Announcements

Data Analysis due tonight at 11:59 pm

Paid tutoring opportunity: student looking for a tutor for ENGR 390: Engineering Economy. Send me an email if you are interested.

Week 8

Analysis of Variance (ANOVA)

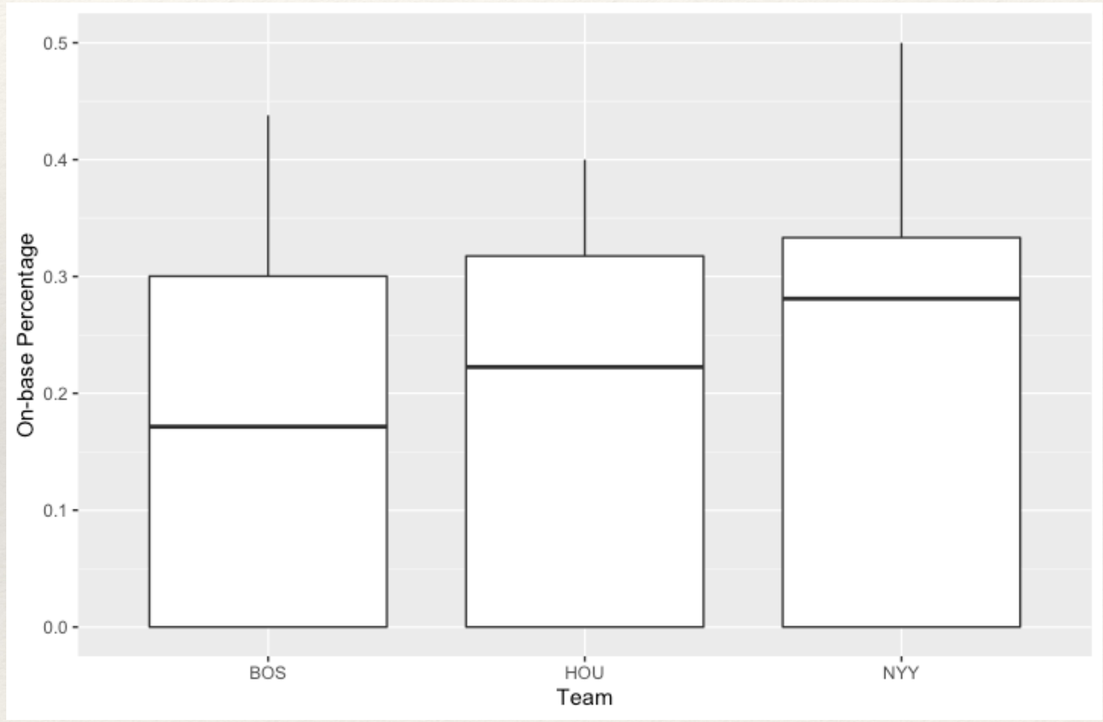
ST 314
Introduction to Statistics for Engineers

Oregon State University

Comparing More than Two Means

In 2018, the Boston Red Sox, the Houston Astros, and the New York Yankees were ranked as the top three teams in the MLB.

The side-by-side boxplot displays the On-base Percentages for 30 randomly selected players from each of the 2018 top three teams.



According to the three graph, is there evidence that average On-base Percentages differ between the ~~two~~ three teams?

Comparing Means with Mean Squares

- To compare more than two group means we compare the average between group variability to the average within group variability.
- Mean squares represents the average variation **between** groups and average variation **within** groups.

Comparing Means with Mean Squares

- Notation
  - $k =$ # of groups
  - $n =$ overall sample size (all groups combined)
  - $\bar{x} =$ overall mean (all groups combined)
  - $n_i =$ sample size for the ith group
  - $\bar{x}_i =$ sample mean for the ith group
  - $s_i =$ sample standard deviation for the ith group

Comparing Means with Mean Squares

Mean Square Between Groups (MSG)	Mean Square Error (MSE)
Average variability between groups	Avgerage variability within groups
$MSG = \frac{1}{k-1}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	$MSE = \frac{1}{n-k}\sum_{i=1}^{k}(n_i - 1)s_i^2$
$df_G = k-1$	$df_E = n-k$

k = # of groups

n = total sample size

$\bar{x}$ = overall mean

n_i = sample size of group i

$\bar{x}_i$ = mean of group i

s_i = standard deviation of group i

Single Factor ANOVA F test

- When to use: want to test for difference between more than 2 group means
- Conditions required for inference:
  - independent observations between + within groups (random sampling)
  - sufficiently large sample sizes for each group
  - constant variance across groups
- Null & Alternative hypotheses:

