IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN:2319-765X. Volume 10, Issue 2 Ver. VI (Mar-Apr. 2014), PP 01-07
www.iosrjournals.org

Identification of Disputed Writings in Tamil Articles Using Multivariate Statistical Techniques

G. Manimannan[*] and R. Lakshmi Priya[**]
[*]*Department of Statistics, Madras Christian College, Tambaram, Chennai*
[**]*Department of Statistics, Dr. Ambedkar Govt. Arts College, Vysarpadi, Chennai*

Abstract: *Many interesting problems are associated with the science of authorship attribution and Stylometry Analysis. By quantifying relevant features related to literary style, it is possible to classify articles written by different writers and also attribute authorship to newly discovered texts. Literary style attracts the opportunity to introduce and utilise many classical multivariate statistical techniques. In this paper, an attempt is made to attribute authorship on the basis of stylistic features of certain articles written on Indian freedom movements published in the magazine* ***India****. Application of Multivariate Hotelling* T^2 *and Cluster Analysis shows that the* ***unknown*** *articles are significantly similar of* ***known*** *articles. This indicates that the Stylometry evidence do place the* ***unknown*** *articles are attributed to* ***known*** *articles. The control article is isolated from* ***known*** *and* ***Unknown*** *articles. Hence the two unattributed articles can be associated with the writings of Bharathiar.The writing style of other authors is also extracted in this study.*

Keywords: *Authorship, Multivariate Hotelling* T^2*, Authorship, Stylometry and Cluster Analysis.*

I. Introduction

Computational literary methods with regard to language as an object for scientific investigation, quantify linguistic variables and provide controlled and extensive mathematical and statistical methods to analyse literary data. These quantitative methods help us to understand the working of linguistic events in a given language and study what is common to all such events in that language, namely, the material and the relations between parts in a text. With introduction of modern computers, the field of computational linguistics is very active and much research work is taking place. As these linguistics studies involve collection, computation and analysis of huge volume of data, computers are needed to execute these works. Brinegar (1963) used distribution properties of word-length with chi-square test to prove that Mark Twain did not write “The Quintus Curtius Snodgrass letters”.

Thomas Bayes (1871) was the first to use statistical theory for solving authorship issues in the federalist papers. Auguste de Morgan as early as in 1851 has suggested the mean length of words as a measure to resolve authorship problem. Identifying the writer of an article on the basis of stylistic character is the author attribution problem in linguistic research. There has been much research covering different aspects of this field. Thisted and Efron (1976) have used distribution theory to identify the authorship of Shakespeare Plays. According to Bailey (1979) the underlying principle for authorship attribution comes from the following premises:

1. The number of putative authors should constitute a well-defined set.
2. The lengths of the writing should be sufficient to reflect the linguistic habits of the author of the disputed text and also each of the candidates.
3. The texts used for comparison should be commensurate with the disputed writing.

II. Review Of Literature

In 1901, Mendenhall reduced the concordances of Shakespeare and Bacon to distribution of word length and plotted these distributions as graph so-called *Characteristic curves*. This served as on early example of the use of graphics in distinguishing authorship. Mendenhall looked at the differences in the shapes of curves and concluded that Bacon probably did not write any of Shakespeare’s works. In 1975, Williams reproduced some of Medenhall’s surveys and noted that he was mistaken in some of his conclusions and that there was little evidence for or against the theory that some works written by Shakespeare could have been written by Bacon. Holmes (1992) has used hierarchical cluster analysis to detect changes in authorship in Mormon scripture. He also used various measures of vocabulary richness to conduct analysis.

Holmes and Forsyth (1995) used genetic algorithms to determine authorship of the disputed federalist papers. Tweedie *et al*. (1996) used a standard feedforward artificial neural network multi-layer perception to attribute authorship to the disputed Federalist papers. Holmes (1998) has used QSUM chart to settle some authorship problems. Many early attempts of quantify style relied on concordances, or inventions of the frequency of every word in text. The authorship problem is solved by searching the features of a given writer, features of which the writer is probably unaware and which can be measured quantitatively in order to have a basis for comparison with other writers, solves the authorship problem. If the number of possible writers of an unattributed work is limited, it may be possible to discover individual traits that identify one of these persons as the most likely author from that group.

