Statistical Analysis in Fin Mkts

MSF 502

Li Cai

CHAPTER 13

Analysis of Variance

Chapter 13 Learning Objectives (LOs)

LO 13.1: Provide a conceptual overview of ANOVA.

LO 13.2: Conduct and evaluate hypothesis tests based on one-way ANOVA.

LO 13.3: Use confidence intervals and Tukey's HSD method in order to determine which means differ.

LO 13.4: Conduct and evaluate hypothesis tests based on two-way ANOVA with no interaction.

LO 13.5: Conduct and evaluate hypothesis tests based on two-way ANOVA with interaction.

How Much Does Using Public Transportation

Boston	New York	San Francisco	Chicago
$\bar{x}_1 = \$12{,}622$	$\bar{x}_2 = \$12{,}585$	$\bar{x}_3 = \$11{,}720$	$\bar{x}_4 = \$10{,}730$
$s_1 = \$87.79$	$s_2 = \$80.40$	$s_3 = \$83.96$	$s_4 = \$90.62$
$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n_4 = 5$

- Environmental and economic concerns have led to an upswing in the use of public transportation.
- Research analyst Sean Cox looked at study results from a *Boston Globe* article that claimed commuters there topped the nation in cost savings from public transportation.
- He wants to know if the average savings significantly differ among these cities.

13.1 One-Way ANOVA

LO 13.1 Provide a conceptual overview of ANOVA.

- Analysis of Variance (ANOVA) is used to determine if there are differences among three or more populations.

- One-way ANOVA compares population means based on one categorical variable.

- We utilize a completely randomized design, comparing sample means computed for each treatment to test whether the population means differ.

LO

ANOVA Assumptions

The assumptions are extensions of those we used when comparing just two populations:

1. The populations are normally distributed.

2. The population standard deviations are unknown but assumed equal.

3. Samples are selected independently from each population.

Here we compare a total of *c* populations, rather than just two.

The Hypothesis Test

LO 13.2 Conduct and evaluate hypothesis tests based on one-way

- The competing hypotheses for the one-way ANOVA:

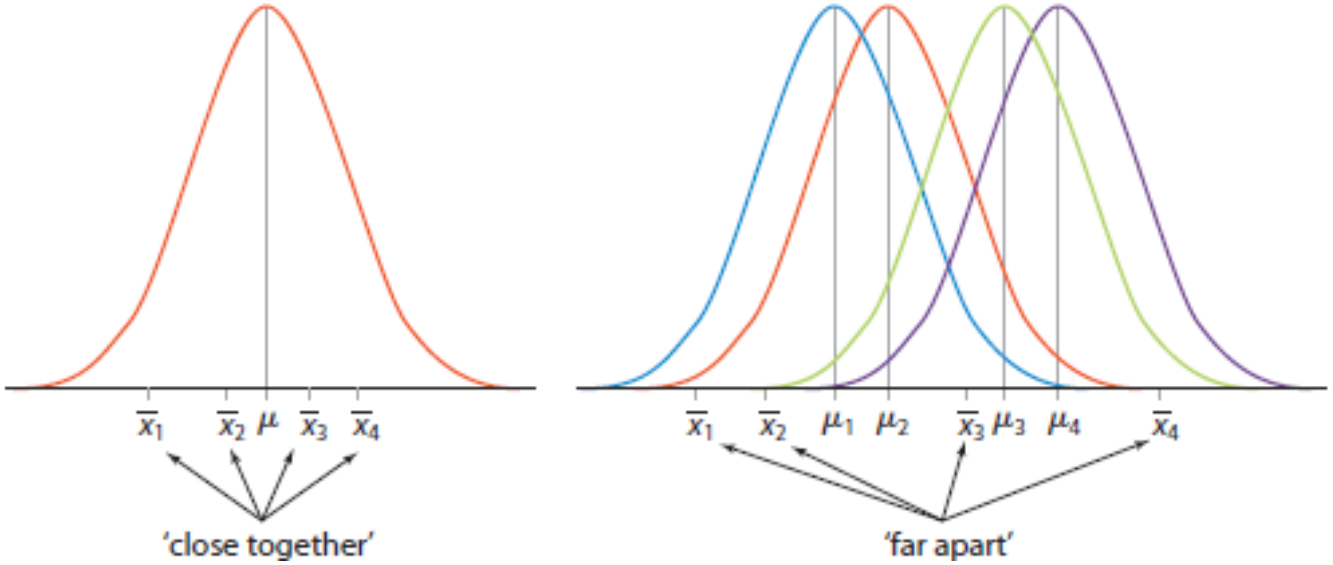
H_0: $\mu_1 = \mu_2 = \ldots = \mu_c$

H_A: Not all population means are equal

LO 13.2

The ANOVA Concept

- The competing hypotheses are displayed graphically belo



- The left graph depicts the null hypothesis, where all sample means are drawn from the same distribution.
- On the right, the distributions, and population means, differ.

LO

Methodology

- We first compute the amount of variability *between* the sample means.
- Then we measure how much variability there is *within* each sample.
- A ratio of the first quantity to the second forms our test statistic which follows the $F_{(df_1, df_2)}$ distribution.

LO 13.2

Between-Treatments Estimate

- To measure between-treatments variability, we compare the sample means to the overall mean, sometimes called the grand mean.

- To compute the grand mean, simply average all the values from the dataset:

$$\overline{\overline{x}} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n_i} x_{ij}}{n_T}$$

LO

Between-Treatments Estimate

- First, we compute the sum of squares due to treatments, *SSTR*:

$$SSTR = \sum_{i=1}^{c} n_i(\bar{x}_i - \bar{\bar{x}})^2$$

- Then, we compute the mean square for treatments, *MSTR*:

$$MSTR = SSTR/(c - 1).$$

- *MSTR* is our measure of variability between samples.

LO

Transportation Example

Summary statistics for our example include:

Boston	New York	San Francisco	Chicago
$\bar{x}_1 = \$12{,}622$	$\bar{x}_2 = \$12{,}585$	$\bar{x}_3 = \$11{,}720$	$\bar{x}_4 = \$10{,}730$
$s_1 = \$87.79$	$s_2 = \$80.40$	$s_3 = \$83.96$	$s_4 = \$90.62$
$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n_4 = 5$

The grand mean: $\bar{\bar{x}} = 287{,}760/24 = 11{,}990$

$$SSTR = 5(12622-11990)^2+8(12585-11990)^2 + 6(11720-11990)^2+8(10730-11990)^2 = 13{,}204{,}720$$

$$MSTR = 13{,}204{,}720/(4-1) = 4{,}401{,}573.$$

LO

Within-Treatments Estimate

- The denominator of our test statistic measures the within-sample variability. It really is an extension of the pooled-sample variance that we used in a two-sample comparison.

- First, we compute the error sum of squares, *SSE*:

$$SSE = \sum_{i=1}^{c} (n_i - 1)s_i^2$$

- Then, we compute the mean squared error, *MSE*:

$$MSE = SSE/(n_T - c)$$

LO

Transportation Example - continued

Boston	New York	San Francisco	Chicago
$\bar{x}_1 = \$12{,}622$	$\bar{x}_2 = \$12{,}585$	$\bar{x}_3 = \$11{,}720$	$\bar{x}_4 = \$10{,}730$
$s_1 = \$87.79$	$s_2 = \$80.40$	$s_3 = \$83.96$	$s_4 = \$90.62$
$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n_4 = 5$

- For the transportation example, we compute:

$$SSE = (5 - 1)(87.79)^2 + (8 - 1)(80.40)^2 + (6 - 1)(83.96)^2 + (5 - 1)(90.62)^2 = 144{,}180$$

- The mean squared error is:

$$MSE = SSE/(n_T - c) = 144{,}180/(24 - 4) = 7209$$

LO

The *F* Test

- We test whether average cost savings from using public transportation differ between the four cities:

  H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

  H_A: Not all population means are equal

- The value of the test statistic is calculated as

$$F_{(df_1, df_2)} = MSTR / MSE,$$

  where $df_1 = (c\text{-}1)$ and $df_2 = (n_T - c)$.

