International Journal of
Science and Engineering Investigations vol. 6, issue 68, September 2017

ISSN: 2251-8843

A New Approach to Testing the Distribution Type of Life Data

Zheng Li
Ph.D., Department of Computer Science and Technology, North China Electric Power University (Baoding), China
(yeziperfect@163.com)

Abstract- In order to calculate reliability indices of some important equipment, the life data type of equipment or product needs to be ascertained firstly. Aiming to this question, the linear regression and correlation coefficient method is suggested. The linearization of common distribution like as exponential distribution is complied with the principle of linear regression. But for three-parameter Weibull distribution, the location parameter usually cannot be ascertained. A binary search algorithm is programmed to locate the location parameter firstly. Then through the comparison of correlation coefficient, the most fitted distribution type can be found out quickly. The result also shows that the inferred distribution type is in accord with the facts. Example illustrates the effectiveness and validness of linear regression and correlation coefficient for the inference of distribution type of life data.

Keywords- *Linear Regression, Correlation Coefficient, Life Data, Three-Parameter Weibull Distribution, Reliability*

I. INTRODUCTION

There are two categories of method to predict the life of equipment and/or products. One is statistical hypothesis tests and the other is probability paper method. No matter which method is to be adopted, the preliminary analysis for the equipment's life data is first necessary. The most frequently used tool for data analysis is histogram which can be helpful to find out the distribution type of life data. According to the shape of frequency histogram the most possible life data distribution type could be figured out.

For every possible distribution type, the principle of hypothesis test should be used to give a decision of acceptation or rejection on the null hypothesis H_0 . The Kolmogorov-Smirnov (K-S) test is a simple but effective technology that has already been studied by a number of researchers using a wide range of approaches [1-3]. Chi-square χ^2 test was proposed by Pearson in 1900 and had already been employed for various situations [4, 5]. There can be several distributions meeting the hypothesis during the test, the most suitable one would be choose based on the personal experience. The probability paper method plots the data point on the special paper where the ordinate axis has a non-linear scale corresponding to the cumulative probability of failure. When these life data points are to be plotted on certain type probability paper, it is necessary to arrange these data in ascending order and then assign a cumulative probability of failure $F(i,n)$ to each point. If a best straight line may be fitted through such points, these data can be seen as belonging to the distribution type above mentioned. At the same time, the unknown parameters can be estimated from the intercept. The researches on how to use probability paper are summarized in [6, 7]. When there are several distribution types meeting the straight line by eye, how to distinguish which one is better fitted such data is a challenge.

There have been numerous procedures developed in the literature for determining whether a random sample comes from a specific distribution. All these procedures have been broadly classified as goodness-of-fit tests. Aiming to the deficiency that is unable to choose a better distribution, how to choose quantitatively the best distribution is a crucial issue. Linear regression technology has been employed in many applications [8, 9]. In this paper, the linear regression technology had been applied in the judgment of distribution type.

This paper is organized as follows. Section II presents the principle of linear regression and how to apply the correlation coefficient to the reliability engineering. In section III, how to linearize several typical distributions is shown. In section IV, a number of experiments are conducted to demonstrate the linear regression method for two types of distributions. Finally, the conclusions drawn from this study are given in Section V.

II. I. LINEAR REGRESSION AND CORRELATION COEFFICIENT

If random variable Y and X are of linear relation, their n observation values $(x_i, y_i), i=1,2,\ldots,n$ are independent, then let

$$y_i = a + bx_i + \varepsilon_i \quad \text{(i=1,2,...,n)} \tag{1}$$

where, a,b are unknown parameters, and called regression coefficient. The random variable ε_i subjected to s-Normal distribution $N(0,\sigma^2)$ and shows effect on y_i .All above mentioned is a linear regression model. Regression analysis is to calculate the estimation value $\hat{a},\hat{b}$ of regression coefficient a,b according to the experimental data. Given x ,

$$\hat{y} = \hat{a} + \hat{b}x \quad , \tag{2}$$

could be seen as estimation of y=a+bx. Equation (2) is called as linear regression equation $\mu(x)$, and whose pictures are named as regression lines. Once the estimation value $\hat{a},\hat{b}$

could be found out, the explicit form of (2) could be used to calculate the value of variable $\hat{y}$ given x.

