ORIGINAL RESEARCH PAPER

Volume - 10 | Issue - 06 | June - 2021 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr

INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

PROBABILITY DISTRIBUTION ANALYSIS OF PATIENTS ARRIVING AT REGIONAL REFERRAL HOSPITAL DODOMA, TANZANIA (2017-2018).

Statistics

Prof. Ramkumar T. Balan* Asso.Prof., Dept.of Mathematics & Statistics ,CNMS, University of Dodoma, Tanzania. *Corresponding Author

Julius Moinget Loibor Scholar, Dept.of Mathematics & Statistics ,CNMS, University of Dodoma, Tanzania.

ABSTRACT

Health care is essential to the general welfare of society. Studying the hospital patients' data distribution through the probability distribution analysis model is very important in the health care system. This study has examined the hospital inpatients and outpatients' daily data for two years taken from DRRH through the hospital electronic health management information system. This study seeks to identify comprehensively the appropriate statistical distributions on inpatient and outpatient data of the DRR hospital. The primary fitting of the distributions to inpatient and outpatient data was performed by the Easyfit 5.5 Profession statistical software. The software deals with 61 continuous distributions, including three goodness of fit test for raw data and two for frequency data. Kolmogorov- Smirnov test, Anderson- Darling test and Chi-Square test only for raw data. The parameters of the selected distributions were estimated by the maximum likelihood method. The final selection of fittest distribution was done with respect to the minimum calculated value of log-likelihood and hence AIC and BIC values. The research work revealed that Generalized Extreme Value distribution is the best-fit distribution model for the hospital inpatient daily data. Also, the Dagum distribution followed by Log logistic (3P) distribution was selected to be the best-fit distribution model representing the hospital outpatients' daily data.

KEYWORDS

Probability distribution, hospital data, Likelihood function, Best fit, Minimum AIC and BIC crierion

INTRODUCTION

Hospitals in general and healthcare systems in particular, are the major determinant of the quality of healthy life. In Tanzania, both government hospitals and private hospitals provide health care services to different groups of people, but most of the common people rely on public health services. The need for health care services has been rising in recent years due to the growth of both preventable and communicable diseases. As a result of lack of sufficient hospitals, human resources in the field of health, well organized operating systems, patients in the government hospitals are suffering a lot. They have to spend excessively long times to get attended and fast dispensing due to overcrowded waiting for treatment so that patients are dissatisfied and grumbling. The outpatient and inpatient departments are core units in the hospital regulating the patients arriving at the hospital and so a scientific evaluation is needed to control the system. Hospitals generally differ from other types of health centers offering medical services by their ability to admit and care for inpatients. In a hospital, the entry point starts at the outpatient department and the clients' moderator will set the standard facility that the patient can receive at the outpatient department. Globally, it is accepted that the first intuition will go parallel to the way people's perceptions about things happen and the place they reached. Long waiting health services in outpatient departments in hospitals has become a problem in the healthcare setting all over the world (Kelaniya, 2014).

Among several composite provisions of health care services, admission of a number of patients, the number of arrivals to the hospital was one of the very significant measures of healthcare utilization. In the brightness of these challenges, a need for review and improvement of our healthcare practices has become evident. The outcome of statistical probability distribution analysis helps the management to plan the system on infrastructure, healthcare services, medicine and suggest future requirements. Probability distribution study of the hospital will help to enhance the allocation of the exchequer for infrastructure, assignment of doctors, specialists, and other medical care servants, the volume of medicine needed as well as the transport and other facilities for the patients. It will reduce queuing of patients in the outpatient department and facilitate emergency services. The inpatient statistics and its distribution are more essential to get proper accommodation and health service for all needful patients. In our daily life, the knowledge of probability distributions is essential in different fields such as agricultural science, medicine, education, economics and engineering sciences (Machekposhti & Sedghi, 2019). The use of Probability distribution approaches to fit and select the appropriate best fitting statistical distribution model to a set of given data has largely gained momentum recently among researchers (Langat, Kumar, and Koech, 2019). The distribution analysis has been used in many other fields such as Health insurance, Industry, Banking, Transportation, Planning and development, etc. A lot of studies use, distribution models to investigate actuarial problems by fitting the appropriate probability distribution for the given data.

