Statistics on microcomputers

A non-algebraic guide to their appropriate use in biomedical research and pathology practice

3 Analysis of variance and distribution-free methods

R A BROWN, J SWANSON BECK*

*From the Departments of Mathematical Sciences and *Pathology, University of Dundee*

Comparison of several groups where distributions approximate to normal and where variances are similar

It is unusual for a biomedical research study to involve only comparison between two groups; more often the same observation is made on each individual in several groups with the aims of (i) detecting group differences and (ii) estimating their probable magnitudes. An example of this type of study is a comparison between haemoglobin concentrations in normal men (group A), normal women not receiving medication (group B), normal women taking the contraceptive pill (group C), and women with menometrorrhagia (group D) (fig 1).

A common but inappropriate method of analysis of such data entails the comparison of each pair of groups using the two sample *t* test. Thus group A is compared in turn with groups B, C and D; group B is compared with groups C and D; and finally group C is compared with D, resulting in a total of six *t* tests. This approach is basically flawed as it ignores the inherent associations in multiple comparisons and will inflate the chance of obtaining a misleading result. Remember that in the *t* test for comparison of two groups of data, the *p* value is the probability of obtaining a result at least as extreme as the observed difference between the sample means when there is no difference between the corresponding population means. Consequently, when applying a single *t* test, the investigator has a 5% chance of being misled into concluding that there is evidence of a difference when the population means

are equal and the observed extreme result is attributable entirely to the random variability of the responses. If two independent *t* tests are performed—for example, comparing A with B and C with D—in a situation where the corresponding population means are equal, the probability that both will correctly indicate this is 0.95×0.95 or 0.9025 (90.25%), and so there is a 9.75% chance that at least one of the pair of tests will be misleading. If six comparisons were made (as in the haemoglobin study) and the corresponding population means were equal then the chance of a correct conclusion in all six tests is about $(0.95)^6$ or

Fig 1 Box and whisker plots of observed distribution of haemoglobin concentrations in random samples of 100 subjects from each of the following groups: (a) healthy men; (b) healthy women not on medication; (c) healthy women taking the contraceptive pill; (d) women with menometrorrhagia. All distributions seem to be symmetrical and variability is about the same in all groups.

0.735. In other words there is a greater than 1 in 4 chance that the result of at least one of the six tests will be misleading.

The probability of at least one misleading conclusion in a series of tests is known as the *error rate* of the series. The tendency for the error rate to rise rapidly as the number of groups increases means that the technique of pairwise comparisons cannot be recommended. The table indicates the approximate error rate for comparison of several groups by pairwise *t* tests when it is proposed to report all results with *p* values of less than 0.05. Being more cautious and reporting only those results with *p* values of less than 0.01 alleviates but does not entirely eliminate the problem.

The appropriate technique, *analysis of variance*,¹² not only takes full account of the multiple comparison nature of the problem but also uses the information in all samples to provide standard errors for the differences between pairs of population means (the pairwise *t* test method uses only information from the pair of samples being compared to determine the standard error of the difference).

The method is available in both Statgraphics and Minitab. It is based on the following assumptions: (i) the variability within each group is about the same; (ii) within each population, observed values are more or less normally distributed.

The basic principle is quite simple. If there are real differences between some of the population means this will be reflected as differences between the corresponding sample means which will be significant if the study is large enough. In any investigation involving several groups the overall variability of the responses can be separated into two components: (i) variability between sample means, indicating possible differences between population means; and (ii) variability of responses within each group, indicating natural variability within each population. If the population means are equal then differences between the sample means will be entirely due to chance variations in the individual observations within each group so that the two types of variability (i) and (ii) ought to be comparable. If there are substantial differences between the population means, however, variability between sample

Table Overall error rates when several groups are compared by performing two sample *t* tests on all pairs of groups

No of groups	No of pairs	Overall error rate when the chosen significance level of each test is:	
		5%	1%
3	3	14%	3%
4	6	26%	6%
5	10	40%	10%
6	15	54%	14%
10	45	90%	36%

means (i) ought to be considerably greater than variability within groups (ii).

