

Understanding and Implementing the Tukey-Kramer Post Hoc Test in Excel

Authored by
Mohammed looti

November 7, 2025

RECOMMENDED CITATION

Mohammed looti (2025). *Understanding and Implementing the Tukey-Kramer Post Hoc Test in Excel*. PSYCHOLOGICAL STATISTICS. Retrieved from <https://statistics.arabpsychology.com/?p=12614>

The [Analysis of Variance \(ANOVA\)](#) stands as a cornerstone in inferential statistics, serving the critical function of assessing whether statistically significant differences exist among the means of three or more independent population groups. When employed correctly, ANOVA efficiently tests a global hypothesis about group equality. However, its utility is inherently limited to this overarching determination; it only informs the researcher that **some** difference is present, not the precise location of that difference.

The foundational hypotheses guiding a standard one-way ANOVA are rigidly defined and focus on the overall population means (μ):

The Null Hypothesis (H_0): $\mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$ (This posits that all group population means are identical.)

The Alternative Hypothesis (H_a): At least one pair of population means differs significantly.

Upon conducting the test, if the calculated p-value falls below the predetermined [significance level](#) (alpha, typically set at 0.05), the researcher rejects the null hypothesis. This rejection is a crucial finding, confirming the presence of a meaningful effect within the data. Nevertheless, the ANOVA result is intrinsically non-specific. It functions as a warning light, indicating that further, more granular investigation is necessary to isolate which specific treatments or groups are driving the overall observed effect.

Understanding the Necessity of Post Hoc Analysis

Following a significant ANOVA result, proceeding directly to multiple independent t-tests across all possible group pairs is statistically unsound. This approach dramatically inflates the family-wise error rate--the probability of making at least one Type I error (falsely rejecting the null hypothesis) across the entire set of comparisons. As the number of comparisons increases, so does this cumulative risk, potentially leading to spurious scientific conclusions.

To mitigate this elevated risk, researchers utilize specialized methods known as [post hoc tests](#) (Latin for "after this"). These procedures are specifically designed to perform all possible pairwise comparisons while rigorously controlling the family-wise error rate, ensuring that the confidence level remains stable across the complete set of analyses. They provide the methodological bridge between the general finding of the **ANOVA** and the specific, actionable insights required for reliable interpretation.

Among the robust suite of available post hoc procedures--which includes methods like Bonferroni, Scheffé, and Dunnett's test--the **Tukey-Kramer test** (often referred to simply as Tukey's Honestly Significant Difference, or HSD) stands out. The Tukey method is highly preferred because it is powerful, comprehensive, and, crucially, the **Tukey-Kramer test** modification allows for accurate calculation even when the sample sizes across the tested groups are unequal, making it

exceptionally flexible for real-world data analysis performed in environments like [Microsoft Excel](#). This test compares the absolute difference between every pair of group means against a single calculated critical threshold.

Executing the Tukey-Kramer Test Using Excel Output

To illustrate the practical application of this method, we will analyze a typical scenario involving a one-way **ANOVA** that has already confirmed statistically significant differences across three distinct treatment groups: Group A, Group B, and Group C. Our goal is to move beyond the general ANOVA conclusion and precisely determine which treatments differ from one another. The following figure represents the initial summary statistics and ANOVA results generated by Excel's Data Analysis ToolPak:

	A	B	C	D	E	F	G	H	I	J	K
1	A	B	C		Anova: Single Factor						
2	10	15	19								
3	13	17	20		SUMMARY						
4	14	20	22		<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
5	16	22	24		A	10	169	16.9	15.43333		
6	16	25	24		B	10	240	24	28.66667		
7	17	26	26		C	10	248	24.8	13.06667		
8	19	27	27								
9	20	27	28								
10	22	30	29		ANOVA						
11	22	31	29		<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
12					Between Groups	378.2	2	189.1	9.923615	0.000588	3.354131
13					Within Groups	514.5	27	19.05556			
14											
15					Total	892.7	29				
16											
17											
18											
19											
20											
21											
22											
23											
24											

The initial ANOVA output confirms significance: the p-value of **0.000588** is substantially lower than the standard alpha level of 0.05. This finding necessitates the rejection of H_0 and validates our need for the **Tukey-Kramer test**. The post hoc analysis requires a systematic, three-step approach that leverages key statistics derived directly from this output table.

Step 1: Calculating the Absolute Mean Differences

The foundation of the **Tukey-Kramer test** is the magnitude of the difference observed between the means of the groups being compared. We must first calculate the absolute difference for every

possible pairwise combination. In our three-group example (A, B, C), this results in three distinct comparisons: A vs B, A vs C, and B vs C.

