Chapter 13

Analysis of Variance

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.

LO 13.1 **Provide a conceptual overview of ANOVA.**

13.1 One-Way 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.

13.1.1 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.

13.1.2 The Hypothesis Test

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

- The competing hypotheses for the one-way ANOVA:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_c$$
$$H_A : \text{Not all population means are equal}$$

13.1.3 The ANOVA Concept

- The competing hypotheses are displayed graphically below.

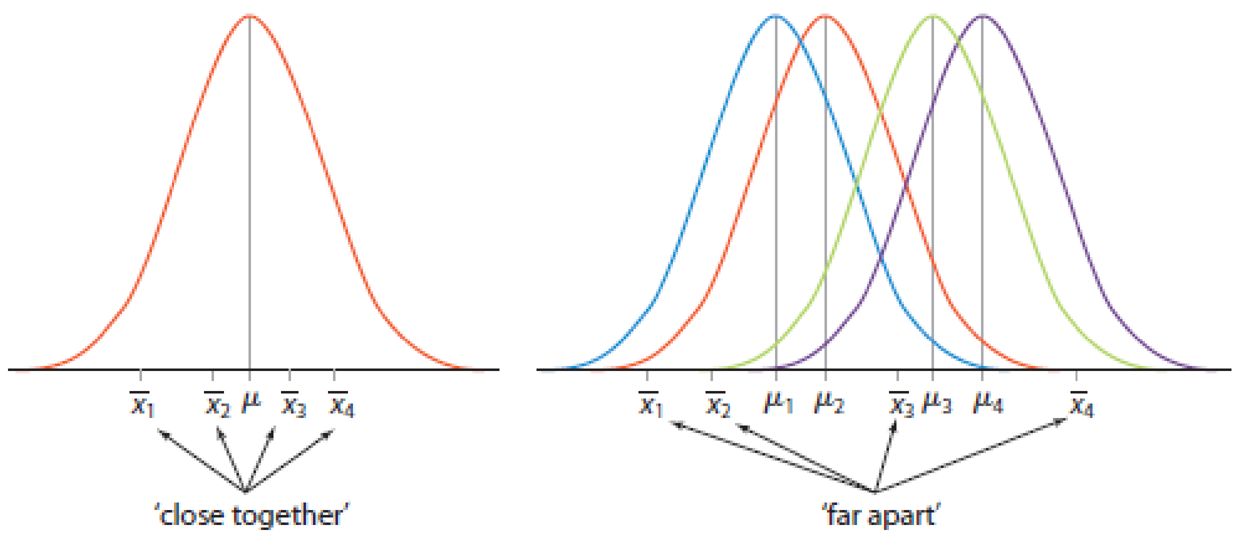

Figure 13.1: The ANOVA Concept.

- 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.

13.1.4 Methodology

- We first compute the amount of variability between the sample means.
- Then we measure how much variability these 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.

13.1.5 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 $\bar{\bar{x}}$, simply average all the values from the dataset:

$$\bar{\bar{x}} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{n_i} x_{ij}}{n_T} \tag{13.1}$$

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

$$SSTR = \sum_{i=1}^{c} n_i \left(\bar{x}_i - \bar{\bar{x}}\right)^2 \tag{13.2}$$

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

$$MSTR = \frac{SSTR}{c-1} \tag{13.3}$$

- MSTR is our measure of variability between samples.

13.1.6 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:

$$SSE = \sum_{i=1}^{c} (n_i - 1) s_i^2 \tag{13.4}$$

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

$$MSE = \frac{SSE}{n_T - c} \tag{13.5}$$

13.1.7 The F Test

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

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_A : \text{Not all population means are equal}$$

- The value of the test statistic is calculated as

$$F_{(df_1, df_2)} = \frac{MSTR}{MSE}, \tag{13.6}$$

where $df_1 = c - 1$ 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.

13.1.8 The F distribution

- $F_{(df_1,df_2)}$ distribution is a family of distributions, each one is define 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.
- $F_{\alpha,(df_1,df_2)}$ 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.

