Good analytical practice: statistics and handling data in biomedical science. A primer and directions for authors. Part 1: Introduction. Data within and between one or two sets of individuals

A. D. BLANN* and B. R. NATION†

*Haemostasis, Thrombosis and Vascular Biology Unit, University Department of Medicine, City Hospital, Birmingham, B18 7QL; and †Institute of Biomedical Science, 12 Coldbath Square, London EC1 5HL, UK

Accepted: 21 August 2008

ABSTRACT

The biomedical scientist is bombarded on a daily basis by information, almost all of which refers to the health status of an individual or groups of individuals. This review is the first of a two-part article written to explain some of the issues related to the presentation and analysis of data. The first part focuses on types of data and how to present and analyse data from an individual or from one or two groups of persons. The second part will examine data from three or more sets of persons, what methods are available to allow this analysis (i.e., statistical software packages), and will conclude with a statement on appropriate descriptors of data, their analyses, and presentation for authors considering submission of their data to this journal.

KEY WORDS: Correlation of data.
Data interpretation, statistical.
Mann-Whitney U text.
Statistics.
Student's t-test.

Introduction

Many biomedical scientists are involved in the generation of numerical or descriptive data derived directly from human body tissues. For haematologists, the data generated may be the number of red and white blood cells in a sample of venous blood, while for biochemists the data may be the level of glucose in that blood sample. Microbiologists may collect data on which particular microbes are present in a sample of sputum from a patient with a lung problem, and histologists may study lung tissues from that patient: the data they handle may determine the presence or absence of cancer. Data may not only be at a single time point, as in some cases we are interested in how a molecule (perhaps cholesterol) changes as someone undergoes a change in their diet and lifestyle.

In biomedical science – and, indeed, in almost all branches of science – information can be described as one of two types. **Quantitative data** are information that have been assigned a directly measurable numerical value. Examples of this are height, weight, age, red blood cell count, temperature, serum potassium, or perhaps the proportion or percentage of patients with a particular problem (e.g., an infection, a risk factor or a cancer) or a particular ABO blood group. In almost all cases, the value of the data is defined not by an individual but by an objective observation or scale, itself often derived from a machine. **Qualitative data**, on the other hand, are information gathered and analysed primarily as words (singly or in phrases), possibly in narrative form or with descriptive quotations. The proponents of qualitative research quite reasonably argue that not all information can or should be reduced to a number. For example, how can one reasonably quantify (place a numerical value on) someone's attitudes or opinions, such as a belief, fear or love? Qualitative data are often obtained from people in surveys, interviews or observations, generally from a validated or structured questionnaire. An alternative would be to observe and record verbal or written comments from a defined focus group. However, as the vast majority (if not all) of the information collected and presented to the biomedical scientist will be quantitative, we focus on these types of data.

The purpose of this document is to provide a broad overview of the issues surrounding data handling and statistical analysis as it applies to biomedical science. We will explore the different ways in which data can be described and analysed, and consider issues that relate to best practice when handling data.

Quantitative data

Broadly speaking, almost all quantitative data fall into one of two types: that which is categorical and that which is continuous. **Categorical data** fits into one of any number of discrete boxes or categories – there are no 'in betweens'. An example of this is the number of men and the number of women in a particular group. Another example is children. In real life, you cannot have a fraction or proportion of a child, only whole numbers. This concept can be extended to more than two discrete groups. For practical purposes,

Correspondence to: Dr. Andrew Blann
Email: a.blann@bham.ac.uk

almost all people belong to one of four ABO blood groups (A, B, AB and O). Microbiologists can say with a good degree of confidence that an organism (e.g., methicillin-resistant *Staphylococcus aureus* [MRSA]) is present or absent – there should be no in-between. Immunologists can say 'yes' or 'no' to the presence of antibodies to microorganisms and whether or not a patient with rheumatoid arthritis is seropositive or seronegative. Histopathologists will tell us that a tissue either is or is not invaded by a cancer, while cytologists report the presence or absence of certain abnormal cells.

Continuous data include factors such as height, weight, blood pressure, and just about all haematology and biochemistry results (e.g., the number of red and white blood cells and levels of triglycerides and sodium in the serum). The data, consisting of individual numbers, can be described by almost any figure in a given range (e.g., age, which can be anywhere between 0 and 100 years plus). The presentation and analysis of continuous data is more complex than if the data were categorical. When the data are continuous we need to consider three aspects: the central point or tendency, the variance, and the distribution.

Central point

The most important statistical aspect of data that has a continuous distribution is the central point. Some statistician prefer to use the word 'tendency' (which may seem rather inexact) instead of 'point', which is not unreasonable as there are actually two such values within a single data set – the average value and the middle value. These two different central points are the mean and the median, respectively. We obtain the **mean** point of a set of data simply by adding up all the data and then dividing by the number of individual data points. So, when we talk about an 'average' we are often referring to the mean. An alternative way of arranging a set of numbers is to list them in rank order (i.e., from the lowest to the highest). If we do this, the data point in the middle of the entire series is the **median**. So, the median value is arrived at by a completely different set of rules than is the mean value. Thus, each set of data has both a mean and a median. Sometimes they are the same number (or are very close together), but in other data sets they may be very different.