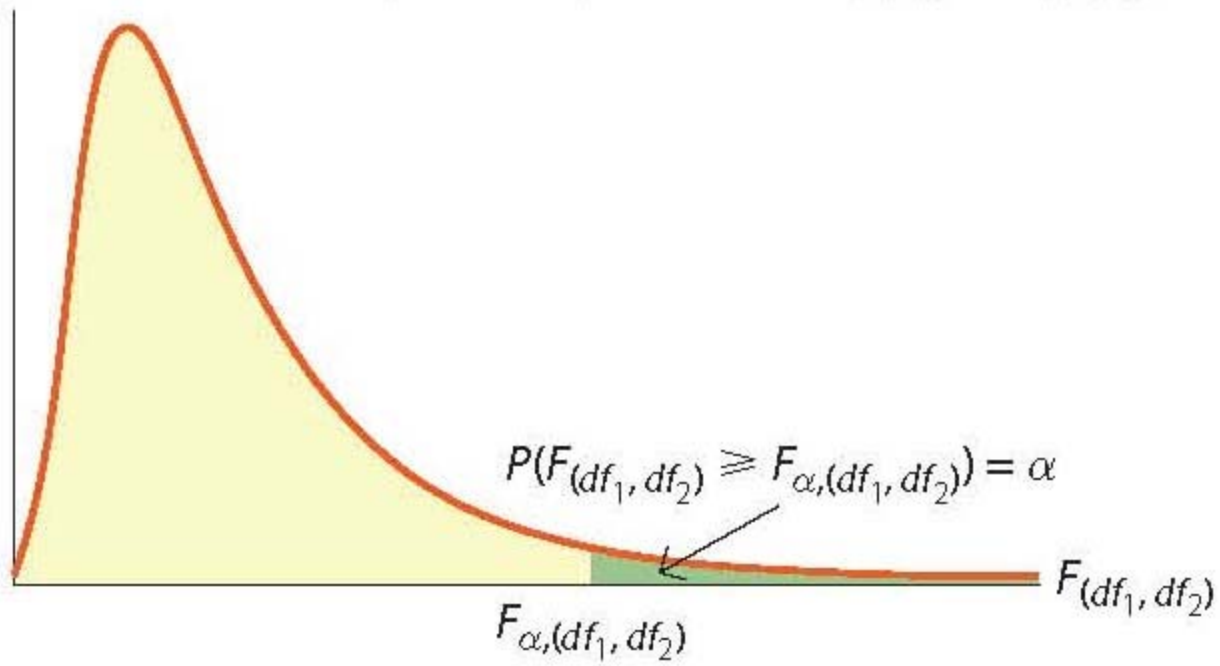
- For $c = 4$ and $n_T = 24$, we use the $F_{(3,20)}$ distribution. At the 5% significance level, the critical value is 3.10.

The *F* Distribution

- $F(df_1, df_2)$ distribution is a family of distributions, each one is defined by two degrees of freedom parameters, one for the numerator and one for the denominator.

- More details of F distribution can be found in Chapter 11.

The *F* Distribution



- $F_{\alpha,(df_1,df_2)}$ represents a value such that the area in the right tail of the distribution is α.
- With two *df* parameters, *F* tables occupy several pages.

Right-tail Values

Denominator Degrees of Freedom, df_2	Area in Upper Tail, α	Numerator Degrees of Freedom, df_1		
		6	7	8
6	0.10	3.05	3.01	2.98
	0.05	4.28	4.21	4.15
	0.025	5.82	5.70	5.60
	0.01	8.47	8.26	8.10
7	0.10	2.83	2.78	2.75
	0.05	3.87	3.79	3.73
	0.025	5.12	4.99	4.90
	0.01	7.19	6.99	6.84
8	0.10	2.67	2.62	2.59
	0.05	**3.58**	3.50	3.44
	0.025	4.65	4.53	4.43
	0.01	**6.37**	6.18	6.03

With $df_1 = 6$ and $df_2 = 8$, 5% of the area falls above 3.58.

