Application of one-way ANOVA in completely randomized experiments

Zaharah Wahid¹, Ahmad Izwan Latiff² and Kartini Ahmad³

1, ³ Department of Science in Engineering, Faculty of Engineering, International Islamic University Malaysia, P.O. Box 10, Jalan Gombak, 53100, Kuala Lumpur, Malaysia.
2 Department of Biotechnology in Engineering Faculty of Engineering, International Islamic University Malaysia, P.O. Box 10, Jalan Gombak, 53100, Kuala Lumpur, Malaysia.

E-mail : ¹ zaharahwahid@iium.edu.my

Abstract. This paper describes an application of a statistical technique one-way ANOVA in completely randomized experiments with three replicates. This technique was employed to a single factor with four levels and multiple observations at each level. The aim of this study is to investigate the relationship between chemical oxygen demand index and location on sites. Two different approaches are employed for the analyses; critical value and p-value. It also presents key assumptions of the technique to be satisfied by the data in order to obtain valid results. Pairwise comparisons by Turkey method are also considered and discussed to determine where the significant differences among the means is after the ANOVA has been performed. The results revealed that there are statistically significant relationship exist between the chemical oxygen demand index and the location on sites.

1. Introduction

The analysis of variance, often abbreviated as ANOVA is a statistical technique to compare three or more means. It is referred to a one-way ANOVA when the number of independent variable in the analysis of variance test is only one. The null hypothesis the means are all alike using the F distribution. The alternative hypothesis is that at least one of the means is different from the others. With the F-test, all the means are compared simultaneously. Even though one is comparing three or more means in the use of the F-test, variances are used in the test instead of means. The t and z tests are not used to compare the means of three or more samples since they tend to increase the estimated errors [1, 2].

The primary assumptions in applying ANOVA are that the response must follow a normal distribution. The variance of the response is constant and they are independent and randomly distributed.

The analysis of variance was originally invented by Ronald Fisher in the 1920’s and has its root in agriculture sector. It has been further developed and has been applied in various fields such as engineering, chemical, business, management and human resource [3, 4, 5]. The basic idea remains the same.
2. Application of one-way ANOVA

2.1. Completely randomized experiments
The order of conducting the experimental trials can affect the result. The order is selected randomly or randomly assign by allocating each trial a number and using random numbers to choose the order. If there is only one factor, the experimental design is referred to as a completely randomized design. In this study the factor being investigated is the location where the leachates are collected. The location is subdivided into four groups or levels. The response variable of interest is the chemical oxygen demand index (COD). In the case of one way classification analysis of variance, data obtained from a designed experiment as described for the completely randomized experiments, CRD are usually considered to represent by a fixed effect model, since treatment of interests are included [1, 2]. The experiments were replicated three times.

2.2. Completely randomized experiments
The ANOVA test is computed based on decomposition of the total variances, $\sigma^2$ into two distinct components. The first component is known as the between group variance, and it involves finding the variance of the means. The second component, the within group or (levels) variance, by computing the variance using all data and is not affected by differences in the means, [6]. If there is no difference in the COD means, the between group variance estimate will be approximately equal to the within group variance estimate, and the F statistics test will be equal to 1. The null hypothesis will not be rejected.

To detect differences in the four means of the chemical oxygen demand indexes, we test the null hypothesis, $H_0$ against the alternative hypothesis, $H_1$ at a significant level, $\alpha = 0.05$. The sum square between group SSB ($S_{B}^2$) and within group, SSW ($S_{W}^2$) can be computed as follow:-

The grand mean, GM

$$\bar{X}_{GM} = \frac{\sum x}{N}$$

(1)

Computational of the between –group variance

Sum square between groups, SSB

$$SSB = \sum n_i (\bar{X}_i - \bar{X}_{GM})^2$$

(2)

The mean square between, MSB is an estimate of the between groups mean square

$$MSB = \frac{\sum n_i (\bar{X}_i - \bar{X}_{GM})^2}{N - 1}$$

(3)

The within-group variance, SSW

$$SSW = \sum (n_i - 1)S_i^2$$

(4)

Mean Square within, MSW is an estimate of the expected with within groups mean square

$$MSW = \frac{\Sigma (n_i-1)S_i^2}{\Sigma (n_i-1)}$$

(5)
Compute the value of the test statistic, $F$

$$F_{cal} = \frac{MSB}{MSW}$$

When the $F_{cal}$ calculated which is the ratio of equation (3) and (5) is larger than the $F$ value obtained from the table, there is enough evidence to reject the null hypothesis and conclude that at least one of the means is different from the others. The p-value approached could be another alternative to analyze the ANOVA result. Both techniques should provide the same conclusion.

