European Journal of Business and Management
ISSN 2222-1905 (Paper) ISSN 2222-2839 (Online)
Vol 4, No.17, 2012

FITTING THE STATISTICAL DISTRIBUTION FOR DAILY RAINFALL IN IBADAN, BASED ON CHI-SQUARE AND KOLMOGOROV- SMIRNOV GOODNESS-OF-FIT TESTS.

OSENI, B. Azeez[1] and Femi J. AYOOLA.[2*]

1. Department of Mathematics and Statistics, The Polytechnics, Ibadan.

2. Department of Statistics, University of Ibadan, Ibadan, Nigeria.

Corresponding e-mails: ayoolafemi@yahoo.com

Abstract

This paper presents several types of statistical distributions to describe rainfall distribution in Ibadan metropolis over a period of 30 years.The exponential, gamma, normal and poisson distributions are compared to identify the optimal model for daily rainfall amount based on data recorded at rain guage station at Forestry Research Institute of Nigeria, Jericho, Ibadan (FRIN). The models are evaluated based on chi- square and kolmogorov-smirnov tests. Overall, this study has shown that the exponential distribution is the best model followed by normal and poisson model that has the same estimated rainfall amount for describing the daily rainfall in Ibadan metropolis.

Keywords: scale parameter, asymptotically, exponential distribution, gamma distribution, poisson and kolmogorov-smirnov.

Introduction

Modeling of daily rainfall using various mathematical models has been done throughout the world to give better understanding about the rainfall pattern and its characteristics which involve the study on the sequence of dry and wet days and also the rainfall amount on the wet days. Markov chain models have been widely used in modelling the sequence of dry and wet days, Gabriel and Neumann(1962 ; Roldan and Woolhiser(1982); Stern and Coe(1984); Jimoh and Webster(1996).

On the otherhand, the gamma with two parameters distribution is often used in fitting rainfall amount because it represents the large sizes of drop size distribution better than simple exponential, Ison et al(1971); Katz(1977); Bruishand(1978); Aksoy(2000) and May(2004).

Some other theoretical distributions that have been employed in the analysis of rainfall are the exponential, the mixed exponential, the weibull and the skew normal. These mathematical models of rainfall have been employed in various applications. Mostly used in the study of agriculture and crop planning. Sharda and Das(2005) compared the two and three parameters probability distributions in order to identify the most suitable distributions that best describe the weekly rainfall data. The results from the model have been used to study the effect of rainfall variability during the cropping season in India. The same kind of study has also been conducted in Uganda, Rugumayo et al(2003). The first-order Markov chain, the gamma and Weibull distributions have been selected as the models to generate the daily data such as the studies done in Argentina,Castellvi et al(2004) and in western Australia, Mummery and Battaglia(2004).

In addition, these mathematical models of rainfall are also play important role to serve other purpose. For example, Yoo et al(2005) have used the results from the parameter estimations of the mixed gamma distribution to study the effect of global warming in Korea. As a result of these, the importance of these mathematical models in rainfall studies should not be neglected. In the Ibadan metropolis as a case study received less attention. The studies that have been conducted in Nigeria are more on the general aspects such as pattern, trend and variability of rainfall. Most of the data are outdated and not analysed comprehensively especially in the area of statistics.

The purpose of this study is to find the most appropriate distribution(s) that best describes the daily rainfall of Ibadan metropolis and to test for the goodness-of-fit tests of the data using two methods viz: chi-square and kolmogorov-smirnov

Materials and Methods

Rainfall data, consisting of rainfall frequency, recorded daily intervals collected from Agro- meteorological data for forestry research institute of Nigeria, Jericho Ibadan(FRIN) . The periods of data was 30 year(1979 - 2008). The data which covered period of 30 years was examined and missing records were removed. The original data were tabulated in the Appendix A.

Modelling Rainfall Amount: In this study four models were tested. These four models are described below with their probability mass function (pmf) or probability density function (pdf). In these models, we let X to be a random variable representing the daily rainfall amount.

Model 1: The exponential distribution, with one parameter which represents the scale parameter determines the variation of rainfall amount series that is given in the same unit as the random variable X is given by:

$$f(x) = \frac{1}{\beta} exp\left(\frac{-x}{\beta}\right), x \geq 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The maximum likelihood estimation (mle) for β is given as β =$\bar{x}$.

