CONF-9302243--2

WSRC-MS--92-522

DE93 006152

STATISTICAL SOFTWARE FOR RISK ANALYSIS AT THE SAVANNAH RIVER SITE (U)

by J. H. Weber

Westinghouse Savannah River Company
Savannah River Site
Aiken, South Carolina 29808

Other Authors:

A paper proposed for Presentation/Publication
at/in the Data Banks for Risk Assessment Workshop
Augusta, GA
02/02-03/93

DISCLAIMER

WSRC-MS-92-522

STATISTICAL SOFTWARE FOR RISK ANALYSIS AT THE SAVANNAH RIVER SITE (U)

BY

J. H. Weber
Westinghouse Savannah River Company
Savannah River Site
Aiken, SC 29802

A paper proposed for presentation at
Data Banks for Risk Assessment Workshop
Augusta, Georgia
February 2-3, 1993

WSRC-MS-92-

STATISTICAL SOFTWARE FOR RISK ANALYSIS AT THE SAVANNAH RIVER SITE (U)

BY

J. H. Weber
Westinghouse Savannah River Company
Savannah River Site
Aiken, SC 29802

A paper proposed for presentation at
Data Banks for Risk Assessment Workshop
Augusta, Georgia
February 2-3, 1993

STATISTICAL SOFTWARE FOR RISK ANALYSIS AT THE SAVANNAH RIVER SITE (U)

ABSTRACT

This paper describes statistical software developed at the Savannah River Site (SRS) to analyze event time of occurrence data extracted from fault tree data banks and/or user defined input data files. Five different distributions can currently be fit to the empirical data: normal, lognormal, exponential, Weibull and loguniform. Two goodness of fit tests, the Kolmogorov-Smirnov one-sample test and the Chi-squared test, are used to determine how well a particular distribution fits the observed data. In addition, a comparison across fitted distributions is done to determine the most likely distribution fitting the data. A number of graphics can be generated illustrating the important characteristics of the data and how well each theoretical distribution fits the data. The theoretical distribution which best fits the observed data, the expected occurrence rate, and the probability of occurrence are used in fault tree analyses. Results from the SRS developed software were compared with commercially developed and tested software, SAS®.

INTRODUCTION

The Safety Analysis and Risk Management Department of the Savannah River Technology Center performs safety studies requiring the occurrence rate and distribution characteristics for different types of events. Consequently, Savannah River Site maintains several data banks for unusual events, equipment failure and replacement, accidents, etc. which are used in fault tree analyses. In the late 1970's a computer code (STATPAC) was developed to analyze the data from one data bank, SEPR, a database for unusual events in the Separations areas. STATPAC has been modified several times over the years, can accept as data input, data extracted from any of the fault tree data banks at SRS, and can be operated on a mainframe IBM under the MVS operating system and in the VM environment.

STATPAC was designed specifically to be used with the data extracted from SRS fault tree data banks but also allows data files to be input directly by the user. The user can define the data file by listing times between occurrences or dates of occurrences. Table 1 illustrates the data input file as extracted from a data bank. Table 2 shows the user defined data file with dates of occurrences and Table 3 shows the user defined data file with times between occurrences. A second user defined options file is required which describes the type and format of the data file, distributions to be fit, plots, descriptive titles and output labels.

INPUT DATA AND OPTION FILES

The input data file can be one of three types. The first and most general usage is for the data to be extracted from one of the fault tree data banks. An example of the data from the data banks is given in Table 1. The data includes the date of occurrence, an event identification number, codes indicating source(s) of the incident, area, facility, type of equipment, type of incident, and a description of the incident. STATPAC converts the date of occurrence to time between occurrences. If the data is not available from one of the data banks or if the user wants to input his own data file, then two formats are available: a list of dates of event occurrences or a list of time between occurrences. Each occurrence date is

TABLE 1: EXAMPLE OF DATA EXTRACTED FROM A FAULT TREE DATA BANK

****
200 AREA FAULT TREE DATA STORAGE AND RETRIEVAL SYSTEM
****

NO.	SOURCE	DATE	OCCURRENCE
111111	; ; ; ;07	12-31-60	DUMMY
6535	04	03-13-61	ERRONEOUS CHEMICAL ADDITION, EMULSIFICATION WHEN 1200# OF CARBONATE SOLUTION INADVERTENTLY ADDED TO THE ACID WASHER
2838	01	05-06-61	CATION RESIN MADE UP INSTEAD OF ANION RESIN. FAILED TO UNDERSTAND PROCEDURE .
6359	04	06-28-61	ADJUSTMENT DIFFICULTIES BECAUSE OF POOR SAMPLING.
5848	04	06-30-61	CONCENTRATE FROM MISTAKENLY DUMPED TO PRC FEED. ROSE TO
2797	04	07-13-61	WATER DILUTION OF ERRONEOUSLY OMITTED. SLIGHT PU BUILDUP.
2798	04	08-04-61	USE OF OUT OF SPEC RESIN .
2799	04	11-14-61	DISSOLVED MATERIAL ANALYZED AS ACID DEFICIENT RETURNED TO DISSOLVER AND ACID ADDED . LATER ANALYTICAL REPORTED ORIGINAL ANALYSIS IN ERROR.
6914	04	01-02-62	HIGH ALUMINUM IN MAKEUP WATER. ION BED INSTALLED BUT PLUGGED WITH RESIN FINES
2860	04	01-18-62	CALIBRATION FOR IN ERROR BY 10%. MATERIAL BALANCE DISCREPANCY.
2800	04	03-13-62	LOW ACID ADJUSTMENT, HIGH LOSS DUE TO DIFFICULTY WITH SPECIFIC GRAVITY INSTRUMENTATION.
2491	01	03-30-62	6000 LBS 26% SODIUM NITRATE TO TO - OPERATOR NOT PAYING ATTENTION - ON TOP OF RAW METAL SOLUTION.
2839	01;04	04-14-62	VALENCE STABILIZER NOT ADDED TO FEED. INADEQUATE TRAINING. VALVE NOT OPENED.
2840	01	05-07-62	FERROUS SULFAMATE ADDED TO FRAME PRIOR TO ACID ADJUSTMENT . ACID ADJUSTMENT NOT MADE WHEN REQUIRED AND ERROR MADE IN FESA CALCULATION.
2861	04	06-05-62	ERROR IN UNIDENTIFIED MATERIAL CHARGED TO WOULD NOT DISSOLVE. ZR OR STAINLESS STEEL.
2841	01	06-05-62	FILLED WITH WATER INSTEAD OF 60% NITRIC ACID.
2842	01;04	07-01-62	2190 LBS ULTRASENE AND 1310 LBS OF TBP INADVERTENTLY ADDED TO SOLVENT SYSTEM. VALVING ERROR, PROCEDURAL INADEQUACY.

in MMDDYY format with ten dates in each line sorted in chronological order. An example is given in Table 2. STATPAC converts the date of occurrence to time between occurrences. An example of a data input file for the time between event occurrences is given in Table 3. Two types of formats are allowed, first, twelve floating point numbers per line (12F6.1), with each number right justified in its six character wide field and second, six numbers per line with each number right justified and expressed in E notation (6E12.6). The units for the length of time between occurrences must be the same for all entries in the data file and must be expressed in days or years either as integers or decimal fractions.

