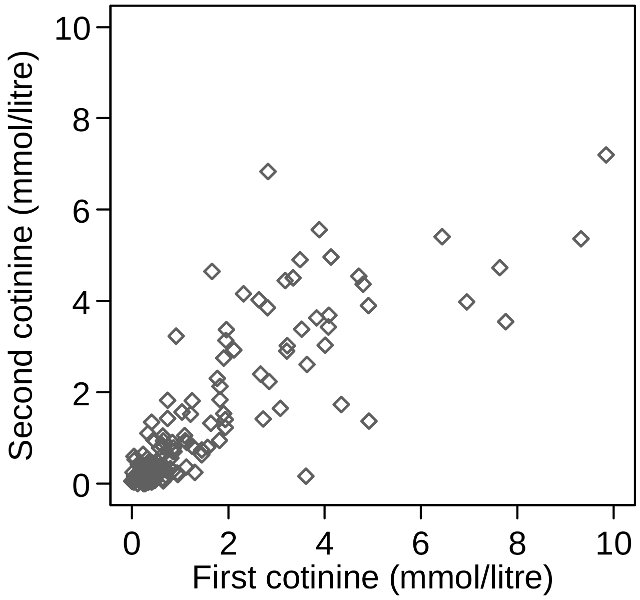


Figure 4. Second against first measurement of plasma cotinine, data of Table 3.

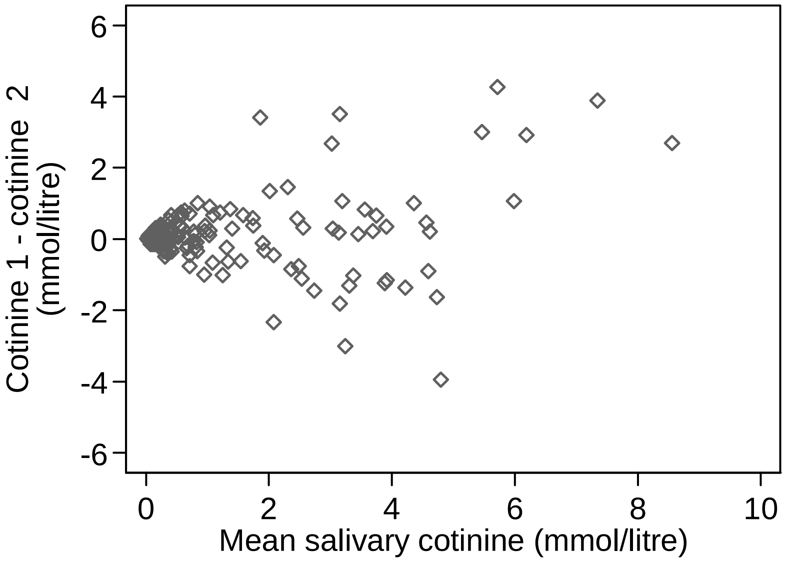


Figure 5. Difference between salivary cotinine measurements against average for the child

Repeatability dependent on the magnitude of the variable

Sometimes the precision of a measurement is related to the magnitude of the quantity being measured. For example, Table 3 shows pairs of measurements of salivary cotinine, a metabolite of nicotine. Figure 4 shows a plot of one measurement against the other. The variability appears to increase as the cotinine increases. Figure 5 shows a plot of difference against average. We will need a measure of variability which is related to the size of the cotinine being measured, in practice to the measurement itself.