Model 2: The gamma distribution with two parameters, α and β denote the shape and scale parameters respectively defined as

$$f(x) = \frac{\beta^{-\alpha} x^{\alpha-1}}{\Gamma(\alpha)} exp\left(\frac{-x}{\beta}\right), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The shape parameter governs the shape of the rainfall distribution and the scale parameter determines the variation of rainfall amount unit as the random variable X. For α <1, the rainfall distribution tends to be positively skewed and the maximum of gamma density is located at 0mm/day. For α = 1, the rainfall distribution exhibit an exponential shape and the probability density function approaches 0 mm/day asymptotically. For α>1, the gamma density exhibits a single mode at x = β(α-1) and will result in the distribution to be less skewed and the probability density function will be shifted to the right. The maximum likelihood estimation (m.l.e) for α and β is given as

$$\alpha = \frac{\bar{x}^2}{S^2} \quad and \quad \beta = \frac{S^2}{\bar{x}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Table 1: The behaviours of shape and scale parameter of gamma distribution is shown below.

For α < 1	+vely skewed	0(zero) mm per day
For α = 1	Exponential	Approaches 0(zero) per day asymptotically
For α>1	Less skewed to the right	$x = \beta(\alpha - 1)$

Model 3: The Normal distribution.

The probability density function (pdf) of a normal random variable X with mean μ and standard deviation σ is given by

$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]; \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots ..(4)$$

$-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0.$

If X is a normal random variable with mean μ and standard deviation σ, the

$$P(X \leq x) = P\left(Z \leq \left(\frac{x-\mu}{\sigma}\right)\right) = \int_{-\infty}^{\left(\frac{x-\mu}{\sigma}\right)} exp\left(\frac{t^2}{2}\right)dt$$

$$= \Phi\left(\frac{x-\mu}{\sigma}\right) \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots ..(5)$$

Here, the mean μ, is the location parameter , and the standard deviation σ is the, scale parameter. The normal distribution is the most commonly used distribution to model univariate data from a population or from an experiment

Model 4: Poisson Distribution .

Let X denote the number of events in a unit interval of time or in a Poisson random variable with mean number of events λ in a unit interval of time. The probability mass function of a Poisson distribution with mean λ is given by

$$f(k;\lambda) = P(X = k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}; k = 0,1,2, \dots \dots \dots \dots \dots \dots \dots \dots (6)$$

The Poisson distribution can also be developed as a limiting distribution of the binomial, in which $n \to \infty$ and $p \to 0$ so that np remain a constant. In other words, for large n and small p, the binomial distribution can be approximated by the Poisson distribution with mean $\lambda = np.$

Goodness -of- Fit Tests:

There are several tests of goodness of – fit tests that exist in determining which distribution is the best model to describe the daily rainfall. Among them were the Emperical Distribution Function Statistics (EDF) which includes Kolmogorov – Smirnov, Anderson – Darling and so on.

In this paper, two goodness – of – fit tests were conducted at $\alpha = 5\%$ level of significance. The tests are as follows:

Test 1: Chi-Squared Test (x^2) : C – S

This test simply compares how well theoretical distribution fits the empirical distribution (PDF). The statistic is defined as

$$x^2 = \sum_{i}^{k} \frac{(O_i - E_i)^2}{E_i} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (6)$$

Where O_i = observed frequency for bin i

E_i = the expected frequency for bin i

K = the number of classes

and $E_i = F(X_2) – F(X_1)$ where X_1 and X_2 are the lower and upper limit for bin i.

If the observed frequencies are close to the corresponding expected frequencies, the Chi-Square value will be small indicating a good fit otherwise, it is a poor fit. A good fit leads to the acceptance of Ho whereas a poor fit leads to its rejection.

Test 2: Kolmogorov – Smirnov Test (K-S)

This test statistic is used to decide if a sample comes from a hypothesized continuous PDF. It is a test that is based on the largest vertical difference between the theoretical and empirical CDF. The test is defined as, for a random variable X and samples $(x1, x2, \dots\dots\dots\dots, xn)$, the empirical CDF of X

$$F_n(X) \text{ is given by}$$

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (7)$$

where, I is (condition) = 1 if true and 0, otherwise.

Equation (7) can be simply written as

$$\delta_i(x) = \begin{cases} 1, & \text{if } x_i \leq x \\ 0, & \text{if otherwise} \end{cases} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (8)$$

and that,

$$S_n(x) = \frac{\sum \delta_i(x)}{n}, \text{ implies that}$$

$$nS_n(x) = \sum \delta_i(x) \sim b\big(n, F(x)\big), \qquad where$$

$S_n(x)$ is an unbiased estimator of F(x) and it is also consistent.

