IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STUDY OF SPECIES SPECIFIC PROBABILITY DISTRIBUTIONS OF FEATURES OF IRIS FLOWERS

[1]Nanda Pandharikar,[2]Chintamani Kasture

[1]Associate Professor,[2]MCA student at IMS-CDR, Ahmednagar

[1]Department of Mathematics and Statistics,[1]Sydenham College of Commerce and Economics, Mumbai,India,

Abstract: Multivariate data consisting of four features namely, Sepal Length, Sepal Width, Petal Length and Petal Width of three species namely *Iris_Setosa*, *Iris_Versicolor* and *Iris_Verginica* of iris flower is taken under study .In taxonomy classification of flowers is important but mostly the approach is subjective leading to misclassification. Being objective in nature data analysts can provide better solution with quantifiable misclassification. The data set of Iris flowers fetching the attraction of many data science learners .In this paper probability distributions of four features of all the three species of Iris flower are fitted. It is observed that all the four features have different statistical distributions.

Index Terms - Iris flower, Probability Distributions, skewness.

Introduction: The data studied by Fisher on Iris flowers is still very interesting for learners of statistics and machine learning. This multivariate data consists of four features namely, Sepal Length, Sepal Width, Petal Length and Petal Width of three species namely *Iris_Setosa*, *Iris_Versicolor* and *Iris_Verginica* on 50 flowers of each species. From taxonomy consideration, it is important for classification of another flower. Taxonomists can apply their experience and taxonomical aspects to classify the given another flower. This classification approach is mostly subjective in nature and may lead to errors in classification. Even error of misclassification cannot be computed in this taxonomical subjective approach. Therefore, data analysts are playing very important role in solving this problem of classification. The approach of data analyst is data based and hence is more objective in nature. Even errors of misclassifications are quantifiable. Therefore, presently, in the world of data science this data is fetching the attraction of many budding data science learners.

In this paper, probability distributions are fitted to four features of three species. It means in all 12 statistical models are fitted to the data. This study in general gives the overall view of the structure of set of flowers. The difference in probability distributions in terms of location, variation skewness and kurtosis will reflect more light on the physical structure of set flowers. This difference can also be observed by collecting handful set of flowers and after having keen insightful observation.

This statistical probability distribution fitting is carried out on XLSTAT software in which users can analyze, customize and share results within Microsoft excel. For each sample of size 50, tables of summary statistics, automatic fit summary, estimated parameters, Histogram and descriptive statistics for interval are generated. Summary statistics includes minimum, maximum, mean, standard deviation for each variable. Automatic fit summary gives the list of distributions fitted by the software along with p-values and the conclusion regarding, distribution that fits best data for goodness of fit test. Descriptive statistics for the intervals contains lower and upper bound along with frequency and relative frequency. Log-likelihood**,** **Atkaine** Information Criterion (AIC), Bayesian Information Criterion(BIC) are used for fitting. Also Kolmogorov- Smirnov test is applied to test goodness of fit with α equal to 0.05.

Following tables give some extracted summary of analysis results which includes variable names, its probability distribution, estimated parameters and mean, variance, skewness and kurtosis of the distribution which are the important tools to study any probability distribution.

Table1: Summary Statistics for *Iris_Verginica*

Feature	Probability Distribution	Parameters	P -value for K-S Test	Mean	Variance	Skewness	Kurtosis
Petal width	Logistic	μ=1.327,s=0.115	0.246	1.327	0.044	0.418	1.200
Petal Length	Beta4	(6.405,2.639,1.965,5.205)	0.905	4.259	0.216	0.454	-0.118
Sepal Width	Weibull(2)	(10.431,2.906)	0.802	2.769	0.103	0.215	0.616
Sepal Length	Lognormal	(1.777,0.086)	0.830	5.936	0.263	0.193	0.120

It is seen from Table-I that, all the features of species *Iris_Verginica* are positively skewed, it means majority of flowers have lesser Sepal and Petal measurements. Also, in general both Sepal and Petals are bigger than respective widths. Kurtosis of Petal length is platykurtic, means flat top and heavy tails. Other three features have picked top with low weight tails. It is further seen that all the four features have different statistical distributions. Petal length have four parameter Beta distribution, whereas other features have two parameters based, Logistic Weibull and Lognormal distributions. This distinction between the probability distribution needs to correlate it with taxonomical properties so that data scientist can throw more light on the insight of the data in hand.

Table2: Summary Statistics for *Iris_Versicolor*

Feature	Probability Distribution	Parameters	P value for K-S Test	Mean	Variance	skewness	Kurtosis
Petal width	Logistic	μ=1.327,sigma=0.115	0.246	1.327	0.044	0.000	1.200
Petal Length	Beta4	(6.405,2.639,1.965,5.205)	0.905	4.259	0.216	-0.526	-0.118
Sepal Width	Weibull(2)	(10.431,2.906)	0.802	2.769	0.103	-0.655	0.616
Sepal Length	Lognormal	(1.777,0.086)	0.830	5.936	0.263	0.260	0.120

It is seen from Table-II that, Petal Length and Sepal Width of species *Iris_Versicolor* are negatively skewed, whereas Petal Width and Sepal Length are positively skewed. It reflects the structural difference between the

features of two species *Iris_Verginica* and *Iris_versicolor* Also, in general both Sepal and Petals are bigger than respective widths. Kurtosis of Petal length is platykurtic, means flat top and heavy tails. Other three features have picked top with low weight tails. It is further seen that all the four features have different statistical distributions. Petal length have four parameter Beta distribution, whereas other features have two parameters based, Logistic Weibull and Lognormal distributions. This distinction between the probability distribution needs to correlate it with taxonomical properties so that data scientist can throw more light on the insight of the data in hand.

