World Applied Sciences Journal 15 (8): 1181-1185, 2011
ISSN 1818-4952

Determination of Plotting Position Formula for the Normal, Log-Normal, Pearson(III), Log-Pearson(III) and Gumble Distributional Hypotheses Using The Probability Plot Correlation Coefficient Test

[1]*Fuladipanah Mehdi and* [2]*Jorabloo Mehdi*

[1]Department of Civil Engineering, Ramhormoz Branch, Islamic Azad University, Ramhormoz, Iran
[2]Department of Engineering, Garmsar Branch, Islamic Azad University, Garmsar, Iran

Abstract: The selection of an appropriate plotting plot formula for each statistical distribution is the most important step that is generally chosen by using the goodness of fit tests. Probability Plot Correlation Coefficient (PPCC) test which was developed by Filliben for normality is a powerful tool among the goodness of fit tests. PPCC test is based on probability plot and correlation coefficient. In this paper, PPCC test was used for data with Normal, Log-Normal, Pearson, Log-Pearson and Gumbel distributions. This statistical test was used for sample of length n=10, 20, 30, 40, 50 and 100. Using Chow's formula, frequency factor was determined California, Hazen, Weibul, Blom, Chegodayov, Tukey and Gringorton plotting position formulas. Critical points of the test statics were provided for samples of mentioned length. Results showed that Blom, Gringorton and Weibul plotting position formula are appropriate for the Normal(Log-normal), the Gumbel and PearsonIII (Log-pearsonIII) distributions, respectively.

Key words: Probability Plot Correlation Coefficient Test · Statistical Distributions · Plotting Position Formula · Critical Points

INTRODUCTION

Accurate analysis of recorded data plays an important role in the decision-making process in hydrological and hydraulic projects. It have been developed many goodness of fit tests in literatures such as the Kolmogorov-Smirnov test, the Cramer von Mises test and the Chi-square test which are popular especially [1]. Filliben developed powerful probability plot correlation coefficient (PPCC) tests for normality which have following attractive features: 1) the test statistics is conceptually easy to understand because it combines two fundamentally simple concepts: the probability plot and correlation coefficient, 2) the test is computationally simple since it is only requires computation of a simple correlation coefficient, 3) the test statistics is readily extendible for testing some non-normal distribution hypotheses, 4) the test compares favorably with seven other tests of normality on the basis of empirical power studies performed by Filliben, 5) the test is invariant to the parameter estimation procedure employed to fit the probability distribution and 6) the test allows a comparison of the results in both a graphical (probability plot) and a numerical (correlation coefficient) form [2, 3]. Probability plots have been used widely in many resources investigations. While analytic approaches for fitting probability distributions to observed data are, in theory, more efficient statistical procedures than graphical curve fitting procedures, many hydrologists wouldn't make engineering decisions without the use of a graphical display (probability plot). Probability plots were recommended by the National Research Council as a basis for extrapolation of flood frequeccy curve in dam safety evaluations [1]. Similarly, The Federal Emergency Management Agency recommends the use of probability plots in determination of the probability distribution of annual maximum flood elevations which arises from the combined effects of ice jam and storm-induced flodding [1]. Looney and Gulledge applied various plotting position formulas to normal distribution and chose the Blom plotting position formula for the derivation of normal PPCC test statistics [4]. Vogel proposed the PPCC test statistics for the Gumbel distribution [3] and Vogel and Kroll derived the PPCC test statistics for the 2-parameter

Corresponding Author: Fuladipanah Mehdi, Department of Civil Engineering, Ramhormoz Branch, Islamic Azad University, Ramhormoz, Iran.

