ELSEVIER

International Journal for Parasitology 32 (2002) 1065–1070

The Australian Society for Parasitology Inc.
INTERNATIONAL Journal for PARASITOLOGY

www.parasitology-online.com

Letter to the Editor

How to improve statistical analysis in parasitology research publications

Glantz (1997) reports that the most common error made during the statistical analysis of data in medical journals has been the use of *t*-tests (or their non-parametric equivalent, the Mann–Whitney *U*-test) to compare more than two groups of observations. This particular problem has appeared in many areas of biology, and my own work suggests that this is probably the biggest problem in parasitology research publications as well.

My aim here is to introduce this particular problem, to explain how to avoid it, and to discuss two simple examples (from the recent parasitological literature) of the consequences of failing to deal with this issue. Put simply, the ultimate consequence of employing any inappropriate statistical analysis is that you are likely to come to the wrong conclusion from your data analysis. The most common outcome is a claim that a particular biological pattern or process exists when there is actually no evidence for it in the experiment.

This letter is therefore an attempt to highlight one particular statistical problem and to point out the solutions. I am doing this in the hope that the high quality of much of the research in parasitology will be carried forward from the experimental phase into the analysis phase as well. The quality of any experiment depends on both the experimental protocol *and* the data analysis. There is little to be gained by collecting high-quality data and then analysing it in some inappropriate manner.

There are two inter-related problems with the use of *t*-tests; or, more precisely, there is one fundamental problem that manifests itself in one or both of two ways. The fundamental problem is how the statistical probabilities are calculated and interpreted, and it manifests itself when (i) a large number of statistical tests are performed on a particular data set, and/or (ii) when all possible comparisons of groups of observations are tested.

The *t*-test and the *U*-test are conceptually simple tests that are easy to perform. This seems to be their appeal to biologists, especially those who feel less confident with mathematical analyses. However, their application is very limited, and biologists seem to want to apply them to situations where they are inappropriate. There are more suitable alternatives for these situations, although these alternatives are somewhat more complicated. The point is: do you want to use an inappropriate but simple test or an appropriate but more complicated test? These days, with the calculations almost universally being performed by computer programs, there is little excuse for restricting your data analysis to simple tests.

Both the *t*-test and the *U*-test are designed to compare two groups of observations. They estimate statistically whether the central locations of the two samples differ from each other with respect to the characteristic that has been measured – for the *t*-test the central location is the mean of the observations while for the *U*-test it is the median. Each of the tests produces a probability, a small probability (e.g. $P < 0.05$) being interpreted to indicate that the two groups are unlikely to have the same central location (the difference between the two groups is 'statistically significant') and a large probability indicating that the locations are the same.

When calculating the statistical probability, the *t*-test and the *U*-test assume that the comparison being tested is independent of all other tests. This means that the outcome of a particular test is not affected by the outcome of any other test. If the outcome *is* affected by another test then the calculation of the probability will be wrong – the probability produced by the test will be smaller than the true value. So, when all possible comparisons of groups are tested for only two groups, the *t*-test or the *U*-test is appropriate because there is only one comparison. However, when there are more than two groups there is more than one comparison, and the multiple comparisons of the groups within any one data set will not be independent of each other. Hence, both the *t*-test and the *U*-test are inappropriate for testing all possible pairwise comparisons.

The other problem with probabilities appears when performing a large number of statistical tests. The probability quoted with statistical tests is used to accept or reject the null hypothesis being tested – for a *t*-test or *U*-test the null hypothesis states that the two central locations being tested are not different from each other. By convention, we would reject the null hypothesis at a level of $P < 0.05$. This probability tells us the likelihood of making a mistake if we reject the null hypothesis – statistically, this mistake is referred to as a Type I error. So, we cannot be absolutely certain that we are right to reject the null hypothesis, but our probability tells us something about how much confidence we can have when we reject it. Small probabilities indicate greater confidence because there is less chance that we have made a mistake. However, the consequence is that we are prepared to make a mistake 5% of the time if we use the conventional significance level. The more tests that we perform then the more inevitable it is that some of the

PII: S0020-7519(02)00064-4

time we have made a mistake. Hence, it is inappropriate to perform large numbers of *t*-tests or *U*-tests.

