220
10-26

ANCR-1024

1955
ANCR-1024
UC-32

MASTER

THE EVALUATION OF
FAILURE AND FAILURE RELATED DATA

BY
W. E. VESELY

Aerojet Nuclear Company

NATIONAL REACTOR TESTING STATION

Idaho Falls, Idaho — 83401

DATE PUBLISHED—AUGUST 1971

PREPARED FOR THE

U. S. ATOMIC ENERGY COMMISSION

IDAHO OPERATIONS OFFICE UNDER CONTRACT AT(10-1)-1375

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Printed in the United States of America
Available from
National Technical Information Service
U. S. Department of Commerce
5285 Port Royal Road
Springfield, Virginia 22151
Price: Printed Copy \$3.00; Microfiche \$0.95

LEGAL NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

ANCR-1024

ANCR-1024
Mathematics and Computers
TID-4500

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Atomic Energy Commission, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

**THE EVALUATION OF
FAILURE AND FAILURE RELATED DATA**

BY

W. E. Vesely

AEROJET NUCLEAR COMPANY

Date Published -- August 1971

PREPARED FOR THE U. S. ATOMIC ENERGY COMMISSION
IDAHO OPERATIONS OFFICE
UNDER CONTRACT NO. AT(10-1)-1375

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED
fey

ACKNOWLEDGMENTS

Acknowledgment is given to the Reliability Branch for their valuable assistance in the use and interpretation of the results obtained. W. J. Suitt, W. H. Rettig, R. E. Narum, and M. A. Lintner, all of the Computer Science Branch, developed the computer programs for the analyses.

ABSTRACT

Techniques for evaluating failure data are described which yield a wide range of significant results. The failure data need not be particularly accurate and can be quite indirect in form; trouble call reports, failure logs, maintenance histories and other such data sources generally suffice to give the required information. From these basic types of data sources, failure rates can be determined, wear-out and degradations can be uncovered, environments causing unwarranted failures can be detected, objective quantitative assessments can be made on repair and maintenance effectiveness, and sub-par performers can be located. These results are determined by means of quantitative evaluations which allow the uncovering of deficiencies before they are obvious. The particular techniques are discussed and data requirements and practical applications are given. Computer programs are briefly discussed to show results of actual evaluations and to show the capability of automating the techniques.

CONTENTS

	PAGE
ACKNOWLEDGMENTS	ii
ABSTRACT.....	iii
I. INTRODUCTION	1
II. THE DATA INFORMATION REQUIRED FOR QUANTITATIVE EVALUATIONS..	3
III. EVALUATIONS AND THE EXTRACTED RESULTS.....	6
A. The Obtainment of Failure Rates.....	7
B. The Determination of Abnormal Environments and Stresses.	14
C. The Assessment of Repair Effectiveness.....	17
D. Evaluations of Maintenance Effectiveness	21
E. The Detection of Upgrades and Downgrades.....	24
F. The Identification of Deviate Performances.....	30
IV. SUMMARY	36

FIGURES

1. Percent of time for failure free operation.....	12
2. Down time percentage versus average system repair time.....	13
3. Failure frequency versus time between failure for random and caused failures.....	14
4. Failure frequency versus time between failure for effective and ineffective repair.....	18
5. Percent of subsystem failures due to Mode A.....	22
6. Comparison levels for mean time between failure changes.....	26

TABLES

1. Mean Time Between Failure Results.....	10
2. Exponential Test Results Showing an Abnormality.....	16
3. Test Results Indicating Ineffective Repair or Replacement.....	20
4. Audit of Failure Changes.....	23
5. Results of the Performance Test Showing a Downgrade.....	27
6. Results of the Performance Test Showing both Upgrades and Downgrades.....	29
7. Deviate Performance Results.....	35

THE EVALUATION OF FAILURE AND FAILURE RELATED DATA

I. INTRODUCTION

In the general field of failure reporting, a wealth of data exists pertaining to failures of components, subsystems and systems. The reported failures include minor and major accidents, safety and reliability failures, and electrical and mechanical equipment failures. The failures incorporate a wide range of identification, ranging from drastic-type failures to maintenance-type troubles. These failures are recorded, sometimes in great detail, and are then stored, catalogued, cross-referenced, summarized, and in certain cases, filed in an automatic computerized information center.

This systematizing of past history is extremely useful for qualitative, detailed-type information retrieval. However, a significant amount of critical information is lost in this process since the data is rarely evaluated to extract quantitative results which objectively characterize not only the failure, but the causes of the failure. If any evaluation is performed, then failure rates are usually the only information extracted. This, however, represents only a fraction of the information obtainable. A more complete list of quantitative results which can be yielded by evaluation includes:

1. The obtainment of component and system failure rates.
2. The determination of abnormal environments and stresses causing unwarranted failures.
3. The assessment of testing and repair effectiveness including the detection of unwarranted burn-in failures.
4. The detection of degradations in component and system performance including the onset of wear-out.
5. The evaluation of maintenance effectiveness.
6. The identification of components or systems which are deviating in performance from their peers.

These results can be obtained from any data which records repetitive-type failures, that is, the failures can be repaired or if they are nonrepairable, then the components or systems failing have other identical or similar counterparts in existence. (It is rare to find a component or system which

does not have some similarities with other existing units.) Furthermore, the above results can be obtained not only from the drastic, consequential-type failure recordings, but can be extracted from the inconsequential-type failure records ("trouble" records), such as operating maintenance histories.

The above extractable information offers the reliability or safety engineer a powerful means of auditing, predicting and correcting failures. In particular, the extractable information offers the following potentials:

1. A means of objectively evaluating system performance with regard to reliability and safety.
2. A means of objectively assessing the efficiency of a maintenance and repair program.
3. A means by which deficiencies in existing systems can be pinpointed and corrected - in many cases before the deficiencies actually cause system failure^[a].
4. A means of quantitatively assessing the impacts of design modifications and maintenance changes.
5. Finally, a means by which realistic data can be obtained to be used to predict future reliability and safety performance.

These potentials are extremely practical and hence extremely attractive since they pertain to an existing, real life situation; the evaluations are not done on some mathematical model, but are done on a physical system or phenomenon in order to yield physically applicable results.

This discussion will attempt to describe the types of evaluations which yield the above quantitative information and will describe the applications of these results in order to realize the above potentials. In the evaluation of any failure data, the general objective is to extract physically significant behaviors from the random-type behavior which is recorded, that is, to extract the signal from the noise. If a conclusion is to be reached from the evaluation of the data, then one wants to have confidence that his conclusion is physically justified and is not based on a "random perturbation", particularly when that conclusion calls for an expenditure of manpower or money. This discussion will attempt to explain the methods by which the physical behaviors are isolated thereby yielding physical justification for action.

[a] This correcting of deficiencies before system failure is particularly applicable to redundancy situations, where component degradation does not immediately cause system failure.

