BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2012.02206.x
www.brjpharmacol.org

EDITORIAL

Type I: families, planning and errors

Gordon B Drummond[1] and Sarah L Vowler[2]

[1]*Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, Edinburgh, UK, and* [2]*Cancer Research UK, Cambridge Research Institute, Li Ka Shing Centre, Cambridge, UK*

Correspondence
Dr Gordon B. Drummond, Department of Anaesthesia and Pain Medicine, University of Edinburgh, Royal Infirmary, 51 Little France Crescent, Edinburgh EH16 4HA, UK. E-mail: g.b.drummond@ed.ac.uk

This article is being published in *The Journal of Physiology*, *Experimental Physiology*, the *British Journal of Pharmacology*, *Advances in Physiology Education*, *Microcirculation*, and *Clinical and Experimental Pharmacology and Physiology*.

Gordon Drummond is Senior Statistics Editor for *The Journal of Physiology*.

Sarah Vowler is Senior Statistician in the Bioinformatics Core at Cancer Research UK's Cambridge Research Institute.

This article is the 11th in a series of articles on Best Practice in Statistical Reporting. All the articles can be found at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm.

Key points

- Comparisons propose no difference, and then ask 'How probable?'
- Misclassification is inevitable from time to time: false conclusions result
- Families of observations are best only tested once
- The more comparisons, the more likely is misclassification
- For several comparisons in one family, test criteria should be more stringent

Scientists frequently want to answer the question 'has this treatment had an effect?' Most are unaware that the tests they usually use do not directly address this question. These tests usually pose a different question, based on the possibility that *nothing* has happened. The question becomes: 'how probable are these data, if there were NO difference between the original populations from which the data have been randomly drawn?' (In fact, for most laboratory experiments, this supposition is patently false – the experiment has been conducted using a pre-ordained sample, possibly randomly divided into treatment and control groups, but certainly not randomly sampled.)

However, if we continue to consider the usual analysis that is used, we have to assume that we have random samples, from the same population. Such samples will always differ to some extent. Occasionally, the difference might be substantial, large enough to suspect that they might not have come from the same source population. The usual context in which we use this test is that the data are already 'under suspicion': we usually don't want to believe the null hypothesis at all, and we are testing to see if the data are unlikely to be consistent with this hypothesis. To assess how 'suspicious' our results can be, we estimate how frequently we might obtain results like ours:

- if the 'null hypothesis' were true
- if we were to repeatedly sample the population
- if the results are workings of chance.

Generally, we reject the null hypothesis if chance alone could yield data like ours less than 1 time in 20 (or, equivalently, 95 times out of 100), an arbitrary and probably unnecessarily inflexible value (Ziliak and McCloskey, 2008). We believe our suspicions are justified, and we can then accept the alternative hypothesis: the samples are not from the same population. We rarely employ the same cautious vocabulary as the statistician, who might qualify this interpretation. The researcher assumes, wrongly, a possibility of 0.05 (i.e. 5%, or 1 in 20) to indicate that the null hypothesis is false, there is

therefore a genuine difference, and that the result can be reliably replicated (Cohen, 1994). The more cautious statistician would argue that the findings are consistent with that conclusion, but are not unequivocal. Indeed, this is so: if a single experiment just meets the level of significance, it is just as likely NOT to give a significant result if the same experiment were to be repeated. It's a bit like exams: the marks of students that just fail are often considered carefully, in case the examiners have been too severe, but the students that scrape through are illogically allowed to pass without further scrutiny. In testing our results, we accept the fact that we may conclude that an intervention has had a 'real effect' when in fact we may be wrong 5% of the time: this is the type I error rate (Figure 1).

Usually scientists do one experiment at a time, or at least they think they do: the experiment asks the question 'has this treatment had an effect?', and if $P < 0.05$, we accept that we have found an effect. However, in many experiments, a variety of factors such as time, expense, resources and concern about animal use may lead to a single study asking several questions. Each of these may require statistical testing. As soon as more answers are sought from the same data, the trustworthiness of the answers can change. The error rate in the experiment as a whole (experiment-wise error rate) will increase and become greater than the error rate in each comparison (comparison-wise error rate).

