International Journal of Innovative Research and Advanced Studies (IJIRAS)
Volume 4 Issue 8, August 2017

ISSN: 2394-4404

Data Analysis On Bartlett's Test Of Homogeneity Of Variances

C. Vipin

S. Vaishnavi

Coimbatore Institute of Technology, Coimbatore

***Abstract:** This study aims to analyse if there are unique variations in the iris flowers namely Setosa, Virginica and Versicolor in terms of its features like length and width of petals and sepals. This analysis is done on the iris data set using the statistical technique namely the Bartlett's test of homogeneity of variances.*

***Keywords:** Bartlett's Test, Pooled Variance, Data Analytics, Statistical Analysis, Business Intelligence.*

I. INTRODUCTION

Data analysis is a process for obtaining raw data and converting it into information useful for decision-making by users. Data is collected and analysed to answer questions, test hypotheses or disprove theories[1].Big Data Analytics is to provide smartness on many sectors such as health care, water and also with virtually unlimited computing and storage resource which are to be the natural places for big data analytics and can provide easy management for IoT services[2]. The benefits of data Analysis are: allows for the identification of important trends, helps business identity performance problems that require some sort of action and it can provide a company with an edge over their competitors [3]. The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher. The data set consists of 50 samples from each of these species: Iris Setosa, Iris Virginica and Iris Versicolor. Four features were measured from each sample: the length and width of the petals and sepals. *In statistics, Bartlett's test (see Snedecor and Cochran, 1989) is used to test if k samples are from populations with equal variances* [4]. Bartlett's test is a chi-square test statistic with (k-1) degrees of freedom, where k is the number of categories in independent variable [5].

GOAL

The goal of this study was to find if there were any variations in the four features of the three samples namely Setosa, Virginica and Versicolor respectively.

II. METHODOLOGY

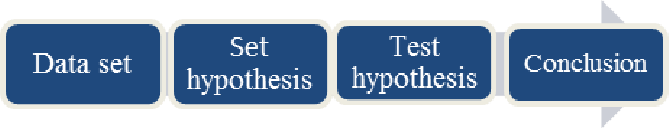

Figure 1: Process to analyse data set

Data set had been downloaded and stored as excel (.xls) file.

TYPE	PW	PL	SW	SL
Setosa	2	14	33	50
Setosa	2	10	36	46
Setosa	2	16	31	48
Setosa	1	14	36	49
Setosa	2	13	32	44
Setosa	2	16	38	51
Setosa	2	16	30	50
Setosa	4	19	38	51
Setosa	2	14	30	49
Setosa	2	14	36	50
Setosa	4	15	34	54
Setosa	2	14	42	55

Setosa	2	14	29	44

Table 1: Sample data set of Iris Flower Setosa

TYPE	PW	PL	SW	SL
Virginica	24	56	31	67
Virginica	23	51	31	69
Virginica	20	52	30	65
Virginica	19	51	27	58
Virginica	17	45	25	49
Virginica	19	50	25	63
Virginica	18	49	27	63
Virginica	21	56	28	64
Virginica	19	51	27	58
Virginica	18	55	31	64
Virginica	15	50	22	60
Virginica	23	57	32	69
Virginica	20	49	28	56

Table 2: Sample data set of Iris Flower Virginica

TYPE	PW	PL	SW	SL
Versicolor	13	45	28	57
Versicolor	16	47	33	63
Versicolor	14	47	32	70
Versicolor	12	40	26	58
Versicolor	10	33	23	50
Versicolor	10	41	27	58
Versicolor	15	45	29	60
Versicolor	10	33	24	49
Versicolor	14	39	27	52
Versicolor	12	39	27	58
Versicolor	15	42	30	59
Versicolor	13	44	23	63
Versicolor	15	49	25	63

Table 3: Sample data set of Iris Flower Versicolor

A null hypothesis is a hypothesis that says there is no statistical significance between the two variables in the hypothesis. An alternative hypothesis simply is the inverse, or opposite, of the null hypothesis [6].