Variance

A second important concept in a set of data is its **variance**. This index gives us information such as the highest and lowest points in a particular group of data, and the extent to which the data cluster tightly around the central point of the mean or the median. If it is tightly clustered, the data are said to have low variance: if more diverse (spread out), it has high variance. We use one of two particular measures of variance depending on the nature of the central point (the mean or the median). When we use the mean, we use standard deviation to describe the variance. However, when using the median then the variance is described in terms of the inter-quartile range.

The **standard deviation** (SD) provides an idea of the degree to which the data are clustered close about the mean value or are more spread out (i.e., the variance). For example, in a data set with a mean of 100 and a SD of 20, perhaps two-thirds of the data points lie within one SD either side of the mean (i.e., between 80 and 120). However, we can go further,

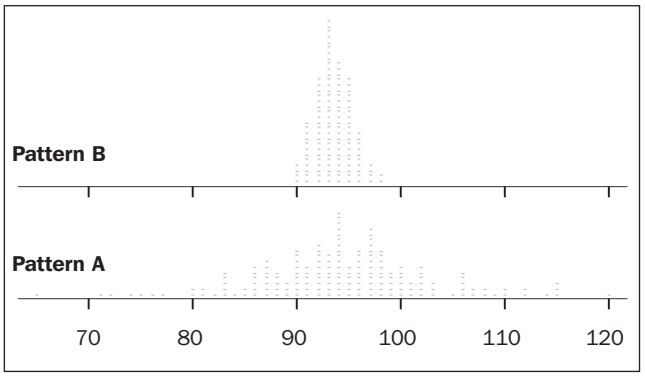

Fig. 1. Variance. Pattern A and pattern B both have the same central point value of about 95. However, the two patterns have very different variances – pattern A (lower) covers a much greater range (from about 65 to 120) compared to pattern B (upper), which runs from 90 to about 98. So, in both cases the mean is 95, but the SD of pattern A is 9, while the SD of pattern B is only 2.

and often find that nearly all the data points (about 95% of them) are between 60 (i.e., the mean minus two SDs) and 140 (i.e., the mean plus two SDs). However, if another data set has a similar mean of 100 but a much smaller SD (e.g., 8) then 95% of the data points should range between 84 (i.e., 100 minus 16) and 116 (i.e., 100 plus 16). So, a small SD tells us that the set of data is tightly clustered near to the mean, and when the SD is large it means that the data are more spread out. The mean and SD of the first data set is described as 100 (20), although some (erroneously) use the notation 100 ± 20. Similarly, the second data set can be written as 100(8). Figure 1 illustrates this concept.

The second method of describing variance, the **inter-quartile range** (IQR), is derived from the data points that are one-quarter and three-quarters of the way into the complete data set when it is ranked from lowest to highest. So, starting from the lowest point, and working our way up, after we have looked at a quarter (i.e., 25%) of the data points, we have the 25th percentile. Continuing to work up the data set, the half-way point (i.e., after we have looked at half [50%] of the data points) is the 50th percentile (i.e., the median – as we have arrived at the middle of the data set). Continuing up the rank, after we have seen 75% of the data points, we have arrived at the 75th percentile. Finally, the highest value is the 100th percentile, as we will have assessed all (100%) of the data points. The IQR, like the SD, also gives us an idea of the spread of the data. We would write a summary of this type of data of the median and IQR as 77 (49–148). Note that the median value here (77) is certainly not in the middle of the IQR of 49 to 148. Clearly, 77 is much closer to 49 than it is to 148, and this fact is an important feature of these types of data. However, the data set 77 (70–100), with the same median, has a much smaller spread (range) of data points than the set 77 (49–148).

Distribution

We bring the concepts of the central point and variance together to describe the distribution of a set of data. The vast majority of data in biomedical science are generally of a **normal distribution**, and include many laboratory and physiological indices such as haemoglobin, albumin, height and body mass index. This distribution may be described as 'bell shaped'. An important component of data with a

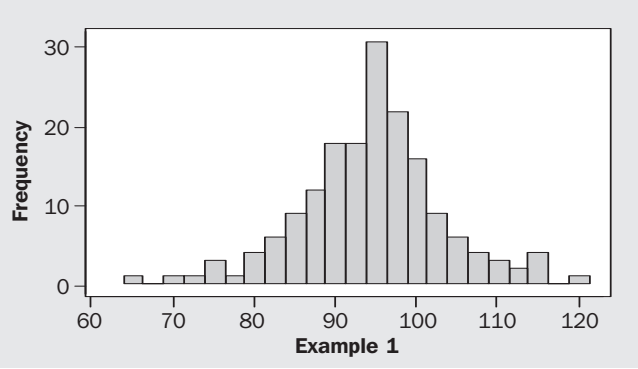

Fig. 2. Data distributed normally. This shows the distribution of a set of data arranged as a series of columns (histograms). The 'height' of each histogram (described as the frequency on the vertical Y axis) depends on the number of data points in each particular histogram. A 'tall' histogram has more data points than a 'short' histogram. The data (on the horizontal X axis) run from the lowest of 65 to the highest of 120, with a central point value of about 95. The tallest histogram column (with a value of about 95) is roughly in the middle of the entire set of data, with roughly an equal number of histogram columns above (nine columns to the right) and below (eleven columns to the left) the highest column. This means that we can be fairly confident that the data are distributed normally, even without knowledge of the mean and SD. However, we can safely predict that the mean value is about 95.

