Determination of the Best Probability Plotting Position for Predicting Parameters of the Weibull Distribution

Ahmad Shukri Yahaya
Chong Suat Yee
Nor Azam Ramli
Fauziah Ahmad

Clean Air Research Group
School of Civil Engineering
Universiti Sains Malaysia
Engineering Campus, 14300 Nibong Tebal
Seberang Perai Selatan, Pulau Pinang.

Abstract

Probability plotting position methods was compared to estimate the shape and scale parameters of the Weibull distributions. Simulation technique was used to obtain random variables for twelve different Weibull distributions. Ten probability plotting positions were compared namely Hazen, California, Weibull, Blom, Gringorten, Chegodayev, Cunnane, Tukey, Beard and Median formulas. To determine the best probability plotting positions, three error measures were used which are the normalized absolute error, mean absolute error and the root mean square error. This study shows that the Gringorten formula performs the best for all sample sizes.

Keywords: Weibull distribution, Probability plotting positions, Error measures

1. Introduction

Regression method is one of several techniques to estimate parameters of a distribution such as the Weibull distribution. To estimate the parameters, probability plotting positions are used to represent the cumulative distribution function of the chosen distribution. The slope and the intercept of the regression line were then used to estimate the parameters of the distribution. Weibull distribution have been used to fit distributions in air pollution studies and to determine return periods(Nor Faizah Fitri et al., 2010; Seinfeld and Pandis, 1998; Maffeis ,1998). This distribution have also been used successfully in fitting distributions for wind speed (Shoji , 2005; Jaramillo and Borja,2004 and Ahmad Shukri Yahaya et al. (2007). Weibull distribution has also been used in life testing and reliability theory. Wang and Keats (1995) describe the use of four plotting positions to estimate parameters of the Weibull distribution. They used simulation method to obtain the time to the r^{th} failure and the proportion of the distribution of the failure. They found that the proposed method is useful and have good properties for estimation of Weibull parameters when used for censored and uncensored data.

Ross (1994) described a technique to obtain parameters of the Weibull distribution in which three types of plotting positions are reviewed that are median probability, mean probability and mean plotting position. The mean type gives unbiased estimators by linear regression estimation, thus it was proposed that the best parameter estimators for the Weibull distribution is the mean plotting formulae. Zhang *et al.* (2006) used the Bernard's median rank estimator and Herd-Johnson estimator to fit regression models of the form of Y on X and X on Y. They found that a better model can be determined by the value of shape parameter, β. When β <1, LS Y on X will provide a better model. Otherwise, LS X on Y will give a better model. Other regression methods using probability positions using different distributions are discussed in Shabri (2002), Lund *et al*., (1998), Adeboye and Alatise (2007), Looney and Gulledge (1985). This paper compares ten types of probability plotting positions to determine the best probability plotting position for the Weibull distribution. The Weibull distribution was chosen because it is one of the most widely used distributions in air pollution modeling and in reliability analyses.

2. Materials and Methods

2.1 The Weibull distribution

The Weibull distribution is widely used in studies of air pollution modeling, reliability and wind load studies.

In probability theory and statistics, the Weibull distribution is a continuous probability distribution with the probability density function (pdf) defined on the interval $[0,\infty]$. A continuous random variable x has a Weibull distribution (Bury, 1999) if its probability density function (pdf) can be expressed as

$$f(x)=\frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1}\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right], x>0, \alpha>0, \beta>0 \quad (1)$$

and the cumulative distribution function (cdf) takes the form

$$F(x)=1-\exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right], x>0, \alpha>0, \beta>0 \quad (2)$$

where α is the scale parameter and β is the shape parameter. The scale parameter controls the spread of the distribution, and the shape parameter controls the form of the distribution.

2.1.1 Estimating parameters using regression method

A common practice among engineers is to plot the observation against the estimated cumulative distribution function of the Weibull distribution on a Weibull probability paper and then fit the line to the data point (Zhang *et al*., 2006).The intercept and gradient of the straight line will give the estimated values for the scale and shape parameters of the Weibull distribution respectively.

