AD NUMBER
AD447677
LIMITATION CHANGES
TO: **Approved for public release; distribution is unlimited.**
FROM: **Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; JUL 1964. Other requests shall be referred to Office of Naval Research, Washington, DC 20360.**
AUTHORITY
onr via cna ltr 2 dec 1970

CNA

ESTIMATING CUMULATIVE PROBABILITY FROM AGGREGATED TRUNCATED DATA

By H. Spitz

Research Contribution No. 56

CONTRACT NONR 3732(00)

Research Contribution

OPERATIONS EVALUATION GROUP

Center for Naval Analyses

WASHINGTON 25, D. C.

6 July 1964

RESEARCH CONTRIBUTION

Operations Evaluation Group

CENTER FOR NAVAL ANALYSES

ESTIMATING CUMULATIVE PROBABILITY FROM AGGREGATED TRUNCATED DATA

By H. Spitz

Research Contribution No. 56

H Spitz

ABSTRACT

Cumulative survival, failure, or detection probabilities cannot in general be precisely estimated from truncated samples if only data grouped in successive time intervals is available. Mathematical models of the failure rate and abort rate within the time intervals are postulated from which estimates may be obtained from grouped data when the models are valid. An easily calculable approximation formula can be used in the earlier time intervals where the sample size is relatively large. This can provide data for verifying or rejecting a given model prior to making calculations in later intervals where the smaller sample size would otherwise diminish the reliability of the resulting probabilities.

(REVERSE BLANK)

INTRODUCTION

The cumulative probability as a function of time of the occurrence of some significant event, such as the failure of an element or the detection of a target, is often useful in describing the effectiveness of a system. This probability may be estimated from data giving the times of occurrence of the event in a sample consisting of a number of observations of the operation of the system or of similar systems. When no observations in the sample are terminated except when the event of interest occurs, the sample may be described as "uncensored" or "non-truncated." In such cases the appropriate estimate of the cumulative probability of occurrence of the event to time T is the fraction of the sample in which the event has occurred prior to T, reference (a). However, if the sample is truncated because it contains aborted observations which terminated prior to the occurrence of the event, the appropriate estimate of the cumulative probability is not so readily obtainable. Methods given in references (a) and (b) are not rigorous because they depend on a derivation of the expected value of the probability at the time of the first occurrence of the event which excludes the possibility of abort. The present paper suggests other methods of deriving cumulative probability estimates from truncated samples.

DEFINITION OF PROBLEM

Find the cumulative survival probability, P_s, or the cumulative failure probability, $P_f = 1 - P_s$, as a function of time from data given in the form: In successive time intervals Δt_i, $\Delta t_i = t_i - t_{i-1}$, $(t_o = 0)$

N_i elements were present at the beginning of the interval, i.e., at time t_{i-1}

r_i elements failed during the interval

a_i elements aborted from the sample during the interval but prior to failing.

Data is usually grouped in equal time intervals, but the Δt_i's need not be equal in this formulation. For application to problems involving detection of targets, "non-detection" and "detection" may be substituted for "survival" and "failure" respectively in the above definition.

CUMULATION FORMULAS

In general,

$$P_s(t_i) = P_s(t_{i-1})p_{si} \tag{1}$$

with $P_s(t_i)$ = cumulative probability of surviving to the end of Δt_i, and

p_{si} = probability of not failing within Δt_i. This may also be written as

$$P_s(t_i) = \prod_{j=1}^{i} p_{sj} \ . \tag{2}$$

Alternatively, in terms of failure probability, where $P_f(t_i) = 1 - P_s(t_i)$, and $p_{fi} = 1 - p_{si}$:

$$P_f(t_i) = P_f(t_{i-1}) + \left[1 - P_f(t_{i-1})\right]p_{fi} \tag{3}$$

is equivalent to equation (1) and

$$P_f(t_i) = 1 - \prod_{j=1}^{i} (1 - p_{fj}) \tag{4}$$

to equation (2).

NON-TRUNCATED CASE

If there are no aborts, $a_i = 0$ and the usual definition of p_{si} applies:

$$p_{si} = \frac{N_i - r_i}{N_i} \ . \tag{5}$$

For a set of data which is not truncated $N_i = N_{i-1} - r_{i-1}$ so that substitution of equation (5) into equation (2) gives

$$P_s(t_i) = \prod_{j=1}^{i} \frac{N_j - r_j}{N_j}$$

$$= \left[\frac{N_1 - r_1}{N_1}\right]\left[\frac{N_1 - r_1 - r_2}{N_1 - r_1}\right] \cdots \left[\frac{N_1 - r_1 - \cdots - r_{i-1}}{N_1 - r_1 - \cdots - r_{i-2}}\right]\left[\frac{N_i - r_1 - \cdots - r_i}{N_1 - r_1 - \cdots - r_{i-1}}\right]$$

$$P_s(t_i) = 1 - \frac{\sum_{j=1}^{i} r_j}{N_1}$$

the conventional result that the cumulative failure probability is the ratio of failures to trials.

