01 Jan 1982

Stochastic Modeling of Individual Resource Consumption during the Programming Phase of Software Development

Daniel G. McNicholl

Kenneth Magel

Submitted to IEEE Transactions on Software Engineering

STOCHASTIC MODELING OF INDIVIDUAL RESOURCE CONSUMPTION DURING THE PROGRAMMING PHASE OF SOFTWARE DEVELOPMENT

Daniel G. McNicholl[1] and Kenneth Magel[1]

CSc-82/1

Department of Computer Science

University of Missouri-Rolla

Rolla, Mo. 65401 (314)341-4491

[1]Work supported in part by NSF Grant MCS 8002667

ABSTRACT

In the past several years there has been a considerable amount of research effort devoted to developing models of individual resource consumption during the software development process. Since many conditions affect individual resource consumption during the software development process, including several which are difficult if not impossible to quantify, it is our contention that a stochastic model is more appropriate than a deterministic model.

In order to test our hypothesis we conducted an experiment based upon several student programming assignments. Data from this experiment is used to demonstrate that the two parameter Log-Normal distribution is appropriate for describing the probabilistic behavior of the random variable 'resource consumption'. In addition we present a theoretic argument for the applicability of the Log-Normal distribution based on the concept of a proportional effects model for the growth of a program.

TABLE OF CONTENTS

I - INTRODUCTION.

There is a consensus among software engineers [5], [11], [12] that one of the more important aspects of software development management is the prediction of resource consumption during the software development process. Given this impetus there has been a marked increase in research efforts during the past several years towards developing accurate predictive models of individual resource consumption with only moderate degrees of success. Almost without exception these efforts have used deterministic models of the software development process. We believe the underlying nature of software development is inherently stochastic. The success of a such a stochastic predictive model of software development will depend on two factors: (1) the adequacy of the distribution chosen to explain the probabilistic behavior of the process; and (2) the ability of the model to determine the parameters of the distribution for a particular development. Only the first of these two criteria will be addressed in this paper.

Since we want to develop precise rather than qualitative results, we have concentrated on a single phase of the software development process, q.e. the Programming Phase. Although historically this phase has only accounted for approximately 10 to 20 percent of the total resources

consumed during a software development process, it is a natural starting point since it is one of the most mechanical and therefore measurable phases of the process. Given the somewhat loose use of terminology within the software engineering discipline we state here our working definition of the 'programming phase' to avoid confusion.

> The PROGRAMMING PHASE of a software development process is identified as that task which: 1) has as it's inputs a complete description of the developing program's inputs, outputs and the functional dependencies of the output on the input; and 2) produces a correct program in a machine readable language.

The immediate goal of our investigation was to determine the probabilistic behavior, as realized by a probability distribution, of various resource consumption quantities during the programming phase of software development. Our approach was to devise an experiment which would allow accurate measurements of all the relevant conditions and outcomes; see Section II. From the collected data evidence is shown in Section III to justify our premise that a stochastic model is needed. From the resultant experimental data we then tested the goodness-of-fit of various postulated theoretic distributions as described in Section IV. Equally important to us was the derivation of a theoretic rationale for the empirical results; this is presented in Section V. In Section VI our results are

summarized and follow-up research is outlined.

II - DESCRIPTION OF THE EXPERIMENT.

Sixty students in two sections (same instructor) of a college sophomore - level class on PL/I were given six written programming assignments. These students had taken two previous computer science courses using Fortran and one using assembly language. The assignments, which were relatively simple, contained a complete description of the program's inputs and outputs as well as the functional dependencies of the output on the input. The resources which the students consumed were collected via two mechanisms. During the students' attempts to solve the assignments certain data were automatically collected by the computer's accounting system. When processed this data yielded three resource usage meters:

- #RUNS - The total number of computer runs made.
- CPU-TIME - The total cpu time, as measured in seconds, used to develop the program.
- CPU-COST - The total cpu charges, as measured in dollars, incurred during the assignment.

In addition, after the assignment was finished the students completed a questionnaire on each assignment from which the following resource usage meter was extracted.

- #HOURS - The total number of hours of expended effort.

Information on the individual characteristics of the students was collected at the beginning of the course through a written survey. Tables I, II and III present some simple univariate descriptive statistics of the data collected from the experiment. The number of PL/I statements in the resultant programs are summarized in Table IV to give some idea of the size of each assignment.

RESOURCE METERS	PROGRAM IDS	SAMPLE STATISTICS		
		SIZE	MEAN	STD DEV
#HOURS	TRI	56/53	8.0/ 7.1	7.2/ 5.7
	RKT	50/46	13.7/14.0	8.4/ 8.3
	INS	42/35	18.6/19.0	17.7/19.1
	SRT	44/42	10.8/11.1	10.0/10.2
	MST	37/34	16.3/16.5	15.2/15.9
	MML	40/32	21.6/22.3	19.0/21.0
#RUNS	TRI	59/56	13.9/13.7	7.4/ 7.4
	RKT	53/48	20.9/21.3	16.9/17.5
	INS	53/41	29.4/28.7	21.5/20.5
	SRT	52/48	21.1/21.6	18.1/18.4
	MST	48/40	24.5/25.2	17.5/18.2
	MML	48/37	28.0/26.6	21.0/15.8
CPU-TIME	TRI	59/56	3.6/ 3.5	1.9/ 1.8
	RKT	53/48	8.4/ 8.4	6.7/ 6.7
	INS	53/41	13.5/13.6	13.3/13.8
	SRT	52/48	13.0/13.6	13.2/13.4
	MST	48/40	22.4/23.5	22.4/23.9
	MML	48/37	21.3/20.3	19.5/12.8
CPU-COST	TRI	59/56	0.7/ 0.6	0.4/ 0.4
	RKT	53/48	1.9/ 2.0	1.8/ 1.8
	INS	53/41	3.0/ 3.0	2.5/ 2.5
	SRT	52/48	1.4/ 1.5	1.2/ 1.2
	MST	48/40	2.7/ 2.8	2.7/ 2.9
	MML	48/37	3.7/ 3.8	3.0/ 2.9

TABLE I - Descriptive Statistics of the Resource Usage Data: Where two values are given the first value corresponds to the statistic for all students who attempted the assignment, the second applies to only those students who successfully completed the assignment.

INDIVIDUAL CHARACTERISTIC	SAMPLE STATISTICS			
	SIZE	MEAN	MEDIAN	STD DEV
EDLEVEL	60	67.7	63	29.9
EDLOAD	60	14.9	15	2.86
EMPLOAD	60	4.17	0	10.5
CAMPEXP	60	3.82	3	3.26
DEPTEXP	60	3.57	3	2.54
WORKEXP	60	21.0	11	28.1
DPEXP	60	4.27	0	7.64
CGPA	59	2.95	2.9	0.6
AGE	60	22.4	21.1	3.75

(A)

INDIVIDUAL CHARACTERISTIC	DISTRIBUTION OF RESPONSES(%)					
	VERY HIGH	HIGH	GOOD	FAIR	LOW	NONE
INTEREST	45.8	40.7	8.5	3.4	0.0	1.7
ABILITY	3.4	33.9	49.2	13.6	0.0	0.0
ENJOYMENT	22.0	44.1	20.3	13.6	0.0	0.0

(B)

TABLE II - Descriptive Statistics of the Individual Characteristic Data: See Table III for definitions of Individual Characteristics.

GLOSSARY OF INDIVIDUAL CHARACTERISTICS	
EDLEVEL	The number of college credit hours completed prior to the course.
EDLOAD	The number of credit hours the student was taking simultaneously with the PL/I course.
EMPLOAD	The number of hours per week that a student works at a job.
CAMPEXP	The number of semesters completed at the campus.
DEPTEXP	The number of CSC courses taken at the campus.
WORKEXP	The number of full-time months , or equivalent for part-time, worked at any job since graduating from high school.
DPEXP	Same as WORKEXP except for DP related jobs only.
CGPA	Cumulative Grade Point Average.
AGE	Student's age in years at beginning of the course.
INTEREST	The student's rating of his interest in Computer Science.
ABILITY	The student's rating of his ability as a programmer.
ENJOYMENT	The student's rating of his enjoyment of programming.

TABLE III - Definition of Individual Characteristics.

PROGRAM	MEAN	STD DEV	LO	MED	HI
TRI	22	8.0	14	20	71
RKT	68	15.5	49	66	134
INS	73	16.5	51	70	127
SRT	47	10.8	34	45	96
MST	85	15.1	61	82	125
MML	116	24.9	73	112	181

TABLE IV - Descriptive Statistics of the Number of Statements Program Size Meter: includes all types of PL/I statements.

III - JUSTIFICATION FOR THE STOCHASTIC MODEL.

If we assume for the moment that there exists a deterministic model of individual resource consumption during the programming phase then there must exist a function g_x such that:

$$R_x = g_x(Z_1, Z_2, \ldots, Z_n)$$

where:

- R_x is the amount of resource X consumed by a specific individual on a specific programming assignment.
- $Z_1, Z_2, \ldots, Z_n$ are meters of all the relevant sources of variation in the programming process, e.g. programmer, program assignment, environment, etc..

In the experiment described in the previous section there are only three potential sources of variation: the programming assignment, the programmer, and the sequence of assignments. We might then hypothesize that there must exist a function g_x such that:

$$R_{x,ij} = g_x(I_i, P_j, S_j) \qquad \text{[EQ1]}$$

where:

- $R_{x,ij}$ is the amount of resource X consumed by individual i on programming assignment j.
- I_i is some set of characteristic measurements for individual i which account for the variation among individuals' resource usage.

