RNSH SERT Institute

Surgical | Education | Research | Training

GUIDES TO UNDERTAKING RESEARCH

3.5 Making multiple comparisons in datasets

One of the foundations upon which a medical study rests is the statistical testing of its data. This can unfortunately be undermined by a problem seen even in highly respectable peer-reviewed literature. This problem arises from making *multiple comparisons* in the data and it is one source of statistical errors that result in a lot of studies being hard to reproduce. Since a study that cannot be reproduced is a waste of time it is essential for researchers understand this issue, since it often has a relatively fix.

Making single comparisons

Statistical tests can assess whether or not patterns observed in a dataset (e.g., outcomes from patients undertaking a therapy) only arise from random variations. Working on random numbers is of no interest to anyone, but since the human mind is bad at spotting random patterns (and good at spotting non-random patterns that aren't there) there are objective methods that have been developed to assist, namely the techniques of statistical analysis.

Student's t-test

A commonly used statistical approach is Student's t-test which can be used to compare data from two groups, such as treatment and placebo groups. Thus, to see the effects of a drug on a patient parameter, two groups are set up: patients taking the drug (the treatment group) and patients taking a fake drug (the placebo group). Here, we can imagine that the data relates to systolic blood pressure, with measurements taken from both groups of patients. The question here is whether the drug reduces patient blood pressure, which is to say the group receiving the treatment has significantly reduced their average (mean) blood pressure compared to that of the placebo group.

Significantly significant

The word 'significant' is a code word here, it doesn't mean "a lot" or "important" but that there is detectable difference that is unlikely to be due to random chance. Since random data fluctuations cause variations in the group means, we want to know a reduction in mean blood pressure is *not* explained by such random fluctuations. The t-test can do this job, but it does not return a yes/no answer, rather it returns a p-value, which we then have to interpret.

P-values

The t-test p-value is an estimate of the probability that the observed difference between placebo and treatment groups (or a larger difference than that seen) is due to random variation. A p-value less than 5% is often used to establish that a drug effect is non-random; thus a p-value below 5% is often designated as 'statistically significant'. Note that a p-value above 5% does not prove that the drug has no effect, it simply means that we have no evidence for an effect; it may be there but we can't see it. That 5% threshold for the p-value is not very stringent (lower thresholds are often used and are preferable) but it is often employed so we will use it here.

A key point is that a p-value of 5% represents a 1 in 20 chance that this conclusion is wrong, i.e., it was a false positive, since the observed decrease in blood pressure was actually due to random noise after all. Leaving aside that little issue, here lies the source of another really important problem when we want to test not just one, but several parameters in the same patients.

Multiple tests can spoil everything

The t-test and the p-values it gives are widely understood, but multiple tests are often needed. Suppose not just blood pressure but 20 different parameters are tested in patients taking the drug (or placebo) and that t-tests are performed for all of these parameters. Furthermore, imagine every one of those 20 t-tests returned a p-value just below 5% so would appear to be statistically significant. Having 20 comparisons each of which has a 1 in 20 chance of being wrong suggests that something will be wrong. Indeed, something is wrong; the chance they are *all* non-random true effects is only about 1 in 3. That is a problem: we will overstate the number of true positive results Worse, it is not at all clear which of those measured parameters might be giving a false positive result.

Ronald Fisher's potatoes

Fisher was one of the founders of statistical science and invented many of its core concepts, including p-values. In one of his studies conducted in the early 1920s he had data on the yield of 12 different varieties of potatoes and wanted to find which gave the most productive crop. For this he used his newly developed ANOVA (or analysis of variance) model.

In this model, the mean and variance of the measurements of potato yield (weight per crop) for each potato variety was calculated and the ANOVA test determined whether all the means were the same or if there was some difference detected. Thus, ANOVA compared whether mean and variance of yield between the individual potato varieties were similar to that seen when the overall mean and variation data from all varieties were pooled together. Fisher's potato data indicated there were indeed differences between the potato varieties. However, the test did not reveal which varieties were best.

This is a type of *omnibus* test that alerts us that something is going on with the our parameters of interest but does not tell us which parameter. Such testing only gives useful general information about the dataset as a whole but it is important. Moving away from the potato fields for a moment, imagine a biscuit jar in a share house that one day is found to be unusually empty. This information indicates that biscuit consumption has suddenly increased but does incriminate a particular over-consuming housemate. Clearly it is best to determine that the biscuit jar really is lower than usual before leveling accusations. This illustrates that are thus two stages to the enquiry

1. detecting a biscuit deficit then
2. culprit identification.

In the same way ANOVA indicates that there is some statistical signal which can be followed up by specific tests.

Threshold correction

Thus, after ANOVA finds a signal a *post hoc* test is needed. There are various approaches used, and it not always clear what is the most appropriate (it depends on the dataset) but a common approach is to adjust the p-value threshold to compensate for multiple testing. This will reduce the number of false positives.

One approach is to use a False Discovery Rate, an algorithmic method used for large datasets. Also common is the use of pairwise t-tests used to compare each pair of means then use a Bonferroni correction to the p-value threshold. In this case the threshold (here, 5%) is divided by the number of tests done and that adjusted p-value is then used to determine significance. The p-value then has to be lower than this adjusted value. Other *post hoc* tests exist, including

Tukey's, which is more conservative than Bonferroni, i.e., less to obtain a result designated as significant.

1-way and 2-way ANOVA

Fisher's fun with potatoes did not end with potato varieties since there were also different types of fertilisers used with the potatoes, and the varieties varied in their response to the fertilisers. This provides an extra dimension (or category) to the data results in a 'two-way' comparison, since the potato yield can vary due both to potato type and fertiliser. This cannot be done with the simple ANOVA as described above and requires a *2-way ANOVA* test. The details need not detain us here, as software will deal with this.

Similarly to those potatoes, in clinical data there may be both treatment and patient sex to take into account because there may be a difference in treatment response between the two sexes. 2-way is more work to calculate than simple 1-way ANOVA (horrifyingly so back in the 1920s) but for computer software the work is only marginally more.

Not using ANOVA

The main assumption underlying ANOVA is that like t-test there is an assumption that the data is normally distributed. If this is not true, a test such as Kruskal-Wallis with paired Mann-Whitney tests can be used. These test data ranks rather than the data itself. Which test to use may depend on the software available but it would be wise to get some advice on this. There are other approaches to examining multiple comparisons, such as methods based on Bayesian statistics, but this is a large subject in itself.

Author: Julian Quinn ***Version*** 2.2 (Dec 2020)
Thanks to Dr Richard Piper and Dr Fenglian Xu for reviewing and critiquing this article.

Royal North Shore Hospital, DIVISION of SURGERY and ANAESTHESIA