

## ***Understanding Blocking in AP Statistics***

**Dan Teague**

**North Carolina School of Science and Mathematics**

In *Introduction to the Design and Analysis of Experiments*, George Cobb (1998) 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, while blocking turns unplanned, systematic variation into planned, systematic variation.

The management of these three sources of variation is the essence of experimental design. To focus the discussion, consider the variation inherent in the following experimental setting. To keep the computations simple and clear, the sample size is unrealistically small. We hope the gain in simplicity and clarity from this example outweighs the obvious problem with sample size.

**Example Experiment:** Compare two kinds of rabbit food on weight gain (in ounces) from the age of two weeks to the age of six weeks of life. We want to know if the rabbits will gain more weight on one diet than on the other. We have space to house eight rabbits for this experiment.

### **Sources of Variation**

The most obvious is, perhaps, that the rabbits are all different rabbits, and so they will all grow at different rates. If different breeds of rabbit are used, then we will have an additional source of variation. Young California rabbits do not grow at the same rate as young Florida White rabbits, for example. The environment in which the rabbits live will not be exactly the same. They cannot all live in the same location; some will be in slightly warmer areas while others will be in areas with more light. They will not all have exactly the same amount of exercise or sleep. The food will be carefully weighed before it is given to the rabbits, but there will inevitably be measurement error in the amount of food given to each rabbit. Similarly, the rabbits will be weighed before the experiment begins and after the experiment ends. Measurement error (hopefully small) will occur in both these weighings. All of the aspects of the experimental setting mentioned so far can be considered natural chance-like variation.

There is another source of essential variation: the systematic difference in the rate of growth that is a result of the different diets. This is a variation we want to investigate. One way to think about the goal of the experiment is that we want to know if the variation that is a result of the diet is larger than the variation that is due to all the natural variation inherent in rabbit growth. In designing our experiment, we want to accentuate this planned, systematic variation, while reducing the natural chance-like variation.

This chapter considers some of the methods the experimenter has for managing these three sources of variation in the example experiment. The methods are Control, Randomization, Replication, and Blocking.

**Control:** We control the experiment by organizing the structural components of the experiment to remove as many sources of chance-like variation as possible. We want to keep the experimental units (in this example, the rabbits in their cage) as similar as possible. We would like to use only one breed of rabbit. We might also prefer to use just one gender of rabbit, since male and female growth patterns may differ. We certainly want to keep the cages in a single location so that the effects of heat, light, air flow, and other unknowable affects will be as consistent as possible. As much as possible, we want the only difference among the rabbits to be the diet they are receiving.

To control the measurement error, we want to use the same scale when measuring the food each day. Similarly, it is important to use the same scales to measure the weights of the rabbits before and after the

experimental regimen. We would prefer to use the same technician as well. In the end, no matter how much control we have in our experiment, some chance-like variation remains. It is the natural variation in average weight gain for our rabbits. By control, we try to make sure that our estimate of this variation is as accurate as possible and is not inflated by being combined with other extraneous sources of variation.

In the end, we want to do a calculation like a  $t$ -test, and the standard error in the denominator of this  $t$ -test will be our estimate of the natural chance-like variation in average weight gain for rabbits. The smaller we make our estimate of this chance-like variation, the more likely we are to detect a difference in the two varieties of rabbit food if a difference really exists. This is called increasing the power our test. We can gain power through controlling the experiment.

### **The Price of Control: Scope of Inference**

The scope of inference refers to the population to which inference can reasonably be drawn based on the study. This population is the population from which the random sample used in the study was drawn. If only one breed of rabbit and one gender (male) is used, we might consider the results a random sample of results possible with this breed of male rabbit. We can comment only on this breed of male rabbit. If someone suggests that other breeds or females would behave differently with the diets, we have no counter-argument. We only have information about the breed of male rabbits we considered.

If we had taken a random sample of rabbits from several breeds, we introduce the variation inherent in those breeds. Some breeds are smaller and more active than others, while others are larger and more sedate. This variation makes it more difficult for us to find a difference in weight gain due to diet if one exists. However, if we do find a significant difference in weight gain, we can say something about rabbits of different breeds, not just one special breed. Similarly, if we used males and females, our population of inference would be rabbits of either gender.

Through control, the experimenter attempts to accentuate or make as visible as possible the planned, systematic variation between treatment groups, while at the same time reducing or removing as much chance-like variability as possible. The smaller the population of inference, often, the greater the control we have.

### **Randomization**

A second approach to handling the chance-like variability is through randomization. Clearly, it is not possible to remove all chance-like variation through our methods of control. Rabbits are still different, even if they are the same breed and gender; some will grow faster than others, regardless of the diet. Measurement error is always present even if the same scales and technicians are used.

By randomly assigning the rabbits to the treatment groups, we will spread the chance-like variation among the treatment groups. This adds to the variation in each group, but it removes the bias that would otherwise doom the experiment. This random assignment of experimental unit to treatment group is essential for an experiment and distinguishes it from an observational study. In an observational study, the experimental units or subjects are not randomly assigned to the treatment groups. As an example, if we looked at rabbits at Farm A who are presently being fed Diet A and rabbits at Farm B who are presently being fed Diet B, we might find that the average weight gain for the rabbits at Farm A (on Diet A) had a greater average weight gain. However, the diet and farm are inextricably confounded. We could never say that Diet A was better than Diet B, since any weight gain could be a result of other aspects of rabbit life at the two farms. If, in the end, we want to say that one diet produces greater average weight gain in rabbits than the other, then randomization is essential.

Randomization plays other roles in our experiment. We should place the rabbit cages in our designated locations through a random process. This keeps the lurking variables of heat, light, air flow, and other unknowable variables from biasing the results. These unknown variables can produce systematic, unplanned variation if randomization is not used. The effects of these variables, if any, is distributed to the two treatment groups by this randomization process. This form of randomization is our protection against bias (unplanned, systematic variation) in the experiment.

Finally, a randomization process creates the probability models we use for the basis of hypothesis tests. If the null hypothesis is true (diet has no effect on average weight gain), then the variation we see between treatment groups must all be of the chance-like variety. We have estimated the size of this variation, and we build our hypothesis tests around it.

Both of these types of randomization are essential to a good experimental design. A third type of randomization, random sample of experimental units from the population of inference, is not essential and is often not possible. However, if a random sample is taken, we can make inferences to the population once we have our result.

### Replication

Replication in an experiment just means using more than one experimental unit. In this context, it does not mean repeating the whole experiment multiple times. Each additional rabbit is called a replicate. We must have a way to estimate the size of the chance variation and we need at least two values to compute a standard deviation. Without replication, there is no way for the experimenter to estimate the chance-like variation to compare to the systematic, planned variation between treatment groups. The more rabbits used for each diet, the more accurate is the estimate of the natural variation in weight gain. There is a second benefit of using more rabbits; the greater the number of replicates in each treatment group, the smaller the standard error used in the  $t$ -test, since the estimated variance of the mean weight gain of  $n$  rabbits is the estimated variance for single rabbits divided by  $n$ . This corresponds to a reduction in the estimate for the size of the chance-like variation in the mean increase in weight.

### Blocking

The final method for managing the variability inherent in an experiment is through blocking. Blocking is more complicated than control, randomization, and replication. To understand and appreciate the effect of blocking in an experiment, it is essential to have a general knowledge of the statistical technique of Analysis of Variance (ANOVA). So, before discussing the effect of blocking in experimental design, we need to consider ANOVA.

### A Completely Randomized Design

To develop the technique of ANOVA, let's think about a particular experimental setup for the rabbit diet example. We have two different diets that we want to compare. The diets are labeled Diet A and Diet B. The two diets will define two treatment groups. We are interested in how the diets affect the weight gain of rabbits. We have eight male Florida White rabbits available for the experiment, so we will randomly assign four to each diet. How should we use randomization to assign the rabbits to the two treatment groups? The eight rabbits arrive and are placed in a large compound until you are ready to begin the experiment, at which time they will be transferred to cages.

