The Use of Parametric and Nonparametric Methods to Study the Effects of Smoking on High-Density Lipoprotein Cholesterol

Layla A. Ahmed
Department of Mathematics, College of Education, University of Garmian, Kurdistan Region – Iraq

Received: 9/5/2020 Accepted: 27/7/2020

Abstract
Analysis of variance (ANOVA) is one of the most widely used methods in statistics to analyze the behavior of one variable compared to another. The data were collected from a sample size of 65 adult males who were nonsmokers, light smokers, or heavy smokers. The aim of this study is to analyze the effects of cigarette smoking on high-density lipoprotein cholesterol (HDL-C) level and determine whether smoking causes a reduction in this level, by using the completely randomized design (CRD) and Kruskal-Wallis method. The results showed that the assumptions of the one-way ANOVA are not satisfied, while, after transforming original data by using log transformation, they are satisfied. From the results, a significantly decreased level of HDL-C in smokers as compared to non-smokers is indicated.

Keywords: completely randomized design, Kruskal-Wallis method, normality, homogeneity of variances, and independent of means and variances.

1. Introduction
Smoking can negatively impact health in many different ways, including the possible impacts on blood cholesterol. Having high cholesterol levels and smoking can be a dangerous combination for the function of the heart [1]. Smoking is now increasing rapidly throughout the developing world [2]. It is correlated to increases in the concentrations of serum total cholesterol and high-density lipoprotein

Email: layla.aziz@garmian.edu.krd
cholesterol, which are in turn positively associated with the risk of coronary heart disease [3]. For the above reasons, the analysis of high-density lipoprotein cholesterol requires implementing scientific methods, both parametric and non-parametric.

Analysis of variance (ANOVA) is one of the most frequently used statistical methods [4, 5]. It allows comparing the mean values of more than two groups in a continuous response variable [6]. It can be thought of as an extension of the t-test for two independent samples to more than two groups [7]. The development of analysis of variance is due to the work of Ronald A. Fisher (1925). Much of the early work in this area dealt with agricultural experiments [4, 7].

The valid application of ANOVA depends on three preconditions: independence of samples, normal distribution of error, and homogeneity of variances. Dependence can be eliminated by an appropriate model [5-9].

The appropriate statistical methods for analyzing the data depend on the selected measurement scale and experimental design [10]. Analysis of variance is a robust test against the normality assumption, but it may be inappropriate when the assumption of homogeneity of variance has been violated [6].

The Kruskal–Wallis test is a popular rank-based statistical method of analysis and is the non-parametric equivalent of the one-way ANOVA [10].

Transformation of data is another technique used to solve the problems of non-normality [11] and inhomogeneous variances [6]. Several data transformation techniques are available to normalize data from a non-normal form [11]. The most commonly used transformations are square roots, logarithms, and arcsine transformations to reduce heterogeneity and normalize distributions [12]. Singh [2] studied the effects of cigarette smoking on lipid profile among smokers who had smoked for more than 20 years. He found that high-density protein was significantly higher in non-smokers compared to smokers.

The purpose of this study is to analyze the effects of cigarette smoking on high density lipoprotein cholesterol and determine whether smoking causes a reduction in its blood level, by using completely randomized design and Kruskal–Wallis method.

2. Completely Randomized Design

CRD is a parametric method used to compare more than two groups and its mathematical model is [8, 13]

\[
y_{ij} = \mu + \tau_i + \varepsilon_{ijk} \tag{1}
\]

where \(y_{ij}\) is the \(j_{th}\) observation of the \(i_{th}\) treatment, \(\mu\) is the population mean, \(\tau_i\) is the treatment effect of the \(i_{th}\) level, and \(\varepsilon_{ijk}\) is the random error. The equation (1) can be rewritten as

\[
y_{ij} = \bar{y} + \bar{y}_i + (y_{ij} - \bar{y}_i) \tag{2}
\]

Sum of squares for one-way design can be written as

\[
\sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y})^2 = \sum_{i=1}^{t} \left[ \sum_{j=1}^{r} (y_{ij} - \bar{y}_i) + (y_{ij} - \bar{y}_i) \right]^2 \tag{3}
\]

In order to analyze the differences among the group means, the total variability of the \(y_{ij}\) observations (SS\(_{Total}\)) is calculated and partitioned into components: treatment sum of squares (SS\(_{Treat}\)) and error sum of squares (SS\(_{Error}\)) [14].

\[
SS_{Total} = \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y})^2 \tag{4}
\]

\[
SS_{Treat} = \sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y}_i)^2 \tag{5}
\]

\[
SS_{Error} = SS_{Total} - SS_{Treat} \tag{6}
\]

The F-test is

\[
F - test = \frac{SS_{Treat}}{SS_{Error}} \times \frac{t-1}{t-1} = \frac{MS_{Treat}}{MS_{Error}} \tag{7}
\]

where \(MS_{Treat}\) denotes the mean square for treatment and \(MS_{Error}\) denotes the mean square for error. The F-test is distributed as F-distribution with \(d.f_{Treat}=(t-1)\), which is the degree of freedom for treatment and \(d.f_{Error}=(tr-1)\) which is the degree of freedom for error term.

