Notes on multiple comparisons in statistical hypothesis testing

Carolyn Johnston

February 20, 2020

Introduction

Suppose that we have a Completely Randomized Design (CRD) experiment set up, in which we are studying $p = 4$ different treatment groups with means $\mu_{i}{}_{i=1}^{4}$, each group having r replicates. We would like to test whether the differences between each pair are significant at the $\alpha = 0.05$ level. Our first, naive thought might be to do $\binom{4}{2} = 6$ different hypothesis tests, each at significance level $\alpha = 0.05$, in order to figure out which pairs of means differ. In this case, we are testing 6 individual null hypotheses H_{0c}, for pairs of comparisons $c = 1, ..., 6$, against 6 individual alternative hypotheses H_{ac} at significance level 0.05.

The problem is that, if we follow this procedure, the probability of our falsely concluding that at least one pair of means are different is much higher than $\alpha = 0.05$. If we do 6 pairwise hypothesis tests groups, and each has a type I error rate of .05, then assuming all the outcomes are independent, the probability that there will be at least one type I error in the group of 6 comparisons is actually $1 - 0.95^6 = .26$, much higher than we would like.

Having a handle on Type I error rates, and understanding the cost to the statistical power of experiments, is at the heart of statistical practice and experimental design. Standard statistical practice is to fix the Type I error at some value α and build a design around that criterion. This writeup is about practical approaches in statistical practice for controlling Type I error; i.e., bounding various Type I error metrics below some value in cases when an experiment involves multiple hypothesis tests.

Different multiple-comparison hypothesis testing situations require different definitions for Type I error, some more stringent than others. The more conservative Type I error metrics (such as EERP and MEER) are controlled only at a severe cost to the statistical power of an experiment.

In the next section, I discuss some important error metrics used in measuring Type I error for multiple comparisons. In the section after that, I discuss some Type I error control approaches in experimental practice, and discuss the Type I error metrics that they control (or don't).

Type I error metrics

Suppose we are testing a null hypothesis H_0 vs. a defined alternative hypothesis H_a. H_0 might look something like "the two group means are equal", and H_a might look like "the two group means are not equal" (this is a two-sided alternative), or "group mean 1 is larger than group mean 2" (a one-sided alternative).

1. Probability of Type I error ("false positive"): $p(H_0$ rejected$|H_0$ true). We want this to be small.
2. Probability of Type II error ("false negative"): $p(H_0$ not rejected$|H_a$ true).
3. Power: $p(H_0$ rejected$|H_a$ true)=1-(probability of type II error). In the design process, we want to make this as large as possible.

In classical hypothesis testing, standard practice is to control Type I error at a known level; i.e., we restrict the probability of a type I error to be less than 0.05. Generally, controlling Type I errors comes at the expense of statistical power; we fail to reject the null hypothesis more often, at the expense of failing to detect the alternative hypothesis when it is true.

Comparisonwise error rate, CER

As defined in references [1] and [2], this is the ratio of type I error rates to the total number of comparisons made, C.

$$CER = \frac{\Sigma_{c=1}^{C} I\{H_{0c}\text{was rejected when } H_{0c}\text{was true}\}}{\text{total number of comparisons } C},$$

where I is the indicator function:

$$I(\text{statement}) = \begin{cases} 1 & \text{statement is true} \\ 0 & \text{statement is false} \end{cases}.$$

In the scenario with 4 treatment means, if all 6 possible comparisons are made, and exactly one of them results in a type I error, then the CER would be 1/6.

Experimentwise error rate under a complete null hypothesis, EERC

"Complete H_0" refers to the situation in which all treatment means are equal: $\mu_1 = .. = \mu_4$. If we run E independent experiments on the same population, for which complete H_0 holds, then the EERC is defined to be the number of times out of E in which we incorrectly rejected some H_{0c} *at least once:*

$$EERC = \frac{\Sigma_{e=1}^{E} I\{H_{0c} \text{ falsely rejected for any } c \text{ during experiment } e\}}{\text{total number of experiments } E},$$

given that H_{0c} is true for all comparisons c. For experimentwise error, the 6 comparisons in the example above are regarded as a single experiment*;* and a "type I error" occurs for the overall experiment if even one of the rejections of H_{0c} for some comparison c is false.

