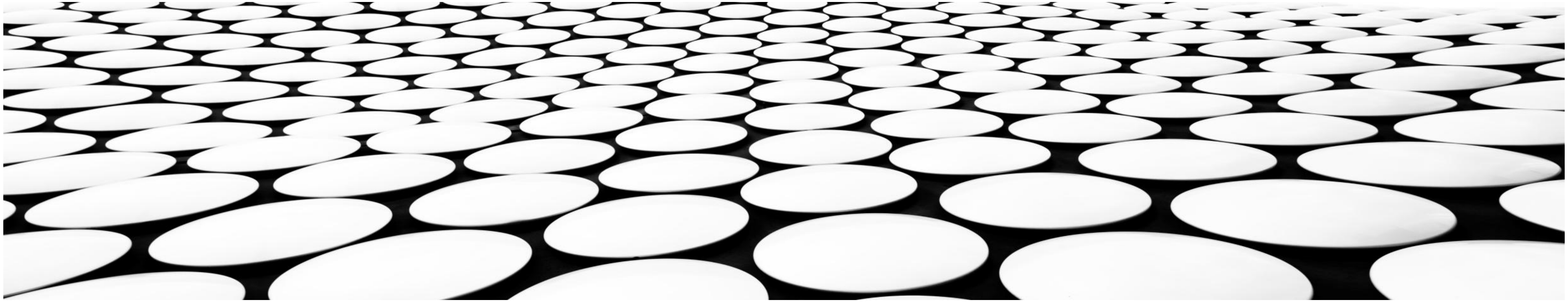
DATA MINING AND PREDICTIVE DATA ANALYTICS

CHAPTER-6

MULTIVARIATE STATISTICS



INTRODUCTION

- **Multivariate statistical analysis helps examine:**
  - Relationships between two variables (bivariate analysis)
  - Relationships between a target variable and multiple predictors
  - Joint behavior of several variables in a dataset
- A key use in data mining is **validating training–test splits**.

  Datasets are typically divided into:
  - **Training set** – for model building
  - **Test set** – for performance evaluation
- For cross-validation to be valid, both sets must represent the **same population**.
- If key variables (means or proportions) differ significantly between them, the partition is biased and the model will not generalize well.

INTRODUCTION

- Data miners use bivariate hypothesis testing to compare:
  - Means of continuous variables
  - Proportions of binary (flag) variables
  - Category distributions of multinomial variables
- These tests check whether the **training and test datasets are statistically similar**.

Type of Variable	Appropriate Test
Continuous variable	Two-sample t-test for difference in means
Binary (flag) variable	Two-sample Z-test for difference in proportions
Multinomial variable	Test for homogeneity of proportions (Chi-square test)

- In practice, only a few randomly chosen variables need to be tested.
- If these are similar, typically the whole dataset is consistent.

TWO-SAMPLE T-TEST FOR DIFFERENCE IN MEANS

- Used when comparing **population means** for a continuous variable across two independent samples.
- **Purpose:** To test whether: $\mu_1 = \mu_2 \quad \text{or} \quad \mu_1 \neq \mu_2$
  - where:
    - μ1 = mean of population 1 (training set)
    - μ2 = mean of population 2 (test set)
- **Test Statistic**
  - For two independent samples:

$$t_{\text{obs}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where:

- $\bar{X}_1, \bar{X}_2$ = sample means
- s_1^2, s_2^2 = sample variances
- n_1, n_2 = sample sizes

TWO-SAMPLE T-TEST FOR DIFFERENCE IN MEANS

- **Distribution**
  - The statistic approximately follows a **t distribution** with: $df = \min(n1-1, n2-1)$
  - This is an acceptable approximation when:
    - Both populations are **normally distributed**, or
    - Both sample sizes are **large** ($n \geq 30$)

TWO-SAMPLE T-TEST FOR DIFFERENCE IN MEANS

- **Example:** Customer Service Calls - Churn Dataset
  - churn data set is partitioned into a training set of 2529 records and a test set of 804 records.
  - We are to assess the validity of the partition by testing whether the population mean number of *customer service calls* differs between the two data sets.
  - Given Summary Statistics

