One-Way ANOVA

11/15/2021 15:50

# Abstract

Outcome variable: *year*  
Predictor (grouping) variable: *continent*

### One-Way ANOVA

ANOVA tests the difference among the means of multiple groups. It can be used only for continuous outcome data. For ANOVA 1-way, there are 1 continuous outcome variable and 1 categorical predictor variable.

# Descriptive Statistics

Table 1 gives the basic information of the analyzing data set. Observations with missing values are removed when calculating. It shows that there is no missing value in the data.

Completeness of data.

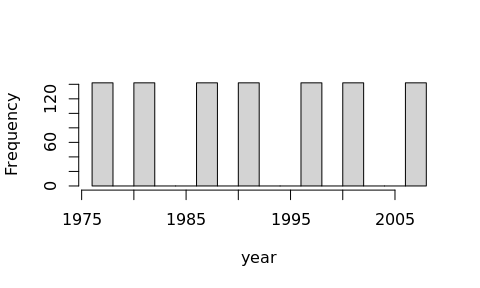
|  |  |  |
| --- | --- | --- |
|  | Observation | Incomplete Observation (not used) |
| Number | 994 | 0 |

Table 2 gives the descriptive statistics.

*Descriptive statistics.*

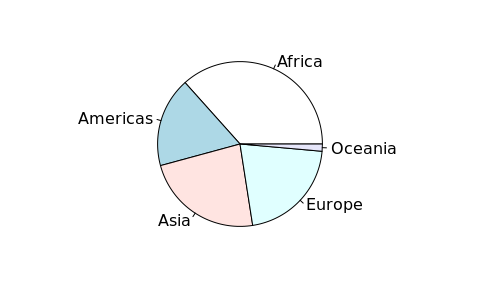
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| continent | n | Percent | year mean | sd |
| *Africa* | 364 | 36.620 | 1992 | 10 |
| *Americas* | 175 | 17.606 | 1992 | 10 |
| *Asia* | 231 | 23.239 | 1992 | 10 |
| *Europe* | 210 | 21.127 | 1992 | 10 |
| *Oceania* | 14 | 1.408 | 1992 | 10 |
| *All* | 994 | 100.000 | 1992 | 10 |

For the response (or dependent variable), the histogram graph is given in Figure 1.



Histogram of year.

For the factor group, the pie chart is given in Figure 2.



Pie chart of continent.

# Results

Table 3 gives the results of ANOVA.

Results of ANOVA test.