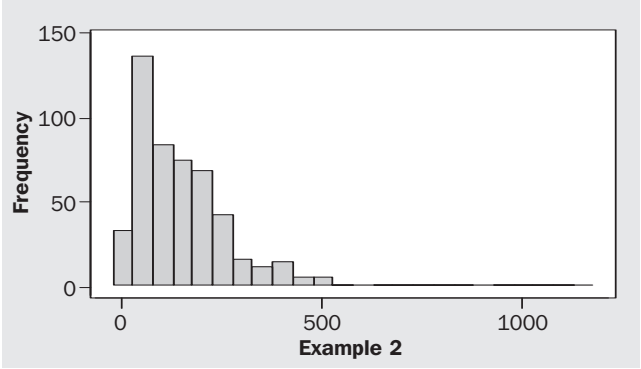

Fig. 3. Data distributed non-normally. This set of data has some aspects in common with the data in Figure 2. The data are presented as histogram columns and the 'height' of the histogram is also related to the number of data points present. However, many other differences are present. Perhaps the most obvious is that the 'tallest' histogram is not in the centre of the set of histograms, but is over to the left hand side. The data values run from zero on the left to 500 on the right, although the computer software that drew this plot has taken the X axis to 1000. The most crucial aspect of this dataset is that, unlike the normally distributed data set in Figure 2, there is only one histogram to the left of the tallest histogram, but nine histograms to the right. Therefore, we can say that the data are skewed to the left. This means that we can be fairly confident that the data are distributed non-normally. Other analyses, such as the size of the SD compared to the mean, or the difference between the mean and median values, would be needed to confirm this. Nevertheless, we can predict with a reasonable degree of confidence that the median value is around 50.

normal distribution is that we take the mean to be the central point of the data set. When data are normally distributed, the mean and median are close together (e.g., 26.6 and 27), and the SD is far smaller than the mean (e.g., 100[20]). Figure 2 provides an illustration of a data set that is normally distributed. Data that has a **non-normal distribution** is less common, perhaps the best example from biochemistry being levels of serum triglycerides in a healthy population, and from physiology the duration of a pregnancy. These types of data may also be described as 'skewed'. An important characteristic of data with a non-normal distribution is that the mean and the median are always far apart (e.g., a mean of 35.8 and a median of 25), so that the average of the dataset is not the middle point. Furthermore, when data are non-normally distributed, the SD is often quite large compared to the mean. Thus, a mean (SD) of 60 (45) is very likely to be of non-normal distribution. So, in these cases we use the median to define the central point and the IQR to define the variance. These points are illustrated in Figure 3.

The nature of the distribution is important because it is needed to be able to apply the correct statistical test to the data, or decide if a particular result from a particular patient warrants attention. Choice of the wrong statistical test may lead to an incorrect answer. Broadly speaking, there are a few simple rules about whether or not data are distributed normally or non-normally. If the SD is much smaller than the mean then the data are very likely to have a normal distribution (e.g., a mean of 100 with an SD of 15). However, if the SD is large compared to the mean (e.g., 100[90]), then the data are very likely to be distributed non-normally. The question, therefore, is how much smaller does the SD have to be compared to the mean for it to be normally distributed? A frequently cited rule of thumb is generally about a third. So, if the SD is up to a fifth of the mean (as above), a normal distribution is expected. If the SD exceeds the mean then the distribution is unquestionably non-normal.

If the mean and median are close together, the data are likely to be normally distributed (e.g., mean 26.6 and median 27 [a difference of only 1.5%]). However, if the mean and median are far apart (e.g., 35.8 and 25 [a difference of 43%]) then the data are very likely to be non-normally distributed. A big problem with these two methods is that they work very well only for data where the nature of the distribution (e.g., mean 34, SD 2.5) is reasonably obvious and/or you have a great deal of experience. Fortunately, all good statistical software packages have programs that determine distribution. Details of these statistical packages will be presented in Part 2 of this article, to be published in the next issue of *BJBS*.

Analysis and interpretation of data

Biomedical scientists use scientific methods to understand disease processes with a view to providing a diagnosis and then monitoring treatment. For many, this will be to compare the pathology result from an individual whose health status is unclear with that from a group of individuals we know to be healthy. Other studies (possibly in research) may compare data from a group of individuals (e.g., with a particular complaint, symptom or frank disease [often referred to as the cases]) with that from another group generally considered to be free of the problem (the controls – hence, a case-control study). In both types of analysis, the control group provides the normal or **reference** range.

Reference range

We need to know about distribution so that we can accurately determine if there is a statistically significant difference between different sets of data, or if the result from a particular individual is abnormal when compared to a large number of healthy persons. Illustrating the former, the blood pressure of one group with mean (SD) of 156 (25) mmHg is clearly different from that of another group of 132 (21) mmHg. Indeed, this difference of about 18% seems large and so may be significantly different. However, precise statistical tests must determine whether or not such a difference does not merely 'seem' significant but is genuinely statistically significant.

A second 'need to know' about the distribution of an index is in the ability to define a particular result from a certain individual (e.g., a patient with a serum triglyceride result of 2.1 mmol/L) as normal or abnormal. This is important as a high level may indicate a particular syndrome, the presence of a particular disease, or an increased risk of a heart attack. In this case we need to compare the triglyceride level from that individual with the results of the same test from a large number of people we know to be healthy. Data from this large group make up what is called the reference range. We need to know the distribution of triglyceride data in this healthy population in order to be able to make a judgement about whether or not a result from an individual is within or outside this reference range.

