Good analytical practice: statistics and handling data in biomedical science: A primer and directions for authors. Part 2: Analysis of data from three or more groups, and instructions for authors

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

Biomedical scientists are bombarded daily by information, almost all of which refers to the health status of an individual or groups of individuals. This article is the second of a two-part review written to explain some of the issues related to presentation and analysis of data. In the first part (*Br J Biomed Sci* 2008; **65**: 209–17) we focused on types of data, and how to analyse and present the data from an individual or from two groups of persons. Here, we will continue with an examination of 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 submitting their data to this journal.

KEY WORDS: Analysis of variance.
Correlation of data.
Kruskall-Wallis test.
Statistical analysis.
Statistics.

Introduction

The first part of this two-part article provided a primer on basic data presentation and analysis. We first explained the difference between the two major types of information – that which is numerical (i.e., quantitative) and that which is described in words (i.e., qualitative). We then focused on quantitative data, which can in turn be subclassified into that which is continuous (e.g., age, haemoglobin, serum potassium) and that which is categorical (e.g., the numbers of men and women in a particular group, or those with, or free of, heart disease). Presentation and analysis of these two types of data are completely different.

Data with a continuous variation can be described in terms of the central point and the variance. The mean is the central point when all the data points (e.g., 15, 7, 10, 5 and 25) are added together and the sum (i.e., 62) is divided by the number of data points (i.e., 5) to give the 'average' of 12.4. The median value is the data point in the middle when they are all ranked from lowest to highest (i.e., in this case, 10). Variance gives an idea of the scatter of the data points around the central point. When we use the mean then the variance is described in terms of the standard deviation (SD). When we use the median then the variance is described in terms of the interquartile range (IQR). Thus, for the data set described above, the mean and SD are 12.4 (7.99), while the median and IQR are 10 (6–20).

We bring the concepts of the central point and variance together to describe the distribution of a set of data – which can be normal or non-normal. When the data set is normally distributed we describe it in terms of mean (SD), but if it is non-normally distributed then we use median (IQR). This is important because distribution governs analysis. If two data sets have a normal distribution then any difference is sought using Student's *t*-test. However, if one or both sets of data have a non-normal distribution then it is analysed by the Mann-Whitney U test.

Differences between data sets which are categorical should be sought using a test such as the χ^2 (pronounced chi-squared) test. If possible, research questions should be formed in terms of an original hypothesis that, if possible, should be quantified and 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 then 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 by using 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 the probability of chance to be less than 5% (i.e., $P<0.05$). It follows that we require the probability to exceed 95% for us to be confident that the difference is genuine. This general rule of probability applies for any number of sets of data. However, the precise method by which the *P* value is arrived at differs considerably when more than two sets of data are analysed.

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

Analysis of three or more sets of data

Data that are continuously variable

Analysis becomes increasingly complicated once we move away from two groups into three or more groups. The biggest difference is that the simple *t*-test and the Mann-Whitney U test are appropriate only in the analysis of two groups. We must use a different test when we have three or more groups to analyse.

Analysis of variance (commonly abbreviated to **ANOVA**) is the correct test to use when all the groups have a normal distribution. However, we have to use the **Kruskall-Wallis test** if the distribution of even one of these multiple groups has a non-normal distribution.

Figure 1 illustrates these points with three sets of data: A, B and C. A cursory glance at the spread of these data gives the impression that they are normally distributed, with most of the individual data points clustered around the central section of each of the data sets. However, it is imperative that the distribution of each set of data be determined formally, generally using a statistical software package. Once we are assured that each data set has a normal distribution, we must use the ANOVA test to look for differences in these data.

In the example provided, ANOVA tells us that there is indeed a significant difference between the three sets, at a level of $P=0.001$. What this means in practice is that the probability of this difference being due to chance is only 0.1%. So, put the other way, this means it is highly likely (to 99.9%) that there is a real difference between some of the three individual data sets. However, ANOVA does not tell us exactly where the difference is to be found. Is it between A and B, or between A and C, or is it between B and C? There may also be a difference between all three sets. What we cannot do is a series of different *t*-tests between each pair in turn (i.e., A versus B, then A versus C, and then B versus C). This may well lead to error. In order to discover where the difference is, we must use a second test – a **post-hoc test**. There are several to choose from, but Tukey's post-hoc test is one of the most popular (another is Dunnett's test). Using this test, we find that there is a significant difference between A and B, between B and C, and also between A and C, each with a probability of $P<0.05$.

Table 1. Analysis of three sets of categorical data.