Experimentwise error rate under a partial null hypothesis, EERP

"Partial H_0" refers to any situation in which at least one pair of treatment means are equal, but not all pairs are required to be equal. For example, the case $\mu_1 = \mu_2 = \mu_3 = 0$ and $\mu_4 > 0$ is a partial null hypothesis, and so is $\mu_1 = \mu_2 < \mu_3 = \mu_4$.

If we run E independent experiments on such a population, then the EERP is defined to be the number of times out of E in which we incorrectly rejected H_0 *at least once:*

$$EERP = \frac{\Sigma_{e=1}^{E} I\{H_{0c} \text{ falsely rejected for any } c \text{ during experiment } e\}}{\text{total number of experiments } E},$$

given that H_{0c} is true for at least one comparison c.

Maximum experimentwise error rate, MEER

In reference [3], MEER is defined to be the probability of making at least one Type I error if *any* complete or partial null hypothesis is true.

$$MEER = \sup_{\text{all complete or partial hypotheses } H_0} \{EERP\}.$$

In other words, pick your worst scenario: how likely are you to incorrectly reject H_0 at least once in *that* case? For example, suppose $\mu_1 < \mu_2 < \mu_3 = \mu_4$. Suppose the true differences $|\mu_i - \mu_j|$, for $i, j = 1, 2, 3$, are close to the noise threshold, so that they are fairly close to the sample value $|\hat{\mu}_3 - \hat{\mu}_4|$ for the pair of means that are equal. This is a case in which it would be extremely easy to reject that one true null hypothesis in order to detect the differences between the other means. The MEER is at least as large as the rate at which you falsely reject the null, using your protocol in the most unfavorable situation you can imagine. As a result, unsurprisingly, controlling MEER comes at great expense to statistical power. Note that we always have MEER $>$ EERC and MEER $>$ EERP, by definition.

False Discovery Rate, FDR

A 'discovery' is a term for the rejection of H_0 in a given comparison:

$$FDR = \frac{\text{number of false rejections of } H_{0c}, \text{ for } c = 1, ..., C}{\text{number of rejections of } H_{0c}, \text{ for } c = 1, ..., C},$$

i.e., FDR is defined to be the ratio of false discoveries to total discoveries within an experiment.

Methods for controlling Type I error metrics

"Controlling an error" means that we are following a procedure that will force the error metric to be bounded at less than or equal to some value. For example,

"controlling EERC" means that we are following a hypothesis testing protocol that will force EERC to be less than or equal to some predetermined value α. Below I discuss some testing approaches for controlling the Type I error metrics listed in the previous section.

Least significant difference (LSD)

LSD is the protocol we would use naively: it involves using a pairwise t-test to test each hypothesis H_{0c} in a set of C comparisons, and rejecting at level α if $|t| > t_{\alpha/2,dfe}$.

This protocol controls only CER. Suppose we do a number E of experiments, each involving $C = 6$ comparisons of 4 treatment means, and count up the number of comparisons in which we falsely reject the null; by construction, this ratio is α.

F-protected LSD (FLSD)

F-protected LSD is almost as easy as the LSD protocol. Assume that all treatment means are equal in our experiment. Then we add one step: before we begin testing means pairwise for equality, we first do an overall F-test of the null hypothesis that all treatment means are equal, at level α. By construction, the overall F-test will falsely reject H_0 only $100 \cdot \alpha\%$ of the time. If we do not reject the overall F-test H_0, then we do not go on to do comparison-wise tests. If we do reject the overall F-test H_0, then we go on to do comparison-wise tests of H_{0c} using the LSD protocol; if these null hypotheses are all accepted, then we accept the experiment-level H_0.

Here is a proof that this protocol controls EERC (experimental Type I error in the case where all treatment means are actually equal). Let E_{H_0} be the event that all the group treatment means are equal, let R_0 be the event that the experiment-level H_0 is falsely rejected, and let OF_0 be the event that the overall F-test is falsely rejected. Then we have:

$$EERC = P(R_0|E_{H_0})$$
$$= P(R_0|OF_0, E_{H_0}) \cdot P(OF_0|E_{H_0}) + P(R_0|\neg OF_0, E_{H_0}) \cdot P(\neg OF_0|E_{H_0}).$$

$P(R_0|OF_0, E_{H_0})$ is the probability that the experiment-level H_0 (all treatment means are equal) is rejected, given that the overall F-test is falsely rejected. This is less than 1, since it is possible that even if the overall F-test is falsely rejected, each of the pairwise comparisons $c = 1, ..., C$ will result in accepting the null hypotheses H_{0c}.

