EGG-RAAM--10563

DE93 007146

DATA AND STATISTICAL METHODS FOR ANALYSIS OF TRENDS AND PATTERNS

**C. L. Atwood
C. D. Gentillon
G. E. Wilson**

Published November 1992

**Idaho National Engineering Laboratory
EG&G Idaho, Inc.
Idaho Falls, ID 83415**

**Prepared for the
Office of Analysis and Evaluation of Operational Data
U.S. Nuclear Regulatory Commission
Washington, DC 20555
Under DOE Field Office, Idaho
Contract DE-AC07-76ID01570
FIN L1420**

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

ABSTRACT

This report summarizes topics considered at a working meeting on data and statistical methods for analysis of trends and patterns in U.S. commercial nuclear power plants. This meeting was sponsored by the Office of Analysis and Evaluation of Operational Data (AEOD) of the Nuclear Regulatory Commission (NRC). Three data sets are briefly described: Nuclear Plant Reliability Data System (NPRDS), Licensee Event Report (LER) data, and Performance Indicator data. Two types of study are emphasized: screening studies, to see if any trends or patterns appear to be present; and detailed studies, which are more concerned with checking the analysis assumptions, modeling any patterns that are present, and searching for causes. A prescription is given for a screening study, and ideas are suggested for a detailed study, when the data take of any of three forms: counts of events per time, counts of events per demand, and non-event data.

SUMMARY

At the request of the Trends and Patterns Section of the Nuclear Regulatory Commission (NRC) Office of Analysis and Evaluation of Operational Data (AEOD), a working meeting was held on Sept. 21-24, 1992. The objective of this meeting was to outline methods and guidance for data analysts who are concerned with trends and patterns in typical NRC data. This report includes methods that were mentioned in a pre-meeting discussion paper, and ideas that were suggested at the meeting.

The three primary data sources are reviewed: the Nuclear Plant Reliability Data System (NPRDS), Licensee Event Report (LER) data, and Performance Indicator (PI) data. The NPRDS data deals with components and their failures (excluding failures caused by human error). It can be used to investigate trends over time, but not differences between plants because of inconsistent reporting by plants. If possible, a second data source should be used to supplement NPRDS data. The LER data can be used for events at the system level, rather than the component level or train level. The PI data draws events from various sources, and are now used primarily by the PI program rather than by the Trends and Patterns Branch.

Data may consist of counts of events per time, counts of events per demand, or non-event data. The working meeting concentrated on counts of events per time.

For a screening study of events per time, the standard assumptions are that the event counts are independent Poisson random variables with a single constant event rate. To investigate this model, the data should be partitioned into disjoint time cells, and a chi-squared test performed. The cells should be chosen so that the average count per cell is at least 5, and close to 5 if possible. A graphical analysis should supplement this test: for each cell, plot the estimated event rate, with a confidence interval or a Bayes posterior distribution. This plot can be examined to determine the kind of departure, if any, from the initial assumption of a constant event rate.

For events per demand or non-event data, a similar mixture of tests and plots

is recommended.

If the initial assumption of constant failure rate is rejected, a more detailed study may be called for. This differs from the screening study in that it checks the validity of underlying assumptions, it probably tries to model the form of the event rate. Many techniques were suggested at the working meeting, too disparate to list here.

This report concludes with references that give more detail about many of the methods briefly mentioned in the report, and a list of the meeting participants.

CONTENTS

ABSTRACT	iii
SUMMARY	v
1. INTRODUCTION	1
1.1 Background	1
1.2 Working Meeting on Data and Statistical Methods	2
2. PRINCIPAL DATA BASES	4
2.1 NPRDS	4
2.2 LER Data	6
2.3 Performance Indicator Data	7
3. TYPES OF TRENDING ANALYSES	8
3.1 General Study Considerations	9
3.2 Screening Studies	9
3.3 Detailed Studies	10
3.4 Special Studies	11
4. STATISTICAL ASPECTS OF A SCREENING STUDY	11
4.1 Counts of Events per Time	13
Simple Model	13
Screening Methods	14
Example	15
4.2 Counts of Events per Demand	17
Simple Model	18
Screening Methods	18
Example	19
4.3 Non-Event Data	21
Simple Model	21
Screening Methods	21
4.4 Presentation	22
5. STATISTICAL ASPECTS OF A DETAILED STUDY	23
5.1 Checks of Assumptions.	24
5.2 Bayesian Methods	25
5.3 Other Possible Ideas	26
5.4 Counts of Events per Time	29
Poisson Processes to Model Trends	29
Comparisons of Counts from Different Populations	30
Bayesian Priors	32
5.5 Counts of Events per Demand	32
Modeling Trends	32
Comparisons of Counts from Different Populations	32
Bayesian Priors	33
5.6 Non-Event Data	34
The Logarithmic Transformation	34
Checks of Assumptions	34
Bayesian priors	35

REFERENCES	37
APPENDIX Participants in the Working Meeting	39

DATA AND STATISTICAL METHODS FOR ANALYSIS OF TRENDS AND PATTERNS

1. INTRODUCTION

1.1 BACKGROUND

The Nuclear Regulatory Commission's (NRC) Office of Analysis and Evaluation of Operational Data (AEOD) provides, as one of its primary roles, an independent capability to analyze operational data from commercial power reactors. AEOD reviews and evaluates operating experience to identify:

- Portentous events and their associated safety concerns and root causes,
- Trends and patterns displayed by the events,
- The adequacy of corrective actions taken to address safety concerns, and
- The generic applicability of portentous events to all nuclear power plants.

AEOD successfully uses a number of existing programs and tools to assess reactor performance, yet recognizes the desirability of developing new methods to mitigate existing shortcomings of the methods and more fully utilize the information contained in the performance data. Consequently, AEOD has established several initiatives to explore methods and approaches that could provide the ability to focus on risk relevant concerns more quickly, and provide a more thorough culling of operating reactor experience data than is now in routine use. In support of the AEOD improvement initiatives, the Trends and Patterns Analysis Branch (TPAB) has been reassessing its mission and the types of analyses it performs.

To aid the TPAB mission reassessment, a workshop (AEOD 1992a) was conducted in January 1992 to investigate how Probabilistic Risk Assessment (PRA) insights could help AEOD perform enhanced reactor safety evaluations. The workshop objective was to explore innovative approaches and methods related to:

- Use of preexisting PRA and Individual Plant Examination (IPE) results in performing routine event evaluations,
- Methods and approaches for evaluating industry risk profiles (trends),
- Innovative uses of existing/available data sources to identify risk significant trends and safety issues,
- Potential new plant-specific risk-based performance indicators (PIs), data needs and availability,
- Risk-based approaches for selecting and analyzing plant-specific and "generic" trends and patterns, including common cause failure, system interactions, and human performance concerns, and
- Analytical methods, software and procedures which could be used, adapted, modified, improved, or developed to enhance operational data analysis.

Following the workshop, and in partial support of the furtherance of several insights gained therein, the Trends and Patterns Section formulated a plan (AEOD 1992b) to address trends and patterns analysis related to hardware and safety issues. The primary objectives of the tasks defined in the plan are to specifically identify important initiating events and accident sequences, risk important components, risk important systems and safety and regulatory issue for trend analysis. The potential uses of trending analysis are illustrated by Figure 1, taken from the plan.

1.2 WORKING MEETING ON DATA AND STATISTICAL METHODS

One of the initial efforts in the TPAB reassessment was to characterize the types of analysis performed by the Trends & Patterns Section, particularly those using statistical techniques. In support of the analysis characterization a working meeting on statistical methods was conducted on Sept. 21-24, 1992. A draft paper was used as background for that working meeting. This report updates the draft to include the recommendations from the working meeting.

