Handout 06: 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 issue.

Consider a specific project that has has k different statistical tests resulting in a sequence of k different p-values. We assume that these p-values are created using a valid technique for each individual test.

Some of the issues with multiple comparisons are social.