Number the rabbits 1-8. In a bowl put eight strips of paper each with one of the integers 1-8 written on it. In a second bowl put eight strips of paper, four each labeled A and B. Select a number and a letter from each bowl. The rabbit designated by the number is given the diet designated by the letter. Repeat without replacement until all rabbits have been assigned a diet. Each rabbit must have its own cage.

Suppose we run this experiment, and the weight gains for the eight rabbits are as given in the table below with each measurement in ounces.

<b>Diet A (ounces)</b>	52	60	56	52
<b>Diet B (ounces)</b>	44	50	52	42

Table1: Weight Gain in Ounces and Diet Type

### The Standard Two-sample Analysis

First, let's analyze the results using a two-sample  $t$ -test with pooled variance. The hypotheses we are interested in testing are

$$H_0: \mu_A = \mu_B$$

$$H_a: \mu_A \neq \mu_B$$

where  $\mu_A$  is the mean weight gain for the rabbits on Diet A and  $\mu_B$  is the mean weight gain for the rabbits on Diet B. We will use the decision criterion  $\alpha = 0.05$  throughout this discussion. We are assuming that the response variable, weight gain in ounces, has a distribution that is approximately normal. With only four data points in each group, this assumption of normality is difficult to assess, but we rely on the fact that  $t$ -tests are robust against violations of normality and there is no evidence to suggest non-normality in the data. We see

that there are no outliers, and we have been careful in our design to insure that the result for each rabbit is independent of the others. So, there is no reason to call into question the use of the  $t$ -test in this situation. There are only four data points in each set, but the lack of realism hopefully can be made up by the clarity of the example.

From the sample data we have

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

$$\text{So, } t_6 = \frac{(55 - 47) - (\mu_A - \mu_B)}{\sqrt{18.667} \sqrt{\frac{1}{4} + \frac{1}{4}}} = \frac{8}{3.055} = 2.6186.$$

With six 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 only chance-like variation. We believe that Diet A leads to greater weight gain in male Florida White rabbits.

### The Short Overview of the ANOVA Approach

We will repeat the analysis above using the method of ANOVA. When you look at an ANOVA table, all you see are sums of squares. The discussion that follows will focus on where those sums of squares come from, and how they are used to compare the mean weight gain.

First, create a vector of the data  $[52, 60, 56, 52, | 44, 50, 52, 42]^T$ . A vertical bar (|) has been placed in the vector to separate the values of the two sets of data, Diet A and Diet B. A more convenient method used by Box, Hunter, and Hunter (1978) is to write the vector in a matrix format as shown below. This helps keep the data separate and will help clarify the ANOVA technique. Even though the data is written using matrix notation, it can be operated on as the vector that it actually is.

$$Y = \begin{array}{cc} & \begin{array}{cc} \mathbf{A} & \mathbf{B} \end{array} \\ \begin{bmatrix} 52 & 44 \\ 60 & 50 \\ 56 & 52 \\ 52 & 42 \end{bmatrix} \end{array}$$

The mathematical model for ANOVA is the simple additive model  $Y = \mu + \tau + \varepsilon$ . The observations,  $Y$ , are decomposed into three partitions, the grand mean represented by  $\mu$ , the effect of the treatment represented by  $\tau$ , and the random error represented by  $\varepsilon$ . 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 population values  $\mu$ ,  $\tau$ , and  $\varepsilon$ , and Roman letters for their sample estimates based on the data collected.  $M$  represents the sample estimate for the mean vector  $\mu$ ,  $T$  the sample estimate for the treatment vector  $\tau$ , and  $E$  the sample estimate of the error vector  $\varepsilon$ .

The grand mean is the average of all of the sample data. For these eight numbers, this average is 51. So  $M$ , the sample estimate of  $\mu$ , is given by

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

The effect of Diet A or Diet B can be estimated by comparing the mean for each diet to the grand mean of 51. In general, these rabbits had a 51 ounce weight gain. However, rabbits in the first column of the matrix,

that is, rabbits fed Diet A, have a mean of 55. This suggests that they are expected to have an additional four ounces of weight gain over what is typical for all rabbits. For rabbits in the second column (those fed Diet B), then they have four fewer ounces of weight gain than was typical of all rabbits. The *effect* of being fed Diet A is to add four ounces of weight gain from what is typical and the effect of being Diet B is to subtract four ounces.

So we use the matrix  $T = \begin{bmatrix} 4 & -4 \\ 4 & -4 \\ 4 & -4 \\ 4 & -4 \end{bmatrix}$  to represent the treatment effect.

Now, what about the errors? The error, or residual, vector contains whatever values are necessary to make  $Y = M + T + E$  a valid equation. Remember, vectors are added by adding corresponding elements. This corresponds to adding corresponding values in our matrix notation as well.

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

What values should be in the last matrix to create a valid statement? In the first element,  $52 = 51 + 4 + E_1$ , 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 vector  $E$  sum to zero.

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

Remember that  $M$ ,  $T$ , and  $E$  are actually vectors written in matrix format to keep the treatments easily identified. 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. Because the three vectors are mutually perpendicular, the Pythagorean Theorem must hold (this is similar to using the Pythagorean Theorem to find the length of a diagonal in a rectangular room using the lengths of the walls and the height to the ceiling). The sums of squares in ANOVA come from the general Pythagorean Theorem. The sums of the squares in ANOVA are the squared lengths of the vectors  $Y$ ,  $M$ ,  $T$  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  $Y$ . In this example, we can verify that this is true.

$$\begin{aligned} \sum Y_i^2 &= 52^2 + 60^2 + \dots + 52^2 + 42^2 = 12,048 \\ &\text{while} \\ \sum M_i^2 &= 51^2 + 51^2 + \dots + 51^2 = 20,808 \\ \sum T_i^2 &= 4^2 + 4^2 + \dots + (-4)^2 + (-4)^2 = 128 \\ &\text{(this is known as the sums of square for treatment SST)} \\ &\text{and} \\ \sum E_i^2 &= (-3)^2 + 5^2 + \dots + 5^2 + (-5)^2 = 112 \end{aligned}$$

(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 to consider 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  $Y$ , the data will be whatever they are going to be. This matrix has eight degrees of freedom.
- In the  $M$  matrix, all of the entries are the same. If you know any one entry, then you know all entries. This uses one 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 one degree of freedom.
- This leaves six degrees of freedom for  $E$ . Since each column of  $E$  must separately sum to zero, knowing any three entries 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 ( $Y = M + T + E$ ), but also do the sums of squares ( $\sum Y^2 = \sum M^2 + \sum T^2 + \sum E^2$ ) and the degrees of freedom.

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

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

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

$MSE = \frac{112}{6} = 18.667$ . The ratio of these two mean squares is called the  $F$  statistic, with one and six degrees

of freedom. In this example, we have  $F_{1,6} = \frac{128}{18.667} = 6.857$ . The  $p$ -value associated with this  $F$ -value is

$p = 0.0397$ . Notice that the  $p$ -value is the same value given by our pooled two-sample approach. Moreover, notice that the value of  $F$  is the square of the value of  $t$ ,  $2.6186^2 = 6.857$ . Even more importantly, notice that the  $MSE$  is exactly the same as the pooled variance in the  $t$ -test,  $s_p^2 = MSE = 18.667$ . These are not chance occurrences. These relationships between the pooled  $t$ -test and the ANOVA  $F$ -test will always hold for a balanced (equal number of experimental units for each treatment) two sample design.

The two-sided, two-sample  $t$ -test using pooled variance and this ANOVA  $F$ -test are variations on the same theme. ANOVA is a generalization of the pooled two-sample  $t$  procedure, and the matrix structure used in this example can be extended to compare more than two means and to include blocking variables as well.

