

Analysis of Variance

Daniel J. Teague

NCSSM Statistics Leadership Institute

In this section we consider Analysis of Variance (ANOVA) as an extension of the two-sample t -test. We will use a vector/matrix representation described in *Statistics for Experimenters*, by Box, Hunter, and Hunter (see References). This vector/matrix representation attempts to clarify the sums of squares computations and illustrates nicely how blocking reduces the variability.

Example Problem: Two types of popcorn are being compared. Five hundred kernels of one type are placed in an automatic popcorn popper. The popper is turned on for five minutes and the number of unpopped kernels is counted. Sufficient time is left between treatments to allow the popper to cool down. Four bags of both types are popped one at a time, with the order determined at random. This is a completely randomized design. The results are given in the table below and will be analyzed using a two-sample t -test with pooled variance.

Type A	52	60	56	52
Type B	44	50	52	42

Two-sample Analysis

The hypotheses were are to test are

$$H_0: \mathbf{m}_A = \mathbf{m}_B$$

$$H_a: \mathbf{m}_A \neq \mathbf{m}_B$$

We will use the decision criterion $\alpha = 0.05$ throughout this discussion. We are assuming that the response variable, number of unpopped kernels, has a distribution that is not significantly non-normal. That is, there is no reason to call into question the use of the t -test in this situation. There is no evidence of outliers in either set. There are only 4 data points in each set, but we hope this lack of realism is made up by the clarity of the example.

From the data we have

$$\begin{aligned} \bar{x}_A &= 55, s_A = 3.8297 \\ \bar{x}_B &= 47, s_B = 4.7610 \end{aligned} \quad \text{with } s_p^2 = \frac{3(3.8297)^2 + 3(4.7610)^2}{6} = 18.667.$$

So,

$$t_6 = \frac{(55 - 47) - (\mathbf{m}_A - \mathbf{m}_B)}{\sqrt{18.667} \sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{8}{3.055} = 2.6186.$$

With 6 degrees of freedom, the p -value for this two-sided test is $p = 0.0397$. Based on this low p -value, we reject the null hypothesis of no difference in population means. The observed means are too disparate to reasonably be considered the results of a random draw from a common distribution.

The ANOVA Approach

Now we will consider the same problem, using the ANOVA approach. We create a vector of the data $[52, 60, 56, 52, | 44, 50, 52, 42]^T$. We have put in a vertical bar (|) to separate the values of the two sets of data. A more convenient method is to write the vector in a “matrix” format as shown below. This helps us keep the data separate and will help clarify the ANOVA technique. Even though we write the data using matrix notation, we operate on it as the vector that it actually is.

$$X = \begin{matrix} & \mathbf{A} & \mathbf{B} \\ \begin{bmatrix} 52 & 44 \\ 60 & 50 \\ 56 & 52 \\ 52 & 42 \end{bmatrix} \end{matrix}$$

ANOVA Model

The mathematical model for ANOVA is $X = \mathbf{m} + \mathbf{t} + \mathbf{e}$. What we see, X , is decomposed into three partitions, the grand mean represented by \mathbf{m} , the effect of the treatment represented by \mathbf{t} , and the error represented by \mathbf{e} . This is a vector equation in which corresponding elements of the vectors are added. It is standard notation in statistics to use Greek letters for the model parameters \mathbf{m} , \mathbf{t} , and \mathbf{e} , and Roman letters their sample estimates based on the data collected. We will use \mathbf{M} for the sample estimate for the mean vector \mathbf{m} , \mathbf{T} for the sample estimate for the treatment vector \mathbf{t} , and \mathbf{E} for the sample estimate of the the error vector \mathbf{e} .

The grand mean is the average of all of the data. For these eight numbers, this average is 51. So \mathbf{M} , our sample estimate of \mathbf{m} , is given by

$$\mathbf{M} = \begin{bmatrix} 51 & 51 \\ 51 & 51 \\ 51 & 51 \\ 51 & 51 \end{bmatrix}.$$

The effect of being Brand A or Brand B can be determined by comparing their means to the grand mean of 51. In general, there are 51 unpopped kernels in the bag. However, if you are in the first column of our “matrix”, that is, if you are Brand A, your mean is 55. This means that you are expected to have an additional 4 unpopped kernels in the bag. If you are in the second column (Brand B), then you have 4 fewer unpopped kernels in the bag. The effect of being Brand A is to add 4 kernels from what is typical and the effect of being Brand B is to subtract 4 kernels.

