British Standards Institution Study Day

Types of data and how they can be analysed

Martin Bland
Prof. of Health Statistics
University of York
http://martinbland.co.uk

1. Types of data

In this lecture we shall look at the statistical principals involved in the presentation of the results of research studies. We shall illustrate this with a look at how people analysed data in a leading health research journal: *The British Medical Journal*, or *BMJ*. The BMJ is published both on paper and on-line, but it now uses the on-line journal as the key publication. For this lecture, we shall look at the four research papers published in the week starting between 30 May 2011 and 5 Jun 2011. You can find these on

http://www.bmj.com/archive/online/2011/05-30

The first is by Steckelberg *et al.*: Effect of evidence based risk information on "informed choice" in colorectal cancer screening: randomised controlled trial (*BMJ* 2011; **342**: d3193). In the Abstract under Results we have:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; P<0.001). A "positive attitude" towards colorectal screening prevailed in both groups but was significantly lower in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%, −5.9% to −0.3%; P<0.01). The intervention had no effect on the combination of actual and planned uptake (72.4% (568) v 72.9% (577); P=0.87). . . .'

We can pick some terms out of this passage:

'. . . **345/785 (44.0%) participants in the intervention group made an informed choice**, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). More intervention group participants had **"good knowledge"** (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; P<0.001). A **"positive attitude" towards colorectal screening** prevailed in both groups but was significantly lower in the intervention group (93.4% (733) v 96.5% (764); difference −3.1%, −5.9% to −0.3%; P<0.01). The intervention had no effect on the combination of **actual and planned uptake** (72.4% (568) v 72.9% (577); P=0.87). . . .'

These are some of the outcome variables in this study:

- making an informed choice,
- having 'good knowledge',
- having a 'positive attitude' towards colorectal screening,
- uptake of screening.

In a research study, the outcome variables are those which we are trying to explain or to influence. All these are 'yes or no' variables. A variable which classifies individuals into

groups is called qualitative or categorical. When we have only two categories we say the variable is dichotomous.

Next consider the paper by Johansson *et al.*: Longer term effects of very low energy diet on obstructive sleep apnoea in cohort derived from randomised controlled trial: prospective observational follow-up study (*BMJ* 2011; **342**: d3017). In the Abstract under Results we read:

'. . . After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001). . . . At one year, 30/63 (48%, 95% confidence interval 35% to 60%) no longer required continuous positive airway pressure and 6/63 (10%, 2% to 17%) had total remission of obstructive sleep apnoea (apnoea-hypopnoea index <5 events/hour).'

We can highlight the outcome variables:

'. . . After one year the **apnoea-hypopnoea index** had improved by −17 events/hour (−13 to −21) and **body weight** by −12 kg (−10 to −14) compared with baseline (both P<0.001). . . . At one year, 30/63 (48%, 95% confidence interval 35% to 60%) no longer **required continuous positive airway pressure** and 6/63 (10%, 2% to 17%) had **total remission of obstructive sleep apnoea** (apnoea-hypopnoea index <5 events/hour).'

The outcome variables include:

- body weight (Kg),
- apnoea-hypopnoea index = the total number of complete cessations of breathing (apnoea) and partial obstructions (hypopnoea) for at least 10 seconds during sleep (events/hour).

These are quantitative. They have a numerical value which is not just a code. As they can take any value within a range, they are described as continuous.

The outcome variables also include:

- required continuous positive airway pressure,
- had total remission of obstructive sleep apnoea.

These are both qualitative and as they have only two categories they are dichotomous.

In the paper by Guttmann *et al.,* Association between waiting times and short term mortality and hospital admission after departure from emergency department: population based cohort study from Ontario, Canada (*BMJ* 2011; **342**: d2983), the outcome variables are as highlighted:

'. . . The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay ≥6 *v* <1 hour the adjusted odds ratio (95% confidence interval) was 1.79 (1.24 to 2.59) for **death** and 1.95 (1.79 to 2.13) for **admission** in high acuity patients and 1.71 (1.25 to 2.35) for **death** and 1.66 (1.56 to 1.76) for **admission** in low acuity patients) . . .'.

Death and hospital admission are both qualitative and have two categories so are dichotomous.

