Department of AI & DS
CSE and CS&IT

COURSE NAME: PROBABILITY, STATISTICS AND QUEUING THEORY

COURSE CODE: 23MT2005

Topic

Central Limit Theorem and its applications, ANOVA One way and Two way

Session - 19

AIM OF THE SESSION

To familiarize students with the concept of Central limit theorem and ANOVA

INSTRUCTIONAL OBJECTIVES

This Session is designed to:

1. Define Central limit Theorem and its applications
2. List out ANOVA and its assumptions
3. Apply the procedure of ANOVA for one-way classified data
4. Apply the procedure of ANOVA for two-way classified data

LEARNING OUTCOMES

At the end of this session, you should be able to:

1. Define Central limit theorem and ANOVA
2. Describe the procedure of ANOVA I way and II way

Central Limit Theorem

- If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ, the sample means, $\bar{x}$, are approximately normally distributed for sufficiently large sample sizes ($n \geq 30$) regardless of the shape of the population distribution.

- If the population is normally distributed, the sample means are normally distributed for any size sample.

- The Central Limit Theorem (CLT) is a fundamental concept in statistics. It states that, given a sufficiently large sample size from a population with any shape of distribution, the distribution of the sample mean will approach a normal distribution (bell-shaped curve), regardless of the population's original distribution.

Key points for the Central Limit Theorem

Sample Size:

The theorem applies when the sample size is large enough. A common rule of thumb is a sample size of at least 30.

Population Distribution:

The population from which samples are drawn can have any shape (e.g., skewed, uniform, or even bimodal).

As sample size increases, the sample mean distribution becomes normal.

Mean and Standard Deviation:

- The mean of the sampling distribution will be equal to the population mean μ
- The standard deviation of the sampling distribution (called the standard error) is given by:

$$\sigma_{mean} = \frac{\sigma}{\sqrt{n}}$$

- σ is the population standard deviation and *n*
- n is the sample size.

Example

A shipping company handles packages with weights that are normally distributed with a mean of 10 kg and a standard deviation of 2 kg. If a random sample of 25 packages is selected: What is the expected mean and standard deviation of the sample mean?

Solution:

Population mean, μ=10 kg
Population standard deviation, σ=2 kg
Sample size, n=25
Expected Mean and Standard Deviation of the Sample Mean:

The Central Limit Theorem states that for a large enough sample size, the sampling distribution of the sample mean will be approximately normal, with:

Mean of the sample mean:
Standard deviation of the sample mean (standard error):

Thus, the sample mean has an expected mean of 10 kg and a standard deviation of 0.4 kg.

Applications of Central Limit Theorem (CLT)

The Central Limit Theorem is a fundamental concept in statistics with broad applications in various fields. Here are some key applications:

1. **Hypothesis Testing**

Purpose: To test claims or hypotheses about a population parameter.

Application: The sampling distribution of the test statistic (e.g., sample mean) is assumed to be normal under the null hypothesis, thanks to the CLT.

Example: Testing if a new medication has a different effect from the standard treatment using sample data.

2.Election Polling

Purpose: To predict election outcomes based on sample polls.

Application: Pollsters rely on the CLT to assume that the sample mean (or proportion of votes) will be normally distributed, enabling them to calculate margins of error.

Example: Predicting the percentage of votes each candidate will receive.

ANOVA

The Analysis of variance (ANOVA) is a powerful statistical tool for tests of significance. The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. When we have three or more samples to consider at a time an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population, i.e, they have the same mean. The basic purpose of the analysis of variance is to test the homogeneity of several means.

The term ‘Analysis of Variance’ was introduced by Prof. R. A. Fisher in 1920’s to deal with problem in the analysis of agronomical data.

The total variation in any set of numerical data is due to a number of causes which may be classified as: (i) Assignable causes and (ii) Chance causes.

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

ANOVA

Definition of ANOVA: According to Prof. R. A. Fisher, Analysis of Variance is the “Separation of variance ascribable to one group of causes from the variance ascribable to other group”.

Assumption of ANOVA test: ANOVA test is based on the test statistic F (or Variance Ratio).