Note that: for given two cumulative probability functions $F(x_1)$ and $F(x_2)$, the K-S test statistics (D_n, D_n^+ and D_n^-) is given by

$$D_n = sup_x \quad /F(x_1) - F(x_2)/ \quad for\ H_1: F(x_1) \neq F(x_2)$$

$$D_n^+ = sup_x\big(F(x_1) - F(x_2)\big)\ for\ H_1: F(x_1) > F(x_2)$$

$$D_n^- = sup_x\big(F(x_2) - F(x_1)\big)\ for\ H_1: F(x_1) < F(x_2)$$

Results and Discussion

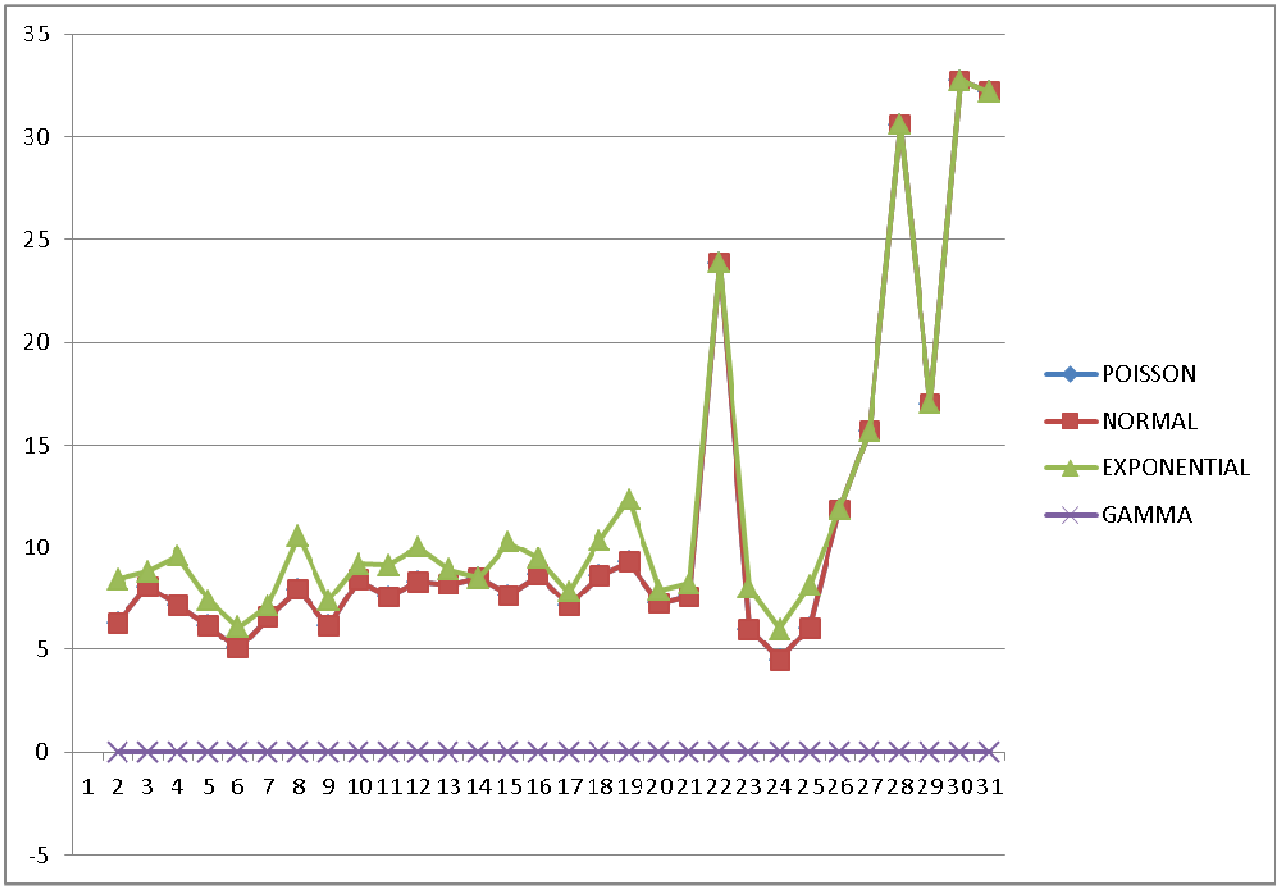
In the first place, we will have a brief discussion on the descriptive statistics for each of the thirty years and then proceed to comment on the results of fitting distributions that are based on chi-square test and kolmogorov- smirnov test. Finally, the remarks on the estimated parameters for the best model will be made.

