Statistical Analysis of Indian Monsoon Data

Rohit Toshniwal,
Department of Computing ,
Dublin City University, Dublin, Ireland,
rohit.toshniwal2@mail.dcu.ie
Student ID: 19211138

Raj Singh,
Department of Computing ,
Dublin City University, Dublin, Ireland,
raj.singh5@mail.dcu.ie
Student ID: 19210213

***Abstract*— The monthly rainfall data for 140 years is used to understand normal rainfall, deficit rainfall, Excess rainfall and Seasonal rainfall of the selected areas in India. Further various plotting position formulae available are used to evaluate the return period of monthly, seasonally and annual rainfall. This analysis will provide helpful information for water resources planners, farmers and urban designers to evaluate the availability of water and create the storage accordingly. Some methods of statistical analysis are described and are illustrated using monthly rainfall records of years from 1870 to 2016 for India. The mean, standard deviation, and coefficient of variation of monthly and annual rainfall are calculated to check the rainfall variability. From the calculated results, the rainfall pattern is found to be inconsistent. Finding some pattern in the occurrence of Indian rainfall will surely help in predicting rainfall.**

Keywords—Monsoon, India, Rainfall, Months, JJAS, ANN

I. INTRODUCTION

The primary source of agricultural production for most of the country is rainfall. Efficient Utilization of rainfall improves crop growth and development. About 80% of the world and 60% of Indian Agriculture is rain-dependent. Frequency or probability distribution helps to relate the magnitude of the extreme events like floods, droughts and severe storms with their number of occurrences such that their chance of occurrence with time can be predicted easily.

Flood during the monsoon season is a continuing problem in some parts of India. To manage flood including its prediction and analysis, we need to understand the characteristics of monsoon distribution over the country and the trend it follows. Besides that, nearly 80% of the total annual rainfall occurs during this period. As Indian economy is mostly dependent upon rainfall, we need to understand the nature of the distribution of monsoon rain to decide the agricultural strategy. It should also be stressed that many industries need water and to decide the location of such industries, we should have a clear idea about the amount of rainfall India can expect every year.

We have monthly data of the last 120 years, about 80% of the rainfall in India occurs during the four monsoon months (June–September) with large spatial and temporal variations over the country, so we are analyzing JJAS data. Suchit Kumar Rai et al., studied the change, variability and rainfall probability for crop planning in few districts of Central India [1]. Nyatuame et al. [2] performed the statistical analysis and studied the variability in the distribution of rainfall. Rajendran et al. [3] carried out the frequency analysis of rainy days and studied the rainfall variation.

II. RELATED WORK

Incidentally, some works have already come out featuring different aspects of monsoon rain in India like Probability and frequency analysis of rainfall data enable us to determine the expected rainfall at various chances (Bhakkar et al., 2008). Studies carried out by several investigators have shown that the trend and magnitude of warming over India/the Indian sub-continent over the last century is broadly consistent with the global trend and magnitude (Hingane, 1995; Pant & Kumar, 1997, Arora et al., 2005, Dash et al., 2007). Pant & Kumar (1997) analyzed the seasonal and annual air temperatures from 1881–1997 and have shown that there has been an increasing trend of mean annual temperature, at the rate of 0.57C per 100 years.

Some past studies relating to changes in rainfall over India have concluded that there is no clear trend of increase or decrease in average annual rainfall over the country (Mooley & Parthasarathy, 1984; Sarker & Thapliyal, 1988; Thapliyal & Kulshrestha, 1991; Lal, 2001). Though no trend in the monsoon rainfall in India is found over a long period, particularly on the all-India scale, pockets of significant long-term rainfall changes have been identified (Koteswaram & Alvi, 1969; Jagannathan & Parthasarathy, 1973; Raghavendra, 1974; Chaudhary & Abhyankar, 1979; Kumar et al., 2005; Dash et al., 2007; Kumar & Jain, 2009).