The completely randomized design has several assumptions that need to be fulfilled, including the normality, homogeneity of variance, and independent of mean and variance.

Ahmed
2.1 Assumptions of CRD

In this section, some assumptions of the CRD are given.

2.1.1 Normality

Most of the parametric tests require that the assumption of normality be met. Normality means that the distribution of the test is normally distributed and the assumption of normality is derived under the hypotheses:

\[ H_0: \varepsilon_{ijk} = N(0, \sigma^2) \]
\[ H_a: \varepsilon_{ijk} \neq N(0, \sigma^2) \]  

(8)

To test the assumption of normality, the following tests are used:

1. The Shapiro – Wilk test is a test for normality that was developed by Shapiro and Wilk (1965) [15] and is the most powerful test in most situations [15-17]. The statistic for this test is

\[ S, W = \frac{\left( \sum_{i=1}^{n} a_i X_i \right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \]  

(9)

where \( X_i \) is the \( i \)th largest order statistic, \( \bar{X} \) is the sample mean, and \( n \) is the number of observations.

If the p-value is over 0.05, we fail to reject the null hypothesis that the sample comes from a normal distribution.

2. The Kolmogorov–Smirnov test is another test for normality, which was first proposed by Kolmogorov (1933) and then developed by Smirnov (1939) [9]. The statistic for this test is

\[ D = |F(x) - S(x)| \]  

(10)

where \( F(x) \) is the function of the random variable \( x \) (expected) and \( S(x) \) is the observed frequency of the variable \( x \) from the sample.

If the resulting D statistic is significant, then the hypotheses that the sample comes from a normally distributed population is rejected.

2.1.2 Homogeneity of Variances

Levene’s test was used as a preliminary check of the equal variance (homogeneity of variances) assumption in ANOVA [7, 11]. Levene’s (1960) [7] original article was motivated by the k-sample problem. Before comparing the sample means, one should check that the underlying populations have a common variance. The test hypotheses are

\[ H_0: \text{all variances are equal} \]
\[ H_a: \text{at least one variance is not equal} \]  

(11)

The test statistic is

\[ W = \frac{(N-k) \sum_{i=1}^{k} n_i (z_i - \bar{z})^2}{(k-1) \sum_{i=1}^{k} n_i \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2} \]  

(12)

where

\[ z_{ij} = |y_{ij} - \bar{y}_i|, N = \sum_{i=1}^{k} n_i, \bar{z}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} z_{ij}, \bar{z} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} z_{ij} \]

2.1.3 Independence of Means and Variances

The independence of means and variances is the other assumption in the assumptions of analysis of variance, and we use a simple correlation coefficient to determine the relationship between the mean and variance. The tested hypotheses are

\[ H_0: \rho = 0 \]
\[ H_a: \rho \neq 0 \]  

(13)

where \( \rho \) is the correlation coefficient, the significance of which is tested through the t test. The statistic for this test is

\[ t = \frac{\hat{\rho} \sqrt{k-2}}{\sqrt{1-\rho^2}} \]  

(14)

If the p-value is over the level of significance, we fail to reject the null hypothesis that the correlation coefficient between the mean and variance is significant.

1. **Kruskal-Wallis Test**

The Kruskal-Wallis test is a nonparametric method for testing whether samples originate from the same distribution. Since it is a nonparametric test, it does not assume that the response variable is normally distributed.

The hypothesis tests are...
When there are no ties, the test statistic is given by
\[ H = H^* = \frac{12}{m(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(n + 1) \]
where \( k \) is the number of samples, \( n_i \) is the number of observations in the \( i^{th} \) sample, \( n \) is the number of observations in all samples combined, and \( R_i \) is the sum of the rank in the \( i^{th} \) sample.

If there are ties, the test statistic is given by
\[ H = \frac{H^*}{1 - \frac{\sum_{i=1}^{k} t_i^2}{n^3-n}} \]
where \( g \) is the number of groups with tied values and \( t_i \) is the number of observations with tie in \( i^{th} \) group, \( i = 1,2,\ldots,g \) for a level test, reject \( H_0 \) if \( K > X_{k-1,1-\alpha}^2 \) or \( P(X_{k-1}^2) < \alpha \).