So we have vector $T = \begin{bmatrix} 4 & -4 \\ 4 & -4 \\ 4 & -4 \\ 4 & -4 \end{bmatrix}$.

Now, what about the error vector? The error vector contains whatever values are necessary to make $X = m + t + e$ a valid equation. Remember, we add vectors by adding corresponding elements.

$$\begin{matrix} X & M & T & E \\ \begin{bmatrix} 52 & 44 \\ 60 & 50 \\ 56 & 52 \\ 52 & 42 \end{bmatrix} & = & \begin{bmatrix} 51 & 51 \\ 51 & 51 \\ 51 & 51 \\ 51 & 51 \end{bmatrix} & + & \begin{bmatrix} 4 & -4 \\ 4 & -4 \\ 4 & -4 \\ 4 & -4 \end{bmatrix} & + & \begin{bmatrix} - & - \\ - & - \\ - & - \\ - & - \end{bmatrix} \end{matrix}$$

What values do we put in the last vector to create a valid statement? In the first element, we have $52 = 51 + 4 + E_1$, so $E_1 = -3$. Continuing in like manner, we find the elements of vector E . Notice that the entries of each column in the matrix representation of E sum to zero.

$$\begin{matrix} X & M & T & E \\ \begin{bmatrix} 52 & 44 \\ 60 & 50 \\ 56 & 52 \\ 52 & 42 \end{bmatrix} & = & \begin{bmatrix} 51 & 51 \\ 51 & 51 \\ 51 & 51 \\ 51 & 51 \end{bmatrix} & + & \begin{bmatrix} 4 & -4 \\ 4 & -4 \\ 4 & -4 \\ 4 & -4 \end{bmatrix} & + & \begin{bmatrix} -3 & -3 \\ 5 & 3 \\ 1 & 5 \\ -3 & -5 \end{bmatrix} \end{matrix}$$

Now, remember that M, T , and E are vectors. What do you get if you compute the dot products $M \cdot T$, $M \cdot E$, and $T \cdot E$? It should be clear that $M \cdot T = 0$ and $M \cdot E = 0$, since M is a constant vector and the entries of T and E sum to zero. Also, the columns of T are constant

and the columns of E sum to zero, so $T \cdot E = 0$ as well. If the dot product of two vectors is zero, then the vectors are perpendicular. We have three mutually perpendicular vectors, which means the Pythagorean Theorem must hold. This is where we get the sums of squares in ANOVA. The sums of the squares in ANOVA are the squared lengths of the vectors X, M, T , and E . The sums of squares of the elements of M, T , and E must equal the sum of the squares of the elements of X . In this example, we can verify that this is true.

$$\sum X_i^2 = 52^2 + 60^2 + 52^2 + 42^2 = 21,048$$

while

$$\sum M_i^2 = 51^2 + 51^2 + 51^2 = 20,808$$

$$\sum T_i^2 = 4^2 + 4^2 + (-4)^2 + (-4)^2 = 128$$

(this is known as the sums of square for treatment SST)

and

$$\sum E_i^2 = (-3)^2 + 5^2 + 5^2 + (-5)^2 = 112$$

(this is known as the sums of squares for error SSE).

The Pythagorean Theorem holds since $21,048 = 20,808 + 128 + 112$.

We also have an issue with degrees of freedom. One way to think about this in the context of the “matrix” structure is to consider how many of the values in each “matrix” you must be given before you can determine all the others.

- For the initial “matrix” X , the data will be whatever it is going to be. This “matrix” has 8 degrees of freedom. In the M matrix, all of the entries are the same. If you know any one of them then you know them all. This uses 1 degree of freedom.
- In the T “matrix”, all entries in each column are the same, so once you know that the first entry is 4, you know all the rest in the first column are 4. Moreover, the sum of the entries must be zero, since the mean deviation from the mean is always zero, and that is what we are measuring here. So if the first column is all 4's, the second column must be all -4 's. Thus, “matrix” T also has only 1 degree of freedom.
- This leaves 6 degrees of freedom for E . Since each column of E must separately sum to zero, knowing any three in each column is sufficient; it has six degrees of freedom.

So, this is our additive ANOVA structure. Not only do the entries add up ($X = M + T + E$), but also the sums of squares ($\sum X^2 = \sum M^2 + \sum T^2 + \sum E^2$) and the degrees of freedom.

We are interested in comparing the sums of squares and the degrees of freedom for “matrix” T and “matrix” E .

