Lecture 30

Two-Way ANOVA Theory

In this lecture we will discuss the theory behind two-way ANOVA. We will describe the theory in the case of a completely balanced design first and then describe how one deals with an unbalanced design. The notation becomes hairier because of the additional variables involved.

Two-Way ANOVA

The data for a two-way ANOVA fits into a table with row treatments and column treatments. As an example, consider the following data concerning the percentage of iron absorbed based on the type of iron used and the dose used (data from Rice's text: Mathematical Statistics and Data Analysis):

Type	Dose in millimolars		
	10.2	1.2	0.3
Fe^{3+}	0.71 1.66 2.01 2.16 2.42 2.42 2.56 2.6 3.31 3.64 3.74 3.74 4.39 4.5 5.07 5.26 8.15 8.24	2.2 2.93 3.08 3.49 4.11 4.95 5.16 5.54 5.68 6.25 7.25 7.9 8.85 11.96 15.54 15.89 18.3 18.59	2.25 3.93 5.08 5.82 5.84 6.89 8.5 8.56 9.44 10.52 13.46 13.57 14.76 16.41 16.96 17.56 22.82 29.13
Fe^{2+}	2.2 2.69 3.54 3.75 3.83 4.08 4.27 4.53 5.32 6.18 6.22 6.33 6.97 6.97 7.52 8.36 11.65 12.45	4.04 4.16 4.42 4.93 5.49 5.77 5.86 6.28 6.97 7.06 7.78 9.23 9.34 9.91 13.46 18.4 23.89 26.39	2.71 5.43 6.38 6.38 8.32 9.04 9.56 10.01 10.08 10.62 13.8 15.99 17.9 18.25 19.32 19.87 21.6 22.25

The data are referenced as $Y_{i,j,k}$ where i refers to the i^{th} row treatment, j refers to the j^{th} column treatment, and k refers to the k^{th} individual or data point in that row-column category. The model is then

$$Y_{i,j,k} = \mu + \alpha_i + \beta_j + \delta_{i,j} + \epsilon_{i,j,k}$$

where μ is the overall average, α_i is the effect of the i^{th} row treatment, β_j is the effect of the j^{th} column treatment, $\delta_{i,j}$ is the interaction of the two, and $\epsilon_{i,j,k}$ is the residual error for the k^{th} individual in this group. The assumption for ANOVA is that these residuals are independent, identically distributed, and normal. Thus, the variance in the data in each row-column treatment combination should be the same. The design is balanced if there are K individuals in each row-column treatment combination. If there are I row treatments and J column treatments and K individuals in each, then the total sample size will be $n = I \times J \times K$. Balance will make your results more robust.

Notation:

$$\begin{aligned}
\hat{\mu} &= \bar{Y}_{\cdots}, \text{ (the overall average value for the response, } Y) \\
\hat{\alpha}_i &= \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots}, \text{ (the specific effect of being in the } i^{\text{th}} \text{ row treatment)} \\
\hat{\beta}_j &= \bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots}, \text{ (the specific effect of being in the } j^{\text{th}} \text{ column treatment)} \\
\hat{\delta}_{ij} &= \bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots}, \text{ (the interaction effect of being in the } (i,j) \text{ cell)}
\end{aligned}$$

With this notation, we can parse the total sums of squares as:

$$\begin{aligned}
SS_{\text{Total}} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{ijk} - \bar{Y}_{\cdots})^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\underbrace{Y_{ijk} - \bar{Y}_{ij\cdot}}_{\text{error}} + \underbrace{(\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdots})}_{\delta_{ij}} - \underbrace{(\bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdots})}_{\alpha_i} - \underbrace{(\bar{Y}_{\cdot j\cdot} - \bar{Y}_{\cdots})}_{\beta_j})^2 \\
&= \text{SSE} + \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}(\alpha_i + \beta_j - \delta_{ij})^2 \\
&= \text{SSE} + JK\sum_{i=1}^{I}\alpha_i^2 + IK\sum_{j=1}^{J}\beta_j^2 + K\sum_{i=1}^{I}\sum_{j=1}^{J}\delta_{ij}^2 \\
&= \text{SSE} + \text{SS}_{\alpha} + \text{SS}_{\beta} + \text{SS}_{\alpha\beta}
\end{aligned}$$

In an additive model, the interaction term is assumed to be zero and is absent in the above formulation.

The formula above works for a balanced design. If the design is not balanced, there is no theoretical problem but your results are more sensitive to violations in the equal variances assumption. A full two-way ANOVA, in either the balanced case or the unbalanced case, is equivalent to running a regression with indicator variables for each row and treatment group (with one used as reference) and with interactions as products of the row/column indicator variables. The main complication is that there are different methods for calculating the sums of squares in an unbalanced design:

Type I or Sequential Sums of Squares are based on the order the variables are added in the model. The Type I or Sequential sum of squares for a given variable is the additional reduction in the sums of squares due to error due to that variable after the previous variables added to the model are accounted for.

Type III or Adjusted Sums of Squares is the unique sums of squares for that variable given all of the other variables in the model. It does not depend on the order the variables are included in the model and is the correct choice of sums of squares to look at in most situation.

Type II Sums of Squares measures the reduction in the sums of square due to error when the variable is included in the full model versus when all variables besides the current one are added to the model.

Minitab reports both the sequential and adjusted sums of squares but uses the adjusted sums of squares when determining the significance of the variable.

The Two-Way ANOVA menu in Minitab only analyzes balanced designs. If you have an unbalanced design, use the General Linear Models dialog box instead.

We will discuss the "pictures" for main effects and interactions as well. If there is time, we will use Minitab to analyze the Iron absorbtion data above.

Exercises for Lecture 30

1. –

2. –