In the paper by Kumar *et al.*, Effect of weekly vitamin D supplements on mortality, morbidity, and growth of low birthweight term infants in India up to age 6 months:

randomised controlled trial (*BMJ* 2011; **342**: d2975), the outcome variables are as highlighted:

'. . . Between group differences were not significant for **death** or **hospital admissions** (92 among 1039 infants in the vitamin D group *v* 99 among 1040 infants in the placebo group; adjusted rate ratio 0.93, 95% confidence interval 0.68 to 1.29; P=0.68), or referral to the outpatient clinic for moderate morbidity. Vitamin D supplementation resulted in better vitamin D status as assessed by **plasma calcidiol levels at six months**.'

We have one outcome variable which is qualitative, with two categories making it dichotomous: death or hospital admission.

We have one outcome variable which is quantitative. It can take any value within a range, making it continuous: plasma calcidiol levels at six months.

We have only two types of outcome data in the abstracts of these studies:

- qualitative, dichotomous,
- quantitative, continuous.

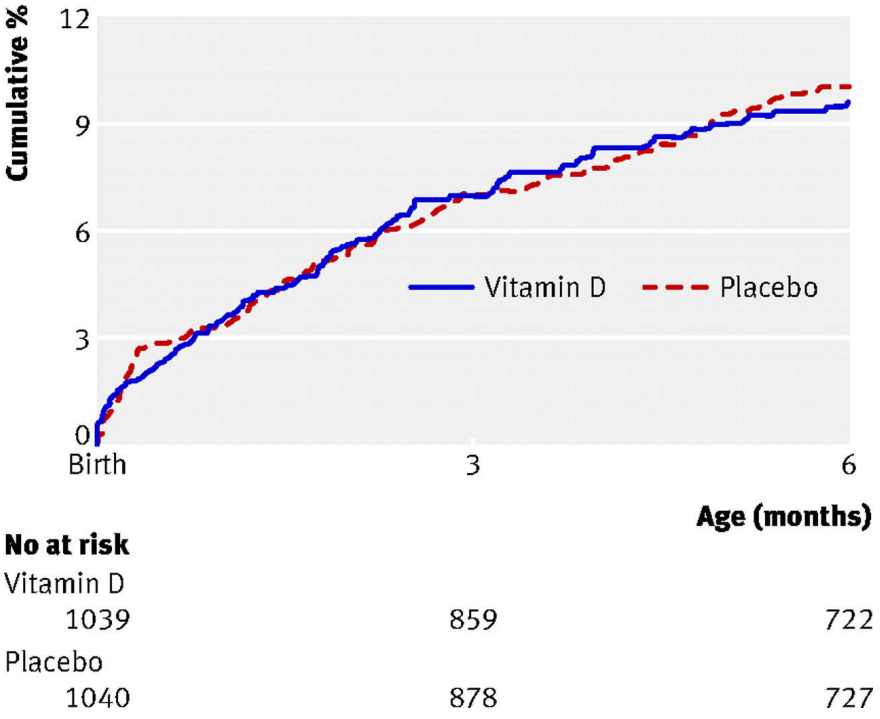
These are the kinds of data most often encountered in health research. There are other types of data:

- time to event, which combines a quantitative element, the time, with a qualitative element, whether the event happens (examples include wound healing, recurrence of a tumour, admission to hospital, or death).
- quantitative, discrete, where only certain values are possible, such as number of falls, number of attendances at hospital,
- qualitative but ordered, where we have more than two categories which have a logical order, such as physical condition or satisfaction with service rated as excellent, good, fair, poor,
- qualitative, multinomial, with more than two categories which are not ordered, such as single, married, divorced, widowed.

These are the kinds of data most often encountered in health research.

In the Vitamin D study of Kumar *et al.*, we have an example of a time to event variable: time from birth to death or admission to hospital. Such are data are analysed by a special set of statistical methods outside the scope of this module, called time-to-event or survival analysis. Figure 1 shows the analysis by Kumar *et al*. Kumar *et al.* also presented an ordered qualitative variable, Vitamin D status classified as adequate (>50 nmol/L), mildly deficient (25-50 nmol/L), or severely deficient (<25 nmol/L).

Figure 1. Time to event analysis by Kumar *et al.* (2011) showing the estimated cumulative proportion of children either died or admitted to hospital, against age.



2. How data can be analysed

If you ever need to analyse data, you are advised to

- never collect data which you don't know how to analyse,
- if in doubt, get advice before you collect anything,
- be sure you have suitable software which you know how to use.