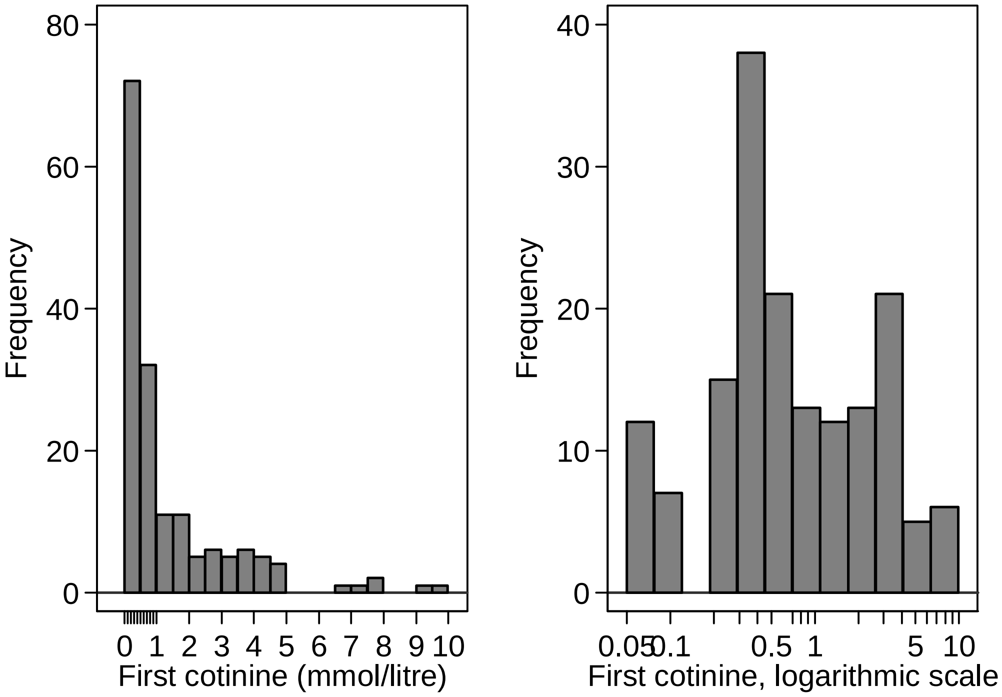


Figure 6. Histograms showing the cotinine data on a natural and a logarithmic scale

How precisely should we record data?

This depends to some extent on the purpose for which the data are to be recorded. Any data which are to be subjected to statistical analysis should be recorded as accurately as possible. A study can only be as good as the data, and data are often very costly and time-consuming to collect. The accuracy to which data are to be recorded and all other procedures to be used in measurement should be decided in advance and stated in the protocol, the written statement of how the study is to be carried out.

In the cotinine measurements of Table 3, the second lowest measurable observation was 0.1 and the second highest was 9.3. Both are recorded to one decimal place, but 0.1 has one significant figure and 9.3 has two. The second observation is recorded more accurately. This becomes very important when we come to analyse the data.

(You can omit this paragraph, a diversion into statistics, if you find it does not mean anything to you.) We do not always analyse data on the scale which we originally use to collect them, but sometimes we manipulate the data in some way first. A typical transformation is to take the logarithm, to analyse data on a logarithmic scale (to be discussed more in the statistical component of research methods0. Figure 6 shows the cotinine data on a logarithmic scale. The greater inaccuracy of recording at the lower end of the scale is magnified by the transformation.

The accuracy of recording depends on the number of significant figures recorded, not the number of decimal places. Significant figures are those which follow the first non-zero digit and which precede the point where there are no non-zero digit. Thus 32000000, 0.32, 32, and 0.00032000 are all recorded to two significant figures. We should make sure that there as many significant figures as are meaningful for the measurement in question.

Table 4. Answers to the question: 'Have you ever smoked a cigarette?', by Derbyshire school children

		Interview		
		Yes	No	Total
Self-administered	Yes	61	2	63
questionnaire	No	6	25	31
Total		67	27	94

Table 5. Answers to the question: Do you leak any urine/water when you don't mean to? That means anything from a few drops to a flood during the day or night?', by Leicestershire women (data of Kate Williams)

		Self-administered questionnaire		
		Yes	No	Total
Interview	Yes	21	3	24
	No	1	9	10
Total		22	12	34

Digit preference

Sometimes in measurement there is uncertainty in the last digit. Observers will often have some values for this last digit which they record much more often than others. Many observers are much more likely to record a terminal zero than a nine or a one, for example. In the 210 blood pressure measurements of Table 1, there are 62 which end in zero, such as 130. If observers did not record zero in preference to other digits we would expect one in ten observations, 10%, to end in zero. In fact, 62/210 = 30%, not 10%.

Digit preference decreases the precision with which data are recorded and it should be avoided. If possible readings should be taken to sufficient significant figures for the last digit to be unimportant. Observer training and awareness of the problem help to minimize digit preference.

Non-numerical data

We also find measurement error in non-numerical data. For example, Table 5 shows two answers to a question about cigarette smoking by school-children. Unlike the FEV and cotinine measurements, the questions were not asked in quick succession. The answers to the second question would be certain to be influenced by the answer to the first if we did. To avoid this problem, we have to separate the questions in time so that the respondents may have forgotten what they said the first time round. In this case, a survey of several hundred children used a self-administered questionnaire and a sample of these were interviewed a few days later.