However, the context of experiments and tests can vary, and context affects the logic of the tests. Yet again, statistical terminology can be confusing, and clear definitions are often lacking. Suppose we go back to our Californian frogs, and choose to study samples supplied by a dealer. She assures us that these are random samples from seven different counties: Alpine, Butte, the well-known Calaveras County, Del Norte, El Dorado, Fresno and Humboldt (California lacks a county with the initial G). We measure how far they can jump, and use ANOVA to assess the possibility that performance may differ according to origin (Figure 2).

The ANOVA test assesses the possibility that all these samples have come from a single population. This is an 'omnibus' test, considering all the results together and comparing the variation (and potential differences) between and within all the groups. It's a good name: we load all the sets of data into the omnibus and test them together. Subsequently, we may choose to make a 'family' of comparisons between the data sets present in the 'omnibus'. 'Family' is a more difficult statistical concept, and often loosely used, different authors expressing different opinions. Ludbrook suggested, 'A family of hypotheses is all those actually tested on the results of a single experiment' and also that a family is 'all those experimental observations that could be analysed statistically by a global procedure' (such as an omnibus test) (Ludbrook, 1998). Perhaps, it's as well to bear in mind that data families, like social ones, can breed trouble.

Looking at our results, we're surprised that ANOVA suggests there is no evidence of a difference between the groups. We could have sampled data like this, or even more extreme, about 40% of the time. This can happen: differences can be detected with further tests that are not shown up by an omnibus test. We conducted this study with the suspicion that there could be a difference here because one group comes from the Calaveras County. So, we start to compare the groups in pairs. (We ignore more complex possibilities; it's possible that we might want to compare northern counties with southern counties, and so on.) There are 21 ways to conduct simple pairwise comparisons: the family is shown in Figure 3.

Here, we have 21 comparisons (the comparison of A vs. B will yield the same result as B vs. A, so this doesn't need a

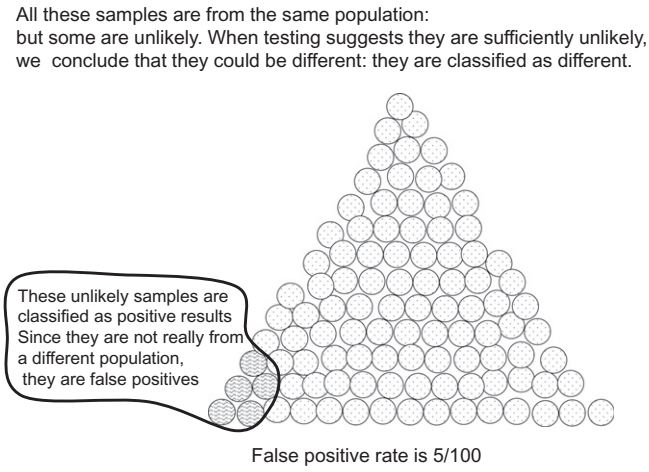

Figure 1

In the long run, if we accept $P = 0.05$ as a threshold possibility, we make the correct decision 95 times in 100 experiments. We will misclassify 5% of experiments as positive when this is not the case. The type I error rate is the rate of classification of results as positive (different) when in fact there is no difference: the false positive rate. In the context of 'trust' in results, a false positive rate of 5% is generally considered acceptable.