- P_j is some set of complexity/size metrics for programming assignment j which will account for the variation of resource usage between assignments.

- S_j is the sequence of programming assignment j which possibly could account for some additional inter-program variation.

Contrasted to the deterministic model represented by equation 1 is a stochastic model in which we hypothesize that there are so many unmeasurable conditions affecting an individual's resource usage that we can not completely determine it. Therefore we settle for describing the probabilistic behavior of the resource consumption in terms of it's probability density (p.d.f.) or cumulative distribution functions (c.d.f):

$$\text{e.g. } R_{x/j} \sim f(r_{x/j}; \theta_{x/j_1}, \theta_{x/j_2}, \ldots, \theta_{x/j_n}) \qquad \text{[EQ2]}$$

where:

- $R_{x/j}$ is a random variable defined as the amount of resource X consumed by any individual on programming assignment j.

- f is the probability density function of resource usage.

- $\theta_{x/j_1}, \theta_{x/j_2}, \ldots, \theta_{x/j_n}$ are parameters of the p.d.f. for resource X and programming assignment j.

Furthermore we hypothesize that the values of the parameters are predictable from some measurements of all the non-individual sources of variations. In our experiment the mathematical model for the parameters would be:

$$\theta_{x/jk} = h_{x/k}(P_j, S_j) \qquad \text{[EQ3]}$$

i.e. there exists a function h for each parameter of each resource consumption random variable distribution which can be determined from the given assignment and order.

To recap, the deterministic model of equation 1 assumes that we can measure all the relevant individual characteristics which will allow us to predict the actual resource consumption of that individual on a programming assignment. The stochastic model of equations 2 and 3 assumes that there is no possibility of measuring all the relevant individual characteristics and therefore we must build our model such that when given all the relevant assignment characteristics it will be able to describe the probabilistic behavior of the resource consumption. In the stochastic model we are not able to predict individual resource usage but we can predict such statistics as the expected values, confidence ranges, etc..

To determine whether a deterministic model is appropriate we need to verify whether or not there exists individual characteristics which account for the inter-individual variation of resource usage. If we assume for the moment that a deterministic model does exist, we can propose the model:

$$\Phi_{x/ij} = g(I_i) \qquad \text{[EQ4]}$$

where:

$$\Phi_{x/ij} = [R_{x/ij} - \mu(R_{x/j})]/\sigma(R_{x/j})$$

- $\Phi_{x/ij}$ is the standardized amount of resource X consumed by individual i on program j. In other words $\Phi_{x/ij}$ is the number of standard deviations in terms of resource X for individual i on program j.

- $\mu(R_{x/j})$ is the mean amount of resource X consumed by all students on program j.

- $\sigma(R_{x/j})$ is the standard deviation of the resource X consumption on program j.

- g is a function which when given the appropriate individual characteristic(s) will yield the individual's standardized amount of resource X on assignment j.

The validity of this model depends on two assumptions. The first is that an individual's standardized resource usage is independent of the program meters, i.e. if an individual is productive for large programs he will be for small ones and vice versa. The second assumption is that either the individual characteristics do not change with time, or they change relatively uniformly for all individuals.

Using the data that was collected on individual characteristics, (see Section II), a series of linear correlation analyses were performed using the model in equation 4. The results of these correlation analyses show that none of the individual characteristics measured in this experiment account for much of the inter-individual variation of resource usage, (see Table V). Other

researchers including Chrysler [2] have also indicated difficulty in determining individual characteristic meters which are predictive of individual resource consumption. Although it has been shown by DeNelsky & McKee [4] and by McNamara & Huges [8] that certain programmer's aptitude tests are moderately predictive of job performance as measured by supervisory ratings, there has been no evidence that these tests predict individual resource variations.

An argument could be made that the failure to find an adequate measure of individual resource variation does not imply that one does not exist. However when a canonical regression analysis is performed to determine the ability of the individual characteristics as a group to explain the inter-individual variation they account for little of the variation, (see Table VI). In addition, if we examine the ranges[1] of the standardized resource consumption amounts, $\Phi_{x,j}$, of the data collected from our experiment most individuals vary significantly in their standardized resource consumption, (see Figure 1 and Table VII).

In previous research [9] the authors have shown that the perceived complexity of a programming task is a highly

---

[1]The range for an individual is defined to be the absolute difference between the maximum and minimum $\Phi_{x,j}$ for all j.

subjective entity and thus the deterministic prediction of individual resource consumption depends upon understanding the psychology of the individual - a formidable, if not impossible task. All these points would seem to weigh against the possibility of finding a set of individual characteristics adequate for the deterministic model.

It therefore seems to us that the assumption of a deterministic model is not supportable. In the next section we shall examine the validity of the stochastic model.

INDIVIDUAL CHARACTERISTICS	RESOURCE USAGE METERS			
	#HOURS	#RUNS	CPU-TIME	CPU-COST
EDLEVEL	.001	.000	.000	.000
EDLOAD	.000	.000	.001	.007
EMPLOAD	.032	.000	.000	.000
CAMPEXP	.027	.007	.002	.002
DEPTEXP	.144	.004	.000	.000
WORKEXP	.002	.013	.008	.011
DPEXP	.002	.002	.003	.007
CGPA	.078	.173	.108	.134
AGE	.001	.004	.001	.003
INTEREST	.030	.001	.002	.012
ABILITY	.121	.003	.003	.003
ENJOY	.018	.017	.024	.020

TABLE V - Results of the correlation analysis of the individual characteristics with the standardized resource usage variables. Results are in terms of the coefficient of determination(R^2).

RESOURCE METER	R^2
# HOURS	.256
# RUNS	.276
CPU-COST	.247
CPU-TIME	.185

TABLE VI - Results of the canonical correlation of individual characteristics as a group with the standardized resource usage variables. Results are in terms of the coefficient of determination.

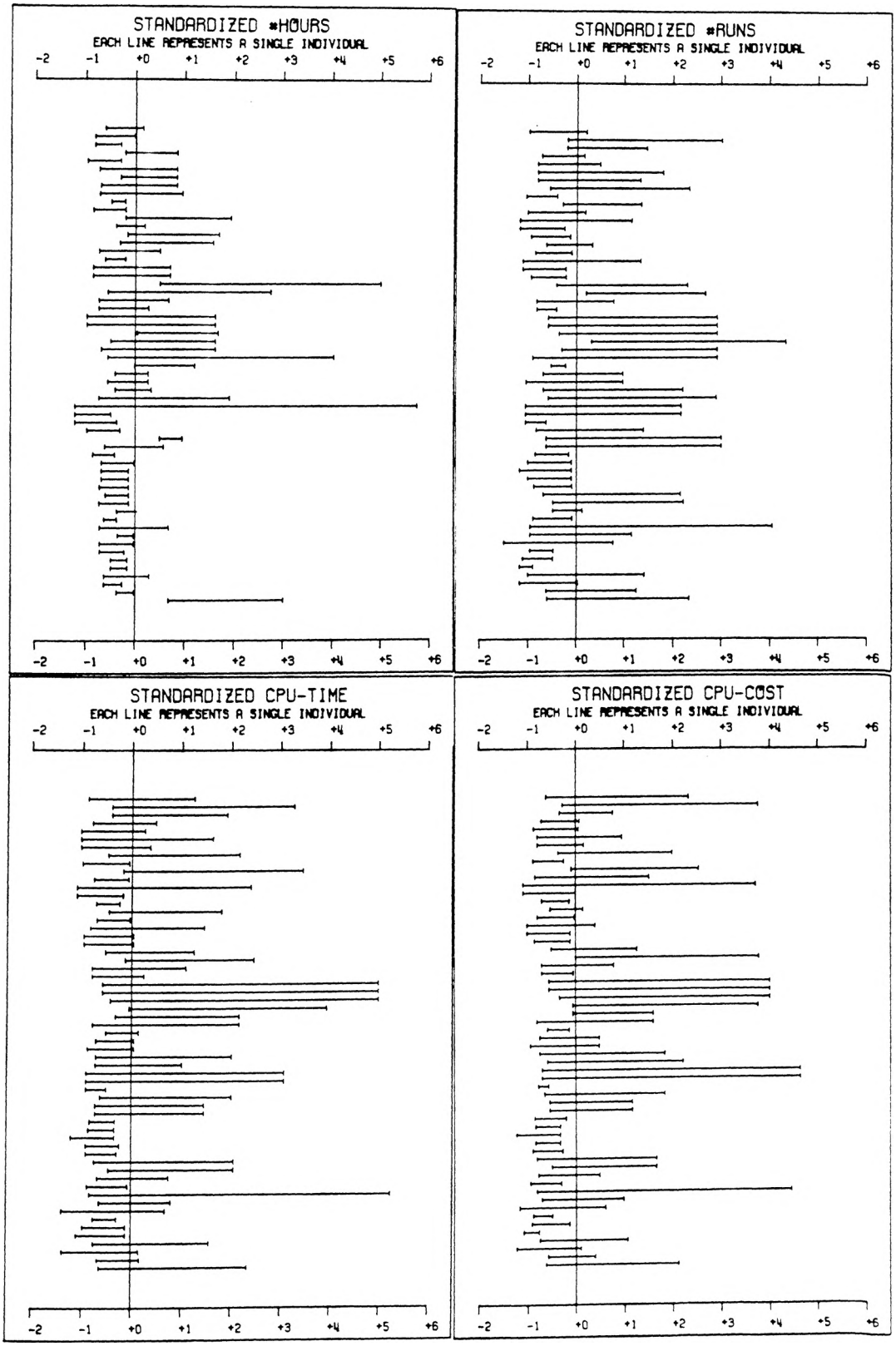


FIGURE 1 - Deviation plots of the individuals' standardized resource consumption variables.