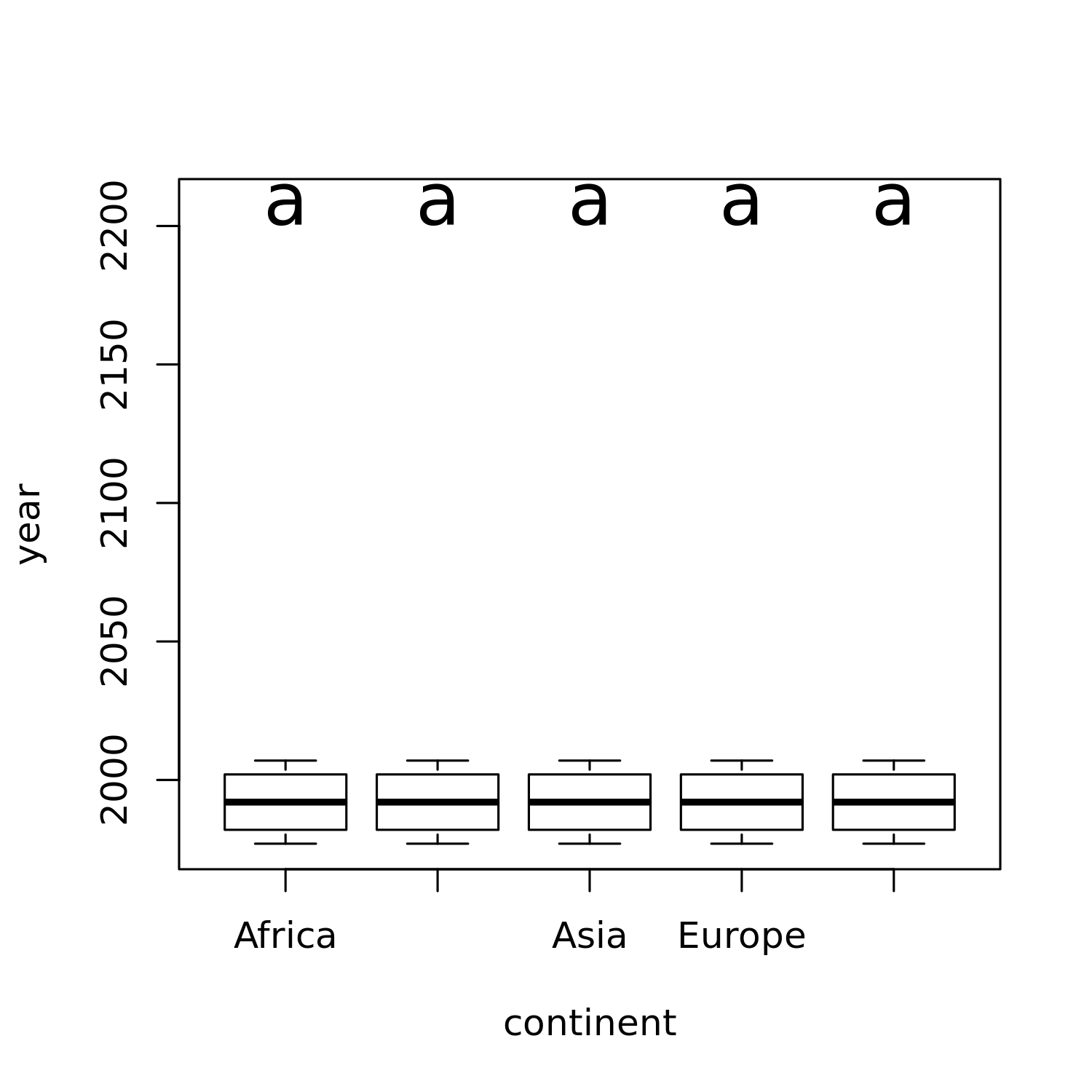
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Df | Sum Sq | Mean Sq | F Value | Pr(>F) |
| continent | 4 | 0 | 0.000 | 0 | 1 |
| Residuals | 989 | 99400 | 100.506 |  |  |

It is the result of ANOVA. The degree of freedom (Df), sum of squares (Sum Sq), mean square (Mean Sq), F value, and p-value (Pr>F) are given. A small p-value indicates that there is a difference between the group means. The commonly used significance level is 0.05.

For the given data set, the *p*-value of the F-test is 1. Since the *p*-value is larger than 0.05, we accept the null hypothesis, hence the differences between the group means are not statistically significant.