Table3: Summary Statistics for *Iris_Setosa*

Feature	Probability Distribution	Parameters	P value for K-S Test	Mean	Variance	skewness	Kurtosis
Petal width	Arcsine	0.383	1	0.383	0.118	0.453	-1.270
Petal Length	Logistic	(1.464,0.094)	0.227	1.464	0.029	0.000	1.200
Sepal Width	Logistic	(3.406,0.210)	0.752	3.406	0.145	0.000	1.200
Sepal Length	Logistic	(5.005,0.200)	0.616	5.005	0.132	0.0000	1.200

It is seen from Table-III that, all features of *Iris-Setosa* species are positively skewed, It reflects the structural difference between the features of three species *Iris_Verginica* and *Iris_versicolor* and *Iris-Setosa.* Also, in general both Sepal Length and Width is bigger than Petal Length and Petal Width. Kurtosis of Petal Width is platykurtic, means flat top and heavy tails. Other three features have picked top with low weight tails. It is further seen that all the four features have different statistical distributions. Petal length have four parameter Beta distribution, whereas other features have two parameters based, Logistic Weibull and Lognormal distributions. This distinction between the probability distribution needs to correlate it with taxonomical properties so that data scientist can throw more light on the insight of the data in hand. Following table gives the mean of all the variables for all the three species.

Table 4:Species wise Mean for different features

	SepalLength	SepalWidth	PetalLength	PetalWidth
Iris_Setosa	5.006	3.418	1.464	0.244
Iris_Versicolor	5.936	2.77	4.26	1.326
Iris_Verginica	6.588	2.974	5.552	2.026

The shape of the distribution is a fundamental characteristic of the sample. Histograms are an excellent tool for identifying the shape of the distribution. To understand the distributions of all the four attributes,histograms are plotted. Histograms are informative about the summary statistics.

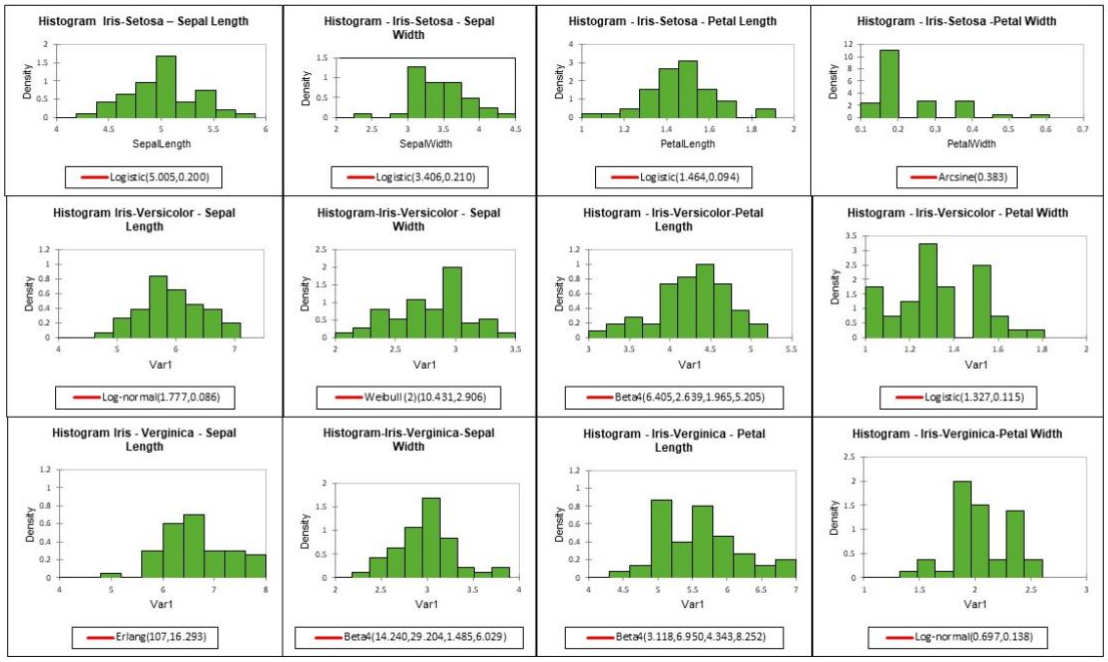

fig.1 Histograms

Conclusion

Using XLSTAT software probability distributions are fitted to four features of three species. Twelve statistical models are fitted to the data. It is observed that the variables have different distributions with different parameters. The difference in probability distributions in terms of location, variation skewness and kurtosis reflect more light on the physical structure of set flowers. Also mean and variance calculated from the data and corresponding estimated values are very much close for all the distributions.

REFERENCES

[1] Fisher,R.A.,The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936)

[2] Duda,R.O., & Hart,P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.[Web Link]