Let us suppose that 100 healthy people (referred to as the sample size) provide overnight fasting blood for a triglyceride test, and that the results have a non-normal distribution, with 95% of the results lying between 0.5 mmol/L and 1.9 mmol/L. Under these conditions the result for the individual in question (2.1 mmol/L) only just exceeds the top of this reference range, and so may possibly require additional investigation. However, if the reference range has a normal distribution, it will provide different values that we would consider to be healthy (e.g., between 0.8 mmol/L and 1.5 mmol/L). If so, the result from the individual patient of 2.1 mmol/L is of greater significance because it is much higher than the top of the reference range and so may have significant clinical repercussions.

Hypothesis, sample size and power

A key step in research is to be able to describe what it is you want to find out in terms of a question, and then turn it into a statement, generally called an **hypothesis**. Indeed, much of the research process is called hypothesis testing. People with an interest in heart disease may ask "Does the cholesterol level in one group differ from that in another group?" This can be turned into a formal hypothesis statement such as "the mean cholesterol level of one group is 0.5 mmol/L higher than that in a different group". Other researchers may form their hypothesis as "the mean systolic blood pressure of a group of patients will be reduced by 10 mmHg if they regularly take a particular drug". Once the hypothesis has been formed, a precise mathematical calculation is required to determine how many people (the sample size) are to be recruited in order to ensure that any findings are reliable – this is couched in terms of power. If a study is underpowered (i.e., has not recruited sufficient people – the sample size is too small), then the difference in levels of cholesterol between the two groups, although seeming large, may not be significantly large. A statistician should be consulted at an early stage to define the correct number of subjects to be recruited (i.e., to perform a **power calculation**). A general rule of thumb is the equation 'significance = difference x power'. What this means in practice is that a large difference between two data sets many not be statistically significant because of low power (i.e., the sample size is too small).

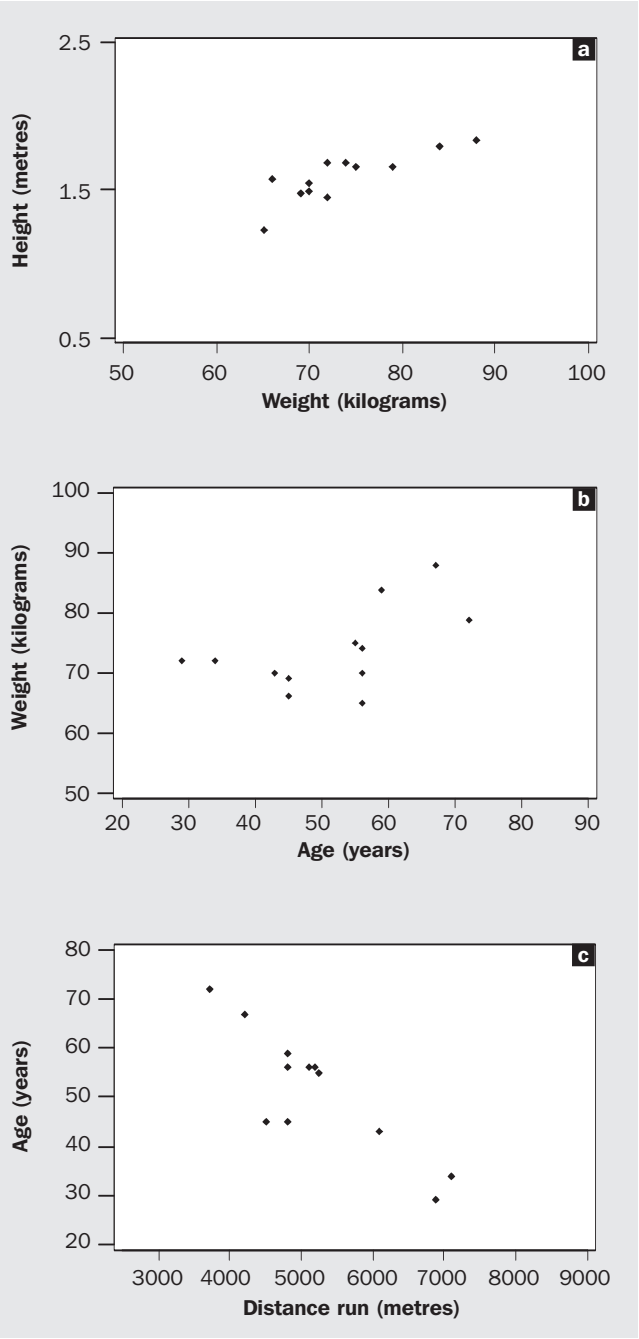

Fig. 4. Correlation. a) Relationship between height and weight. b) Relationship between weight and age. c) Relationship between age and distance run. See legend to Table 1 for an explanation.

Probability

It is a well-established fact that normal, healthy adult men are taller than normal, healthy adult women. But this may not be the case in all populations; for example, women who have grown up with high levels of growth hormone are very

Table 1. Correlation. Consider these data from 12 people: their height, weight, age and distance that they can run in a certain fixed time period such as 30 minutes.

Subject	Height (metres)	Weight (kilograms)	Age (years)	Distance run (metres)
1	1.56	70	43	6100
2	1.23	65	56	5200
3	1.70	72	29	6900
4	1.81	84	59	4800
5	1.46	72	34	7100
6	1.50	70	56	4800
7	1.59	66	45	4500
8	1.66	79	72	3700
9	1.70	74	56	5100
10	1.48	69	45	4800
11	1.85	88	67	4200
12	1.66	75	55	5250

By plotting the height and weight for each person, a graph is obtained (Fig. 4a). Because the two indices have a normal distribution, Pearson's method is used. In doing so, $r=0.85$ is obtained, which indicates a strongly positive relationship. It also happens that this is statistically significant ($P=0.001$).

