STAT 111, Week 8 - Multiple Comparisons

Leia Donaway — Swarthmore College

March 21, 2025

This week we are working from the k-sample Normal Week 8 guidelines from Professor Everson which are on Moodle. The fourth part is Multiple Comparisons. These notes make reference to Blitzstein & Hwang (Introduction to Probability) 4.4.3 and Rice 12.2.2.2, as well as resources made available on Moodle which include Phil's Wordle data and the Stein's Paradox in Statistics article (Effron & Morris, 1997).

A) With m independent tests, we can compute the exact 'family-wise' error rate (FWER). For example, imagine administering drug tests to $m = 10$ members of a sports team, none of whom have used drugs. If each test has false positive rate α, what is the probability of at least one false positive? What level α should be used to make the family-wise rate equal to 0.05? Show this is greater than 0.05/m.

Note: The significance level α for a single test indicates the probability we will reject H_0 when H_0 is true. With multiple tests, the probability of at least one false rejection grows with the number of tests.
With m independent tests, we can compute the exact "Family-wise" error rate (FWER).
Example: Sport team drug tests on $m = 10$ players. We know that in reality none of the players have taken drugs. Each test has a false positivity rate α, what is the probability of at least one false positive?

$$\mathbb{P}(\text{at least one false positive}) = 1 - \mathbb{P}(\text{no false positives}) = 1 - (1-\alpha)^m = 1 - (1-\alpha)^{10}$$

What level α should be used to make the family-wise rate equal to 0.05?

$$\text{Set } 1 - (1-\alpha)^{10} = 0.005$$

$$(1-\alpha)^{10} = 0.95$$

$$1 - \alpha = 0.95^{\cdot 1}$$

$$\alpha = 1 - 0.95^{\cdot 1} \approx 0.00512$$

Note: this result is greater than $\frac{0.05}{m} = \frac{0.05}{10} = 0.005$

B) For multiple testing without independence, we can use the Bonferroni adjustment. Prove the finite version of Boole's inequality:

$$\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \Sigma_{i=1}^{n} \mathbb{P}(A_i)$$

Explain how this leads to the Bonferroni adjustment for multiple tests and interval estimates that may or may not be independent.

Proof of Boole's inequality found in Blitzstein & Hwang 4.4.3 (page 152).
For any finite collection of events $A_1, A_2, \ldots, A_n$ and $I(A_i) = 1$ if A_i occurs and $I(A_i) = 1$ if A_i does not occur.

$$I(A_1 \cup A_2 \cup \cdots \cup A_n) \leq I(A_1) + \cdots + I(A_n)$$

We know we can assert this because if the LHS $= 0$, then we know $\Sigma I(A_i) = 0$ and if the LHS $= 1$ then we know $\Sigma I(A_i) \geq 1$ is at least 1.
Taking the expectation of both sides, noting that $\mathbb{E}[I(A)] = \mathbb{P}(A)$, we get

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \mathbb{P}(A_1) + \cdots + \mathbb{P}(A_n)$$

$$\implies \mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \Sigma_{i=1}^{n} \mathbb{P}(A_i)$$

With this slick proof of Boole's inequality, we are lead to the Bonferroni Adjustment for multiple tests and interval estimates. There is a brief section about this in Rice 12.2.2.2 which summarizes, but the Wikipedia page is more in-depth if you're curious.
Boole's Inequality provides us an upper bound on the probability of at least one false positive across multiple tests in order to counteract the "multiple comparisons problem". For m hypothesis tests of significance level α_i, take A_i to be the event that H_0 is rejected for the i-th test when H_0 is true ($A_i =$ the event of a false positive). Apply Boole's; if all tests are conducted at the same significance level α, then

$$\mathbb{P}(\bigcup_{i=1}^{m} A_i) \leq m\alpha$$

$$\text{Set } m\alpha \leq \alpha_{FWER}$$

so that the total family-wise error rate doesn't exceed some desired α_{FWER} rate. Solving for α, we get the Bonferroni-adjusted significance level $\alpha = \frac{\alpha_{FWER}}{m}$ to guarantee a type I Error rate of at most α.
Bonferroni controls the FWER by setting a stricter per-test significance level. This is good for when multiple tests are not independent (or we are unsure).

C) For k independent samples, there are $\binom{k}{2}$ possible pairwise tests of equal means. Explain why these do not represent independent tests. Use my Wordle data as an example, with $k = 5$ start words.

As an example, let's take Phil's Wordle data with $k = 5$ starting words. As we can imagine, there is a bit of luck involved over the 10 days each starting word is used, like one day the answer is found in 1 guess, not indicating strength of the word. Comparisons were performed for each possible word pair $\binom{5}{2} = 10$, and two were found to be significantly different at an $\alpha = .05$ (simple t-test).
AUDIO v CRANE and SAUTE v CRANE have p-values $p = 0.0066 < \alpha$. We are unsurprised to see CRANE in both significant results, since it is the same data, meaning these don't represent independent tests (we compare to the same data with each comparison to any one word).

D) For an ANOVA F test, describe the estimates made for the k unknown means when we reject and fail to reject Ho. In 1962 Charles Stein showed that, for estimating $k \geq 3$ mean parameters, using the k sample averages is inadmissible. Explain what this means and why it is called "Stein's Paradox in Statistics". Describe the improved James-Stein "shrinkage" estimates.

For an ANOVA F test for difference among k means
$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$, H_A : At least one mean differs
When we fail to reject H_0, we conclude that the means aren't significantly different and assume the overall average $\bar{Y} = \frac{1}{k}\Sigma_{i=1}^{k}\bar{Y}_i$ for all k means.
When we do reject H_0, we can only conclude that at least one group's mean differs from the others , leaving us with group $\bar{Y}_i$ mean as an estimate. However, in 1962, Charles Stein showed that, for estimating $k \geq 3$ mean parameters, using the k sample averages when we reject H_0 is "inadmissible." By "inadmissible," he meant that there is another estimator that performs better (has lower mean square error). The better estimate involves shrinking the estimates toward a common value. Using individual sample means is unbiased, but they are high in variance. By borrowing information across all groups, we can reduce the total estimation error. This leads to the James-Stein "shrinkage" estimate.

$$\hat{\theta}_i = \hat{B}\bar{Y} + (1 - \hat{B})Y_i$$

$$\hat{B} = \frac{(k-3)V}{\Sigma(Y_i - \bar{Y})^2}, \text{for } k \geq 3, \text{with V a common known variance among groups}$$

Read more about this in the "Stein's Paradox" article on Moodle. An example used in the article was estimating batting averages partway through the season. We know that early in the season, some players may be hitting amazing (or perhaps subpar) average, but by the end of the season, with more data collected, we expect the means to converge or shrink, rather than to remain so varied. This is called "Stein;s Paradox in Statistics." To explain this consider that if we added a group mean totally unrelated (in the article, they use proportion of foreign cars passing the field in a fixed amount of time). This would add a point to the

graph, causing other estimates to be slightly effective. It is not an actual paradox. It is called one because the estimator allows for unrelated data to change the related estimates, but this is more of a user error, because we shouldn't include what is essentially noise.
However, the James-Stein shrinkage estimates are also considered "inadmissible," because a better (improved) version is know, where $\hat{B}^* = \min(1, \hat{B})$ so we can ensure the shrinkage factor does not over-correct. We do not want the shrinkage to be greater than 1, because then estimates start moving in the other direction.