Searching and adapting fittest probability distribution on a data is in need as lack of accuracy is imminent in some kind of data where high tendencies of variation, tailing and unstable occurrence is found. To evaluate patient arrival and stay in the hospital, the data analysis is insufficient with descriptive statistics and more intrinsic study is developed with the help of underlined statistical distributions. Mehrannia and Pakgohar (2014) had shown the advantage of using software to fit data and interpreting probability. They are able to fit automatically a variety of known distribution patterns simultaneously on data using algorithms and programming. These techniques are preferred, especially in cases where little or no information about the base distributions /pattern in the data is not available and desired to find the best fit distribution (Mehrannia and Pakgohar, 2014). Though, to our knowledge, there are little distributions suitable for hospital patients' data of Tanzania especially for Dodoma Regional Referral Hospital (DRRH). We are starting with a bunch of continuous probability distributions from a broad area of statistical distributions and from them, appropriate distribution is attained with statistical accuracy. The elementary befitting distributions were found using the EasyFit statistical software asserting the goodness of fit with Chi-Square, Anderson and Kolmogorov tests. The primary selected distribution analysis is conducted by P-P plot and Q-Q plot and the best fit is determined by the use of AIC or BIC or Log-likelihood values. There are sixty-one well developed statistical distributions entertained in the EasyFit software and to a numerical continuous data, the best fit can be evaluated with suitable logic.

This study points out the following continuous probability distributions-Beta, Burr, Cauchy, Dagum, Gamma (3P), Generalized Extreme Value, Generalized Gamma (4P), Johnson SB, Log logistic (3P), Nakagami, Rayleigh, Rayleigh (2P) and Weibull distributions- as the basis to find the best fit probability distribution for number of inpatients and outpatients of DRRH. It attempts to find probability distribution for the patient daily arrival data to the hospital as well as probability distribution of daily stay in hospital.

METHODOLOGY

Basic concepts in statistics and probabilistic are probability distributions. The results of statistical daily hospital patients' data and their probabilities of occurrence are connected with probability distributions. Generally, the study involves the following procedures, research approach, research design, collection of daily hospital patients data, selection of the best-fitted distributions for daily hospital patients data and parameters estimation. This study was used both quantitative research approach and a retrospective research design using secondary data of patients arriving at the Dodoma regional referral hospital through the electronic health management information systems. The secondary daily hospital patients' data from Dodoma regional referral hospital were evaluated with thirteen (13) probability distribution models to find the best fit model. At the first and before modeling of daily hospital patients' data, we need to have check adequacy, accuracy and relevance conditions. The probability distribution modeling process of the selected distributions after initial analysis of the hospital patients' data, conducted through the use of EasyFit software was described. That is the assessment of primary statistical distributions befitting the daily patient data subjected to sixty-one (61) continuous probability distributions suggested in the EasyFit software.

Selection of Best fit Distributions.

Selection of suitable probability distribution is an essential step before any statistical analysis and definitely, it is central to the pursuit of science in general (Kadane&Lazar, 2004). Most empirical economic research, that involves the specification, estimation, and evaluation of the statistical models based on certain criteria, will choose an appropriate distribution for the data. So here for the first step of selection of probability distributions, and its initial analysis of Dodoma regional referral hospital (DRRH) data, it was conducted through EasyFit 5.5 Professional software. Running the program, the software will give the ranks of fit of each probability distribution with its ML parameter estimates and goodness of fit of the distributions. From this list, one can draw the suitable number of distributions needed as per ranks. For the raw data, three tests were applied to find the goodness of fit of the distributions -Chi-Square test, Anderson Darling test, and Kolmogorov Smirnov test -and the ranks were given in their respective order separately. For the frequency data, the Chi-Square test is not adopted in this software. From the suggested distributions the selection of first five ranked probability distributions was done on each goodness of fit criterion and finally, a list of suggested distributions is made. Now it is necessary to find the likelihood function and hence Akaike Information criteria and Bayesian Information criteria based on the sample to distinguish the best fit.

