International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 6, Issue 5, Aug – Sep 2017; 81-90

International Academy of Science,
Engineering and Technology
Connecting Researchers; Nurturing Innovations

PATTERN RECOGNITION OF CHILDREN STORIES WITH SPECIAL REFERENCE TO "JACK TALES"

MENAKA SIKDAR[1] & PRANITA SARMAH[2]

[1]Research Scholar, Department of Statistics, Gauhati University, Guwahati, Assam, India

[2]Professor, Department of Statistics, Gauhati University, Guwahati, Assam, India

ABSTRACT

This paper presents a study for two languages, namely Indian and European. The main objective of this study is to pattern recognition of children's stories with special reference to Jack tales, in order to find the distinction among these languages. We consider only the Jack tales, because, they have the same type of ideology among the characters of the stories. In this article, Jack tales are classified into two categories, namely Indian (Assamese and Bengali) and European (English) including thirty seven stories in total. Detailed statistical analyses have been performed by quantifying the texts and presenting them graphically. Non-parametric approaches have been used to test the significant differences among the texts under consideration. It has been shown that there exist significant differences among the writing patterns of these stories written by different authors in both the languages. The run test, confidence bands for empirical distribution functions, Smirnov test and Friedman Test are applied to recognize their writing patterns.

KEYWORDS: Confidence Bands, Empirical Distribution Function, Friedman Test, Non-Parametric Tests, Run Test and Smirnov Test

1. INTRODUCTION

Indian stories are similar to the European stories because in ancient time, all the Aryans resided in the same place of middle Asia and these stories were verbally communicated at that time. But after some time they split all over the world, some of them went to Asian countries and others went to European countries. Their behavior and style were changed according to the time and the places as well as these stories were reshaped by the customs, rituals and culture of that place and time. The structures of the stories may be changed, but the skeleton of the stories along with the moral lesson that are conveyed to the society remain all the more same over time and place. Conventional Indo- European stories may be structurally different in many subjects, but their bases are same. Because of this reason, similarity among the stories under different languages is visible. Children's stories are mainly two types. Some of them are written only for giving moral lessons to our kids- Panchatantra, Hitopadesha are the examples of this category. Others are written to bring them into the world of fantasy as well as they give enormous pleasure to all the persons of different ages, mainly kids. The variation in the writing patterns of stories with respect to different languages, cultures and authors seems to be quite natural. The skeleton of the stories being similar, it is difficult to comment whether there exists any significant difference amongst them in their first look. The children's stories are mainly classified into four different categories, namely fables, fairy tales, formula tales and Jack tales. Understanding of the complexity associated with the pattern of writing offered by different authors in various languages is not an easy task. All kinds of mathematical tools were adopted to gain understanding of the complexity of human language. Quite a number of research scholars, namely Marco Turchi et al (2006), Agarwal et al (2014), Bahl et al (1983), Peter et al (1992), Sveta zinger (2006), Brown et al. (1990) and Mays et al (1990) has attempted

to work on pattern recognition on language models. Sikdar and Sarmah (2017A,2017B 2017 C, 2017 D) presented four articles in three languages, namely Assamese, Bengali and English for pattern recognition of language model with special reference to children's stories. This article is about Jack tales. According to Wikipedia, Jack is an archetypal Cornish and English hero and stock character appearing in legends, fairy tales and nursery rhymes. A Jack tale is a category of 'folk tale', in which they usually have a character portrayed as a young adult called Jack, who appears lazy or stupid but actually wins in the end because he is 'tricky. In this way, he may resemble a trickster. These types of stories are enjoyed by people all over the world, irrespective of their age. In this article Jack tales are classified into two categories, namely Indian (Assamese and Bengali) and European (English) including thirty seven stories in total. The parameters that will actually help us to recognize the pattern of the children's stories with reference to Jack tales are- (i) total number of words contained in a story (ii) total number of sentences contained in a story (iii) mean number of words per sentence of a story (iv) range of the size of sentences of a story. The next section of this article is about objectives of the study. Section 3 includes the source of data, whereas section 4 of this article includes material and methods for statistical presentation the texts under consideration. Finally section5 is devoted to Statistical analyses.