Similarly, a plot of each person's weight against their age (Fig. 4b) also gives the impression that as someone's age increases, so does their weight. Indeed, the correlation coefficient for these data is $r=0.56$, which would normally be quite respectable. However, the probability that these data are a true reflection of the relationship between age and weight just fails to reach statistical significance as $P=0.06$ (i.e., the likelihood of a difference being genuine is 'only' 94%). It is very probable that the addition of three or four more data points to this set, with little change to the correlation coefficient, would reduce the P value (e.g., to $P=0.045$) and so would be significant purely because of the increased power. This is because the strength of the probability = difference (r value) x power (sample size). It follows that a sample size of 12 in this case may not be large enough, and so the analysis is underpowered.

The relationship between age and the distance run is presented in Figure 4c. Analysis points to an excellent correlation coefficient where $r=0.87$, slightly better than the relationship between height and weight. Indeed, the probability that these data truly reflect a real association is highly significant, with $P<0.001$ (i.e., greater than 99.9%). However, as the relationship is inverse, we have to place a minus sign before the correlation coefficient (i.e., $r = -0.87$). If all the data points were to be found on a straight line, the correlation coefficient would be –1. In practice, however, this is almost never found in biomedical science.

likely to be taller than men who have grown up with low levels of growth hormone. Despite this, at the practical level, we can predict that the average height of even a small group of men is likely to be taller than the average height of a similar sized group of women. However, statisticians use the word **probability** (abbreviated to P) to give a more scientific and secure basis to this likelihood. In this setting, we are keen to establish whether or not a difference in two sets of data is genuinely due to, for example, a pathological process, or is due simply to chance.

Statisticians have developed a consensus which says that we are prepared to accept a difference as real if the probability of it being a real effect (i.e., not due to chance) is greater than 95% (i.e., 19 times out of 20). We express 95% in the decimal form as 0.95. So, in accepting that there is a 95% probability that the difference is real, then we also accept that there is a 5% (i.e., one out of 20) probability that the difference could be coincidental – that is, due to chance. We express 5% in decimal form as 0.05. Hence, our requirement is for P to be less than 0.05 (i.e. $P<0.05$). It follows that if $P=0.06$ (i.e., we have a likelihood of 94% that the difference is genuine), we do not consider this to be of sufficient reliability, and so describe it as statistically not significant.

It follows that if the chance of an effect being spurious is only 1 in 10 (i.e., $P=0.1$) then this difference is not statistically significant because 0.1 is greater than 0.05. What would be even more significant would be if a difference is so large that the probability of such a difference being spurious or coincidental in only 1 in 200 (i.e., 0.5%), which means that the chances of the effect being real are 199 in 200 (i.e., 99.5%). Thus, a probability of one in 200 gives $P=0.005$, a result that is considerably less (10 times less) than 0.05. Overall, the smaller the P value, the greater is the likelihood that the difference is real and not due to chance.

Analysis of data from an individual

Is a serum total cholesterol result of 6.1 mmol/L of concern? This depends on the age, gender and some other health issues. But an initial question may be "Is this result within the reference range?" As discussed, the reference range is composed of results from hundreds or even thousands of supposedly healthy people (e.g., blood donors). Using the example of cholesterol, we would expect data from a large pool of healthy people (e.g., 400) to have a normal distribution, with a mean of perhaps 4.5 mmol/L and an SD of maybe 0.5 mmol/L. As already described, an important component of the relationship between the mean and SD of a normally distributed index such as cholesterol is that the results from 95% of those people (i.e., about 380) will lie between 3.5 and 5.5 mmol/L – these numbers being derived from the mean value plus two SDs and the mean value minus two SDs.

The figure of 95% of the data points (as opposed to, for example, 70% or 80%) is chosen because it is most likely to provide representative and reliable information. So, if 95%

Table 2. Data at two time points.

	Level of creatinine (mmol/L)		
Patient number	Before the drug	After the drug	Difference
1	150	126	–24
2	225	196	–29
3	176	168	–8
4	166	172	+6
5	145	140	–5
6	226	196	–30
7	189	168	–21
8	168	173	+5
9	173	146	–27
10	149	144	–5
11	171	155	–16
12	156	149	–7
Mean (SD)	174 (27)	161 (21)	–13 (13)
Median (IQR)	169 (151–186)	161 (144–172)	–12 (5–26)

These illustrative data show the effect of a new drug on the function of the kidney. We need to obtain a blood sample before the drug is issued, and then place the patients on the new drug for perhaps several months, and then take a second sample of blood. Note that in 10 of the 12 patients, the level of creatinine has fallen. However, in two patients (numbers 4 and 8) levels have increased. Nevertheless, overall there has been a decrease in creatinine that, assuming no other changes in the lives of the patients, implies that the drug may be active in alleviating the renal disease in these patients.

To be fully confident, however, we need to apply the correct statistical test to the data. The choice is between a paired *t*-test and the Wilcoxon test. Although the mean difference (–13) is the same as the SD (13), it is virtually the same as the median difference (–12), so we can be fairly confident that the data are normally distributed and so the appropriate test would be a paired *t*-test. This test will give $P=0.004$, which is considerably smaller than the required cut-off point for statistical significance ($P<0.05$). In fact, this *P* value tells use we can be 99.6% confident that the effect is real, and only 0.4% confident that the difference is due to chance.

