Which of These Things is Not Like the Others?

How Multiple Comparison Tests Work

Dennis R. Helsel

PracticalStats.com

Objectives of the 'Which of These Things' webinar

1. To understand why to use multiple comparison tests, and how they work
2. To demonstrate how the False Discovery Rate adjustment works and why it should be used
3. To highlight one of the many new items in the new 2nd edition of Statistical Methods in Water Resources, soon available at http://practicalstats.com/info2use/books.html

Outline: How Multiple Comparison Tests Work

1. What are multiple comparison tests used for?
2. What's new since the 1950-1960s?
3. How does the False Discovery Rate work, and why should I care?

Poll

Do you know source of the song "Which of these things is not like the others?" ?

Billy Joel
Sesame Street
ABBA
The Cat in the Hat
~~(original) Lion King~~
Bill Nye the Science Guy

Why Multiple Comparison Tests? When are They Used?

- **After Analysis of Variance (ANOVA) and the Kruskal-Wallis (K-W) tests**. These two tests do not determine which groups differ from others. Multiple comparison tests come along afterwards to determine which groups differ from others, with the goal of setting the probability of making an error in the pattern of group orderings equal to the alpha level used in the ANOVA or K-W test.
- **To control the site-wide false positive rate**. At a ground water monitoring site, perhaps 10 different chemicals at 8 wells are tested over time to see if concentrations remain below the legal standard. This requires performing 10*8=80 tests. If each test were run at an alpha = 0.05, there would be a 98% probability that at least 1 test was a false positive, so falsely requiring remediation. Multiple comparison procedures are adopted to set the probability of false positives across the entire site.
- **Trend analysis at many sites in a region** or any situation where there are multiple tests being run as a group, if each is run with a specific alpha false positive rate such as 5%, the overall error rate for the entire group is much higher. Using the adjustments from multiple comparison tests allow the scientist to set the overall false positive rate for the entire group (entire region / collection of sites / all seasons, etc.)

The Goal of Multiple Comparison Tests

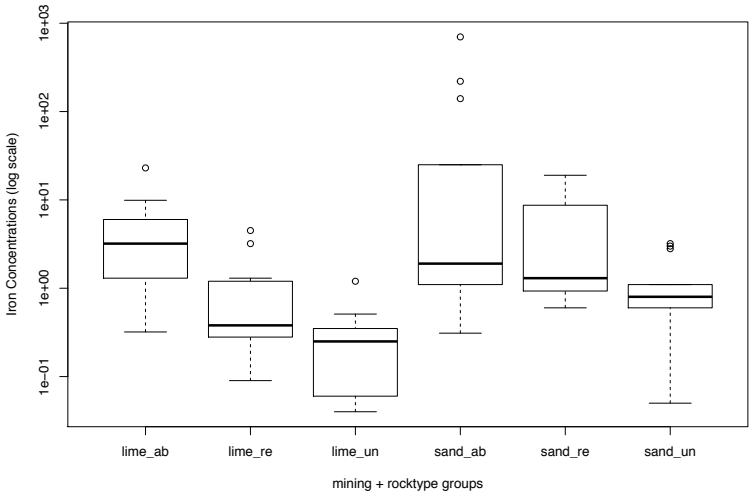
- Multiple comparison methods determine whether there are signals (differences between pairs of groups; any significance test) for a collection of tests, while controlling (setting) the error rate for the entire group.
- For example, when testing for differences in means between 6 groups there are (6*5)/2 = 15 tests performed.
- Together, these tests produce a pattern of which groups differ from others, such as group A > group B > group C = group D > group E = group F
- A common goal is to specify the overall (pattern, or family) error rate, so that there is no more than a 5% (or whatever amount you choose) probability of making an error in the pattern