TRUNCATED CASE WITH DISCRETE DATA

Another special case of the general problem defined above arises when the data gives the exact time of each failure and abort. In this case the events, both failures and aborts, may be ordered chronologically and t_i chosen to be the time of occurrence of the i-th event. Then, if the i-th event is a failure, $r_i = 1$, $a_i = 0$, $p_{si} = \frac{N_i - 1}{N_i} = \frac{N_1 - i}{N_1 - i + 1}$; similarly, $r_i = 0$, $a_i = 1$, $p_{si} = 1$, if the i-th event is an abort. Substitution of these values of p_{si} into equations (2) or (3) gives an estimate of the cumulative probabilities at the time of occurrence of each failure or abort. A computational shortcut is available when a sequence of n failures is uninterrupted by aborts since repeated application of equation (1) shows that

$$P_s(t_i) = P_s(t_{i-n}) \frac{N_1 - i}{N_1 - i + n}$$

when event i and the n - 1 preceding events are all failures. This might also be derived from consideration of equation (5).

TRUNCATED CASE WITH GROUPED DATA

With grouped truncated data the definition of p_{si} given by equation (5) does not hold unless the assumption is made that all aborts occur at the end of the time interval. If, on the other hand, it is assumed that all aborts occur at the beginning of Δt_i the equivalent form of equation (5) is

$$P_{si} = \frac{N_i - a_i - r_i}{N_i - a_i}. \tag{6}$$

As a third hypothesis, assume that all aborts occur simultaneously somewhere within the time interval, so that r' failures occur prior to the aborts and the remaining $r_i - r'$ after the aborts. Then

$$P_{si} = \frac{N_i - r'}{N_i} \cdot \frac{N_i - a_i - r_i}{N_i - r' - a_i}. \tag{7}$$

Thus, the value of p_{si} depends on when the aborts occur. It is assumed that this is not known for the grouped data case. Nevertheless, it is possible to place limits on the value of p_{si} since equation (7) always gives values between those of equations (5) and (6). Thus,

$$\frac{N_i - a_i - r_i}{N_i - a_i} \leq p_{si} \leq \frac{N_i - r_i}{N_i} \tag{8}$$

or alternatively

$$\frac{r_i}{N_i} \leq p_{fi} \leq \frac{r_i}{N_i - a_i}. \tag{9}$$

MATHEMATICAL MODELS OF BEHAVIOR DURING Δt_i

Since, within the limits given by equations (8) and (9), the values of the survival and failure probabilities during Δt_i depend on the history of the failures and aborts within the time interval, it is appropriate to compare the results which arise from various reasonable assumptions about

this history. Dr. Joseph H. Engel of the Operations Evaluation Group has proposed the following model: For convenience of notation define a new time variable $\theta = (t - t_{i-1})/\Delta t_i$ such that $\theta = 0$ at $t = t_{i-1}$, the beginning of Δt_i, and $\theta = 1$ at $t = t_i$, the end of Δt_i.

Let $f'(\theta)$ = the unforestalled failure probability density, that is, the rate of failure assuming there is no abort mechanism in operation

$w'(\theta)$ = the unforestalled abort probability density (rate of aborts assuming there is no failure mechanism in operation).

Then $f(\theta) = \int_0^\theta f'(s)ds$ = the unforestalled cumulative probability of failure to time θ, and probability of failure within Δt_i is

$$p_f = f(1) \tag{10}$$

(The subscript i is omitted here and below where it is understood that only the i-th interval is under consideration.)

Then, if the failure and abort mechanisms are statistically independent, it follows that the probability of failure during Δt_i, allowing for the probability of failure being forestalled by aborts, is

$$F = \int_0^1 \left[1 - w(s)\right] f'(s)ds. \tag{11}$$

Similarly, the probability of abort during Δt, with the probability of abort being forestalled by failure included, is

$$W = \int_0^1 \left[1 - f(s)\right] w'(s)ds. \tag{12}$$

Then with N elements in the sample at the beginning of the interval the expected number of failures (with forestalling by aborts accounted for) is NF. This expected number of failures may be set equal to the observed number of failures:

$$NF = r \tag{13}$$

and similarly

$$NW = a. \tag{14}$$

Equations (13) and (14) provide unbiased estimates of F and W.

Exponential Rates of Failures and Aborts

Assuming

$$f'(\theta) = be^{-b\theta} \tag{15}$$

$$w'(\theta) = ce^{-c\theta} \tag{16}$$

produces from equations (11) and (12)

$$F = \frac{b}{b+c}\left[1 - e^{-(b+c)}\right] \tag{17}$$

and

$$W = \frac{c}{b+c}\left[1 - e^{-(b+c)}\right]. \tag{18}$$

Solving these simultaneously with equations (13) and (14) gives

$$b = \begin{cases} \frac{r}{r+a} \log_e \frac{N}{N - r - a}, & \text{for } r + a > 0 \\ 0, & \text{for } r = a = 0. \end{cases} \tag{19}$$

