Handout 08: Multiple Comparisons Problem

The results we have established for hypothesis testing concern a single, specific test. Things become more complex when looking at multiple tests all at once. Consider for example running hypothesis tests from 100 experiments with a significant level of 0.95. Even if the null hypothesis is in fact true in each experiement, if we cherry-pick the lowest p-value, we would expect on average to have 5 of the tests erroneously show up as significant. Let's see some approaches to addressing this.

Consider a sequence of m null hypothesis $H_1, \ldots, H_m$ and the corresponding p-values defined by $p_1, \ldots, p_m$. The probability of incorrectly rejecting at least one of the hypotheses is called the **family-wise error rate (FWER)**.[1] Assume that all of the null hypothesis are in fact true; what is the probability that we incorrectly reject one of them using a cut-off significance level of α? Well, we have that:

$$\text{FWER} = \mathbb{P}\left[\bigcup_{j=1}^{m} \{p_j \leq \alpha\}\right] \leq \sum_{j=1}^{m} \mathbb{P}\left[p_j \leq \alpha\right] = \sum_{j=1}^{m} \alpha = m \cdot \alpha.$$

So the probability of making at least one mistake could be as high as m times the confidence level. This means that if we actually want an error rate of α, we need to consider a cut-off value m times smaller than this error rate. Alternatively, we could adjust the p-values by multiplying them by the number of tests. That is, consider $p'_j = p_k \cdot m$. This gives:

$$\text{FWER} = \mathbb{P}\left[\bigcup_{j=1}^{m} \left\{p'_j \leq \alpha\right\}\right] \leq \sum_{j=1}^{m} \mathbb{P}\left[p_j \cdot m \leq \alpha\right] = \sum_{j=1}^{m} \frac{\alpha}{m} = \alpha.$$

Adjusting the p-values tends to be a better choice because they can be presented on their own terms without having to also specify the adjustment.

The above procedure shows that we will control the FWER if all of the hypotheses are true. What happens when only $k < m$ of the null hypotheses are in fact true? It turns out in this case the same equation above holds, but with the intersection taken over only the k true null hypothesis. Our correction will be valid, though overly conservative. So, the p-value adjustment provides valid control over the FWER.[2] The procedure outlined here is called the **Bonferroni correction**. There is a slightly stronger and equally valid version called the **Holm-Bonferroni correction** that we will see on today's worksheet. This procedure should be used to adjust p-values whenever you are running a large number of tests and want to be able to reliably trust any of the individual findings in isolation.

[1] There are other less conservative ways of controlling the error rates that arise when doing testing of many procedures. These control other quantities, usually something such as the **false discovery rate (FDR)**. We will not cover these directly this semester, but it would be a good project topic for the final project if you are interested.

[2] Note that by the error rate we refer only to what some sources call the 'Type I' error, falsely rejecting the null hypothesis. Controlling the the power of the test to detect false null-hypothesis depends on the specific structure of the data and sampling that we do not have access to in this setup.