Lecture 10

One-way ANOVA theory

This lecture will be an introduction to the theory behind ANOVA. Despite its name (Analysis of Variance), ANOVA is a test of whether multiple means are equal or not. The procedure gets its name because it breaks up the overall variance in the combined data into two parts: the variance within the groups and the variance between the groups. If the variance between the groups is significantly larger than the variance within the groups, then we conclude that the groups have significantly different means. Here is a picture:

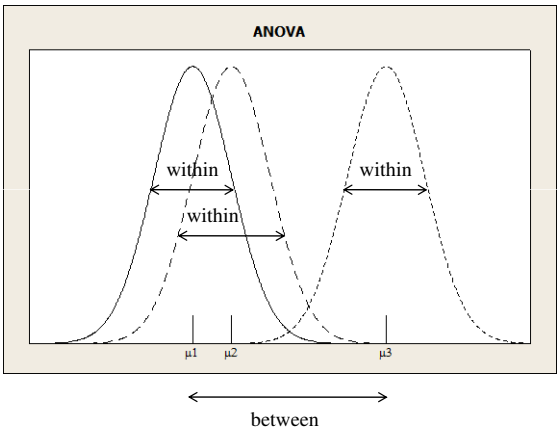

Figure 10.1: Depiction of ANOVA

How does ANOVA work? It is based on a mathematical trick (or on a miracle, or on the beauty and power of mathematics, depending on your point of view). The trick involves noticing that

$$\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

due to the fact that the sample mean is the sum of the observations divided by the sample size.

After that we need some notation (of course). It will involve a lot of subscripts and some bars and dots... Bars mean means. Dots are place holders to tell you which variables you are summing over the get the mean. Subscripts denote the population and observation within the population. In more detail:

$Y_{i,j}$ = j^{th} data point from population i.
$1 \le i \le I$ where I is the number of different groups/populations.
$1 \le j \le n_i$ where n_i is the sample size from population i
$\bar{Y}_{i,\cdot}$ = the sample mean of the data from population i, sometimes written by me as $\bar{Y}_i$
$\bar{Y}_{\cdot,\cdot}$ = the sample mean of all the data combined, sometimes written by me as $\bar{Y}$.

ANOVA breaks up the total sum of squares (the sum of the square deviations of each $Y_{i,j}$ from the overall mean $\bar{Y}_{\cdot,\cdot}$ into within population and between population sum of squares as follows:

$$
\begin{aligned}
&\text{Total Sum of Squares} \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{i,j}-\bar{Y}_{\cdot,\cdot}\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left((Y_{i,j}-\bar{Y}_{i,\cdot})+(\bar{Y}_{i,\cdot}-\bar{Y}_{\cdot,\cdot})\right)^2 \\
&= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{i,j}-\bar{Y}_{i,\cdot}\right)^2 + 2\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{i,j}-\bar{Y}_{i,\cdot}\right)\left(\bar{Y}_{i,\cdot}-\bar{Y}_{\cdot,\cdot}\right) + \sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(\bar{Y}_{i,\cdot}-\bar{Y}_{\cdot,\cdot}\right)^2 \\
&= \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{n_i}\left(Y_{i,j}-\bar{Y}_{i,\cdot}\right)^2}_{\text{Sum of Squares Within}} + 2\sum_{i=1}^{I}\left(\bar{Y}_{i,\cdot}-\bar{Y}_{\cdot,\cdot}\right)\underbrace{\sum_{j=1}^{n_i}\left(Y_{i,j}-\bar{Y}_{i,\cdot}\right)}_{\text{sums to } 0} + \underbrace{\sum_{i=1}^{I} n_i\left(\bar{Y}_{i,\cdot}-\bar{Y}_{\cdot,\cdot}\right)^2}_{\text{Sum of Squares Between}} \\
&= \text{Sums of Squares Within Populations} + \text{Sum of Squares Between Populations}
\end{aligned}
$$

The Total Sum of Squares is abbreviated simply as SS. The Sum of Squares Within Populations is also known as Sum of Squares due to Error and is often abbreviated SSE. Sum of Squares Between Populations is variously known as Sum of Squares due to the Model (SSM), Sum of Squares due to Treatment (SST), and Sum of Squares for the Groups (SSG). The fundamental equation for ANOVA then becomes SS = SSM + SSE.

The question of whether the means of the groups is different or not becomes a question of whether the Sum of Squares due to Groups/Treatment is large compared to the Sum of Squares due to error. The appropriate test statistic is an F-test as follows:

$$\begin{aligned} F &= \frac{\text{SSM/d.f. Model}}{\text{SSE/d.f. Error}} \\ &= \frac{\text{SSM}/(I-1)}{\text{SSE}/(n-I)} \text{ where } n = n_1 + n_2 + \cdots + n_I \\ &= \frac{\text{MS Model}}{\text{MS Error}} \text{ where the M stands for mean and MS stands for Mean Sum of Squares} \end{aligned}$$

The resulting F-statistic has $I-1$ numerator degrees of freedom and $n-I$ denominator degrees of freedom.

Assumptions for ANOVA: the data a normal, the variances within each population are equal, every observation is independent of every other observation. With larger sample sizes, normality is not so critical. With a balanced design (equal samples sizes in each group), equal variances are not so critical. For an unbalanced design, equal variances are crucial. Independence is crucial but it should result from the experimental design.

ANOVA from a modeling viewpoint

A different way of thinking about ANOVA is by thinking about models for the data. ANOVA corresponds to the comparisons of two models for the data - a Full model or Separate Means model where every population gets its own mean and a Reduced model or Equal Means model where there is one common mean for each population. These models can be represented visually/mathematically as

Full Model (separate means):	μ_1	μ_2	$\cdots$	μ_I
Reduced Model (equal means):	μ	μ	$\cdots$	μ

In this formulation, we identify the Sum of Squares due to the Model (Treatment, Group) with the Extra Sum of Squares that the reduced model has in its error as compared to what the full model has in its error:

$$\begin{aligned} \text{SS Model} &= \text{Extra SS} \\ &= \text{Residual sum of squares from the reduced model} \\ &\quad - \text{Residual sum of squares from the full model} \end{aligned}$$

The model degrees of freedom $(I-1)$ is also the Extra degrees of freedom that the reduced model has as compared to the full model:

$$\text{Extra d.f.} = (n-1) - (n_1 - 1 + n_2 - 1 + \cdots + n_I - 1) = I - 1$$

and is also simply the number of parameters in the full model minus the number of parameters in the reduced model.

Thinking in terms of models can be useful. The above generalizes, for instance to a model $(\mu_2 = \mu_3 = \cdots = \mu_I)$, where all the means are equal except that the first mean can be different.

For ANOVA, you should check normality and apply a log transform to all the data to correct skewness when appropriate. For a balanced design, ANOVA is robust to violations in the assumption of equal variances. If the design is unbalanced and the variances are unequal, you should use not use ANOVA.

Exercises for Lecture 10

1. –

2. –