The objective of the working meeting on statistics, and consequently of this report, was to outline methods and guidance that can be used in the development of procedures for the analysis of operational event data. Thus, the working meeting was to identify:

Figure 1. Overview of planned Trends and Patterns activities.

- The assumptions underlying the different types of analyses,
- Statistical methods for testing the validity of the assumptions, and
- Key questions an analyst should consider in performing an analysis. These questions may be partitioned into initial questions posed at the start, and those that may follow as the analysis develops.

The remainder of this report describes data bases and statistical methods that were discussed at the meeting, emphasizing those considered most appropriate by the group. The participants in the meeting are listed at the end of this report in an appendix. This summary, however, is the responsibility of the authors.

2. PRINCIPAL DATA BASES

The AEOD staff analyzes and evaluates the operational experience of nuclear power plants using several major data sources:

- Voluntary reports of component failures submitted by the utilities to the Nuclear Plant Reliability Data System (NPRDS), operated by the Institute of Nuclear Power Operations (INPO),
- Licensee event reports (LERs), (10 CFR 50.73, Licensee Event Report System),
- Immediate notification reports (10 CFR 50.72, Immediate Notification Requirements for Operating Nuclear Power Reactors),
- Plant operating data contained in licensee monthly operating reports, and
- Plant radiation exposure data collected and reported by INPO.

An overview of the NPRDS is provided in Section 2.1. The LER data source is reviewed in Section 2.2. Selected data from all of the above sources are collected, analyzed and reported as part of the NRC's Performance Indicator (PI) program executed by the Idaho National Engineering Laboratory (INEL) under contract to AEOD. The aspects of the PI program applicable to objectives of this report are reviewed in Section 2.3.

2.1 NPRDS

The NPRDS is an industry-wide system for monitoring the performance of selected nuclear power plant components that are important to safe and reliable plant operation. The performance information is obtained from reports submitted by participating utilities when the designated components fail to perform their intended functions. Participation is "voluntary" but INPO and the NRC strongly encourage participation, and therefore every operating plant participates. The NPRDS data base contains component failure and system and component engineering data.

The NPRDS data is used to satisfy a number of objectives. Those of interest here include determination of equipment failure rates for use in maintenance

program analysis, identification of potential generic issues, and PRA. Because the system was designed by the Architect/Engineers and the Nuclear Steam System Supplier (NSSS) vendors, the system is weak in some areas of interest to PRA analysts. For example, transient initiators stemming from non-safety-class equipment were not originally included. Other problems for PRA analysts are itemized more fully below. The data base currently contains approximately 140,000 component failure records and approximately 565,000 engineering records. These records are related to 35 types of components in 112 units.

Since 1984, NPRDS has been operated by INPO. INPO periodically audits the failure data, visiting each plant about every five years. The auditors compare selected plant records with the information in NPRDS, and meet with the plant personnel responsible for NPRDS reporting. In contrast to the failure data, the engineering data (information about the components, rather than about their failures) have never been audited, except for a few cases.

Problems when using the NPRDS for PRA purposes include the following:

- NPRDS reporting is commonly not given high support or budget by utility management. Submitters may be new employees, who move to new positions frequently. Consequently, when INPO receives a submittal it checks it for obvious errors or omissions, and sends about 50% back to the plants for correction. During an outage everyone at a plant is working long overtime to get the plant up, so NPRDS gets ignored. Audits of the records show that about 70% to 80% of the reportable events are actually reported; this can vary from as low as 20% at some plants to as high as 99% at others. INPO will not reveal these plant reporting completeness rates to the NRC.
- Events resulting directly from human error are not reportable.
- If a component is replaced by a different component, this information is not always evident in NPRDS. This is especially the case when the components have the same model number. Often the engineering data is simply updated, with the information such as the serial number of the former component erased. In the process, the in-service date of the new component may be lost.

- The number of successful demands is not reported. This is a universal problem, not special to NPRDS.

Unquantifiable incompleteness of the records was of great concern to the meeting participants. The most promising approach suggested was to ask INPO for the names of the best ten or so plants — the plants with the most complete reporting, not necessarily those with the lowest event rates; INPO might be willing to reveal this much of their reporting data, without feeling that it is incriminating the other plants. Base subsequent NPRDS data analyses on the data from these plants, as if they were a random sample from all the plants in the industry. This approach assumes that a plant's completeness of reporting is consistent over time, and that the completeness is independent of the event rate. A second approach would be to use the plant-specific (and often proprietary) data from well-done PRAs, or IPE data, to estimate the NPRDS reporting completeness for those plants. This approach would require matching events, or at least failure modes, in the PRA data with those in NPRDS.

There is room for guarded optimism, because AEOD issued a very critical evaluation of NPRDS last year. INPO, the industry, and the NRC are trying to work out a more satisfactory system. For now, an analyst should always try to use a second data source, such as maintenance records, as well as NPRDS, never trusting NPRDS completely.

2.2 LER DATA

Beginning with implementation of the LER rule in 1984, reactor events meeting certain threshold criteria are reported to the NRC for each licensed commercial power reactor. The information contained in the LERs includes:

- LER identifier, event date and LER date,
- Event description,
- Event cause(s),
- Event analysis,
- Corrective actions.

Depending on the time required to perform the event analysis, supplements to an LER may be issued. The magnitude of the LER data is illustrated in the following table.

Year	Total LERs	No. of Operating Reactors	Average No. of LERs per Reactor
1985	3002	96	31
1986	2866	105	27
1987	2919	110	27
1988	2500	111	23
1989	2362	113	21
1990	2128	111	19
1991	1858	111	17

To facilitate the storage and retrieval of information contained in the LERs, AEOD has established the Sequence Coding and Search System (SCSS) at the Oak Ridge National Laboratory. On average, the SCSS contains 150 pieces of related data for each LER.

The LERs are easily available to the public, and therefore are often deliberately not very detailed. They are meant to capture system-level information, unlike entries in NPRDS, which capture component-level information. Train-level failures are not necessarily in either data base.

2.3 PERFORMANCE INDICATOR DATA

The LERs, immediate notification reports, licensee monthly operating reports, and INPO radiation exposure information serve as sources for selected data collected, analyzed and reported quarterly in the PI program for each reactor. The PI data contain information relating to:

- Occurrences (counts) of Automatic Scrams While Critical (Scrams), Safety System Actuations (SSA), Significant Events (SE), Safety System Failures (SSF), and Cause Codes,
- Rates of Forced Outage (FOR), Equipment Forced Outages per 1000 Commercial Critical Hours (EFO), and Radiation Exposure, and
- Plant specific and industry trends for most of the above.

The PI data can be sorted into sets that relate to plant types, peer groups, NSSS vendors, and NRC regions. The records cover data reported from 1985 to the present.

3. TYPES OF TRENDING ANALYSES

The Trends and Patterns Section plans to perform three different types of analyses: 1) screening studies (expected to constitute the majority of the work), 2) detailed studies of those items identified for further analysis from a screening study, and 3) special studies which are primarily requests to examine items of special interest. General considerations for all three study types are given in Section 3.1. Overviews of the screening, detailed and special studies are provided in Sections 3.2, 3.3 and 3.4. Further details of statistical methods are presented separately for the screening and detailed studies in Sections 4 and 5. It should be noted that not all the methods described in this report were discussed in the working meeting: the meeting concentrated on failure rates (with units 1/time), while the other settings considered in this report were discussed in the background paper that was prepared before the meeting.

In this report, a *trend* means a steady increase or decrease over time, for example, a trend in a failure probability. A *pattern* means any deviation from an initial hypothesis, resulting from some cause more fundamental than mere randomness of the data; examples could be a step change in reporting rates, caused by issuance of a rule, or a high failure rate in one year, caused by discovery of a generic problem and a backlog of previously unrecognized degraded components.