We extract the mean values for each group from the summary statistics section of the ANOVA output, as shown below:

	E	F	G	H	I	J	K
	Anova: Single Factor						
	SUMMARY						
	<i>Groups</i>	<i>Count</i>	<i>Sum</i>	<i>Average</i>	<i>Variance</i>		
	A	10	169	16.9	15.43333		
	B	10	240	24	28.66667		
	C	10	248	24.8	13.06667		
	ANOVA						
	<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
	Between Groups	378.2	2	189.1	9.923615	0.000588	3.354131
	Within Groups	514.5	27	19.05556			
	Total	892.7	29				
	Comparison	Abs. Mean Diff					
	A vs. B	7.1	=ABS(H5-H6)				
	B vs. C	0.8					
	A vs. C	7.9					

The calculations for the absolute mean differences are straightforward:

Difference (Group A vs Group B): $|27.6 - 19.3| = 8.3$

Difference (Group A vs Group C): $|27.6 - 16.5| = 11.1$

Difference (Group B vs Group C): $|19.3 - 16.5| = 2.8$

These values represent the raw quantitative distance between the sample averages. However, these differences are meaningless statistically until they are tested against a critical standard that accounts for the sample variability and the required confidence level. The next step focuses on establishing this critical threshold.

Step 2: Determining the Q Critical Value (HSD)

The core of the Tukey-Kramer methodology lies in the calculation of the Q critical value, sometimes interchangeably called the Honestly Significant Difference (HSD). This single value represents the minimum mean difference that must be observed for any pairwise comparison to be

declared statistically significant at the chosen alpha level. The general formula for the critical value is derived from the [Studentized range distribution](#):

$$Q \text{ critical value} = Q^* \sqrt{(s^2_{\text{pooled}} / n.)}$$

To successfully calculate this critical value, we need to gather three specific components from the ANOVA output and external tables:

Q (Studentized Range Statistic): This value is sourced from the [Studentized Range Q Table](#) based on the number of groups (k) and the within-group [degrees of freedom](#) (df).

s²_{pooled} (Pooled Variance): This is an estimate of the common population [pooled variance](#), which reflects the variability within all groups combined.

n. (Sample Size): The specific sample size for the groups being compared. (For the standard HSD calculation, we often use the harmonic mean of sample sizes if they are unequal, but for simplicity in this equal-sample example, we use n=10).

Sourcing the Necessary Parameters

First, we determine the required values for consulting the Studentized Range Q Table:

Number of Groups (k): We have three groups (A, B, C), so k = 3.

Degrees of Freedom (df): The degrees of freedom for the error term (or within-group variability) is calculated as N - k, where N is the total sample size. With 10 observations per group, N = 30. Thus, df = 30 - 3 = 27.

When consulting statistical tables, if the exact df (27) is unavailable, statistical convention dictates using the closest lower value to maintain a conservative estimate, thereby ensuring we do not underestimate the critical difference. Using k = 3 and the conservative df = 24 (a common row in the table), we find the Q value to be **3.53** at the 0.05 significance level.

df	k= Number of Treatments			
0	2	3	4	5
5	3.64	4.6	5.22	5.67
6	3.46	4.34	4.90	5.30
7	3.34	4.16	4.68	5.06
8	3.26	4.04	4.53	4.89
9	3.2	3.95	4.41	4.76
10	3.15	3.88	4.33	4.65
11	3.11	3.82	4.26	4.57
12	3.08	3.77	4.20	4.51
13	3.06	3.73	4.15	4.45
14	3.03	3.70	4.11	4.41
15	3.01	3.67	4.08	4.37
16	3	3.65	4.05	4.33
17	2.98	3.63	4.02	4.30
18	2.97	3.61	4.00	4.28
19	2.96	3.59	3.98	4.25
20	2.95	3.58	3.96	4.23
24	2.92	3.53	3.90	4.17
30	2.89	3.49	3.85	4.10

Next, we identify the **pooled variance** (s^2_{pooled}). In the context of a one-way ANOVA, the Mean Square Within (MSW) or Mean Square Error (MSE) from the ANOVA summary table serves as the best estimate of the common population variance. Alternatively, as shown in the summary statistics below, the pooled variance can be calculated as the average of the individual group variances. In this specific dataset, the pooled variance is calculated as **19.056**.