Example: iron concentrations in 6 groups

Kruskal-Wallis test
p-value = 2.633e-06

There are differences between the group percentiles (cdfs)

lime = limestone
sand = sandstone



The overall (pattern or family) error rate

- If each pair of k groups is tested for difference, the number of pairwise tests equals $c = k(k-1)/2$. For k=6 groups, there are 15 tests run.
- If each of the c tests is run using a pairwise (or individual) error rate $\alpha_{pairwise}$ = 0.05, there is an overall error rate (the chance of making at least one error in the pattern) equal to $\alpha_{overall} = 1 - (1 - \alpha_{pairwise})^c$ = 0.54.
- One of the pioneers of solutions to this multiple comparison problem was Olive Dunn. In the 1960s she developed multiple comparison methods to follow both ANOVA (1961) and the Kruskal-Wallis test (1964). Dunn's 1964 method was the standard procedure for nonparametric tests until quite recently.
- Multiple comparison tests allow you to set the overall error rate, say to 0.05, and then declare groups different when a two-group test's p-value is below an individual error rate smaller than 0.05.
- Different multiple comparison tests differ in how they get down to the individual error rate while keeping the overall rate to your specification.

Simplest adjustment method: Bonferroni

- Bonferroni's method is the simplest method to keep the overall error rate below the desired amount (0.05, for example).
- It is perhaps the least powerful method –> the individual error rate is the smallest of these type of tests, and therefore less able to see differences.
- $\alpha_{Bonferroni} = \frac{\alpha_{Overall}}{c}$. For 6 groups with an overall error rate of 0.05, c=15 and so

$\alpha_{Bonferroni} = \frac{0.05}{15} = 0.0033$

Each of the two-group tests run manually with a t- or rank-sum test must have a p-value less than 0.0033 to be declared different using the Bonferroni adjustment. This is a very restrictive approach, making it difficult to see differences that are there.

Minimizing the Overall Error Rate (Bonferroni)

1. Set the overall error rate to desired level (say 0.05). Compute the c (for 6 groups, c=15) two-group tests.
2. Sort the resulting p-values from high to low.
3. Compare each p-value to the pairwise limit $\alpha_{overall}/c$. For c=15, the pairwise limit = 0.0033

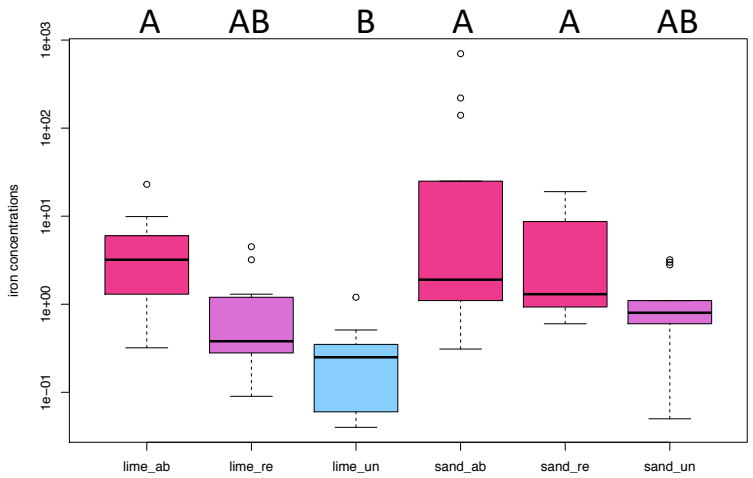
p	limit	reported	p	limit	reported	p	limit	reported	p	limit	reported
1.000	> 0.0033		0.0544	> 0.0033		0.0096	> 0.0033		0.0016	< 0.0033	
0.644	> 0.0033		0.0378	> 0.0033		0.0065	> 0.0033		0.0002	< 0.0033	
0.397	> 0.0033		0.0256	> 0.0033		0.0034	> 0.0033		0.0000	< 0.0033	
0.270	> 0.0033		0.0148	> 0.0033					0.0000	< 0.0033	

4. If the p-value is below its limit it is significant.
5. The Bonferroni method finds 4 p-values below 0.0033. The Holm method is a more modern standard, and usually finds more tests that are significant for the same objective (minimize overall error rate).

Illustrate significant difference with colors

based on Bonferroni multiple comparison results.

Purple boxes are a gradation rather than a hard break. Only red and blue are significantly different.