Dataset	Mean ( )	SD (s)	Sample Size (n)
Training set	1.5714	1.3126	2529
Test set	1.5361	1.3251	804

TWO-SAMPLE T-TEST FOR DIFFERENCE IN MEANS

- Hypotheses:
  - This is a **two-tailed two-sample t-test** for the difference in means.

$$H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$$

  - Test Statistic

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t_{\text{obs}} = \frac{1.5714 - 1.5361}{\sqrt{\frac{1.3126^2}{2529} + \frac{1.3251^2}{804}}}$$

$$\frac{s_1^2}{n_1} = \frac{1.7229}{2529} = 0.000681$$

$$\frac{s_2^2}{n_2} = \frac{1.7558}{804} = 0.002184$$

$$\sqrt{0.000681 + 0.002184} = \sqrt{0.002865} = 0.05352$$

$$t_{\text{obs}} = \frac{0.0353}{0.05352} = 0.6595$$

TWO-SAMPLE T-TEST FOR DIFFERENCE IN MEANS

- **p-Value (Two-Tailed)**
  - *Degree of Freedom, Df = min* (n1−1,n2−1) = *min* (2528,803) = 803
  - Compute: $p = 2 \cdot P(t > |0.6595|)$
  - Using df = 803, $p = 2 \cdot (1 - 0.7449) = 0.5098$
  - Since p = 0.5098, which is much larger than 0.05, we <u>fail to reject the null hypothesis</u>.
  - There is no evidence of a difference in the mean number of *customer service calls* between the training and test datasets.
  - For this variable, the training–test partition appears valid and representative.

TWO-SAMPLE Z-TEST FOR DIFFERENCE IN PROPORTIONS

- We could turn to the two-sample Z-test for the difference in proportions for a flag variable (0/1) across two independent samples.

Let

- x_1, x_2 = number of "successes" (value = 1)
- $p_1 = x_1/n_1$, $p_2 = x_2/n_2$ = observed proportions
- p = pooled proportion: $$p = \frac{x_1 + x_2}{n_1 + n_2}$$

- Test Statistic: $$Z_{\text{obs}} = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

TWO-SAMPLE Z-TEST FOR DIFFERENCE IN PROPORTIONS

- Example: Voice Mail Plan Membership –Churn Dataset
  - We compare the proportion of customers enrolled in the Voice Mail Plan between the training and test sets.

Given

- Training set: $x_1 = 707, \; n_1 = 2529$
- Test set: $x_2 = 215, \; n_2 = 804$

Compute observed sample proportions:

$$p_1 = \frac{x_1}{n_1} = \frac{707}{2529} = 0.279557 \quad \text{(rounded 0.2796)}$$

$$p_2 = \frac{x_2}{n_2} = \frac{215}{804} = 0.267413 \quad \text{(rounded 0.2674)}$$

TWO-SAMPLE Z-TEST FOR DIFFERENCE IN PROPORTIONS

- Pooled proportion

$$p_{\text{pooled}} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{707 + 215}{2529 + 804} = \frac{922}{3333} = 0.276628 \quad (\text{rounded } 0.2766)$$

- Hypotheses (two-tailed test):

$$H_0: \; \pi_1 = \pi_2 \qquad H_1: \; \pi_1 \neq \pi_2$$

- Test statistic (two-sample Z):

$$Z_{\text{obs}} = \frac{p_1 - p_2}{\sqrt{p_{\text{pooled}}(1 - p_{\text{pooled}})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