The least square method (LSM) is often employed to get the estimation value of regression coefficient. Generally speaking, $y_i \neq \hat{y}_i$ because the random variable ε_i . The approach degrees of theoretical value y and factual value $\hat{y}$ can be represented by

$$Q(\hat{a},\hat{b}) = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 \tag{3}$$

So, the less value of $Q(\hat{a},\hat{b})$, the higher of fit degree. Consequently, according to the principle of LSM, the estimation value of regression coefficient can be gotten when (3) reaches the minimum. It can be described by

$$Q(\hat{a},\hat{b}) = \min Q(a,b) \tag{4}$$

According to the principle of derivatives, extreme values can be reached by solve the following equations

$$\begin{cases} \dfrac{\partial Q}{\partial a} = -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i) = 0 \\ \dfrac{\partial Q}{\partial b} = -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)x_i = 0 \end{cases} \tag{5}$$

After simplification of the (5),

$$\begin{cases} \hat{b} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \hat{a} = \overline{y} - \hat{b}\overline{x} \end{cases} \tag{6}$$

with $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. Introducing the following notations

$$\begin{cases} S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2 \\ S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2 \\ S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) \end{cases} , \tag{7}$$

(6) can be changed into the following forms

$$\begin{cases} \hat{b} = \dfrac{S_{xy}}{S_{xx}} \\ \hat{a} = \overline{y} - \hat{b}\overline{x} \end{cases} \tag{8}$$

As a consequence of the above, the estimation values of regression coefficient can be obtained by the method of LSM.

Computing correlation coefficient is the most commonly used method to measure the linear degrees between variables that are linearly related. Correlation coefficient is defined by

$$\gamma = \frac{S_{xy}}{\sqrt{S_{xx} \times S_{yy}}} \tag{9}$$

γ is a dimensionless statistics and its absolute value is less than or equal 1. When $|\gamma| = 1$, it shows that all the experimental data are all located on the straight line. Then the variables x, y are of linear relation at the probability value 1. If $|\gamma| \neq 1$, then the bigger of the $|\gamma|$, the better of the linear degree between the variables; the less of the $|\gamma|$, the worse of the linear degree between the variables.

III. LINEARIZATON OF TYPICAL DISTRIBUTION

In reliability engineering, a lot of equipment's reliability level and life data is not of linear relation. On the contrary, there are some following relations, e.g. exponential distribution and Weibull distribution, between those two variables. As a result, the correlation coefficient cannot be employed directly to judge the life data law. By variable transformation, the relation between reliability level and life data may be changed into linear. Then linear regression theory can be used to judge which distribution is better than other. At the same time, the unknown parameters of one distribution can be estimated by this method.

The detailed procedure is as follows. The life data can be seen as random variables T , and be ordered by ascendant sequence, e.g. $t_1 \le t_2 \le \cdots t_i \le \cdots \le t_n$. Each life data t_i has a corresponding failure level $F(t_i)$ which can be described by the following approximate median ranks formula [10]

$$F(t_i) = P(T \le t_i) = \frac{\text{i-0.3}}{\text{n+0.4}}, (\text{n} \le 20) \tag{10}$$

Then using (10), the functional value of each life data t_i can be obtained and the data point $(t_i, F(t_i))$ will be transformed properly into $(\sigma(t_i), \xi(F(t_i)))$ so that the latter is of linear relation. Thereby, the correlation coefficient can be calculated and compared quantitatively in order to choose a better distribution type which reflects the equipment's life data laws.