## Tukey Honest Significant Differences

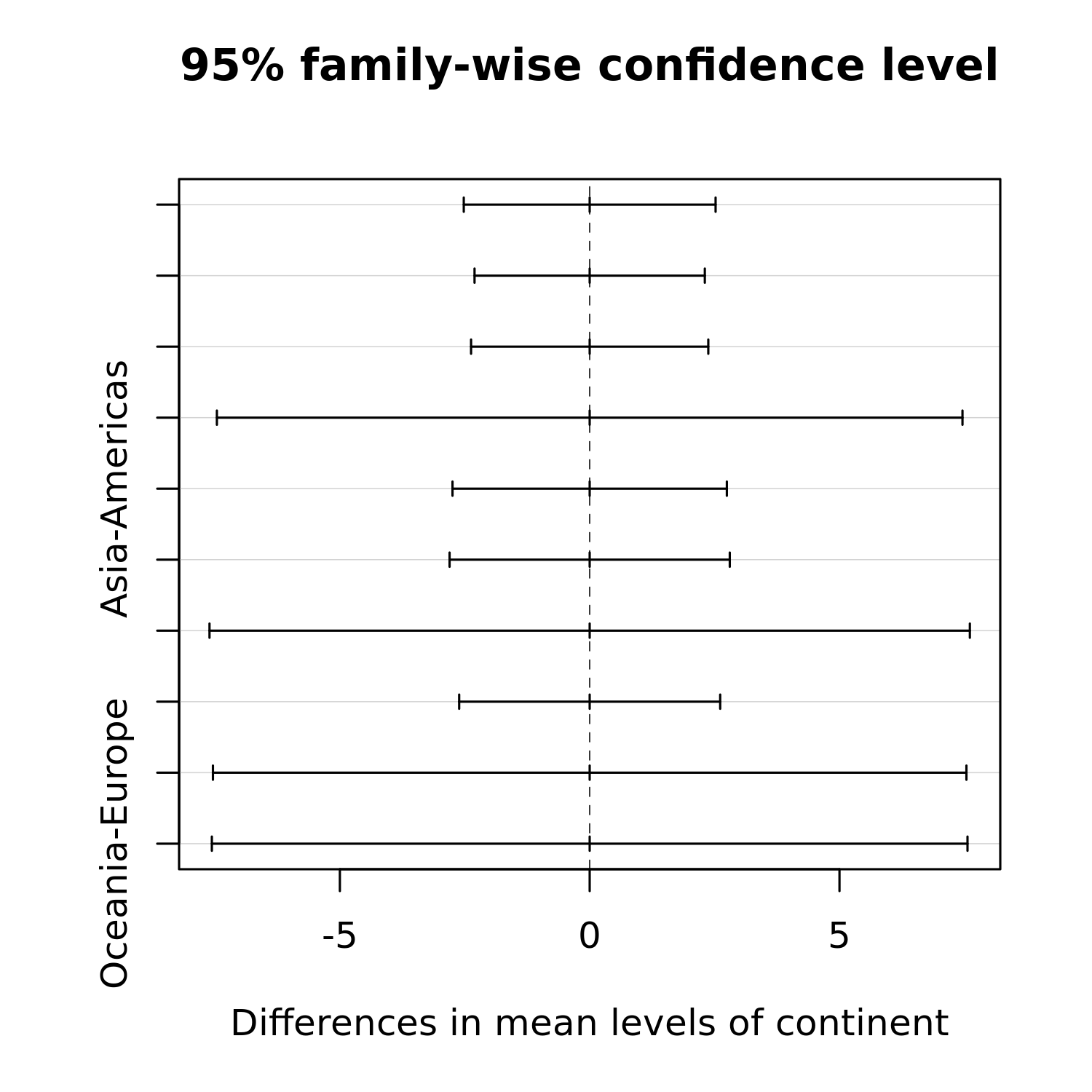
This section creates a set of confidence intervals on the differences between the means of different groups with the specified family-wise probability of coverage. The intervals are based on the Studentized range statistic, Tukey’s ‘Honest Significant Difference’ method (Miller 1981; Yandell 2017).



Box plots of each group with the Tukey test result.

The box plot with the Tukey test result is to show the shape of the distribution, its central value, and its variability for each level of the factor. In a box, the ends of the box are the upper and lower quartiles. Boxes don’t have the same letter on the top means there is a significant difference at the level of 0.05 according to the Tukey test.

The difference in the observed means and the 95% confidence intervals are given in Figure 4. Table 4 shows the details.



95% family-wise confidence interval.

This figure shows a set of confidence intervals on the differences between the means of the levels of a factor with the specified family-wise probability of coverage.

Tukey honest significant differences.

