Identification of appropriatedistribution in Reliability Analysis

Sanjeev Kumar Jha
Ph.D. Research Scholar,
School of Computer Science and IT
Singhania University Rajasthan

Shivlok Singh
Ph.D. Research Scholar,
School of Computer Science and IT
Singhania University Rajasthan

Dr. Amod Tiwari
Dean, ,PSIT Kanpur,India

Abstract: *Distribution fitting is the method of selecting appropriate distributions among the list of distributions to be fitted on given set of data. The main aim of distribution fitting is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval. There are many probability distributions of which some can be fitted more closely to the observed frequency of the data than others, depending on the characteristics of the phenomenon and of the distribution. The distribution giving a close fit is supposed to lead to good predictions .In distribution fitting, therefore, one needs to select a distribution that suits the data well. Here in this paper we discuss an optimum technique of selecting distributions. Goodness of fit method for identification of distribution is used.*

***Keywords: Goodness of Fit, Open source software (OSS) , Apache; Kolmogrove,Software reliability model; Software architecture; Reliability growth model**;*

I. INTRODUCTION

The aim of distribution fitting is to predict the probability or to forecast the frequency of occurrence of the magnitude of the phenomenon in a certain interval. In this paper with the help of data collected for Apache Web Server is used.

There are different life data distributions which can be used for reliability analysis. Single distributions cannot be used for reliability modeling of all samples under study. Thus for reliability modeling best distribution is to be selected for each sample under study. After collecting data and converting it into appropriate format for analysis, Goodness of fit test is used for selection of best distributions. There is different goodness of fit techniques available. As MLE's are MVUE. Thus in this research Likelihood Ratio is used for goodness of fit test. In this technique we will have maximum accuracy and minimum error. Different softwares are available for statistical distribution and analysis. In this research for statistical calculation and analysis purpose Microsoft Excel 2007 with Easy Fit5.5 is used.

In distribution fitting, therefore, one needs to select a distribution that suits the data well. The goodness of fit [1] of a statistical model describes how well it fits a set of observations. Measures of goodness of fit typically summarize the discrepancy between observed values and the values expected under the model in question. Such measures can be used in statistical hypothesis testing, e.g. to test for normality of residuals, to test whether two samples are drawn from identical distributions or whether outcome frequencies follow a specified distribution. There are various methods used for goodness of fit test. Most important among them are as given below:

- Kolmogorov-Smirnov
- Anderson-Darling
- Chi-Squared

In this research all the methods for distribution identification is discussed and finally on the basis of collected sample distributions are identified.

II. KOLMOGOROV-SMIRNOV TEST [4]

This test is used to decide if a sample comes from a hypothesized continuous distribution. It is based on the empirical cumulative distribution function (ECDF). Assume that we have a random sample x1,x2..., xn from some distribution with CDF F(x).

The empirical CDF is denoted by

$$F_n(x) = \frac{1}{n}[\text{Number of Observations} \leq x] \qquad (1)$$

The Kolmogorov-Smirnov statistic (D) is based on the largest vertical difference between the theoretical and the empirical cumulative distribution function:

$$D = \max_{1 \leq i \leq n}\left(F(x_i) - \frac{i-1}{n}, \frac{i}{n} - F(x_i)\right)$$

(2)

Hypothesis Testing

The null and the alternative hypotheses are:

- H0: the data follow the specified distribution;
- HA:the data do not follow the specified distribution.

The hypothesis regarding the distributional form is rejected at the chosen significance level (α) if the test statistic, D, is greater than the critical value obtained from a table. The fixed values of α (0.01, 0.05 etc.) are generally used to evaluate the null hypothesis (H0) at various significance levels. A value of 0.05 is typically used for most applications, however, in some critical industries; a lower (α) value may be applied. The standard tables of critical values used for this test are only valid when testing whether a data set is from a completely specified distribution. If one or more distribution parameters are estimated, the results will be conservative: the actual significance level will be smaller than that given by the standard tables and the probability that the fit will be rejected in error will be lower.

P-Value

The P-value, in contrast to fixed α values, is calculated based on the test statistic, and denotes the threshold value of the significance level in the sense that the null hypothesis (H0) will be accepted for all values of α less than the P-value. For example, if P=0.025, the null hypothesis will be accepted at all significance levels less than P (i.e. 0.01 and 0.02), and rejected at higher levels, including 0.05 and 0.1. The P-value can be useful, in particular, when the null hypothesis is rejected at all predefined significance levels, and you need to know at which level it could be accepted.

