MULTIPLE COMPARISON

Multiple testing (multiple comparisons): what's the problem?

If we perform n hypothesis tests, each at level α, then the probability that at least one of them will give a false positive result (a Type-I error) is greater than α even when all involved null hypotheses are true.

In the context of comparing group means, with more than two groups, when all group means are equal, the chance of making at least one type I error is greater than α.

Exercise. Suppose you toss a biased coin n times, the probability of getting a tail is 0.05, and all tosses are independent. What is the probability of getting at least one tail out of n tosses?

If we perform n hypothesis tests, each at level alpha, then the probability that at least one of them will give a false positive result (a Type-I error) is greater than alpha even when all involved null hypotheses are true.

In the context of comparing group means, with more than two groups, when all group means are equal, the chance of making at least one type I error is greater than alpha.

To think about the effect of multiple testing. We can do this exercise. Suppose you toss a biased coin n times, the probability of getting a tail is 0.05 and all tosses are independent. What is the probability of getting at least one tail out of n tosses?

The answer to the exercise,

The answer to the exercise is $1 - (0.95)^n$, since the probability of getting all heads is $(0.95)^n$.

If we were to perform n independent hypothesis test each with an alpha level of 0.05, then the probability of making at least one type I error can be as high as $1 - (0.95)^n$ even when all n null hypotheses are true. Below we list the values of $1 - (0.95)^n$ for some n-values:

```
##         n         p
## [1,]    1 0.0500000
## [2,]    2 0.0975000
## [3,]    3 0.1426250
## [4,]    4 0.1854938
## [5,]    5 0.2262191
## [6,]   10 0.4012631
## [7,]  100 0.9940795
## [8,] 1000 1.0000000
```

The answer to the exercise is (1 - (0.95)^n), since the probability of getting all heads is ((0.95)^n).

If we were to perform n independent hypothesis test each with an alpha level of 0.05, then the probability of making at least one type I error can be as high as (1 - (0.95)^n) even when all null hypotheses are true.

Below we list the values of (1 - (0.95)^n) for some values of n.

Possible solutions to the multiple-testing or multiple-comparison problem

1. Make no adjustments, but understand the consequence and implications.
2. Use a more stringent α level. For example, the Bonferonni correction uses a significant level of α/n in each individual test to control the probability of making at least one type I error, where n is the number of tests/comparisons.
3. Allow some number of false positives: control false discovery rate (FDR, q-value).FDR is the expected proportion of false positive results among all positive results.
4. Use a "global" test statistic that summarizes the entire "experiment"—consisting of all comparisons. The Tukey method is one such method. It looks at the distribution of the difference between the largest group mean and the smallest group mean.

In this class, we will briefly discuss approaches 1 and 4.

In practice, the FDR approach (approach 3) is widely used when we need to repeat the same test a large number of times.

- For example, in gene expression analysis, we often need to perform tens of thousands tests on a single data set.

Here I listed four possible solutions to the multiple-testing or multiple-comparison problem:

1. Make no adjustments, but understand the consequence and implications.

2. Use a more stringent alpha level. For example, the Bonferonni correction uses a significant level of (alpha over n) in each individual test to control the probability of making at least one type I error where n is the number of tests/comparisons.

3. Allow some number of false positives: e.g., control false discovery rate (FDR, q-value). FDR is the expected proportion of false positive results among all positive results.

4. Use a "global" test statistic that summarizes the entire "experiment"—consisting of all comparisons. The Tukey method is one such method. It looks at the distribution of the difference between the largest group mean and the smallest group mean.

In this class, we will briefly discuss approaches 1 and 4.

In practice, the FDR approach, approach 3, is widely used when we need to repeat the same test a large number of times.

For example, in gene expression analysis, we often need to perform tens of thousands tests on a single data set.

Inflated per-experiment error-rate in all pairwise comparisons

The more hypothesis tests we do, the greater the chance of obtaining at least one significant result, even if the null hypothesis is true.

Suppose we do all pairwise comparisons among t treatment means and define

Per-comparison (or 'comparison-wise') error rate:

$$\alpha_C = \text{Type I error rate in a single comparison} = Pr(\text{reject } H_0 | H_0 \text{ true})$$

Per-experiment (or 'experiment-wise') error rate:

$$\alpha_E = Pr(\text{at least one significant test result| all } H_0 \text{ true})$$

If $\mu_1 = \mu_2 = \cdots = \mu_t$, and we do all pairwise comparisons using $\alpha_C = 0.05$, then the per-experiment error rates are as in the following table.