Some of the reasons given for the existing data not being evaluated are that the data are quantitatively "incomplete", that they are not in a systematic format, and that too long of a recording period is necessary for significant quantitative results. These reasons generally have little substance. The recorded data can be quite simple with respect to quantitative information given; failure reports, maintenance histories, failure logs, and trouble call histories are usually sufficient to give most or all of the above list of evaluated information. In fact, the data can be quite "dirty" and contain only a percentage of true failures with the remaining data being unconfirmed trouble calls, erroneous failure reports, etc. (It is not absolutely necessary to even be able to separate out the true failures.) Furthermore, since rather simple data are needed, no sophisticated format is necessary for the extraction of useful quantitative results. In fact, in order to be immediately useful, the present diverse formats of the failure data must be admissible in a quantitative evaluation. This poses no serious problem since numerous techniques exist for reformatting and transforming the data before the evaluation is begun. Finally, with regard to the nature of data, an inordinately long period of failure data collection is not necessary to yield significant quantitative results; a year or so of recording time is usually sufficient and for certain of the results, several months are all that are required. This discussion will begin in its context by describing the type of simple information which must exist in the recorded failure data in order for quantitative evaluations to be performed.

II. THE DATA INFORMATION REQUIRED FOR QUANTITATIVE EVALUATIONS

The basic piece of data needed for a quantitative evaluation is the time of the occurrence of the failure^[a]. The precise minute and hour of the failure is not necessary and, in fact, is generally useless for the previously described evaluations. The day of the failure is usually the finest resolution needed and is only needed when failures occur frequently (on the order of once a week). A finer resolution given for the time of the failure, of course, poses no problem since the desired resolution can always be obtained. A favorable

[a] Here, and for the rest of the report, a "failure" means any "abnormality", covering the spectrum from a consequential failure to merely minor trouble.

circumstance in data recording is that the resolution needed for the time of the failure decreases as the number of failures decrease; for those failures which occur on the order of once a month the approximate day (plus or minus several days) is needed and for those failures which occur less frequently the week or month only of the failure is needed.

In addition to the time of the failure, the only other basic data needed are the identification of the failure. This identification should optimally be a categorized identification with succeeding finer resolution given in the lower category levels. The general nature of a categorized identification is given below:

1. The system in which the failure occurred.
2. The subsystem(s) of the failure occurrence.
3. The component suffering the failure.
4. The criticality and mode of failure.

For finer resolution of identification, the above categories would be subdivided into sublevels and for simple accessing, the categories would be keyed by some indexing scheme.

The categorized identification of the failures is optimal since it allows various detail in the evaluation, from a system to a mode of failure examination. However, any identifiers of the failure will yield the same type of results as listed in the INTRODUCTION. The resolution of the identification does not affect the nature of the results obtained nor does it affect the amount of information obtained. The resolution only affects the discriminating ability of the results, i.e., their "fineness".

As an example of the effect of identification, consider the case where failures are only identified to a subsystem level, where the subsystem is either functional or otherwise. The subsystem would be the basic unit evaluated and results would be obtained for the subsystem in exactly the same way as they would be obtained for a basic component. Failure rates, determinations of abnormal environments, degradations, repair and maintenance efficiencies, and deviate performances would still be obtained, but on a subsystem level. If one of the essential components of the subsystem had suffered a degradation, then the evaluations would show that one of the components had degraded. The evaluations would not disclose the particular component suffering degradation, but it would show that a degradation had occurred.

The important point of the above example is that the evaluations lose little of their potential because of the coarseness of failure identification. One usually desires to evaluate the data to as fine a level as permitted by the identifiers of the failures; however, at times too little data exist at the detailed level in order to make valid quantitative conclusions from the evaluators. The amount of detail allowed in an evaluation is automatically resolved since the evaluations incorporate statistical methods which determine the resolution permitted in order for definite results to be obtained. In essence, the evaluations automatically determine the amount of detail discrimination allowed. Later sections will elaborate upon this capability of the evaluations to determine the depth to which analyses can be performed.

With regard to the categorized identification listed earlier, it should be noted that if the components or systems are immediately replaceable or repairable, then the time of installation of the component or system is not necessarily needed as data for the evaluations. Here, "immediately replaceable or repairable" means that the failed component or system is replaced or repaired in a time shorter than its average time between failures. The installation date is only needed when the repair-replacement time becomes comparable to the average time between failures.

The installation date and manufacturers information are needed, of course, when evaluations are performed on a unique instrument level in order to determine differences in manufacturer quality, individual component performance, etc. However, when this information is not known, different units (components or systems) can be lumped together for the evaluations and unique identifiers are not necessary. Later sections will describe tests which, only from the times between failures, determine whether the failure characteristics have changed in any time period. If they have not, any new equipment is similar to the old and no distinction is needed. If they have changed, the test will tell this and will give the point at which the change occurred. The information that new equipment has been installed can then be used to uncover the cause of this change; it is however not needed in the basic evaluations.

Of a similar nature to the replacement or repair time is any period or periods of time in which the component or system is not operating. If this nonoperational time is smaller than the mean time between failure, then it is not needed. Even if it is on the order of the mean time, and it is not recorded, many of the evaluated results can still be obtained.

If preventative maintenance or test checking is being performed on the equipment the failure data and accompanying evaluations will automatically encompass this fact. The failure rates and other determined characteristics will include any maintenance schedule being adhered to, simply because the failure times reflect this maintenance (i.e., the failures are occurring, and are being recorded, under the particular maintenance "environment"). The actual maintenance or testing schedule is therefore not needed for the evaluations. This schedule is useful for additional double correlation tests which determine the effects of changes of this schedule on the reliability and safety. It is also useful in uncovering the causes of the various effects determined by the evaluations. However, knowledge of this schedule is not directly needed for the results listed in the INTRODUCTION.

From the above statements, it can be said in summary that any identification will yield pertinent results of the form listed in the INTRODUCTION. Additional identification detail will increase the discriminating ability of the evaluations and in this respect it is extremely useful, but it is not necessary. If additional detail is desired, and it is not in the failure records, the general evaluation methods have the attractive feature that this detail can enter at various stages of the analysis and can enter from other sources. In essence, any identification scheme or any failure record can be analyzed by suitable tailoring of the evaluations.

III. EVALUATIONS AND THE EXTRACTED RESULTS

The following sections briefly describe the types of evaluations which are used to obtain the extracted results listed earlier. Incorporated within certain of the evaluations are statistical tests which screen the noise from the physically significant behaviors, and the nature of these tests is also described. The sections also discuss the interpretations to be given to the results so that they can be transformed to physically meaningful actions. In all of the sections, the emphasis will not be on the mathematical details, but will be on the applications and particular powers of the evaluations.

A. The Obtainment of Failure Rates

The failure rate is the basic parameter which characterizes the reliability or safety of a unit (or incident). If λ is the failure rate and R is the probability that the unit will suffer no failure to time t , then R is simply given by the relation

$$R = e^{-\lambda t} \quad (1)$$

In reliability engineering, the quantity R is termed the "reliability" and is simply the percentage of time the unit will suffer no failure when it is operated for a period of t hours (or days). Where the failure is not associated with a unit, such as an incident, R simply gives the probability that the incident will not occur in the time period t .

Since t is arbitrary, knowledge of the parameter λ allows one to predict the future performance of the unit for any period t or allows one to predict the occurrence chance of an incident. Knowledge of λ allows more complex applications such as in fault tree and block diagram analysis. Because the failure rates are determined from actual operating experience and are not extrapolated from a "textbook", all of these applications have a physical meaning and pertain directly to the circumstance being analyzed.