2. OBJECTIVES OF THE STUDY

The main objectives of our study are

a) To recognize the patterns of writings Jack tales presented by different authors in Indian and European languages.

b) To obtain the confidence bands for the empirical distribution functions of the random variables (i),(ii),(iii) and (iv) (as mentioned in **section I**).

c) To study the significant differences between the distributions functions of the random variables (i), (ii), (iii) and (iv) under the above mentioned Indian and European languages.

d) To study the significant differences between the effects of various types of sentences in the written text in two different languages.

3. SOURCES OF DATA

For our statistical analysis purpose, eighteen Indian (Assamese and Bengali) Jack tales and nineteen European (English) Jack tales are considered. The data have been collected from the stories by Sahityo –rothi Laxminath Bezbarua, Upendrakishore Roy Choudhury,the Gmimm brothers and Hans Christian Andersen. Thirteen Assamese Jack tales from "Burhi Aai'r Xaadhu" (literary translated to Grandma's Tales), five Bengali Jack tales from "Tuntunir Boi" (Book of the tailor-bird), sixteen English Jack tales from Grimm's fairy tales and three English Jack tales from Andersen's fairy tales are selected to recognize their patterns.

4. MATERIALS AND METHODS

Nonparametric techniques are used extensively to understand the patterns of Jack tales written in Indian and European languages. Run test has been conducted to test the randomness of the data obtained from different languages. The concept of empirical distribution functions along with confidence bands is studied to understand the probabilistic structures of the distributions under study. The Smirnov test (Two sample kolmogorov Smirnov test) is adopted to study the significant differences between the distributions under two languages. The Friedman test is applied to understand the effect of various types of sentences in the written texts.

4.1 Some Definitions

A word is a basic element in every language with proper combination of letters arranged in such a manner that they should represent either objects or ideas.

Let $w_{ij}(k,l)$ be the j^{th} word in the i^{th} sentence of k^{th} Jack tale under l^{th} language $\forall i=1,2,.....p\ j=1,2....q,\ k=1,2...n$ and $l=1,2,...,t$.

Here $w_{kl}=\sum_i\sum_j w_{ij}(k,l)$ is **the total number of words** in k^{th} Jack tale described under l^{th} language.

$w_i(kl)=\sum_j w_{ij}(k,l)$ is the total number of words in i^{th} sentence of the k^{th} Jack tale described under l^{th} language. Hence $w_{kl}=\sum_i w_i(k,l)$

A sentence is a function of words which makes complete sense. Placing of words at different positions of the sentence, use of proper part of speech, use of phrases common to a culture, place and language present the writing style of a story which adds a flavor to the story. Sentences in children's stories can be filtered into three categories, namely simple, complex and direct speech.

a) **Simple Sentence**→ Sentences having single action.

b) **Complex Sentence**→ Sentences having more than one action.

c) **Direct Speech**→ Direct speech refers to the quoted words of a character created by the narrators.

S_{kl} is the **total number sentences** in the k^{th} Jack tale under l^{th} language.

$\frac{w_{kl}}{s_{kl}}=m_{kl}$ $(k=1,2,...,n$ and $l=1,2,...,t)$ is **the mean number of words per sentence** of the k^{th} Jack tale under l^{th} language.

Range of the size of sentences in the k^{th} Jack tale under l^{th} language is the difference between the maximum and minimum size of sentences of that particular Jack tale.

Here, $w_i(kl)=\sum_j w_{ij}(k,l)$ is the size of the i^{th} sentence in the k^{th} Jack tale under l^{th} language.

Therefore, $r_{kl}=Max\ w_i(kl)-Min\ w_i(kl),\ i=1,2,...,p,k=1,2,...,n$ and $l=1,2,...,t$ represents **range of the size of sentences** of k^{th} Jack tale under l^{th} language.

4.2 Some Statistical Techniques for Analyzing the Patterns of Jack Tales

4.2.1 Run Test

Run test has been applied to test the randomness of the distributions under study. For the distribution of total number of words of different Jack tales under l^{th} language, our data consist of set of observations

$w_{kl};\ k=1,2...n$ and $l=1$(Indian),2(European). However, when the distribution of the observed data is unknown, the hypotheses may be considered as

H_0: Sequence obtained from total number of words of different Jack tales under l^{th} language is random.

H_1: Sequence is not random.