	X	M	T	E
SS	21048	= 20808	+ 128	+ 112
df	8	= 1	+ 1	+ 6

The ratio of the sums of squares to the degrees of freedom is called the mean square. So the mean square for treatment is $MST = \frac{128}{1} = 128$ while the mean square for error is $MSE = \frac{112}{6} = 18.667$.

The ratio of these two mean squares is the value of the F statistic, with 1 and 6 degrees of freedom. In this example, we have

$$F_{1,6} = \frac{128}{18.667} = 6.857.$$

The p -value associated with this F -score is $p = 0.0397$. This is the same value given by our pooled two-sample approach. Moreover, notice that the F -score is the square of the t -score, $2.6186^2 = 6.857$. Even more importantly, notice that the MSE is exactly the same as the pooled variance in the t -test, $s_p^2 = MSE = 18.667$.

As we see, the two-sided, two-sample t -test using pooled variance and this ANOVA F -test are variations on the same theme. However, the “matrix” structure used in this example can be extended to compare more than two means. Also note that F -tests are always two-sided, since they involve squares.

Analysis of Variance is often described as a comparison of signal to noise. The MST is the measure of the strength of the signal while the MSE is the measure of the noise, or the natural variability of the process under study. (In books by David Moore, et. al., the notation MSG (mean square for groups) is used to denote the mean square for treatment.) In statistics texts, the MST is generally described as a measure of variability among or between treatment groups while MSE is described as the measure of variation within treatment groups. Consequently, Analysis of Variance compares the between-sample variability to the within-sample variability.

To my students, this between-sample variability to the within-sample variability concept is often very confusing. Rather than use that terminology, we concentrate on correctly partitioning the results into measures of the treatment effect (signal), using “matrix” T , and error (noise), using “matrix” E . The F statistic is then a measure of how much “stronger” the signal is than the noise. If this ratio is large enough, we say the effects of treatment are statistically significant.

Testing More Than Two Means

Suppose we now add a third brand of popcorn, Brand C. If we again use a completely randomized design and select the order in which the popcorn is popped at random, using the same popper each time with a sufficient cooling period between treatments to maintain independence, we can analyze the results with a one-way ANOVA procedure.

Suppose the results are as shown below, with Brand C added as the 3rd column. The mean for Brand C is 57 unpopped kernels.

A	B	C
52	44	60
60	50	58
56	52	60
52	42	50

Here the null hypothesis is $H_0: \mathbf{m}_A = \mathbf{m}_B = \mathbf{m}_C$, all means are equal. The alternative hypothesis is H_a : at least one mean different. We still have the same additive model, $X = \mathbf{m} + \mathbf{t} + \mathbf{e}$, so we set out to compute the entries in the “matrices”. Adding Brand C changes the overall, or grand mean. With C added, the grand mean is 53. The means of the columns are 55, 47, and 57. The effect of being Brand A is to add 2 unpopped kernels per bag, while Brand B subtracts 6 kernels, and Brand C adds 4 kernels. We can use these values to find the elements of \mathbf{E} as before. So our “matrices” are:

$$\begin{array}{c}
 \mathbf{X} \\
 \left[\begin{array}{ccc} 52 & 44 & 60 \\ 60 & 50 & 58 \\ 56 & 52 & 60 \\ 52 & 42 & 50 \end{array} \right]
 \end{array}
 =
 \begin{array}{c}
 \mathbf{M} \\
 \left[\begin{array}{ccc} 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \end{array} \right]
 \end{array}
 +
 \begin{array}{c}
 \mathbf{T} \\
 \left[\begin{array}{ccc} 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \end{array} \right]
 \end{array}
 +
 \begin{array}{c}
 \mathbf{E} \\
 \left[\begin{array}{ccc} -3 & -3 & 3 \\ 5 & 3 & 1 \\ 1 & 5 & 3 \\ -3 & -5 & -7 \end{array} \right]
 \end{array}$$

Notice that the columns of \mathbf{T} are constant and the rows sum to zero, while for \mathbf{E} , the columns sum to zero. The vectors \mathbf{M}, \mathbf{T} , and \mathbf{E} are again perpendicular, so the Pythagorean Theorem can be invoked. We have

$$\begin{aligned}
 (\sum X^2 = \sum M^2 + \sum T^2 + \sum E^2) \\
 \text{with} \\
 34,112 = 33,708 + 224 + 180.
 \end{aligned}$$

Notice that the sums of squares for \mathbf{E} could be found by subtraction,

$$\sum E^2 = \sum X^2 - (\sum M^2 + \sum T^2).$$

We really didn't need to find all the individual elements of \mathbf{e} to compute its sums of squares.