Weibull and uniform distributions in frequency analysis for low flow data [5]. In addition, the PPCC test statistics of 5% significance level for gamma distribution are studied by Vogel and McMartin [6] and the PPCC test statistics for the GEV distribution are provided by Chowdhury *et al*. [7]. Recently, Heo *et al*., proposed the regression equations to estimate the test statistics for normal, gamma, Gumbel, GEV and Weibull distributions [8]. Sooyoung *et al*., used PPCC test for generalized logistic distribution. They used PPCC test extreme value distribution. They found that this test is very strong tool to find appropriate distribution [9]. Goda had studied different plotting plot formulas and their application in various statistical distributions [10]. Cook conducted a study about the effect of sampling error on plotting position formula in extreme values. His study was about normal distribution [11].

In this paper, the Chow formula was used to apply the PPCC test for the Normal, Log-normal, PearsonIII, Log-pearsonIII and Gumbel distributions. The Chow formula is as following [1]:

$$X_{TR} = XMEAN + K.S \quad (1)$$

Which X_{TR} is amount of variable with return period TR, XMEAN is mean of variable X_i, K is frequency factor and S is standard deviation of X_i. In this equation, K is a function of probability occurrence. There are different relations for calculation of empirical probability occurrence which are listed in Table 1 [1]. Therefore, the selection of the best relationship is very important. The critical points (or significance levels) of distributions of r (correlation coefficient) were obtained by using the empirical sampling procedure.

MATERIALS AND METHODS

If the sample to be tested is actually distributed as hypothesized, one would expect the plot of the ordered observations versus the order statics means or medians to be approximately linear. Thus the product moment correlation coefficient which measures the degree of linear association between two random variables is an appropriate test static. Filliben's PPCC test is a simply formalization of a technique used by statistical hydrologists for many decades; that is, it determines the linearity of a probability plot. In this study, in order to select appropriate plotting position formula for Normal, Log-Normal, PearsonIII, Log-pearsonIII and Gumbel distributions the following steps were used:

I Start.
II Generate random data.
III Select statistical distribution.
IV Fit statistical distribution (step III) on the generated data (step II).
V Classify fitted data (step IV) to samples of length n.
VI Fit statistical distribution (step III) on classified data (step IV) using equation (1).
VII Calculate r between data and frequency factor.
VIII Determine appropriate plotting position formula for statistical distribution (step III) using maximum r values in PPCC test.
IX Determine critical points of r using PPCC test for statistical distribution (step III).
X End.

This algorithm was applied for the following statistical distribution:

- Normal and Log-Normal distribution
- Pearson type III and Log-Pearson type III distribution
- Gumbel distribution

Detailed of calculation for each distribution is presented in the following subsections.

Normal and Log-nomral Distributions: To determine appropriate plotting position formula for normal distribution, the following steps were considered:

I Random data (x_i) of length 1,000,000 were generated in interval [0,1].
II Data were divided into samples of length n: 10, 20, 30, 40, 50, 100. For example, there will be 100000 classes for sample of length length n=10.
III The normal distribution was fitted on data (step 2) using the following equations [6]:

$$\text{If } x_i \geq 0.5 \rightarrow r_i = \sqrt{\frac{1}{x_i^2}} \rightarrow x_{Ni} = r_i - \frac{2.515517+0.802853r_i + 0.010328r_i^2}{1+1.432788r_i + 0.189269r_i^2 + 0.001308r_i^2} \quad (2)$$

$$\text{If } x_i < 0.5 \rightarrow r_i = \sqrt{\frac{1}{(1-x_i)^2}} \rightarrow x_{Ni} = -[r_i - \frac{2.515517+0.802853r_i\, 0.010328r_i^2}{1+1.432788r_i + 0.189269r_i^2 + 0.001308r_i^3} \quad (3)$$

Which x_i is generated data in step 1, r_i is a coefficient and x_{Ni} is normalized value of x_i.