However, you will need to know enough about data analysis to read published research in your own field and what follows should get you started.

Consider this quotation from the paper by Steckelberg *et al.* on 'informed choice' in colorectal cancer screening:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001). More intervention group participants had "good knowledge" (59.6% (n=468) v 16.2% (128); difference 43.5%, 37.8% to 49.1%; P<0.001). . .'

This is a comparison of two groups, intervention and control. Compare the study by Johansson *et al.* on obstructive sleep apnoea:

'. . . After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001). . . . At one year, 30/63 (48%, 95% confidence interval 35% to 60%) no longer required continuous positive airway pressure and 6/63 (10%, 2% to 17%) had total remission of obstructive sleep apnoea (apnoea-hypopnoea index <5 events/hour).'

Here the comparison is a change within one group over a year. We have two types of comparison:

- comparison of two groups,
- change within one group.

These are likely to be the main comparisons in health research studies. We also have two main types of data:

- qualitative, dichotomous,
- quantitative, continuous.

We shall look at how we can carry out each type of comparison for each of these types of data.

3. Compare two groups, dichotomous data

First we shall look at the comparison of two groups using continuous data. Consider this quotation from the paper by Steckelberg *et al.* on 'informed choice' in colorectal cancer screening:

'. . . 345/785 (44.0%) participants in the intervention group made an informed choice, compared with 101/792 (12.8%) in the control group (difference 31.2%, 99% confidence interval 25.7% to 36.7%; P<0.001).'

Here we are looking at the difference between two proportions. The proportions are 44.0% and12.8%, difference = 31.2%, percentage points. As Steckelberg *et al.* have done, we can calculate a confidence interval for the difference. This is quite straightforward. There is an algebraic formula for this, but we shall just say that it works and that it has been included in statistical software by programmers who have read the books and who understand this sort of thing. What we have to know is that the observations must be independent and that we need a 'large' data set for the method to be valid. By independent, we mean that the observations are not related to one another and that knowing whether one person made an informed choice doesn't tell us anything about the choice of another. As these observations are all on different people, this should be fine. By 'large', we mean that there are at least 5 'yes's and 5 'no's per group. If there are fewer than this, the difference between the proportions is unlikely to be very meaningful and we would choose a different approach.

The most popular test of significance for comparison of two proportions is called a chi-squared test. It, too, needs independence of observations and a 'large' data set. It will work if we have at least 5 'yes's and 5 'no's per group, approximately, but the usual condition is a bit more complicated and a bit more liberal. If we have a sample which is too small, we can use an alternative test of significance called Fisher's exact test. This, too, requires independence and it can be used with any sample size.

In the study by Guttmann *et al.* on waiting times and short term mortality and hospital admission, the comparison of two proportions is presented, not as a difference, but as an odds ratio:

'. . . The risk of adverse events increased with the mean length of stay of similar patients in the same shift in the emergency department. For mean length of stay $\geq$6 *v* <1 hour the adjusted **odds ratio** (95% confidence interval) was 1.79 (1.24 to 2.59) for death and 1.95 (1.79 to 2.13) for admission in high acuity patients and 1.71 (1.25 to 2.35) for death and 1.66 (1.56 to 1.76) for admission in low acuity patients) . . .'.

To describe an odds ratio, we must first explain 'odds'. In statistics, the odds of an event is equal to the number experiencing it divided by the number not experiencing it:

$$\text{Odds for admission} = \frac{\text{number admitted}}{\text{number not admitted}}$$

The odds ratio (OR) is then the ratio of two odds. The odds ratio for admission, long wait vs. short wait, is given by:

$$\text{Odds ratio} = \frac{\text{odds of admission, long wait}}{\text{odds of admission, short wait}}$$

You will see a lot of odds ratios in health research. They have several very useful mathematical properties. The down side is that very few of us have a good intuitive insight into what the actual number means. However, we can remember that no difference is represented by OR = 1.0.

This odds ratio is described as 'adjusted'? What does this mean? Adverse events might be related to the particular emergency department. They might be related to patient characteristics including age group, sex, calendar month, weekend/holiday versus weekday, time of day or night, average income level of the patient's neighbourhood and whether rural or urban, number of visits made to an emergency department in the past year, and main complaint. We estimate the odds ratio of adverse events for patients who are the same on all of these but have different waits. The method to do this is called logistic regression, and you will see many logistic regressions used in health research. We need only remember that it is the adjustment method we use when the outcome variable is dichotomous. It is easy when you know how!