3.1 GENERAL STUDY CONSIDERATIONS

For the studies mentioned above, a basic assumption to be tested is that a simple process (for example, independent failures with a single failure rate or a single probability of failure on demand) produced the various data subsets (for example, data from different plants or different time periods). It is already recognized that the different plants do not necessarily report to NPRDS with the same completeness. Therefore true differences between plants cannot be evaluated with NPRDS data alone. The discussion below therefore concentrates on pooling data across selected plants, and testing whether the failure rate, demand failure probability, or other quantity of interest is constant over time. If it is not, the type of departure from the assumption and its cause become the items of interest.

The basic data upon which all studies are dependent are most commonly of the forms:

- Discrete event counts in fixed time periods (for example, number of scrams while the reactor is critical, or number of significant events during a quarter or year)
- Discrete event counts per demand (for example, number of safety system failures when demanded)
- Measurements of continuous quantities (for example, radiation exposure by quarter or year, or forced outage rate).

3.2 SCREENING STUDIES

A screening study consists of a "quick look" analysis to determine if a trend in the operational experience data may exist, and if there is a need for more detailed evaluation. The study will generally include a summary of the major causes of failure-related study parameters. Every item present in an issues list will be subjected to screening analyses. The expected level of effort per analysis is approximately one staff-month.

It is intended that screening analyses will be performed periodically using

data summarized, for example, by year. Simple statistical tests can be expected to be applied to models of failure rates or failure probabilities.

Reports documenting screening analyses will consist of narratives, graphs and tables explaining the significant findings and formatted in the following sections:

- Purpose of study
- Description of parameter of interest
- Data sources
- Method of analysis
- Results
- Conclusions
- References

3.3 DETAILED STUDIES

A detailed analysis will be performed if a screening analysis indicates one is warranted, or there is special interest in a particular question which can only be answered by a detailed analysis. The level of effort for a detailed analysis can range from two staff-months to over a staff-year, depending on the complexity of the item under study and the supporting data base.

The purpose of a detailed analysis is to examine the data in depth from an engineering perspective. The use of more detailed and sophisticated statistical methods to quantify trends and differences may be expected. For example, failure dates may be used in addition to counts in the statistical analysis. Details of failure events may need to be verified, or it may be necessary to acquire additional technical information from equipment manufacturers.

The documentation reports will contain an Executive Summary, Background Information, and Appendices in addition to those listed for the screening studies.

3.4 SPECIAL STUDIES

Special studies are expected to be similar to detailed studies in documentation format, magnitude of content, and required resources. However, the focus, statistical methods, and questions of interest of the special studies may be uniquely different from the detailed studies.

4. STATISTICAL ASPECTS OF A SCREENING STUDY

Any screening study should involve the following activities:

- Identify the issue to be studied.
- Decide what data would be relevant for this issue (including the time frame, such as 1987 and later.)
- Collect a data set.
- Frame the issue as a hypothesis (such as "constant failure rate").
- Create plots to investigate possible departures from the hypothesis.
- Perform a standard statistical test of the hypothesis.
- Present the conclusions of the study.

It does not matter whether the plotting or hypothesis testing is done first; both should be done. Sections 4.1 through 4.3 elaborate on some of these steps for event data in time, for event data per "demand", and for non-event data, in that order. Within each subsection, the topics considered are the standard assumptions and estimators, and screening methods, concentrating on graphical methods and on tests of homogeneity. A condensed example is also given for each of the two discrete data cases. Following this treatment for the different data types, Section 4.4 discusses methods of presenting results; the presentations are similar for all data types, and so are discussed together.

For a screening study, the data are typically grouped into summary tables, often by year. Therefore, grouped data are emphasized here and below.

Some screening studies are performed every year, using the same data as before with one new year added. When additional data have been collected in this way, the working group believed that all the data should be used. The oldest data

should not be discarded, i.e. a moving time window should not be used, unless there is statistical evidence of a change at some time point, and a reason for this change (for example, implementation of a new rule) is understood. Then the data after the change point should be analyzed separately from the data before the change.

Possible trends and patterns are presented in Figure 2. The trends in 2B and 2C are probably of most interest to AEOD, but 2D and 2G are also important and may yield engineering insights. If an apparent trend or pattern is seen, and identified as statistically significant, caution should be used in extrapolation. The bon mot was quoted at the meeting: "Prediction is always difficult, particularly of the future." Based on a few years' data, it may be difficult to distinguish between a trend or pattern with little variance (such as cases 2C and 2D) and a constant mean with large variability (for example, case 2F). To help

Figure 2. Types of trends and patterns over time.

confirm statistical evidence, engineering or programmatic explanations must be sought for any apparent trend or other pattern. The statistical exploration points out things to investigate; true knowledge can come only after the investigation.

What p-value (attained significance level when testing the hypothesis) should be used to trigger more detailed investigation? This depends on how many false alarms AEOD is willing to accept and how important it is to identify true patterns. The value 0.10 was suggested at the meeting as a trigger for considering a more detailed study.

4.1 COUNTS OF EVENTS PER TIME

Simple Model

The standard assumptions are: (1) The events occur independently in nonoverlapping time intervals. (2) The event rate λ is unchanging over time. (3) The data set consists of all the events in a fixed predetermined time period. Then the number of events in any fixed time period t is Poisson(λt).

Suppose that n events occur in time t . The event rate λ is estimated by n/t , and the standard deviation of the estimator is $(\lambda/t)^{1/2}$. An exact formula for a confidence interval for λ is given in Section 4.6.2 of Johnson and Kotz (1969):

$$\begin{aligned}\lambda_L &= (1/2t)\chi^2_{\alpha/2}(2n) \\ \lambda_U &= (1/2t)\chi^2_{1-\alpha/2}(2n+2) .\end{aligned}$$

Here λ_L and λ_U are the lower and upper ends of a $100(1-\alpha)\%$ confidence interval, and $\chi^2_a(b)$ denotes the a th quantile of a chi-squared distribution with b degrees of freedom.

The Bayes posterior distribution for λ , based on the Jeffreys noninformative prior is

$$(1/2t)\chi^2(2n+1) .$$

(See Box and Tiao 1973.) This distribution is often written as a gamma distribution, but we use the χ^2 notation here to show the relation between the Bayesian distribution and the confidence intervals.

Screening Methods

To explore the relations between subsets of the data, such as different years, mark the subsets on one axis. For each of these subsets of the data, plot an estimate of λ and a confidence interval for λ against the other axis. If preferred, plot a Bayes posterior density for λ for each of the subsets. Patterns such as trends, outlying subsets, or large scatter between subsets are then made visible. An example is given below.

In addition, if the individual event times are known, plot the times between events in sequence; patterns indicate changes in λ . Optionally, also plot the cumulative event count against time; a straight line indicates no change in λ over time. These two plots are analogues of familiar plots in regression analysis, the first corresponding to a residual plot and the second to a plot of the fitted vs. measured values. This can be done for the whole data set; it can also be done within subsets of the data, such as events with different failure causes or failure modes.

If several populations (such as different time periods) produce independent Poisson counts, the joint distribution of the counts, conditional on the total count, is multinomial. Tests of simple relations among the populations can be based on this fact. The simplest case is a one-way layout, where under the null hypothesis the i th population is assumed to correspond to a Poisson count with parameter proportional to a known c_i . (For example, if the population correspond to different time periods, c_i might be the number of plant operating hours in the i th time period.)