The analysis of variance differentiates between these situations by making a direct comparison between (i) and (ii). For technical reasons the measure of spread used is not the standard deviation but its square, the variance. The method uses (a) the variance of the set of sample means and (b) an estimate of the natural variability based on the variances of the responses in each group, known as the "pooled" within groups variance, to calculate the variance ratio (or *F*-ratio) which is merely (a) divided by (b).

The evidence for or against differences between the population means of the various groups is assessed by determining the probability that a variance ratio at least as extreme as that actually observed could occur by chance if all the population means were equal—that is, the *p* value. A variance ratio which has a low *p* value is taken as evidence that real differences exist: by itself it does not indicate which sets are different from the remaining sets. The output from both packages is much the same and takes the form of a standard analysis of variance table (fig 2), which is part of the Statgraphics analysis of the data displayed in fig 1.

The analysis of variance is basically a significance testing procedure, so an *F*-ratio which has a *p* value of less than 0.05 is conventionally taken as evidence of genuine differences between some of the population means. As usual the smaller the *p* value the stronger is the evidence for differences. Thus in the haemoglobin example there is very strong evidence of differences between mean haemoglobin concentration in some of the groups. Note that the analysis of variance, like all significance testing procedures, merely gives an indica-

Source of variation	Sum of squares	d.f.	Mean square	F-ratio	Sig level
Between groups	865.4021	3	288.4674	43.8941	0.0000
Within groups	2602.4706	396	6.5719		
Total (corrected)	3467.8727	399			

Fig 2 Analysis of variance table produced by Statgraphics for the data summarised in fig 1. *F*-ratio is highly significant (the value 0.0000 displayed under "Sig. level" means that the *p* value is less than 0.00005), indicating that the data provide extremely strong evidence of differences between the population means of some groups. It is not possible to tell from this table which groups differ.

Fig 3 Analysis of variance table provided by Minitab also includes a graphical follow up analysis. 95% confidence intervals for the population means of the four groups are interpreted in the usual way: intervals which do not overlap indicate a difference between the corresponding population means.

tion of the strength of the evidence; it does not indicate which population means do differ or whether the differences are of practical, as distinct from statistical, significance.

FOLLOW UP ANALYSIS

To answer such questions it is essential to report the results of a "follow up" analysis. This can be done graphically by presenting appropriate confidence intervals for the population means. There are various methods of doing this. Minitab displays individual 95% confidence intervals for the means using essentially the same method as used for a two sample investigation, but with the important proviso that the standard error for the difference between any pair of means is derived from information on variability obtained from *all* samples and not merely from the two samples being compared. The analysis of variance and follow up produced by Minitab are displayed in fig 3. This shows that (i) healthy men have a higher mean haemoglobin concentration than women; (ii) women taking the contraceptive pill have a mean haemoglobin concentration which cannot be distinguished from that of healthy women not receiving medication; and (iii) women with menometrorrhagia have a mean haemoglobin concentration which is significantly lower than that of healthy women.

The Statgraphics package produces a similar follow up analysis on request but tabulates the confidence limits rather than displaying them graphically. An extra follow up feature of Statgraphics is its ability to indicate those sample means for which there is no evidence of a difference between the corresponding population means (fig 4). In the section of the table headed "homogeneous groups" asterisks in the same column alongside the means for groups B and C indicate that they do not differ significantly. The technique used for these calculations is known as a *multiple comparisons procedure*³ and several alternative methods are available within Statgraphics, differing mainly in the details of the calculations

required to produce the final result. Which one is used is largely immaterial as all will give very similar results, what is important is that some form of follow up analysis should always be performed after obtaining a significant F-ratio in the initial analysis of variance.

