Chapter 8
Analysis of variance: single factor

Department of Statistics and Operations Research

December 3, 2019

Plan

Plan

Assumptions and Hypotheses in One-Way ANOVA

It is assumed that the k populations are independent and normally distributed with means $\mu_1, \mu_2, \ldots, \mu_k$ and common variance σ^2.
We wish to derive appropriate methods for testing the hypothesis

$$\begin{cases} H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k \\ H1: A\ least\ two\ of\ the\ means\ are\ not\ equal. \end{cases}$$

Let y_{ij} denote the j^{th} observation from the i^{th} treatment and arrange the data as in Table 1. Here, $y_{i.}$ is the total of all observations in the sample from the i^{th} treatment, $\overline{y}_{i.}$ is the mean of all observations in the sample from the i^{th} treatment, $y_{..}$ is the total of all n_k observations, and $\overline{y}_{..}$ is the mean of all n_k observations.

Table 1

Treatment	1	2	...	i	...	k	
	y_{11}	y_{21}	...	y_{i1}	...	y_{k1}	
	$\vdots$	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$	
	y_{1b}	y_{2b}	...	y_{ib}	...	y_{kb}	
Total	$y_{1.}$	$y_{2.}$	...	$y_{i.}$	...	$y_{k.}$	$y_{..}$
Mean	$\overline{y}_{1.}$	$\overline{y}_{2.}$	...	$\overline{y}_{i.}$	...	$\overline{y}_{k.}$	$\overline{y}_{..}$

Each observation may be written in the form

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

where ε_{ij} measures the deviation of j^{th} the observation of the i^{th} 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 models.

Theorem

We have

1. $$SST = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2 = b\sum_{i=1}^{k}(\overline{y}_{i} - \overline{y}_{..})^2 + \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i.})^2,$$
2. $SST = SSA + SSE.$

It will be convenient in what follows to identify the terms of the sum-of-squares identity by the following notation (3 important measures of variability):

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{..})^2 = \textit{total sum of squares},$$

$$SSA = b\sum_{i=1}^{k}(\overline{y}_{i} - \overline{y}_{..})^2 = \textit{treatment sum of squares},$$

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{b}(y_{ij} - \overline{y}_{i.})^2 = \textit{error sum of squares}.$$

F-Ratio for Testing Equality of Means

Source of variation	Sum of squares	Degrees of freedom	Mean square	Computed
Treatments	SSA	$k-1$	$s_1^2 = \frac{SSA}{k-1}$	$f = \frac{s_1^2}{s^2}$
Error	SSE	$k(b-1)$	$s^2 = \frac{SSE}{k(b-1)}$	
Total	SSTO	$kb-1$		

When H_0 is true, the ratio $f = \frac{s_1^2}{s^2}$ is a value of the random variable F having the F-distribution with $k-1$ and $k(b-1)$ degrees of freedom.

The null hypothesis H_0 is rejected at the α-level of significance when

$f > f_\alpha(k-1, k(b-1))$

Example

Suppose in an industrial experiment that an engineer is interested in how the mean absorption of moisture in concrete varies among 5 different concrete aggregates. The samples are exposed to moisture for 48 hours. It is decided that 6 samples are to be tested for each aggregate, requiring a total of 30 samples to be tested. The model for this situation may be set up as follows. There are 6 observations taken from each of 5 populations with means $\mu_1, \mu_2, \ldots, \mu_5$ respectively. We may wish to test

$$\begin{cases} H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_5 \\ H1: A\ least\ two\ of\ the\ means\ are\ not\ equal. \end{cases}$$

Table 2: Absorption of Moisture in Concrete Aggregates

Aggregate	1	2	3	4	5	
	551	595	639	417	563	
	457	580	615	449	631	
	450	508	511	517	522	
	731	583	573	438	613	
	499	633	648	415	656	
	632	517	677	555	679	
Total	3320	3416	3663	2791	3664	16854
Mean	553.33	569.33	610.5	465.17	610.67	561.8

Test the hypothesis $\mu_1 = \mu_2 = \ldots = \mu_5$ at the 0.05 level of significance for the data of Table 2 on absorption of moisture by various types of cement aggregates.

