A hack for the multiple comparison problem

E-mail distributed on 27-07-2023

Dear all,

For today's topic, what better moment than the dead of summer to return to the problem of **multiple comparisons**, also known as the **multiple testing problem**. I have previously shared flowcharts for structured Bonferroni corrections in factorial designs. Today I want to discuss the topic more in-depth. You may be surprised to learn that, as other areas in statistics like setting sample size, correction for multiple comparison can be more art than science.

Family of tests

To recap, if the significance level of inferential testing is 0.05, then on average 5 in 100 independent tests will be significant as a false positive. While one can choose to adopt this false positive rate raw, statisticians have suggested instead to control the **family-wise rate of false positives**. Thus, instead of controlling the probability of a single test being a false positive, we would like to control the probability of *at least one false positive in a family of tests*. For short, a family of tests is typically equated to a family of *p*-values.

This approach requires defining what is a family of *p*-values in any given analysis. For example, consider a study with 3 between-subjects groups, and 4 moderately correlated outcome measures. One could define the following levels of testing, with a standard Bonferroni correction for each:

1. MANOVA of Group on all 4 outcomes simultaneously (1 *p*-value; no correction)
2. After a significant MANOVA, individual ANOVAs of Group on each of the 4 outcomes separately (4 *p*-values; multiply by 4)
3. After significant ANOVAs, 3 pairwise Group comparisons for the significant outcomes (3 *p*-values; multiply by 3)

While this seems defensible, each of these levels has its own controversies. Level 1 is often skipped by researchers, who start with uncorrected individual ANOVAs per outcome. As well, a MANOVA may not be possible when the outcomes have different distributions (e.g., reaction times and accuracies). At level 2, the Bonferroni correction will certainly be too strict, because it assumes independent *p*-values, when in fact the outcomes are correlated. At level 3, non-independence is also an issue, and one could also argue that the pairwise Group comparisons across all outcomes should be a single family of *p*-values, which makes between 0 and 12 *p*-values potentially. A completely different approach would be to consider the *p*-values of all levels of testing as a single family, thus potentially 17 total.

Independence

Assuming the researcher has defined their families of *p*-values, the question still remains how to correct. As mentioned, Bonferroni is too conservative due to assuming independent tests, which is almost never the case. Modifications have therefore been proposed, such as step-down methods (e.g., Bonferroni-Holm) that incrementally decrease the multiplication factor on each successive significant *p*-value. However this method still requires that the most significant *p*-value survives the strictest correction.

For certain special cases there are historical solutions known to be optimal, such as **Tukey's Honest Significant Difference (HSD)** for pairwise comparisons after a one-way between-subjects ANOVA. In many other cases, an optimal correction will not be available, especially for non-standard tests and models, or a mix of multiple methods (e.g., correlation analysis followed by multilevel regressions). Calculating the precise dependence between all these *p*-values would only be feasible by computational simulation, which is currently not widely used. Instead, a much simpler solution could be applied.

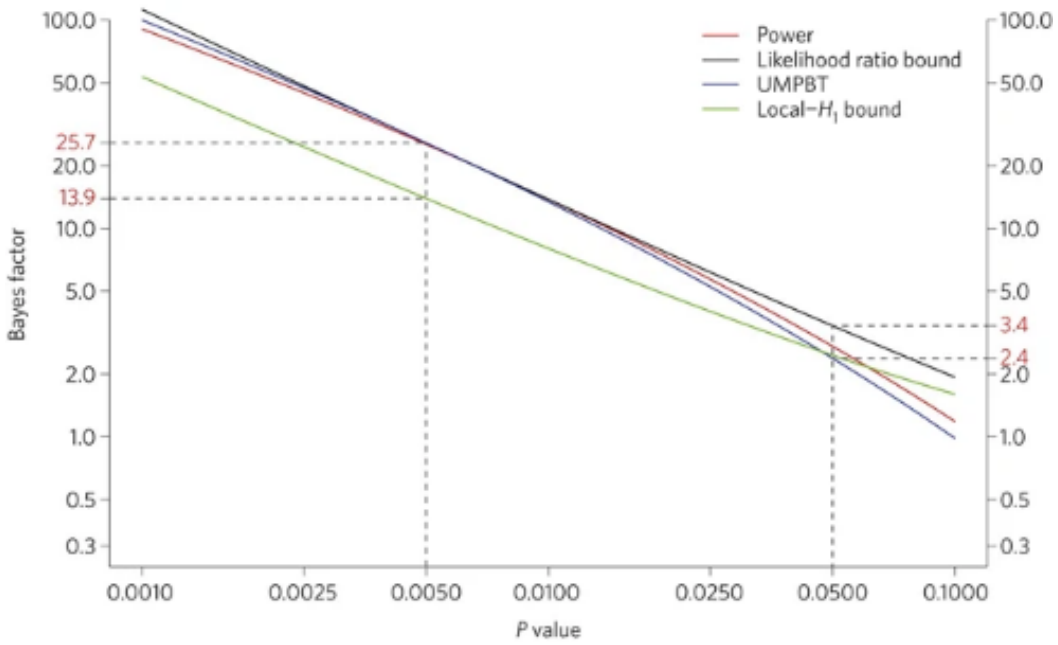

Figure 1. Relationship between the P.value and the Bayes factor (Benjamin et al., 2017)

Reduce your significance level!

Rather than making many subjective choices regarding *p*-value families and correction methods, I have been applying a general purpose solution to all my data analyses in recent years, which is to reduce the significance level to 0.005 or even 0.001. This approach was advocated/endorsed by a large number of researchers—including Eric Wagenmakers—in Nature (Benjamin et al., 2017). Their recommendation was formulated foremost for *new discoveries* in social sciences, on grounds that for

such *p*-values, values in the range [0.05,0.005] are on average associated with weak Bayes factors, and strong Bayes factors below 0.005 (see Figure 1).

Although they caution that the multiple testing problem may need to be considered on top of the reduced threshold, I believe it goes some way to providing a good correction for the multiple testing problem in any case. In fact, reducing the significance threshold comes with numerous advantages:

- It is equivalent to a Bonferroni correction for 10 independent *p*-values, which is still relatively conservative for some analyses.
- Larger effect sizes and Bayes factors can be expected under $p < 0.005$.
- No need to define families of *p*-values or which correction method to apply.
- *P*-values can be reported raw, only their significance interpretation changes.
- *P*-values between [0.05,0.005] can optionally be interpreted as suggestive or as trends (Benjamin et al., 2017), which is less risky than interpreting [0.1,0.05] as such.

In sum, if you experience difficulties deciding how to correct for multiple comparisons, then reducing your significance level is a very simple solution with solid justifications. I caution of course that other good practices still need to be respected in parallel (e.g., do not proceed with follow-up tests after a non-significant omnibus test). As well, specialized fields have developed their own standards for *p*-value correction (e.g., neuroscience, replication studies, confirmatory studies), which should be preferred over the solution presented here.

Note that, at the end of the day, choosing to correct *p*-values at all is a choice you make, not a strict requirement of inferential testing!

Best,
Ben

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79