Then from equations (10) and (15)

$$p_f = \begin{cases} 1 - \left(1 - \frac{r+a}{N}\right)^{\frac{r}{r+a}}, & \text{for } r + a > 0 \\ 0, & \text{for } r = a = 0 \end{cases} \tag{20}$$

or

$$p_s = \begin{cases} \left(1 - \frac{r+a}{N}\right)^{\frac{r}{r+a}}, & \text{for } r + a > 0 \\ 1, & \text{for } r = a = 0. \end{cases} \tag{21}$$

Constant Rates of Failures and Aborts

Assuming

$$f'(\theta) = h, \tag{22}$$

$$w'(\theta) = k, \tag{23}$$

gives

$$F = h\left[1 - \frac{k}{2}\right] \tag{24}$$

$$W = k\left[1 - \frac{h}{2}\right] \tag{25}$$

which may be solved simultaneously with equations (13) and (14) to produce

$$p_f = h = 1 + \frac{r-a}{2N} - \sqrt{\left(1 + \frac{r-a}{2N}\right)^2 - 2\frac{r}{N}}\,, \tag{26}$$

or

$$p_s = \frac{a-r}{2N} + \sqrt{\left(1 - \frac{a-r}{2N}\right)^2 - 2\frac{r}{N}}\,. \tag{27}$$

Other functional forms could be postulated for $f'(\theta)$ and $w'(\theta)$ and, as long as they involve exactly two constants, it is theoretically possible to solve the simultaneous equations (11) through (14) for these constants and thus derive expressions for p_f and p_s as above. Since the two sets of assumptions on failure and abort rates already examined are as reasonable as many others that might be postulated, it does not appear worthwhile to pursue this approach further here. However, the expressions (21) and (27) are somewhat cumbersome to evaluate, especially in the absence of computational aids, so that consideration of simpler expressions approximating these equations may be fruitful.

Average of Limits Approximation

One such approximation is the arithmetic mean between the limits of equations (8) and (9). These may be written as

$$p_s = \frac{1}{2}\left[\frac{N-a-r}{N-a} + \frac{N-r}{N}\right], \qquad (28)$$

$$p_f = \frac{1}{2}\left[\frac{r}{N} + \frac{r}{N-a}\right]. \qquad (29)$$

Average Sample Size Approximation

A simpler expression from the point of view of computational ease may be derived by substituting a/2 for a in equation (6) giving

$$p_s = \frac{N - \frac{a}{2} - r}{N - \frac{a}{2}}, \qquad (30)$$

$$p_f = \frac{r}{N - \frac{a}{2}}. \qquad (31)$$

These last two equations may be thought of as the result of assuming that the average number of elements in the time interval is the number at the beginning decreased by half the number of aborts.

COMPARISON OF RESULTS FROM VARIOUS MODELS

Figure 1 shows in graphical form how the failure probabilities derived from the four expressions arrived at above behave as a function of r/N, the fraction failing within time interval Δt, for the particular case in which one-fifth of the initial elements abort during Δt. However, the following observations apply for all values of a/N:

(a) The p_f value from the Exponential Rates Model exceeds the p_f from the Constant Rates Model from r = 0 to r = a and falls short of it thereafter.

(b) For any value of r/N the Average Sample Size value of p_f is always less than all the others. Thus, this gives a "conservative" estimate of the failure probability.