Recent studies (Khan et al., 2000; Shrestha et al., 2000; Mirza, 2002; Lal, 2003; Min et al., 2003; Goswami et al., 2006; Dash et al., 2007) show that, in general, the frequency of more intense rainfall events in many parts of Asia has increased, while the number of rainy days and total annual amount of precipitation has decreased. A change detection study using monthly rainfall data for 306 stations distributed across India was attempted by Rupa Kumar et al. (1992). They showed that areas of the northeast peninsula, northeast India and northwest peninsula experienced a decreasing trend in summer monsoon rainfall.

III. DATASET AND EXPLORATORY ANALYSIS

3.1 Dataset Information

Sub-divisional monthly rainfall data of India prepared by the Indian Institute of Tropical Meteorology (IITM: htpp://www.tropmet.res.in) were used in this study. We gathered our Rainfall data from the official site of IITM(Indian Institute of Tropical Meteorology). Before releasing the data, the IITM carries out quality checks to ensure that error-free data are used in analysis and design. Thus the quality of this data set is very good and it is one of the most reliable long series of data. We are using monthly data for the past around 140 years from 1870 to 2016. To investigate the changes in rainfall for different seasons, a year was divided into four seasons: winter (December–February), pre-monsoon (March–May), monsoon (June–September), and post-monsoon (October–November). We are using only Monsoon season data for the statistical analysis, so we could predict and analyse the rainfall. Also, we are using total of monsoon season as JJAS and annual data for better understanding and comparison purpose. In case of monthly or seasonal study, we are considering only that data set which has full data for the entire season.

3.2 Exploratory Data Aanalysis

3.2.1 Basic statistical properties

Basic statistics, such as minimum, maximum, mean, count, standard deviation of monsoon(June–September), JJAS, annual rainfall of data set is given in Tables 1.

Interquartile Range (IQR): This helps us detect where most of the rainfall data lies. IQR is expressed as:

IQR = Q1−Q3

Generally, it is preferred to use IQR instead of the mean or median when trying to find out where most of the rainfall data lies. The IQR gives us a measurement of how spread out the entirety of our data set is. The interquartile range, which tells us how far apart the first and third quartile are, indicates how spread out the middle 50% of our set of data is.

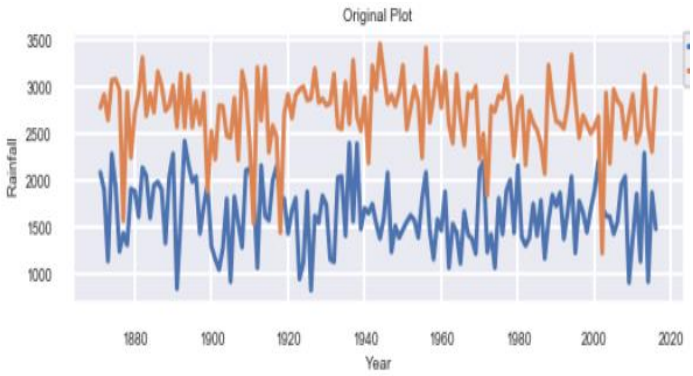
Note: JJAS is nothing but sum of June, July, August and September rainfall data and ANN is annual rainfall data

YEAR	JUN	JUL	AUG	SEP	JJAS	ANNUAL
Count	146.000	146.000	146.000	146.000	146.000	146.000
Mean	1630.575	2725.3493	2422.3	1703.397	8481.62	10859.09
Std	63.9908	374.8062	376.96	367.1681	834.522	1013.681
Min	815.00	1213.00	1441.00	773.00	6040.0	8109.00
Q1	1384.00	2555.000	2169.75	1398.500	7932.750	10163.75
Q2	1613.500	2788.00	2416.5	1697.500	8585.500	10879.50
Q3	1896.500	2944.50	2711.2	1986.75	9054.50	11554.75
IQR	512.5	389.5	501.45	588.25	1121.75	1391.00
Max	2416.00	3460.00	3393.0	2678.00	10202.0	13470.00