TABLE 2: EXAMPLE OF INPUT DATA FILE WITH DATES OF OCCURRENCE

```
123176 012377 012377 012477 020177 020577 020877 021077 021677 022177
022777 022777 022877 030277 030377 030577 031977 042377 040777 041177
041377 041477 041677 041877 042177 042277 042377 032377 042477 042777
050677 050777 050777 050977 051077 060877 061877 062477 062477 062677
070577 070777 070877 071177 072777 072977 072977 073077 082477 102977
103177 111077 112877 012678 032178 041178 051178 051678 051678 060778
070578 071378 072678 072678 082478 091778 091878 092178 102178 102178
110278 110978 112678 121378 122678 122978 122978 010379 010479 012679
020179 020579 021379 030179 030379 030679 031479 031579 032179 052679
080179 080679 080979 082479 082979 093079 112379 120979 010180
```

TABLE 3: EXAMPLE OF INPUT DATA FILE WITH TIME BETWEEN OCCURRENCES

FORMAT 12F6.1

```
  1.0    1.0    1.0    2.0    2.0    3.0    3.0    3.0    4.0    5.0    6.0    7.0
  8.0    8.0    9.0    9.0   10.0   10.0   11.0   12.0   12.0   13.0   13.0   13.0
 23.0   23.1   48.9   48.9   50.0   55.0   70.0   90.5   95.0  100.0  105.0  106.0
107.0  108.0  110.0  111.0  120.0  150.0  151.0  151.0  152.0  153.0  154.0  154.0
155.0  160.0  165.0  166.0  167.0
```

FORMAT 6E12.6

```
1.000000D00 1.000000E00 1.000000E00 2.000000E00 2.000000E00 3.000000E00
3.000000E00 3.000000E00 4.000000E00 5.000000E00 6.000000E00 7.000000E00
8.000000E00 8.000000E00 9.000000E00 9.000000E00 1.000000E01 1.000000E01
1.100000E01 1.100000E01 1.200000E01 1.200000E01 1.300000E01 1.300000E01
1.300000E01 2.300000E01 2.310000E01 4.890000E01 4.890000E01 5.000000E01
5.500000E01 7.000000E01 9.050000E01 9.500000E01 1.000000E02 1.050000E02
1.060000E02 1.070000E02 1.080000E02 1.100000E02 1.110000E02 1.200000E02
1.500000E02 1.510000E02 1.510000E02 1.520000E02 1.530000E02 1.540000E02
1.540000E02 1.550000E02 1.600000E02 1.650000E02 1.660000E02 1.670000E02
```

The options file gives labels and titles to be used with the output graphics, describes the type of data in the data file, number of entries for list of times and dates, format for list of times, the units for reporting (days or years), length of time intervals for graphics, the distributions to be fit to the data, and the plots to be printed. The input file is described in Table 4 with an example in Table 5.

TABLE 4: DESCRIPTION OF OPTIONS INPUT FILE

Line Number	Columns	Variable	Format	Meaning	Comment
1	1	NEOD	I1	Specifies when the Data File was prepared: NEOD = 0: after 6/16/82 NEOD = 1: before 6/16/82	
2	1-8	COMPN	A8	The content of this field is written into the label boxes on the QMS plotter graphs.	Historically, this line was used in writing repair time and failure rate to the Failure Rate Data Bank.
3	1-72	DESCR	72A1	A description of the type of event covered in the Data File. The content of this field is used as a subtitle on the QMS plotter graphs.	
4	1-4	INOPT	I4	Specifies the type of data in the Data File: INOPT=1: A list of times between occurrences INOPT=2: A list of dates of incident INOPT=3: Data extracted from a data bank	In the most common usage, INOPT=3.
	5-8	INUM	I4	If INOPT=1, then this number should be the number of times between occurrences listed in the Data File. If INOPT≠1, then this number should be 1.	
	9-12	IUN	I4	The FORTRAN input unit number used to access the Data File.	Traditionally, this field has a value of 63.
	13-16	IFOR	I4	Specifies the format by which times between occurrences are listed in the Data File: If IFOR=651, then the Data File format is 12F6.1 If IFOR=652, then the Data File format is 6E12.6	Used only when INOPT=1. Appendix D explains both of these formats.
5	1-6	JL	I6	If INOPT=2, then this field should contain the number of dates listed in the Data File. The format for the Data File when INOPT=2 is described in Appendix C.	If INOPT≠2, then this line should not exist, and lines 6 and 7 become lines 5 and 6.
6	1-6	DELT	F6.2	Length of time intervals into which data is grouped for graphing. This field affects the appearance of all graphs which show the observed distribution or the observed cumulative distribution.	For best results, this number should be (time span covered by data in days ÷ number of data entries) x 0.25.
	7-8	UNIT	I2	Specifies the units in which DELT is expressed: If UNIT=1, then DELT is expressed in years. If UNIT≠1, then DELT is expressed in days.	
	9-14	DELFRQ	F6.2	Length of time intervals over which data is averaged for the frequency of occurrences vs. time graphs.	
	15-16	UNIT1	I2	Specifies the units in which DELFRQ is expressed: If UNIT1=1, then DELFRQ is expressed in years. If UNIT1≠1, then DELFRQ is expressed in days.	
	17-22	DELTOT	F6.2	Length of time intervals into which data is grouped for the number of occurrences vs. time graphs.	
	23-24	UNIT2	F6.2	Specifies the units in which DELTOT is expressed: If UNIT2=1, then DELTOT is expressed in years. If UNIT2≠1, then DELTOT is expressed in days.	