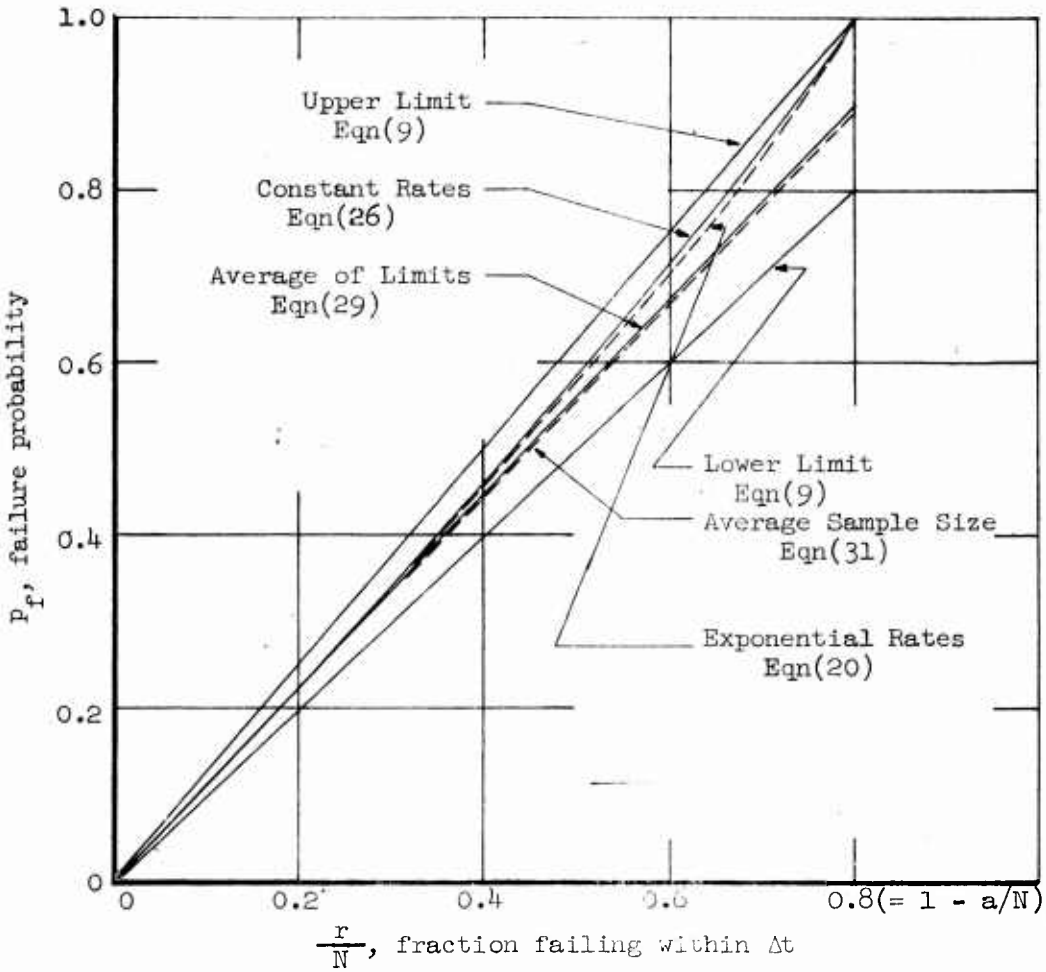

FIG. 1: COMPARISON OF FAILURE PROBABILITIES DERIVED FROM VARIOUS MODELS

(Fraction aborting, $\frac{a}{N} = 0.2$)

(c) When r/N is less than about .25 and a > r, the p_f value from the Average of Limits computation exceeds all the others. This estimate is therefore not "conservative" in this region.

(d) All the estimates of p_f considered lie very close together when r/N and a/N are small.

In order to quantify this last observation the maximum absolute differences between the Constant Rates or Exponential Rates values and each of the Average Sample Size and Average of Limits values were calculated. Figure 2 shows curves of constant differences between p_f (or p_s) values from the Exponential Rates Model or the Constant Rates Model, whichever is larger, and the values from the Average Sample Size formula. From curves of this type, figure 3 was derived showing maximum absolute differences as a function of $\frac{r+a}{N}$, the fraction of the initial number withdrawn from the sample during Δt, either by failure or abort. For values of this fraction up to 0.4, using the most easily calculable approximations, equations (30) and (31) will produce differences no greater than .0032. Since probabilities are ordinarily quoted to only 2 decimal places this approximation will usally suffice. When $\frac{r+a}{N}$ exceeds 0.4, the limiting values on p_s and p_f, equations (8) and (9), are so far apart that the confidence interval on an estimate from any model would be large unless the model could be verified. Ordinarily this value of $\frac{r+a}{N}$ will be exceeded only in later time intervals Δt_i when the sample size N_i has become small. At this point one could plot the cumulative survival probabilities already obtained and also plot cumulative non-abort probabilities derived from formulas analogous to equations (2) and (30):

$$P_{na}(t_i) = \prod_{j=1}^{i} p_{naj} \tag{32}$$

$$p_{na} = \frac{N - \frac{r}{2} - a}{N - \frac{r}{2}} . \tag{33}$$

If these plots appear to fit a straight line on semi-logarithmic paper the assumption of Exponential Rates of Failures and Aborts is appropriate and one may proceed confidently with that model. This method is illustrated in the following numerical example.

