5.2: Multiple Comparisons

When the null hypothesis is rejected by the F-test, we believe that there are significant differences among the *k* population means. So, which ones are different? Multiple comparison method is the way to identify which of the means are different while controlling the experiment-wise error (the accumulated risk associated with a family of comparisons). There are many multiple comparison methods available.

In **The Least Significant Difference Test**, each individual hypothesis is tested with the student t-statistic. When the Type I error probability is set at some value and the variance s2 has *v* degrees of freedom, the null hypothesis is rejected for any observed value such that |to|>tα/2, v. It is an abbreviated version of conducting all possible pair-wise t-tests. This method has weak experiment-wise error rate. Fisher's Protected LSD is somewhat better at controlling this problem.

Bonferroni inequality is a conservative alternative when software is not available. When conducting n comparisons, αe≤ n αc therefore αc = αe/n. In other words, divide the experiment-wise level of significance by the number of multiple comparisons to get the comparison-wise level of significance. The Bonferroni procedure is based on computing confidence intervals for the differences between each possible pair of μ's. The critical value for the confidence intervals comes from a table with (N – *k*) degrees of freedom and *k*(*k* – 1)/2 number of intervals. If a particular interval does not contain zero, the two means are declared to be significantly different from one another. An interval that contains zero indicates that the two means are NOT significantly different.

Dunnett's procedure was created for studies where one of the treatments acts as a control treatment for some or all of the remaining treatments. It is primarily used if the interest of the study is determining whether the mean responses for the treatments differ from that of the control. Like Bonferroni, confidence intervals are created to estimate the difference between two treatment means with a specific table of critical values used to control the experiment-wise error rate. The standard error of the difference is Image37382.PNG.

Scheffe's test is also a conservative method for all possible simultaneous comparisons suggested by the data. This test equates the F statistic of ANOVA with the t-test statistic. Since t2 = F then t = √F, we can substitute √F(αe, v1, v2) for t(αe, v2) for Scheffe's statistic.

Tukey's test provides a strong sense of experiment-wise error rate for all pair-wise comparison of treatment means. This test is also known as the *Honestly Significant Difference*. This test orders the treatments from smallest to largest and uses the studentized range statistic

$$q = \frac{\bar{y}(largest) - \bar{y}(smallest)}{\sqrt{MSE/r}} \tag{5.2.1}$$

The absolute difference of the two means is used because the location of the two means in the calculated difference is arbitrary, with the sign of the difference depending on which mean is used first. For unequal replications, the Tukey-Kramer approximation is used instead.

Student-Newman-Keuls (SNK) test is a multiple range test based on the studentized range statistic like Tukey's. The critical value is based on a particular pair of means being tested within the entire set of ordered means. Two or more ranges among means are used for test criteria. While it is similar to Tukey's in terms of a test statistic, it has weak experiment-wise error rates.

Bonferroni, Dunnett's, and Scheffe's tests are the most conservative, meaning that the difference between the two means must be greater before concluding a significant difference. The LSD and SNK tests are the least conservative. Tukey's test is in the middle. Robert Kuehl, author of *Design of Experiments: Statistical Principles of Research Design and Analysis* (2000), states that the Tukey method provides the best protection against decision errors, along with a strong inference about magnitude and direction of differences.

Let's go back to our question on mean rain acidity in Alaska, Florida, and Texas. The null and alternative hypotheses were as follows:

H_0: μA = μF = μT	H_1: at least one of the means is different

The p-value for the F-test was 0.000229, which is less than our 5% level of significance. We rejected the null hypothesis and had enough evidence to support the claim that at least one of the means was significantly different from another. We will use Bonferroni and Tukey's methods for multiple comparisons in order to determine which mean(s) is different.

Bonferroni Multiple Comparison Method

A Bonferroni confidence interval is computed for each pair-wise comparison. For *k* populations, there will be *k*(*k*-1)/2 multiple comparisons. The confidence interval takes the form of:

$$For\ \mu_1 - \mu_2 : (\bar{x_1} - \bar{x_2}) \pm (Bonferronit\ critical\ value)\sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$$

$$For\ \mu_{k-1} - \mu_k : (\bar{x_{k-1}} - \bar{x_k}) \pm (Bonferronit\ critical\ value)\sqrt{\frac{MSE}{n_{k-1}} + \frac{MSE}{n_k}}$$

Where MSE is from the analysis of variance table and the Bonferroni *t* critical value comes from the Bonferroni Table given below. The Bonferroni *t* critical value, instead of the student *t* critical value, combined with the use of the MSE is used to achieve a simultaneous confidence level of at least 95% for all intervals computed. The two means are judged to be significantly different if the corresponding interval does not include zero.