TABLE 4: DESCRIPTION OF OPTIONS INPUT FILE (CONTINUED)

Line Number	Columns	Variable	Format	Meaning	Comment
7	1-10	IPLT(I)	5I2	Specifies if the Ith theoretical distribution will be calculated: If IPLT(I)=1, then the Ith distribution is calculated. If IPLT(I)=0, then the Ith distribution is not calculated.	The 1st distribution is Normal. The 2nd distribution is Lognormal. The 3rd distribution is Exponential. The 4th distribution is Weibull. The 5th distribution is Loguniform.
	11-12	IPLT(6)	I2	Specifies if the frequency of occurrences vs. time graphs will be created: If IPLT(6)=1, then the graphs will be created. If IPLT(6)=0, then the graphs will not be created.	In the previous version of STATPAC, this graph did not print, regardless of the value of IPLT(6).
	13-14	IPLT(7)	I2	Specifies if the number of occurrences vs. time graphs will be created: If IPLT(7)=1, then the graphs will be created. If IPLT(7)=0, then the graphs will not be created.	
	15-16	IPLT(8)	I2	Specifies if any of the graphs will be created. If IPLT(8)=1, then the graphs will be created as specified in IPLT(1) through IPLT(7). If IPLT(8)=0, then the no graphs of any kind will be created.	

TABLE 5: EXAMPLE OF OPTIONS INPUT FILE

```
0
CHEMAD
CHEMICAL ADDITION ERRORS
   3      1   63
   23.    0    1.  1    1.    1
1  1  1  1  1  0  1  1
```

ANALYSES

The primary objective of STATPAC is to determine which distribution best fits the time between occurrence data. STATPAC can fit the data to any of five distributions. These are: normal, lognormal, exponential, Weibull, and loguniform. Since the loguniform distribution is rarely used, only the first four distributions will be described. In order to fit a distribution to the data, certain parameters associated with the distribution must be estimated. STATPAC uses the maximum likelihood (ML) method to estimate the distribution parameters from the empirical data. For some distributions, an unbiased function of the ML estimate is used. For example, if $f(x:\theta)$ is the density function for an assumed distribution with parameter vector θ, then the ML estimator of θ is the vector which maximizes either $\prod f(x_i:\theta)$ or $\log[\prod f(x_i;\theta)]$, where $\{x_1, x_2, \ldots, x_n\}$ are the data values which are assumed to be a random sample of values from the distribution. Most ML estimators have good statistical properties at least for "large" samples and usually also for small samples. The ML estimator is consistent (converges in probability, as the sample size increases, to the parameter it is estimating), asymptotically unbiased, and has an asymptotic normal distribution. Among the five distributions that can be fit using STATPAC, only the ML estimator for the Weibull parameters cannot be expressed as a closed mathematical form.

The distributions, ML estimators and the STATPAC estimators for the four distributions: exponential, normal, lognormal and Weibull are given as well as estimates of the mean, median, standard deviation, and error factor.

EXPONENTIAL:

The probability density function for outcome $y \geq 0$ is given by

$$f(y) = (1 \div \theta)\exp(-y \div \theta),\ y \geq 0,$$

where the parameter θ is called the distribution mean which must be positive and expressed in the same units as y. This density can also be written as

$$f(y) = \lambda\exp(-\lambda y),\ y \geq 0,$$

where the parameter $\lambda = 1 \div \theta$ is called the failure rate.

The sample log likelihood is given by

$$L(\theta) = \Sigma[-\ln(\theta) - (y_i \div \theta)].$$

The maximum likelihood estimate for θ is

$$\hat{\theta} = \Sigma y_i \div n,$$

where n is the number of failures. This estimate is unbiased and is the one used by STATPAC.

The mean and standard deviation of the distribution are both estimated by $\hat{\theta}$. The median of the distribution is estimated by

$$\text{median} = -\hat{\theta}\ln(0.5).$$

The error factor, EF, is defined as

$$EF = [\ln(0.05) \div \ln(0.95)]^{1/2}.$$

NORMAL:

The normal probability density is

$$f(y) = (1 \div \sigma)(2\pi)^{-1/2} \exp[-(y - \mu)^2 \div (2\sigma^2)],\ -\infty < y < \infty.$$

The sample log likelihood is given by

$$L(\mu,\sigma) = -(n \div 2)[\ln(2\pi)] - n\ln(\sigma) - \Sigma(y_i - \mu)^2 \div (2\sigma^2)].$$

The maximim likelihood estimates for μ and σ are given by

$\hat{\mu} = \Sigma y_i \div n$ and

$\tilde{\sigma} = [\Sigma(y_i - \hat{\mu})^2 \div n]^{1/2}$ which is a biased estimate. The unbiased estimate is

$\hat{\sigma} = [n\tilde{\sigma}^2 \div (n - 1)]^{1/2} = [\Sigma(y_i - \hat{\mu})^2 \div (n - 1)]^{1/2}$ which is used by STATPAC. $\hat{\mu}$ and $\hat{\sigma}$ are the estimates of the distribution mean and standard deviation. The median is estimated from $\hat{\mu}$ also.

The error factor is defined as

$EF = X(0.95) - X(0.05)$,

where $F[X(0.95)] = 0.95$ and $F[X(0.05)] = 0.05$ as calculated from the distribution after estimating the mean and standard deviation.

LOGNORMAL:

$$f(y) = (1 \div \sigma y)(2\pi)^{-1/2} \exp[-(\ln(y) - \mu)^2 \div (2\sigma^2)], \quad -0 < y < \infty,$$

where μ is the log mean and σ is the log standard deviation. These, like y, are dimensionless numbers.

The log likelihood function is

$$L(\mu,\sigma) = -(n \div 2)[\ln(2\pi)] - (n)\ln(\sigma) - \Sigma\ln(y_i) - \Sigma(\ln(y_i) - \mu)^2 \div (2\sigma^2)].$$

The maximum likelihood estimates in log units are

$\hat{\mu}_L = \Sigma\ln(y_i) \div n$ and

$\tilde{\sigma}_L = [\Sigma(\ln(y_i) - \hat{\mu}_L)^2 \div n]^{1/2}$, a biased estimate. The unbiased estimate is

$\hat{\sigma}_L = [n\tilde{\sigma}^2 \div (n - 1)]^{1/2} = [\Sigma(\ln(y_i) - \hat{\mu}_L)^2 \div (n - 1)]^{1/2}$, which is used by STATPAC.

These are in the log space *not* linear space.

The mean, median and standard deviation of the distribution are estimated by

$\text{mean} = \exp[\hat{\mu}_L + (\hat{\sigma}_L^2) \div 2]$,

$\text{median} = \exp[\hat{\mu}_L]$, and

$\text{standard deviation} = \exp(\hat{\mu}_L)\{\exp(2\hat{\sigma}_L^2) - \exp(\hat{\sigma}_L^2)\}^{1/2}$.

The error factor, EF, is defined as

$$EF = \exp [1.645\ \hat{\sigma}_L].$$

WEIBULL:

The probability density function is given by

$$f(y) = (\beta \div \alpha) y^{(\beta-1)} \exp [-(y)^{\beta} \div \alpha], \text{ where } y > 0.$$

An alternate form is $\alpha = (\alpha')^{\beta}$.

