SHORT COMMUNICATION

BULLETIN OF ENVIRONMENTAL STUDIES

Open Access Freely available online

# Non-parametric versus parametric methods in environmental sciences

**Muhammad Riaz[1], Tahir Mahmood[1], Muhammad Arslan[2,3]**

1. *Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia.*
2. *College of Petroleum and Geoscience Department, King Fahd University of Petroleum and Minerals, Dhahran 31261, Saudi Arabia.*
3. *Environmental Biotechnology Division, Helmholtz Centre for Environmental Research, Leipzig, Germany.*

## Abstract

This current report intends to highlight the importance of considering background assumptions required for the analysis of real datasets in different disciplines. We will provide comparative discussion of parametric methods (that depends on distributional assumptions (like normality)) relative to non-parametric methods (that are free from many distributional assumptions). We have chosen a real dataset from environmental sciences (one of the application areas). The findings may be extended to the other disciplines following the same spirit.

**Keywords:** F-Test; Kruskal-Wallis Test; Parametric Methods; Non-Parametric Methods

**Article Information**

***Correspondence:**
Muhammad Riaz, Associate Professor,
Email: riaz76qau@yahoo.com
Phone# 00966505714271

## Introduction

Statistical techniques are developed under certain assumptions that need to be fulfilled for a valid application. A particular statistical method is not applicable everywhere unless we ensure the validity of its background assumptions. The statistical methods are mainly classified into two types namely parametric and non-parametric. The former need the strict assumptions about the shape of the probability distribution of the data such as normality and the latter are free from any such distributional assumptions. In different application areas including environmental sciences, we have noticed that the parametric methods are more popular even if the distributional assumptions (like normality) are not satisfied. In this report, we will highlight that the non-parametric methods are better alternatives for the real applications in environmental studies where data may not always be normally distributed. Some relevant literature on the topic may be seen in Anderson (2001), Mumby (2002), Sheskin (2007), Montgomery (2012) and the references therein.

## Materials and Methods

For the said purposes, we have used a dataset related to drinking water that analyzes the microbiological quality of public water supply. The drinking water samples were collected from a public groundwater serving water to various domestic localities of Lahore, Pakistan. Five replicate samples were collected from each sampling station and all the samples were collected, preserved and stored in accordance with the standard methods APHA 9060 A, APHA 9060 B of American Public Health Association (2005). All of the collected samples were analyzed within 24 hours of sampling to avoid unpredictable changes in the microbial population. The heterotrophic colony count was determined by "Pour Plate Method" following APHA 9215 B standard method. Similarly, total coliform in water samples was determined by "Membrane Filter Analyses" in accordance with APHA 9222B. Furthermore, faecal coliform count, i.e. Escherichia coli was assessed in the samples considering APHA 9222D under "Fecal Coliform Membrane Filter Procedure". The individual population of *Citrobacter*, *Enterobacter* and *Klebsiella* was also detected after screening on selective media. Finally, enumeration was performed by using Fotodyne's TotalLab Quant Analysis software. The resulting dataset on colony forming units is given in the Table 1.

## Statistical Analysis

In order to see if there significant differences among different types of bacteria in forming up pathogens colonies we may

**Table 1:** Pathogens Colony Counts from Drinking Water Supply

| Heterotrophic colony count | Total coliform | Fecel coliform | *Citrobacter* | *Enterobacter* | *Klebsiella* |
|---|---|---|---|---|---|
| 780 | 448 | 251 | 117 | 108 | 81 |
| 448 | 547 | 134 | 9 | 161 | 54 |
| 278 | 206 | 197 | 81 | 143 | 90 |
| 233 | 206 | 197 | 72 | 287 | 923 |
| 251 | 547 | 146 | 134 | 197 | 125 |

formulate our hypotheses as:

- $H_0$: All the six types of bacteria contribute equally to the microbial contamination;
- $H_1$: At least one bacteria type contributes significantly different as compared to others.

To test this hypothesis, we use analysis of variance (ANOVA) approach at a particular level of significance (say α). A popular approach is to use the usual F-test that strictly depends on the normality assumption. If the assumption is not fulfilled for a dataset, the usual F-test may lead us to incorrect conclusions. We have applied the said F-test for the dataset under discussion and the resulting analysis is given in Table 2 (MINITAB output):

**Table 2:** One-way ANOVA: Bacterial Counts versus Bacteria Type

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Bacteria Type | 5 | 394971 | 78994 | 2.04 | 0.109 |
| Error | 24 | 929826 | 38743 | | |
| Type | 29 | 1324797 | | | |

The p-value 0.109 indicates insignificant differences among bacteria types. This is a misleading conclusion with reference to the background theory of the topic. The reason being the non-normality that may be seen from the probability plot of the residuals (Figure 1). We can see a significant deviation of the red dots from the straight line (in blue color), which is an indication of departure form normality assumption. We have also performed Anderson-Darling test of normality that gave a p-value < 0.005.

