L13.2 ANOVA

2020年4月30日 星期四 下午6:00

Lecture 13_ch12...

4/26/2020

Math 661
Applied Statistics
Lecture Note 2
Chapter 12

Instructor: Padma Nalanjan

Introduction to the Practice of Statistics
Moore-McCabe, and Craig

1

Chapter 12
ANOVA

- The one-way analysis of variance is used to test the equality of population means from several populations to using one sample.
- Analysis of variance is usually abbreviated **ANOVA**.
- Hypotheses:
  $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$ (all population means are equal)
  H_1: means are not all equal (or at least one pair is not equal).

test equality of population means

3 pop.

3

ANOVA

We look at the following:

- the variance **between** samples (or treatments/groups)
  - the spread between sample means in the sample between group variance or treatment variance
- the variance **within** samples (or treatment/group)
  - the spread between observations within each observation in the sample and the sample means,

and draw conclusions from the results:

4

ANOVA

- The between samples sum of squares is computed by summing the squared differences between each sample mean and the overall mean.
- The within sample sum of squares is computed by summing the squared differences between each observation in the sample and the sample mean.

5

ANOVA

Necessary Conditions

- All k (k ≥ 3) sets of observations are random.
- All k populations are independent, normal distribution or approximately normal.
- All k populations are approximately normal (or approximately normal by CLT).
- All k samples are independent.
- All k population variances are equal (or can be assumed to be) $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

6

Test Statistic

$$F = \frac{\text{Variance between samples (Mean square between)}}{\text{Variance within samples (Mean square within)}}$$

where F_{df_1, df_2} (the number of treatments/groups and the the total sample size.

$df_1 = k - 1$
$df_2 = N - k$

$$MSE = \frac{\text{Variance between samples (Mean square between)}}{\text{Variance within samples (Mean square within)}}$$

1

4/26/2020

F-distribution

1. The F distribution is an asymmetric distribution that takes on values greater than or equal to zero
2. The F distribution has two numbers of degrees of freedom, df_1 for the numerator and df_2 for the denominator.

F_{df_1, df_2}

First number denotes the df for the numerator (df_1)

Second number denotes the df for the denominator (df_2)

7

Degrees of freedom

- The degrees of freedom for the F test are
  $df_{num} = k - 1$ and $df_{den} = N - k$

where k is the number of treatments/groups and N is the total sample size.

8

ANOVA

- The greater the variation between between samples (or treatments/groups) in comparison with the variation around within samples(or treatments/groups)
- the more likely it is that there are "real" differences between the population means, μ_1, μ_2, μ_3 and μ_4.
- we may conclude that there is sufficient evidence to enable us to reject the null hypothesis
  $H_0: \mu_1 = \mu_2 = \mu_3$

9

ANOVA test procedure

- The ANOVA test procedure is as follows
  - Find the value of F
  - Find the number of degrees of freedom for both the numerator and denominator
  - Obtain the critical value from the table
- Reject H_0 or fail to reject H_0 depending on the value of F compared with the critical value (from table)

10

Finding critical values from Table E

1. Identify the level of significance, α
2. Determine the degrees of freedom(df_1) for the numerator
3. Determine the degrees of freedom(df_2) for the denominator
4. Use Table E to find the **critical value**
5. Choose the F value (**critical value**) corresponding to the given numerator (df_1) and denominator (df_2)

11

Performing a One-Way ANOVA Test

1. State the null and alternative hypotheses. Normally and H_1
2. Specify the level of significance. Identify α
3. Identify the degrees of freedom: k - number of groups. $df_1 = k - 1$
   $df_2 = N - k$
4. Determine the **critical value**. Use Table E

12

4/26/2020

Performing a One-Way ANOVA Test

5. Choose the **statistic**.
   Formula $F = \frac{MS_{Between}}{MS_{Within}}$
6. Calculate the **statistic**.
7. Make a decision to reject or fail to reject the null hypothesis.
   If F is in the rejection region, reject H_0.
   Otherwise, fail to reject H_0.