The following statistics are computed from the differences \((r_i - r_j)\) of the rank averages \((i \neq j)\):
\[ z_{ij} = \frac{r_i - r_j}{\sqrt{\frac{n(n+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \]
Let \( z_{1-\alpha/t(t-1)} \) be the \((1 - \alpha/t(t-1))\) quintile of the \(N(0,1)\).
\[ H_0: \mu_i = \mu_j \text{ for all } i,j \]
\[ H_1: \mu_i \neq \mu_j \text{ for at least one pair } i,j. \]
If \( |z_{ij}| > z_{1-\alpha/t(t-1)} \), then the \( H_0 \) is rejected on the level of significance.

2. Practical Part
In this study, the effect of tobacco smoking on HDL-C level is analyzed. The data were collected from a sample of 65 adult males who were nonsmokers, light smokers (one cigarette/day), or heavy smokers (more than one cigarette/day), from the chemical analysis laboratory (Aya Lab) in Chamchamal city, North Iraq. The data were analyzed with Minitab software v. 17.

Table 1 - Results of HDL-C test

<table>
<thead>
<tr>
<th>HDL-C Level (unit)</th>
<th>Nonsmokers</th>
<th>Light smokers</th>
<th>Heavy smokers</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.00</td>
<td>60.00</td>
<td>42.50</td>
<td>58.80</td>
</tr>
<tr>
<td>49.00</td>
<td>43.00</td>
<td>36.50</td>
<td>55.50</td>
</tr>
<tr>
<td>60.00</td>
<td>56.00</td>
<td>28.50</td>
<td>33.70</td>
</tr>
<tr>
<td>58.00</td>
<td>61.00</td>
<td>24.70</td>
<td>30.90</td>
</tr>
<tr>
<td>66.00</td>
<td>42.00</td>
<td>20.30</td>
<td>37.70</td>
</tr>
<tr>
<td>58.00</td>
<td>41.56</td>
<td>27.20</td>
<td>25.00</td>
</tr>
<tr>
<td>74.00</td>
<td>54.65</td>
<td>31.60</td>
<td>29.30</td>
</tr>
<tr>
<td>62.00</td>
<td>40.75</td>
<td>24.70</td>
<td>29.50</td>
</tr>
<tr>
<td>68.00</td>
<td>35.20</td>
<td>101.70</td>
<td>50.80</td>
</tr>
<tr>
<td>66.00</td>
<td>31.78</td>
<td>130.10</td>
<td>37.20</td>
</tr>
<tr>
<td>45.00</td>
<td>26.01</td>
<td>37.50</td>
<td>22.50</td>
</tr>
<tr>
<td>41.00</td>
<td>26.70</td>
<td>32.02</td>
<td>30.60</td>
</tr>
<tr>
<td>46.00</td>
<td>24.30</td>
<td>22.00</td>
<td>35.80</td>
</tr>
<tr>
<td>51.00</td>
<td>38.90</td>
<td>34.63</td>
<td>23.10</td>
</tr>
<tr>
<td>53.00</td>
<td>50.40</td>
<td>31.23</td>
<td>19.86</td>
</tr>
</tbody>
</table>

Mean=46.679         Mean=45.952         Mean=32.598
S. D=14.099         S. D=30.569         S. D=8.551
4.1. Parametric ANOVA Results

Before carrying out any tests, the data must be tested to determine whether these assumptions are satisfied. One of the first steps in using CRD is to test assumptions. To test the assumption of the normality, Shapiro-Willk and Kolmogorov-Smirnov tests were used and the calculations were made according to the aforementioned equations. The following results were obtained.

**Kolmogorov – Smirnov test** = 0.121, with **p-value** = 0.019

And

**Shapiro – Wilk test** = 0.844, with **p-value** = 0.00

Since the **p-value** of both Kolmogorov-Smirnov (0.019) and Shapiro-Wilk (0.00) tests is less than the value of the level of significance (0.05), this implies that the null hypothesis in (8) cannot be accepted, and that the data do not have normal distribution.

Also, to test the assumption of homogeneity of variance, Levene's test is used and the result is as follows

**Levene's test** = 6.639, with **p-value** = 0.002

The **p-value** of Levene's test is less than that of the level of significance (0.05), which implies that the null hypothesis in (11) cannot be accepted and there is problem of homogeneity of variances.

To test the assumption of independence of means and variances, simple correlation coefficient is used.

The value resulted from the test is equal to 0.661 with a **p-value** of 0.54, which is greater than that assigned for the level of significance (0.05), which implies that the null hypothesis in (13) cannot be rejected and that the means and variances are independent.

Hence, the two assumptions were tested and both of them were not met.

