Multiple comparisons: The issues as I see them

Roger Newson
Asthma Club, 7 February 2008

Figure 1. Test statistics for 110 comparisons using 3 confounder sets

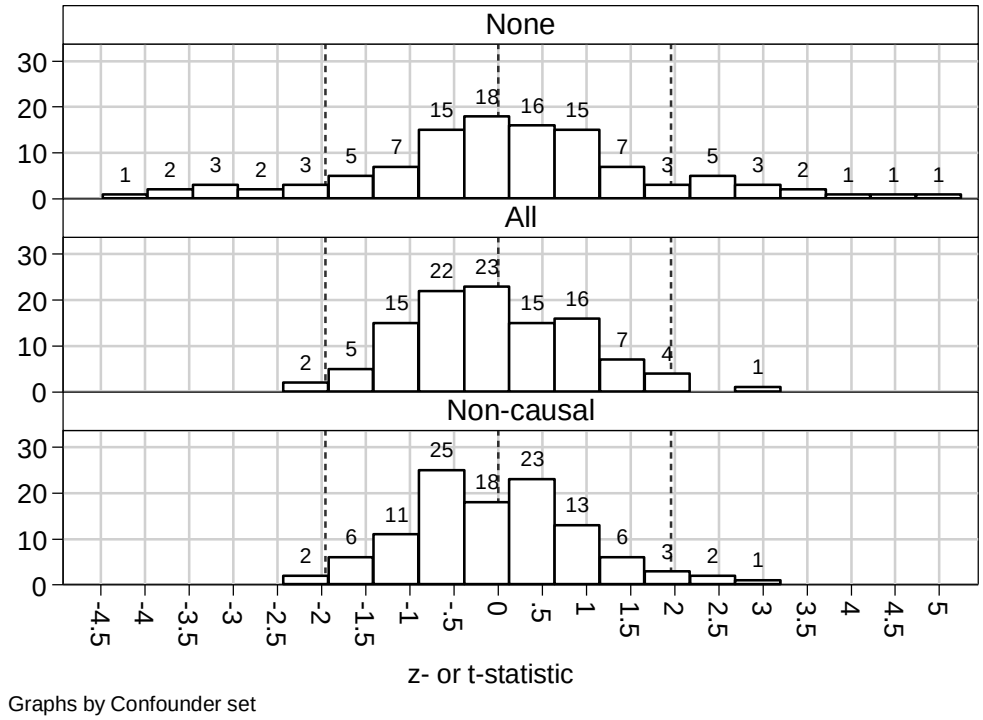

Figure 1 summarizes a recent analysis from the ALSPAC child cohort, which (as luck would have it) illustrates the issues involved in multiple comparisons particularly well. We measured 110 associations, corresponding to 5 maternal dietary score exposures combined with 22 child disease outcomes. Each of these associations was measured adjusting for 3 confounder sets, which we have named "None" (the empty set, corresponding to an unadjusted analysis), "All" (the full set of 35 confounders), and "Non-causal" (all confounders except for 3, which some epidemiologists consider to be possibly on the causal pathway between diet and disease). The histograms show the distributions of the *z*-statistics or *t*-statistics, which, for samples of over 1000 such as ours, are expected to have a very nearly standard Normal sampling distribution, if the null hypothesis is true.

In the frequentist scheme of things, if there are multiple comparisons and they are all equally *a priori*, then there is automatically a multiple comparisons issue. We addressed this issue by carrying out the Simes procedure, controlling the false discovery rate (FDR) at 0.05, on the 110 *P*-values for the 110 comparisons adjusted for each confounder set (separately for the 3 confounder sets). The Simes procedure defines a "discovery set" of comparisons, with the feature that we can be 95% confident that *some* of the corresponding null hypotheses are false, or 90% confident that *most* of the corresponding null hypotheses are false. In general, a frequentist multiple-test procedure extends the concept of confidence regions by defining a confidence region, not for a scalar parameter, and not for a vector parameter, but for a

set-valued parameter, namely "the set of null hypotheses that are true". Frequentist multiple-test procedures are discussed in Newson *et al.* (2003).

In the unadjusted associations, the top subgraph of Figure 1 shows that the distribution of the test statistics is *not* standard Normal. The Simes procedure defined a critical *P*-value threshold of .00727273 and a discovery set of 16 associations, with *P*-values below that threshold, and test statistics forming the tails of the distribution. In cases such as this, we say that we are "data-mining in rich paydirt".

In the adjusted associations, the lower 2 subgraphs of Figure 1 show that, with both confounder sets, the distribution of the test statistics is approximately standard Normal. The number of "nominally significant" comparisons ($P \leq 0.05$) is 5 with both confounder sets, and is similar to the number that we would expect, assuming *all* null hypotheses to be true. The Simes procedure defined a much stricter critical *P*-value threshold of .00045455, and an empty discovery set, for both confounder sets. In cases such as these, we say that we are "data-mining in poor paydirt".

Note that the Simes procedure (unlike the Bonferroni and Sidak procedures) is *NOT* "blindly adjusting for the number of comparisons". The Simes critical *P*-value threshold depends, not only on the number of *P*-values, but also on the *P*-values themselves. That is why it is *much* higher for 110 comparisons in rich paydirt than for 110 comparisons in poor paydirt. With the Bonferroni and Sidak procedures, the critical *P*-value threshold invariably falls towards zero as the number of comparisons increases. (And, therefore, so does the power to detect a difference of a given size with a given sample number.) With the Simes procedure, the critical *P*-value threshold *typically* falls, as the number of comparisons increases, towards a minimum that depends on the richness of the paydirt in which we are data-mining. In rich paydirt, that minimum will be positive. However, in totally sterile paydirt, that minimum will be zero. This is because, in order to do its job, the Simes procedure *must* produce an empty discovery set in at least 95% of samples, when all null hypotheses are true. In the lower 2 subgraphs of Figure 1, where the discovery sets are empty, the Simes threshold is in fact equal to the Bonferroni threshold. This feature of the Simes procedure often makes epidemiologists depressed, but is a necessary consequence of being evidence-based, objectivist and frequentist.

In the authority-based subjectivist Bayesian scheme of things (as I understand it), the rules are slightly different. There, the epidemiologists may impose an authority-based prior probability, and claim to *KNOW* that at least 5 percent of their hypotheses *must* be true. Under these circumstances, the role of the statisticians is not to question this, but to estimate *which* 5 percent. The epidemiologists are therefore guaranteed a stream of "positive" results in the long run, even if the paydirt that they are mining is totally sterile. Unfortunately, the authority of such epidemiologists does not always extend beyond the epidemiology sector, or even beyond their own Departments.

References

Newson R, The ALSPAC Study Team. Multiple-test procedures and smile plots. *The Stata Journal* 2003; **3(2)**: 109-132. Download pre-publication draft from http://www.imperial.ac.uk/nhli/r.newson/papers.htm