For the validity of the F-test in ANOVA, the following assumptions are made:

(i) The observations are independent.
(ii) Parent population from which observations are taken is normal, and
(iii) Various treatment and environmental effects are additive in nature

ANOVA is a method of dividing the variation observed in experimental data into different parts, each part assignable to a known source, cause or factor therefore

$$F = \frac{\text{Variance between the groups}}{\text{Variance within the groups}} \qquad F = \frac{\sigma^2 \text{between the groups}}{\sigma^2 \text{within the groups}}$$

In which σ^2 is the population variance and it never be negative.

The Procedure of Analysis of Variance:

Before we discuss the procedure of analysis of variance , it is to be noted here that when we have taken a large group or a finite population, to represent its total units the symbol ‘N’ will be used.

When the large group is divided into two or more than two sub groups having equal number of units, the symbol ‘n’ will be used and for number of groups the symbol ‘k’ will be used.

Here we apply one way analysis of variance and two way analysis of variance.

One Way Analysis of Variance: In this case, only one variate under study.

$$\text{Model: } Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

Where ϵ_{ij} measures the deviation of the jth observation of the ith sample from the corresponding treatment mean. The ϵ_{ij} term represents random error and plays the same role as the error terms in the regression model, μ is the grand mean of all the μ’s

ANOVA

$$\mu = \frac{1}{k}\sum_{i=1}^{k} \mu_i$$

and α_i is called the effect of the ith treatment.

The null hypothesis that the k population means are equal against the alternative that at least two of the means are unequal may now be replaced by equivalent hypothesis.

$H_0: \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$

H_1: Atleast one of the $\alpha_i's$ is not equal to zero.

Step 1: Set up null hypothesis

Step 2: Set up alternative hypothesis

Step 3: Obtain all the sum of raw scores and the squares of raw scores. Write them at the end of each column.

Step 4: Obtain grand sums of raw scores as and square of raw square as $\sum x^2$

Step 5: Calculate the correction factor (c.f.) by using the formula

c.f.= $\frac{G^2}{N}$; Where G= Grand total of all the observations

ANOVA

Step 6: Calculate the sum of squares due to the treatments (S. S. A) or sum of squares among the groups by using the formula

S. S. A. = $(\sum x^2/N)$- c.f. (or) $\text{S.S.A.}=\frac{\left(\sum x_1^2\right)^2}{n_1}+\frac{\left(\sum x_2^2\right)^2}{n_2}+\frac{\left(\sum x_3^2\right)^2}{n_3}+\cdots+\frac{\left(\sum x_k^2\right)^2}{n_k}-\text{c. f.}$

Step 7: Calculate the total sum of squares (S. S. T.)

S.S. T. = $\sum\sum x_{ij}^2$-c.f.

Step 8: Calculate the sum of square with in the groups or Error sum of square

S. S. E. = S. S. T- S.S.A

Step 10: Calculate the degrees of freedom as

Degrees of freedom for S.S.A = k-1(where k is the number of groups)

Degrees of freedom for Error S. S. E. =N-k (where N is the total number in the group)

Degrees of freedom for total S. S. T= N-1

Step 11: Find the value of Mean sum of squares of two variances as

ANOVA

$$\text{Mean sum of square between the group M.S.S.A} = \frac{\text{S.S.A}}{k-1}$$

$$\text{Mean sum of square within the group M.S.S.E} = \frac{\text{S.S.E}}{N-k}$$

Step 12: Prepare ANOVA table

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Variance ration	
				F-calculated value	F-tabulated value
Between the groups	k-1	S.S.A	$M.S.S.A=F_1$	$M.S.S.A/M.S.S.E=F_1/F_2$	$F_1/F_2 \sim F_{(k-1,\ N-k)}$
Within the groups	N-k	S.S.E.	$M.S.S.E=F_2$	-	-
Total	N-1	S.S.T	-	-	-

Step 13: State your conclusion

If F calculated value is less than F-table value then accepts the null hypothesis at ‘α%’ level of significance. Otherwise we reject the null hypothesis.

ANOVA TWO WAY CLASSIFICATION:

ANOVA two classification identifies two factors: treatments and blocks –both of which affect the response.

