Two-Factor Analysis of Variance (ANOVA)

Researchers were interested in studying the effect of three varieties of corn (A_1, A_2, and A_3) and two fertilizer types (B_1 and B_2) on crop yield. A field was divided into 24 plots of equal size. Each of the 6 combinations of variety and fertilizer (A_1B_1, A_1B_2, A_2B_1, A_2B_2, A_3B_1, A_3B_2) were assigned to 4 plots selected at random. Data on yield (in bushels per acre minus 100) are presented in the table below.

	Fertilizer Type							
Variety	B_1				B_2			
A_1	37	40	45	46	47	45	42	40
A_2	40	43	47	42	37	47	41	47
A_3	41	46	52	45	51	55	52	57

There are many questions that we might want to answer with the help of this data. Do the different varieties of corn have significantly different yields? Do the different fertilizer types result in significantly different yields? Is the effect of fertilizer type on yield the same for all three varieties? It will be helpful to discuss some terminology, basic assumptions, and notation before attempting to answer these and related questions.

Some Terminology

1. A *factor* is an explanatory variable studied in an investigation. We will use capital letters A, B, C, etc. to denote the names of generic factors. Give specific names to the two factors studied in the example above.
2. The different values of a factor are called *levels*. In this example, factor A has 3 levels (A_1, A_2, A_3), and factor B has 2 levels (B_1, B_2).
3. A combination of one level from each factor is a *treatment*. This example has how many treatments?
4. Treatments are applied to *experimental units*. What are the experimental units in the example above?
5. The measures of the *response variable* are used to make comparisons among treatments. What is the response variable in the example considered here?

Two-Factor ANOVA Model

- We are studying the effect of two factors A and B on a response variable.
- Factor A has a levels. Factor B has b levels.
- There are $r > 1$ experimental units for each of the ab treatments.
- y_{ijk} denotes the measurement of the response variable for the kth experimental unit exposed to the ith level of factor A and the jth level of factor B.
- For the treatment defined by the ith level of factor A and the jth level of factor B,

$$y_{ij1}, \ldots, y_{ijr} \sim N(\mu_{ij}, \sigma^2) \quad \text{or} \quad y_{ijk} = \mu_{ij} + e_{ijk} \text{ where } e_{ijk} \sim N(0, \sigma^2).$$

- All observations are independent.

6. Find a, b, r, Y_{113}, Y_{123}, Y_{311}, Y_{224} for the yield experiment.

Two-Factor ANOVA Notation

The ijth sample $(y_{ij1}, \ldots, y_{ijr})$ has mean $\bar{y}_{ij\cdot} = \frac{1}{r}\sum_{k=1}^{r} y_{ijk}$, and variance $s_{ij}^2 = \frac{1}{r-1}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij\cdot})^2$.

$\bar{y}_{i\cdot\cdot} = \frac{1}{br}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$ $\quad \bar{y}_{\cdot j\cdot} = \frac{1}{ar}\sum_{i=1}^{a}\sum_{k=1}^{r} y_{ijk}$ $\quad \bar{y}_{\cdots} = \frac{1}{abr}\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r} y_{ijk}$

$\bar{\mu}_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij}$ $\quad \bar{\mu}_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij}$ $\quad \mu = \bar{\mu}_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$

$\alpha_i = \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot\cdot}$ $\quad \beta_j = \bar{\mu}_{\cdot j} - \bar{\mu}_{\cdot\cdot}$ $\quad (\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot}$

7. What are $\bar{y}_{21\cdot}$, $\bar{y}_{3\cdot\cdot}$, $\bar{y}_{\cdot 1\cdot}$, and s_{21}^2 for the data from the yield experiment?

8. Show that $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$. It follows that we may write our model for the data as

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + e_{ijk}, \text{ where } e_{ijk} \text{ independent } N(0, \sigma^2).$$

Partitioning the Total Sum of Squares

Source	D.F.	Sum of Squares	Mean Squares	F	Expected Mean Square
Treatment	$ab-1$	$SSTR$	$MSTR$	$\frac{MSTR}{MSE}$	$\sigma^2 + \frac{r\sum_{i=1}^{a}\sum_{j=a}^{b}(\mu_{ij}-\mu)^2}{ab-1}$
Factor A	$a-1$	SSA	MSA	$\frac{MSA}{MSE}$	$\sigma^2 + \frac{rb\sum_{i=1}^{a}\alpha_i^2}{a-1}$
Factor B	$b-1$	SSB	MSB	$\frac{MSB}{MSE}$	$\sigma^2 + \frac{ra\sum_{j=1}^{b}\beta_j^2}{b-1}$
Interaction	$(a-1)(b-1)$	SSI	MSI	$\frac{MSI}{MSE}$	$\sigma^2 + \frac{r\sum_{i=1}^{a}\sum_{j=a}^{b}(\alpha\beta)_{ij}^2}{(a-1)(b-1)}$
Error	$ab(r-1)$	SSE	MSE		σ^2
Total	$abr-1$	$SSTO$			

$$SSTO = SSTR + SSE \qquad SSTR = SSA + SSB + SSI$$

$$SSTO = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{\cdots})^2 \quad SSTR = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{\cdots})^2 \quad SSE = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij\cdot})^2$$

$$SSA = br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 \quad SSB = ar\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2 \quad SSI = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

Two-Factor ANOVA (continued)

A Test for Interaction between Factors A and B

Factors A and B are said to *interact* if the effect of factor A depends on the level of factor B, or, equivalently, if the effect of factor B depends on the level of factor A.

