Intro to ANOVA

DS705

Analysis of Variance

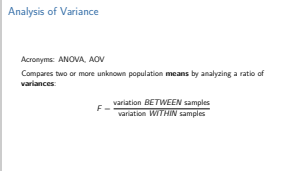
Acronyms: ANOVA, AOV

Compares two or more unknown population **means** by analyzing a ratio of **variances**:

$$F = \frac{\text{variation } \textit{BETWEEN} \text{ samples}}{\text{variation } \textit{WITHIN} \text{ samples}}$$

Intro to ANOVA

Analysis of Variance



ANOVA is used to compare two or more unknown population means. That is, when there is a quantitative response variable and a categorical explanatory variable (called a factor).

When there is only one factor, it is called a One-Way ANOVA.

The ANOVA Table

Source	df	SS	MS	F	P-value
Treatment	$t-1$	SST	$MST = \frac{SST}{t-1}$	$F_0 = \frac{MST}{MSE}$	$P(F_{t-1,n_T-t} > F_0)$
Error	$n_T - t$	SSE	$MSE = \frac{SSE}{n_T-t}$		
Total	$n_T - 1$	TSS			

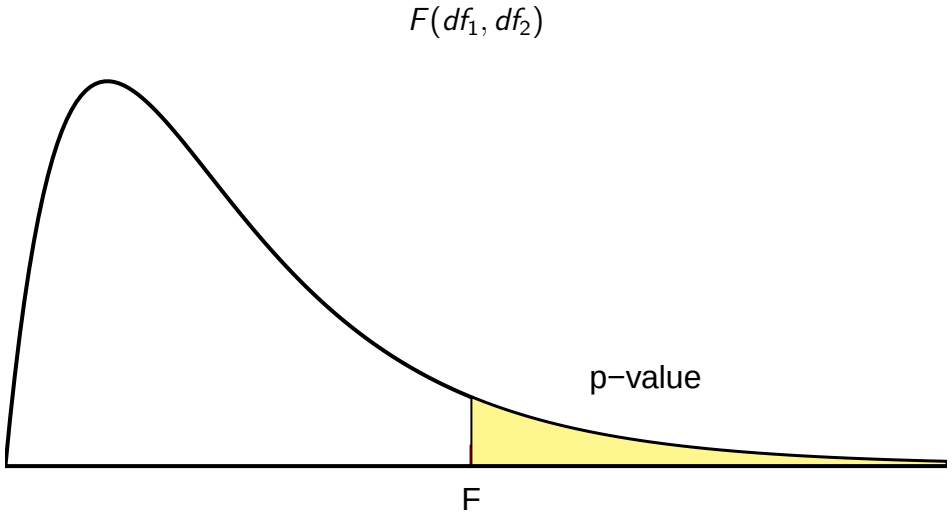
The ANOVA Table

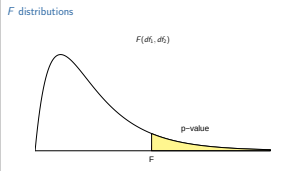
Source	df	SS	MS	F	P-value
Treatment	$t-1$	SST	$MST = \frac{SST}{t-1}$	$F_0 = \frac{MST}{MSE}$	$P(F_{t-1,n_T-t} > F_0)$
Error	$n_T - t$	SSE	$MSE = \frac{SSE}{n_T-t}$		
Total	$n_T - 1$	TSS			

The ANOVA table is a great way to organize the information and see how the test statistic is computed. The first column is labeled "Source," and what that really means is "Source of variation in the response variable." The values of the response variable are not all identical, and we can mathematically attribute that variation to different causes or sources.

Some of the variation is due to random variation that is a part of random sampling. This type of variation is typically labeled as "Error" or "Residual" (as in R) or "Within" (as in Ott's book). The other potential source of variation in the response variable is the fact that they come from populations with different means. This source may be labeled as Treatment, Factor, Between, or by the name of the categorical factor that defines the populations.

F distributions





F distributions are continuous, right-skewed distributions with non-negative values identified by two parameters called numerator degrees of freedom (labeled as dfN or df1) and denominator degrees of freedom (labeled as dfD or df2). The p-value in analysis of variance is the probability of seeing a value in the associated F distribution that is at least as big as the observed test statistic F.

Large values of F provide more evidence against the null hypothesis. Notice, the larger F is, the smaller the p-value will be.

The Model for One-Way ANOVA

$$y_{i,j} = \mu + \tau_i + \epsilon_{i,j}$$

where $i = 1, \ldots, t$ and $j = 1, \ldots, n_i$

and $\epsilon_{i,j}$ are independent and $N(0, \sigma_\epsilon)$

The Model for One-Way ANOVA

$$y_{i,j} = \mu + \tau_i + \epsilon_{i,j}$$

where $i = 1, \ldots, t$ and $j = 1, \ldots, n_i$

and $\epsilon_{i,j}$ are independent and $N(0, \sigma_\epsilon)$

Y sub i,j, represents the value of the response variable for the jth individual in the ith sample. There are a total of t independent samples, corresponding to the t populations and t poplation means we wish to compare. The ith sample contains a total of n sub i onservations.

Mu is the overall population mean if all populations were combined into one big population. Sometimes this is called the grand mean.

Tau is the effect of the ith population. If all tau's are zero, then all population means are equal. A positive value of tau indicate that that particular population mean is larger than the overall mean and a negative value would show that a particular mean is less than the overall mean.