1. All possible pairwise comparisons

The characteristic of independence in statistical tests is dealt with by the degrees-of-freedom that are reported with all such tests. For the comparison of groups, every degree-of-freedom indicates an independent test of a statistical hypothesis – that is what the degrees-of-freedom are for, to represent the number of independent pieces of information there are in a statistical analysis. For each *t*-test and *U*-test there is thus a single degree-of-freedom for the groups. (N.B. there is another set of degrees-of-freedom associated with the number of observations within each group, but this is not relevant to the discussion here, so please do not confuse the two issues.) There is always one less degree-of-freedom than there is entities being analysed. So, this is why the *t*-test or the *U*-test is appropriate for comparing two groups: there are $k = 2$ groups and therefore $k - 1 = 1$ degrees-of-freedom. The data analysis thus matches the assumption of the statistical test.

For larger numbers of groups there are clearly more available degrees-of-freedom. For instance, if there are $k = 30$ groups being analysed then there are $k - 1 = 29$ degrees-of-freedom. It thus seems that more tests are possible under these circumstances, which is true. However, for comparisons of larger numbers of groups it is inappropriate to use the *t*-test or the *U*-test to compare all possible pairs of groups. The problem with statistically testing the experimental groups in all possible pairwise combinations is that for any number of groups greater than two there are more pairwise combinations than there are degrees-of-freedom. There are $k \times (k - 1)/2$ possible pairwise comparisons for k groups, which means that there will always be $k/2$ times as many comparisons as there are degrees-of-freedom.

For example, for three groups there are two degrees-of-freedom but three possible pairwise comparisons (i.e. group 1 vs. group 2, group 1 vs. group 3, group 2 vs. group 3). This means that one of these pairwise comparisons cannot be treated as independent of the other two. This can be made obvious by considering the following possible scenario: (1) use a *t*-test to compare the means of group 1 vs. group 2 and conclude that $1 > 2$; (2) compare group 2 vs. group 3 and conclude that $2 > 3$. It is not necessary for us to now compare group 1 vs. group 3, because $1 > 3$ is the only logical outcome for this particular scenario. The first two comparisons exhaust our two degrees-of-freedom, and the third comparison cannot be independent of these two comparisons. If this third comparison was to be made using a *t*-test, then the assumption of independence would be violated. This situation becomes progressively more extreme as the number of groups increases (e.g. for 10 groups there are nine degrees-of-freedom but 45 possible pairwise comparisons).

So, *t*-tests are only designed to deal with two groups because each test is assumed to have a single degree-of-freedom. The same is true of the *U*-test. If there are more than two groups then these two types of test are inappropriate, because they assume that there are more degrees-of-freedom available than there really are – they treat each comparison as independent of every other comparison, which is an invalid assumption. This will result in the estimated probabilities being artificially much smaller than they really are.

The appropriate statistical test for analysing more than two groups is analysis of variance, if you want a parametric test, or the Kruskal–Wallis test, if you want a non-parametric test. These tests are described clearly in all reputable statistical texts for biologists (e.g. Sokal and Rohlf, 1994; Glantz, 1997; Zar, 1999), and most commercially available computer statistical programs will perform the calculations. These tests take into account the correct degrees-of-freedom for multiple groups when calculating the probability. This is why the degrees-of-freedom for the number of comparison groups needs to be reported for these two tests (as well as the degrees-of-freedom associated with the number of observations within each group).

However, these tests only indicate whether or not there is a statistically significant difference somewhere among the groups – that is, that at least one of the experimental groups is different from at least one other experimental group. This may or may not be sufficient information to answer the biological question being asked. If more information is required, indicating exactly which groups differ from which other groups, then a multiple comparison test is required. This test follows *after* the analysis of variance or Kruskal–Wallis test, using some of the results of that analysis to make the more detailed comparison. The multiple comparison test makes every possible pairwise comparison of the experimental groups, but it attempts to take into account the correct degrees-of-freedom when calculating the probabilities associated with each pairwise comparison. The estimated probabilities will then be much closer to the true values.

