

/AN EVALUATION OF VARIOUS PLOTTING POSITIONS/ 205

by

MALGORZATA JOANNA RYS

M.S., Technical University of Wroclaw, Wroclaw
Poland, 1979

A MASTER'S REPORT

submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1986

Approved by:


Major Professor

LD
2668
.24
1986
K97
c.2

A11202 664089

TABLE OF CONTENTS

	Page
ACKNOWLEDGEMENTS.	iii
LIST OF FIGURES	iv
LIST OF TABLES.	v
1.0 INTRODUCTION	1
1.2 Definitions	2
1.2.1 Reliability measures	2
1.2.2 Hazard function models	5
1.2.3 Mean Square Error.	9
2.0 DEVELOPMENT OF THE RANK.	11
2.1 Order statistics.	11
2.2 The multinominal distribution	12
2.3 The probability density of the rth order statistic. . .	13
3.0 DEVELOPMENT OF THE RANK DISTRIBUTION	16
4.0 DEVELOPMENT OF THE PLOTTING POSITIONS.	19
4.1 The mean plotting position.	19
4.2 The median and median approximation plotting position .	19
4.3 The midpoint plotting position.	21
4.4 An approximate mean value plotting position	21
5.0 PLOTTING POSITION ON PROBABILITY PAPER	26
6.0 THEORETICAL DISCUSSION ABOUT PLOTTING POSITIONS.	28
7.0 PROCEDURE.	32
8.0 RESULTS.	37
9.0 CONCLUSIONS.	40

TABLE OF CONTENTS (continued)

	Page
APPENDICES	
A. The Beta Distribution.	42
B. The Weibull Distribution	44
C. The Exponential Distribution	47
D. Rank Tables.	49
E. Examples of Nelson's procedure [5] for making a probability plot	51
F. Regression analysis program for exponential distribution . .	54
G. Regression analysis program the Weibull distribution	57
H. Program for calculating the MSE for exponential distribution	61
I. Program for calculating the MSE for the Weibull distribution	62

ACKNOWLEDGEMENTS

I would like to express my thanks and gratitude to Professor Doris Grosh for her help, guidance and constant support. I would also like to thank Dr. L. E. Grosh and Dr. J. J. Higgins for serving on my M.S. committee. Finally, I owe my greatest debt to my husband Andrzej, who has been a constant source of encouragement, help, and understanding.

LIST OF FIGURES

	Page
Figure 1. The Bathtub Curve.	6
Figure 2. Hazard functions for the Weibull distribution. . .	10
Figure 3. The rank distribution for a sample of size 10. . .	18
Figure 4. The plots of a Weibull sample using different plotting positions	31
Figure 5. The Weibull distribution for different values of β	46
Figure 6. Exponential probability plot of the 34-kV data . .	53

LIST OF TABLES

	Page
Table 1. Partitioning of population	16
Table 2. Normal distribution; root α_{in} of equation (4.9).	23
Table 3. Normal distribution; calculation of Ex_1	24
Table 4. Kimball's comparison of plotting positions for estimating the shape parameter β	30
Table 5. Empirical values for $F(t_i)$	32
Table 6. Results for exponential distribution.	36
Table 7. Results for the Weibull distribution.	37
Table 8. Times to breakdown of a insulating fluid.	49

1.0 INTRODUCTION

It has been stated that the hazard function could be used for distinguishing between various failure models such as the gamma, Weibull, exponential, normal, etc.. Comparing the strikingly different curve shapes of various hazard functions such a belief appears reasonable.

The results presented by D. Grosh [1] were in disagreement with the above belief, due to the fact that few of the plots looked anything like their theoretical curves. The procedure used by Grosh can be found in [1]. In the next approach Grosh used the old fashioned probability plotting technique. This method uses appropriate scaling to convert the cumulative distribution function (CDF) into a straight line. Although perfect selecting was not experienced, the discrimination was certainly much better than with the hazard function plots.

In order to use the latter method in its best form, different probability plotting positions should be considered; the question arises - what should be used as a plotting position? The problem is particularly pertinent where a probability paper is used, such that a straight line plot of the data indicates conformance to some type of distribution.

Bradford F. Kimball [6] has done limited study comparing some of the possible plotting positions. The types of distributions Kimball considered were the normal and the extreme-value. The results were based on complete samples of size 6. This work is an extension of the problems studied by Kimball, however a different approach will be used.

The following different plotting positions will be discussed and analyzed in this report:

(Let r be the r th ordered observation, n be the total number of observations)

- (1) "mean" plotting position $r/(n+1)$
- (2) "median" (Appendix D)
- (3) "median approximation" $(r-0.3)/(n+0.4)$
- (4) "midpoint" $(r-0.5)/n$
- (5) "mean approximation" $(r-0.375)/(n+0.25)$

Separate programs will be written for each of the studied distributions (exponential, Weibull) using an IMSL random variable generator. Samples of some chosen size will be created and sorted, and the IMSL Regression Analysis will be used to evaluate the estimators based on the different plotting position. Mean Square Error (MSE) will be used as a criterion for determining the best plotting position for each distribution.

1.2 Definitions

1.2.1 Reliability measures

Let T be non-negative continuous random variable which represents the useful life (or time to failure) of a component. The failure law for the component can be described in any of several ways. The most fundamental formulation is in terms of $F(t)$, the cumulative distribution function (or CDF), defined as the probability that the system will fail by time t , and which can be stated

$$F(t) = P(T \leq t) \quad t \geq 0$$

In other words, $F(t)$ is the failure distribution function also called the unreliability function. If we define reliability as the probability of success, or the probability that the system (or component) lives for at least time t , we can write

$$R(t) = P(T > t) = 1 - F(t)$$

The failure laws can be also described in terms of the density function

$$f(t) = F'(t)$$

which must have the properties that

$$f(t) > 0$$

$$\int_0^{\infty} f(t)dt = 1$$

The function $f(t)$, although it describes how the failure probability is spread over time, is not itself a probability, but a probability density. It can be used to compute the approximate probability over a short interval when the exact calculation is inconvenient or analytically intractable.

The probability of failure of a system in a given time interval $[t_1, t_2]$ can be expressed in terms of the reliability function as

$$\int_{t_1}^{t_2} f(t)dt = \int_{t_1}^{\infty} f(t)dt - \int_{t_2}^{\infty} f(t)dt = R(t_1) - R(t_2).$$

The rate at which failures occur in a certain time interval $[t_1, t_2]$ is called the failure rate during that interval. It is defined as the probability that a failure per unit time occurs in the interval, given

that a failure has not occurred prior to t_1 , the beginning of the interval.

Thus the failure rate is

$$\frac{R(t_1) - R(t_2)}{(t_2 - t_1) R(t_1)}$$

If we redefine the interval as $[t, t+\Delta t]$, the above expression becomes

$$\frac{R(t) - R(t+\Delta t)}{\Delta t R(t)}$$

The hazard function is defined as the limit of the failure rate as the interval limit approaches zero. Thus the hazard function is the instantaneous failure rate. The hazard function is defined by:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{R(t) - R(t+\Delta t)}{\Delta t \cdot R(t)} = \frac{1}{R(t)} \left[- \frac{d}{dt} R(t) \right] = \frac{f(t)}{R(t)}$$

The quantity $h(t)dt$ represents the probability that a component (or system) of age t will fail in the small interval of time t to $t + dt$. The importance of the hazard function is that it indicates the change in the failure rate over the life of a population of systems (components, devices). For example, two designs may provide the same reliability at a specific point in time; however, the failure rates up to this point in time can differ.

The hazard function $h(t)$ will change over the lifetime of a population of products somewhat as shown in Figure 1. The first interval of time, t_0 to t_1 , represents early failures due to material or manufacturing defects. Quality control and initial product testing usually eliminate many substandard components, and thus avoid this higher initial failure rate. This phase of the curve is called "infant

mortality", where the death rate is high for newborns. The second phase of the curve from t_1 to t_2 represents chance failures caused by sudden stress, extreme conditions. In actuarial terms this could be equated to the accidents encountered by a population of individuals on a day-to-day basis. In mechanical devices such as an automobile these failures can be eliminated only by overdesigning for the extreme conditions. The portion of the curve beyond t_2 represents wearout failures. Here the hazard rate increases as equipment deteriorates. If time t_2 could be predicted with certainty, then equipment could be replaced before this wearout phase begins.

The Bathtub Curve presented above is probably an idealization, since real components almost never display the type of behavior depicted. Human beings probably fit the model best. In some instances, it may be more convenient to select a hazard function model than a distribution model.

1.2.2 Hazard function models

In drawing conclusions from test data about the behavior of a lot of similar components it is necessary to fit the failure data with a mathematical model. Sometimes the model is to be used in studying the reliability of the particular component in question. Then the model should be accurate enough to reveal the detailed reliability behavior of

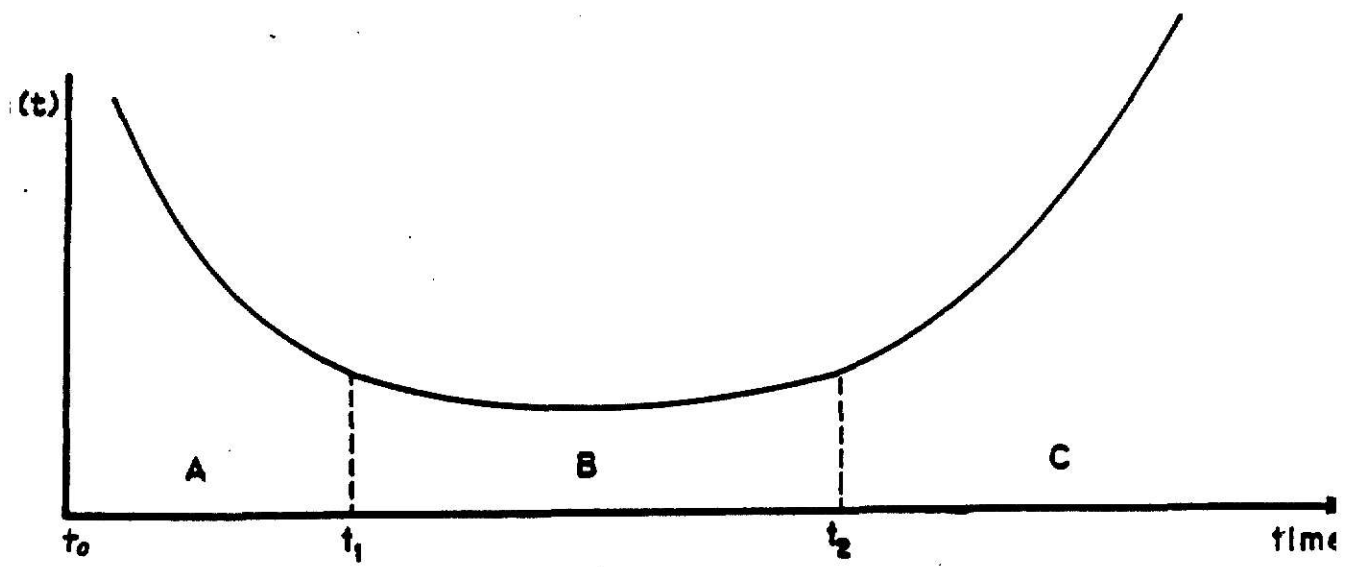


Figure 1. The Bathtub Curve [1].

the part and should suggest various underlying causes and modes of failure. In many cases, the study of component reliability is only an intermediate step in the discussion of system reliability. In such cases the chosen model must give the overall result that is easy to interpret and use.

Constant Hazard

Let us first consider a model in which the hazard function is constant. The constant hazard rate represents the portion of the hazard function over product life between times t_1 and t_2 . During this time chance or random failures occur, and failure cannot be attributed to deterioration.

Let the hazard function be given by

$$h(t) = \lambda \text{ failures/unit time}$$

where λ is a constant. The formula

$$f(t) = h(t) \exp\left[-\int_0^t h(\tau) d\tau\right]$$

in this case gives

$$f(t) = \lambda e^{-\lambda t}.$$

The reliability function is given by

$$R(t) = f(t)/h(t) = e^{-\lambda t}$$

Thus a constant hazard rate leads to the exponential density function of the time to failure random variable. A constant hazard model implies

that the failure probability in any time interval remains constant throughout the component lifetime.