For grouped data, a chi-squared test is recommended as an initial tool to decide whether the cells can be pooled. Do not worry about distinguishing between possible alternatives such as upward trends and downward trends. Just use the chi-squared test as an all-purpose tool. As for grouping, if feasible divide the data into N cells with $N \approx (\text{total count})/5$, so that under the null

hypothesis each cell has expected count approximately equal to 5. The reason for wanting the expected count ≥ 5 is the usual rule of thumb: the chi-squared distribution can be used if the expected cell sizes are all 5 or larger. The reason for wanting the expected count to be not much larger than 5 is that information is thrown away if small cells are collapsed into larger cells unnecessarily: the test loses power. Therefore, aim for an expected cell count of about 5 if possible.

If a statistically significant departure from the null hypothesis is found, proceed to check the model assumptions, as described in Section 5.4. Engineering or programmatic explanations for the departures must be sought. Engineering explanations could include aging of the equipment or changing maintenance practices; programmatic explanations could include new rules or reporting requirements. After such explanations have been tentatively identified it may be reasonable to model a process that differs from the original null hypothesis, such as a trend in the failure rate. Ways of doing this are mentioned in Section 5.4.

Example

An example with lightning strikes is shown in Table 1. A 95% confidence interval for λ was found for each year. The calculations were performed by SAS (1988) using the confidence interval formula based on the χ^2 distribution, and the plot was typed in for this example. In actual practice, the plot would be drawn either by hand or by a graphics package.

Because the exact event times are not known for this example, the recommended plots (between-event times or cumulative counts) are not shown. In this example, such plots might reveal a seasonal pattern.

Now consider the chi-squared test for this example. If the seven years could be grouped into 4-month periods, there would be 112 events in 21 cells, resulting in an average of 5.33 events per cell. This is slightly larger than 5 per cell, so it would be recommended if the data were available. Seasonal differences might be apparent, depending on how closely the 4-month periods coincided with lightning seasons. Therefore, the annual summary data might be preferred over

Table 1. Lightning strikes per reactor-year

YR	EVENTS	REACTORS	LAMLO	LAMHAT	LAMUP				
85	16	97	0.09428	0.16495	0.26787		(---*---)		
86	21	106	0.12264	0.19811	0.30284		(---*---)		
87	20	111	0.11006	0.18018	0.27827		(---*---)		
88	12	110	0.05637	0.10909	0.19056		(---*---)		
89	9	112	0.03674	0.08036	0.15254	(---*---)			
90	11	111	0.04947	0.09910	0.17732	(---*---)			
91	23	111	0.13135	0.20721	0.31091		(---*---)		
tot	112	758	0.12166	0.14776	0.17779		(---)		

the 4-month data for investigating longer term trends. In any case, such detailed data are not available for this example. Therefore, a chi-squared test is performed here based only on the annual data of Table 1. The seven cells correspond to years. Under the null hypothesis of no difference between years, the probability of a lightning strike in year i is $c_i \lambda$, where c_i is the number of reactor-years for the i th year, and λ has units $1/(\text{reactor-year})$.

The calculations are shown in Table 2, using the same data as Table 1. The term labeled CHISQ equals $(\text{Observed} - \text{Expected})^2 / \text{Expected}$, the contribution of the cell to the chi-squared test statistic. The p-value is 0.067, which does not quite lead to rejection at the 0.05 level. It does, however, lead to rejection at the 0.10 level, which was suggested at the beginning of Section 4 as a trigger level for initiating a detailed study.

It is of some interest that when the lightning strikes are normalized by critical years rather than by reactor years, the p-value drops below 0.05. To choose the more correct normalizer, one must consider the physical process. Presumably, the probability of a lightning strike depends on whether the reactor is standing, not whether the reactor is at power or not. If this is the case one should normalize by reactor-years, not by critical years.

Table 2. Chi-squared analysis of lightning strikes per reactor-year

YR	EVENTS	REACTORS	E(EVENTS)	CHISQ
85	16	97	14.3325	0.19401
86	21	106	15.6623	1.81911
87	20	111	16.4011	0.78973
88	12	110	16.2533	1.11304
89	9	112	16.5488	3.44342
90	11	111	16.4011	1.77863
91	23	111	16.4011	2.65508

				11.79302

p-value = 0.066749, based on 6 degrees of freedom

If a detailed study were started based on the p-value in the screening study, we would have to investigate whether the observed different event rates in different years were caused by differences in the weather or differences in the plants. The narratives of the events could be important sources of information on why the differences occurred.

4.2 COUNTS OF EVENTS PER DEMAND

It can be hard to find this kind of data, because often the number of demands is unknown; then time is commonly used as a surrogate for the number of demands. We use the word "demand" even when the application is to another setting, such as the number of failures with some particular cause, out of all the failures that occur.

If the number of demands is known and large, use exactly the same methods discussed in Section 4.1 for event counts in time, replacing "hour" by "demand." Therefore the discussion below assumes that the number of demands is moderate or small.

Simple Model

A screening analysis makes the following simple assumptions: (1) Given a demand, an event (such as some type of failure) occurs with some probability p . (2) Occurrences of events for different demands are independent; that is, the event probability for one demand is not affected by what has happened on other demands. (3) The data consist of the count of events in a known number of demands. Under these assumptions, the number of events in n demands has a binomial(n, p) distribution. Suppose k events occur in n demands. The event probability p is estimated by k/n , and the standard deviation of the estimator is $[p(1-p)/n]^{1/2}$. Johnson and Kotz (1969, Section 3.7.2), give an exact formula for a $100(1 - \alpha)\%$ confidence interval for p : The lower limit is

$$p_L = v_1 F / (v_2 + v_1 F)$$

where $v_1 = 2k$, $v_2 = 2(n-k+1)$, and F equals the $\alpha/2$ quantile of the F distribution with v_1 and v_2 degrees of freedom. Similarly, the upper limit is

$$p_U = v_1 F / (v_2 + v_1 F)$$

where $v_1 = 2(k+1)$, $v_2 = 2(n-k)$, and F equals the $(1 - \alpha/2)$ quantile of the F distribution with v_1 and v_2 degrees of freedom.

The Bayes posterior distribution for p , based on the Jeffreys noninformative prior, is beta($k + \frac{1}{2}$, $n - k + \frac{1}{2}$), as shown by Box and Tiao (1973). This results in an interval estimate that is numerically close to the confidence interval, but a little shorter at each end.

Screening Methods

To explore the relations between subsets of the data, such as different years or plants, mark the subsets on one axis. For each of these subsets of the data, plot an estimate of p and a confidence interval for p against the other axis. Patterns such as trends, outliers, or large scatter are then made visible.

Other graphical plots parallel those in Section 4.1. Two such are a

sequential plot of the number of demands until the next event and a cumulative plot of the event count against the number of demands. These plots can show patterns indicating changes in p . This can be done for the whole data set or for subsets.

When performing statistical tests, the natural format for the data is a contingency table. The rows correspond to the kinds of event: for example, the first row could show system failures and the second row system successes; or in a different context, each row could correspond to a different type of cause for a failure. The columns correspond to some other attribute of the events: different time periods, plant types, or whatever. The table entries are the counts of the events for each cell. A chi-squared test can test whether the rows and columns are independent, for example, whether system failures have the same probability for each level of the column attribute. If the sample size is so small that more than one expected sample size is less than 5, either cells must be pooled or a method from Section 5.5 must be used.

As in Section 4.1, use the chi-squared test during the screening. If the null hypothesis is rejected, examine the modeling assumptions, as described in Section 5.4, and search for engineering or programmatic explanations. If there appears to be a trend over time, use logistic regression, also described in Section 5.5.

Example

Consider data on turbine train failures for the years 1987 through 1991, shown below. A 95% confidence interval for p was found for each year. The calculations were performed by SAS using the confidence interval formula based on the F distribution, and the plot was typed in for this example. One year appears a bit higher than the others, but the confidence intervals show that the difference could have resulted from random variation. It appears acceptable to pool the data from the five years.

