2.3: Tukey Test for Pairwise Mean Comparisons

If (and only if) we reject the null hypothesis, we then conclude at least one group is different from one other (importantly, we do *not* conclude that all the groups are different).

If it is the case that we reject the null, then we will want to know *which* group or groups are different. In our example we are not satisfied knowing at least one treatment level is different, we want to know where the difference is and the nature of the difference. To answer this question, we can follow up the ANOVA with a mean comparison procedure to find out which means differ from each other and which ones don't.

You might think we could not bother with the ANOVA and proceed with a series of t-tests to compare the groups. While that is intuitively simple, it creates inflation of the type I error. How does this inflation of type I error happen? For a single test,

$$\alpha = 1 - (.95) \tag{2.3.1}$$

The probability of committing a type I error (by random chance) for two simultaneous tests follows from the Multiplication Rule for independent events in probability. Recall that, for two independent events A and B the probability of A and B both occurring is $P(\text{A and B}) = P(\text{A}) * P(\text{B})$. So for two tests, we have

$$\alpha = 1 - ((.95) * (.95)) = 0.0975 \tag{2.3.2}$$

which is now larger than the α that we originally set. For our example, we have 6 comparisons, so $\alpha = 1 - (.95^6) = 0.2649$ which is a much larger (inflated) probability of committing a type I error than we originally set.

The multiple comparison procedures compensate for the type I error inflation (although each does so in a slightly different way).

There are several comparison procedures that can be employed, but we will start with the one most commonly used, the Tukey procedure. In the Tukey procedure, we compute a "yardstick" value based on the MS_{Error} and the number of means being compared. If any two means differ by more than the Tukey w value, then they are significantly different.

Step 1: Compute Tukey's w value

$$w = q_{\alpha(p, df_{Error})} \cdot s_{\bar{Y}} \tag{2.3.3}$$

where q_α is obtained from a table of Tukey q values

p = the number of treatment levels

$s_{\bar{Y}}$ = standard error of a treatment mean = $\sqrt{MS_{Error}/r}$

r = number of replications

Show Tukey q Values Table

df for Error Term	α	p = Number of Treatments								
		2	3	4	5	6	7	3	9	10
5	0.05 0.01	3.64 5.70	4.6 6.98	5.22 7.80	5.67 8.42	6.03 8.91	6.33 9.32	6.58 9.67	6.80 9.97	6.99 10.24
6	0.05 0.01	3.46 5.24	4.34 6.33	4.90 7.03	5.30 7.56	5.63 7.97	5.90 8.32	6.12 8.61	6.32 8.87	6.49 9.10
7	0.05 0.01	3.34 4.95	4.16 5.92	4.68 6.54	5.06 7.01	5.36 7.37	5.61 7.68	5.82 7.94	6.00 8.17	6.16 8.37
8	0.05 0.01	3.26 4.75	4.04 5.64	4.53 6.20	4.89 6.62	5.17 6.96	5.40 7.24	5.60 7.47	5.77 7.68	5.92 7.86
9	0.05 0.01	3.20 4.60	3.95 5.43	4.41 5.96	4.76 6.35	5.02 6.66	5.24 6.91	5.43 7.13	5.59 7.33	5.74 7.49