13

ANOVA Summary Table

It is a table is a convenient way to summarize the results in an ANOVA test.

Variation	Sum of Squares	Degrees of freedom	Mean squares	F
Between (Groups)	$SS_{Between}$	$k - 1$	$MS_{Between} = \frac{SS_{Between}}{k-1}$	$F = \frac{MS_{Between}}{MS_{Within}}$
Within (Error)	SS_{Within}	$N - k$	$MS_{Within} = \frac{SS_{Within}}{N-k}$	
Total	SS_{Total}	$N - 1$		

14

Example 1

- **Testing Differences in Mean Time to Pain Relief Among 3 Treatments** (Data from Ott and Longnecker, Statistics is not under the stress different at $\alpha = 0.05$)

Drug A	Drug B	Drug C
26	28	35
30	26	31
34	29	29
28	31	32

$\alpha = 0.05$

15

Solution 1

$$SS_{Total} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = (26-30.5)^2 + \cdots + (29-30.5)^2 = 175.5$$

$$SS_{Between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2 = 4(28-30.5)^2 + 4(29-30.5)^2 + 4(31-30.5)^2 = 33.5$$

$$SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = (26-28)^2 + \cdots + (29-29)^2 = 142$$

16

Solution 1

- **Hypotheses:**
  $H_0: \mu_1 = \mu_2 = \mu_3$ (population means are equal)
  H_1: means are not all equal (or at least one pair is not equal or $k \geq 3$)
- **Test statistic:**

$$F = \frac{MS_{Between}}{MS_{Within}}$$

17

Solution 1

- **Decision Rule:**
  Overall Ideas
  Some procedures are used the same! mean can be $\bar{X}$ & $\bar{X}$ can be $\bar{X}_j$ & $\bar{X}$ computed by using mean of the sample means!
- **Between-treatment sum of squares:**

$$SS_{Between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$$

$$SS_{Between} = 4(28-30.5)^2 + 4(29-30.5)^2 + 4(31-30.5)^2 = 33.5$$

- **Within-treatment sum of squares:**

$$SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = (26-28)^2 + \cdots + (29-29)^2 = 142$$

18

4/26/2020

Solution 1

- Decision Rule:
  $df_{num} = k-1 = 2$
  $df_{den} = N-k = 9-3 = 6$
  Right tailed test
  The critical value of F with 2 and 12 degrees of freedom, relative to 5% level of significance, $F_{2,6,0.05} = 5.14$
  Decision rule is Reject H_0 if $F > 5.14$
  Do not Reject H_0 if $F \leq 5.14$
- Compute the test statistic:

19

Solution 1

- **Overall Ideas**
  Some procedures are used the same! mean can be $\bar{X}$ & $\bar{X}$ can be $\bar{X}_j$ & $\bar{X}$ computed by using mean of the sample means!
- **Between-treatment sum of squares:**

$$SS_{Between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$$

$$SS_{Between} = 4(28-30.5)^2 + 4(29-30.5)^2 + 4(31-30.5)^2 = 33.5$$

- **Within-treatment sum of squares:**

$$SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = (26-28)^2 + \cdots + (29-29)^2 = 142$$

20

Solution 1

Mean square

Drug A		Drug B		Drug C	
X_{1A}	$(X_{1A} - \bar{X}_A)^2$	X_{1B}	$(X_{1B} - \bar{X}_B)^2$	X_{1C}	$(X_{1C} - \bar{X}_C)^2$
26	16	28	4	35	12.25
30	0	26	9	31	0.25
34	16	29	1	29	6.25
28	4	31	4	32	2.25
Sum =	36	Sum =	18	Sum =	20.75
Mean =	30	Mean =	28	Mean =	31

$$SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = 36 + 18 + 20.75 = 74.75$$

21

Solution 1

$$SS_{Total} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = (26-30.5)^2 + \cdots + (29-30.5)^2 = 175.5$$