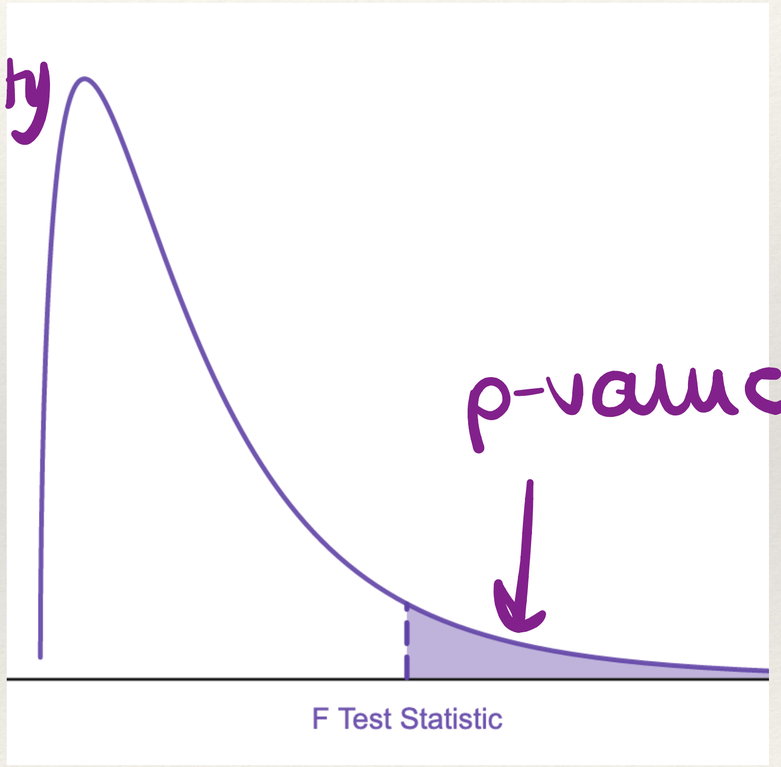
$H_0: \mu_1 = \mu_2 = \dots = \mu_K$

H_A: At least one group's mean differs from the others

Single Factor ANOVA F test

- Test Statistic: Ratio of the average between group variability to the average within group variability

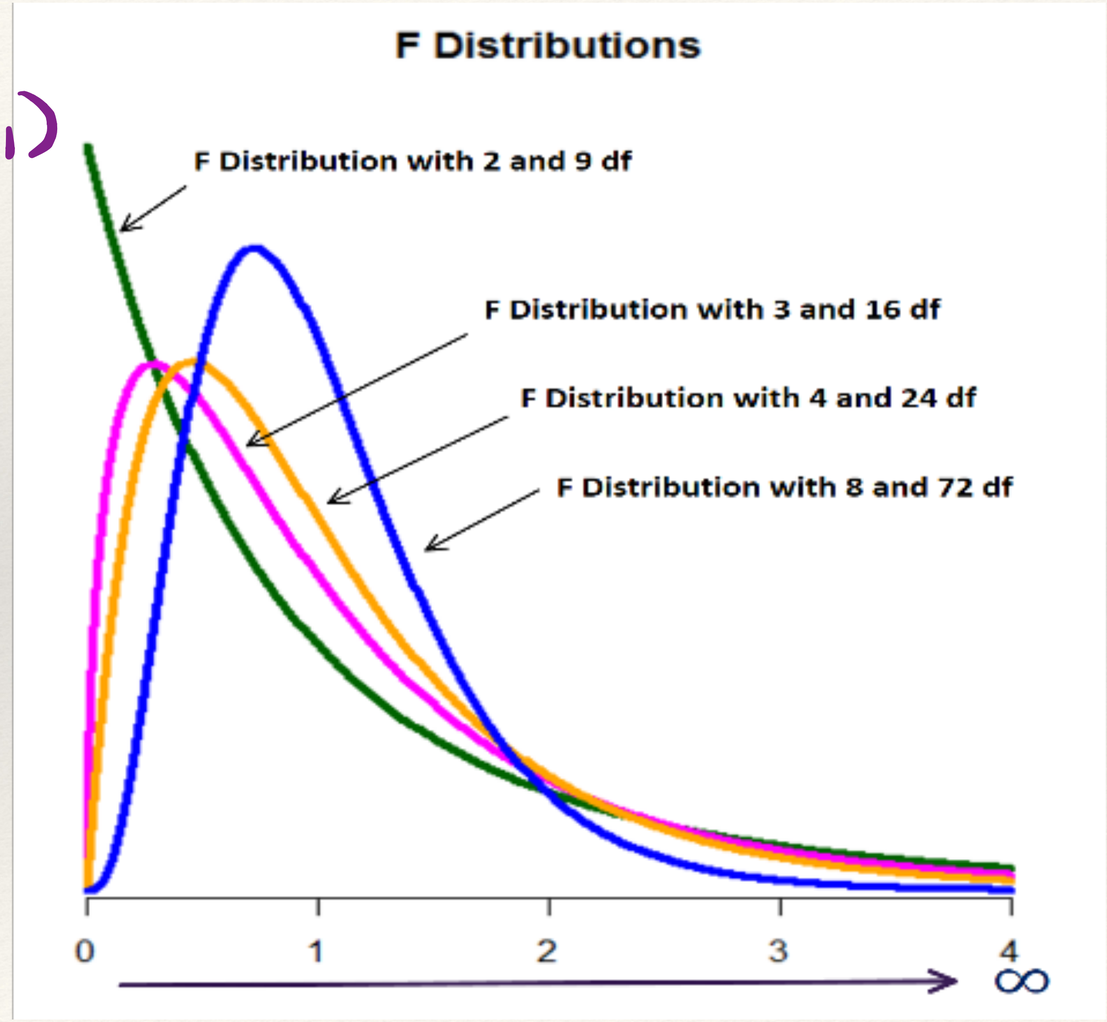
$$F = \frac{MSG}{MSE} \sim F \text{ distribution with } k-1 \text{ and } n-k \text{ degrees of freedom}$$