III. ANDERSON-DARLING TEST [5]

The Anderson-Darling procedure is a general test to compare the fit of an observed cumulative distribution function to an expected cumulative distribution function. This test gives more weight to the tails than the Kolmogorov-Smirnov test.

The Anderson-Darling statistic (A^2) [6] is defined as

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\cdot[\ln F(X_i) + \ln(1-F(X_{n-i+1}))] \quad (3)$$

HYPOTHESIS TESTING [6]

The null and the alternative hypotheses are:

- H0: the data follow the specified distribution.
- H_A: the data do not follow the specified distribution.

The hypothesis regarding the distributional form is rejected at the chosen significance level (α) if the test statistic, A^2, is greater than the critical value obtained from a table. The fixed values of α (0.01, 0.05 etc.) are generally used to evaluate the null hypothesis (H0) at various significance levels. A value of 0.05 is typically used for most applications however; in some critical industries a lower α value may be applied. In general, critical values of the Anderson-Darling test statistic depend on the specific distribution being tested. However, tables of critical values for many distributions (except several the most widely used ones) are not easy to find. The Anderson-Darling test implemented in Easy Fit uses the same critical values for all distributions. These values are calculated using the approximation formula, and depend on the sample size only. This kind of test (compared to the "original" A-D test) is less likely to reject the good fit, and can be successfully used to compare the goodness of fit of several fitted distributions.

IV. CHI-SQUARED TEST [7]

The Chi-Squared test is used to determine if a sample comes from a population with a specific distribution. This test is applied to binned data, so the value of the test statistic depends on how the data is binned. This test is used for continuous sample data only. Although there is no optimal choice for the number of bins (k), there are several formulas which can be used to calculate this number based on the sample size (N). For example, Easy Fit employs the following empirical formula: k=1+log2N. The data can be grouped into intervals of equal probability or equal width. Each bin should contain at least 5 or more data points, so certain adjacent bins sometimes need to be joined together for this condition to be satisfied.

Definition: The Chi-Squared statistic is defined as:

$$\chi^2 = \sum_{i=1}^{n}\frac{(O_i - E_i)^2}{E_i} \quad (4)$$

where Oi is the observed frequency for bin i, and Ei is the expected frequency for bin i calculated by E i = F(x2)-F(x1) ,where F is the CDF of the probability distribution being tested, and x1, x2 are the limits for bin i.

HYPOTHESIS TESTING

The Null and Alternative Hypothesis are given by

- H_0: the data follow the specified distribution;
- H_A: the data do not follow the specified distribution.

The hypothesis regarding the distributional form is rejected at the chosen significance level (α) if the test statistic is greater than the critical value defined as $\chi^2_{1-\alpha,k-1}$ meaning the Chi-Squared inverse CDF with k-1 degrees of freedom and a significance level of α. Though the number of degrees of freedom can be calculated as k-c-1 (where c is the number of estimated parameters), Easy Fit calculates it as k-1 since this kind of test is least likely to reject the fit in error. The fixed values of α (0.01, 0.05 etc.) are generally used to evaluate the null hypothesis (H_0) at various significance levels. A value of 0.05 is typically used for most applications, however, in some critical industries; a lower α value may be applied.

P-Value

The P-value, in contrast to fixed α values, is calculated based on the test statistic, and denotes the threshold value of the significance level in the sense that the null hypothesis (H0) will be accepted for all values of α less than the P-value. For example, if P=0.025, the null hypothesis will be accepted at all significance levels less than P (i.e. 0.01 and 0.02), and rejected at higher levels, including 0.05 and 0.1. The P-value can be useful; in particular, when the null hypothesis is rejected at all predefined significance levels, and you need to know at which level it could be accepted. Easy Fit displays the P-values based on the Chi-Squared test statistics (χ2) calculated for each fitted distribution.

V. DATA COLLECTION AND PREPROCESSING

Here Data collection is associated with collecting a data related to failure of Apache. In the bug-collection step, the online bug-repository systems are used to collect the failure data. For this purpose web site http://www.bugzilla.apache.org is used. Data is collected for Apache2, which is latest available stable version of Apache. This version of Apache2 came into existence in year 2000. Bugs are collected from 26/8/2002 to May 2011. Data are extracted directly from the web site. Bugs reported might be duplicates, provide incomplete information, or may not represent real defects. Therefore, during the bug preprocessing such noises are removed from the bugs gathered in the first step. Finally, in the third step, the preprocessed data is stored in Mysql database. Initially data was in csv (comma separated value) format. Mysql is an open source data base system. It is freely available and very secure. Total of 1,250 records were extracted from repository and after preprocessing finally 501 records were stored in Mysql Table. During preprocessing following records were deleted:

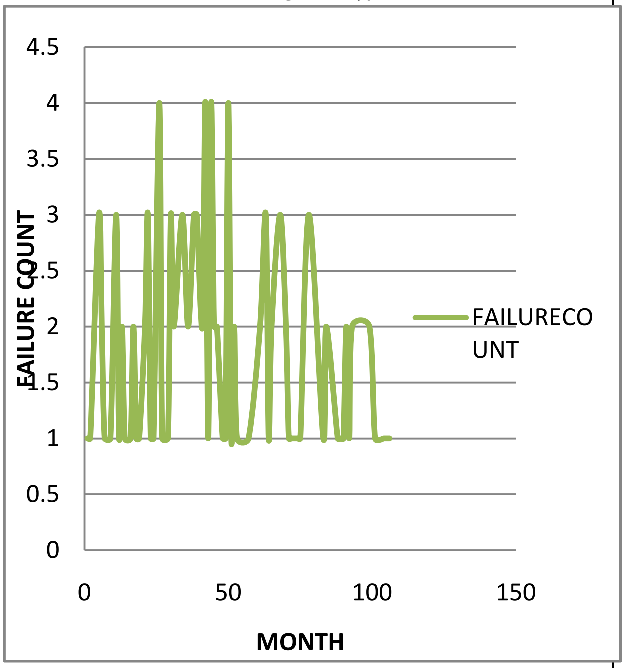
- Versions other than Apache2.Because here only Apache2 with major subversions are considered, thus versions other than Apache2 are not considered.
- The records whose status was Need Info. These types of records may or may not be considered as failure records.
- The records having low severity were deleted.
- Some of the records were irrelevant means there open date were less than that of the release date. All those records were deleted.
- Some of the records were without any versions, all those records were deleted.

VI. RESEARCH METHODOLOGY USED AND IDENTIFICATION OF DISTRIBUTION

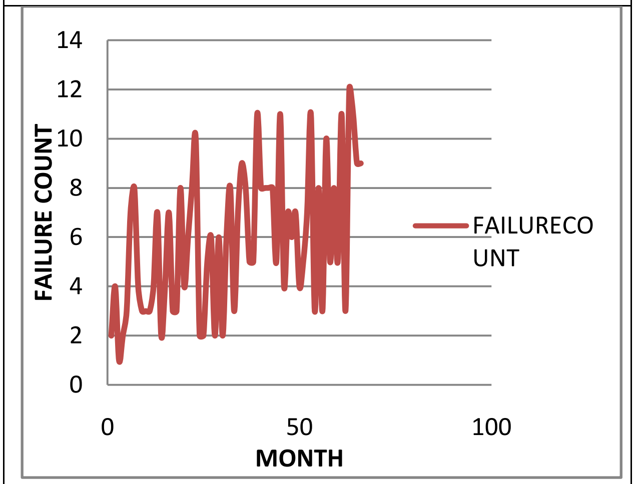
Before applying goodness of fit test on data collected for major sub versions of Apache2 series bug frequency corresponding to time to failure in week is plotted and shown in Table 1. Further from figure in Table 1 nothing can be concluded regarding its distribution as well as trend, thus Goodness of fit Test is applied for all the samples and best distribution is identified. These data are tested by three tests for 23 life data distributions.

Table 1: Failure Count of Apache 2.0 and Apache 2.2

APACHE 2.0



Apache 2.2



A. APACHE 2.0

From Mysql table on the basis of version, data related to Apache 2.0 version is extracted and stored in a separate table. By using appropriate sql query Time to Failure in terms of week is calculated and stored. For Apache 2.0 we have total of 112 preprocessed failure records.

Goodness of Fit Test

Goodness of Fit Test using Easy Fit is performed and result is stored in Table2.

On the basis of goodness fit test result in above table following distributions are found suitable for time to failure data of Apache 2.0 sample

- Gen. Extreme Value Distribution.
- Rayleigh (2P) Distribution and
- Weibull (3P) Distribution

Among these distributions on the basis of their test statistic ranking and detail result of goodness of fit test Gen. Extreme value Distribution is identified as best distribution to be fitted. The Detail of Goodness of fit test of this distribution is as shown in Table3.