The degrees of freedom are 12 for \mathbf{X} , 1 for \mathbf{M} (since the elements are all the same), 2 for \mathbf{T} (since the columns are all the same and the rows must add to zero), and 9 for \mathbf{E} (because that's

all that's left or because the columns must add separately to zero). The $MSE = \frac{SSE}{df}$ is the pooled variance we get by the standard formula. In this example, $MSE = \frac{180}{9} = 20$.

Now, the $MST = \frac{224}{2} = 112$ and $MSE = \frac{180}{9} = 20$, so $F_{2,9} = \frac{112}{20} = 5.6$. The p -value associated with this F -score is $p = 0.0263$. With this p -value, we reject the null hypothesis of equal population means. The evidence suggests that at least one mean is different from another. To determine which are different we need some additional statistical reasoning. We will delay this development until after we have done a few more examples. At this point, we can certainly say that the Brands associated with the most extreme means are significantly different, that is, Brand C is different from Brand B. We do not know if either is statistically different from Brand A.

Notice that we did not use the sums of squares of \mathbf{X} or of \mathbf{M} in computing F . We are only interested in the sums of squares of \mathbf{T} and \mathbf{E} , but we need the others to find them. If we use a statistical package (JMP-IN) to do this same problem, we generate the following output:

Kernels By Brand
 Oneway Anova

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	224.00000	112.000	5.6000
Error	9	180.00000	20.000	Prob>F
C Total	11	404.00000	36.727	0.0263

Means for Oneway Anova

Level	Number	Mean	Std Error
A	4	55.0000	2.2361
B	4	47.0000	2.2361
C	4	57.0000	2.2361

Std Error uses a pooled estimate of error variance

In the table, notice that in JMP-IN the treatment sums of squares is called the Model Sums of Squares. It is the same 224 we computed from our \mathbf{T} “matrix”. The mean squares and F Ratio are the same as those we computed. The Std Error is the standard error of the sample means and is computed as $\frac{s}{\sqrt{n}} = \frac{\sqrt{MSE}}{\sqrt{n}} = \frac{\sqrt{20}}{\sqrt{4}} = 2.2361$. Notice also that no attention was paid to the sums of squares of \mathbf{X} and \mathbf{M} . The Total Sums of Squares in the print-out is just the 404 that represent the difference $\sum \mathbf{X}^2 - \sum \mathbf{M}^2$. How these 404 sums of squares are partitioned between the signal as measured by MST and noise as measured by MSE is the essence of ANOVA.

Blocking to Reduce Variability

Suppose we had done the experiment differently. Suppose we didn't have enough time to use only one popper and let it cool between treatments. So instead, we used 4 different poppers, and made sure that we popped one bag of each brand of popcorn in each of the poppers. The order in which this was done was randomized. This experimental design would be a randomized block design. Suppose the results were the same as before.

$$\mathbf{X} = \begin{array}{c} \text{I} \\ \text{II} \\ \text{III} \\ \text{IV} \end{array} \begin{array}{ccc} \text{A} & \text{B} & \text{C} \\ \left[\begin{array}{ccc} 52 & 44 & 60 \\ 60 & 50 & 58 \\ 56 & 52 & 60 \\ 52 & 42 & 50 \end{array} \right] \end{array}$$

The null and alternative hypotheses remain the same regardless of blocking. If we do not acknowledge the difference in the poppers, if we don't block, we generate the same sums of squares as before. What happens when we block? Assume that all entries in the first row of \mathbf{X} used Popper I, all entries in the second row of \mathbf{X} used Popper II, all entries in the third row of \mathbf{X} used Popper III, and all entries in the fourth row of \mathbf{X} used Popper IV.

By blocking we modify our model. The additive model we use is $X = \mu + \tau + \mathbf{b} + \mathbf{e}$, with “matrix” \mathbf{b} representing the effect of the blocking variable. This is the effect of being in a particular row of \mathbf{X} . Recall that the overall average for this data is 53 unpopped kernels. The average for Popper I is 52, so this popper leaves one fewer unpopped kernel per bag. The effect of being in Row 1 is -1 . The average for Popper II is 56, so in general, this popper left an extra three kernels unpopped. The mean for Popper III is also 56, and for Popper IV is 48, representing an average decrease of 5 kernels. The structure of “matrix” \mathbf{B} , our sample estimate of \mathbf{b} , is

$$\mathbf{B} = \begin{bmatrix} -1 & -1 & -1 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \\ -5 & -5 & -5 \end{bmatrix}$$

