Post Hoc Analysis: One vs. the Rest

Neil J. Hatfield

3/6/2021

In discussions revolving around the Clinical Drug Trial example, we wrestled with the following question:

> 2) If so [the model with the drug factor being the better fit], which drugs are statistically different from the others?

During class, I steered us towards making the $\binom{12}{2} = 66$ pairwise tests. However, the question was raised "Why can't we just do one drug at a time vs. the rest of the drugs?"

On the surface, this looks like a good idea: 12 acts of statistical inference vs. 66 acts. Fewer acts means fewer opportunities to make a Type I error. Let's explore both options; we'll refer to them as the Pairwise Approach and the One vs. the Rest Approach.

Setting the Stage

To help us with our explorations, we're going to make use of a simplified and generic context. This is so that we can make a definitive statement of what is True.

For our simplified context we will deal with a single response and a single factor, fixed, with four levels. This results in a one-way ANOVA context where we will use a completely randomized and balanced design with four treatments (A, B, C, and D) and 50 replicates in each treatment.

The data completely adhere to the assumptions of the parametric shortcut in that $y_{ij} \overset{iid}{\sim} \mathcal{N}(\mu_{\cdot\cdot} + \alpha_i, 100), \forall \alpha_i \neq 0$. Further, the parametric shortcut (One-way ANOVA F test) should detect that there is an effect due to the factor at the 95% confidence level (i.e., $UT = 0.05$)

The final constraint we will put in place is that only treatments A and D will have treatment effects which are sufficiently different enough from one another at the $UT = 0.05$ level.

With this simpler data set, we will test following SRQs:

1) Does the factor have an impact on the response?
2) Which treatments are statistically different from the others?

(While I have not mentioned how, I have put into place appropriate guards against Type I errors.)

One-way ANOVA *F* Test Results (SRQ 1)

Table 1: ANOVA Table for Demostration Data

	df	SS	MS	F	p-value
factor	3	773.6715	257.8905	3.0197	0.031
Residuals	196	16739.0714	85.4034		

From Table 1, we can see that our p-value is lower than our Unusualness Threshold. Thus, we will reject the null hypothesis. The results of this test are in line with what we know to be True, given how we constructed the data.

The Pairwise Approach to SRQ 2

If we were to answer the second research question using the pairwise approach, we would get the results displayed in Table 2. Notice that for all of the comparisons except D vs. A (third row), 0 is between the Lower and Upper bounds

of 95% confidence interval and the p-value is greater than our Unusualness Threshold ($UT = 0.05$). This set of six decisions do not reflect any Type I or Type II errors, as we know that only treatments A and D are sufficient different from one another; all other pairings should (and do) result in us stating that the treatments are not statistically different from one another.

Table 2: Pairwise Results

Comparison	Difference	Lower Bound	Upper Bound	p-value
B vs. A	-3.2694	-8.0586	1.5199	0.2915
C vs. A	-2.5288	-7.3180	2.2605	0.5208
D vs. A	-5.5097	-10.2990	-0.7204	0.0169
C vs. B	0.7406	-4.0487	5.5299	0.9782
D vs. B	-2.2403	-7.0296	2.5490	0.6199
D vs. C	-2.9809	-7.7702	1.8083	0.3738

One vs. the Rest Approach to SRQ 2

Table 3 displays the results of doing the four comparisons of one treatment versus the rest. Notice here, that all of the 95% confidence intervals (lower and upper bounds) contain 0 and all of the p-values are greater than our Unusualness Threshold ($UT = 0.05$). This approach has failed to detect the difference between Treatment A and Treatment D (a Type II error). In essence, we masked the difference between Treatments A and D when we pooled treatments together.

Table 3: One vs. Rest Results

Comparison	Difference	Lower Bound	Upper Bound	p-value
A vs. B,C,D	3.7693	-0.4863	8.0249	0.1042
B vs. A,C,D	-0.5899	-4.8455	3.6657	0.9848
C vs. A,B,D	0.3976	-3.8580	4.6532	0.9952
D vs. A,B,C	-3.5770	-7.8326	0.6786	0.1355

What's Going On?

From our simulation, we know that the Pairwise Approach resulted in neither Type I nor Type II errors (Table 2). Further, we know that the One vs. the Rest Approach resulted in a Type II error but no Type I errors (Table 3). There are three things that are going on:

1) In the One vs. the Rest Approach, the individual treatments effects are replaced with new treatment effect for the pool,
$$\gamma = \frac{\sum_{i=1}^{k-1} n_i \alpha_i}{\sum_{i=1}^{k-1} n_i}$$
While we used a balanced design, if the n_i's weren't all the same (i.e., we have an imbalanced design), some groups might greatly overpower others in the pooling.
2) We are running into a Type III Error.
3) We are running into a Type IV Error.

Both the Type III and Type IV errors deserve a closer inspection.

The Type III Issue

Recall that a Type III Error goes by the title "Right Answer, Wrong Question". We end up in such a situation when we correctly interpret the results of our test/method BUT what we aren't realizing is that the test/method we used does not actually answer our research question. As an extreme example, imagine using a One-sample Z test to answer our question of whether the type of drug has an effect on hours of pain relief. While you could certainly run all of the data through the One-sample Z test, this method tells nothing about groups nor helps us to answer the question posed.