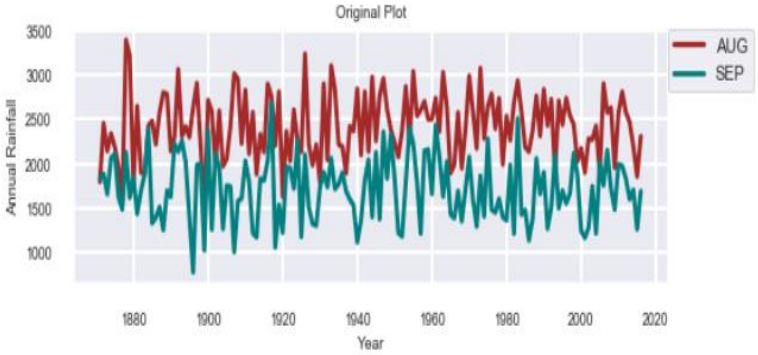
<Table 1: Details of statistical properties of Monthly, JJAS and Annual rainfall>

3.2.2 Frequency distribution

Figure 1 and 2 shows the frequency distribution of rainfall data for June to September for the country and the entire study area. So, for this distribution graph, X-axis is used for Year and Y-axis for rainfall values in mm. It is giving a clear visualization of variation and fluctuations of rainfall in each month of monsoon season.



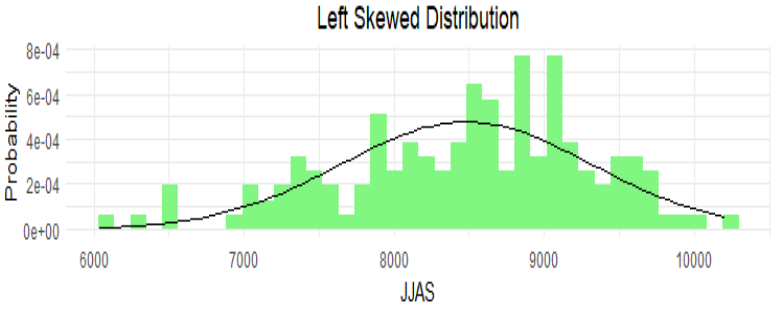
<Figure 1: Rainfall frequency distribution for June and July month>



<Figure 2: Rainfall frequency distribution for August and Sept month>

3.2.3 Skewness and Kurtosis

Table 2 shows that highest value of Skewness as well as Kurtosis occurs in September. Kurtosis varies between -0.5 and 2.5 for the other five divisions. So except September, the density distribution for the seasonal rainfall for the remaining is more towards normal distribution. This is also evident from the value of Skewness. The Figure 3 indicates that rainfall in the month of JJAS is left skewed. It is an indication that both the mean and the median are less than the mode of the data set. So, the rainfall tends to be high in for the months of JJAS in the dataset.



<Figure 3: Skewness distribution of overall monsoon (JJAS)>

Division	Skewness	Kurtosis	Comment
June	-0.05736	-0.61906	Significant
July	-1.18178	2.63707	Not Significant
August	0.023047	-0.43227	Not Significant
Sept	0.107345	-0.52665	Not Significant
JJAS	-0.51715	-0.04576	Significant

<Table 2 Skewness and Kurtosis of seasonal rainfall>

3.2.4 Confidence interval for mean seasonal rainfall

Finally, a comparison of the lengths of 95% confidence intervals [Table 3] for mean rainfall, for the mentioned four months and JJAS suggests that the best interval has been obtained for June, though the lengths of the respective confidence intervals for the other months are not very poor. In this respect, it is worth mentioning that the length of the interval can be used as an inverse measure of precision of the interval estimate; of the two confidence intervals, the one having the smaller length is obviously preferable.

Month	*95 % CI For*	*LOWER LIMIT (mm)*	*UPPER LIMIT (mm)*	*Range(mm)*
JUNE	*MEAN*	*1571*	1690.1	*119.1*
JULY	*MEAN*	*2664*	2786.7	*122.7*

AUGUST	***MEAN***	**2360.7**	**2484**	***123.3***
SEPTEMBER	***MEAN***	**1643.3**	**1763.5**	***120.2***
JJAS	***MEAN***	**8345.1**	**8618.1**	***273***

<Table 3 Confidence interval for mean Monsoon period rainfall >

3.2.5 Correlation of JJAS and ANN using different methods

We have used Correlation to measure the strength and direction of the association between two variables. Two variables are JJAS and ANN. There are two methods of correlation: the Pearson product moment correlation and the Spearman rank order correlation. As our data is not linearly related, so we used Spearman rank order correlation. Define relation between JJAS and ANN.