STANDARDIZED # RUNS		
RANGES	FREQ %	CUM FREQ %
0.00 - 0.25	0.0	0.0
0.25 - 0.50	8.5	8.5
0.50 - 1.00	25.4	33.9
1.00 - 2.00	16.9	50.8
2.00 - 3.00	27.1	78.0
3.00 - 4.00	18.6	96.6
4.00 - 5.00	1.7	98.3
5.00 - 6.00	1.7	100.0
6.00 - 7.00	0.0	100.0

(A)

STANDARDIZED # HOURS		
RANGES	FREQ %	CUM FREQ %
0.00 - 0.25	0.0	0.0
0.25 - 0.50	13.6	13.6
0.50 - 1.00	37.3	50.9
1.00 - 2.00	28.8	79.7
2.00 - 3.00	6.8	86.4
3.00 - 4.00	8.5	94.9
4.00 - 5.00	3.4	98.3
5.00 - 6.00	0.0	98.3
6.00 - 7.00	1.7	100.0

(B)

TABLE VII - Distribution of individuals' standardized resource usage ranges. Part 1 of 2 parts.

STANDARDIZED CPU-COST		
RANGES	FREQ %	CUM FREQ %
0.00 - 0.25	1.7	1.7
0.25 - 0.50	5.1	6.8
0.50 - 1.00	30.5	37.3
1.00 - 2.00	27.1	64.4
2.00 - 3.00	18.6	83.1
3.00 - 4.00	3.4	86.4
4.00 - 5.00	8.5	94.9
5.00 - 6.00	5.1	100.0
6.00 - 7.00	0.0	100.0

(C)

STANDARDIZED CPU-TIME		
RANGES	FREQ %	CUM FREQ %
0.00 - 0.25	0.0	0.0
0.25 - 0.50	5.1	5.1
0.50 - 1.00	25.4	30.5
1.00 - 2.00	22.0	52.5
2.00 - 3.00	30.5	83.1
3.00 - 4.00	10.2	93.2
4.00 - 5.00	0.0	93.2
5.00 - 6.00	5.1	98.3
6.00 - 7.00	1.7	100.0

(D)

TABLE VII continued.

IV - DETERMINATION OF THE RESOURCE CONSUMPTION DISTRIBUTION.

Given the premise that the programming task of a software development process is essentially a stochastic process, it becomes necessary to define the resource consumption quantities as random variables. In the experiment described in the previous section let the observable outcomes of interest be designated as R_h, R_r, R_t and R_c and defined as the resource consumption of a successfully completed programming task as measured in man-hours, computer-runs, cpu-seconds, and cpu-dollars respectively. In the following discussion the abstract random variable R will be used to denote any of these specific random variables.

Since the distinguishing property of a random variable is the probability value associated with each event of a measurement of that random variable, the probabilistic behavior of the outcomes can be completely described by identifying their probability distributions.

In order to derive a set of postulated theoretic distributions which might describe the empirical data we examined the histograms of R_h R_r, R_t, and R_c. A representative sample of these histograms is given in Figure 2. It is readily apparent that the histograms all display a general unimodal shape and positive skewness. Five

theoretic distributions were selected as postulates based on their ability to assume the desired shape and skewness. These distributions were the Log-Normal, Beta, Gamma, Weibull and Type I Extreme Value Maxima. In addition the Normal distribution was used for comparison. The exact forms of the distributions selected are shown in Appendix A.

In order to investigate the adequacy of the six postulated distributions to explain the empirical distributions of the collected data, a series of Chi-Square goodness-of-fit tests were performed. The methods used to estimate the parameter values of the theoretic distributions from the sample are given in Appendix A. Figure 3 depicts the six theoretic p.d.f.'s overlaid on a sample resource histogram. The results of the Chi-Square tests are shown in Table VIII.

The log-Normal distribution best fits the empirical data in most instances. Possible reasons for the appropriateness of this distribution will be examined in Section V.

Although the foregoing goodness-of-fit tests were performed correctly there was one major, but intentional, omission. In the conduct of the experiment it was likely that some individuals would either not complete an assignment or complete it unsuccessfully. In either case

these individuals were excluded from the samples of R_h, R_r, R_t and R_c in the previous goodness-of-fit analysis because they did not meet the definition of the random variables - i.e. they were not "successfully completed". The presence of this multiple random censoring causes complications because the censored data do not provide complete information i.e. tell us when the task was successfully completed. Yet the censored data do provide partial information which is that up to the point of censoring successful completion did not take place. To incorporate this partial information available from the multiple censored data, and therefore construct a more accurate model, we can make use of a procedure[1] outlined by Bury[1].

Let us define a function of each of the resource consumption random variables called the 'Completion Intensity Function' or the 'Instantaneous Rate of Completion' and designate it as v(r)[2] This function will yield the conditional probability of successfully completing

---

[1] The applicability of this and other procedures from the area of reliability theory is easily understood if one considers the parallel nature of that theory and our research. The principal unknown random variable in reliability theory is TIME TO FAILURE, in our research the unknown random variable is similar but opposite, q.e. RESOURCE USAGE TO COMPLETION.

[2]Remember that the random variable R is an abstract one representing the actual random variables R_h, R_r, R_t and R_c.

the programming task with resource consumption R=r given that R>=r. Let us define f(r) as the probability density function (p.d.f) of R and F(r) as the cumulative distribution function (c.d.f) of R. Clearly f(r)=P(R=r) and F(r)=P(R<=r) and therefore the instantaneous rate of completion can be obtained by:

$$v(r) = f(r) \,/\, [1-F(r)]$$

To obtain the actual instantaneous rates of completion from the empirical data we will use the following. Let the measurement domain of the random variable R be divided into adjoining intervals $\Delta r_1, \Delta r_2, \ldots, \Delta r_n$. Denote the number of occurrences in Δr_i as $c(r_i)$. The relative frequency of occurrences of measurements in Δr_i, denoted by $\eta(r_i)$, is then the ratio of the number of occurrences to the sample size, i.e. $\eta(r_i)=c(r_i)/n$; where n includes both completed and censored data points. Since theoretically it is possible to obtain larger and larger samples while decreasing the interval size such that $n*\Delta r_i$ remains finite then the relative frequency of observations per interval approaches the probability density of R.

$$\text{i.e. } \lim_{\substack{n\to\infty \\ \Delta r_i \to dr}} \eta(r_i) \,/\, \Delta r_i = f(r_i)$$

A like argument can be followed to show that:

$$\lim_{\substack{n\to\infty \\ \Delta r_j \to dr}} \sum_{j=1}^{i} \eta(r_j) = F(r_i)$$

We can now restate the formula for the instantaneous rate of completion using the above results as:

$$v(r_i) = \lim_{\substack{n\to\infty \\ \Delta r_j \to dr}} [c(r_i)/(n*\Delta r_i)] \;/\; [1-\sum_{j=1}^{i} c(r_j)/n]$$

If we now order all the observations according to the amount of resource R they consumed at completion or censoring we can then define the 'Cumulative' instantaneous rate of completion as:

$$V(r_i) = \sum_{j=1}^{i} v(r_j)*\Delta r_j$$

or:

$$V(r_i) = \lim_{\substack{n\to\infty \\ \Delta r_j \to dr}} \sum_{j=1}^{i} [c(r_j) / (n-\sum_{k=1}^{j} c(r_k))] \qquad [EQ5]$$

The denominator of the above equation now represents the total number of non-completions just prior to the resource consumption r_i including those data that were subsequently censored.

Since we desire the distribution of R, i.e. F(r), we need to establish a relationship between it and V(r). By

definition we know that:

$$v(r) = f(r) \;/\; [1-F(r)]$$

by a simple transformation we obtain the following:

$$v(r) = [d(F(r))/dr] \;/\; [1-F(r)]$$

and thus:

$$v(r)\, dr = d[-LN(1-F(r))]$$

integrating the above from a truncation point r_0 to r yields:

$$\int_{r_0}^{r} v(r)\, dr = -LN[1-F(r)] \Big|_{r_0}^{r}$$

Since $F(r_0) = 0$:

$$\int_{r_0}^{r} v(r)\, dr = -LN[1-F(r)]$$

Hence:

$$F(r) = 1 - EXP\left[-\int_{r_0}^{r} v(r)\, dr \right]$$

And since by definition:

$$V(r) = \int_{r_0}^{r} v(r)\, dr$$

then:

$$F(r) = 1 - EXP[-V(r)] \qquad [EQ6]$$

Summarizing the above we now know that we can develop an empirical c.d.f. using all the available information from the sample by means of equations 5 and 6. It is now

possible to test the postulated distributions' c.d.f's for their adequacy in explaining this empirical c.d.f..

A series of non-linear regression analyses were performed to determine the adequacy of the six postulated distributions' c.d.f.'s to explain the empirical c.d.f's of the experimental data formed from equations 5 and 6. The results of these analyses are shown in Table IX. Sample plots of the theoretic and empirical c.d.f.'s are displayed in Figure 4. Once again the Log-Normal distribution appears to best explain the empirical data. The values of the parameters of the Log-Normal distributions as determined by these regression analyses are presented in Table X.

To statistically verify the significance of the theoretic Log-Normal c.d.f. a series of Komolgorov goodness-of-fit tests were performed using the parameter values from Table X. The results of these tests are shown in Table XI and demonstrate that the Log-Normal distribution is statistically significant as a theoretic distribution for the empirical data.