Left-tail values

- $F_{1-\alpha,(df_1,df_2)}$ represents a value such that the area in the left tail of the distribution is α.
- $F_{1-\alpha,(df_1,df_2)} = \frac{1}{F_{\alpha,(df_2,df_1)}}$
- For an $F_{(6,8)}$ distribution, find the value such that the area in the left tail is 5%, or $F_{0.95(6,8)}$.
- First find $F_{0.05,(8,6)}$. This value is 4.15.
- $F_{0.95(6,8)} = \frac{1}{4.15} = 0.24.$

LO

Do savings differ by city?

- We have computed *MSTR* = 4,401,573 and *MSE* = 7,209.

- Our test statistic is then:

$$F_{(3,20)} = 4{,}401{,}573 / 7{,}209 = 610.57.$$

- This greatly exceeds the critical value of 3.10, so we conclude that the cost savings differ across cities.

- The ANOVA test does not tell us which cities have different cost savings, but later in the chapter we will develop techniques to help answer these questions.

LO

Excel and One-Way ANOVA

- By choosing **Data > Data Analysis > ANOVA: Single Factor**, Excel will open a dialog box that will enable one-way ANOVA.
- After choosing the data for the *Input Range*, Excel reports the following output:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	13204720	3	4401573	**610.57**	**7.96E-20**	3.098
Within Groups	144180	20	7209			
Total	13348900	23				

LO 13.2

Excel and One-Way ANOVA

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Between Groups	13204720	3	4401573	**610.57**	**7.96E-20**	3.098
Within Groups	144180	20	7209			
Total	13348900	23				

- In addition to using the critical value approach to perform the hypothesis test, we can also easily use the *p*-value approach with the Excel output. Since the *p*-value is very close to zero, we again reject the null hypothesis.

13.2 Multiple Comparison Methods

LO 13.3 Use confidence intervals and Tukey's HSD method in order to determine which means differ.

- When the one-way ANOVA finds significant differences between the population means, it is natural to ask which means differ.
- In this section we show two techniques for performing this follow-up analysis:
  - Fisher's Least Difference Method
  - Tukey's Honestly Significant Differences Method

LO

Multiple Comparison Methods

- When comparing two population means, we compute:

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2, n_i + n_j - 2} \sqrt{s_p^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)},$$

where s_p^2 is the pooled sample variance.

- Here we will improve upon the precision of this estimate by substituting *MSE* from the ANOVA test for s_p^2.

LO

Fisher's Confidence Intervals

- For comparing population means μ_i and μ_j as a follow-up to the ANOVA test, we can form Fisher's confidence interval:

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2, n_T - c} \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- The *t*-distribution has (n_T - *c*) degrees of freedom regardless of which two means we are comparing.

LO

Transportation Example

Relevant summary statistics for our example include:

Boston	New York	San Francisco	Chicago
$\bar{x}_1 = \$12{,}622$	$\bar{x}_2 = \$12{,}585$	$\bar{x}_3 = \$11{,}720$	$\bar{x}_4 = \$10{,}730$
$s_1 = \$87.79$	$s_2 = \$80.40$	$s_3 = \$83.96$	$s_4 = \$90.62$
$n_1 = 5$	$n_2 = 8$	$n_3 = 6$	$n_4 = 5$

We have $n_T = 24$ with $MSE = 7{,}209$. For a 95% confidence interval the appropriate t-statistic is $t_{0.025,20} = 2.086$.

The confidence interval for the difference between Boston and New York is:

$$(12622 - 12585) \pm 2.086\sqrt{7209\left(\frac{1}{5} + \frac{1}{8}\right)} = 37 \pm 100.97$$

LO 13.3

Interval Interpretation

The interval for μ_{Boston} - $\mu_{New\ York}$ is -63.97 to 137.97. Since the interval includes 0, we cannot conclude that Bostonians save more on transportation cost as compared to New Yorkers.

