Analysis of Variance (ANOVA)

Sessions 19-20

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: $F_{STAT} = \frac{MSC}{MSE}$
- 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

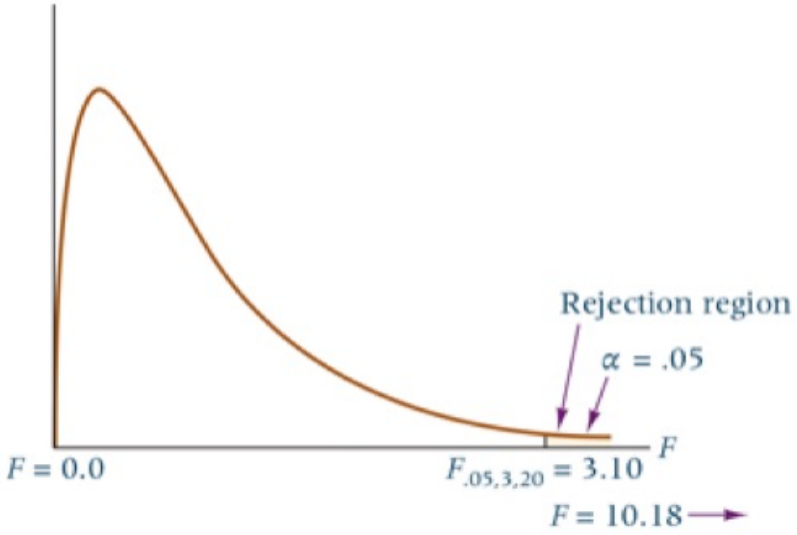
Anova: Single Factor

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 Variat	*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				



Application

A company has three manufacturing plants, and company officials want to determine whether there is a difference in the average age of workers at the three locations. The following data are the ages of five randomly selected workers at each plant. Perform a one-way ANOVA to determine whether there is a significant difference in the mean ages of the workers at the three plants. Use $\alpha = .01$ and note that the sample sizes are equal.

Employee Ages		
Plant 1	**Plant 2**	**Plant 3**
29	32	25
27	33	24
30	31	24
27	34	25
28	30	26

ANOVA and t-test

- Assume: α= 0.05
- For one 2 sample t-test: Probability of Type I error = (1 - α) = 0.95
- For two tests: Probability of Type I error = 0.95 x 0.95 = 0.9025
- Probability of Type I error increases as the combinations of t-tests ${}_nC_x$ increases

ANOVA

Which pair of groups are significantly different?

Tukey-Kramer Procedure:

- Compute:

$$q_{\alpha,C,N-C}\sqrt{\frac{MSE}{n}}$$

- Compare this with absolute difference in means of each pair of groups eg. $|\bar{x}_1 - \bar{x}_2|$
- If the absolute difference in means is greater than critical value – difference between groups is significant
- In case of groups of different sizes

$$q_{\alpha,C,N-C}\sqrt{\frac{MSE}{2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

ANOVA Experiment Designs

- **Completely Randomized Design (One-Way ANOVA):** Experiment with one factor
- **Randomized Block Design:** When the researcher controls for one more variable – which is not the treatment variable – but is likely to impact the relationship between Dependent and independent variables.

$$SST = SSC + SSR + SSE$$

SST = Total Sum of Squares

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

SSR = Sum of Squares Rows (Blocking)

SSE = Sum of Squares Error (Within)

Randomized Block Design

- For Treatment effects:

  H_0: $\mu_{.1} = \mu_{.2} = \mu_{.3} = \ldots = \mu_{.C}$

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

- For the blocking effects, they are

  H_0: $\mu_{1.} = \mu_{2.} = \mu_{3.} = \ldots = \mu_{R.}$

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

Randomized Block Design

$$\mathrm{SST} = \sum_{j=1}^{C}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}\right)^2$$

$$\mathrm{SSC} = n\sum_{j=1}^{C}\left(\bar{x}_j - \bar{x}\right)^2$$

$$\mathrm{SSR} = C\sum_{i=1}^{n}\left(\bar{x}_i - \bar{x}\right)^2$$

$$\mathrm{SSE} = \sum_{j=1}^{C}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j - \bar{x}_i + \bar{x}\right)^2$$

N = Total number of observations

C = number of treatment levels

n = number of observations in each treatment level (number of blocks

Randomized Block Design

$$\mathrm{MSC} = \frac{\mathrm{SSC}}{C-1}$$

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{n-1}$$

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{N-n-C+1}$$

$$F_{\text{treatments}} = \frac{\mathrm{MSC}}{\mathrm{MSE}} \qquad F_{\text{blocks}} = \frac{\mathrm{MSR}}{\mathrm{MSE}}$$

Check the table value of F_{stat} at appropriate degrees of freedom

If $F_{calc} < F_{stat}$ → Fail to reject H_0

Application

A tire company has developed a new tire. The company conducted tread-wear tests on the tire to determine whether there is a significant difference in tread wear if the average speed with which the automobile is driven varies.