DIAGNOSTIC PLOTS

The analysis of variance is based on assumptions of homogeneity of spread of responses in the various groups and normality of the distribution of responses. Failure of the responses to conform to these assumptions either directly or via the presence of outliers may result in the analysis of variance being misleading. The assumptions may be checked by examining the *residuals*, which are the deviations of the responses from the corresponding group means. The most useful plots are (i) a plot of residuals against the group means and (ii) a normal plot (article 2) of the residuals. Residuals which are particularly large may indicate outliers. Repeating the analysis with suspect observations omitted will indicate whether they play a large part in determining the conclusions (article 1).

The plot of residuals against the group means (an additional option is Statgraphics) is very similar to the plot of the raw data which should be the starting point of the analysis of variance except that each group of residuals is centred on zero and so visual comparison is easier. The danger sign to look out for is either one group of residuals with a much greater spread than the rest, or a tendency for the spread to vary appreciably and systematically over the groups. If either pattern

Multiple range analysis for haemoglobin by group			
Method group	95% Scheffe count	Average	Homogenous groups
A	100	15.659	
C	100	14.156	*
B	100	13.943	*
D	100	11.552	

Fig 4 Statgraphics provides an alternative follow up analysis based on multiple comparison methods.

Statistics on microcomputers

appears then the assumption of uniform spread is not tenable and some corrective action such as a transformation of the data may be needed. The normal plot of the residuals should be nearly a straight line if the data conform to the normality assumption, but curvature can indicate non-normality, inhomogeneity of spread, or the presence of outliers, it is therefore wise to look at both plots. Both packages calculate residuals and fitted values on demand and so the two diagnostic plots can easily be obtained.

"Distribution-free" methods

It has been pointed out that the various statistical techniques described in this and previous articles are based on assumptions concerning the distribution of responses. In particular, outliers or extreme non-normality may seriously influence the results of any analysis which relies on the normality assumption. Normal plots of the raw data in each sample or of the residuals can be used to assess how well the data conform, provided the sample sizes are sufficiently large (article 2). This method is insensitive when small samples are studied, however, when only the most extreme forms of non-normality will be detected. Fortunately, both the methods based on the *t* test and the analysis of variance are pretty robust as far as moderate departures from normality are concerned.

Some types of non-normality may be corrected by an appropriate transformation such as the logarithmic or the square root transformation. If a transformation to correct non-normality cannot be found then an alternative approach is available. A significance test and the appropriate confidence interval may be determined by using distribution-free (or non-parametric) statistical methods.⁴⁵ These methods do not assume that the population distributions are "normal" but assume instead that the distribution shape is about the same in each population being compared. The widespread belief that non-parametric methods can be applied to all sets of data is not valid.

DISTRIBUTION-FREE EQUIVALENT OF THE T TEST

The median is a more appropriate indicator of location than the mean for appreciably asymmetrical non-normal distributions so the two sample distribution-free significance test assesses the evidence for a difference between the population medians: the distribution-free confidence interval represents a range of plausible values for this difference.

The relevant method for two sample comparisons is known as the *Mann-Whitney procedure*. It is based on the principle of replacing the actual data values by their ranks. First, the two samples are combined and sorted into ascending order. Then the observations are

ranked giving rank 1 to the smallest, rank 2 to the next smallest, and so on. If two or more observations are equal then the average of the ranks which would be assigned is allocated to each one.

The test statistic used to assess evidence for a difference between the population medians is the sum of the ranks in the first sample and is often designated by the letter *W*. The value of *W* reflects the relative position of the two sets of sample values; if the values in the first sample tend to be higher than those in the second sample then the larger ranks will occur predominantly in the first sample and *W* will be large. Obviously *W* will be small if the values in the first sample tend to be smaller than those in the second sample. For example, box plots of the survivals of two groups of patients with breast cancer are shown in fig 5. Group A comprises patients with 40% or more of malignant cells in sections of neoplastic tissue and group B patients with less than 40% of malignant cells. The plots suggest that survival may be related to percentage of malignant cells in the sense that survival seems to be somewhat longer in group B. The output from Minitab consisting of a basic description of the data, the Mann-Whitney test, and a confidence interval for the difference between the population medians is displayed in fig 6. Although the median of group B is larger than the median of group A, the probability of obtaining a difference in medians at least as large as this when the population medians are equal is 0.25 and consequently the difference is not

Fig 5 Box and whisker plots of survivals of breast cancer patients in a retrospective study of the use of morphometric features as predictors of survival. The distributions are extremely asymmetrical and so median survival is more informative than mean survival. This degree of asymmetry rules out the use of methods based on the normal distribution.

