NUREG/CP-0063

Proceedings of the

1984 Statistical Symposium on National Energy Issues

Held at
Edgewater Inn
Seattle, WA
October 16-18, 1984

Compiled by: Robert Kinnison, Pamela Doctor

Sponsored by
Division of Risk Analysis and Operations
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
and
Pacific Northwest Laboratory
Operated by
Battelle Memorial Institute

NOTICE

These proceeedings have been authored by a contractor of the United States Government. Neither the United States Government nor any agency thereof, or any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for any third party's use, or the results of such use, of any information, apparatus, product or process disclosed in these proceedings, or represents that its use by such third party would not infringe privately owned rights. The views expressed in these proceedings are not necessarily those of the U.S. Nuclear Regulatory Commission.

Available from

Superintendent of Documents
U.S. Government Printing Office
P.O. Box 37082
Washington, D.C. 20013-7982

and

National Technical Information Service
Springfield, VA 22161

NUREG/CP-0063

Proceedings of the

1984 Statistical Symposium on National Energy Issues

Held at
Edgewater Inn
Seattle, WA
October 16-18, 1984

Compiled by: Robert Kinnison, Pamela Doctor

Manuscript Completed: March 1985
Date Published: July 1985

Sponsored by
Division of Risk Analysis
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555
and
Pacific Northwest Laboratory
Richland, WA 99352
NRC FIN B2386

PREFACE

The 1984 Statistical Symposium on National Energy Issues was the tenth in a series of annual symposia bringing together statisticians and other interested parties who are actively engaged in the pursuit of solving the nation's energy problems. Initially the symposium was sponsored by U.S. Department of Energy (DOE) and named the DOE Statistical Symposium. The symposium is organized by a steering committee made up of representatives from the national laboratories: Pacific Northwest Laboratory, Lawrence Livermore National Laboratory, Sandia National Laboratory, Oak Ridge National Laboratory, Brookhaven National Laboratory, and Los Alamos National Laboratory. Each laboratory has taken turns hosting the symposium. Now it is sponsored by the U.S. Nuclear Regulatory Commission (NRC) and the hosting institution.

The 1984 symposium was hosted by Pacific Northwest Laboratory (PNL). Members of the steering committee for this symposium were chairman Pamela Doctor (PNL), Dale Rasmuson (NRC), David Margolies (Lawrence Livermore Laboratory), Ronald Iman (Sandia National Laboratory), Bill Lever (Oak Ridge National Laboratory), Sam Kao (Brookhaven National Laboratory), Maurice Bryson (Los Alamos National Laboratory).

This years symposium was organized around four special topical sessions: (1) Assessing and Assuring High Reliability, (2) Spatial Statistics, (3) Quantification of Informed Opinion, and (4) Health Effects of Energy Technologies. These were chosen by the steering committee as topics currently of high importance in energy research and data analysis. Much of the value of the symposium to the participants was in the informal discussions generated by the topical sessions presentations.

The 1984 symposium furthered communications among statisticians, the government community, scientists from the national laboratories who are concerned with energy research, the academic community, and industries involved in energy related research. The contributions of all those who participated in and support this symposium are greatly appreciated. Several of the papers that were presented at the symposium are not included in the proceedings because the authors elected not to publish at this time.

PROGRAM

Special Topical Session

Assessing and Assuring High Reliability
Coordinator: Dr. David Nelson, Boeing Computer Services

A Review of the Statistical Aspects of Accelerated Life Testing
William Meeker, Iowa State University

Fault Tolerance as a Means to Achieve High Reliability
David Rose, Boeing Computer Services

The Role of Reliability Modeling in the Design of a Fault-Tolerant Feedwater Controller
Blake F. Putney, Jr., SAI, Palo Alto

Statistical Uncertanties and Unrecognized Relationships
John P. Rankin, Boeing Computer Services International

Special Topical Session

Spatial Statistics
Coordinators: Drs. Pamela Doctor and Tony Olsen

Uncertainty and Sensitivity Analysis of Environmental Transport Models
Timothy S. Margulies, U. S. Nuclear Regulatory Commission

Kriging: An Investigation of Aquifer Pressure on Long Island
Neal Oden, Brookhaven National Laboratory

Empirical Evaluation of Simple and Universal Kriging
Jeanne Simpson, University of Washington

A Review and Assessment of Environmental Contaminant Sampling Plans
Richard Gilbert, Pacific Northwest Laboratory

Special Topical Session

Quantification of Informed Opinion
Coordinator: Dr. Dale Rasmuson, U. S. Nuclear Regulatory Commission

Illiciting and Aggregating Subjective Judgments - Some Experimental Results
Harry F. Martz, Los Alamos

Human Factors Affecting Subjective Judgments
Mary A. Meyer, Los Alamos National Laboratory

Use of Informed Opinion in Seismic Hazard Characterization
R. Mensing, Lawrence Livermore National Laboratory

Contents

Contributed Papers

Special Topical Session

Health Effects of Energy Technologies

Coordinator: Dr. David Margolies, Lawrence Livermore National Laboratory

Statistical Issues Within the National Plutonium Workers Study
Gary Tietjen, Los Alamos National Laboratory

The Hanford Study - A Review of Its Limitations and Controversial Conclusions
Ethyl S. Gilbert

Melanoma Among Lawrence Livermore National Laboratory Employees: An Effect Without a Cause
Dan Moore, Lawrence Livermore National Laboratory

Contributed Papers

Coordinator: Dr. Robert Kinnison, Pacific Northwest Laboratory

Hazard Function Modeling of Early Effects Mortality Risk Associated with Light Water Nuclear Reactor Accidents
Bobby Scott, Lovelace Research Institute

Baysean Methods Using Informed Opinion
Ronald Canfield, Utah State University

Risk Evaluation in High-Altitude Level Flight
James Lechner, National Bureau of Standards

Ranking Alternatives by Stochastic and Relative Importance Judgements
Lee Abramson, Nuclear Regulatory Commission

A Simple Method for Assigning Uncertainties to Weights of Characteristics Based on Paired Comparison
V. R. R. Uppuluri, Oak Ridge National Laboratory

Laplace's Law of Succession and Prediction Intervals
David Rubenstein, Nuclear Regulatory Commission

Comparison of Field In-Situ and Laboratory Measurement Methods for Ra-226
P. R. Engelder, Bendix Field Engineering Corporation

The Variance of Measurements from a Calibration Function Derived from Data Which Exhibit Run-to-Run Differences
A. M. Liebetrau

Sampling Inspection of Nuclear Power Plants
Julius Goodman, Bechtel Power Corporation

Special Topical Session

Assessing and Assuring High Reliability

SESSION ON ASSESSING AND ASSURING HIGH RELIABILITY:

INTRODUCTORY REMARKS

David L. Nelson, Session Coordinator
Boeing Computer Services Company

No set of introductory remarks on reliability needs to contain very much in the way of a defense of the importance of reliability as a topic of study. We all look for as much reliability as we can get for the money spent, whether it be in our personal goods, the equipment we use at work, or the items bought or constructed with the taxes or utility charges we pay. Reliability and the other associated "ilities" are very much in the public consciousness, probably even more than another current topic of popular study, statistical quality control. Suffice it to say that a session concerning the achievement of high reliability, and the assessment of whether you've achieved it when you think you have, is surely an important one when statisticians gather to discuss national energy issues.

In looking at this session, I think it's good to keep in mind a definition from the preface of the book by Barlow and Proschan (1965): "Reliability theory ... is a body of ideas, mathematical models, and methods directed toward the solution of problems in predicting, estimating, or optimizing the probability of survival, mean life, or life distribution of components or systems; other problems considered in reliability theory are those involving the probability of proper functioning of the system at either a specified or an arbitrary time, or the proportion of time the system is functioning properly." (I added the underlining for emphasis.) In this session, we will look at methods of optimizing probabilities by various methods, some not yet well known, and also at tools that can be used to assess the level of reliability that can be expected from a particular system design or type of component.

The application areas in which the papers in this session have a great deal of relevance would seem to include the following:

- Electric turbines and power generation equipment
- Computerized control systems
- Pumps, valves, bearings and other mechanical devices
- Transformers, overload mechanisms and other electronic equipment
- Oil drilling and pumping machinery

Listeners or readers can supply a lot more examples from their own work experience.

The papers given here cover a broad collection of reliability attainment processes. Bill Meeker's paper on accelerated life testing contains the results of decades of work in this area by himself and other researchers. In fact, a whole session at this symposium could have been devoted to just accelerated life testing, given the wealth of useful methodologies that have been developed. We have two papers concerned with an area that is being advanced in aerospace, but has not received much attention as

yet in energy fields -- fault tolerance. The paper by David Rose gives an introduction to fault tolerance as a means of achieving high reliability (showing that it is not merely a different way to model redundancy), while Blake Putney shows a concrete example of the modeling of a digital feedwater controller in a nuclear power plant as a fault tolerant system. Finally, John Rankin gives an overview of the important topics of sneak circuit and common cause failure analyses, along with some interesting examples of each. All of these papers contain some discussion of reliability assessment as an associated tool for the methodologies being presented.

REFERENCE

Barlow, R. E., and F. Proschan (1965), Mathematical Theory of Reliability, John Wiley & Sons, New York.

A REVIEW OF THE STATISTICAL ASPECTS OF ACCELERATED LIFE TESTING

William Q. Meeker, Jr.
Department of Statistics
Iowa State University

ABSTRACT

This paper briefly describes various statistical techniques that are used in accelerated life testing. Topics covered include accelerated life test models, methods for data analysis, test planning, and some special topics such as multiple failure modes and step-stress experiments. Emphasis has been placed on work in the literature that appears to have direct practical value.

1. INTRODUCTION

The purpose of this paper is to provide an introduction to the basic concepts of and the practical statistical literature on accelerated life testing (ALT) and some other closely related topics.

1.1 General Methodology

The basic idea of ALT is to subject units of a product or specimens of a material to higher than usual levels of stress to induce early failures and thus reduce the time that it takes to get information on product life. Results from an ALT are used, through a statistical model, to extrapolate and estimate (or predict) product life at design or use conditions. Yurkowski, Schafer, and Finkelstein (1967), though now somewhat outdated, gives an overview of ALT methodology and reviews most of the literature that was available before 1967. Section 11 of General Electric (1975), which was written by Wayne Nelson, is a compendium of practical methods, important references, and examples for planning and analyzing ALT's. Derringer (1982) and Ahmad and Sheikh (1983) review some aspects of ALT's. Kalbfleisch and Prentice (1980) and Lawless (1982) describe the theory and methods for some of the important statistical methods that can be used to analyze ALT data. In particular, they cover both parametric models and the Cox (1972) proportional hazard model (a quasi-nonparametric model) for regression analysis of censored data. Many of their examples are, however, with applications that are different from ALT's.

1.2 Some Applications

ALT's have been used in many different applications. Some representative examples are:

- Life of capacitors as a function of voltage (Levenbach (1957)).
- Life of an insulating fluid as a function of voltage (Nelson (1970, 1972a)).
- Life of electric motor insulation as a function of temperature (Nelson (1974a, 1982)).
- Life of a pressurized vessel as a function of pressure (Barlow (1982).
- Life of integrated circuits as a function of temperature (see Peck (1975) and Reynolds (1977)).
- Life of incandescent light bulbs as a function of voltage.
- Shelf life of food products as a function of temperature.
- Life of mechanical switches as a function of switching rate.

For other applications, see the bibliography of Meeker (1980).

2. DATA AND MODELS

2.1 Example

Nelson (1970, 1972a) gives failure times (in minutes to breakdown) of insulating fluid samples subject to constant levels of voltage. The experiment consisted of tests at 26, 28, 30, 32, 34, 36, and 38 kV. The purpose of the experiment was to investigate the relationship between life and voltage level and to explore the possibility of using ALT's to predict life of similar products. For purposes of discussion, we will assume that it was also desired to estimate life at 20 kV. We will use these data to illustrate some of the statistical techniques that are reviewed in this paper.

2.2 Censored Data

One of the distinguishing features of life data is that they are usually censored. That is, for some observations, an analyst may not know the exact time of failure. Typically an "observation" is one of the following:

- An exact failure if failure has occurred and the failure time has been recorded.

- A right censored observation if failure has not occurred when the data are analyzed.

- A left censored observation if failure occurred before the first time that a unit was inspected.

- An interval censored observation if all that is known is that the unit failed between two points in time (usually two inspections).

Most ALT's result in some right censored observations. Such observations do not cause any insurmountable problems, as a variety of statistical methods and computer programs are available to analyze censored data. Nelson's insulating fluid data are not censored, as all test units failed before the end of the experiment.

2.3 Models for Accelerated Life Testing

In the most widely used ALT models, a parameter (e.g., median life) of a particular parametric life distribution is assumed to be functionally related to a stress variable like voltage, temperature, pressure, or cycling rate. In most practical applications, the form of the relationship is assumed to be known and unknown parameters in the relationship are estimated from the available data. The most common models for ALT's use a distribution like the lognormal or Weibull and one of the following life-stress relationships:

- The Inverse Power Law model which is used to model life as a function of stresses like voltage, pressure, and cycling rate.

- The Arrhenius and Eyring relationships which are used to model life as a function of temperature.

Chapter 9 of Mann, Schafer, and Singpurwalla (1974) gives formulae, other details, and references to some applications for these models. Singpurwalla (1975) surveys other physical models that have been used in ALT applications. Also see Goba (1969) and Grange (1971).

The most commonly used life distributions for ALT's are the exponential (see Mann, Schafer, and Singpurwalla (1974), Chapter 9), the Weibull (see Nelson (1970)), the Lognormal (see Nelson (1971)), and the inverse Gaussian distribution (see Bhattacharyya and Fries (1981) and Ahmad and Sheikh (1983)). These parametric models, when used correctly, use the available data most efficiently. Kalbfleisch and Prentice (1980) and Lawless (1982) give general theory for estimating the parameters of the parametric models and the Cox (1972) quasi-nonparametric proportional hazard model.

For the insulating fluid example, the inverse power law/Weibull distribution model is assumed. This model is commonly used for dielectric materials (e.g., capacitors and electrical insulation) under voltage stress. Then the assumed model for time to failure T is

$$P(T<t) = F_T(t) = 1 - \exp[(t/\alpha)^{\beta}], \quad t>0$$

where α and β are positive scale and shape parameters respectively, and the scale parameter α is an inverse power function of applied stress V. That is

$$\alpha = \alpha(V) = \gamma_0 V^{\gamma_1}$$

where γ_0, γ_1, and β are usually unknown parameters to be estimated from the available data.

For this model, it is more convenient to work with the logarithms of failure time data. If T follows a two-parameter Weibull distribution, then $Y = \ell n(T)$ follows a smallest extreme value distribution with cdf

$$P(Y<y) = F_Y(y) = 1 - \exp\{-\exp[(y-u)/\sigma]\}, \quad -\infty < y < \infty \qquad (2.1)$$

where $\mu=\ell n(\alpha)$ and $\sigma=1/\beta$ are location and scale parameters, respectively. Then the inverse power law model can be written

$$\mu = \gamma_0 + \gamma_1 x \qquad (2.2)$$

where $x=\ell n(V)$.

2.4 Modeling Difficulties

The most difficult problem facing the users of ALT's is that of finding a suitable model that will allow extrapolation to lower levels of stress. Many research papers (especially in the statistical literature) state that they are assuming that the specified model is true and proceed without any mention of what might happen if the assumption is not true. Most experiments do not provide enough information to check for anything but very pronounced departures from the assumed model, even within the range of one's data (see Meeker (1984)), and usually there is little of no empirical information to check for the assumed model over the range of extrapolation. In practice, there is no easy way around this problem. Typically, investigators will use information from previous experiments with similar products, pilot experiments, and models based on knowledge of the physics and chemistry of the failure mechanism(s), to justify an assumed model. In any case, it is best to keep experiments as simple as possible, to use test plans and methods of estimation that are robust to departures from model and other assumptions, and to limit the amount of applied stress. See Meeker (1981, 1984) and Meeker and Hahn (1984) for further discussion.

3. ANALYSIS OF ACCELERATED LIFE TEST DATA

3.1 Graphical Analysis

Nelson (1972a) gives methods for graphical analysis of ALT data. Graphical display and estimation of ALT data is useful for

- Getting an overall easy-to-explain picture of one's data,
- Obtaining simple subjective estimates of model parameters and distribution percentiles as a function of stress,
- Subjectively assessing model assumptions.

The basic idea is to treat the sub-experiments at each level of stress as separate samples and to plot the nonparametric cdf estimates for each on a single hazard or probability plot. Parametric estimates can also be added to the plot. The estimates from the individual analyses can be compared with those obtained with an assumed life-stress relationship.

3.2 Maximum Likelihood Estimation

Maximum likelihood (ML) is the most popular analytical method for estimating the parameters of ALT models. The main reasons for this are:

- ML is versatile, being able to handle many different models and kinds of data.
- ML is powerful, yielding point estimates, approximate confidence intervals, and tests of hypotheses.
- ML estimators have desirable large sample properties when compared to competing estimators.
- There are computer programs available for ML estimation of life data.

3.3 Checking Model Assumptions

Seriously incorrect inferences can result from an inadequate model. The most important assumption is that of the relationship between life and stress. To make inferences at the design stress generally requires extrapolation outside the range of one's data. This requires faith in the assumed ALT model, often justified (correctly or incorrectly) on the basis of a physical model or previous empirical experience with similar test units. Users of ALT's should carefully scrutinize their data for departures from the assumed model. When doint this, it should be remembered that

- A statistical model can only be expected to be an approximation to reality.

- A model is satisfactory if it is accurate enough to consistently (allowing for statistical variability) provide useful answers.

- Typically, only the most pronounced departures can be detected with small samples.

Nelson (1973) shows how to analyze residuals with censored data and applies the techniques to perform diagnostic checks for ALT data and models. See Nelson (1972a, 1972c) and chapter 6 of Lawless (1982) for other methods for detecting departures and for more discussion and suggestions.

3.4 Computer Programs for Analyzing Accelerated Life Test Data

The following user-oriented programs can be used to obtain graphical displays of nonparametric estimates of life distributions, to produce probability plots, and to fit parametric models to ALT data:

1. STATPAC (Nelson, Morgan, and Caporal, 1983).

2. CENSOR (Meeker and Duke, 1981).

3. SURVREG (Preston and Clarkson, 1983).

4. STAR (Buswell, Meeker, and Myers, 1984).

STATPAC allows the most flexibility for choosing parametric life-stress models. Both CENSOR and STATPAC can be used to conduct Monte Carlo simulations. SURVREG and STAR can be used to fit Cox (1972) proportional hazard models. Some other well known statistical packages including BMDP and SAS also have procedures for fitting Cox's proportional hazard model.