Analysis of Variance is often described as a comparison of signal to noise. The  $MST$  is the measure of the strength of the signal. The signal is the planned, systematic variation we are after. The  $MSE$  is the measure of the noise, or the natural chance-like variability of the process under study. We partitioned our observations 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.

### Reading Computer Output

If you perform an Analysis of Variance on the data in Table 1 using statistical software, the results will be a standard ANOVA table as shown below from JMP-INTRO.

### Oneway Analysis of Weight Gain By Diet

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Diet	1	128.0000	128.0000	6.8571	0.0397
Error	6	112.0000	18.667		
C. Total	7	240.0000			

Notice that the results in the ANOVA table are the same as in the vector computations. The sums of squares are 128, 112, and 240. The Total Sums of Squares of 240 is simply the difference in the sums of squares for  $\mathbf{Y}$  and  $\mathbf{M}$ . We see the two mean squares,  $MST$  and  $MSE$ , and the  $F$ -ratio and  $p$ -values. Now we can see how these values were computed and what they mean. For a more complete discussion of the vector/matrix representation of ANOVA for multiple comparisons, see *NCSSM Statistics Leadership Institute Notes* (1999).

### How Two Estimates of Variance 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 close to one. The farther the ratio of variances is 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 the data are

<b>Diet A (ounces)</b>	52	60	56	52
<b>Diet B (ounces)</b>	44	50	52	42

Table1: Weight Gain in Ounces and Diet Type

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

$$s_p^2 = \frac{3(14.667) + 3(22.667)}{6} = 18.667.$$

Notice this is the Mean Square Error in the earlier computation. The  $MSE$  is an estimate of the natural, chance-like variation in the response variable.

A second way to estimate the variance  $\sigma^2$  is to infer the value of  $\sigma^2$  from  $s_{\bar{Y}}^2$ , where  $s_{\bar{Y}}^2$  is the observed variance of the sample means. We calculate this by considering the means of the two treatment groups, A and B. The two means,  $\bar{y}_A = 55$  and  $\bar{y}_B = 47$ , 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  $\sigma^2$ . The variance of the observed means, 55 and 47, is  $s_{\bar{Y}}^2 = 32$ . So 32 is a estimate of the value of  $\frac{\sigma^2}{4}$ . This gives us a second estimate of  $\sigma^2$  that is equal to 128. Notice that this is the Mean Square Treatment. When the null hypothesis is true, the  $MST$  is also an estimate of the natural, chance-like variation in our response variable.

So, if the null hypothesis is true, then both 18.667 and 128 are estimates of the same population variance  $\sigma^2$ . The  $F$ -score compares these two estimates of variance, and the more these two estimates differ (the larger the  $F$ -score), the more evidence there is against the null hypothesis.

### Blocking

Suppose, due to availability, we were forced to use two different breeds of rabbits instead of just one. We have four Californian and four Florida White rabbits for use in this experiment. We believe that the Californian will grow faster than the Florida White and so the weight gains for these four rabbits will be larger than that of the other four, regardless of the diet. For example, we might expect the average weight gain for

Florida White rabbits to be about 10 ounces less than the average gain for the larger Californian rabbits. However, we don't think there will be an interaction between breed and diet. This means that the effect of each diet on the rabbits' growth will be the same additive amount. We might suspect that, for example, Diet A will add 6 ounces to the weight gain for both California and Florida White rabbits. The variability due to the two breeds is not chance-like; it is systematic, unplanned variation. We can turn this variation into chance-like variation by our random assignment process, but the variation caused by having two different breeds will be included in the estimate of chance variation, inflating it and reducing the power of the test.

A better solution comes from the process called blocking. We will use the breed as a blocking variable. We are not really interested in the effect of breed, so we think of breed as a nuisance variable. We want to estimate the amount of variation added by having two different breeds, and remove it from our estimate of the chance-like variation in our calculations. Now that we have laid the groundwork for the computations of ANOVA, we can see how this would work in practice.

In our randomization scheme, we will randomly assign two of each breed to the two Diets A and B. This is known as a randomized complete block design. For the sake of comparison, we will reconsider the results of the previous example.

	FW	C	C	FW
Diet A (ounces)	52	60	56	52
Diet B (ounces)	44	50	52	42

Table 2: Weight Gain in Ounces with Diet and Rabbit Breed

For ease of reading, we have organized the table so the Florida White rabbits are at the end of the table and the California rabbits are in the two center columns. We can now repeat the analysis as before, only now we have a slightly modified model. The mathematical model for ANOVA is  $Y = \mu + \tau + \beta + \varepsilon$ . We see that  $Y$  is decomposed into four partitions, the grand mean represented by  $\mu$ , the effect of the treatment represented by  $\tau$ , the effect of the blocking variable  $\beta$ , and the random error represented by  $\varepsilon$ . As before, we will use  $M$  for the sample estimate for the mean vector  $\mu$ ,  $T$  for the sample estimate for the treatment vector  $\tau$ ,  $B$  for the blocking vector, and  $E$  for the sample estimate of the error vector  $\varepsilon$ . So, the model is  $Y = M + T + B + E$ .

The grand mean is the average of all of the data, and this has not changed. Also, the effect of being Diet A or Diet B is the same as before, so we have the following vector equation.

$$\begin{bmatrix} 52 & 44 \\ 60 & 50 \\ 56 & 52 \\ 52 & 42 \end{bmatrix} = \begin{bmatrix} 51 & 51 \\ 51 & 51 \\ 51 & 51 \\ 51 & 51 \end{bmatrix} + \begin{bmatrix} 4 & -4 \\ 4 & -4 \\ 4 & -4 \\ 4 & -4 \end{bmatrix} + \begin{bmatrix} - & - \\ - & - \\ - & - \\ - & - \end{bmatrix} + \begin{bmatrix} - & - \\ - & - \\ - & - \\ - & - \end{bmatrix}$$

To determine the elements in the blocking vector we ask, what is the effect of being a particular type of rabbit? The mean of the Florida White rabbits is 47.5, while the mean of the California rabbits is 54.5. Since the average for all rabbits is 51, the effect of being a California rabbit is to raise the average weight gain by 3.5 ounces, while the effect of being a Florida White rabbit is to reduce the average weight gain by 3.5 ounces. This is the same kind of comparison we made when estimating the treatment effects.

$$\begin{bmatrix} 52 & 44 \\ 60 & 50 \\ 56 & 52 \\ 52 & 42 \end{bmatrix} = \begin{bmatrix} 51 & 51 \\ 51 & 51 \\ 51 & 51 \\ 51 & 51 \end{bmatrix} + \begin{bmatrix} 4 & -4 \\ 4 & -4 \\ 4 & -4 \\ 4 & -4 \end{bmatrix} + \begin{bmatrix} -3.5 & -3.5 \\ 3.5 & 3.5 \\ 3.5 & 3.5 \\ -3.5 & -3.5 \end{bmatrix} + \begin{bmatrix} - & - \\ - & - \\ - & - \\ - & - \end{bmatrix}$$

As before, the vectors  $M, T, B,$  and  $E$  are mutually perpendicular, so the Pythagorean Theorem holds. The sums of squares of the elements of  $M, T, B,$  and  $E$  must equal the sum of the squares of the elements of

$Y$ . We are interested in comparing the sums of squares and the degrees of freedom for matrix  $T$  and matrix  $E$ . Notice, we didn't really need to compute the components in the error vector, since we only want the sums of squares, and we can find those by subtraction.

$$\begin{array}{rcccccc} & & \mathbf{Y} & & \mathbf{M} & \mathbf{T} & & \mathbf{B} & & \mathbf{E} \\ \text{SS} & & 21048 & = & 20808 & + & 128 & + & 98 & + & 14 \\ \text{df} & & 8 & = & 1 & + & 1 & + & 1 & + & 5 \end{array}$$

The mean square for treatment is still  $MST = \frac{128}{1} = 128$  while the mean square for error has been reduced to

$$MSE = \frac{14}{5} = 2.8. \text{ By blocking we reduced the sum of squares error by 98 while using one degree of freedom.}$$

The ratio of these two mean squares is the value of the  $F$  statistic, with one and six degrees of freedom. In this example, we have  $F_{1,5} = \frac{128}{2.8} = 45.7$ , a much stronger signal to noise ratio. The  $p$ -value associated with this  $F$ -value is essentially zero.