MTB > describe "Group A" "Group B"								
	<i>N</i>	<i>Mean</i>	<i>Median</i>	<i>STDEV</i>	<i>Min</i>	<i>Max</i>	<i>Q1</i>	<i>Q3</i>
Group A	30	39.03	30.00	35.81	3.00	142.00	10.75	64.75
Group B	32	64.70	41.00	59.40	0.00	191.00	9.20	121.20
MTB > Mann-Whitney "Group A" "Group B"								
Mann-Whitney confidence interval and test								
Group A	N = 30	Median = 30.000						
Group B	N = 32	Median = 41.000						
Point estimate for ETA1-ETA2 is - 11.9986								
95.1 PCT C.I. for ETA1-ETA2 is (- 50, 5)								
W = 862.0								
Test of ETA1 = ETA2 vs ETA1 N.E. ETA2 is significant at 0.2452								
Cannot reject at alpha = 0.05								

Fig 6 Minitab display of basic descriptive statistics and the distribution-free Mann-Whitney analysis of the data summarised in fig 5. "ETA1" and "ETA2" are Minitab jargon for the population medians for the two groups. The 95% confidence interval for the true difference between the population medians includes zero, indicating that there is insufficient evidence to claim a real difference between median survival. This is confirmed by the *p* value (0.2452) associated with *W*. The phrase "cannot reject at alpha = 0.05" indicates that a significance test at the conventional level of 5% fails to reject the hypothesis that the population medians are equal.

significant. The conclusion is clear: the data fail to provide statistically acceptable evidence of a difference between population median survivals for the two groups.

The distribution-free confidence interval for the difference between the population medians is based on the calculation of all differences between the survivals in the first group and those in the second group. As this procedure is not widely known an example of how the calculations are performed is given in the appendix. The observed difference in median survival is 12 months for the data in fig 5. The confidence interval calculated by Minitab stretches from less than five months to more than 51 months, which brackets zero and so is in line with the conclusion derived from the Mann-Whitney significance test. Statgraphics is not so useful in this situation: it is capable of performing the Mann-Whitney test, but does not provide the corresponding confidence interval.

A similar distribution-free technique for the *paired sample comparison* is known as the *Wilcoxon technique*. Recall that a typical paired comparison investigation involves two measurements made on each subject, and the appropriate analysis involves examining the differences between the paired observations.

First, the observed differences are ordered ignoring their signs. Ranks are then assigned in the usual way and each rank is then given the same sign as the corresponding difference. The test statistic for assessing evidence that the median of the population distribution of differences is non-zero is the sum of the signed ranks. A value close to zero indicates no real difference between the pairs of measurements, whereas a large value (positive or negative) may be evidence of genuine differences. If the *p* value associated with the

signed rank statistic is less than 0.05 then this is taken as evidence of genuine differences between the first and second response. There is a corresponding method of calculating a confidence interval for the population median difference between the paired responses. Both Minitab and Statgraphics perform the Wilcoxon signed rank test, but only Minitab provides the confidence interval. The appendix describes the method for calculation of the confidence interval for the benefit of Statgraphics users.

As the distribution-free techniques do not take account of the actual data values but merely their relative positions they are much more resistant to the effects of outliers than the corresponding techniques based on the *t* statistic. They thus provide at least a partial remedy to the thorny problem of whether to omit apparent outliers from an analysis.

