QM353: Business Statistics

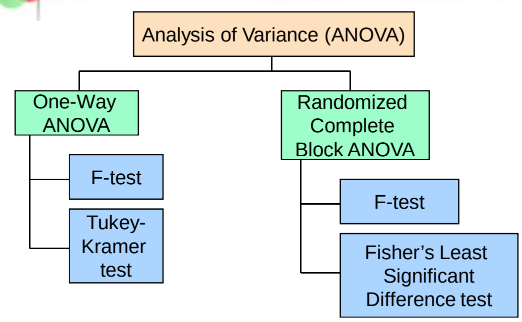
Chapter 5
Analysis of Variance (ANOVA)

Chapter Goals

After completing this chapter, you should be able to:

- Recognize situations in which to use analysis of variance
- Understand different analysis of variance designs
- Perform a single-factor hypothesis test and interpret results
- Conduct and interpret post-analysis of variance pairwise comparisons procedures
- Set up and perform randomized blocks analysis

Chapter Overview



Logic of Analysis of Variance

- Investigator controls one or more independent variables
  - Called factors (or treatment variables)
  - Each factor contains two or more levels (or categories/classifications)
- Observe effects on dependent variable
  - Response to levels of independent variable
- Experimental design: the plan used to test hypothesis

Completely Randomized Design

- Experimental units (subjects) are assigned randomly to treatments
- Only one factor or independent variable
  - With two or more treatment levels
- Analyzed by
  - One-factor analysis of variance (one-way ANOVA)
- Called a Balanced Design if all factor levels have equal sample size

One-Way Analysis of Variance

- Evaluate the difference among the means of three or more populations

> Examples: • Accident rates for 1st, 2nd, and 3rd shift
> • Expected mileage for five brands of tires

- Assumptions
  - Populations are normally distributed
  - Populations have equal variances
  - Samples are randomly and independently drawn
  - Data's measurement level is interval or ratio

Hypotheses of One-Way ANOVA

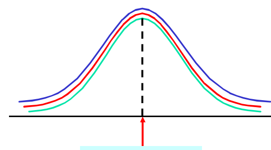
- $H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$
  - All population means are equal
  - i.e., no treatment effect (no variation in means among groups)
- H_A : Not all of the population means are the same
  - At least one population mean is different
  - i.e., there is a treatment effect
  - Does not mean that all population means are different (some pairs may be the same)

One-Factor ANOVA

$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_A : Not all μ_i are the same

All Means are the same:
The Null Hypothesis is True
(No Treatment Effect)



$\mu_1 = \mu_2 = \mu_3$

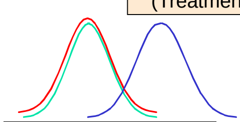
One-Factor ANOVA

(continued)

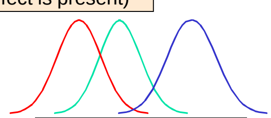
$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$

H_A : Not all μ_i are the same

At least one mean is different:
The Null Hypothesis is NOT true
(Treatment Effect is present)



or



$\mu_1 = \mu_2 \neq \mu_3$ $\mu_1 \neq \mu_2 \neq \mu_3$

Partitioning the Variation

- Total variation can be split into two parts:

SST = SSB + SSW

SST = Total Sum of Squares (total variation)
SSB = Sum of Squares Between (variation between samples)
SSW = Sum of Squares Within (within each factor level)

Partitioning the Variation

(continued)

SST = SSB + SSW

Total Variation (SST) = the aggregate dispersion of the individual data values across the various factor levels

Between-Sample Variation (SSB) = dispersion among the factor sample means

Within-Sample Variation (SSW) = dispersion that exists among the data values within a particular factor level

Partition of Total Variation

Total Variation (SST)

= **Variation Due to Factor (SSB)** + **Variation Due to Random Sampling (SSW)**

Commonly referred to as:
- Sum of Squares Between
- Sum of Squares Among
- Sum of Squares Explained
- Among Groups Variation

Commonly referred to as:
- Sum of Squares Within
- Sum of Squares Error
- Sum of Squares Unexplained
- Within Groups Variation

Total Sum of Squares

SST = SSB + SSW

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{\bar{x}})^2$$

Where:

SST = Total sum of squares

k = number of populations (levels or treatments)

n_i = sample size from population i

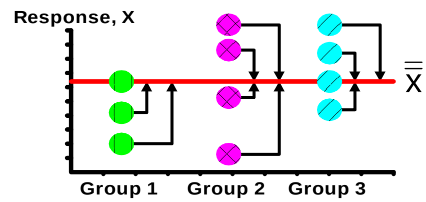
x_{ij} = j^{th} measurement from population i

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Total Variation

(continued)

$$SST = (x_{11} - \bar{\bar{x}})^2 + (x_{12} - \bar{\bar{x}})^2 + \ldots + (x_{kn_k} - \bar{\bar{x}})^2$$



Sum of Squares Between

SST = SSB + SSW

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

Where:

SSB = Sum of squares between

k = number of populations

n_i = sample size from population i

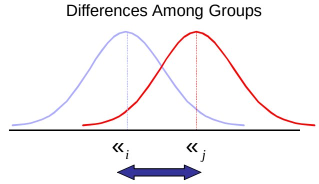
$\bar{x}_i$ = sample mean from population i

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Between-Group Variation

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

Variation Due to Differences Among Groups



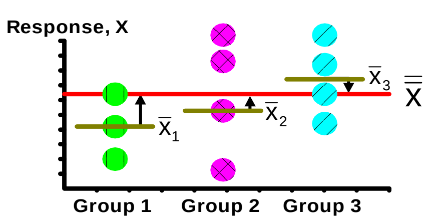
$$MSB = \frac{SSB}{k-1}$$

Mean Square Between = SSB/degrees of freedom

Between-Group Variation

(continued)

$$SSB = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \ldots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$



Sum of Squares Within

SST = SSB + SSW

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_j}(x_{ij} - \bar{x}_i)^2$$

Where:

SSW = Sum of squares within

k = number of populations

n_i = sample size from population i

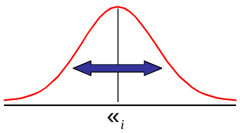
$\bar{x}_i$ = sample mean from population i

x_{ij} = j^{th} measurement from population i

Within-Group Variation

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_i)^2$$

Summing the variation within each group and then adding over all groups



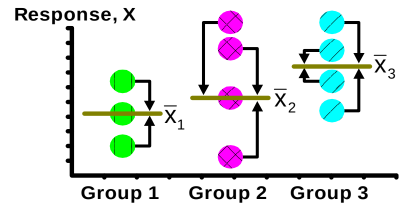
$$MSW = \frac{SSW}{n_T - k}$$

Mean Square Within = SSW/degrees of freedom

Within-Group Variation

(continued)

$$SSW = (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_2)^2 + \ldots + (x_{kn_k} - \bar{x}_k)^2$$



One-Way ANOVA Table

Source of Variation	SS	df	MS	F ratio
Between Samples	SSB	k - 1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Samples	SSW	n_T - k	$MSW = \frac{SSW}{n_T - k}$	
Total	SST = SSB+SSW	n_T - 1		

k = number of populations
n_T = sum of the sample sizes from all populations
df = degrees of freedom

One-Factor ANOVA F Test Statistic

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
H_A: At least two population means are different

- Test statistic

$$F = \frac{MSB}{MSW}$$

MSB is mean squares between variances
MSW is mean squares within variances

- Degrees of freedom
  - $df_1 = k - 1$ (k = number of populations)
  - $df_2 = n_T - k$ (n_T = sum of sample sizes from all populations)

Interpreting One-Factor ANOVA F Statistic

- The F statistic is the ratio of the between estimate of variance and the within estimate of variance
  - The ratio must always be positive
  - df_1 = k -1 will typically be small
  - df_2 = n_T - k will typically be large

The ratio should be close to 1 if H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ is true

The ratio will be larger than 1 if H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ is false

ANOVA Steps

1. Specify parameter of interest
2. Formulate hypotheses
3. Specify the significance level, α
4. Select independent, random samples
   - Compute sample means and grand mean
5. Determine the decision rule
6. Verify the normality and equal variance assumptions have been satisfied
7. Create ANOVA table
8. Reach a decision and draw a conclusion

One-Factor ANOVA F Test Example

You want to see if three different golf clubs yield different distances. You randomly select five measurements from trials on an automated driving machine for each club. At the 0.05 significance level, is there a difference in mean distance?