The Weibull Model

Let $h(t)$ have the general power function or polynomial form

$$h(t) = c_1 t^{c_2}$$

where c_1 is positive and c_2 may be positive, negative, or zero. The resulting curve shapes for $h(t)$ are increasing failure rate, decreasing failure rate, or constant failure rate. The quantity needed for the exponent of $R(t)$ is

$$H(t) = \int_0^t h(\tau) d\tau = c_1 \int_0^t \tau^{c_2} d\tau = \frac{c_1 t^{c_2+1}}{c_2+1}$$

Let us make the substitutions

$$\frac{c_1}{c_2+1} = \alpha$$

$$c_2 + 1 = \beta \text{ (the "Weibull slope").}$$

Substituting into the above equation yields

$$h(t) = \alpha \beta t^{\beta-1}$$

which leads to

$$R(t) = e^{-\alpha t^\beta}$$

and

$$f(t) = \alpha \beta t^{\beta-1} e^{-\alpha t^\beta}$$

When $\beta = 1$ the negative exponential distribution results. The hazard curves which result from various values of β are shown in Figure 2.

1.2.3 Mean Square Error.

The mean square error (MSE) for a statistic $\hat{\theta}$ which is to be used to estimate the parameter θ is defined as the average-square deviation of $\hat{\theta}$ from the true value θ . Mathematically it is defined as $E[(\hat{\theta} - \theta)^2]$. The mean square error is a variance-like quantity and for unbiased estimators where $E(\hat{\theta}) = \theta$ it is identical with $\text{Var}(\hat{\theta})$. In the more general case we have

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E[\{\hat{\theta} - E(\hat{\theta})\} - \{\theta - E(\hat{\theta})\}]^2 \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + [\theta - E(\hat{\theta})]^2 \\ &= \text{Var}(\hat{\theta}) + [\theta - E(\hat{\theta})]^2 \end{aligned}$$

Thus the mean-squared error is the sum of two nonnegative quantities.

The term $\theta - E(\hat{\theta})$ is called the bias of the estimator and can be either positive, negative, or zero. If an estimator can be found with bias close to zero and such that $\text{Var}(\hat{\theta})$ is small, the mean square error will be small.

The magnitude of the MSE is useful as a criterion for selecting the best among several competing estimators.

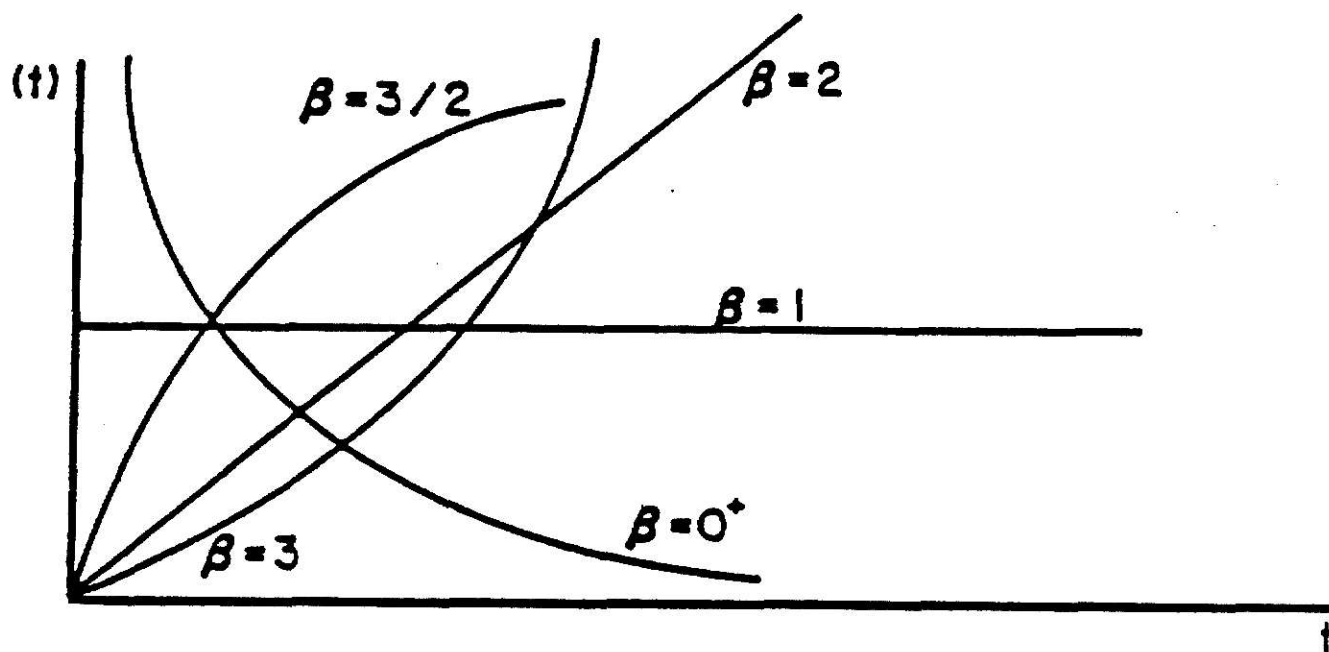


Figure 2. Hazard functions for the Weibull distribution [1].

2.0 DEVELOPMENT OF THE RANK DISTRIBUTION.

2.1. Order Statistics

Given a random sample of size n from an infinite population with a continuous density function, assume that we arrange the values $x_1, x_2, \dots, x_n$ according to size and look upon the smallest of the x 's as a value of random variable y_1 , the next largest as a value of random variable y_2 , ..., and the largest as a value of random variable y_n . The random variables thus defined are referred to as order statistics; in particular, y_1 is the first order statistic, y_2 is the second order statistic and so on. (This discussion is limited to infinite populations with continuous densities).

To be more specific, let us consider the case where $n=2$ and the relationship between the values of x_1 and x_2 and y_1 and y_2 is:

$$y_1 = x_1 \text{ and } y_2 = x_2 \text{ when } x_1 < x_2$$

$$y_1 = x_2 \text{ and } y_2 = x_1 \text{ when } x_2 < x_1$$

Similarly, for $n=3$ the relationship between the values of the respective random variables is:

$$y_1 = x_1, y_2 = x_2 \text{ and } y_3 = x_3 \text{ when } x_1 < x_2 < x_3$$

$$y_1 = x_1, y_2 = x_3 \text{ and } y_3 = x_2 \text{ when } x_1 < x_3 < x_2$$

$$y_1 = x_3, y_2 = x_2 \text{ and } y_3 = x_1 \text{ when } x_3 < x_2 < x_1$$

Before we derive the probability of the r th order statistics let us review the Multinomial Distribution.

2.2. The Multinomial Distribution.

It arises as a generalization of the binomial distribution when each trial has more than two possible outcomes. The probabilities of the respective outcomes are the same for each trial and the trials are all independent. This would be the case, for example, when persons interviewed by an opinion poll are asked whether they are for a candidate, against, or undecided, or when samples of manufactured products are rated excellent, average or bad.

Let us consider the general case where there are n independent trials with k mutually exclusive outcomes whose respective probabilities

are $\theta_1, \theta_2, \dots, \theta_k$ (with $\sum_{i=1}^k \theta_i = 1$). Referring to the outcomes as

being of the first kind, the second kind, ..., and the k th kind, we are interested in the probability of getting x_1 outcomes of the first kind, x_2 outcomes of the second kind, ..., and x_k outcomes of the k th kind.

Proceeding as in the derivation of the formula for the binomial distribution [2, p. 166], we first find that the probability of getting x_1 outcomes of the first kind, x_2 outcomes of the second kind, ..., and

x_k outcomes of the k -th kind in a specific order is $\theta_1^{x_1}, \theta_2^{x_2}, \dots, \theta_k^{x_k}$.

To get the corresponding probability for that many outcomes of each kind in any order, we have to multiply the probability for any specific order by [2, p. 12]

$$\left[\begin{matrix} n \\ x_1, x_2, \dots, x_k \end{matrix} \right] = \frac{n!}{x_1! \cdot x_2! \cdot \dots \cdot x_k!}$$

Now we are ready for definition [2]:

The random variables $x_1, x_2, \dots, x_k$ have the multinomial distribution if and only if their joint probability distribution is given by

$$f(x_1, x_2, \dots, x_k; n, \theta_1, \theta_2, \dots, \theta_2, \dots, \theta_k) \\ = \left[\begin{matrix} n \\ x_1, x_2, \dots, x_t \end{matrix} \right] \cdot \theta_1^{x_1} \cdot \theta_2^{x_2} \cdot \dots \cdot \theta_k^{x_k}$$

for $x_i = 0, 1, \dots, n$ for each i , where $\sum_{i=1}^k x_i = n$ and $\sum_{i=1}^k \theta_i = 1$

The numbers of outcomes of the different kind are random variables having the multinomial distribution with the parameters $n, \theta_1, \theta_2, \dots, \theta_k$.

Note: The name "multinomial" derives from the fact that for various values of the x_i , the probabilities equal corresponding terms of the multinomial expansion of $(\theta_1 + \theta_2 + \dots + \theta_k)^n$.

2.3. The probability density of the r th order statistic.

THEOREM: For random samples of size n from an infinite population, whose density has the value $f(x)$ at x , the probability density of the r th order statistic y_r is given by:

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!} \left[\int_{-\infty}^{y_r} f(x) dx \right]^{r-1} f(y_r) \left[\int_{y_r}^{\infty} f(x) dx \right]^{n-r}$$

Proof:

Suppose that the real axis is divided into three intervals, one from $-\infty$ to y_r , a second from y_r to $y_r + h$ (where h is a positive constant), and the third from $y_r + h$ to ∞ . Then, if the density of the population we are sampling has the value $f(x)$ at x , the probability that $r-1$ of the sample values fall into the first interval, one falls into the second interval, and $n-r$ fall into the third interval is

$$\frac{n!}{(r-1)!(n-r)!} \left[\int_{-\infty}^{y_r} f(x)dx \right]^{r-1} \left[\int_{y_r}^{y_r+h} f(x)dx \right]^1 \left[\int_{y_r+h}^{\infty} f(x)dx \right]^{n-r}$$

Using the mean value theorem from calculus, we have

$$\int_{y_r}^{y_r+h} f(x)dx = f(\xi) \cdot h \text{ where } y_r < \xi < y_r+h$$

and if we let $h \rightarrow 0$, we finally get

$$g_r(y_r) = \frac{n!}{(r-1)!(n-r)!} \left[\int_{-\infty}^{y_r} f(x)dx \right]^{r-1} f(y_r) \left[\int_{y_r}^{\infty} f(x)dx \right]^{n-r}$$

for $-\infty < y_r < \infty$ for the probability density of the r -order statistic. In particular, the distribution of y_1 , the smallest value in a random sample of size n , is given by

$$g_1(y_1) = n \cdot f(y_1) \left[\int_{y_1}^{\infty} f(x)dx \right]^{n-1} \text{ for } -\infty < y_1 < \infty$$

while the distribution of y_n , the largest value in a random sample of size n , is given by

$$g_n(y_n) = n \cdot f(y_n) \left[\int_{-\infty}^{y_n} f(x) dx \right]^{n-1} \quad \text{for } -\infty < y_n < \infty$$

Also, in a random sample of size $n = 2m + 1$ the sample median $\tilde{x}$ is y_{m+1} ,

so that its sampling distribution is given by

$$h(\tilde{x}) = \frac{(2m+1)!}{m!m!} \left[\int_{-\infty}^{\tilde{x}} f(x) dx \right]^m f(\tilde{x}) \left[\int_{\tilde{x}}^{\infty} f(x) dx \right]^m$$

3.0 DEVELOPMENT OF THE RANK DISTRIBUTION

Let us designate the r th ordered observation as x_r and the fraction of the population below this observation as p_r , so that $p_r = F(x_r)$. Since the value of x_r varies from sample to sample, p_r is a random variable. Thus we must consider the distribution of p_r .

Given an ordered random sample $x_1, x_2, \dots, x_n$ of size n from a population having a cumulative distribution function $F(x)$, where x is continuous, we will determine estimators for $F(x_1), F(x_2), \dots, F(x_n)$. Let us define

${}_n p_r$ is the fraction of the population failing prior to the r th ordered observation in a sample of size n . Next we partition the population into three regions.