Method type: Spearman correlation

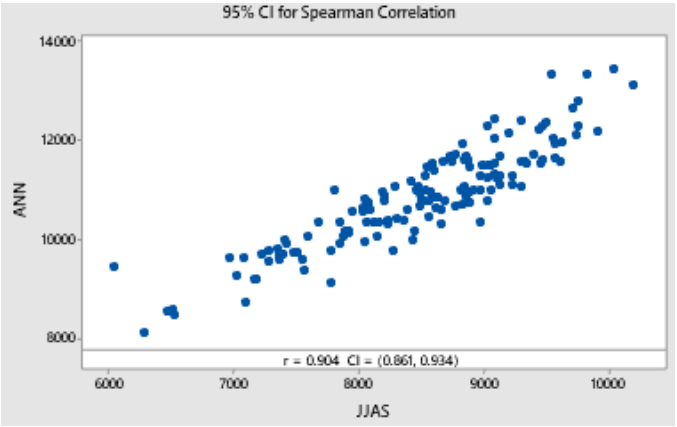
Number of rows used: 146

Correlations:

	JJAS	**Comment**
ANN	**0.904**	**A very Strong Correlation**

<Table 4 Correlation value of JJAS and ANN>

If the correlation value is between 0.90 and 1.00, then there is a very strong correlation between the mentioned variables. So, as per the Table 4, correlation of JJAS and ANN is very strong in strength. Figure 4 shows the matrix plot of JJAS and ANN for 95% confidence interval, as after calculation.



<Figure 4 Scatter plot of correlation between JJAS and ANN>

3.2.6 Regression analysis of rainfall for JJAS and ANN

Linear regression is the most widely used statistical technique. It is a way to model relationship between two sets of variables. We used Annual rainfall in ANN to predict rainfall in JJAS. We calculated a scatter plot to check if our data fits in roughly with a line. Because regression will always give you an equation, and it may not make any sense if your data is scattered exponentially. Our regression equation is:

Regression Equation

$$JJAS = 372 + 0.7468\ ANN$$

Linear regression equation is represented by Y = a +bX where Y is dependent variable, and X is independent variable (predictor), b is the slope of the line and a is y-intercept.

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Term	Coefficient	P-Value
Constant	a = 372	0.239
ANN	b = 0.7468	0.000

For this, R^2 (R-sq.) = 82.28%, which means that the independent variable, ANN, explains 82.28% of the variability of the dependent variable, JJAS. Adjusted R^2 is also an estimate of the effect size, which at 72.1%, is indicative of a large effect size according to Cohen's (1988) classification. In this example, the regression model is statistically significant, p < .0005. This indicates that, overall, the model applied can statistically significantly predict the dependent variable, JJAS.

R^2 is calculated by the using the formula:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

ANN Values	**JJAS predicted values as per Equation**
11493	**11493**
10949	**8548.76**
11582	**9021.47**
11092	**8655.55**
11567	**9010.27**

<Table 5 JJAS Predicted values from ANN values>

IV. HYPOTHESES AND RESEARCH QUESTION

Research Question: Can we predict India's monsoon period i.e. JJAS rainfall data using ANN i.e. annual rainfall data**? (Refer 3.2.6)**

Hypothesis Testing 1

There are two types of hypothesis the Null Hypothesis and the Alternative Hypothesis.

Null Hypothesis (H0): We don't know if there is any correlation between rainfall and JJAS and ANN. It means there is no association between these two variables. Hence,

$\mathbf{(JJAS)_{corr} \neq (ANN)_{corr}}$

Alternative Hypothesis (H1): Alternative hypothesis is opposite to our Null Hypothesis, so there is a correlation between JJAS and ANN. There is an association between these two variables and hence, the variables are dependent.