Solution

The hypotheses are

$$\begin{cases} H_0 : \quad \mu_1 = \mu_2 = \ldots = \mu_5 \\ H1 : A \text{ least two of the means are not equal.} \end{cases}$$

$\alpha = 0.05$.
Critical region: $f > 2.76$ with $\nu_1 = 4$ and $\nu_2 = 25$. The sum-of-squares computations give

$$\begin{aligned} SST &= \sum_{i=1}^{5}\sum_{j=1}^{6}(y_{ij} - \bar{y}_{..})^2 = (551 - 561.8)^2 \\ &+ (595 - 561.8)^2 + \cdots + (555 - 561.8)^2 + (639 - 561.8)^2 \\ &= 209\,377 \end{aligned}$$

$$
\begin{aligned}
SSA &= b\sum_{i=1}^{5}(\bar{y}_{i.} - \bar{y}_{..})^2 = 6\Big[(553.33 - 561.8)^2 + (569.33 - 561.8)^2 \\
&+ (610.5 - 561.8)^2 + (465.17 - 561.8)^2 + (610.67 - 561.8)^2\Big] \\
&= 85\,356
\end{aligned}
$$

and

$$SSE = SST - SSE = 209\,377 - 85\,356 = 124\,021.$$

The ratio $f = \frac{s_1^2}{s^2} = 4.3$. These results and the remaining computations are exhibited in the next figure in the SAS ANOVA procedure.

The GLM Procedure

Dependent Variable: moisture

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	85356.4667	21339.1167	4.30	0.0088
Error	25	124020.3333	4960.8133		
Corrected Total	29	209376.8000			

R-Square	Coeff Var	Root MSE	moisture Mean
0.407669	12.53703	70.43304	561.8000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
aggregate	4	85356.46667	21339.11667	4.30	0.0088

Decision: Reject H_0 and conclude that the aggregates do not have the same mean absorption. The P-value for $= 4.3$ is 0.0088 which is smaller than 0.05.

Plan

where y_{11} represents the response obtained by using treatment 1 in block 1, y_{12} represents the response obtained by using treatment 1 in block 2, . . ., and y_{34} represents the response obtained by using treatment 3 in block 4.
Let us now generalize and consider the case of k treatments assigned to b blocks.

Treatment	1	2	Block ...	j	...	b	
1	y_{11}	y_{12}	...	y_{1j}	...	y_{1b}	$\bar{y}_{1.}$
2	y_{21}	y_{22}	...	y_{2j}	...	y_{2b}	$\bar{y}_{2.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	y_{i1}	y_{i2}	...	y_{ij}	...	y_{ib}	$\bar{y}_{i.}$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
k	y_{k1}	y_{k2}	...	y_{kj}	...	y_{kb}	$\bar{y}_{k.}$
Mean	$\bar{y}_{.1}$	$\bar{y}_{.2}$	...	$\bar{y}_{.j}$	...	$\bar{y}_{b.}$	$\bar{y}_{..}$

1. y_{ij}= the observations in the $(ij)^{th}$ cell,
2. $\bar{y}_{i.}$= mean of the observations for the i^th level,
3. $\bar{y}_{.j}$= mean of the observations for the j^th= block,
4. $\bar{y}_{..}$ = mean of all kb observations.

Plan

1. One Way ANOVA
2. Comparing a Set of Treatments in Blocks
3. Model and Hypotheses

The analysis of variance model is based on the following

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad i = 1, \ldots, k \textit{ and } j = 1, \ldots, b$$

where y_{ij} represents the observation of the i^{th} treatment in the j^{th} block, μ_{ij} is the mean response and ε_{ij} are errors which are independent $N(0, \sigma^2)$. The response mean for the level i of the factor is $\mu_{i.} = E(\overline{y}_{i.})$ and the response mean for the block j is $\mu_{.j} = E(\overline{y}_{.j})$ and the overall mean is $\mu_{..} = E(\overline{y}_{..})$.
The hypothesis to be tested is as follows:

$$\begin{cases} H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k \\ H1: \textit{A least two of the means are not equal}. \end{cases}$$

Theorem (Sum of squares Identity)