3.5 Analysis of Nelson's Insulating Fluid Data

For the insulating fluid example, STAR was used to obtain separate nonparametric estimates and to fit separate Weibull distributions to the data at each level of voltage. Although the detailed computer output is not shown here, the nonparametric estimates (plotted points) and the ML estimates (plotted lines) are shown on Weibull probability paper in Figure 1 for the subexperiments at 26, 30, 34, and 38 kV. The plotted points do not deviate significantly from linearity, indicating that the Weibull distribution provides a reasonable fit for these data over this range of voltage. Also, the slopes of the fitted lines do not deviate significantly from each other, so there is no strong evidence against the assumption of a common Weibull shape parameter over this range of

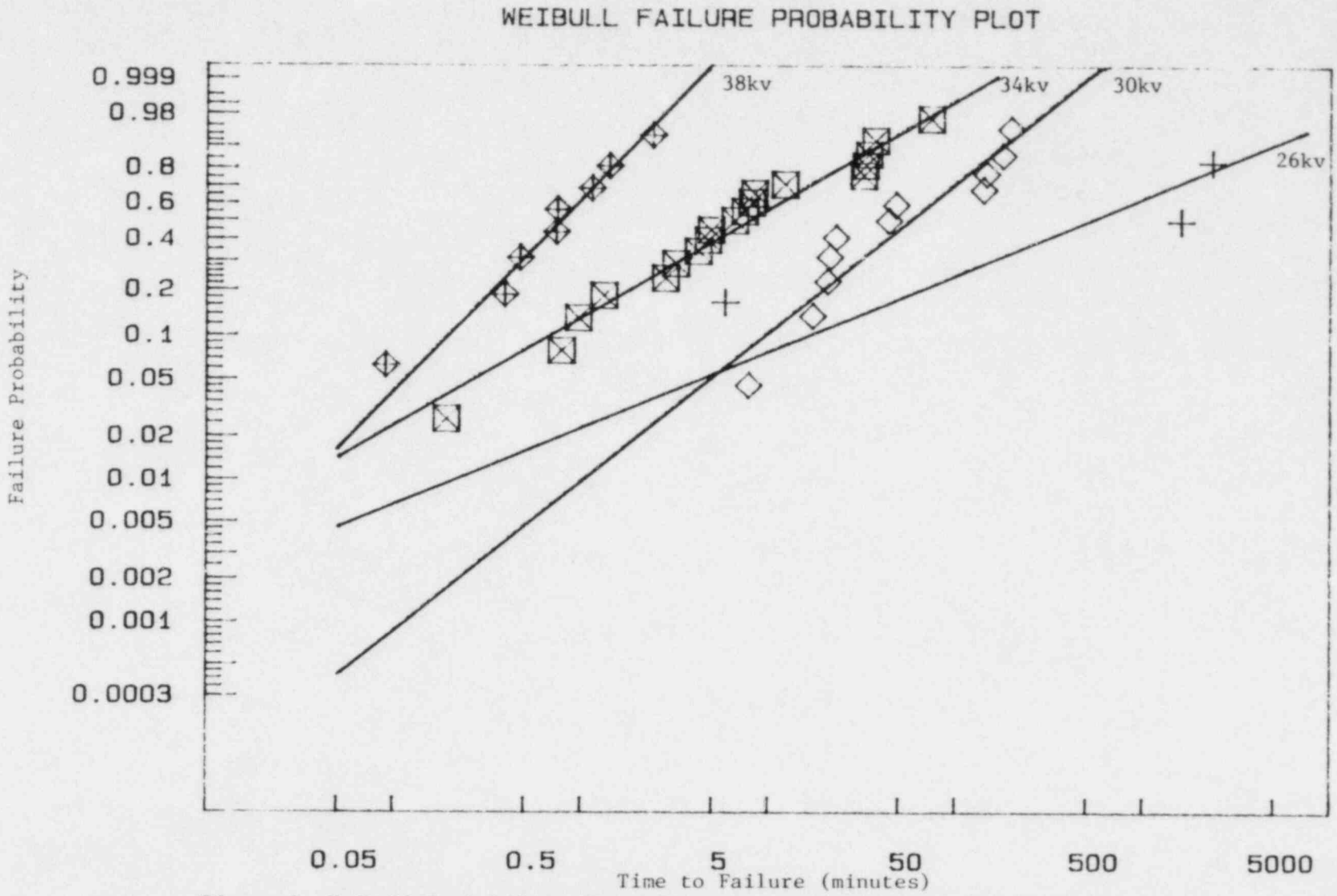

Figure 1 Weibull Probability Plot of Insulating Fluid Data with Individual Maximum Likelihood Estimates.

voltage. The smallest observation at 26 kV appears to be an outlier, but analysis shows that it is statistically consistent with the assumed model and the rest of the data.

We also used STAR to fit the inverse power law model (equations (2.1) and (2.2)) to these data. A summary of the results and a table of estimated percentiles of the life distribution at the design voltage of 20 kV are given in Figure 2. Figure 3 gives another probability plot with the straight lines now depicting the model fit (notice the common slope of the lines corresponding to the assumed common Weibull shape parameters). The line labeled 20 kV was extrapolated from the model fit. Figure 4 shows the residuals from the regression model, plotted on Weibull probability scale. If there had been a significant departure from linearity in this plot, it would have been an indication that the residuals do not follow the assumed Weibull distribution.

3.6 Other Methods of Estimation

Although maximum likelihood is the most widely used technique for estimating the parameters of ALT models, a number of other methods have appeared in the literature. Hahn and Nelson (1974) compare graphical, ML, and linear (based on observed order statistics) methods of estimation for censored regression data. Schmee and Hahn (1979) present and evaluate an iterative least squares method and Bhattacharyya and Soejoeti (1981) investigate the properties of "modified least squares" estimators. Barlow (1982) gives a very interesting example, using Bayesian-likelihood methods for ALT data analysis. Nonparametric models and corresponding methods of estimation have been presented, for example, by Shaked, Zimmer, and Ball (1979), Proschan and Singpurwalla (1980), Basu and Ebrahimi (1982), and Shaked and Singpurwalla (1980). There is a large amount of literature on the Cox proportional hazards model. See Kalbfleisch and Prentice (1980) and Lawless (1982) for references.

4. PLANNING ACCELERATED LIFE TESTS

ALT's provide timely information on product life. Well designed experiments can improve the precision with which inferences are made. Although standard experimental design techniques like randomization and blocking are important in planning ALT's, the presence of censored data and the need for extrapolative inferences complicates the planning of experiments. Nelson (1972b, 1974b) gives general guidance for planning ALT's. Also see Little and Jebe (1974).

Summary Report of: Insulating Fluid ALT

Distribution is Weibull

The natural logs of these observations follow a[n] extreme-value distribution.

Intercept term included in model.

Maximum value of the log-likelihood is -137.7476

Parameter estimates for the extreme-value distribution:

			95.0% Confidence Limits	
	Estimate	Std Error	Lower	Upper
Scale	1.287739	0.1133354	1.083667	1.530239
Intercept	64.84719	5.619756	53.83025	75.86414
VOLTAGE	-17.72958	1.606833	-20.87961	-14.57955

Variance-Covariance Matrix:

	Scale	Intercept	VOLTAGE
Scale	0.01284491	-0.00888925	0.000924538
Intercept	-0.008889254	31.58166	-9.026538
VOLTAGE	0.000924538	-9.026538	2.581914

Correlation Matrix:

	Scale	Intercept	VOLTAGE
Scale	1.000000	-0.01395668	0.00507678
Intercept	-0.01395668	1.000000	-0.9996150
VOLTAGE	0.00507678	-0.9996150	1.000000

Insulating Fluid ALT

WEIBULL QUANTILE ESTIMATES WITH 95% CONFIDENCE LIMITS

QUANTILE	MINUTES	STD ERROR	LOWER CL	UPPER CL
0.01	333.728	333.521	47.04691	2367.302
0.05	2722.430	2465.738	461.1517	16071.99
0.10	6879.012	6009.730	1240.865	38135.30
0.20	18080.41	15319.34	3434.279	95187.78
0.30	33075.44	27611.80	6438.064	169924.5
0.40	52528.29	43464.10	10373.33	265991.7
0.50	77819.17	64022.72	15510.97	390421.7

Figure 2 Summary of Maximum Likelihood Estimation of the Inverse Power Law Model Fitted to the Insulating Fluid Data

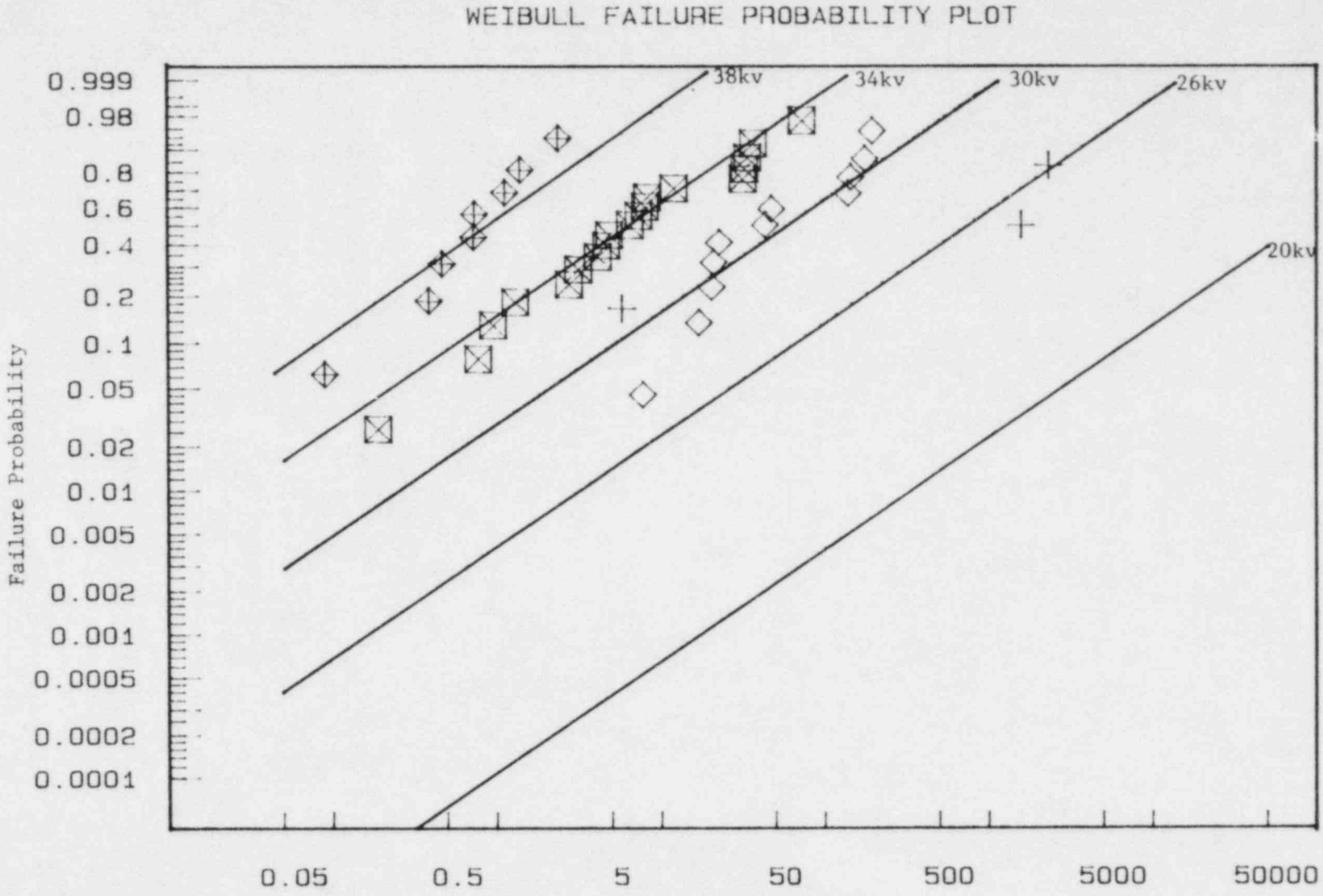

Figure 3 Weibull Probability Plot of Insulating Fluid Data With Inverse Power Law Maximum Likelihood Estimates

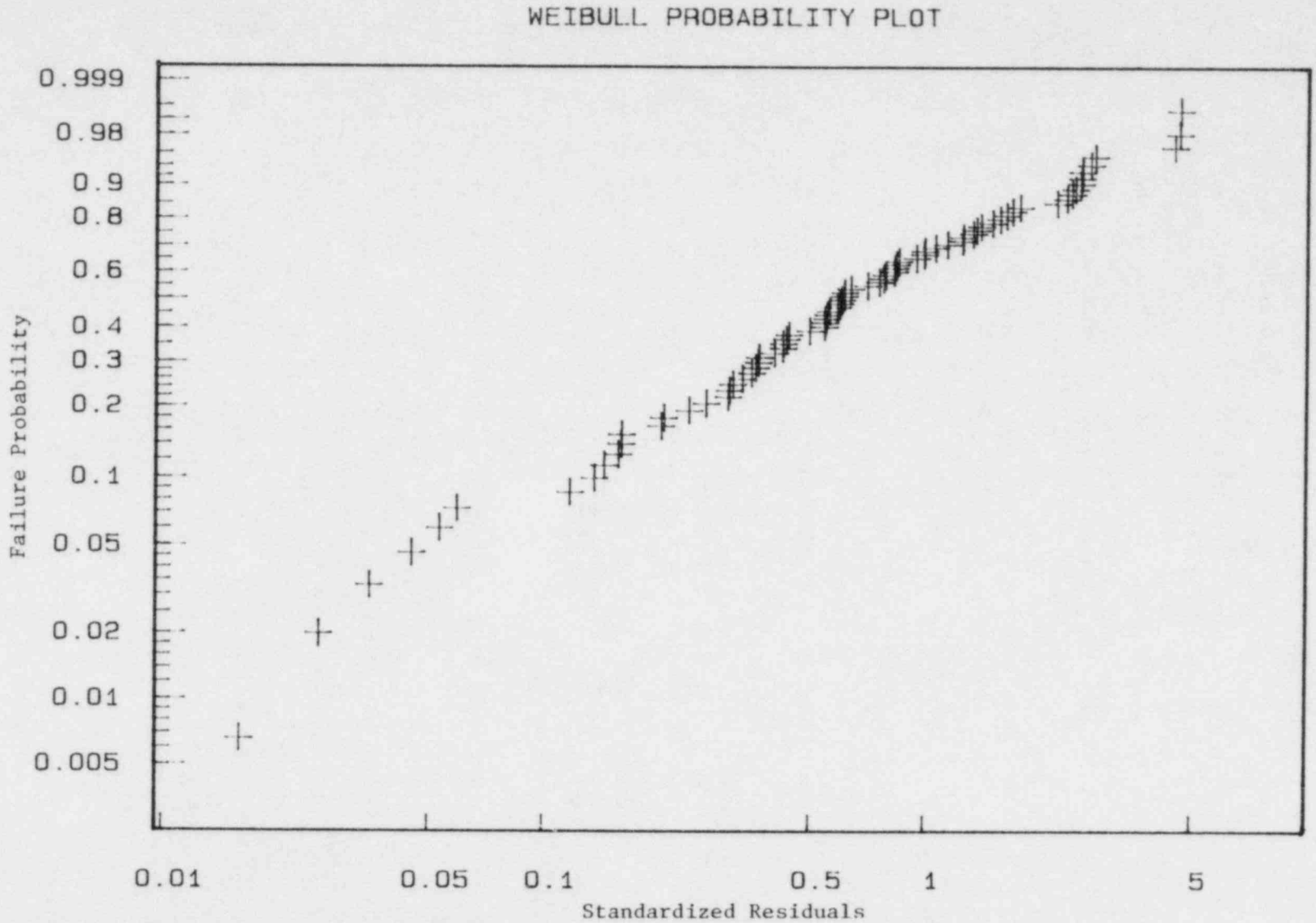

Figure 4 Weibull Probability Plot of Residuals from the Inverse Power Law Model

4.1 Test Plans and Constraints

There are many different ways to conduct an ALT. Most practical experiments use tests at 3 or more different levels of stress. Simultaneous testing gives the quickest results. However, when there is a limit on the number of test positions, some form of sequential testing can be used. The length of the test is often fixed in advance, leading to Type I right censoring. In other cases, tests are run until a specified number of failures have been observed, leading to Type II right censoring. Type II censoring provides better control over the amount of precision that an experiment will provide. More frequently, some combination of these must be used because of practical constraints on time, the number of test positions, and desired precision.

4.2 Optimum Test Plans

Chernoff (1962) gives optimum ALT plans for the exponential distribution and a variety of testing situations and types of censoring. For the Weibull and lognormal distributions respectively, Meeker and Nelson (1976) and Kielpinski and Nelson (1976) give charts for finding optimum ALT plans for Type I censoring. These papers show how to choose the levels of stress and the allocations of test units to minimize the variance of the maximum likelihood estimator of a percentile of the life distribution at design conditions. Corresponding theory for developing these plans is given in Nelson and Kielpinski (1976) and Nelson and Meeker (1978). Mann (1972) and Escobar and Meeker (1984) give theory for optimum ALT's with Type II censoring.

4.3 Compromise Test Plans

Optimum test plans have practical limitations as they a) only use 2 levels of stress, b) typically require much extrapolation, and c) tend not to be robust to departures from the assumptions needed to derive them. Thus, in practice it is usually necessary to sacrifice some statistical efficiency and use a compromise plan. Nelson and Kielpinski (1976) suggested some methods for finding compromise plans. Meeker (1984) outlines criteria for comparing ALT plans, gives several methods for obtaining practical compromise test plans, and compares these with optimum test plans. Meeker and Hahn (1984) use some of the results of Meeker (1984) and further numerical evaluations to develop a set of easy-to-use rules that can be used to obtain test plans that meet practical constraints and that have good statistical properties.

4.4 Evaluation and Comparison of Test Plans

It is possible to use a computer program to compare alternative test plans and, for example, to directly evaluate the trade-offs involved when going from an optimum plan to a compromise plan. A computer program to perform this task is currently being developed.

4.5 Simulating Accelerated Life Test Experiments to Evaluate Test Plans

During the planning stage of an ALT it is often useful to generate and analyze simulated data. The data can be generated either by hand or with the computer. Such pseudo-experiments can be used to see if reasonable results can be obtained with the available resources and to get a rough idea of the properties of the estimates that will be obtained from the real experiment. Repeating the pseudo-experiments a large number of times (40 to 60 or more) provides information on the sampling distributions of estimators. These simulations and analyses can be automated within some statistical packages (e.g., STATPAC and CENSOR). Nelson (1983) gives an example using of simulations to help plan a complicated ALT.

4.6 Choosing the Sample Size

Meeker and Nelson (1976) show how to choose the sample size needed to estimate a particular percentile of the life distribution at design conditions with a specified degree of precision. This formula can also be used with the compromise plans given in Meeker (1984) and Meeker and Hahn (1984).