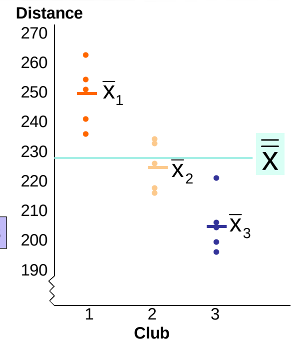
Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

One-Factor ANOVA Example: Scatter Diagram

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$ $\bar{x}_2 = 226.0$ $\bar{x}_3 = 205.8$

$\bar{\bar{x}} = 227.0$



One-Factor ANOVA Example Computations

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

$\bar{x}_1 = 249.2$ $n_1 = 5$

$\bar{x}_2 = 226.0$ $n_2 = 5$

$\bar{x}_3 = 205.8$ $n_3 = 5$

$\bar{\bar{x}} = 227.0$ $n_T = 15$

$k = 3$

SSB = $5 [(249.2 – 227)^2 + (226 – 227)^2 + (205.8 – 227)^2] = 4716.4$

SSW = $(254 – 249.2)^2 + (263 – 249.2)^2 + \ldots + (204 – 205.8)^2 = 1119.6$

MSB = 4716.4 / (3-1) = 2358.2

MSW = 1119.6 / (15-3) = 93.3

$$F = \frac{2358.2}{93.3} = 25.275$$

One-Factor ANOVA Example Solution

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

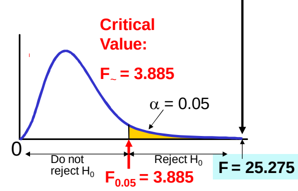
H_A: μ_i not all equal

~ = 0.05

$df_1 = 2$ $df_2 = 12$

Critical Value:

F~ = 3.885



Test Statistic:

$$F = \frac{MSB}{MSW} = \frac{2358.2}{93.3} = 25.275$$

Decision:

Reject H_0 at ~ = 0.05

Conclusion:

There is evidence that at least one μ_i differs from the rest

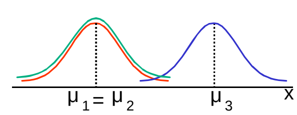
ANOVA -- Single Factor: Excel Output

EXCEL: tools | data analysis | ANOVA: single factor

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Club 1	5	1246	249.2	108.2		
Club 2	5	1130	226	77.5		
Club 3	5	1029	205.8	94.2		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	**4716.4**	**2**	**2358.2**	**25.275**	4.99E-05	**3.885**
Within Groups	**1119.6**	**12**	**93.3**			
Total	5836.0	14				

The Tukey-Kramer Procedure

- Tells which population means are significantly different
  - e.g.: $\mu_1 = \mu_2 \neq \mu_3$
  - Done after rejection of equal means in ANOVA
- Allows pair-wise comparisons
  - Compare absolute mean differences with critical range



Tukey-Kramer Critical Range

$$\text{Critical Range} = q_{1-\alpha}\sqrt{\frac{\text{MSW}}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}$$

where:

$q_{1-\alpha}$ = Value from standardized range table with k and n_T - k degrees of freedom for the desired level of α

MSW = Mean Square Within

n_i and n_j = Sample sizes from populations (levels) i and j

The Tukey-Kramer Procedure: Example

Club 1	Club 2	Club 3
254	234	200
263	218	222
241	235	197
237	227	206
251	216	204

1. Compute absolute mean differences:

$$|\bar{x}_1 - \bar{x}_2| = |249.2 - 226.0| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = |249.2 - 205.8| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = |226.0 - 205.8| = 20.2$$

2. Find the q value from the table in appendix J with k and n_T - k degrees of freedom for the desired level of α

$$q_{1-\alpha} = 3.77$$

The Tukey-Kramer Procedure: Example

3. Compute Critical Range:

$$\text{Critical Range} = q_{1-\alpha}\sqrt{\frac{\text{MSW}}{2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)} = 3.77\sqrt{\frac{93.3}{2}\left(\frac{1}{5}+\frac{1}{5}\right)} = 16.285$$

4. Compare:

$$|\bar{x}_1 - \bar{x}_2| = 23.2$$

$$|\bar{x}_1 - \bar{x}_3| = 43.4$$

$$|\bar{x}_2 - \bar{x}_3| = 20.2$$

5. All of the absolute mean differences are greater than critical range. Therefore there is a significant difference between each pair of means at 5% level of significance.