Unfortunately, there is a range of such tests available, and there is no real consensus among statisticians about which one (or ones) should be used. Do not let this fact put you off – these tests are still far superior to the inappropriate use of *t*-tests and *U*-tests. For parametric analyses, the Tukey test and the Newman–Keuls (or SNK) test are the ones that are most commonly recommended. They should produce similar results, but they will not always be identical. For non-parametric analyses, there are equivalent tests to these two tests but they will only deal with groups that have equal sample sizes; the Dunn test is usually the recommended test for data with unequal group sizes. These tests are described clearly in several statistical texts for biologists (e.g. Sokal and Rohlf, 1994; Glantz, 1997; Zar, 1999), and most commercially available computer statistical programs will perform the calculations for at least one of the para-

metric tests. Sadly, many of the commercially available computer statistical programs do not make any of the non-parametric multiple comparison tests available.

It is important to note that a multiple comparison test is not as effective as the original analysis of variance of the same data, and so their results may differ slightly from the results of the analysis of variance. For example, the analysis of variance may indicate that there is a statistically significant difference somewhere among the groups but the multiple comparison test may fail to find it. Alternatively, the analysis of variance may indicate that there is no statistically significant difference anywhere among the groups but the multiple comparison test may indicate that there is. In both cases, the analysis of variance is the superior procedure, and its results should be preferred. For example, if the analysis of variance finds a significant difference among the groups then at a minimum the groups with the largest and the smallest means must be different, irrespective of what the multiple comparison test says. A similar set of comments applies to the Dunn test with respect to the Kruskal–Wallis test, although the likelihood of this happening is less.

Finally, if you were to do an analysis of variance, discover that it indicates a significant difference among the groups, and then investigate where this difference lies using a series of *t*-tests, then you would actually be doing what is called the Fisher Least Significant Difference (LSD) test. This procedure was first proposed in 1949, and it is never recommended these days as an appropriate multiple comparison procedure because it still underestimates the true probability for each pairwise comparison. This comment is relevant here because this same procedure is sometimes recommended for non-parametric testing. That is, a series of *U*-tests is performed after a Kruskal–Wallis test is found to be statistically significant, to examine all possible pairwise comparisons. I cannot agree with this recommendation – the use of the Dunn test or the non-parametric equivalents of the Tukey or SNK tests will always be more appropriate.

As a specific example of this problem in the parasitological literature, Matthews et al. (2001) describe an experiment in which the worm (*Dictyocaulus viviparus*) burden was measured in the lungs of calves subjected to one of four experimental treatments. One group of calves was immunised with L3 (a radiation-attenuated larval vaccine), one group was immunised with native adult excretory/secretory (ES) products (including acetylcholinesterase), one group was immunised with a recombinant parasite acetylcholinesterase (N2 fusion protein), and the final group of calves consisted of the challenge control. The statistical differences in the worm burdens of the calves in each of these groups (measured 28–30 days post-challenge) was compared by a series of non-parametric Mann–Whitney *U*-tests, using all possible pairwise comparisons of the experimental groups.

In this analysis there are $k = 4$ groups and therefore $k - 1 = 3$ degrees of freedom. This means that only three of the $k \times (k - 1)/2 = 6$ possible pairwise comparisons of the four groups can be independent of each other. Any data analysis that involves more than three 'independent' comparisons will be inappropriate, and will result in probabilities that are much smaller than they should be. The results of the six *U*-tests are shown in Table 1. At the conventional level of $P < 0.05$ for statistical significance, three of the six tests can be deemed to be statistically significant, resulting in the conclusion that the worm burden in the L3-treated calves is different from that of the other three experimental groups.