If we do not adjust the odds ratio, the significance test is the chi-squared or Fisher's exact test, as for the difference.

In the study by Kumar *et al.* on vitamin D supplements, we have:

'Between group differences were not significant for death or hospital admissions (92 among 1039 infants in the vitamin D group *v* 99 among 1040 infants in the placebo group; **adjusted rate ratio** 0.93, 95% confidence interval 0.68 to 1.29; P=0.68), . . .'.

Rather than the odds ratio, the ratio of two proportions is sometimes presented. This is also called the risk ratio or relative risk (RR). The proportions are 92 / 1039 = 0.0885 = 8.85% and 99 / 1040 = 0.0952 = 0.952%. The risk ratio = 0.0885 / 0.0952 = 0.93. The risk ratio is much easy to have an intuitive understanding of than is the odds ratio. We can find a confidence interval for the rate ratio, risk ratio, or relative risk (RR). It needs the same conditions as the confidence interval for the difference. The significance test is as for the difference.

It is difficult to adjust risk ratios and odds ratios are usually preferred if this is needed. In this study, the ratio presented is not actually a risk ratio, despite numerical identity, but a ratio of rates over time and was calculated and adjusted by a very complicated statistical method, which we shall not consider further.

4. Compare two groups, time to event data

There is another ratio, often used in the analysis of data, which is very similar to the risk ratio. This is the hazard ratio, used for comparisons involving time-to-event data.

Table 1. Baseline characteristics of obese men with moderate to severe obstructive sleep apnoea

	Mean (SD)	Range
Age (years)	48.7 (7.3)	33-61
Weight (kg)	113.1 (14.2)	86.9-139.9
Height (m)	1.80 (0.08)	1.65-2.03
Body mass index (BMI)	34.8 (2.9)	30.2-40.4
(63 men)		

There were no examples in my chosen week in the *BMJ*, but there was a good one in the following week. Sarah Cockayne and her colleagues from this very Department of Health Sciences published the study 'Cryotherapy versus salicylic acid for the treatment of plantar warts (verrucae): a randomised controlled trial' (*BMJ* 2011; **342**: d3271). The Abstract contains:

'There was no evidence of a difference between the salicylic acid and cryotherapy groups in self reported clearance of plantar warts at six months (29/95 (31%) v 33/98 (34%), difference –3.15% (–16.31 to 10.02), P=0.64) or in time to clearance (hazard ratio 0.80 (95% CI 0.51 to 1.25), P=0.33).'

We give the word 'hazard' a special meaning: it is the rate at which events happen. In health research they are usually bad events, hence the choice of word, but not in this case. Clearance of the wart is what we want. The hazard can change over time. For example, easy-to-clear warts might clear quickly early on in the follow-up and we might be left with much more difficult warts which clear at a much slower rate. Our model for the data is that this process will be the same in each group and anything which increases the rate of clearance will do so in the same ratio throughout the follow-up. We can check this, called the proportional hazards assumption, in several ways. The hazard ratio is the ratio of the rate of clearance in the cryotherapy group to the rate of clearance in the salicylic acid group. We can adjust this as we did for an odds ratio, using a method called Cox proportional hazards regression.

We seldom have differences within one group for time-to-event data, as events may only be able to happen once or the chance of subsequent events may be very different, but it can be done.

5. Change within one group, dichotomous data

We can find an estimate and its confidence interval for either the difference between proportions or the odds ratio, called a conditional odds ratio. The test of significance is called McNemar's test. We rarely see this in health research.

6. Continuous data

The paper by Johansson *et al.* on obstructive sleep apnoea includes Table 1. What do these terms mean? By "Mean" we mean the arithmetic mean or average value. By "Range" we mean the smallest and largest observed values. What about "SD"? "SD" stands for "Standard Deviation".

