Chapter V

Analysis of Variance

D. 5. 1. (***One-Way ANOVA***)
One-way analysis of variance (*ANOVA*), is a statistical technique that can be used to evaluate whether there are differences between the means of three or more populations.

R. 5. 1.
The analysis of variance procedure can also be used to compare two population means. However, there are more efficient procedures to do so.

R. 5. 2. (***Assumptions***)
The following *assumptions* must hold to use one-way ANOVA:

1. The population from which the samples are drawn are (approximately) normally distributed.
2. The population from which the samples are drawn have the same variance (or standard deviation).
3. The samples drawn from different populations are random and independent.

ALG. 5. 1. (***One-Way ANOVA***)

Step 1 (Formulation of the Hypotheses):

$$H_0: \mu_1 = \mu_2 = ... = \mu_k; \quad H_1: \mu_i \neq \mu_j, \text{ for at least one } i \neq j, \; i, j = 1, 2, ..., k, \qquad k \geq 3.$$

Step 2 (Calculation of the Test Statistic)

We define:

x_{ij}: Value of the observation i for treatment j
n_j: Number of observations for treatment j
$\bar{x}_j$: Sample mean for treatment j
s_j^2: Sample variance for treatment j
s_j: Sample standard deviation for treatment j .

Formulas:

$$\bar{x}_j = \frac{\sum_{i=1}^{n_j} x_{ij}}{n_j}$$

$$s_j^2 = \frac{\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)}{n_j - 1}$$

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n_T} \quad ; \quad n_T = n_1 + n_2 + ... + n_k. \quad (\textit{Overall sample mean})$$

(If the size of each sample is n, then $n_T = k \cdot n$ und

$$\bar{\bar{x}} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{kn} = \frac{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}}{n}}{k} = \frac{\sum_{j=1}^{k} \bar{x}_j}{k}.)$$

$$SSTR := \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{\bar{x}})^2 \quad (\underline{S}\textit{um of }\underline{S}\textit{quares due to }\underline{T}\textit{rea}\underline{T}\textit{ments})$$

$$MSTR := \frac{SSTR}{k-1} \quad (\textit{Mean Square due to TreaTment})$$

$$SSE = \sum_{j=1}^{k} (n_j - 1) s_j^2 \quad (\underline{S}\textit{um of }\underline{S}\textit{quares due to }\underline{E}\textit{rror})$$

$$MSE := \frac{\sum_{j=1}^{k} (n_j - 1) s_j^2}{n_T - k} = \frac{SSE}{n_T - k} \quad (\underline{M}\textit{ean }\underline{S}\textit{quare due to }\underline{E}\textit{rror})$$

$$\boxed{F = \frac{MSTR}{MSE}}$$

Step 3 (Decision)

$p - value -$ approach : Reject H_0 if $p - Value \le \alpha$

Critical-Value-approach: Reject H_0 if $F \ge F_\alpha$

(F_α is based on an F distribution with $k-1$ numerator degree of freedom and $n_T - k$ denominator degree of freedom.)

Ex. 5. 1.
The manager of a bank wanted to check whether the four tellers at a branch of this bank serve, on average, the same number of customers per hour. He observed each of the four tellers for a certain number of hours. The following table gives the number of customers served by the four tellers during each of the observed hours:

Teller A	Teller B	Teller C	Teller D
19	14	11	24
21	16	14	19
26	14	21	21
24	13	13	26
18	17	16	20
	13	18	

At the 5% significance level, test the hypothesis that the mean number of customers served per hour by each of these four tellers is the same. Assume that all the assumptions required to apply the one-way *ANOVA* procedure hold true.