p-value in a ANOVA F test is ALWAYS the area under the curve to the right of the F statistic

F Distributions

- Positively skewed distribution ranging from 0 to ∞
- The shape is defined by numerator df (v_1) and denominator df (v_2)
- Denoted: F_{v_1, v_2}
- For a single factor ANOVA F test, the p-value is always the area under the distribution curve to the right of the F statistic in the distribution:

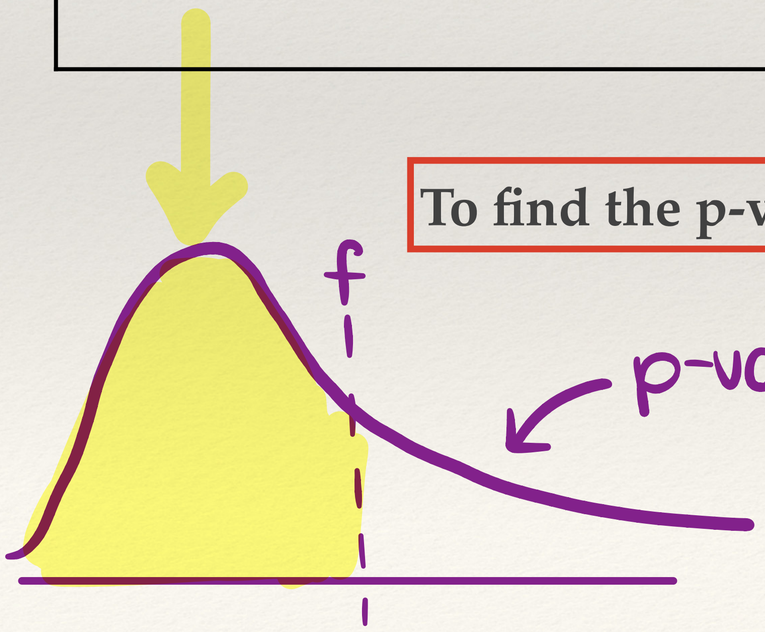


Finding *F* p-values Using `pf( )`

For an ANOVA F test, the p-value is the area under the curve of an *F* distribution with v_1 and v_2 degrees of freedom, to the right of the F statistic.