Minimizing the Overall Error Rate (Holm)

1. Set the overall error rate to desired level (say 0.05). Compute the c (for 6 groups, c=15) two-group tests.
2. Sort the resulting p-values from high to low.
3. Compare each p-value to the pairwise limit $\alpha_{overall}/(c\text{-}i)$, starting at i = c-1 and ending at I = 0.

p	limit	p	limit	p	limit	p	limit
1.000	$\frac{1}{1}0.05 = 0.05$	0.054	$\frac{1}{5}0.05 = 0.010$	0.0096	$\frac{1}{9}0.05 = 0.0055$	0.0016 <	$\frac{1}{12}0.05 = 0.004$
0.644	$\frac{1}{2}0.05 = 0.025$	0.037	$\frac{1}{6}0.05 = 0.008$	0.0065	$\frac{1}{10}0.05 = 0.005$	0.0002 <	$\frac{1}{13}0.05 = 0.0038$
0.397	$\frac{1}{3}0.05 = 0.017$	0.026	$\frac{1}{7}0.05 = 0.007$	0.0034 <	$\frac{1}{11}0.05 = 0.0045$	0.0000 <	$\frac{1}{14}0.05 = 0.0036$
0.270	$\frac{1}{4}0.05 = 0.013$	0.015	$\frac{1}{8}0.05 = 0.006$			0.0000 <	$\frac{1}{15}0.05 = 0.0033$

4. The first two-group p-value to go below its pairwise limit is significant. So are all tests with lower p-values.

Holm finds 5 differences, one more than did Bonferroni. Bonferronni compared all of these p-values to 0.0033. Holm finds more differences for the same objective (while still minimizing the overall error rate).

Issues with Multiple Comparison Tests

- Earliest tests in the 1950s and 1960s (Duncan's, Least Significant Range tests) set individual error rates, and did not control (set) the overall error rate, which could get quite high.
- Some common tests assume normality for each of the groups (Tukey's, Dunn's following ANOVA). This is rarely true. Non-normality results in lower ability (power) to see the differences with parametric tests.
- Controlling the overall error rate (probability of making at least one error in the pattern) costs power if the adjustment (especially Bonferroni) is not efficient at getting down to the individual error rate, so differences may not be seen. These adjustments (particularly Holm's) have been the standard tests until recently but their loss of power has led to some people recommending not to use them.
- For example, in analysis of microarray gene expression data researchers sometimes perform as many as hundreds to thousands of comparison tests. A Bonferroni adjustment would require pairwise error rates of 0.00001 or smaller, making it extremely unlikely that any significant differences present would be seen.

What's new? Is the Overall Error Rate the best objective?

- Often the overall error rate is not the main issue. The primary interest is often whether an erroneous rejection, a false statement of difference or "false discovery", is made.
- This is a subset of the overall error rate, which was the probability of "making at least one error in the pattern".
- Minimizing only false positives is called the False Discovery Rate (FDR), measured using the Benjamini-Hochberg (BH) adjustment from their paper in 1995.
- The FDR provides a large gain in the power of seeing differences, as compared to the overall error rate used by Bonferroni or Holm adjustments.
- The increase in power by setting the FDR instead of the overall error rate usually results in a larger number of significant differences. It does not control the number of false no-differences. As there are usually fewer observed no-differences using the FDR, this often seems unimportant in comparison to obtaining more power to see differences.

Minimizing the False Discovery Rate (BH adjustment)

1. Set the FDR to desired level (say 0.05)
2. Compute the c (for 6 groups, c=15) two-group tests. Sort them by their p-values from high to low.
3. Compare each p-value to the limit i/c *(FDR), starting at i=15 for the largest p-value.