In this type of study we should also measure creatinine in a similar group of patients who have not been taking this particular drug over the same time period. We call these subjects the control group – this is a vital consideration when designing and carrying out experiments.

of the results from this population are within two SDs either side of the mean, what about the remaining 5%? There are likely to be as many at the bottom end of the scale as at the top end of the scale, and in our example it means that 10 people will have a cholesterol level below 3.5 mmol/L, and another 10 will have a result over 5.5 mmol/L. It is important to point out that these people are not immediately considered to be in ill-health.

So, given the above criteria, a serum cholesterol of 6.1 mmol/L is certainly above the top of the reference range of 3.5 to 5.5 mmol/L. But is this person in ill-health? This we cannot say as more details are needed before we give the individual a potential diagnosis, which in this case would be hypercholesterolaemia. Certainly, we feel that the individual should be made aware of this high cholesterol result and generally instructed on the risk factors for atherosclerosis and their contribution to heart attack and stroke. Although the cholesterol result of 6.1 mmol/L may be markedly raised compared to the normal range, it is generally inappropriate to say it is 'significantly' raised in the statistical sense of the word. However, it is possible to define the exact probability that a single laboratory result is outside the normal range, but this analysis is complex and is beyond the scope of this document. We reserve the use of expressions such as significance and probability for situations in which there is a different mode of analysis, such as that of data from groups of individuals or sets of data.

Analysis of data from two groups of individuals

In looking at data from two groups of individuals, we apply different statistical tests depending on the nature of the data and what hypotheses we are testing.

Continuously-variable data from two different groups of subjects

It is accepted that total cholesterol data are normally distributed, and so data would be presented as mean and SD. The correct test to use in this case is **Student's *t*-test**. If the data from one group of perhaps 30 persons is 5.9 (0.7) mmol/L, and that for a different group of 30 is 5.1 (0.6) mmol/L, then by applying the *t*-test we get a *P* value of 0.025. The power calculation (giving a sample size of 30 persons per group) will ensure that this number of subjects is large enough to provide confidence that the result will be reliable. The *P* value of 0.025 is less than our cut-off point of $P<0.05$, so we can say with confidence (i.e., $P=0.975$, or 97.5%) that the difference is statistically real and is not spurious. By contrast, serum triglycerides have a non-normal distribution and would be presented with a median and IQR. Such data would be analysed by the **Mann-Whitney U test**. If the results for the first group is 1.7 (1.2–2.8) mmol/L, and the second is 1.1 (0.9-1.6) mmol/L, then application of the Mann-Whitney U test would give a probability value of $P=0.002$. This is a highly significant difference, as we can say that the

probability that the difference is real is 99.8%, while the probability that the difference is spurious is only 0.2%. Therefore, there is a difference in total cholesterol levels of $P=0.025$, and a difference in triglyceride levels of $P=0.002$. Thus, the difference in the triglyceride data is greater than the cholesterol data. The questions that follow may include "why are these differences present?" and, perhaps later, "if these differences have important consequences, do we need to act on either of them?"

Two sets of continuously variable data from a single group of subjects

Not only can we compare a single laboratory index (e.g., serum cholesterol) from two different populations (as above) but we can also look at two different sets of data (e.g., serum cholesterol and blood glucose) within a single population. We may predict (or perhaps hypothesise) that the two are related in that those people who have a high cholesterol will also have raised glucose, and those who have low cholesterol will also have low glucose. We are therefore predicting a **correlation** between glucose and cholesterol, and in doing so we seek a **correlation coefficient**, represented by the Greek letter Rho (r). A very strong association between two sets of indices would be represented by a correlation coefficient where r is close to 1 (e.g., 0.92), whereas we would consider the relationship to be weak if we obtained a small r value of perhaps 0.15. Once more we need to perform a power calculation to ensure that the number of data points (sample size) is large enough to provide meaningful data. For example, a correlation coefficient (r) of 0.65 may seem good, but the probability that this association is genuine (i.e., $P<0.05$) will only occur if the sample size is sufficiently large. Conversely, large epidemiology studies often provide very significant P values (e.g., $P<0.001$) on correlations that seems poor (e.g., $r = 0.15$), merely because the sample size is of thousands of people.

We must be cautious about interpreting correlations. The fact that two indices correlate strongly does not necessarily mean that one causes the other. A crucial comment is always to recall that correlation does not imply causation. An excellent example of this is the relationship between height and weight. It is quite well established that, in general, tall people are heavier than short people, but it is just as well known that two people of the same height can have very different weights (and vice versa). It is also evident that, in general, height correlates with weight. But is this because the taller you are, the heavier you become, or is it that the heavier you are, the taller you become? The former seems more likely. A better example in biomedical science is from human epidemiology and clinical studies where we know that systolic blood pressure in a large population generally correlates very strongly with diastolic blood pressure. But the increased systolic value does not cause the diastolic value to rise – the factors that act to increase systolic blood pressure also act to cause diastolic blood pressure to rise. As a result, both systolic blood pressure and diastolic blood pressure rise in parallel, but do so independently of one another.

