Analysis of Variance

October 16, 2019

ANOVA and the F-test

- Question: is the variability in the sample means so large that it seems unlikely to be from chance alone?
- We call this variability the **mean square between groups** (MSG) or **mean square for treatment** (MST).

Mean Square Between Groups

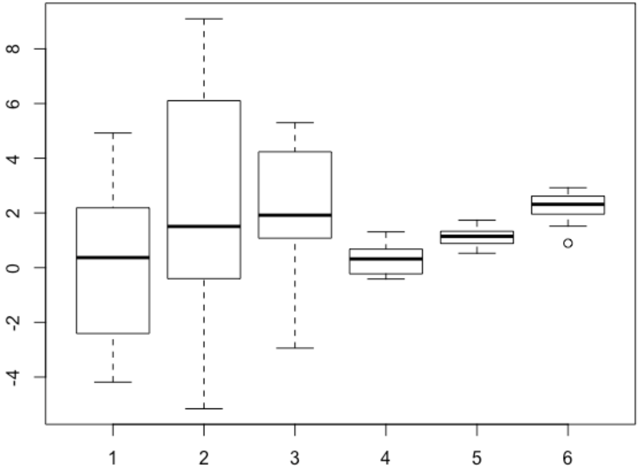
- This acts as a measure of variability for the k group means.
- It has degrees of freedom $df_G = k - 1$.
- If H_0 is true, we expect this variability to be small.

Mean Square Between Groups

$$\begin{aligned} MSG &= \frac{1}{df_G} SSG \\ &= \frac{1}{k-1} \sum_{i=1}^{k} (\bar{x}_i - \bar{x})^2 \end{aligned}$$

where SSG is the sum of squares between groups.

Mean Square Between Groups



...but MSG isn't very useful on its own.

Mean Square Error

- We need an idea of how much variability would be expected (or normal) if H_0 were true.
- This is done using a pooled variance estimate, called the **mean square error** (MSE).
- This is a measure of variability within groups.
- MSE has degrees of freedom $df_E = n - k$

Mean Square Error

$$
\begin{aligned}
MSE &= \frac{1}{df_E} SSE \\
&= \frac{1}{n-k} \sum_{i=1}^{k} (n_i - 1) s_i^2
\end{aligned}
$$

where SSE is the sum of squares for error and s_i is the standard deviation for the observations in group i.

Sum of Squares Total

It's also useful to think of a **sum of squares total** (SST)

$$SST = SSG + SSE$$

and total degrees of freedom

$$\begin{aligned} df_T &= df_G + df_E \\ &= k - 1 + n - k \\ &= n - 1 \end{aligned}$$

Mean Square Total

If we were to find the mean square total,

$$\begin{aligned} MST &= \frac{1}{df_T} SST \\ &= \frac{1}{n-1}(SSG + SST) \\ &= \frac{1}{n-1}\sum_{j=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

we would get the variance across all observations!

ANOVA

The ANOVA breaks the variance down into

- within group (random) variability (MSE).
- between group (means) variability (MSG).

ANOVA

We want to know how much variability is due to differences in groups *relative to the within groups variability*.

So our test statistic is

$$F = \frac{MSG}{MSE}$$

Example

For our baseball example,

	OF	IF	C
Sample size (n_i)	160	205	64
Sample mean ($\bar{x}_i$)	0.320	0.318	0.302
Sample sd (s_i)	0.043	0.038	0.038

$MSG = 0.00803$ and $MSE = 0.00158$.

Find the degrees of freedom and the F statistic.

The F Test

With our F distribution comes the **F-test**. Using the F-distribution, we calculate

- $F_\alpha(df_1, df_2)$ critical values.
- p-values

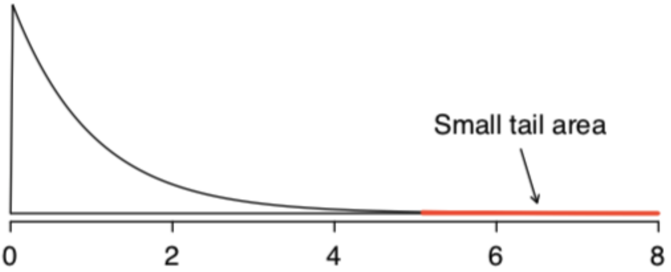
The F Test

If the between-group variability is high relative to the within group variability,

- $MSG > MSE$
- F will be large.
- Large values of F represent stronger evidence against the null.

The F Test

This is the F(2, 426) distribution from our baseball example.



- F-test p-values will always be from the upper tail area.
- We no longer have one- or two-sided tests to worry about.
- The critical value is $F_{0.05}(2, 426) = 3.0169$.

Example

What can we conclude about the baseball field positions?

Recall $F_{0.05}(2, 426) = 3.0169$.

Reading an ANOVA Table

- Typically we will run ANOVA using software.
- Fortunately there is a standard output for this analysis.

Let's take some time to write out the ANOVA table.

Reading an ANOVA Table from Software

This is the ANOVA from `R` for the MLB example.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
position	2	0.0161	0.0080	5.0766	0.0066
Residuals	426	0.6740	0.0016		

What can we conclude based on the table?

Example

Suppose we have 10 data points from each of 5 groups of interest.

Source	df	SS	MS	F
Group	______	______	3	______
Error	______	______	______	
Total	______	20		

Fill in the missing information from the ANOVA table.

Graphical Diagnostics for ANOVA

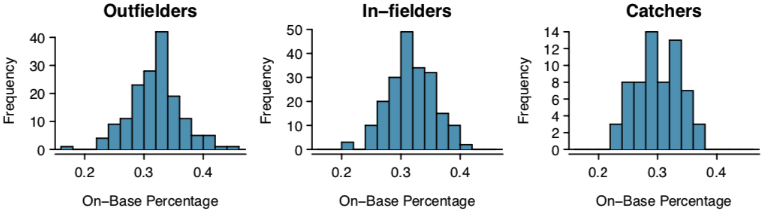
There are three conditions for ANOVA:

1. Independence
2. Approximate normality
3. Constant variance

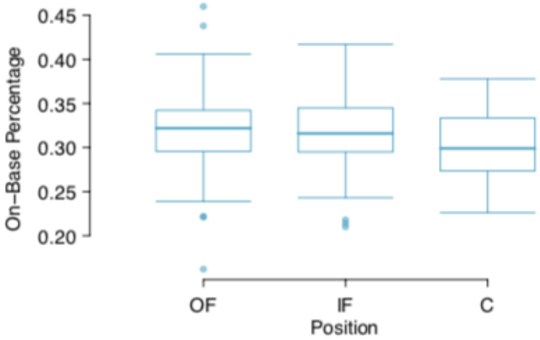
ANOVA Diagnostics: Independence

- It is reasonable to assume independence if the data are a simple random sample.
- If the data are not a random sample, consider carefully.
  - In the MLB example, no clear reason why a player's batting stats would impact another player's batting stats.

ANOVA Diagnostics: Normality



- Normality is especially important for small samples.
- For large samples, ANOVA is *robust to* deviations from normality.



- We can check this visually or by examining the standard deviations for each group.
- Constant variance is especially important when the sample sizes differ between groups.