The data from Table 3 can be analyzed with a contingency table, as shown in Table 4. This analysis was performed by SAS. The largest contribution to the chi-square statistic is 4.1521, for the failures in 1989. This property of 1989 is consistent with the plot shown in Table 3. To assess the significance level,

Table 3. Confidence limits on p, for AFW turbine train.

YR	FAILS	DEMANDS	PLO	PHAT	PUP					
87	6	62	0.036350	0.09677	0.19883	(---*---)				
88	2	40	0.006114	0.05000	0.16920	(--*---)				
89	7	32	0.092772	0.21875	0.39973	(-----*-----)				
90	3	35	0.018038	0.08571	0.23058	(---*---)				
91	2	25	0.009840	0.08000	0.26031	(---*---)				
Total	20	194	0.064118	0.10309	0.15472	(-*--)				

Table 4. Contingency table analysis of AFW turbine train failures.

Frequency Expected Cell Chi-Square	87	88	89	90	91	Total
failures	6 6.3918 0.024	2 4.1237 1.0937	7 3.299 4.1521	3 3.6082 0.1025	2 2.5773 0.1293	20
successes	56 55.608 0.0028	38 35.876 0.1257	25 28.701 0.4773	32 31.392 0.0118	23 22.423 0.0149	174
Total	62	40	32	35	25	194

STATISTICS FOR TABLE OF FAILSUCC BY YR

Statistic	DF	Value	Prob
Chi-Square	4	6.134	0.189
Likelihood Ratio Chi-Square	4	5.408	0.248
Mantel-Haenszel Chi-Square	1	0.036	0.849
Fisher's Exact Test (2-Tail)			0.249
Phi Coefficient		0.178	
Contingency Coefficient		0.175	
Cramer's V		0.178	

Sample Size = 194

WARNING: 40% of the cells have expected counts less than 5. Chi-Square may not be a valid test.

use Fisher's exact test, because the expected cell counts are too small for the chi-squared distribution to be accurate. By the exact test, the p-value is 0.249, too large to warrant rejection of the hypothesis of independence of rows and columns. We conclude that there is no strong evidence of a difference between years, just as we suspected after looking at the plots in Table 3.

4.3 NON-EVENT DATA

The above sections deal with discrete quantities, based on counts. Now consider quantities that can have values in a continuum, such as a radiation exposure or forced outage rate. This material was not considered by the working group, but was part of the discussion paper prepared before the group met.

Simple Model

The simplest models use normal distributions, either of the measured quantity or of some transformation such as the logarithm of the quantity. It is assumed that the observed measurements constitute a random sample from that population. The usual summary statistics provide estimates of the parameters. Warning: if the original measurements are highly skewed, do not try to describe the populations in terms of their moments; use quantiles if at all possible. See the discussion in Section 5.6.

Screening Methods

Histograms and summary statistics for single variates can suggest whether transformations should be applied to obtain normal variates from non-normal variates.

Scatter plots are a simple way to display relationships between pairs of variates. For example, plot the measurements against time, against reactor type, etc.

The tools of Tukey's Exploratory Data Analysis, such as stem-and-leaf diagrams

and box plots, are designed for such continuous data. These tools have now found their way into the major commercial statistical software packages.

Trends can be modeled with great flexibility by regression software, but this may go beyond the scope of a screening study.

A t -test (with or without the assumption of common variance) can compare two populations. Most software will also provide a confidence interval for the difference of the means, using Satterthwaite's approximation if the populations are not assumed to have a common variance. For comparison of more than two populations with a common variance but different means, use analysis of variance.

4.4 PRESENTATION

Graphical methods seem to be well received: bar charts, scatter plots with fitted lines, pictures of confidence intervals and of posterior distributions. One useful kind of bar chart is an "engineering plot," explained by Crosby (1992); it is a bar chart, with a confidence interval on each bar, constructed so that if two intervals do not overlap, the two populations are declared different at the $100(1-\alpha)\%$ confidence level. This can be done when the data are balanced (e.g., equal numbers of demands for the different populations.)

In presenting results of the analyses, we need to distinguish between practical significance (a big enough difference between populations to have a practical consequence) and statistical significance (enough data to produce high confidence of a difference, although possibly a small one, between populations.) Engineering plots have this feature; they are plotted on an absolute scale, so practical differences can be judged, and the associated intervals show statistically significant differences. For AEOD purposes, statistical significance should be considered first: If there is no statistically significant difference, only a quantification of the random variability is necessary. If, on the other hand, there is a statistically significant difference, the difference can be modeled and quantified in a detailed study, to determine if it is also of practical significance.

In some cases, the data may be further grouped for presentation purposes. Suppose, for example, that the analysis has used data grouped by quarter, because this produced an expected cell count of about 5. For the final presentation, it still might be clearer to present the data grouped by year. The presenter must decide what gets the essential message across most clearly.

5. STATISTICAL ASPECTS OF A DETAILED STUDY

During the working group meeting, detailed studies did not receive as much attention as screening studies. Therefore many of the methods given here are the suggestions of the authors of this report, not necessarily of the group. Certain ideas that were suggested by individual members of the group, but not discussed thoroughly, are collected in Section 5.3.

As mentioned several times in Section 4, a detailed study is not purely statistical; the underlying engineering or programmatic causes must also be investigated. This report, however, only deals with the statistical aspects.

This section differs from Section 4 in several ways. It assumes that the screening study has already been done, so it does not repeat the material on standard models and on simple plots and tests. It also does not repeat the material on presentation of the results. A detailed study goes beyond a screening study in two ways: it spends substantial effort in checking the assumptions of the model, and in building a better model if necessary; and it tries to extract all the available information out of the data, even at the cost of complexity and computing effort. A detailed study may also go beyond a screening study by trying to apply the results to more complex analysis, for example, by producing results that can be used in a PRA. In this case, a Bayesian posterior distribution is probably necessary. Some general considerations are given in Sections 5.1 and 5.2 on checks of assumptions and on Bayesian methods, respectively. Section 5.3 mentions ideas that could have wide applicability but that were not considered in detail at the group meeting. Then Sections 5.4 through 5.6 consider the three specific types of data: events per time, events per demand, and non-event data.

5.1 CHECKS OF ASSUMPTIONS.

An implicit assumption is that there are no erroneous entries in the data set. These errors often occur not from transcription but from decisions that the data collector made. If possible, the analyst should read the narrative descriptions of the events. If this is not feasible, the analyst should try to talk with the persons who collected the data. It is distressingly common for data collectors to change their minds about which events should be included. It is a good investment to try to solidify the data set at the outset, before analyzing it thoroughly.

In a detailed study, the analyst has the responsibility to list the modeling assumptions and to try to challenge them from the data. Several general types of assumptions are considered here, following Atwood (1992b). In each paragraph below, a graphical check and a formal hypothesis test are mentioned.

If an assumption says that several subpopulations (e.g., components, systems, or time periods) have the same value for a parameter θ , estimate θ for the i th subpopulation by $\hat{\theta}_i$ and for the data pooled across all other subpopulations estimate θ by $\hat{\theta}_{-i}$. For each i , also calculate a confidence interval for θ_i ; plot these side by side to spot outliers. The standardized difference is

$$(\hat{\theta}_i - \hat{\theta}_{-i}) / [\text{var}(\hat{\theta}_i) + \text{var}(\hat{\theta}_{-i})]^{1/2} .$$

Under the assumption that the i th population is like all the others, this has mean 0 and variance 1. If the distribution of the difference is approximately normal, an approximate significance level can be calculated. Use the Bonferroni correction to account for the fact that many such significances are being examined simultaneously. That is, if 10 subpopulations are being compared, and the smallest significance level achieved for any of these comparisons is 0.02, the overall significance level for all the comparisons is $10 \times 0.02 = 0.20$. This is not strong evidence against homogeneity of the subpopulations.

