Post-hoc Tests
Multiple Testing Corrections

Pamela Wu
IBB2014

Post-hoc tests

- Latin for “after this”
- Why do it?
  - After a statistically significant ANOVA, how do you know which subgroups had the differing means?
- Pairwise T-tests not stringent
- Need to control for Type I error (too many false positives)

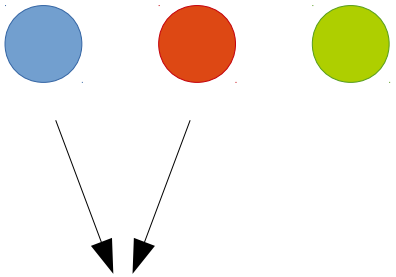
Multiple Testing Correction

- Post-hoc tests are essentially a multiple testing correction on a series of pairwise tests of significance
- Many correction algorithms to choose from, depending on your needs
- Today we will cover some famous ones:
  - Fisher's LSD and Tukey's HSD (pairwise only)
  - Bonferroni and Benjamini-Hochberg (general multiple testing)
- For a full list of other correction procedures, go to the Wikipedia page for post-hoc tests

Fisher's Least Significant Difference

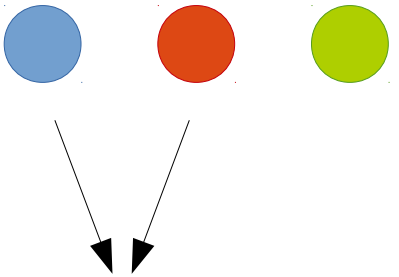
- Algorithm consists of performing all of the pairwise tests of significance, but instead of using the pooled standard deviation of the two subgroups being considered for each test, you use the pooled standard deviation for **all** subgroups for each test
- This makes **each** test operate with a much larger standard deviation, which makes this procedure a lot more stringent
- Only works for pairwise comparisons

Regular T-Test



s^2 = blue and orange

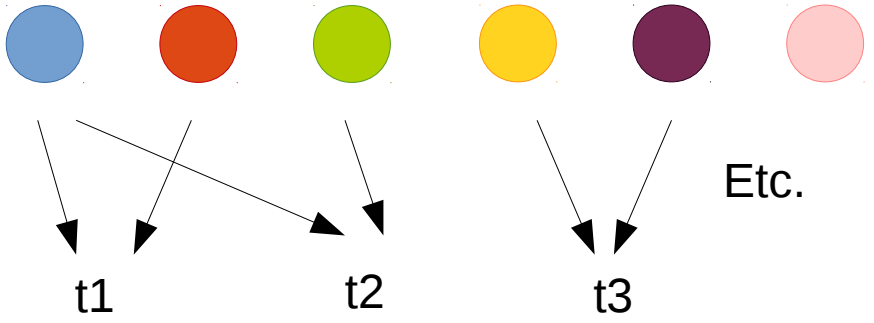
LSD T-Test



s^2 = blue and orange and green

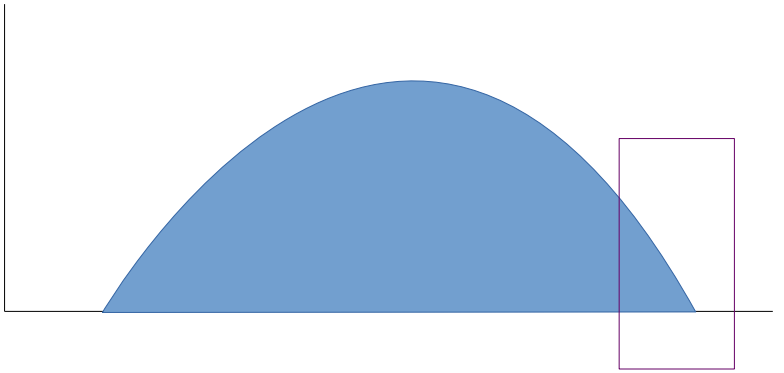
Tukey's Honest Sig. Diff.

- Tukey's HSD assumes that the T-statistics of many pairwise T-tests themselves form a normal distribution, and where each T-statistic falls into the overall distribution will give an adjusted p-value
- Unlike the LSD, the pooled standard deviation for each pairwise T-test is not pooled from all subgroups



Get t-scores for all pairwise comparisons

Distribution of all t-statistics



$$q = \frac{(\overline{y}_{max} - \overline{y}_{min})}{S\sqrt{2/n}}$$

Only these t-statistics Have a significant t-statistic

Multiple Testing Corrections

- For biomedical researchers, there are two main cases where we want to use multiple testing corrections
  - Any pairwise comparisons (like before)
  - Too many independent comparisons being positive (significant genes in microarray where 50,000/300,000 turn up as “significant”)
- The technical definition of a p-value for a T-test is “the probability that a difference this big or bigger can be observed by random chance.”
- What is the probability of observing at least one significant result by chance alone with cut-off at 0.05 and only 30 trials?
  - P(>= 1 significant) = 1 – P(no significant)^trials
  - P(>= 1 significant) = 1 – P(1 – 0.05)^30
  - P(>= 1 significant) = 0.785

Bonferroni Correction

- These next two don't care how you got the p-values because they consider them independently after the multiple testing
- The Bonferroni is the simplest and also generally the most stringent of all multiple testing corrections
- This is how it works:
  - If your initial cut-off was 0.05 and you have 30 comparisons, your new p-value cut-off (or alpha) is 0.05/30
  - q = p/m, where m is the number of hypotheses
  - P(>= 1 significant) = 1 – P(1 – 0.05/30)^30
  - P(>= 1 significant) = 0.0488
- But you subject yourself to a much higher Type II error (missed true positives)

Benjamini-Hochberg or False Discovery Rate

- Maybe instead of guaranteeing that the probability of having ***one* false discovery** is around 0.05 (or insert own cut-off), we can guarantee that **only around 0.05 of *all* discoveries is false**
- The P-values are first sorted and ranked. The smallest value gets rank 1, the second rank 2, and the largest gets rank N. Then, each P-value is multiplied by N and divided by its assigned rank to give the adjusted P-values.
- X = [0.001, 0.01, 0.05, 0.15, 0.2, 0.35, 0.7, 0.85, 0.9, 0.95]
- X(adj) = [0.01, 0.05, 0.16666667, 0.375, 0.4, 0.58333333, 1.0, 1.0625, 1.0, 0.95]