The common statistical distributions in daily total inpatient data and outpatient data were the following. 1)Beta distribution, 2)Burr distribution, 3)Cauchy distribution,4)Dagum distribution,5)Gamma 3P distribution, 6)Generalized Extreme Value distribution, 7)Generalized Gamma 4P distribution, 8)Johnson SB distribution, 9)Log logistic 3P distribution, 10)Nakagami distribution, 11)Rayleigh distribution,12)Rayleigh 2P distribution and 13)Weibull distribution.

Estimation of Parameters of Probability Distributions.

Estimation of the parameters for each of the chosen statistical probability distributions for hospital patients' data was done after an initial analysis of the data through EasyFit 5.5 Professional software. After estimating the parameters of the chosen probability distributions, further analysis of the fitted distribution will be allowed to proceed. In this present study, the parameters of the selected statistical probability distributions were estimated using the method of the maximum likelihood technique with the help of EasyFit software. The maximum likelihood estimate requires primarily the likelihood function of a random variable. The likelihood function L (θ) is the joint probability density function of the observed data expressed as a function in terms of parameter θ. Given that $x_1, x_2, x_3, \ldots, x_n$ have a joint density function $f(x_i/\theta)$ for every observed sample of independent observations i=1,2,3,…,n, then the likelihood function is given by $L(\theta) = \prod_{i=1}^{n} f(\chi_i, \theta)$ (Omari, Nyambura and Mwangi 2018). The MLE $\hat{\theta}$ of a parameter θ is obtained by maximizing the likelihood function L (θ). It is needed to derive the MLE's for the set of parameters of the probability distributions by solving the normal equations obtained by maximization of logL (θ). But ML estimate of the parameters for the given set of data is readily available with the help of EasyFit 5.5 software and SAS 9.4 software. With the calculated Likelihood value for the given data, one can find Akaike's Information Criterion and Bayesian Information Criterion and it is useful to compare the goodness of fit of the models.

Goodness of fit test.

The goodness of fit describes how best the statistical distribution fit a dataset observation. The evaluation of the goodness of fit is very important in the process of selecting the best distribution. It measures the compatibility of a random sample with a theoretical probability distribution. Therefore to assess the goodness of fit of the selected thirteen distributions various tests are applied. The goodness of fit tests is effectively based on either two distribution functions or probability functions in order to test the null hypothesis that the unknown distribution or probability function follows a specified function or other. The methods considered for finding the appropriateness of the fitted distributions are, Kolmogorov-Smirnov test and its modified named Anderson Darling test and Chi-Square test. For assessing whether or not patients' dataset follows chosen distributions such as Beta, Burr, Cauchy, Dagum, Gamma 3P, Generalized Extreme Value, Generalized Gamma 4P, Johnson SB, Log logistic 3P, Nakagami, Rayleigh, Rayleigh 2P, and Weibull distribution the above criterion were used.

The Anderson Darling test is used to test if a sample data come from a population with a specified distribution. It is the modification of the Kolmogorov Smirnov test and gives more weight to tail than does the KS test. The Kolmogorov Smirnov test compares a hypothetical or fitted cumulative distribution function with an empirical cumulative distribution in order to assess the goodness of fit of a given data set to a theoretical distribution. The goodness of fit test statistic of the Kolmogorov Smirnov (KS) and the Anderson Darling test was obtained using the EasyFit 5.5 Professional software.

Chi-square goodness fit test.

This is one of the goodness of fit test used in this study. The chi-square test allows us to test whether the data approves the observed and proposed frequencies. The Chi-square statistic measures how well the expected frequency of the fitted distributions compares with the observed frequency of histogram of the observed data. The chi-square statistic is calculated using this formula:

$$\chi^2_{cal} = \sum_{i=1}^{n} \left[\frac{(\theta_i - E_i)^2}{E_i} \right] \square \ \chi^2_{n-1\,df} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where θ_i is the observed frequency in the i^{th} cell, E_i is the expected frequency in the i^{th} cell and I is the number of observations (1, 2, 3…, k). The Chi-square goodness of fit test statistic was obtained using the EasyFit 5.5 Professional software.