$P(OF_0|E_{H_0})$ is the probability that the overall F-test will be falsely rejected, given that all treatment means are equal. The F-test at the beginning of the FLSD protocol ensures that this probability is capped at α.

$P(R_0|\neg OF_0, E_{H_0})$ is the probability of the event that the experiment-level H_0 is falsely rejected, given that the overall F-test is not falsely rejected, and

the treatment means are all equal. This probability is actually 0, since we stop the protocol if we do not reject the overall F-test null hypothesis.

Putting all these pieces together, we have:

$$EERC = P(R_0|E_{H_0}) < 1 \cdot \alpha + 0 = \alpha.$$

This protocol controls EERC at level α, but not EERP. To see why, recall that EERP gives the experimentwise error rate calculated over experiments where *some,* but perhaps not all, treatment means are equal. If not all treatment means are equal, then the overall F-test will (should!) reject H_0 at an uncontrolled rate. Since then we continue to do the pairwise hypothesis tests part of the protocol, we once again rely on the LSD method to determine pairwise significant differences, and the LSD method does not control experimentwise Type I error.

Tukey's Honestly Significant Difference (HSD)

This protocol is designed to be conservative; that is, it rejects H_{0c} less frequently than many other multiple comparison methods. But it controls MEER, and it has better power than many other testing procedures that control MEER.

For each pair of experimental means, we calculate differences $\hat{D}_{i,j} = \bar{y}_{i\cdot} - \bar{y}_{j\cdot}$ and standard errors $\text{se}(\hat{D}_{ij}) = \sqrt{MSE(\frac{1}{r_i} + \frac{1}{r_j})}$. Then we calculate the statistic $q_{ij} = \hat{D}_{ij}/se(\hat{D}_{ij})$. These look like t-test statistics, but instead of using a t-distribution for the test, we will use a different distribution defined below.

Suppose we take a size N i.i.d. sample $\{y_n^k\}_{n=1}^N$ from each of K populations with the same distribution $N(\mu, \sigma^2)$, and calculate their K sample means $\hat{\mu}_k = \bar{y}_{\cdot}^k$. Let $\bar{y}_{min}$ and $\bar{y}_{max}$ be the minimum and maximum from the set of K sample means. Then, under the null hypothesis H_0 that all population means are equal to μ, the *studentized range distribution* (with group parameter $k = K$ and degrees of freedom $\nu = (N-1)K$) is the distribution of the statistic

$$q = \frac{\bar{y}_{max} - \bar{y}_{min}}{s/\sqrt{N}},$$

where s is the pooled sample standard deviation from these samples (i.e., throw all the samples y_n^k into a common basket and calculate their sample variance).

To apply the test, reject H_0 if $q_{ij} > q_{(K,\nu),\alpha}$ for significance level α.

It is a theorem from mathematical statistics, originally due to Tukey in the case where all group sizes are equal [4], that this protocol controls the MEER at level α.

Bonferroni's method

Bonferroni is the first multiple-comparison Type I error control method that most people learn about in their mathematical statistics classes.

Given C individual tests of null hypotheses H_{0c} to be made, the Bonferroni correction rejects the null hypothesis for a p-value p_c if $p_c < \frac{\alpha}{C}$ (rather than α

as in the LSD case). In this protocol, the experimental error rates (EERP or EERC) are controlled at level α, since

$$EER = P\cup_{i=1}^{C}(p_c \leq \alpha/C) \leq \Sigma_{i=1}^{C} P(p_c \leq \alpha/C) = C \cdot \frac{\alpha}{C} = \alpha.$$

The Bonferroni method is free of distribution assumptions and controls MEER at α, but it is extremely conservative and causes a significant reduction in power. Tukey's HSD is better for the 'all-pairs comparison' problem.

Student-Newman-Kuels (SNK) method

The SNK method tests for equivalent sets of means within a set of T experimental means, using Tukey's HSD method at each stage.

Suppose we are testing 6 treatment groups, with r replicates within each group, and we have obtained experimental means with $\hat{\mu}_1 < \hat{\mu}_2 < ... < \hat{\mu}_5 < \hat{\mu}_6$.