Up to this point we have seen only empirical justification for the applicability of the Log-Normal distribution to the resource consumption during the programming phase. In the next section we will present both

an informal and a formal argument to explain why this applicability might exist.

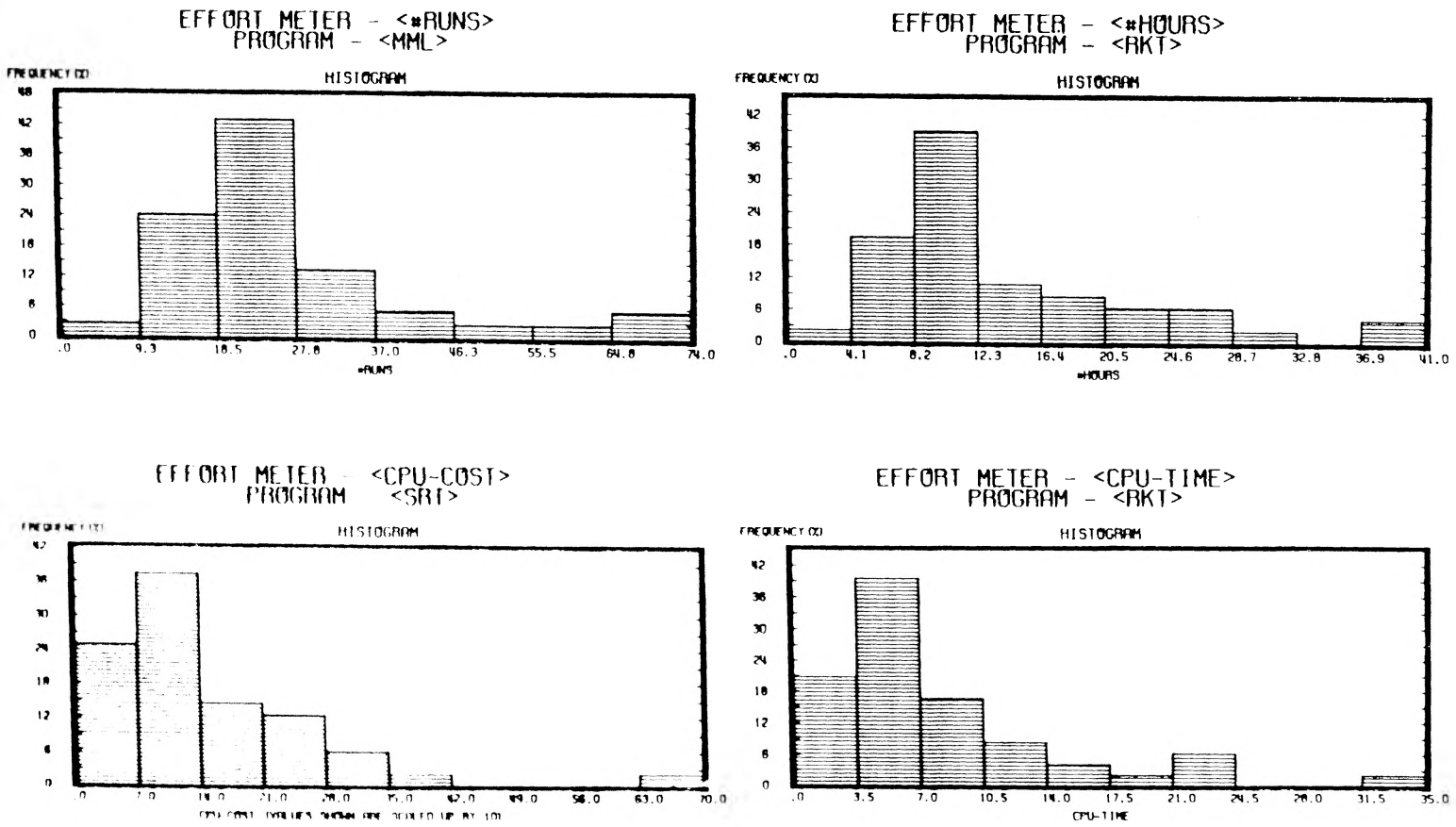


FIGURE 2 - Sample Histograms of the Resource Consumption Random Variables.

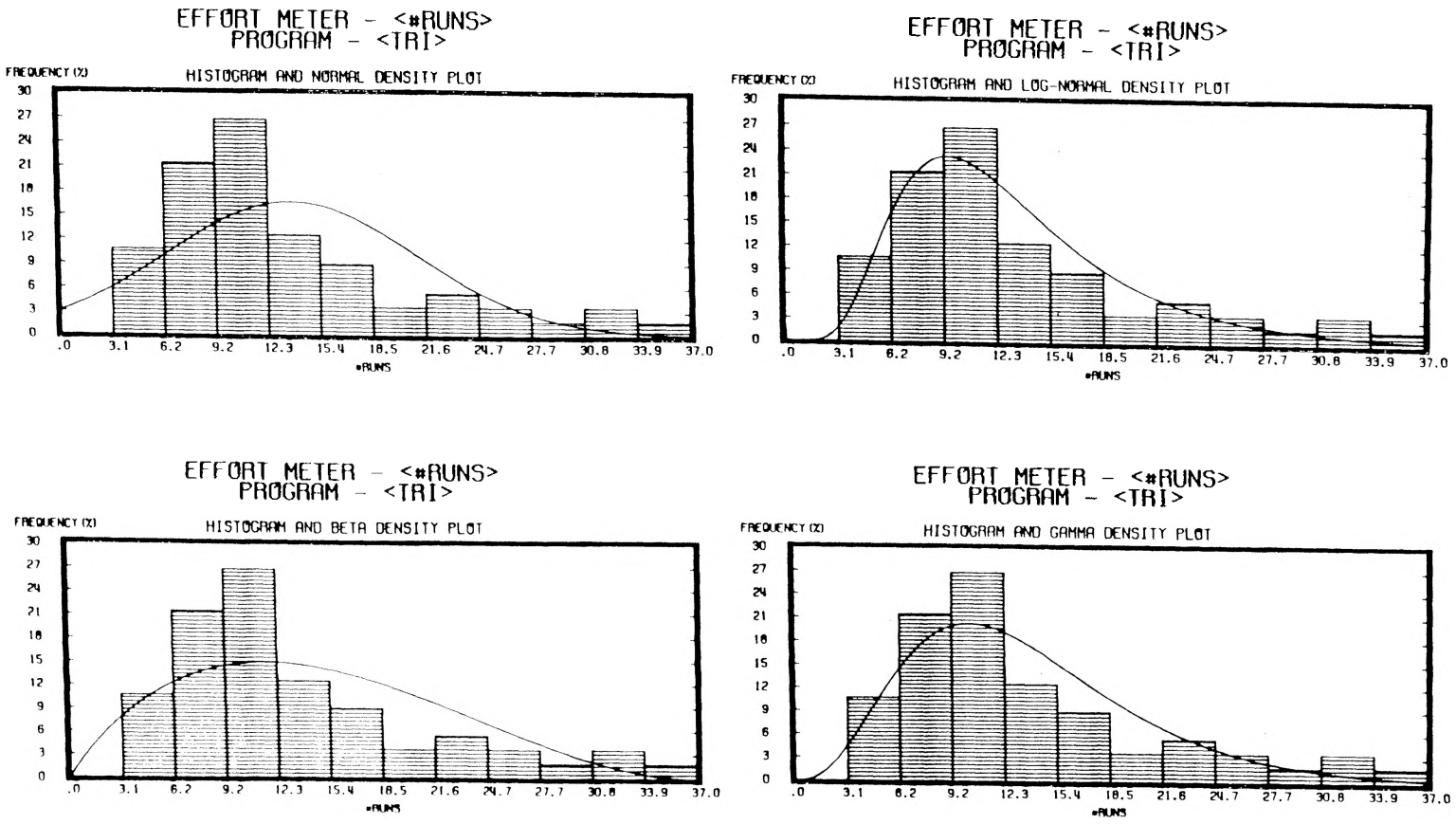
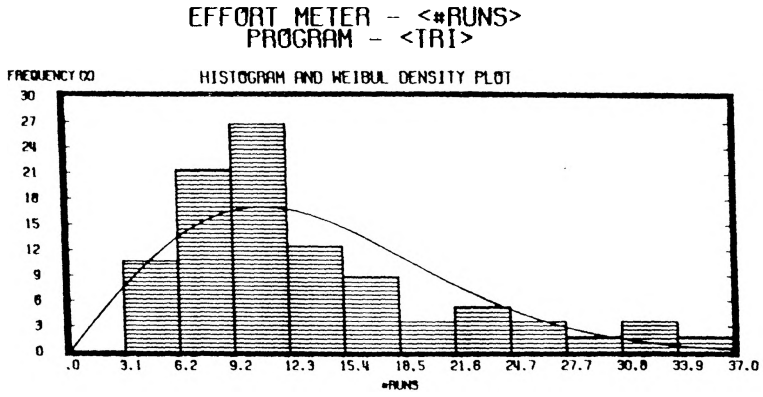


FIGURE 3 - Postulated distributions overlaid on a sample histogram. Part 1 of 2 parts.