The Akaike's Information Criterion.

The Akaike's Information Criterion (AIC) is another goodness of fit statistic used in this study. The AIC is a means of selecting a model from a set of models. It is an estimation of Kullback- Leibler information or distance which attempts to select a good approximating model for assumption based on the principle of cost-cutting (Anderson and Burnham, 2004). This AIC was derived based on the concept that truth is very complex and that no "true models" exits (Anderson and Burnham, 2004). The AIC can be defined as, AIC = -2(maximized log-likelihood) +2(the number of parameters estimated). For the fitted statistical distributions which will appear to have the smallest value of AIC is chosen to be the most appropriate distribution for fitting the patients' data set.

The Bayesian Information Criterion.

The Bayesian Information Criterion is another criterion for model selection which was proposed by (Schwarz, 1978) in order to identify the appropriate best-fit distribution model for a given set of data. The formula for calculating the BIC was given below:

$$BIC = -2\log L + \kappa \log n \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2).$$

Where n is the number of observation and κ is the number of parameters in the distribution model. And also L is the maximized log-likelihood function.

In this present study, both values of AIC and BIC for all suggested statistical distributions were computed after the values of the log-likelihood for the individual distribution has been obtained. The researcher used estimates of parameters of the fitted distribution model to find log-likelihood for individual statistical distribution and hence both values of AIC and BIC were computed manually using Excel software.

RESULTS AND DISCUSSIONS

Discussion: In order to determine the suggestions of distributions, we make use of EasyFit 5.5 Professional software on raw data for inpatients and frequency data for outpatients (as raw data does not provide consistent suggestions). Anderson test, Kolmogorov test, and chi-square test are applied in the first data and first two tests are valid in the second data. From table 1.0 below By Kolmogorov test and Chi-square test of goodness of fit, the ranks of this distribution are one among 61 options for continuous distributions. Also, both test accepts the null hypothesis that the inpatient data follows Generalized Extreme Value distribution at 1%, 2%, 5%, 10%, and 20% level of significance. In addition to these, Anderson test also accepts the same hypothesis at all these levels of significances, even though there exist another 10 distributions better than this. Hence the number of inpatients per day at DRRH follows Generalized Extreme Values distribution. The preliminarily selected distributions for inpatient daily data were: Beta, Burr, Gamma (3P), Gen. Gamma (4P), and Gen. Extreme value distribution.

For the outpatient raw data, the fitting of distribution with respect to 61 continuous distributions cannot identify any distribution uniquely by means of the three tests. So conveniently the data is arranged in terms of frequency data and then the software will test the goodness of fit by Anderson and Kolmogorov test. The possible distributions coming under rank 5 of both tests were considered to determine the plausible basic distributions for the outpatient data. Thus there are 9 distributions were under consideration for the second stage analysis. From table 2.0 below, The outpatient frequency data is supposed to follow Dagum distribution under the null hypothesis and it is accepted by both tests at 1%,2%, 5%,10% and20% level of significance. By Kolmogorov test, this distribution is of rank 1 while it is rank 8 for Anderson test. Generalizing the result, the number of outpatients per day at DRRH follows Dagum distribution. The selected preliminary distributions for outpatient daily data were: Burr, Cauchy, Dagum, Johnson SB, Nakagami, Log-logistic (3P), Rayleigh, Rayleigh (2P), and Weibull distribution.