Model: $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$

Where ϵ_{ij} measures the deviation of the jth observation of the ith sample from the corresponding treatment mean. The ϵ_{ij} term represents random error and plays the same role as the error terms in the regression model, μ is the grand mean of all the μ's, α_i is called the effect of the ith block, β_j is called the effect of the jth treatment.

H_{01}: The null hypothesis that there is no difference between the blocks.

H_{02}: There is no significant difference between the treatments.

H_{01}: $\alpha_1 = \alpha_2 = \cdots = \alpha_b = 0$; H_{02}: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: Atleast one of the $\alpha_i' s$ is not equal to zero.

Step 1: Set up null hypothesis

Step 2: Set up alternative hypothesis

Step 3: Obtain all the sum of raw scores and the squares of raw scores. Write them at the end of each column.

Step 4: Obtain grand sums of raw scores as and square of raw square as $\sum x^2$

ANOVA

Step 5: Calculate the correction factor (c.f.) by using the formula

c.f.= $\frac{G^2}{N}$; Where G= Grand total of all the observations

Step 6: Calculate the sum of squares due to the treatments (S. S. T) or sum of squares among the groups by using the formula

Sum of squares due to treatments (Tr.S.S.). = $(\sum T_i{}^2/b)$- c.f.

Step 7: Calculate the sum of squares due to the blocks (S. S. B) or sum of squares among the groups by using the formula

Sum of squares due to treatments (SSB.). = $(\sum B_j{}^2/k)$- c.f.

Step 8: Calculate the total sum of squares (S. S. T.)

S.S. T. = $\sum\sum x_{ij}^2$-c.f.

Step 9: Calculate the sum of square with in the groups or Error sum of square

S. S. E. = Total sum of squares-SSB-SST

ANOVA

Step 10: Prepare ANOVA table

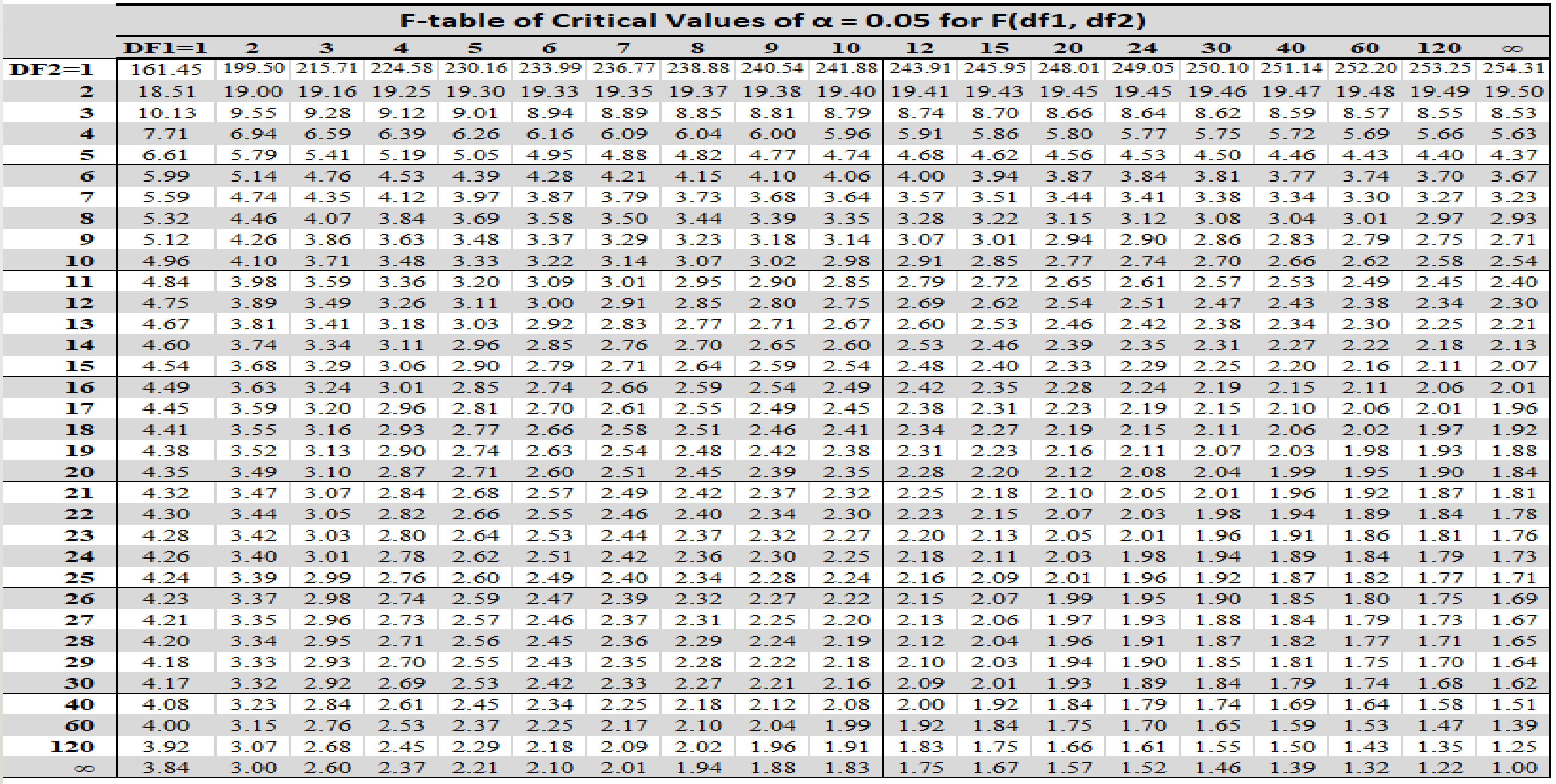
Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Variance ration	
				F-calculated value	F-tabulated value
Treatments	k-1	S.S.T	MST=SST/(k-1)	$M.S.T/M.S.E=F_1/F_3$	$F_1/F_3 \sim F_{(k-1,\ (b-1)(k-1))}$
Blocks	b-1	S.S.B	MSB=SSB/(b-1)	$M.\ S.B/M.S.E=F_2/F_3$	$F_2/F_3 \sim F_{(b-1,\ (b-1)(k-1))}$
Error	(b-1)(k-1)	SSE	MSE=SSE/((b-1)(k-1))		-
Total	bk-1	Total SS	-	-	-