Brinegar (1980) has used T^2- statistic to date Shakespeare's plays with disputed dates on the basis of lexical variables and other variables such as average verse, line length in words, the percentage of split lines and certain types-token relationship. Sundari (1997) has used Hotelling's T^2- statistic to compare social and historical novels of the great Tamil novelist Akilan. Mannion and Dixon (1997) have used T^2- statistic in attributing some articles to Oliver Goldsmith.

Cluster analysis was used to identify of twenty three disputed articles of Mahakavi Bharathiar in Tamil literature (G. Manimannan, 2009). Clustering techniques are used to cluster the social and historical novels of the Tamil writer Kalki on the basis of sentence-length. Mannion and Dixon (1997) to attribute some essays of Oliver Goldsmith used nearest neighbourhood clustering technique. Bailey (1979) and Boreland Galloway (1980) have used clustering techniques for literary data analysis. Bhatacharrya (1974) in his statistical study of word-length in Bengali prose found that the word-length distributions reveal historical trends in average word-length and give dimensional ideas of word-length in different fields of literature. Much work has been done in the literary study of *Tamil* language in the last fifty years.

III. Database

The present study deals with the literary work of the famous Tamil Poet Subramanya Bharathiar, popularly known as Mahakavi Bharathiar (1882-1921). He was a well-known poet and freedom fighter of the nineteenth century. He was the editor of ***India***, a magazine in the year 1906. In this magazine Bharathiar and other writers have written anonymously articles, editorials and short stories. Ilasai Maniyan (1975) has compiled all these articles and has brought out a book entitled ***Bharathi Dharisanam***. Two articles from this compiled book are taken up for consideration for author attribution in this study on the basis of literary style. To identify Bharathiar's style, it is also necessary to identify and set aside elements, which have the common stylistic characters of the writers of the same period. In this connection we have considered three articles written by the contemporaries of Tamil poet, namely V. Kalyansundaram, U.V. Swaminatha Iyer and V.O. Chidhambaram,and four articles written by the poet himself on the same topic in other magazines of the same period are considered for this study. All these nine articles deal with the common topic, namely, Freedom Movement of India.

Our study was based on nine articles. Out of these nine articles, four were written by the poet Bharathiar himself which we took as ***knowns***. Two were selected at random from Ilasai Maniyan's edited book, which are referred as ***unknowns*** and of the remaining three articles, each one was selected at random from the works of three different authors of the same period, which are known as ***controls***. Each sentence considered as a sample from each article. The numbers of sentences selected from each of the nine articles are given in table 1.

Table 1. Number of Sample Sentences Selected for the Study

Articles	Samples
Knowns	
Article 1	144
Article 2	192
Article 3	116
Article 4	222
Controls	
Article 5	103
Article 6	148
Article 7	112
Unknown	
Article 8	52
Article 9	96
Total Samples	1185

We have considered twenty-nine linguistic variables to measure different components of style (Grammar and syntax). These variables are listed with abbreviations in Table 2. These twenty-nine variables are identified for each sentence. If we have n sentences and p identifiable variables from each sentence, it gives rise to a data matrix of size $n \times p$. Thus each article was converted as a data matrix that forms basis for the literary data analysis of this study. As there are nine articles, it was converted in nine different matrices. The main objective of this study is to explore the authorship attributions of the ***unknown*** articles with those of ***known*** and ***controls,*** we preferred comparison between them.

Summary statistics like average values and standard deviation are calculated for each linguistic variable and they are given in Table 3 and 4. From this table we find that the standard deviations for some variables are zero for all the articles. This indicates that these nine articles do not differ from one another in terms of these variables and the remaining twenty seven variables are considered for further analysis.

III. Mehodology And Result

Two methods are considered for this study. The first method assumes the population distribution is known and it is normal. Hotelling's T^2-statistic is used to draw proper conclusion. In the second method, it is assumed that the populations are distribution-free and clustering techniques are used to compare the above said nine articles.

3.1 METHOD I

Let us assume the mean vectors of two ***unknown*** articles as mean vectors of two different normal populations. Then the mean vectors of ***known*** and of ***control*** articles are compared assuming the sentences of articles of observations from different normal populations. Treating this problem as a single sample problem, Hotelling's T^2-statistic is used to compare the mean vectors of the ***unknowns*** with those of ***knowns*** and ***controls*** on the basis of twenty-seven linguistic variables and the results are given in Table 5. The results show that both the ***unknowns*** articles are significantly attributed to ***known*** articles and hence it can be attributed to the poet Bharathiar and to other authors considered as ***controls*** in this study.