Since the mean square of  $E$  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. Without blocking, the sums of squares for the error term was 112. By blocking on rabbit breed, we estimated that 98 of those sums of squares was a result of having two breeds. This nuisance variation was systematic, unplanned variation, which we turned into systematic, planned variation by blocking. We were able to mathematically remove this variation from our analysis, and consequently had a better estimate of the true, chance-like variation to use in our probability model.

### Reading Computer Output with Blocking

If you perform an Analysis of Variance on the data in Table 2 using statistical software, the results will be a standard ANOVA table.

Oneway Analysis of Weight Gain By Diet

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Diet	1	128.0000	128.0000	45.7143	0.0011
Breed	1	98.0000	98.0000	35.0000	0.0020
Error	5	14.0000	2.8000		
C. Total	7	240.0000			

Notice that, as before, the results in the ANOVA table are consistent with the vector computations. You will also see an  $F$ -ratio and  $p$ -value for the blocking variable Breed. These are computed exactly as for the treatments.

*Important Note:* We have used two different analyses on the same set of data. This was for pedagogical reasons. It allowed us to show how blocking can reduce our estimate of variability. In practice, however, you only do one experiment. The type of analysis performed on the data from the experiment depends upon the design used to create the data. An essential mantra for experimental design is: *How you randomize, how you analyze.*

One way to interpret the  $p$ -value in the hypothesis test is: if we repeated the experiment over and over again under the null hypothesis, we would obtain a test statistic ( $F$ -score) that is as or more extreme just by chance with probability  $p$ . The expression "repeated the experiment" is crucial here. It includes the randomization process. In a completely randomized experiment, some of those trials would include three Diet A's with the two center columns. This could not happen in the blocked randomization process. So, after the fact analyses are frowned upon. Once you decide how you randomize, you have determined how you must analyze the data.

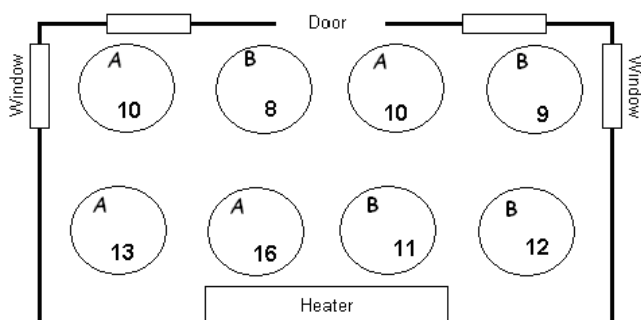
### Blocking Well and Blocking Poorly

Any time we add a blocking variable, we will reduce the sums of squares in our error vector. So, why not block on every conceivable variable? The price you pay to block is in reduced degrees of freedom. Both the numerator and denominator of our MSE is reduced by blocking. If you block inappropriately, on a variable that

does not add significantly to the variation, you will find that your MSE will increase due to the loss of degrees of freedom.

In this next example, we again want to compare two types of rabbit food for weight gain. The rabbits have already been randomly placed in eight cages in a room. The difficulty is that some of the cages must be placed near a heater, which could affect the weight gain for the rabbits near the heater. In the diagrams below, **A** represents a cage in which the rabbits were fed Diet A, while **B** represents a cage in which the rabbits were fed Diet B. The mean increase in weight is given for the rabbit in each cage. In each variation of the problem, we have chosen a different structure to our randomization of tanks to treatment. Since our analysis is tied to our randomization process, we will compare the results using the appropriate analysis for that design. In each case, we are looking at the question, “if we had used this design and the results were as given, what could we conclude?” In each situation, the null hypothesis is equal mean weight gain with the alternative unequal weight gain. All necessary conditions for the hypothesis tests are met.

a) Suppose the design used was a completely randomized design. We randomly assigned the eight cages, four cages for each treatment. If the results were as shown below, is there evidence of a difference in mean weight gain for the two types of food?



In this situation, we could use a standard two-sample *t*-procedure to analyze the completely randomized design. However, we will use the ANOVA technique simply to reinforce the procedure and for comparison with the blocked designs to follow. The mean weight gain for all rabbits is 11.125 ounces, while the mean weight gain for rabbits on Diet A is 12.25 ounces and for rabbits on Diet B is 10 ounces. Being on Diet A adds approximately 1.125 ounces to the weight gain while being on Diet B reduces the weight gain by that amount.

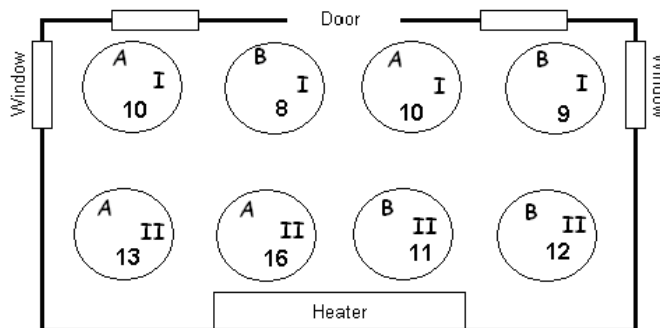
Repeating the ANOVA analysis described above, we find the following results. Notice that we didn't take the time to compute the elements of the error vector *E*, since all we really want are the appropriate sums of squares.

$$\begin{array}{c}
 \begin{matrix} & \mathbf{A} & \mathbf{B} \\ \begin{bmatrix} 10 & 8 \\ 10 & 9 \\ 13 & 12 \\ 16 & 11 \end{bmatrix} & = & \begin{bmatrix} 11.125 & 11.125 \\ 11.125 & 11.125 \\ 11.125 & 11.125 \\ 11.125 & 11.125 \end{bmatrix} & + & \begin{bmatrix} 1.125 & -1.125 \\ 1.125 & -1.125 \\ 1.125 & -1.125 \\ 1.125 & -1.125 \end{bmatrix} & + & [E] \\
 \text{SSquares } 1035 & = & 990.125 & + & 10.125 & + & 34.75 \\
 \text{df } 8 & = & 1 & + & 1 & + & 6
 \end{matrix}
 \end{array}$$

Since our  $MST = 10.125$  and  $MSE = \frac{34.75}{6}$ , we have  $F_{1,6} = \frac{\left(\frac{10.125}{1}\right)}{\left(\frac{34.75}{6}\right)} = 1.748$ , which corresponds to a *p*-

value of  $p = 0.23$ . This is the same *p*-value we would get using a 2-sample *t*-test. From this analysis, we fail to reject the null hypothesis of equal mean weight gain and conclude that there is no evidence to sustain a belief that the mean weight gain differs between the two diets.

b) The second design considers the affect of the heater to be an important contributor to the variation among the weight gains. This design blocks on the nuisance variable “close to heater”, by randomly assigning two of each diet to the four tanks close to the heater (bottom row) and two of each diet to the four tanks farther from the heater (top row). The top row is labeled Block I and the bottom row Block II. If the results were as shown below, is there evidence of a difference in mean weight gain for the two types of food? Is there evidence of an effect of the temperature gradient?



Since we used a blocked design, we must add the blocking variable to our model. The mean for Block I is 9.25 ounces and for Block II is 13 ounces. Being in Block I tends to lower the average weight gain by 1.875 ounces, while being in Block II tends to raise the average weight gain by 1.875 ounces. The estimated effect of Diets A and B remain unchanged.

