derivation and explanation of the Bonferroni Correction.Rmd

Alan E. Berger March 17, 2020

This explanation / derivation of the Bonferroni correction follows the Wikipedia article (as of March 16, 2021): https://en.wikipedia.org/wiki/Bonferroni_correction with some commentary added. We are doing m statistical tests (and for the Bonferroni correction we do not need to assume they are independent). We are going to "declare" that, for any i, if the **p-value**, p_i for the i^{th} test is $\leq \tilde{\alpha}$ then that result is **statistically significant**. We want to choose $\tilde{\alpha}$ such that **the probability of one or more false positives among the m tests is $\leq \boldsymbol{\alpha}$** (a false positive here meaning the null hypothesis was true for test i but $p_i \leq \tilde{\alpha}$ so we incorrectly declared this test result to be statistically significant). This is what is meant by **controlling the family-wise error rate (FWER) at level $\boldsymbol{\alpha}$**. Following the Wikipedia article, we calculate an upper bound for the probability of getting one or more false positives in the following way. By the definition of the p-value, the probability of getting $p_i \leq \tilde{\alpha}$ just by random chance when the null hypothesis is true is $\tilde{\alpha}$ (assuming the conditions for validity of the statistical test being used are satisfied). If the null hypothesis is true for only some $m_0 < m$ of the statistical tests we would only need to control for false positives for that smaller number of tests, but unless we have a good estimate for m_0 we will take the conservative approach of using the number m of all the tests being done. We will see that the appropriate choice to have the probability that one or more false positives occur be $\leq \alpha$ is to define

$$\tilde{\alpha} = \alpha/m$$

This is the Bonferroni correction procedure. To verify that this is sufficient, assume the null hypothesis is true for all m tests. Let T_i be the event that the p-value for test i is $\leq \tilde{\alpha}$ (giving a false positive test result). Then whether or not the tests are independent of each other, the probability that one or more of the T_i occur is

$$P(\cup_{i=1}^{m} T_i) \leq \sum_{i=1}^{m} P(T_i) = m * \tilde{\alpha}$$

The inequality in the equation above is true by the properties of probability measures (see for example https://en.wikipedia.org/wiki/Boole%27s_inequality), and the equality in the equation above follows by the definition of a p-value. Hence if we set

$$\tilde{\alpha} = \alpha/m,$$

then the probability of one or more false positives is bounded by α, so we have controlled the FWER at level α. So one can either check, for each i, if $p_i \leq \alpha/m$ (and declare that test i is statistically significant if that is true) **OR** equivalently, one can multiply p_i by m and declare test i to be statistically significant if $m * p_i \leq \alpha$.

Hope this helps explain how and why Bonferroni correction works.