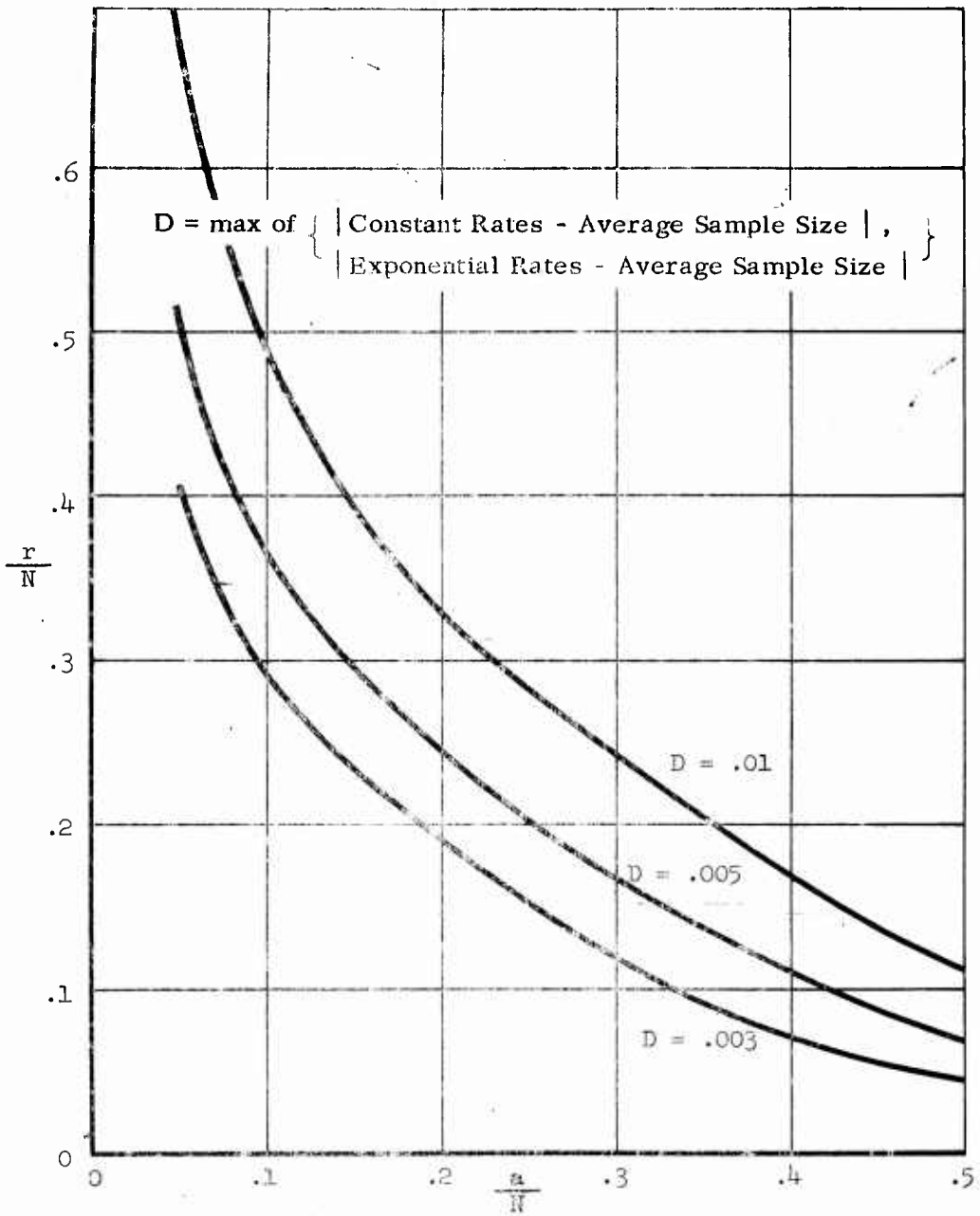

FIG. 2: CURVES OF CONSTANT DIFFERENCE BETWEEN VALUES OF p_s or p_f CALCULATED FROM CONSTANT RATES OR EXPONENTIAL RATES

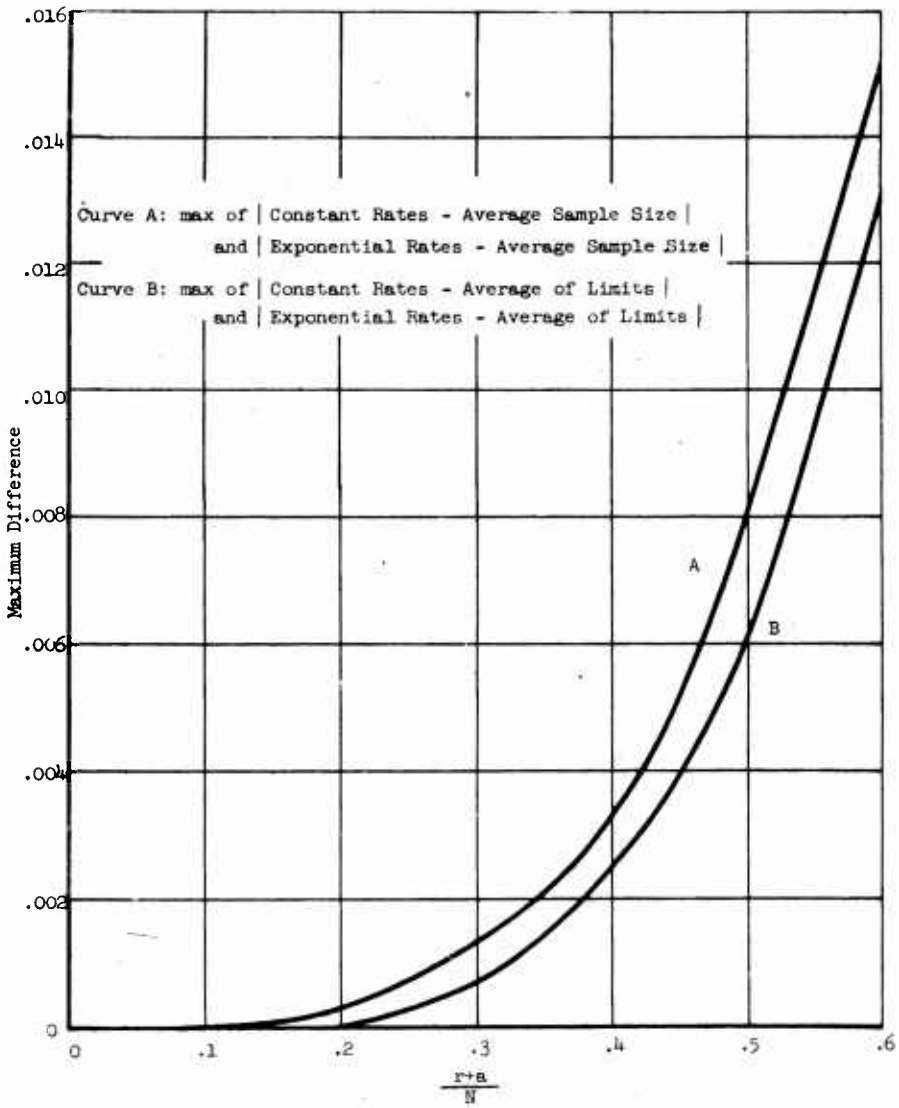

FIG. 3: MAXIMUM ABSOLUTE DIFFERENCES BETWEEN p_s OR p_f CALCULATED FROM CONSTANT RATES OR EXPONENTIAL RATES MODELS AND AVERAGE OF LIMITS AND AVERAGE SAMPLE SIZE FORMULAS

NUMERICAL EXAMPLE

Columns (3), (4), and (5) of table I give hypothetical data for equal consecutive time intervals of length T:

N_i is the sample size at the beginning of the i-th interval;

r_i is the number of failures within the interval;

a_i is the number of aborts within the interval.

Column (6) gives the empirical probability of surviving to the end of the interval on condition of being present at the beginning of the interval. These are calculated from equation (30) except in the cases indicated by asterisks where $r_i + a_i > 0.4N_i$. In these cases equation (21) is used. Column (7) gives the cumulative survival probability to the end of the i-th interval obtained from equation (2). Column (8) gives empirical probability of not aborting within the interval on condition of being present at the beginning of the interval obtained from equation (33). Column (9) gives cumulative non-abort probability from equation (32).