Table 5. Bonferroni t-critical values.

df	2	3	4	5	6	10
2	6.21	7.65	8.86	9.92	10.89	14.09
3	4.18	4.86	5.39	5.84	6.23	7.45
4	3.50	3.96	4.31	4.60	4.85	5.60
5	3.16	3.53	3.81	4.03	4.22	4.77
6	2.97	3.29	3.52	3.71	3.86	4.32
7	2.84	3.13	3.34	3.50	3.64	4.03
8	2.75	3.02	3.21	3.36	3.48	3.83
9	2.69	2.93	3.11	3.25	3.36	3.69
10	2.63	2.87	3.04	3.17	3.28	3.58
11	2.59	2.82	2.98	3.11	3.21	3.50
12	2.56	2.78	2.93	3.05	3.15	3.43
13	2.53	2.75	2.90	3.01	3.11	3.37
14	2.51	2.72	2.86	2.98	3.07	3.33
15	2.49	2.69	2.84	2.95	3.04	3.29
16	2.47	2.67	2.81	2.92	3.01	3.25
17	2.46	2.66	2.79	2.90	2.98	3.22
18	2.45	2.64	2.77	2.88	2.96	3.20
19	2.43	2.63	2.76	2.86	2.94	3.17
20	2.42	2.61	2.74	2.85	2.93	3.15
21	2.41	2.60	2.73	2.83	2.91	3.14
22	2.41	2.59	2.72	2.82	2.90	3.12
23	2.40	2.58	2.71	2.81	2.89	3.10
24	2.39	2.57	2.70	2.80	2.88	3.09
25	2.38	2.57	2.69	2.79	2.86	3.08

df	2	3	4	5	6	10
26	2.38	2.56	2.68	2.78	2.86	3.07
27	2.37	2.55	2.68	2.77	2.85	3.06
28	2.37	2.55	2.67	2.76	2.84	3.05
29	2.36	2.54	2.66	2.76	2.83	3.04
30	2.36	2.54	2.66	2.75	2.82	3.03
40	2.33	2.50	2.62	2.70	2.78	2.97
60	2.30	2.46	2.58	2.66	2.73	2.91
120	2.27	2.43	2.54	2.62	2.68	2.86

For this problem, $k = 3$ so there are $k(k-1)/2 = 3(3-1)/2 = 3$ multiple comparisons. The degrees of freedom are equal to $N - k = 18 - 3 = 15$. The Bonferroni critical value is 2.69.

$$For \mu_A - \mu_F : (5.033 - 4.517) \pm (2.69)\sqrt{\frac{0.1011}{6} + \frac{0.1011}{6}} = (0.0222, 1.0098)$$

$$For \mu_A - \mu_T : (5.033 - 5.537) \pm (2.69)\sqrt{\frac{0.1011}{6} + \frac{0.1011}{6}} = (-0.9978, -0.0102)$$

$$For \mu_F - \mu_T : (4.517 - 5.537) \pm (2.69)\sqrt{\frac{0.1011}{6} + \frac{0.1011}{6}} = (-1.5138, 0.5262)$$

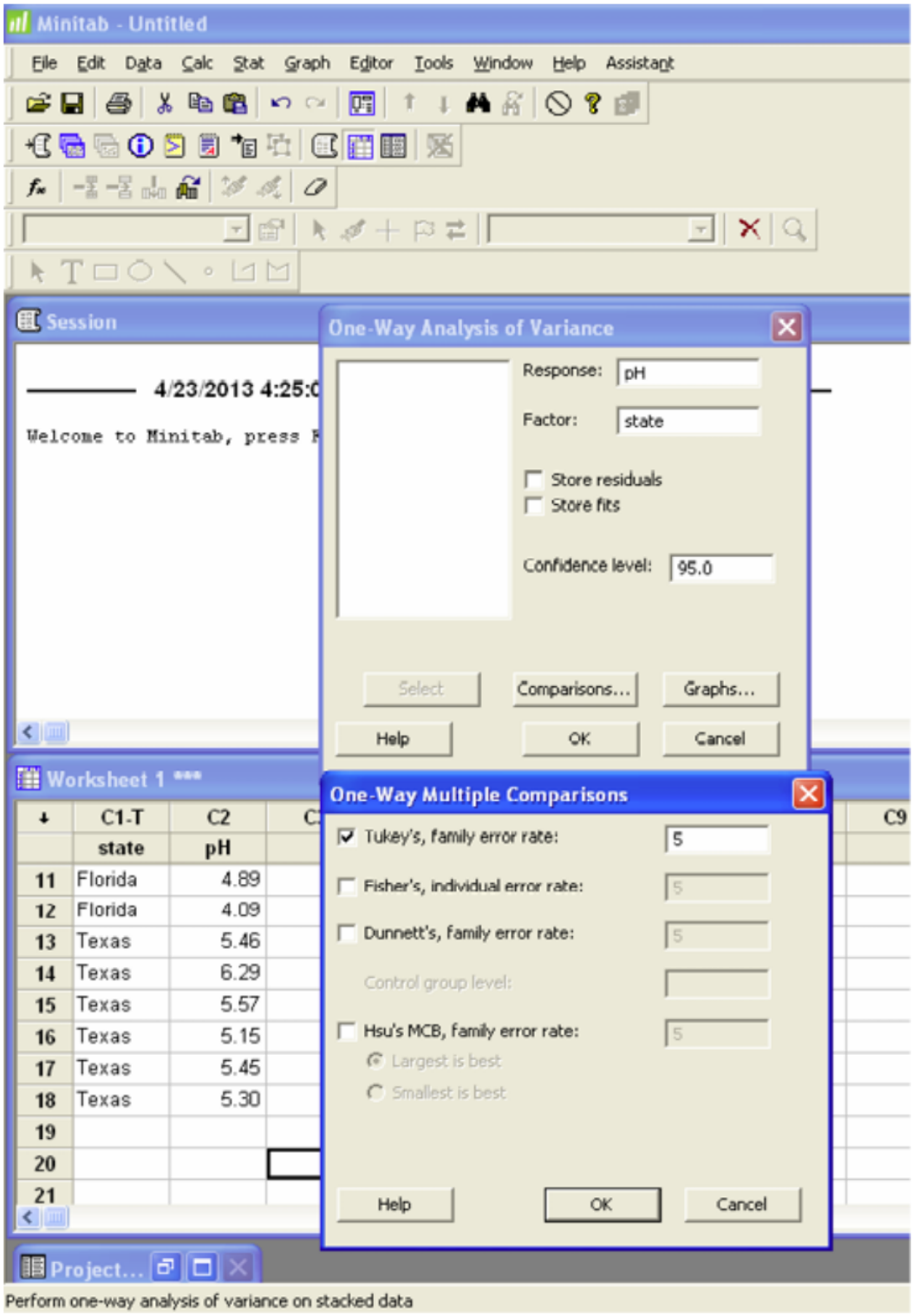
The first confidence interval contains all positive values. This tells you that there is a significant difference between the two means and that the mean rain pH for Alaska is significantly greater than the mean rain pH for Florida.