The log likelihood function is

$$L(\alpha,\beta) = n\ln(\beta) - n\ln(\alpha) + (\beta-1)\Sigma\ln(y_i) - \Sigma(y_i)^{\beta} \div \alpha.$$

The maximum likelihood estimates can not be expressed in a closed mathematical form but can be estimated by solving the following two equations:

$$\frac{\partial(L(\alpha,\beta))}{\partial\alpha} = -n \div \alpha + (\Sigma y_i^{\beta}) \div \alpha^2 = 0 \text{ and}$$

$$\frac{\partial(L(\alpha,\beta))}{\partial\beta} = (n \div \beta) + \Sigma\ln(y_i) - \Sigma(y_i)^{\beta} \ln(y_i) \div \alpha = 0.$$

The solutions to the ML equations can be shown to satisfy the following equations:

$$\hat{\alpha} = \{\Sigma y_i^{\hat{\beta}} \div n\} \text{ and}$$

$$\hat{\beta} = n\{[\Sigma y_i^{\hat{\beta}} \ln(y_i) \div \hat{\alpha}] - \Sigma\ln(y_i)\}^{-1}.$$

STATPAC solves the above equations iteratively for $\hat{\alpha}$ and $\hat{\beta}$.

The mean, median, and standard deviation of the distribution are estimated by

$$\text{mean} = (\hat{\alpha})^{1/\hat{\beta}} \Gamma(1+1 \div \hat{\beta}),$$

$$\text{median} = (\hat{\alpha})^{1/\hat{\beta}} (-\ln(0.5))^{1/\hat{\beta}}, \text{ and}$$

$$\text{standard deviation} = (\hat{\alpha})^{1/\hat{\beta}} \{\Gamma(1+2 \div \hat{\beta}) - (\Gamma(1+1 \div \hat{\beta}))^2\}^{1/2}.$$

The error factor, EF, is defined as

$$EF = \{(\ln(0.05)/\ln(0.95)]^{1/\hat{\beta}}\}^{1/2}.$$

GOODNESS OF FIT TESTS

Once the different distributions have been fit to the data, the problem is to determine which distribution best represents the empirical data. This is referred to as hypothesis testing with the estimated parameters determining the true distribution;

$$H_o: F(x) = F_o(x),$$

where $F_o(x)$ is the distribution function estimated from the data. The problem of testing is known as a goodness-of-fit problem. Any test of the null hypothesis is called a test of fit. Simple hypothesis are when $F_o(x)$ is completely specified; e.g., the hypothesis that the n observations have come from a normal distribution with specified mean and variance. A composite hypothesis assumes the parameters are estimated from the data. STATPAC computes three "goodness-of-fit" statistics which can be used to determine which of the five distributions best fits the data.

CHI-SQUARE STATISTIC

The range of the variate y is arbitrarily divided into k mutually exclusive classes. Then, since $F_o(y)$ is specified, the probability of an observation falling in each class can be calculated. If these are denoted by p_{oi}, i = 1,2, ..., k and the observed frequencies in the k classes by n_i ($\sum n_i = n$), the n_i are multinomially distributed. The classical Chi-square statistic is defined as

$$X^2 = \sum(n_i - np_{oi})^2/(np_{oi}),$$

with degrees of freedom equal to its rank, k-1.

H_o is rejected when X^2 is large (upper-tail test). However, since only rarely are the distribution and parameters specified in advance rather than estimated from the data, the effect of estimating the unknown parameters on the asymptotic distribution of the X^2 statistic must be considered. When the parameters are estimated using ML estimates based on the n observations (and not the k intervals), the X^2 statistic no longer has an asymptotic Chi-square distribution. The distribution of X^2 is bounded between a Chi-square with (k-1) degrees of freedom and a Chi-square with (k-s-1) degrees of freedom where s is the number of parameters estimated. There is a partial recovery of the s degrees of freedom when the ML parameters are estimated from all the data. As k becomes large, these are so close together that the difference can be ignored. For small k, the effect of using the Chi-Square with (k-s-1) degrees of freedom distribution for test purposes may lead to serious error. When ordinary ML estimation is used, X^2 should exceed both critical values of $X^2(k-s-1)$ and $X^2(k-1)$ before rejecting. The Chi-square test assumes that the k classes were determined without reference to the observations. However, it is common practice to determine the class boundaries and sometimes even to determine k itself, after examining the data. STATPAC determines the number of classes based on the number of observations in the data set and estimates the class boundaries using the equal-probabilities method. This rule was suggested by Mann and Wald (1942) and Gumbel (1943); given k, choose the classes so that the hypothetical probabilities, p_{oi}, are all equal to 1/k. The equal-probabilities method may result in loss of sensitivity at the extremes unless k is rather large.

KOLMOGOROV'S D_n

Kolmogorov's D_n test is based on the cumulative distribution of the sample defined by

$$S_n(x) \begin{cases} = 0 & x < x(1) \\ = r/n & x(r) \le x < x(r+1), \\ = 1 & x(n) \le x, \end{cases}$$

where the x(r) are the order statistics; i.e. $x(1) \le x(2) \le \ldots \le x(n)$. $S_n(x)$ is simply the proportion of the observations not exceeding x. If $F_o(x)$ is the true distribution function, fully specified, from which the observations come, then

$$\lim P\{S_n(x) - F_o(x)\} \text{ as } n \to \infty = 1.$$

One sided forms of the Kolmogorov test can be defined as

$$D^+_n = \sup \{S_n(x) - F_o(x)\}, \text{ and}$$

$$D^-_n = \sup \{F_o(x) - S_n(x)\}.$$

The maximum absolute difference is

$$D_n = \sup|S_n(x)-F_o(x)| = \max(D^+_n, D^-_n).$$

STATPAC computes D^+_n, D^-_n, and D_n and computes the probability of accepting the proposed distribution based on a normal approximation.

C. A. Williams (1950) and Massey (1951) compared the values of D for which the large-sample powers of the X^2 and D_n tests are at least 0.5. For test size $\alpha = 0.05$, the D_n test can detect with power 0.5 a D about half the magnitude of that which the X^2 test can detect with this power; even with n = 200. D_n is a much more sensitive test for the fit of a continuous distribution than the X^2 test. For the composite hypothesis with unspecified parameters, Kolmogorov-type tests were investigated by Durbin (1975). Durbin tabulated the percentage points of D^+_n, D^-_n, and D_n up to n = 100.