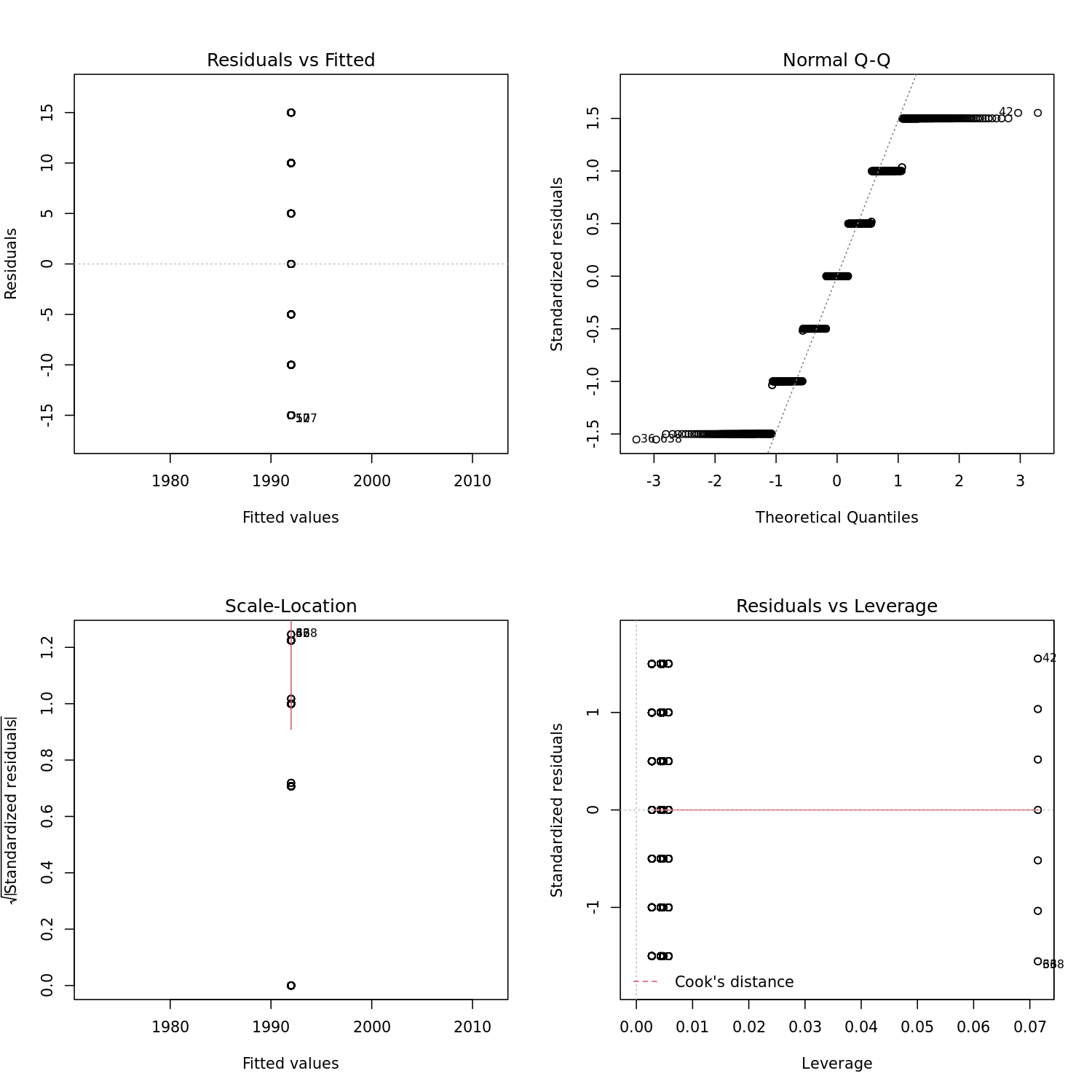
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | diff | lwr | upr | p-adj |
| Americas-Africa | 0 | -2.520 | 2.520 | 1 |
| Asia-Africa | 0 | -2.305 | 2.305 | 1 |
| Europe-Africa | 0 | -2.374 | 2.374 | 1 |
| Oceania-Africa | 0 | -7.462 | 7.462 | 1 |
| Asia-Americas | 0 | -2.746 | 2.746 | 1 |
| Europe-Americas | 0 | -2.804 | 2.804 | 1 |
| Oceania-Americas | 0 | -7.609 | 7.609 | 1 |
| Europe-Asia | 0 | -2.612 | 2.612 | 1 |
| Oceania-Asia | 0 | -7.541 | 7.541 | 1 |
| Oceania-Europe | 0 | -7.562 | 7.562 | 1 |

*diff* gives the difference in the observed means, *lwr* gives the lower endpoint of the interval, *upr* gives the upper endpoint and *p-adj* is the p-value after adjustment for the multiple comparisons.

According to Table 4, the mean difference between groups, Americas-Africa,Asia-Africa,Europe-Africa,Oceania-Africa,Asia-Americas,Europe-Americas,Oceania-Americas,Europe-Asia,Oceania-Asia,Oceania-Europe, aren’t statistically significant.

## Residual Plots

Four plots on the residuals are given, including Residuals vs Fitted, Scale-Location, Normal Q-Q, and Residuals vs Leverage. The possible outliers are marked out if any.



Residual diagnostic plots.

There are four plots, a plot of Residuals vs Fitted, a Scale-Location plot, a Normal Q-Q plot, and a plot of Residuals vs Leverage. Residuals vs Fitted is a scatter plot of residuals and fitted values (estimated responses). The plot is used to detect non-linearity, unequal error variances, and outliers. When residuals bounce randomly around line 0, it suggests that the assumption that the relationship is linear is reasonable. When the residuals roughly form a horizontal band around line 0, this suggests that the variances of the error terms are equal. When no one residual stands out from the basic random pattern of residuals, it suggests that there are no outliers. The Scale-Location plot shows whether the residuals are spread equally along with the predictor range, i.e. homoscedastic. We want the line on this plot to be horizontal with randomly spread points on the plot. The Normal Q-Q plot is used to learn whether it is reasonable to assume that the error terms are normally distributed. If the resulting plot is approximately linear, we proceed to assume that the error terms are normally distributed. The Residuals vs Leverage plots help to identify influential data points. The points are values in the upper right or lower right corners, which are outside the red dashed Cook’s distance line. These are points that would be influential in the model and removing them would likely noticeably alter the regression results.