The failure rate is related to the mean time between failure T by the equation,

$$\lambda = \frac{1}{T} \quad (2)$$

The quantity T is simply an average of the times between failure occurrences. One may start at any failure for this averaging, and hence, as discussed earlier, the installation date is not necessary when the repair or replacement time is small compared to T .

The above two equations serve as the basis for determining the failure rate from the failure data recordings^[a]. A complete statistical treatment shows that in order to determine λ any one of three types of data can be used:

1. The number of failure occurrences,
2. The time of the last failure,

or

3. The times between the failures.

[a] As stated earlier, the failure data recordings can be in the form of failure reports, maintenance histories, trouble call recordings, etc.

Corresponding to the three types of data used, three methods exist by which the failure rate can be determined. Each of the above methods has its own merits and advantages depending upon the particular circumstances; however, any one is sufficient to yield the failure rates. Every one of the methods yields an unbiased and sufficient value for the failure rate and every one will give the same results (within the experimental errors). The capability of determining the failure rate in a number of ways offers the advantage of the evaluation being able to be adapted to any peculiar data. Furthermore, any of the methods require simple information; for example, in the first method the number of failures is only required with no times of failure or times between failures needed. In fact, if no failures are observed, a failure rate is still obtained.

In addition to the failure rates, each of the above methods also yields confidence limits, or error bars, for the failure rates determined. Each of the errors determined are of the same order and for 90% confidence decrease from a factor of 20 if one failure is observed to a factor of 2 if five failures are observed. For more failures, the errors decrease to ten to twenty percent. The failures may occur in any time period with the errors essentially being dependent only on the number of failures observed.

For a large number of reliability and safety calculations, accurate values are not needed for the failure rates, and hence, large errors can be tolerated. If, for example, the safety of a unit is being calculated, then the concern is with the exponent of the result, i.e., whether the probability of failure is 10^{-2} or 10^{-6} . In this and in many computations, errors of the order of a factor of 10 can thus be tolerated and any data will usually yield the accuracy needed. In addition to the failure rates themselves, the 90% or 95% upper bounds of these failure rates are of importance since these upper bounds give the worst that can be expected. Any failures, or no failures, will generally give useable values for these upper bounds.

Two common situations in which accurate failure rates are desired are those of frequent failure occurrence, where the failures are not individually critical, but are collectively costly, and those situations where a large number of similar units are evaluated. These situations usually resolve the accuracy problems themselves since by their nature more data exist on failures giving accurate failure rates.

For those other situations in which accurate failure rates are desired, but little data exist batching techniques can be used. In a later section of this paper, tests are described which compare the failure characteristics of different components or systems. If these tests show no difference in characteristics, then the failures of a number of units can be lumped together to yield a more accurate failure rate.

For any circumstance, the failure rates can be subdivided into any level of detail desired. Failure rates for system failure, subsystem failure, component failure and mode of component failure can generally all be obtained from the same data record. These various levels of detail allow a complete scrutiny of the failures, from a system level to a mode of failure level. For these various types of failure rates, associated error bounds are always determined.

The above discussions, and the general utilization of failure rates assume that the failure occurrences follow an exponential distribution. The evaluations described later actually test this assumption to determine if it is valid. If it is, as is the usual case, then the failure rates can be used without any limitations. If it is not a valid assumption, these evaluations can be used to determine which parameters and hence which distributions actually apply. These new parameters (such as the Weibull parameters) can then be used in the reliability and safety calculations analogously to the way the failure rate is used.

Table 1 depicts the type of failure rate printout which is yielded by a straightforward computer program. In this particular instance, the mean time between failure was output (the failure rate is simply the inverse of the mean time between failure). Actual failure data records were evaluated by the computer program and the table illustrates part of the results of this evaluation. The format of the table is analogous to the format of the computer output. In the run approximately 1400 component failure records were evaluated. (The computer time was negligible for the run; on an IBM 360/75 computer a total of 40 seconds was required.)

In Table 1, the "Group Number" and "Unit Number" denote the particular indexing used in the failure record. The group number identifies the subsystem and the unit number identifies the component within the subsystem.

TABLE 1. MEAN TIME BETWEEN FAILURE RESULTS (in days)

Group No. 28

Unit No.	1969	1970	1971 (3 Mo.)	2.25-Year	1970 (90%)	1971 (90%)	2.25 (90%)
06	121.7	45.6	90.0	68.3	(25.3, 91.7)	(19.0, 1754.7)	(42.2, 118.4)
07	121.7	91.3	45.0	91.1	(39.9, 267.1)	(14.3, 253.3)	(52.2, 174.6)
08	121.7	45.6	90.0	68.3	(25.3, 91.7)	(19.0, 1754.7)	(42.2, 118.4)
09	91.3	60.8	90.0	74.5	(30.8, 139.7)	(19.0, 1754.7)	(45.0, 132.9)
10	60.8	33.2	45.0	43.2	(20.0, 59.2)	(14.3, 253.3)	(29.4, 66.0)
11	91.3	36.5	90.0	54.7	(21.5, 67.3)	(19.0, 1754.7)	(33.5, 88.7)
12	365.0	121.7	90.0	273.3	(47.1, 446.4)	(30.0, -----)	(105.8, 1002.8)
13	121.7	36.5	12.9	41.0	(21.5, 67.3)	(6.8, 27.4)	(28.1, 62.0)
14	60.8	73.0	90.0	68.3	(34.7, 185.3)	(19.0, 1754.7)	(42.2, 118.4)
15	365.0	121.7	90.0	205.0	(47.1, 446.4)	(30.0, -----)	(89.6, 600.2)
16	182.5	60.8	90.0	91.1	(30.8, 139.7)	(19.0, 1754.7)	(52.2, 174.6)

The first three columns give the mean time between failure for the corresponding component for 1969, 1970, and for the first three months of 1971. The column denoted by "2.25-Year" gives the average mean time between failure for this 2.25 year period. The last three columns give the 90% confidence bounds (error bounds) for the mean time between failures. (The error bounds for 1969 were not requested in the computer printout.) A dash in an error bounds column denotes the fact that no failures occurred in the particular period and hence no upper bound could be obtained. As noted from the table, the mean times between failure give the most concise description of any failure behavior.

To conclude this section, certain basic uses of the failure rate will be described in order to show its potential power. Figure 1 on the next page depicts the use of the failure rate to obtain the percent of component failure free operation versus operating time. The percent of failure free operation is simply the probability of no failure, R , given by Equation (1). From the figure, for example, if the component is operated for 60 hours, then it has a probability of approximately 92% of not failing in this 60 hours (the "best estimate curve"). The 90% upper limit and lower limit curves merely depict the possible error spread in the calculations (due to the possible errors in the failure rate). These results were obtained by using the mean times between failure which were printed out and described in the preceding paragraph.

Figure 2 illustrates the use of the failure rate to determine the down-time percentage versus the repair time needed. The failure rate used in this computation was again taken from the computer printout described earlier. In this case, the failure rate was a system failure rate.