There is a common misconception that distribution-free methods make no assumptions whatsoever about the data to which they may be applied and so there is a tendency to apply them in all situations, even when they are inappropriate. It is commonly stated that almost all biomedical data are non-normally distributed and so methods which have a built-in assumption of normality are of no value. This exaggerates the true situation and also ignores the fact that when data conform reasonably well with the assumption of normality, the *t* based techniques are not seriously misleading and are in fact slightly more *powerful*, in the sense of being able to detect smaller differences. Moreover, when they are appropriate, *t* based confidence intervals will be narrower than distribution-free intervals. Thus our strong recommendation is to use the appropriate technique: if there is no evidence that the data are seriously out of line with the usual

assumptions use the *t* based methods and reserve distribution-free techniques for remedial cases.

DISTRIBUTION-FREE ANALYSIS OF VARIANCE

As with the two sample situations, there is a distribution-free technique which may be used in the event of serious non-conformity with the normality assumption either through asymmetry or outliers which cannot be corrected. The method is the *Kruskal-Wallis test* and like the Mann-Whitney test it is based on ranks. The data from all groups are combined and ranked, then the mean rank is calculated for each group. The test statistic measures how far apart these mean ranks are and statistical theory indicates how large the Kruskal-Wallis statistic is likely to be when there are no real differences between the corresponding population medians. To illustrate the technique the Kruskal-Wallis method is applied to the haemoglobin data of fig 1 (even though these data are normally distributed), and the corresponding output from Statgraphics is displayed in fig 7. The mean ranks for the four groups are somewhat different and the extremely small *p* value of the Kruskal-Wallis statistic indicates that there are differences between some of the groups. Unfortunately, the utility of the Kruskal-Wallis method is somewhat reduced as neither package provides a follow up analysis and there is no easy way of obtaining distribution-free confidence intervals for the group medians. One way round this is to apply either the signed rank method (see appendix) to produce individual confidence intervals for each population median or to use the Mann-Whitney based method to obtain confidence intervals for differences between all pairs of group medians.

The former is marginally simpler to apply than the latter, but either will only be useful when the sample sizes are reasonably large (more than 10 in each group as a rough guide), and even in this case the procedure will tend to be rather conservative as there is nothing analogous to the residual mean square for gaining extra precision by using information from all samples in determining the standard error as in the analysis of variance. In the event of there being serious doubts about whether the standard method is applicable then the best advice is to consult your friendly local statistical guru.

The broader problem of multiple testing

In many investigations various observations are made on individuals belonging to different categories and these may then be analysed in a number of ways to suggest fruitful areas for concentration of further research effort. With a large and complex data set there are likely to be several hundred ways in which various combinations of variables may be analysed.

Kruskal-Wallis analysis of haemoglobin by group		
Group	Sample size	Average rank
A	100	270.9000
B	100	204.8000
C	100	216.3000
D	100	109.9000
Test statistic = 100.5000		Significance level = 3.702E-6

Fig 7 Statgraphics display for the Kruskal-Wallis analysis of the data summarised in fig 1. The average ranks for the four groups differ appreciably. The *p* value of the Kruskal-Wallis statistic *t* is extremely small ($3.702\text{E}-6$ means 3.702×10^{-6}) so the analysis indicates extremely strong evidence of differences between groups, which agrees with the analysis of variance and follow up analyses displayed in figs 2-4.

The problem of overall error rates and multiple comparisons is obviously important here for if every result which has a *p* value of 0.05 or less is reported as corresponding to a genuine difference then 1 in 20 (5%) of the reported results may be misleading in the sense of indicating a difference where one does not exist. Partly the problem arises from the use of a significance test as an exploratory tool in sifting through a large body of data: a function for which it is not ideally suited. The proportion of spuriously significant results among a large group of tests may be reduced by resisting the temptation to report every result with *p* value of less than 0.05, reporting only those results with *p* values less than 0.01 or some similar figure: using 0.05 divided by the square root of the number of tests performed gives a *p* value which is a reasonable compromise between spurious significance and being too stringent. Alternatively, one may adopt the viewpoint that the statistical tests performed in an exploratory situation are merely being used to generate interesting hypotheses which must be confirmed in a further investigation of at least the same size as that from which they were derived.