Table 1. Partitioning of population.

Region	1	2	3
Limits of region	0	$x_r - 1/2 dx_r$	$x_r + 1/2 dx_r$
Probability of obtaining an outcome in region	$F(x_r - 1/2 dx_r)$	$f(x_r) dx_r$	$1 - F(x_r + 1/2 dx_r)$

If we ignore terms of the order of $1/2 dx_r$ then we have

Number of outcomes	Probability
$(r-1)$	$F(x_r)$
1	$f(x_r) dx$
$(n-r)$	$1 - F(x_r)$

Thus the probability of $r-1$ outcomes falling in region 1, one falling in region 2, and $n-r$ falling in region 3, is given by:

$$\frac{n!}{(r-1)!(1)!(n-r)!} [F(x_r)]^{r-1} f(x_r) dx [1-F(x_r)]^{n-r} \quad (3.2)$$

This is precisely the distribution of x_r , the value of the r th ordered observation, which has been derived previously.

Now we know that

$$F(x_r) = {}_n p_r \quad (3.3)$$

which by differentiating yields

$$dF(x_r) = f(x_r) dx = d {}_n p_r \quad (3.4)$$

Hence the probability element becomes:

$$g({}_n p_r) d {}_n p_r = \frac{n!}{(r-1)!(1)!(n-r)!} {}_n p_r^{r-1} (1-{}_n p_r)^{n-r} d {}_n p_r \quad (3.5)$$

where $0 < {}_n p_r < 1$

This distribution is named the RANK DISTRIBUTION when used in the life testing context. Actually the PDF of the random variable ${}_n p_r$ is the Beta Distribution [Appendix A]. Figure 3 on page 18 shows the PDF for different values of r for a sample of size $n = 10$ [2].

In the derivation of rank distribution the only assumption made was that $F(x)$ was differentiable. Thus the rank distribution does not apply exclusively to the Weibull distribution [Appendix B], but would be equally valid for the normal or the exponential [Appendix C].

Since ${}_n p_r$ is a random variable we face the problem of selecting a single value representative of ${}_n p_r$ to use in plotting. In this effort we will consider various alternatives.

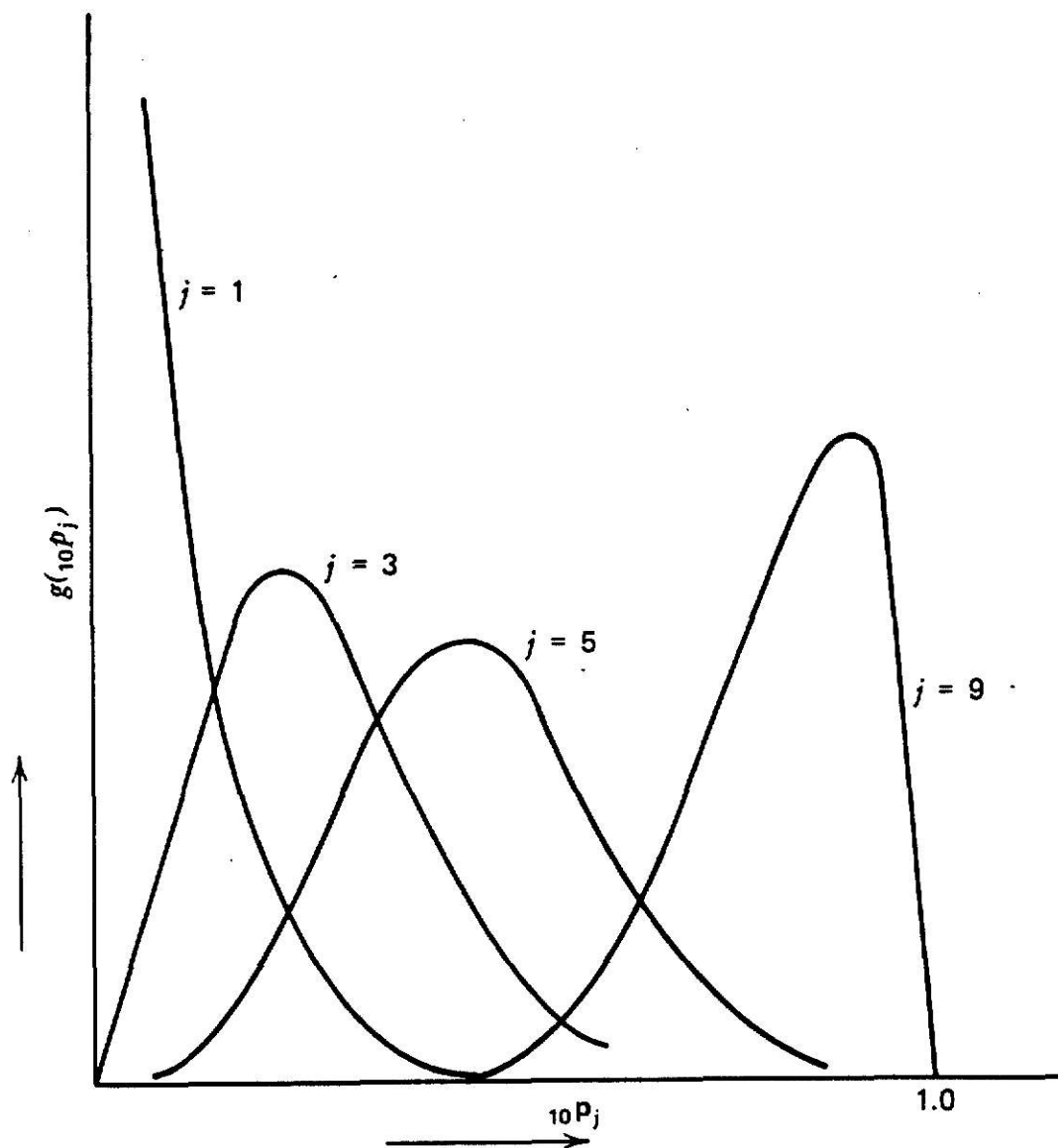


Figure 3. The rank distribution for a sample of size 10 [2].

4.0 DEVELOPMENT OF THE PLOTTING POSITIONS

4.1. The mean plotting position

The mean value of the random variable ${}_n p_r$ for which probability density function (PDF) is the beta distribution is given by

$$\begin{aligned} E({}_n p_r) &= \int_0^1 {}_n p_r \frac{n!}{(r-1)!(n-r)!} {}_n p_r^{r-1} (1-{}_n p_r)^{n-r} d{}_n p_r \\ &= \frac{n!}{(r-1)!(n-r)!} \int_0^1 {}_n p_r (1-{}_n p_r)^{n-r} d{}_n p_r \end{aligned} \quad (4.1)$$

A Beta function [Appendix A] in the standard form is given by

$$B(m,k) = \int_0^1 x^{m-1} (1-x)^{k-1} dx = \frac{\Gamma(m)\Gamma(k)}{\Gamma(m+k)} \quad (4.2)$$

which, when applied to equation (4.1) yields

$$E({}_n p_r) = \frac{n!}{(r-1)!(n-r)!} \frac{\Gamma(r+1)\Gamma(n-r+1)}{\Gamma(n-r+1+r+1)} = \frac{r}{n+1} \quad (4.3)$$

This means that the mean value of the fraction failing prior to the r th ranked observation in a sample of size n equals $r/(n+1)$.

4.2. The median and median approximation plotting position

The median value, commonly called the median rank, is more difficult to find. Consider the integral

$$P = \int_0^{\tilde{x}} g({}_n p_r) d{}_n p_r \quad (4.4)$$

where $g({}_n p_r)$ is defined in equation (3.5). The value of $\tilde{x}$ for which $P = 0.5$ is the desired median value.

The incomplete Beta function is given by:

$$B_x(m, n) = \int_0^x y^{m-1} (1-y)^{n-1} dy \quad (4.5)$$

Recalling the definition of the beta function as given in equation (4.2) the integral in equation (4.4) becomes

$$I_{\tilde{x}}^{\sim}(n, n-r+1) = \frac{B_{\tilde{x}}^{\sim}(r, n-r+1)}{B(r, n-r+1)} \quad (4.6)$$

The incomplete beta ratios $I_x(m, k)$ are extensively tabled (e.g. see Karl Pearson, "Tables of the Incomplete Beta Function", Cambridge, England, 1932). However, for convenience, the median rank values have been summarized in the tables and can be found in many statistics books. Appendix D from [2] is an example of such a table.

Let ${}_n p_r$ be monotone function of x_r (${}_n p_r = F(x_r)$). According to [10] an approximation to the median x can be found by substituting the value of ${}_n p_r = \frac{r-0.3}{n+0.4}$ in the inverse function $x_r = F^{-1}({}_n p_r)$. Thus an approximation to the median rank is given by [2]

$$\tilde{x}_r = F^{-1}\left(\frac{r-0.3}{n+0.4}\right) \quad (4.7)$$

where r is the r th ordered observation and n is the sample size.

4.3. The midpoint plotting position

The smallest of n observations may be thought of as representation of the first $(100/n)\%$ (fractile) of the population. The midpoint of this interval is the first plotting position $(100(1-0.5)/n)\%$. Similarly, the second smallest of n observations represents the second $(100/n)\%$ of the population, that is the portion of the population between $(100/n)\%$ and $(100(2/n))\%$. The midpoint of this interval is the second plotting position $(100(2-0.5)/n)\%$. This pattern continues through the n -th (largest) observation, which represents the last $(100/n)\%$ of the population, that is, that between $(100(n-1)/n)\%$ and $(100(n/n))\%$. The midpoint of this interval is the n -th plotting position $(100(n-0.5)/n)\%$.

In general, for the r -th observation the midpoint plotting position is

$$\frac{100(r-0.5)}{n} \% \quad \text{or} \quad \frac{r-0.5}{n} . \quad (4.8)$$

4.4. An approximate mean value plotting position

A numerical study of the equation

$$Ex_i = \Phi^{-1} \left(\frac{i-\alpha}{n-2\alpha+1} \right) \quad (4.9)$$

was made by Bloom [4]. For any fixed pair of values i and n , this equation in α has generally a single root $\alpha = \alpha_{in}$. When i varies from 1 to n , a sequence of roots $\alpha_{in}, \dots, \alpha_{nn}$ is obtained, where $\alpha_{in} = \alpha_{n-i+1,n}$. The smallest and the greatest of the roots form endpoints of an interval

$$\Phi^{-1} \left(\frac{i-0.39}{n+0.22} \right) < Ex_i < \Phi^{-1} \left(\frac{i-0.5}{n} \right) \quad (4.10)$$

The root $\alpha = \alpha_{in}$ of this equation has been determined for some values of i and n by means of normal order statistics (Godwin, 1949). The result is given in Table 2 [4]. It can be seen that α_{in} does not change very much with i and n when $n < 20$, the smallest and largest values being 0.33 and 0.39 respectively. When n goes to infinity the roots tend to some point within the interval I considered above. Thus we can say that α_{in} is always located in the interval (0.33, 0.50). This would be equivalent to saying that the inequality which is obtained by replacing $(i-0.39)/(n+0.22)$ by $(i-0.33)/(n+0.33)$ in equation (4.10) is approximately true for any i and n . The validity of this statement has however not been analytically verified.