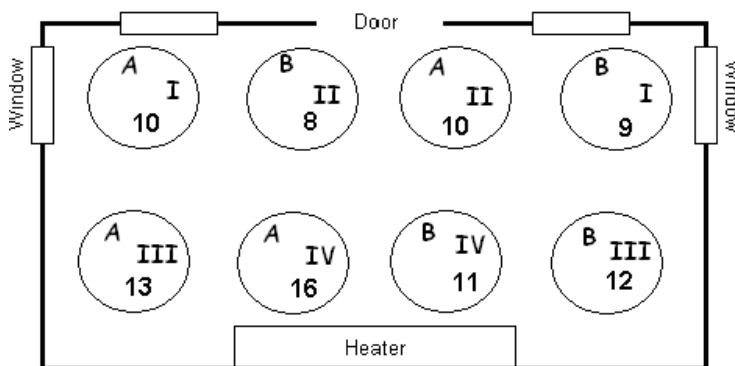
$$\begin{matrix} & \mathbf{A} & \mathbf{B} \\ \begin{bmatrix} 10 & 8 \\ 10 & 9 \\ 13 & 12 \\ 16 & 11 \end{bmatrix} & = & \begin{bmatrix} 11.125 & 11.125 \\ 11.125 & 11.125 \\ 11.125 & 11.125 \\ 11.125 & 11.125 \end{bmatrix} & + & \begin{bmatrix} 1.125 & -1.125 \\ 1.125 & -1.125 \\ 1.125 & -1.125 \\ 1.125 & -1.125 \end{bmatrix} & + & \begin{bmatrix} -1.875 & -1.875 \\ -1.875 & -1.875 \\ 1.875 & 1.875 \\ 1.875 & 1.875 \end{bmatrix} & + & [E]
 \end{matrix}$$

$$\begin{matrix} \text{SSquares} & 1035 & = & 990.125 & + & 10.125 & + & 28.125 & + & 6.625 \\ \text{df} & 8 & = & 1 & + & 1 & + & 1 & + & 5 \end{matrix}$$

Since our *MST* remains 10.125 and *MSE* is now  $\frac{6.625}{5}$ , we have  $F_{1,5} = \frac{\left(\frac{10.125}{1}\right)}{\left(\frac{6.625}{5}\right)} = 7.642$  and

$p = 0.039$ . From this analysis, we reject the null hypothesis of equal means. There is sufficient evidence to support the belief that the mean weight gain under Diet A is larger than under Diet B. We see that proximity to the heater did indeed contribute significantly to the variation in the observed weight gains. Although the degrees of freedom for error is reduced to five, we have more than made up for that with the reduction in sums of squares that was achieved.

c) In this scenario, we notice that there are really four different conditions in the room. Some of the cages are near the windows, while others near the door. The middle two cages on the bottom row are next to the heater, while the end cages on the bottom row are farther from the heater and away from the light. Suppose we block according to the relative positions in the room, taking into account both proximity to the heater and the effect of being in the light or dark. The third design also acknowledges the possible effects of windows and doors. Now we have four blocks labeled I, II, III, and IV. If the results were as shown below, is there evidence of a difference in mean weight gain for the two types of food? Does this design appear to be better than the one above?



The design we have just described is a matched pairs design and can be analyzed with a *t*-test on mean difference of the two treatments in each block. Our ANOVA analysis is equivalent, and we can see how the reduction in sums of squares interacts with the reduction in degrees of freedom.

$$\begin{matrix} & \mathbf{A} & \mathbf{B} \\ \begin{bmatrix} 10 & 8 \\ 10 & 9 \\ 13 & 12 \\ 16 & 11 \end{bmatrix} & = & \begin{bmatrix} 11.125 & 11.125 \\ 11.125 & 11.125 \\ 11.125 & 11.125 \\ 11.125 & 11.125 \end{bmatrix} & + & \begin{bmatrix} 1.125 & -1.125 \\ 1.125 & -1.125 \\ 1.125 & -1.125 \\ 1.125 & -1.125 \end{bmatrix} & + & \begin{bmatrix} -1.625 & -1.625 \\ -2.125 & -2.125 \\ 1.375 & 1.375 \\ 2.375 & 2.375 \end{bmatrix} & + & [E] \\ \text{SSquares} & 1035 & = & 990.125 & + & 10.125 & + & 29.375 & + & 5.375 \\ \text{df} & 8 & = & 1 & + & 1 & + & 3 & + & 3 \end{matrix}$$

In this analysis, the degrees of freedom for the blocks is three, since knowing the first three numbers in the first column allows us to complete the components in the vector. The sums of squares for the blocking variable is significant, 29.375, but this is only marginally larger than the sums of squares when only proximity to heater was used. The loss in degrees of freedom is a greater concern. We see now that

$$F_{1,3} = \frac{\left(\frac{10.125}{1}\right)}{\left(\frac{5.375}{3}\right)} = 5.65, \text{ with } p = 0.0979. \text{ From this test, we would fail to reject the null hypothesis in the}$$

belief that there is no difference in mean weight gain for the two diets. Because we blocked on a variable that did not add appreciably to the variation, we missed an effect that was really there. Our test lost power.

Again, it is important to note that we must make our decision on if and how to block at the beginning of the experiment. We can't block as in example c) and then analyze the results as if we had blocking like example b). The randomization process for b) would allow both cages next to the door to be Diet A, while randomization process c) would not. One essential lesson is that you should think very carefully about each blocking variable to decide if the nuisance variation it contributes is significant and important, or if it make only a minor contribution to the observed variation.

**Conclusion**

In AP Statistics students are taught that a good experiment requires Control, Replication, and Randomization. Each of these attributes offers the experimenter a way to manage the inescapable variability inherent in the experimental process.

The planned, systematic variability is emphasized by controlling extraneous sources of variation, while the chance-like variability is managed by randomization and reduced by replication and control. Finally, the unplanned, systematic variability that can destroy our results can be managed by blocking when its causes are recognized prior to the experiment and through randomization in any event.

The experimenter must always pay careful attention to the design of the experiment, since the method of analysis is determined by the manner in which the experimental units are randomized to treatments. The way you randomize is the way you analyze.

### ANOVA: Testing More Than Two Means

Suppose we now add a third brand of rabbit food, Brand C. If we again use a completely randomized design 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 ounces.

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 differs from another. We still have the same additive model,  $Y = \mu + \tau + \varepsilon$ , 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 now 53. The means of the columns are 55, 47, and 57. The effect of being Brand A is to add 2 ounces per rabbit, while Brand B subtracts 6 ounces per rabbit, and Brand C adds 4 ounces per rabbit. We can use these values to find the elements of  $E$  as before. So our “matrices” are:

$$\begin{matrix}
 \mathbf{Y} & & \mathbf{M} & & \mathbf{T} & & \mathbf{E} \\
 \begin{bmatrix} 52 & 44 & 60 \\ 60 & 50 & 58 \\ 56 & 52 & 60 \\ 52 & 42 & 50 \end{bmatrix} & = & \begin{bmatrix} 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \end{bmatrix} & + & \begin{bmatrix} 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \end{bmatrix} & + & \begin{bmatrix} -3 & -3 & 3 \\ 5 & 3 & 1 \\ 1 & 5 & 3 \\ -3 & -5 & -7 \end{bmatrix}
 \end{matrix}$$

Notice, as before, the columns of  $T$  are constant and the rows sum to zero, while for  $E$ , the columns sum to zero. The vectors  $M, T$ , and  $E$  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

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

The sums of squares for  $E$  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  $E$  to compute its sums of squares.

The degrees of freedom are 12 for  $Y$ , 1 for  $M$  (since the elements are all the same), 2 for  $T$  (since the column entries are all the same and the rows must add to zero), and 9 for  $E$  (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 differs 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  $Y$  or of  $M$  in computing  $F$ . We are only interested in the sums of squares of  $T$  and  $E$ , 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
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

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.

