Analysis of Variance (ANOVA)

Sessions 17-18

Application

An experiment is designed to compare five different add-on product offerings. A sample of 30 outlets, taken from a larger group, is randomly assigned to the five advertisements (so that there are 6 outlets to each advertisement). The *average daily sales of the outlets in the week for which the advertisements were put-up* for the 30 outlets, are as follows:

At the 0.05 level of significance, is there evidence of a difference in the mean sales following exposure to five advertisements?

Terminology

- **Experimental Design** is a plan and a structure to test hypotheses in which the researcher either controls or manipulates one or more variables. It contains independent and dependent variables
- **Independent variable (Factors)** may be either a treatment variable or a classification variable.
- **Treatment variable** is a variable the experimenter controls or modifies in the experiment.
- **Classification variable** is some characteristic of the experimental subject that was present prior to the experiment and is not a result of the experimenter's manipulations or control
- **Levels**, or classifications, of independent variables are the subcategories of the independent variable used by the researcher in the experimental design.
- **Dependent variable** is the response to the different levels of the independent variables

ANOVA

- Analysis of Variance (ANOVA) allows statistical comparison across many groups of data
- Is advertisement a *'factor'* in determining sales?
- In this example we have 5 levels of the factor (A, B, …, E)

$$H_0: \mu_a = \mu_b = \mu_c = \mu_d = \mu_e$$

$$H1: \mu_a \neq \mu_b \neq \mu_c \neq \mu_d \neq \mu_e$$

Completely Randomized Design

Subjects are assigned randomly to treatments

- Study of tire-quality
  - Treatment levels: low, medium, and high quality.
  - Dependent variable: number of miles driven before the tread fails state inspection.
- Sales volumes for Walmart stores
  - Treatment levels: inner-city stores, suburban stores, stores in medium-sized cities, and stores in small towns.
  - Dependent variable: Sales dollars.

Application

Suppose a manufacturing organization produces a valve that is specified to have an opening of 6.37 cm. Quality controllers within the company might decide to test to determine how the openings for produced valves vary among four different machines on three different shifts.

Independent variables: (i) Type of machine (ii) Work shift.

Levels

Type of machine: 4

Shift: 3

Application

In the valve example, suppose only Machine Type is relevant. The data on 24 randomly sampled valves from 4 machines is given below. Is there a significant difference between valves produced by 4 machines.

Method 1:

Compare using 2-sample t-test

Number of t-tests: ${}_4C_2 = \frac{4!}{(4-2)!2!} = 6$

Method 2:

Use ANOVA

Machine Type			
1	**2**	**3**	**4**
6.33	6.26	6.44	6.29
6.26	6.36	6.38	6.23
6.31	6.23	6.58	6.19
6.29	6.27	6.54	6.21
6.40	6.19	6.56	
	6.50	6.34	
	6.19	6.58	
	6.22		

ANOVA

- $H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_j$

  H_1 : At least one of the means is different from the others.

- Partitioning the total variance of the data into the following two variances.
  - The variance resulting from the treatment (columns)
  - The error variance, or that portion of the total variance unexplained by the treatment

$$SST = SSC + SSE$$

$$\sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{C} n_j(\bar{x_j} - \bar{x})^2 + \sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_{ij} - \bar{x_j})^2$$

SST = Total Sum of Squares

SSC = Sum of Squares Columns (Across / Treatment / Explained)

SSE = Sum of Squares Error (Within)

ANOVA

$$SST = SSC + SSE$$

$$\sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})^2 = \sum_{j=1}^{C} n_j(\bar{x_j} - \bar{x})^2 + \sum_{j=1}^{C}\sum_{i=1}^{n_j}(x_{ij} - \bar{x_j})^2$$

- i = particular member of treatment level
- j = treatment level
- C = number of treatment levels
- n_j = number of observations in the j^{th} treatment level
- $\bar{x}$ = grand mean
- $\bar{x}_j$ = mean of j^{th} treatment group or level
- x_{ij} = individual value

ANOVA Procedure

- Step 1: Grand mean = Sum of all values / Total Sample size = $\bar{x}$
- Step 2: Calculate $\bar{x}_j$ (individual group means)
- Step 2: Sum of squares total (SST):$\Sigma\Sigma$ (x_{ij} - $\bar{x}$)^2 for all observations
- Step 3: Sum of squares among groups (SSC): Σ ($\bar{x}_j$ - $\bar{x}$)^2 x number of observations in the group (n_j)
- Step 4: Sum of squares within groups (SSE): Σ Σ (x_{ij} - $\bar{x}_j$)^2

ANOVA Procedure

- Step 5: MSC = $\frac{SSC}{(C-1)}$

- Step 6: MSE = $\frac{SSE}{(N-C)}$

  where N = Total number of observations

- Step 7: F_{STAT} = $\frac{MSC}{MS}$

- Compare this with critical value or check the p-value *(Remember: F-stat is associated with numerator and denominator degrees of freedom)*

- If p-value less than the level of significance → reject H_0

ANOVA Excel Output

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	5	31.59	6.318	0.00277
Column 2	8	50.22	6.2775	0.01107857
Column 3	7	45.42	6.48857143	0.01011429
Column 4	4	24.92	6.23	0.00186667

ANOVA

Source of Variati	SS	df	MS	F	P-value	F crit
Between Gr	0.23658012	3	0.07886004	10.1810252	0.00027858	3.09839121
Within Grou	0.15491571	20	0.00774579			
Total	0.39149583	23