Using the median (M_d) of all observations as a focal point, the dichotomy for the ordinary run tests compares each observation with the median. Two types of ordered sequences with $n_1\ (\geq M_d)$ and $n_2(<M_d)$ are considered such that

$n = n_1 + n_2$. Let r_1 and r_2 be number of runs obtained from the observations n_1 and n_2 respectively. Here the test statistics is $R\ (=r_1+r_2)$. The probability distribution of $R\ (=r_1+r_2)$ is given by

$$f(r) = 2\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}\Big/\binom{n_1+n_2}{n_1} \text{ if } r \text{ is even}$$

$$=\left[\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_1-1}{(r-3)/2}\binom{n_2-1}{(r-1)/2}\right]\Big/\binom{n_1+n_2}{n_1} \text{ if } r \text{ is odd}$$

When both n_1 and n_2 are large, we can get the normal approximation for the null distribution of $R\ (=r_1+r_2)$, for which mean and variance are given by (Gibbons 1971)

$$\mu = 2n_1n_2/n+1 \quad \text{and} \quad \sigma^2 = 2n_1n_2(2n_1n_2-n)/n^2(n-1)$$

If $n_1 \leq 20$, the run test has been conducted by using the critical values found in the Run Test Table.

4.2.2 Empirical Distribution Function and Confidence Bands

The empirical distribution function is based on a random sample that may be used to estimate the unknown population distribution function. In case of the distribution of total number of words of different Jack tales under l^{th} language, our data consist of a random sample w_{kl}; k=1, 2…n. The empirical distribution function, $F_l(w)=(number\ of\ w_{kl} \leq w)/n$, where $k=1,2,\ldots,n,\ l=1,2$.

Confidence Bands: This is the application of the Kolmogorov-Smirnov statistic in forming the confidence bands for an unknown distribution function $F_l(w)$. To form a confidence band for $F_l(w)$, we basically need to find a confidence interval for each value of $w.100(1-\alpha)\%$ confidence bands for the unknown distribution function $F_l(w)$ is given by $U_l(w)$ and $L_l(w)$

$U_l(w)=F_l(w)+d_{1-\alpha}$ if $F_l(w)+d_{1-\alpha} \leq 1$ and $U_l(w)=1$ if $F_l(w)+d_{1-\alpha} > 1$

$L_l(w)= F_l(w)-d_{1-\alpha}$ if $F_l(w)-d_{1-\alpha} \geq 0$ and $L_l(w)=0$ if $F_l(w)-d_{1-\alpha} < 0$

Here $d_{1-\alpha}$ is the $1-\alpha$ quantile of the Kolmogorov statistic for sample size n from table *13* for two-sided test. [Table *13* is given in "Conover W.J. *(2006)*"]

The resulting probability statement is $P=[L_l(w) \leq F_l(w) \leq U_l(w)]$, for all $w] \geq 1-\alpha$

Similarly we can obtain the empirical distribution functions and confidence bands for the distributions namely (ii), (iii) and (iv) under different languages.

4.2.3 Smirnov Test (Also Known as Two Sample Kolmogorov and Smirnov Test)

The Smirnov test has been developed to examine whether the distribution functions obtained from two different populations are identical or not. In this case the maximum vertical distance between two empirical distribution functions as a measure of how well the functions resemble each other are used. Let $w_{k1};\ k=1,2,\ldots,\ n_1$ and $w_{k2};\ k=1,2,\ldots,\ n_2$ be the two samples of total number of words of different Jack tales under both the languages which are associated with some unknown distribution functions $Q_1\ (w)$ and $Q_2\ (w)$ respectively.

Test Statistic

Let $F_1(w)$ and $F_2(w)$ be the empirical distribution functions of the total number of words of different Jack tales under language one and two respectively. The test statistic for the two-sided test is defined as

$$T = \sup_{w} |F_1(w) - F_2(w)|$$

Null Distribution

The exact null distribution of T is obtained by considering all ordering of w_1's and w_2's to be equally likely under the null hypothesis and computing T, as appropriate, for each ordering. Quantiles of the null distribution are given in table *19* for equal sample sizes and table *20* for unequal sample sizes. [Table *19* and Table *20* are given in "Conover W.J. (2006).]

Hypothesis: (Two sided test)

$\boldsymbol{H_0}$: $Q_1(w) = Q_2(w)$ for all w from $-\infty$ to $+\infty$

$\boldsymbol{H_1}$: $Q_1(w) \neq Q_2(w)$ for at least one value of w

Above procedure and hypotheses are adopted for the distributions, namely (ii), (iii) and (iv) respectively.