Because $r_9 + a_9 > 0.4N_9$, $P_s(t_i)$ and $P_{na}(t_i)$ for i = 1 to 8 were plotted as shown in figure 4 to validate the Constant Rates of Failure and Abort Model before proceeding further with the calculations.

Figure 5 shows the fit of the resulting $P_s(t_i)$ points to $P_s(t) = e^{-t/m}$ where m is the mean-time-to-failure derived from the original data by the following method.

MEAN-TIME-TO-FAILURE

In the discrete case where exact time of each failure and abort is known, the mean-time-to-failure (MTF) is

$$m = \frac{\sum_i (r_i + a_i)t_i}{\sum_i r_i} \tag{34}$$

where t_i is the time of the i-th event with $r_i = 1$ and $a_i = 0$ if this event is a failure or $r_i = 0$ and $a_i = 1$ if the i-th event is an abort.

TABLE I

ILLUSTRATIVE NUMERICAL EXAMPLE

(1) i	(2) t_i	(3) N_i	(4) r_i	(5) a_i	(6) p_{si}	(7) $P_s(t_i)$	(8) p_{nai}	(9) $P_{na}(t_i)$
1	T	100	15	10	.842	.842	.892	.892
2	2T	75	12	7	.832	.701	.898	.801
3	3T	56	6	6	.887	.622	.887	.710
4	4T	44	7	3	.835	.519	.926	.657
5	5T	34	7	1	.791	.411	.967	.635
6	6T	26	3	4	.875	.360	.837	.531
7	7T	19	2	3	.886	.319	.833	.442
8	8T	14	3	0	.786	.251	1.0	.442
9	9T	11	4	1	.616*	.155	-	-
10	10T	6	0	2	1.0	.155	-	-
11	11T	4	3	1	0*	0	-	-

* Calculated from equation (21) rather than equation (30) because $r_i + a_i > 0.4N_i$.