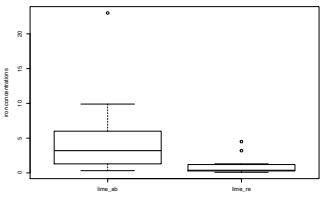
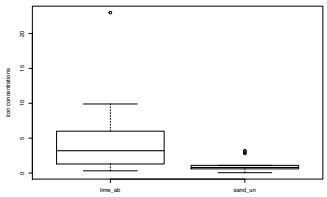
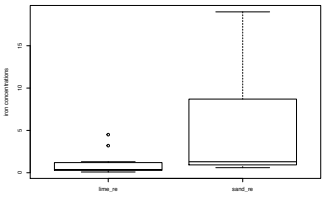
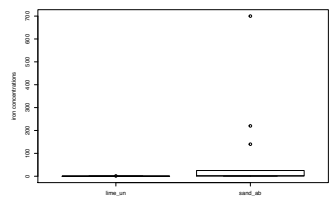
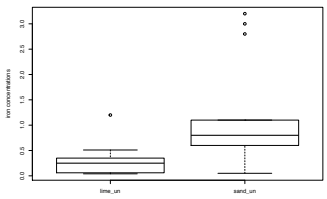
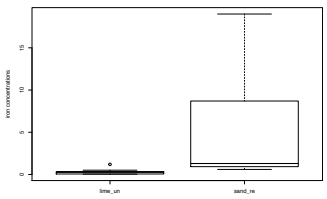
p	limit	p	limit	p	limit	p	limit
1.00	$\frac{15}{15}0.05 = 0.05$	0.054	$\frac{11}{15}0.05 = 0.037$	0.009	$\frac{7}{15}0.05 = 0.023$	0.0016	$\frac{4}{15}0.05 = 0.013$
0.64	$\frac{14}{15}0.05 = 0.047$	0.038	$\frac{10}{15}0.05 = 0.033$	0.006	$\frac{6}{15}0.05 = 0.020$	0.0002	$\frac{3}{15}0.05 = 0.010$
0.39	$\frac{13}{15}0.05 = 0.043$	0.025	$\frac{9}{15}0.05 = 0.030$	0.003	$\frac{5}{15}0.05 = 0.017$	0.0000	$\frac{2}{15}0.05 = 0.007$
0.27	$\frac{12}{15}0.05 = 0.040$	0.015	$\frac{8}{15}0.05 = 0.027$			0.0000	$\frac{1}{15}0.05 = 0.0033$

4. The first two-group p-value to go below its limit is significant. So are all tests with lower p-values.
5. The BH adjustment usually finds more significant comparisons than Bonferroni or Holm; It has a different objective, the False Discovery Rate.

15

Comparisons Finding a Difference using the BH adjustment

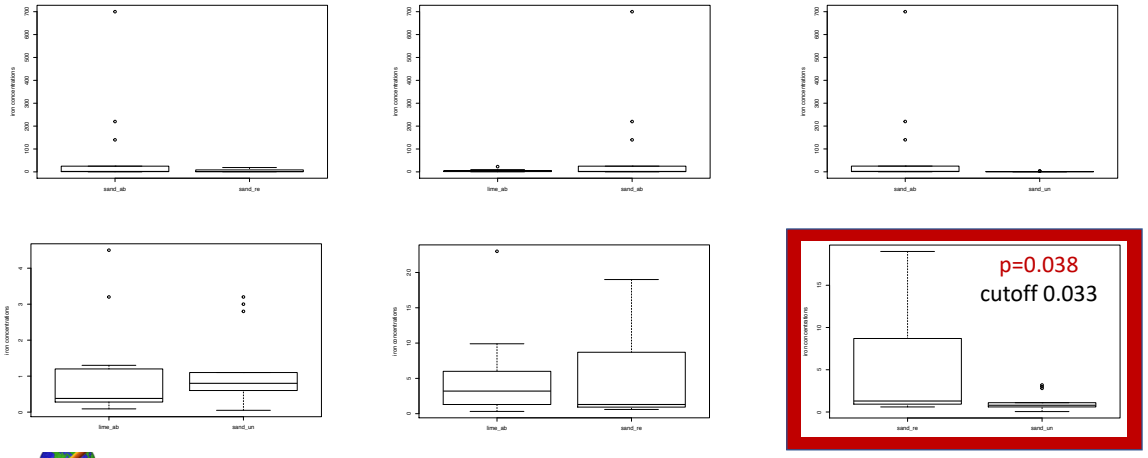
(6 of the 9 two-group tests found different)

16

Comparisons Not Finding a Difference using the BH adjustment

(The 6 two-group tests that found no difference at α=0.05)



Reporting results of multiple comparison tests

1. Each two-group test's p-value is re-scaled by software before reporting it to the user, so that the user can compare its value to their overall error or false discovery rate (say of 0.05).
2. For example, individual p-values for c=15 tests were all compared to Bonferroni's 0.05/15 = 0.0033 are multiplied by 15 before reporting them back to the user. A p-value just under 0.0033 would therefore be reported as just under 0.05. The user then compares the reported p-values to their overall error rate. So the highest significant p-value of 0.0016 would be reported in the output as p = 0.0016*15, or 0.024, to compare to the overall error rate.
3. BH p-values are multiplied by the 1/fraction used to adjust the FDR. In our example for the first significant p-value, the fractional multiplier of 9/15 becomes a multiplier of 15/9 = 1.667. The obtained p-value of 0.0256 is then reported as 0.0256*1.667 = 0.0426 to the user, who compares it to their false discovery rate of 0.05.
4. These re-adjusted p-values are often reported to the user in a triangular matrix.