4.2.4 Friedman Test

In order to understand the effect of various types of sentences in writing text, we have adopted the Friedman test which is a two-way analysis of variance on ranks. In our analysis, we have considered the different Jack tales as blocks, and various types of sentences namely simple, complex and direct speech as treatments. Let X_{ij} be the total number of sentences associated with type (treatment) j in i^{th} Jack tale (block) under a particular language, $j=1,2,3$ and $i=1,2,\ldots,n$.

The 'n' blocks are arranged as follows

Block (Jack Tale)	Number of Sentences		
	Treatment(Types of Sentences)		
	Simple(1)	Complex(2)	Direct Speech(3)
1	X_{11}	X_{12}	X_{13}
2	X_{21}	X_{22}	X_{23}
3	X_{31}	X_{32}	X_{33}
…	…	…	…
n	X_{n1}	X_{n2}	X_{n3}

Let $\rho(X_{ij})$ be the rank,from *1* to *3*, assigned to X_{ij} within block (story) i, i.e $\rho(X_{ij}) = j$, if the value of the rank is j, $j=1,2,3$. Average ranks are used in case of ties.

Then the sum of ranks for each treatments are obtained and given as

$\rho_j = \sum_{i=1}^{n} \rho(X_{ij})$ for $j=1,2,3$.

Test Statistic

Friedman suggested using the statistic

$T_1 = \frac{12}{nk(k+1)} \sum_{j=1}^{k} \left(\rho_j - \frac{n(k+1)}{2}\right)^2$, where k is the number of treatments.

The statistic T_1, adjusted for the presence of ties, becomes

$$T_1 = \frac{(k-1)\left[\sum_{j=1}^{k} \rho_j^2 - nC_1\right]}{A_1 - C_1} = \frac{(k-1)\sum_{j=1}^{k}\left(\rho_j - \frac{n(k+1)}{2}\right)^2}{A_1 - C_1}$$

Where A_1 be the sum of squares of the ranks and C_1 be the correction factor.

$$A_1 = \sum_{i,j}\left[\rho(X_{ij})\right]^2 \text{ and } C_1 = nk(k+1)^2/4$$

The preferred test statistics is $T_2 = \frac{(n-1)T_1}{n(k-1)-T_1}$

Null Distribution

The approximate distribution of T_1 is the chi-squared distribution with *k-1* degrees of freedom. The approximate distribution of T_2 is the *F* distribution with K_1= *k-1* and k_2= *(n-1) (k-1)* degrees of freedom, when the null hypothesis is true.

Hypotheses

H₀: The effects of three types of sentences within the Jack tales are identical.

H₁: At least one type of sentence tends to be written more in the Jack tales than others.

Multiple Comparisons

When the null hypothesis is rejected, we may use the following procedure to determine which pairs of treatments (types of sentences) tend to differ. The treatments say i and j seem to be different if the following inequality is satisfied

$$\left|\rho_i - \rho_j\right| > t_{1-\alpha/2}\left[\frac{(A_1 - C_1)2n}{(n-1)(k-1)}\left(1 - \frac{T_1}{n(k-1}\right)\right]^{\frac{1}{2}}$$

Where $t_{1-\alpha/2}$ is the *1-α/2* quantile of the *t* distribution with *(n-1)(k-1)* degrees of freedom.

If there are no ties, A_1 and (A_1-C_1) simplify to $A_1=nk(k+1)(2k+1)/6$ and $A_1-C_1=nk(k+1)(k-1)/12$

5. STATISTICAL ANALYSES AND RESULTS

Results of the run test are obtained by using SPSS software and given in the following table

Table 1: Results of the Run Test

Language	Distribution	size	Median(M_d)	n_1 ($\geq M_d$)	n_2 ($< M_d$)	Number of Runs	Critical Value at 0.05 Level	Critical Value at 0.01 Level
Indian	(i)	18	780	9	9	10	[5,15]	[4,16]
	(ii)	18	66	9	9	10	[5,15]	[4,16]
	(iii)	18	12	12	6	4	[4,13]	[3,-]
	(iv)	18	30	10	8	8	[5,15]	[4,16]
English	(i)	19	1333	10	9	8	[5,16]	[4,17]
	(ii)	19	55	11	8	10	[5,15]	[4,16]
	(iii)	19	23	10	9	9	[5,16]	[4,17]
	(iv)	19	56	10	9	8	[5,16]	[4,17]