$\mathbf{(JJAS)_{corr} = (ANN)_{corr}}$

Calculating the Spearmans correlation co-efficient r

Value of Coefficient(r)	**Meaning**
0.00 to 0.19	**A very weak correlation**
0.20 to 0.39	**A weak correlation**
0.40 to 0.69	**A moderate correlation**
0.70 to 0.89	**A strong correlation**
0.90 to 1.00	**A very strong correlation**

<Table 6 the strength of a correlation >

A p-value close to 1 suggests no correlation other than due to chance and that the null hypothesis assumption is correct. If the p-value is close to 0, the observed correlation is unlikely to be due to chance and there is a very high probability that the null hypothesis is wrong. As we define correlation based on the Value of Coefficient(R), so from Table 4, we can see that our r is 0.904 and from Table 5, we can show that the correlation

between JJAS and ANN is a very strong correlation. For further analysis we will be calculating z value from the r-value which will be used to calculate the p value. p value for us comes as 0 which supports the r value showing strong co-relation between ANN and JJAS. In this case, we are rejecting the null (H0) hypothesis and failed to reject the alternative hypothesis (H1).
Detailed explanation in METHOD USED AND WHY(5.1)

Hypothesis Testing 2

In this, we are performing a significance test to decide whether the ratio of standard deviation of JJAS and standard deviation of ANN is equal to one or not based on the population sample
To do this we test the null hypothesis, H0, that the ratio of Std of JJAS and ANN is equal to 1 in the population against the alternative hypothesis, H1, that it is not equal to 1; our data will indicate which of these opposing hypotheses is most likely to be true. We can thus express this test as:

Method

σ_1: standard deviation of JJAS
σ_2: standard deviation of ANN
Ratio: σ_1/σ_2
The Bonnett and Levene's methods are valid for any continuous distribution.

Descriptive Statistics

Variable	N	StDev	Variance	95% CI for σ
JJAS	146	834.523	696428.457	(747.545, 944.297)
ANN	146	1013.681	1027549.674	(904.924, 1150.960)

Ratio of Standard Deviations

Estimated Ratio	95% CI for Ratio using Bonett	95% CI for Ratio using Levene
0.823260	(0.696, 0.975)	(0.695, 1.001)

Test

Null hypothesis	$H_0: \sigma_1 / \sigma_2 = 1$
Alternative hypothesis	$H_1: \sigma_1 / \sigma_2 \neq 1$
Significance level	$\alpha = 0.05$

Method	Test Statistic	DF1	DF2	P-Value
Bonnet	4.98	1		0.026
Levene	3.84	1	290	0.050

Key Result: P-Value
In these results, the null hypothesis states that the ratio in the standard deviations of rainfall data between two periods is not equal to 1. Because both the p-value is less than or equal to the significance level of 0.05, here we rejected the null hypothesis and conclude that the standard deviations of the JJAS and ANN between the rainfall data are different.

V. METHODS USED AND WHY

5.1 Spearman correlation coefficient and P-value for Hypothesis Testing 1

First, we have used Spearman correlation coefficient to find out strength and direction of linear relationship between JJAS and ANN. Spearman coefficient r value is a statistical measure of the strength of link or relationship between two sets of data.
Spearman's correlation has many common uses in geography including analysis of changes in economy, social, environmental variables. We can categorize the type of correlation by considering as one variable increase what happens to the other variable:

- Positive correlation – the other variable also has a tendency to increase;
- Negative correlation – the other variable also has a tendency to decrease;
- No correlation – the other variable does not tend to either increase or decrease.

The starting point for this analysis is to create a scatter plot as show in <Figure 4> and find a trend. Here, the scatter plot suggests a definite positive relationship between rainfall in JJAS and ANN. We also note that there appears to be linear relationship between two variables. After this we define the Hypothesis H0 and H1 as in Hypothesis test 1.

The coefficient (r) is calculated using formula,

$$r = 1 - \left(\frac{6\Sigma d^2}{n^3 - n}\right)$$

For sample size n, each variable of n → X,Y for a particular observation are converted into ranks depending upon its value. Here, d is difference between ranks of X,Y. So, $\Sigma d^2 = (Xi - Yi)^2$. Applying our data to the formula gives us the value of r as 0.904.