**Blocking to Reduce Variability**

Suppose we had done the experiment differently. Suppose we couldn’t find 12 of the same breed of rabbit, but could get three rabbits each of four different breeds. Now, we need to randomly assign one of each breed to each type of food. This experimental design would be a randomized block design. To facilitate comparisons, suppose the results were the same as before.

$$Y = \begin{matrix} & \mathbf{A} & \mathbf{B} & \mathbf{C} \\ \mathbf{I} & \begin{bmatrix} 52 & 44 & 60 \end{bmatrix} \\ \mathbf{II} & \begin{bmatrix} 60 & 50 & 58 \end{bmatrix} \\ \mathbf{III} & \begin{bmatrix} 56 & 52 & 60 \end{bmatrix} \\ \mathbf{IV} & \begin{bmatrix} 52 & 42 & 50 \end{bmatrix} \end{matrix}$$

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  $Y$  are Breed I, all entries in the second row of  $Y$  are Breed II, all entries in the third row of  $Y$  are Breed III, and all entries in the fourth row of  $Y$  are Breed IV.

By blocking we modify our mathematical 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 ounces. The average for Breed I is 52, so this Breed gains one fewer ounce per rabbit than the overall average. The effect of being in Row 1 is  $-1$ . The average for Breed II is 56, so in general, this breed gained an extra three ounces per rabbit. The mean for Breed III is also 56, and for Breed IV is 48, representing an average decrease of 5 ounces. The structure of “matrix”  $B$ , our sample estimate of  $\beta$ , is

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

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

$$\begin{bmatrix} 52 & 44 & 60 \\ 60 & 50 & 58 \\ 56 & 52 & 60 \\ 52 & 42 & 50 \end{bmatrix} = \begin{bmatrix} 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \\ 53 & 53 & 53 \end{bmatrix} + \begin{bmatrix} 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \\ 2 & -6 & 4 \end{bmatrix} + \begin{bmatrix} -1 & -1 & -1 \\ 3 & 3 & 3 \\ 3 & 3 & 3 \\ -5 & -5 & -5 \end{bmatrix} + [E]$$

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

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

with

SS	34,112	=	33,708	+	224	+	132	+	48.
df	12	=	1	+	2	+	3	+	6

Notice that blocking did not affect the sums of square of  $Y, M$ , or  $T$ . The additional sum of squares for  $B$  must come out of  $E$ . Since the mean square of  $E$  is a measure of variability, we can see how blocking reduces variation. The degrees of freedom also change. The “matrix”  $B$  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,  $H_0: \mu_I = \mu_{II} = \mu_{III} = \mu_{IV}$  and the alternative is  $H_a$ : at least one mean is different.

$$\text{Our } F\text{-score is } F_{3,6} = \frac{MSB}{MSE} = \frac{\left(\frac{132}{3}\right)}{\left(\frac{48}{6}\right)} = \frac{44}{8} = 5.5. \text{ The associated } p\text{-value is } p = 0.0371. \text{ There is}$$

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: Ounces

Effect Test

Source	Nparm	DF	Sum of Squares	F Ratio	Prob>F
Brand	2	2	224.00000	14.0000	0.0055
Breed	3	3	132.00000	5.5000	0.0371

Whole-Model Test

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	356.00000	71.2000	8.9000
Error	6	48.00000	8.0000	Prob>F
C Total	11	404.00000		0.0096

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

Breed

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  $Y = \mu + (\tau + \beta) + \varepsilon$ . This model considers  $(\tau + \beta)$  as a single factor and combines the sums of squares and degrees of freedom of  $T$  and  $B$  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 Breed 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 can be obtained 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.

**Extending Blocking: A 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
$\bar{y}_A = 18$	$\bar{y}_1 = 19$	$\bar{y}_I = 23$
$\bar{y}_B = 22$	$\bar{y}_2 = 20$	$\bar{y}_{II} = 24$
$\bar{y}_C = 21$	$\bar{y}_3 = 19$	$\bar{y}_{III} = 15$
$\bar{y}_D = 19$	$\bar{y}_4 = 22$	$\bar{y}_{IV} = 18$

Our additive model for ANOVA is  $Y = \mu + \tau + \beta_D + \beta_C + \varepsilon$ , and the “matrix” structure is given below:

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

The sums of squares and degrees of freedom are now computed, with  $\varepsilon$  found by subtraction.

	<b><i>Y</i></b>	=	<b><i>M</i></b>	+	<b><i>T</i></b>	+	<b><i>B<sub>D</sub></i></b>	+	<b><i>B<sub>C</sub></i></b>	+	<b><i>E</i></b>
SS	6696		6400		40		24		216		16
df	16		1		3		3		3		6

**Additives:** We can now test additives.  $H_0: \mu_A = \mu_B = \mu_C = \mu_D$   
 $H_a$ : at least one mean differs.

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 of equal means for the additives. At least one population mean differs from another.

**Drivers:** We can also test drivers.  $H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$   
 $H_a$ : at least one mean differs

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  $p = 0.1170$ . We fail to reject the null

hypothesis. Our observations are consistent with random variation. There is not sufficient evidence to conclude that the drivers differ on the basis of mean emission of pollutants.

**Cars:** We can also test cars.  $H_0: \mu_I = \mu_{II} = \mu_{III} = \mu_{IV}$   
 $H_a$ : at least one mean differs.

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 of the cars has a different mean level of pollution emission from the others.

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 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*. 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  $MSE$ . Note that the  $LSD$  procedure is used only when the  $F$ -test indicates a significant difference exists.

**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 ounces gained per rabbit, but that Brand A and C are indistinguishable. In short, Brand B is worse on average than both A and C if you want bigger rabbits.

**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 breeds, we now can conclude that any treatment means more than 4.9 ounces 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 Breeds. We found a significant difference in the mean weight gain between breeds of rabbit. The averages for the breeds were 52, 56, 56, and 48, for Breeds 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 breeds more than 5.65 ounces apart are considered significantly different. The maximum difference between two means for the breeds is 8, so Breed IV differs significantly from Breeds II and III with respect to the mean number of ounces gained. In this sense, Breed IV is the smaller rabbit breed. However, we cannot claim that Breed IV differs significantly from Breed I since the difference in means is 4, which is smaller than the computed  $LSD$ . For the same reason, we cannot distinguish Breed I from Breed 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 and Brands D and C. However, the mean levels of pollutant for Brands C and B are significantly larger than for Brand A while the mean levels of pollution for Brands A and D are significantly 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.

## Two-Factor Designs

We have seen how to use a two-way ANOVA to compare two or more treatments in the presence of one or more nuisance variables. By blocking on the nuisance variables, we were able to remove the variation attributed to those variables from the sum of squared errors (SSE) and consequently, from our estimate of the natural variation of the process under study. In this section, we will consider a variation of the two-way ANOVA that will allow us to compare two explanatory variables and to assess the level of interaction between the two variables.

**Battery Problem:** Students want to compare the performance of three kinds of batteries in a scientific calculator and a TI-89 Titanium calculator. Six each kind of calculator were used in the study. Four calculators (two of each type) were filled with Duracell, Eveready, or Wal-Mart brand batteries. They were programmed to perform some simple computations in an infinite loop. They were timed until the program failed. The time in minutes was recorded.

This is a two factor design, with Battery Type at 3 levels and Calculator Type at 2 levels.

	Scientific	TI-89 Titanium
Duracell	27.6, 28.2	22.4, 25.0
Eveready	23.9, 24.5	25.4, 26.6
Wal-Mart	18.4, 19.0	15.5, 17.7

The vector of observed values is structured to make it easy to keep track of the two treatments Calculator Type and Battery Type.

	<i>S</i>	<i>TI</i>
<i>D</i>	27.6	22.4
<i>D</i>	28.2	25.0
<i>E</i>	23.9	25.4
<i>E</i>	24.5	26.6
<i>W</i>	18.4	15.5
<i>W</i>	19.6	17.7

The model we are using is a variation of the additive model used previously. Now, we have  $Y = \mu + \tau_C + \tau_B + I_{CB} + \epsilon$ . The observed values are composed of the grand mean, the effect of calculator, the effect of battery, the interaction between calculator and battery, and the error or residual. As before, the MSE is our estimate of the natural variation or noise.