We transform original data using log transformation to reduce heterogeneity and to normalize distributions. After transformation, the values of tests are:

**Kolmogorov – Smirnov test** = 0.059, with **p-value** = 0.2

And

**Shapiro – Wilk test** = 0.0977, with **p-value** = 0.259

After transforming the original data by using log transformation, the **p-values** of both Kolmogorov-Smirnov test (0.2) and Shapiro-Wilk (0.259) are greater than that of the level of significance (0.01), which implies that the null hypothesis in (8) cannot be rejected and that the data are normally distributed.

**Levene's test** = 2.45, with **p-value** = 0.095

The value of Levene’s test is equal to (2.45) with a **p-value** of 0.095. The **p-value** is greater than that of the level of significance (0.01), which implies that the null hypothesis in (11) cannot be rejected and there is no problem of homogeneity of variances. Hence, the assumptions were tested and met.

The parametric analysis of variance is run on this transformed dataset, as given in Table-2

<table>
<thead>
<tr>
<th>S. O. V</th>
<th>Sum of Squares</th>
<th>d. f.</th>
<th>Mean Square</th>
<th>F</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SS_{treat}$</td>
<td>0.230</td>
<td>2</td>
<td>0.115</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$SS_{error}$</td>
<td>1.567</td>
<td>62</td>
<td>0.025</td>
<td>4.545*</td>
<td>0.014</td>
</tr>
<tr>
<td>$SS_{total}$</td>
<td>1.797</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Significant at 0.05

From the above Table-2, the **p-value** is 0.014, from which we conclude that the effect of smoking is significant at the level of significance of 0.05.

To compare the means of two selected treatments, suppose we want to test the hypotheses:

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$  \hspace{1cm} (21)

The appropriate test statistic is the least significant difference (LSD):

$$LSD = t \left( 1 - \frac{a}{2}, df \right) \sqrt{\frac{2MSE}{r}}$$  \hspace{1cm} (22)

Table 2- ANOVA Table for Data after Transformation

Ahmed
Table 3- Multiple Comparison of LSD Test

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Mean Difference (1-J)</th>
<th>Std. Error</th>
<th>p- value</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoking(i)</td>
<td>Smoking(j)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.0465</td>
<td>0.049</td>
<td>0.347</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>0.1479*</td>
<td>0.049</td>
<td>0.004</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0.1015</td>
<td>0.058</td>
<td>0.085</td>
</tr>
</tbody>
</table>

*. The mean difference is significant at the 0.05 level.

From Table-3, we clearly observe that the mean difference between 1 and 2 is non-significant at 0.05, the mean difference between 2 and 3 is non-significant at 0.05, and the mean difference between 1 and 3 is significant at both 0.01 and 0.05 levels of significance.

4.2 Non-Parametric Results
The value of Kruskal-Wallis test is calculated using Minitab17. The Kruskal–Wallis statistic is: $H = 10.81$, with $p = 0.004$

The p-value is less than the value of the level of significance (0.01), then we conclude that the effect of smoking is significant at the level of significance of 0.01.

We can now conduct the multiple comparisons of the pairwise differences. The test statistic is:

Table 4- Multiple Comparisons of Non-Parametric Analysis.

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Median</th>
<th>Ave. Rank</th>
<th>$z_{ij}$</th>
<th>$z_{1−α/2(t−1)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>46</td>
<td>39.6</td>
<td>1.651</td>
<td>2.13</td>
</tr>
<tr>
<td>1-3</td>
<td>33.7</td>
<td>29.7</td>
<td>3.01*</td>
<td></td>
</tr>
<tr>
<td>2-3</td>
<td>31.7</td>
<td>20.9</td>
<td>1.15</td>
<td></td>
</tr>
</tbody>
</table>

From a Table-4, it is observed that the differences between 1 and 2 and between 2 and 3 are non-significant at the level of significance of 0.01, while the difference between 1 and 3 is significant.

3. Conclusions
From the results of the present study, it is concluded that the assumptions of the one-way ANOVA are not satisfied, while these assumptions, after transforming original data by using log transformation, are satisfied. There was a significant decrease in the level of HDL-C in smokers in comparison to that in non-smokers. The mean values of HDL-C level for non-smokers, light smokers, and heavy smokers were 46.679, 45.952, and 32.598, respectively. If the assumptions of the one-way ANOVA F-test are not met, then we can use ANOVA F-test after the transformation of the data and, hence, we can use the non-parametric Kruskal–Wallis rank test for the original data.

References
5. Moder, K. 2007. How to Keep the Type I Error Rate in ANOVA if Variances are Heteroscedastic, Austrian journal of Statistics, 36(3): 179-188.