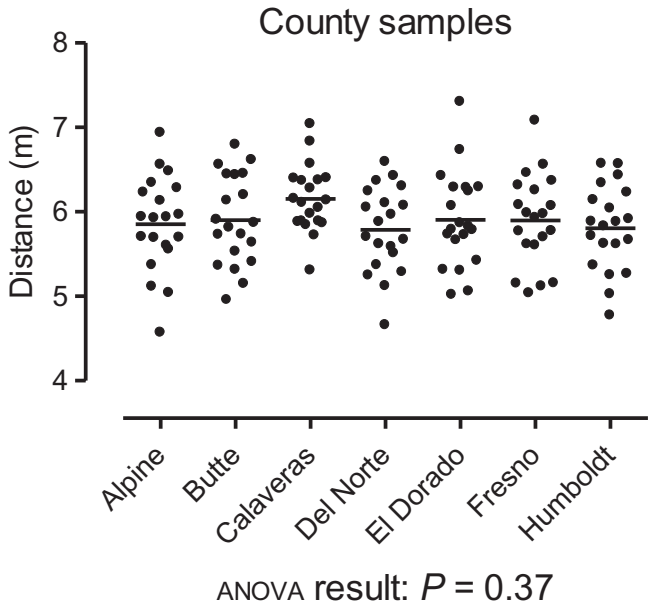

Figure 2

The samples we have obtained, and the test result.

A v B B v C C v D D v E E v F F v H

A v C B v D C v E D v F E v H

A v D B v E C v F D v H

A v E B v F C v H

A v F B v H

A v H

$$\text{Number of pairs} = \frac{n!}{2\,(n-2)!}$$

Figure 3
Family of pairwise comparisons.

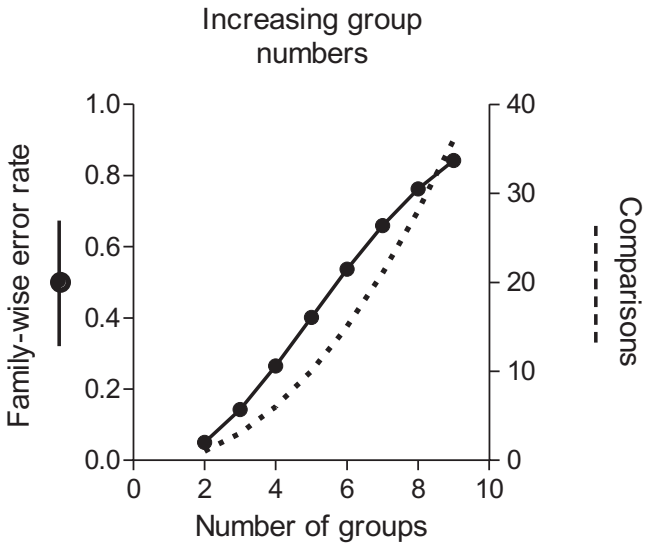

Figure 4
The number of pairwise comparisons that can be made (dashed line) increases with the number of groups available. As the number of comparisons increases, the error rate (rate of false positive conclusions) also increases (continuous line).

separate comparison), and repeat comparisons alter the overall error rate. Table 1 gives the results in order of *P*-value.

Two of the comparisons in the left-hand column would be 'significant' if they had been tested individually. In a single comparison, when we take $P = 0.05$ as a threshold probability, we know our conclusion could be wrong. Nevertheless, we accept this possibility since, in the long run, our decision would only be wrong once out of 20 times. This is the 'comparison-wise' error rate, and it is the same as the probability that we set as our threshold for 'significance'. In contrast, if we make 21 comparisons, the risk that at least one of these several comparisons could lead to an error increases substantially. The risk of error, in several comparisons, is shown in Figure 4.

The likely error in a family of tests increases with each set of data, up to the point where with seven sets of data, there

Table 1
Original *P*-values for all the possible comparisons, in order of size

Original *P*-value	Comparison	Critical *P*-value
0.015	C versus D	0.002
0.020	C versus H	0.005
0.059	A versus C	0.007
0.097	C versus F	0.010
0.098	B versus C	0.012
0.119	C versus E	0.014
0.488	D versus E	0.017
0.494	B versus D	0.019
0.502	D versus F	0.021
0.555	E versus H	0.024
0.563	B versus H	0.026
0.572	F versus H	0.029
0.704	A versus D	0.031
0.766	A versus E	0.033
0.782	A versus H	0.036
0.782	A versus B	0.038
0.791	A versus F	0.041
0.913	D versus H	0.043
0.967	E versus F	0.045
0.976	B versus E	0.048
0.991	B versus F	0.050