4.7 Test Planning Example

Meeker (1984) gives an example in which Nelson's insulating fluid experiment was treated as a pilot experiment to obtain information to help plan a longer and larger experiment to estimate the 10th percentile of the life distribution of the insulating fluid at 15 kV. The maximum test voltage for the future experiment was to be 25 kV and the experiment was to be completed in 28 days. Table 3 compares

a) the optimum test plan (only 2 sub-experiments),

b) several compromise plans (each with 3 sub-experiments), and

For each plan, the sample size was chosen to be large enough to estimate the 10th percentile of the life distribution at 15 kV to within a factor of 2 with 95 percent confidence. Although the compromise plans require somewhat larger samples, these test plans are more robust to departures

Table 1

Test Plans for the Insulating Fluid Example

(Sample sizes chosen for equal precision in estimating the 10th percentile of life at 15kV)

Statistically Optimum Test Plan

Voltage kV	Proportion of Test Units Allocated	Number of Test Units Allocated	Probability of Failure	Expected Number of Test Units Failing
19.8	0.77	197	0.30	58.6
25.0	0.23	60	1.00	59.9
		257		

Best Standard Test Plan

Voltage kV	Proportion of Test Units Allocated	Number of Test Units Allocated	Probability of Failure	Expected Number of Test Units Failing
18.3	0.33	113.	0.12	13.4
21.4	0.33	113.	0.65	74.1
25.0	0.33	113.	1.00	113.4
		339		

Best Compromise Test Plan

Voltage kV	Proportion of Test Units Allocated	Number of Test Units Allocated	Probability of Failure	Expected Number of Test Units Failing
19.3	0.58	174.	0.23	39.2
22.0	0.20	60.	0.78	46.4
25.0	0.22	64.	1.00	64.3
		298		

from the assumptions that were used to find the plans. See Meeker (1984) for further details and Meeker and Hahn (1984) for a different example.

5. SOME SPECIAL TOPICS

5.1 Comparing Products with Accelerated Tests

ALT's are sometimes used to compare two competing products, materials, or manufacturing methods. Such comparisons can be relatively straightforward if stress affects both products in roughly the same way (i.e., similar models, activation energies, etc.). If, however, stress affects one product more than another (e.g., a material with a higher activation energy is used in one product), it is possible for a product (say product A) to perform better than another product (say product B) at the accelerated levels of stress but for the comparison between the products at the design stress to favor product B. Justifying conclusions based on an ALT with results like this is difficult. See Nelson (1972c) for further discussion on comparing products with ALT's.

5.2 Accelerated Life Tests and Infant Mortality

Infant mortality is an important problem in reliability, particularly in the electronics manufacturing industry. Most infant mortality is caused by a small proportion of defective units, within a product population, that fail early in life. ALT's can be used to assess the extent of infant mortality in a product population. Peck (1978) discusses this subject.

5.3 Step-Stress and Progressive-Stress Experiments

In some ALT's, stress is changed as a function of time. See Allen (1959) and Nelson (1980) for theory and data analysis methods for continuously increasing and step increasing stress, respectively. Miller and Nelson (1983) give methods for planning simple step-stress experiments with the exponential distribution.

5.4 Multiple Failure Mode

There is often more than one potential cause of product failure. The failure mode for failed units in a life test may or may not be known. Sometimes testing at accelerated conditions will cause failure modes that would never occur at the design conditions.

Most ALT experiments involving more than one cause of failure assume that

a) cause of failure is known for all units that failed and

b) the causes act independently.

These assumptions greatly simplify data analyses and interpretation. However, they are difficult to check and desired inferences may not be robust to departures from them.

Basic theory and methods for "competing risk" problems are covered, for example, by David and Moeschberger (1978), Birnbaum (1979), Nelson (1982), and Lawless (1982). Nelson (1975) gives graphical methods for analyzing ALT data with more than one failure mode. Nelson (1974a) gives ML methods for the same problem. Similar methods are given by Klein and Basu (1982) for the exponential distribution and by Klein and Basu (1981) for the Weibull distribution.

Multiple failure modes can cause serious problems when making inferences from an ALT if

1. The cause of failure cannot be determined and stress has different effects on the different failure modes.

2. An important failure mode at the design stress is not accelerated and is thus not observed in the ALT (see Allen (1963)).

If the different causes of failure are not independent, the problem becomes much more complicated and a parametric model is generally required. For further information, see Chapter 4 of David and Moeschberger (1978), Birnbaum (1979), and Chapter 7 of Kalbfleisch and Prentice (1980).

5.5 More Than One Accelerating Stress

Most ALT's use only one accelerating stress. If more than one accelerating stress is used (implying extrapolation in two or more dimensions), there might be some implied gains in precision for multiple-stress experiments (see Derringer (1982)), but modeling problems become very complicated, especially if there is any chance that there is important interaction between the stresses. See Meeker (1981) for further discussion.

A somewhat easier problem, conceptually, is that of a single accelerating stress (implying extrapolation in a single dimension) with other experimental factors. These experiments are often motivated by a desire to model life as a function of environmental stress(es) and are common in practice. Extrapolation in only one dimension simplifies the modeling process.

Methods of estimation can be extended in a straightforward manner to handle multiple-stress ALT experiments. The difficult problems are in modeling and, to a lesser extent, in test planning.

5.6 Accelerated Life Testing with Binary Responses

In some applications, failure times are unobservable or a product might simply operate or fail to operate with some probability that is a function of stress. For example, in some experiments the only available information is whether or not a unit failed within a specified mission time. Data from such experiments are called quantal-response data. See Easterling (1975) for an example. Models for quantal-response data express the probability of failure as a function of stress. Cox (1970) gives models and data analysis methods for quantal-response data. Meeker and Hahn (1977) discuss methods of planning ALT experiments for quantal response data and a logistic model. Meeker and Hahn (1978) compare optimum and various compromise plans for this particular model.

5.7 Accelerated Testing With Measured Degradation

In some reliability testing situations it is possible to measure the amount degradation on individual test units. In such cases failure might be defined as a particular level of degradation and the failure time is the first time at which a unit crosses that threshold. Although life data analysis techniques can be used to analyze time-to-failure data when there is a suitable definition of failure, there may be important information in the degradation measurements themselves. This is particularly true when a large proportion of units is censored. Nelson (1979) discusses the analysis of data from an accelerated test in which destructive testing was required so that only one measurement could be made on each unit. Often it is possible to make a sequence of degradation measurements over time. Hooper and Amster (1982), and Amster and Hooper (1983) discuss accelerated testing with a sequence of degradation measurements. Growth curve modeling techniques play an important role in analyzing measured degradation data. See Draper and Smith (1981) and Amster and Hooper (1983) for a review of the literature in this area. Like other types of ALT's, the most difficult problem facing data analysts is that of finding and validating statistical models.

ACKNOWLEDGEMENTS

Many of the ideas presented in this paper were developed through conversations and work done with Gerry Hahn and Wayne Nelson of the General Electric Corporate Research and Development Center, and Sig Amster, Blan Godfrey, and Jeff Hooper, among others, at AT&T Bell Laboratories. Luis Escobar made a number of helpful comments on an earlier draft of this paper.

REFERENCES

Ahmad, M. and Sheikh, A. K. (1983), "Accelerated Life Testing," Paper presented at the annual meeting of the American Statistical Association, Toronto, Canada.

Allen, W. R. (1959), "Inference from Tests with Continuously Increasing Stress," Operations Research 7, 303-312

Allen, W. R. (1963) "A Pitfall in Accelerated Life Testing," Naval Research Logistics Quarterly 10, 271-273.

Amster, S. J. and Hooper, J. H. (1983), "Accelerated Life Tests with Measured Degradation and Growth Curve Models," Paper presented at the annual meeting of the American Statistical Association, Toronto, Canada.

Barlow, R. E. (1982), "Accelerated Life Tests and Information," Radiation Research 90, 90-97.

Basu, A. P. and Ebrahimi, N. (1982), "Nonparametric Accelerated Life Testing," IEEE Transactions on Reliability R-31, 432-435.

Bhattacharyya, G. K. and Fries, A. (1981), "Inverse Gaussian Regression and Accelerated Life Tests," Technical Report No. 659, Department of Statistics, University of Wisconsin.

Bhattacharyya, G. K. and Soejoeti, Z. (1981), "Asymptotic Normality and Efficiency of Modified Least Squares in some Accelerated Life Test Models," Sankhya Series B 43, 18-39.

Birnbaum, Z. W. (1979), On the Mathematics of Competing Risks, U.S. Department of HEW, Publication No. (PHS) 79-1351.

Buswell, G.D., Meeker, W.Q., and Myers, D.H. (1984), "STAR--Statistical Analysis of Reliability Data." Internal AT&T Bell Laboratories Document.

Chernoff, H. (1962), "Optimal Accelerated Life Designs for Estimation," Technometrics 4, 381-408.

Cox, D. R. (1970), Analysis of Binary Data, London: Methuen.

Cox, D. R. (1972), "Regression Models and Life Tables," (with discussion), Journal of the Royal Statistical Society, Series B34, 187-220.

David, H. A. and Moeschberger, M. L. (1978), The Theory of Competing Risks, Charles Griffin & Co., London.

Derringer, G. C. (1982), "Considerations in Single and Multiple-Stress Accelerated Life Testing," Journal of Quality Technology 14, 130-134.

Draper, N. R. and Smith, H. (1981), Applied Regression Analysis, Second Edition, New York: John Wiley and Sons, Inc.

Easterling, R. G. (1975), Reliability Estimation and Sensitivity Testing." Microelectronics and Reliability 14, 141-152.

Escobar, L. A. and Meeker, W. Q. (1984), "Optimum Accelerated Life Tests with Type II Censored Data," Department of Statistics, Iowa State University.

General Electric (1975), "Reliability Manual for Liquid Metal Fast Breeder Reactors." General Electric Company Corporate Research and Development Report SRD-75-064, General Electric Company, Schenectady, NY. Copies available from Mr. Richard Gilchrist, Fast Breeder Reactor Department, General Electric Company, De Guigne Drive, Sunnyvale, CA 94806.

Grange, J. M. (1971), "Study on the Validity of Electronic Parts Stress Models," IEEE Transactions on Reliability R-20, 136-142.

Goba, F. A. (1969), "Bibliography on Thermal Aging of Electrical Insulation," IEEE Transactions on Electrical Insulation IE-4, 31-58.

Hahn, G. J. and Nelson, W. (1974), "A Comparison of Methods for Estimating Relationships between Product Life and Stress from Censored Data," IEEE Transactions on Reliability R-23, 321-332.

Hooper, J. H. and Amster, S. J. (1982), "Accelerated Life Testing with Measured Degradation," Paper presented at the annual meeting of the American Statistical Association, Cincinatti, Ohio.

Kalbfleisch, J. D. and Prentice, R. L. (1980), The Statistical Analysis of Failure Time Data, New York: John Wiley and Sons, Inc.

Kielpinski, T. J., and Nelson, W. (1975), "Optimum Censored Accelerated Life Tests for Normal and Lognormal Life Distributions," IEEE Transactions on Reliability R-24, 310-320.

Klein, J. P., and Basu, A. P. (1981), "Weibull Accelerated Life Tests when there are Competing Causes of Failure," Communications in Statistics Theory and Methods A10, 2073-2100.

Klein, J. P. and Basu, A. P. (1982), "Accelerated Life Testing Under Competing Exponential Failure Distributions," IAPQR Transactions 7, 1-20.

Lawless, J. F. (1982), Statistical Models and Methods for Lifetime Data, New York: John Wiley and Sons, Inc.

Levenbach, G. J. (1957), "Accelerated Life Testing of Capacitors," IEEE Transactions on Reliability and Quality Control RQC-10, 9-20.

Little, R. E. and Jebe, E. H. (1974), Statistical Design of Fatigue Experiments, New York: Halsted Press.

Mann, N. R. (1972), "Design of Over-Stress Life-Test Experiments when Failure Times have the Two-Parameter Weibull Distribution," Technometrics 14, 437-451.

Mann, N. R., Schafer, R. E., and Singpurwalla, N. D. (1974), Methods for Statistical Analysis of Reliability and Life Data, New York: John Wiley and Sons, Inc.

Meeker, W. Q. (1980), "Bibliography on Accelerated Testing," Department of Statistics, Iowa State University.

Meeker, W.Q. (1981), "Accelerated Life Testing--Problems and Prospects." Section 9 of Experimental Techniques for Investigating the Degradation of Electrical Insulation, Electric Power Research Institute Technical Report EL-1854.

Meeker, W. Q. (1984), "A Comparison of Accelerated Life Test Plans for the Weibull and Lognormal Distributions and Type I Censored Data," Technometrics 26 157-171.

Meeker, W. Q. and Duke, S. D. (1981), "CENSOR--A User-Oriented Computer Program for Life Data Analysis," American Statistician 35, 112.

Meeker, W. Q. and Hahn, G. J. (1977), "Asymptotically Optimum Over-stress Tests to Estimate the Survival Probability at a Condition with a Low Expected Failure Probability," Technometrics 19, 381-399.

Meeker, W. Q. and Hahn, G. J. (1978), "A Comparison of Accelerated Test Plans to Estimate the Survival Probability at a Design Stress," Technometrics 20, 245-247.

Meeker, W.Q. and Hahn, G.J., (1984) "Improved Simple Accelerated Life Test Plans." Presented at the 144st Annual Meetings of the American Statistical Association, Philadelphia, PA, August 1984.

Meeker, W. and Nelson, W. (1976), "Optimum Accelerated Life Tests for Weibull and Extreme Value Distributions," IEEE Transactions on Reliability R-24, 321-332.

Miller, R. and Nelson, W. (1983), "Optimum Simple Step-Stress Tests for Accelerated Life Testing," IEEE Transactions on Reliability R-32, 59-65.

Nelson, W. B. (1970), "Statistical Methods for Accelerated Life Test Data--the Inverse Power Law Model," General Electric Company TIS Report 71C001, Schenectady, N.Y.

Nelson, W. (1971), "Analysis of Accelerated Life Test Data--Part 1: The Arrhenius Model and Graphical Methods," IEEE Transactions on Electrical Insulation EI-6, 165-181.

Nelson, W. (1972a), "Graphical Analysis of Accelerated Test Data with the Inverse Power Law Model," IEEE Transactions on Reliability R-21, 2-11, correction on page 155 of the same volume.

Nelson, W. (1972b), "Analysis of Accelerated Life Test Data--Part 2: Numerical Methods and Test Planning," IEEE Transactions on Electrical Insulation EI-7, 36-55.

Nelson, W. (1972c), "Analysis of Accelerated Life Test Data--Part 3: Product Comparisons and Checks on the Validity of the Model and Data," IEEE Transactions on Electrical Insulation EI-7, 99-119.

Nelson, W. (1973), "Analysis of Residuals from Censored Data," Technometrics 15, 697-715.

Nelson, W. B. (1974a), "Analysis of Accelerated Life Test Data with a Mix of Failure Modes by Maximum Likelihood," General Electric CR&D TIS Report 74CRD160, Schenectady, N.Y.

Nelson, W. (1974b), "A Survey of Methods for Planning and Analyzing Accelerated Tests," IEEE Transactions on Electrical Insulation, EI-9, 12-18.

General Electric (1975), "Reliability Manual for Liquid Metal Fast Breeder Reactors." General Electric Company Corporate Research and Development Report SRD-75-064, General Electric Company, Schenectady, NY. Copies available from Mr. Richard Gilchrist, Fast Breeder Reactor Department, General Electric Company, De Guigne Drive, Sunnyvale, CA 94806.

Nelson, W. (1975), "Graphical Analysis of Accelerated Life Test Data with a Mix of Failure Modes," IEEE Transactions on Reliability R-24, 230-237.

Nelson, W. B. (1979), "Analysis of Performance Degradation Data from Accelerated Tests," General Electric Company TIS Report 79CRD217, Schenectady, N.Y..

Nelson, W. B. (1980), "Accelerated Life Testing - Step-Stress Models and Data Analysis," IEEE Transactions on Reliability R-29, 103-108.

Nelson, W. (1982), Applied Life Data Analysis, New York: John Wiley and Sons.

Nelson, W. B. (1983), "Monte Carlo Evaluation of Accelerated Life Test Plans," Paper presented at the 143rd Annual Meeting of the American Statistical Association, Toronto, Canada.

Nelson, W. and Kielpinski, T. (1976), "Theory for Optimum Accelerated Life Tests for Normal and Lognormal Distributions," Technometrics 18, 105-114.

Nelson, W. and Meeker, W. Q. (1978), "Theory for Optimum Accelerated Censored Life Tests for Weibull and Extreme Value Distributions," Technometrics 20, 171-177.

Nelson, W. B., Morgan, C., and Caporal, P. (1983), "1983 STATPAC Simplified - a Short Introduction to how to Run STATPAC, a General Purpose Statistical Package for Data Analysis," General Electric CR&D TIS Report 83CRD146, Schenectady, N.Y.

Peck, D. S. (1975), "Practical Applications of Accelerated Testing - Introduction," Proceedings of the 13th Annual Reliability Physics Symposium, 253-254.

Peck, D. S. (1978), "New Concerns About Integrated Circuit Reliability," Proceedings of the 16th Annual Reliability Physics Symposium, 1-6.

Preston, D. and Clarkson D. (1983), "SURVREG: A Program for Interactive Analysis of Survival Regression Models," The American Statistician 37, 174.

Proschan, F. and Singpurwalla, N. D. (1980), "A New Approach to Inference from Accelerated Life Tests," IEEE Transactions on Reliability R-29, 98-102.

Reynolds, F. H. (1977), "Accelerated-Test Procedures For Semiconductor Components," Proceedings of the 15th Annual Reliability Physics Symposium, 253-254.

Schmee, J. and Hahn, G. J. (1979), "A Simple Regression Procedure for Regression Analysis from Censored Data with Applications to Accelerated Life Testing," Technometrics 21, 417-432.

Shaked, M. and Singpurwalla, N. D. (1982), "Nonparametric Estimation and Goodness-of-Fit Testing of Hypotheses for Distributions in Accelerated Life Testing," IEEE Transactions on Reliability R-31, 69-74.

Shaked, M., Zimmer, W. J., and Ball, C. A. (1979), A Nonparametric Approach to Accelerated Life Testing," Journal of the American Statistical Association 79, 694-699.

Singpurwalla, N. D. (1975), "Annotated Bibliography on some Physical Models in Accelerated Life Testing and Models for Fatigue Failure," George Washington University Technical Memorandum TM-64901.

Yurkowski, W., Schafer, R. E., and Finkelstein, J. M. (1967), "Accelerated Testing Technology," Vol. 1 and 2, Technical report No. RADC-TR-67-420, November 1967, Rome Air Development Center Air Force Systems Command Griffiss Air Force Base, New York.

THE ROLE OF RELIABILITY MODELING IN THE DESIGN OF A FAULT-TOLERANT FEEDWATER CONTROL SYSTEM

Blake F. Putney, Jr., and Laurence A. Carmichael
Science Applications International Corporation

ABSTRACT

The use of reliability models in the design and development of a digital feedwater control system is described. Fault tree models, data bases, and their impact on system design processes are discussed.