Because of the small variation of α_{in} it may be expected that equation (4.9) will be fairly accurate when used with constant value of α . As a compromise between different possible values Bloom [4] suggested $\alpha = 3/8$, which leads to the formula

$$Ex_i \sim \Phi^{-1} \left(\frac{i-0.375}{n+0.25} \right) \quad (4.11)$$

Table 2. Normal distribution; root α_{in} of equation (4.9) [4].

i	Sample-size n						
	2	4	6	8	10	15	20
1	0.330	0.347	0.355	0.360	0.364	0.370	0.374
2		0.359	0.368	0.374	0.378	0.385	0.390
3			0.370	0.375	0.379	0.386	0.391
4				0.375	0.379	0.386	0.390
5					0.379	0.384	0.389
6						0.383	0.387
7						0.382	0.387
8							0.386
9							0.386
10							0.386

This expression is quite useful as an overall approximation to the mean. Table 3 contains accurate values taken from existing tables and values computed from the expressions:

$$Ex_i \sim \Phi^{-1} \left(\frac{i}{n+1} \right) \quad (4.12)$$

$$Ex_i \sim \Phi^{-1} \left(\frac{i-1/2}{n} \right) \quad (4.13)$$

Table 3. Normal distribution; calculation of Ex_i [4].

n	i	Exact value	0	α 3/8	1/2
5	1	1.163	0.967	1.180	1.282
	2	0.495	0.431	0.497	0.524
10	1	1.539	1.335	1.547	1.645
	2	1.001	0.908	1.000	1.036
	3	0.656	0.605	0.655	0.674
	4	0.376	0.349	0.375	0.385
	5	0.123	0.114	0.123	0.126
15	1	1.736	1.534	1.739	1.834
	2	1.248	1.150	1.241	1.282
	3	0.948	0.887	0.946	0.967
	4	0.715	0.674	0.714	0.728
	5	0.516	0.489	0.515	0.524
	6	0.335	0.319	0.335	0.341
	7	0.165	0.157	0.165	0.168

All entries should be given negative signs.

It can be seen from Table 3 that formula (4.11) is quite accurate for small sample sizes. Formula (4.12) is unsatisfactory when n is small;

for example when $n=10$, it underestimates the true value by 7 to 13 percent, depending on i . For $n = 10$, the expression (4.13) overestimates the mean by 3 to 7 percent.

5.0 PLOTTING POSITION ON PROBABILITY PAPER

Probability paper transforms a curvilinear CDF into a straight line. If the approach is to be purely analytical there is no point in using the special scale paper. It is important to have in mind the purpose served by plotting the observed data on the special graph paper. In general (according to Kimball [6]) there are three rather different kinds of purposes that might be served. These are:

1. A test as to whether or not the sample data indicate that the universe is of the prescribed type. One argues that the universe is of the prescribed type only if the plotted points tend to lie along a straight line.
2. The graphical method may be used as a short cut in estimating the standard deviation of the distribution, which in turn is directly determined from the slope of the fitted straight line.
3. Graphical extrapolation at one of the extremes. This is the purpose most commonly served in plotting data from an extreme-value universe.

These objectives can overlap. When accenting (2) one may well have a weather eye on objective (1). For example, in examining a batch of samples taken from different populations one may have inferred that the universes are normal and so accent objective (2). However, some errant populations may deviate considerably from the normal and so one may also give some weight to the graphical test of normality. Similarly in accenting objective (3), objective (2) is important. Furthermore, there are situations where the data over the lower range are known not to follow the prescribed distribution, in which case the data on the upper

range are given greater weight in fitting a straight line. Thus objective (1) is involved in indicating what part of the range conforms to the prescribed universe.

Nelson [5] suggested the following seven steps to make a probability plot.

1. Order the n failure times from smallest to largest.
2. Assign a rank to each failure (give the earliest failure rank 1, the second failure rank 2, etc.).
3. Calculate probability plotting positions. Chapter IV of this report describes commonly used plotting positions.
4. Plot the failure times on probability paper. There are probability papers for many distributions, including the exponential, normal, lognormal, Weibull, extreme value and chi-square distributions. The distribution may be chosen based on experience or understanding the physical nature of the problem.
5. Assess the data and assumed distribution. If the plotted points tend to follow a straight line, then the chosen distribution appears to be adequate. A curved plot indicates that the distribution does not fit the data; one should then try plotting the data on other paper. An outlier point extremely out of line with the rest is suspect and might be discarded, but its cause should be sought since that information may help improve the product or data collection.
6. Draw a straight line through the plotted data. Draw the line by eye so as to minimize the deviations between the line and the plotted points. Depending on how the line will be used, it

can be fitted to the whole sample, the center of data, the lower tail, or whatever is appropriate, ignoring or taking into account outliers.

7. Obtain the desired information.

In Appendix E one can see example of the above procedure taken from [5].

6.0 THEORETICAL DISCUSSION ABOUT PLOTTING POSITIONS

At this point we have five possible candidates for the value of the plotting position. They are:

- (1) The mean value $r/(n+1)$
- (2) The median rank (Appendix D)
- (3) An approximation to the median rank $(r-0.3)/(n+0.4)$
- (4) The midpoint $(r-0.5)/n$
- (5) An approximate mean value $(r-0.375)/(n+0.25)$

The mean value is used because the mean is commonly taken as the representative value of sample from a distribution. However, in highly skewed distributions, as are most rank distributions, the median may be better descriptor [2]. The rank distributions are systematically skewed from right to left as one proceeds from the lowest to the highest sample value. Thus the probability of a sample value falling below the mean value is highest for the first ordered sample value and lowest for the last ordered sample value. This implies that if mean rank plotting positions are used, the rank assigned to the first observation will probably be too high and the rank assigned to the last observation too low with successive changes in error for intermediate observations. When a straight line is fitted to these observations using mean ranks, the line will most likely be rotated clockwise, and the slope will be underestimated [2].

Figure 4 illustrates how the three plotting positions generate different lines and hence different estimators of shape for a Weibull sample.

Kimball [6] has done a limited study comparing some of the possible plotting positions. His results are summarized in Table 4. The results are based on complete sample of size 6 and consider only the estimation of the shape parameter β . Based on Kimball's results for estimating β , the plotting position $(r-0.375)/(n+0.25)$ appears to be the best. The popular median rank plotting position appears to be inferior to the mean approximation plotting position, but better than mean ranks.

Table 4. Kimball's comparison of plotting positions for estimating the shape parameter β [2].

Plotting Conventions	Bias/ β	Mean Square
		Error/ β^2
Mean = $i/(n+1)$	0.2284	0.3011
Median	0.0632	0.1928
$(i - \frac{3}{8})/(n + \frac{1}{4})$	0.0220	0.1758
$(i - \frac{1}{2})/n$	-0.0547	0.1543

THREE PLOTS OF A WEIBULL SAMPLE OF SIZE 20 USING DIFFERENT PLOTTING POSITIONS

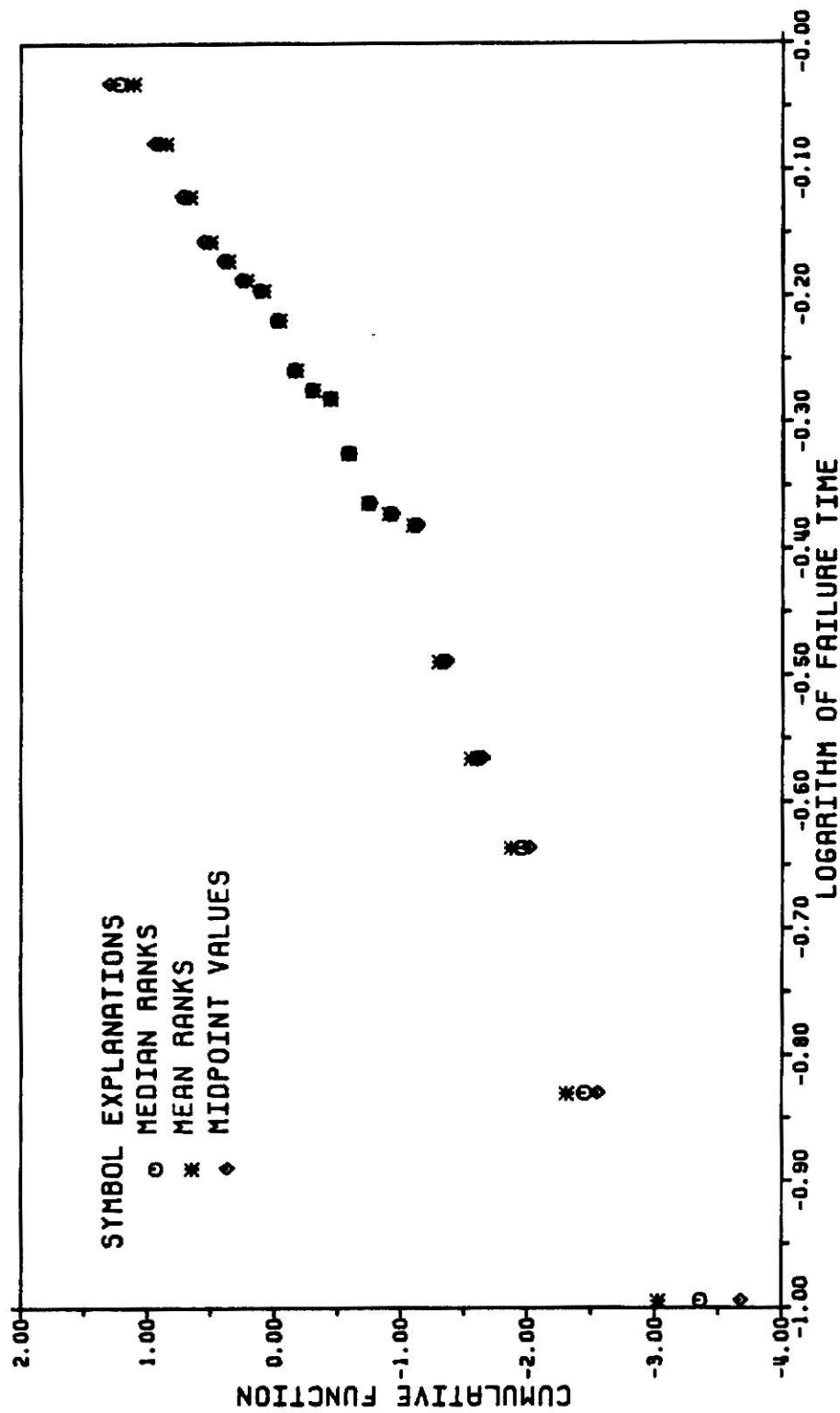


Figure 4. The plots of a Weibull sample using different plotting positions [1].

7.0 PROCEDURE

There are several different methods for recognizing the exponential failure law. One of the methods suggested by Epstein [11] uses graphical procedures for checking exponentiality.

Suppose that the underlying distribution is really exponential, then the cumulative distribution function (CDF), $F(t)$ is given by

$$F(t) = \begin{cases} 1 - e^{-t/\theta} & t > 0, \theta > 0 \\ 0 & \text{elsewhere} \end{cases} \quad (8.1)$$

From (8.1)

$$y = \log \left(\frac{1}{1 - F(t)} \right) = t/\theta \quad (8.2)$$

If we plot $y = \log [1/(1-F(t))]$ against t , we get a straight line with slope $1/\theta$. This suggests the following procedure for testing departures from exponentiality: suppose n items are put on life test and $t_1 < t_2 < \dots < t_n$ are the times when the items fail. Of course $F(t_i)$ is not known precisely so in plotting the graph, let us use the estimate (empirical value)

$$\hat{F}(t_i) = i/n+1 \quad (8.3)$$

This is done because $E(F(t_i)) = i/n+1$. Substituting $F(t_i)$ in the equation (8.2) by $i/n+1$, then y assumes the form

$$y = \log \frac{n+1}{n+1-i} \quad (8.4)$$

If the exponential assumption holds, then the plotted points can be fitted well by a straight line passing through the origin.

The actual procedure used in this work is based on assumption that $F(t)$ is not known, as described earlier in this paragraph. Therefore it

is proposed here to use five different plotting positions, which were stated in chapter four. Substituting the five chosen positions in place of $F(t)$ into equation (8.2) leads to the construction of Table 5.

Table 5. Empirical values for $F(t_i)$.

Position Number	Position Description	$F(t_i)$	y_i
1	Mean plot. pos.	$i/n+1$	$\log \frac{n+1}{n+1-i}$
2	Median approx.	$(i-0.3)/(n+0.4)$	$\log \frac{n+0.4}{n+0.7-i}$
3	Mean approx.	$(i-0.375)/(n+0.25)$	$\log \frac{n+0.25}{n+0.625-i}$
4	Midpoint plot. pos.	$i-0.5/n$	$\log \frac{n}{n+0.5-i}$
5	Median plot. pos. [2] Table C-1 p. 237		$\log \frac{1}{1-\text{values}}$ from Table C-1

To determine which of these positions is the best one for a particular distribution the following procedure was carried out.