For testing normality (normal and lognormal distributions), Shapiro and Wilk give a test based on the regression of the order-statistics upon their expected values. Shapiro et al. (1968) and Stephens (1974) make power comparisons from extensive sampling experiments and show that W is usually superior to most other tests when the distribution is normal or lognormal. In addition, Stephens (1974) compared two other tests which can be used for both normal (lognormal) and exponential (Weibull) distributions. These are the Anderson-Darling statistic, A^2, and the Cramer-von Mises statistic, W^2. Stephens concluded when the distribution is the normal or lognormal with parameters estimated from the data, A^2 and W^2 do a better job the Kolmogorov- D and Chi-Square and have powers roughly comparable with the Wilk-Shapiro statistic.

SUM OF SQUARED DEVIATIONS

STATPAC uses as its primary selection criteria the average of the squared deviations (ASD) between the fitted cumulative distribution and the empirical data normalized to 1.0.

Let $F(Y_i)_j$ be one of the four cumulative distributions under consideration, then

$$ASD^2_j = \{\Sigma [F(Y_i)_j - (i \div n)]^2\} \div n.$$

STATPAC selects the distribution with the smallest ASD^2_j as the "best" fitting distribution among the four candidates.

OUTPUT

The following descriptive statistics and plots are available from STATPAC to graphically determine the optimum distribution for the data. STATPAC prints the data in time between each incident and total time to the incident. In addition, the following statistics are computed on the actual data before fitting any distributions: mean, median, standard deviation, and the number of data values. The following option values are printed: step size, units (days or years), and the time dependent frequency distribution averaging time.

For each distribution the user specifies to be fit to the data, STATPAC prints the parameter estimates, mean, median, standard deviation, error factor, the results of the Kolmogorov-D one-sample test for both D, D^+, and D^-, the normalized Z statistic and the probabilities of exceeding Z, the actual counts observed in each of the equiprobable intervals for the Chi-square, and each of the components, $[(n_i - p_{oi})^2/p_{oi}]$, the Chi-square sum, degrees of freedom, probability of exceeding the computed X^2, and ASD. The distribution with the smallest ASD is printed as the "Best" distribution. Table 6 gives the above output for the normal distribution and Table 7 for the Weibull distribution.

Plots are available for both the line printer and plotter. The line printer is convenient but the plotter gives clearer graphics. Examples of the available plots are shown in Figures 1-4. Figures 1 and 2 show the frequency of time between occurrences with the fitted distribution overlaid on the actual data for the normal and Weibull distributions respectively. Figure 3 shows the number of occurrences over the observation period with an estimated slope for the empirical data. Figure 4 shows the empirical cumulative probability distribution.

COMPARISON WITH SAS

Several data sets were analyzed by both STATPAC and SAS® QC Procedure Capability. SAS® QC uses ML methodology to estimate the parameters for the following distributions: normal, lognormal, exponential, Weibull, gamma, and beta. SAS provides estimates of the parameters, mean, median, standard deviation, and 9 percentiles (1%, 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99%), the Chi-square goodness-of-fit statistic (with degrees of freedom and probability of exceeding the statistic). For the normal, lognormal, exponential and Weibull, both the Anderson-Darling and the Cramer-von Mises goodness of fit statistics are computed with probabilities of exceeding the computed value. The Kolmogorov-D (but not D+ and D-) are computed for the normal, lognormal and exponential. For the normal and lognormal, the Wilk-Shapiro statistic is computed with probability of getting a smaller value. In addition, the SAS®procedure produces plots of the cumulative distribution with the observed data, and a histogram with the probability density overlaid.

Both SAS and STATPAC computed the same estimates of mean, median, standard deviation, and percentiles for each of the four distributions: normal, lognormal, exponential

TABLE 6: OUTPUT STATISTICS FOR THE NORMAL DISTRIBUTION

***** NORMAL DISTRIBUTION PARAMETERS *****
Mean = 96.914 Sig squared = 14851.699

*** KOLMOGOROV-SMIRNOV ONE-SAMPLE TEST ***
D (Max) = 0.21395 D (Plus) = 0.20377 D (Minus) = 0.21395
Statistic (Z) used to obtain probabilities = 2.11802
Probability of statistic exceeding Z (one-sided) = 0.12690e-03
Probability of statistic exceeding Z (two-sided) = 0.00025

*** CHI-SQUARE TEST ***
*** Counts of observations which fall into 8 equiprobable categories
0.0000E+00 0.2500E+02 0.3100E+02 0.6000E+01 0.1100E+02 0.7000E+01
0.6000E+01 0.1200E+02
*** Components of Chi-Square Statistic ***
0.1225E+02 0.1327E+02 0.2870E+02 0.3189E+01 0.1276E+00 0.2250E+01
0.3188E+01 0.5091E-02
Chi-Square Statistic (CS) = 0.630E+02 Degrees of Freedom = 5
Probability (Q) of Chi-Square Statistic exceeding CS = 0.0000E+00.

Range of dates	:	12-31-60 to 1- 1-87
Description	:	CHEMICAL ADDITION ERRORS
Component Name	:	CHEMAD
# of Incidents	:	98
** DISTRIBUTION**	:	NR
STATPAC-2	:	1.4645E-02
CHI-SQUARE	:	6.2979E+01
MEDIAN	:	9.6914E+01
SIGMA	:	1.2187E+02
MEAN	:	9.6914E+01
ERROR FACTOR	:	1.9986E+02

TABLE 7: OUTPUT STATISTICS FOR THE WEIBULL DISTRIBUTION

***** WEIBULL DISTRIBUTION PARAMETERS *****
Eta for Weibull = 0.738E+00 Sig for Weibull = 0.255E+02 Eps (zero) = 0.000E+00

*** KOLMOGOROV-SMIRNOV ONE-SAMPLE TEST ***
D (Max) = 0.09886 D (Plus) = 0.09886 D (Minus) = 0.05556
Statistic (Z) used to obtain probabilities = 0.97870
Probability of statistic exceeding Z (one-sided) = 0.14724E+00
Probability of statistic exceeding Z (two-sided) = 0.29448

*** CHI-SQUARE TEST ***
*** Counts of observations which fall into 8 equiprobable categories
0.1000E+02 0.1500E+02 0.2000E+02 0.3000E+01 0.1200E+02 0.1100E+02
0.1100E+02 0.1600E+02
*** Components of Chi-Square Statistic ***
0.4133E+00 0.6174E+00 0.4903E+01 0.6985E+01 0.5103E-02 0.1276E+00
0.1276E+00 0.1148E+01
Chi-Squared Statistic (CS) = 0.143E+02 Degrees of Freedom = 5
Probability (Q) of Chi-Square Statistic Exceeding CS = 0.13666E-01.