We have

$$SST = SSA + SSB + SSE,$$

where

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{b} (y_{ij} - \overline{y}_{..})^2$$

$$SSA = b * \sum_{i=1}^{k} (\overline{y}_{i.} - \overline{y}_{..})^2$$

$$SSB = k * \sum_{j=1}^{b} (\overline{y}_{.j} - \overline{y}_{..})^2$$

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{b} (y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..})^2.$$

Source	Degrees of freedom	Sum of Squares
Treatment	$k-1$	SSA
Block	$b-1$	SSE
Residual	$(k-1)(b-1)$	SSB
Total	$bk-1$	SST

The null hypothesis of no treatment effect difference

$$H_0: \quad \mu_1 = \mu_2 = \ldots, = \mu_k$$

can be tested can be tested by using the F statistic

$$F = \frac{SSA/(k-1)}{SSE/((k-1)(b-1))} = \frac{MSA}{MSE},$$

where SSTR and SSE are the treatment and error sums of squares. The F test rejects H_0 at level α if the F value in exceeds

$$F_\alpha[k-1, (k-1)(b-1)]$$

.

Example

Operator (block)

Machine	1	2	3	4	5	6	Total	Means
1	42.5	39.3	39.6	39.9	42.9	43.9	247.8	41.3
2	39.8	40.1	40.5	42.3	42.5	43.1	248.3	41.38
3	40.2	40.5	41.3	43.4	44.9	45.1	255.4	42.57
4	41.3	42.2	43.5	44.2	45.9	42.3	259.4	43.23
Total	163.8	162.1	164.9	169.8	176.2	174.1	1010.9	
Means	40.95	40.525	41.225	42.45	44.05	43.525		

Analysis of variance for the previous data

Source of variation	Sum of squares	Degrees of freedom	Mean square	Computed
Machines	15.93	3	5.31	$f = 3.34$
Operators	42.09	5	8.42	
Error	23.84	15	1.59	
Total	81.86	23		

We have

$$
\begin{aligned}
\overline{y}_{..} &= \frac{1}{k*b} \sum_{4}^{i=1} \sum_{6}^{j=1} y_{ij} \\
&= \frac{1}{6}\Big(163.8 + 162.1 + 164.9 + 169.8 + 176.2 + 174.1\Big) \\
&= \frac{1010.9}{24} = 42.12
\end{aligned}
$$

$$
\begin{aligned}
SSB &= 4 * \sum_{j=1}^{6} (\overline{y}_{.j} - \overline{y}_{..})^2 \\
&= (40.95 - 42.12)^2 + (40.525 - 42.12)^2 + (41.225 - 42.12)^2 \\
&+ (42.45 - 42.12)^2 + (44.05 - 42.12)^2 + (43.525 - 42.12)^2 \\
&= 42.09
\end{aligned}
$$

$$
\begin{aligned}
SSA &= 6 * \sum_{i=1}^{4} (\overline{y}_{i.} - \overline{y}_{..})^2 \\
&= (41.3 - 42.12)^2 + (41.38 - 42.12)^2 + (42.57 - 42.12)^2 \\
&+ (43.23 - 42.12)^2 = 15.93
\end{aligned}
$$

$$
\begin{aligned}
SST &= \sum_{i=1}^{4} \sum_{j=1}^{6} (\overline{y}_{ij} - \overline{y}_{..})^2 \\
&= (42.5 - 42.12)^2 + (39.3 - 42.12)^2 + \ldots + \\
&= (42.3 - 42.12)^2 = 81.86
\end{aligned}
$$

$$SSE = SST - SSA - SSB = 81.86 - 15.93 - 42.09 = 23.84$$

$$k = 4 \quad and \quad b = 6.$$

The null hypothesis is rejected at the $\alpha = 0.05$-level of significance since

$$F = 3.34 > f_\alpha[k-1,(k-1)(b-1)] = f_{0.05}[3,15] = 3.29.$$

Similarly for the null hypothesis is

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_6$$

we can compute the value of the F statistic $F_\alpha[k-1, k-1)(b-1)]$ for testing the difference between blocks:

$$F = \frac{8.42}{1.59} = 5.30.$$