

ANOVA: The Tukey Test: A Post Hoc Test (Section 13.2)

Post hoc is a Latin phrase meaning “after this” or “after the event”. Although I was quite certain this was named after the feast served at Thanksgiving, it appears to be named for the statistician John Tukey (1915-2000). So boring. Can’t we name anything after birds?

It is also called the, and this is way better, **Honestly Significant Difference Test** or the **Wholly Significant Difference Test**. When we look for it in StatCrunch, it will be labeled as Tukey HSD.

The idea is that once we determine (as in the previous section through ANOVA) that at least one population of many is not equal to the others, we would like to know which means differ significantly. Sometimes the researcher will have an idea (probably by looking at the boxplots, which coincidentally were invented by Tukey.)

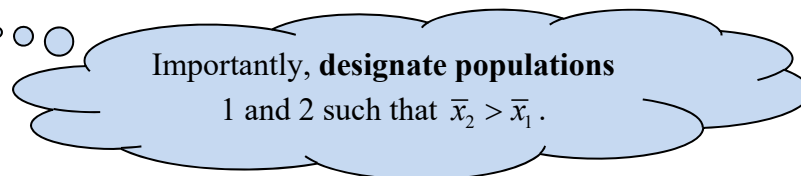
Tukey’s test follows the same logic that we used when we compared two means from independent samples. However, we are *not* just doing Welch’s test as we will see.

Null and Alternative Hypotheses:

We will test the following for two means we suspect are different.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

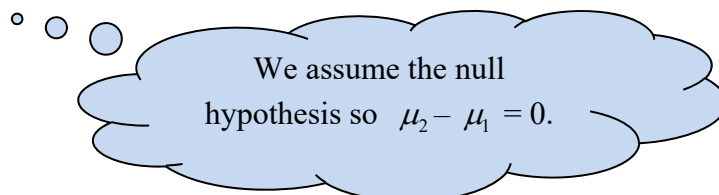


Test Statistic for Tukey’s Test:

The test statistic uses a different calculation for the standard error (denominator below) as when we use Welch’s test. It is $q_0 = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{s^2}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$. Here, s^2 is the mean square

$$q_0 = \frac{(\bar{x}_2 - \bar{x}_1) - (\mu_2 - \mu_1)}{\sqrt{\frac{s^2}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{(\bar{x}_2 - \bar{x}_1)}{\sqrt{\frac{s^2}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

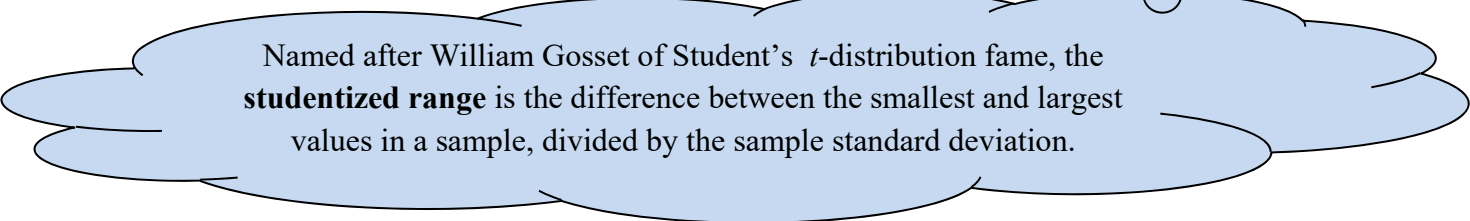
error (MSE) from the ANOVA work done earlier. Recall that was $MSE = \frac{\sum_{i=1}^k (n_i - 1) s_i^2}{n - k}$ where n is the total sample size, n_i are the individual sample sizes, and k is the number of treatments.



So, what do we do with this statistic?

Studentized Range Distribution:

Our q -test statistic follows the **Studentized range distribution**. This distribution's shape depends on two factors, the error degrees of freedom, v calculated as $n - k$, and the total number of means being compared, k . We will compare our test statistic to a critical value based on the level of significance α . We will label this **critical value** $q_{\alpha, v, k}$.



Named after William Gosset of Student's t -distribution fame, the **studentized range** is the difference between the smallest and largest values in a sample, divided by the sample standard deviation.