Rearranging terms and taking natural logarithms from the cdf of the Weibull distribution, a straight line relation is obtained which is given by

$$\ln\left(x_{(i)}\right)=\ln\alpha+\frac{1}{\beta}\ln\left(-\ln\left[1-\hat{F}\left(x_{(i)}\right)\right]\right) \quad (3)$$

where $x_{(i)}$is the ordered observations for x. A regression equation will be obtained with $\ln\left(x_{(i)}\right)$ as the dependent variable and $\ln\left(-\ln\left[1-\hat{F}\left(x_{(i)}\right)\right]\right)$ as the independent variable. The intercept and gradient of the regression line will give the estimated values for the scale and shape parameter of the Weibull distribution respectively.

2.1.2 Probability plotting positions

Probability plots for the Weibull distribution was used to estimate the shape and scale parameters. The best quantile estimate made from the plotting formula should be unbiased and should have the smallest root means error among all such estimates (De, 2000).

In this study, ten types of probability plotting positions that are commonly used namely Hazen, California, Weibull, Blom, Gringorten, Chegodayev, Cunnane, Tukey, Beard and Foster (Akbar, 2006; Shabri, 2002). These probability plotting positions are used to estimate the cumulative distribution function of the Weibull distribution. All of the plotting position formulas in this study are summarized in Table 1.

Table 1. Potting Position Formulas (Source: Rao and Hamed,2000)

Case	Plotting Formula	Probability Plotting Position
1	Hazen	$\frac{(i-0.5)}{n}$
2	California	$\frac{i}{n}$
3	Weibull	$\frac{i}{(n+1)}$
4	Blom	$\frac{(i-3/8)}{(n+1/4)}$
5	Gringorten	$\frac{(i-0.44)}{(n+0.12)}$
6	Chegodayev	$\frac{(i-0.3)}{(n+0.4)}$
7	Cunnane	$\frac{(i-0.4)}{(n+0.2)}$
8	Tukey	$\frac{(i-1/3)}{(n+1/3)}$
9	Beard	$\frac{(i-0.31)}{(n+0.38)}$
10	Median	$\frac{(i-0.3175)}{n+0.365}$

2.2 Error Measures

Three error measures were used to determine the best estimator. The three error measures are the normalized absolute error (NAE), mean absolute error (MAE) and root mean square error (RMSE) (Armstrong and Collopy, 2000). Table 2 below gives the formulae for the three error measures.

Table 2. Error measures

Error Measures	Formula
NAE	$\frac{\sum_{i=1}^{N}\lvert P_i - \mathrm{O}_i\rvert}{\sum_{i=1}^{N}\mathrm{O}_i}$
MAE	$\frac{\sum_{i=1}^{N}\lvert P_i - \mathrm{O}_i\rvert}{N}$
RMSE	$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i - \mathrm{O}_i)^2}$

*P is the predicted value, O is the observed value and N is the number of data

3. Data

Simulation of random variables for Weibull distribution was used to compare the performance of the probability plotting positions. Twelve different Weibull distributions (represented by the $W(\alpha,\beta)$) were used the random variables namely W(0.7,1.7), W(1,1.2), W(3.2,1.4), W(13,1.3), W(14,1.8), W(17,2), W(26,2), W(30,1.5),W(40,1.1), W(80,1.6), W(685,2) and W(233,1.9). The values of the shape and scale parameters were chosen to represent various forms of the Weibull distribution. Figure 1 shows the probability density function (pdf) of nine of the selected Weibull distribution.