Figure 2. Histogram of heights for 54 men obtained by Research methods students in 2011

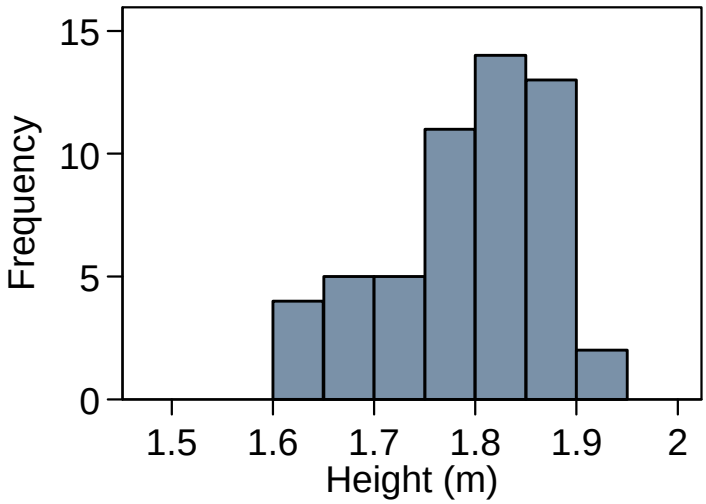
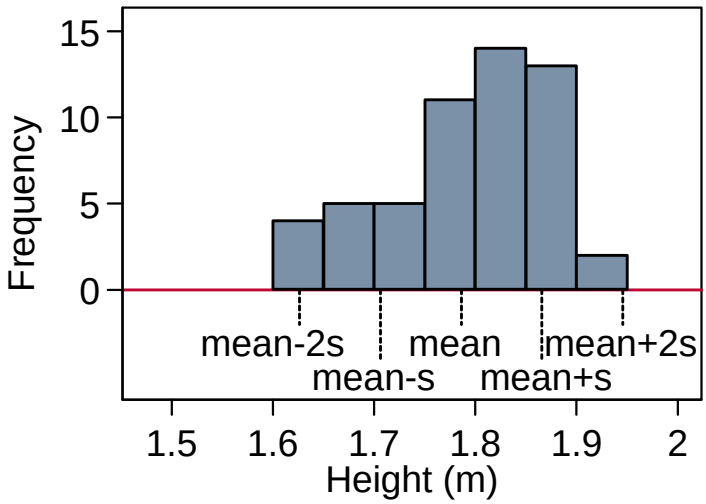


Figure 3. Histogram of heights for 54 men showing the position of the mean height and of the mean minus or plus one and two standard deviations



Standard deviation is a measure of variability. To illustrate it, we shall use some data very similar to those in the *BMJ* paper. When students collected data using the Research Methods module questionnaire designed in 2011, the sample included 54 men who reported their heights. This sample was very similar to the *BMJ* sample:

	Mean (SD)	Range
BMJ sample: height (m)	1.80 (0.08)	1.65-2.03
Research Methods, 2011 sample:	1.79 (0.08)	1.60-1.93

We can illustrate these data further with a histogram (Figure 2). This is a diagram where for each little range of values on the horizontal axis the number of men is shown on the vertical axis.

The mean height is 1.79 m. The position of this in relation to the histogram is shown in Figure 3. As we might expect, it is near the centre of the heights. The standard deviation is 0.08 m and the mean minus one standard deviation is 1.79 – 0.08 = 1.71 m. The mean plus one standard deviation is 1.87 m. From Figure 3, we can see that most of the heights are between the mean minus one standard deviation and the mean plus one standard deviation. The mean minus two standard deviations is 1.79 – 2×0.08 = 1.63 m and the mean plus one standard deviation is 1.79 + 2×0.08 = 1.95 m. We can see that nearly all the observations are between these two values. In general, the majority of observations will be within one standard deviation from the mean and nearly all, about 95%, will be with two standard deviations from the mean. You will see a great many pairs of mean and standard deviation used to summarise data in published research papers.

7. Compare two groups, continuous data

In the paper by Kumar et al. (2011) on vitamin D supplements, the authors say that 'vitamin D supplementation resulted in better vitamin D status as assessed by plasma calcidiol levels at six months.' The following data are given:

	Vitamin D group (n=216)	Placebo group (n=237)
Mean (SD) calcidiol level (nmol/L)	55.0 (22.5)	36.0 (25.5)

The authors say that infants in the vitamin D treatment group had significantly higher plasma calcidiol levels at six months; crude mean difference 19.0 nmol/L (95% confidence interval 14.7 to 23.5; P<0.001). For the difference between two means, we can find a confidence interval or test of significance by two methods: the large sample Normal method (z method) or two sample t method (small samples). The Normal method uses something called the Normal distribution, a theoretical, mathematical entity which has many remarkable properties. We shall learn more in later sessions.