In this “matrix”, the entries in each row is the same and the columns add to zero. This will be important in determining the number of degrees of freedom we have for \mathbf{B} . So, our completed model is

$$\begin{array}{c}
 \mathbf{X} \qquad \qquad \mathbf{M} \qquad \qquad \mathbf{T} \qquad \qquad \mathbf{B} \qquad \qquad \mathbf{E} \\
 \begin{bmatrix} 52 & 44 & 60 \\ 60 & 50 & 58 \\ 56 & 52 & 60 \\ 52 & 42 & 50 \end{bmatrix} = \begin{bmatrix} 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \end{bmatrix} + \begin{bmatrix} 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -1 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \\ -5 & -5 & -5 \end{bmatrix} + [\mathbf{E}]
 \end{array}$$

We did not take the time to compute the entries of \mathbf{E} because all we really want are the sum of squares. Instead, we found the sums of squares of the other “matrices” and computed $\sum \mathbf{E}^2$ by subtraction. If we did compute these entries, we would find (as expected) that $\mathbf{M}, \mathbf{T}, \mathbf{B}$, and \mathbf{E} are all mutually perpendicular vectors and the Pythagorean Theorem will again be employed. So, we have

$$\begin{array}{l}
 \sum \mathbf{X}^2 = \sum \mathbf{M}^2 + \sum \mathbf{T}^2 + \sum \mathbf{B}^2 + \sum \mathbf{E}^2 \\
 \text{with} \\
 \text{SS} \quad 34,112 = 33,708 + 224 + 132 + 48. \\
 \text{df} \quad 12 = 1 + 2 + 3 + 6
 \end{array}$$

Notice that blocking did not affect the sums of square of \mathbf{X}, \mathbf{M} , or \mathbf{T} . The additional sums of squares for \mathbf{B} must come out of \mathbf{E} and the mean square of \mathbf{E} is a measure of variability. This is how blocking reduces variation. The degrees of freedom also change. The “matrix” \mathbf{B} has three degrees of freedom since each column must add to zero and the row entries are constant.

Now, we can compute $MST = \frac{224}{2} = 112$ and $MSE = \frac{48}{6} = 8$. The signal (112) is just as strong as before but the noise has been reduced from $s^2 = 20$ to $s^2 = 8$. Our F -score is $F_{2,6} = \frac{112}{8} = 14$ with a p -value of $p = 0.0055$. Again, we reject the null hypothesis of equal means in the belief that at least one of the population means differs from another.

Testing the Blocking Variable

There is no difference in the structure of the treatment “matrix” and the blocking “matrix”. We could just as easily test to see if there is any significant difference in the poppers, or whether the differences we see are examples of chance variation. In this case, $H_0: \mathbf{m}_I = \mathbf{m}_{II} = \mathbf{m}_{III} = \mathbf{m}_{IV}$ and the alternative is H_a : at least one mean different.

$$\text{Our } F\text{-score is } F_{3,6} = \frac{MSB}{MSE} = \frac{\left(\frac{132}{3}\right)}{\left(\frac{48}{6}\right)} = \frac{44}{8} = 5.5. \text{ The associated } p\text{-value is } p = 0.037101.$$

There is sufficient evidence to reject the null hypothesis of equal means for the poppers. The differences we see are too disparate to be considered examples of random variation. We think there is a real difference in the poppers.

Again using JMP-IN, we get the following printout.

Response:	Kernels				
Effect Test					
Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Brand	2	2	224.00000	14.0000	0.0055
Popper	3	3	132.00000	5.5000	0.0371
Whole-Model Test					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Ratio	Prob>F
Model	5	356.00000	71.2000	8.9000	
Error	6	48.00000	8.0000		0.0096
C Total	11	404.00000			
Brand					
Effect Test					
Sum of Squares	F Ratio	DF	Prob>F		
224.00000	14.0000	2	0.0055		
Least Squares Means					
Level	Least Sq Mean	Std Error	Mean		
A	55.00000000	1.414213562	55.0000		
B	47.00000000	1.414213562	47.0000		
C	57.00000000	1.414213562	57.0000		
Popper					
Effect Test					
Sum of Squares	F Ratio	DF	Prob>F		
132.00000	5.5000	3	0.0371		
Least Squares Means					
Level	Least Sq Mean	Std Error	Mean		
I	52.00000000	1.632993162	52.0000		
II	56.00000000	1.632993162	56.0000		
III	56.00000000	1.632993162	56.0000		
IV	48.00000000	1.632993162	48.0000		