	Number of obese people in three different locations			
	A	B	C	Total
Obese people	15	28	10	53
Non-obese people	72	69	84	222
Total	87	97	94	278

If we simply look at the percentages of the obese people in each of the three groups (A, B and C), we find 17%, 29% and 11%, respectively. The χ^2 test tells us that there is a high probability (i.e., 99.5%) that this difference is statically significant (i.e., $P=0.005$).

Figure 2 shows three other data sets (X, Y and Z), each of which has a non-normal distribution. This can be noted by a simple examination of the spread of these data, which are clearly distributed in a different manner to groups A, B and C in Figure 1. In each of the sets X, Y and Z, most of the data points are clustered not in the central part of the entire set (as they are in A, B and C), but are skewed over to the far left of each group, suggesting that each set has a non-normal distribution, which indeed is correct. Just as it is inappropriate to use a series of individual *t*-tests when three groups of data are normally distributed (as in A, B and C), we cannot simply perform a series of individual Mann-Whitney U tests between X and Y, between X and Z and then finally between Y and Z. Instead, the correct test to use is the Kruskall-Wallis test, which in this case gives a very high probability (i.e., $P=0.001$) that there is a genuine difference between some of the groups. However, like the ANOVA test, the Kruskall-Wallis test does not tell us exactly which pairs of data are significantly different from each other. To achieve this we must use an additional test (e.g., Tukey's test).

The problem is that Tukey's test is designed to analyse data that have a normal distribution, and our data sets (X, Y and Z) all have a non-normal distribution. One way of getting around this is to convert data that are non-normally distributed into data that are normally distributed. This can be achieved by transforming the data into a logarithmic form that generally (but not always) gives them a normal distribution. This is called **log transformation**. So, once the data sets X, Y and Z have been log transformed, and have effectively become normal, ANOVA followed by the Tukey test can be performed. In this case, we find a statistically significant difference between X and Z and between Y and Z

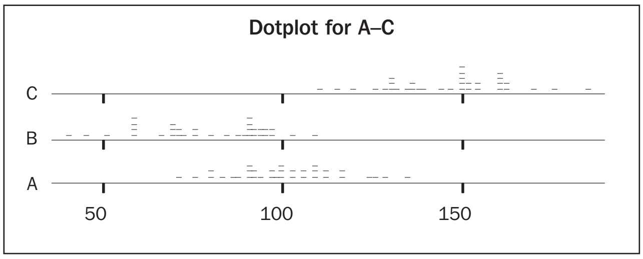

Fig. 1. Analysis of three groups of data, each of which has a normal distribution. Data set A 'seems' to have a central point of about 100 units. In B, the central point is not so easy to define, but looks to be perhaps 75. The central point of set C is about 150. In each case, there is approximately an equal number of data points above and below the central point, so we can say with a modest degree of confidence that the data are normally distributed, and that each central point is the mean value.

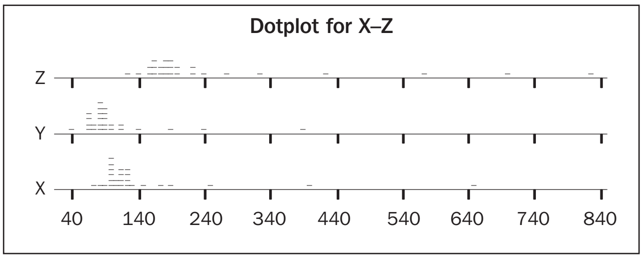

Fig. 2. Analysis of three groups of data, each of which has a non-normal distribution. In each of these data sets, most of the data points are clustered around the 100–200 units region. However, in each data set, there are several data points of considerably greater value. As the bulk of data is clustered to the left, then each set would seem to be non-normally distributed. The median values would seem to be perhaps 180, 80 and 110 for sets X, Y and Z, respectively.

Table 2. Analysis of three or more sets of linked data. Consider these data from 10 people: the platelet count (units: 10^9/L) has been measured before and on three weekly occasions after the introduction of a new drug.

Subject	Before drug use	After one week on the drug	After two weeks on the drug	After three weeks on the drug
1	299	278	284	256
2	314	325	311	289
3	287	281	294	271
4	379	376	366	356
5	269	252	261	241
6	312	311	335	326
7	325	329	356	313
8	378	311	278	251
9	267	275	261	258
10	412	387	404	356
Mean	324	312	315	292
SD	50	44	49	43

Although the mean platelet count falls from 324 to 312 in just a week, is this a significant drop? Note that levels then rise marginally to a mean of 315 after two weeks, then fall to a mean of 292 after three weeks. Overall, there is a significant change in the mean values with a likelihood of P=0.002. However, note also the variation in results from different subjects.