The more hypothesis tests we do, the greater the chance of obtaining at least one significant result, even if the null hypothesis is true.

For example, suppose we do all pairwise comparisons among t treatment means and define

Per-comparison (or 'comparison-wise') error rate is the Type I error rate in a single comparison.

Per-experiment (or 'experiment-wise') error rate is the probability that there is at least one type I error.

If all mean parameters are actually equal, and we do all pairwise comparisons using (alpha C = 0.05), then the per-experiment error rates are as in the following table.

t	No. of tests $[\ t(t-1)/2\ ]$	Prob. at least one significant $(\ \alpha_E\ )$
2	1	0.05
3	3	0.11
5	10	0.28
10	45	0.64

Per-experiment error rate for all-pairwise comparisons

Possible solutions

To avoid inflated per-experiment error rates, many statisticians suggest reducing the per-comparison error rate when doing multiple comparisons. One of many possible approaches is Tukey's method, which I'll compare to a method (the least significant difference) that makes no adjustment of error rates. Other multiple-comparison methods are discussed by Kuehl, pp. 94-115.

To avoid inflated per-experiment error rates, many statisticians suggest reducing the per-comparison error rate when doing multiple comparisons. One of many possible approaches is Tukey's method, which I'll compare to a method (the least significant difference) that makes no adjustment of error rates. Other multiple-comparison methods are discussed by Kuehl, pp. 94-115.

Least significance difference (LSD)

When testing $H_0: \mu_1 = \mu_2 = \cdots = \mu_t$, the least significance difference (LSD) method can be used **if the ANOVA F-test shows significance.**

The LSD method makes no additional adjustment for multiple comparisons.

The LSD corresponds to a two-sample t-test based on the MSE:

$$LSD_{ij} = t_{\alpha/2,N-t} \cdot SE(\overline{y}_{i\cdot} - \overline{y}_{j\cdot}) = t_{\alpha/2,N-t} \cdot \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

Reject $H_0: \mu_i = \mu_j$ if $|\overline{y}_{i\cdot} - \overline{y}_{j\cdot}| > LSD_{ij}$.

Note: an ordinary two-sample t-test uses only data from two groups to estimate the variance parameter. In the LSD method, the population variance is estimated by MSE which combines the variability information from all groups.

When testing the null hypothesis that all means are the same, the least significance difference (LSD) method can be used **if the ANOVA F-test shows significance.**

The LSD method makes no additional adjustment for multiple comparisons.

The LSD corresponds to a two-sample t-test based on the MSE.

Note that an ordinary two-sample t-test uses only information from two groups to estimate the variance parameter. In the LSD method, the population variance is estimated by MSE which combines the variability information from all groups.

Tukey's Honestly Significant Difference (HSD)

Tukey's HSD method adjusts for all pairwise comparisons: This method gives a Type I error rate of α for all pairwise comparisons on a per-experiment basis.

The honestly significant difference for comparing μ_i and μ_j is

$$HSD_{ij} = \frac{q_{\alpha,t,N-t}}{\sqrt{2}} \cdot \sqrt{MSE\left(\frac{1}{r_i} + \frac{1}{r_j}\right)}$$

where q is a quantile from **the Studentized range distribution**, obtained in R as: $\texttt{qtukey}(1 - \texttt{alpha}, \texttt{t}, \texttt{N} - \texttt{t})$. (Replace $t_{\alpha/2,N-t}$ in LSD_{ij} by $\frac{q_{\alpha,t,N-t}}{\sqrt{2}}$.)

If the two sample means are separated by more than the HSD, we conclude that they are statistically different.

A $100(1-\alpha)\%$ **simultaneous confidence interval** for all $\mu_i - \mu_j$ is given by

$$\overline{y}_{i\cdot} - \overline{y}_{j\cdot} \pm HSD_{ij}.$$

Tukey's Honestly Significant Difference or HSD method adjust for all pairwise comparisons. This method gives a Type I error rate of alpha for all pairwise comparisons on a per-experiment basis.

The honestly significant difference for comparing (mu i) and (mu j) is given by this formula. Comparing this formula to the LSD, we see that the only difference is that the t critical value is replaced by (q alpha, t, N-t) over square root of 2.

