ONLINE MASTERS IN **DATA SCIENCE**

DSC 215 - PROBABILITY AND STATISTICS FOR DATA SCIENCE

COMPARING MANY MEANS WITH ANOVA

UC San Diego

COMPUTER SCIENCE & ENGINEERING

HALICIOĞLU DATA SCIENCE INSTITUTE

What is ANOVA?

- **Previously:** Constructed confidence intervals, and conducted hypothesis tests for:
  - A single mean
  - The difference of two means
- **Now:** We will consider comparing means *across many groups.*
- **Question:** Why not just do pairwise comparisons of all the means?
  - If you have *k* groups, you'll need k(k-1)/2 comparisons, so if k is big, there will be a high chance that you'll find some difference just by luck.
- This is where ANOVA comes in.

What is ANOVA?

- **Analysis of variance (ANOVA)** is a method that uses a *single hypothesis test* to check whether the means across groups are equal.

- Here, our hypotheses are:

  - H_0 : The mean is the same across all groups, i.e., $\mu_1 = \mu_2 = \ldots = \mu_k$.

  - H_1 : At least one mean is different, i.e., $\mu_i \neq \mu_j$, for some $i \neq j$.

- To test our hypotheses, we'll introduce a new test statistic: F-statistic

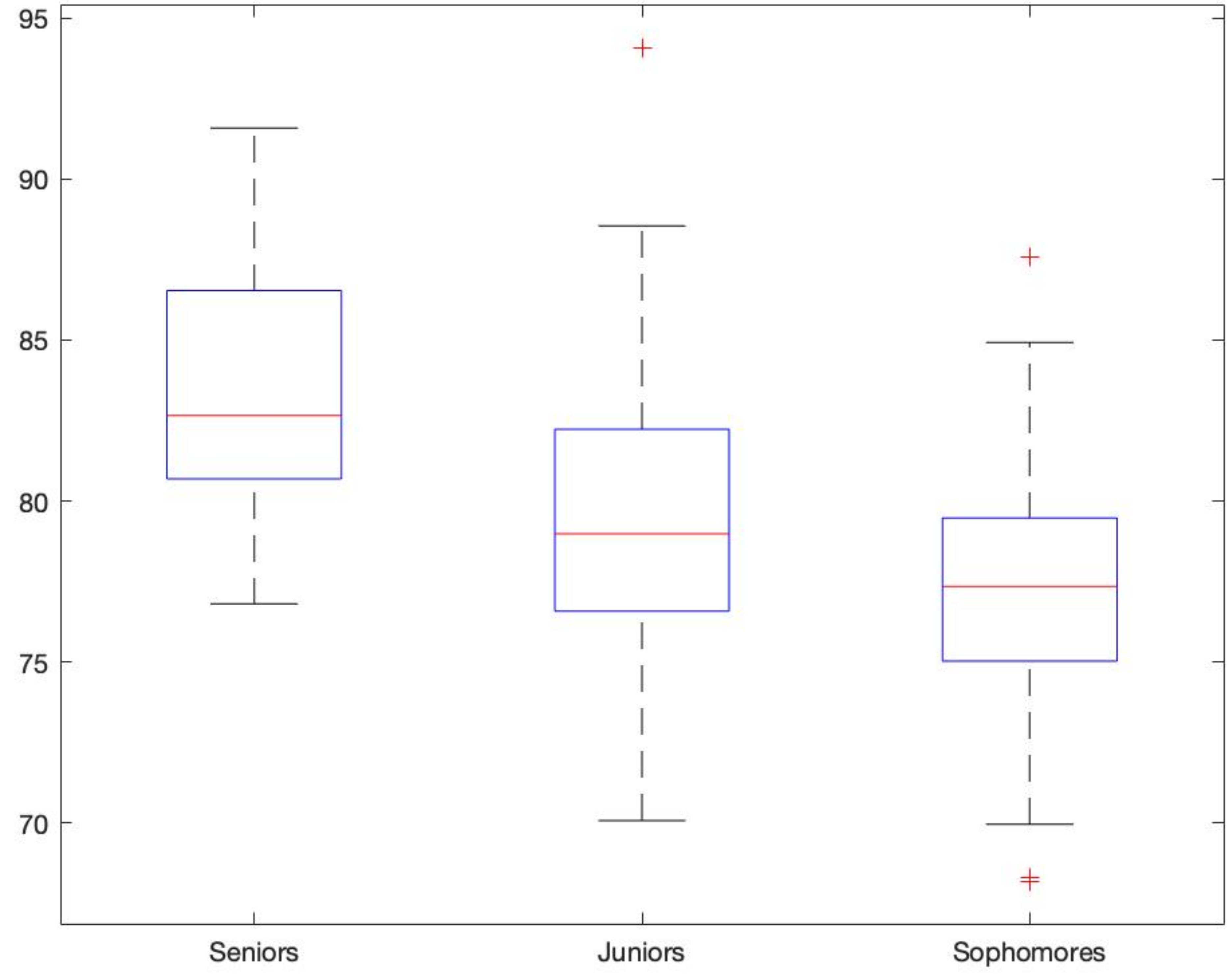
Analysis of Variance (ANOVA): Conditions