The two mean vectors of ***unknowns*** are compared as a two-sample problem of Hotelling's T^2-statistic and the result is given in Table 5. The result indicates that the two articles do similar from *known* in terms of the linguistic variables of this study and hence they can be attributed to a single author. Also the mean vectors of the whole ***knowns*** articles of Bharathiar's are compared statistically.They are similar to *unknown* articles from one another as for as these twenty seven linguistic variable are considered.

3.2 METHOD II

To validate these established results, the means of all the nine articles (***knowns***, ***controls*** and ***unknowns***) are clustered on the basis of twenty-seven linguistic variables assuming the population distribution of these nine articles are unknown. Six agglomerative hierarchical clustering techniques are used to cluster the nine articles treating the data matrices of these as inputs. The patterns of cluster formation of the six methods are given in Figures 1 through 6 as dendrogram of these nine articles respectively. All these clustering methods provide the same clusters and we treat them as natural clusters.

Fig. 1 to 6 DENDROGRAM FOR NINE ARTICLES

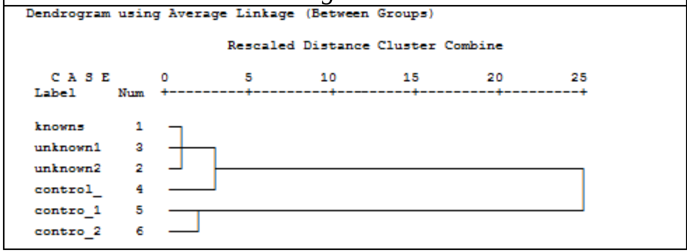
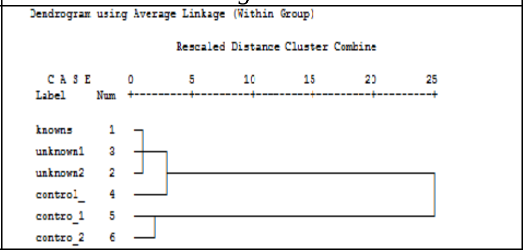
Figure 1	*Figure 2*
	

Figure 3

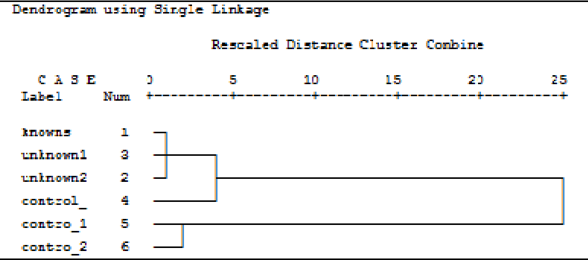


Figure 4

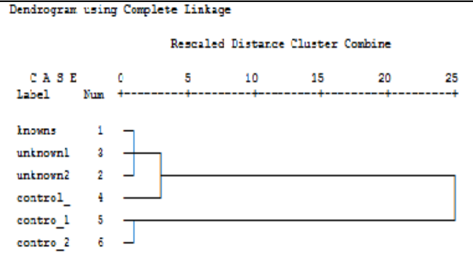


Figure 5

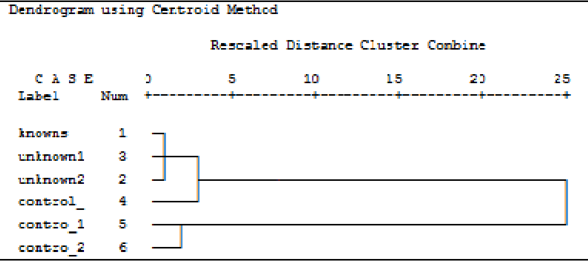
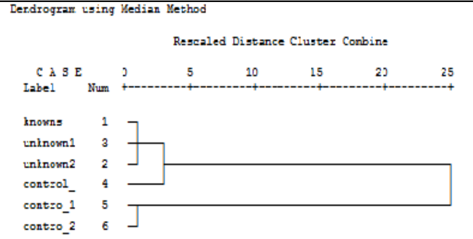


Figure 6



We get four major clusters, one cluster three objects, one Singleton clusters, and a cluster of two objects. We name them as clusters one, two and three respectively. Cluster one consists of Bharathiar's four articles and also the unknown articles. This indicates that the writing styles of Bharathiar and that of unknown articles are similar. The existence of the singleton cluster establishes that the writing styles of U. V. Saminatha Iyer. Cluster three consists of both the control articles and they can be attributed to a single author. Bharathiar's writing style is similar of unknown articles and different from those of Kalyanasundaram and Chidambararanar. This shows that the unknown articles of two may be written by Bharathiar.