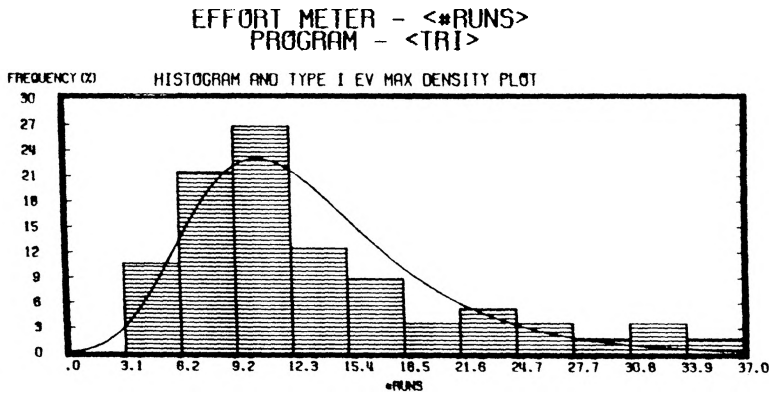


FIGURE 3 - continued.

EFFORT METER - # RUNS

PGM IDS	POSTULATED DISTRIBUTIONS						D F	TAB χ^2
	NORM	LN	BETA	GAM	WEIB	MAX		
TRI	33.6	29.3	24.1	13.5	29.6	17.5	8	15.5
RKT	31.1	4.9	19.1	18.0	16.9	13.5	6	12.6
INS	20.5	0.9	9.1	3.3	6.8	2.9	5	11.1
SRT	27.7	2.6	22.5	6.0	14.2	6.7	6	12.6
MST	30.0	10.8	14.8	8.0	5.6	8.0	5	11.1
MML	18.1	9.7	15.4	12.0	12.8	7.5	4	9.5

EFFORT METER - # HOURS

PGM IDS	POSTULATED DISTRIBUTIONS						D F	TAB χ^2
	NORM	LN	BETA	GAM	WEIB	MAX		
TRI	56.6	20.8	29.8	18.5	44.2	20.0	7	14.1
RKT	33.8	5.7	31.1	11.5	17.0	19.3	6	12.6
INS	24.8	4.4	25.2	15.6	17.6	7.6	4	9.5
SRT	22.0	9.4	21.6	12.9	15.9	12.1	5	11.1
MST	25.3	5.9	26.7	13.3	11.2	11.5	3	7.8
MML	28.4	8.5	24.6	13.4	13.4	11.5	3	7.8

TABLE VIII - Results of the χ^2 goodness-of-fit tests for the six postulated distributions with the resource usage data (not including censored data points). Results are in terms of the calculated χ^2 values. Part 1 of 2 parts. See Part 2 for legend.

EFFORT METER - CPU-TIME

PGM	POSTULATED DISTRIBUTIONS						D	TAB
IDS	NORM	LN	BETA	GAM	WEIB	MAX	F	χ^2
TRI	27.7	8.4	28.9	19.8	27.7	14.7	8	15.5
RKT	29.2	10.9	22.9	13.5	15.0	16.1	6	12.6
INS	29.8	2.1	18.5	2.5	7.6	4.5	5	11.1
SRT	36.0	3.7	23.2	9.7	6.3	11.6	6	12.6
MST	21.6	5.2	16.8	9.2	12.8	7.2	5	11.1
MML	25.2	14.3	33.2	12.8	17.3	11.2	4	9.5

EFFORT METER - CPU-COST

PGM	POSTULATED DISTRIBUTIONS						D	TAB
IDS	NORM	LN	BETA	GAM	WEIB	MAX	F	χ^2
TRI	29.2	10.0	28.9	18.2	21.0	13.9	8	15.5
RKT	55.5	6.4	27.0	10.1	16.9	16.1	6	12.6
INS	24.4	2.9	9.1	4.5	5.2	8.0	5	11.1
SRT	19.5	4.1	18.0	6.0	9.0	4.9	6	12.6
MST	16.0	7.2	20.0	11.2	18.0	8.0	5	11.1
MML	26.8	9.7	23.4	12.8	20.3	10.1	4	9.5

LEGEND

DF: Degrees of Freedom.

TAB χ^2: Tabulated χ^2 values at the .05 significancy level.

≡XX.X≡: A calculated χ^2 value indicating that the distribution is not rejected at the .05 significancy level.

TABLE VIII - continued.

EFFORT METERS	PGM IDS	POSTULATED DISTRIBUTIONS					
		NORM	LN	BETA	GAM	WEIB	MAX
#RUNS	TRI	.072	.037	.063	.053	.057	.049
	RKT	.074	.029	.051	.042	.049	.051
	INS	.066	.018	.053	.101	.034	.044
	SRT	.064	.020	.041	.032	.038	.041
	MST	.059	.037	.040	.046	.035	.043
	MML	.079	.047	.065	.070	.068	.059
#HOURS	TRI	.071	.033	.052	.072	.052	.052
	RKT	.079	.041	.066	.053	.062	.056
	INS	.083	.035	.058	.064	.055	.062
	SRT	.057	.026	.040	.079	.041	.041
	MST	.043	.040	.037	.093	.038	.039
	MML	.091	.056	.076	.069	.076	.074
CPU-TIME	TRI	.073	.040	.063	.051	.062	.049
	RKT	.073	.028	.051	.074	.048	.150
	INS	.065	.019	.036	.065	.036	.043
	SRT	.054	.021	.031	.083	.032	.035
	MST	.053	.035	.037	.080	.034	.042
	MML	.096	.060	.076	.071	.078	.074
CPU-COST	TRI	.069	.040	.055	.155	.060	.323
	RKT	.067	.024	.045	.075	.045	.046
	INS	.064	.021	.041	.067	.039	.042
	SRT	.060	.023	.036	.090	.035	.038
	MST	.049	.043	.037	.058	.037	.039
	MML	.070	.049	.063	.057	.066	.056

TABLE IX - Results of the non-linear regression analyses of the postulated distributions' C.D.F.'s ability to describe the empirically derived C.D.F.. Results are in terms of the standard error of the estimate (s.e.e.).

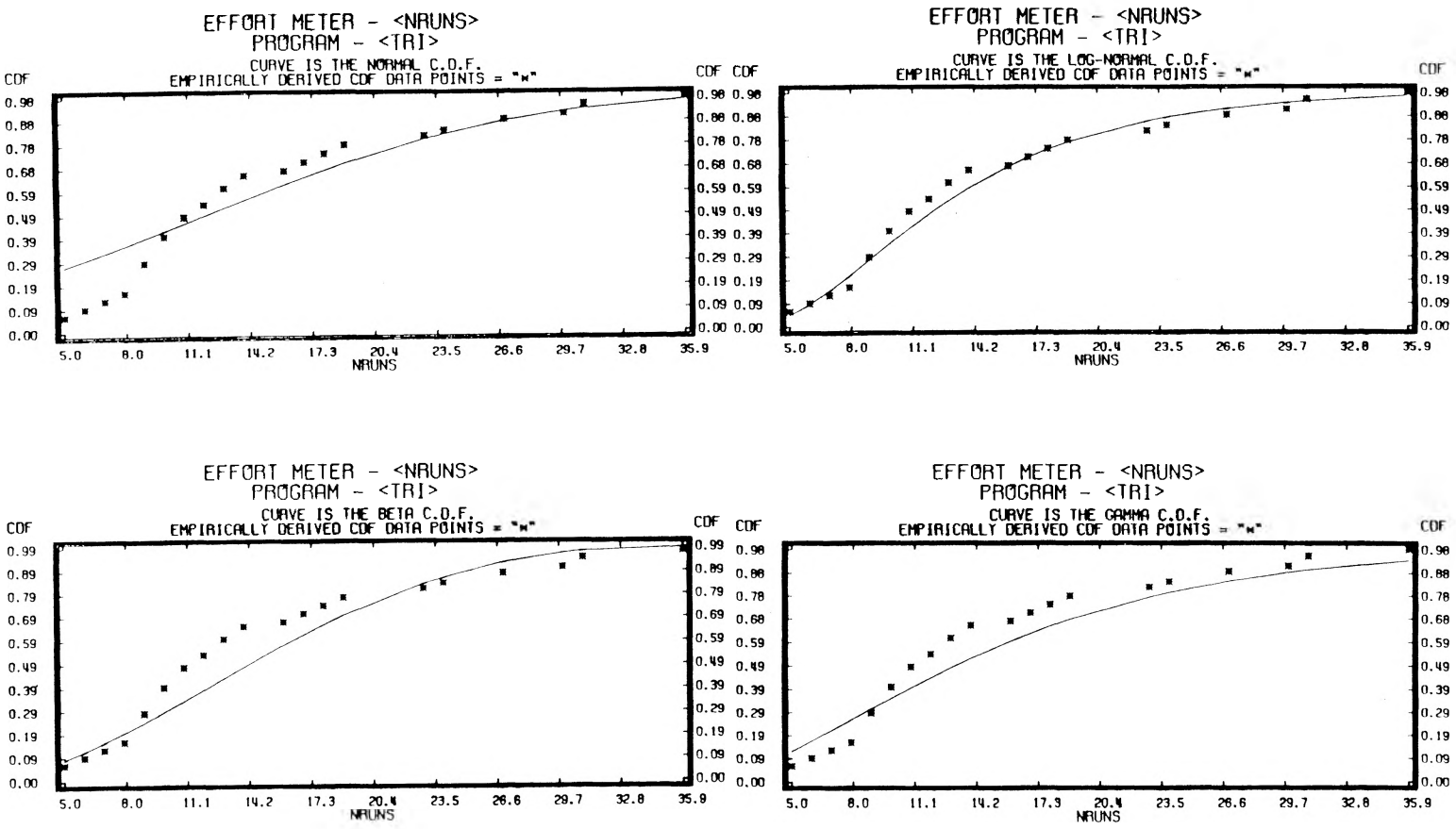


FIGURE 4 - Sample plots of the theoretic distributions' c.d.f.'s and the empirical c.d.f. formed from all the data. Theoretic parameters derived from the non-linear regressions analyses. Part 1 of 2 parts.