Descriptive Statistics: The descriptive statistics of the daily rainfall amount for each of the thirty (30) years are summarized in table 1, where the mean , standard deviation, coefficient of variations, skewness, kurtosis number of wet days and maximum amount of daily rainfall of each year are given. Year 2007 received the highest mean rainfall amount followed by 2008, 2005,2006 in that order, while year 2001 received the lowest followed by 2000 and so on. The irregularity of the daily rainfall between years is represented by coefficient of variation (CV) which is evident in all cases that the 100 percent is clearly less than. Those 25 years have high coefficient of variations which ranged between 65 percent to 95 percent compared to other 5 years that ranged between 3 percent to 55 percent. In terms of co-efficient of skewness statistics, the shape of the rainfall distribution for most of the years is strongly skewed. This may be due to the effect of extreme values of the rainfall amount time series of the maximum amount of rainfall that were recorded at those years. We also observed that only few years are weakly skewed Finally, we could summarized that the differences among the years is mostly affected by the climate change in different years

Fitting Distributions Based On Chi-Square test and Kolmogorov –Smirnov test:

The values for chi-Square and Kolmogorov-Smirnov have been calculated and the results are tabulated in table 3. Of these four models tested, the exponential model is found to be the best fitting distribution for all the years studied.This is followed closely by the Poisson and normal models that has the same estimated values throughout the 30 years. The least and poor among the four models tested is gamma model. The graph below illustrate better.

COMPARISON OF ESTIMATED CURVES OF THE FITTED MODELS


35
30
25
20
15
10
5
0
-5
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
POISSON
NORMAL
EXPONENTIAL
GAMMA

The estimated amount of rainfall models were tabulated in the table 2 below

YEAR	POISSON	NORMAL	EXPONENTIAL	GAMMA
1979	6.333333333	6.33333333	8.444444444	-1.10E-07
1980	8.083333333	8.083333333	8.818181818	-2.89E-08
1981	7.166666667	7.166666667	9.555555556	-2.41E-07
1982	6.166666667	6.166666667	7.4	-2.38E-08
1983	5.083333333	5.083333333	6.1	-9.19E-07
1984	6.583333333	6.583333333	7.181818182	-1.55E-08
1985	7.916666667	7.916666667	10.55555556	-1.27E-06
1986	6.166666667	6.166666667	7.4	-1.59E-07
1987	8.416666667	8.416666667	9.181818182	-6.04E-06
1988	7.583333333	7.583333333	9.1	-1.47E-07
1989	8.333333333	8.333333333	10	-2.73E-08
1990	8.166666667	8.166666667	8.909090909	-9.05E-08
1991	8.5	8.5	8.5	-1.78E-08
1992	7.666666667	7.666666667	10.22222222	-3.01E-07
1993	8.666666667	8.666666667	9.454545455	-1.46E-08
1994	7.166666667	7.166666667	7.818181818	-4.81E-08
1995	8.583333333	8.583333333	10.3	-1.60E-08
1996	9.25	9.25	12.33333333	-1.57E-08
1997	7.25	7.25	7.909090909	7.46E-11
1998	7.583333333	7.583333333	8.272727273	-1.75E-08
1999	23.83333333	23.83333333	23.83333333	#DIV/0!
2000	6	6	8	-4.51E-07
2001	4.5	4.5	6	-6.36E-07
2002	6.083333333	6.083333333	8.111111111	-1.67E-08
2003	11.83333333	11.83333333	11.83333333	-9.77E-09
2004	15.66666667	15.66666667	15.66666667	1.96E-10
2005	30.58333333	30.58333333	30.58333333	0
2006	17	17	17	5.05E-15
2007	32.75	32.75	32.75	#NUM!
2008	32.17	32.17	32.17	#NUM!

At 0.05 level of significance kolmogorov- smirnov test are significance for normal, poisson and exponential distribution while the test is not significant in case of chi-square test.