The average platelet counts over the whole of the experiment in subjects 5 and 9 both seem to be considerable less (256 and 265, respectively) than the averages for subjects 4 and 10 (369 and 390, respectively). Indeed, the probability that the difference between subjects is statistically significant is P=0.001

Thus, it may be that any difference in the overall result may be due to changes in only some of the subjects, which may not be truly representative of the entire data set. The complicated and powerful mathematics of the two-way ANOVA adjusts for any possible differences in particular groups.

Technically, from the research perspective, we would also obtain platelet counts at roughly the same time points from 10 patients who have not changed their mediation.

(both with a probability of $P<0.05$), but the difference between X and Y is not significant.

Once more, such analyses are virtually impossible to perform without a good statistical software package.

Correlation of multiple groups

The ability to correlate three of more continuously variable sets of data at the same time is extremely complex and is very rarely required in biomedical science. Consequently, it is beyond the scope of this article.

Analysis of three or more groups of categorical data

In analysing categorical data, generally we use the χ^2 test. Almost exactly the same analyses are performed when we look at three or more groups as for two groups. An example of this could be the frequency of obesity (defined as body mass index greater than 30 kg/m^2) among the inhabitants of three different suburbs of the same town: A place, B place and C place. Suppose that of 87 people in A place, 15 are obese (i.e., about 17%), in B place, 28 of 97 are obese (i.e., about 29%), and in C place, 10 of 94 (about 11%) are obese. Are these proportions different? These data are summarised in Table 1.

Table 3. Examples of the need for statistical software.

● Determination of the nature of the distribution of continuously variable data (i.e., normal or non-normal).
● Determination of the number of subjects to be recruited in order to test an hypothesis adequately (i.e., a power calculation).
● Analysis of continuously variable data (e.g., using Student's *t*-test or the Mann-Whitney U test for two groups; ANOVA or the Krukall-Wallis test for three or more groups).
● Analysis of categorical data (e.g., the χ^2 test).
● Generation of a correlation coefficient and a graph showing the relationship between two indices.

If we apply the χ^2 test to these data we get P=0.005. This tells us that there is a high probability (i.e., 99.5%) that there is a meaningful difference, and only a 0.5% chance that the differences are spurious. So it appears that one or more of the groups have a different rate of obesity than one or both of the other groups. But where is the difference? As for the ANOVA and Kruskall-Wallis test, we cannot tell exactly where this difference lies. In order to do this we must break up the data sets and perform individual χ^2 tests on each pair of data points. Two-group χ^2 testing was explained in the first part of the review (*Br J Biomed Sci* 2008; **65**: 209–17).

First, we would compare the 15 obese and 72 non-obese (17% obese) in A place with the 28 obese and 68 non-obese (29% obese) in B place in their own χ^2 test. Although this difference (29% versus 17%) seems large, we actually get P=0.063, which is therefore (surprisingly) not significant. In practice, this means there is a 6.3% chance that this difference is spurious, and only a 93.7% chance that it is genuine. In order to be sure that the data are reliable, we demand a greater than 95% chance (probability) that the difference is genuine, and therefore a less then 5% chance (probability) that such a difference is spurious (i.e., $P<0.05$).

Similarly, comparing the 15 obese/72 non-obese (17%) in A place with 10 obese/84 non-obese (11%) in C place in a second χ^2 test gives P=0.198 (i.e., there is a 19.8% chance that the difference is spurious, and only an 80.2% chance it is genuine). Finally, therefore, we compare the 28 obese/69 non-obese (29%) in B place with the 10 obese/84 non-obese (11%) in C place. A χ^2 test of these data gives P=0.002 (i.e., a 99.8% probability that the difference is genuine). This seems acceptable, given the crude rates of obesity of 11% versus 29%.

Analysis of three or more sets of linked data

We have already seen (*Br J Biomed Sci* 2008; **65**: 209–17) how to analyse paired samples of plasma taken at two different time points or linked together in some other way. Similar analyses are possible for the assessment of data obtained at three or more time points or linked by three or more factors. Consider the development of a new drug designed to reduce the numbers of platelets in the blood. It is important to know how long the drug takes to have an impact – would it be days, weeks or months? Such an experiment would have to obtain blood samples on different occasions, have them analysed by the platelet count, and then submit the data to **two-way analysis of variance**, although this may also be known as repeated-measures analysis of variance. It is described as two-way ANOVA because we need to be able to factor in the possible effects of differences between patients, and also between the different stages of the experiment. This is described in Table 2.