Appendix

THE MANN-WHITNEY CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO MEDIANS

A simple example will make the method clear. Suppose two samples of responses are (in ascending order),

group A 56, 57, 58, 61, 62, 63, 66, 69 – median 61.5
group B 42, 48, 50, 51, 53, 54, 59 – median 51.0

All possible arithmetical differences between responses in group A and group B are calculated. This is most easily done by constructing a table bordered by the ordered responses with the minimum values in the "north-west" corner, entering the corresponding differences in the body of the table as shown below.

		Group B							
		42	48	50	51	53	54	59	
Group A	56	14	8	6	5	3*	3*	-3*	
	57	15	9	7	6	4*	3*	-2*	
	58	16	10	8	7	5	4*	-1*	
	61	19*	13	11	10	8	7	2*	
	62	20*	14	12	11	9	8	3*	
	63	21*	15	13	12	10	9	4	
	66	24*	18*	16	15	13	12	7	
	69	27*	21*	19*	18*	16*	15	10	

The 95% confidence interval is formed by eliminating the m smallest and m largest values from this table, where the value of m is determined by consulting a table of percentage points of the Mann-Whitney distribution.⁶ In this example the 10 smallest and 10 largest values must be eliminated. These values are marked with asterisks in the table. The smallest of the remaining values is 4 and the largest is 16 so the 95% confidence interval for the difference between group medians stretches from 4 to 16.

The calculations are rather tedious done by hand but they can be shortened by noting that with the above type of tabulation the largest and smallest differences will always appear in the "north-east" and "south-west" corners of the table and so it is not necessary to calculate all differences. With a little ingenuity the calculations can be done with a spreadsheet program, readers with some knowledge of BASIC programming should be able to construct a small program to perform the calculations.

WILCOXON SIGNED RANK BASED INTERVAL

Suppose the differences between paired responses in each of eight individuals were:

-4, 1, 5, 7, 9, 10, 12, 16.

The confidence interval for the median difference is obtained in a somewhat similar manner to the Mann-Whitney interval. A table is constructed bordered by

the ordered differences with the minimum value in the north-west corner and the entries in the table are the averages of the corresponding row and column values, as shown below (for example, the entry in row 3 column 5 is $(5+9)/(2=7)$). It is necessary only to complete the upper half of the table to obtain all the necessary values as the lower half is merely a repetition of the upper portion. The 95% confidence interval is obtained by eliminating the m smallest and m largest averages: the value of m is obtained from a table of the percentage points of the Wilcoxon signed ranks distribution,⁶ the interval stretches from the smallest to the largest of the remaining averages. In this case the appropriate value of m is 3 and so the 95% confidence interval stretches from 1 up to 12.5.

	-4	1	5	7	9	10	12	16
-4	-4.0*	-1.5*	0.5*	1.5	2.5	3.0	4.0	6.0
1		1.0	3.0	4.0	5.0	5.5	6.5	8.5
5			5.0	6.0	7.0	7.5	8.5	10.5
7				7.0	8.0	8.5	9.5	11.5
9					9.0	9.5	10.5	12.5
10						10.0	11.0	13.0*
12							12.0	14.0*
16								16.0*

References

- 1 Armitage P. *Statistical methods in medical research*. Oxford: Blackwell, 1971.
- 2 Bailey NTJ. *Statistical methods in biology*. London: Hodder and Stoughton, 1981.
- 3 Miller RG. *Simultaneous statistical inference*. New York: Springer-Verlag, 1981.
- 4 Conover WJ. *Practical nonparametric statistics*. New York: John Wiley, 1980.
- 5 Sprent P. *Quick statistics*. London: Penguin Group, 1981.
- 6 Neave HR. *Elementary statistics tables*. London: Allen and Unwin, 1981.

Requests for reprints to: Professor J Swanson Beck, Department of Pathology, Ninewells Hospital and Medical School, PO Box 120, Dundee DD1 9SY, Scotland.