Two computer programs were written for each distribution (exponential and Weibull). The first program uses the IMSL (The International Mathematical and Statistical Library) subroutine GGEXN to generate the exponential random numbers.

CALL GGEXN (DSEED, XM, NR, R)

This is the invocation of the random number generator

XM is the mean of the distribution

NR is the sample size

R is the output reactor of random exponential values

To invoke the random number generator for the Weibull distribution GGWIB subroutine was used.

CALL GGWIB (DSEED, BETA, NR, R)

Beta is the shape parameter

NR is the sample size

R is the output vector

The number of samples (NS) to be generated for both distributions was chosen to be 100 (if more or fewer samples are desired, simply change NS to the desired number). The next subroutine VSRTA is the sort procedure, which orders the random numbers (failure times).

CALL VSRTA (R, NR)

This is the IMSL sort procedure which orders the failure times

NR as above

R - is the output vector of ordered failure times

The third IMSL subroutine Regression Analysis (RLONE) calculates the estimate of the regression coefficient based on the different plotting positions.

CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA, STAT,
PRED, IP, NN, IER)

XY(.,.) is the matrix containing the independent and dependent
variable

DES (.) is the output vector containing basic descriptive statistic for the independent and dependent variable

ANOVA (.) is the output vector containing standard analysis of variance table

STAT (.) is the output vector containing the estimators for the model

The detailed description of those subroutines can be found in Appendix F for the exponential distribution and in Appendix G for the Weibull distribution.

The results (estimates of the regression coefficient for each plotting position) are passed to the second program (Appendix H for exponential, Appendix I for the Weibull), in which the variance and mean square error of the population parameter based on each plotting position are calculated.

The following formula was used to compute the variance

$$\text{VAR} = (\text{SS} - (\text{SB}^2/\text{NS}))/\text{NS}$$

SS - sum of squares for the estimate of the regression coefficient

SB - sum of the estimate for the regression coefficient

NS - sample size

Then using the MSE definition given in section 1.2.3, we write it as

$$\text{MSE} = (\text{SS}/\text{NS}) - 2a(\text{SB}/\text{NS}) + a^2$$

SS, NS - same as above

a - real value of a parameter in the model

or

$$\text{MSE} = \text{Var} + (a - \text{SB}/\text{NS})^2$$

Var, a, SB, NS - same as above

The best plotting position for each distribution was chosen based on the size of MSE.

8.0 RESULTS

MSE was chosen to be the criterion for determining the best plotting position for each distribution. The best plotting position is the one with the smallest mean square error. Variance, MSE and BIAS, (which is the difference between MSE and variance) were calculated for each plotting position. Typical results obtained from the second computer program (Appendix H for exponential distribution) are listed in Table 6 below.

Table 6. Results for exponential distribution.

Position Number	Position Description	MSE	Variance	Bias
1	Mean Plot. Pos.	0.4339E-05	0.3474E-05	0.8646E-06
2	Median Approx.	0.5460E-05	0.5458E-05	0.1716E-08
3	Mean Approx.	0.6409E-05	0.6393E-05	0.1662E-07
4	Midpoint Plot. Pos.	0.4962E-05	0.4947E-05	0.1476E-07
5	Median Plot. Pos.	0.4297E-05*	0.4161E-05	0.1363E-06

* indicates the best plotting position

The same parameters (Variance, MSE, Bias) were computed for the Weibull distribution (Appendix I). Two different values of the scale parameter (Alpha) and seven different values of the shape parameter (Beta) were selected. The results are presented in Table 7.

Table 7. Results for the Weibull distribution.

Parameters		Position	MSE-BETA	VAR-BETA	BIAS
Alpha	Beta	Number			
10.0	0.1	1	0.5428E-03	0.3427E-03	0.2000E-03
		2	0.5234E-03	0.5227E-03	0.7236E-06
		3	0.4540E-03*	0.4506E-03	0.3470E-05
		4	0.5132E-03	0.5131E-03	0.4400E-07
		5	0.5015E-03	0.4542E-03	0.4723E-04
5.0	0.25	1	0.3393E-02	0.2142E-02	0.1250E-02
		2	0.3273E-02	0.3268E-02	0.4539E-02
		3	0.2838E-02*	0.2817E-02	0.2165E-04
		4	0.3208E-02	0.3208E-02	0.2480E-04
		5	0.3135E-02	0.2840E-02	0.2952E-03
5.0	1.0	1	0.5428E-01	0.3427E-01	0.2000E-01
		2	0.5235E-01	0.5227E-01	0.7267E-04
		3	0.4541E-01*	0.4506E-01	0.3456E-03
		4	0.5132E-01	0.5132E-01	0.3397E-05
		5	0.5015E-01	0.4542E-01	0.4722E-02
5.0	2.0	1	0.2171E+00	0.1371E+00	0.8002E-01
		2	0.2093E+00	0.2091E+00	0.2874E-03
		3	0.1816E+00*	0.1802E+00	0.1384E-02
		4	0.2052E+00	0.2052E+00	0.1770E-04
		5	0.2005E+00	0.1816E+00	0.1889E-01
5.0	4.0	1	0.8685E+00	0.5484E+00	0.3201E+00
		2	0.8378E+00	0.8366E+00	0.1160E-02
		3	0.7267E+00*	0.7211E+00	0.5553E-02
		4	0.8213E+00	0.8212E+00	0.6348E-04
		5	0.8025E+00	0.7270E+00	0.7557E-01
5.0	6.0	1	0.1954E+01	0.1234E+01	0.7202E+00
		2	0.1885E+01	0.1882E+01	0.2617E-02
		3	0.1635E+01*	0.1623E+01	0.1249E-01
		4	0.1848E+01	0.1848E+01	0.1373E-03
		5	0.1806E+01	0.1636E+01	0.1700E+00
10.0	10.0	1	0.5428E+01	0.3427E+01	0.2001E+01
		2	0.5235E+01	0.5227E+01	0.7223E-02
		3	0.4541E+01*	0.4506E+01	0.3463E-01
		4	0.5132E+01	0.5132E+01	0.4091E-03
		5	0.5014E+01	0.4542E+01	0.4722E+00

* indicates the best plotting position

One can see that the best plotting position for the exponential distribution is the median, and for the Weibull distribution is the mean approximation plotting position. The median plotting position for the Weibull distribution appears to be inferior to the mean approximation, but better than mean plotting position.

9.0 CONCLUSIONS

The purpose of this master's report was to evaluate the various plotting positions. To accomplish this the IMSL Regression Analysis was used to evaluate the estimators based on the different plotting position. Mean Square Error (MSE) was used as a criterion for determining the best plotting position for each of the studied distributions (exponential, Weibull).

The results of the computer program for the exponential distribution indicate that the smallest mean square error is associated with the median plotting position. The mean approximation plotting position was the best one for the Weibull distribution.

The results are basically the same as presented by Kimball, with the exception that for the exponential distribution the best plotting position appears to be the median.

REFERENCES

- [1] Grosh, D., "A Primer of Reliability Theory", Unpublished manuscript, 1985.
- [2] Kapur, K. C. and Lamberson, L. R., "Reliability in Engineering Design", John Wiley & Sons, New York, 1977.
- [3] Freund, J. E. and Walpole, R. E., "Mathematical Statistics", Prentice-Hall, Inc., New Jersey, 1980.
- [4] Bloom, G., "Statistical Estimates and Transformed Beta-Variables", John Wiley & Sons, New York 1958.
- [5] Nelson, W., "Applied Life Data Analysis", John Wiley & Sons, New York, 1982.
- [6] Kimball, B. F., "On the choice of plotting positions on probability paper", Journal of the American Statistical Association, Vol. 55, September, 1960.
- [7] Ipsen, J. and Jerne, N. K., "Graphical Evaluation of the Distribution of Small Experimental Series", Acta Pathologica, 21, p. 343.
- [8] Hazen, A., "Flood Flows", John Wiley & Sons, New York, 1930.
- [9] Bliss, C. I., "The Calculation of the Time-Mortality Curve", Ann. Appl. Biol. 24, p. 815, 1937.
- [10] Bernard, A. and Bos-Levenbach, E. C., "The plotting of observations on probability paper", Statistica 7, p. 163, 1953 (in Dutch).
- [11] Epstein, B., "Tests for the Validity of the Assumption that the Underlying Distribution is Exponential. Part I", Technometrics, Vol. 2, No. 1, p. 83-101, Feb. 1960.
- [12] Epstein, B., "Tests for the Validity of the Assumption that the Underlying Distribution is Exponential. Part II", Technometrics, Vol. 2, p. 167-183, May 1960.
- [13] Shooman, M. L., "Probabilistic Reliability", McGraw-Hill, 1968.
- [14] Lee, Elisa T., "Statistical Methods for Survival Data Analysis", Lifetime Learning Publications, Belmont, California, 1980.
- [15] Mood, A. M., and Graybill, F. A., "Introduction to the Theory of Statistics", McGraw-Hill, 1963.

Appendix A. Beta Distribution

A.1. The Beta Function $B(\cdot, \cdot)$

Consider the function defined by

$$B(m, k) = \int_0^1 x^{m-1} (1-x)^{k-1} dx$$

This is a constant whose value is determined by m and k . An alternative form is obtained by the substitution $x = \sin^2 \theta$, $1-x = \cos^2 \theta$ $dx = 2 \sin \theta \cos \theta d\theta$

$$B(m, k) = 2 \int_0^{\pi/2} \sin^{2m-1} \theta \cos^{2k-1} \theta d\theta$$

To evaluate this we use a trick based on the gamma function, demonstrated in [1],

$$\begin{aligned} \Gamma(m)\Gamma(k) &= \left[2 \int_0^\pi z^{2m-1} e^{-z^2} dz \right] \left[2 \int_0^\infty w^{2k-1} e^{-w^2} dw \right] \\ &= 4 \int_0^\infty \int_0^\infty z^{2m-1} w^{2k-1} e^{-(z^2+w^2)} dz dw \end{aligned}$$

$$\text{Let } z = r \sin \theta \quad z^2 + w^2 = r^2$$

$$w = r \cos \theta \quad dz dw \rightarrow r dr d\theta$$

$$\begin{aligned} \Gamma(m)\Gamma(k) &= 4 \int_0^{\pi/2} \int_0^\infty r^{2m-1} (\sin \theta)^{2m-1} r^{2k-1} (\cos \theta)^{2k-1} e^{-r^2} r dr d\theta \\ &= \left\{ 2 \int_0^\infty r^{2(m+k)-1} e^{-r^2} dr \right\} \left\{ 2 \int_0^{\pi/2} \sin^{2m-1} \theta \cos^{2k-1} \theta d\theta \right\} \\ &= \Gamma(m+k) B(m, k) \end{aligned}$$

$$\text{Thus } B(m, k) = \frac{\Gamma(m)\Gamma(k)}{\Gamma(m+k)}$$

A.2. The Beta Family of Random Variables.

Let x be a random variable with density function

$$f(x) = k x^{m-1} (1-x)^{n-1} \text{ for } 0 \leq x \leq 1$$

$$= 0 \text{ elsewhere}$$

It can easily be proved that the normalizing constant must be $K = \frac{1}{B(m,n)}$

The moments

$$\begin{aligned} \mu'_r &= E(x^r) = \int_0^1 x^r f(x) dx = \int_0^1 \frac{x^{r+m-1} (1-x)^{n-1}}{B(m,n)} dx \\ &= \frac{B(r+m,n)}{B(m,n)} = \frac{\Gamma(r+m)\Gamma(m)}{\Gamma(m+n+r)} \frac{\Gamma(m+n)}{\Gamma(m)\Gamma(n)} = \\ &= \frac{\Gamma(m+r)\Gamma(m+n)}{\Gamma(m+n+r)\Gamma(m)} \end{aligned}$$

In particular

$$\begin{aligned} \mu &= E(x) = \frac{\Gamma(m+1)\Gamma(m+n)}{\Gamma(m+n+1)\Gamma(m)} = \frac{m \Gamma(m)\Gamma(m+n)}{(m+n)\Gamma(m+n)\Gamma(m)} = \frac{m}{m+n} \\ \mu'_2 &= E(x^2) = \frac{m(m+1)}{(m+n)(m+n+1)} \\ \sigma &= \mu'_2 - \mu^2 = \frac{mn}{(m+n)^2(m+n+1)} \end{aligned}$$

Mode of distribution

$$\text{Set } y = k x^{m-1} (1-x)^{n-1}$$

$$\text{then } y' = k x^{m-2} (1-x)^{n-2} [-x(n+m-2) + m-1]$$

$$\text{so the mode is } x_0 = \frac{m-1}{m+n-2}.$$

Setting $y'' = 0$ gives inflection points at

$$\frac{m-1}{m+n-2} \pm \frac{1}{m+n-2} \sqrt{\frac{(m-1)(n-1)}{m+n-3}}$$

Appendix B. The Weibull Distribution

B.1. The Weibull Distribution

The cumulative distribution for random variable distributed as the three-parameter Weibull is given by

$$F(x; \theta, \beta, \sigma) = 1 - e^{-\left(\frac{x - \sigma}{\theta - \sigma}\right)^\beta} \quad x > \sigma \quad (B.1)$$

where $\beta > 0, \theta > 0$ and $\sigma > 0$. The parameter β is called the shape parameter or the Weibull slope; the parameter θ is called the scale parameter or the characteristic life; and the parameter σ is called the location parameter or the minimum life.