- As usual, we'll have to check some conditions before applying our test:
  - **Independence:** observations are independent within and across groups
  - **Normality:** the data within each group is nearly normal
  - **Variability:** the variability across groups is comparable
- Why a variability condition?
  - Strong evidence favoring the alternative hypothesis in ANOVA is described by unusually large differences among the group means.
  - Assessing the variability of the group means relative to the variability among individual observations within each group is key to ANOVA's success.

ANOVA: an Example Scenario

- **Example:** A certain course is taken by Sophomores, Juniors, and Seniors. The students all take the same exam, and their grades have the following summary statistics, and box-plot.

	n	Sample mean	Standard Deviation
Seniors	82	83.31	4.43
Juniors	66	78.99	4.39
Sophomores	115	77.07	3.51



- The conditions for ANOVA are satisfied. (Why?)

ANOVA: the Basic (Rough) Idea

- **Idea:** If the populations means are different, the variance between groups must be larger than the variance within the groups.

- So, if the ratio,

$$\frac{\text{“variance between groups”}}{\text{“variance within groups”}}$$

is large, the means are different.

ANOVA: the Test Statistic

- Denote $\bar{x}_i$ as the mean of group i, $\bar{x}$ as the mean across all groups, and n as the sum of n_i.
- Then calculate the numerator, called Mean Square between Groups (MSG):

$$MSG = \frac{1}{k-1} \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

- And the denominator, often called the Mean Square Error (MSE):

$$MSE = \frac{1}{n-k} \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 - \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 \right)$$

- Finally

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

ANOVA: Back to Our Example

- In our example, we can calculate $\bar{x} = 79.5$, so

$$MSG = \frac{1}{3-1}\sum_{i=1}^{3} n_i(\bar{x}_i - \bar{x})^2$$
$$= \frac{1}{2}\left(82(83.31 - 79.5)^2 + 66(78.99 - 79.5)^2 + 115(77.07 - 79.5)^2\right)$$
$$= 945.49$$
$$MSE = \frac{1}{n-k}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 - \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2\right)$$
$$= \ldots$$
$$= \frac{1}{263-3}\left(6138.34 - 1891.81\right) = 16.336$$

ANOVA: Back to Our Example

- Continuing, we can calculate $F = \dfrac{MSG}{MSE} \approx 57.877.$

- But how do decide whether to reject the null?

- Like in our previous hypothesis tests, we have to calculate

$$\begin{aligned} p - value &= \mathbb{P}(\text{data at least as extreme as our statistic} \,|\, H_0) \\ &= \mathbb{P}(\geq F \,|\, H_0) \end{aligned}$$

- But what is the right distribution for our **F-test**? The F-distribution.