As before, we need to compute the overall mean ( $\bar{x} = 22.9$ ) and the means for each subset of the data. This includes the mean for Scientific Calculators ( $\bar{x}_{Sci} = 23.7$ ), the mean for TI-89 Calculators ( $\bar{x}_{TI} = 22.1$ ), the mean for Duracell Batteries ( $\bar{x}_D = 25.8$ ), the mean for Eveready Batteries ( $\bar{x}_E = 25.1$ ), and the mean for Wal-Mart Batteries ( $\bar{x}_W = 17.8$ ). We create our mean and treatment vectors as before. The overall mean is 22.9 and the mean for Scientific Calculators is 23.7, so the effect of being a Scientific Calculator (1<sup>st</sup> column) is to add 0.8 minutes to the time. Consequently, the effect of TI-89 is to subtract 0.8. Similarly, the effects of the Battery types are computed. These are called the Main Effects of the treatments.

<i>S</i>	<i>TI</i>	<i>Mean</i>	<i>Calc</i>	<i>Batt</i>	<i>Inter Resid</i>
----------	-----------	-------------	-------------	-------------	--------------------

$$\begin{array}{l}
 D \\
 D \\
 E \\
 E \\
 W \\
 W
 \end{array}
 \begin{bmatrix}
 27.6 & 22.4 \\
 28.2 & 25.0 \\
 23.9 & 25.4 \\
 24.5 & 26.6 \\
 18.4 & 15.5 \\
 19.6 & 17.7
 \end{bmatrix}
 =
 \begin{bmatrix}
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9
 \end{bmatrix}
 +
 \begin{bmatrix}
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8
 \end{bmatrix}
 +
 \begin{bmatrix}
 2.9 & 2.9 \\
 2.9 & 2.9 \\
 2.2 & 2.2 \\
 2.2 & 2.2 \\
 -5.1 & -5.1 \\
 -5.1 & -5.1
 \end{bmatrix}
 + [I] + [\varepsilon]$$

To assess the interaction between Battery Type and Calculator Type, we also need to include some new subsets in our calculations. We want to know the mean for Scientific Calculators on Duracell Batteries

( $\bar{x}_{D/S} = 27.9$ ), the mean for TI-89 Calculators on Duracell Batteries ( $\bar{x}_{D/TI} = 23.7$ ), the mean for Scientific Calculators on Eveready Batteries ( $\bar{x}_{E/S} = 24.2$ ), the mean for TI-89 Calculators on Eveready Batteries ( $\bar{x}_{E/TI} = 26.0$ ), the mean for Scientific Calculators on Wal-Mart Batteries ( $\bar{x}_{W/S} = 19.0$ ), and the mean for TI-89 Calculators on Wal-Mart Batteries ( $\bar{x}_{W/TI} = 16.6$ ).

The computation of the effect of the interaction between Calculator Type and Battery Type is a bit more complicated. Without thinking too deeply about it, one might perform the same computation as with the two main effects. To see that this is not correct, we will perform that operation and see what goes wrong.

Since the mean for Scientific Calculators on Duracell Batteries is  $\bar{x}_{D/S} = 27.9$ , we could estimate the effect of this interaction to be +5. Continuing in like manner, we can complete the vector to arrive at the vector equation shown below. What is wrong with this equation?

$$\begin{array}{l}
 D \\
 D \\
 E \\
 E \\
 W \\
 W
 \end{array}
 \begin{bmatrix}
 S & TI \\
 27.6 & 22.4 \\
 28.2 & 25.0 \\
 23.9 & 25.4 \\
 24.5 & 26.6 \\
 18.4 & 15.5 \\
 19.6 & 17.7
 \end{bmatrix}
 =
 \begin{bmatrix}
 Mean \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9 \\
 22.9 & 22.9
 \end{bmatrix}
 +
 \begin{bmatrix}
 Calc \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8 \\
 0.8 & -0.8
 \end{bmatrix}
 +
 \begin{bmatrix}
 Batt \\
 2.9 & 2.9 \\
 2.9 & 2.9 \\
 2.2 & 2.2 \\
 2.2 & 2.2 \\
 -5.1 & -5.1 \\
 -5.1 & -5.1
 \end{bmatrix}
 +
 \begin{bmatrix}
 Inter \\
 5.0 & 0.8 \\
 5.0 & 0.8 \\
 1.3 & 3.1 \\
 1.3 & 3.1 \\
 -3.9 & -6.3 \\
 -3.9 & -6.3
 \end{bmatrix}
 + [\varepsilon]$$

There is a fundamental flaw in this calculation. Our vector of interaction effects is not perpendicular to the other vectors. We can see this in two ways. The dot products are not zero and the sum of square of vectors *M*, *C*, *B*, and *I* are already greater than the sum of squares of vector *Y*. So, something is wrong.

To compute the elements of the Interaction vector, we compare the average of the two entries for Scientific Calculators on Duracell Batteries (27.9) with the partial sum of the Main Effects ( $22.9 + 0.8 + 2.9 = 26.6$ ). The difference is 1.3. So, we estimate the effect of this interaction to be 1.3. The average of the two entries for Scientific Calculators on Eveready Batteries (24.2) is compared with the partial sum of the Main Effects ( $22.9 + 0.8 + 2.2 = 25.9$ ). The difference is  $-1.7$ . So, we estimate the effect of this interaction to be  $-1.7$ . The other elements of the vector are computed in the same way. We actually have all the information we need with these two calculations, since the rows and columns must add to zero, so this vector has two degrees of freedom leaving six for the residual vector.

	<i>S</i>	<i>TI</i>	<i>Mean</i>	<i>Calc</i>	<i>Batt</i>	<i>Inter</i>	<i>Resid</i>
--	----------	-----------	-------------	-------------	-------------	--------------	--------------

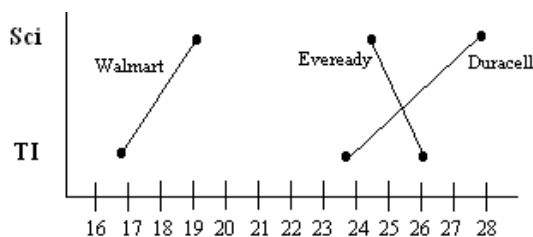
$$\begin{matrix} D \\ D \\ E \\ E \\ W \\ W \end{matrix} \begin{bmatrix} 27.6 & 22.4 \\ 28.2 & 25.0 \\ 23.9 & 25.4 \\ 24.5 & 26.6 \\ 18.4 & 15.5 \\ 19.6 & 17.7 \end{bmatrix} = \begin{bmatrix} 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \end{bmatrix} + \begin{bmatrix} 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \end{bmatrix} + \begin{bmatrix} 2.9 & 2.9 \\ 2.9 & 2.9 \\ 2.2 & 2.2 \\ 2.2 & 2.2 \\ -5.1 & -5.1 \\ -5.1 & -5.1 \end{bmatrix} + \begin{bmatrix} 1.3 & -1.3 \\ 1.3 & -1.3 \\ -1.7 & 1.7 \\ -1.7 & 1.7 \\ 0.4 & -0.4 \\ 0.4 & -0.4 \end{bmatrix} + \begin{bmatrix} -0.3 & -1.3 \\ 0.3 & 1.3 \\ -0.3 & -0.6 \\ 0.3 & 0.6 \\ -0.6 & -1.1 \\ 0.6 & 1.1 \end{bmatrix}$$

SS: 6484.2 = 6292.92 + 7.68 + 157.04 + 18.96 + 7.6  
 df: 12 = 1 + 1 + 2 + 2 + 6

If there is a significant interaction, we do not consider Main Effects. We will look more carefully at this shortly. Always consider the interaction first. The null hypothesis is that there is no interaction between the factors Battery Type and Calculator Type while the alternative is that there is an interaction. Our calculations

proceed as before:  $F_{2,6} = \frac{(\frac{18.96}{2})}{(\frac{7.6}{6})} = 7.48$  and  $p = 0.023$ .