INTRODUCTION

This paper describes the development and use of reliability models in the design process of a fault-tolerant digital feedwater control system for a Boiling Water Reactor (BWR). The reliability studies described here are part of a project sponsored by the Electric Power Research Institute, Northern States Power and SAIC. This project resulted from a feasibility study (Carmichael 1984) that showed that in addition to operational improvements, design of a more reliable feedwater control system would reduce the number of inadvertent plant trips.

The overall program objective is to design, develop and install a replacement for an existing analog feedwater control system with a new system that has increased reliability. This objective is to be accomplished with minimal changes to the existing plant design and operation. Although the reliabilty requirements of the Digital Feedwater Controller (DFC) were based on the controller hardware, the role of the reliability analysis for the project is to identify areas of the design where additional redundancy or fault detection can provide significant reliability improvements for the entire system. This role was accomplished through the analysis of a basic control system design to identify and rank system failure modes. These failure modes were used to support system design, including hardware, fault detection, and redundancy decisions. The final objective of the reliability analysis is to provide a high level of assurance that the control system will meet its design goal of less than two (2) plant trips in ten years.

The reliability analysis was made up of three elements, System Definition, Component Failure Definition, and Modeling. The following sections describe these elements and the results of the study.

SYSTEM DEFINITION

The purpose of the feedwater control system (see Figures 1 and 2) is to ensure that the proper inventory is maintained in the reactor vessel during all modes of operation of the reactor. The normal operating range for reactor vessel water level control is very narrow in a BWR. The capability must be provided in the level control system for programming of the water level setpoint according to reactor load. Programming the level in the reactor between 10% and 100% load serves to:

- Provide minimum carryover and carryunder throughout the range of plant conditions during normal operation.
- Provide an adequate margin between reactor water level and the top of the steam separators such that an increase in level resulting from recirculation pump trips will not cause the separators to be flooded.
- Provide an adequate margin between reactor water level and the bottom of the steam separators such that the decrease in level resulting from a reactor scram does not cau the separators to become uncovered.

Carryover is defined as the amount of water (percent total weight) carried over with the steam to the turbine. Steam leaving the reactor vessel should contain less than 0.1% carryover. Excessive carryover results in turbine damage due to water erosion of the turbine blades. Carryunder is defined as the amount of steam (weight percent) entrained in the liquid flowing into the vessel downcomer region. Excessive carryunder may result in recirculation pump cavitation and may also result in excessive fuel element temperatures due to decreased core inlet flow subcooling.

The feedwater control system consists of two 55% capacity electric motor driven feedpumps. At reactor loads higher than about 20 percent, control is provided by two valves, whose positions are automatically set by the operating feedwater controller. For operation at reactor load below 20 percent power, a low-flow control valve is placed in parallel with the main control valves, to provide precise flow control during startup and shutdown operations.

Correct water level in the reactor vessel can be maintained automatically or manually. When the feedwater system is manually controlled, remote manual controls enable the operator to vary the positions of the throttling valves directly. A manual/auto transfer station for each main control valve is located on the control room benchboard and set in the manual mode. When the manual/auto transfer station is set in the "Auto" mode, the operating level controller has control of the valve. A similar arrangement exists for the low flow valve.

There are two types of automatic level control available, "single-element" and "three-element" control. Single-element control uses only the

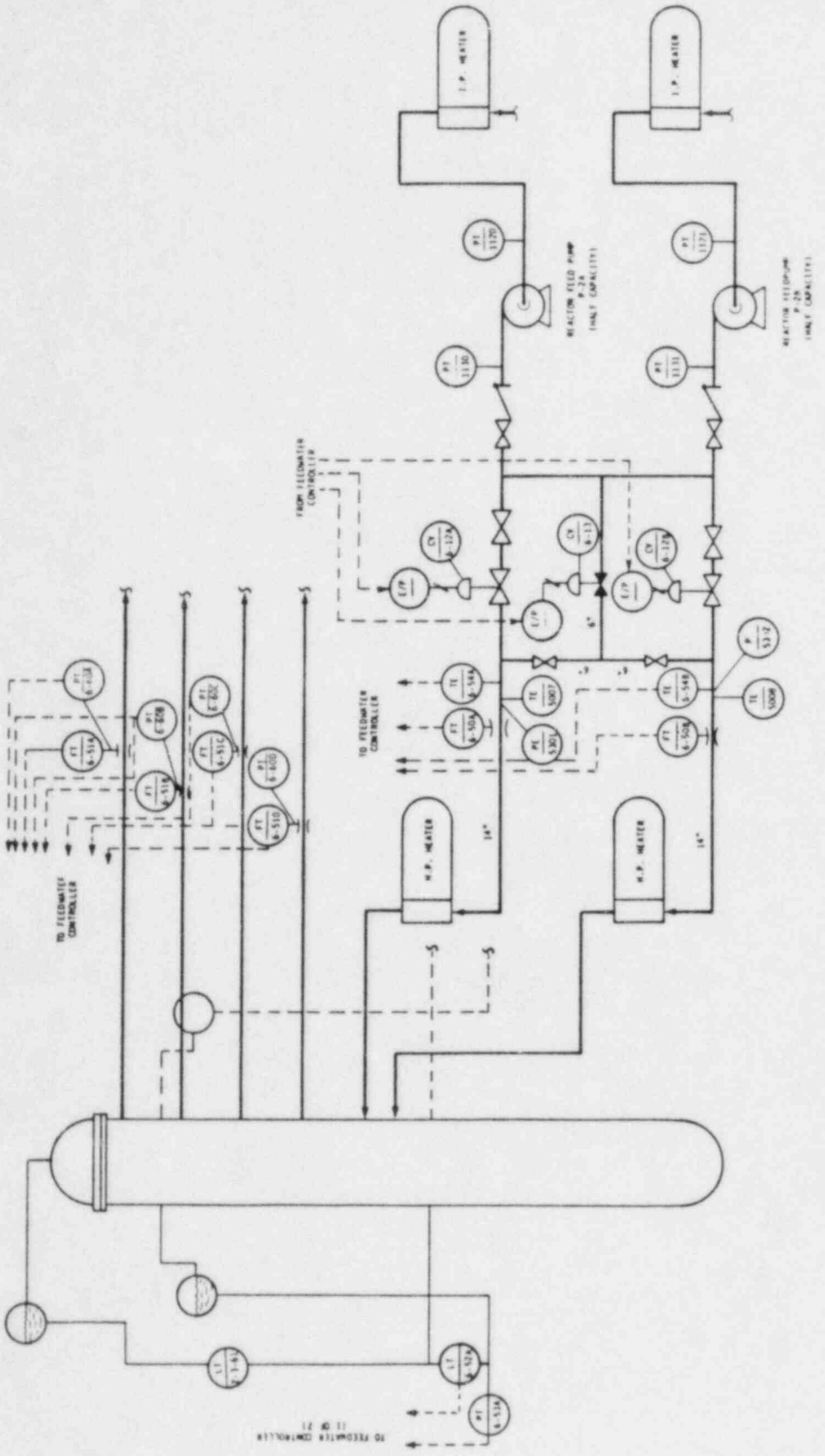

Figure 1 Monticello Feedwater Control System

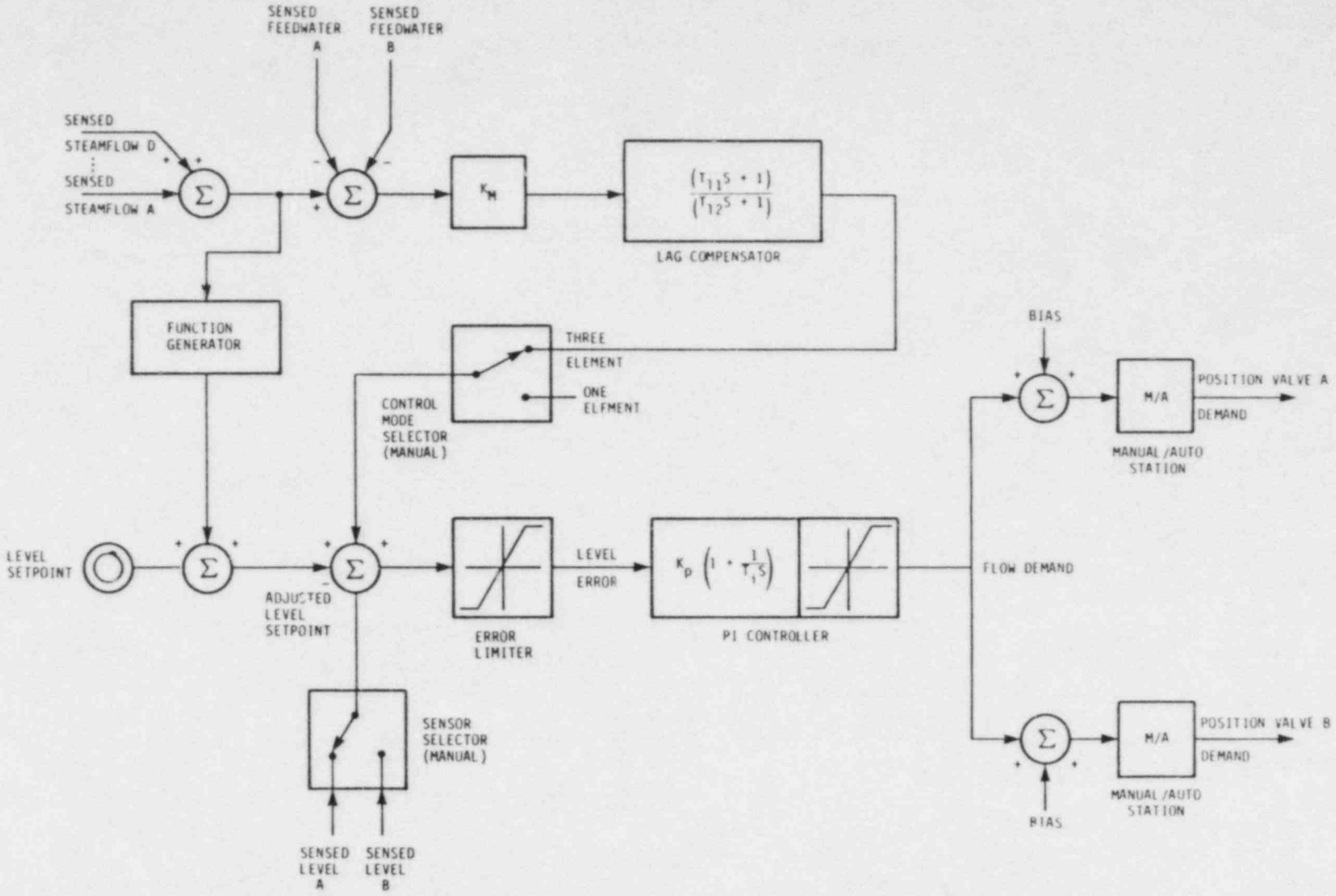

Figure 2 Monticello Analog Feedwater Controller Transfer Function

reactor water level signal as a control variable, while the three-element control uses the reactor level, steam flow and feed flow signals. The main feed water controller can operate in either mode. The startup or low-flow feedwater controller, which controls flow through the low-power throttling valve, is limited to single-element control only. This is because the steam and feedwater flow signals are at the low end of their scale at low-flow operating conditions.

The existing system contains multiple sensing elements for each sensed variable, with manual switchover in the event of a failure. However, the response time available to the operator for switchover is very short, and operator response to sensor failure is generally not quick enough to prevent a reactor trip. The replacement system, the DFC, was specified to include enough additional redundant sensor inputs to detect sensor failures reliably and in a timely manner.

One of the major design requirements for the DFC was that it look and act like the analog control system it was replacing, and that it require minimal plant modification to install. In addition to operational improvements, a baseline system design was developed using existing sensors, power supplies and control elements, with a new digital controller and a minimal set of fault detection algorithms. This system represents a minimal impact on the existing plant design and operations. Analysis of this system provided the basis for design decisions in determining potential modifications that would have significant impact on overall system reliability.

Once the design process is complete, a new baseline system will be defined and used for further study in order to assure that the reliability goals can be met. Figure 3 illustrates the reliability allocation for the new system, in its operating modes. This design results in a single-failure-proof system for single-element control for normal flow, with no individual failure modes that dominate system unreliability.

COMPONENT FAILURE DEFINITION

Once the baseline system design was completed, the components in the system were examined to determine their failure modes and their associated failure rates. Due to budget constraints, plant specific data gathering on system components was not possible. Therefore, potential data sources were examined to obtain failure rate information. Unfortunately, no single source of data contained data for all the system's components and their failure modes, and the existing data bases (IEEE 1983; Henley and Kumamito 1982) contained failure rate estimates that were significantly different for similar components and failure modes. The component failure data base was normalized through the use of a factor derived by setting similar components in each data base equal to each other and solving for a normalization factor to be used on component failures appearing only in Henley and Kumamito (1982). The resulting component data base was presented to the plant staff to assure that the failure rates were reasonable. This data base was then used as a basis for model solution.

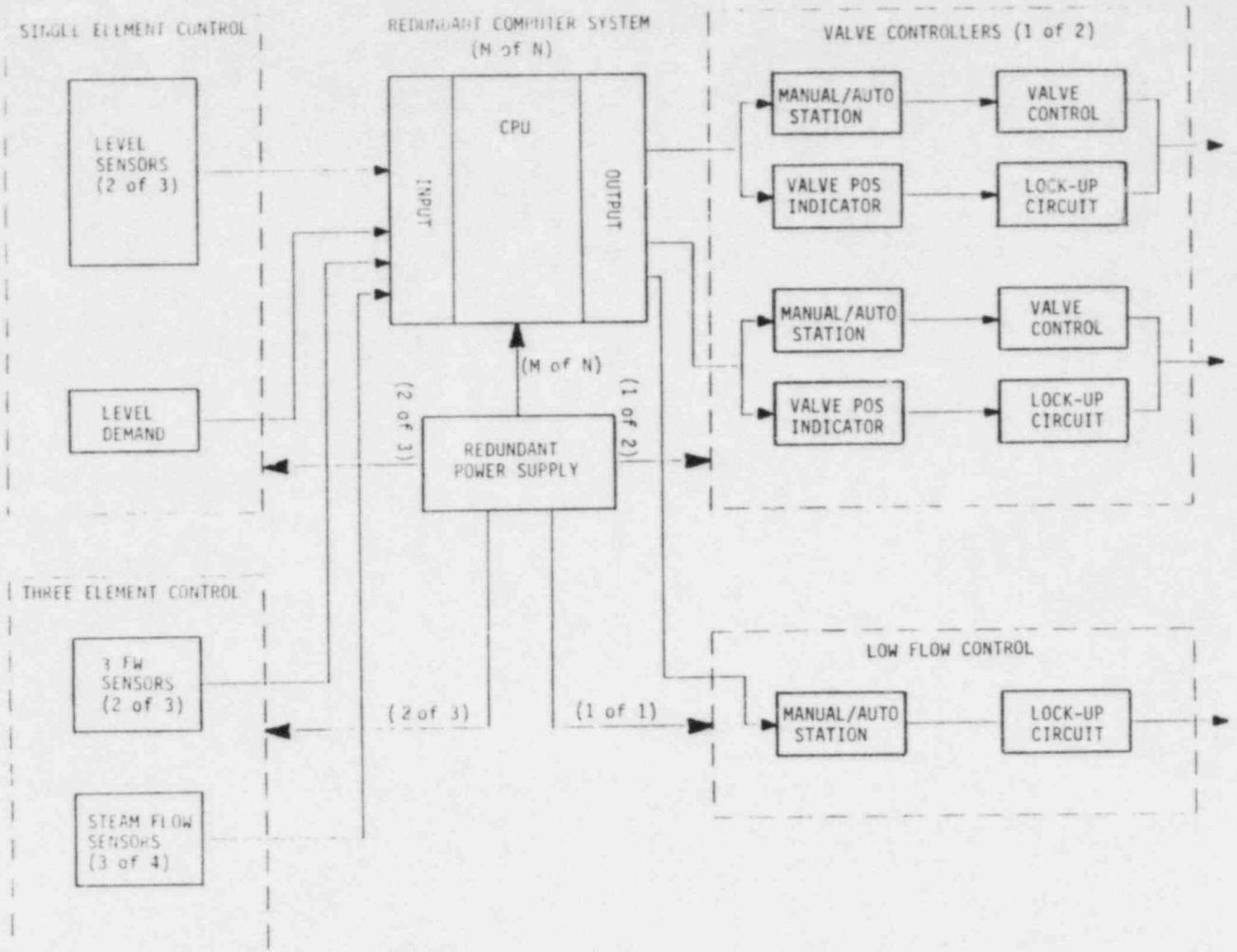

Figure 3 Reliability Allocation for Digital Feedwater Control System

MODELING

A number of techniques were considered for modeling system reliability for the DFC. Two existing computer codes, NASA's CARE III (available through Boeing Computer Services) and GRAMP (available from Systems Control Technology), were evaluated for use on this project. Both codes have extensive application in the modeling of fault tolerant systems. These codes at first appeared to be directly applicable to the DFC design. However, the codes assume instantaneous repair of a system using a finite number of on-line spares (i.e., a satellite computer system with 10 computers), while the DFC will have a mean time to repair of 24 hours, and a theoretically infinite supply of replacement components. Although models could be developed that would be able to use the codes, a simpler fault tree approach was chosen.

The fault tree approach that was used combines component failure modes and system states that result in loss of feedwater control. These failure modes were quantified and used as the basis for determining design changes that would improve system reliability. The most difficult modeling areas involved the fault detection and voting schemes used in the DFC, as well as modeling fault response and operator interfaces. System design allows the CPU and its input/output subsystems to be analyzed separately. Similar reliability models of these systems will be used in establishing requirements for the CPU system.

RESULTS

The cut sets for the system model have provided a direct means of defining and evaluating potential design changes that can significantly improve the reliability of the system. These changes have included reconfiguration of the system power supplies, additional signal redundancy, additional positional feedback for control actuators, new fault detection schemes, and operator interfaces that are less likely to cause system failures.

REFERENCES

Carmichael, L. A. (1984), Digital Feedwater Controller for a BWR: A Conceptual Design Study, EPRI-NP3323, Interim Report.

Henly, J., and H. Kumamito (1982), Reliability Engineering and Risk Assessment, Prentice-Hall, Englewood Cliffs, N. J.

IEEE (1983), IEEE-Std 500-1984, IEEE Guide to the Collection and Presentation of Electrical Electronic Sensing Component and Mechanical Equipment Reliability Data for Nuclear Power Generation Station.

STATISTICAL UNCERTAINTIES AND UNRECOGNIZED RELATIONSHIPS

John P. Rankin
Boeing Services International

ABSTRACT

This paper deals with "missing links" in evaluations of system reliability. It reflects the author's experience with various types of product assurance analyses, including those applied to nuclear energy systems. Examples are given of subtle system-specific design relationships that were identified by some of the author's unique analyses. These same design relationships had remained unrecognized by more traditional methods commonly employed in statistical reliability evaluations.