## Assumption Checking

In this section, we will check the assumptions of the model. Firstly, we perform a normality test for the residuals, and then Bartlett’s test to check the assumption of homogeneity of variances (Bartlett 1937).

Table 5 shows the result of normality for the residuals.

Residual normality test.

|  |  |  |
| --- | --- | --- |
|  | Statistics | p-values |
| Kolmogorov-Smirnov test | 0.127 | 0 |
| Shapiro-Wilk test | 0.918 | 0 |

It shows the Kolmogorov-Smirnov test and the Shapiro-Wilk normality test for the residuals. A small p-value (less than 0.05) means the normality assumption doesn’t hold.

The results of Bartlett’s test are given in Table 6.

Barlett’s test of homogeneity of variances.

|  |  |  |
| --- | --- | --- |
| Bartlett’s K-squared | df | p-value |
| 0.033 | 4 | 1 |

It shows Bartlett’s test of the null that the variances in each of the groups (samples) are the same. A small p-value (less than 0.05) means that the variances are different among the groups.

For the given data set, the *p*-values of the normality test and Bartlett’s test are 0 and 1.

The *p*-value of the normality test is less than or equal to 0.05, hence we reject the null hypothesis, i.e., the normality assumption for the residuals is not satisfied.

We suggest using the following result of the Kruskal-Wallis test (non-parametric analog of the one-way ANOVA) for further analysis.

Kruskal-Wallis rank-sum test.

|  |  |  |
| --- | --- | --- |
| Kruskal-Wallis chi-squared | df | p-value |
| 0 | 4 | 1 |

It is the result of the Kruskal test, which is a non-parametric analog of the one-way ANOVA. Kruskal-Wallis chi-squared, degree of freedom (Df), and p-value are given. A small p-value indicates that there is a difference between the group means. The commonly used significance level is 0.05.

# Conclusions

We can get the following conclusions:

* There is at least one important assumption of ANOVA that doesn’t hold. We suggest making some data transformation to meet the assumption and refit the model or use the result of the Kruskal-Wallis test for further analysis.
* There is no sufficient evidence to say there is a difference among the groups based on the result of Kruskal-Wallis test (Table 7) at the significance level of 0.05.

# Terminologies

***Degrees of Freedom (df)***: The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called a number of degrees of freedom. In other words, the number of degrees of freedom can be defined as the minimum number of independent coordinates that can specify the position of the system completely (Good 1973).

***Normality***: Refers to the distribution of the data. The assumption is that the data follows the normal distribution.

***p-value***: In statistical hypothesis testing, the p-value is, for a given statistical model, the probability that, when the null hypothesis is true, the statistical summary would be greater than or equal to the actual observed results (Wasserstein, Lazar, and others 2016).

***Residuals***: The residual of an observed value is the difference between the observed value and the estimated value of the quantity of interest.

***F-Value***: Used with the two df values to determine the p-value, calculated by dividing the between-subjects *MS* by the residuals *MS*.

***Mean Square(MS)***: Used to determine the F value; calculated by dividing the *SS* by *df*

***Sum of Squares (SS)***: Used with the df to determine the *MS*.

***Barlett’s test***: Used to test if the samples are from populations with equal variances. Equal variances across populations is called homoscedasticity or homogeneity of variances

***Tukey Test***: Tukey test is a single-step multiple comparison procedure and statistical test. It is a posthoc analysis, which means that it is used in conjunction with an ANOVA.

# References

Bartlett, Maurice Stevenson. 1937. “Properties of Sufficiency and Statistical Tests.” *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* 160 (901): 268–82.

Good, IJ. 1973. “What Are Degrees of Freedom?” *The American Statistician* 27 (5): 227–28.

Miller, Rupert G. 1981. “Normal Univariate Techniques.” In *Simultaneous Statistical Inference*, 37–108. Springer.

Wasserstein, Ronald L, Nicole A Lazar, and others. 2016. “The Asa’s Statement on P-Values: Context, Process, and Purpose.” *The American Statistician* 70 (2): 129–33.

Yandell, BrianS. 2017. *Practical Data Analysis for Designed Experiments*. Routledge.