Solution:
Step 1:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4, \qquad H_1: \mu_i \neq \mu_j \text{ for at least one } i \neq j,\ i, j = 1,2,3,4$$

Step 2:

$$\bar{x}_1 = \frac{19+21+26+24+18}{5} = 21.6, \qquad \bar{x}_2 = \frac{14+16+14+13+17+13}{6} = 14.5,$$

$$\bar{x}_3 = \frac{11+14+21+13+16+18}{6} = 15.5, \qquad \bar{x}_4 = \frac{24+19+21+26+20}{5} = 22.0.$$

$$s_1^2 = \frac{(19-21.6)^2+(21-21.6)^2+(26-21.6)^2+(24-21.6)^2+(18-21.6)^2}{4} = \frac{45.2}{4} = 11.3$$

$$s_2^2 = \frac{(14-14.5)^2+(16-14.5)^2+(14-14.5)^2+(13-14.5)^2+(17-14.5)^2+(13-14.5)^2}{5} = \frac{13.5}{5} = 2.7$$

$$s_3^2 = \frac{(11-15.5)^2+(14-15.5)^2+(21-15.5)^2+(13-15.5)^2+(16-15.5)^2+(18-15.5)^2}{5} = \frac{65.5}{5} = 13.1$$

$$s_4^2 = \frac{(24-22.0)^2+(19-22.0)^2+(21-22.0)^2+(26-22.0)^2+(20-22.0)^2}{4} = \frac{34}{4} = 8.5$$

$$\bar{\bar{x}} = \frac{5 \cdot 21.6 + 6 \cdot 14.5 + 6 \cdot 15.5 + 5 \cdot 22.0}{22} = 18.0909091 \approx 18.09.$$

$$SSTR = 5 \cdot (21.6 - 18.09)^2 + 6 \cdot (14.5 - 18.09)^2 + 6 \cdot (15.5 - 18.09)^2 + 5 \cdot (22 - 18.09)^2 = 255.6182.$$

$$MSTR = \frac{255.6182}{4 - 1} = 85.2060667$$

$$SSE = (5-1) \cdot 11.3 + (6-1) \cdot 2.7 + (6-1) \cdot 13.1 + (5-1) \cdot 8.5 = 158.2$$

$$MSE = \frac{158.2}{22 - 4} = 8.7889$$

$$F = \frac{85.2060667}{8.7889} \approx 9.695.$$

Step 3:

$\alpha = 0.05.$

Numerator degree of freedom: $k - 1 = 4 - 1,$
Denominator degree of freedom: $n_T - k = 22 - 4 = 18.$

$F_{0.05}(df_1 = 3; df_2 = 18) = 3.16.$

Because of $F = 9.695 > 3.16 = F_{0.05}$ we reject the null hypothesis and conclude that the mean number of customers served per hour by each of the four tellers is not the same. In other words, at least one of the four means is different from the other three.

R. 5. 3. (***ANOVA Table***)
The results of the preceding calculations for our example can be displayed conveniently in a table referred to as *ANOVA Table:*

ANOVA Table

Source of Variation	Sum of Squares	Degree of Freedom	Mean Square	F
Treatments	255.6182	3	85.2060667	9.695
Error	158.2000	18	8.7889	
Total	413.8182	21		

R. 5. 4. (***Confidence Intervals for Each of the Population Means***)
Confidence interval estimates for each of the *k* population means can be developed using

$$\mu_j \in \left[\bar{x}_j - t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j},\ \bar{x}_j + t_{\alpha/2} \cdot \frac{\sqrt{MSE}}{n_j} \right], \quad j = 1, 2, ..., k,$$

The degrees of freedom for the *t* value are the degrees of freedom associated with the within-treatments of σ^2, namely $n_T - k$.