From the figure, if it requires two hours to detect and repair a failure (on the x-axis), then the system will be down 0.45% of the time (the y-axis value). Consequently, for this two hour repair and detection time, if the system is needed at a particular time (such as for an incident occurrence), then there is a probability of 0.45% that the system will not be operational at this time. Equivalently, for this repair and detection type, the system is available 99.55% of the time (the up-time percentage). For different values of the repair and detection time, the down-time percentage would have different values as obtained from the figure. This type of failure rate utilization is pertinent when the effect of speed of repair is investigated or when an increase in up-time percentage is desired.

Average Detection Plus Repair Time (Hours)

In addition to the above example, the failure rate can be used in a number of other practical applications. Knowledge of the failure rate allows one to predict the number of spare parts which should be kept in inventory. Criticality of components, maintenance program evaluations, manufacturer comparisons, and effects of possible design changes can all be investigated with knowledge of the failure rates. Possible errors in the failure rates pose no problem since they are known and hence can be incorporated in these investigations. In essence, the failure rates quantify past failure experience, allowing quantification in further investigations. The quantification is meaningful since actual operating experience forms the basis for these applications.

B. The Determination of Abnormal Environments and Stresses

The statistical test described in this section determines any failures which are not occurring randomly. From basic reliability and safety theory, if failures occur randomly, then the times between failures follow an exponential distribution. These random failures are to be expected and have no prescribable cause as to their occurrence. If the failures are not random and are due to some physical cause, then departures from the exponential will be observed. For example, if the failures are due to some abnormal environment condition (such as heat build-up), then the times between failure will show peak characteristics in its distribution.

FIGURE 3. FAILURE FREQUENCY VERSUS TIME BETWEEN FAILURE
FOR RANDOM AND CAUSED FAILURES

For the statistical test, termed an "exponential test", times between failures are needed as the only data. On the average, ten failures are needed in order to determine abnormal causes for the failures. However, for more extreme abnormalities, four failures suffice to show these causes. The statistical test, which is essentially a rejection-type test, screens the expected perturbations from the unexpected, caused perturbations.

Even if the failures are occurring randomly, the distribution will not exactly follow an exponential. One can expect "bouncing" in the data, some failures occurring almost together and others occurring quite far apart. The merit of the exponential test is that it screens this noise-type of bouncing from the deviations which are abnormal and have a physical cause. In the test, the confidence levels are systematically varied so that a ranking of departure is obtained. The most severe abnormal departures are obtained first, then the lesser departures are yielded, with the test automatically giving the degree of severity.

In practical application, the exponential test is used to determine abnormal conditions which exist and which cause unwarranted failures. Stresses within the component or system, nondesigned operating environments, and in general, any abnormal impressed condition can be uncovered by the test. The test is also useful in determining whether a failure rate (λ) validly defines the failure occurrences; for these failures which follow an exponential distribution, the failure rate can be used in further applications as discussed in the previous section. The exponential test can be applied to any category of failure identification, allowing structured levels of investigation. For the abnormal conditions uncovered, the reliability or safety engineer may decide that these abnormalities (exponential departures) must be lived with and are to be considered "normal", such as Weibull behaviors. In many cases, however, the abnormalities can be simply corrected giving substantially improved reliability or safety.

Table 2 shows the results of the exponential test applied to physical failure data. The results were extracted from the output of a computer program which automatically tests all the data in the failure records. Each of the individual failure records is classified as to whether the failures are occurring randomly or not, with added information given for the nonrandom cases. For the particular case illustrated in the table, maintenance was also concurrently

TABLE 2. EXPONENTIAL TEST RESULTS SHOWING AN ABNORMALITY (90% Confidence)

Group No. = 39 Unit No. = 46
Total Deviation = 1.4 (8 Failures)

End of Intervals (days)	Expected Percentage	Observed Percentage	Deviation
8.3	14.3	0.0	0.0
18.0	14.3	14.3	0.0
30.0	14.3	0.0	0.0
45.4	14.3	14.3	0.0
67.1	14.3	57.1	6.3
104.2	14.3	14.3	0.0
-	14.3	0.0	0.0

being performed on the component. Eight failures had occurred over a two year period and an abnormality was detected.

In the table, the "Group Number" and "Unit Number" identify the subsystem and the particular component within the subsystem. Thus, an abnormality condition was detected for Component 46 in Subsystem 39. The departure from an exponential distribution was significant at the 90% confidence level and the "Total Deviation" (1.4) simply gives the magnitude of the departure (no departure is zero deviation). This deviation number can be used to further rank the abnormalities obtained. In the computer program, the times between failure are grouped into various intervals and the "End of the Intervals" column gives the interval limits. Thus, the failures occurring from 0.0 to 8.3 days apart were grouped together and counted, those occurring from 8.3 days to 18.0 days apart were grouped and counted, etc. The dash symbol in the last column denotes that the limit of the last interval is infinity.

From the numbers of failures occurring within each interval, the observed distribution of the times between failure was obtained. The intervals were chosen such that the expected percentage was the same for each interval (14.3%). The expected percentage was that which would be obtained if an exponential distribution was followed. The observed percentage in each interval is shown in the third column, which was the actual distribution followed. The deviation

of the observed from the expected percentage in each interval is shown in the last column.

The deviation number for each interval denotes the amount that interval departs from an exponential distribution. These deviation numbers can be used to rank the abnormalities of the various intervals. For no significant deviation from an exponential behavior, this number is zero. Even if the observed percentage does not coincide with the expected percentage, this number will be zero if the difference is within the noise level (the limit of natural data bouncing). As observed in the table, the abnormality thus exists within the interval of 45.4 days to 67.1 days. Of the total number of failures, 57% occurred from 45.4 to 67.1 days apart compared to 14% which should have occurred if the random, exponential distribution was followed. This difference was beyond the natural bouncing behavior and thus was physically caused.

The exponential test consequently showed that this particular component was experiencing an environment which caused it to depart from a random failure behavior. The environment incurred a peaking in the interval from 45.4 days to 67.1 days, causing four times more failures than expected. Correcting this environment maximally would reduce the number of failures occurring in the interval by a factor of four which in turn would reduce the total number of failures occurring by 40%.

C. The Assessment of Repair Effectiveness

The exponential test of the previous section can be simply extended to determine ineffective repairs and unwarranted burn-in failures. From basic precepts of reliability theory, if the repair or replacement is effective, then the time between failure distribution will be exponential. Similarly, an exponential distribution will result if excessive burn-in failures ("sudden deaths") are not being experienced. For ineffective repair or excessive burn-in, on the other hand, the distribution will exhibit a peak for small values of the times between failure.

FIGURE 4. FAILURE FREQUENCY VERSUS TIME BETWEEN FAILURE
FOR EFFECTIVE AND INEFFECTIVE REPAIR

If the failures are repaired when they occur, this initial peak in the distribution simply means that one is having to come back too soon to repair the item again because it was not effectively repaired the first time. This ineffective repair may be due not particularly to actual repair inefficiency, but to undiagnosed troubles or may be due to hardware effects such as poor quality parts, unusually severe usage, etc. If the items which fail are replaced instead of repaired, then the initial high values in the distribution depict a high percentage of burn-in failures occurring, such as when the items have not undergone burn-in within the initial quality control program. In either case, repair or replacement, the large number of failures experienced within short intervals denotes an ineffectiveness in operation which can often times be corrected.