Table 1: Plotting Position Formulas

Formula	Year	Description
California	1923	$P(x \geq X)=\frac{m}{N}$
Hazen	1930	$P(x \geq X)=\frac{2m-1}{2N}$
Weibul	1939	$P(x \geq X)=\frac{m}{N+1}$
Blom	1954	$P(x \geq X)=\frac{m-0.375}{N+0.25}$
Chegodayov	1955	$P(x \geq X)=\frac{m-0.3}{N+0.4}$
Tukey	1962	$P(x \geq X)=\frac{3m-1}{3N+1}$
Gringorton	1963	$P(x \geq X)=\frac{m+0.12}{N-0.44}$

IV Data (step 3) in each class were arranged in the descending order. Probability of each data in each class was calculated using the formulas were listed in the Table 1.

V Corresponding frequency factors were calculated using following equation for the normal distribution [6]:

$$\text{If } P \geq 0.5 \rightarrow w=\sqrt{\frac{1}{P^2}} \rightarrow K_N = w-\frac{2.515517+0.802853w+0.010328w^2}{1+1.432788w+0.189269w^2+0.001308w^3} \quad (4)$$

$$\text{If } P < 0.5 \rightarrow w=\sqrt{\frac{1}{(1-P)^2}} \rightarrow K_N = -[w-\frac{2.515517+0.802853+0.010328w^2}{1+1.432788w+0.189269w^2+0.001308w^3} \quad (5)$$

Which P is probability, w is a coefficient and K_N is frequency factor for normal distribution.

I Correlation coefficients were calculated between each class in steps 3 and 7. These coefficients (for example, if n=10 then the number of r equals 100000) were used for two purposes:

II The maximum values of r were applied to select appropriate plotting position formula.

III The average of r corresponding to appropriate probability formula was calculated as critical points.

The obtained results for normal distribution can be developed to Log-normal distribution; because the logarithm of normalized data has normal distribution. Final results are presented in Table 2.

Table 2: Critical points of r which r is the Normal Probability Plot Correlation Coefficient

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.99579	0.99297	0.99174	0.98922
20	0.99640	0.99503	0.99380	0.99264
30	0.99630	0.99071	0.99514	0.99445
40	0.99717	0.99671	0.99641	0.99553
50	0.99710	0.99669	0.99652	0.99574
100	0.99812	0.99812	0.99812	0.99721

Pearson Type III and Log-Pearson Type III Distributions: Six mentioned steps in subsection Normal ang Log-Normal were used for pearson type III distribution, too. However, to fit Pearson III distribution on generated data in step 3, the following equation was used [6]:

$$xp_i = \frac{2}{g}\{[\left(x_{Ni}-\frac{g}{6}\right)\frac{g}{6}+1]^3-1\} \quad (6)$$

Which x_{Ni} is normalized data, g is skewness coefficient and x_{Pi} is transformed data to pearson distribution. The amount of g for Pearson III distribution fitted data is often in interval [-4,4]; thus, x_{Pi} were calculated for g= -4,-3,-2,-1,1,2,3 and 4. Another difference is in step 5. Frequency factor for pearsonIII distribution is calculated as following equation [6]:

$$Kp = -\frac{2}{g}\{[\left(K_N-\frac{g}{6}\right)\frac{g}{6}+1]^3-1\} \quad (7)$$

Which K_P and K_N are frequency factor of pearson and normal distribution respectively, g is skewness coefficient. Amounts of K_P were calculated for g= -4,-3,-2,-1,1,2,3 and 4.

The obtained results for normal distribution can be developed to Log-normal distribution; because the logarithm of pearson distribution fitted data have pearson distribution. Final results are presented in Tables 3 to 10.