In the first step of the method, we test the null hypothesis that the largest subgroup of ordered sample means, the full set of $k = 6$, are equivalent, using Tukey's HSD method. That is, we compare the statistic $q_{16} = \hat{D}_{16}/\text{se}(\hat{D}_{16})$, where $\hat{D}_{16} = |\hat{\mu}_6 - \hat{\mu}_1|$, with the critical value $q_6^* = q_{(6,6(r-1)),\alpha}$. If $q_{16} > q_6*$, then we declare the difference between the least and greatest group mean to be significant. Otherwise, if we cannot reject H_0, we declare all means to be equivalent, and end the SNK test.

If we do not reject H_0 at this level, then we continue to the next step of the SNK test. The full group of experimental means contains two ordered subgroups of size $K = 5$: $\hat{\mu}_1 < \hat{\mu}_2 < ... < \hat{\mu}_5$ and $\hat{\mu}_2 < \hat{\mu}_2 < ... < \hat{\mu}_6$, and in this step we apply the Tukey's HSD method to each of these two subgroups of 5.

For example, first we would compare the statistic $q_{15} = \hat{D}_{15}/\text{se}(\hat{D}_{15})$ to $q_5^* = q_{(5,5(r-1)),\alpha}$ (note that q_5^* is a different critical value (it is smaller) than the one used for the full group of 6 treatment means). If $q_{15} > q_5^*$, then we reject the null hypothesis that all treatment means in that subgroup of 5 are equal. If we do not reject H_0, we conclude that the treatment means in that set form a set of 5 equivalent treatment means, and we will not test any smaller ordered subgroups within that set. We test the subgroup $\hat{\mu}_2 < \hat{\mu}_2 < ... < \hat{\mu}_6$ similarly.

Finally, we continue testing smaller ordered subgroups of size $k = 4, 3, 2$, from within any larger subgroup for which the HSD test H_0 is rejected, until we have completed all tests, and have identified all subsets of equivalent means of size $k = 6, 5, 4, 3, 2$ within the full set of experiment means.

You can think of the SNK process as forming a tree of tests, in which every level contains 'leaves' representing ordered subsets of size k of the original K treatment means. At the top level, $k = 6$, we test the full set of 6 means using a Tukey HSD test. If H_0 is not rejected, we stop the testing at level 6; but if H_0 is rejected, then we add two branches coming from the top level to two nodes representing two ordered subsets of size 5. Each ordered subset of size 5 is tested using a Tukey (HSD) test; and again, if H_0 is rejected for either of these subsets, then we add two branches coming from that test to its two ordered subsets of size 4. The tree continues to grow until all its nodes terminate, either

because H_0 was not rejected for that node, or because the node contains an ordered subset of size 1 that is not a part of any other subgroup for which H_0 was not rejected.

Figure 1 shows a diagram of such a tree for a $K = 6$ example. The group sample means are $(11, 13, 15, 16, 17, 21)$. At the top, we perform an HSD test for $K = 6$, and H_0 is rejected; so the top ordered list is subdivided into two $K = 5$ tests, each of which rejects H_0 and is subdivided again into two $K = 4$ lists each. Note that at level 4, the list $(13, 15, 16, 17)$ is duplicated on both sides of the tree, but this test would actually only be performed once (on the left hand side of the tree). H_0 is not rejected for the list $(13, 15, 16, 17)$, and so the tree would terminate at that node.

In Figure 1, terminal nodes are denoted by purple boxes. Dashed boxes indicate tests that would not actually be performed, either because they were performed somewhere else in the tree, or because the ordered lists are subsets of a larger list for which H_0 was not rejected at a higher level. For example, H_0 is not rejected at level $k = 4$ for the list $(13, 15, 16, 17)$, and so we do not actually test $(13, 15, 16)$ at level 3, or $(13, 15)$ at level 2. In the end, the SNK tree defines a partial order on the means defined by $\mu_{11} < \mu_{15}$ and $\mu_{17} < \mu_{21}$. Put another way, it defines three semi-equivalence classes: $(11, 13)$, $(13, 15, 16, 17)$, and (21).

At each step in the SNK method, we are testing ordered subgroups of size $k = 6, .., 2$ against a critical value $q^*_K = q_{(K,K(r-1)),\alpha}$, which is changing, growing smaller, as k decreases from 6 to 2. This critical value is the same for all ordered subgroups in the tree at a given level k.

The SNK method, applied at level α, controls EERC at below α. It does not control MEER, unlike the HSD method, but it has more power than the HSD method.