What is a small sample? My rule of thumb for this is that a sample should be treated as small if there are fewer than 50 observations in either group. For large samples using the Normal or z method we need only have observations which are independent. For small samples using the two sample t method there are other conditions the data also must meet:

- each sample, i.e. the observations themselves, must come from a Normal distribution,
- We must have the same variability in both populations.

There are several ways to check these assumptions. A histogram of our data is one way to do this. Figure 4 shows a histogram for a sample of 1,603 birthweights for babies born at >37 weeks gestation at St. George's Hospital, London. This also shows the Normal distribution which corresponds to these data, showing how well the birthweights fit to it.

If we want to use the two sample t method but our data do not meet the conditions, we can:

- find a mathematical transformation of the data which does,
- use other methods which don't need them (not so good).

Figure 4. Birthweights of 1603 singleton term births showing the Normal distribution with the same mean and standard deviation

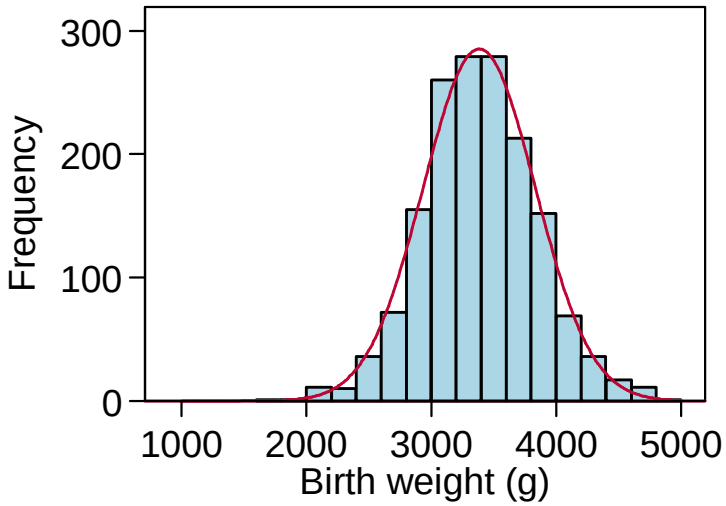


Figure 5. Histograms for 86 serum cholesterol measurements in stroke patients, before and after logarithmic transformation, with corresponding Normal distribution curves

Figure 5 shows Histograms for 86 cholesterol measurements in stroke patients, before and after logarithmic transformation, with corresponding Normal distribution curves. The data fit the curve much better after transformation. Almost all concentrations measured in blood behave like this and we often analyse their logarithms.

Kumar *et al.* (2001) say:

'After **adjustment** for sunlight exposure and for factors associated with not having a result for calcidiol, the adjusted mean difference was 18.7 nmol/L (14.2 to 23.5; P<0.001).'

This adjustment is similar to the one described in Section 3 for the study by Guttmann *et al.* on waiting times. That was for dichotomous data using odds ratios and logistic regression. For continuous data like the calcidiol concentrations of Kumar *et al.*, the appropriate method is called multiple regression or ordinary least squares regression, two names for the same thing.

8. Change within one group, continuous data

In the study by Johansson *et al.* on obstructive sleep apnoea, the authors say:

'After one year the apnoea-hypopnoea index had improved by −17 events/hour (−13 to −21) and body weight by −12 kg (−10 to −14) compared with baseline (both P<0.001).'

Here they have found the mean difference and a confidence interval and P value for it. As for two samples, we can use a large sample Normal or z method or, for small samples, we can use the one sample t method. Here a guide to what is a small sample is one with size less than 100.

For the large sample, the only condition is that the pairs of measurements are independent. For the t method, there are additional conditions the data must meet:

- Normal distribution for differences,
- mean and variability for differences are the same throughout the scale.

There are several ways to check that these conditions apply, similar to those for two samples. We shall leave them for another lecture.

9. Still to come in Research Methods:

- how to analyse data in practice using SPSS,
- data entry,
- histograms, means, standard deviations and other statistics,
- comparing means,
- comparing proportions, odds ratios and risk,
- regression and adjustment.