Example: Reporting results of Bonferroni adjustment

1. Compute the all possible (c) two-group tests.
2. Multiply the obtained p-values by c to obtain the p-value reported to the user.
3. For example, the Wilcoxon rank-sum p-value of 0.0016*15 = 0.0240

p	limit	reported	p	limit	reported	p	limit	reported	p	limit	reported
1.000	> 0.0033	1.000	0.0544	> 0.0033	0.8158	0.0096	> 0.0033	0.1435	0.0016	< 0.0033	0.0240*
0.644	> 0.0033	1.000	0.0378	> 0.0033	0.5666	0.0065	> 0.0033	0.0981	0.0002	< 0.0033	0.0027
0.397	> 0.0033	1.000	0.0256	> 0.0033	0.3836	0.0034	> 0.0033	0.0516	0.0000	< 0.0033	0.0011
0.270	> 0.0033	1.000	0.0148	> 0.0033	0.2333				0.0000	< 0.0033	0.0009

4. The user compares the reported p-values to their Overall Error Rate.

Example: iron concentrations in 6 groups

Bonferroni adjustment

```
> kruskal.test(fe~group)
         Kruskal-Wallis rank sum test
data:  fe by group
Kruskal-Wallis chi-squared = 33.781, df = 5, p-value = 2.633e-06

> pairwise.wilcox.test (fe, group, p.adjust.method = "bonferroni")
         Pairwise comparisons using Wilcoxon rank sum test
data:  fe and group
        lime_ab  lime_re lime_un  sand_ab sand_re
lime_re 0.05157  -       -        -       -
lime_un 0.00105  0.38355 -        -       -
sand_ab 1.00000  0.09812 0.00270  -       -              4 of 15 tests significant
sand_re 1.00000  0.14352 0.00094  1.00000 -
sand_un 0.22231  1.00000 0.02404* 0.81582 0.56664
```

Example: Reporting results of BH adjustment

1. Compute the all-possible (c) two-group tests.
2. Multiply the obtained p-values by c/i (= 1/fraction) to obtain the p-value reported to the user.
3. For example, the Wilcoxon rank-sum p-value of 0.0016*15/4 = 0.0006

p	limit	reported	p	limit	reported	p	limit	reported	p	limit	reported
1.000	> 0.050	1.000	0.0544	> 0.037	0.0742	0.0096	< 0.023	0.0205	0.0016	< 0.013	0.0060*
0.644	> 0.047	0.690	0.0378	> 0.033	0.0567	0.0065	< 0.020	0.0164	0.0002	< 0.010	0.0009
0.397	> 0.043	0.458	0.0256	< 0.030	0.0426	0.0034	< 0.017	0.0103	0.0000	< 0.007	0.0005
0.270	> 0.040	0.337	0.0148	< 0.027	0.0278				0.0000	< 0.0033	0.0005