IV. Conclusion

This paper deals with the attribution of authorship problem and also quantifies the style of a writer. During the Indian Freedom Movement, the poet Bharathiar has written a number of articles by attributing his name and sometimes anonymously in the magazine ***India***. In this study an attempt was made to attribute the authorship of Bharathiar of two randomly selected unattributed articles on the basis of stylistic features.

Four articles written by Bharathiar, referred to as *knowns* and two unattributed articles called as *unknowns* were collected from Ilasai Maniyan has compiled all these articles and has brought out a book entitled *Bharathi Dharisanam*. Also three articles referred to as *controls*, written by three different authors, namely Kalayana Sundram, U. V. Swaminatha Iyer and Chindrambram of the same period were considered for identification of the common stylistic features of the same period and to identify the distinct stylistic features of Bharathiar.

First of all twenty-nine linguistic variables were considered for this study. Standard deviations of two variables, namely, percentage of sentences ending with a verb (PEVE_1) and occurrences of verbs (VER1) were zero. This indicates that all the nine articles of this study do not differ from one another as far as these two variables are considered. The remaining twenty-seven variables were considered for attribution problem. Applications of Hotelling's T^2-statistic and clustering techniques have shown that the *unknown* articles are significantly similar of *knowns.* The rest of *control* articles vary from *known* and *unknown*. This indicates that the stylometry evidence do place the *unknowns* attributed to *knowns* articles. Hence the two unattributed articles attributed to Bharathiar and not to the other authors of the same period considered in this study. Clustering techniques have also shown that the attributed articles are close *unknowns* articles than to *Control* articles. Also the two unknown articles and known articles form a single cluster and it is significantly similar from one another and hence these articles may be attributed to Mahakavi Bharathiar.

References

[1] Bailey, R. W.(1979). Authorship Attribution in a Forensic Setting. *In advances in Computer-Aided Literary and Linguistic Research. Proceedings of the Fifth International Symposium on Computers in Literary and Linguistic Research.Eds, D. E. Ager, F. E. Knoles and J. Smith*. Birminham, pp. 1-15.

[2] Bhattacharya, N. C. (1974). A statistical study of word-length in Bengali Prose, Sankaya: *The Indian Journal of Statistics society*, series B, Pt.4, pp.323-347.

[3] Boreland, H. and Galloway. P, (1980). Authorship, Discrimination and Clustering: Timoneda, Montesine anonymous poems. Ass. for *Lit. and Lingust. Comput.* Bull, 8,pp. 125-151.

[4] Brainerd, B.(1980). On the distribution between a novel and a romance: a discriminant analysis. *Computer and the Humanities*, 7: pp. 259-270.

[5] Brinegar, C. S. (1963). Mark Twain and the Quintus Curtius Snodgrass Letters: A Statistical Test of Authorship. *Journal of American Statistical Association*, 58, pp. 85-96.

[6] David Mannion and Peter Dixon (1997). Authorship Attribution: the case of Oliver Goldsmith, *The Statistician*, 46, No.1, pp. 1-18.

[7] Holmes, D. I. (1985). The Analysis of literary style: a review, *Journal of Royal Statistical Society*, (Series A), 148: pp. 328-341.

[8] Holmes, D. I. (1992), A Stylometric Analysis of Mormon Scripture and Related Texts, *Journal of Royal Statistical Society*, Series A, 155: pp. 91-120.

[9] Holmes, D. I. and Forsyth, R. J. (1995). 'The Federalist' Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, 10: pp. 111-119.

[10] Holmes, D. I. (1998). The evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing,* 13: pp.111-117.

[11] Manimannan G., (1999). *Authorship Attribution: The Case of Bharathiar,* . M. Phil Thesis, Department of Statistics, Madras Christian College, Chennai, India.

[12] Thisted, B. Efron, R. (1987). Did Shakespeare write a newly discovered poem? *Biometrica*, pp. 445-55.