The second confidence interval contains all negative values. This tells you that there is a significant difference between the two means and that the mean rain pH of Alaska is significantly lower than the mean rain pH of Texas.

The third confidence interval also contains all negative values. This tells you that there is a significant difference between the two means and that the mean rain pH of Florida is significantly lower than the mean rain pH of Texas.

All three states have significantly different levels of rain pH. Texas has the highest rain pH, then Alaska followed by Florida, which has the lowest mean rain pH level. You can use the confidence intervals to estimate the mean difference between the states. For example, the average rain pH in Texas ranges from 0.5262 to 1.5138 higher than the average rain pH in Florida.

Now let's use the Tukey method for multiple comparisons. We are going to let software compute the values for us. Excel doesn't do multiple comparisons so we are going to rely on Minitab output.



One-way ANOVA: pH vs. state

Source	DF	SS	MS	F	P
state	2	3.121	1.561	15.4	0.000
Error	15	1.517	0.101		

Source	DF	SS	MS	F	P
Total	17	4.638			
S = 0.3180		R-Sq = 67.29%		R-Sq(adj) = 62.93%	

We have seen this part of the output before. We now want to focus on the *Grouping Information Using Tukey Method.* All three states have different letters indicating that the mean rain pH for each state is significantly different. They are also listed from highest to lowest. It is easy to see that Texas has the highest mean rain pH while Florida has the lowest.

Grouping Information Using Tukey Method

state	N	Mean	Grouping
Texas	6	5.5367	A
Alaska	6	5.0333	B
Florida	6	4.516	C
Means that do not share a letter are significantly different.			

This next set of confidence intervals is similar to the Bonferroni confidence intervals. They estimate the difference of each pair of means. The individual confidence interval level is set at 97.97% instead of 95% thus controlling the experiment-wise error rate.

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of state
Individual confidence level = **97.97%**

state = Alaska subtracted from:							
state	Lower	Center	Upper	———+———+———+———+			
Florida	-0.9931	-0.5167	-0.0402		(—*—)		
Texas	0.0269	0.5033	0.9798			(—*—)	
				———+———+———+———+			
				-0.80	0.00	0.80	1.60

state = Florida subtracted from:							
state	Lower	Center	Upper	———+———+———+———+			
Texas	0.5435	1.0200	1.4965			(—*—)	
				———+———+———+———+			
				-0.80	0.00	0.80	1.60

The first pairing is Florida – Alaska, which results in an interval of (-0.9931, -0.0402). The interval has all negative values indicating that Florida is significantly lower than Alaska. The second pairing is Texas – Alaska, which results in an interval of (0.0269, 0.9798). The interval has all positive values indicating that Texas is greater than Alaska. The third pairing is Texas – Florida, which results in an interval from (0.5435, 1.4965). All positive values indicate that Texas is greater than Florida.

The intervals are similar to the Bonferroni intervals with differences in width due to methods used. In both cases, the same conclusions are reached.

When we use one-way ANOVA and conclude that the differences among the means are significant, we can't be absolutely sure that the given factor is responsible for the differences. It is possible that the variation of some other unknown factor is responsible. One

way to reduce the effect of extraneous factors is to design an experiment so that it has a completely randomized design. This means that each element has an equal probability of receiving any treatment or belonging to any different group. In general good results require that the experiment be carefully designed and executed.

Additional Example:

https://youtu.be/BMyYXc8cWHs