To test whether this ineffectiveness exists, on the order of three or more failures are needed. Of course, the more failures the finer is the discriminating ability of the test (which is a virtue since the more often occurring failures are those toward which upgrade goals are usually directed). From the failure recordings, the times between failures are the only values needed and are used in an analogous way as for the exponential test. The failures which occur within short times of one another are separated from the

rest of the failures and are tabulated. This separation is performed by extracting those times between failures which are smaller than some maximum value which is the indicator of effective repair. This maximum value is obtained directly from the data and is not any "handbook" value.

After extracting these more recurrent-type failures, a comparison is made of the number which actually occur to the expected number which should occur. The expected number is again obtained directly from the data. If the actual number is significantly larger than the expected number, a repair ineffectiveness is flagged. As for the exponential test, a statistical screening is incorporated in the evaluation to determine the true departures from the random perturbations. The confidence levels in the screening are again systematically varied to obtain a ranking of the ineffectiveness.

This test for repair weakness is quite flexible since it can be applied to a variety of failure identification levels. Applied to a mode of failure identification, the test will depict those types of failures which are most ineffectively being repaired. Applied to subsystem failures, the test has the ability of uncovering failures being blamed on the wrong component (this component is corrected and the system soon fails again due to the true cause not being corrected). In all of these applications, the test is intended to aid engineering, quality control, and maintenance by uncovering areas which can be improved.

When a repair defect is uncovered, the particular operational circumstances may dictate that this ineffectiveness cannot be corrected and is a "fact of life". This is the case, for example, when repair is very difficult and several attempts are needed (such as O-ring alignments). It is also the case when burn-ins must be tolerated (e.g., when destructive testing is the only means of investigation). The test results will simply point out these difficulties. In a number of other cases, however, the ineffectiveness can be corrected to yield sizeable returns, particularly since the correction will significantly increase the mean time between failure.

Table 3 shows the results of the test when applied. The evaluation was part of the automated program described in the previous section. All the failure records were analyzed as to the repair-replacement effectiveness and the ineffective repair-replacements were signaled. (The effective repair-replacements were also denoted.)

TABLE 3. TEST RESULTS INDICATING INEFFECTIVE REPAIR OR REPLACEMENT
(90% Confidence)

Percentage of Failures Occurring Within Three Days Apart	
EXPECTED	OBSERVED
14%	53%

Mean Time Between Failure Consequences	
EXPECTED	OBSERVED
33.2 Days	19.2 Days

For this failure record, a three day interval length was chosen by the program as the maximum time between failure value (the ineffective repair indicator), and the failures occurring within three days apart were extracted from the record. The expected number which should occur within three days was then next computed from the entire number of times between failure. This was then compared to the actual number which had occurred.

Of the total number of failures occurring, 53% had times between failure less than or equal to three days; this meant that in 53% of the repairs, one had to return within three days because the item had failed again. From the data, one should only expect 14% of the repairs having to be repaired again within three days. The test determined that this discrepancy was in fact real and an ineffectiveness existed. As in the exponential test, the deviation was also given in order to further rank the ineffectiveness.

To illustrate the effects of correcting this ineffectiveness, the test also gives the consequences in mean time between failure. The observed mean time between failure is 19.2 days which incorporates the degradation due to the large number of failures occurring within three days apart (these failures have approximately three days as their mean time between failure). Correcting the ineffectiveness would give 33.2 days as the mean time between failures, which is the expected value. Thus, correcting the repair ineffectiveness would not only decrease the failures by 38% (from 53% to 14%), but would also approximately double the performance of the unit (from 19.2 days to 33.2 days mean time between failures).

D. Evaluations of Maintenance Effectiveness

Statistical evaluations can be performed on the failure data to answer the following questions: "Has a change in maintenance been effective in reducing failures"?, "Have the number of failures significantly increased or decreased in a given period"?, and "Has a design change effectively reduced failures"? Since failure behavior exhibits noise-like characteristics, simple observations cannot always be made to obtain answers to these questions. Even though a unit is maintained at constant performance, with no upgrade or downgrade occurring, it can suffer more failures in one period than in another period. This is true simply because of the random nature of failure behavior. This random-like behavior occurs to even more of an extent when maintenance is being performed on the unit; at times maintenance is better than average and at other times it is worse than average.

Using the failure data, a statistical test can be performed to determine if any failure changes are beyond the "noise limit" of behavior. If significant changes are detected, then the changes are interpreted as being physically caused; a true increase or decrease has occurred. The statistical test thus is an objective means of measuring the true changes from the apparent, random changes.

The number of failures occurring in different time periods is the only data needed for the test. The particular test which uses these data is a straightforward chi-squared test in which a comparison is made of the changes in failure number. If the change is greater than a particular noise value (confidence value), then a true increase or true decrease has occurred. This confidence value therefore acts as a filter in screening the noise from the true changes and can be systematically raised or lowered to obtain the criticality of the increase or decrease. The most significant changes will be first obtained with the lesser changes being next obtained and having a lower ranking.

Figure 5 illustrates the percent of time a subsystem failed in a particular mode in the years 1969 and 1970. (The particular mode was given the keying title of "Mode A".) In 1969, 71% of the subsystem failures were due to this mode of failure and in 1970, 21% were due to this mode.

FIGURE 5. PERCENT OF SUBSYSTEM FAILURES DUE TO MODE A

The chi-squared analysis showed that this was a significant upgrade and was not simply a random decrease in the percentage of failures. Tracing the cause for this upgrade showed that in 1970 tube-type units were replaced by solid state devices. The test thus showed that this replacement caused significant improvement, decreasing the failures by more than a factor of three.

In the above example, the change was perhaps obvious; however, the test substantiated that the decrease was real. There are times when apparent changes are simply fluctuating-type behavior in the number of failures, and no physical change has actually occurred. For these apparent, but unreal changes, if one recorded data longer, then the fluctuations would eventually settle to the same average value. The chi-squared test is an objective method in which existing data need only be used to separate the true failure changes from these unreal changes.

Table 4 shows the chi-squared test applied to the yearly number of failures occurring in a system. The results are extracted from a computer run in which all the failure records were analyzed.

TABLE 4. AUDIT OF FAILURE CHANGES
(to 90% Confidence)

KEY

- * Significant Decrease In Failures From 1969 to 1970
- / Significant Increase In Failures From 1969 to 1970
- ↑ (No Symbol Denotes Steady State Behavior)

Unit No.	No. of Failures in 1969	No. of Failures in 1970	Change From 1969 to 1970
01	5	0	↓
02	5	0	↓
03	3	1	
04	4	1	
05	6	15	↑
06	2	1	
07	0	1	
08	0	4	↑

The unit numbers in the table are individual component identifiers for this particular failure record. As observed from the table, the true changes are straightforwardly differentiated from the noise behaviors. For example, Unit 03 failures decreased from three in 1969 to one in 1970, but the test showed that this could not be attributed to any true decrease. No significant change was therefore given to the change in number of Unit 03 failures. In addition to the above failure change evaluations, the program has the capability of also obtaining the dominant failure contributors of a subsystem, or system, i.e., those which are causing more than their share of failures.