The best-fitting distribution on these data is determined by log-likelihood statistic, AIC (Akaike, 1974) and BIC values of suggested distributions. Since the criteria of selection method is based on the log-likelihood values of distributions on sample values, it is calculated after estimating the parameters and hence their AIC and BIC values. The AIC criterion was derived based on the concept that the truth is very complex and that no "true model" exists for any sample data set (Anderson and Burnham, 2004). Thus, for given suggested distributions, it was possible to estimate which distribution was nearby to the unknown true model. In this present study, both values of AIC and BIC were computed after finding the values of the log-likelihood for given data with respect to individual distributions suggested. The MLE estimation method is used to find the parameters in each model and then log-likelihood is determined manually using Excel software for each distribution suggested. The distribution which will be appeared to have the smallest value of the AIC and BIC is the one that will be appropriate for fitting the daily hospital patients' data.

From table 3.0 below There are five distributions suitable to present the inpatient data of DRRH (See Anderson, Kolmogorov, and Chi-Square test tables). Beta, Burr, Generalized Extreme Value, Gamma (3P), Generalized Gamma (4P) distributions were the feasible suggestions using the statistical test criteria. The result from Table 3.0 shows that from the distributions suggested the AIC, BIC, and Log-likelihood values were minimum for the Generalized Extreme Value distribution. The values were consistently small for this distribution satisfying all three criteria and so it is the best fit distribution for the inpatient data of DRRH. Burr distribution and three parametric Gamma distribution are coming under second and third suggestions but their AIC, BIC, and Log-likelihood values are far different from Generalized Extreme Value distribution so that for probability analysis of the inpatient data it is uniquely suggested the Generalized Extreme Value distribution.

For daily outpatients data, There are nine statistical distributions shown in Table 4.0 satisfy the Kolmogorov test and Anderson test of goodness of fit for the outpatient frequency data suggesting the basic feasible probability distributions. The AIC and BIC values are derived from the maximum log-likelihood value for the sample values for each distribution and illustrated in Table 4.0. From Table 4.0 Dagum distribution is the most befitted distribution for the outpatient data as the AIC, BIC, and Log-likelihood values stood uniquely minimum compared to all other distributions. Three parametric Log logistic distribution and Cauchy distributions are the second and third appropriate fitted curves to avail the probability information.

RESULTS.

Table 1. 0: showing the test statistic for Gen. Extreme Value distribution of the daily total number of inpatients hospital data.

Gen. Extreme Value Distribution.					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	730 0.02746 0.63037 1				
Significance level	0.2	0.1	0.05	0.02	0.01
Critical Value	0.03971	0.04527	0.05026	0.05618	0.06029
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	730 0.79848 11				
Significance level	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No
Chi-Squared					
Deg. of freedom Statistic P-Value Rank	9 1.3019 0.99837 1				
Significance level	0.2	0.1	0.05	0.02	0.01
Critical Value	12.242	14.684	16.919	19.679	21.666
Reject?	No	No	No	No	No

Table 2. 0: Showing the test statistic for the Dagum distribution of the daily total number of outpatients' hospital data.

Dagum Distribution.					
Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	13 0.0933 0.99917 1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.2847	0.32549	0.36143	0.40362	0.43247
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	13 0.96788 8				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	1.3749	1.9286	2.5018	3.2892	3.9074
Reject?	No	No	No	No	No

Table 3. 0: Model selection for a total number of inpatients in DRRH.

Distribution	AIC	BIC	Log-likelihood
Beta	44064.34	44067.79	-22028.17
Burr	6054	6067	-3023.80
Gen. Extreme Value	**1067.96**	**1067.408**	**-529.98**
Gamma (3P)	6054.31	6068.09	-3124.16
Gen. Gamma (4P)	21713.0	21716.62	-10852.58

Table 4. 0: Model selection for a total number of outpatients in DRRH.

Distributions	AIC	BIC	Log-likelihood
Burr	176.967	178.662	-85.483
Cauchy	82.985	81.213	-39.493
Dagum	**78.455**	**75.797**	**-36.228**
Johnson SB	182.317	178.317	-87.159
Log logistic (3P)	81.975	79.317	-37.988
Nakagami	950084.112	950082.340	-475040.056
Rayleigh	152.430	151.544	-75.215
Rayleigh (2P)	99.530	97.758	-47.765
Weibull	174.965	176.095	-85.482

Summary of the results of the study.

