

## A Closer Look at Blocking

Dan Teague

NC School of Science and Mathematics

In George Cobb's excellent text, *Introduction to the Design and Analysis of Experiments*, he describes the variability inherent in an experiment in the following way: Any experiment is likely to involve three kinds of variability:

1. Planned, systematic variability. This is the kind we want since it includes the differences due to the treatments.
2. Chance-like variability. This is the kind our probability models allow us to live with. We can estimate the size of this variability if we plan our experiment correctly.
3. Unplanned, systematic variability. This kind *Threatens Disaster!* We deal with this variability in two ways, by randomization and by blocking.
  - Randomization turns unplanned, systematic variation into planned, chance-like variation.
  - Blocking turns unplanned, systematic variation into planned, systematic variation.

In this paper, we will take a closer look at how blocking turns unplanned, systematic variation into planned, systematic variation.

### ANOVA: Testing More Than Two Means

**Example Problem:** Three types of popcorn are being compared. Five hundred kernels of each type are placed in an automatic popcorn popper. The popper is turned on for five minutes and the number of unpopped kernels remaining is counted. Sufficient time is left between treatments to allow the popper to cool down. Four bags of each type are popped one at a time, with the order determined at random. This is a completely randomized design. We can analyze the results with a one-way ANOVA procedure.

Suppose the results are as shown below. The mean for Brand A is 55 unpopped kernels, for Brand B is 47 unpopped kernels, and 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  $H_0: \mu_A = \mu_B = \mu_C$ , all means are equal. The alternative hypothesis is  $H_a$ : at least one mean is different. The basic model for ANOVA is the additive model,  $Y = \mu + \tau + \varepsilon$ .

Our observations,  $Y$ , is decomposed into three partitions, the grand mean represented by  $\mu$ , the effect of the treatment represented by  $\tau$  (this is the signal we want to measure), and the random error represented by  $\varepsilon$  (this is the noise that makes it difficult to find the signal). The equation  $Y = \mu + \tau + \varepsilon$  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  $\mu$ ,  $\tau$ , and  $\varepsilon$ , and Roman letters their sample

estimates based on the data collected. We will use  $\mathbf{M}$  for the sample estimate for the mean vector  $\mu$ ,  $\mathbf{T}$  for the sample estimate for the treatment vector  $\tau$ , and  $\mathbf{E}$  for the sample estimate of the error vector  $\varepsilon$ .

To make it easier to read, we will write the vectors in a matrix format. Our “matrices” are:

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

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

$$\left( \sum Y^2 = \sum M^2 + \sum T^2 + \sum E^2 \right)$$

with

$$34,112 = 33,708 + 224 + 180.$$

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

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

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

The degrees of freedom are 12 for  $\mathbf{Y}$ , 1 for  $\mathbf{M}$  (since the elements are all the same), 2 for  $\mathbf{T}$  (since the column entries 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  $MST = \frac{224}{2} = 112$ , 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 (with  $\bar{x}_C = 57$ ) is different from Brand B (with  $\bar{x}_B = 47$ ). We do not know if either is statistically different from Brand A (with  $\bar{x}_A = 55$ ).

Notice that we did not use the sums of squares of  $\mathbf{Y}$  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:

#### Analysis of Variance

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

Notice that in the JMP-IN table the treatment sums of squares is called the Model Sums of Squares. It is the same 224 we computed from our  $T$  "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  $Y$  and  $M$ . The Total Sums of Squares in the print-out is just the 404 that represent the difference  $\sum Y^2 - \sum M^2$ . 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.

### How Two Estimates of Variance Can Compare Means

Analysis of variance gets its name because it uses two different estimates of the variance to compare means. If the null hypothesis is true, and there is no treatment effect, then the two estimates of variance should be comparable, that is, their ratio should be one. The farther is the ratio of variances from one, the more doubt is placed on the null hypothesis. Here is the basic idea behind this comparison.