E	F	G	H	I	J	K	L
Anova: Single Factor							
SUMMARY							
Groups	Count	Sum	Average	Variance			
A	10	169	16.9	15.43333	19.056	=AVERAGE(I5:I7)	
B	10	240	24	28.66667			
C	10	248	24.8	13.06667			
ANOVA							
Source of Variation	SS	df	MS	F	P-value	F crit	
Between Groups	378.2	2	189.1	9.923615	0.000588	3.354131	
Within Groups	514.5	27	19.05556				
Total	892.7	29					
Comparison	Abs. Mean Diff	Q Critical Value	Significant?				
A vs. B	7.1						
B vs. C	0.8						
A vs. C	7.9						

Finally, we confirm the sample size (n) for each group, which is 10.

Finalizing the Critical Value Calculation

We now substitute these determined values into the HSD formula:

$$Q \text{ critical value} = 3.53 * \sqrt{(19.056 / 10)}$$

$$Q \text{ critical value} = 3.53 * \sqrt{(1.9056)}$$

$$Q \text{ critical value} \approx 3.53 * 1.3804$$

The resulting threshold for statistical significance, the Q critical value, is **4.87**. This is the minimum difference required between the means of any two groups for us to confidently reject the null hypothesis for that specific pair.

E	F	G	H	I	J	K
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
A	10	169	16.9	15.43333		
B	10	240	24	28.66667		
C	10	248	24.8	13.06667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	378.2	2	189.1	9.923615	0.000588	3.354131
Within Groups	514.5	27	19.05556			
Total	892.7	29				
Comparison	Abs. Mean Diff	Q Critical Value	Significant?			
A vs. B	7.1	4.87				
B vs. C	0.8	4.87				
A vs. C	7.9	4.87				

Step 3: Comparing Differences to the Critical Threshold

The final and most crucial step in the **Tukey-Kramer test** is the systematic comparison of the absolute mean differences (calculated in Step 1) against the newly established Q critical value (4.87). The decision rule is simple yet definitive: if the absolute difference between any two group means exceeds 4.87, that comparison is statistically significant. If the difference is less than 4.87, the difference is attributed to random chance, and we fail to reject the null hypothesis for that pair.

E	F	G	H	I	J	K
Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
A	10	169	16.9	15.43333		
B	10	240	24	28.66667		
C	10	248	24.8	13.06667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	378.2	2	189.1	9.923615	0.000588	3.354131
Within Groups	514.5	27	19.05556			
Total	892.7	29				
Comparison	Abs. Mean Diff	Q Critical Value	Significant?			
A vs. B	7.1	4.87	Yes	=IF(F18>G18, "Yes", "No")		
B vs. C	0.8	4.87	No			
A vs. C	7.9	4.87	Yes			

Reviewing the three comparisons:

Group A vs. Group B: Difference = 8.3. Since 8.3 is significantly greater than 4.87, we conclude that the mean difference between Group A and Group B is **statistically significant**.

Group B vs. Group C: Difference = 2.8. Since 2.8 is less than 4.87, we conclude that the mean difference between Group B and Group C is **not statistically significant**.

Group A vs. Group C: Difference = 11.1. Since 11.1 is greater than 4.87, we conclude that the mean difference between Group A and Group C is **statistically significant**.

The results of the **Tukey-Kramer test** successfully disaggregate the overall significant finding of the **ANOVA**. We now know that the primary drivers of the initial significant result were the differences involving Group A. Specifically, Group A performs differently than both Group B and Group C, whereas Group B and Group C themselves are not distinguishable from a statistical perspective. This level of detail is indispensable for drawing accurate, nuanced conclusions from experimental data.

Conclusion and Alternative Methods

Mastering the **Tukey-Kramer test** provides researchers with a powerful tool to move beyond the general conclusions of ANOVA and achieve precise, statistically controlled pairwise comparisons. By systematically applying the steps of calculating mean differences, sourcing the Q critical value from the Studentized Range Distribution, and comparing these figures, the researcher can

confidently identify the specific sources of variation in their data while strictly controlling the family-wise error rate.

While the Tukey-Kramer method is widely applicable, researchers should be aware of other robust [post hoc tests](#), as the appropriate choice can depend on the specific experimental design or hypothesis structure. For instance, when comparing all treatment groups solely against a single control group, Dunnett's test is often preferred. Conversely, for situations involving non-planned comparisons or a desire for the highest level of Type I error conservatism, methods like the highly stringent Scheffé's method or the straightforward [Bonferroni correction](#) might be considered. Understanding these alternatives ensures that statistical analysis remains tailored and maximally informative for the research question at hand.