It is felt that these "hidden" relationships in specific designs directly contribute to inaccuracies in reliability assessments. Uncertainty factors at the system level may sometimes be applied in attempts to compensate for the impact of such unrecognized relationships. Often "uncertainty bands" are used to relegate unknowns to a "miscellaneous" category of low-probability occurrences. However, experience and modern analytical methods indicate that perhaps the dominant, most probable and significant events are sometimes overlooked in statistical reliability assurances. We tend to call it "Murphy," but Murphy has a face and can be identified.

The utility of two unique methods of identifying the otherwise often unforeseeable system interdependencies for statistical evaluations will be discussed. These methods are sneak circuit analysis and a checklist form of common cause failure analysis. Unless these techniques (or a suitable equivalent) are also employed along with the more widely-known assurance tools, high reliability of complex systems may not be adequately assured. This concern will be indicated by specific illustrations. The reader can then arrive at his own conclusions of statistical coverage for the impact of unrecognized relationships in traditional approaches.

INTRODUCTION

Criticisms of numerical assessments of high-reliability design typically focus on statistical uncertainties in the source data. This may be because

of the relative unbelieveability of the ultimate system reliability prediction plus the associated indefensibility of the source data as opposed to the concrete mechanics of the technology of the numeric manipulations. However, questions regarding the technology of reliability calculations or accuracy of recording or interpreting failure rates relative to specific usage environments are not the issue, per se, in this paper. Rather, the intent is to show that often unrecognized relationships between the design elements may be a significant, but unmeasured, factor in degrading the validity of the final system reliability prediction. The paper is written from a management perspective, without numbers, rather than attempting to guide technologists in refinements of numerical manipulations or comparisons. The paper is also not a tutorial on how to perform sneak circuit analysis or common cause failure analysis. These subjects have been covered in the referenced papers.

Experience of many a project manager or operator has shown that what was not included in reliability predictions became far more important than what was included. Perhaps this viewpoint arises from the fact that the things which were covered perform very well, but it's the "accidents" which happen (ovesights, perhaps?) that stick in the memory of management. While they have no numbers for comparison, many project managers continue to distrust numeric assurances because everybody knows of things that go wrong when the "experts" say it won't or can't happen.

My premise is that many such situations are rooted in previously unrecognized system element interrelationships--relationships that were not identified to the numerical assessment experts or others for inclusion into the evaluations. This leads to broad "uncertainty bands" in attempts to cover unanticipated events in the model and to make theory or empirical extrapolations agree with field experience memories. Uncertainty bands, however, are seen by some as an admission of error or guesswork. They may actually tend to weaken credibility. In all probability, use of uncertainty factors may not be the best approach. The numbers should simply state what is known. Management should then apply their additional judgement and/or derating factors based upon system modeling familiarity or the confidence in attempts to analytically ascertain all system interrelationships. The final credibility level of reliability prognostications can thereby be tailored to each project. Public acceptance of judgement calls will, of course, always be subject to hazard and phobias, but lessening the surprises in operations will eventually help.

Now let's examine a few examples of some surprising system design relationships. These surprises were identified analytically in mature systems, long after the traditional reliability assessments had been completed and operations initiated.

OVERSIGHT IN FUNCTIONAL CRITICALITY BY DEFINITION

An illustration of limitations in the application of traditional assurance technology is provided in Figure 1. This circuitry was analyzed as part of a topological sneak circuit demonstration effort described by Rankin (1972) on a shutdown system for the Savannah River (nuclear) Plant in South Carolina.

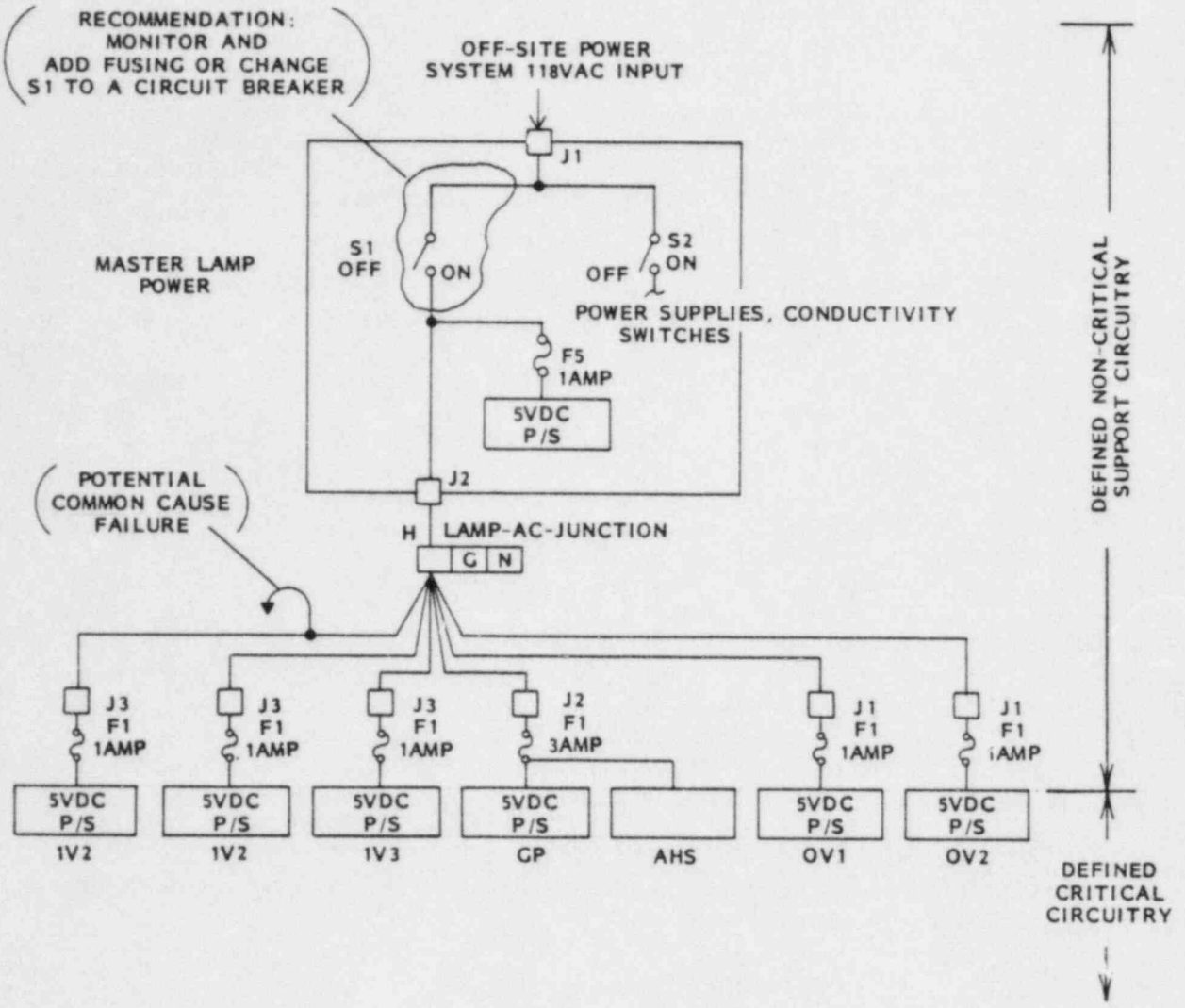

Figure 1. Potential Common Cause Failure Point, Savannah River Plant Shutdown System.

The system had been operational for several years prior to the sneak circuit analysis. The scope of the system to be analyzed was specified as beginning with the power supplies shown as boxes at the bottom of Figure 1. During the topological analysis, a request was made for data depicting the inputs to the power supplies, even though this information was not considered by the client to be in scope. The circuitry shown in Figure 1 was considered by the client to be part of a "non-critical" support system per existing plant definitions. Yet, the circuitry (not shown) below the boxes was defined by the client as critical; that is, its proper operation was essential to plant safety.

This philosophy of criticality had been used by the client in past applications of traditional nontopological assurance analyses. The topological approach recognized no valid technical reason for a defined criticality boundary to terminate circuit continuity, so it continued into the "non-critical" area shown in Figure 1.

An auxiliary finding of the sneak circuit analysis resulted from corollary common cause failure checks made with regard to the possible effects of shorted nodes. In this case, shorting the node at pin H of the LAMP-AC-JUNCTION disables all of the power supplies driving the "critical" system (not shown) below the boxes. The fault would also be propagated to the system 118 VAC input and negate any annunciation of the condition. The system had been intended to be extremely reliable and to annunciate any adverse operating condition in the plant as it took action toward safety. However, the analysis showed that the system itself could be insidiously disabled for long periods of time before routine maintenance checks would be expected to discover its condition. Meanwhile, the system may not be available to assure safety of the plant.

This potential failure and its consequence had not been considered during the analyses performed by the design organization simply because it is predicated upon a specific, detailed failure in areas designated as "non-critical." The earlier analyses were not performed to include non-critical areas. The fault did not lie in the technology of the earlier analyses. The problem was a scoping of efforts according to arbitrary system functional criticality definitions. Engineering efforts were apparently bounded by these definitions without adequate regard to detailed circuit implementation features.

PERCEPTION PROBLEMS DUE TO NATURE OF DIAGRAMS

Perhaps one of the largest root causes of oversight in analyzing implementation of intended system design relationships is encountered in the layout of engineering drawings. This is especially true for electrical/electronics control systems. The drawings found in industry typically do not clearly display the electrical network topology. Yet, all of our college texts teach electrical engineers to analyze networks topologically. Industrial schematics are more concerned with connections and wire routing, and they therefore obscure the topology.

Sneak circuit analysis is predicated upon topological display of the network trees, as described by Rankin (1973, 1981-2). Such a display greatly enhances the ability to perceive the possible electrical current flow paths under all conditions. Pattern recognition clues even further assure that previously-obscured paths will be identified in the design implementation connections. By these means, literally thousands of sneak circuits have been found after they had been latently designed into systems such that they either had escaped, or would be likely to escape, detection in testing and initial operations. They also were typically not recognized when other analyses (such as fault tree and failure modes and effects) were performed, due to the obscured topology. A couple of examples will illustrate the benefits of topological presentation.

The circuit of Figure 2 is a topological representation of rod control and setback switching. It was generated during sneak circuit analysis of the N reactor at Hanford, Washington, as reported by Rankin, et al. (1974). Arrows have been drawn in the figure to indicate possible sneak flow paths which would activate the ROD OUT SLOW relay (3K41), possibly at a time when emergency reactor shutdown (all rods in) had been demanded by the automatic actions of opening interlock contacts 2K35 and 1K4. In this case, the manual switch 1H41 would have to be selected to its 2, 10, 14, and 12 contact-make position at the time of the scram (emergency shutdown demand). Such an inadvertent override of the scram demand would cause affected rods to actually back out or stay out when intended to insert for shutdown.

Because it was obscured in industrial schematics of the plant circuitry, the possibilities for this event had never before been recognized. Neither had the particular combination of events and operating configurations leading to the possible scram override ever happened, even though the N reactor had already been in service for many years at the time of the sneak circuit analysis which found the unexpected design relationship. Prior to the sneak circuit analysis, it is likely that any assessment of the probability of rods backing out or failing to insert during a scram demand would not have realistically accounted for this unrecognized and unexperienced possibility. Yet, it apparently was quite reasonably possible.

Another example from the same sneak circuit analysis of the N reactor illustrates latent failure effects. Figure 3 topologically represents control rod travel command circuitry. For an operational scenario with circuit breaker CB108 failed open, the arrows are drawn to indicate that sneak current can flow throughout the circuit. As shown, even with the open breaker, full voltage and current can accivate the ROD OUT SLOW relays (3V5R and 403K5), while equally activating the ROD IN SLOW relays (3V5F and 203K5). The industrial schematics in use prior to the sneak circuit analysis obscured this possibility, making it appear that the ROD OUT function could not be activated with circuit breaker CB108 open. While it could happen, operating experience had never established the conditions to make it occur. Neither had the possibility been discovered in other analyses. Again, assessment of the probability of such an event (based upon an unrecognized design capability) would likely miss the mark without the topological analysis.

N Reactor Rod Control and Setback Switching Circuit
PWR CX29 120V AC

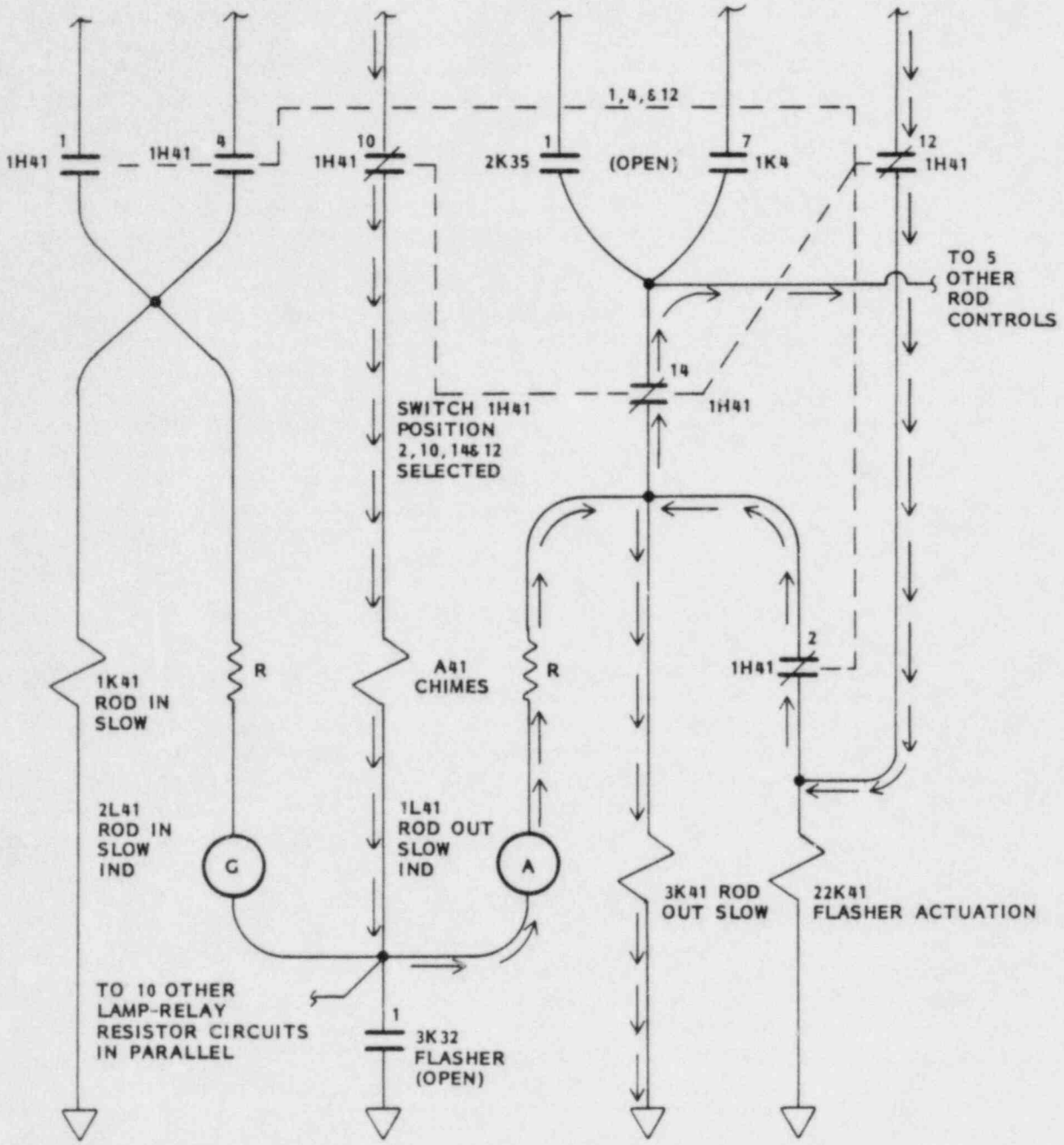

Figure 2. Sneak Path to Rod Out Relay 3K41 (et al.) with Open Scram Relays 2K35 and 1K4.

N Reactor Control Rod Travel Command Circuit

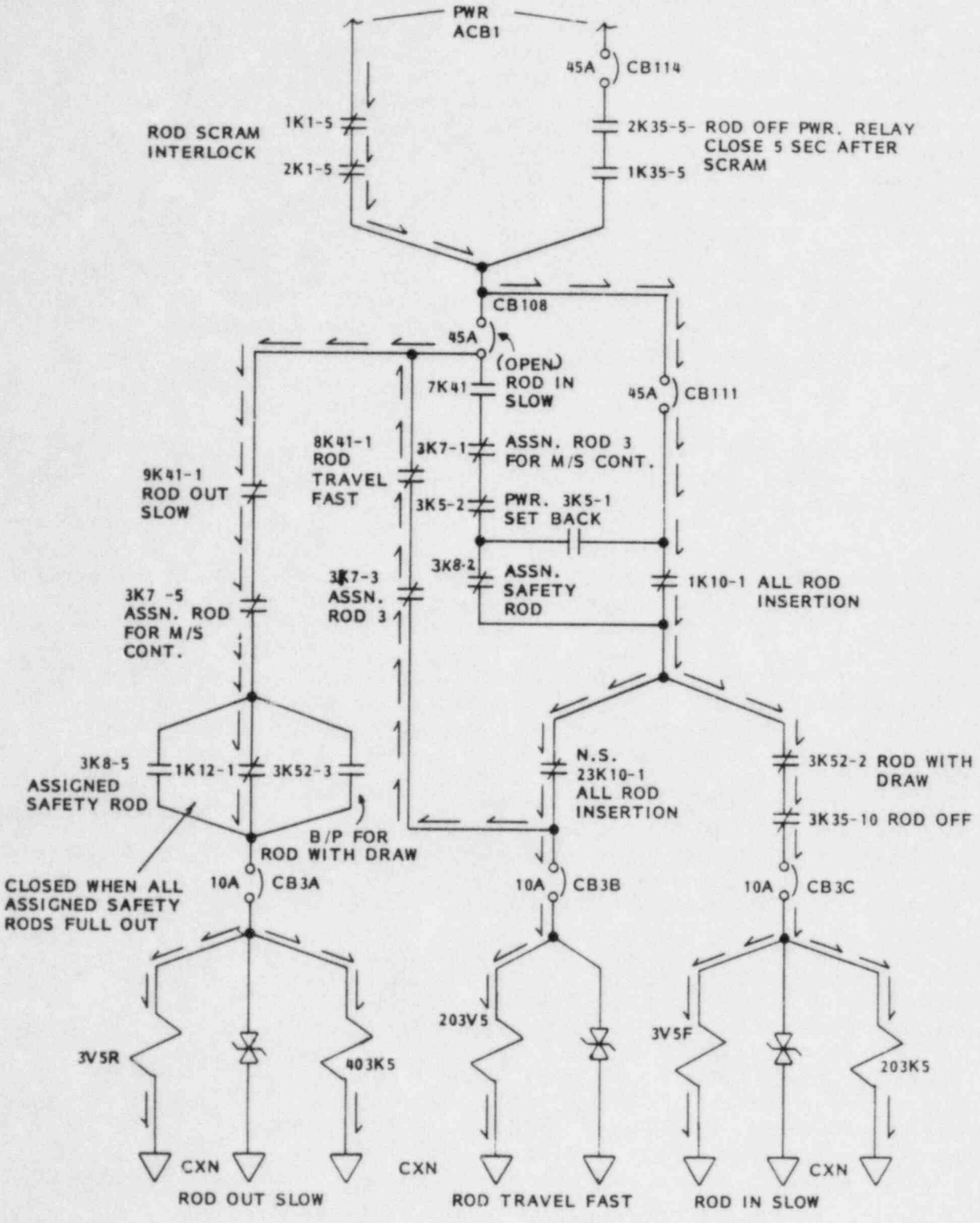

Figure 3. Conflicting Commands Sneak Path.