Gumbel Distribution: All steps mentioned in subsection Normal and Log-Normal are applied for the Gumbel distribution, too. There are two differences in steps 3 and 5. The following equation is used to fit the Gumbel distribution on generated data in step 3 [6]:

$$x_{Gi} = \left(-\frac{\sqrt{6}}{\pi}\right)[0.5772+\text{Ln}\left(\text{Ln}\left(\frac{1}{1-x_i}\right)\right)] \quad (8)$$

Table 3: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g= -4

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.97734	0.96566	0.94971	0.93065
20	0.94778	0.93751	0.91506	0.89667
30	0.93421	0.92521	0.89613	0.87157
40	0.91625	0.91016	0.88895	0.86952
50	0.91180	0.90072	0.87234	0.85708
100	0.90219	0.85785	0.85683	0.84428

Table 4: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g= -3

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.98572	0.97975	0.97220	0.96270
20	0.96914	0.96260	0.94870	0.93484
30	0.96122	0.95932	0.93998	0.91817
40	0.94817	0.94291	0.93239	0.91704
50	0.93872	0.93748	0.91252	0.90567
100	0.93377	0.90372	0.89670	0.88494

Table 5: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g= -2

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.99170	0.98977	0.98534	0.98043
20	0.98671	0.98359	0.97857	0.97145
30	0.98545	0.98370	0.97706	0.96667
40	0.98004	0.97770	0.97322	0.96328
50	0.97649	0.97419	0.96458	0.95893
100	0.97108	0.95994	0.95716	0.94596

Table 6: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g= -1

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.99426	0.99222	0.98933	0.98601
20	0.99424	0.99245	0.99075	0.98825
30	0.99430	0.99386	0.99201	0.98974
40	0.99639	0.99445	0.99241	0.98942
50	0.99451	0.99396	0.99225	0.98983
100	0.99322	0.99322	0.98947	0.98785

Table 7: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g=1

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.99303	0.99152	0.98896	0.98612
20	0.99405	0.99328	0.99100	0.98781
30	0.99460	0.99354	0.99140	0.98950
40	0.99597	0.99357	0.99150	0.98810
50	0.99524	0.99412	0.99160	0.98811
100	0.99058	0.98940	0.98681	0.98520

Table 8: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g=2

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.99202	0.98939	0.98505	0.97867
20	0.98846	0.98689	0.97666	0.96986
30	0.98538	0.97914	0.97000	0.96377
40	0.98114	0.97355	0.96436	0.95560
50	0.97234	0.96959	0.96250	0.95511
100	0.94968	0.94655	0.94377	0.93961

Table 9: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g=3

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.98614	0.97895	0.96867	0.95674
20	0.96932	0.96099	0.94329	0.93224
30	0.95128	0.94532	0.92488	0.91243
40	0.94245	0.93338	0.92315	0.90044
50	0.92619	0.92456	0.90805	0.89835
100	0.88725	0.88261	0.87964	0.87460

Table 10: Critical points of r which r is the PearsonIII and log-Pearson Probability Plot Correlation Coefficient for g=4

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.97277	0.96363	0.94298	0.92173
20	0.95185	0.92635	0.90548	0.88849
30	0.91641	0.90808	0.88739	0.86942
40	0.91160	0.90040	0.87637	0.85456
50	0.88983	0.88621	0.87302	0.85144
100	0.83847	0.83762	0.83217	0.82568

Table 11: Critical points of r which r is the Gumbel Probability Plot Correlation Coefficient

	Significance Level			
Class Length	99%	98%	95%	90%
10	0.99400	0.99280	0.99060	0.98730
20	0.99480	0.99450	0.99220	0.99070
30	0.99640	0.99600	0.99430	0.99350
40	0.99710	0.99580	0.99450	0.99330
50	0.99840	0.99670	0.99540	0.99450
100	0.99890	0.99750	0.99700	0.99640

Which x_i is random generated data and x_{Gi} is corresponding value of x_i with gumbel distribution. Also, the following equation is used to calculate frequency factor in the Gumbel distribution in step 5 [6]:

$$K_G = \left(-\frac{\sqrt{6}}{\pi}\right)\left[0.5772 + \mathrm{Ln}\left(\mathrm{Ln}\left(\frac{1}{1\text{-}P}\right)\right)\right] \quad (9)$$

Which P is occurrence probability.The final results are presented in Table 11.