We might call the level of significance the **experiment-wise** or **familywise error rate**. Recall this is the probability of rejecting the null hypothesis when it's actually true.

The book provides Table X to look up these critical values for $\alpha = 0.01$ and 0.05 . If the value of v is *not* in the table, use the closest you can. You can look up how to use this table but we will focus on using technology.

Tukey's Test By Hand:

After rejecting the null hypothesis in an ANOVA test, we will do the following to compare pairs of means.

Step 1: Arrange the sample means in increasing order.

Step 2: Compute *each* pairwise difference $\bar{x}_i - \bar{x}_j$ where $\bar{x}_i > \bar{x}_j$. (For example, if there are four samples, we would do this a total of three times for the lowest mean.) Form a table with these differences in decreasing order (for ease in step 5).

Step 3: Compute the test statistic $q_0 = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{\frac{s^2}{2} \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$ for *each* pairwise difference. (Again, for

four samples, we would do this $3 + 2 + 1$ or 6 times.)

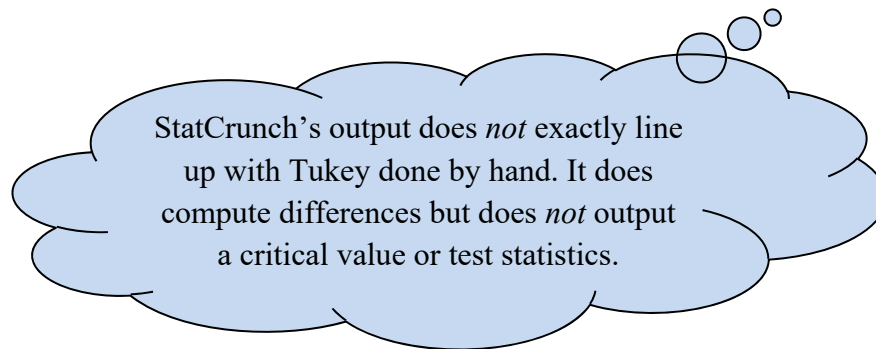
Step 4: Determine the critical value $q_{\alpha, v, k}$ which is the same for all.

Step 5: If $q_0 \geq q_{\alpha, v, k}$ for a pair, reject the null hypothesis that those two population means are equal. We conclude that our evidence shows that those two means are significantly different. (If you arranged step 2 differences in decreasing order, and you do *not* reject a null hypothesis, you can *automatically not* reject those under it, as it will have a smaller difference and so should *not* be rejected too.)

Step 6: State which means are significantly different and which are *not*. We will group means that are *not* significantly different by underlining.

Instructions for StatCrunch:

1. Enter the data for each sample or treatment in a separate column. Label the columns. As an alternative, you can put all values in one column and then use a second column for indicator variables for each sample.
2. Select **Stat > ANOVA > One Way**.
3. If you used multiple columns for each sample, choose to **Compare: Selected columns**. If you used the alternative way to enter data, you select **Values in a single column**. In either case, tell it which columns contain the data. (Under **Options**, select **Compute Tukey HSD** and enter the **Level** as $1 - \alpha$.) Click **Compute!**.
4. The output will show each difference between a mean and means greater than it. You will also see **Lower** and **Upper** numbers given as well as a **P-value** for each difference. What's all that about?



Interpreting the StatCrunch Output:

Lower and **Upper**: These are the lower and upper bounds of a confidence interval for the difference of the two means. **Recall, that if this interval contains zero, then we do *not* reject the null hypothesis** (as our evidence shows the difference could be 0 and therefore the means are equal). Inversely, if the interval does *not* contain zero, then we reject the null hypothesis and we conclude that the means are significantly different.



P-value: Alternatively, you can use the *P*-values given, recalling that we reject a null hypothesis if this *P*-value is less than α .

expl 1: A farmer performed a randomized experiment to see which of three types of plots used for planting corn is best. She randomly picked six rows of corn using each of the three different plots and then measured the mean number of plants for each plot as given on the right.