TWO-SAMPLE Z-TEST FOR DIFFERENCE IN PROPORTIONS

- Compute Components
  1. $p_1 - p_2 = 0.279557 - 0.267413 = 0.012144.$
  2. $1/n_1 + 1/n_2 = \frac{1}{2529} + \frac{1}{804} = 0.00039532 + 0.00124378 = 0.00163910.$
  3. $p_{\text{pooled}}(1 - p_{\text{pooled}}) = 0.276628 \times 0.723372 = 0.200176.$

$$Z_{\text{obs}} = \frac{0.012144}{0.018115} = 0.67054 \quad (\text{rounded } 0.6705)$$

- p-value: $p = 2 \cdot P(Z > |Z_{\text{obs}}|)$ $\quad 2 \times 0.251256 = 0.50251$
- Because the p-value: p ≈ 0.5025 ≫ 0.05, we **fail to reject H_0**.
- There is **no evidence** that the proportion of *Vmail_Plan* members differs between the training and test datasets. For this binary variable, the partition appears valid.

TEST FOR THE HOMOGENEITY OF PROPORTIONS

- When categorical data have more than two categories (**multinomial data**), we may need to verify whether two independent datasets have the same proportions across categories.
- Multinomial data : categorical variable can take k > 2 categories (Example- marital status - single/married/divorces)
- This type of question arises often in data mining when checking whether a training set and a test set are drawn from the same underlying population.
- The **Test for the Homogeneity of Proportions** allows us to determine whether the <u>multinomial proportions of two or more groups are equal</u>.

TEST FOR THE HOMOGENEITY OF PROPORTIONS

- **Example:**
  - A multinomial variable marital status with categories: Married/ Single/ Other
  - Suppose we have
    - a training set of 1000 people,
    - a test set of 250 people,
  - with the frequencies shown below.

Data Set	Married	Single	Other	Total
Training set	410	340	250	1000
Test set	95	85	70	250
Total	505	425	320	1250

TEST FOR THE HOMOGENEITY OF PROPORTIONS

- To determine whether significant differences exist between the multinomial proportions of the two data sets, we could turn to the test for the homogeneity of proportions
- This is a test of homogeneity.
- Hypotheses

$$H_0 : \begin{cases} p_{\text{married, training}} = p_{\text{married, test}} \\ p_{\text{single, training}} = p_{\text{single, test}} \\ p_{\text{other, training}} = p_{\text{other, test}} \end{cases}$$

$$H_a : \text{At least one equality in } H_0 \text{ is false.}$$

TEST FOR THE HOMOGENEITY OF PROPORTIONS

- **Expected Frequencies**: To determine whether the observed frequencies differ significantly, we compute the expected frequencies under the assumption that the overall proportions apply to both groups.
- General expected frequency formula: $E = \dfrac{(\text{row total})(\text{column total})}{\text{grand total}}$
- Example: $\text{Expected}_{\text{married, training}} = \dfrac{(1000)(505)}{1250} = 404$
- Applying this to all cells gives

Data Set	Married	Single	Other	Total
Training set	404	340	256	1000
Test set	101	85	64	250
Total	**505**	**425**	**320**	**1250**

TEST FOR THE HOMOGENEITY OF PROPORTIONS

- Chi-Square Test Statistic: Observed frequencies (O) are compared with Expected frequencies (E) using: $\chi^2_{\text{data}} = \sum \frac{(O - E)^2}{E}$
- Large deviations give a large chi-square value → small p-value → reject Ho

Cell	Observed Frequency	Expected Frequency	$(O - E)^2 / E$
Married, training	410	404	0.09
Married, test	95	101	0.36
Single, training	340	340	0
Single, test	85	85	0
Other, training	250	256	0.14
Other, test	70	64	0.56
			Sum = 1.15

TEST FOR THE HOMOGENEITY OF PROPORTIONS

- Degrees of Freedom: $df = (r-1)(c-1)$
- Here: rows = 2 (training, test) and columns = 3 (married, single, other)
- Thus, $df = (2-1)(3-1) = (1)(2) = 2$
- p-Value: $p\text{-value} = P(\chi^2 > 1.15) = 0.5627$

  (This matches the value from the chi-square distribution table.)