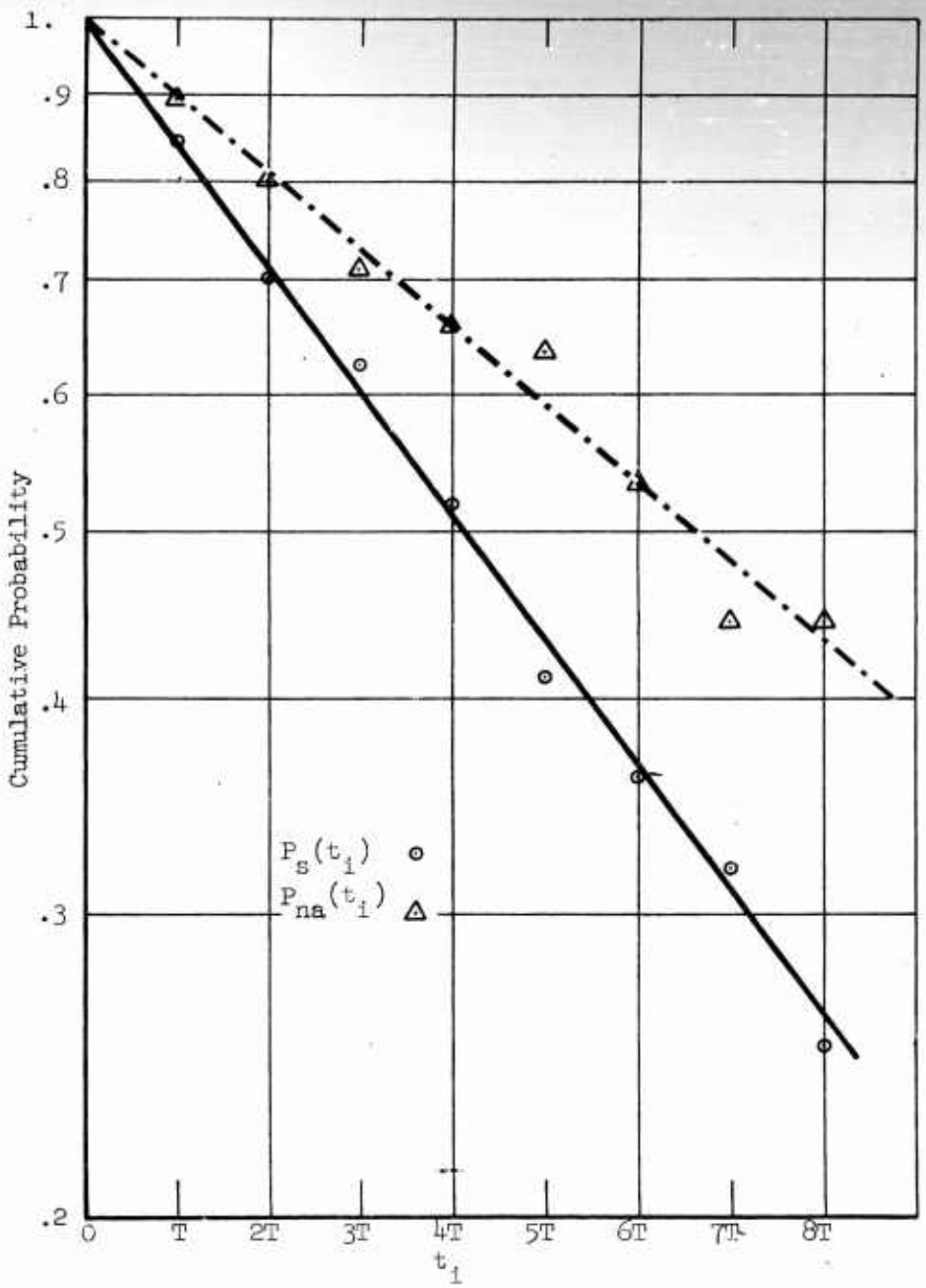

FIG. 4: PLOT OF CUMULATIVE SURVIVAL AND NON-ABORT PROBABILITIES TO VERIFY EXPONENTIAL RATES ASSUMPTION

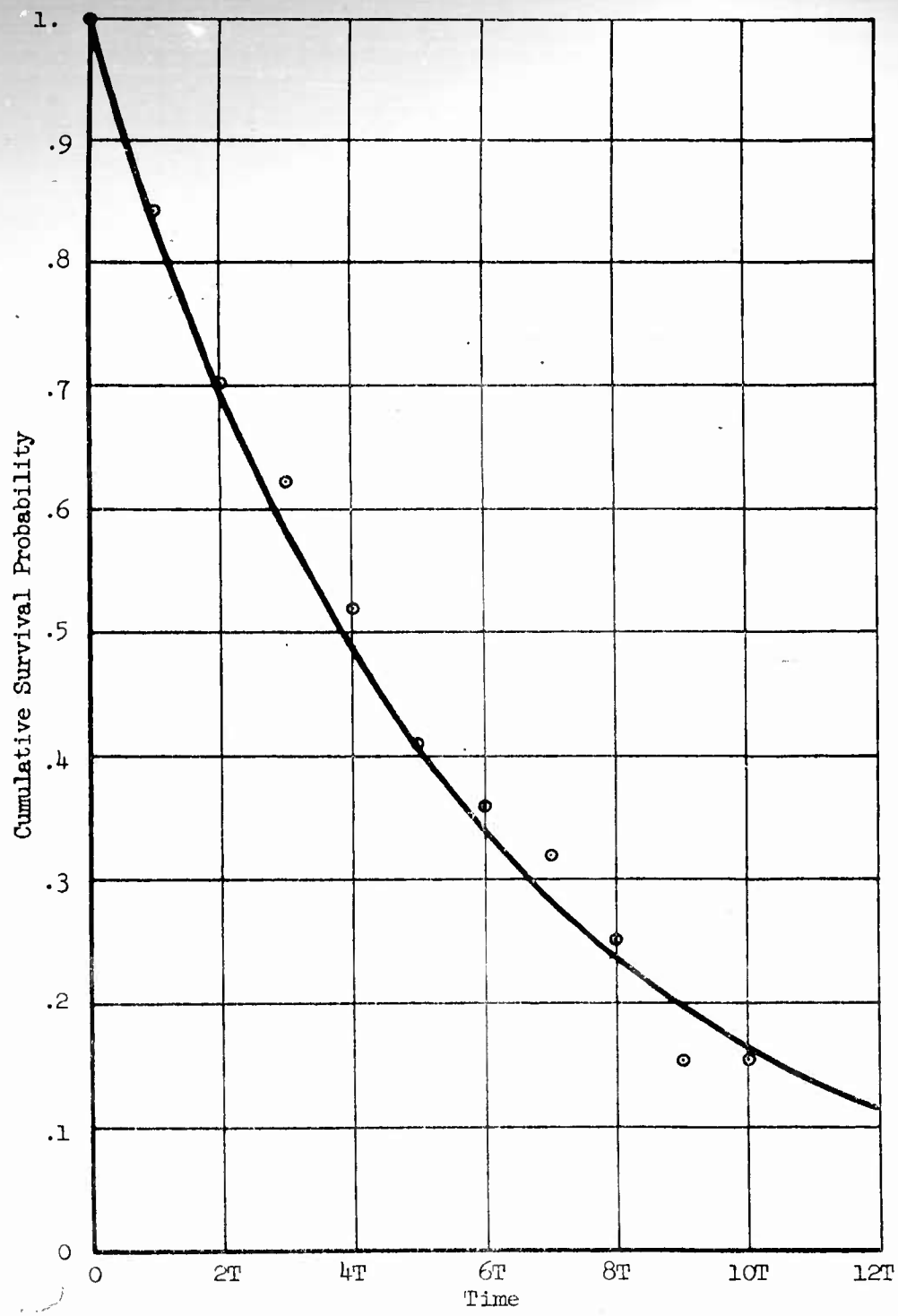

FIG. 5: FIT OF CALCULATED PROBABILITIES TO MEAN-TIME-TO-FAILURE EXPONENTIAL

Reference (a) uses this formula also for data grouped in time intervals Δt_i with t_i the time at the end of the i-th interval and r_i and a_i the number of failures and aborts, respectively, within the interval. For equal time intervals, $\Delta t_i = T$, equation (34) can be written in the more convenient form

$$m' = \frac{T \sum_i N_i}{\sum_i r_i} \tag{35}$$

where m' indicates that this is only a first approximation to MTF for data grouped in equal time intervals. This estimate is generally too high because it assumes that the sample size N_i at the beginning of the i-th interval persists throughout the interval. Assuming exponential rates of failure and abort, a correction factor may be derived:

Let probability of failure by time t be

$$P_f(t) = 1 - e^{-t/m} \qquad (m = \text{MTF}) \tag{36}$$

and probability of abort by time t be

$$P_a(t) = 1 - e^{-t/u} \tag{37}$$

with u = mean-time-to-abort. Then probability of withdrawal from the sample by either failure or abort by time t is