Randomized Complete Block ANOVA

- Like One-Way ANOVA, we test for equal population means (for different factor levels, for example)...
- ...but we want to control for possible variation from a second factor (with two or more levels)
- Used when more than one factor may influence the value of the dependent variable, but only one is of key interest
- Levels of the secondary factor are called blocks

Randomized Complete Block ANOVA

(continued)

- **Assumptions**
  - Populations are normally distributed
  - Populations have equal variances
  - The observations within samples are independent
  - The date measurement must be interval or ratio
- **Application examples**
  - Testing 5 routes to a destination through 3 different cab companies to see if differences exist
  - Determining the best training program (out of 4 choices) for various departments within a company

Partitioning the Variation

- Total variation can now be split into three parts:

SST = SSB + SSBL + SSW

SST = Total sum of squares
SSB = Sum of squares between factor levels
SSBL = Sum of squares between blocks
SSW = Sum of squares within levels

Sum of Squares for Blocking

SST = SSB + SSBL + SSW

$$SSBL = \sum_{j=1}^{b} k(\bar{x}_j - \bar{\bar{x}})^2$$

Where:

k = number of levels for this factor

b = number of blocks

$\bar{x}_j$ = sample mean from the j^{th} block

$\bar{\bar{x}}$ = grand mean (mean of all data values)

Partitioning the Variation

- Total variation can now be split into three parts:

SST = SSB + SSBL + SSW

SST and SSB are computed as they were in One-Way ANOVA

SSW = SST – (SSB + SSBL)

Mean Squares

$$MSBL = \text{Mean square blocking} = \frac{SSBL}{b-1}$$

$$MSB = \text{Mean square between} = \frac{SSB}{k-1}$$

$$MSW = \text{Mean square within} = \frac{SSW}{(k-1)(b-1)}$$

Randomized Block ANOVA Table

Source of Variation	SS	df	MS	F ratio
Between Blocks	SSBL	b - 1	MSBL	$\frac{MSBL}{MSW}$
Between Samples	SSB	k - 1	MSB	$\frac{MSB}{MSW}$
Within Samples	SSW	(k–1)(b-1)	MSW	
Total	SST	n_T - 1		

k = number of populations
b = number of blocks
n_T = sum of the sample sizes from all populations
df = degrees of freedom

Blocking Test

$H_0: \mu_{b1} = \mu_{b2} = \mu_{b3} = \ldots$

H_A: Not all block means are equal

$$F = \frac{MSBL}{MSW}$$

- Blocking test: $df_1 = b - 1$, $df_2 = (k - 1)(b - 1)$

Reject H_0 if $F > F_\alpha$

Main Factor Test

$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

H_A: Not all population means are equal

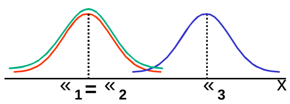
$$F = \frac{MSB}{MSW}$$

- Main Factor test: $df_1 = k - 1$, $df_2 = (k - 1)(b - 1)$

Reject H_0 if $F > F_\alpha$

Fisher's Least Significant Difference Test

- To test which population means are significantly different
  - e.g.: $\mu_1 = \mu_2 \neq \mu_3$
  - Done after rejection of equal means in randomized block ANOVA design
- Allows pair-wise comparisons
  - Compare absolute mean differences with critical range



Fisher's Least Significant Difference (LSD) Test

$$LSD = t_{\alpha/2}\sqrt{MSW}\sqrt{\frac{2}{b}}$$

where:

$t_{\alpha/2}$ = Upper-tailed value from Student's t-distribution for $\alpha/2$ and $(k - 1)(b - 1)$ degrees of freedom

MSW = Mean square within from ANOVA table

b = number of blocks

k = number of levels of the main factor

NOTE: This is a similar process as Tukey-Kramer

Fisher's Least Significant Difference (LSD) Test

(continued)

$$LSD = t_{\alpha/2}\sqrt{MSW}\sqrt{\frac{2}{b}}$$

Compare:

Is $|\bar{x}_i - \bar{x}_j| > LSD$?

$|\bar{x}_1 - \bar{x}_2|$

$|\bar{x}_1 - \bar{x}_3|$

$|\bar{x}_2 - \bar{x}_3|$

etc...

If the absolute mean difference is greater than LSD then there is a significant difference between that pair of means at the chosen level of significance

Chapter Summary

- Described one-way analysis of variance
  - The logic of ANOVA
  - ANOVA assumptions
  - F test for difference in k means
  - The Tukey-Kramer procedure for multiple comparisons
- Described randomized complete block designs
  - F test
  - Fisher's least significant difference test for multiple comparisons