- **Conclusion**
  - Since **p = 0.5627 is large**, we **fail to reject H$_0$**.
  - There is **no evidence** that the marital-status proportions are significantly different between the training and test datasets.
  - Thus, **for this variable, the partition is valid**.

CHI-SQUARE TEST FOR GOODNESS OF FIT OF MULTINOMIAL DATA

- The chi-square goodness-of-fit test allows us to determine whether in multinomial data an observed sample follows a specified population distribution.
- **Example**
  - Suppose a multinomial variable *marital status* takes the values married, single, and other
  - suppose that we know that 40% of the individuals in the population are married, 35% are single, and 25% report another marital status.
  - We are taking a sample and would like to determine whether the sample is representative of the population

CHI-SQUARE TEST FOR GOODNESS OF FIT OF MULTINOMIAL DATA

- We are given the population proportions:

$$p_{\text{married}} = 0.40, \qquad p_{\text{single}} = 0.35, \qquad p_{\text{other}} = 0.25$$

- We draw a sample of n=100 (36 married, 35 single, 29 others)
- Hypotheses:

$$H_0 : p_{\text{married}} = 0.40, \quad p_{\text{single}} = 0.35, \quad p_{\text{other}} = 0.25$$

$$H_a : \text{At least one of the proportions in } H_0 \text{ is wrong.}$$

- Observed Frequencies:
  - The sample yields the following observed frequencies:

$$O_{\text{married}} = 36, \quad O_{\text{single}} = 35, \quad O_{\text{other}} = 29$$

CHI-SQUARE TEST FOR GOODNESS OF FIT OF MULTINOMIAL DATA

- Expected Frequencies
  - Under H_0, expected frequencies are: $E = n \times p$
- Test Statistic
  - The chi-square test statistic is: $\chi^2_{\text{data}} = \sum \frac{(O - E)^2}{E}$

$$E_{\text{married}} = 100 \times 0.40 = 40$$

$$E_{\text{single}} = 100 \times 0.35 = 35$$

$$E_{\text{other}} = 100 \times 0.25 = 25$$

Marital Status	Observed Frequency (O)	Expected Frequency (E)	(O − E)² / E
Married	36	40	0.4
Single	35	35	0
Other	29	25	0.64

$$\chi^2_{\text{data}} = 0.4 + 0 + 0.64 = 1.04$$

CHI-SQUARE TEST FOR GOODNESS OF FIT OF MULTINOMIAL DATA

- Degrees of Freedom:
  - For a multinomial goodness-of-fit test: $df = k - 1$

$$df = 3 - 1 = 2$$

- p-value: $p\text{-value} = P(\chi^2 > 1.04) = 0.5945$

  (using the chi-square distribution with 2 degrees of freedom)
  - Interpretation: The p-value is large (0.5945), meaning the observed sample frequencies do not significantly differ from what is expected under Ho
  - Conclusion: There is no evidence that the sample proportions differ from the population proportions.
  - Thus, the sample is representative of the population with respect to marital status.

ANALYSIS OF VARIANCE

- Analysis of Variance (ANOVA) is used when comparing the means of three or more groups to determine whether they come from populations with the same mean.
- ANOVA extends the two-sample t-test to multiple groups.
- We test whether the mean value of a continuous variable is identical across all subsets.

ANALYSIS OF VARIANCE

- **Example 1 — Groups A, B, and C**
  - We have samples from three groups (A, B, C), with four observations each.
  - The continuous variable measured is *age*.
  - Sample ages for Groups A, B, and C