As mentioned above, some distribution types' data point $(t_i, F(t_i))$ is not of linear relation and certain kinds of transformation need to do. Aiming to different distribution type, the linearization's procedure is also different.

A. Exponential Distribution

The cumulative distribution function (CDF) of exponential distribution is as follow:

$$F(t)=1-e^{-\frac{t-\varphi}{\eta}},(t\geq\varphi) \tag{11}$$

And the parameter φ is location parameter. Obviously, the data point $(t_i,F(t_i))$ is not of linear relationship. After taking logarithm of (11), the following linear equation is obtained

$$\ln\frac{1}{1-F(t)}=\frac{1}{\eta}t-\frac{\varphi}{\eta} \tag{12}$$

The converted variables can be constructed as follow

$$\left.\begin{aligned} y_i&=\ln\frac{1}{1-F(t_i)}\\ x_i&=t_i \end{aligned}\right\} \tag{13}$$

Then the variables (x_i,y_i) are of linear relationship.

B. Weibull Distribution with three-parameter

Weibull distribution is widely used to model the variability in the fracture properties of ceramics and metals, where the concept of weakest link applies. For three-parameter Weibull distribution, its CDF is expressed by [11]

$$F(t)=1-e^{-\left(\frac{t-\varepsilon}{\eta}\right)^m} \tag{14}$$

The difference between two-parameter and three-parameter lies in location parameter $\varepsilon\neq 0$. In the similar way, after taking logarithm twice a linear equation will yield

$$\ln\ln\frac{1}{1-F(t)}=m\ln(t-\varepsilon)-m\ln\eta \tag{15}$$

Let

$$\left.\begin{aligned} y_i&=\ln\ln\frac{1}{1-F(t_i)}\\ x_i&=\ln(t_i-\varepsilon) \end{aligned}\right\} \tag{16}$$

The data points (x_i,y_i) are of linear relation. But the location parameter ε is not an exact value, it means the value of x_i cannot be got directly.

It is well known that many researchers have long been pursuing algorithm to estimate the parameter. Commonly used methods are the maximum likelihood estimation and the least squares estimation [11, 12]. However, these methods are tedious and fallible. Here a new idea is proposed to estimate the shape parameter. x_i can be seen as a function about variable ε, then the correlation coefficient is also a function about ε and it can be depicted by

$$\gamma(\varepsilon)=\frac{S_{xy}(\varepsilon)}{\sqrt{S_{xx}(\varepsilon)\times S_{yy}}} \tag{17}$$

The function $\gamma(\varepsilon)$ will be as an objective function which is used to calculate its maximum values. When $\gamma(\varepsilon)$ reaches its extreme value, the value of ε is our demanded naturally.

A numerical analysis method, binary search program, is employed by Matlab. The detailed procedures are as follows:

Step1: In order to simplify the calculation, solving the derivative should be first to do. It means that the following equation is solved firstly.

$$d(\gamma^2(\varepsilon))/d\varepsilon=0 \tag{18}$$

After algebraic simplification, the following equation can be gotten.

$$E(\varepsilon)=\frac{1}{S_{xy}}\frac{dS_{xy}}{d\varepsilon}-\frac{1}{2S_{xx}}\frac{dS_{xx}}{d\varepsilon}=0 \tag{19}$$

Step2: If $E(0)\leq 0$, then $\hat{\varepsilon}=0$, turn to the Step5; If $E(0)>0$, a threshold value such as $\delta=1*e-10$ need to be set and turned to the next step.

Step3: Set up an initial interval (t_0,t_1), where $t_0=0$. The middle value $t_{\mathrm{mi}d}=(t_1-t_0)/2$ of the interval will be calculated. If $E(t_{mid})>\delta$, then go to the next step; else $\hat{\varepsilon}=t_{mid}$.

Step4: If $E(t_{mid})>0$, then set $t_0=t_{mid}$. If $E(t_{mid})<0$, then set $t_1=t_{mid}$. Return to Step3.

Step5: The end.