The two-parameter Weibull has a minimum life of zero, and the cumulative distribution is given by

$$F(x; \theta, \beta) = 1 - e^{-(x/\theta)^\beta} \quad x > 0 \quad (B.2)$$

The p.d.f. is obtained by differentiating equation (B.2) as

$$f(x; \theta, \beta) = \frac{\beta}{\theta} \left(\frac{x}{\theta}\right)^{\beta-1} e^{-(x/\theta)^\beta}, \quad x > 0$$

The k th moment of the Weibull distribution is found as follows:

$$\mu'_k = E(x^k) = \int_0^\infty x^k \frac{\beta}{\theta} \left(\frac{x}{\theta}\right)^{\beta-1} e^{-(x/\theta)^\beta} dx$$

Using the transformation

$$u = \left(\frac{x}{\theta}\right)^\beta \rightarrow x = \theta \cdot u^{1/\beta}$$

then

$$du = \frac{\beta}{\theta} \left(\frac{x}{\theta}\right)^{\beta-1} dx$$

and we have

$$\mu'_k = \theta^k \int_0^{\infty} u^{k/\beta} e^{-u} du$$

which is recognizable as a gamma function and yields

$$\mu'_k = \theta^k \Gamma(1 + \frac{k}{\beta})$$

Setting $k = 1$ yields the mean of the Weibull distribution

$$\mu = \theta \Gamma(1 + \frac{1}{\beta})$$

and the variance is

$$\sigma^2 = \theta^2 [\Gamma(1 + \frac{2}{\beta}) - \Gamma^2(1 + \frac{1}{\beta})]$$

The shape parameter β , as the name implies, determines the shape of the distribution. Figure 5 shows various forms of the Weibull distribution where the scale parameter (θ) is taken as unity.

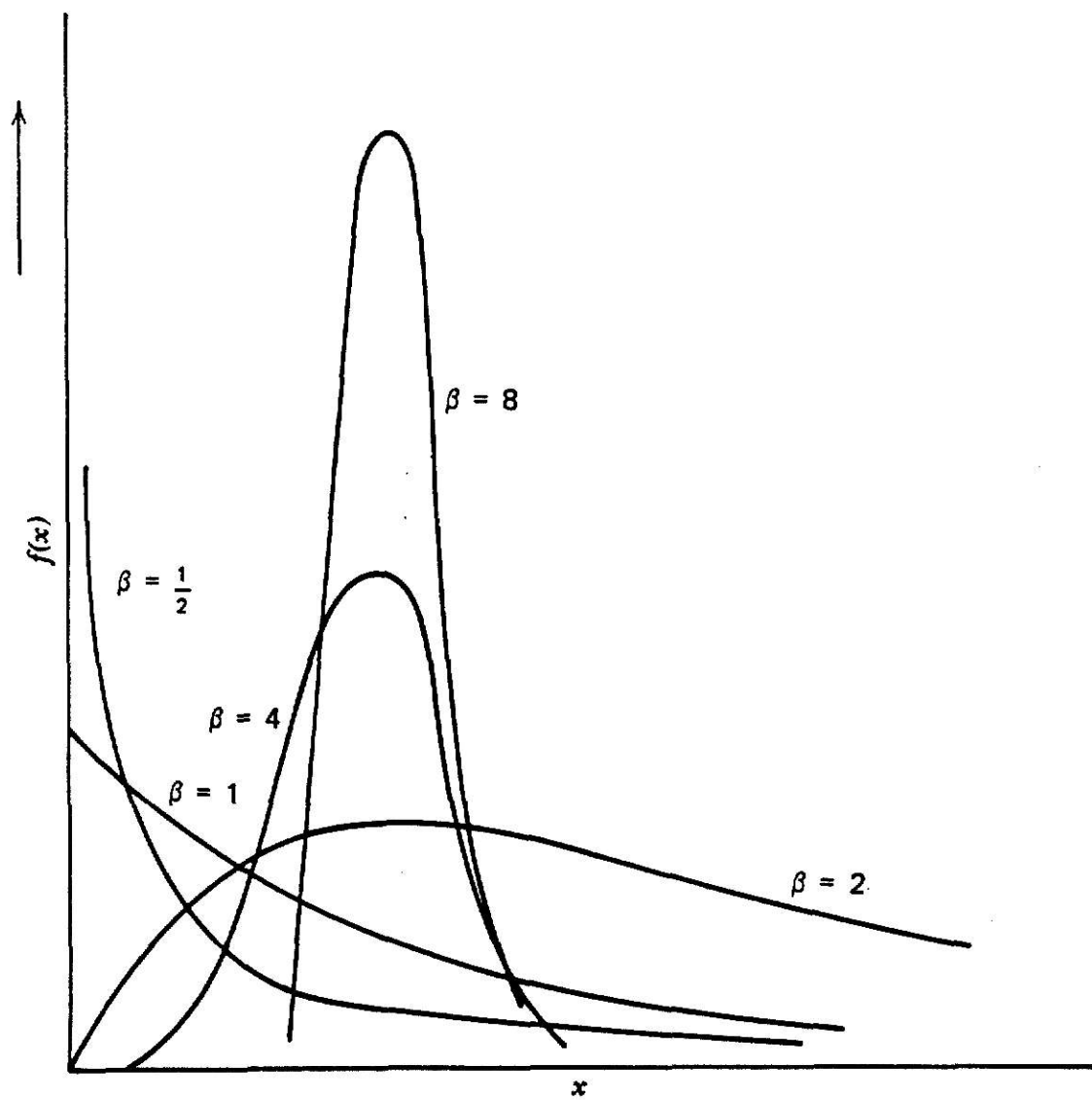


Figure 5. The Weibull distribution for different values of β .

APPENDIX C. The Exponential Distribution

A random variable X has an exponential distribution, and it is referred to as an exponential random variable, if and only if its probability density is given by

$$f(x) = \begin{cases} 1/\theta e^{-x/\theta} & \text{for } x > 0, \theta > 0 \\ 0 & \text{otherwise} \end{cases}$$

Let us refer to the following example [3]. Suppose, for instance, that $f(x,t)$ is the probability of getting x successes during the time interval of length t when

- i) the probability of success during a very small time interval from t to $t+\Delta t$ is $\alpha \cdot \Delta t$,
- ii) the probability of more than one success during such a time interval is negligible, and
- iii) the probability of success during such a time interval does not depend on what happened prior to time t .

Let us determine the probability density of the random variable y , the waiting time until the first success. Using the formula for the Poisson distribution, we find that the probability of zero successes during a time interval of length y is

$$\frac{(\lambda y)^0 e^{-\lambda y}}{0!} = e^{-\lambda y}$$

and the probability of getting the first success prior to time y (the CDF of y) is:

$$F(y) = \begin{cases} 1 - e^{-\lambda y} & \text{for } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Having thus found the distribution function of y , differentiation with respect to y yields

$$f(y) = \begin{cases} \lambda e^{-\lambda y} & \text{for } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

This is an exponential PDF with respect to $\theta = 1/\lambda$

The exponential PDF is a special case of general Gamma PDF in which $\alpha = 1$ and β has been replaced by $1/\lambda$.

Appendix D. Rank Tables - (median, 5%, 95%)[2]

Median Ranks

j\n	Sample Size									
	1	2	3	4	5	6	7	8	9	10
1	50.000	29.289	20.630	15.910	12.945	10.910	9.428	8.300	7.412	6.697
2		70.711	50.000	38.573	31.381	26.445	22.849	20.113	17.962	16.226
3			79.370	61.427	50.000	42.141	36.412	32.052	28.624	25.857
4				84.090	68.619	57.859	50.000	44.015	39.308	35.510
5					87.055	73.555	63.588	55.984	50.000	45.169
6						89.090	77.151	67.948	60.691	54.831
7							90.572	79.887	71.376	64.490
8								91.700	82.038	74.142
9									92.587	83.774
10										93.303

Medium Ranks

j\n	Sample Size									
	11	12	13	14	15	16	17	18	19	20
1	6.107	5.613	5.192	4.830	4.516	4.240	3.995	3.778	3.582	3.406
2	14.796	13.598	12.579	11.702	10.940	10.270	9.678	9.151	8.677	8.251
3	23.578	21.669	20.045	18.647	17.432	16.365	15.422	14.581	13.827	13.147
4	32.380	29.758	27.528	25.608	23.939	22.474	21.178	20.024	18.988	18.055
5	41.189	37.853	35.016	32.575	30.452	28.589	26.940	25.471	24.154	22.967
6	50.000	45.951	42.508	39.544	36.967	34.705	32.704	30.921	29.322	27.880
7	58.811	54.049	50.000	46.515	43.483	40.823	38.469	36.371	34.491	32.795
8	67.620	62.147	57.492	53.485	50.000	46.941	44.234	41.823	39.660	37.710
9	76.421	70.242	64.984	60.456	56.517	53.059	50.000	47.274	44.830	42.626
10	85.204	78.331	72.472	67.425	63.033	59.177	55.766	52.726	50.000	47.542
11	93.893	86.402	79.955	74.392	69.548	65.295	61.531	58.177	55.170	52.458
12		94.387	87.421	81.353	76.061	71.411	67.296	63.629	60.340	57.374
13			94.808	88.298	82.568	77.525	73.060	69.079	65.509	62.289
14				95.169	89.060	83.635	78.821	74.529	70.678	67.205
15					95.484	89.730	84.578	79.976	75.846	72.119
16						95.760	90.322	85.419	81.011	77.033
17							96.005	90.849	86.173	81.945
18								96.222	91.322	86.853
19									96.418	91.749
20										96.594

Median Ranks

j\n	Sample Size									
	21	22	23	24	25	26	27	28	29	30
1	3.247	3.101	2.969	2.847	2.734	2.631	2.534	2.445	2.362	2.284
2	7.864	7.512	7.191	6.895	6.623	6.372	6.139	5.922	5.720	5.532
3	12.531	11.970	11.458	10.987	10.553	10.153	9.781	9.436	9.114	8.814
4	17.209	16.439	15.734	15.088	14.492	13.942	13.432	12.958	12.517	12.104
5	21.890	20.911	20.015	19.192	18.435	17.735	17.086	16.483	15.922	15.397
6	26.574	25.384	24.297	23.299	22.379	21.529	20.742	20.010	19.328	18.691
7	31.258	29.859	28.580	27.406	26.324	25.325	24.398	23.537	22.735	21.986
8	35.943	34.334	32.863	31.513	30.269	29.120	28.055	27.065	26.143	25.281
9	40.629	38.810	37.147	35.621	34.215	32.916	31.712	30.593	29.550	28.576
10	45.314	43.286	41.431	39.729	38.161	36.712	35.370	34.121	32.958	31.872
11	50.000	47.762	45.716	43.837	42.107	40.509	39.027	37.650	36.367	35.168
12	54.686	52.238	50.000	47.946	46.054	44.305	42.685	41.178	39.775	38.464
13	59.371	56.714	54.284	52.054	50.000	48.102	46.342	44.707	43.183	41.760
14	64.057	61.190	58.568	56.162	53.946	51.898	50.000	48.236	46.592	45.056
15	68.742	65.665	62.853	60.271	57.892	55.695	53.658	51.764	50.000	48.352
16	73.426	70.141	67.137	64.379	61.839	59.491	57.315	55.293	53.408	51.648
17	78.109	74.616	71.420	68.487	65.785	63.287	60.973	58.821	56.817	54.944
18	82.791	79.089	75.703	72.594	69.730	67.084	64.630	62.350	60.225	58.240
19	87.469	83.561	79.985	76.701	73.676	70.880	68.288	65.878	63.633	61.536
20	92.136	88.030	84.266	80.808	77.621	74.675	71.945	69.407	67.041	64.852
21	96.753	92.488	88.542	84.912	81.565	78.471	75.602	72.935	70.450	68.128
22		96.898	92.809	89.013	85.507	82.265	79.258	76.463	73.857	71.424
23			97.031	93.105	89.447	86.058	82.914	79.990	77.265	74.719
24				97.153	93.377	89.847	86.568	83.517	80.672	78.014
25					97.265	93.628	90.219	87.042	84.078	81.309
26						97.369	93.861	90.564	87.483	84.603
27							97.465	94.078	90.865	87.896
28								97.555	94.280	91.186
29									97.638	94.468
30										97.716

APPENDIX E. Example of Nelson's procedure [5] for making a probability plot.