Group A	Group B	Group C
30	25	25
40	30	30
50	40	50
60	55	45

Sample Means

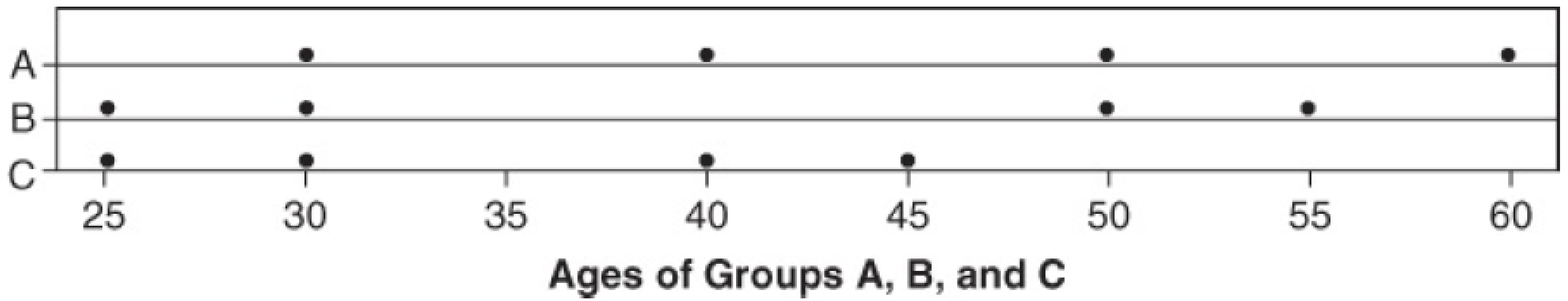
- $\bar{x}_A = 45$
- $\bar{x}_B = 40$
- $\bar{x}_C = 35$

ANALYSIS OF VARIANCE

- The Hypotheses are: $H_0 : \mu_A = \mu_B = \mu_C$

$$H_a : \text{At least one population mean differs}$$

- Dotplot Interpretation:
  - Considerable overlap among A, B, C
  - Despite differing means, the spread within each group is large
  - Conclusion: No evidence to reject H_0 in this example.



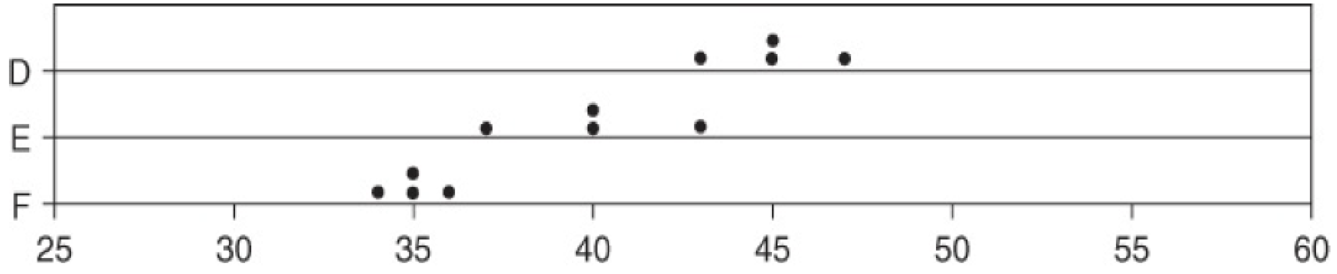
ANALYSIS OF VARIANCE

- **Example 2 — Groups D, E, and F**

Group D	Group E	Group F
43	37	34
45	40	35
50	45	36
47	46	35

Sample Means

- $\bar{x}_D = 45$
- $\bar{x}_E = 40$
- $\bar{x}_F = 35$



- **Interpretation:**
  - Very little overlap among D, E, F
  - Within-group spread is small
  - Suggests strong evidence against H_0

ANALYSIS OF VARIANCE

- Example-1 shows no evidence of differences in group means, while Example-2 shows clear differences, even though their sample means are the same.
- This contrast arises from how much the groups overlap, which is determined by the spread within each group.
- In Example-1, the spread within each group is large, causing the group means to look similar.
- In Example-2, the spread is small, so even the same differences in sample means appear large.

ANALYSIS OF VARIANCE

- **How ANOVA Works:**
  - ANOVA compares two different sources of variability:
  - (A) Between-Group Variability (Treatment Variation)
    - This measures how far each group mean is from the overall mean.
    - If group means differ substantially → between-group variance is large → groups likely come from different populations.
  - (B) Within-Group Variability (Error Variation)
    - This measures how much individual observations within each group vary around their group mean.
    - If within-group variation is small → samples are consistent inside groups.