Used in the above surveying-type manner, the test serves as a simple tool for auditing the number of failure occurrences, showing the dominant failure contributors, the significant increases, decreases and steady state performances. The "bad actors" (increases and dominant contributors) are flagged for possible further investigations and the "good actors" (decreases and steady actors) show effective modifications and effective maintenance.

A computer program, such as the one used, can give the audit yearly, quarterly, monthly, or at any other periodic interval. This audit gives an up-to-date report on failure behavior and enables both manager and engineer to stay atop the overall performance picture.

E. The Detection of Upgrades and Downgrades

During the course of operation, certain of the components and systems may suffer a degradation such that their chance of failure increases (for example, components wearing out). These units will subsequently start to fail more often. Where the failure is not hardware associated, but is a general incident occurrence, the degradation may take the form of conditions changing such that the chance of the incident occurring increases. Whether the degradations are hardware associated or incident associated, when such degradations occur, performance and safety decrease and cost expenditures increase. From both an economic and safety point of view, these degradations need be identified and if the degradation is critical, they need be corrected.

Of equal importance to downgrade detection is the identification of upgrades in performance. Due to some change in maintenance practice, some new type of unit installed, or some change in environment, less failures are being experienced. The detection of an upgrade determines that this change is indeed beneficial and indeed improves performance. The upgrade evaluations thus allow one to objectively assess the value of any change with respect to its effect on the failures experienced.

The returns obtained in identifying downgrades and upgrades are thus quite obvious. The problem with identifying these downgrades and upgrades again lies in the random-type behavior of failures. There will be times when failures occur more often or less often and a true degradation or upgrade has not occurred. In these instances, simple inspection of the data would simply lead to erroneous conclusions. The test described in this section is an objective means by which true downgrades and upgrades are confidently identified and are separated from noise-like fluctuations.

In the test, termed a performance test, allowance is made for natural fluctuations in the failure occurrences. However, when the changes in failure

occurrence exceed a certain maximum "noise level", then a degradation or upgrade is signaled. This signaling simply says that the experienced change is beyond that which would be expected if a steady state performance was existing. Hence, the change is physically caused and represents a true upgrade or downgrade with the test telling which occurred. The performance test also gives the magnitude of the downgrade or upgrade and gives the approximate time at which it occurred. The test is valuable in that it is capable of detecting changes which are just beginning to occur. If these changes are degradations, then they can be corrected before causing a more disastrous effect, such as a plant shutdown or a major accident.

For the test, the times of the failure occurrences are used as the basic data. On the order of six failures are needed as minimum data. These failures may have any detail of identification; if the failures are classified only to a subsystem level, then the subsystem performance will be evaluated for a downgrade or upgrade. If the failures are identified to a component mode of failure level, then the performance of this mode of failure will be evaluated. The performance test is thus quite versatile in its range of applicability.

The failures analyzed may be repairable or they may be nonrepairable (in the nonrepairable case, the failed units are thus simply replaced by new units). Furthermore, the evaluations are independent of whether preventative maintenance is being performed on the units. Hence, any maintenance schedule is not needed in the analysis. Where units are replaced, the installation date is not necessary. All of this additional information is not needed since the only data required are the times of the failures. These are the only data which will show any downgrades or upgrades. When a downgrade or upgrade has been detected by the test, this additional information will then be required to help uncover the cause of the change.

As an example of the use of this additional information, consider the case where a downgrade has been detected. If no intermittent maintenance is being performed on the unit, then a hardware wear-out is indicated by the downgrade detection. If preventative maintenance is being performed, the degradation may be due to the maintenance itself and not directly due to the hardware deterioration. If a new unit had been installed the degradation can, in cases, be traced to this new unit being defective. In all of these instances, the same degradation occurred and was detected from the same data,

the times of the failures. The additional information further identifying the failures did not enter into the degradation detection, but entered into determining the causes of the degradation.

The performance test used in detecting downgrades or upgrades is a straightforward adaption of the standard statistical F-test. From the failure times, the mean times between failure are determined for various operating periods. The change in mean time between failure is then compared to the maximum change which would be expected under a steady state failure occurrence. If the observed change is greater than this maximum change, a downgrade or upgrade has occurred.

FIGURE 6. COMPARISON LEVELS FOR MEAN TIME BETWEEN FAILURE CHANGES

The maximum change (i.e., the maximum noise level value) is determined from the confidence one desires in the evaluations (usually 90% or 95%). In reliability terms, this confidence simply denotes the amount of noise attenuation; for a 90% confidence, 90% of the noise is eliminated. One is then 90% certain that the change is not due to a noise fluctuation. By systematically lowering the noise level values, a ranking is obtained of the criticality of the downgrades and upgrades. The downgrades and upgrades obtained with the highest noise level value will be the most significant and those obtained at lower noise levels will be accordingly less significant. The downgrades obtained at the lower noise levels are of importance since they are just beginning to occur and therefore can be corrected in this early stage.

If a combination of several downgrades and upgrades has occurred, the test will indicate each of these changes and the approximate times at which they occurred. As for a single change, the approximate time is determined from the times of the failures which separate the downgrades and upgrades. (This will be further illustrated in the examples.) The detection of several

changes is important since it gives a continuous history of the performance changes. (For example, an upgrade experienced after an initial downgrade was effectively corrected.) Finally, the test is capable of differentiating step upgrades or downgrades from continuous-type changes. This particular ability is useful for classifying the type of failure changes and aids in the uncovering of the cause of the downgrade or upgrade. For example, a continuous downgrade indicates a hardware wear-out or environment change while a step downgrade may indicate an ineffective maintenance at a particular time.

Table 5 shows the results of the performance test applied to actual failure records. The table was extracted from the output of a computer program which analyzes the failure records of a data source and detects any downgrades or upgrades. Any components and subsystems which have changed in their failure behavior are flagged and information is given about these changes. The program is quite flexible and can be applied to any type of failure records whether they be hardware associated or not.

TABLE 5. RESULTS OF THE PERFORMANCE TEST SHOWING A DOWNGRADE
(90% Confidence)

Group No. = 48

Unit No. = 66

	Failure Times M/D/Y	Times Between Failures (Days)	Type of Change ^[a]	Ratio of Change	T ₁ (Days)	T ₂ (Days)
1	5/19/69					
2	5/28/69	9				
3	6/17/69	20				
4	11/10/69	146				
5	11/12/69	2				
6	5/11/70	180	Downgrade	3.88	71.4	18.4
7	6/07/70	27				
8	7/01/70	24				
9	7/02/70	1				
10	7/07/70	5				
11	8/11/70	35				

^[a]No symbol in the column denotes steady state behavior.

In Table 5, the Group Number (48) identifies the subsystem in which the component is located and the Unit Number (66) identifies the particular component. This indexing scheme was used in the particular failure records analyzed, the program is, however, simply adaptable to any identification scheme. In addition to the information given in Table 5, the computer program prints out the component code name and any further failure identifiers. The format of the printout is also a bit plainer than Table 5 since page space necessitated a compressing of this computer information.

