

Treatment →	A	B	C	D
	5881.0			
	5842.0			
	6578.0			
	5613.0			
	5677.0			
	6696.0			
	6514.0			
	7474.0			
	7874.0			
	6373.0			
	7191.0			
	6756.0			
	6128.0			
	5542.0			
	4978.0			
	6465.0			
	4346.0			
	7588.0			
	5414.0			
	5507.0			
	4407.0			
	6302.0			

Descriptive statistics of your $k=4$ independent treatments:

Treatment →	A	B	C	D	Pooled
observations N	9038	6949	5282	5619	26887
sum $\sum x_i$	56,183,673.0000	30,706,506.0000	28,425,056.0000	27,763,544.0000	143,078,779.0000
mean \bar{x}	6,216.3834	4,418.8381	5,381.4949	4,941.0116	5,381.4949
sum of squares $\sum x_i^2$	357,328,625,999.0000	144,612,524,166.0000	157,567,452,774.0000	144,585,166,482.0000	804,093,769,779.0000
sample variance s^2	892,926.2878	1,284,606.4112	870,698.5818	1,318,115.7729	1,589,426.2878
sample std. dev. s	944.9478	1,133.4048	933.1123	1,148.0922	1,260.7208
std. dev. of mean $SE_{\bar{x}}$	9.9397	13.5964	12.8391	15.3161	12.6072

One-way ANOVA of your $k=4$ independent treatments:

source	sum of squares SS	degrees of freedom ν	mean square MS	F statistic	p-value
treatment	13,732,293,658.0729	3	4,577,431,219.3576	4,243.7067	1.1102e-16
error	28,998,153,830.0634	26884	1,078,639.8538		
total	42,730,447,488.1362	26887			

Conclusion from Anova:

The p-value corresponding to the F-statistic of one-way ANOVA is lower than 0.05, suggesting that the one or more treatments are significantly different. The Tukey HSD test, Scheffé, Bonferroni and Holm multiple comparison tests follow. These post-hoc tests would likely identify which of the pairs of treatments are significantly different from each other.

Tukey HSD Test:

The p-value corresponding to the F-statistic of one-way ANOVA is lower than 0.01 which strongly suggests that

one or more pairs of treatments are significantly different. You have $k = 4$ treatments, for which we shall apply Tukey's HSD test to each of the 6 pairs to pinpoint which of them exhibits statistically significant difference.

We first establish the critical value of the Tukey-Kramer HSD Q statistic based on the $k = 4$ treatments and $\nu = 26884$ degrees of freedom for the error term, for significance level $\alpha = 0.01$ and 0.05 (p-values) in the Studentized Range distribution. We obtain these critical values for Q , for α of 0.01 and 0.05 , as $Q_{critical}^{\alpha=0.01, k=4, \nu=26884} = 4.4033$ and $Q_{critical}^{\alpha=0.05, k=4, \nu=26884} = 3.6334$, respectively. These critical values may be verified at several published tables of the inverse Studentized Range distribution, such as [this table at Duke University](#).

Next, we establish a Tukey test statistic from our sample columns to compare with the appropriate critical value of the studentized range distribution. We take the Tukey-Kramer confidence limits as documented in the [NIST Engineering Statistics Handbook](#) and make simplifying algebraic transformation. We calculate a parameter for each pair of columns being compared, which we loosely call here as the Tukey-Kramer HSD Q -statistic, or simply the Tukey HSD Q -statistic, as:

$$Q_{i,j} = \frac{|\bar{x}_i - \bar{x}_j|}{s_{i,j}}$$

where the denominator in the above expression is:

$$s_{i,j} = \frac{\hat{\sigma}_\epsilon}{\sqrt{H_{i,j}}} \quad i, j = 1, \dots, k; \quad i \neq j.$$

The quantity $H_{i,j}$ is the [harmonic mean](#) of the number of observations in columns labeled i and j . Note that when the sample sizes in the columns are equal, then their harmonic mean is simply the common sample size. When the sample sizes of columns in a pair being compared are different, the harmonic mean lies somewhere in-between the two sample sizes. The relevant harmonic mean is required for applying the Tukey-Kramer procedure for columns with unequal sample sizes. The quantity $\hat{\sigma}_\epsilon = 1,038.5759$ is the square root of the Mean Square Error $= 1,078,639.8538$ determined in the precursor one-way ANOVA procedure. Note that $\hat{\sigma}_\epsilon$ is same across all pairs being compared. The only factor that varies across pairs in the computation of $s_{i,j} = \frac{\hat{\sigma}_\epsilon}{\sqrt{H_{i,j}}}$ is the denominator, which is the harmonic mean of the sample sizes being compared.