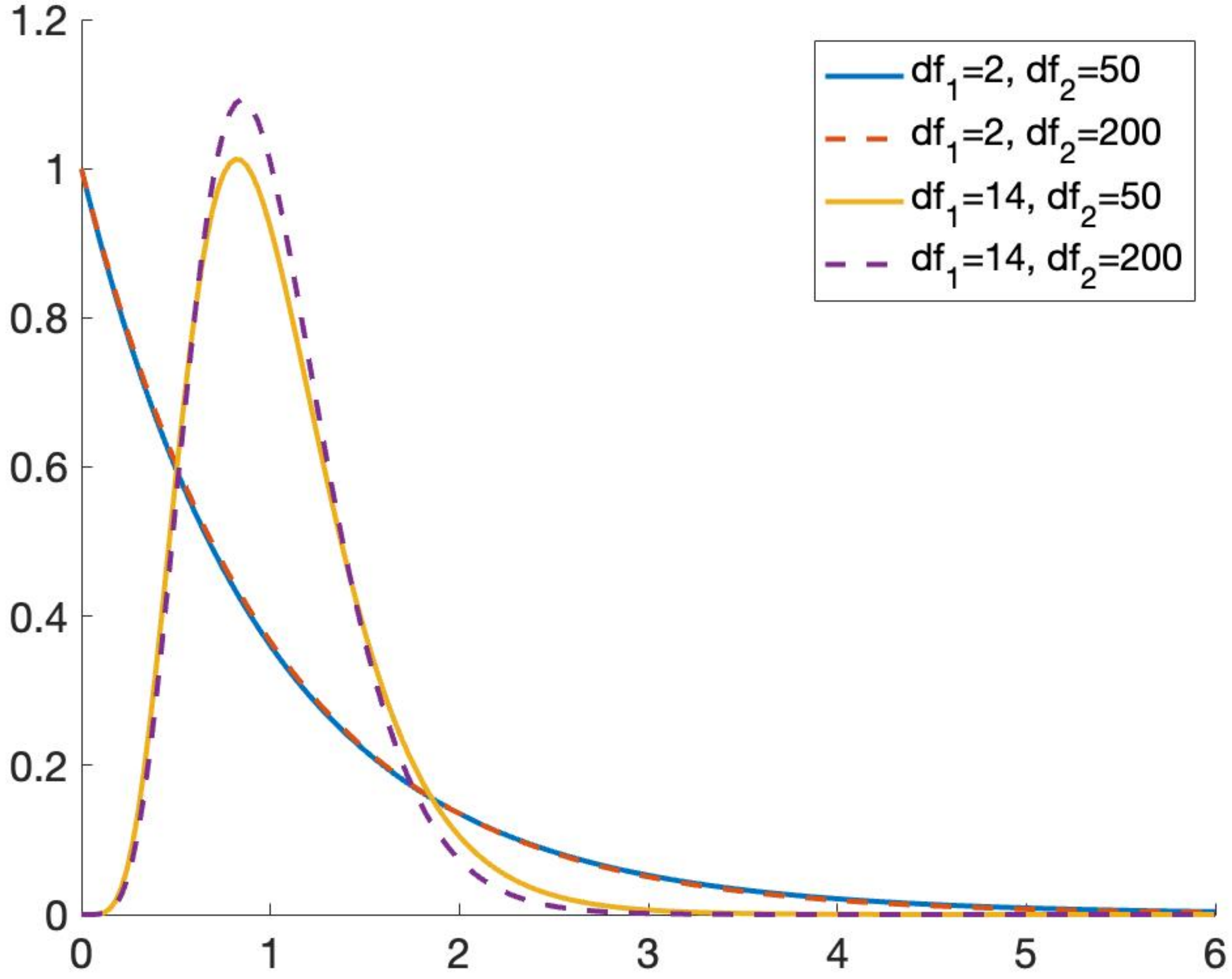
- As usual, we calculate the probability using tables or software.

ANOVA: the F-Distribution

- **The F-distribution:** parametrized by two parameters

  - $df_1 = k - 1$
    The degrees of freedom for MSG

  - $df_2 = n - k$
    The degrees of freedom for MSE



ANOVA: Back to Our Example (One Last Time... For Now)

- We calculated
  - $F = \dfrac{MSG}{MSE} \approx 12.18.$
  - $df_1 = k - 1 = 2$
  - $df_2 = n - k = 260$
- Using tables, or software we obtain
  - p-value $< 1 \times 10^{-6}$
- So, we reject the null at the $\alpha = 0.05$ level (or any reasonable level).

Multiple Comparisons and Controlling Type-1 Errors

- Having rejected the null in the previous example, we may wonder:
  - Which of the groups have different means?
- We could compare the groups pairwise, using t-tests:
  - Sophomores to Juniors
  - Juniors to Seniors
  - Seniors to Sophomores
- **The issue:** Because we are doing 3 tests (in this example) instead of 1, the odds of spuriously rejecting the null (Type I error) in at least one test is higher.
- **The fix:** Use a modified significance level, and a pooled estimate of the standard deviation across groups (where we calculate the pooled estimate using $df_2 = n - k$).

Multiple Comparisons and the Bonferroni Correction

- Testing many pairs of groups is called **multiple comparisons**.

- **The Bonferroni correction** uses a more conservative significance level is more for these tests:

$$\alpha^* = \frac{\alpha}{K} \quad \text{where} \quad K = \frac{k(k-1)}{2}.$$

- Why does this make sense? Probability! The p-values for the i-th comparison satisfy

$$\mathbb{P}(\cup_{i=1}^{K} \left(p_i \leq \alpha/K\right)) \leq \sum_{i=1}^{K} \mathbb{P}\left(p_i \leq \alpha/K\right) \leq \alpha.$$

Multiple Comparisons: Final Thoughts

- It is possible to reject the null hypothesis using ANOVA and then not identify differences in the pairwise comparisons.
- This *does not* invalidate the ANOVA conclusion.
- We interpret this to mean that we have not been able to successfully identify which specific groups differ in their means.