As in comparing two groups of data from different populations, where we use Student's t-test if the data are normally distributed and the Mann-Whitney U test if they are non-normally distributed, we have a choice of different tests of correlation. If the two sets of data have a normal distribution (e.g., height and weight), we use **Pearson's correlation method**. However, if one or both sets of data have a non-normal distribution (e.g., total cholesterol and triglycerides) we have to use **Spearman's correlation method**. These aspects are illustrated in Table 1 and Figure 4.

Table 3. Key aspects of good analytical practice.

• Quantitative data are generally of two forms: that which is continuously variable and that which is categorical.
• Data which are continuously variable must be subjected to a formal test to define distribution. The two most common forms of distribution are normal and non-normal.
• Data which are normally distributed should be presented as mean and standard deviation. Differences between two data sets of normal distribution should be sought using Student's t-test.
• Data which are non-normally distributed should be presented as median and inter-quartile range. Differences between two data sets of non-normal distribution, or between one set of data normally distributed and one set non-normally distributed, should be sought using the Mann-Whitney U test.
• Differences between data sets which are categorical should be sought using a test such as the Chi-squared (χ^2) test.
• If possible, research questions should be formed in terms of an original hypothesis. If possible, such an hypothesis should be quantified. A quantified hypothesis should be supported by a power calculation to define the sample size.
• Any relationship between two sets of data may be sought by correlation. For data normally distributed, Pearson's method is appropriate. If one or both sets of data are non-normally distributed, Spearman's method is appropriate.
• Differences in directly linked pairs of data (e.g., serial data) should be sought using a paired t-test (if the difference between the linked pairs has a normal distribution) or Wilcoxon's method (if the difference has a non-normal distribution).
• In order to be assured that any difference we have is genuinely ascribable to a defined pathology, and is not due to chance, we require that the probability of chance is less than 5% (i.e., $P<0.05$). It follows that we require the probability to exceed 95% in order to be confident that the difference is genuine.

Analysis of continuous data obtained at two time points

So far we have been looking at data that are generally assessed at a single time point. However, it is also important to be able to assess whether or not there is a change in a particular index in a group of people at two time points. These types of data are called 'paired' because there are always pairs of figures (e.g., before and after), never just a single point. In clinical medicine, as part of the development of new therapeutics, we need to know if a new drug genuinely influences the biological system it is designed to act upon. For example, to be convinced that a new drug designed to lower blood pressure actually does so, we need data on the reduction of systolic blood pressure and/or diastolic blood pressure from the same people before they are placed on the drug, and then again later, once they have been taking the drug for perhaps weeks or months. Exactly how much the drug can be expected to reduce blood pressure will determine, via a power calculation, the

minimum number of patients to be recruited. Table 2 provides the opportunity to explain these points.

The exact type of statistical test depends on the nature of the distribution of the data: if the difference between the two sets of data is normally distributed then a **paired *t*-test** is appropriate. However, if the difference between the two sets of data is distributed non-normally then **Wilcoxon's test** should be used. It is important to note that in both cases it is not the distribution of the original data or the follow up that is important, but the difference between them. Paired analyses need not be linked in time but may be linked in other ways (e.g., the systolic blood pressure in the left arm compared with that in the right arm in the same person, or the difference in a component of blood measured in serum compared to plasma from the same person).

Analysis of two sets of categorical data

An example of this class of data is the number of people who fall into only one of a small number of well-defined categories (e.g., being male or female, whether or not you regularly smoke cigarettes). In pathology, alternatives may be the presence of a certain cancer, the presence of a condition such as diabetes, or a history of heart disease. Taking the last as an example, an appropriate hypothesis may be "people with diabetes have twice as much heart disease as people without diabetes". In order to test this hypothesis we need to examine the presence of heart disease in two groups of people – those with diabetes and those free of diabetes. We also need to define heart disease precisely (e.g., history of angina, having had a heart attack or heart surgery). In this comparison, we describe those free of diabetes as the control group.

The next step would be to find out how many people we need to recruit in order to have sufficient statistical power for our result to be reliable. If we expect a frequency of heart disease of 10% in the healthy control group, then our hypothesis would suggest a doubling in the frequency (i.e., to 20%) in the group with diabetes. The power calculation calls for a sample size of 200 (i.e., 100 people in each group). As with the definition of a normal or a non-normal distribution, statistical programs are available that will define precisely the exact number of people a study requires (i.e., the sample size).

Following ethics committee approval, data collection may begin. Perhaps the best source of the 100 patients with diabetes would be a diabetes out-patient clinic, and the data itself may be the simple answer of 'yes' to the presence of one of the definitions of heart disease previously outlined.

GLOSSARY

Categorical data
Data that fits into exact, unambiguous units with no data points between. Examples include gender (where effectively everybody is either male or female) and marital status (everybody is either single or married). In both cases there is no third option, no in-betweens.

Chi-squared (χ^2) test
Used to analyse the frequency/proportion of categorical data in different groups.

Continuous data
Numbers that flow freely from zero to very, very large, with an almost infinite number of steps between each data point (e.g., age, weight, height, virtually all plasma molecules).

Correlation and correlation coefficient
A method for examining possible relationships between two sets of continuous data from the same groups of subjects (e.g., height/weight/age). It follows that you cannot correlate, for example, gender and smoking or gender and height. Correlation does not imply causation. It derives a correlation coefficient (*r*, or Rho) where 0.9 is excellent, and implies a strong and significant association, and 0.1 is poor, implying a weak, non-significant association.