Table 3: Test of Goodness – Fit- Tests

		K-S		
YEAR	**CHI-SQR**	**NORMAL**	**EXPONENTIAL**	**POISSON**
1979	2.666666667	0.63596433	1.507781775	1.109581449
1980	1.333333333	0.723520855	0.841889902	1.304618374
1981	3	0.763471818	1.562251334	1.352871236
1982	1.5	0.584788846	1.257252161	0.969003286
1983	1.333333333	0.622670892	1.007821057	0.853823184
1984	3	0.734865106	0.77494637	1.118363843
1985	2.666666667	0.512388569	1.451073693	1.103857754
1986	1.333333333	0.640255558	1.109283577	0.82838039
1987	2.666666667	0.867340728	0.88464691	1.621573259
1988	1.5	0.601818324	1.311431876	0.924197256
1989	1	0.78849329	1.338055053	1.287095863
1990	1.5	0.686773418	0.694605771	1.312259016
1991	6	0.734646003	0.819155028	1.148004605
1992	2.666666667	0.506957001	1.515685974	1.026389831
1993	3	0.743070723	0.70145285	1.255796487
1994	1.333333333	0.646817629	0.788430578	1.188832203
1995	1.333333333	0.622902115	1.183312401	1.124598025
1996	2.666666667	0.546613094	1.697085525	0.991566561
1997	3.166666667	0.695324847	0.825424928	1.069749649
1998	3	0.629803505	0.691633438	1.006543426
1999	2	0.895398497	2.144383242	1.402088279
2000	3	0.685543513	1.52132183	1.228709438
2001	2	0.545125435	1.566626809	0.943045728
2002	3.166666667	0.596932984	1.655749939	0.858125305
2003	3	0.727098727	0.993590259	1.3651883
2004	2.666666667	0.71011351	0.94650584	1.422035504
2005	0.833333333	0.813519857	2.031317425	0.81159942
2006	0.833333333	0.436723814	0.887943508	0.935436714
2007	1.333333333	0.642625082	2.03512699	0.80636485
2008	1.333333333	0.530911014	2.05789052	0.916513217

Conclusion

Basic statistical characteristics of daily rainfall for Ibadan metropolis have been obtained using daily data recorded at FRIN over a period of 30 years. Four probability distributions namely, Poisson, normal, exponential and gamma were tested to model the distributions of the daily rainfall and kolmogorov-smirnov and chi-squared goodness-of-fit tests were used to evaluate the best fit at 0.05 level of significance. Exponential model is found to be the most

suitable distribution for modelling the daily rainfall amount. Basen on these findings it is recommended as the best model for describing the daily rainfall in Ibadan metropolis.

References

1.Gabriel, K.R and J.Neumann(1962):A Markov chain model for daily rainfall occurrence at Tel. Aviv. Quarterly Journal of Royal meteorology society 8: 90-95.

2. Roldan, J and D.A , Woolhier(1982): Stochastic daily precipitation models 1: A comparison of occurrence process. Water resources research, 18(5):1451-1459.

3. Stern, R.D and R. Coe,(1984): A model fitting analysis of daily rainfall data. Journal of royal statistical society series A,147.1-34.

4.Jimoh,O.D and P .Webster,(1996):Optimum order of Markov chain for daily rainfall in Nigeria Journal of hydrology,185:45-69.

5. Ison,N.T; A.M.Feyerherm and L.Dean Bark,(1971):Wet period precipitation and the gamma distribution .Journal of Applied Meteorology,10:658-665.

6. Katz, R.W(1977):Precipitation as chain-dependent process. Journal of Applied Meteorology, 16: 671-676.

7. Buishand,T.A.(1978):Some remarks on the use of daily rainfall models .Journal of Hydrology, 36:295-308.

8. Aksoy, H.(2000): Use of gamma distribution in hydrological analysis .Turkey Journal of Engineering Environmental sciences, 24: 419-428.

9.Woolhiser, D.A. and J. Roldan(1982): Stochastic daily precipitation Models 2. A comparison of distribution of Amounts. Water Resources Research, 18(5): 1461-1468.

10. Sharda , V. N. And P . K. Das (2005) : Modelling Weekly Rainfall Data for crop planning in a sub –humid climate of India. Agricultural Water Management, 76 ; 120-138.

11.Rugumay , A .I., N .Kiiza ,J. Shima (2003):Rainfall reliability for crop production .A case study in Uganda. Diffuse pollution conference Dublin .

12.Castellvi , F. ,I. ,Mormeneo and P.J. Perez (2004):.Generation of daily amounts of precipitation from standard climatic data: a case study for Argentina. Journal of hydrology, 289: 286- 302.

13.Mummery , D . and M . Battaglia (2004) Significance of rainfall distribution in predicting eucalypt plantation growth , management options and risk assessment using the process – based model CABALA . Forest Ecology and Management 193: 283- 296.

14.Yoo, C. , K .S.Jung and W . K .Tae(2005):Rainfall frequency analysis using a mixed Gamma distribution : evaluation of the global warming effect of daily rainfall . Hydrological Processes, 19; 3851 – 3861.