Now, the majority of Type III errors are nowhere near as extreme or obvious as this last example. This is what makes Type III errors much more difficult to detect. Let's return to the question posed in the Clinical Drug Trial situation:

2) If so, which drugs are statistically different from the others?

To help us guard against making a Type III error, we must first understand what the question is actually asking. To do so, let's first consider the drug Aspirin. We want to know if Aspirin is statistically different from Ibuprofen, from Naproxen, from Acetaminophen, etc. Once we're done with Aspirin, we'll move to the next drug (Ibuprofen) and ask whether Ibuprofen is different from Naproxen, from Acetaminophen, etc. (We don't have to do Ibuprofen and Aspirin as we've already done that comparison.)

Notice what happened: we replaced "which drugs" with a specific drug (first Aspirin, then Ibuprofen) AND we replaced "from the others" with each of the other drugs. This highlights the special nature of the second research question: this research question is a common place shorthand expression for asking the following:

1) Is Aspirin statistically different from Ibuprofen?
2) Is Aspirin statistically different from Naproxen?
3) Is Aspirin statistically different from Acetaminophen?
4) Is Aspirin statistically different from Tramadol?
5) Is Aspirin statistically different from Hydrocodone?
6) Is Aspirin statistically different from Dicolfenac?
7) Is Aspirin statistically different from Oxycodone?
8) Is Aspirin statistically different from Fentanyl?
9) Is Aspirin statistically different from Celecoxib?
10) Is Aspirin statistically different from Benzocaine?
11) Is Aspirin statistically different from Hydrocortisone?
12) Is Ibuprofen statistically different from Naproxen?
13) Is Ibuprofen statistically different from Acetaminophen?
14) Is Ibuprofen statistically different from Tramadol?
15) Is Ibuprofen statistically different from Hydrocodone?
16) Is Ibuprofen statistically different from Dicolfenac?
17) Is Ibuprofen statistically different from Oxycodone?
18) Is Ibuprofen statistically different from Fentanyl?
19) Is Ibuprofen statistically different from Celecoxib?
20) Is Ibuprofen statistically different from Benzocaine?
21) Is Ibuprofen statistically different from Hydrocortisone?
22) Is Naproxen statistically different from Acetaminophen?
23) Is Naproxen statistically different from Tramadol?
24) Is Naproxen statistically different from Hydrocodone?
25) Is Naproxen statistically different from Dicolfenac?
26) Is Naproxen statistically different from Oxycodone?
27) Is Naproxen statistically different from Fentanyl?
28) Is Naproxen statistically different from Celecoxib?
29) Is Naproxen statistically different from Benzocaine?
30) Is Naproxen statistically different from Hydrocortisone?
31) Is Acetaminophen statistically different from Tramadol?
32) Is Acetaminophen statistically different from Hydrocodone?
33) Is Acetaminophen statistically different from Dicolfenac?
34) Is Acetaminophen statistically different from Oxycodone?
35) Is Acetaminophen statistically different from Fentanyl?
36) Is Acetaminophen statistically different from Celecoxib?
37) Is Acetaminophen statistically different from Benzocaine?
38) Is Acetaminophen statistically different from Hydrocortisone?
39) Is Tramadol statistically different from Hydrocodone?
40) Is Tramadol statistically different from Dicolfenac?
41) Is Tramadol statistically different from Oxycodone?
42) Is Tramadol statistically different from Fentanyl?
43) Is Tramadol statistically different from Celecoxib?
44) Is Tramadol statistically different from Benzocaine?
45) Is Tramadol statistically different from Hydrocortisone?
46) Is Hydrocodone statistically different from Dicolfenac?
47) Is Hydrocodone statistically different from Oxycodone?

48) Is Hydrocodone statistically different from Fentanyl?
49) Is Hydrocodone statistically different from Celecoxib?
50) Is Hydrocodone statistically different from Benzocaine?
51) Is Hydrocodone statistically different from Hydrocortisone?
52) Is Dicolfenac statistically different from Oxycodone?
53) Is Dicolfenac statistically different from Fentanyl?
54) Is Dicolfenac statistically different from Celecoxib?
55) Is Dicolfenac statistically different from Benzocaine?
56) Is Dicolfenac statistically different from Hydrocortisone?
57) Is Oxycodone statistically different from Fentanyl?
58) Is Oxycodone statistically different from Celecoxib?
59) Is Oxycodone statistically different from Benzocaine?
60) Is Oxycodone statistically different from Hydrocortisone?
61) Is Fentanyl statistically different from Celecoxib?
62) Is Fentanyl statistically different from Benzocaine?
63) Is Fentanyl statistically different from Hydrocortisone?
64) Is Celecoxib statistically different from Benzocaine?
65) Is Celecoxib statistically different from Hydrocortisone?
66) Is Benzocaine statistically different from Hydrocortisone?