We reject the null hypothesis of no interaction in favor of the alternative. We believe that the mean life of the battery changes with calculator or the survival time of the calculator changes with battery. Indeed, the diagram below illustrates the interaction. The Scientific Calculator lasts longer on Duracell batteries while the TI-89 last longer on Eveready Batteries.



Since there is a significant interaction, we don't consider the individual main effects. The LSD is  $2.45\sqrt{\frac{7.6}{6}(\frac{1}{2} + \frac{1}{2})} = 2.75$ . Any difference in means greater than 2.75 are considered significant at 0.05 level.

We can conclude that Wal-Mart batteries do not perform significantly differently in the two types of calculators, but they perform significantly below all other batteries. The Duracell-Scientific Calculator combination performs significantly better than the Duracell-TI and the Eveready-Scientific combination. There is no difference in mean time to failure among Duracell-TI, Eveready-Scientific, and Eveready-TI combinations.

**The Importance of Replicates to Assess Interaction**

Suppose for the moment that we had only one measurement for each of the combinations. Without the replicates in each cell, we would not be able to compute an estimate of the interaction effect.

	Scientific	TI-89 Titanium
Duracell	27.6	22.4
Eveready	23.9	23.0
Wal-Mart	18.4	15.5

$$\begin{bmatrix} 27.6 & 22.4 \\ 23.9 & 23.0 \\ 18.4 & 15.5 \end{bmatrix} = \begin{bmatrix} 21.8 & 21.8 \\ 21.8 & 21.8 \\ 21.8 & 21.8 \end{bmatrix} + \begin{bmatrix} 1.5 & -1.5 \\ 1.5 & -1.5 \\ 1.5 & -1.5 \end{bmatrix} + \begin{bmatrix} 3.2 & 3.2 \\ 1.65 & 1.65 \\ -4.85 & -4.85 \end{bmatrix} + \begin{bmatrix} 1.1 & -1.1 \\ -1.05 & 1.05 \\ -0.05 & 0.05 \end{bmatrix} + [\epsilon]$$

If we now compute the interaction effect by comparing the mean of each cell (with one entry, that's just the value of the entry) to the partial sum. For Scientific Calculators on Duracell Batteries (27.6), the partial sum of the Main Effects is  $21.8 + 1.5 + 3.2 = 26.5$ . The difference is 1.1. So, we estimate the effect of this

interaction to be 1.1. For Scientific Calculators on Eveready Batteries (23.9), the partial sum of the Main Effects is  $21.8 + 1.5 + 1.65 = 24.95$ . The difference is  $-1.05$ . So, we estimate the effect of this interaction to be  $-1.05$ . The other elements of the vector are computed in the same way or by subtraction. Now, what's wrong with that? The residual vector must be all zeros! Our interaction vector is confounded with the residual vector and we can't separate the two without at least one replicate for each combination.

**Alternative Completely Randomized Design Approach**

We could have designed this experiment as a Completely Randomized Design with six treatments using the six different battery-calculator combinations with two replicates in each. The analysis would be a standard ANOVA as shown below:

$$\begin{matrix} DS & DTI & ES & ETI & WS & WTI \\ \begin{bmatrix} 27.6 & 22.4 & 23.9 & 25.4 & 18.4 & 15.5 \\ 28.2 & 25.0 & 24.5 & 26.6 & 19.6 & 17.7 \end{bmatrix} & = & [22.9] & + & \begin{bmatrix} 5 & 0.8 & 1.3 & 3.1 & -3.9 & -6.3 \\ 5 & 0.8 & 1.3 & 3.1 & -3.9 & -6.3 \end{bmatrix} & + & [R] \end{matrix}$$

**SS:**  $6484.2 = 6292.92 + 183.68 + 7.6$   $F_{2,6} = \frac{\left(\frac{18.96}{2}\right)}{\left(\frac{7.6}{6}\right)} = 7.48.$

**df:**  $12 = 1 + 5 + 6$   $p = 0.023$

We see that we have the same *F*-score and *p*-value and the conclusions would have been the same as in the earlier analysis.

**Calculator/Battery Problem without Considering the Interaction Effect**

Suppose we had failed to consider the interaction between Battery and Calculator. What conclusions would our analysis have led us to? The computations are straightforward.

	<i>S</i>	<i>TI</i>	<i>Mean</i>	<i>Calc</i>	<i>Batt</i>	<i>Resid</i>
<i>D</i>	27.6	22.4	22.9	22.9	0.8	-0.8
<i>D</i>	28.2	25.0	22.9	22.9	0.8	-0.8
<i>E</i>	23.9	25.4	22.9	22.9	0.8	-0.8
<i>E</i>	24.5	26.6	22.9	22.9	0.8	-0.8
<i>W</i>	18.4	15.5	22.9	22.9	0.8	-0.8
<i>W</i>	19.6	17.7	22.9	22.9	0.8	-0.8

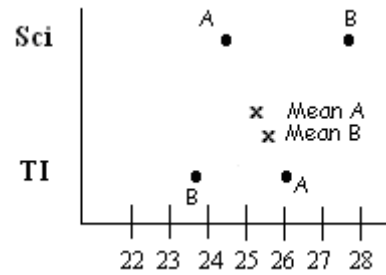
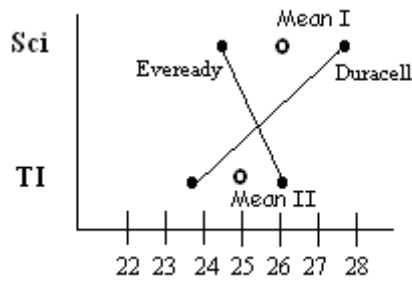
$$= \begin{bmatrix} 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \\ 22.9 & 22.9 \end{bmatrix} + \begin{bmatrix} 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \\ 0.8 & -0.8 \end{bmatrix} + \begin{bmatrix} 2.9 & 2.9 \\ 2.9 & 2.9 \\ 2.2 & 2.2 \\ 2.2 & 2.2 \\ -5.1 & -5.1 \\ -5.1 & -5.1 \end{bmatrix} + [\varepsilon]$$

**SS:**  $6484.2 = 6292.92 + 7.68 + 157.04 + 26.56$   
**df:**  $12 = 1 + 1 + 2 + 8$

For the factor Calculator Type, we have  $F_{1,8} = \frac{\left(\frac{7.68}{1}\right)}{\left(\frac{26.56}{8}\right)} = 2.31$  and  $p = 0.167$ . There is no evidence of

a difference in mean times for the two calculators. For the factor Battery Type, we have  $F_{2,8} = \frac{\left(\frac{157.04}{2}\right)}{\left(\frac{26.56}{8}\right)} = 23.67$

and  $p = 0.00043$ . There is ample evidence of a difference in mean times for the two 3 batteries. The difference in the two calculator types is hidden by the interaction between battery and calculator. Consider just Batteries A and B.



**References:**

Box, George, William Hunter, and J. Stuart Hunter, *Statistics for Experimenters*, John Wiley & Sons, New York, New York, 1978.

Cobb, George W., *Introduction to the Design and Analysis of Experiments*, Springer-Verlag, New York, New York, 1998.

Iman, Ronald, L., *A Data-Based Approach to Statistics*, Duxbury Press, Belmont, California, 1994.

Sall, John, and Ann Lehman, *JMP Start Statistics*, Duxbury Press, Belmont, California, 1996.

Snedecor, George W., and William G. Cochran, *Statistical Methods, 6th*, The Iowa State University Press, Ames, Iowa, 1967.

NCSSM Statistics Leadership Institute Notes, [http://courses.ncssm.edu/math/Stat\\_Inst/Notes.htm](http://courses.ncssm.edu/math/Stat_Inst/Notes.htm), 1999.