The appropriate non-parametric statistical test for these four experimental groups is the Kruskal–Wallis test. The result of this statistical test is $H = 7.48$, $P = 0.058$. This test indicates that there is actually not much evidence that the worm burden of any of the four experimental groups differs from any of the others (if we use the same level of $P < 0.05$ for statistical significance). This clearly contradicts the results of the *U*-tests, thus highlighting the problem of inappropriately using statistical tests – it is quite likely that you will reach a false conclusion.

The reason for the different results can be clearly seen if the multiple-comparison Dunn test is applied to the data (the

Table 1
Probabilities derived from three separate analyses of the four groups of worm burdens

Comparison of worm burden groups[a]	Non-parametric analyses		Parametric analysis
	Mann–Whitney *U*-tests[b]	Dunn tests[c]	Tukey tests[d]
Control vs. L3	0.033	0.286	0.127
Control vs. ES	0.749	0.992	>0.999
Control vs. Fusion	0.423	0.966	0.921
L3 vs. ES	0.033	0.156	0.143
L3 vs. Fusion	0.019	0.053	0.043
ES vs. Fusion	0.631	0.985	0.894

[a] The four groups are: challenge control, L3 infection, adult ES products, fusion protein.

[b] These tests are performed independently for each comparison. The analyses are based on the data in Table 1 of Matthews et al. (2001), and the probabilities differ slightly from those reported in that paper.

[c] These tests are multiple comparisons following after a Kruskal–Wallis test, which produced $H = 7.48$, $P = 0.058$.

[d] These tests are multiple comparisons following after a one-factor analysis of variance, which produced $F = 3.04$, $P = 0.056$.

sample sizes are unequal for this data set, and so this is the appropriate non-parametric test). This test takes into account the correct degrees-of-freedom when calculating the probabilities associated with each pairwise comparison. The results of this analysis are also shown in Table 1. Note that for every comparison the probability from the Dunn test is much larger than the one produced by the equivalent *U*-test. This is what always happens with Mann–Whitney *U*-tests and with *t*-tests: the probabilities are artificially lowered because these tests assume that there are more degrees-of-freedom available than there really are. Under these circumstances, if you use these tests in this manner then you are thus in danger of concluding that there are statistically significant patterns in your data when there are not.

This same pattern of results occurs if these data are analysed with a parametric analysis. In fact, there seems to be no reason why such an analysis cannot be validly performed – the variances of the groups are approximately equal (Bartlett test: $\chi^2 = 4.87$, $P = 0.218$) and the data are normally distributed (Kolmogorov–Smirnov test of the residuals: $D = 0.11$, $P = 0.761$). The result of the analysis of variance is $F = 3.04$, $P = 0.056$, and the results of the Tukey multiple comparison test are shown in Table 1. These results match closely the results of the non-parametric analysis, thus confirming that the evidence for any difference among the worm burdens is doubtful. Note also that one of the results of the Tukey analysis contradicts that of the analysis of variance, as mentioned above – that is, it indicates a statistically significant difference between groups for one of the comparisons while the analysis of variance suggests that there are none.

2. Multiple tests

If we use the conventional significance level of $P = 0.05$ when we test hypotheses statistically, then we are saying that we are prepared to make a mistake 5% of the time when we reject the null hypothesis. Or, put the other way around, when we reject a null hypothesis we are likely to be making a mistake 5% of the time, on average. So, even for completely random data, in which there is no pattern at all, statistical tests will find a statistically significant pattern 5% of the time, if we test hypotheses at $P = 0.05$.

This fact is not a problem if we are only testing a few hypotheses. However, clearly, the more tests that we perform then the more inevitable it is that some of the time we have made a mistake (i.e. a Type I error). For example, if we perform 50 tests then we are likely to make $50 \times 0.05 = 2.5$ mistakes, on average. So, when we look at the result of a large series of statistical tests, and we observe that some of the tests are statistically significant at $P < 0.05$, some of these significant results may represent real biological patterns and some may represent Type I errors. If they are real biological patterns then we should reject the null hypothesis, and go on to interpret the biological meaning of the patterns, but if they are mistakes then we should not reject the null hypothesis.