The platelet count falls steadily from a mean (SD) of 324 (50) at baseline, to 312 (44), then to 315 (49) and finally to 292 (43). It is tempting to consider that the fall from 324 to 312 is significant. However, it seems unlikely that the change from 312 to 315 is meaningful. However, close observation of the data set reveals that all the results from patient 9 (ranging from 275 to 258) are considerably smaller than all the results from patient 10 (ranging from 412 to 356). Hence, one might argue that these two individual sets of data may influence the entire data set unduly, which indeed may happen. However, the whole point of two-way analysis (indeed, its strength) is that two-way ANOVA will compensate for differences between and within the groups.

Data need not be linked by time in order for them to be analysed by this method. For example, we would also use a two-way ANOVA to analyse levels of a component of the blood from the same person that has been measured in serum, in plasma anticoagulated with sodium citrate, and also in plasma anticoagulated with EDTA.

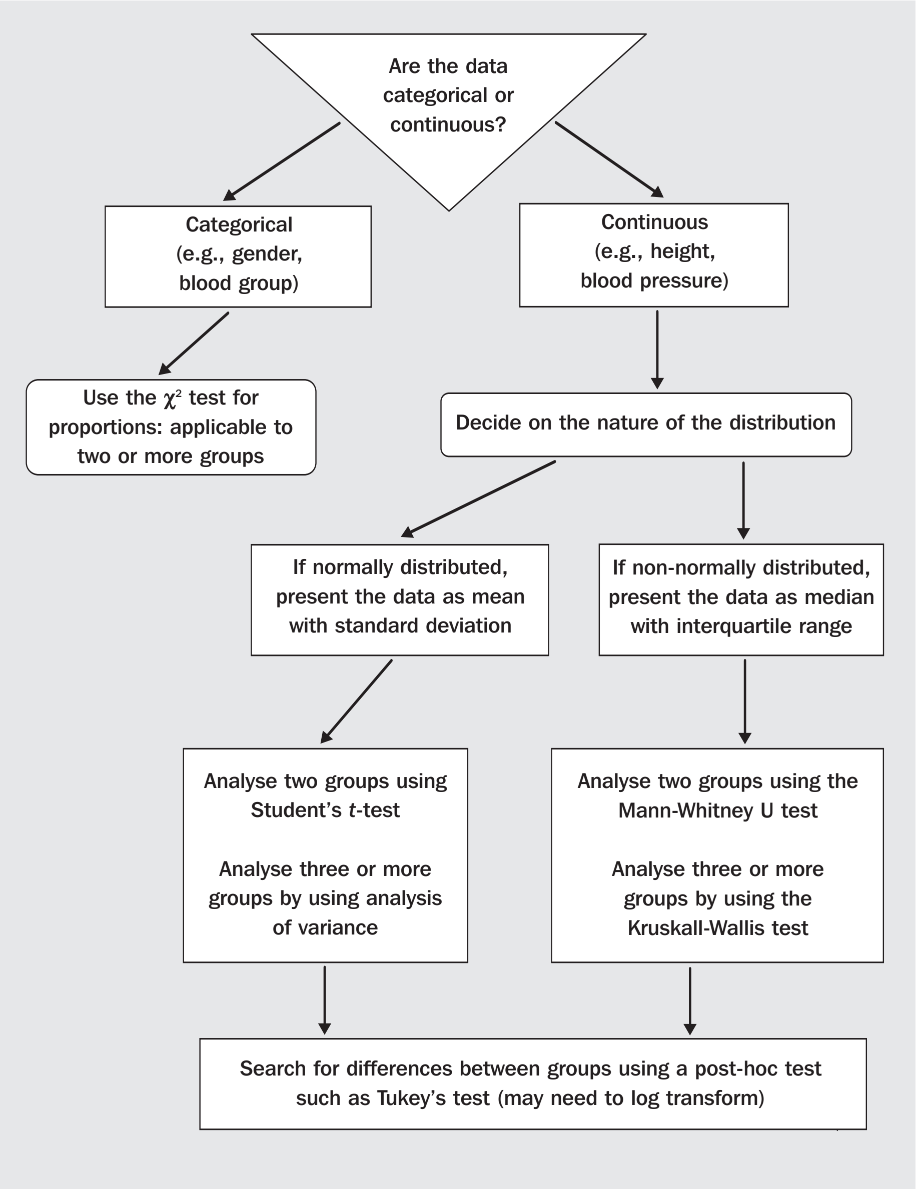
Are the data categorical or continuous?
Categorical (e.g., gender, blood group)
Continuous (e.g., height, blood pressure)
Use the χ^2 test for proportions: applicable to two or more groups
Decide on the nature of the distribution
If normally distributed, present the data as mean with standard deviation
If non-normally distributed, present the data as median with interquartile range
Analyse two groups using Student's *t*-test
Analyse three or more groups by using analysis of variance
Analyse two groups using the Mann-Whitney U test
Analyse three or more groups by using the Kruskall-Wallis test
Search for differences between groups using a post-hoc test such as Tukey's test (may need to log transform)