If the null hypothesis is true and all samples can be considered to come from one population, we can estimate the variance in a couple of ways. Both assume that the observations are distributed about a common mean  $\mu$  with variance  $\sigma^2$ .

Recall that our data is

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

One method of estimating the variance  $\sigma^2$  is to pool the estimates from each of the samples of 4 that our null hypothesis assumes have been taken from a single population. In this example, we have the  $s_A^2 = 14.667$ ,  $s_B^2 = 22.667$ ,  $s_C^2 = 22.667$ . The pooled estimate of  $\sigma^2$  is

$$s_p^2 = \frac{3(14.667) + 3(22.667) + 3(22.667)}{9} = 20.00.$$

Notice that this is the Mean Square Error in our earlier computation.

A second way to estimate the variance  $\sigma^2$  is to infer the value of  $\sigma^2$  from  $s_Y^2$ , where  $s_Y^2$  is the observed variance of the sample means. We calculate this by considering the means of the three treatment groups,  $A$ ,  $B$ , and  $C$ . The three means,  $\bar{y}_A = 55$ ,  $\bar{y}_B = 47$ ,  $\bar{y}_C = 57$ , are expected to have a variance of  $\frac{\sigma^2}{4}$  since they are the means of samples of size 4 drawn at random from a population with variance. The variance of the data set 55, 47, and 57 is  $s_Y^2 = 28$ . So 28 is a estimate of the value of  $\frac{\sigma^2}{4}$ . This gives us a second estimate of  $\sigma^2$  that is equal to 112. Notice that this is the Mean Square Treatment.

If the null hypothesis is true, then both 20 and 112 are estimates of the same population variance  $\sigma^2$ . The more these two estimates differ (the larger their ratio), the more evidence this gives against the null hypothesis.

### 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.

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

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

By blocking we modify our model. The additive model we use is  $Y = \mu + \tau + \beta + \varepsilon$ , with “matrix”  $\beta$  representing the effect of the blocking variable. This is,  $\beta$  represents the effect of being in a particular row of  $Y$ . 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. So, our completed model is

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

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

$$\sum Y^2 = \sum M^2 + \sum T^2 + \sum B^2 + \sum E^2$$

$$\begin{array}{rcl} & & \text{with} \\ \text{SS} & 34,112 = & 33,708 + 224 + 132 + 48. \\ \text{df} & 12 = & 1 + 2 + 3 + 6 \end{array}$$

Notice that blocking did not affect the sums of square of  $\mathbf{Y}$ ,  $\mathbf{M}$ , or  $\mathbf{T}$ . The additional sum of squares for  $\mathbf{B}$  must come out of  $\mathbf{E}$ . *This is the way in which blocking “reduces variation.” Since the mean square of  $\mathbf{E}$  is a measure of variability, we can see how blocking reduces variation. Blocking allows us to estimate the contribution to variance of the blocking variable and remove it from our analysis.* 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.

### Determining Which Means are Different

We have detected some differences in each of the examples we have considered. How do we 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. 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 in introductory courses already understand.

### Fisher’s Least Significant Difference Procedure

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

$$\sigma_D^2 = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}.$$

In the case of ANOVA, we do not know  $\sigma^2$ , 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 number of degrees of freedom for  $t^*$  is always that of  $MSE$ . Note that the *LSD* procedure is used only when the  $F$ -test indicates a significant difference exists.

### 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 treatment 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 average number of unpopped kernels, but that Brand A and C are indistinguishable. In short, Brand B is better on average 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 treatment 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.

### References:

- Box, George ., William Hunter, and J. Stuart Hunter, *Statistics for Experimenters*, John Wiley & Sons, New York, 1978.  
 Cobb, George W., *Introduction to the Design and Analysis of Experiments*, Springer-Verlag, New York, New York, 1998.  
 Snedecor, G., and W. Cochran, *Statistical Methods*, 6th, The Iowa State University Press, Ames, Iowa, 1967.