The failure times in the table denote the times of occurrence of the failure, given as the month of the failure (M), the day of the month (D), and the year (Y). The failures are numbered sequentially and these are the left-most numbers. The times between failures (days) are simply the intervals of time between the subsequent failure occurrences. For this component, the test detected a significant downgrade in performance. This downgrade resulted from a comparison of the mean time between failure for the first six failures to that for the remaining five failures, and this is the reason the "downgrade" symbol is beside the sixth failure. For the first six failures, the mean time between failure (T_1) was 71.4 days and for the remaining five failures, it was 18.4 days (T_2). If one would assign a beginning time to the degradation it would thus be between the six and seventh failure occurrence. The ratio of change in the table is the ratio of T_1 to T_2 . The mean time between failure thus had decreased by a factor of 3.88 which meant that on the average the failures were occurring 3.88 times as often as before. The factor of 3.88 was greater than the noise level value and hence the change was a true degradation.

In this particular instance the performance test thus signaled a component undergoing a degradation from the many component failure records analyzed. The magnitude of the degradation was given which was a decrease of a factor of 3.88 in mean time between failure. The test determined that this decrease was not simply a noise fluctuation and hence one could now take action, confidently knowing that the degradation was real.

Table 6 illustrates the results of the performance test when a combination of upgrades and downgrades is detected. The symbols in Table 6 are the same as those used for Table 5. The symbol T_1 always denotes the mean time between failure for the earlier occurring failures and T_2 for the latter occurring

failures. The numbers in parenthesis beside the "Upgrade or Downgrade" denotations simply give the rank (significance) of the change, that is, how far it is beyond the noise level. The computer program also gives the magnitude of the departures from the noise level which can be used for finer rankings (these are not given in the table).

TABLE 6. RESULTS OF THE PERFORMANCE TEST SHOWING BOTH UPGRADES AND DOWNGRADES (90% Confidence)

Group No. = 2 Unit = 13

	Failure Times M/D/Y	Times Between Failures (Days)	Type of Change	Ratio of Change	T ₁ (Days)	T ₂ (Days)
1	4/03/69					
2	4/18/69	15				
3	4/22/69	4	Upgrade (2)	5.77	9.5	54.8
4	11/07/69	199				
5	3/17/70	130	Downgrade (5)	2.64	87.0	32.9
6	3/19/70	2				
7	4/02/70	14				
8	9/11/70	162	Downgrade (3)	3.48	75.1	21.6
9	9/21/70	10	Downgrade (4)	2.85	67.0	23.5
10	9/28/70	7	Downgrade (8)	2.25	60.3	26.8
11	11/16/70	49	Downgrade (6)	2.79	59.2	21.2
12	12/16/70	30	Downgrade (7)	3.08	56.6	18.3
13	2/06/71	52	Downgrade (1)	37.44	56.2	1.5
14	2/08/71	2				
15	2/09/71	1				

In the table, an initial upgrade was experienced at the third failure. This meant that for the first three failures, the mean time between failures ($T_1 = 9.5$ days) was significantly smaller than the mean time for the remaining twelve failures ($T_2 = 54.8$ days). The change in the mean time between failure was 5.77 which was larger than the maximum noise value as determined by the test. Thus, the mean time between failure increased by a factor of 5.77 and compared to initial operation, failures were occurring 5.77 times less frequently.

After this initial upgrade, there was a more or less steady state of operation from the third failure to the eighth failure (encompassing approximately one year of operation time). One downgrade was experienced in this period, but otherwise no changes were experienced (where there is no "Downgrade" or "Upgrade" signaled in the table, no change is experienced and a steady state performance exists). After the eighth failure, a steady, continuous degradation set in.

At the thirteenth failure in this deterioration period, the maximum downgrade occurred (corresponding to (1) in the ranking beside the downgrade). For the thirteen failures, the mean time between failure was 56.2 days and for the remaining failures, the mean time was 1.5 days. The failures were thus occurring 37 times as often as previously. The component was in its maximum degraded state and remained there for the last few failures (hence there were no further downgrades for failures 14 and 15).

For this component, a complete history of its performance was thus obtained. The upgrade, steady state period, and continuous downgrade was explicitly defined by the test (it is interesting to note that this performance is characteristic of the common "bathtub" behavior, i.e., upgrade, steady state, and downgrade). Without the test, any such behavior would not have been uncovered. With the test it is now obvious that the component need be corrected or replaced. In fact, if the test were being concurrently performed as the failures were occurring, the first several downgrades would have signaled a deterioration which could have been corrected at an earlier stage.

F. The Identification of Deviate Performances

The last test described in this paper is one of the most versatile tests for evaluating failure data. From a selected group of components or systems, the test will determine those particular units which are behaving differently from the rest. Those units which are significantly worse than the others and those which are significantly better than the others will be obtained. These deviate performers will be identified because of their failing more frequently or less frequently, but will also be identified because of their different failure behavior. The deviate identification due to frequency of failure allows

one to locate the good performers and bad performers of the group, those which have better or poorer mean times between failures. The deviate identification due to behavior allows one to investigate the actual distribution of failures, enabling one to determine how units are failing and to determine the causes for their different behavior.

If general incidents are of concern which are not directly hardware oriented, the test may still be used. From a collection of incident occurrences, the test will differentiate those incidents which are occurring more frequently or less frequently than the others and will determine the different behaviors of occurrence. In all of the differentiations, whether hardware associated or not, any collection of items may be selected for the comparisons and any time periods may be used in the comparisons (the use of different time periods allows one, for example, to compare behaviors of the same unit at different times). As in all the previous tests, numerical characteristics are obtained which effectively describe the deviate performances.

The identification of deviate performers sounds simple; one merely needs to examine the failure data and "pick out" the differences. It is not this simple, however, because of the random fluctuations in data with which all the previous tests, as well as this one, have been concerned. As before, simple examination of the data will only lead to erroneous conclusions. In this case, the natural data jumpings will lead one to erroneously select differences which in fact are not true differences, but are simply due to natural fluctuations in one set of data. Analogous to the previous tests, the test which identifies the deviate performers takes into account the natural fluctuations and separates these from the true differences which may exist.

For the deviate performance test, a group of items are first selected for the comparisons, where the items may be components, subsystems, incident occurrences, etc. Each item in the group, representing one series of failures, is then compared with the other items of the group. The failure times of each item are used as the data in the comparisons. A minimum of two failures is needed for each item and the items compared may each have a different number of failures. As is the general case for all of the tests, the more failures recorded, the finer is the resolution of the test.

In the first phase of the test, the failure times of each item are used to compute the mean time between failure for that item. This mean time between failure is compared with the average mean time between failure for all the other items of the group. If the mean time between failure of the item is significantly different than the average for the other items, a deviate performance is signaled. The test cycles and performs the above comparison on each item in the selected group. A noise screening is built into the test which accounts for the natural fluctuations in the data. This noise screening works in the same manner as previously described for the other tests. Upon signaling of a deviate performer, the test gives quantitative details about the deviation, such as the mean times between failure for the performer and for the rest of the group.

If, for example, each item is a component, then the above evaluation would be made on the various components in the group chosen to be intercompared. The mean time between failure of one component would be compared with the average of the mean times between failure of the other components. If the component mean time between failure is significantly different than the average for the other components this component will be signaled as a deviate performer. Upon cycling and comparing each component in the above manner, the test will yield those components with significantly lower mean times between failures (the bad actors) and those with significantly higher mean times between failure (the good actors). It will also yield those components which have essentially equal mean times between failure.