TABLE 7: OUTPUT STATISTICS FOR THE WEIBULL DISTRIBUTION (CONTINUED)

Range of Dates	:	12-31-60 to 1- 1-87
Description	:	CHEMICAL ADDITION ERRORS
Component Name	:	CHEMAD
# of Incidents	:	98
** DISTRIBUTION**	:	WI
STATPAC-2	:	1.1478E-03
CHI-SQUARE	:	1.4327E+01
MEAN	:	9.7103E+01
MEDIAN	:	4.8958E+01
SIGMA	:	1.3377E+02
X-RMS	:	1.0744E+02
ERROR FACTOR	:	1.5737E+01

and Weibull. Only two goodness-of-fit tests are common between the two packages. These are the Chi-Square and the Kolmogorov-D. Chi-Square is very dependent on the number of intervals and the boundary points for the intervals. The two software packages compute the Chi-Square differently with different probabilities of accepting the distribution as true particularly for the lognormal distribution. For the examples compared, STATPAC accepted the lognormal while SAS rejected the lognormal at the 0.05 level.

Both packages gave the same estimate for the Kolmogorov-D statistic; however for the comparison examples, STATPAC gave significantly different probabilities of accepting the distribution than SAS. SAS does not calculate the Kolmogorov-D statistic for the Weibull. The STATPAC average of squared deviations (ASD) agrees the best with the SAS Cramer-von Mises and the Anderson-Darling statistics. The disadvantage with the STATPAC ASD test is that no probabilities are computed. The results of the goodness-of-fit tests should be evaluated by a statistician whether using SAS or STATPAC. Comparison of STATPAC with other commercial risk analyses software is continuing as well as the evaluation of an optimum goodness-of-fit test.

SUMMARY

STATPAC appears to correctly fit a normal, lognormal, exponential, and Weibull distribution to time between occurrence data when compared with the well-established commercial software package SAS®. However, there is disagreement between the two software packages in determining the probabilities of accepting the proposed distribution as "best". The Chi-square statistic is dependent on the number and boundary values for the intervals so the two might be expected to disagree. The degrees of freedom depend on whether the parameters are estimated from the intervals or from the entire data. SAS correctly computes the degrees of freedom while STATPAC does not. The method used by STATPAC for calculating the probability of accepting the distribution based on the Kolmogorov-D statistic gives significantly different results than obtained using SAS. The STATPAC average of squared deviations (ASD) agrees the closest with the SAS Cramer-von Mises and the Anderson-Darling statistics. A disadvantage with the STATPAC ASD statistic is the probability of getting a larger value is not computed. Most STATPAC users choose the "best" distribution based on the ASD. Additional commercial software comparisons with STATPAC are planned as well as determining the optimum goodness-of-fit test.

REFERENCES

Dubovsky, P. J. (1990) Savannah River Site Memo WSRC-TR-90-481, "STATPAC-3, a VM Version of the STATPAC Statistical Package'.

Durbin, J. (1975). "Kolmogorov-Smirnov Tests when Parameters are Estimated with Application to Tests of Exponentiality and Tests on Spacing. Biometrika, 62, p. 5.

Gumbel, (1943). "On the Reliability of the Classical Chi-Square Test." Ann. Math. Statist., 14, p. 253.

Hsu, Y. S. & Huang, J. C. (1983). Savannah River Site Memo, DPST-83-793, "STATPAC-2: A New Version of STATPAC".

Mann, Nancy R., Schafer, Ray. E. and Singpurwalla, Nozer D. (1974). Methods for Statistical Analysis of Reliability and Life Data, John Wiley & Sons, N. Y.

Mann, H. B., & Wald, A. (1942). "On the Choice of the Number of Intervals in the Application of the Chi-Square Test." Ann. Math. Statist., 13, p. 306.

Massey, F. J., Jr. (1951). "The Kolmogorov-Smirnov Test of Goodness of Fit", J. Amer. Statist. Ass., 46, p. 68.

Nelson, Wayne (1982). Applied Life Data Analysis, John Wiley & Sons, New York, pp 634.

SAS QC manual, SAS Institute, Inc. Cary, NC Version , 1st Edition,

Shapiro, SS., Wilk, M. B. & Chen, H. J. (1968). " A Comparative Study of Various Tests for Normality". Amer. Statist. Ass. 63, p. 1343.

Stephens, M. A. (1974). "EDF Statistics for Goodness of Fit and Some Comparisons", J. Amer. Statist. Ass., 69, No. 347. p. 730-737.

Stuart, Alan & Ord, J. Keith (1991): Kendall's Advanced Theory of Statistics, Vol. 2, 5th Edition, Oxford University Press, N.Y.

Williams, C. A., Jr. (1950). "On the Choice of the Number and Width of Classes for the Chi-Square Goodness of Fit", J. Amer. Statist. Ass., 45, p. 71.