Sum of Squares

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS) (Variance)	F
Between (Treatment)	33.5	k - 1 = 3 - 1 = 2	MSB = 33.5/2 = 16.75	F = 16.75/8.3958 = 1.9952
Within (Error)	142	N - k = 12 - 3 = 9	MSW = 142/9 = 15.7778	
Total	175.5	N - 1 = 12 - 1 = 11		

Conclusion: Reject H_0 as 33.5 > 9.88
We have significant evidence, α = 0.01 to show that the mean time to pain relief is not the same for people under the three drugs A, B, and C at α = 0.01 level.

22

Solution 1

F=10.56
df_{num}=2
df_{den}=12
α= 0.05
P-value: 0.003 (From Minitab)
0.001 < P-value < 0.01 (From Table)
P-value < α
Reject H_0

23

Minitab: Stat > Anova > One-way

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	2	42.1	21.0	10.56	0.003
Error	12	23.9	2.0		
Total	14	67.1			

24

4/26/2020

Example 2

- *Testing Difference in Mean Age Completion 8th Grade Among 3 School Districts.* The following data, collected from three school districts (Hamilton, Jefferson, and Lincoln schools Dr.S) reflect age of students at completion of eighth grade.
- Test if there is a significant difference in the mean age at completion of eighth grade for n_j=6 students, and eighth students using α 5% level of significance.

Hamil	14	14	14	14	15	15
Jeffer	13	13	14	14	14	14
Lincoln	13	13	13	13	14	14

3 groups

25

Solution 2

Hamil	14	14	14	14	15	15		
X_{iH}	14	14	14	14	15	15	14	$(X_{iH} - \bar{X}_H)^2$
$(X_{iH} - \bar{X}_H)^2$	0	0	0	0	1	1		
Jeffer	13	13	14	14	14	14	13.8	
X_{iJ}	13	13	14	14	14	14		$(X_{iJ} - \bar{X}_J)^2$
$(X_{iJ} - \bar{X}_J)^2$	0.64	0.64	0.04	0.04	0.04	0.04		
Lincoln	13	13	13	13	14	14	13.5	
X_{iL}	13	13	13	13	14	14		$(X_{iL} - \bar{X}_L)^2$
$(X_{iL} - \bar{X}_L)^2$	0.25	0.25	0.25	0.25	0.25	0.25		

$\sum X_{iH} = 87$ $\sum (X_{iH} - \bar{X}_H)^2 = 2.435$
$\sum X_{iJ} = 82.8$ $\sum (X_{iJ} - \bar{X}_J)^2 = 0.2752$
$\sum X_{iL} = 81.5$ $\sum (X_{iL} - \bar{X}_L)^2 = 1.5000$

26

Solution 2

- Hypotheses:
  $H_0: \mu_1 = \mu_2 = \mu_3$ (population means are equal)
  H_1: means are not all equal (or at least one pair is not equal or $k \geq 3$).
- Test statistic:

$$F = \frac{MS_{Between}}{MS_{Within}}$$

27

Solution 2

- Overall Ideas
  Some procedures are not equal!
- Between-treatment sum of squares:

$$SS_{Between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$$

$$SS_{Between} = 6(14.5 - 14.4)^2 + 6(13.8 - 14.4)^2 + 6(13.5 - 14.4)^2 = 2.4$$

- Within-treatment sum of squares:

Method 1

$$SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = (14-14.5)^2 + \cdots + (15-14.5)^2 + (13-13.8)^2 + \cdots + (14-13.8)^2 + (13-13.5)^2 + \cdots + (14-13.5)^2 = 2.4$$

28

Solution 2

- Decision Rule:
  $df_{num} = k-1 = 2$
  $df_{den} = N-k = 18-3=15$
  Right tailed test
  The critical value of F with 2 and 15 degrees of freedom, relative to 5% level of significance, $F_{2,15,0.05} = 3.68$
  Decision rule is Reject H_0 if $F > 3.68$
  Do not Reject H_0 if $F \leq 3.68$
- Compute the test statistic:

29

Solution 2

- Overall Ideas
  Some procedures are not equal!
- Between-treatment sum of squares:

$$SS_{Between} = \sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2$$

$$SS_{Between} = 6(14.5 - 14.4)^2 + 6(13.8 - 14.4)^2 + 6(13.5 - 14.4)^2 = 2.4$$

- Within-treatment sum of squares:

Method 1

$$SS_{Within} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = (14-14.5)^2 + \cdots + (15-14.5)^2 + (13-13.8)^2 + \cdots + (14-13.8)^2 + (13-13.5)^2 + \cdots + (14-13.5)^2 = 2.4$$

30

4/26/2020

Hamil	14	14	14	14	15	15		
X_{iH}	14	14	14	14	15	15	14	$(X_{iH} - \bar{X}_H)^2$
$(X_{iH} - \bar{X}_H)^2$	0	0	0	0	1	1		
Jeffer	13	13	14	14	14	14	13.8	
X_{iJ}	13	13	14	14	14	14		$(X_{iJ} - \bar{X}_J)^2$
$(X_{iJ} - \bar{X}_J)^2$	0.64	0.64	0.04	0.04	0.04	0.04		
Lincoln	13	13	13	13	14	14	13.5	
X_{iL}	13	13	13	13	14	14		$(X_{iL} - \bar{X}_L)^2$
$(X_{iL} - \bar{X}_L)^2$	0.25	0.25	0.25	0.25	0.25	0.25		

$\sum X_{iH} = 87$ $\sum (X_{iH} - \bar{X}_H)^2 = 2.435$
$\sum X_{iJ} = 82.8$ $\sum (X_{iJ} - \bar{X}_J)^2 = 0.2752$
$\sum X_{iL} = 81.5$ $\sum (X_{iL} - \bar{X}_L)^2 = 1.5000$

31

Solution 2

$$SS_{Total} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = (14-14.4)^2 + \cdots + (15-14.4)^2 + (13-14.4)^2 + \cdots + (14-14.4)^2 + (13-14.4)^2 + \cdots + (14-14.4)^2 = 5.4$$

Sum of Squares

Source of Variation	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS) (Variance)	F
Between (Treatment)	2.4	k - 1 = 3 - 1 = 2	MSB = 2.4/2 = 1.2	F = 1.2/0.1833 = 6.5455
Within (Error)	3.0	N - k = 18 - 3 = 15	MSW = 3/15 = 0.2	
Total	5.4	N - 1 = 18 - 1 = 17		

Conclusion: Reject H_0 as 4.99 > 3.29
We have significant evidence, α=0.05 to show that the mean ages at completion of eighth grade are not all equal

32

Solution 2

F=6.99
df_{num}=2
df_{den}=15
α= 0.05
P-value: 0.015 (From Minitab)
0.01 < P-value < 0.025
P-value < α
Reject H_0

33

Minitab: Stat > Anova > One-way

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	2	0.24	0.120	6.55	0.015
Error	15	0.275	0.0183		
Total	17	0.515			

34

Obtaining Test Statistic for a One-Way ANOVA

1. Find the mean of each sample (or group) $\bar{X}_j$
   (n-sample size)
2. Find the overall mean (the mean of the means) $\bar{X}$
   (N-total sample size)
3. Find the sum of squares between samples (groups): $SS_{Between}$
4. Find the sum of squares within the sample (or group): SS_{Within}

Review of Introduction to Probability and Statistics for Engineers and Scientists, Third Edition, Sheldon M. Ross, Copyright © 2004, Elsevier Inc. All rights reserved. Chapter 10 Comparison of Treatments, Section 10.2 Comparing Sample Means

35

Obtaining Test Statistic for a One-Way ANOVA

5. Find the variance between the samples (groups): (k = number of sampled groups)
6. Find the variance within the samples (groups):
7. Find the test statistic F

36

6