In the second phase of the deviate performance test, the failure times of each item are used to compute the failure probability distribution for that series of failures. This distribution is compared with the average distribution of the other items in the selected group. If the difference in distributions is larger than the natural fluctuation limit, a deviate performer with regard to failure behavior is signaled. Each individual distribution is evaluated in this manner until all the items (failure histories) are processed which are to be compared.

For the previous example of each item being a component, the distribution comparison would give those components deviating in behavior from the rest. These deviate performers are components which have a different failure

distribution as compared to the average distribution of the other components. The different distributions show either a difference in the units themselves, such as differences caused by quality defects, or show a difference in environments experienced, such as stresses being experienced by one component.

The deviate performance test has a countless number of uses. Similar components of different manufacturers can be compared for quality control checking. Subsystems with the same functional output can be compared to detect sub-par performance. Similar components or subsystems can be compared to determine any different environments being experienced. A group of parallel functioning units can be intercompared to determine any bad actors. If the items compared are failures of one unit at different time periods, comparisons can be made to detect any changes in failure behavior with respect to time. If different failure histories are to be coalesced to yield numerical results with less statistical error (such as failure rate determinations), the deviate performance test can even be used to determine whether the histories are indeed the same and can be coalesced.

In all of the various applications of the deviate performance test, the fundamental principle is the determination of deviate items from a selected group of items which is intercompared. The group which is chosen to be intercompared need not always have basic similarities, such as components having the same functional outputs. Completely different units may wish to be compared to determine any deviations. One application of this mixed grouping is the determination of the weakest links in a system, e.g., selecting out those components having the smallest mean time between failure.

An advantage of the test is that the group can be of any size; for example, two components may be intercompared or two hundred components may be intercompared. Each of these components, furthermore, may have a completely different number of failures recorded. For more than two quantities compared each individual quantity is always compared to the average of the remaining quantities (the average not including the one which is being compared). Since the size can be arbitrary, a systematic variation can be made of the number of quantities compared. A large group may first be intercompared and after finding the deviate performers these in turn may be intercompared to find the worst deviates.

Table 7 illustrates the output of a computer program which utilizes the deviate performance comparisons. In this particular instance, comparisons were made of similar components located in the same and in different subsystems. The test was performed to identify any bad actors, either due to hardware defects or to adverse environments experienced. The input to the program is quite simple, requiring simply the components to be compared. The program has the flexibility of performing any number of intercomparisons and allows any number of quantities in each intercomparison. For each intercomparison, the program cycles through all the quantities to be compared and prints out a simple, legible output as illustrated in the table.

In the table, analogous to the previous examples, the "Group Number" identifies the subsystem and the "Unit Number" identifies the individual component within the subsystem. The group number here is simply another identifier of the failure history and does not refer to the group being inter-compared. One intercomparison (and hence one "grouping") is being evaluated. The "Individual MTBF" is the individual mean time between failure for the particular component and the "Average MTBF" is the average mean time between failure for the remaining components. The mean time between failure is in units of days. The "MTBF Ratio" is simply the ratio of the individual to the average.

In the column labeled "MTBF Comparison", the word "worse" means that the component mean time between failure is significantly worse than the rest and "better" means it is significantly better than the rest. No descriptor in the column denotes no significant difference from the rest. The "Distribution Comparison" column has a similar interpretation with no descriptor indicating it has the same distribution. The mean time between failure of the entire ensemble is the average value for the entire collection of components. The program prints the output in a more descriptive and detailed manner; however, even from the table, it is observed that an explicit description is given of the components which deviate in performance from the rest.

Since these deviate performers represent true deviations and not simply random fluctuations, action can now be taken. The components which have mean time between failures which are worse than the rest can be replaced or upgrade action can be taken. One notes that, in general, the Components of Group 1 are worse than those in Group 2 (all the "worse" components are in Group 1

TABLE 7. DEVIATE PERFORMANCE RESULTS
(90% Confidence)

Group No.	Unit No.	Individual MTBF	Average MTBF	MTBF Ratio	MTBF Comparisons	Distribution Comparisons
1	1	48.0	105.1	0.46	Worse	Different
1	3	23.0	104.4	0.22	Worse	Different
1	5	128.4	97.7	1.31		
1	7	92.8	99.3	0.93		
1	9	136.4	97.4	1.40		
1	16	57.4	102.7	0.56	Worse	
1	17	397.0	96.5	4.11	Better	
2	3	92.0	99.4	0.93		
2	10	451.0	96.1	4.69	Better	
2	14	95.8	99.6	0.96		
2	15	114.5	98.7	1.16		
2	18	248.0	96.5	2.57	Better	
2	20	82.0	99.2	0.83		
2	23	181.0	96.3	1.88	Better	

The MTBF of the Entire Ensemble = 99.1 Days

and Group 2 has all "better" components). Hence, a subsystem degradation is indicated. Action can also be taken on this definitive result.

In the pursual of this action, the tests of the previous sections can be used to help uncover the causes for the bad actors and for the subsystem degradation. For example, the exponential test may be used to show any abnormal stresses on the bad actors. For a majority of the bad actors, the distributions are also significantly different than the rest which shows a different failure behavior being experienced, such as high-burn or wear-out.

The different distribution results denote the fact that these different failure behaviors are critical enough to cause a deviate performance. The individual distributions can now be obtained to determine the specific causes for the deviations. As discussed previously, if a high peak exists at small times between failure, ineffective repair is indicated. If the times between failure are becoming successively smaller, a wear-out may be indicated. A number of the previous tests evaluated the data for these types of characteristics and their output can be used to help uncover the causes for the deviations. Used in this manner, the set of tests described in this paper form a single, consistent set of techniques which compliment one another and which together form a powerful team for evaluating failure data.

IV. SUMMARY

The tests and evaluations described in this paper represent a set of methods for analyzing failure data. Each test evaluates a particular aspect of the data and therefore each has particular utilization. The technique for obtaining failure rates gives concise parameters of the failure behavior. The exponential test uncovers abnormal environments and the related repair effectiveness test assesses the amount of repair performed. The chi-squared test audits significant failure changes in various time periods. The performance test probes for increases and decreases in failure frequency, and the deviate performance test extracts below par and above par performers. Because of the different features of each test, the tests together are an effective means of auditing, probing, and correcting failures.

In addition to the particular utilizations given in this paper, the tests have the added virtue of being able to be extended to further applications. Different classes of results can be obtained and different types of data can be analyzed. In this respect, an entirely new series of tests can be generated from the basic features of the tests described. In essence, the evaluations can be tailored to any particular need and to any particular type of data.

Finally, the tests can be straightforwardly incorporated into a computer program which automatically evaluates any size data source. Computer programs were described in this paper which automatically performed the individual evaluations. At the time of this writing, these programs are being incorporated into one final program which will allow flexibility as to the type of evaluations performed, the data records to be examined, and the time periods over which the failures are to be analyzed. Automatic updating is included and the program is simple to run with word input used. The final computer program represents the automated manner in which these tests can be performed to efficiently give timely evaluations of failure data.