[13] Tweedie, F. J. Singh, S. Holmes, D. I. (1996). Neural Network Application in Stylometry: *The Federalist paper*. *Computer and the Humanities,* 30, pp.1- 10

TABLE 2. VARIABLE NAME AND ABBREVIATIONS

S. No	Variable Name	Abbreviations
1	AD_AJ1	Occurrence of Adverbs and Particle adjuncts
2	ADJ1	Occurrence of Adjectives
3	APO_1	Occurrence of apostrophe
4	CO_CON1	co-ordinating conjunction followed by noun
5	CO_DET1	co-ordinating conjunction followed by determiner
6	CON_IT1	Occurrence of Conjunctions, interrogatives
7	D_1	Mean difference in length between consecutive sentences
8	DE_NUM1	Determiners and numerals
9	DOUB1	Doublet formation (Percentage)
10	IN_IE1	Intensifiers and Intersections
11	INPA_1	Infinities and Participles
12	MA_PRO1	Mark of punctuation followed by pronoun
13	MA_SU1	Ratio of main clauses to subordinate clauses
14	NO_AD1	Noun followed by adverb
15	NO_AU1	Noun followed by Auxiliary Verb
16	NO_CO1	Noun followed by co-ordinating Conjunction
17	NO_PRO1	Nouns and Pronouns
18	OM_RE1	Ratio of occurrence of omitted relative markers
19	PER_C1	Percentage of sentences containing *though. yet, nevertheless,* or *however*
20	PEVE_1	Percentage of sentences ending with a verb
21	PHA_1	Occurrences of Word or Phrase
22	PREPO_1	Postpositions
23	SY_1	Mean length of the sentences (in syllables)
24	THER1	Ratio of occurrences of *therefore* and *thus*
25	THIS1	Ratio of occurrences of *this, these, that* and *those*
26	VER1	Verbs
27	W_1	Mean sentences length (in words)
28	WHI1	Occurrence of *Which*
29	NO_OF1	Ratio of occurrences of the construction *noun* followed by *determiner*

TABLE 3. MEAN VALUE OF THE LINGUISTIC VARIABLES

Variable Name	**Combined Mean for Knowns**	**Controls**			**Unknowns**	
	1 2 3 4	1	2	3	1	2

D_AJ1	2.337	2.195	1.426	1.054	2.20	1.43
ADJ1	0.104	0.250	0.162	0.179	0.25	0.16
APO_1	0.109	0.135	0.196	0.071	0.14	0.20
CO_CON1	0.196	0.038	0.068	0.116	0.04	0.07
CO_DET1	0.080	0.096	0.108	0.420	0.10	0.11
CON_IT1	0.974	1.827	0.804	0.750	1.83	0.80
D_1	3.952	4.317	3.588	2.607	4.32	3.59
DE_NU1	0.261	0.173	0.142	0.170	0.17	0.14
DOUB1	0.036	0.000	0.020	0.027	0.00	0.02
IN_IE1	0.020	0.000	0.000	0.009	0.00	0.00
INPA_1	0.415	0.173	0.493	0.705	0.17	0.49
MA_PRO1	1.681	0.769	0.669	0.179	1.77	1.67
MA_SU1	0.545	0.721	0.703	0.865	0.72	0.70
NO_AD1	1.865	0.029	0.122	0.188	1.03	1.12
NO_AU1	0.421	0.029	0.182	0.134	0.03	0.18
NO_CO1	0.429	0.077	0.095	0.179	0.08	0.10
NO_OF1	1.433	0.192	0.135	0.107	1.19	1.14
NO_PRO1	0.934	1.163	0.959	0.384	1.16	0.96
OM_RE1	0.391	0.654	0.392	0.607	0.65	0.39
PER_C1	0.275	0.202	0.101	0.071	0.20	0.10
PEVE_1	1.000	1.000	1.000	0.991	1.00	1.00
PHA_1	0.902	0.452	0.743	0.589	0.45	0.74
PREPO1	1.377	1.337	1.358	1.000	1.34	1.36
SY_1	21.845	25.923	22.973	18.482	22.92	21.97
THER1	0.173	0.096	0.061	0.036	0.10	0.06
THIS1	0.136	0.115	0.264	0.232	0.12	0.26
VER1	1.000	1.000	1.000	1.000	1.00	1.00
W_1	8.791	9.231	9.095	7.330	8.23	8.10
WHI1	0.007	0.019	0.027	0.054	0.02	0.03