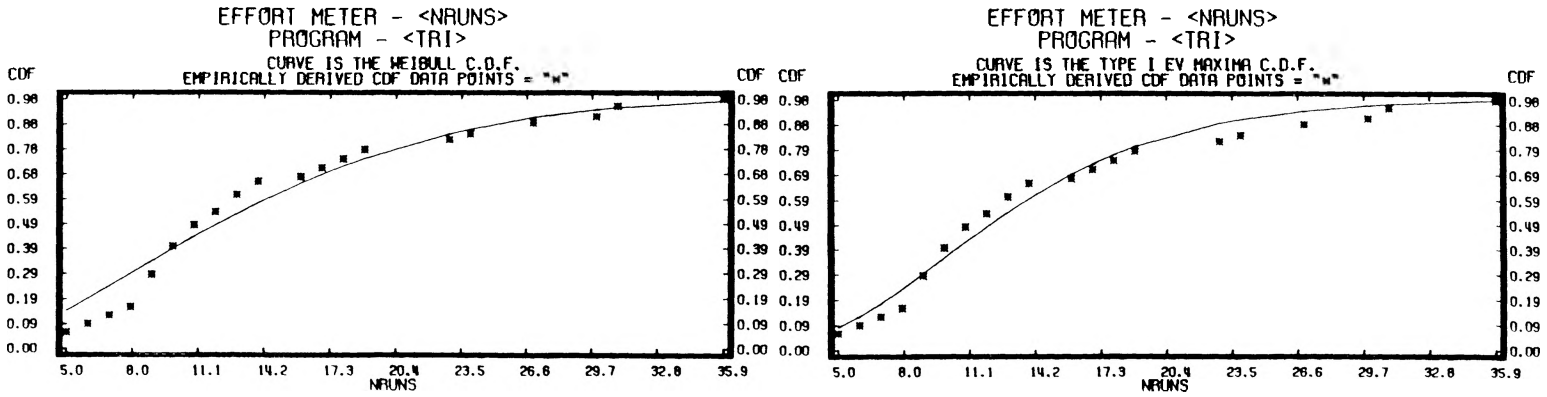


FIGURE 4 - continued.

EFFORT METER	PGM IDS	θ_1	θ_1 ASYMPTOTIC STD. ERROR	θ_2	θ_2 ASYMPTOTIC STD. ERROR
# RUNS	TRI	2.483	.017	0.578	.026
	RKT	2.831	.015	0.759	.025
	INS	3.285	.009	0.867	.016
	SRT	2.795	.009	0.780	.016
	MST	3.088	.021	0.863	.034
	MML	3.224	.019	0.603	.036
# HOURS	TRI	1.759	.024	0.720	.038
	RKT	2.468	.019	0.613	.031
	INS	2.683	.022	0.834	.038
	SRT	2.105	.016	0.736	.027
	MST	2.581	.018	0.504	.028
	MML	2.866	.031	0.731	.056
CPU-COST	TRI	-.584	.011	0.516	.019
	RKT	0.397	.009	0.745	.015
	INS	0.979	.009	0.793	.016
	SRT	0.146	.009	0.775	.016
	MST	0.795	.021	0.897	.035
	MML	1.193	.015	0.549	.029
CPU-TIME	TRI	1.150	.012	0.516	.020
	RKT	1.891	.011	0.782	.020
	INS	2.436	.008	0.827	.015
	SRT	2.293	.008	0.802	.014
	MST	2.867	.019	1.057	.035
	MML	2.906	.022	0.661	.042

TABLE X - Parametric results of Log-Normal non-linear regression analysis of the empirical c.d.f. formed from all data.

EFFORT METERS	PGM IDS	KCAL	N	KTAB
# HOURS	TRI	.076	53	.187
	RKT	.076	46	.200
	INS	.110	35	.224
	SRT	.065	42	.210
	MST	.099	34	.227
	MML	.160	32	.234
# RUNS	TRI	.077	56	.183
	RKT	.073	48	.196
	INS	.066	41	.212
	SRT	.049	48	.196
	MST	.073	40	.210
	MML	.129	37	.218
CPU-COST	TRI	.073	56	.183
	RKT	.055	48	.196
	INS	.051	41	.212
	SRT	.056	48	.196
	MST	.088	40	.210
	MML	.120	37	.218
CPU-TIME	TRI	.073	56	.183
	RKT	.068	48	.196
	INS	.070	41	.212
	SRT	.052	48	.196
	MST	.063	40	.210
	MML	.123	37	.218

KCAL - Calculated K statistic.
N - Sample size.
KTAB - Tabulated K statistic at the appropriate degrees of freedom based on N and at the .05 significancy level.

TABLE XI - Results of the Komolgorov goodness-of-fit tests of the Log-Normal distribution to the empirical c.d.f. formed from all the data. In all cases the hypothesis was not rejected at the 0.05 significancy level.

V - RATIONALE FOR THE LOG-NORMAL DISTRIBUTION.

In the previous section we have shown by empirical analyses that the Log-Normal theoretic distribution is a reasonable one to adopt in our attempt to describe the probabilistic behavior of individual resource consumption during the programming phase of a software development process. Shortly we will present a formal theoretic argument for the applicability of this distribution but first it will be insightful to informally examine the characteristics of the postulated distributions to determine whether they match expectations based on our intuitive understanding of the programming process.

From our knowledge of the programming process we know that there must exist a lower bound to the resource consumption random variable at R=0. In order for distributions such as the Normal or Type I Extreme Value Maxima, which theoretically have no lower bounds, to be applied to our process we must impose an artificial truncation below the point R=0. However, the Log-Normal, Beta, Gamma, and Weibull distributions do have theoretic lower bounds at the appropriate point making them more appropriate then those that do not.

We also know that there does not exist an upper bound to resource consumption, at least theoretically. Of the six postulated theoretic distributions all but the Beta distribution have no upper bounds.

Intuitively we would expect the distribution of individual resource consumption to be such that there are a small number of highly productive individuals, i.e. individuals who use less resources, a larger number of individuals who are moderately productive and a small number of individual who display low productivity. In other words we would expect the distribution to be unimodal. All six of the postulated distributions possess the ability to assume a unimodal shape given appropriate values for their parameters.

Another facet we can examine is the nature of the Completion Intensity Function, (see Section IV). We would expect that as time or any other resource quantity is expended the chances of an individual completing an assignment correctly, given that he has not already completed it, would increase - up to a point. At some point in the resource spectrum it seems natural that this conditional probability should start to decrease due to such factors as falling motivation, getting 'lost' on the assignment, etc.. Of all the postulated distributions only

the Log-normal has a Completion Intensity Function which is non-monotonically increasing. Figure 5 shows some sample Completion Intensity function plots based on our experimental data for the six postulated distributions.

Table XII recaps the preceding discussion and demonstrates that as far as intuition is concerned the Log-Normal distribution once again seems appropriate. While intuition alone is often misleading, in this case it is supported by the empirical evidence.

It is appropriate to mention at this point that prior to our data analyses the authors did not expect the Log-Normal to be the most suitable of the postulated distributions. When the empirical data convinced us that it was we felt it necessary to attempt a theoretic argument for the Log-Normal distribution in order to better understand it's applicability. Once again we drew upon the discipline of reliability theory to develop a possible theoretic argument for the applicability of the Log-Normal distribution. The following argument closely follows that of Mann, et al [7] where it was applied to the applicability of the Log-Normal to a fatigue-life model.

The central theme of our argument may be stated informally as relying on a 'proportional effects model' for

the 'growth' of a program. This implies that the growth of a program at each 'stage' of it's development is randomly proportional to it's degree of 'maturation' at the previous stage. The program starts out at some minute, but non-zero, degree of maturation and continues to mature until it reaches some predefined completion degree of maturation. The following is a more formal expression of this argument.

Let $M_0 < M_1 < \ldots < M_n$ be a sequence of random variables that denote the degree of maturation of the program at successive stages of growth. The degree of maturation of a program can be thought of as its percentage of completion. A stage of growth can be viewed in micro terms using the principle of cognitive psychology that information is processed in stages [3].

Let ΔM_i be the increment that the maturity of the program was increased by in stage i, i.e. $\Delta M_i = M_i - M_{i-1}$. Given the proportional effects model we know:

$$\Delta M_i = \rho_i M_{i-1}$$

where ρ_i, the constant of proportionality, is a random variable assumed to be independent of all other ρ_j's, but not necessarily with the same distribution.

Let M_0 be the program's degree of maturation at the beginning which is some minute but non-zero value. This can

be thought of as the conceptual 'seed' from which the program grows.

Let M_n be some predefined degree of maturation which has been specified to be the acceptable minimum maturity of the completed program.

Then:

$$\rho_i = [\Delta M_i / M_{i-1}] \qquad i=1,2,\ldots,n$$

and thus:

$$\sum_{i=0}^{n} \rho_i = \sum_{i=0}^{n} [\Delta M_i / M_{i-1}]$$

Given that ΔM_i can get smaller and smaller while n increases such that $\Delta M_i n$ remains finite then:

$$\lim_{\substack{\Delta M_i \to 0 \\ n \to \infty}} \sum_{i=0}^{n} \rho_i = \int_{M_0}^{M_n} M^{-1} dM = \ln(M_n) - \ln(M_0)$$

or:

$$\ln(M_n) = \sum_{i=1}^{n} \rho_i + \ln(M_0)$$

Since by assumption the ρ_i's are independent random variables it follows from the central limit theorem that:

$$\sum_{i=1}^{n} \rho_i \sim \text{NORMAL}(\mu_1, \sigma_1^2)$$

and thus:

$$\ln(M_n) \sim \text{NORMAL}(\mu_2, \sigma_2^2)$$

therefore:

$$M_n \sim \text{LOG-NORMAL}(\mu_2, \sigma_2^2)$$

If we assume that the resource consumption, R_i, at any stage is a function of the work performed:

i.e. $R_i = h(\Delta M_i)$

and further assume that the function h is distribution preserving then:

$$R_n \sim \text{LOG-NORMAL}(\theta_1, \theta_2)$$

In summary we have argued that if we assume a proportional effects model for the growth of a program and in addition assume that resource usage is a distribution preserving function of the program growth then we reach the conclusion that the resource consumption random variable is Log-Normal. In this argument we have had to make a number of assumptions which to the authors seem reasonable.