From Table one, it has been noticed that the observed values of the total number of runs for the distributions of (i),(ii),(iii)(except for Indian Jack tales) and (iv) under both the languages are lies between the intervals of critical values at *5%* level of significance. Therefore, we may accept our null hypotheses at *the 5 % level* of significance and may conclude

that these distributions are random. On the other hand, observed value of a run for the distribution (iii) under Indian Jack tales is equal to the small value of the pair of critical values at *the 5 % level* of significance, hence we reject the null hypothesis at *the 5 % level* of significance. Again, it is greater than the smallest value of the pair of critical values at *the 1 % level* of significance. The null hypothesis may be accepted at *0.01* levels. Therefore the distribution (iii) under Indian Jack tales is also random.

The confidence bands for empirical distribution functions are given in the following table

Table 2: Confidence Bands

Language	Distribution	Sample size	$d_{0.95}$	95% Confidence Bands for Empirical Distribution Function
Indian	(i) Total number of words	18	0.309	$F_1(w)\pm0.309$
	(ii) Total number of sentences	18	0.309	$F_1(s)\pm0.309$
	(iii) Mean number of words per sentence	18	0.309	$F_1(m)\pm0.309$
	(iv) Range of the size of sentences	18	0.309	$F_1(r)\pm0.309$
English	(i) Total number of words	19	0.301	$F_2(w)\pm0.301$
	(ii) Total number of sentences	19	0.301	$F_2(s)\pm0.301$
	(iii) Mean number of words per sentence	19	0.301	$F_2(m)\pm0.301$
	(iv) Range of the size of sentences	19	0.301	$F_2(r)\pm0.301$

The empirical distribution functions $F_1(i)$ and $F_2(i)$ for $i=w, s, m, r$ along with their *95%* confidence bands are presented below.

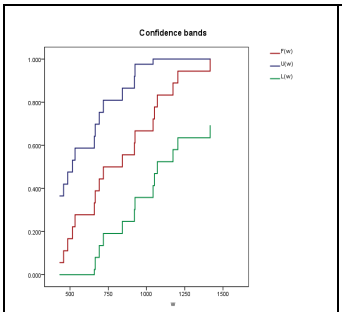

Figure 1

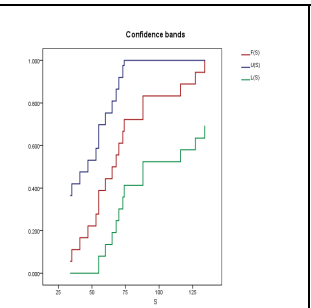

Figure 2

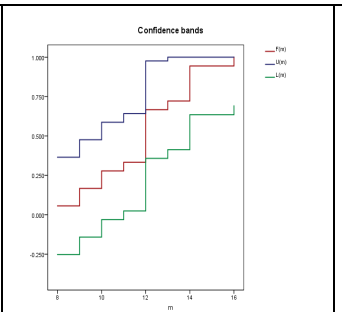

Figure 3

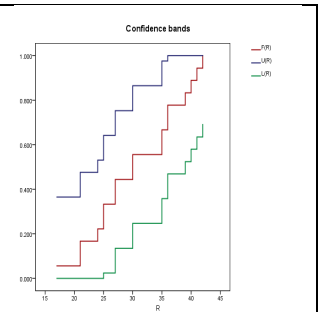

Figure 4

[The empirical distribution functions along with *95%* confidence bands of the distributions of (i),(ii),(iii) and (iv) under Indian Jack tales are represented in Figure *1,2,3* and *4* respectively obtained by using SPSS soft ware.]

