Analysis of Variance

In an experiment study, various treatments are applied to test subjects and the response data is gathered for analysis.

A critical tool for carrying out the analysis is the Analysis of Variance (ANOVA). It enables a researcher to differentiate treatment results based on easily computed statistical quantities from the treatment outcome.

The statistical process is derived from estimates of the population variances via two separate approaches.

The first approach is based on the variance of the sample means, and the second one is based on the mean of the sample variances.

Under the ANOVA assumptions as stated below, the ratio of the two statistical estimates follows the F distribution.

Hence we can test the null hypothesis on the equality of various response data from different treatments via estimates of critical regions.

- The treatment responses are independent of each other.
- The response data follow the normal distribution.
- The variances of the response data are identical.

In the following tutorials, we demonstrate how to perform ANOVA on a few basic experimental designs.

Completely Randomized Design

In a completely randomized design, there is only one primary factor under consideration in the experiment. The test subjects are assigned to treatment levels of the primary factor at random.

Example

A fast food franchise is test marketing 3 new menu items. To find out if they the same popularity, 18 franchisee restaurants are randomly chosen for participation in the study. In accordance with the completely randomized design, 6 of the restaurants are randomly chosen to test market the first new menu item, another 6 for the second menu item, and the remaining 6 for the last menu item.

Problem

Suppose the following table represents the sales figures of the 3 new menu items in the 18 restaurants after a week of test marketing. At .05 level of significance, test whether the mean sales volume for the 3 new menu items are all equal.

```
 Item1 Item2 Item3
    22    52    16
    42    33    24
    44     8    19
    52    47    18
    45    43    34
    37    32    39
```

Solution

The solution consists of the following steps:

1. Copy and paste the sales figure above into a table file named "fastfood-1.txt" with a text editor.

2. Load the file into a data frame named df1 with the read.table function. As the first line in the file contains the column names, we set the header argument as TRUE.

   ```
   > df1 = read.table("fastfood-1.txt", header=TRUE); df1
     Item1 Item2 Item3
   1    22    52    16
   2    42    33    24
   3    44     8    19
   ```

```
4    52    47    18
5    45    43    34
6    37    32    39
```

3. Concatenate the data rows of df1 into a single vector r .

   ```
   > r = c(t(as.matrix(df1)))        # response data
   > r
    [1] 22 52 16 42 33 ...
   ```

4. Assign new variables for the treatment levels and number of observations.

   ```
   > f = c("Item1", "Item2", "Item3")   # treatment levels
   > k = 3                              # number of treatment levels
   > n = 6                              # observations per treatment
   ```

5. Create a vector of treatment factors that corresponds to each element of r in step 3 with the gl function.

   ```
   > tm = gl(k, 1, n*k, factor(f))     # matching treatments
   > tm
    [1] Item1 Item2 Item3 Item1 Item2 ...
   ```

6. Apply the function aov to a formula that describes the response r by the treatment factor tm.

   ```
   > av = aov(r ~ tm)
   ```

7. Print out the ANOVA table with the summary function.

   ```
   > summary(av)
               Df Sum Sq Mean Sq F value Pr(>F)
   tm           2    745     373    2.54   0.11
   Residuals   15   2200     147
   ```

Answer

Since the p-value of 0.11 is greater than the .05 significance level, we do not reject the null hypothesis that the mean sales volume of the new menu items are all equal.

Exercise

Create the response data in step 3 above along vertical columns instead of horizontal rows. Adjust the factor levels in step 5 accordingly.

Randomized Block Design

In a randomized block design, there is only one primary factor under consideration in the experiment. Similar test subjects are grouped into blocks.

Each block is tested against all treatment levels of the primary factor at random order. This is intended to eliminate possible influence by other extraneous factors.

Example

A fast food franchise is test marketing 3 new menu items. To find out if they have the same popularity, 6 franchisee restaurants are randomly chosen for participation in the study. In accordance with the randomized block design, each restaurant will be test marketing all 3 new menu items. Furthermore, a restaurant will test market only one menu item per week, and it takes 3 weeks to test market all menu items. The testing order of the menu items for each restaurant is randomly assigned as well.

Problem

Suppose each row in the following table represents the sales figures of the 3 new menu in a restaurant after a week of test marketing. At .05 level of significance, test whether the mean sales volume for the 3 new menu items are all equal.

Item1	Item2	Item3
31	27	24
31	28	31
45	29	46
21	18	48
42	36	46
32	17	40

Solution

The solution consists of the following steps:

1. Copy and paste the sales figure above into a table file named "fastfood-2.txt" with a text editor.

2. Load the file into a data frame named df2 with the read.table function. As the first line in the file contains the column names, we set the header argument as TRUE.

   ```
   > df2 = read.table("fastfood-2.txt", header=TRUE); df2
     Item1 Item2 Item3
   1    31    27    24
   2    31    28    31
   3    45    29    46
   ```

```
4   21   18   48
5   42   36   46
6   32   17   40
```

3. Concatenate the data rows in df2 into a single vector r .

   ```
   > r = c(t(as.matrix(df2))) # response data
   > r
    [1] 31 27 24 31 28 ...
   ```

4. Assign new variables for the treatment levels and number of control blocks.

   ```
   > f = c("Item1", "Item2", "Item3")   # treatment levels
   > k = 3                  # number of treatment levels
   > n = 6                  # number of control blocks
   ```

5. Create a vector of treatment factors that corresponds to the each element in r of step 3 with the gl function.

   ```
   > tm = gl(k, 1, n*k, factor(f))   # matching treatment
   > tm
    [1] Item1 Item2 Item3 Item1 Item2 ...
   ```

6. Similarly, create a vector of blocking factors for each element in the response data r.

   ```
   > blk = gl(n, k, k*n)             # blocking factor
   > blk
    [1] 1 1 1 2 2 2 3 3 3 4 4 4 5 5 5 6 6 6
   Levels: 1 2 3 4 5 6
   ```

7. Apply the function aov to a formula that describes the response r by both the treatment factor tm and the block control blk.