Is the effect of fertilizer type on yield the same for all three varieties? To answer this question we test for significant interaction between variety and fertilizer type. If there is significant interaction between variety and fertilizer type, then the effect of fertilizer type on yield will not be the same for all three varieties. If there is no interaction, the effect of fertilizer type is the same for all three varieties.

The test for interaction is based on the statistic $\frac{MSI}{MSE}$.

The null hypothesis says there is no interaction between A and B. In symbolic form, the null hypothesis is

$$H_0 : (\alpha\beta)_{ij} = 0 \text{ for all } i = 1, \ldots, a \text{ and } j = 1, \ldots, b.$$

When the null hypothesis is true, $\frac{MSI}{MSE}$ has an F distribution with $(a-1)(b-1)$ and $ab(r-1)$ degrees of freedom.

We reject H_0 for large values of $\frac{MSI}{MSE}$. To see why this makes some sense note that $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}$ is a natural estimate of $(\alpha\beta)_{ij} = \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdots}$. Thus the quantities $\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots}$ should be close to zero when H_0 is true. It follows that

$$MSI = \frac{r}{(a-1)(b-1)} \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2$$

should be close to zero when H_0 is true. What is meant by "close to zero" will depend on σ^2. If σ^2 is large, we might see quite large values of MSI even when H_0 is true. By examining the expected mean squares, we can see that MSI should be around the same size as MSE when H_0 is true.

A Test for Factor A Main Effects

When there is no important interaction between factors A and B, it makes sense to test for significant factor A main effects. In the crop yield example, we would want to know if the different varieties have different mean yields.

In symbolic form, the null hypothesis is

$$H_0 : \bar{\mu}_{1\cdot} = \bar{\mu}_{2\cdot} = \cdots = \bar{\mu}_{a\cdot} \qquad \text{or} \qquad H_0 : \alpha_i = 0 \text{ for all } i = 1, \ldots, a.$$

When the null hypothesis is true, $\frac{MSA}{MSE}$ has an F distribution with $a-1$ and $ab(r-1)$ degrees of freedom.

We reject H_0 for large values of $\frac{MSA}{MSE}$. To see why this makes some sense note that $\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}$ is a natural estimate of $\alpha_i = \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdots}$. Thus the quantities $\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots}$ should be close to zero when H_0 is true. It follows that

$$MSA = \frac{br}{a-1} \sum_{i=1}^{a} (\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$$

should be close to zero when H_0 is true. What is meant by "close to zero" will depend on σ^2. If σ^2 is large, we might see quite large values of MSA even when H_0 is true. By examining the expected mean squares, we can see that MSA should be around the same size as MSE when H_0 is true.

A Test for Factor B Main Effects

When there is no important interaction between factors A and B, it makes sense to test for significant factor B main effects. In the crop yield example, we would want to know if the different fertilizer types have different mean yields.

In symbolic form, the null hypothesis is

$$H_0 : \bar{\mu}_{\cdot 1} = \bar{\mu}_{\cdot 2} = \cdots = \bar{\mu}_{\cdot b} \qquad \text{or} \qquad H_0 : \beta_j = 0 \text{ for all } j = 1, \ldots, b.$$

When the null hypothesis is true, $\frac{MSB}{MSE}$ has an F distribution with $b-1$ and $ab(r-1)$ degrees of freedom.

We reject H_0 for large values of $\frac{MSB}{MSE}$.

A General Analysis Strategy

1. Test for interaction between A and B.
2. If interaction is significant and important, make comparisons among the levels of A for each level of factor B and/or make comparisons among the levels of B for each level of factor A. If interaction is not significant or significant but not important, continue with step 3.
3. Test for factor A main effects. Test for factor B main effects. If either or both sets of main effects are significant continue with step 4.
4. Make comparisons among the levels of each of the factors with significant main effects. For example, compare the mean yields of varieties A_1, A_2, and A_3 to one another if $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ is rejected.

Comparison of Factor Levels

Suppose, for example, that variety and fertilizer type interact in the corn yield study. We could compare the yields of the three varieties to one another for fertilizer type B_1. The same comparisons among the varieties could be made separately for fertilizer type B_2. Because of significant interaction, we might find, for example, that variety A_1 is best when fertilizer B_1 is used although variety A_3 is best when fertilizer B_2 is used.

We can use results on linear combinations of means to make the desired comparisons just as in one-way ANOVA. Because we have two factors, the notation is a bit more complicated. We can test hypotheses about $C = \sum_{i=1}^{a}\sum_{j=1}^{b} k_{ij}\mu_{ij}$ or get confidence intervals for $C = \sum_{i=1}^{a}\sum_{j=1}^{b} k_{ij}\mu_{ij}$ by making use of the following facts:

$$\hat{c} = \sum_{i=1}^{a}\sum_{j=1}^{b} k_{ij}\bar{y}_{ij\cdot} \text{ estimates } C, \qquad \mathrm{SE}(\hat{c}) = \sqrt{\frac{MSE}{r}\sum_{i=1}^{a}\sum_{j=1}^{b} k_{ij}^2}, \qquad \frac{\hat{c} - C}{\mathrm{SE}(\hat{c})} \sim t \text{ with } ab(r-1) \text{ d.f.}$$

Give the formula for the t-statistic used to compare the mean for variety A_1/fertilizer B_1 to the mean for variety A_2/fertilizer B_1.

When there is no significant or no important interaction but significant main effects, we can compare the main effects for any pair of levels. Suppose, for example, that there is no interaction between variety and fertilizer type but significant variety main effects. It is natural to make comparisons among the varieties. Give the formula for a 95% confidence interval for the difference between variety A_1 and variety A_3 main effects. (This is equivalent to comparing the mean yield of variety A_1, averaged over fertilizer types, to the mean yield of variety A_3, averaged over fertilizer types.)