HIDDEN HUMAN FACTORS

While the system design may have latent relationships between its components, it may also aggravate human factors evaluations. Consider the situation described in N Reactor Sneak Circuit Report N-10, R1, pages 1 and 2, provided as Figures 4-a and 4-b. In this case, the plant operators can easily be misled to false conclusions about system status. This could occur even in the absence of component failures because of the unusual relationships between indicator labels and design detail implementation. As before, the functional relationships were obscured in the industrial schematics, but the topological analysis approach brought them to light. Would a human factors study have adequately allowed for the possible resultant (erroneous) operator actions due to misleading design features? We don't need to design ambiguous systems that can confuse operators. However, unless the sneak circuit analysis is done, how will we know? By the way, topological sneak circuit analysis apparently is still relatively unknown in the power industry in the United States.

OVERWHELMING COMPLEXITY OF DETAIL

Another example from N reactor portrays an additional aspect of modern assurance analysis that is effective at identifying previously unrecognized design relationships. Susceptibility of the scram system to common cause, cascaded, or latent coexistent specific component failures was discovered by an early outgrowth of the sneak circuit analysis technique. The outgrowth approach subsequently was refined into a simple but highly effective method of detailed common cause failure analysis by checklists, as described by Rankin (1980, 1981-b, 1982).

Even in the topological representation of Figure 5, the N reactor ROD SCRAM circuit is complicated. However, because of this arrangement of the circuit diagram supplemented with analytical clues in a checklist, the susceptibility of scram system override was found. The problem arises when pairs of blocking diodes fail in association with a control rod that is WITHDRAWN and/or OFF for maintenance purposes during reactor operations. The arrows in Figure 5 show the resultant sneak path that renders ineffective the isolation of the electrical OPERATIONS BUS from the electrical MAINTENANCE BUS. By this path through the failed diodes, which are undetectable failure points, rod scram would be inhibited on all rods.

The credibility of failure of the diode "quads" -- the four blocking diodes associated with each rod are encapulated in a single, inaccessible module called a "quad" -- was described by Gallagher (1971). That is, it actually happened before the analysis was performed. Likewise, it is anticipated that it certainly will happen again. At the time of the already experienced failure, with a rod withdrawn for maintenance and a scram initiated, the rods remained out. Reactor shutdown was achieved by "poison balls" from a separate system. (N reactor is believed to be the only U.S. production reactor to have the poison ball backup scram system.) After the event, modifications were implemented in the circuit design (resulting in the

SNEAK CIRCUIT REPORT N-10, R1 Page 1 of 2

DATE 7-3-74

ENGINEER W. S. Brown

TITLE Command Function Used to Indicate Performance

REFERENCES

1. Drawing H-1-32065, Rev. 6, "Electrical Elementrary Diagram, Emergency Cooling Water System, Index 33".

2. Drawing H-1-32070, Rev. 8, Same title.

MODULE/EQUIPMENT

Control Room/Indicating Lamp

EXPLANATION

Please refer to Figure 1, extracted from references 1 and 2. Relay 66K33 is operated when reactor cooling water declines to a pressure below 200 PSI and a temperature of less than 250°F. (1A). Contacts of this relay are used to permit V23 valves 10V33 and 11V33 in Emergency Cooling Water (ECW) System 1 to operate to the "Low Pressure Diversion" position (1B). 134K33 is used for ECW System 2. Another contact of 66K33 is used to energize indicator lamps whose titles are "LOW PRESSURE DIVERSION OPERATED" (1C). 66K33 contacts in the valve circuits, however, only permit operation of these controls if the circuit breakers are closed, an enabling switch is operated, and the High Pressure Diversion limit switch is satisfied. It is possible, therefore, for relay 66K33 to be energized, providing activation of lamps 35L33 & 50N33-L, without actual occurrence of low pressure diversion due to lockouts of the breakers, the enabling switch, or the High Pressure Diversion limit switch.

Inasmuch as control and performance are energized from different breakers, and indication is from a different source entirely, another sneaky aspect exists. With CD18 open, Low Pressure Diversion may be enabled and indicated without operation. If CJ2 is opened, operation may be enabled without indication. The division of operation at 120-125 volts and indication at 24 volts is general in the system.

POTENTIAL IMPACT:

False indication of operation.

RECOMMENDATION

1. Provide additional normally open position switch contacts for Low Pressure Diversion valves 10V33 A & B and 11V33 A & B, connect these in series with lamps 35L33 and 50N33-L.

2. Change the title on 35L33 and 50N33-1 to "LOW PRESSURE DIVERSION PERMITTED".

Figure 4-a. N Reactor Sneak Circuit Report N-10, R1, Page 1.

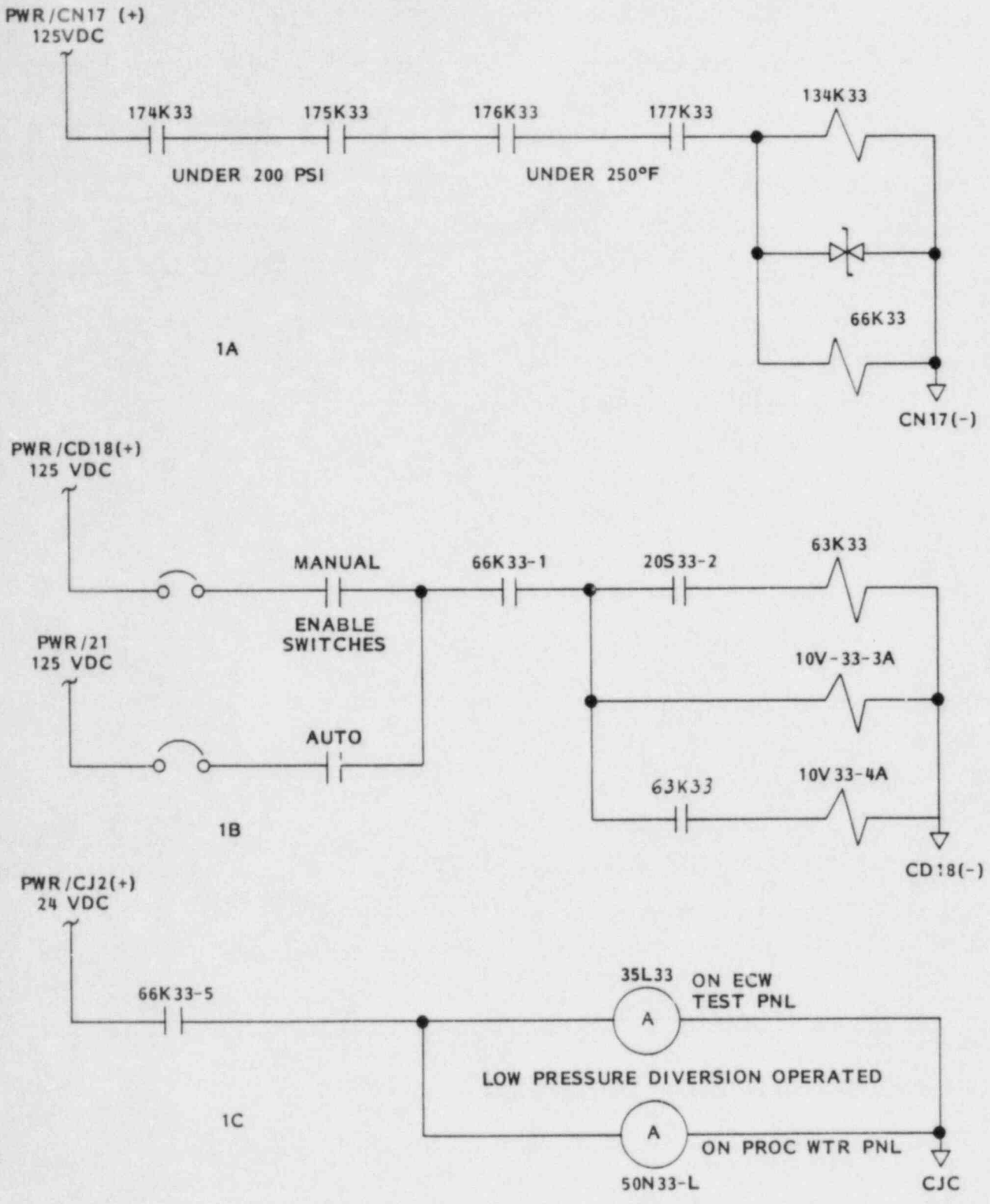

Figure 4-b. Figure 1 of N Reactor Sneak Circuit Report N-10, R1; Page 2 of 2

N Reactor Scram Circuit

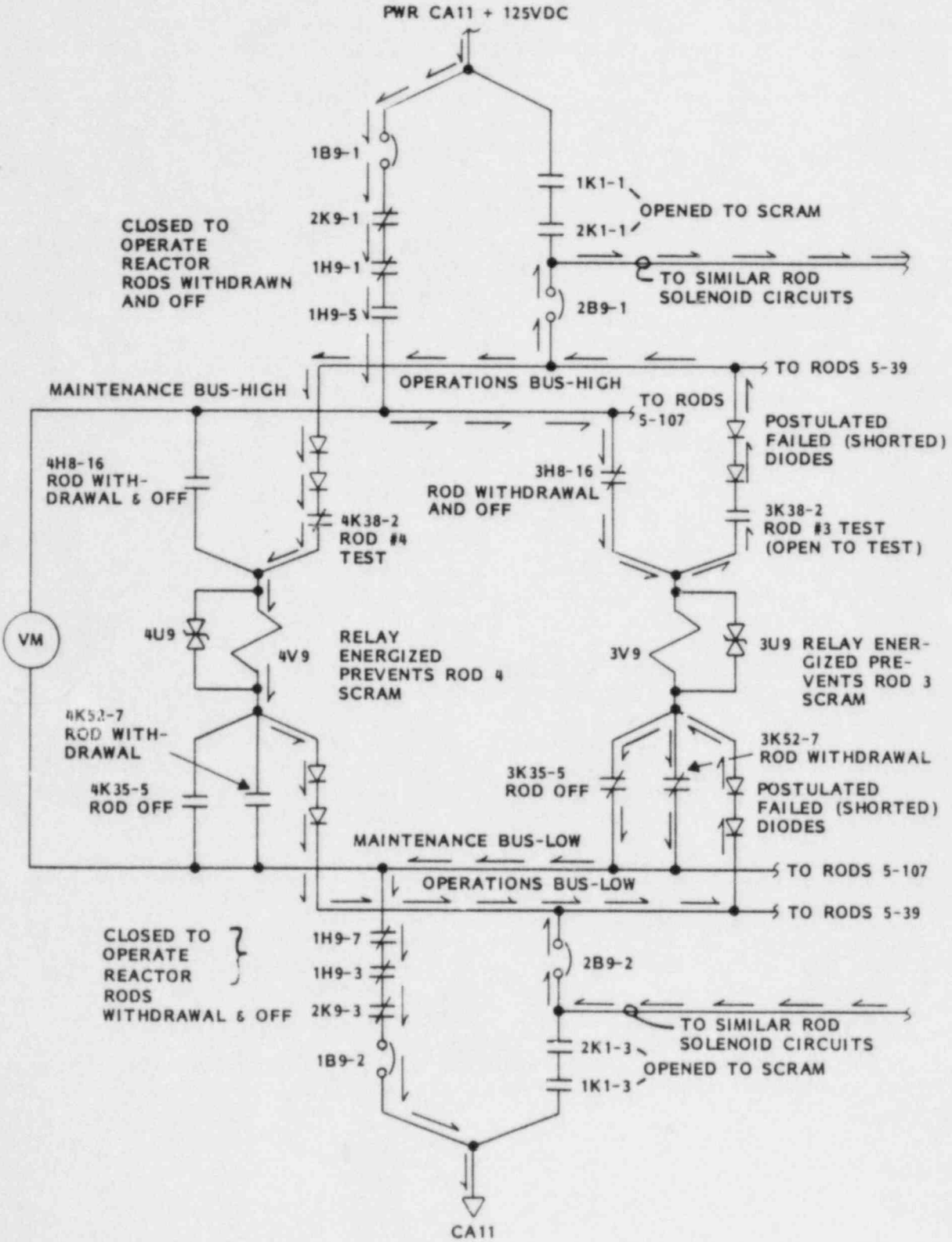

Figure 5. Sneak Path in Rod Scram Solenoid Circuit Due To Postulated Failed Diode Quads (Common Cause Failures).

circuit of Figure 5) by which it was believed that the susceptibility to rod scram inhibit (even with failed diodes) was completely removed. After incorporation of the new circuit design of Figure 5 and based upon available industrial drawings, the probability of such an occurrence was supposed to be zero or nearly so. However, the findings of the topological analysis done 4 years later and depicted in Figure 5 indicate that susceptibility to the unplanned event was still designed into the system. In fact, the probability of occurrence was ever-increasing with time, due to diode quad aging and other effects as discussed with project management and engineering personnel upon completion of the topological analysis.

CONCLUSIONS

While the illustrations provided may appear unique or highly selected, they are not. Actually, the author himself has been involved in the application of the topological analysis techniques to about 100 projects since 1969. From these efforts, many thousands of such hidden design features were identified. Quite a number of them involved sneak paths and common cause failures of greater subtlety and more significant potential impact on their projects than those presented in the illustrations of this paper. Most of these analyses were performed on projects that had already been scrutinized with fault tree and/or failure modes and effects analysis techniques. It is logical to conclude therefore, that the topological techniques should also be applied in order to enhance assurance that all implemented design features are understood in models for evaluations of risk.

It appears that often the specific design hookups and implementations of a system provide not only what was functionally intended, but also a few things that were not intended. Unfortunately, these unintended functional capabilities and some subtle common cause failure points are prone to lie dormant through routine testing and operations. They seem to await the most inopportune moments for activation or failure occurrence. Because they are unintended and/or unrecognized, they are not adequately treated in the models used for probability evaluations. This omission leaves gaps in the credibility of such calculations; and when the unplanned events or failures finally occur, the probability model may be discredited, according to the significance of the incident upon operations and safety.

Politically, coverage by uncertainty bands is seen as inappropriate because "the experts" shouldn't admit that they don't know something about their area of expertise. (That is culturally unacceptable.) Besides, the unplanned events may be more probable to occur than an "acceptable" uncertainty band would allow -- or even more probable than what is selected as a "dominant sequence." It is the author's belief that probabilistic risk assessments will likely continue to experience low levels of public trust -- not so much due to lack of valid failure rate data, but because of incomplete or inaccurate treatments with regard to unrecognized relationships in the models unless topological analyses are also employed.

REFERENCES

Gallagher, G. R., (November-December 1971), "Failure of N Reactor Primary Scram System," Nuclear Safety, Volume 12, No. 6, pp. 608-614.

Rankin, J. P., et al., (December 1972), "Sneak Circuit Analysis of the AEC - DuPont Savannah River Plant Automatic Incident Action System," The Boeing Company, Houston, Texas 77258. (Performed under agreements with the AEC Division of Operational Safety and DuPont.)

Rankin, J. P.,(September-October 1973), "Sneak-Circuit Analysis," Nuclear Safety, Volume 14, No. 5, pp. 461-468.

Rankin, J. P., et al., (July 31, 1974), "Sneak Circuit Analysis of N Reactor," D2-118542-1, The Boeing Company, Houston, Texas 77258. (Available from Government Technical Information Center as document RLO-75-1).

Rankin, J. P., (April 1980), "Common Cause Failure Analysis -- Why Interlocked Redundant Systems Fail," Turbine Powered Executive Aircraft Meeting, Phoenix, Arizona, Society of Automotive Engineers, Inc., Warrendale, Pennsylvania 15096. (SAE Paper 800631, contained in Volume 89 of the 1980 SAE Transactions.)

Rankin, J. P., (April 1981-a), "Common Cause Failure Analysis of Instrumentation and Control Systems," Reliability Conference for the Electrical Power Industry, Portland, Oregon.

Rankin, J. P., (October 21-23, 1981-b), "Identification of Common Cause Failures in Instrumentation and Control Systems," IEEE 1981 Symposium on Nuclear Power Systems, San Francisco, California.

Rankin, J. P., (January 26-28, 1982), "Common Cause Hazard Analysis for Random Glitches," Reliability and Maintainability Symposium, Los Angeles, California.

Special Topical Session

Spatial Statistics

UNCERTAINTY AND SENSITIVITY ANALYSIS OF ENVIRONMENTAL TRANSPORT MODELS

Timothy S. Margulies and Leslie E. Lancaster
U.S. Nuclear Regulatory Commission, Washington, D.C. 20555

ABSTRACT

An uncertainty and sensitivity analysis has been made of the CRAC-2 (Calculations of Reactor Accident Consequences) atmospheric transport and deposition models. Robustness and uncertainty aspects of air and ground deposited material and the relative contribution of input and model parameters were systematically studied. The underlying data structures were investigated using a multiway layout of factors over specified ranges generated via a Latin hypercube sampling scheme. The variables selected in our analysis include: weather bin, dry deposition velocity, rain washout coefficient/rain intensity, duration of release, heat content, sigma-z (vertical) plume dispersion parameter, sigma-y (crosswind) plume dispersion parameter, and mixing height. To determine the contributors to the output variability (versus distance from the site) step-wise regression analyses were performed on transformations of the spatial concentration patterns simulated.

INTRODUCTION

An important part of performing a radiological risk assessment for an electric generating facility is quantifying the uncertainties - those associated with the probabilities of accident scenarios and those associated with the (off-site) consequence estimates. This paper describes the process of performing an integrated uncertainty and sensitivity analysis of environmental transport and deposition models in the atmosphere. The study objectives are (1) to measure the overall uncertainty of the output variables of interest (2) to rank and quantify the inputs (or model parameters) according to their contributions to output uncertainty and (3) to compare several uncertainty approaches and ways of displaying results (e.g., stochastic and ignorance uncertainties). The analysis will also indicate modeling and research needs.

STATISTICAL APPROACH

A variety of techniques are available to quantify the uncertainty in complex models for assessing radiological impact upon man and the environment that may include nonlinearities and time-varying phenomena[1]. These include - the Monte Carlo[2], fractional factorial design[3], Latin hypercube sampling[4-6], response surface[7-8] and differential sensitivity analysis (e.g., adjoint[9,10]) methodologies. A preferred technical approach would be flexible, economical to use, easy to implement, provide a capability to estimate an output distribution function and rank input variables by different criteria.