Primary fitting of the distributions to inpatient and outpatient data was performed by the Easyfit 5.5 Profession statistical software. The software is dealing with 61 continuous distributions with three goodness of fit tests for raw data and two for frequency data. Kolmogorov Smirnov, Anderson- Darling are common tests and Chi-Square test for raw data. Preliminary distributions for inpatient data was organized by observing repeated distributions within a rank of 5 in 3 tests. There are 5 distributions found for inpatient distributions and all of them show similar characteristics on mean, SD, skewness and kurtosis. For outpatient data, there are 9 probability distributions satisfactory but their characteristics are a little different.

The parameters of the selected distributions were estimated by the maximum likelihood method. The location, scale and shape parameters were detected and estimated for 13 cases with the help of software. To identify the best fitting distribution from the selected list, 3 statistical criterions were adopted. Log-Likelihood was determined using Excel or SAS for the given set of inpatient and outpatient data and the AIC and BIC is deduced by the linear relations. Generalized Extreme Value distribution is the fittest among the 5 for inpatient data and Dagum distribution is appropriate for outpatient data. The most befitting distribution is detected by finding the least values formed for the 3 criterions. The computed values of AIC, BIC and value of log-likelihood for Generalized Extreme Value distribution were 1067.96, 1067.408 and -529.98 which is minimum in each column. Similarly, AIC, BIC, and the log-likelihood value for Dagum distribution were 78.455, 75.797 and -36.228 respectively and they are minimum in respective columns.

CONCLUSION

This study has examined the total hospital inpatients and total outpatients daily data for two years from DRRH obtained through the hospital electronic health management information system and has fitted all the selected statistical distributions. After following the distributions fitting process using the statistical software known as EasyFit 5.5 Professional software with tested accuracy the study proposes Generalized Extreme Value distribution as the best-fit distribution model for fitting the total hospital inpatients data. Also, the Dagum distribution followed by Log logistic (3P) distribution was selected to be the best fit distribution models for representing the total hospital outpatients' data. The final selection of these distributions was done by using the minimum calculated values of both AIC value, BIC value and the value of the log-likelihood. The present set up of inpatient and outpatient accommodation may cause insecurity and complaining in-hospital care for the Dodoma people as the probability of a number of inpatients and outpatients is more comparing to available infrastructure facilities and healthcare resources of experts.

REFERENCES

1. Anderson, D.R and Burnham, K.P. (2004). "Multimodel infrence "Understanding A.I.C &B.I.C in model selection". *Colorado Cooperative Fish&Wildlife Research Unit (USGS-BRD).*"
2. Akaike, H. (1974). "A new look at the statistical model identification". *IEEE Transaction on Automatic Control vol. 6*, 716-725.
3. Cousineau D, Brown S, Heathcote A. (2004). Fitting distributions using maximum likelihood methods and packages.
4. G., S. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461-464.
5. Kadane, J. B., & Lazar, N. A. (2004). Methods and criteria for model selection. *Journal of the American Statistical Association, 99 (465)*, 279-290.
6. Karim Hamidi Machekposhti, Hossein Sedghi. (2019). Determination of the Best Fit Probability Ditribution For Annual Rainfall in Karkheh River at Iran. *International Journal of Environmental and Ecological Engineering.*
7. Kelaniya, S. (2014). A Simulation approach for reduced outpatient waiting time,. 4-9.
8. Mehrannia Hossein and Pakgohar Alireza. (2014). Using EasyFit Software for Goodness-of-Fit Test and Data Generation. *International Journal of Mathematical Archve-5(1)*, 118-124.
9. Omari,C.O., Nyambura, S.G. and Mwangi, J.M.W. (2018). Modeling the Frequency and Severity of Auto Insurance Claims Using Statistical Distributions. *Journal of Mathematical Finance, 8,*, 137-160.
10. Philip Kibet Langat, Lalit Kumar &Richard Koech. (2019). Identification of the Most Suitable Probability Distribution Models for Maximum, Minimum and Mean Streamflow.