This potential problem is dealt with by recognising the distinction between the individual error rate of a single statistical test and the setwise (or groupwise) error rate of a set of tests. The probability resulting from a particular test is the individual error rate, and its calculation is based on the assumption that the outcome of this single test is the only thing of interest. When there is a series of tests being performed, each separate test has its own individual error rate, but we are *also* interested in the cumulative probability associated with the series of tests as a whole, which is the setwise error rate. The individual error rate tells us the probability of having made a mistake for this particular test on its own, while the setwise error rate tells us how likely we are to have made at least one mistake among all of the tests in the set as a whole.

For example, for $j = 10$ hypothesis tests, each performed at an individual error rate of $p = 0.05$, the setwise error rate can be calculated from the formula: $1 - (1 - p)^j = 0.40$. So, while we have only a 5% chance of making a Type I error for each individual test, we have a 40% chance of making at least one Type I error in the entire collection of tests. From a practical point of view, what happens in most data analyses is that the experimenter chooses the individual error rate and then lets the setwise error rate be whatever it turns out to be – that is, they choose to control the individual error rate but not the setwise error rate. This is an inappropriate thing to do.

The solution to this problem is therefore to choose the setwise error rate rather than the individual error rate. That is, we choose $P = 0.05$ (for example) for the setwise error rate as our criterion for statistical significance, and then we calculate what individual error rate we need to use for the individual tests in order to achieve this setwise error rate. We thus control both the setwise error rate *and* the individual error rate.

The simplest and best-known procedure for controlling both the setwise and individual error rates is to use the Bonferroni inequality (e.g. Bland and Altman, 1995; Glantz and Slinker, 2001). In this procedure, we choose the setwise error rate, p', and then calculate the individual error rate as $p = p'/j$. We then reject the null hypothesis for a particular test only if $P < p$. For example, if we are carrying out $j = 10$ hypothesis tests and we want to use the conventional significance level of 0.05, we would set $p' = 0.05$ for the setwise error rate and then use $p = 0.05/10 = 0.005$ as the individual error rate for each of the 10 tests. We would only declare an individual test as significant if we get $P < 0.005$ for that test. In this way we guarantee that there will, on average, be less than one mistake among the set of tests. This procedure is implemented in most commercially available computer statistical programs, but even if it is not then it is obviously easy to carry out by hand.

The only potential problem with this procedure is that it is

statistically conservative. This means that it slightly over-corrects the individual error rates, and thus the setwise error rate will actually be less than our specified level. This is not a particularly big problem, and many commentators recommend ignoring it. However, several modifications to the procedure have been proposed to deal with this conservatism, notably those of Holm and of Hochberg, which modify the probabilities sequentially rather than simultaneously. Alternatively, one can use the Dunn–Sidák adjustment: $p = 1 - (1 - p')^{1/j}$. These modified procedures are described by Glantz and Slinker (2001).

It is worth noting that the problem of uncontrolled setwise error rates exists in addition to the problem of non-independence discussed above. For example, if there are 50 groups being analysed then there are 49 independent tests that could be performed, but there will still be 2.45 Type I errors, on average, among these tests. If we were to perform all 1,225 possible pairwise comparisons, then there will be 61.25 Type I errors, on average. This means that performing all possible pairwise comparisons has two sources of potential statistical problems. It is sometimes suggested that the Bonferroni correction will deal with both problems, but this is clearly not so.

As a specific example of this problem in the parasitological literature, Muller et al. (2001) describe an experiment in which one group of mice was experimentally infected with the helminth *Schistosoma mansoni* and compared to a control group of uninfected mice. The neutral lipid content of the liver, ileum and serum was quantified in both groups of mice at three times post infection (prior to egg deposition, during acute disease, during established disease). Three types of neutral lipid were measured in the liver, three types of lipid were measured in the ileum, and four types were measured in the serum, giving a total of 10 data variables. The differences between the control and infected mice were analysed for each of these 10 data variables at each of the three times using a series of *t*-tests, giving a total of $10 \times 3 = 30$ statistical tests.