If an assumption says that all the measurements should be independent and follow a certain univariate distribution (possibly with estimated parameters),

construct a plot of the observed vs. the theoretical quantiles (called a "Q-Q" plot), and see if the plot appears to show random points around a straight line. A formal goodness-of-fit test can also be used to see if the data follow the theoretical distribution: the Shapiro-Wilk test for normality; a Kolmogorov-Smirnov test, or (better) an Anderson-Darling test for an arbitrary distribution; a chi-squared test for a discrete or binned distribution with large sample sizes.

If an assumption says that each measurement should have a distribution that depends on known quantities (regression is the prototypical example), calculate each residual, the difference between an observation and its estimated mean. Plot them to show patterns. Divide each by its estimated standard deviation to obtain standardized residuals. A rough significance level can be calculated based on the largest standardized residuals in absolute value; are they much larger than would be expected if the residuals were randomly sampled from a standard normal distribution?

When the measurements, residuals, or estimated θ 's come in an ordered sequence (for example, corresponding to times), plot them against sequence number or against time, and look for patterns. Formal tests depend on the specific situation.

5.2 BAYESIAN METHODS

In the introduction to Section 5, it was mentioned that Bayesian methods may be necessary if the results of the analysis are to be used in a PRA. An important issue in Bayesian analysis is the choice of a prior distribution. The basis for the prior must be documented. It can be based on *careful* elicitation of expert opinion, through a formal documented process; or it can be taken from an analysis of generic data; or a "noninformative" prior can be used. See Box and Tiao (1973), for one presentation of noninformative priors. A noninformative prior is not uniform. Rather, it is the distribution that transforms to a uniform distribution when the model is transformed to one in which the parameter is a location parameter. The concept works well when there is one unknown parameter, less well when there are many unknown parameters. When many parameters are modeled simultaneously to produce a single answer, mechanical use

of a noninformative prior for each parameter can produce a misleading conclusion.

A Bayesian posterior interval for a single parameter based on a noninformative prior is numerically close to a (non-Bayesian) confidence interval for the same parameter with the same confidence level. Pictures of posterior densities communicate information effectively to non-statisticians, successfully hiding the difficult and subtle issues while displaying the conclusions pictorially.

5.3 OTHER POSSIBLE IDEAS

This section collects ideas that were suggested at the working group meeting, but which were not discussed or evaluated thoroughly. They are not listed in any particular order, although items pertaining to one topic are grouped together.

Investigate not only the mean failure rate as a function of time, but also the (between-plant) variance of the failure rate as a function of time. Changes in the variance could suggest insights into the process, leading both to better understanding of the behavior of the mean and to better approaches to fixing any undesirable characteristics of the process. With Poisson counts based on a common exposure time for each plant, the between-plant variance can be estimated by subtracting the contribution of the variances of the individual plants from the measured variance of the data, as follows. Let λ be the failure rate, treated as random between plants, and let X be the count in time t . Then

$$\begin{aligned}\text{var}(X) &= \text{var}[E(X)] + E[\text{var}(X)] \\ &= \text{var}(\lambda t) + E(\lambda t) \\ &= t^2 \text{var}(\lambda) + E(\lambda t).\end{aligned}$$

Therefore $t^2 \text{var}(\lambda)$ is estimated by $s^2 - \bar{X}$, where s^2 and $\bar{X}$ are the sample variance and sample mean of the counts at time t . This avoids the need for modeling the distribution of λ , although one way to model this distribution, as a gamma distribution, is given below in Section 5.4. Once the variance has been estimated for each time period, the contributions of the different plants to the variance can be investigated by leaving out one plant at a time from the calculations.

For one kind of failure, see if the failure rates corresponding to the different failure causes are constant over time.

Adapt control chart techniques to monitor trends in failures. Adapt time-series methods (generally using normal distributions) to model failure patterns. Both of these suggestions would presumably use failure counts in time intervals rather than individual failure dates.

Adapt methods from medical literature, such as longitudinal methods that follow individual human beings over time. Note, some medical studies consider only survival, and are appropriate only for nonrepairable components, while other might be adaptable to repairable components/systems/plants, for which repeated repairable failures are possible. One such method is the Kaplan-Meier estimator, available in SAS. It estimates the failure rate function nonparametrically, for nonrepairable components. Cox's proportional hazards model also might be applicable.

Use Kalman filtering to analyze data that is continually being augmented, when a non-constant failure rate is assumed. A tutorial article is by Martz, Campbell, and Davis (1977). If, instead, a constant failure rate is assumed, continual Bayes updates could be used.

Estimate the fraction of missing data, for example caused by underreporting from the plant. Use this value to correct the final answer. Then estimate the uncertainty in the fraction, and propagate this uncertainty through the algebraic calculations along with the uncertainty caused by the randomness of the data, to obtain a total uncertainty on the final answer.

Look at all repair actions, not just those corresponding to true failures. The incipient failures (degraded conditions that initiated repair even though nothing actually failed) contain more information than the failures alone. Therefore, an analysis of them might be fruitful. Note, however, that NPRDS data should not be used for such an effort, because incipient failures are not reportable in NPRDS, and therefore are only occasionally included in the data. Plant records would have to be used instead.

Failures are not necessarily independent. For example, so-called repairs that do not really correct the problem can result in a sequence of recurring failures in quick succession. In truth there was one failure with a long diagnosis-and-repair time, but the data will show several failures, which are not independent. If such a situation is modeled, simulations could be used to determine the properties of statistical procedures, such as the power of a chi-squared test that assumes independent failures with constant failure rate.

Although not discussed at the working meeting, two recent useful references on common cause failures are Mosleh et al. (1988) and Paula and Parry (1990).

To model a time-dependent failure rate, consider including more than one explanatory variable. For example, let x be the plant age and y be the component age, and model the failure rate of the component as

$$\lambda(x, y) = \exp(\alpha + \beta x + \gamma y) .$$

For quantifying the uncertainty in the failure rate of a complex system, the virtually universal method for PRAs is propagation of the Bayesian distributions of the basic event failure rates. A non-Bayesian competitor is the "Maximus" method, described by Bohn et al. (1988). Bootstrapping gives another non-Bayesian method, as exemplified by Dalal et al. (1989).

If a chi-squared test has rejected the null hypothesis, the individual cell contributions to the statistic show which cells have the greatest departure from the overall mean. Typically, these will be cells with small expectations and large counts. When approaching the event narratives, these cells are a good place to start reading. If something unusual was happening, these are the cells where it may have been happening the most.

When looking at standby systems or components, such as diesel generators, it is sometimes important to realize that a portion of the failure probability comes from failures that occur in time and sit there waiting for the test to reveal them, and other failures are caused by the test itself.

The AEOD Trends and Patterns Section should keep a file of all its analyses,

for later meta-analysis. That is, within its archival system they should record the p-value that resulted in each decision to accept or reject the null hypothesis. In theory, if the null hypothesis is true the p-values are uniformly distributed. Therefore, if an engineering reason is found for disbelieving the null hypothesis, the p-value for that issue should be put in one drawer. If no such reason can be found, the p-value should be put in a different drawer. Occasionally, all the p-values from this second drawer should be examined to see if they appear uniformly distributed.

5.4 COUNTS OF EVENTS PER TIME

When the data consist of event counts per time, this section considers ways to model trends and ways to test for differences between subpopulations. It then has a brief section giving the most commonly used Bayesian facts.