Let's compare San Francisco and Chicago:

$$(11720-10730)\pm 2.086\sqrt{7209\left(\frac{1}{6}+\frac{1}{5}\right)}=990\pm 107.25$$

This implies San Franciscans do save more than their Windy City counterparts.

LO

Overall Error Rate

- When we use a 95% confidence interval it has a 5% Type I error rate.
- If we form two intervals simultaneously, we could err on either one. The combined error rate is much larger than 5%.
- In the transportation example there are 4 cities, for a total of 6 possible pairwise comparisons. Forming all six would greatly increase the overall error rate.

LO

The Tukey HSD Procedure

- Therefore, if you know ahead of time you will want to make many comparisons, **Tukey's HSD method** reduces the incidence of Type I Error.

- It controls the combined error rate by making each interval slightly wider than the Fisher counterpart.

- Instead of a *t*-distribution multiplier, the HSD procedure uses a value from a new distribution.

LO

The Tukey HSD Procedure

- Tukey's method generates confidence intervals of the form:

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha,(c,n_T-c)} \sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- The *q*-statistic comes from the **studentized range distribution** with c and (n_T – c) degrees of freedom.

LO

Applying Tukey's HSD Procedure

- Suppose that the results from an ANOVA test showed that prices for a generic drug differed across three regions of California, with an *MSE* of 3.86.
- Let's use Tukey's HSD method to determine which regions' means differ.
- The sample statistics for the three regions are:

SUMMARY					
Groups	*Count*	*Sum*	*Average*	*Variance*	
Region 1	10	350	35.0	3.78	
Region 2	10	342	34.2	5.07	
Region 3	10	395	39.5	2.72	

LO

Applying Tukey's HSD Procedure

- First, with α = 0.05, $c = 3$, and $(n_T - c) = 30 - 3 = 27$, the appropriate *q*-statistic is $q_{0.05,(3,27)} = 3.51$.

- Comparing Regions 1 and 2 yields:

$$(35.0 - 34.2) \pm 3.51\sqrt{\frac{3.86}{2}\left(\frac{1}{10} + \frac{1}{10}\right)} = 0.80 \pm 2.18$$

- For Regions 1 and 3, we obtain -4.5 ± 2.18. For Regions 2 and 3, we obtain -5.3 ± 2.18. Prices in Region 3 significantly differ from Regions 1 and 2.

13.3 Two-Way ANOVA (No Interaction)

LO 13.4 Conduct and evaluate hypothesis tests based on two-way ANOVA with no interaction.

- We now consider problems where the data are categorized by two factors.
- For example, we may want to determine if the brand of a hybrid car and the octane level of gasoline influence average miles per gallon.
- Using a two-way ANOVA, we are able to assess the effect of each factor while controlling for the other one.

LO

Two-Way ANOVA Example

- An undergraduate student is trying to decide what career to pursue.
- She interviews four workers in three fields and determines their salaries.

Educational Services	Financial Services	Medical Services
35	58	110
18	90	62
75	25	26
46	45	43

- The average salaries of 43.50, 54.50 and 60.20 are not significantly different ($F_{(2,9)} = 0.33$) using a one-way ANOVA.

LO

Two-Way ANOVA Example

- If the education level of the 12 workers is considered, a different story emerges.

Education Level (Factor B)	Field of Employment (Factor A)			Factor B Means
	Educational Services	Financial Services	Medical Services	
High School	18	25	26	$\bar{X}_{high\ school} = 23.00$
Bachelor's	35	45	43	$\bar{X}_{bachelor's} = 41.00$
Master's	46	58	62	$\bar{X}_{master's} = 55.33$
Ph.D.	75	90	110	$\bar{X}_{Ph.D.} = 91.67$
Factor A Means	$\bar{X}_{education} = 43.50$	$\bar{X}_{financial} = 54.50$	$\bar{X}_{medical} = 60.25$	$\bar{\bar{X}} = 52.75$

- It is clear that education also impacts wage.

LO

The Randomized Block Design

- This type of two-way ANOVA is called a **randomized block design**.
- The term "block" refers to a matched set of observations across the treatments.
- In the salary example, the treatments are the three fields of employment.
- The blocks are the education levels. Until we account for them, we cannot capture the employment field effects.