Note that the authors were interested in the *set* of statistical analyses rather than in any particular individual *t*-test. They found that three of their tests were statistically significant at $P < 0.05$, and they thus rejected the null hypothesis for each of these three tests. They concluded from this set of tests that schistosomiasis causes "significant changes in the lipid profiles of the liver and ileum".

However, for 30 statistical tests, each conducted at the conventional Type I error rate of $P < 0.05$, there will be $30 \times 0.05 = 1.5$ mistakes, on average. The number of significant results in this set of tests differs very little from this expectation, and so there is certainly not very strong evidence among these *t*-tests for any effect of schistosomiasis on the mice. That is, the three significant results are no more than would be expected by random chance.

If we use the Bonferroni procedure to choose the individual error rate, specifying $p' = 0.05$ for the setwise error rate, we would use $p = 0.05/30 = 0.0017$ as the individual

Table 2
Results of a separate two-factor analysis of variance for each of the 10 data sets from the lipid experiment

Source of variation	Free sterols		Free fatty acids[a]		Triacylglycerols		Cholesteryl esters[b]	
	F	*P*	*F*	*P*	*F*	*P*	*F*	*P*
Liver[c]								
Time[d]	7.68	0.002			14.68	<0.001	6.77	0.004
Treatment[e]	9.11	0.006			7.11	0.013	7.98	0.009
Time × Treatment[f]	0.74	0.485			0.88	0.425	0.05	0.956
Ileum[g]								
Time	14.13	<0.001	7.26	0.003	16.60	<0.001		
Treatment	1.45	0.239	3.01	0.095	1.44	0.241		
Time × Treatment	1.35	0.278	2.24	0.127	0.16	0.856		
Serum[h]								
Time	11.74	<0.001	8.17	0.002	3.22	0.056	0.27	0.767
Treatment	5.49	0.027	0.04	0.839	0.07	0.801	1.07	0.310
Time × Treatment	1.39	0.266	0.76	0.476	0.61	0.552	0.27	0.767

[a] Data not collected from the liver.
[b] Data not collected from the ileum.
[c] Based on the data in Table 1 of Muller et al. (2001).
[d] For this factor there are three groups, one for each time (14 days p.i., 45 days p.i., 75 days p.i.). The test assesses whether there is any difference in the response of the mice among these three groups (averaged across the two experimental treatments).
[e] For this factor there are two groups, one for the *S. mansoni*-infected mice and one for the control mice. The test assesses whether there is any difference in the response of the mice between these two groups (averaged across the three times).
[f] This is the interaction between the two main factors. The test assesses whether the schistosomiasis treatment produces different effects on the mice at different times compared to the control group.
[g] Based on the data in Table 2 of Muller et al. (2001).
[h] Based on the data in Table 3 of Muller et al. (2001).

error rate for each of the 30 tests. None of the *t*-tests is significant at this probability, and so we would not reject any of the null hypotheses. There is thus no evidence at all in this modified data analysis that schistosomiasis influences the lipid profile of the mice. This analysis clearly contradicts the results of the previous analysis, thus highlighting the problem of inappropriately using statistical tests – it is quite likely that you will reach a false conclusion.

As a further important point about this set of data analyses, it is actually unnecessary to be using a large series of *t*-tests at all. There are two separate sources of experimental variation in this experiment: the two groups of mice, and the three times of measurement. This means that there are six groups of observations being analysed for each lipid type (two types of mice × three times). These six groups have been analysed by three *t*-tests, whereas the point has been made above that when comparing more than two groups of observations an analysis of variance is more appropriate.

In this particular example, there will be 10 analyses of variance, one for each of the 10 lipid data variables. Each of these analyses will be a two-factor orthogonal analysis of variance, consisting of a test for equality of the two mouse groups (averaged across the three times), a test for equality of the three times (averaged across the two mouse groups), and a test for the interaction between mouse-type and time. It is this last test that is of primary interest in the experiment, because it assesses whether the difference between the two mouse groups is the same at each of the three times – for example, the two mouse groups might not be different in lipid content at the first time measurement but might become more and more different as the course of the disease progresses. It was solely this interaction that the authors were interested in examining with their series of *t*-tests.