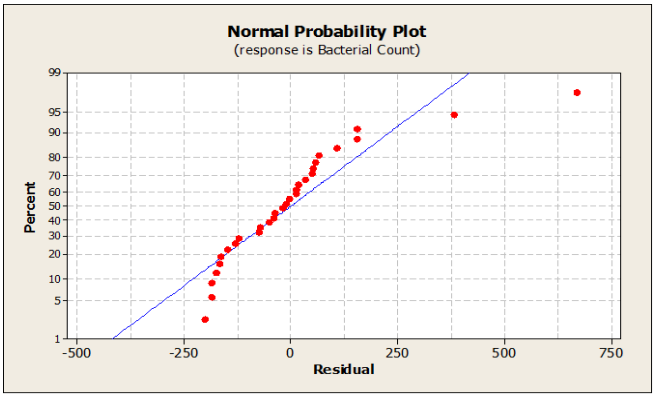


**Figure 1.** Normal Probability plot for the Bacterial Count

In such situations, a correct choice is to use Kruskal Wallis test for the testing of above sated hypotheses $H_0$ versus $H_1$. We have analyzed the datasets using the Kruskal Wallis testing procedure and the resulting output is shown in Table 3 (MINITAB output):

**Table 3:** Kruskal-Wallis Test: Bacterial Count versus Bateria Type

| H | DF | P |
|---|---|---|
| 16.95 | 5 | 0.005 |

The p-value 0.005 advocates significant differences among the bacteria types causing microbial contamination. This is in accordance with the theory of the topic. The reason being the distribution free nature of Kruskal Wallis procedure.

However, if the typical assumptions are met then parametric methods will be the best choices in terms of efficiency. To support this statement, we pick a book problem from Montgomery (2012). The statement of the problem states that "A semiconductor manufacturer has developed three different methods for reducing particle counts on wafers. All three methods are tested on five different wafers and the after treatment particle count obtained". The data are shown in the form of a table and is available as exercise problem 3.29 on page 135 of Montgomery (2012). The objective of the experiment is described as: *Do all methods have the same effect on mean particle count?* The hypotheses may be stated as:

- $H_0$: All the three methods have the same effect on mean particle count;
- $H_1$: At least one method has significantly different effect as compared to others.

Now this may be tested by the usual F-test assuming normality. We have tested the normality of this dataset using Anderson-Darling test of normality. We got a p-value of 0.136 for this test that means normality is not seriously affected and hence the usual F-test is applicable. The findings (MINITAB output) of F-test are reported in Table 4. We have also applied Kruskal Wallis test on this dataset and the MINITAB analysis outputs are given in the Table 5.

**Table 4:** One-way ANOVA: Bacterial Counts versus Bacteria Type

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Method | 2 | 8964 | 4482 | 7.91 | 0.006 |
| Error | 12 | 6796 | 566 | | |
| Type | 14 | 15760 | | | |

**Table 5:** Kruskal-Wallis Test: Count versus Method

| H | DF | P |
|---|---|---|
| 8.54 | 2 | 0.014 |

From the above analysis, it may be seen that the F-test has smaller p-value than that of the Kruskal Wallis test. It means that F-test rejects the null hypothesis more strongly. The reason being the normality of the dataset and hence appropriateness of the F-test as more efficient choice.

# Concluding Remarks

From the above discussion we conclude that we should be careful in applying the parametric procedures that rely on the distributional assumptions. If the data do not fulfil the required

assumptions we should prefer non-parametric methods of analysis (such as Kruskal Wallis test), as supported by the analysis of the data on "Pathogens Colony Counts from Drinking Water Supply" that appeared as non-normal dataset. However, for normally distributed datasets the parametric methods (such as F-test) are more efficient choices to reject incorrect hypothesis, as supported by a normal dataset taken form Montgomery (2012).

# References

American Public Health Association, 2005. Standard Methods for the Examination of Water and Wastewater (20th ed.). American Public Health Association, New York.

Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. Austral ecology, 26(1), 32-46.

Montgomery, D.C., 2012. Design and Analysis of Experiments. 8th edition, Wiley, New York.

Mumby, P.J., 2002. Statistical power of non-parametric tests: A quick guide for designing sampling strategies. Marine pollution bulletin, 44(1), 85-87.

Sheskin, D.J, 2007. Handbook of parametric and nonparametric statistical procedures, (4th ed.) Chapman and Hall/CRC, Boca Raton.

*Citation: Riaz, M., Mahmood, T., and Arslan, M., 2016. Non-Parametric versus Parametric Methods in Environmental Sciences. Bulletin of Environmental Studies 1:1 36-38.*