LO

The ANOVA Layout

Source of Variation	SS	df	MS	F
Rows	SSB	$r-1$	$MSB=\frac{SSB}{r-1}$	$F_{(df_1,df_2)}=\frac{MSB}{MSE}$
Columns	SSA	$c-1$	$MSA=\frac{SSA}{c-1}$	$F_{(df_1,df_2)}=\frac{MSA}{MSE}$
Error	SSE	$n_T-c-r+1$	$MSE=\frac{SSE}{(n_T-c-r+1)}$	
Total	SST	n_T-1		

There are now three sources of variation:

1. Row variability (due to blocks or Factor B),
2. Column variability (due to treatments or Factor A), and
3. Variability due to chance or *SSE*

LO

ANOVA for the Salary Example

- After selecting **Data > Data Analysis > ANOVA: Two Factor Without Replication** and choosing the data for the *Input Range*, Excel reports the following output:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Rows	7632.92	3	2544.31	56.58	8.6E-05	4.76
Columns	579.50	2	289.75	6.44	0.03207	5.14
Error	269.83	6	44.97			
Total	8482.25	11				

- Note that the procedure includes the value for each $F_{(df_1, df_2)}$ test statistic, as well as the associated *p*-values and critical values.

LO 13.4

Interpreting the Results

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Rows	7632.92	3	2544.31	56.58	8.6E-05	4.76
Columns	579.50	2	289.75	6.44	0.03207	5.14
Error	269.83	6	44.97			
Total	8482.25	11				

- The *p*-value for column variation is about 0.03. So, after controlling for educational level, the average salaries differ by field at the 5% significance level.

- Salary also differs by educational level, as indicated by the very small *p*-value.

13.4: Two-Way ANOVA with Interaction

LO 13.5 Conduct and evaluate hypothesis tests based on two-way ANOVA with interaction.

- Now we will look at data categorized by two factors, but with two or more values observed in each "cell."
- In two-way ANOVA with interaction, we partition the total variability of the data set into four components: *SSA*, *SSB*, *SSAB*, and *SSE*.

LO 13.5

What is Interaction?

- Interaction means that the effect of one factor depends on the level of the other factor.

- For example, perhaps education impacts salaries in the financial sector, but not in professional sports. The two categories, employment sector and education, interact differently depending on the sector.

LO

Another Salary Example

Education Level (Factor B)	Field of Employment (Factor A)			Factor B Means
	Educational Services	Financial Services	Medical Services	
High School	22.33	25.67	25.00	24.33
Bachelor's	33.00	46.00	43.33	40.78
Master's	47.67	54.67	59.33	53.89
Ph.D.	77.00	92.33	98.33	89.22
Factor A Means	45.00	54.67	56.50	$\bar{\bar{x}} = 52.06$

- Each interior cell entry represents the average salary of 3 employees who fit the categories.
- By choosing **Data > Data Analysis > ANOVA: Two Factor With Replication**, we are able to obtain Excel output.

LO

Two-way ANOVA with Interaction

- Here is the Excel output for a two-way ANOVA with interaction:

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Sample (Rows)	20524	3	6841	658.5	3.58E-23	3.009
Columns	916.2	2	458.1	44.1	9.18E-09	3.403
Interaction	318.4	6	53.07	5.109	0.001659	2.508
Within (Error)	249.3	24	10.39			
Total	22008	35				

- Now, we have values for three $F_{(df_1, df_2)}$ test statistics, as well as corresponding *p*-values and critical values.

LO

Interpreting the Output

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*p-value*	*F crit*
Sample (Rows)	20524	3	6841	658.5	3.58E-23	3.009
Columns	916.2	2	458.1	44.1	9.18E-09	3.403
Interaction	318.4	6	53.07	5.109	0.001659	2.508
Within (Error)	249.3	24	10.39			
Total	22008	35				

- If the output indicates that interaction between the factors is significant (as is the case here), then interpretation of the main effects is complicated.
- When interaction is significant, we should be careful about making conclusions.

End of Chapter