The test of whether the NIST Tukey-Kramer confidence interval includes zero is equivalent to evaluating whether $Q_{i,j} > Q_{critical}$, the latter determined according to the desired level of significance α (p-value), the number of treatments k and the degrees of freedom for error ν , as described above.

post-hoc Tukey HSD Test Calculator results:

$k = 4$ treatments

degrees of freedom for the error term $\nu = 26884$

Critical values of the Studentized Range Q statistic:

$$Q_{critical}^{\alpha=0.01, k=4, \nu=26884} = 4.4033 \quad Q_{critical}^{\alpha=0.05, k=4, \nu=26884} = 3.6334$$

We present below color coded results (red for insignificant, green for significant) of evaluating whether $Q_{i,j} > Q_{critical}$ for all relevant pairs of treatments. In addition, we also present the significance (p-value) of the observed Q -statistic $Q_{i,j}$. The algorithm used here to calculate the critical values of the studentized range distribution, as well as p-values corresponding to an observed value of $Q_{i,j}$, is that of [Gleason \(1999\)](#). This is an improvement over the [Copenhaver-Holland \(1988\) algorithm](#) deployed in [the R statistical package](#).

Tukey HSD results

treatments pair	Tukey HSD Q statistic	Tukey HSD p-value	Tukey HSD inference
A vs B	153.4159	0.0010053	** p<0.01
A vs C	65.6401	0.0010053	** p<0.01
A vs D	102.2249	0.0010053	** p<0.01
B vs C	71.8088	0.0010053	** p<0.01
B vs D	39.6323	0.0010053	** p<0.01
C vs D	31.2969	0.0010053	** p<0.01

Scheffé multiple comparison

We define a statistic named T as the ratio of unsigned contrast mean to contrast standard error, as explained in the [NIST Engineering Statistics Handbook page for Scheffe's method](#). It can be shown that for contrasts that are treatment pairs (i, j) with unit coefficients,

$$T_{i,j} = \frac{Q_{i,j}}{\sqrt{2}}$$

where $Q_{i,j}$ is the Q -statistic that was created for the Tukey HSD test. This T -statistic has interesting properties.

The same [NIST Engineering Statistics Handbook page for Scheffe's method](#) provides a formula which directly leads to the Scheffé p-value corresponding to an observed value of T as:

$$1 - F\left(\frac{T^2}{k-1}, k-1, \nu\right)$$

where $F()$ is the cumulative F distribution with its two degrees of freedom parameters $k-1$ and ν . Note that k is the number of treatments and ν is the degrees of freedom of error that were established earlier.

The Scheffé p-value of the observed T -statistic $T_{i,j}$ is shown below for all relevant pairs of treatments, along with color coded Scheffé inference (red for insignificant, green for significant) based on the p-value.

Scheffé results

treatments pair	Scheffé T -statistic	Scheffé p-value	Scheffé inference
A vs B	108.4814	1.1102e-16	** p<0.01
A vs C	46.4145	1.1102e-16	** p<0.01
A vs D	72.2839	1.1102e-16	** p<0.01
B vs C	50.7765	1.1102e-16	** p<0.01
B vs D	28.0243	1.1102e-16	** p<0.01
C vs D	22.1303	1.1102e-16	** p<0.01

Bonferroni and Holm multiple comparison

The same statistic T for the Scheffé method, along with the number of contrasts (pairs) q being simultaneously compared, leads to the Bonferroni formula. The [NIST Engineering Statistics Handbook page for Bonferroni method](#) provides a formula which directly leads to the Bonferroni p-value corresponding to an observed value of T in the context of simultaneous comparison of q contrasts as:

Bonferroni p-value: $P_{i,j}^{\text{Bonferroni}} = P_{i,j}^{\text{unadjusted}} q$ where

$$P_{i,j}^{\text{unadjusted}} = \left[1 - t\left(\frac{T^2}{k-1}, \nu\right) \right]^2$$

and where $t()$ is the cumulative Student's t distribution with its degree of freedom parameter ν . Note that ν is the degrees of freedom of error that were established earlier. Also note that the p-value of Bonferroni simultaneous comparison is directly proportional to q , the number of contrasts (pairs) being simultaneously compared.

The Holm procedure described in [Aickin and Gensler \(1996\) review paper](#) requires sorting the $P_{i,j}^{\text{unadjusted}}$ as above in *ascending* order and determining the sort rank $R_{i,j}$ of each unique pair (i, j) . These sort ranks run from 1 through q . The Holm p-value for comparing a given pair (i, j) in the context of multiple comparison of q such pairs simultaneously is:

Holm p-value: $P_{i,j}^{\text{Holm}} = P_{i,j}^{\text{unadjusted}} (q - R_{i,j} + 1)$

In this first combined Bonferroni and Holm table below, we consider all possible contrasts (pairs) for simultaneous comparison, thus $q=6$. The Bonferroni and Holm p-values of the observed T -statistic $T_{i,j}$ for all relevant $q=6$ pairs of treatments is shown below, along with color coded Bonferroni and Holm inferences (red for insignificant, green for significant) based on the p-value.