ANALYSIS OF VARIANCE

- **The ANOVA Formulas**
  - ANOVA decomposes total variation in the data into:

    Total Variation = Between-Group Variation + Within-Group Variation

  - These are measured through **Sum of Squares**.
- Sum of Squares for Treatment (Between-Group)

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

Meaning of each term

- n_i: weight of each group
- $\bar{x}_i$: group mean
- $\bar{x}$: overall mean

k = number of groups
(e.g., 3 groups: A, B, C)
n_i = sample size in group i
(e.g., 4 observations each group)

  - Interpretation:
  - If groups differ strongly, their means will be far from the grand mean → SSTR increases.

ANALYSIS OF VARIANCE

- Sum of Squares for Error (Within-Group)

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

Meaning of each term

- $n_i - 1$: degrees of freedom within each group
- s_i^2: variance inside group i

  - Interpretation: If samples within each group are very spread out → SSE becomes large.

- Total Sum of Squares:

SST = SSTR + SSE

ANALYSIS OF VARIANCE

- **Degrees of Freedom**
- 1. Degrees of freedom for treatment

$$df_1 = k - 1$$

- 2. Degrees of freedom for error

$$df_2 = N - k$$

- 3. Degrees of freedom total

$$df_{total} = N - 1$$

k = number of groups

(e.g., 3 groups: A, B, C)

n_i = sample size in group i

(e.g., 4 observations each group)

$N = \sum n_i$ = total sample size

(for 3 groups × 4 observations = 12)

ANALYSIS OF VARIANCE

- Mean Squares: MSTR and MSE
- Mean Square Treatment: $\text{MSTR} = \frac{\text{SSTR}}{df_1}$
  - This is an estimator of variance between group
- Mean Square Error: $\text{MSE} = \frac{\text{SSE}}{df_2}$
  - It is the average within-group variance.
- ANOVA F-Statistic:

$$F_{\text{data}} = \frac{\text{MSTR}}{\text{MSE}}$$

Interpretation

- If **MSTR >> MSE**, then
  → F becomes large
  → group means differ significantly
  → reject H_0
- If **MSTR ≈ MSE**, then
  → F small
  → group means not different
  → fail to reject H_0

ANALYSIS OF VARIANCE

- Example -1 ANOVA Results (Groups A, B, C)
  - ANOVA Results for $H_0 : \mu_A = \mu_B = \mu_C$

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square	F	P-value
Treatment	2	200	100	0.64	0.548
Error	9	1400	156		

Test statistic: $F_{\text{obs}} = 0.64$.

$\text{p-value} = P(F_{df_1, df_2} > F_{\text{obs}})$

  - Because $p \approx 0.55$ (much larger than typical $\alpha = 0.05$, we fail to reject H_0)
  - There is no evidence that the three population means differ — this matches the earlier dotplot conclusion (considerable overlap among groups).

ANALYSIS OF VARIANCE

- Example -2 ANOVA Results (Groups A, B, C)
  - ANOVA Results for: $H_0: \mu_D = \mu_E = \mu_F$

Source of Variation	Degree of Freedom	Sum of Squares	Mean Square	F	p
Treatment	2	200.00	100.00	32.14	0.000
Error	9	28.00	3.11		

  - Because $p \approx 0.00008 \ll 0.05$, we strongly reject H_0 .
  - There is very strong evidence that not all population means are equal — consistent with the dotplot showing little overlap across groups D, E, and F.

ANALYSIS OF VARIANCE

- Quick recap of the procedure
  - Compute **SSTR** and **SSE** (sum of squares for treatment and error).
  - Compute **MSTR = SSTR / (k-1)** and **MSE = SSE / (N-k)**.
  - Compute **F= MSTR / MSE**.
  - Compute **p-value**
  - If $p < \alpha$ (e.g., 0.05), reject H_0