Notice that the 404 sums of squares are now partitioned into 3 components, those associated with the treatment T , those associated with the block B , and those associated with the error E . Also observe the sums of squares, the degrees of freedom, the F -ratios, and the p -values are the same as we computed using our “matrices”. The Whole-Model Test uses the

model $X = \mathbf{m} + (\mathbf{t} + \mathbf{b}) + \mathbf{e}$. This model considers $(\mathbf{t} + \mathbf{b})$ as a single factor and combines the sums of squares and degrees of freedom of \mathbf{T} and \mathbf{B} in a single “matrix”. The standard error for the Brands in the print-out is $\frac{s}{\sqrt{n}} = \frac{\sqrt{8}}{\sqrt{4}} = 1.414$ and for Popper is $\frac{s}{\sqrt{n}} = \frac{\sqrt{8}}{\sqrt{3}} = 1.633$.

So, even though most of the ANOVA print-outs have a lot of information we may not need, we should see that the sums of squares, mean squares, and F -scores come from our “matrix” structure. We are simply decomposing the observed values into orthogonal partitions, and using the Pythagorean Theorem to measure the length of the vectors representing the signal and the noise. These vectors are “normalized” in a sense by the degrees of freedom and we measure the ratio of the signal to the noise. If the signal is sufficiently larger than the noise, that is, the differences in the means are sufficient to reject they are a result of chance variation, we reject the null hypothesis and say there is a significant difference.

Latin Square Design

Statistics for Experimenters gives a very nice example of a Latin square design which uses two blocking variables. Suppose that four cars and four drivers are employed in a study of gasoline additives. Four different additives, A, B, C, and D are to be used in each of the cars and with each of the drivers. The cars are labeled I, II, III, and IV and the drivers 1, 2, 3, 4. The cars were driven on a test track and the level of a pollutant in the exhaust measured. Higher scores mean a greater amount of the pollutant.

The results of the study are given in the table below.

	Driver 1	Driver 2	Driver 3	Driver 4
Car I	A 21	B 26	D 20	C 25
Car II	D 23	C 26	A 20	B 27
Car III	B 15	D 13	C 16	A 16
Car IV	C 17	A 15	B 20	D 20

The mean for each additive, car, and driver are calculated and shown below:

Additives	Drivers	Cars
$\bar{x}_A = 18$	$\bar{x}_1 = 19$	$\bar{x}_I = 23$
$\bar{x}_B = 22$	$\bar{x}_2 = 20$	$\bar{x}_{II} = 24$
$\bar{x}_C = 21$	$\bar{x}_3 = 19$	$\bar{x}_{III} = 15$
$\bar{x}_D = 19$	$\bar{x}_4 = 22$	$\bar{x}_{IV} = 18$

Our additive model for ANOVA is $X = \mathbf{m} + \mathbf{t} + \mathbf{b}_D + \mathbf{b}_C + \mathbf{e}$, and the “matrix” structure is given below:

$$\begin{bmatrix} 21 & 26 & 20 & 25 \\ 23 & 26 & 20 & 27 \\ 15 & 13 & 16 & 16 \\ 17 & 15 & 20 & 20 \end{bmatrix} = \begin{bmatrix} 20 & 20 & 20 & 20 \\ 20 & 20 & 20 & 20 \\ 20 & 20 & 20 & 20 \\ 20 & 20 & 20 & 20 \end{bmatrix} + \begin{bmatrix} -2 & 2 & -1 & 1 \\ -1 & 1 & -2 & 2 \\ 2 & -1 & 1 & -2 \\ 1 & -2 & 2 & -1 \end{bmatrix} + \begin{bmatrix} -1 & 0 & -1 & 2 \\ -1 & 0 & -1 & 2 \\ -1 & 0 & -1 & 2 \\ -1 & 0 & -1 & 2 \end{bmatrix} + \begin{bmatrix} 3 & 3 & 3 & 3 \\ 4 & 4 & 4 & 4 \\ -5 & -5 & -5 & -5 \\ -2 & -2 & -2 & -2 \end{bmatrix} + [\mathbf{e}]$$

The sums of squares and degrees of freedom are now computed, with \mathbf{e} found by subtraction.

	X	=	M	+	T	+	B_D	+	B_C	+	E
SS	6696		6400		40		24		216		16
df	16		1		3		3		3		6

Additives: We can now test additives. $H_0: \mathbf{m}_A = \mathbf{m}_B = \mathbf{m}_C = \mathbf{m}_D$
 H_a : at least one mean different.