The Pearson correlation coefficient value of 0.904 confirms what was apparent from the graph, i.e. there appears to be a positive very strong correlation between the two variables. However, we need to perform a significance test to decide whether based upon this sample there is any or no evidence to suggest that linear correlation is present in the population. To do this we need to test the hypothesis H_0 and H_1. For this we need to find the p value from the r value. The p (or probability) is a measure of how likely or probable it is that any observed correlation is due to chance. P-values range between 0 (0%) and 1 (100%). A p-value close to 1 suggests no correlation other than due to chance and that your H_1 hypothesis assumption is correct. If your p-value is close to 0, there is a very high probability that your null hypothesis is wrong, in this case you must accept the alternative (H1) hypothesis and that there is a correlation between your data sets.

To calculate p-value we need to find t value with n-2 degree of freedom. The formula for finding t value is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The calculated value of t is then given in the tdist() function in excel to find the associated p value. By using t value, n-2 degree of freedom, and third number of tails either 1 or 2 we get the p value as 0.

P value	p-value %	Evidence for rejecting H_0
More than 0.1	>10%	Very weak to none
Between 0.1-0.05	10%-5%	Weak
Between 0.05 – 0.01	5%-1%	Strong
Less than 0.01	<1%	Very strong

< Table 7 P-value and evidence for rejecting the H0 null hypothesis>

So, by the table we can clearly say that we failed to reject H1 and there is a strong correlation between rainfall in JJAS and ANN.

5.2 Levene's method and Bonett's method for Hypothesis Testing 2

We have used these methods in Hypothesis Testing 2. By default, the 2 variances test displays the results for Levene's

method and Bonett's method. Bonett's method is usually more reliable than Levene's method. However, for extremely skewed and heavy tailed distributions, Levene's method is usually more reliable than Bonett's method. We Choose the test on the properties of the distribution of the data, as follows:

- Bonett's test is accurate for any continuous distribution and does not require that the data are normal. Bonett's test is usually more reliable than Levene's test.
- Levene's test is also accurate with any continuous distribution. For extremely skewed and heavy tailed distributions, Levene's method tends to be more reliable than Bonett's method.
- The F-test is accurate only for normally distributed data. Any small deviation from normality can cause the F-test to be inaccurate, even with large samples. However, if the data conform well to the normal distribution, then the F-test is usually more powerful than either Bonett's or Levene's test.

VI. RESULTS AND FINDINGS

India is a developing country and its economy is heavily dependent upon agriculture. So, analyzing and understanding Indian monsoon rainfall is very important factor. As per our analysis here are some of our observations and findings.

- As per the quartile ranges, 25% of the time the rainfall in months of June, July, August, September(JJAS) was below 7932 mm, and 25% of the time the rainfall was above 9054.50 mm. The maximum recorded rainfall for this period is 10202.0 mm and minimum is 6040 mm.
- From the frequency distribution<Figure 1> <Figure 2> we can see that the rainfall in India varies a lot and the frequency distribution is very noisy.
- From <Figure 3> we can see that rainfall in JJAS is a bit skewed to the left with a Skewness value of -0.51715.
- From <Table 3> the 95% confidence interval for rainfall in JUNE, JULY, AUGUST, SEPTEMBER, JJAS are 1571 to 1690.1, 2664 to 2786.7, 2360.7 to 2484, 1643.3 to 1763.5 and 8345.1 to 8618.
- From 3.2.5 a scatter plot of rainfall in JJAS and the ANN (annual) is created which is showing a linear relationship between the two.
- In 3.2.6 a linear regression equation is calculated between rainfall in JJAS and ANN after finding that there is a linear positive relation between the two by using Spearman correlation coefficient. The equation is JJAS = 372 + 0.7468 ANN. Few values are calculated for the month of JJAS by giving ANN values in the equation. For e.g. if ANN rainfall is 11503 mm then the rainfall in JJAS should be around 8962.48 mm. So, we have an equation that can predict the rainfall for the months of JJAS or Annual rainfall, given we know the value of at least one variable.
- Hypothesis Test 1: With a simple scatterplot of JJAS and ANN we decided our Hypothesis Test 1. In Hypothesis Test 1 we had two assumptions H0 and H1, where:

  $H0 \rightarrow (JJAS)_{corr} \neq (ANN)_{corr}$

  $H1 \rightarrow (JJAS)_{corr} = (ANN)_{corr}$

  The Spearman co-efficient was r was calculated where r = 0.904 which was used to calculate p value, where p = 0. Both these showed a very strong correlation between JJAS and ANN and gave us good evidence to support H1 and therefore we failed to reject H1.
- Hypothesis Testing 2: We performed test of variance for JJAS and ANN variables. We defined two assumptions H0 and H1, where:

  $H_0: \sigma_1 / \sigma_2 = 1$

  $H_1: \sigma_1 / \sigma_2 \neq 1$

  By using Bonett's and Levenes's test, we got our P-value for Hypothesis test. After calculations, we got that ratio of std of JJAS and ANN is not equal to 1, so we rejected the null hypothesis and conclude that the standard deviations of the JJAS and ANN between the rainfall data are different.

CONCLUSION

We have applied various statistical methods for the analysis of the rainfall data and found trends between the JJAS rainfall months and annual rainfall. 75% of the times the rainfall in the months of JJAS was below 9054.50. Also, the rainfall in JJAS is related to Annual rainfall by the linear equation JJAS = 372 + 07468ANN. So just by looking at the results obtained rainfall in JJAS we can predict the Annual rainfall trends in India as they positive linear relationship as proved from Spearman Co-efficient. This information can be used by water resources planners, farmers and urban designers to evaluate the availability of water and create the water storage accordingly.

ACKNOWLEDGMENT

The authors acknowledge the IITM(Indian Institute of Tropical Meteorology) for providing datasets and Dublin City University, Dublin, Ireland for providing the facilities to carry out the work and the encouragement in completing this work.

REFERENCES

[1] Suchit Kumar Rai, Sunil Kumar, Arvind Kumar Rai, Satyapriya and Dana Ram Palsaniya 2014 Climate Change Variability and Rainfall Probability for Crop Planning in Few Districts of CentralIndia. Atmos. Climate Sci. 4 394-403.

[2] Nyatuame M, Owusu-Gyimah V and Ampiaw F 2014 Statistical Analysis of Rainfall Trend for Volta Region in Ghana. Int. J. Atmos. Sci. 67(2) 1-11.

[3] Rajendran V, Venkatasubramani R and Vijayakumar G 2016 Rainfall variation and frequency analysis study in Dharmapuri district (India). Indian J. Geo. Mar. Sci. 45(11) 1560-5.

[4] Hingane, L. S. (1995) Is a signature of socio-economic impact written on the climate? Climatic Change 32, 91–101.

[5] Jagannathan, P. & Parthasarathy, B. (1973) Trends and periodicities of rainfall over India. Monthly Weather Rev. 101, 371–375.

[6] Koteswaram, P. & Alvi, S. M. A. (1969) Secular trends and periodicities in rainfall at west coast stations in India. Current Sci. 38, 229–231.

[7] Mooley, D. A. & Parthasarthy, B. (1984) Fluctuations of all India summer monsoon rainfall during 1871–1978. Climatic Change 6, 287–301.

[8] Khan, T. M. A., Singh, O. P. & Sazedur Rahman, M. D. (2000). Recent sea level and sea surface temperature trends along the Bangladesh coast in relation to the frequency of intense cyclones. Marine Geodesy 23, 103–116.

[9] Mirza, M. Q. (2002) Global Warming and changes in the probability of occurence of floods in Bangladesh and implications. Global Environ. Chang. 12, 127–138.

[10] [10] Rupa Kumar, K., Pant, G. B., Parthasarathy, B. & Sontakke, N. A. (1992) Spatial and sub seasonal patterns of the long term trends of Indian summer monsoon rainfall. Int. J. Climatol. 12, 257–26.