The above list of 66 questions is rather tedious to write (in fact, I wrote code to have R do it for me). However, this is what is meant by asking the question "Which drugs are statistically different from the others?" Many researchers use such phrasing to get around having to write out the $\binom{k}{2}$ possible SRQs about pairwise differences.

When researchers ***do*** want to combine different treatments together for a SRQ, they will do so rather explicitly. For example, SRQs 4–7 in the Clinical Drug Trial situation. In these questions, the researchers did in fact want to combine different treatments (drugs) together into pools. When you use the One vs. the Rest Approach, you are essentially seeking asnswers to the following research questions:

1) Is Aspirin statistically different from the ***pool of drugs containing*** Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
2) Is Ibuprofen statistically different from the ***pool of drugs containing*** Aspirin, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
3) Is Naproxen statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
4) Is Acetaminophen statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
5) Is Tramadol statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
6) Is Hydocodone statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
7) Is Dicolfenac statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Oxycodone, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
8) Is Oxycodone statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Fentanyl, Celecoxib, Benzocaine, and Hydrocortisone?
9) Is Fentanyl statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Celecoxib, Benzocaine, and Hydrocortisone?
10) Is Celecoxib statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Benzocaine, and Hydrocortisone?
11) Is Benzocaine statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, and Hydrocortisone?
12) Is Hydrocortisone statistically different from the ***pool of drugs containing*** Aspirin, Ibuprofen, Naproxen, Acetaminophen, Tramadol, Hydocodone, Dicolfenac, Oxycodone, Fentanyl, Celecoxib, and Benzocaine?

This second set of 12 questions is not equivalent to the first set of 66 questions. Further, there is no overlap and you can't transform one set into the other. Thus, using the One vs. the Rest Approach to answer SRQ 2, is to make a Type III error.

The Type IV Issue

Making a Type IV Error occurs when you over-interpret, over-reach, or otherwise make claims beyond what your method actually supports you in doing. The pop culture saying "Correlation isn't Causation" strikes at the heart of

Type IV errors. Many people take a high value of the correlation between A and B to mean that A causes B (or B causes A; whichever order fits with their beliefs). Measures of association (like [Pearson's] correlation) **can not** tell us anything about causation. What they can tell us is that there might be something going on, but that could be any number of things.

While the risk of making a Type IV error exists in both approaches, there is a greater risk in the One vs. the Rest Approach.

Suppose that we did in fact want to examine whether each drug was statistically different from the pool consisting of the remaining 11 drugs. Now suppose that of those 12 tests, only the one where we looked at Fentanyl vs. the Rest came back as statistically significant. There is a massive temptation to state that

- Fentanyl is statistically different from the other drugs, and
- The other drugs are not statistically different from each other.

Unfortunately, neither of these statements is supported by the One vs. the Rest Approach and are thus both Type IV errors. While you might intend for the statement "Fentanyl is statistically different from the other drugs" to mean *the pool* of the other drugs, your readers are going to understand you to mean "Fentanyl is statistically different from *each* of the other drugs."

The basis for making the second statement is that the other 11 tests did not return a statistically significant result, so each drug isn't different from the pools, therefore they must not be statistically different from each other. Again, this is not supported by the One vs. the Rest Approach. This method only lets you claim that each drug (except Fentanyl) is not statistically different from the *the pool* of other drugs. This approach does not allow you to make inferences about pairs of drugs. Thus, this claim is also a Type IV error.

If we were to return to SRQ 2 as written (i.e., we need to use the Pairwise Approach), then not only would we make a Type III error, but we would also make the Type IV errors just described. This highlights an important fact: Type III and Type IV errors are independent of each other. This is different from Type I and Type II errors; if you make a Type I error on a certain test, you can't also make a Type II error on that same test (and vice versa). Type III and Type IV errors may occur in any combination with Type I and Type II errors.

Final Notes

Method	Type I Error*	Type II Error*	Type III Error†	Type IV Error‡
Pairwise	No	No	No	No
One vs. the Rest	No	Yes	Yes	Yes

For our constructed situation (*), we know that the Pairwise Approach did not result in Type I or Type II errors; the One vs. the Rest Approach did result in a Type II error. In practice, you won't know what the Truth is, thus we will not know for certain whether we've made a Type I or II error.

When we are looking at a research question along the lines of "If so, which drugs are statistically different from the others?" (†), we know that the Pairwise Approach will not result in a Type III error while the One vs. the Rest Approach will.

Finally, we know that barring any other forms of Type IV errors beyond what was discussed(‡), the Pairwise Approach will not lead to those mentioned Type IV errors whereas the One vs. the Rest Approach will lead to those mentioned Type IV errors.

While the One vs. the Rest Approach will result in fewer acts of statistical inference and thus fewer total opportunities for Type I errors, this method does bring an increased risk of Type II, Type III, and Type IV errors when trying to answer the second SRQ. There is a constant dance you must participate in as you work to guard against making inference errors.

Keep in mind that if the research question was about pools, then the One vs. the Rest Approach *might* be a valid approach to use. However, this will depend upon the nature of the SRQ. Everything comes back to the SRQ and what meaning you have given to it.