   ```
   > av = aov(r ~ tm + blk)
   ```

8. Print out the ANOVA table with the summary function.

   ```
   > summary(av)
             Df Sum Sq Mean Sq F value Pr(>F)
   tm         2    539     269    4.96  0.032 *
   blk        5    560     112    2.06  0.155
   Residuals 10    543      54
   ```

Answer

Since the p-value of 0.032 is less than the .05 significance level, we reject the null hypothesis that the mean sales volume of the new menu items are all equal.

Exercise

Create the response data in step 3 above along vertical columns instead of horizontal rows. Adjust the factor levels in steps 5 and 6 accordingly.

Factorial Design

In a factorial design, there are more than one factors under consideration in the experiment.

The test subjects are assigned to treatment levels of every factor combinations at random.

Example

A fast food franchise is test marketing 3 new menu items in both East and West Coasts of continental United States. To find out if they the same popularity, 12 franchisee restaurants from each Coast are randomly chosen for participation in the study. In accordance with the factorial design, within the 12 restaurants from East Coast, 4 are randomly chosen to test market the first new menu item, another 4 for the second menu item, and the remaining 4 for the last menu item. The 12 restaurants from the West Coast are arranged likewise.

Problem

Suppose the following tables represent the sales figures of the 3 new menu items after a week of test marketing. Each row in the upper table represents the sales figures of 3 different East Coast restaurants. The lower half represents West Coast restaurants. At .05 level of significance, test whether the mean sales volume for the new menu items are all equal. Decide also whether the mean sales volume of the two coastal regions differs.

East Coast:

==========

	Item1	Item2	Item3
E1	25	39	36
E2	36	42	24
E3	31	39	28
E4	26	35	29

West Coast:

==========

	Item1	Item2	Item3
W1	51	43	42
W2	47	39	36
W3	47	53	32
W4	52	46	33

Solution

The solution consists of the following steps:

1. Save the sales figure into a file named "fastfood-3.csv" in CSV format as follows.
   ```
   Item1,Item2,Item3
   E1,25,39,36
   E2,36,42,24
   E3,31,39,28
   E4,26,35,29
   W1,51,43,42
   W2,47,39,36
   W3,47,53,32
   W4,52,46,33
   ```

2. Load the data into a data frame named df3 with the read.csv function.
   ```
   > df3 = read.csv("fastfood-3.csv")
   ```

3. Concatenate the data rows in df3 into a single vector r .
   ```
   > r = c(t(as.matrix(df3))) # response data
   > r
    [1] 25 39 36 36 42 ...
   ```

4. Assign new variables for the treatment levels and number of observations.
   ```
   > f1 = c("Item1", "Item2", "Item3") # 1st factor levels
   > f2 = c("East", "West")            # 2nd factor levels
   > k1 = length(f1)                   # number of 1st factors
   > k2 = length(f2)                   # number of 2nd factors
   > n = 4                             # observations per treatment
   ```

5. Create a vector that corresponds to the 1th treatment level of the response data r in step 3 element-by-element with the gl function.
   ```
   > tm1 = gl(k1, 1, n*k1*k2, factor(f1))
   > tm1
    [1] Item1 Item2 Item3 Item1 Item2 ...
   ```

6. Similarly, create a vector that corresponds to the 2nd treatment level of the response data r in step 3.
   ```
   > tm2 = gl(k2, n*k1, n*k1*k2, factor(f2))
   > tm2
    [1] East East East East East ...
   ```

7. Apply the function aov to a formula that describes the response r by the two treatment factors tm1 and tm2 with interaction.
   ```
   > av = aov(r ~ tm1 * tm2)  # include interaction
   ```

8. Print out the ANOVA table with summary function.
   ```
   > summary(av)
   ```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
tm1          2    385     193    9.55  0.0015 **
tm2          1    715     715   35.48 1.2e-05 ***
tm1:tm2      2    234     117    5.81  0.0113 *
Residuals   18    363      20
```

Answer

Since the p-value of 0.0015 for the menu items is less than the .05 significance level, we reject the null hypothesis that the mean sales volume of the new menu items are all equal. Moreover, the p-value of 1.2e-05 for the east-west coasts comparison is also less than the .05 significance level. It shows there is a difference in overall sales volume between the coasts. Finally, the last p-value of 0.0113 (< 0.05) indicates that there is a possible interaction between the menu item and coast location factors, i.e., customers from different coastal regions have different tastes.

Exercise

Create the response data in step 3 above along vertical columns instead of horizontal rows. Adjust the factor levels in steps 5 and 6 accordingly.

Exercise

From Coziol et al. 2011 (2011RMxAA..47..361C) claim that probability for a galaxy to show an AGN characteristic increases with the bulge mass of the galaxy, and find evidence that this trend is really a by-product of the morphology, suggesting that the AGN phenomenon is intimately connected with the formation process of the galaxies. Reproduce the figure 6a, 6b, 9a, 10a, and tests of the table A1 and A2. And the more important the plot of confidence interval that authors forget them.

```
> library(multcomp)

> library(sandwich)

> nelg1= read.table("nelg1_R.dat",h=T)

> boxplot(Sigma ~ Act, data=nelg1, outline=F,notch=T)

> fic1 = aov(Sigma ~ Act, data=nelg1)

> fic1_glht = glht(fic1,mcp(Act="Tukey"),vcov=vcovHC)

> summary(fic1_glht)

> plot(confint(fic1_glht))

> par(mar=c(5,6,4,2))
```