In this case, null hypothesis (H_0) is that the variances in the four features of the given three flowers are equal i.e. $\sigma_1=\sigma_2....=\sigma_k$ where k is the number of samples.

And the alternate hypothesis (H_1) is that the variances in the four features of the three flowers are not equal i.e. $\sigma_1=\sigma_2....\neq\sigma_k$ where k is the number of samples.

Sample selected from the other population. Hence Pooled Variance t test (S_p) used to estimate unknown σ[7]. To test the variances of k samples Bartlett's test is used.

The test statistic is,

$$\mathbf{X^2 = \frac{(N-k)\ \ln\ (S^2{}_p) - \Sigma k_{\ i=1}(n_i-1)\ln(S_i{}^2)}{1+1/3(k-1)\ (\Sigma^{\ k}{}_{i=1}(1/\ (\ n_i-1)-1/N-k)}}$$

Where, $S_p{}^2$ is the pooled variance
N is the total population
n is the sample population
k is the number of samples
$S_i{}^2$ is the variance of the i^{th} sample.

The test statistic is the one used for testing the significance for the small (n<30) samples, which depends on magnitude of n. [8].

Here Chi-Square test had been used.

So, the applications of Chi-Square test are: To test goodness of fit, for independence of attributes, to test if the population has a specified value for the variance σ^2 and to test the equality of several population proportions [9].

If table value > calculated value, H_0 could be accepted.

If table value < calculated value, H_0 could be rejected.

Four samples were given in the iris data set. To find the table value, degrees of freedom (k-1, where k=number of samples) has been used. The level of significance is 0.05 and hence the table value is 7.815.The level of significance is the statistical significance coefficient which is the chance that a relationship a strong or stronger than the one observed due to the chance of random sampling and if level of significance is 5%(0.05) means that there is 5% chance that a correlation as strong or stronger than the observed one would result from an unusual random sampling of data when in fact the correlation was zero [9]. If null hypothesis had been accepted then there is no variability in the four features of the iris flowers and vice versa if null hypothesis had been rejected.

III. RESULTS AND ANALYSIS

SETOSA	VIRGINICA	VERSICOLOR
Calculated Value: 86.998434	Calculated Value: 41.487137	Calculated Value: 53.874814
Table Value: 7.815	Table Value: 7.815	Table Value: 7.815

Table 4: Values obtained through statistical analysis

Table value is less than the calculated value in all the three samples of the iris flower data set.

Therefore, null hypothesis had been rejected in case of all the three flowers namely Setosa, Virginica and Versicolor.

IV. CONCLUSION

In this study, Bartlett's test of homogeneity of variances is applied on the Fisher's iris data set. From the analysis it is found that the features namely length and width of the petals and sepals of the iris flowers Setosa, Virginica and Versicolor are unique. Similarly, this test can be applied on various data sets to check the uniqueness in their variations.

REFERENCES

[1] Judd, Charles and, McCleland, Gary (1989). Data Analysis. Harcourt Brace Jovanovich.

[2] Jianhua He, Jian Wei, Kai Chen, Zuoyin Tang, Yi Zhou, and Yan Zhang, IEEE Internet of Things Journal, Multi-tier Fog Computing with Large-scale IoT Data Analytics for smart cities, DOI 10.1109/JIOT.2017.2724845.

[3] http://www.dashboardinsight.com/articles/new-concepts-in-business-intelligence/data-analysis-overview.aspx.

[4] https://en.wikipedia.org/wiki/Bartlett%/27s_test

[5] David Gurson G (2012) Edition North Carolina State University School of Public and International Affairs, "Testing Statistical Assumptions ".

[6] http://study.com/academy/lesson/what-is-a-null-hypothesis-definition-examples.html

[7] David Levine M, David Stephan F, Kathryn Szabat A (2011), "Pearson Education Statistics for Managers using Microsoft Excel", 8th Edition.

[8] Gupta S.C (2014), "Fundamentals of Statistics", 7th and Enlarged Edition, Himalaya publishing Delhi.

[9] David Gurson G (2012) Edition North Carolina State University, Significance Testing: Parametric and Nonparametric, Statistical Associates Blue Book Series 18 Kindle Edition.