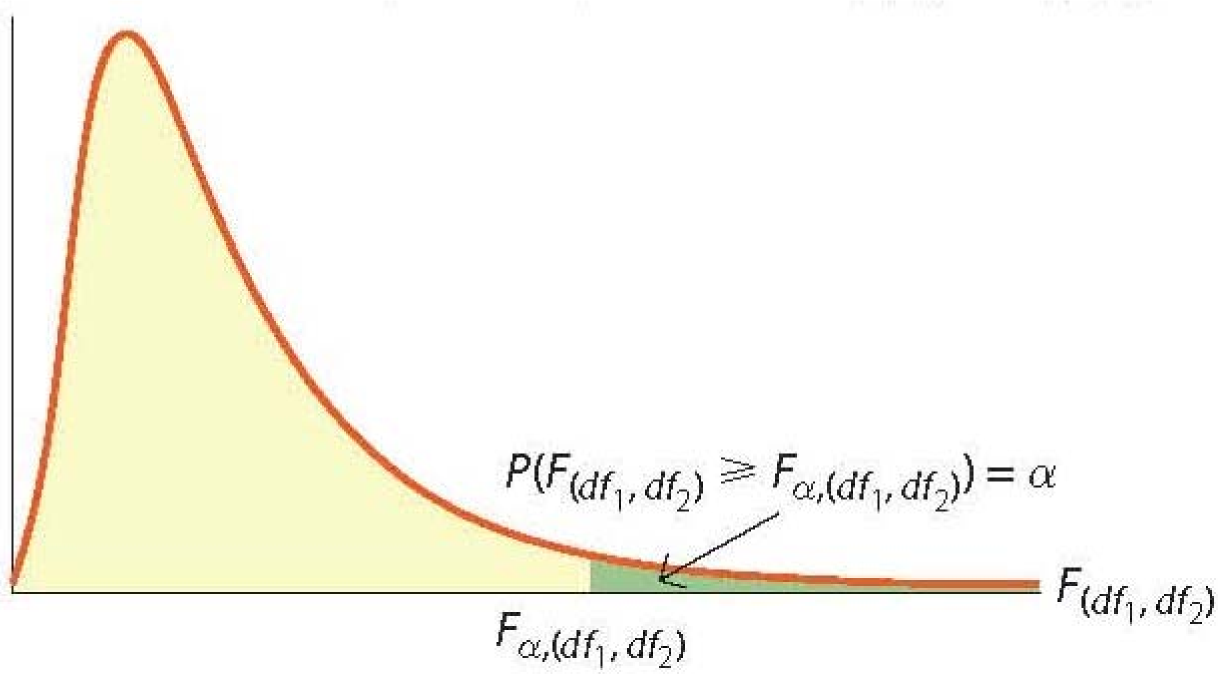

Figure 13.2: The F-distribution

13.1.9 Right-tail Values

- With $df_1 = 6$ and $df_2 = 8$, 5% of the area falls above $\underline{3.58}$.

13.1.10 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.

$$F_{1-\alpha,(df_1,df_2)} = \frac{1}{F_{\alpha,(df_2,df_1)}} \tag{13.7}$$

- 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 $F_{0.05,(8,6)} = 4.15$.

$$\begin{aligned} F_{(0.95,(6,8))} &= \frac{1}{4.15} \\ &= 0.24 \end{aligned}$$

13.1.11 Do savings differ by city?

- We have computed $MSTR = 4,401,573$ and $MSE = 7,209$.
- Our test statistic is then:

$$\begin{aligned} F_{(3,20)} &= \frac{4,401,573}{7,209} \\ &= 610.57 \end{aligned}$$

- The 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.

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

13.3 Two-Way ANOVA with Interactions

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 the 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.

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

Table 13.1: Workers Education Level

	Field of Employment (Factor A)			
Education Level (Factor B)	Educational Services	Financial Services	Medical Services	Factor B M
High School	18	25	26	2
Bachelor's	35	45	43	4
Master's	46	58	62	5
Ph.D.	75	90	110	9
Factor A Means	43.50	54.50	60.25	5

- It is clear that education also impacts wage.

13.3.1 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.

13.3.2 The ANOVA Layout

Table 13.2: 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	$r-1$	$MSA = \frac{SSA}{c-1}$	$F_{(df_1,df_2)} = \frac{MSA}{MSE}$
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 F),
2. Column variability (due to treatments of Factor A), and
3. Variability due to chance or SSE

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.

13.4.1 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.