Step 11: State your conclusion

If F calculated value is less than F-table value then accepts the null hypothesis at ‘α%’ level of significance. Otherwise we reject the null hypothesis.

Error=totaldf-rowdf-columndf=bk-1-(b-1)-(k-1)=bk-1-b+1-k+1=bk-b-k+1=b(k-1)-1(k-1)

ACTIVITIES/ CASE STUDIES/ IMPORTANT FACTS RELATED TO THE SESSION



F-table of Critical Values of α = 0.05 for F(df1, df2)

	DF1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
DF2=1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91	245.95	248.01	249.05	250.10	251.14	252.20	253.25	254.31
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.15	2.07	1.99	1.95	1.90	1.85	1.80	1.75	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.13	2.06	1.97	1.93	1.88	1.84	1.79	1.73	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.12	2.04	1.96	1.91	1.87	1.82	1.77	1.71	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.10	2.03	1.94	1.90	1.85	1.81	1.75	1.70	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39	1.32	1.22	1.00

Example 1

Example: A researcher wanted to analyze whether the mean housing prices are the same regardless of the three crowded places where they are located. A random samples of the prices at which the houses are being purchased at all the three areas are given below:

Observation	Low	Moderate	High
1	120	61	40
2	68	59	55
3	40	110	73
4	95	75	45
5	83	80	64

Calculate H and report your results with the p-value of 0.05

A) Here we analyze the mean housing prices are the same regardless of the three crowded places where they are located. This can be solved by using the ANOVA one-way classification.

Step 1: We set up the null hypothesis.

H_0: There is no significant difference between the mean housing prices of three crowded places.

Example 1

Step 2: H_1: There is a significant difference between the mean housing prices of three crowded places.