Note that the threshold value δ may be adjusted properly according to the actual life data. After the above steps, an estimation $\hat{\varepsilon}$ of shape parameter can be obtained. When the shape parameter is set, the data points (x_i,y_i) are of linear relation from (16) and the principle of linear regression can be applied.

IV. Application Procedure

A. Life Data Preliminary Analysis

Histogram is often used to analyze preliminary life data. The plotting of histogram may be proceeded as follows:

1) To find the maximum $X_{\max}$ and the minimum $X_{\min}$ among one group of data, $X_1,\ldots,X_n$ collected from laboratory or on-site;

2) To group those data, the numbers of group κ can be determined by

$$\kappa=1+3.3\ \log n \tag{20}$$

3) Calculate the interval $\vartriangle t$ as follows:

$$\vartriangle t=(X_{\max}-X_{\min})/\kappa \tag{21}$$

In order to plot conveniently, the interval $\triangle t$ should be adapted properly.

4) To determine the upper limit value and lower limit value of every group.

5) To get the medium value of every group.

6) To count the numbers and frequency of every group.

7) By taking life data as horizontal coordinate and the cumulative frequency as vertical coordinate, plot the frequency histogram.

There are 15 life data X (hour) about certain kinds of electric power equipment $t_i = 1280$, 1430, 1689, 1901, 2046, 2302, 2600, 2673, 2945, 3238, 3521, 3928, 4204, 4625, 5787, $i = 1, 2, \cdots, 15$ Their failure level can be computed by the (10).

Applying the histogram method to the above data, $Min = t_1$, $Max = t_{15}$ the number of group $k = 1 + 3.3 * \log_{10} n = 4.88 \approx 5$, the interval of group $\triangle t = (5787 - 1280)/5 \approx 902$. In order to plot the histogram conveniently, some statistics are listed in the Table 1.

TABLE I. COMPUTE TABLE OF FREQUENCY NUMBER FOR GROUP

Orders(i)	Interval	Middle Value	Count	Frequency
1	1279~2181	1730	5	0.3333
2	2181~3083	2632	4	0.2667
3	3083~3985	3534	3	0.2000
4	3985~4887	4436	2	0.1333
5	4887~5789	5338	1	0.0667
Sum			15	1

In Matlab the command hist (ti,828:902:5338) is used to plot the frequency histogram shown as below.

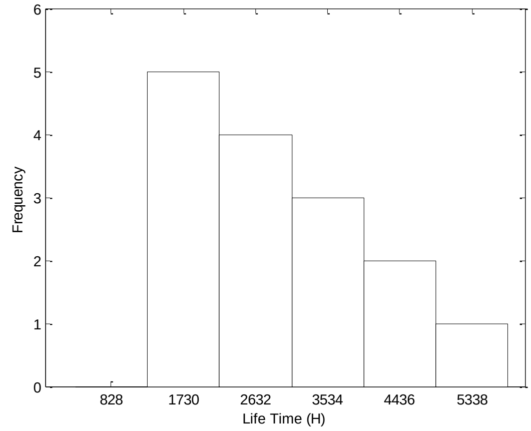

Figure 1. Frequency histogram of life time

Figure 1 can be seen as a probability density function (PDF) of the referred distribution. Through the shape of PDF, it can be concluded that the life data are approximately subjected to exponential distribution and Weibull distribution because there are some similarity shapes to respective PDF figure.

B. Calculation of correlation coefficient

For each possible distribution of the life data, what need to do is to implement the linearization and calculate the correlation coefficient.

If the life data is subjected to exponential distribution, data points $(x_i, y_i), i = 1, \cdots, 15$ will be gotten from (13). According to the principle of linear regression, the regression coefficient estimation value $\hat{a}, \hat{b}$ can be gotten. Then the linear equation is $y = 0.0007x - 1.0067$. Using this equation, the parameter value can be gotten $\eta = 1428, \varphi = 1438$. Correlation coefficient is $\gamma = 0.9863$.