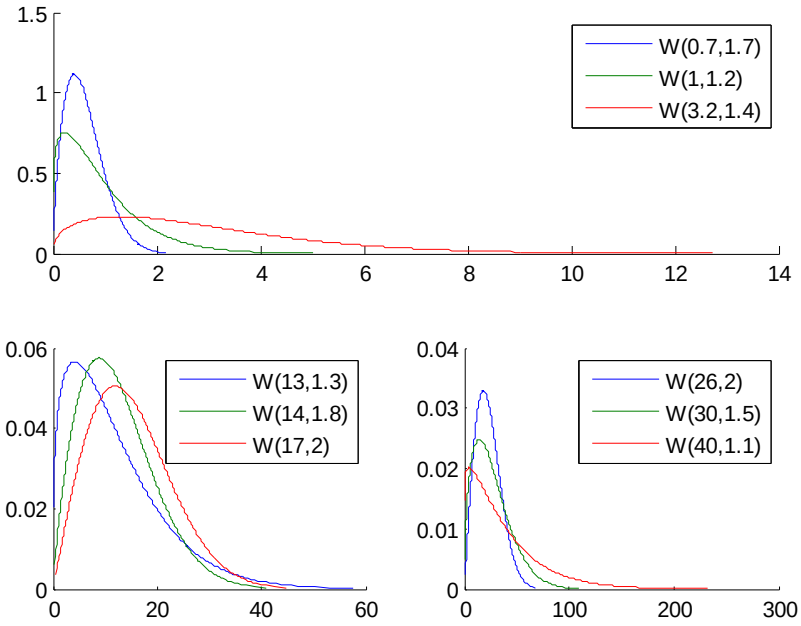

Figure 1. PDF of selected Weibull Distributions

Sample of sizes 10, 20, 50, 60, 80 and 100 were used. Each of the sample sizes were replicated ten times for each Weibull distribution. For each sample sizes, 10000 random variables were generated. Multiplicative congruential generator (Law and Kelton, 2000) of the form

$$X_i = (397204094 \ X_{i-1}) \bmod (2^{31} - 1) \quad (4)$$

was used to simulate the random variables. The above multiplicative congruential generator was found to have good statistical properties.

From these simulations, the predicted values of the random variables for each Weibull distributions of the ten probability plots were obtained. The errors between observed and predicted values were then obtained and were used to determine the best probability plotting position.

4. Results and Discussions

The choice of probability plotting positions to estimate the cumulative distribution function will always give different return periods for a particular dataset (De, 2000). This is because the different probability plotting positions will result in different values of the parameter estimates. Ross (1994) compares three plotting positions namely Weibull, Gringorten and Chegodayev and showed that these three methods can be used with satisfactorily results.

For each Weibull distribution, the error measures of the ten probability plots were obtained. The average values of the error measures from the twelve distributions were obtained. This will determine the best probability plotting positions. Tables 3 to 5 show the average normalized absolute error, mean absolute error and root mean square error.

From Table 3, is can seen that the average NAE will reduce as the sample size increases. This shows that the accuracy of the predictions using the probability plotting positions increases as the sample size increases. Table 1 also shows that the Gringorten method is the best for all sample sizes except when the sample size is 100. For small sample sizes (n = 10 and n = 20) the best plotting position is the Gringorten method and followed by the Blom and Cunnane methods. For medium sample sizes (n = 50 and n = 60), the three best methods are Gringorten, Cunnane and Chegodayev plotting positions. For large sample sizes (n = 80 and n = 100), the three best methods are Cunnane, Gringorten and Hazen plotting positions. Overall, the Gringorten method gives the best estimation and this is followed by the Cunnane and Tukey method respectively.

Table 3. Average Normalized Absolute Error (NAE)

PPP	$n = 10$	$n = 20$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
Weibull	0.4885	0.2549	0.1336	0.1250	0.0893	0.0786
Hazen	0.5107	0.2553	0.1129	0.1193	0.0848	0.0733
Median	0.3933	0.2333	0.1280	0.1151	0.0799	0.0715
Blom	0.3774	0.2293	0.1153	0.1139	0.0783	0.0843
California	0.3827	0.2489	0.1243	0.1188	0.1011	0.0749
Gringorten	0.3610	0.2233	0.1122	0.1104	0.0767	0.0715
Chegodayev	0.3985	0.234	0.1153	0.1137	0.0843	0.0747
Cunnane	0.3785	0.2269	0.1143	0.1131	0.0776	0.0710
Tukey	0.3894	0.2318	0.1149	0.1140	0.0794	0.0717
Beard	0.3911	0.2331	0.1145	0.1155	0.0811	0.0729