In this study the Latin hypercube sampling (LHS) scheme was chosen to be implemented on an environmental transport model in the reactor accident consequence code (called CRAC-2). The advantages and properties of the Latin hypercube sampling techniques are:

- The full range of each input variable is sampled and correlation coefficients between all pair-wise input variables can be specified.

- It provides unbiased estimates of cummulative distribution functions and means for model output under moderate assumptions.

The LHS method is a member of the class of sampling techniques which include Monte Carlo and stratified random sampling. Several risk assessments for nuclear waste repositories[11] have applied LHS techniques. Furthermore, LHS has recently been applied to a multicomponent aerosol physics code[12] and represents a recommended approach for sensitivity and uncertainty analysis of the new integrated risk code (called MELCOR) for severe reactor accident calculations (in-plant and off-site) being developed by the Nuclear Regulatory Commission. It is noted that the weather time series data sampling scheme in CRAC-2 is essentially a random sampling scheme without replacement from a set of weather bins (i.e., a Latin hypercube sampling method). However, many other variables involved in the atmospheric dispersion and deposition models are fixed at a single, best estimate value. A more detailed description of the atmospheric dispersion model and weather times series data sampling method in CRAC-2, will be presented in a following section. First, however we will provide an overview of the steps taken in statistical uncertainty and sensitivity approach.

We remark that one may wish to distinguish between different types of uncertainty associated with modelling of physico-chemical processes, in particular:

1. The statistical uncertainty due to inherent random nature of the processes, and

2. The state (perhaps "lack-of") knowledge uncertainty.

This latter state of knowledge uncertainty may be further subdivided into model and parameter uncertainty. The parameter uncertainty is due to insufficient knowledge about what the input to the encoded model should be. This study documented herein deals with parameter uncertainty. The modeling uncertainty is due to simplifying assumptions and the fact that the models used may not accurately model the true physical process.

The process of conducting an uncertainty and sensitivity analysis starts with an identification of a set of key parameters in the model understudy. For each chosen variable, a set of quantitative information is developed regarding the range of variation, probability distribution, as well as, correlations among the variables. In our study information was primarily based upon expert opinion. Secondly, this information is used as input to the Latin hypercube sampling code[13,14]. LHS is used to generate what is called a design matrix. Specifically, if N computer runs are to be made of the computer code to be analyzed (e.g., the CRAC-2 dispersion model) with k parameters under study, the design matrix has dimensions N x k. Each row of this matrix contains the input valuations of the each of the chosen k parameters.

The next step in the process involves performing a sensitivity analysis on the calculated results of ground (or air) concentration versus distance from the power plant, for example. The aim is to determine and quantify the relative contributions of the k^{th} variable toward the output variability versus distance. This may be achieved by performing step-wise regression analyses on the concentration patterns simulated. Alternatively, one may perform a regression on the ranks of the data, replacing the "raw" data values by their ranks. This may be preferred when highly nonlinear relationships are present between the model outputs and inputs. Both graphical analyses and statistical distribution fitting procedures may also be extremely useful in identifying patterns in the data.

WEATHER DATA SAMPLING METHOD

The transport and dispersion of material associated with a postulated accidental release of radioactivity at a nuclear power plant depends upon the weather conditions at the start of the release from containment through a period of tens of hundreds of hours. The CRAC (Calculations of Reactor Accident Consequences) model which was developed during 1975 Reactor Safety Study (WASH-1400)[15] and subsequently modified (and called CRAC-2)[16] to estimate the offsite health and economic portion of reactor risk to society has several weather data sampling capabilities. The consequence analysis is typically given in terms of a CCDF (cummulative complementary distribution function) representing the range of weather scenarios. The weather data is usually taken from annual records of onsite meteorological tower measurements or nearby National Weather Stations. The data required for CRAC and CRAC-2 consists of hourly averages of windspeed, atmospheric stability class and precipitation. The recommended "importance" sampling method in CRAC-2 was

designed to reduce the variability due to either 1) random 2) stratified random sampling or stratified sampling techniques[17]. For example, risk curves generated with CRAC used a stratified sampling technique that selects data every 4 days ± 13 hours in an attempt to account for diurnal, seasonal and 4-day weather cycles. That is, CRAC for WASH-1400 selected 91 weather sequences to represent the 8760 hours of annual data. Such a sampling scheme is sensitive to the chosen start time and can miss infrequent meteorological conditions (including rain and windspeed slowdowns) with possible corresponding high dose-rate consequences.

In CRAC-2, the annual weather data is sorted into "weather bins" before performing any consequence calculations. The weather bins are defined in Table 1. For example, the first bins refer to whether it begins to rain at a certain distance from the reactor or that the windspeed drops. If neither of these conditions occur the weather sequence is categorized by the stability and windspeed at the start of the accident. Also, a probability is assigned to each bin according to its frequency of occurrence (directional information is available for the nonrain cases only). Four random samples with equally spaced probability are typically chosen from each weather bin to perform the atmospheric transport and deposition calculations in CRAC-2.

ATMOSPHERIC DISPERSION MODEL

Both CRAC and CRAC-2 use similar atmospheric transport and dispersion models - a standard Gaussian-plume formulation[18-20]. Currently, the Gaussian plume atmospheric dispersion model is employed in most consequence-modeling codes for the following reasons: (1) economy in terms of computer time and (2) general lack of availability of the meterological parameters necessary for input to more complicated models. It is recognized that the simple Gaussian model is not appropriate for complex terrain (e.g., mountain-valley) or land (sea) breeze flows. Assuming the material is reflected at the ground, the ground-level, time-integrated concentration for a source of strength Q is given by

$$\chi(x,y,0) = \frac{Q}{\pi\,\sigma_y(x)\,\sigma_z(x)\,u}\exp\left(\frac{-y^2}{2\,\sigma_y^2(x)} + \frac{-h^2}{2\,\sigma_z^2(x)}\right) \quad (1)$$

where $\sigma_y(x)$ and $\sigma_z(x)$, the standard deviations of the crosswind and vertical distributions, respectively, are functions of the downwind distance, x. u is the mean wind speed, and h is the source release height. If Q is in curies, the units of χ are Ci-sec/m^3. In CRAC 2 and CRAC, equation (1) is simplified by replacing the Gaussian crosswind profile (y direction) with a rectangular (or "top hat") function of width 3 σ_y; i.e., in equation (1) the term

Table 1: Weather Importance Bins Applied to One Year of New York City Meteorlogical Data

Weather Bin Definitions

R - Rain starting within indicated interval (miles).
S - Slowdown occurring within indicated interval (miles).
A-C D E F - Stability categories.

1(0-1), 2(1-2), 2(2-3), 4(3-5), 5(GT 5) - Wind Speed intervals (m/s).

Classification (C)	Weather Bin	Number of Sequences	Percent
C1	1 R (0)	697	7.96
	2 R (0-5)	12	.14
	3 R (5-10)	62	.71
	4 R (10-15)	102	1.16
	5 R (15-20)	75	.86
	6 R (20-25)	67	.76
	7 R (25-30)	61	.70
C2	8 S (0-10)	24	.27
	9 S (10-15)	16	.18
	10 S (15-20)	18	.21
	11 S (20-25)	14	.16
	12 S (25-30)	18	.21
C3	13 C 3	168	1.92
	14 C 4	892	10.18
C4	15 D 1	0	0.00
	16 D 2	61	.70
	17 D 3	226	2.58
	18 D 4	948	10.82
	19 D 5	3325	37.96
C5	20 E 1	0	0.00
	21 E 2	27	.31
	22 E 3	167	1.91
	23 E 4	682	7.79
	24 E 5	270	3.08
C6	25 F 1	0	0.00
	26 F 2	116	1.32
	27 F 3	310	3.54
	28 F 4	402	4.59
	29 F 5	0	0.00
		8760	100.00

$$\frac{1}{\sqrt{2\pi}\ \sigma_y(x)} \exp\left(\frac{-y^2}{2\ \sigma_y^2(x)}\right) \qquad (2)$$

is replaced by $1/(3\ \sigma_y(x))$. With this substitution, equation (1) becomes

$$\chi(x,o) = \frac{Q}{3/2\sqrt{2\pi}\ \sigma_y(x)\ \sigma_z(x)u} \exp\left(\frac{-h^2}{2\sigma_z^2(x)}\right) \qquad (3)$$

The amplitude of the top hat is 0.836 of the Gaussian peak: however, the area under the top hat curve is identical to the area under the Gaussian crosswind profile.

For each start-hour selected by the meteorological sampling technique, the CRAC 2 dispersion model uses the subsequent meteorological conditions to predict the dispersion and transport of the released cloud of radioactive material. The sequence of hourly recordings is used to account for changing weather conditions; i.e., wind speed, atmospheric stability, and precipitation may change during plume passage. The wind direction, however, is assumed to be invariant.

Based on the windspeed in each hour, the stability, windspeed, and accumulated precipitation are assigned to all spatial intervals which the plume passes during the hour. If the windspeed for an hour is not sufficient for the plume to fully traverse an interval, the windspeed, stability and accumulated precipitation are averaged for all hours the plume is within that interval (the average of A and C stability is B, of A and B is B). Values of $\sigma_y(x)$ and $\sigma_z(x)$ are calculated for each spatial interval using Pasquill-Gifford curves as provided in Turner[21]. The empirical, best-fit functions for the curves given by Martin and Tikvart[22] are used. Successive growth rates of $\sigma_y(x)$ and $\sigma_z(x)$ for each spatial interval are estimated for the value in the previous spatial interval by calculating the virtual-source distance at the current stability class necessary to give that value, and extrapolating growth to the end of the current spatial interval.

Equation (3) is modified for radioactive decay of each isotope during downwind transport, including build-up of any daughter products, e.g.,

$$Q(x) = Q \exp\left(-\lambda x/u\right)$$

where λ is the radioactive decay constant. Further modifications of equation (3) are incorporated in CRAC 2 to account for the effects of (1) duration of

release, (2) surface roughness, (3) mixing layer depth, (4) dry and wet removal processes, (5) building wake, and (6) plume rise caused by sensible heat buoyancy. Refer to reference [16] for a summary of these effects and submodels in CRAC 2.

Computer Packages

In order to perform the analysis six computer packages were either used, developed or modified. By using an input and output file for each of these computer packages, they were run independently. A brief description of these computer packages will now be given.

Latin Hypercube Sampling (LHS)

The LHS computer packages[13,14] were used to generate probability distributions for each of the eight selected independent variables. For each chosen parameter, probability density functions and any rank correlations must be specified so that the design matrix can be generated.

For a N x k design matrix, each of the k parameters is divided into N equi-probable intervals. Then a random sample is chosen from each interval in a manner that preserves the individual probability density function. In this manner, k N-tuples (an N x 1 vector) of input settings are determined.

The parameter LHS input matrix for our CRAC 2 dispersion study is given in Table 2. The quantitative information was mainly used for demonstration of capability purposes and based primarily upon expert judgement.

CRAC 2 has the weather data sampling option of looking at a discrete set of weather bins as previously discussed. We studied the stabilization within these bins; that is, we performed two calculational cases: the typical four observations per bin and twelve observations per bin. Besides the weather bins, eight other independent variables were selected to perform the CRAC 2 sensitivity and uncertainty analysis. The sample size was determined by the number of observations per bin over the nonzero (e.g., 25 for the New York City data set used) weather bins. Therefore, for the two cases, the sample sizes were 100 and 300, respectively.

CRAC 2

The CRAC 2 computer package[16] was modified to handle the eight independent variables as random variables, with values given by the output of LHS, rather than as deterministic parameters. An output file was created to collect 2400 (7200 for case two) records of the nine independent variables repeated for each distance, the ground and air concentrations, and the 24 corresponding distances chosen.

Table 2: CRAC 2 Variable Ranges and Distributions (Used As Input to LHS code for an SST 1 Release Scenario)

Variable	Range	Distribution	Units
Sensible Heat	10^4 - 3.6 x 10^7	Loguniform	Cal/ sec
Duration of Release	.25 - 1.0 1.0 - 3.0	Uniform p = .5 Uniform p = .5	hour
Dry Deposition Velocity	10^{-3} - 10^{-1}	Loguniform	m/sec
Rain Coefficient	10^{-5} - 10^{-3}	Loguniform	sec^{-1} ($mm\ hr^{-1}$)
Rain Intensity Exponent	0.75 - 1.0	Uniform	1
Mixing Height	250. - 3000.	Uniform	m
Sigma-Y Multiplier	.333 - 3.0	Loguniform	1
Sigma-Z Multiplier	.333 - 3.0	Loguniform	1
Weather Bins	(1.-2.,2.-3.,... 28.-29.)	Uniform	1

Data Base Program (DBPR)

The DBPR computer package was written to prearrange the data file output created by CRAC 2. The file was partitioned on six weather classifications and, within each of these classifications (Table 1) and, the records were rank ordered on distance. For each classification and distance (144 sets in total), the minimum, mean, maximum, and standard deviation were computed and tabled as well as the overall minimum, mean, maximum, and standard deviations. This output file was created to serve as input to the BMDP step-wise-regression computer package and the Weibull computer program.

BIOMEDICAL PROGRAMS (BMDP) - PROGRAM 2R

The UCLA biomedical computer package[23] for performing a step-wise regression was used to get an importance ranking of the eight independent variables. For these runs the dependent variable was the logarithm of ground exposure risk.

Weibull Distribution (WHYD)

To check within and between variations of weighted ground concentrations over the six classifications over the 24 distances, 144 Weibull distributions[24,25] were fitted to corresponding sets of ground concentration values. WHYD was developed and programmed within NRC by using Menon's method of moments to estimate the two-parameter Weibull distribution. Using an iterative scheme, Pittman's estimate was used to estimate the location parameter.

Tell-A-Graph

The fitted Weibull distributions and data were plotted using the Integrated Software System Corporation graphics package[26] in order to do a visual analysis of the within and between variations of ground exposure risk over the six classifications over the chosen distances.

CALCULATED RESULTS

For illustrative purposes, calculations have been performed using CRAC 2 and the LHS design matrix input (based on Table 3) and assuming a severe reactor accident scenario[27] called siting source term one (SST1). This corresponds to the largest releases postulated in the Reactor Safety Study[15] where essentially all installed safety features are assumed to fail. The uncertainty in the ground concentration weighted by the annual frequency of occurrence of that weather class (Ci/m^2) as described by Weibull distributions at 1.25 miles is displayed in Figure 1. The Weibull distributions are

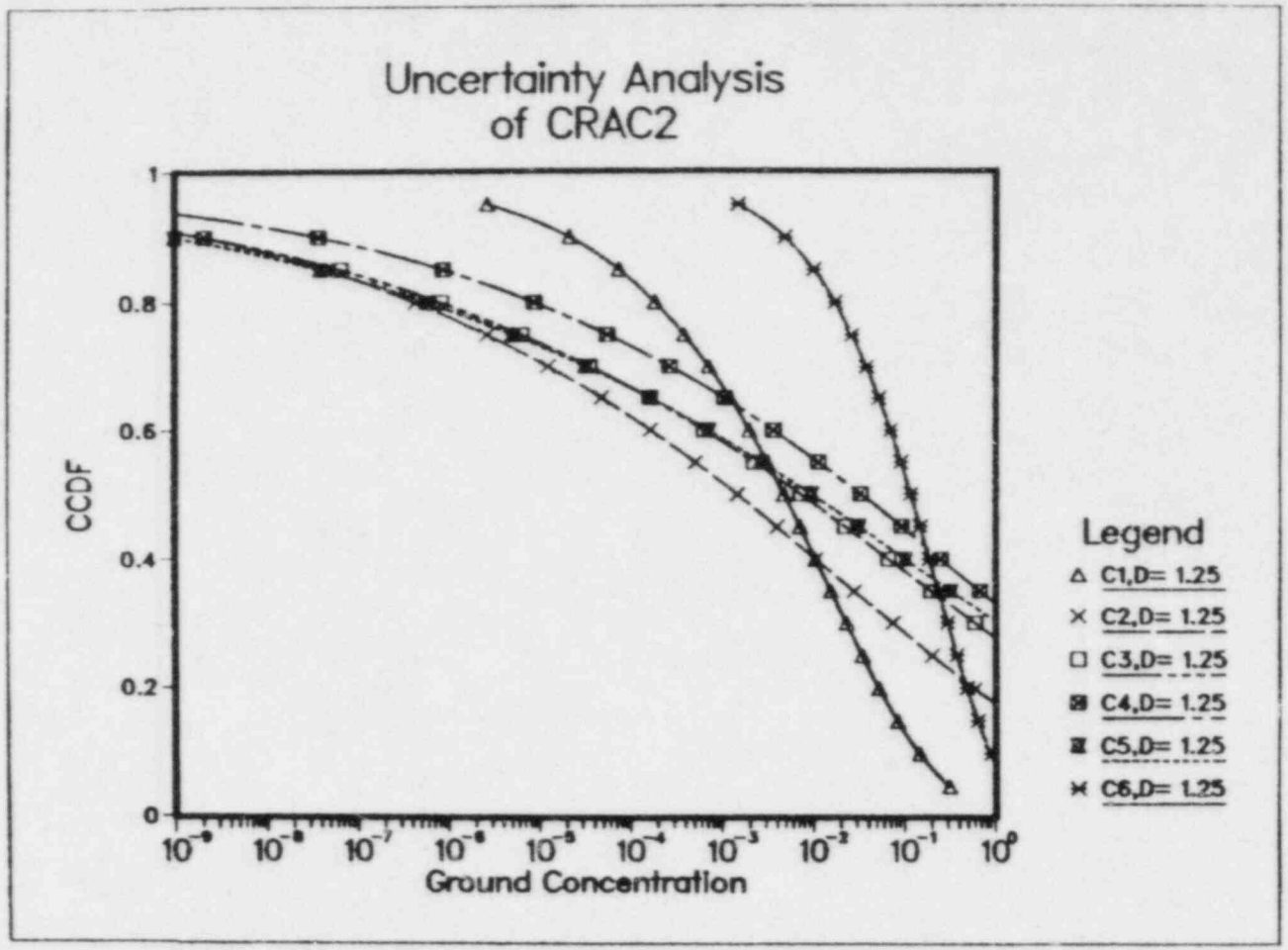

Figure 1 : Cumulative Complementary Distribution Function (CCDF) of Ground Concentration Values (Weighted by Annual Frequency of Each Weather Class) At D = 1.25 miles from the Reactor give an SST1 Release (Ci/m^2).

Refer to Table 1 for definitions of Weather Classes C1 through C6.

shown for each weather class designated C1 through C6 (corresponding to rain, windspeed slowdowns, and stability classes C, D, E and F, respectively). The graphed median values indicate, for example, that the D and F weather classes would result in a higher probability of exceeding a given level of ground deposited radioactive exposure risk than the rain class. For the rain and F stability weather classes the weighted ground concentrations range over several orders of magnitude less than (between CCDF values of 0.1 and 0.9, say) the ranges shown for the other weather classes. Several step-wise regression analysis results are summarized in Table 3. These were performed on logarithmically transformed ground concentration values by weather class. Regression analyses were also performed on transformed ground concentration values at each selected distance as well.