df for Error Term	α	p = Number of Treatments								
		2	3	4	5	6	7	3	9	10
10	0.05 0.01	3.15 4.48	3.88 5.27	4.33 5.77	4.65 6.14	4.91 6.43	5.12 6.67	5.30 6.87	5.46 7.05	5.60 7.21
11	0.05 0.01	3.11 4.39	3.82 5.15	4.26 5.62	4.57 5.97	4.82 6.25	5.03 6.48	5.20 6.67	5.35 6.84	5.49 6.99
12	0.05 0.01	3.08 4.32	3.77 5.05	4.20 5.50	4.51 5.84	4.75 6.10	4.95 6.32	5.12 6.51	5.27 6.67	5.39 6.81
13	0.05 0.01	3.06 4.26	3.73 4.96	4.15 5.40	4.45 5.73	4.69 5.98	4.88 6.19	5.05 6.37	5.19 6.53	5.32 6.67
14	0.05 0.01	3.03 4.21	3.70 4.89	4.11 5.32	4.41 5.63	4.64 5.88	4.83 6.08	4.99 6.26	5.13 6.41	5.25 6.54
15	0.05 0.01	3.01 4.17	3.67 4.84	4.08 5.25	4.37 5.56	4.59 5.80	4.78 5.99	4.94 6.16	5.08 6.31	5.20 6.44
16	0.05 0.01	3.00 4.13	3.65 4.79	4.05 5.19	4.33 5.49	4.56 5.72	4.74 5.92	4.90 6.08	5.03 6.22	5.15 6.35
17	0.05 0.01	2.98 4.10	3.63 4.74	4.02 5.14	4.30 5.43	4.52 5.66	4.70 5.85	4.86 6.01	4.99 6.15	5.11 6.27
18	0.05 0.01	2.97 4.07	3.61 4.70	4.00 5.09	4.28 5.38	4.49 5.60	4.67 5.79	4.82 5.94	4.96 6.08	5.07 6.20
19	0.05 0.01	2.96 4.05	3.59 4.67	3.98 5.05	4.25 5.33	4.47 5.55	4.65 5.73	4.79 5.89	4.92 6.02	5.04 6.14
20	0.05 0.01	2.95 4.02	3.58 4.64	3.96 5.02	4.23 5.29	4.45 5.51	4.62 5.69	4.77 5.84	4.90 5.97	5.01 6.09
24	0.05 0.01	2.92 3.96	3.53 4.55	3.90 4.91	4.17 5.17	4.37 5.37	4.54 5.54	4.68 5.69	4.81 5.81	4.92 5.92
30	0.05 0.01	2.89 3.89	3.49 4.45	3.84 4.80	4.10 5.05	4.30 5.24	4.46 5.40	4.60 5.54	4.72 5.65	4.83 5.76
40	0.05 0.01	2.86 3.82	3.44 4.37	3.79 4.70	4.04 4.93	4.23 5.11	4.39 5.27	4.52 5.39	4.63 5.50	4.74 5.60

For our greenhouse example we get: $w = q_{.05(4,20)}\sqrt{(3.052/6)} = 3.96(0.7132) = 2.824$

Step 2: Rank the means, calculate differences

For the greenhouse example, we rank the means as:

29.20	28.6	25.87	21.00

Start with the largest and second-largest means and calculate the difference, $29.20 - 28.60 = 0.60$, which is *less* than our w of 2.824, so we indicate there is no significant difference between these two means by placing the letter "a" under each:

29.20	28.6	25.87	21.00
a	a		

Then calculate the difference between the largest and third-largest means, $29.20 - 25.87 = 3.33$, which exceeds the critical w of 2.824, so we can label these with a "b" to show this difference is significant:

29.20	28.6	25.87	21.00
a	a	b	

Now we have to consider whether or not the second-largest and third-largest differ significantly. This is a step that sets up a back-and-forth process. Here $28.6 - 25.87 = 2.73$, less than the critical w of 2.824, so these two means do not differ significantly. We need to add a factor of "b" to show this:

29.20	28.6	25.87	21.00
a	ab	b	

Continuing down the line, we now calculate the next difference: $28.60 - 21.00 = 7.60$, exceeding the critical w, so we now add a "c":

29.20	28.6	25.87	21.00
a	ab	b	c

Again, we need to go back and check to see if the third-largest also differs from the smallest: $25.87 - 21.00 = 4.87$, which it does. So we are done.

These letters can be added to figures summarizing the results of the ANOVA.

The Tukey procedure explained above is valid only with equal sample sizes for each treatment level. In the presence of unequal sample sizes, more appropriate is the Tukey-Cramer Method, which calculates the standard deviation for each pairwise comparison separately. This method is available in SAS, R, and most other statistical softwares.