Poisson Processes to Model Trends

Extensive work, now standard, has been done under the assumption that events follow a Poisson process with an event rate that is not necessarily constant; see Cox and Lewis (1966) or Lawless (1982). This work uses the actual event times (as nearly as they can be obtained), not merely the total count per year or other such summary. If the non-constant rate is modeled as a function of a small number of parameters, asymptotic maximum likelihood methods are tractable, using numerical iteration rather than explicit formulas in most cases. Important special cases are when the event rate $\lambda(t)$ is of the form $\lambda_0 e^{\beta t}$, $\lambda_0 t^\beta$, or $\lambda_0(1 + \beta t)$; computer software exists for these models (Atwood 1992a). More complicated functional forms for $\lambda(t)$ would require development of software.

The likelihood ratio test for trend is especially easy for these cases. For example, with the exponential event rate model in an uninterrupted observation period, the test consists of comparing the mean event time with the midpoint of the observation period. Because trends are hard to identify from sparse data, collection of the exact failure times is often not worth the trouble; the graphical results of the screening study should be used to quickly eliminate cases where the likelihood ratio test will probably not find a trend.

Clustering of events is evidence of non-Poisson data; much less theoretical work has been done for such a model. Cyclical change of the environment, as for example in the progression through the operation cycle in a reactor, is another area that would benefit from development of tailored statistical methods.

Comparisons of Counts from Different Populations

If several populations (e.g. different time periods or different plant types) produce Poisson counts, the total count is ancillary, and the joint distribution of the counts, conditional on the total count, is multinomial. In particular, the conditional distribution of any one count is binomial. Tests of simple relations among the populations can be based on this fact. The Pearson chi-squared test is the best known such test.

When the sample sizes are large, most of the usual test statistics have an asymptotic chi-squared distribution. Methods using modern computing power are still being developed when the sample sizes are too small for asymptotic approximations. Agresti (1992) gives a survey of such methods, including a quick review of commercial software, for 2x2 and larger tables. SAS (Proc FREQ) performs what it calls Fisher's exact test for two-way tables of size 2x2 and larger. For the one-way layout, Atwood (1981) wrote a program to calculate the exact significance level of the chi-squared test; unfortunately the program, HOMOG, is no longer operating because it did not keep pace with changing hardware.

The Freeman-Tukey test was suggested in the meeting as an alternative to the usual chi-squared test. A review article by Crow (1983) says that this should work well with expected sample sizes down to size 1.0, but conflicting opinions exist on whether the test is superior to the usual chi-squared test. The test statistic is

$$\chi^2 = \sum [o_i^{\frac{1}{2}} + (e_i+1)^{\frac{1}{2}} - (4e_i+1)^{\frac{1}{2}}]^2$$

where o_i and e_i denote the observed and expected counts, respectively, in the i th cell. The terms in the summation are the squared Freeman-Tukey deviates, each approximately $\chi^2(1)$, and may be used to identify cells that show the greatest

departure from the null hypothesis. This agrees with the use of the cell contributions $(o_i - e_i)^2/e_i$ for the usual chi-squared test.

If two-way or higher-dimensional tables are used, a general framework for the analysis is the log-linear model, a field rich in statistical methods. The ij th population is proportional to $\alpha_i \beta_j$ for parameters α_i and β_j . References on log-linear models are Bishop et al. (1975), a shorter more introductory text by Fienberg (1980), and more recent books by McCullagh and Nelder (1989) and Agresti (1990). The users' manuals for commercial statistical computing packages also give information on this topic.

Whether the table involves a one-way layout or a higher-order layout, great care must be taken with extra-Poisson variation. For example, if comparing BWRs and PWRs, there is probably variation between plants within each reactor type. Although the number of events in one year at all BWRs, say, is Poisson, the number of events per year at a randomly chosen BWR has a larger variance than given by Poisson theory, because the one plant is randomly chosen from a heterogeneous population. This extra variance must be accounted for. It affects all aspects of the analysis: checks for outlying plants; calculation of the significance level for testing whether the two classes of reactor have the same parameters; and the description of each population, such as the relevance of a tolerance interval for the population rather than a confidence interval for the mean. One way to model the extra variance is to assume that the Poisson parameter has a gamma distribution between plants; then the count at a random plant has a negative binomial distribution. This model is used by Atwood (1983).

One approach to the question of pooling/splitting classes is as follows. Account for the variation within each class, and construct a tolerance interval for the population corresponding to each class. Similarly, construct a single tolerance interval for the population when the classes are pooled. Does the pooling result in shorter or longer tolerance intervals? The answer suggests the conceptually more efficient way to model the components of variance.

The comments in Section 5.1 on checking assumptions are also all relevant here.

Bayesian Priors

The conjugate prior for a Poisson parameter λ is a gamma distribution. There are different ways to parameterize a gamma distribution, but we use parameters α and β , corresponding to mean α/β and variance α/β^2 . The posterior distribution corresponding to n events in time t is gamma with parameters $\alpha + n$ and $\beta + t$. The noninformative prior sets $\alpha = 1/2$ and $\beta = 0$. The noninformative prior results in the posterior interval given at the beginning of Section 4.1, expressed there in terms of the χ^2 distribution.

5.5 COUNTS OF EVENTS PER DEMAND

As in Section 4.2, if the number of demands is large, use exactly the same method that is discussed above for counts of events in time. The discussion here assumes that the number of demands is moderate or small.

Like Section 5.4, this section gives ways to model trends and ways to test for differences between subsets of population. It closes with a brief summary of the most commonly used Bayesian facts.

Modeling Trends

If events are rare, there is little hope of trying to find a trend. (Recall, we assumed that the number of demands is not large.) If events eventually become rather frequent one could use logistic regression, although this analysis is not common in a failure-data setting. Logistic regression is a special case of log-linear modeling, mentioned in Section 5.4 (see Bishop et al. 1975, Section 2.2.4.)

Comparisons of Counts from Different Populations

As mentioned in Section 4.2, the natural format for the data is a contingency table. A chi-squared test can test whether the rows and columns are independent, for example, whether system failures have the same probability in all the populations. References are the books cited above for log-linear models. If the

sample size is not large, either cells must be pooled, or a method mentioned in Section 5.4 must be used: the Freeman-Tukey test (which seems to work for expected cell sizes down to 1.0), or an exact method.

An important assumption is that the count for each cell is binomial. If the cells correspond to different years, and the same plants contribute to each year, then the assumption of a binomial count for each year is reasonable. If instead the cells are different reactor types, and the data come from a sample of plants of each reactor type, the analyst must consider the possible extra variance between plants within reactor types. It would be natural to let p have a beta distribution for each reactor type so that the count from a random plant has a beta-binomial distribution. We are not aware of published analyses that do this.

Suppose now that the binomial model is accepted, and consider testing whether p is the same for N plants. A $2 \times N$ contingency table can test the overall hypothesis, following the pattern of Table 4. If instead, the concern is with possible individual outlying plants, a 2×2 contingency table can compare plant i against the pooled collection of all plants except i . This yields a p -value, measuring how different plant i appears from the others, assuming that the others are identical. This would be done for all N of the plants, so the Bonferroni correction must be used to obtain an overall significance level. For example, if there are 20 plants, and the smallest p -value is 0.008, the overall significance level is $0.008 \times 20 = 0.16$, which would probably not cause the analyst to flag any plants as outliers.

Bayesian Priors

Let p be the probability of a failure on demand. The conjugate prior for p is $\text{beta}(a, b)$. The posterior distribution, if there were f failures and s successes is $\text{beta}(a+f, b+s)$. The noninformative prior sets $a = 1/2$ and $b = 1/2$. This says that the prior information is on the order of $1/2$ a failure in one demand. It results in the posterior distribution stated in Section 4.1. When many binomial parameters are estimated simultaneously, mechanical use of the noninformative prior can be misleading.