How closely do the children's answers agree? In this case, closely enough for us to think that we are measuring something, but the agreement is not perfect. Although most children gave the same answer on both occasions, six said 'yes' to the self-administered question and 'no' to the interview question, two said 'no' to the self-administered question and 'yes' to the interview question. We cannot rely on answers to be invariably correct. Here we are assessing the test-retest reliability of the question.

Table 6. Anxiety for a group of osteoarthritis patients as recorded on the HADS scale and diagnosed at clinical interview by a psychiatrist

		Anxiety diagnosed at clinical interview		
		Yes	No	Total
HADS anxiety	Yes	15	7	22
score 8 or more	No	2	30	32
Total		17	37	54

It might be that some children were inhibited from being frank about an activity that the interview might have disapproved. The same is unlikely to be true in the study reported in Table 5, which shows the results of another study using the same design. In a study of urinary incontinence, a group of women were asked the question ‘Do you leak any urine/water when you don’t mean to? That means anything from a few drops to a flood during the day or night?’ A sample of these women were then interviewed and the same question was asked. Although most women gave the same answer on both occasions, one said ‘yes’ to the self-administered question and ‘no’ to the interview question, three said ‘no’ to the self-administered question and ‘yes’ to the interview question. Again we see that questionnaire data contain measurement error.

Table 6 shows a study where two different methods were used to ascertain a patient characteristic, in this case the presence of clinical anxiety. One method was a survey questionnaire, the Hospital Anxiety and Depression Scale (HADS), the other a clinical interview by a psychiatrist. Again there is some agreement: most patients who had a HADS score indicating anxiety were diagnosed in this way by the psychiatrist and most who have HADS scores below the anxiety cut-off were diagnosed as not having anxiety at clinical interview. However, not all those indicated as having anxiety by the questionnaire were so diagnosed and not all those diagnosed with anxiety were picked up by the questionnaire.

There is usually uncertainty in categorical data.

Composite scales

Sometimes we combine a set of items together to make a composite scale. The HADS scale is one of these. For another example, Table 8 shows the depression scale of the GHQ (General Health Questionnaire), with the scoring. We give the score for each answer and add them to get a measure of depression. Such scales are widely used in healthcare research.

One of the questions which we want to ask about such scales is how coherent they are: do they really measure anything useful? They can only do this efficiently if the items all address slightly different aspects of the thing we want to measure. We want them to be fairly closely related, but not identical.

Validity

A measure is valid if it measures what we think it measures or want it to measure. A measure can be reliable or repeatable without being valid. For example, when schoolchildren and their parents were asked ‘Do you (Does your child) usually cough at other times in the day or at night?’ 24.8% of the schoolchildren replied yes, but only 4.5% of their parents did so. The reports might be repeatable but the children and their parents are clearly not reporting the same thing.

Table 8. Depression scale of the GHQ, with scoring

Have you:

been thinking of yourself as a worthless person?	Not at all 0	No more than usual 1	Rather more than usual 2	Much more than usual 3
felt that life is entirely hopeless?	Not at all 0	No more than usual 1	Rather more than usual 2	Much more than usual 3
felt that life isn't worth living?	Not at all 0	No more than usual 1	Rather more than usual 2	Much more than usual 3
thought of the possibility that you might make away with yourself?	Definitely have 3	I don't think so 2	Has crossed my mind 1	Definitely not 0
found at times you couldn't do anything because your nerves were too bad?	Not at all 0	No more than usual 1	Rather more than usual 2	Much more than usual 3
found yourself wishing you were dead and away from it all?	Not at all 0	No more than usual 1	Rather more than usual 2	Much more than usual 3
found that the idea of taking your own life kept coming into your mind?	Definitely have 3	I don't think so 2	Has crossed my mind 1	Definitely not 0

Researchers have identified many types of validity, but the main ones are:

- Criterion validity
  Measurements are closely related to those given by some other, definitive technique, a ‘gold standard’.
- Face validity
  Instrument looks as though it should measure what we want to measure.
- Content validity
  All the items appear relevant to the aim of the index, and all aspects of the thing we wish to measure are covered.
- Construct validity
  Instrument is related to things to which we expect the concept we are trying measure to be related, and independent of those things of which the concept should be independent.

Martin Bland
October 2007