Tsble2: Goodness of Fit – Summary

Distribution	Kolmogorov Smirnov		Anderson Darling		Chi-Squared	
	Statistic	Rank	Statistic	Rank	Statistic	Rank
Gen. Extreme Value	0.04783	1	0.46327	2	2.5651	1
Rayleigh (2P)	0.04862	2	0.51576	3	3.4363	2
Weibull (3P)	0.05254	3	0.45083	1	4.9185	3
Gamma (3P)	0.05935	4	0.54081	4	5.4208	4
Lognormal (3P)	0.05945	5	0.55453	6	6.2789	7
Fatigue Life (3P)	0.05992	6	0.54319	5	5.9276	5
Log-Logistic (3P)	0.06327	7	0.6779	8	6.1952	6
Beta	0.07135	8	0.59535	7	8.2549	13
Gamma	0.07428	9	1.9004	15	7.2914	10
Gumbel Max	0.07618	10	1.151	10	7.764	12
Rayleigh	0.07632	11	2.1334	16	7.0116	9
Gen. Gamma (4P)	0.07646	12	4.4752	19	N/A	
Normal	0.08124	13	1.0723	9	6.7778	8
Logistic	0.08258	14	1.6037	13	8.6466	15
Weibull	0.09517	15	1.8444	14	7.6194	11
Gen. Gamma	0.10421	16	1.4631	12	9.1317	16
Kumaraswamy	0.11334	17	1.3697	11	8.375	14
Lognormal	0.14404	18	3.7415	17	11.774	17
Log-Logistic	0.1484	19	4.1091	18	17.211	19
Gumbel Min	0.15183	20	5.7874	20	16.446	18
Frechet	0.2143	21	9.8675	21	28.736	20
Frechet (3P)	0.21834	22	11.323	22	N/A	
Fatigue Life	0.3087	23	18.386	23	49.855	21

B. Apache 2.2

After study of Apache 2.0, time to failure in terms of week is extracted for Apache 2.2 and stored in excel sheer For this sample we have total of 389 preprocessed data.

Goodness of Fit Test

Goodness of Fit Test is applied for collected sample data mentioned and result is shown in Table4:

Following is the list of distributions which are suitable for time to failure data of Apache 2.2 sample

- Gen. Gamma Distribution.
- Gen. Extreme Value Distribution.

Among these two distributions on the basis of their test statistic ranking and detail result of goodness of fit test Gen. Gamma Distribution is identified as best distribution to be fitted. The goodness of fit test result for Gen. Gamma Distribution is as given in Table 5 .

Table 3: Goodness of fit Detail for Gen. Extreme Value Distribution-Apache 2.0

Kolmogorov-Smirnov					
Sample Size Statistic P-Value Rank	112 0.04783 0.94905 1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.10139	0.11556	0.12832	0.14344	0.15393
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size Statistic Rank	112 0.46327 2				
A	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	6 2.5651 0.86111 1				
A	0.2	0.1	0.05	0.02	0.01
Critical Value	8.5581	10.645	12.592	15.033	16.812
Reject?	No	No	No	No	No

Table4: Goodness of Fit – Summary

	Kolmogorov Smirnov		**Anderson Darling**		**Chi-Squared**	
Distribution	Statistic	Rank	Statistic	Rank	Statistic	Rank
Gen. Gamma (4P)	0.02916	1	0.44158	1	5.5903	2
Beta	0.03503	2	6.1941	9	3.3941	1
Kumaraswamy	0.04667	3	12.519	15	N/A	
Gen. Extreme Value	0.05595	4	2.3652	2	14.406	3
Normal	0.06901	5	4.2274	4	26.363	8
Weibull (3P)	0.06976	6	3.7826	3	27.392	10
Lognormal (3P)	0.06989	7	4.5999	5	26.797	9
Gamma (3P)	0.07182	8	4.8279	8	25.696	6
Fatigue Life (3P)	0.07351	9	4.6869	6	29.321	12
Log-Logistic (3P)	0.0757	10	4.7048	7	20.61	4
Gumbel Min	0.08959	11	7.5283	13	47.409	15
Rayleigh	0.09013	12	6.883	10	22.102	5
Logistic	0.09133	13	7.04	11	39.203	13
Rayleigh (2P)	0.10238	14	7.4245	12	25.998	7