We find that $F_{3,6} = \frac{MST}{MSE} = \frac{\left(\frac{40}{3}\right)}{\left(\frac{16}{6}\right)} = \frac{13.333}{2.6667} = 5$. The p -value is $p = 0.0452$. We reject the null

hypothesis at the 0.05 level of significance. At least one population mean differs from another.

Drivers: We can also test drivers. $H_0: \mathbf{m}_1 = \mathbf{m}_2 = \mathbf{m}_3 = \mathbf{m}_4$
 H_a : at least one mean different.

We find that $F_{3,6} = \frac{MSB_D}{MSE} = \frac{\left(\frac{24}{3}\right)}{\left(\frac{16}{6}\right)} = \frac{8}{2.6667} = 3$. The p -value is $p = 0.117$. We fail to reject the

null hypothesis. Our observations are consistent with random variation. There is no evidence that the drivers differ on the basis of emission of pollutants.

Cars: We can also test cars. $H_0: \mathbf{m}_I = \mathbf{m}_{II} = \mathbf{m}_{III} = \mathbf{m}_{IV}$
 H_a : at least one mean different.

We find that $F_{3,6} = \frac{MSB_C}{MSE} = \frac{\left(\frac{216}{3}\right)}{\left(\frac{16}{6}\right)} = \frac{72}{2.6667} = 27$. The p -value is $p = 0.0007$. We reject the

null hypothesis of equal means for cars. At least one population mean differs from another.

Compare the sums of squares computed in our “matrix” partition and those in the JMP-IN print-out below.

Response: Pollution Level

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Additive	3	3	40.00000	5.0000	0.0452
Car	3	3	216.00000	27.0000	0.0007
Driver	3	3	24.00000	3.0000	0.1170

Whole-Model Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	9	280.00000	31.1111	11.6667
Error	6	16.00000	2.6667	Prob>F
C Total	15	296.00000		0.0037

Additive

Effect Test

Sum of Squares	F Ratio	DF	Prob>F
40.000000	5.0000	3	0.0452

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
A	18.00000000	0.8164965809	18.0000
B	22.00000000	0.8164965809	22.0000
C	21.00000000	0.8164965809	21.0000
D	19.00000000	0.8164965809	19.0000

Car

Effect Test

Sum of Squares	F Ratio	DF	Prob>F
216.00000	27.0000	3	0.0007

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
I	23.00000000	0.8164965809	23.0000
II	24.00000000	0.8164965809	24.0000
III	15.00000000	0.8164965809	15.0000
IV	18.00000000	0.8164965809	18.0000

Driver

Effect Test

Sum of Squares	F Ratio	DF	Prob>F
24.000000	3.0000	3	0.1170

Least Squares Means

Level	Least Sq Mean	Std Error	Mean
1	19.00000000	0.8164965809	19.0000
2	20.00000000	0.8164965809	20.0000
3	19.00000000	0.8164965809	19.0000
4	22.00000000	0.8164965809	22.0000

Determining Which Means are Different

We have detected some differences in each of the examples we have considered. How do you decide which means are significantly different and which are not? If our ANOVA has detected a difference in means, (we have rejected the null hypothesis of equal means in favor of the alternative) we conclude that at least one mean differs from another. Clearly, the most extreme means must be different, but what about the others?

There are a number of ways to decide which means are significantly different and each statistician seems to have their favorite. Some techniques are more conservative than others and some more prone to Type I errors. There is no "universally best" procedure.

One simple approach is to use the *Least Significant Difference (LSD)* criterion. Compared to other methods, the *LSD* procedure is more likely to call a difference significant and therefore prone to Type I errors, but is easy to use and is based on principles that students already understand.

The LSD Procedure

We know that if two random samples of size n are selected from a normal distribution with variance \mathbf{s}^2 , then the variance of the difference in the two means is

$$\mathbf{s}_D^2 = \sqrt{\frac{\mathbf{s}^2}{n} + \frac{\mathbf{s}^2}{n}} = \sqrt{\frac{2\mathbf{s}^2}{n}}.$$

In the case of ANOVA, we do not know \mathbf{s}^2 , but we estimate it with $s^2 = MSE$. So when two random samples of size n are taken from a population whose variance is estimated by MSE , the standard error of the difference between the two means is $\sqrt{\frac{2 \cdot s^2}{n}} = \sqrt{\frac{2 \cdot MSE}{n}}$. Two means will be considered significantly different at the 0.05 significance level if they differ by more than $t^* \sqrt{\frac{2 \cdot MSE}{n}}$, where t^* is the t -value for a 95% confidence interval with the degrees of freedom associated with MSE . The value

$$LSD = t^* \sqrt{\frac{2 \cdot MSE}{n}}$$

is called the *Least Significant Difference*. The *LSD* is used only when the F -test indicates a significant difference exists. If the two samples do not contain the same number of entries, then

$$LSD = t^* \sqrt{MSE} \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}.$$

The number of degrees of freedom for t^* is always that of MSE .