Table 4. Average Mean Absolute Error (MAE)

PPP	$n = 10$	$n = 20$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
Weibull	27.5314	11.4549	7.87725	8.65467	6.48108	4.17283
Hazen	30.4142	10.7352	10.1723	8.1875	5.61117	4.08967
Median	21.9706	9.35558	7.61308	7.96725	5.65567	4.56633
Blom	21.1256	9.0205	8.53667	7.78	5.50775	4.14808
California	23.2603	11.241	7.46225	8.33992	6.40167	4.21025
Gringorten	20.3087	8.71325	7.75367	7.71992	5.31692	4.1075
Chegodayev	22.1601	9.4385	7.55575	7.93525	7.98242	4.14975
Cunnane	20.9895	8.88075	7.66917	7.78783	5.43	4.18675
Tukey	21.6302	9.284	7.76692	7.85717	5.61808	4.17283
Beard	22.5247	9.323	7.87725	7.976	6.29842	4.08967

Table 5. Average Root Mean Square Error (RMSE)

PPP	$n = 10$	$n = 20$	$n = 50$	$n = 60$	$n = 80$	$n = 100$
Weibull	45.3522	15.9183	11.8402	11.5548	9.104	6.8265
Hazen	44.7713	14.4617	10.7482	10.7218	7.85242	5.95267
Median	33.5845	12.4922	13.6397	10.4276	7.70292	5.89517
Blom	31.8565	11.914	10.2843	10.2579	7.43617	5.75983
California	35.8328	14.4076	11.4986	11.1183	8.60383	6.42217
Gringorten	30.0731	11.3902	10.0458	10.0984	7.2115	5.8615
Chegodayev	34.0493	12.2725	10.5279	10.45	10.2923	5.95617
Cunnane	31.3194	11.6765	10.1956	10.2097	7.31083	5.78675
Tukey	32.2348	12.3426	10.3952	10.3265	7.64975	5.86867
Beard	34.3303	12.4267	10.5277	10.4862	8.10217	5.92442

From Tables 4 and 5, the average MAE and RMSE, decreases as the sample size increases for all probability plotting positions. The results from Table 4 show that for small sample sizes the best plotting position is the Gringorten method and followed by the Cunnane and Blom methods. For medium sample sizes, the three best methods are California, Cunnane and Chegodayev plotting positions. For large sample sizes, the three best methods are Gringorten, Blom and Hazen plotting positions. Overall, for the average MAE the best method is the Gringorten method. Using the average RMSE measure, for small and medium sample sizes the three best plotting positions are the Gringorten, Cunnane and Blom methods. For large sample sizes, the three methods have the same values of error measure.

The best probability plotting positions was also obtained for all the error measures simultaneously. It was found that the best method for all sample sizes is the Gringorten plotting position and followed by Cunnane and Blom respectively.

5. Conclusion

Choosing the best probability position is important for estimating parameters of distributions and hence in determining the best estimate for return periods. This paper uses the simulation method to obtain random observations from the twelve different Weibull distributions. Ten Probability plotting positions were used to estimate the cumulative distributions. Three different error measures were used for determining the performance of these plotting positions.

From the results, as the sample size increases, the values of error measures decreases. Given the result of the simulation, it was shown that the accuracy of the estimates of the shape and scale parameters of the Weibull distribution increases when an appropriate plotting position is used. For all sample sizes the Gringorten method produces the best estimate. This is followed by the Cunnane and Blom method.