For Weibull distribution, in virtue of the binary search program the value of location parameter can be obtained $\varepsilon = 1023.6$. Because it is not equal zero, these life data is subject to three-parameter Weibull distribution. From (16) the data points (x_i, y_i) are of linear relation. The linear regression equation is $y = 1.3916x - 10.693$. The remaining unknown parameters are also clear as follow $m = 1.3916, \eta = 2173$. The correlation coefficient is $\gamma = 0.9980$.

It can be found out that equipment's life data is subjected to the three-parameter Weibull distribution because its correlation coefficient is the higher.

V. CONCLUSION

This paper verified that the linear regression method can be employed on the life data distribution's inference. Histogram is also a useful tool for determining the elementary distribution type of life data. For each common distribution, the linear regression method cannot be implemented directly. Linearization is first to do and the detailed procedures is also respective different. During the proceedings of linearization, the unknown parameters of each distribution are also deduced by the way and it is helpful for the next work.

The correlation coefficient had been examined for the comparison of several probable distribution types. For the Weibull distribution, it is needed to estimate the location parameter by the binary search program firstly. The result of correlation coefficient is a quantitative numerical value which is straightforward and convenient to judge which distribution is better fitted for the given life data.

ACKNOWLEDGMENT

The authors wish to thank the reviewers and the Editor for their constructive comments that have helped to improve the article. This work was supported in part by a grant from Hebei

Province Natural Science Foundation (F2014502081) and Fundamental Research Funds for the Central Universities (No. 2015MS128).

REFERENCES

[1] Reschenhofer, E., Generalization of the Kolmogorov-Smirnov test. Computational Statistics and Data Analysis, 24(4):1997, 433-441.

[2] Andrade, F.A., I. Esat, and M.N.M. Badi, A new approach to time-domain vibration condition monitoring: Gear tooth fatigue crack detection and identification by the Kolmogorov-Smirnov test. Journal of Sound and Vibration, 240(5):2001, 909-919.

[3] Drezner, Z., O. Turel, and D. Zerom, A modified kolmogorov-smirnov test for normality. Communications in Statistics: Simulation and Computation, 39(4):2010, 693-704.

[4] Hauschild, T. and M. Jentschel, Comparison of maximum likelihood estimation and chi-square statistics applied to counting experiments. Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 457(1-2):2001, 384-401.

[5] Chen, Y.-T. and M.C. Chen, Using chi-square statistics to measure similarities for text categorization. Expert Systems with Applications, 38(4):2011, 3085-3090.

[6] Fothergill, J.C., Estimating the cumulative probability of failure data points to be plotted on Weibull and other probability paper. IEEE transactions on electrical insulation, 25(3):1990, 489-492.

[7] Takahashi, T., Statistical inference by normal probability paper. Computers and Industrial Engineering, 37(1):1999, 121-124.

[8] Jen-Tzung, C., Quasi-Bayes linear regression for sequential learning of hidden Markov models. IEEE Transactions on Speech and Audio Processing, 10(5):2002, 268-278.

[9] Jen-Tzung, C., Linear regression based Bayesian predictive classification for speech recognition. Speech and Audio Processing, IEEE Transactions on, 11(1):2003, 70-79.

[10] Cacciari, M., G. Mazzanti, and G.C. Montanari, Comparison of maximum likelihood unbiasing methods for the estimation of the Weibull parameters. IEEE Transactions on Dielectrics and Electrical Insulation, 3(1):1996, 18-27.

[11] Hirose, H., Maximum likelihood estimation in the 3-parameter Weibull distribution. A look through the generalized extreme-value distribution. Dielectrics and Electrical Insulation, IEEE Transactions on, 3(1):1996, 43-55.

[12] Zhang, L.F., M. Xie, and L.C. Tang, Bias correction for the least squares estimator of Weibull shape parameter with complete and censored data. Reliability Engineering and System Safety, 91(8):2006, 930-939.