$$\begin{aligned} P_w(t) &= P_f(t) + P_a(t) - P_f(t)P_a(t) \\ &= 1 - e^{-t\left(\frac{1}{m} + \frac{1}{u}\right)} \\ &= 1 - e^{-t/w} \end{aligned} \tag{38}$$

where w, the mean-time-to-withdrawal from the sample is found from

$$\frac{1}{w} = \frac{1}{m} + \frac{1}{u} \; . \tag{39}$$

It follows that the number which have not failed or aborted at time, s, where $s = t - t_{i-1}$ so that $s(t_{i-1}) = 0$ and $s(t_i) = T$, is

$$N(s) = N_i \, e^{-s/w}. \tag{40}$$

Then the average sample size within the i-th time interval is

$$\overline{N_i} = \frac{1}{T} \int_0^T N_i \, e^{-s/w} ds$$

$$\overline{N_i} = \frac{w}{T} \left(1 - e^{-T/w}\right) N_i \, . \tag{41}$$

Substituting $\overline{N_i}$ for N_i in equation (35) gives a better approximation to MTF:

$$m = w \left(1 - e^{-t/w}\right) \frac{\sum_i N_i}{\sum_i r_i} \, . \tag{42}$$

The estimate of w to be used in this equation is obtained from an equation analogous to (35):

$$w' = \frac{T \sum_i N_i}{\sum_i (r_i + a_i)} \tag{43}$$

For the numerical example considered above $\sum_i N_i = 389$, $\sum_i r_i = 62$, $\sum_i a_i = 38$ so that $w' = 3.89T$, $m' = 6.27T$, and $m = 5.54T$.

References: (a) A Revised Course in Reliability Theory and Practice, ARINC, Publication No. 123-7-196, Washington, D. C., 1960

(b) OEG Internal Research Memorandum 28, "FORTRAN Program to Estimate Cumulative Survival Probaiblity and Variance," 31 Oct 1962