RESULTS AND DISCUSSION

The probability plot correlation coefficient test is an attractive and useful tool for testing the normal, log-normal, pearson III, log-pearson III and Gumbel distributions. In this paper, seven plotting position formulas were used for mentioned distributions. The PPCC test was applied for samples of length 10, 20, 30, 40, 50 and 100.

The results show that the PPCC test is flexible, because it is not limited to any sample size. In addition, this test was developed for normality initially, whereas it is extendible readily to non-normal hypotheses. Different plotting position formulas cause different results. Therefore, appropriate plotting position formula plays important role to fit distributions on random data.

Blom, Gringorton and Weibul plotting position formulas are recommended for the Normal (Log-Normal), Pearson III(log-pearson) and Gumble distributions, respectively.

Vogel and Kroll recommendation using Blom plotting position formula for Pearson III and Log-Pearson III is rejected.

Critical points were calculated for samples of different length. They are applicable as at least value to fit statistical distribution on a sample data of certain length.

CONCLUSION

The selection of appropriate plotting position formula and the calculation of critical points for the Normal, log-Normal, Pearson III, log-pearson III and Gumbel distributions were studied in this paper. Filliben's PPCC test was used for this purpose. The test was applied for $g \in [-4,4]$ for the PearsonIII and log-pearsonIII distribution. The results were in the agreement of the other studies with the exception of the PearsonIII and log-pearsonIII distributions. Weibul plotting position formula was recommended for these two distributions unlike Vogel and Kroll recommendation.

REFERENCES

1. Raghunath, H.M., 2007. Hydrology. Principles. Analysis. Design, New Age International Press, New Delhi, pp: 327-336.
2. Filliben, J.J., 1975. The Probability Plot Correlation Coefficient Test for Normality. Technometrics, 17(1): 111-117.
3. Vogel, R.M., 1986. The probability plot correlation coefficient test for the normal, lognormal and Gumbel distributional hypothesis, Water Resources Res., 22(4): 587-590.
4. Looney, S.W. and T.R. Gulledge, 1985. Use the correlation coefficient with normal probability plots. The American Statistician, 39(1): 78-79.
5. Vogel, R.M. and C.N. Kroll, 1989. Low-flow frequency analysis using probability plot correlation coefficients, J. Water Resources Planning and Management, 115(3): 338-357.
6. Vogel, R.M. and D.E. McMartin, 1991. Probability plot goodness-of-fit and skewness estimation procedures for the Pearson type 3 distribution, Water Resources Res., 27(12): 3149-3158.
7. Chowdhury, J.D., J.R. Stedinger and L.H. Lu, 1991. Goodness-of-fit tests for regional generalized extreme value flood distributions. Water Resources Res., 27(7): 1765-1776.
8. Heo, J., Y. Kho, Y.H. Shin, S. Kim and T. Kim, 2007. Regression Equations of Probability Plot Correlation Coefficient Test Statistics from Several Probability Distributions. J. Hydrol., 355(1-4): 1-15.
9. Kim, S., H. Shin, T. Ki and J.H. Heo, 2010. Derivation of the probability plot correlation coefficient test statistics for the generalized logistic distribution. International Workshop of advances in statistical hydrology, May 23-25, Taormina, Italy.
10. Goda, Y., 2010. Plotting postion estimator for the L-moment method and quantile confidence interval for the GEV, GPA and weibull distributiona applied for extreme analysis. Coastal Engineering Journal, 53(2): 111-149.
11. Cook, N., 2011. Comments on Plotting Positions in Extreme Value Analysis. J. Applied Meteorology and Climatol., 50: 255-266.
12. Vito, I., M. Fiorentino, A. Gioia and S. Manfreda, 2010. Best Fit and Selection of Theoretical Flood Frequency Distributions Based on Different Runoff Generation Mechanisms, Water, 2: 239-256.