FIGURE 1: FREQUENCY OF TIME BETWEEN OCCURRENCES
NORMAL DISTRIBUTION

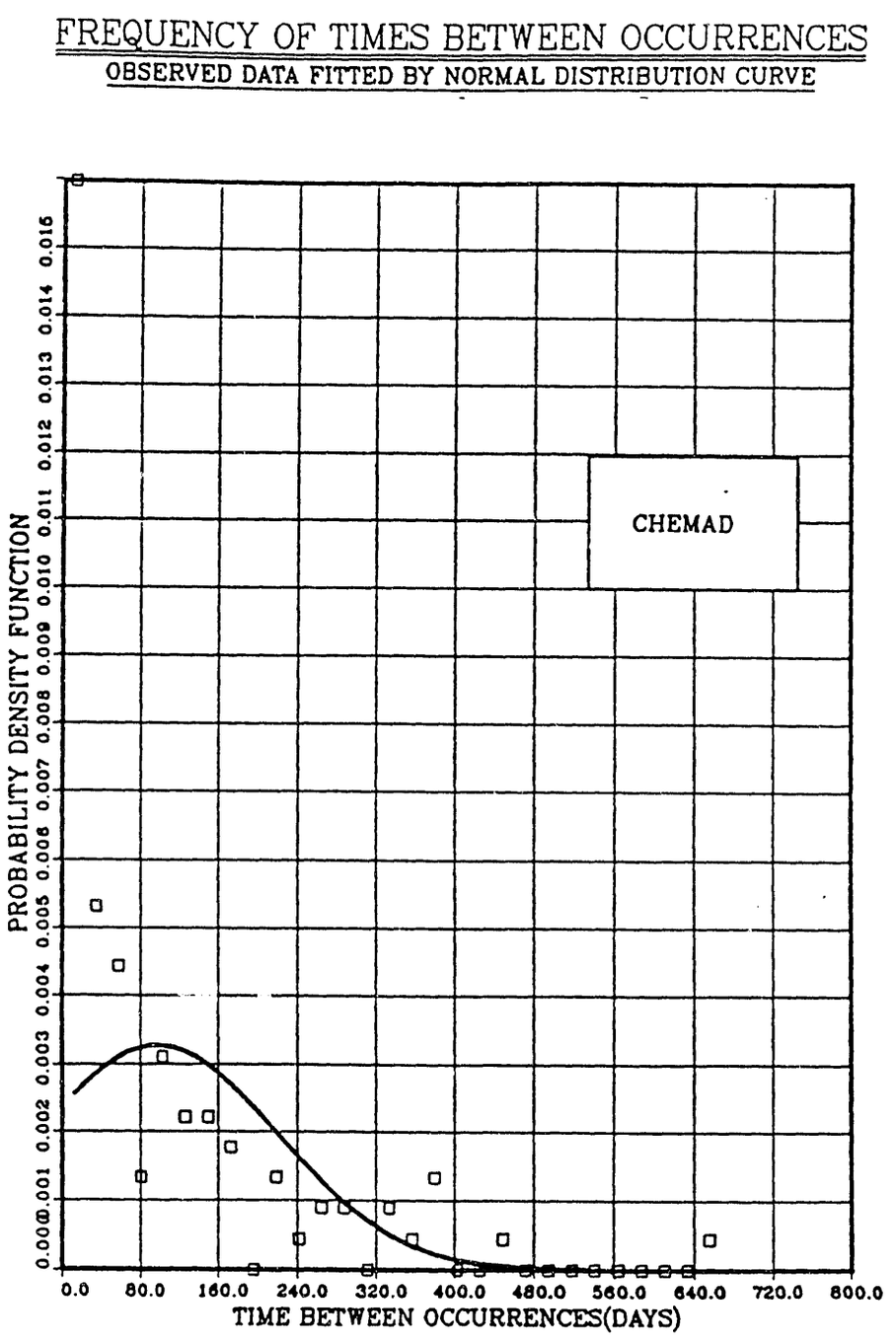


FIGURE 2: FREQUENCY OF TIME BETWEEN OCCURRENCES
WEIBULL DISTRIBUTION

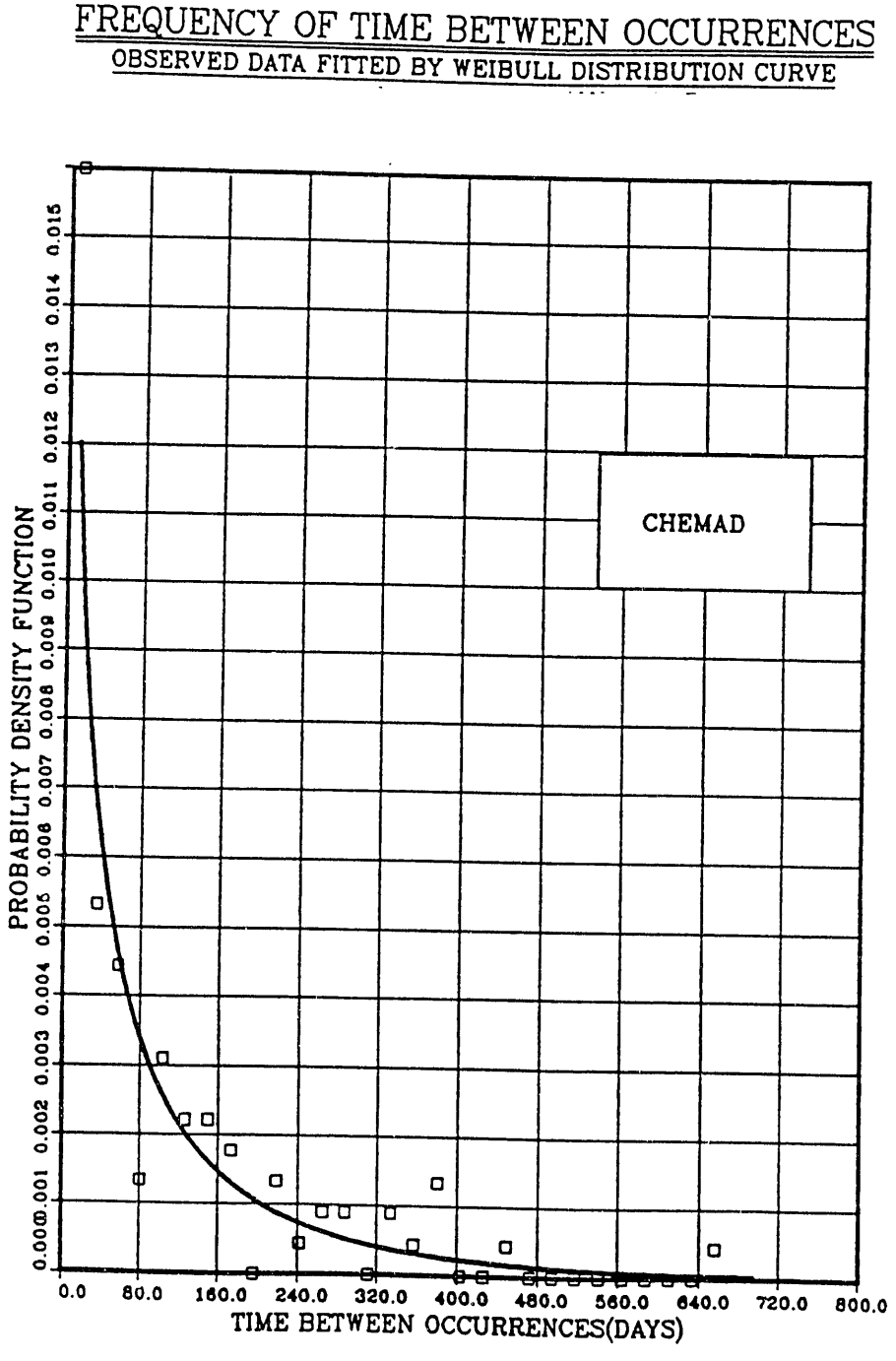