Fig. 3. Simple analysis of quantitative data.

Figure 3 provides a summary and flow path for some of the different methods of analysis, based on the nature of the data, and on how to present the data.

Statistical packages

Efficient handling and analysis of data are impossible without a good statistical package. Such packages can be purchased directly from the manufacturer, although in practice many large organisations such as hospitals and universities will have a site licence, allowing an individual almost immediate access. Indeed, higher and further education institutions will almost certainly have a Department of Mathematics and/or Statistics, from where packages, and advice, can be sourced.

Although claiming to be user-friendly, these packages are very complicated and are certainly not for the untutored. It is essential that those fresh to the subject complete a formal training programme, generally offered by their host organisation (i.e., hospital, university, pharmaceutical company, etc.), although these can also be offered by commercial training organisations.

However, even the most sophisticated statistical package is unable to counter the biggest problem in data management – the reliability of the data that are entered. Even the most powerful computers and statistical packages are unable to correct data that are either derived incorrectly at source or are entered incorrectly by an operator – hence the concept of 'garbage in, garbage out'. This is why the concepts of quality control and quality assurance are so important.

The best known and most widely used statistical packages are (in no particular order) Minitab, SPSS, SAS and Stata. All offer a range of tests, from the simple (e.g., Student's *t*-test, χ^2 test) to the most complex, but some are more user-friendly to the occasional user, while others are designed for the full-time statistician. These packages often come with software to generate plots and figures. Advice generally can be obtained from experienced users, themselves often in formal statistical units. □

GLOSSARY

Analysis of variance (ANOVA)
A test to determine the probability of a difference between three or more sets of normally distributed data.

Kruskall-Wallis test
A test to determine the probability of a difference between three or more sets of non-normally distributed data.

Log transformation
A method used to convert data that are non-normally distributed into a set that is normally distributed.

Post-hoc test
A test to determine any difference between three or more sets of data.

Two-way ANOVA
A test to determine any difference between linked samples (e.g., a series of timed estimations) in a number of individuals, which takes into account any possible differences that may be present in those individuals.

PLEASE NOTE

Part 1 of this article includes a glossary to explain the following terms:
Categorical data, χ^2 test, Continuous data, Correlation and correlation coefficient, Interquartile range (IQR), Hypothesis, Mann-Whitney U test, Mean, Median, Non-normal distribution, Normal distribution, Paired *t*-test, Pearson's correlation method, Power calculation, Probability (*P*), Qualitative, Quantitative, Reference range, *t*-test, Spearman's correlation method, Standard deviation (SD), Variance, Wilcoxon's test.

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.
- Blann AD, Nation B. 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. *Br J Biomed Sci* 2008: **65**; 209–17.
- 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.

INSTRUCTIONS TO AUTHORS

Statistical advice

1 The **Abstract** to an article with original data (i.e., not a Review Article or Biomedical Science in Brief item) generally will include some data but should always include relevant *P* values.

2 Almost all scientific research will test a defined and quantitative hypothesis. An aim, investigation or an objective is of insufficient rigour. The hypothesis is generally stated at the end of the **Introduction**. However, it is recognised that not all research will test a defined hypothesis.

3 The **Materials and methods** section will conclude with a paragraph on statistical methods. This paragraph is likely to include:

- If applicable, reference to the method(s) used to ensure that the sample size is large enough (i.e., a power calculation)
- A statement of the method used to determine the distribution of any data that are continuously distributed
- A declaration of the method for describing such data (i.e., mean [SD] or median [IQR])
- A statement of the method for analysing continuously distributed data
- A statement of the method for analysing categorical data
- The name of the statistical package used for the analysis.

Despite the above, the Editor accepts that not all of these steps will be applicable to all research articles submitted to the *British Journal of Biomedical Science*.