The company set up an experiment in which the independent variable was speed of automobile. There were three treatment levels: slow speed (car is driven 20 miles per hour), medium speed (car is driven 40 miles per hour), and high speed (car is driven 60 miles per hour).

Company analysts realized that several possible variables could confound the study. One of these variables was supplier. The company uses five suppliers to provide a major component of the rubber from which the tires are made. To control for this variable experimentally, the analysts used supplier as a blocking variable.

Fifteen tires were randomly selected for the study, three from each supplier. Each of the three was assigned to be tested under a different speed condition. The data are given in the excel file. These figures represent tire wear in units of 10,000 miles.

Application

Suppose a national travel association studied the cost of premium unleaded gasoline in the United States during the summer of 2019. From experience, association directors believed there was a significant difference in the average cost of a gallon of premium gasoline among urban areas in different parts of the country.

To test this belief, they placed random calls to gasoline stations in five different cities.

In addition, the analysts realized that the brand of gasoline might make a difference.

They were mostly interested in the differences between cities, so they made city their treatment variable. To control for the fact that pricing varies with brand, the analysts included brand as a blocking variable and selected six different brands to participate.

The analysts randomly telephoned one gasoline station for each brand in each city, resulting in 30 measurements (five cities and six brands). Each station operator was asked to report the current cost of a gallon of premium unleaded gasoline at that station. The data are given in the excel file. Determine whether there is a significant difference in the average cost of premium unleaded gasoline by city. Let $\alpha = .01$.

(Check one-way and two-way ANOVA without replication)

Factorial Design - Two-Way ANOVA

- More than one treatment
- Every Level of each treatment is studied under every level of all other treatments
- **Interaction** occurs when the effects of one treatment vary according to the levels of treatment of the other effect. Eg. Education and Socio-economic Status of Family
- Excel treats Randomized Block Design as Two-Way ANOVA with single observation per cell and Two-Way Factorial Design as Two-Way ANOVA with replication

Application

Recall:

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

Two-Way ANOVA

Hypothesis

Row effects:

H_0 : Row means are all equal.

H_1 : At least one row mean is different from the others.

Column effects:

H_0 : Column means are all equal.

H_1 : At least one column mean is different from the others.

Interaction effects:

H_0 : The interaction effects are zero.

H_1 : An interaction effect is present.

Two-Way ANOVA

$$\text{SST} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}\left(x_{ijk} - \bar{x}\right)^2$$

$$\text{SSC} = nR\sum_{j=1}^{C}\left(\bar{x}_j - \bar{x}\right)^2 \qquad \text{MSC} = \frac{\text{SSC}}{C-1}$$

$$\text{SSR} = nC\sum_{i=1}^{R}\left(\bar{x}_i - \bar{x}\right)^2 \qquad \text{MSR} = \frac{\text{SSR}}{R-1}$$

$$\text{SSI} = n\sum_{i=1}^{R}\sum_{j=1}^{C}\left(\bar{x}_{ij} - \bar{x}_i - \bar{x}_j + \bar{x}\right)^2 \qquad \text{MSI} = \frac{\text{SSI}}{(R-1)(C-1)}$$

$$\text{SSE} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{k=1}^{n}\left(x_{ijk} - \bar{x}_{ij}\right)^2 \qquad \text{MSE} = \frac{\text{SSE}}{RC(n-1)}$$

$$F_R = \frac{\text{MSR}}{\text{MSE}} \qquad F_C = \frac{\text{MSC}}{\text{MSE}} \qquad F_I = \frac{\text{MSI}}{\text{MSE}}$$

Two-Way ANOVA

- N = Total number of observations
- C = number of first treatment levels (Columns)
- R = number of second treatment levels (Rows)
- n = number of observations in each cell (LT_1 x LT_2)
- j = columns treatment level
- i = row treatment level
- k = cell member
- x_{ijk} = k^{th} observation in the j^{th} level for treatment 1 and i^{th} level of treatment 2