Table 8 contains times to breakdown of an insulating fluid between electrodes recorded at seven different voltages. For the failure with rank i , the midpoint plotting position was used:

$$F_i = 100(i - 0.5)/n \quad i = 1, 2, \dots, n$$

A test purpose was to assess whether time to breakdown at each voltage has an exponential distribution. If appropriate, the distribution can be used to estimate the probability of fluid breakdown in actual use.

It can be seen in Figure 6 that each failure was plotted (on exponential probability paper) against its time on the data scale and against its plotting position on the cumulative probability scale. For the data in Figure 6 the exponential distribution might be questioned, so the exponential fit is crude. The sample is small (19 observations), so it reveals only large departures from exponential distribution. One should try to plot the data on the Weibull probability paper.

The mean of an exponential distribution is the 63-rd percentile. The graphical estimate of the mean from Figure 5 is 14 minutes.

The failure rate of an exponential distribution is the reciprocal of the mean, which from Figure 5 is $1/14 = 0.07$ failure per minute.

Note:

The first step in this procedure was ordering the n data values from smallest to largest, as can be seen in Table 8. The smallest has rank 1, the next larger has rank 2, ..., and the largest rank n .

Table 8. Times to Breakdown of an Insulating Fluid [5].

26 kV		28 kV		30 kV		32 kV	
Min-utes	Plotting Position	Min-utes	Plotting Position	Min-utes	Plotting Position	Min-utes	Plotting Position
5.79	16.3	68.85	10.0	7.74	4.5	0.27	3.3
1579.52	50.0	108.29	30.0	17.05	13.6	0.40	10.0
2323.70	83.3	110.29	50.0	20.46	22.7	0.69	16.7
		426.07	70.0	21.02	31.8	0.79	23.3
34 kV		1067.60	90.0	22.66	40.9	2.75	30.0
				43.40	50.0	3.91	36.7
				47.30	59.1	9.88	43.3
Min-utes	Plotting Position	36 kV		139.07	68.2	13.95	50.0
				144.12	77.3	15.93	56.7
0.19	2.6	Min-utes	Plotting Position	175.88	86.4	27.80	63.3
0.78	7.9			194.90	95.5	53.24	70.0
0.96	13.2					82.85	76.7
1.31	18.4	0.35	3.3			89.29	83.3
2.78	23.7	0.59	10.0			100.58	90.0
3.16	28.9	0.96	16.7	38 kV		215.10	96.7
4.15	34.2	0.99	23.3				
4.67	39.5	1.69	30.0	Min-utes	Plotting Position		
4.85	44.7	1.97	36.7				
6.50	50.0	2.07	43.3				
7.35	55.3	2.58	50.0	0.09	6.2		
8.01	60.5	2.71	56.7	0.39	18.7		
8.27	65.8	2.90	63.3	0.47	31.2		
12.06	71.1	3.67	70.0	0.73	43.7		
31.75	76.3	3.99	76.7	0.74	56.2		
32.52	81.6	5.35	83.3	1.13	68.7		
33.91	86.8	13.77	90.0	1.40	81.2		
36.71	92.1	25.50	96.7	2.38	93.7		
72.89	97.4						

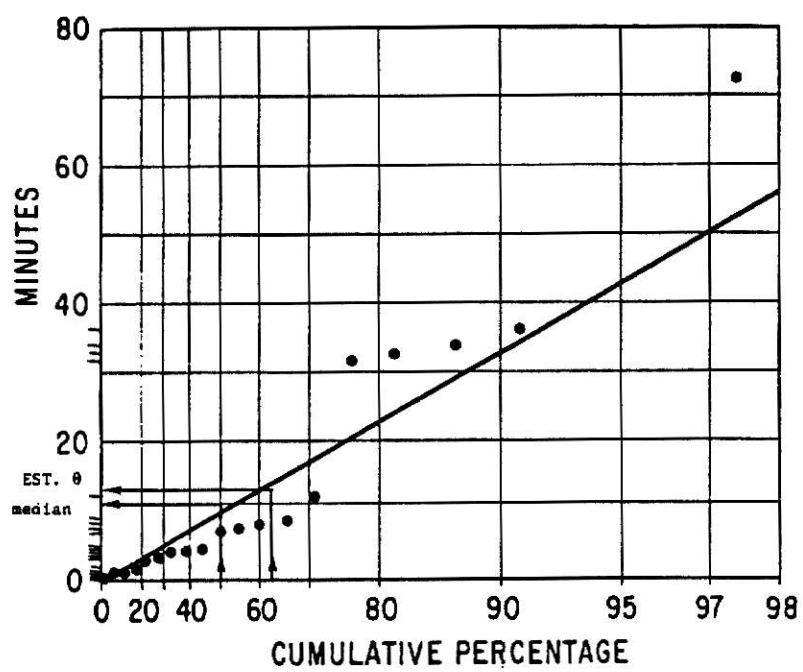


Figure 6. Exponential probability plot of the 34-kV data [5].

APPENDIX F. Regression analysis program for the exponential distribution.

```

/**++ REGION 400K
/**++ PRINT
/**++ LINES 4
/**++ VMMSG
/**++ SERVICE *
// EXEC FORTVCLG
//SYSIN DD *
      INTEGER NR, NS, IX, N, IMOD, IPRED, IP, NN, IER
C      NR IS THE SAMPLE SIZE
C      NS IS THE NUMBER OF SAMPLES TO BE GENERATED
C      IX IS THE NUMBER OF ROWS IN XY MATRIX
C      N IS THE NUMBER OF DATA POINTS
C      IMOD IS THE MODEL OPTION
C      IPRED IS THE PREDICTION ANALYSIS OPTION
C      IP IS THE NUMBER OF ROWS IN PREP MATRIX
C      NN IS THE NUMBER OF POINTS FOR WHICH A PREDICTION ANALYSIS
C      IS DESIRED
      REAL XM, R(20), XY(20,2), ALBAP(3), DES(5), ANOVA(14)
      REAL STAT(9), PRED(1,7), XX(20)
C      XM IS THE MEAN OF THE EXPONENTIAL DISRTBUTION
C      R(.) IS THE OUTPUT VECTOR OF RANDOM EXPONENTIAL VALUES
C      FROM THE IMSL GENERATOR
C      XY(...) IS THE MATRIX CONTAINING THE INDEPENDENT AND
C      DEPENDENT VARIABLE
C      ALBAP(.) IS A INPUT VECTOR CONTAINING RISK LEVELS
C      DES(.) IS AN OUTPUT VECTOR CONTAINING BASIC DESCRIPTIVE
C      STATISTICS
C      ANOVA(.) IS AN OUTPUT VECTOR CONTAINING STANDARD ANALYSIS
C      OF VARIANCE TABLE
C      STAT(.) IS AN OUTPUT VECTOR CONTAINING THE ESTIMATORS
C      FOR THE MODEL PARAMETERS
C      PRED(...) IS AN OUTPUT MATRIX CONTAINING PREDICTION ANAYSIS
      DOUBLE PRECISION DSEED
      DSEED= 123457.DO
      XM = 100.
      NR =20
      NS = 100
      IX=20
      N=20
      IMOD=1
      IPRED=-1
      ALBAP(1) = 0.05
      ALBAP(2) = 0.05
      READ(5,*) (XX(I),I=1,NR)
      DO 200 M=1,NS
      CALL GGEXN (DSEED, XM, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
      CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
      DO 300 I=1,NR
      XY(I,1)= R(I)
300 CONTINUE

```

```

DO 310 I=1,NR
XY(I,2)= LOG(REAL(NR+1)/REAL(NR+1-I))
C      THIS IS THE MEAN PLOTTING POSITION
310 CONTINUE
CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA, STAT,
&          PREP, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
WRITE(6,32) M, STAT(1)
C          M - SAMPLE NUMBER
C          STAT(1) - ESTIMATE OF THE REGRESSION COEFFICIENT
32  FORMAT(5X,I3,5X,F10.6)
200 CONTINUE
DO 201 M=1,NS
CALL GGEXN (DSEED, XM, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
DO 130 I=1,NR
XY(I,1) = R(I)
130 CONTINUE
DO 320 I=1,NR
XY(I,2)= LOG((NR+0.4)/(NR+0.7-I))
C      THIS IS MEDIAN APPROXIMATION PLOTTING POSITION
320 CONTINUE
CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
&          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
WRITE(6,33) M, STAT(1)
33  FORMAT(5X,I3,5X,F10.6)
201 CONTINUE
DO 202 M=1,NS
CALL GGEXN (DSEED, XM, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
DO 119 I=1,NR
XY(I,1) = R(I)
119 CONTINUE
DO 120 I=1,NR
XY(I,2)= LOG((NR+0.375)/(NR+0.625-I))
C      THIS IS MEAN APPROXIMATION PLOTTING POSITION
120 CONTINUE
CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
&          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
WRITE(6,34) M, STAT(1)
34  FORMAT(5X,I3,5X,F10.6)
202 CONTINUE
DO 203 M=1,NS
CALL GGEXN (DSEED, XM, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
DO 181 I=1,NR
XY(I,1) = R(I)

```

```

181  CONTINUE
    DO 182 I=1,NR
      XY(I,2)= LOG(NR/(NR+0.5-I))
182  CONTINUE
C      THIS IS MIDPOINT PLOTTING POSITION
    CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
      &          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
    WRITE(6,35) M, STAT(1)
35   FORMAT(5X,I3,5X,F10.6)
203  CONTINUE
    DO 204 M=1,NS
      CALL GGEXN (DSEED, XM, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
    CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
    DO 140 I=1,NR
      XY(I,1) = R(I)
140  CONTINUE
    DO 142 I=1,NR
      XY(I,2) = LOG(1./(1-XX(I)))
C      THIS IS MEDIAN PLOTTING POSITION
142  CONTINUE
    CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
      &          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
    WRITE(6,36) M, STAT(1)
36   FORMAT(5X,I3,5X,F10.6)
204  CONTINUE
    STOP
    END
//LKED.SYSLIB DD
//          DD DSN=SYS1.IMSL.SPFLIB,DISP=SHR
//GO.SYSIN  DD *