Here the q is a quantile from the Studentized range distribution, and can be obtained in R using the [qtukey] function. The range is the difference between the largest and the smallest group means.

If the two sample means are separated by more than the HSD, we conclude that they are statistically different.

A (100(1 - alpha) percent) simultaneous confidence interval for (mu i - mu j) is given by the corresponding group sample mean difference plus/minus the corresponding HSD.

Tukey's HSD (continued)

The critical value used in the HSD is based on the distribution of the range: The range is the difference between the largest and the smallest group means.

It can be shown that, if $t > 2$, then

$$\frac{q_{\alpha,t,N-t}}{\sqrt{2}} > t_{\alpha/2,N-t},$$

$t_{\alpha/2,N-t}$ is the multiplier we would use if we were not adjusting for multiple comparisons (i.e., the LSD method).

Therefore, tests based on Tukey's HSD are more stringent (give less rejections) than tests based on LSD.

The critical value used in the HSD is based on the distribution of the range: The range is the difference between the largest and the smallest group means.

It can be shown that, if when there is more than two groups, (q alpha, t, N-t) over (square root of 2) is greater than (t alpha/2, N-t).

In other words, tests based on Tukey's HSD is more stringent than tests based on LSD.

Example GPA/residency data

Example. GPA/residency data. Here, $r_1 = r_2 = r_3 = 30$. So:

$$\text{LSD} = t_{\alpha/2,\,N-t} \cdot \sqrt{\text{MSE}\left(\frac{1}{r}+\frac{1}{r}\right)} = t_{0.025,\,87} \cdot \sqrt{0.489\left(\frac{1}{30}+\frac{1}{30}\right)}$$
$$= 1.988 \cdot 0.1806 = 0.359$$

$$\text{HSD} = \frac{q_{\alpha,\,t,\,N-t}}{\sqrt{2}} \cdot \sqrt{\text{MSE}\left(\frac{1}{r}+\frac{1}{r}\right)} = \frac{q_{0.05\,3,\,87}}{\sqrt{2}} \cdot \sqrt{0.489\left(\frac{1}{30}+\frac{1}{30}\right)}$$
$$= (3.372/\sqrt{2}) \cdot 0.1806 = 2.38 \cdot 0.1806 = 0.431\,,$$

In this example, we compute the LSD and HSD for the GPA/residency data.

We see that Tukey's HSD is greater than the LSD.

This implies that when testing pair-wise group mean differences, tests based on Tukey's HSD will be more stringent and have less rejections than the tests based on LSD.

Example GPA/residency data (continued)

where I obtained the needed quantiles in R using `qt(0.975, 87)` and `qtukey(0.95, 3, 87)`.

95% confidence intervals for $\mu_3 - \mu_2$:

$$(\bar{y}_{3\bullet} - \bar{y}_{2\bullet}) \pm \text{LSD} = (3.07 - 2.63) \pm 0.359 = (0.08, 0.80)$$

$$(\bar{y}_{3\bullet} - \bar{y}_{2\bullet}) \pm \text{HSD} = (3.07 - 2.63) \pm 0.431 = (0.01, 0.87)\,. \quad \square$$

Using Tukey's HSD, the resulting confidence interval will be wider.

Discussion: Issues with multiple comparisons

1. What is the "experiment" (the set of comparisons)?

2. How do we choose the per-experiment error rate, and, consequently, the severity of the penalty for multiple comparisons?

The uncertainties and ambiguities of applying multiple-comparison adjustments have led some statisticians (including this one) to avoid them when possible.

Kuehl says (on p. 116), "Choose a test that is consistent with your philosophy" - whatever that means!

What is the "experiment" (the set of comparisons)?

How do we choose the per-experiment error rate, and, consequently, the severity of the penalty for multiple comparisons?

The uncertainties and ambiguities of applying multiple-comparison adjustments have led some statisticians (including this one) to avoid them when possible.

Kuehl says (on p. 116), "Choose a test that is consistent with your philosophy" - whatever that means!

Summary

1. We briefly discussed challenges in multiple-testing adjustments.
2. For comparing multiple groups means, we talked about the Fisher's least significance difference (LSD) method and Tukey's Honestly Significance Difference (HSD) method.

Summary

We briefly discussed challenges in multiple-testing adjustments.

For comparing multiple groups means, we talked about the Fisher's least significance difference (LSD) method and Tukey's Honestly Significance Difference (HSD) method.