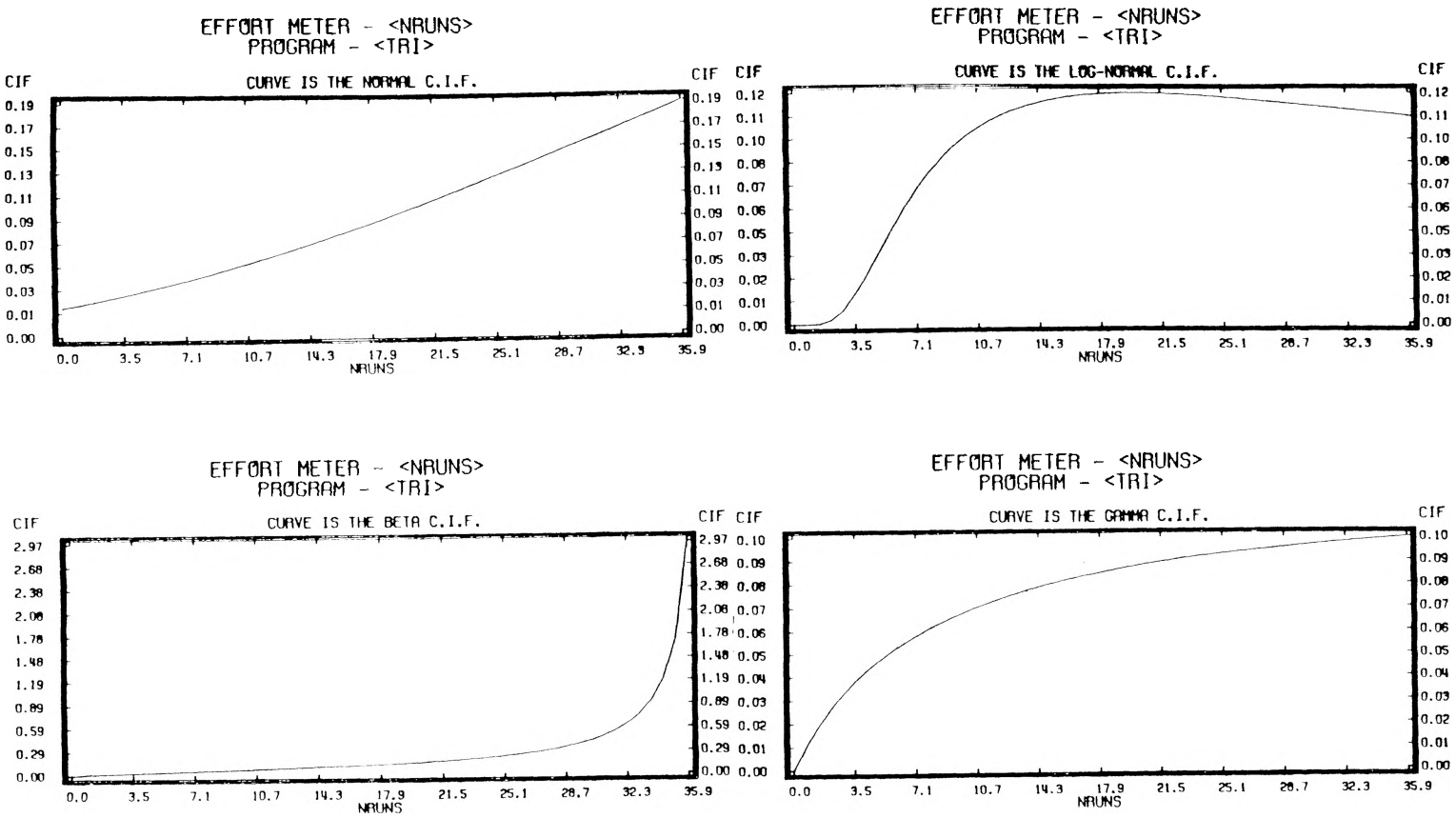


Figure 5 - Sample Completion Intensity Functions for the six postulated distributions. Part 1 of 2 parts.

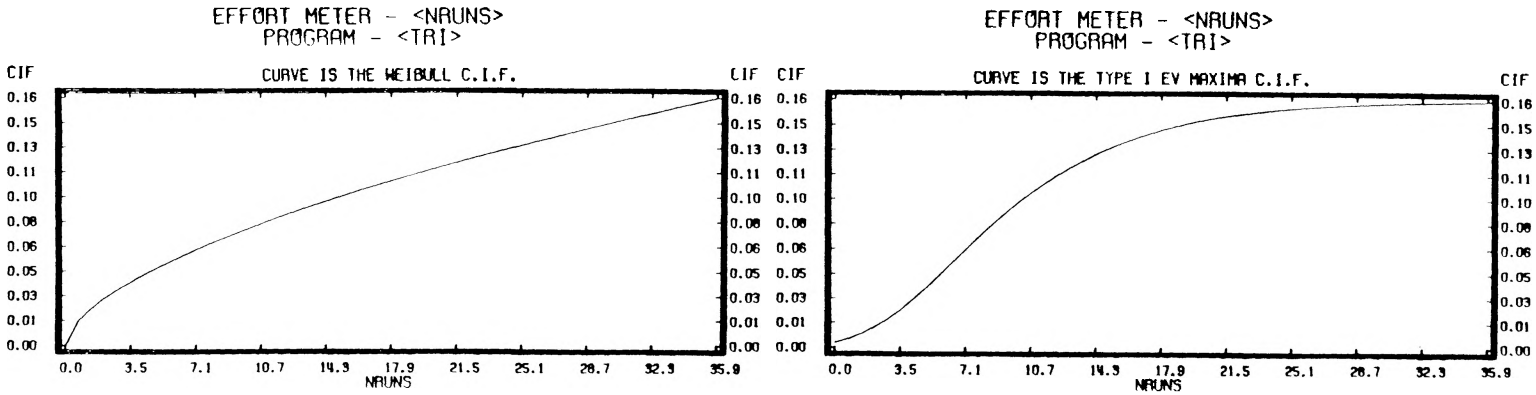


Figure 5 - continued.

POSTULATED DISTRIBUTIONS	CONDITIONS		
	I	II	III
NORMAL	No	Yes	No
LOG-NORMAL	Yes	Yes	Yes
BETA	Yes	No	No
GAMMA	Yes	Yes	No
TYPE I EXTREME VALUE MAXIMA	No	Yes	No
WEIBULL	Yes	Yes	No
IDEAL	Yes	Yes	Yes

CONDITIONS:

I Distribution has a theoretic lower limit at R=0.

II Distribution has no theoretic upper limit.

III Distribution has a non-monotonically increasing Completion Intensity Function.

TABLE XII - Comparison of postulated distributions' characteristics with the 'ideal'.

VI - CONCLUSIONS.

From the preceding analyses we believe that the following conclusions are justified:

- Individual resource consumption is an outcome of a process which is inherently stochastic and should be viewed as a random variable rather than deterministically.
- The Log-Normal distribution appears to be an appropriate model to adopt in order to empirically describe the probabilistic behavior of individual resource consumption during the programming phase of software development.
- The Log-Normal distribution possesses characteristics which seem appropriate for describing resource usage during the programming process and in addition a theoretic argument can be constructed for it which depends on reasonable assumptions.

In addition to the data collected for the experiment described in Section II we have collected data from two other identical experiments which we are currently processing to cross-validate the Log-Normal model.

As with many of the experiments conducted within the area of software engineering, our research is set in the 'academic' environment and therefore care must be taken not to over-generalize the obtained results. Our conclusions need to be validated with professional programmers before they can be applied within the 'industrial' environment.

Yet university research of this type might very likely play a crucial role in the development of predictive models since it is impractical to expect industries to 'repeat' a task.

Our research has concentrated on small programs (less than 500 lines of code) because of our environment. However it should be noted that over 60% of the 'real world' programs fall into this category [6].

The obvious follow-up to the research presented in this paper is an attempt to predict the parameters of the Log-Normal distribution based on some measurement of the assignment specification. The authors are currently investigating this topic with very promising preliminary results[10].

REFERENCES

[1] K.V. Bury, Statistical Models in Applied Science , John Wiley & Sons, N.Y., N.Y., 1975.

[2] E. Chrysler, "Some basic determinants of computer programming productivity," in Commun. Ass. Comput. Mach. , vol. 21, pp. 472-482, 1978.

[3] B. Curtis, Human Factors in Software Development , IEEE Tutorial, COMPSAC November 1981.

[4] G.Y. DeNelsky and M.G. McKee, "Prediction of computer programmer training and job performance using the AABP test," in Personnel Psychology , vol. 27, pp. 129-137, 1974.

[5] J.A. Faraguhar, "A preliminary inquiry into the software estimation process," Tech Report AD 712 052 , Defense Documentation Center, Alexandria, Va, August 1970.