5.6 NON-EVENT DATA

Normal data is so easy to analyze with standard software that it is covered in Section 4.3, except for the following three topics.

The Logarithmic Transformation

Suppose that the measured x 's are positive and strongly skewed to the right. The sample mean, though theoretically unbiased, will be smaller than the population mean for almost all samples, and much larger than the population mean for occasional rare samples. Its mean value is the population mean, but its typical values are smaller.

It is natural to assume that the distribution of x is lognormal (that is, $y = \log x$ is normal). This helps, but it does not eliminate all difficulties. Estimating the parameters of the distribution of y and translating this to estimate the mean and variance of x is dangerous, because the estimators can have very large variances. For highly skewed distributions, therefore, quantiles such as the median and 95th percentile of x have many advantages over moments such as the mean and variance, when the purpose of the analysis is to describe the distribution.

Analysis of variance on the normal y 's assumes that the y 's have a common variance and possibly different means. This assumption is equivalent to assuming that the lognormal x 's having a common shape parameter and possibly different medians. For some facts about the lognormal distribution, see Chapters 1 and 2 of Crow and Shimizu (1988).

Checks of Assumptions

The assumptions most commonly in need of checking are statistical independence, normality, and constant variance. Independence of measurements taken in sequence can be investigated by looking at the autocorrelations. Independence of pairs of measurements can be investigated by looking at their correlation. Normality can be tested by the Shapiro-Wilk test, and investigated graphically

by a Q-Q plot. Constant variance can be investigated graphically by some version of a residual plot.

Bayesian priors

Suppose the observable random variable is normal(μ, σ^2). When σ^2 is known, the conjugate distribution for μ is normal, and the noninformative prior is uniform on the line. The noninformative prior for the pair (μ, σ^2) is $1/\sigma$. For details, see Box and Tiao (1973).

REFERENCES

AEOD (Office for Analysis and Evaluation of Operational Data), 1992a, *Proceedings of the Workshop on the Use of PRA Methodology for the Analysis of Reactor Events and Operational Data*, NUREG/CR-0124, June 1992.

AEOD, 1992b, *Trends and Patterns Analysis Plan*, Draft, August 1992.

Agresti, Alan, 1990, *Categorical Data Analysis*, New York: John Wiley & Sons.

Agresti, Alan, 1992, "A Survey of Exact Inference for Contingency Tables," *Statistical Science*, v. 7, pp. 131-177 (including discussion).

Atwood, Corwin L., 1981, "Tests of a Simple Multinomial Hypothesis When the Sample Is Not Large," *1981 Proceedings of the 1981 Statistical Computing Section, American Statistical Association*, pp. 130-134.

Atwood, Corwin L., 1983, *Data Analysis Using the Binomial Failure Rate Common Cause Model*, NUREG/CR-3437, EGG-2271.

Atwood, Corwin L., 1992a, "Parametric Estimation of Time-Dependent Failure Rates for Probabilistic Risk Assessment," *Reliability Engineering and System Safety*, v. 37, pp. 181-194.

Atwood, Corwin L., 1992b, "Using Data to Check Model Assumptions," *Transactions of the American Nuclear Society*, TANSO 65, pp. 225-226.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland, 1975, *Discrete Multivariate Data: Theory and Practice*, Cambridge, MA: MIT Press.

Bohn, Michael P., Timothy A. Wheeler, and Gareth W. Parry, 1988, *Approaches to Uncertainty Analysis in Probabilistic Risk Assessment*, NUREG/CR-4836, SAND87-0871, January 1988.

Box, George E. P., and George C. Tiao, 1973, *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Cox, D. R. and P. A. W. Lewis, 1966, *The Statistical Analysis of Series of Events*, London: Chapman & Hall.

Crosby, Teri A., with James A. Rutherford, 1992, "Consulting Tips in Industry: Engineering Plots," *SPES News: Newsletter for the Section on Statistics in the Physical and Engineering Sciences of the American Statistical Association*, v. 2, no. 1, p.7.

Crow, Edwin L., 1983, "Freeman-Tukey Test," in *Encyclopedia of Statistical Sciences*, Samuel Kotz and Norman L. Johnson, eds., New York: John Wiley & Sons.

Crow, E. L., and K. Shimizu, eds., 1988, *Lognormal Distributions: Theory and Applications*, New York: Marcel Dekker, Inc.

Dalal, Siddhartha R., Edward B. Fowlkes, and Bruce Hoadley, 1989, "Risk Analysis

of the Space Shuttle: Pre-Challenger Prediction of Failure," *J. Am. Stat. Assoc.*, v. 84, pp. 945-957.

Fienberg, Stephen E., 1980, *The Analysis of Cross-Classified Categorical Data, Second Edition*, Cambridge, MA: MIT Press.

Johnson, Norman L., and Samuel Kotz, 1969, *Discrete Distributions*, New York: John Wiley & Sons.

Lawless, J. F., 1982, *Statistical Models and Methods for Lifetime Data*, New York: John Wiley & Sons.

Martz, H. F., K. Campbell, and H. T. Davis, 1977, "Estimating and Forecasting Failure Rate Processes by Means of the Kalman Filter," in *Theory and Applications of Reliability, with Emphasis on Bayesian and Nonparametric Methods*, vol. 1, Chris P. Tsokos and I. N. Shimi, eds., San Diego: Academic Press, pp. 165-192.

McCullagh, P., and J. A. Nelder, 1989, *Generalized Linear Models, Second Edition*, London: Chapman & Hall.

Mosleh et al., 1988, *Procedures for Treating Common Cause Failures in Safety and Reliability Studies*, NUREG/CR-4780, January 1988.

Paula, H. M., and G. W. Parry, 1990, *A Cause-Defense Approach to the Understanding and Analysis of Common Cause Failures*, NUREG/CR-5460, March 1990.

SAS, 1988, *SAS Language Guide for Personal Computers, Release 6.03 Edition*, Cary, NC: SAS Institute, Inc.

APPENDIX
Participants in the Working Meeting
Gaithersburg, MD, Sept. 21-24, 1992

Cory Atwood, INEL
EG&G Idaho
P.O. Box 1625
Idaho Falls, ID 83415-3421
(208) 526-0431

Gareth Parry, Halliburton NUS
910 Clopper Road
Gaithersburg, MD 20878
(301) 258-2536

Art Buslik, NRC/Research
NLS-372
Nuclear Regulatory Commission
Washington, DC 20555
(301) 492-3958

Dale Rasmuson, NRC/AEOD (Meeting
Organizer)
MNBB-9112
Nuclear Regulatory Commission
Washington, DC 20555
(301) 492-4490

Don Gaver, Naval Postgraduate School
Operations Research Dept.
Naval Postgraduate School
Monterey, CA 93940
(408) 646-2605

Cindy Gentillon, INEL
EG&G Idaho
P.O. Box 1625
Idaho Falls, ID 83415-3421
(208) 526-0431

Ernie Lofgren, SAIC
1710 Goodridge Drive
Tier 2-7-1
McLean, VA 22102
(703) 821-4492

Harry Martz, LANL
Statistics Group (A-1)
Analysis and Assessment Division
Los Alamos National Laboratory
Los Alamos, NM 87545
(505) 667-2687

Ali Mosleh, U. of MD
Building 090
Nuclear Engineering
University of Maryland
College Park, MD 20742-2115
(301) 405-5215

Pat O'Reilly, NRC/AOED
MNBB-9112
Nuclear Regulatory Comm.
Washington, DC 20555
(301) 492-8858

END

DATE
FILMED

5 / 14 / 93