The comparisons are in the second column. In the third column, the critical *P*-values required have been generated by dividing the originally chosen threshold of 0.05 by the number of comparisons remaining as the column is descended. In no case is the original *P*-value less than the adjusted critical value: we conclude that none of the comparisons is significant.

are 21 tests, and the overall risk of error is 0.67. This is the experiment-wise or family-wise error rate. The usual solution proposed to the problem imposed by the multiple tests is to impose a more stringent threshold for 'significance'. The advantage is that we are less likely to have false positive classifications. The disadvantage is equally clear: with a more stringent criterion, we will fail to detect occasions where the null hypothesis is not 'true'. In other words, false negatives will become more common (Figure 5).

In the case we are considering here, we predict that the Calaveras frogs will be better jumpers, so we conduct a complex comparison between Calaveras versus all the others. This would be an *a priori* test. Some authorities would consider a comparison of this sort to be acceptable with no change to the test criterion. We move into considerations of design, motive and the need to balance the risks of confirming bias, missing interesting or important new information, or making a decision with insufficient evidence. In many instances, scientific papers are brimful of two-way comparisons and we cannot be sure that we are not just being pre-

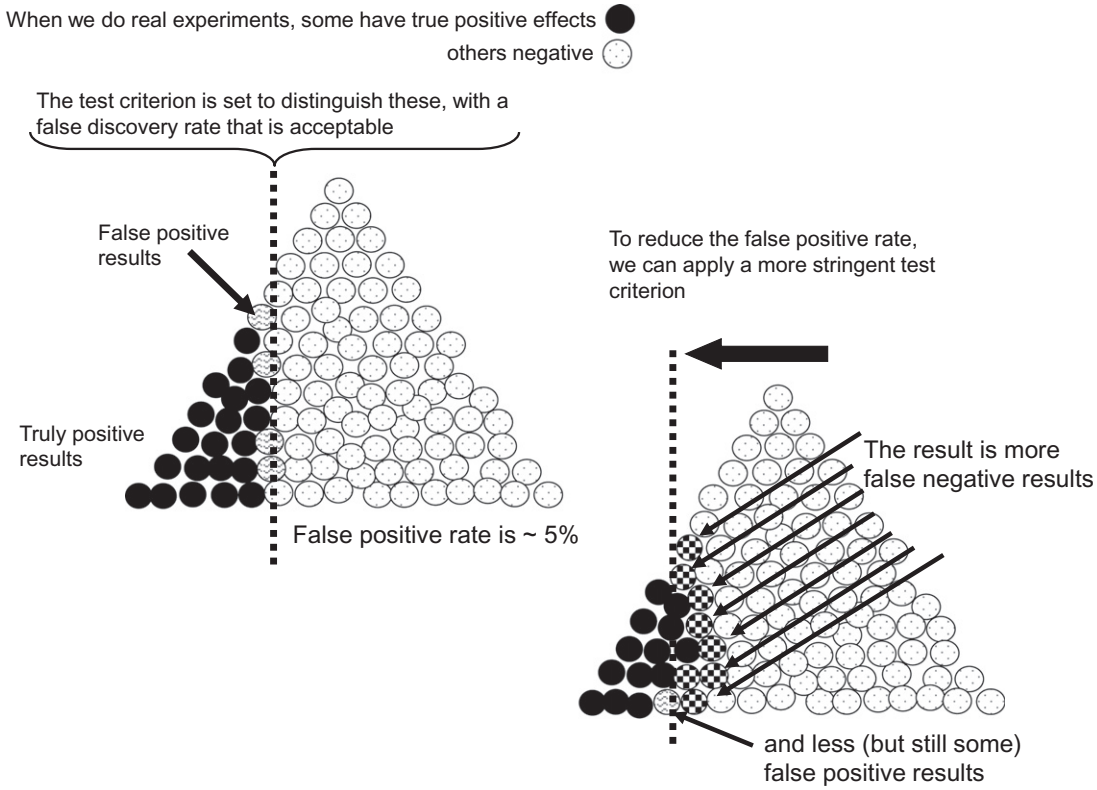

Figure 5

To reduce the experiment-wise false discovery rate, the test criterion can be made more stringent. The result is more false negative findings.

sented with a chance finding that has resulted from a succession of comparisons, as the authors seek for something positive to report.