[6] C. Jones, Programming Productivity: Issues for the Eighties , IEEE Tutorial, COMPSAC November 1981.

[7] N.R. Mann, R.E. Schafer and N.D. Singpurwally, Methods for Statistical Analysis of Reliability and Life Data , John Wiley & Sons, N.Y., 1974.

[8] W.J. McNamara and J.L. Huges, "A review of research on the selection of computer programmers," in Personnel Psychology , vol. 14, pp. 39-41, 1961.

[9] D.G. McNicholl and K. Magel, "The Subjective Nature of Programming Complexity," to be presented at the Human Factors in Computer Systems conference, ACM, March 1982.

[10] D.G. McNicholl and K. Magel, "The Prediction of Resource Consumption during the Programming Phase of Software Development," in preparation.

[11] C.C. Tonies, "Project management fundementals in software engineering," in Software Engineering , Jensen and Tonies, Eds. Prentice Hall, Englewood Cliffs, N.J., 1979.

[12] M.V. Zelkowitz, A.C. Shaw, and J.D. Gannon, Principles of Software Engineering and Design , Prentice Hall, Englewood Cliffs, N.J., 1979.

APPENDIX A - P.D.F.'s, C.D.F.'s, and Parameter Point Estimation Techniques for the Postulated Distributions.

1. Normal Distribution:

(a) Probability density function (p.d.f.)

$$f(x;\theta_1,\theta_2)=[\theta_2\sqrt{2\pi}]^{-1}EXP[-.5((x-\theta_1)/\theta_2)^2]$$

where: $0 < x,\theta_1,\theta_2$

(b) Cumulative distribution function (c.d.f.)

$$F(x;\theta_1,\theta_2)=[\sqrt{2\pi}]^{-1}\int_0^x(\theta_2)^{-1}EXP[-.5((x-\theta_1)/\theta_2)^2]dx$$

(c) Maximum likelihood parameter point estimators

$$\hat{\theta}_1=n^{-1}\sum_{i=1}^{n}X_i$$

$$\hat{\theta}_2=\left[(n-1)^{-1}\sum_{i=1}^{n}(X_i-\hat{\theta}_1)^2\right]^{1/2}$$

2. Log-Normal Distribution:

(a) Probability density function (p.d.f.)

$$f(x;\theta_1,\theta_2)=[x\theta_2\sqrt{2\pi}]^{-1}EXP[-.5((\ln(x)-\theta_1)/\theta_2)^2]$$

where: $0 < x,\theta_2 \qquad -\infty < \theta_1 < +\infty$

(b) Cumulative distribution function (c.d.f.)

$$F(x;\theta_1,\theta_2)=[\theta_2\sqrt{2\pi}]^{-1}\int_0^x(x)^{-1}EXP[-.5((\ln(x)-\theta_1)/\theta_2)^2]dx$$

(c) Maximum likelihood parameter point estimators

$$\hat{\theta}_1=n^{-1}\sum_{i=1}^{n}\ln(X_i)$$

$$\hat{\theta}_2 = \left[(n-1)^{-1} \sum_{i=1}^{n} [\ln(x) - \theta_1]^2 \right]^{1/2}$$

3. Gamma Distribution:

(a) Probability density function (p.d.f.)

$$f(x;\theta_1,\theta_2) = [\theta_1 \Gamma(\theta_2)]^{-1} (x/\theta_1)^{\theta_2 - 1} \text{EXP}[-x/\theta_1]$$

where: $0 < x, \theta_1, \theta_2$
Γ is the gamma function.

(b) Cumulative distribution function (c.d.f.)

$$F(x;\theta_1,\theta_2) = [\Gamma(\theta_2)]^{-1} \int_0^x (x/\theta_1)^{\theta_2 - 1} \text{EXP}[-x/\theta_1](\theta_1)^{-1} dx$$

(c) Maximum Likelihood parameter point estimators

$$\hat{\theta}_1 = \left[n^{-1} \sum_{i=1}^{n} X_i \right] / \hat{\theta}_2$$

The estimate for θ_2 is found by inverse interpolation of the following equation using the moment estimator of θ_2 as a starting value.

$$\ln(\hat{\theta}_2) - \psi(\hat{\theta}_2) - \ln\left[\left(n^{-1} \sum_{i=1}^{n} X_i \right) / \left(\prod_{i=1}^{n} X_i \right)^{1/n} \right] = 0$$

where the moment estimator for θ_2 is:

$$\tilde{\theta}_2 = \left[n^{-1} \sum_{i=1}^{n} X_i \right] / \left[(n-1)^{-1} \sum_{i=1}^{n} \left(X_i - \left[n^{-1} \sum_{i=1}^{n} X_i \right] \right)^2 \right]$$

and ψ is the digamma function.

4. Beta Distribution:

(a) Probability density function (p.d.f.)

$$\theta_3 - 1 \qquad \theta_4 - 1$$

$$f(x;\theta_1,\theta_2,\theta_3,\theta_4)=[B(\theta_3,\theta_4)]^{-1}(z)^{\theta_3-1}(1-z)^{\theta_4-1}[\theta_2-\theta_1]^{-1}$$

where: $B(\theta_3,\theta_4)=\Gamma(\theta_3)\Gamma(\theta_4)/\Gamma(\theta_3+\theta_4)$
$z=(x-\theta_1)/(\theta_2-\theta_1)$
$\theta_1<x<\theta_2 \quad 0<=\theta_1<\theta_2 \quad 0<\theta_3,\theta_4$

(b) Cumulative distribution function (c.d.f.)

$$F(x;\theta_1,\theta_2,\theta_3,\theta_4)=[B(\theta_3,\theta_4)]^{-1}\int_0^z (z)^{\theta_3-1}(1-z)^{\theta_4-1}\,dz$$

(c) Maximum likelihood parameter point estimators

$$\hat{\theta}_1 = 0$$

It is especially difficult to develop an estimator for θ_2 since our process has no theoretic upper limit, see discussion in Section V. For our purposes we will use the following ad hoc estimators:

$$\hat{\theta}_2 = \text{Max}(X_1,X_2,\ldots,X_n)+1 \text{ if } X \text{ is discrete.}$$

$$\hat{\theta}_2 = \text{CEIL}[\text{Max}(X_1,X_2,\ldots,X_n)] \text{ if } X \text{ is continuous.}$$

The estimates for θ_3 and θ_4 must be found by an iterative procedure using the following two equations with the moment estimators as starting values.

$$\psi(\hat{\theta}_3) - \psi(\hat{\theta}_3+\hat{\theta}_4) = \ln\left[\prod_{i=1}^{n} z_i^{1/n}\right]$$

$$\psi(\hat{\theta}_4) - \psi(\hat{\theta}_3+\hat{\theta}_4) = \ln\left[\prod_{i=1}^{n} (1-z_i)^{1/n}\right]$$

where the moment estimators for θ_3 and θ_4 are:

$$\tilde{\theta}_3 = [(\bar{X})^2-\bar{X}M_2]/[M_2-(\bar{X})^2]$$

$$\tilde{\theta}_4 = [\bar{X}-M_2]/[M_2-(M_2)^2] - \tilde{\theta}_3$$

and $\bar{X}$ = the sample average.

$$M_2 = n^{-1}\sum_{i=1}^{n} (X_i)^2$$

5. Type I Extreme Value Maxima Distribution:

(a) Probability density function (p.d.f.)

$$f(x;\theta_1,\theta_2)=(\theta_2)^{-1}EXP[-z-EXP(-z)]$$

where: $z=(x-\theta_1)/\theta_2$

$0 < x,\theta_1 < \infty \quad 0 < \theta_2$

(b) Cumulative distribution function (c.d.f.)

$$F(x;\theta_1,\theta_2) = EXP[-EXP(-z)]$$

(c) Maximum likelihood parameter point estimators

$$\hat{\theta}_1 = -\hat{\theta}_2 ln[n^{-1}\sum_{i=1}^{n} EXP(-x_i/\hat{\theta}_2)]$$

The estimate for θ_2 must be found by inverse interpolation of the following equation using the moment estimator as the starting value.

$$\hat{\theta}_2 - \bar{X} + [\sum_{i=1}^{n} x_i EXP(-x_i/\hat{\theta}_2)][\sum_{i=1}^{n} EXP(-x_i/\hat{\theta}_2)]^{-1} = 0$$

where the moment estimator for θ_2 is:

$$\tilde{\theta}_2 = .7797\sigma_x$$

and σ_x is the standard deviation of the sample.

6. Weibull Distribution:

(a) Probability density function (p.d.f.)

$$f(x;\theta_1,\theta_2)=(\theta_2/\theta_1)(x/\theta_1)^{\theta_2-1} EXP[-(x/\theta_1)^{\theta_2}]$$

where: $0 < x,\theta_1,\theta_2$

(b) Cumulative distribution function (c.d.f.)

$$F(x;,\theta_1,\theta_2) = 1-EXP[-(x/\theta_1)^{\theta_2}]$$

(c) Maximum likelihood parameter point estimators

$$\hat{\theta}_1 = \left[n^{-1} \sum_{i=1}^{n} (x_i^{\hat{\theta}_2}) \right]^{1/\hat{\theta}_2}$$

The estimate for θ_2 must be found by inverse interpolation of the following equation.

$$\left[\sum_{i=1}^{n} x_i^{\hat{\theta}_2} \ln(x_i)\right]\left[\sum_{i=1}^{n} x_i^{\hat{\theta}_2}\right]^{-1} - (\theta_2)^{-1} = n^{-1} \sum_{i=1}^{n} \ln(x_i)$$