Gamma	0.10834	15	17.049	17	57.49	16
Gumbel Max	0.11456	16	17.389	18	66.098	19
Gen. Gamma	0.13441	17	12.999	16	44.27	14
Weibull	0.16383	18	10.855	14	29.052	11
Lognormal	0.16932	19	19.802	19	61.869	18
Log-Logistic	0.17935	20	20.025	20	59.687	17
Frechet	0.22712	21	40.092	21	118.62	20
Frechet (3P)	0.26366	22	40.538	22	N/A	
Fatigue Life	0.27565	23	56.017	23	149.8	21

Table 5: Goodness of Fit Detail of Apache 2.2 TTF Data for Gen. Gamma Distribution

Gen. Gamma (4P)					
Kolmogorov-Smirnov					
Sample Size	389				
Statistic	0.02916				
P-Value	0.88569				
Rank	1				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	0.0544	0.06201	0.06885	0.07697	0.08259
Reject?	No	No	No	No	No
Anderson-Darling					
Sample Size	389				
Statistic	0.44158				
Rank	1				
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
Chi-Squared					
Deg. of freedom	8				
Statistic	5.5903				
P-Value	0.69301				
Rank	2				
α	0.2	0.1	0.05	0.02	0.01
Critical Value	11.03	13.362	15.507	18.168	20.09
Reject?	No	No	No	No	No

This distribution is accepted by all three tests and at all level of significances.

• CONCLUSION:

On the basis of selected data for both versions appropriate distributions. Further from this distribution parameters may be estimated and further model may be constructed. This paper may be useful for researchers of any field.

References

[1] The Problems of Assessing Software Reliability when you really need to depend on it Bev Littlewood Centre for Software Reliability, City University

[2] London, UKJ. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] http://www.mathwave.com

[5] K-Sample Anderson-Darling Tests of Fit, for Continuous and Discrete Cases FW Scholz, MA Stephens in Citeseer (1986)

[6] Guo, H., Honecker, S., Mettas, A., and Ogden, D.,"Reliability Estimation for One-Shot Systems with Zero Component Test Failures, *2010 Annual Reliability and Maintainability Symposium*.

[7] John D. Musa. Software reliability data. Technical report,Data & Analysis Center for Software, January 1980. http://www.dacs.dtic.mil/databases/sled/swrel.shtml [27 Januray 2003].

[8] A Bayesian Model for Predicting Reliability of Software Systems at the Architectural Level Roshanak Roshandel1, Nenad Medvidovic2, Leana Golubchik2,3 Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[9] R. Allen, and D. Garlan. A Formal Basis for Architecture Connection. *ACM Transactions on Software Engineering and Methodology*, 6(3): p.213-249, 1997.

[10] Estimation of the Unknown Parameters of the Generalized Frechet **Distribution** Abd-1 Elfattah, A.M and 2Omima, A.M.

[11] A. Mockus, T.R. Fielding, and J.D. Herbsleb, "Two case studies of open source software development: Apache and Mozilla", ACM Transactions on Software Engineering and Methodology, vol. 11, no. 3, July 2002.

[12] T.R. Moss, "The Reliability Data Handbook", ASME Press, 2005.

[13] J.D. Musa and K. Okumoto, "A logarithmic poisson execution time model for software reliability measurement", 7th Int'l Conference on Software Engineering (ICSE), 1984.

[14] Cobra Rahmani, Harvey Siy, Azad Azadmanesh "An experimental Analysis of open Source Software Reliability".

[15] On line bug repositories Bugzilla: http://www.bugzilla.org.

[16] X distribution, Lifetime Data Analys is , 7: 187-200.H. Goto, Y. Hasegawa, and M. Tanaka, "Efficient Scheduling Focusing on the Duality of MPL Representatives," Proc. IEEE Symp. Computational Intelligence in Scheduling (SCIS 07), IEEE Press, Dec. 2007, pp. 57-64, doi:10.1109/SCIS.2007.357670. **(Article in a conference proceedings)**

Biography:

Sanjeev Kumar Jha, Senior Systems Analyst, NIELIT Chandigarh, Ministry of C&IT Government of India. Bachelors and Masters Degree in Statistics [Honors] from Patna University and presently doing his PhD in Computer Science from Singhania University, Rajasthan. He has attended 2 international Conferences. Area of research interest is Reliability and Open Source Software.

Shivlok Singh , Programmer NIELIT is presently doing his PhD in Computer Science from Singhania University, Rajasthan. He has attended 2 international Conferences. Area of research interest is Reliability and Open Source Software

Dr. Amod Tiwari, Associate Professor, PSIT Kanpur. PhD from IIT Kanpur. Attended many international and national Conferences. Published many papers in international and national journals. Area of research interest is Image Processing, Reliability and Open Source Software.