TABLE 4. STANDARD DEVIATION OF THE LINGUISTIC VARIABLES

Variable Names	Combined Standard Deviation for Knowns	Controls			Unknowns	
	1 2 3 4	1	2	3	1	2
AD_AJ1	1.53	1.981	1.351	1.130	1.37	1.53
ADJ1	0.34	0.570	0.421	0.488	0.44	0.39
APO_1	0.34	0.396	0.462	0.291	0.33	0.26
CO_CON1	0.67	0.309	0.278	0.374	0.92	0.88
CO_DET1	0.28	0.327	0.333	0.639	0.27	0.31
CON_IT1	1.29	1.760	0.967	1.053	1.23	1.28
D_1	5.01	3.586	3.910	3.640	5.93	5.33
DE_NU1	0.50	7.192	0.421	0.482	0.48	0.38
DOUB1	0.18	0.000	0.141	0.162	0.14	0.12
IN_IE1	0.17	0.000	0.000	0.094	0.14	0.10
INPA_1	0.48	0.405	0.502	0.458	0.45	0.43
MA_PRO1	1.17	1.151	0.803	0.429	1.10	1.13
MA_SU1	1.41	0.830	0.742	0.734	1.53	1.44
NO_AD1	1.95	0.168	0.348	0.414	1.92	1.94
NO_AU1	0.53	0.168	0.421	0.342	0.52	0.54
NO_CO1	1.16	0.332	0.356	0.573	1.17	1.54
NO_OF1	1.47	0.504	0.381	0.364	1.38	1.56
NO_PRO1	1.11	1.263	0.910	0.604	1.22	1.21
OM_RE1	0.80	0.883	0.696	0.842	0.80	0.75
PER_C1	0.27	0.470	0.303	0.291	0.29	0.28
PEVE_1	0.00	0.000	0.000	0.094	0.00	0.00
PHA_1	0.26	0.519	0.483	0.494	0.26	0.27
PREPO1	1.30	1.228	1.178	0.849	1.23	1.21
SY_1	12.99	14.265	10.146	8.513	12.77	13.13
THER1	0.27	0.327	0.267	0.186	0.24	0.26
THIS1	0.37	0.350	0.472	0.484	0.35	0.34
VER1	0.00	0.000	0.000	0.000	0.00	0.00
W_1	5.34	4.336	4.304	3.547	5.37	5.44
WHI1	0.22	0.138	0.163	0.263	0.24	0.24

TABLE 5. HOTELLING T^2 - STATISTIC OF NINE ARTICLES

Known's Vs Unknowns				
S. No.	**Articles No.**	**Calculated Values**	**Table Values**	**Hypothesis**

1	1 with 8	1.68	1.71	Accept
2	1 with 9	1.59	1.70	Accept
3	2 with 8	1.65	1.72	Accept
4	2 with 9	1.60	1.69	Accept
5	3 with 8	1.57	1.72	Accept
6	3 with 9	1.62	1.71	Accept
7	4 with 8	1.52	1.69	Accept
8	4 with 9	1.49	1.69	Accept
Controls Vs Unknowns				
9	5 with 8	5.48	1.72	Reject
10	5 with 9	9.05	1.70	Reject
11	6 with 8	7.11	1.71	Reject
12	6 with 9	13.70	1.70	Reject
13	7 with 8	11.56	1.72	Reject
14	7 with 9	20.08	1.71	Reject
Unknowns				
15	8 with 9	1.36	1.73	Accept
Known's				
16	1 with 2	1.46	1.69	Accept
17	1 with 3	1.54	1.70	Accept
18	1 with 4	1.65	1.69	Accept
19	2 with 3	1.62	1.69	Accept
20	2 with 4	1.47	1.69	Accept
21	3 with 4	1.66	1.69	Accept

Table value indicates 5 % level of significance

Author Profile

G. Manimannan received his M. Sc. M. Phil. Ph. D in Statistics from University of Madras, Chennai, India. He received PGDCA (Post Graduate Diploma in Computer Application) from Pondicherry University, Pondicherry, India. He has good research experience by working for many Project Guidance and consultation work in application of Statistics. He has published more than twenty seven research papers in various national and International journals. He is good in many programming languages like, FoxPro, HTML, COBOL, C, C++, VB, DBMS, SPSS, SYSSTAT, STATISTICA, MINITAB, MATLAB and working knowledge in SAS and R.

R. Lakshmi Priya received his M. Sc. M. Phil. in Statistics from University of Madras, Chennai, India. She is Working as Assistant Professor in Statistics, Department of Statistics, Dr. Ambedkar Govt. Arts College, Vyasarpadi,Chennai.She has good knowledge in programming languages like, FORTRAN, PASCAL, COBOL, C++, VB and SPSS.