Ryan, Einot, Gabriel, Welsch (REGWQ) method

The REGWQ method is a stepdown method, rather like SNK, for identifying subgroups of equivalent treatment means. It is more conservative about Type I error, and less powerful, than SNK. The REGWQ method controls MEER at level α.

Suppose again that we are testing 6 treatment groups at level α, with r replicates within each group, and we have obtained experimental means with $\hat{\mu}_1 < \hat{\mu}_2 < ... < \hat{\mu}_5 < \hat{\mu}_6$. As with SNK, we begin by testing first the full ordered group of 6 treatment means, and then the two ordered subgroups of 5 treatment means, against the studentized range statistics $q^*_6 = q_{(6,6(r-1)),\alpha}$ and $q^*_5 = q_{(5,5(r-1)),\alpha}$ respectively. As with SNK, when we fail to reject H_0 for an ordered subset of size p, we do not go on to test any more smaller subsets within that set.

In the REGWQ method, for the succeeding ordered subgroups of size $p = 4, 3, 2$, we lower the value of α at every level k. This contrasts with the SNK method, in which the test level α remains the same for all levels k. For hypothesis testing within ordered subgroups of size p, from a full set of treatment

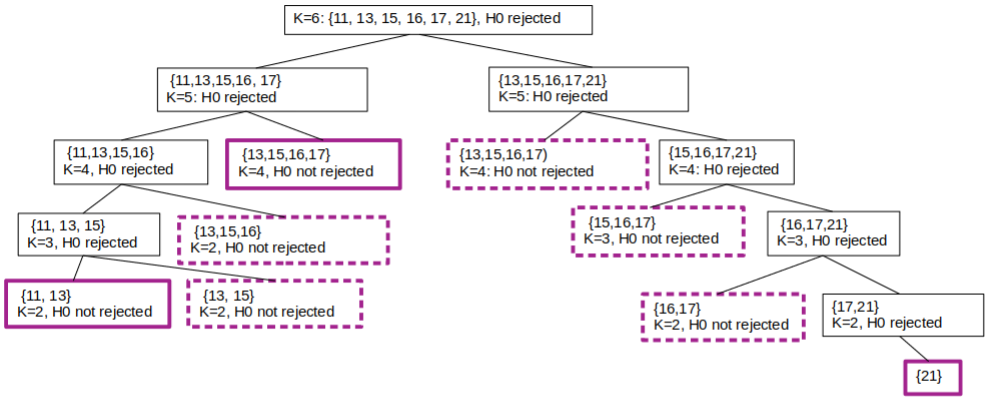

Figure 1: Diagram of tests in an SNK tree.

means of size K, we use the value

$$\alpha_{p,K} = 1 - (1 - \alpha)^{\frac{p}{K}}.$$

Since $\alpha_{p,K} < \alpha$, the criteria for rejecting H_0 for smaller subgroups are tougher for the REGWQ method than they are for the SNK method. Unlike SNK, REGWQ controls MEER.

Conclusion

In this writeup, I've discussed a number of ways that we can measure the Type I error of an experiment involving multiple comparisons of means. These metrics range in stringency; in general, CER and EERC are less than EERP and MEER.

I've also discussed a number of multiple-comparison protocols that control various types of Type I error. These range from mild protocols (such as LSD (controls CER) and FLSD (controls EERC)), to more severe ones (such as HSD and REGWQ, which control MEER). The more severe protocols control the Type I error of an experiment at a great cost to statistical power. Different fields of study have different standards for the acceptability of Type I errors.

I gratefully acknowledge Prof. Jana Anderson's Experimental Design class at Colorado State University for opening my eyes to the importance of error control in experimental design.

References

[1] Kemp, K. *Multiple Comparisons: Comparisonwise Versus Experimentwise Type I Error Rates and Their Relationship to Power.* J Dairy Sci. 1975 Sep;58(9):1374-8.

[2] Boardman and Moffitt, *Graphical Monte Carlo Type I Error Rates for Multiple Comparison Procedures.* 1973.

[3] *Multiple Comparison Procedures.* Compilation available at: http://www.math.montana.edu/jobo/st541/sec2c.pdf.

[4] Tukey, John. *Comparing Individual Means in the Analysis of Variance.* Biometrics. 5 (2): 99–114. JSTOR 3001913. 1949.