```

APPENDIX G. Regression analysis program for the Weibull distribution.

```

/**++ REGION 400K
/**++ PRINT
/**++ LINES 4
/**++ VMMSG
/**++ SERVICE *
// EXEC FORTVCLG
//SYSIN      DD *
      INTEGER NR, NS, IX, N, IMOD, IPRED, IP, NN, IER, K,L,S,J,P
C      NR IS THE SAMPLE SIZE
C      NS IS THE NUMBER OF SAMPLES TO BE GENERATED
C      IX IS THE NUMBER OF ROWS IN XY MATRIX
C      N IS THE NUMBER OF DATA POINTS
C      IMOD IS THE MODEL OPTION
C      IPRED IS THE PREDICTION ANALYSIS OPTION
C      IP IS THE NUMBER OF ROWS IN PREP MATRIX
C      NN IS THE NUMBER OF POINTS FOR WHICH A PREDICTION ANALYSIS
C      IS DESIRED
      REAL ALPHA,BETA, R(20), XY(20,2), ALBAP(3), DES(5), ANOVA(14)
      REAL STAT(9), PRED(1,7), XX(20), COEF
C      ALPHA IS THE SCALE PARAMETER
C      BETA IS THE SHAPE PARAMETER (WEIBULL SLOPE)
C      R(.) IS THE OUTPUT VECTOR OF RANDOM EXPONENTIAL VALUES
C      FROM THE IMSL GENERATOR
C      XY(.,.) IS THE MATRIX CONTAINING THE INDEPENDENT AND
C      DEPENDENT VARIABLE
C      ALBAP(.) IS A INPUT VECTOR CONTAINING RISK LEVELS
C      DES(.) IS AN OUTPUT VECTOR CONTAINING BASIC DESCRIPTIVE
C      STATISTICS
C      ANOVA(.) IS AN OUTPUT VECTOR CONTAINING STANDARD ANALYSIS
C      OF VARIANCE TABLE
C      STAT(.) IS AN OUTPUT VECTOR CONTAINING THE ESTIMATORS
C      FOR THE MODEL PARAMETERS
C      PRED(.,.) IS AN OUTPUT MATRIX CONTAINING PREDICTION ANALYSIS
      DOUBLE PRECISION DSEED
      DSEED= 123457.DO
      ALPHA = 10.0
      BETA = 10.0
      CONST = -1.0/BETA
      COEF = ALPHA**CONST
      NR =20
      NS = 100
      K = 1
      L = 2
      J = 3
      S = 4
      P = 5
      IX=20
      N=20
      IMOD=0
      IPRED=-1
      ALBAP(1) = 0.05
      ALBAP(2) = 0.05

```

```

      READ(5,*) (XX(I),I=1,NR)
      DO 200 M=1,NS
      CALL GGWIB (DSEED, BETA, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
      DO 50 I=1,NR
      R(I) = COEF*R(I)
50    CONTINUE
      CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
      DO 300 I=1,NR
      XY(I,1)= LOG(R(I))
300   CONTINUE
      DO 310 I=1,NR
      XY(I,2)= LOG(LOG(REAL(NR+1)/REAL(NR+1-I)))
C      THIS IS THE MEAN PLOTTING POSITION
310   CONTINUE
      CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA, STAT,
&      PREP, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
      WRITE(6,32) K,M,STAT(1), STAT(5)
C      M - SAMPLE NUMBER
C      STAT(1) - ESTIMATE OF THE REGRESSION COEFFICIENT
C      STAT(5) - ESTIMATE OF THE INTERCEPT
32    FORMAT(3X,I2,3X,I3,3X,F10.6,3X,F10.6)
200   CONTINUE
      DO 201 M=1,NS
      CALL GGWIB (DSEED, BETA, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
      DO 51 I=1,NR
      R(I) = COEF*R(I)
51    CONTINUE
      CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
      DO 130 I=1,NR
      XY(I,1) = LOG(R(I))
130   CONTINUE
      DO 320 I=1,NR
      XY(I,2)= LOG(LOG((NR+0.4)/(NR+0.7-I)))
C      THIS IS MEDIAN APPROXIMATION PLOTTING POSITION
320   CONTINUE
      CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
&      STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
      WRITE(6,33) L,M, STAT(1), STAT(5)
33    FORMAT(3X,I2,3X,I3,3X,F10.6,3X,F10.6)
201   CONTINUE
      DO 202 M=1,NS
      CALL GGWIB (DSEED, BETA, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
      DO 52 I=1,NR
      R(I) = COEF*R(I)
52    CONTINUE
      CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
      DO 119 I=1,NR

```

```

      XY(I,1) = LOG(R(I))
119  CONTINUE
      DO 120 I=1,NR
      XY(I,2)= LOG(LOG((NR+0.375)/(NR+0.625-I)))
C      THIS IS MEAN APPROXIMATION PLOTTING POSITION
120  CONTINUE
      CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
&          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
      WRITE(6,34) J, M, STAT(1), STAT(5)
34   FORMAT(3X,I2,3X,I3,3X,F10.6,3X,F10.6)
202  CONTINUE
      DO 203 M=1,NS
      CALL GGWIB (DSEED, BETA, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
      DO 53 I=1,NR
      R(I) = COEF*R(I)
53   CONTINUE
      CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
      DO 181 I=1,NR
      XY(I,1) = LOG(R(I))
181  CONTINUE
      DO 182 I=1,NR
      XY(I,2)= LOG(LOG(NR/(NR+0.5-I)))
182  CONTINUE
C      THIS IS MIDPOINT PLOTTING POSITION
      CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
&          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
      WRITE(6,35) S, M, STAT(1), STAT(5)
35   FORMAT(3X,I2,3X,I3,3X,F10.6,3X,F10.6)
203  CONTINUE
      DO 204 M=1,NS
      CALL GGWIB (DSEED, BETA, NR, R)
C      THIS IS THE INVOCATION OF THE RANDOM NUMBER GENERATOR
      DO 54 I=1,NR
      R(I) = COEF*R(I)
54   CONTINUE
      CALL VSRTA(R,NR)
C      THIS IS THE IMSL SORT PROCEDURE
      DO 140 I=1,NR
      XY(I,1) = LOG(R(I))
140  CONTINUE
      DO 142 I=1,NR
      XY(I,2) = LOG(LOG(1./(1-XX(I))))
C      THIS IS MEDIAN PLOTTING POSITION
142  CONTINUE
      CALL RLONE (XY, IX, N, IMOD, IPRED, ALBAP, DES, ANOVA,
&          STAT, PRED, IP, NN, IER)
C      THIS IS IMSL ANALYSIS OF A SIMPLE LINEAR REGRESSION
      WRITE(6,36) P, M, STAT(1), STAT(5)
36   FORMAT(3X,I2,3X,I3,3X,F10.6,3X,F10.6)
204  CONTINUE
      STOP

```

```
      END  
//LKED.SYSLIB DD  
//          DD DSN=SYS1.IMSL.SPFLIB,DISP=SHR  
//GO.SYSIN   DD *
```

APPENDIX H. Program for calculating the MSE for the exponential distribution.

```

/**++ REGION 300K
/**++ PRINT VMMSG SERVICE *
/**++ LINES 2
// EXEC FORTVCLG
      INTEGER NS, POS
C      NS - IS THE SAMPLE NUMBER
C      POS- IS THE PLOTTING POSITION
C      1=THE MEAN PLOTTING POSITION
C      2=THE MEDIAN APPROXIMATION PLOTTING POSITION
C      3=THE MEAN APPROXIMATION PLOTTING POSITION
C      4=THE MIDPOINT PLOTTING POSITION
C      5=THE MEDIAN PLOTTING POSITION
      REAL MSE, VAR, X(100), BIA
C      MSE - MEAN SQUARE ERROR
C      VAR - VARIANCE
C      BIA - BIAS
C      X(.) - ESTIMATE OF THE REGRESSION COEFFICIENT
      NS=100
      POS=5
      WRITE(6,*) ' POSITION      MSE      VARIANCE      BIAS'
      WRITE(6,*) '
      DO 30 J=1,POS
      SS=0
      SB=0
      BIA=0
      DO 10 I=1,NS
      READ(5,*) M, NS1, X(I)
C      WRITE(6,*) M, NS1, X(I)
10    CONTINUE
      DO 20 I=1,NS
      SS = X(I)**2 + SS
      SB = X(I) + SB
20    CONTINUE
      VAR = (SS - (SB**2)/NS)/NS
      MSE = (SS/NS) - 0.02*(SB/NS) + 0.01**2
      BIA = MSE - VAR
      WRITE(6,15) M, MSE, VAR, BIA
15    FORMAT(5X,I3,5X,E10.4,5X,E10.4,5X,E10.4)
30    CONTINUE
      STOP
      END
//GO.SYSIN DD *
C      DATA

```

APPENDIX I. Program for calculating the MSE for the Weibull distribution.

```

/*++ REGION 300K
/*++ PRINT VMMSG SERVICE *
/*++ LINES 2
// EXEC FORTVCLG
      INTEGER NS, POS
C      NS - IS THE SAMPLE NUMBER
C      POS- IS THE PLOTTING POSITION
C      1=THE MEAN PLOTTING POSITION
C      2=THE MEDIAN APPROXIMATION PLOTTING POSITION
C      3=THE MEAN APPROXIMATION PLOTTING POSITION
C      4=THE MIDPOINT PLOTTING POSITION
C      5=THE MEDIAN PLOTTING POSITION
C      ALPHA = 10.0
      REAL MSE, VAR, X(100), BIA, Y(100), Z(100), MSI, VAI, BII
C      MSE - MEAN SQUARE ERROR FOR BETA
C      VAR - VARIANCE FOR BETA
C      BIA - BIAS (BETA)
C      X(.) - ESTIMATE OF THE REGRESSION COEFFICIENT
C      Y(.) - ESTIMATE OF THE INTERCEPT
C      Z(.) - NEW VALUE OF THE INTERCEPT ESTIMATE
C              E TO THE Y(.) POWER
C      MSI - MEAN SQUARE ERROR FOR INTERCEPT
C      VAI - VARIANCE FOR INTERCEPT
C      BII - BIAS (INTERCEPT)

      NS=100
      POS=5
      WRITE(6,*) 'BETA EQUALS 10.0'
      WRITE(6,*) '
      WRITE(6,*) 'POSITION   MSE-BETA   VAR-BETA   BIAS-BETA   MSE-
&INT   VAR-INT   BIAS-INT'
      WRITE(6,*) '
      DO 30 J=1,POS
      SS=0
      SB=0
      BIA=0
      SSI=0
      SA=0
      BII=0
      DO 10 I=1,NS
      READ(5,*) M, NS1, X(I), Y(I)
10    CONTINUE
      DO 20 I=1,NS
      SS = X(I)**2 + SS
      SB = X(I) + SB
      Z(I) = EXP(Y(I))
      SSI = Z(I)**2 + SSI
      SA = Z(I) + SA
20    CONTINUE
      VAR = (SS - (SB**2)/NS)/NS
      MSE = (SS/NS) - 20.0*(SB/NS) + 10.0**2
      BIA = MSE - VAR
      VAI = (SSI - (SA**2)/NS)/NS

```

```

      MSI = VAI + (10.0 - (SA/NS))**2
      BII = MSI - VAI
      WRITE(6,15) M, MSE, VAR, BIA, MSI, VAI, BII
15    FORMAT(3X,I2,5X,E10.4,3X,E10.4,3X,E10.4,3X,E10.4,3X,E10.4,
&3X,E10.4)
30    CONTINUE
      STOP
      END
//GO.SYSIN DD *
C      DATA

```

AN EVALUATION OF VARIOUS PLOTTING POSITIONS

by

MALGORZATA JOANNA RYS

M.S., Technical University of Wroclaw, Wroclaw
Poland, 1979

AN ABSTRACT OF A MASTER'S REPORT
submitted in partial fulfillment of the
requirements for the degree

MASTER OF SCIENCE

Department of Industrial Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

1986

ABSTRACT

It has been stated that the hazard function could be used for distinguishing between various failure models such as the gamma, Weibull, exponential, normal, etc. The results presented by D. Grosh [1], were in disagreement with the above belief, due to the fact that few of the plots looked anything like their theoretical curves. In the next approach, Grosh used the old fashioned probability plotting technique; the kind that uses appropriate scaling to convert the cumulative distribution function into the straight line.

In order to use the latter method in its best form the different probability plotting positions should be considered. Bradford F. Kimball [6] has done limited study of comparing some of the possible plotting position. This work is an extension of the problems studied by Kimball. The following different plotting positions were discussed and analyzed in this report: mean plotting position, median, median approximation, midpoint and mean approximation. Separate programs were written for each of the studied distributions (exponential, Weibull) using an IMSL random variable generator. The IMSL regression analysis was used to evaluate the estimators based on the different plotting position. Mean Square Error was used as a criterion for determining the best plotting position for each distribution. It was found that the best plotting position (with the smallest mean square error) is the median for exponential distribution and the mean approximation for the Weibull distribution.