4. The user compares the reported p-values to their False Discovery Rate.

Example: iron concentrations in 6 groups
BH adjustment

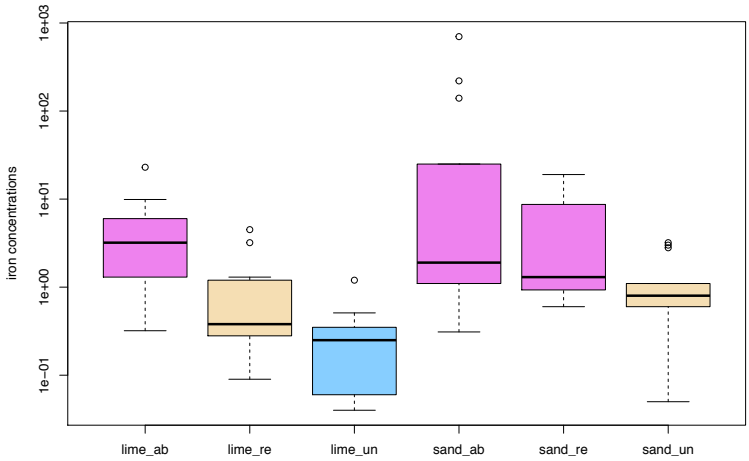
```
> kruskal.test(fe~group)
        Kruskal-Wallis rank sum test
data:  fe by group
Kruskal-Wallis chi-squared = 33.781, df = 5, p-value = 2.633e-06

> pairwise.wilcox.test (fe, group, p.adjust.method = "BH")
        Pairwise comparisons using Wilcoxon rank sum test
data:  fe and group
        lime_ab lime_re  lime_un  sand_ab sand_re
lime_re 0.01031 -        -        -       -
lime_un 0.00052 0.04262  -        -       -
sand_ab 1.00000 0.01635  0.00090  -       -              9 of 15 tests significant
sand_re 0.69026 0.02050  0.00052  0.45843 -
sand_un 0.02779 0.33746  0.00601* 0.07417 0.05666
```

Illustrate significant difference with colors

based on BH multiple comparison results.

Different colors are significantly different (hard breaks).



Example: iron concentrations in 6 groups

Holm adjustment

```
> kruskal.test(fe~group)
        Kruskal-Wallis rank sum test
data:  fe by group
Kruskal-Wallis chi-squared = 33.781, df = 5, p-value = 2.633e-06

> pairwise.wilcox.test (fe, group, p.adjust.method = "holm")
        Pairwise comparisons using Wilcoxon rank sum test
data:  fe and group
        lime_ab lime_re lime_un sand_ab sand_re
lime_re 0.03782 -       -       -       -
lime_un 0.00098 0.17899 -       -       -
sand_ab 1.00000 0.06541 0.00234 -       -              5 of 15 tests significant
sand_re 1.00000 0.08611 0.00094 1.00000 -
sand_un 0.11857 1.00000 0.01923 0.27194 0.22666
```

False Discovery Rate Multiple Comparisons using R

```
>  pairwise.wilcox.test (spcap, rock, p.adjust.method = "BH")
```

after a Kruskal-Wallis test

```
>  pairwise.t.test (spcap, rock, p.adjust.method = "BH")
```

after an ANOVA

```
> cen1way (Thiamethoxam, ThiaCens, SamplingEvent)
```

for a nonparametric test on data with nondetects (BH is the default adjustment method in cen1way)

Summary: Which of these things is not like the others?

- Multiple comparison tests determine which groups differ from others when each test's results will be reported
- Sets the overall (or false discovery) rate, the probability of making an error (or false positive) in the entire set of tests run
- Reported individual p-values are scaled so they can be directly compared to your overall error rate or false discovery rate
- Using the false discovery rate provides more ability (power) to see differences for very little cost than the more traditional Tukey's, Bonferroni, or Holm adjustments

Questions?

Our Next Webinar

Tuesday August 20th 11 am Mountain time

Correlation and Regression for Data with Nondetects

- Is another topic from our new *Nondetects And Data Analysis* online course
- topics for upcoming webinars: Bootstrapping and Permutation tests
  Trend Analysis for Data with Nondetects
  How to include both nondetects and "greater-thans" in analyses
- Online signup for our newsletter/announcement list to directly receive announcements each month is at http://practicalstats.com/news/
- Or check our webinars page periodically at http://practicalstats.com/training/webinar.html to see the announcement and to register.

27

This 'Which of These Things?' webinar will be available Thursday for streaming

- at our Online Training Site
  http://practicalstats.teachable.com/
  (and click the "View all courses" button to see the free webinars)

Let colleagues who missed it know about it.

28

Thank you for attending

- Much of the material is based on the book Statistical Methods in Water Resources, 2nd Edition by Helsel, Hirsch, Archfield, Ryberg and Gilroy. *USGS Techniques and Methods 4-A3* (2019).
- This topic and much more is now covered in our online course Applied Environmental Statistics, on our Training Site.
- All opinions are my own and do not represent those of anyone else you can think of.

Answers to your questions. Some now, all are answered y Thursday in a file that will be on our Downloads page: http://practicalstats.com/info2use/downloads.html

Get in touch!

Dennis Helsel ask@practicalstats.com

Courses & free webinars at our Training Site: http://practicalstats.teachable.com