Hypothesis
A research question reworded to be a statement. Generally, this statement should be quantified (e.g., patients with a certain disease have 20% more substance 'x' in their blood than another group). However, some hypotheses are not easily quantified.

Interquartile range (IQR)
Defines the spread or variance of data that are non-normally distributed, and so is presented with a median. Thus, in a data set of median (IQR) 120 (100–180), 25% of the data points are less than 100, and the next 25% are between 100 and 120. The third quartile is between 120 and 180 and the final 25% are above 180. Thus, the first to third quartiles will include half of the total number of observations.

Mann-Whitney U test
This tests differences between sets of data where at least one is non-normally distributed. It is expressed in terms of median and inter-quartile range (IQR).

Mean
The 'average' of a set of data – obtained by adding up all the data points then dividing the sum by the number of individual data points.

Median
The 'middle' of a set of data – obtained by ranking all the data from lowest to highest.

Non-normal distribution
Data are skewed so that the highest frequency of observations (the median) is not at the central point of the data set but is to one side (generally on the left) of the entire set. The mean and median are far apart, and the standard deviation is large in relation to the mean.

Normal distribution
A bell-shaped curve with the mean generally in the centre and with two equal 'shoulders' on each side. The mean and median are close together, and the standard deviation is small in proportion to the mean.

Collection of data from those free of diabetes may be more problematic, but we would certainly seek to ensure that ultimately there was no significant difference in the mean ages of the two groups or the proportion of the two sexes.

We can perform an analysis once data collection is complete, and the most appropriate test is the **Chi-squared (χ^2) test**. Of the 100 patients with diabetes, suppose that 26 have heart disease, while in the 100 control subjects, heart disease is present in only 12. Effectively, therefore, we compare 26% with 12%. The χ^2 test gives us a probability that the difference is real at $P=0.012$ (i.e., the probability that the results are down to chance is just 1.2%, making it 98.8% likely that the results are real, and are due to some genuine difference). The result of $P=0.012$ is less than the cut-off point of $P=0.05$ so our results support the hypothesis that diabetes is associated with excess heart disease – an established finding in human pathology.

Additional methods of analysis

While generally not part of a biomedical scientist's day-to-day work, data collection, presentation and analysis will be a familiar series of processes. Table 3 summarises the major components of these aspects of good analytical practice. The second part of this article will describe additional methods of analysis that are required by more complex data sets, and will conclude with recommendations for authors who are considering submitting their data to this journal.

Brian Nation is Editor of the British Journal of Biomedical Science. *Andrew Blann is a member of the Editorial Board and a Fellow of the Royal Statistics Society.*

Further reading

- Altman DG. *Practical statistics for medical research*. London: Chapman & Hall, 1991.
- Daly F, Hand DJ, Jones MC, Lunn AD, McConway KJ. *Elements of statistics*. The Open University/Addison Wesley, 1995.
- Holmes D, Moody P, Dine D. *Research methods for the biosciences*. Oxford: Oxford University Press, 2006.
- Petrie A, Sabin C. *Medical statistics at a glance* 2nd edn. Oxford: Blackwell, 2004.
- Swinscrow TDV. *Statistics at square one* 9th edn (revised by MY Campbell). London: BMJ Books, 1996.

Paired t-test
Used to search for a difference between two data sets from the same individual that are linked, perhaps in time (e.g., blood pressure before and after an intervention) or physically (e.g., the circumference of the left versus the right calf muscle, or level of substance 'x' in serum compared to plasma). However, it is important that the difference has a normal distribution.

Pearson's correlation method
Used when attempting to correlate two sets of data that have a normal distribution.

Power calculation
Method of determining how many patients or persons to recruit, or observations to make, in order to be sure that the difference found (if present) will give a reliable outcome.

Probability (P)
This defines whether or not a different is due to a real effect (e.g., of pathology) or is simply due to chance or coincidence. We take a probability of greater than 19 in 20 (i.e., $P>0.95$) to be sufficient evidence that the effect is genuine, and we accept a rate of less than 1 in 20 (i.e., $P<0.05$) that the difference is spurious. It follows that we are convinced that an effect is spurious if the probability is 1 in 10 (i.e., $P=0.1$) but will be convinced if the chance of the difference is 1 in 30 (i.e., $P=0.033$).

Qualitative
Refers to information expressed in words, letters, expressions etc.

Quantitative
Refers to information expressed in numbers.

Reference range
A set of data from a defined group of subjects (often healthy, and can be referred to as controls) against which a data point from an individual or a second group of individuals is compared.

Spearman's correlation method
Used when attempting to correlate two sets of data where one or both have a non-normal distribution.

Standard deviation (SD)
A measure of the spread (variance) of a data set. Put simply, it represents the 'shape' of a set of data that is normally distributed. So, if a data set has a mean of 100 and an SD of 5, is it reasonably 'sharp'. But if the data set has a mean of 100 and an SD of 20, it is much more 'rounded'. Generally, mean±2SD should define 95% of a population.

Student's t-test
Also known as *t*-test. It is used to compare two sets of data whose distribution is normal. Both the mean and standard deviation are required.

Variance
The extent to which continuous data are clustered close to the central point or are more spread out. The measure of variance is the standard deviation when the data are distributed normally. Variance is expressed as the inter-quartile range when the data are non-normally distributed.

Wilcoxon's test
Like the paired *t*-test, this is used to search for a difference between two data sets from the same individual that are linked. However, it is important that the difference between the two sets has a non-normal distribution.