Type of Plot	Sample Mean
Sludge plot	28.8 plants
Spring disc	33.8 plants
No till	26.8 plants

A hypothesis test ($H_0: \mu_{SP} = \mu_{SD} = \mu_{NT}$) was performed and the null hypothesis was rejected. Perform Tukey's test to determine which means should be considered significantly different. Here is the output as given in StatCrunch. (I used the original sample data *not* shown here.)

Options



Column statistics

Column	n	Mean	Std. Dev.	Std. Error
Sludge_Plot	6	28.833333	3.4880749	1.4240006
Spring_Disk	6	33.833333	1.1690452	0.4772607
No_Till	6	26.833333	2.4013885	0.98036274

ANOVA table

Source	DF	SS	MS	F-Stat	P-value
Columns	2	156	78	12.124352	0.0007
Error	15	96.5	6.4333333		
Total	17	252.5			

Tukey HSD results (95% level)

Sludge_Plot subtracted from

	Difference	Lower	Upper	P-value
Spring_Disk	5	1.1962849	8.8037151	0.0101
No_Till	-2	-5.8037151	1.8037151	0.3828

Spring_Disk subtracted from

	Difference	Lower	Upper	P-value
No_Till	-7	-10.803715	-3.1962849	0.0007

To help analyze Tukey's test, write down μ_{SP} , μ_{SD} , and μ_{NT} in order according to their sample means given to the left.

Glance at the ANOVA table and confirm that its P -value is such that the null would have been rejected.

Use the P -values in the topmost Tukey results table to compare μ_{SP} with both μ_{SD} and μ_{NT} . Use $\alpha = 0.05$. Draw a line underlining a pair if the null hypothesis is *not* rejected. (We do *not* have evidence to say those means are *not* equal.) Do the same for the lower Tukey results table to compare μ_{SD} and μ_{NT} .

We will use this underlining notation but also, we will write statements similar to $\mu_{SP} = \mu_{SD} \neq \mu_{NT}$. This is *not* the correct conclusion; write your conclusion from above in this form.

The homework will require you to perform some ANOVA tests and to create boxplots too.

Don't Be a Turkey: Some Cautions Regarding Tukey's Test:

expl 2: Here is the output for the data given in example 3 in Section 13.1 Notes. The data is to the right for reference.

Method I	Method II	Method III	Method IV
79	83	85	82
77	56	60	92
86	83	74	85
75	79	71	59
85	62	67	80
75	41	76	69
79	59		
75	64		
80			
94			

Options				
Columns	5	100.0000	50.0000	50.0000
Error	26	2947.0417	113.34776	
Total	29	4001.2		
Tukey HSD results (95% level)				
Method I subtracted from				
	Difference	Lower	Upper	P-value
Method II	-14.625	-28.478968	-0.77103231	0.0357
Method III	-8.3333333	-23.415623	6.7489563	0.4429
Method IV	-2.6666667	-17.748956	12.415623	0.9617
Method II subtracted from				
	Difference	Lower	Upper	P-value
Method III	6.2916667	-9.4817694	22.065103	0.6961
Method IV	11.958333	-3.8151028	27.731769	0.1862
Method III subtracted from				
	Difference	Lower	Upper	P-value
Method IV	5.6666667	-11.195846	22.529179	0.7934

Analyze the P -values for each of the pairwise differences. Use $\alpha = 0.05$. I have written the population means in order according to the sample means.

$$\mu_{II} \quad \mu_{III} \quad \mu_{IV} \quad \mu_I$$

Notice how the underlinings overlap. We saw that the population means for methods I, III, and IV could be considered equal (but *not* method II) *but at the same time* methods II, III, and IV are considered equal?? That does *not* seem logically possible, does it? The result is ambiguous.

This means that at least one Type II error (*not* rejecting H_0 when it is false) has been committed by the Tukey test. We will conclude that $\mu_{III} = \mu_{IV}$ and $\mu_I \neq \mu_{II}$. We do *not* have conclusive evidence of how μ_{III} and μ_{IV} are related to μ_I and μ_{II} .

A larger sample size may increase the test's power and resolve this ambiguity.

It may also happen that ANOVA rejects the null hypothesis but Tukey does *not* find any pairs to be significantly different. This is because ANOVA is more powerful than Tukey. A larger sample size should help clarify the situation.