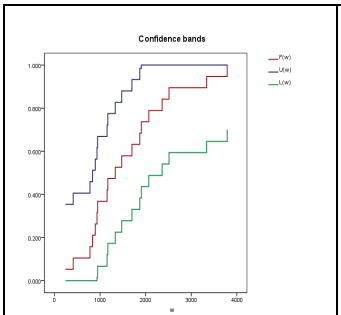

Figure 5

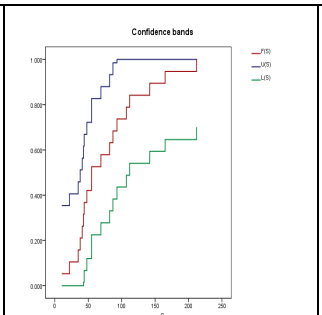

Figure 6

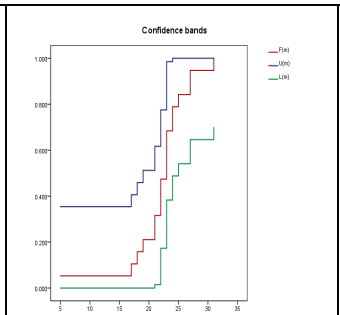

Figure 7

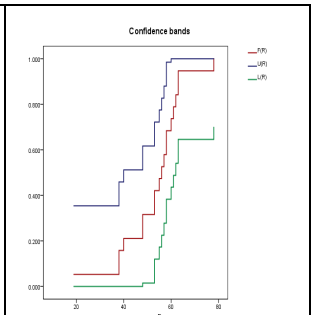

Figure 8

[The empirical distribution functions along with *95%* confidence bands of the distributions of (i),(ii),(iii) and (iv) under English Jack tales are represented in Figure *5,6,7* and *8* respectively obtained by using SPSS soft ware.]

Results of the Smirnov test are obtained by using **SPSS soft-ware** and are given in table*3*

Table 3: Results of Smirnov Test

Language	Distribution	Absolute Difference (D)	Positive Difference (D^+)	Negative Difference (D-)	K.S. Z	p Value
Indian and English	(i) Total number of words	0.474	0.105	-0.474	1.440	0.032
	(ii) Total number of sentences	0.202	0.202	-0.158	0.613	0.846
	(iii) Mean number of words per sentence	0.947	0.053	-0.947	2.880	0.000
	(iv) Range of the size of sentences	0.789	0.000	-0.789	2.400	0.000

From table three, it has been noticed that the *p*-values of the test statistics for the distributions of (i) and (ii) are greater than *0.01* and *0.05* respectively, and our null hypotheses may be accepted at *1%* and *5%* level of significance respectively. Therefore, we may conclude that the distributions of (i) and (ii) under Indian and English languages are not significantly different. On the other hand, the *p*-value of the test statistics for the distributions of (iii) and (iv) are less than *0.01*. Therefore the null hypotheses may be rejected at *the 1 % level* of significance. Hence we may conclude that the distributions of (iii) and (iv) under Indian and English languages are significantly different.

Results of Friedman test are obtained by using SPSS software and are given in table *4.*

Table 4: Results of Friedman Test

Language	Number of Jack Tales	Types of Sentences	Sum of Ranks	Mean Rank	Test Statistic T_1	Test Statistic T_2 (F)	d.f. (k_1,k_2) for T_2	$F_{0.95}(k_1,k_2)$	Inference
Indian	18	Simple	20.5	1.14	20.366	22.146	(2,34)	3.284	Null hypothesis is rejected
		Complex	43	2.39					
		Direct Speech	44.5	2.47					
English	19	Simple	19	1.00	31.684	90.296	(2,36)	3.266	Null hypothesis is rejected
		Complex	42	2.21					
		Direct Speech	53	2.79					

From Table four, it has been noticed that the calculated values of the test statistics T_2 under Indian and English languages are greater than the *0.95* quantile of the *F* distribution with respective degrees of freedom. Therefore, we may reject our null hypotheses at *the 5 % level* of significance and may conclude that the effects of three types of sentences within the Jack tales are not identical in Indian and English languages. However, when such a null hypothesis is rejected, it is a normal practice to perform **Multiple Comparisons Procedure** to determine which pairs of treatments tend to differ.

Calculations for multiple comparisons are given in Table *5***.**

Table 5: Results of Multiple Comparisons under Friedman Test

Language	Type of Sentence(Treatment)	$\lvert\rho_i - \rho_j\rvert$	$t_{1-\alpha/2}\left[\frac{(A_1 - C_1)2n}{(n-1)(k-1)}\left(1 - \frac{T_1}{n(k-1)}\right)\right]^{\frac{1}{2}}$	Inference
Indian	Simple & Complex	22.5	8.2178	Significantly different
	Simple & Direct speech	24		Significantly different
	Complex & Direct speech	1.5		Not significantly different
English	Simple & Complex	23	5.2389	significantly different
	Simple & Direct speech	34		Significantly different
	Complex & Direct speech	11		Significantly different

6. CONCLUSIONS

Jack tales are immensely popular amongst children and elders. Though they have the same type of ideology among the characters of the stories, a striking dissimilarity is observed among the writing patterns of authors under two different languages namely Indian and European. The table three revels that the distributions of total number of words and sentences are not significantly different under both the languages, whereas the distributions of mean number of words per sentences and range of the size of the sentences differ significantly. It is interesting to note that, table, four and five show that for Indian languages, the effects of complex sentences and direct speeches within the Jack tales are not significantly different whereas the result is completely opposite for simple sentences. However, the effects of various types of sentences within the Jack tales in English are significantly different from each other.