3. Results and discussion

Table 3.1 shows the design matrix which is the combinations of factor and levels that make up the experiment. The test results of the experimental run according to a completely randomized design are summarized in table 1. The ANOVA analysis showing the value of the F-statistic and p-value are depicted in table 2 and a boxplot of four different locations is illustrated in figure 1. The plot indicates that location 3 has the highest COD estimated mean value of 611.333. Mathematically,

| Location | Sample | $R_1$ (mg/L) | $R_2$ (mg/L) | $R_3$ (mg/L) | Average(mg/L) |
|----------|--------|--------------|--------------|--------------|---------------|
| 1        | 1      | 256          | 239          | 230          | 241.667       |
| 1        | 2      | 135          | 158          | 124          | 139.000       |
| 1        | 3      | 135          | 141          | 129          | 135.000       |
| 2        | 1      | 202          | 224          | 218          | 214.667       |
| 2        | 2      | 233          | 199          | 141          | 191.000       |
| 2        | 3      | 186          | 181          | 200          | 189.000       |
| 3        | 1      | 591          | 589          | 581          | 587.000       |
| 3        | 2      | 519          | 504          | 486          | 503.000       |
| 3        | 3      | 628          | 574          | 632          | 611.333       |
| 4        | 1      | 260          | 285          | 222          | 255.667       |
| 4        | 2      | 399          | 386          | 375          | 386.667       |
| 4        | 3      | 235          | 231          | 185          | 217.000       |

the hypotheses are stated as follows:

$$H_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1 = Not\ all\ \mu_s\ are\ zero$$

The ANOVA table 2 showed that there is very strong evidence that the location influences the response (COD). Using the critical value approach with $\alpha = 0.05$, the F-test statistic shown in Table 3.2 is 92.09 with, df$_1$=3 and df$_2$=32 degrees of freedom, the $S^2_B$ is many times larger than the $S^2_W$ or error mean square (3198). When F-statistic exceeds the critical value of 3.32, the null hypothesis is rejected and the result is statistically significant at the 5% level. This suggests that there is a difference between the mean chemical oxygen demand at different locations. Similarly using p-value approach, the ($p \leq 0.000$) is less than $\alpha = 0.05$. We have a sufficient evidence to suggest that at least one of the four means is different from one of the others.
Figure 1. Boxplot of chemical oxygen demand index.

Table 2. One-way ANOVA of COD

| Source   | DF | SS    | MS    | F      | P      |
|----------|----|-------|-------|--------|--------|
| Location | 3  | 883469| 294490| 92.09  | 0.000  |
| Error    | 32 | 102334| 3198  |        |        |
| Total    | 35 | 985803|       |        |        |

S = 56.55     R-Square = 89.62%     R-Square (adjusted) = 88.65%

The R²-square (adjusted) value of 88.65% is the coefficient of determination of COD and fit. It indicates that the model fitted to the data can explain 89.5% of the variation occurs in the COD value. In addition, the estimated standard deviation (s =56.55) is used to obtain confidence intervals for the four levels as displayed in Table 3. Next, further investigation was conducted to determine which means are different from the others. In this study, pairwise comparisons were selected by the Turkey method with terms location. Confidence interval with confidence level 95% were performed to decide which levels have significantly different. As displayed in table 3, the first two confidence intervals do not include zero, however, the third interval does. This suggests that there is no evidence of any difference between location 1, 2 and 4. There is a significant difference between location 3 and 4. Based on these findings the locations do influence the oxygen demand index. The adequacy of the model for COD was assessed by graphing the residuals, as shown in figure 1. The structure-less residuals figure 2 and figure 3 verified that the normality assumptions, constant variance and randomly scattered points are validated and therefore the model is adequate.
Table 3. Individual 95% Confidence Interval for Mean (COD)

Tukey 95% Simultaneous Confidence Intervals
All Pairwise Comparisons among Levels of Location

Individual confidence level = 98.87%

Location = 1 subtracted from:

| LOC | Lower  | Center | Upper  |
|-----|--------|--------|--------|
| 2   | -4.873 | -0.400 | 4.073  |
| 3   | -2.873 | 1.600  | 6.073  |
| 4   | 48.727 | 53.200 | 57.673 |

Location = 2 subtracted from:

| LOC | Lower  | Center | Upper  |
|-----|--------|--------|--------|
| 3   | -2.473 | 2.000  | 6.473  |
| 4   | 49.127 | 53.600 | 58.073 |

Location = 3 subtracted from:

| LOC | Lower  | Center | Upper  |
|-----|--------|--------|--------|
| 4   | 47.127 | 51.600 | 56.073 |

Figure 2. Normal probability plot.

Figure 3. Residuals plot.
4. Results and discussion
This paper illustrates the application of one way ANOVA in completely randomized design. The experimental results have provided evidence that the location of the leachate has statistically significant effect on the index of COD. Thus, using the above procedure one can discover which means have significantly different, a common problem facing researchers could be answered.

References
[1] Montgomery, D. C. (2017). Design and analysis of experiments. John Wiley & Sons.
[2] Mead, R., Gilmour, S. G., & Mead, A. (2012). Statistical principles for the design of experiments: applications to real experiments (Vol. 36). Cambridge University Press.
[3] Ellendersen, L. D. S. N., Granato, D., Bigetti Guergoletto, K., & Wosiacki, G. (2012). Development and sensory profile of a probiotic beverage from apple fermented with Lactobacillus casei. Engineering in Life Sciences, 12(4), 475-485.
[4] Wahid, Z., & Tham, M. T. (2012). Single Array and Response Modeling of Robust Design Experiments. International Journal of Applied Physics and Mathematics, 2(5), 359.
[5] Nadir, N., Wahid, Z., Shafie, A. A., Muthalif, A. A., Malek, M. A., Aziz, N. N. A., Zainuddin, M.T. & Islam, N. M. (2013). Development of real time experimental system for investigating photochromic response to UV irradiation. In IOP Conference Series: Materials Science and Engineering (Vol. 53, No. 1, p. 012083). IOP Publishing.
[6] Faiz Elfaki and Zaharah Wahid (2015). Computational Methods and Statistics. IIUM Press. Kuala Lumpur, Malaysia.