6. References

Ahmad Shukri Yahaya, Nor Azam Ramli, Aeizaal Azman Abdul Wahab (2007). Finding The Best Wind Speed Distribution: A Case Study, *World Engineering Congress 2007 (WEC2007)*, 5-9 August 2007, Penang, Malaysia, 14-19.(CD Proceedings)

Adeboye, O.B. and Alatise, M.O.(2007). Performance of Probability Distributions and Plotting Positions in Estimating the Flood of River Osun at Apoje Sub-basin, Nigeria,. *Agricultural Engineering International: the CIGR E-Journal*. Manuscript LW 07 007. Vol. 9

Akbar, N.A. (2006). Flood *Damage Assessment Model Using Cost-Benefit Analysis*. Master Thesis, University Teknology Malaysia. pp. 1-15

Armstrong, J.S. and Collopy, F. (2000). *Another Error Measures For Selection Of The Best Forecasting Method : The Unbiased Absolute Percentage Error* [Online]. Available: http://hops.wharton.upenn.edu/forecast/paperpdf/armstrong-unbiasedAPE.pdf. (August 25, 2008)

Bury, K. (1999). *Statistical Distribution in Engineering*. New York: Cambridge University Press

De, M. (2000). A New Unbiased Plotting Position Formula For Gumbel Distribution. *Stochastic Environmental Research and Risk Assessment*. Vol. 14, pp.1-7

Jaramillo, O.A. and Borja, M.A.(2004). Wind speed analysis in La Ventosa, Mexico: a bimodal probability distribution case. *Renewable Energy*, Vol. 29, pp.1613–1630

Law, A.M. and Kelton, M.D. (2000). *Simulation Modelling and Analysis*, New York: McGraw- Hill International Series

Looney, S.W. and Gulledge JR., T.R. (1985). Probability plotting positions and goodness of fit for normal distribution, *The Statistician*, Vol. 34, pp. 297-303

Lund, J.R., Jenkins, M. and Kalman, O. (1998). *Integrated Planning and Management for Urban Water Supplies Considering Multiple Uncertainties*. Technical Report, Department of Civil and Environmental Engineering University of California, pp.15-16

Maffeis, G. (1998) .Prediction of carbon monoxide acute air pollution episodes. Model formulation and first application in Lombardy. *Atmospheric Environment*, Vol. 33 (23), pp. 3859 – 3872

Noor Faizah Fitri MD Yusof, Nor Azam Ramli, Ahmad Shukri Yahaya, Nurulilyana Sansuddin, Nurul Adyani Ghazali & Wesam Ahmed Al Madhoun (2010). Monsoonal Differences and Probability Distribution of PM_{10} Concentration. *Journal of Environmental Monitoring and Assessment*, Vol. 163 (1), pp.655-667

Rao, A.R. and Hamed, K.H. (2000), *Flood Frequency Analysis*, Boca Raton, Florida: CRC Press

Ross, R. (1994). Graphical Method for Plotting and Evaluating Weibull Distribution Data, *Proceedings of the 4th International Conference on Properties and Application of Dielectric Materials*, 3-8 July 1994, Brisbane Austrialia. pp. 250-253

Seinfeld, J. H. and Pandis, S. N. (1997). *Atmospheric Chemistry and Physics, from Air Pollution to Climate Change*, New York: Wiley

Shabri, A. (2002). A comparison of plotting formulas for the Pearson Type III Distribution. *Journal Teknologi*, Vol. 36(C), pp. 61–74

Shoji T. (2005). Statistical and geo-statistical analysis of wind: A case study of direction statistics, *Computers & Geosciences*, Vol. 32, pp. 1025-1039

Wang, F.K. and Keat, B.J. (1995). Improved percentile estimation for the two-parameter Weibull distribution**,** *Microelectron. Reliab.*, Vol. 35(6), pp. 883-892

Zhang, L.F., Xie, M. and Tang, L.C. (2006). A study of two estimation approaches for parameters of Weibull distribution based on WPP, *Proceedings of the Fourth International Conference on Quality and Reliability*, Quality and Innovation Research Centre, Department of Industrial and Systems Engineering, National University of Singapore, Vol. 92, pp. 360-368.