REFERENCES

1. Duda Richard O., Hart Peter E. Stork David G.*(2000),* Pattern Classification *(2nd ed.)*, John Willey and Sons Inc.
2. Conover W.J. *(2006)*, Practical Nonparametric Statistics *(3rd ed.)*, John Willey and Sons Inc.
3. Mukhopadhyay Parimal *(2000)*, Mathematical Statistics *(2nd ed.)*, Books and Allied (P) Ltd
4. Gun A.M., Gupta M.K., Dasgupta B. *(2005)*,An Outline of Statistical Theory, vol *2 (3rd ed.)*, The World Press Private Limited.
5. Turchi M., Cristianini N. "A Statistical Analysis of Language Evolution" In proceeding of Evolution of Language Sixth International Conference Rome, *12-15* April *2006*
6. Agrawal Shyam S., Mandal Abhimanue, Bansal Shweta, Mahajan Minakshi, "Statistical Analysis of Multilingual Text Corpus and Development of Language Models" In proceeding of the Ninth International Conference on language Resources and Evaluation (LREC) Iceland, *26-31* May,*2014*.
7. Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jennifer C. Lai, Robert L. Mercer, "Class-Based n-gram Models of Natural Language". *Computational Linguistics 18 (4*): *467-479 (1992).*
8. Zinger Sveta, "Statistical Natural Language Processing: N-Gram models", Seminar in Methodology and Statistics, Rijksuniversiteit Groningen, *15th* March, *2006.*
9. Shannon, C. E. *(1951)* "Prediction and entropy of printed English". Bell Systems Technical Journal *(30), 50-64*.
10. Bharthi Akshar, Sangal Rajeev and Bendre Sushma M, "Some Observations Regarding Corpora of Indian Languages" Proceedings of KBCS-*98, 17-19* Dec *1998*, Mumbai.
11. Bansal Shweta, Mahajan Minakshi, Agrawa S.S. l, "Determination of Linguistic Differences and Statistical Analysis of Large Corpora of Indian Languages" OCOCOSDA,Nov. *2013*, Gurgaon, India.
12. *www.wikipedia.org*
13. Bahl L.R.,Jelinek F.,Mercer R.L. (*1983*) " A Maximum Likelihood Approach to Continuous Speech Recognition" IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-*5 (2)*, *179-190.*
14. Mays E., Damerau F.J. and Mercer R.L.(*1990*) " Context-based spelling correction", In proceedings, IBM Natural Language ITL, Paris, France, *517-522*

15. *www.real-statistics.com/non-parametric-tests/one-sample-runs-test*

16. *www2.nau.edu/~shuster/shustercourses/BIO%20682/.../runs%20test%20values.pdf*

17. Sikdar M. & Sarmah P., (*2017A*), "Pattern Recognition In Language Model With Special Reference To Children Stories", International Journal of Innovative Research and Advanced Studies (IJIRAS), Volume*4* Issue*3*.

18. Sikdar M. & Sarmah P., (*2017B*) "Statistical Pattern Classification Of Direct Speeches In Children Stories", International Journal of Applied Mathematics and Statistical Sciences, ISSN (P): *2319-3972*; ISSN (E): *2319-3980*, Vol.*6*, Issue *5*, Aug-Sep *2017; 7-18*.

19. Sikdar M. & Sarmah P. (*2017C*), "Pattern Recognition Of Children Stories based on themes" will be presented in International conference on Recent Trends in Operational Research and Statistics (RTORS), December *28-30, 2017*, Department of Mathematics, Indian Institute of Technology Roorkee, Roorkee, Uttarakhand.

20. Sikdar M. & Sarmah P. (*2017D*), "Pattern Recognition Of Children Stories: A Non-parametric Approach" has been accepted for presentation at International conference on Advancing Frontiers in Operational Research: Towards a Sustainable World' (AFOR*2017*), *21-23* December *2017*, Kolkata. [21]*http://www.ncpedia.org/culture/stories/Jack-tales*