Naturally, many studies present several experiments; for example, there may be experiments to show that a gene has been deleted, does not generate RNA, that a receptor protein is absent, that stimulation is ineffective. These are all separate experiments and can be legitimately assessed without adjusting the test criterion. Indeed, a single *a priori* test is a rare event.

However, multiple testing can often be avoided. There are sensible steps that can be taken to avoid conducting a plethora of comparisons. One is to combine data into a summary statement or expression. A simple and obvious example might be a dose–response curve (Ludbrook, 1998; Festing, 2003), or a growth trajectory, where separate groups of data can be summarized into a single relationship. More subtly, this is the principle found in the interaction term in the multivariate analysis of variance (MANOVA) procedure. However, MANOVA itself is a 'multiple' test since it can yield several *F*-values (each of which represents a test result). Another safeguard to avoid inconsistency is to set a composite hypothesis to avoid contradictory findings: for example, both A and B have to be better, or A has to be better and B not worse. (This latter test is a *non-inferiority* comparison.)

A different approach is to use a procedure that does not control the error rate, but concentrates on the rate at which false positive conclusions are likely. These simple procedures use the *P*-values generated by each test and are particularly useful because the results from different types of tests can be considered together. The primary comparisons are well sustained, even if a lot of additional tests are performed (Curran-Everett, 2000).

Look again at the *P*-values in Table 1. The right-hand column shows more rigorous *P*-value thresholds. The threshold *P*-value chosen for a single test has been divided by the total number of tests. Thus, for all 21 comparisons, the corrected value should be 0.05/21, which is ~0.0025. If the corresponding *P*-value in the left-hand column were less than this, then the next larger *P*-value would be compared with a threshold value corrected using *N*–1, that is, 20. This gives a sequence of thresholds in the right-hand column that is progressively more lenient. We find that there are no significant comparisons in our table. The last *P*-value is of course 0.05, but this is far less than the corresponding *P*-value for the 21st comparison, which is 0.99. (This method is a Ryan-Holm step-down procedure.)

If such techniques are not used, then one of the many methods for dealing with multiple comparisons will be required to reduce the impact of an elevated false discovery rate. There are many of these – too many to describe in a short introduction – and there is far from total agreement over which tests are best. Some books suggest that with unplanned comparisons, the conclusions should be graded according to differences found, into occasions where the null hypothesis should be retained, be rejected, and into an intermediate group of 'not proven' verdicts. Others, more lenient, suggest that if the data are really being 'explored', then cor-

rection for multiple tests may not be needed, or if there is an *a priori* proposal, that particular test need not be corrected, but further tests should be. The most stringent verdict is that all comparisons should be corrected and analysis should be conducted 'independent of expectation'. Horton exemplified the pitfalls of raking through the coals of an experiment to find undiscovered treasures (Horton, 2000). The authors of a paper were asked to conduct an unplanned comparison in subgroups of subjects because the assessors thought there were features of interest. They agreed on the understanding that the first feature they would analyse was the star sign of the participants, and they showed how this could be statistically interpreted as an important factor in drug response.

We have now discovered that our frog dealer was a fraud: all the frogs in our example were sampled randomly from the same population. We should have realized!

References

Cohen J (1994). The earth is round ($p < .05$). Am Psychol 49: 997–1003.

Curran-Everett D (2000). Multiple comparisons: philosophies and illustrations. Am J Physiol Regul Integr Comp Physiol 279: R1–R8.

Festing MF (2003). Principles: the need for better experimental design. Trends Pharmacol Sci 24: 341–345.

Horton R (2000). From star signs to trial guidelines. Lancet 355: 1033–1034.

Ludbrook J (1998). Multiple comparison procedures updated. Clin Exp Pharmacol Physiol 25: 1032–1037.

Ziliak ST, McCloskey DN (2008). The Cult of Statistical Significance. University of Michigan Press: Ann Arbor.