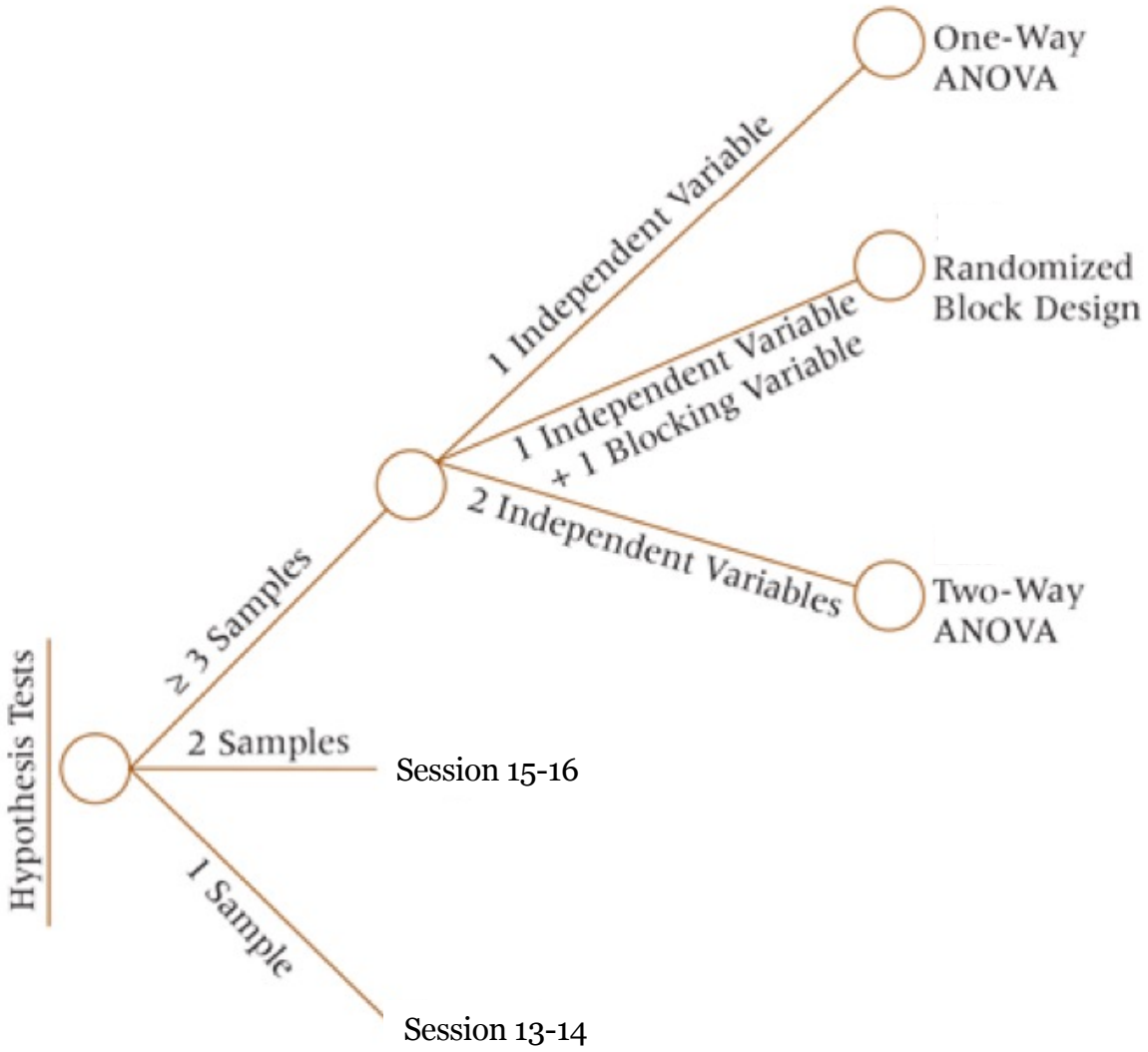
Application

A shoe retailer conducted a study to determine whether there is a difference in the number of pairs of shoes sold per day by stores according to the number of competitors within a 1-mile radius and the location of the store. The company researchers selected three types of stores for consideration in the study: stand-alone suburban stores, mall stores, and downtown stores. These stores vary in the numbers of competing stores within a 1-mile radius, which have been reduced to four categories: 0 competitors, 1 competitor, 2 competitors, and 3 or more competitors. Suppose the following data represent the number of pairs of shoes sold per day for each of these types of stores with the given number of competitors. Use $\alpha = .05$ and a two-way ANOVA to analyze the data.

Application

Consider the valve opening data example discussed earlier. Suppose the data represent valves produced on four different machines on three different shifts and that the quality controllers want to know whether there is any difference in the mean measurements of valve openings by shift or by machine. The data are given here, organized by machine and shift. What are the hypotheses for this problem? Discuss the results obtained. What conclusions might the quality controllers reach from this analysis?

Test Map


Hypothesis Tests
2 Samples
1 Sample
≥ 3 Samples
1 Independent Variable
1 Independent Variable + 1 Blocking Variable
2 Independent Variables
One-Way ANOVA
Randomized Block Design
Two-Way ANOVA
Session 15-16
Session 13-14