Bonferroni and Holm results: all pairs simultaneously compared

treatments pair	Bonferroni and Holm T -statistic	Bonferroni p-value	Bonferroni inference	Holm p-value	Holm inference
A vs B	108.4814	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
A vs C	46.4145	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
A vs D	72.2839	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
B vs C	50.7765	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
B vs D	28.0243	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01

C vs D	22.1303	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
--------	---------	------------	-----------	------------	-----------

In this second Bonferroni and Holm table below, we consider a subset of contrasts (pairs) for simultaneous comparison, of only pairs relative to treatment A. Such a situation may be relevant when treatment A is the control, and the experimenter is interested only in differences of treatments relative to control, thus $q=3$. The Bonferroni and Holm p-values of the observed T -statistic $T_{i,j}$ for $q=3$ relevant pairs of treatments, along with color coded Bonferroni inference (red for insignificant, green for significant) based on the p-value.

Bonferroni and Holm results: only pairs relative to A simultaneously compared

treatments pair	Bonferroni and Holm T -statistic	Bonferroni p-value	Bonferroni inference	Holm p-value	Holm inference
A vs B	108.4814	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
A vs C	46.4145	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01
A vs D	72.2839	0.0000e+00	** p<0.01	0.0000e+00	** p<0.01

How to repeat & verify post-hoc Tukey HSD calculation by hand in Microsoft Excel

Microsoft Excel lacks built-in functions relating to the studentized range distribution, so even though it calculates the Mean Square Error in one-way ANOVA, whose square root is $\hat{\sigma}_\epsilon$, and is aware of all the sample sizes and degrees of freedom, it is unable to conduct the next step of post-hoc Tukey HSD comparison of treatments. To manually conduct post-hoc Tukey HSD test calculation, you would take the mean squared error from the Excel's one-way ANOVA output, then take its square root to determine $\hat{\sigma}_\epsilon$. You would then divide this $\hat{\sigma}_\epsilon$ by the square root of $H_{i,j}$, the harmonic mean of the relevant sample columns being compared, resulting in $s_{i,j}$, for each pair of columns (i,j) . Excel has a [built-in function HARMEAN\(n1, n2, ...\)](#) that calculates the harmonic mean. With these on hand, you would determine $Q_{i,j} = \frac{|\bar{x}_i - \bar{x}_j|}{s_{i,j}}$. Microsoft Excel provides the relevant sample column averages (means) to calculate the numerator. You would finally compare whether $Q_{i,j} > Q_{critical}$. For this comparison, you would obtain the critical values for the appropriate number of degrees of freedom of error (shown in Microsoft Excel) and the number of treatments (columns) from tables of the (inverse) studentized range distribution widely available on the web.


How to repeat & verify post-hoc Scheffé, Bonferroni and Holm calculations by hand in Microsoft Excel

For Scheffé, Bonferroni and Holm steps, calculate the T -statistic $T_{i,j}$ for all pairs by taking $Q_{i,j}$ from the earlier Tukey HSD step and dividing it by $\sqrt{2}$. Calculate the Scheffé p-value of the observed T -statistic by using Excel's [built-in formula for the F distribution](#) which has the form `F.DIST(x, deg_freedom1, deg_freedom2, cumulative)`. Set its first argument x as $\frac{T^2}{k-1}$. The second and third arguments are $k-1$ and ν . The fourth argument is set to 1. The Scheffé p-value is calculated in Excel by the formula as `1-F.DIST(x, k-1, nu, 1)`.

For the Bonferroni and Holm comparison, take the same T -statistic $T_{i,j}$ that was determined for the Scheffé step above. Determine q , the number of pairs that are being simultaneously compared. Calculate the Bonferroni p-value of the observed T -statistic by using Excel's [built-in formula for the t-distribution T.DIST\(x, deg_freedom, cumulative\)](#). The unadjusted p-value $P_{i,j}^{unadjusted}$ is calculated in Excel by the formula as `(1-T.DIST(T, nu, TRUE)) * 2`. The Bonferroni p-value is calculated for each pair (i,j) as $P_{i,j}^{Bonferroni} = P_{i,j}^{unadjusted} q$. The q -element array of the $P_{i,j}^{unadjusted}$ is sorted in *ascending* order to determine the sort rank $R_{i,j}$ for each given pair (i,j) . These sort ranks run from 1 through q . The Holm p-value is calculated for each pair (i,j) as $P_{i,j}^{Holm} = P_{i,j}^{unadjusted} (q - R_{i,j} + 1)$.

R code and Tutorial for conducting Tukey HSD, Scheffé, Bonferroni and Holm methods

A tutorial for the solving the demo example using the free open-source academic-research-grade [R statistical package](#) together with complete [R code and output is provided here](#). The output of the demo example in this web calculator for all three methods of multiple comparison are fully reproduced in R, thus further establishing the validity of the formula and methodology discussed earlier.

Attribution:  2016 Navendu Vasavada
navendu (dot) vasavada (at) alumni (dot) upenn (dot) edu