Ex. 5. 1. (***cont'd***)

$$\mu_1 \in \left[21.6 - 2.101 \cdot \frac{\sqrt{8.7889}}{5},\ 21.6 + 2.101 \cdot \frac{\sqrt{8.7889}}{5} \right] \in [20.354,\ 22.846]$$

$$\mu_2 \in \left[14.5 - 2.101 \cdot \frac{\sqrt{8.7889}}{6},\ 14.5 + 2.101 \cdot \frac{\sqrt{8.7889}}{6} \right] \in [13.462,\ 15.538]$$

$$\mu_3 \in \left[15.5 - 2.101 \cdot \frac{\sqrt{8.7889}}{6},\ 15.5 + 2.101 \cdot \frac{\sqrt{8.7889}}{6} \right] \in [14.462,\ 16.538]$$

$$\mu_4 \in \left[22.0 - 2.101 \cdot \frac{\sqrt{8.7889}}{5},\ 22.0 + 2.101 \cdot \frac{\sqrt{8.7889}}{5} \right] \in [20.754,\ 23.246]$$

R. 5. 5. (***Multiple Comparison Procedures***)
When we use analysis of variance to test whether the means of *k* populations are equal, rejection of the null hypothesis allows us to conclude only that the population means are *not all equal*. In some cases we will want to go a step further and determine where the differences among means occur. In what follows we describe two comparison procedures:

R. 5. 6. (***Fisher's Least Significance Difference, LSD***)

Step 1:
Formulate the hypotheses:

$$H_0 :\ \mu_i = \mu_j \qquad H_1 :\ \mu_i \neq \mu_j \quad `$$

Step 2:
Calculate the test statistic:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}}$$

Step 3:
Decision:

$p-value-$approach : Reject H_0 if $p-Value \le \alpha$

Critical-Value-approach: Reject H_0 if $t \le -t_{\alpha/2}$ or $t \ge t_{\alpha/2}$

(The value of $t_{\alpha/2}$ is based on t distribution with $n_T - k$ degrees of freedom.)

Ex. 5. 1. (***cont'd***)
Step 1:

$$H_0: \mu_1 = \mu_2 \qquad H_1: \mu_1 \neq \mu_2$$

Step 2:

$$t = \frac{21.6 - 14.5}{\sqrt{8.7889 \cdot \left(\frac{1}{5} + \frac{1}{6}\right)}} = \frac{7.1}{1.795159232} \approx 3.955$$

Step 3:

$t_{statistic} = 3.955 > 2.101 = t_{18;0.05}$.

Reject H_0.

R. 5. 7. (***Fisher's Procedure Based on the Test Statistic*** $\bar{x}_i - \bar{x}_j$)
Step 1:
Formulate the hypotheses:

$$H_0: \mu_i = \mu_j \qquad H_1: \mu_i \neq \mu_j$$

Step 2:
Calculate the test statistic:

$$\bar{x}_i - \bar{x}_j$$

Step 3:
Decision:

Reject H_0 if $|\bar{x}_i - \bar{x}_j| > LSD$, where $LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$

Ex. 5. 1. (***Cont'd***)

Step 1:

$$H_0:\ \mu_1 = \mu_2 \qquad\qquad H_1:\ \mu_1 \neq \mu_2$$

Step 2:

$$\bar{x}_1 - \bar{x}_2 = 21.6 - 14.5 = 7.1$$

Step 3:

$$LSD = 2.101 \cdot \sqrt{8.7889 \cdot \left(\frac{1}{5} + \frac{1}{6}\right)} \approx 3.772$$

$$\bar{x}_1 - \bar{x}_2 = 21.6 - 14.5 = 7.1 > 3.772 = LSD\ .$$

Reject H_0.

R. 5. 8. (***Confidence Interval Estimate of the Difference between Two Population Means***)

$$\left(\mu_i - \mu_j\right) \in \left[\ (\bar{x}_i - \bar{x}_j) - LSD,\ \ (\bar{x}_i - \bar{x}_j) + LSD\ \right],$$

where $$LSD = t_{\alpha/2} \cdot \sqrt{MSE \cdot \left(\frac{1}{n_i} + \frac{1}{n_j}\right)}\ .$$

Ex. 5. 1. (***cont'd***)

$$\left(\mu_1 - \mu_2\right) \in \left[7.1 - 3.772,\ 7.1 + 3.772\ \right] = \left[3.328,\ 10.772\right].$$

(Last revised: 09.12.2019)