The use of analysis of variance thus provides a far more comprehensive set of analyses than does the 30 *t*-tests. The tests of the interaction are reduced from 30 tests down to 10 tests, which will help deal with the multiple-test problem. Furthermore, if desired, a further 20 tests can be performed to assess two sets of patterns that were not tested at all by the *t*-tests. The results of these 10 analyses will thus be far more appropriate for this particular experimental design. Performing the actual mathematical calculations is not difficult – the standard statistical texts explain clearly what is going on, and the standard computer programs will do all of the necessary mathematics.

The results of these 10 analyses of variance are shown in Table 2. As far as the interaction is concerned, none of the tests is significant at $P < 0.05$, and so we must conclude that there is no evidence in these data that schistosomiasis influences the lipid profile of the mice in any time-related way. This is the same conclusion that we came to from the set of Bonferroni-corrected *t*-tests above. However, as far as the experimental infection of the mice is concerned (i.e. the 'Treatment' factor shown in Table 2) all three of the tests are significant at $P < 0.05$ for the liver lipids but only one of them is significant for the ileum or serum lipids. There is thus some evidence that there is a time-independent effect of schistosomiasis on the lipid content of the liver but not on that of the ileum or serum. This conclusion was never tested by the set of *t*-tests, and so it is an important contribution of the analyses of variance. Finally, eight of the 10 tests of the 'Time' factor are significant at $P < 0.05$ (with an additional borderline result), indicating that the lipid content of the liver, ileum and serum in the mice changes through time irrespective of the infection with schistosomiasis. This conclusion was also never tested by the *t*-tests.

This leaves us with the final point: should the 30 hypothesis tests in the series of analyses of variance be tested at $P < 0.05$, $P < 0.0017$, or some other probability? We are certainly not performing 30 individual statistical tests, as the hypothesis tests are grouped into 10 analyses of three hypotheses each; so using $P < 0.0017$ is inappropriate. On the other hand, there is definitely a setwise error rate that is greater than the individual error rates; so using $P < 0.05$ is also inappropriate. It is not entirely clear exactly what individual error rate should be used, but $p = 0.05/10 = 0.005$ would be a reasonable value, given that we are interested in the results of all 10 analyses simultaneously. None of the conclusions listed above changes substantially if this revised significance level is used.

References

Bland, J.M., Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. Br. Med. J. 310, 170.

Glantz, S.A, 1997. A Primer of Biostatistics, 4th Edition. McGraw-Hill, New York.

Glantz, S.A., Slinker, B.K., 2001. Primer of Applied Regression and Analysis of Variance, 2nd Edition. McGraw-Hill, New York.

Matthews, J.B., Davidson, A.J., Freeman, K.L., French, N.P., 2001. Immunisation of cattle with recombinant acetylcholinesterase from *Dictyocaulus viviparus* and with adult worm ES products. Int. J. Parasitol. 31, 307–17.

Muller, E., Rosa Brunet, L., Fried, B., Sherma, J., 2001. Effects on the neutral lipid contents of the liver, ileum and serum during experimental schistosomiasis. Int. J. Parasitol. 31, 285–7.

Sokal, R.R., Rohlf, F.J., 1994. Biometry: The Principles and Practice of Statistics in Biological Research, 3rd Edition. Freeman, San Francisco.

Zar, J.H., 1999. Biostatistical Analysis, 4th Edition. Prentice-Hall, Upper Saddle River, NJ.

David A. Morrison*
Molecular Parasitology Unit, University of Technology Sydney, Westbourne Street, Gore Hill, NSW 2065, Australia

* Tel.: +61-2-9514-4159; fax: +61-2-9514-4003.
E-mail address: david.morrison@uts.edu.au (D. A. Morrison).