Popcorn Brands without Blocking

In our first example, without blocking, we had means of 55, 47, and 57 for Brands A, B, and C, respectively. We had a significant difference according to our F -test. The t -score for 95% confidence with 9 degrees of freedom is 2.262. We know that $MSE = 20$ and $n = 4$. Computing LSD , we find

$$LSD = 2.262 \sqrt{\frac{2 \cdot 20}{4}} = 7.15.$$

Any means that differ by more than 7.15 units are considered distinct. So, we can say that Brands A and C are different from Brand B with respect to the number of unpopped kernels, but that Brand A and C are indistinguishable. In short, Brand B is better than both A and C if you want more popcorn to eat and fewer unpopped kernels.

Popcorn Brands with Blocking

When we consider the example using blocking, we have the same means, but new degrees of freedom and MSE , so we have a new LSD . In this case

$$LSD = 2.447 \sqrt{\frac{2 \cdot 8}{4}} = 4.89$$

By removing the variation due to the different poppers, we now can conclude that any means more than 4.9 units apart should be considered different. The results of this analysis are the same as before. The means of Brand A and Brand C are both larger than the mean of Brand B, but Brands A and C remain indistinguishable from each other.

We can also consider the Poppers. We found a significant difference in the number of kernels remaining unpopped. The averages for the poppers were 52, 56, 56, and 48, for poppers I, II, III and IV, respectively. These averages represent the means of 3 observations, so $n = 3$ in our computation.

$$LSD = 2.447 \sqrt{\frac{2 \cdot 8}{3}} = 5.65$$

Any means for poppers more than 5.65 units apart are considered different. The maximum difference between two means for the Poppers is 8, so we believe that Popper IV differs from Poppers II and III with respect to the number of unpopped kernels. In this sense, Popper IV is the better popper. However, we cannot claim that Popper IV differs from Popper I since the difference in means is 4, which is smaller than the computed LSD . For the same reason, we cannot distinguish Popper I from Poppers II and III.

Gasoline Additives

In the car pollution example, we had 6 degrees of freedom in the error “matrix”, so our t -score for LSD is 2.447. Each mean was an average of 4 observations, so $n = 4$. Finally, our $MSE = 2.6667$. The LSD is the same for comparing differences associated with additives and cars.

$$LSD = 2.447 \sqrt{\frac{(2)2.6667}{4}} = 2.83.$$

Any differences larger than 2.83 are considered statistically significant.

For Additives, the means were 18 for Brand A, 19 for Brand D, 21 for Brand C and 22 for Brand B. Additive Brands A and D are indistinguishable, as are Brands C and B. However, the mean levels of pollutant for Brands C and B are larger than for Brand A while the mean level of pollution for Brands A and D are smaller than for Brand B.

There was no significant difference in drivers, so we should not consider LSD in this situation.

For Cars, the means were 15 for III, 18 for IV, 23 for I, and 24 for II. Car III had a lower mean level of pollutant emission than all the rest. Car IV had a higher mean level of pollutant emission than Car III and lower mean level of pollutant emission than both I and II. Cars I and II are indistinguishable on the basis of their mean level of pollutant emission.

References:

- Box, George P., William G. Hunter, and J. Stuart Hunter, *Statistics for Experimenters*, John Wiley & Sons, New York, New York, 1978. ISBN 0-471-09315-7
- Cobb, George W., *Introduction to the Design and Analysis of Experiments*, Springer-Verlag, New York, New York, 1998. ISBN 0-387-94607-1
- Iman, Ronald, L., *A Data-Based Approach to Statistics*, Duxbury Press, Belmont, California, 1994. ISBN 0-534-93317-3
- Sall, John, and Ann Lehman, *JMP Start Statistics*, Duxbury Press, Belmont, California, 1996. ISBN 0-534-26565-0
- Snedecor, George W., and William G. Cochran, *Statistical Methods, 6th*, The Iowa State University Press, Ames, Iowa, 1967.