FIGURE 3: NUMBER OF OCCURRENCES OVER THE OBSERVATION PERIOD

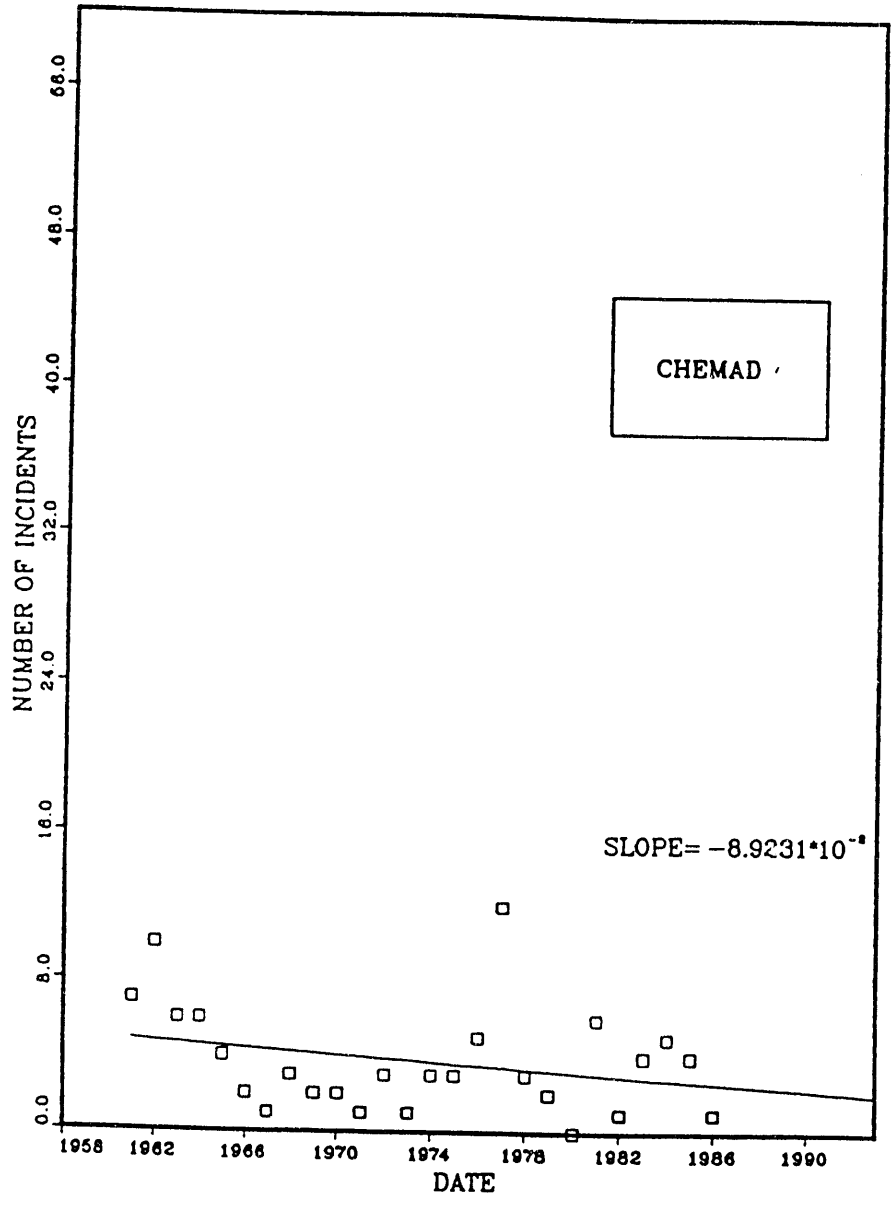
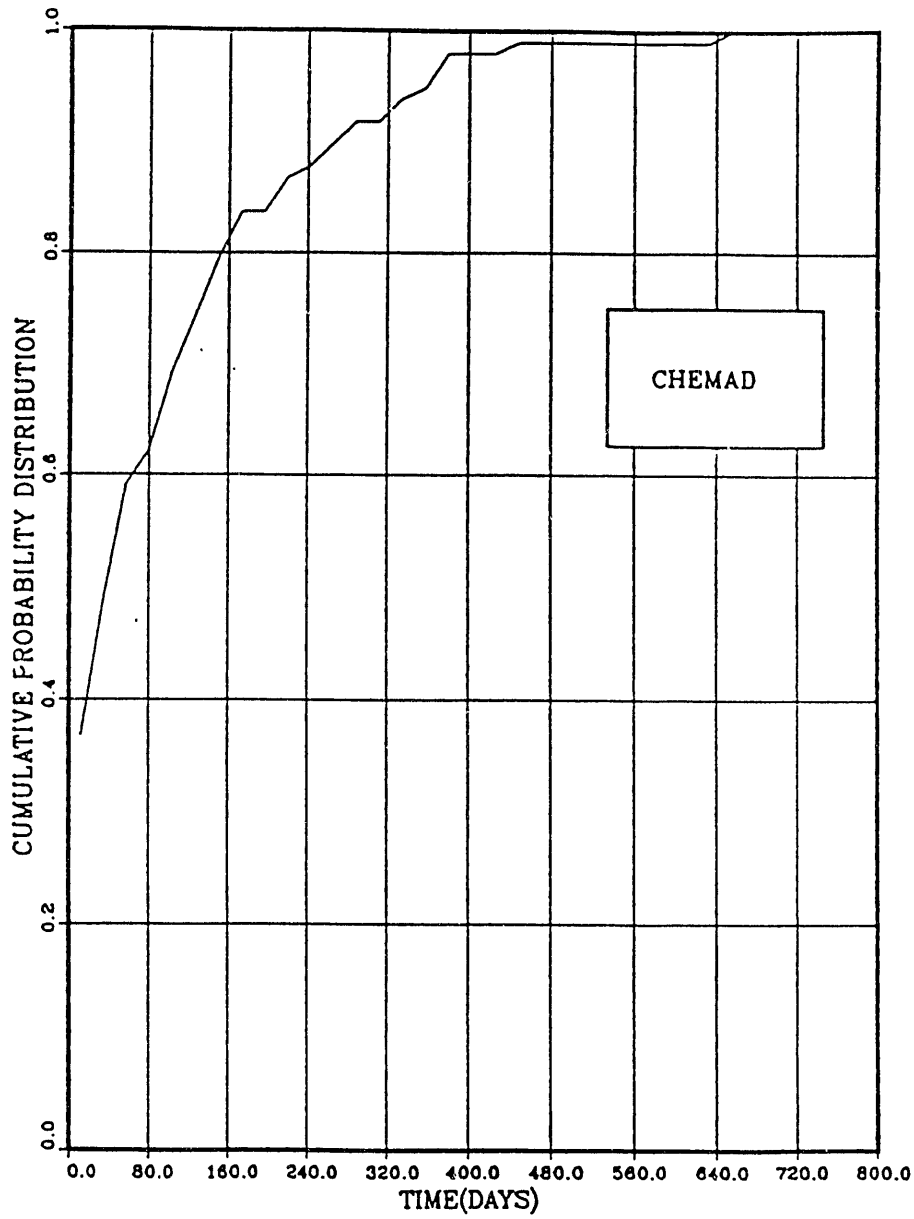


FIGURE 4: THE EMPIRICAL CUMULATIVE PROBABILITY DISTRIBUTIONS



DATE FILMED

4/19/93