Function	Function Values	What does it do?
`pf(f,v1,v2)`	`f` = F statistic `v1` = numerator degrees of freedom `v2` = denominator degrees of freedom	This is the cumulative distribution function for an F distribution. $P(X \leq x)$

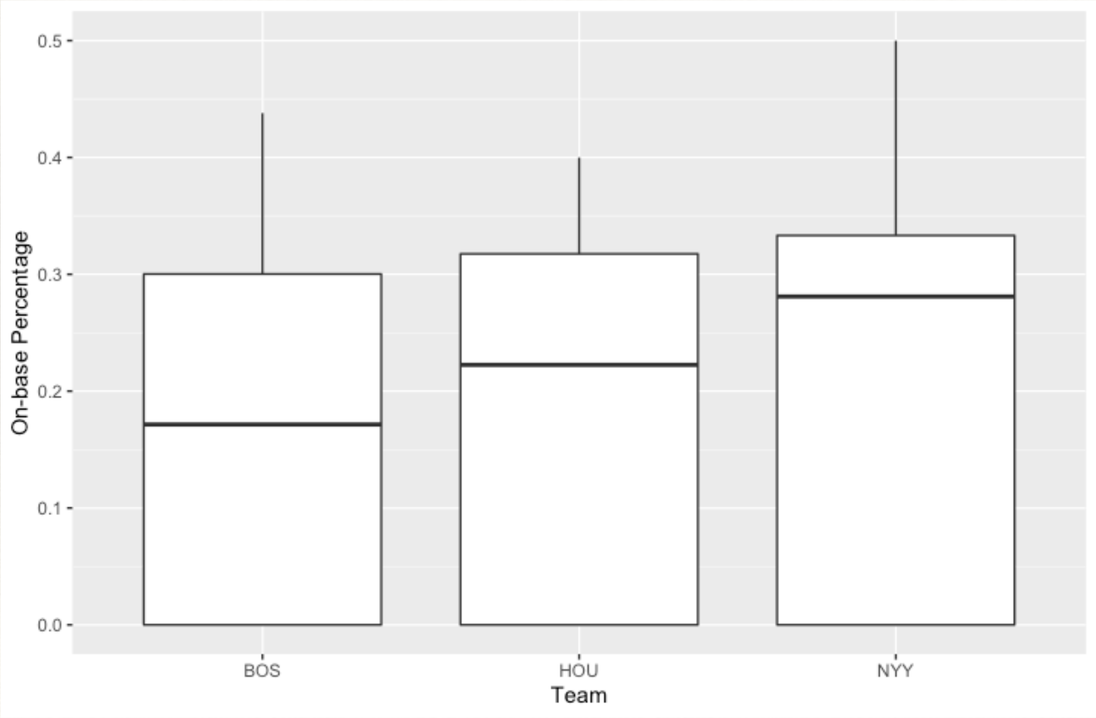
To find the p-value, you need to subtract the above function from 1!



ANOVA *F* test Example

R = red sox
A = astros
y = yankees

Set up the hypotheses to test whether the average On-base Percentage differed significantly between the Red Sox, Astros, and Yankees in 2018.



$H_0: \mu_R = \mu_A = \mu_y$

H_A: At least one team's average on-base percentage differs from the other teams

ANOVA *F* test Example

Performing an ANOVA *F* test by hand is tedious and unnecessary. R can do all the calculations for us in a few easy steps. Below is the R output for this hypothesis test.

	DF	Sum Sq	Mean Sq	*F*	Pr(>F)
Team	**2**	**0.0223**	**0.01113**	**0.405**	**0.668**
Residuals	**87**	**2.3919**	**0.02749**		

Using the R output, determine the df_G, df_E, *MSG*, *MSE*, the *F* statistic, and the p-value.

$df_G = 2$ MSG = 0.01113 F = 0.405 p-value = 0.668

$df_E = 87$ MSE = 0.02749

ANOVA *F* test Example

Performing an ANOVA *F* test by hand is tedious and unnecessary. R can do all the calculations for us in a few easy steps. Below is the R output for this hypothesis test.

	DF	Sum Sq	Mean Sq	*F*	Pr(>F)
Team	**2**	**0.0223**	**0.01113**	**0.405**	**0.668**
Residuals	**87**	**2.3919**	**0.02749**		

Using the output, write a conclusion for the hypothesis test.

There is no evidence to suggest that the average on-base percentages are different between the top three MLB teams in 2018. For any reasonable (or commonly used) significance level, we fail to reject the null hypothesis.

It is possible that we've made a Type II error.

Using R for ANOVA

The following code templates can be used to perform an ANOVA in R.

Function	Function Values	What does it do?
`mod <- aov(response ~ treatment)`	`mod` saves the ANOVA model in an object called mod `response` = vector containing the response variable values `treatment` = vector that defines the groups for each recorded response	Calculates the values necessary to perform an ANOVA F test
`summary(mod)`	`mod` call the saved model stored above	Prints the ANOVA table

Dataset: Baseball

```
mod<-aov(Baseball$onbasc ~ Baseball$team)
summary(mod)
```