Observation	Low	Moderate	High	
1	120	61	40	
2	68	59	55	
3	40	110	73	
4	95	75	45	
5	83	80	64	
Total T_i	T_1=406	T_2=385	T_3=277	G=1068
T_i^2	T_1^2=164836	T_2^2 =148225	T_3^2= 76729	

Step 3: Calculate the correction factor (c.f.)

$$\text{c.f.} = (1068)^2/15=76041.6$$

Example 1

Step 4: We calculate the raw sum of squares

Raw sum of squares = $(120)^2+ (68)^2+(40)^2+(95)^2+(83)^2+(61)^2+(59)^2+(110)^2+(75)^2+$

$(80)^2+ (40)^2+(55)^2+(73)^2+(45)^2+(64)^2=83940$.

Step 5: Calculate the total sum of squares

Total sum of squares (S. S. T) = 83940-76041.6=7898.4

Step 6: Calculate the treatment sum of square (S. S.A.) or between the groups

S. S. A. = (164836/5)+(148225/5)+(76729/5)-76041.6=1916.4

Step 7: Calculate the Error sum of squares

E. S. S. = S.S.T. – S.S.A.=7898.4 - 1916.4=5982

Example 1

Step 8: Calculate the degrees of freedom

Degrees of freedom for S.S.A = k-1(where k is the number of groups)=3-1=2

Degrees of freedom for Error S. S. E. =N-k (where N is the total no. in the group) =14-2=12

Degrees of freedom for total S. S. T= N-1=15-1=14

Step 9: Find the value of Mean sum of squares of two variances as

Mean sum of square between the group M.S.S.A = $\frac{\text{S.S.A}}{\text{k}-1}=\frac{1916.4}{2}$

Mean sum of square within the group M.S.S.E = $\frac{\text{S.S.E}}{\text{N}-\text{k}}$=5982/12

Step 10: Variance ratio

F_1= M.S.S.A/ M.S.S.T

Example 1

Step 11: Prepare ANOVA table

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Variance ratio	
				F-calculated value	F-tabulated value
Between the groups (columns)	3-1=2	1916.4	=1916.4/2= 958.2	1.92	3.88
Within the groups (Error)	14-2= 12	5982	5982/12=498 .5		
Total	15-1=14	7898.4	-	-	-

Step 12: Conclusion F-Calculated value is <a. Therefore it is not significant and we fail to reject H_0 5% level of significance . Hence there is no significant difference between the mean housing prices of three crowded places.

SUMMARY

In this session, ANOVA and its assumptions have described

1. Central limit theorem and its applications
2. ANOVA one way classification
3. ANOVA two way classification

SELF-ASSESSMENT QUESTIONS

Which of the following is Not correct about ANOVA?

A) F Ratio is computed in ANOVA
B) It is a ratio between groups variance and within groups variance
C) The between groups variance is representing as sampling error in the distributions
D) F Ratio is name after Sir Ronald Fisher

Which of the following best describes the purpose of using ANOVA in research?

A) ANOVA is used to compare the means of two groups
B) ANOVA is used to compare the means of more than two groups
C) ANOVA is used to determine the correlation between two variables
D) ANOVA is used to determine the interaction effect between dependent variables

TERMINAL QUESTIONS

1. Explain ANOVA classified data with one observation per cell

2. Explain ANOVA two way classified data with one observation per cell

3. Define ANOVA and also write its assumptions

4. In an experiment to determine the effect of nutrition on the attention spans of elementary school students, a group of 15 students were randomly assigned to each of three meal plans: no breakfast, light breakfast, and full breakfast. Their attention spans (in minutes) were recorded during a morning reading period and are shown in the following table. Construct the analysis of variance table for this experiment.

Attention Spans of students after three meal plans

No breakfast	Light breakfast	Full breakfast
8	14	10
7	16	12
9	12	16
13	17	15
10	11	12

REFERENCES FOR FURTHER LEARNING OF THE SESSION

Reference Books:

1. William Feller, An Introduction to Probability Theory and Its Applications: Volime 1, Third Edition, 1968 by John Wiley & Sons,Inc.
2. Alex Tsun, Probability & Statistics with Applications to Computing (Available at: http://www.alextsun.com/files/Prob_Stat_for_CS_Book.pdf)

3. Richard A Johnson, Miller& Freund's Probability and statistics for Engineers, PHI, New Delhi, 11th Edition (2011).

Sites and Web links:

1. https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/more-significance-testing-videos/v/small-sample-hypothesis-test

THANK YOU

Team – PSQT EVEN SEMESTER 2024-25