The calculated R values provide a measure of importance of each variable to the uncertainty and may be used to rank the variables. It is also seen in Table 3, for example, that approximately half of the variability in weather class F has been accounted for (based on interpreting the R^2 value). The contributions of each of the variables studied are shown. The relative rankings of the important variables depends upon the weather class under consideration. Further analysis could consider including interaction terms or other independent variables. It is remarked that there are several ways to display uncertainty - of various types. It may be useful to display uncertainty due to the stochastic weather data as well as uncertainty due to both meterological data variations and parameter lack-of-knowledge. See Figure 2 which shows the mean and peak values of these two types of uncertainty in the weighted ground concentration versus distance from release point. The mean values of the distributions differ by approximately 3 to 5 while the "worst-case" values are (not surprisingly) about an order of magnitude apart.

CONCLUSIONS

The process of conducting an integrated uncertainty and sensitivity analysis was described and applied to the CRAC 2 atmospheric transport and deposition model using a suite of mainly existing codes. These include codes to implement Latin hypercube sampling of prescribed distributions and correlations for the variables understudy, to fit statistical distributions and graphically display the information and to perform regression analyses. All in all, the experience was worthwhile and results informative and useful. The present uncertainty and sensitivity study may be built upon to encompass the complete radiological consequence assessment, while still focusing on the display of different uncertainties. Also, other models that specifically address complex terrain and land-sea breeze flows should be examined along with available experimental data.

Table 3: Sensitivity Analysis of CRAC 2 Ground Exposure Results Versus Weather Data Class
(Fractions indicate R values from a step-wise regression of transformed CRAC 2 ground exposure risk results; numbers in parentheses indicate parameter ranking)

CRAC 2 Weather Bins	Weather Frequency (Percent)	Dry Deposition Velocity	Rain Intensity Exponent	Rain Intensity Coefficient	Sigma-Y Multiplier	Sigma-z Multiplier	Duration of Release	Sensible Heat	Mixing Height
Rain (Bins 1-7)	12.29	0.44 (4)	0.38 (1)	0.40 (2)	0.44 (8)	0.44 (7)	0.44 (6)	0.44 (5)	0.42 (3)
Windspeed Slowdowns (8-12)	1.03	0.35 (2)	0.43 (4)	0.43 (6)	0.43 (8)	0.39 (3)	0.28 (1)	0.43 (7)	0.43 (5)
C Stability (13,14)	12.10	0.86 (6)	-	0.83 (4)	0.80 (3)	0.75 (2)	0.86 (5)	0.86 (7)	0.45 (1)
D Stability (15-19)	52.06	0.33 (1)	0.50 (2)	0.68 (7)	0.58 (3)	0.68 (6)	0.64 (4)	0.66 (5)	0.69 (8)
E Stability (20-24)	13.09	0.66 (5)	0.68 (8)	0.61 (3)	0.39 (1)	0.59 (2)	0.67 (6)	0.68 (7)	0.64 (4)
F Stability (25-29)	9.45	0.74 (7)	0.7 (3)	0.58 (1)	0.71 (5)	0.72 (6)	0.74 (8)	0.70 (4)	0.67 (2)

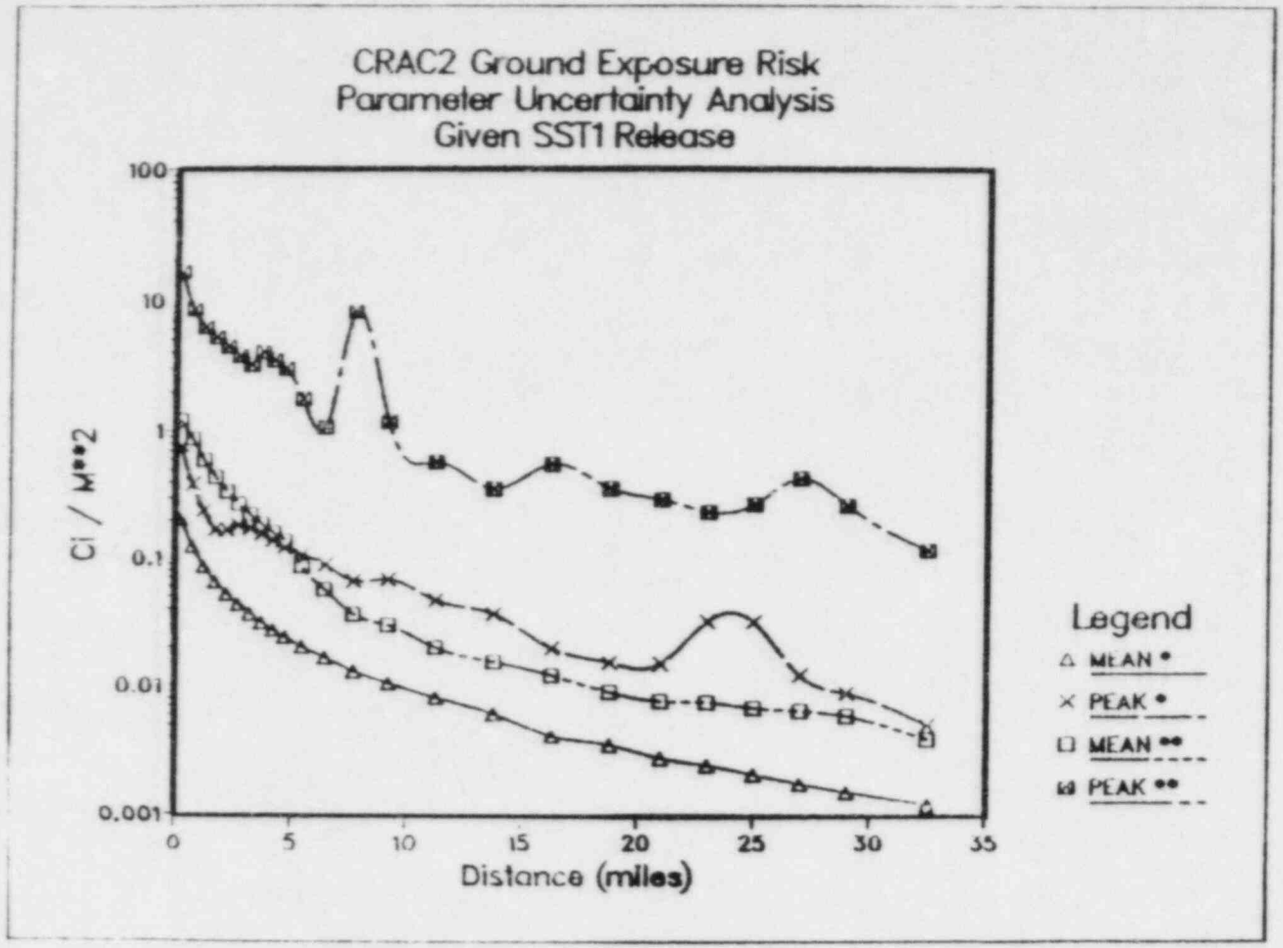

Figure 2: Mean and Peak Values for Stochastic Variation (due to Meteorology) and Full Latin Hypercube (Stochastic and Parameter Lack-of-Knowledge) Uncertainty Analysis of Ground Exposure Risk Versus Distance From Reactor Given an SST1 Release.

* Stochastic Variation
** Stochastic and Lack-of-Knowledge Variation

REFERENCES

1. D. C. Cox and Paul Baybutt, Risk Analysis, Vol. 1, No. 4, p. 251, 1981.

2. Halton, J. H. SIAM Review, 12, 1 (1961).

3. Cochran, William G., Sampling Techniques, John Wiley and Sons, 1963.

4. Cranwell, R. M. and Helton, J. C., "Uncertainty Analysis for Geologic Disposal of Radioactive Waste," pp. 131-144 in Proceedings of Symposium on Uncertainties Associated with the Regulation of the Geologic Disposal of High-Level Radioactive Waste, Gatlingburg, Tennessee, March 9-13, 1981, CONF-810372, ed. D.C. Kocher.

5. Iman, R. L. and W. J. Conover, "The Use of the Rank Transform in Regression," Technometrics 21, p. 499-509, 1979.

6. McKay, M. D., Conover, W. J. and Beckman, R. J., "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code," Technometrics 21, p. 239-245, 1979.

7. R. H. Myers, Response Surface Methodology, Allen and Bacon, Inc. Boston (1971).

8. P. Baybutt, D. C. Cox, and R. E. Kurth, "Methodology for Uncertainty Analysis of Light Water Reactor Meltdown Accident Consequences," Topical Report from Battelle Columbus Laboratories to U.S. Nuclear Regulatory Commission (May 1981).

9. Oblow, E. W., Nuclear Science and Engineering, Vol. 65, p. 187, 1978.

10. D. G. Cacuci, C. F. Weber, E. M. Oblow, and J. H. Marable, "Sensitivity Theory for General System of Nonlinear Equations," Nuclear Science and Engineering 75, 88 (1980).

11. J. E. Campbell, R. L. Iman and M. Reeves, "Risk Methodology for Geologic Disposal of Radioactive Waste: Transport Model Sensitivity Analysis," NUREG/CR-1377, SAND 80-0644, Sandia National Laboratories, Albuquerque, New Mexico.

12. J. C. Helton, R. L. Iman, J. D. Johnson, C. D. Leigh, "Uncertainty and Sensitivity Analysis of a Model for Multicomponent Aerosol Dynamics," SAND 84-1807, Sandia National Laboratories, Albuquerque, New Mexico.

13. Iman, R. I. and M. J. Shortencarier, "A FORTRAN 77 Program and User's Guide for the Generation of Latin Hypercube and Random Samples for Use with Computer Models," U. S. Nuclear Regulatory Commission, NUREG/CR-3624, March 1984.

14. Iman, R. L. and Davenport, J. M., "An Iterative Algorithm to Produce a Positive Definite Correlation Matrix from an Approximate Correlation Matrix (With a Program User's Guide)," Technical Report SAND 8101376, Sandia National Laboratories, Albuquerque, New Mexico, 87185.

15. U.S. Nuclear Regulatory Commission, "Reactor Safety Study, Appendix VI: Calculation of Reactor Accident Consequences, NUREG-75/014, 1975.

16. Ritchie, L. T., J. D. Johnson, R. M. Blond, "Calculations of Reactor Accident Consequences Version 2, CRAC 2: Computer Code, U.S. Nuclear Regulatory Commission, NUREG/CR-2326, February 1983.

17. Ritchie, L. T., Aldrich, D. C. and Blond, R. M., "Weather Sequence Sampling for Risk Calculations," Trans. Am. Nuc. Soc., Vol. 38, 1981.

18. U.S. Nuclear Regulatory Commission, "Environmental Transport and Consequence Analysis," Chapter 9 in "PRA Procedures Guide," Vol. 2, NUREG/CR-2300, 1982.

19. Gifford, F., "Atmospheric Dispersion Models for Environmental Pollution Applications," in Lectures on Air Pollution and Environmental Impact Analysis, D. A. Hanghen, ed. AMS, Boston, 1975.

20. Slade, D. H. (Ed.), Meterology and Atomic Energy, U.S. Atomic Energy Commission TID-24190, 1968.

21. Turner, D. B., "Workbooks of Atmospheric Dispersion Estimates," U.S. EPA Office of Air Programs, Publication No. 999-AP-26, Research Triangle Pork, NC, 1970.

22. Martin, D. O. and J. A. Tikvart, "A General Atmospheric Diffusion Model for Estimating the Effects of Air Quality of one or More Sources," presented at the 61st Annual Meeting of the Air Pollution Control Association.

23. Dixon, W. J. (Chief Editor), BMDP Statistical Software, Berkley, University of California Press, 1981.

24. Sarah, A. E. and B. G. Greenberg, Contributions to Order Statistics, John Wiley and Sons, Inc., 1982.

25. Cohen, A. C., Jr., Technometrics, p. 579, 1965.

26. Tell-A-Graf User's Manual, Version 4.0, Integrated Software System Corporation (ISSCO), San Diego, California, 1981.

27. Blond, R. M., M. Taylor, T. Margulies, M. Cunningham, P. Baranowsky, R. Denning and P. Cybulskis, "The Development of Severe Reactor Accident Source Terms: 1957-81," U.S. Nuclear Regulatory Commission, NUREG-0773, November 1982.

KRIGING: ESTIMATING AREAS OF AQUIFER RECHARGE ON LONG ISLAND

ABSTRACT

N. Oden, A. Meinhold, M. Hauptmann, E. Kaplan

Brookhaven National Laboratory
Upton, New York 11973

Contour maps were produced of the hydrostatic head in the two aquifers of Long Island that are important sources of drinking water. The upper aquifer is known to be contaminated in certain areas. Estimated head-difference maps show areas of likely recharge from the upper aquifer to the lower. A conservative confidence interval is constructed about the difference maps, and areas where recharge is extremely likely are identified. The hypothesis of no recharge on Long Island is formally rejected. Recharge areas are stable over the years when estimated using the same localities in each year.

Key Words

Kriging, Ground Water, Contour Maps, Recharge, Sole-source Aquifer, Confidence Interval, Long Island.

Special Topical Session

Quantification of Informed Opinion

QUANTIFICATION OF INFORMED OPINION

Dale M. Rasmuson

Division of Risk Analysis and Operations
U. S. Nuclear Regulatory Commission
Washington, D. C. 20555

The objective of this session, Quantification of Informed Opinion, is to provide the statistician with a better understanding of this important area. The NRC uses informed opinion, sometimes called engineering judgment or subjective judgment, in many areas. Sometimes informed opinion is the only source of information that exists, especially in phenomenological areas, such as steam explosions, where experiments are costly and phenomena are very difficult to measure.

There are many degrees of informed opinion. These vary from the weatherman who makes predictions concerning relatively high probability events with a large data base to the phenomenological expert who must use his intuition tempered with basic knowledge and little or no measured data to predict the behavior of events with a low probability of occurrence.

The first paper in this session will provide the reader with an overview of the subject area. The second paper will provide some aspects that must be considered in the collection of informed opinion to improve the quality of the information. The final paper contains an example of the use of informed opinion in the area of seismic hazard characterization. These papers should be useful to researchers and statisticians who need to collect and use informed opinion in their work.

ELICITING AND AGGREGATING SUBJECTIVE JUDGMENTS-- SOME EXPERIMENTAL RESULTS

Harry F. Martz, Maurice C. Bryson, and Ray A. Waller
Los Alamos National Laboratory

ABSTRACT

An introductory review of the literature on eliciting and aggregating subjective judgments is provided. Six direct numerical methods for eliciting subjective probabilities over continuous variables are compared using four types of stimuli. The comparison is based on an experiment conducted at Los Alamos. Six mathematical aggregation rules for providing consensus point and confidence interval judgments are also compared. Some simple conclusions are stated.

INTRODUCTION

Much research has been directed toward the study of subjective judgments about uncertain quantities. This research is concentrated in the area known as behavioral decision theory. Einhorn and Hogarth (1981) and Slovic, Fischhoff, and Lichtenstein (1977) provide comprehensive reviews of this general area. A major specialization concerns the practice of eliciting or encoding subjective probabilities. This area can be further partitioned into two subareas. The first concerns elicitation of individual opinions, while the second concerns the formation of an aggregate or combined judgment from a group of individual judgments.

Eliciting Subjective Probabilities

The uncertainty inherent in subjective judgments is usually expressed by means of subjective or personal probability (Hampton, Moore, and Thomas 1973). Such probability measures the "degree of belief" in the assessed event. Hogarth (1975), Spetzler and Staël von Holstein (1975), and Stillwell, Seaver, and Schwartz (1982) provide comprehensive reviews on the assessment of subjective probability.

Four substantive areas have been studied in detail with regard to the quality and performance of subjective probability assessments: (1) military and intelligence, (2) business, (3) medicine, and (4) weather prediction. There is conflicting evidence as to the quality of the assessments within these four areas. Weather forecasters are remarkably good at providing subjective probability assessments (Murphy and Winkler 1974, 1977); bankers, security analysts, and stock analysts are poor (Staël von Holstein 1972, Bartos 1969); medical practitioneers are mixed, but somewhat good (Ludke, Strauss, and Gustafson 1977, Lusted 1977); and intelligence analysts are somewhat good (Zlotnick 1972).

Previous research supports the contention that various task characteristics such as response mode, order, variability and amount of information presented, payoffs, training, and feedback all influence subjectively assessed probabilities (Hogarth 1975, Stillwell, Seaver, and Schwartz 1982). This is consistent with the much larger phychological finding that the same question, when asked in two different ways, wi'l often yield two different answers. Humans also generally have difficulty understanding probabilistic phenomena and in using probability notions to describe uncertain events. Cohen and Hansel (1955) and Kahneman and Tversky (1972) found the concept of a statistical distribution to be lacking among 10-18 year-olds as well as undergraduates. There is also evidence that humans have difficulty with concepts of statistical independence and randomness. (Cohen 1960, Wagenaar 1970, 1972). Also, when faced with the task of estimating parameters intuitively from data, humans have been shown to be fairly accurate at guessing values of central tendency but not the variance of the data (Beach and Swensson 1966, Spencer 1963). Furthermore, evidence indicates that when people think of variability, it is not in the statistical sense of variance (Hofstätter 1939, Lathrop 1967, Beach and Scopp 1968). Such findings suggest that naive persons have difficulty in subjectively assessing probability distributions. The evidence supports the thesis stated by Hogarth (1975) that "... man, as a selective, step-wise information processing system with limited capacity, is ill-suited to the task of assessing probability distributions within the framework of the more common statistical models."

The notion of secondary validity of subjective probability assessments was first introduced by Winkler and Murphy (1968). This property states that subjective probability assessments, such as intervals or ranges, should yield relative frequencies close to those associated with them. For example, 90% intervals over continuous variables should yield about 90% coverage of the corresponding true values. Because naive subjects have difficulty in accurately assessing the variability of their subjective distributions, this property is usually not satisfied with regard to the spread of subjective distributions. Anderson (1980), Hogarth (1975), Holloway (1979), Lichtenstein, Fischhoff,and Phillips (1977), and Slovic, Fischhoff, and Lichtenstein (1980) cite many references of this phenomenon. For example, Alpert and Raiffa (1969) found that experienced adults often overestimate the degree of certainty of their estimates and thus are too sure of their information. They observed that 98% interval estimates frequently corresponded to 50-60% intervals. Furthermore, when informed of the inaccuracy of their intervals, subjects were still not able to make accurate assessments in future intervals. Similar results are also reported by Leonardz and Staël von Holstein (1967). Tversky and Kahneman (1974) attribute such bias in part to "anchoring". Kahneman and Tversky (1979), Wallsten and Budescu (1980), and von Winterfeldt (1980) list and give examples of various other biases in probabilistic judgments as well as explanations of the heuristics that may be the causes. Also, assessing such intervals as 98% intervals is difficult because it requires the estimation of events with one chance in a hundred of occurring. Subjectively assessing such small probabilities is difficult. Because comparisons do not readily come to mind and also because humans tend to avoid using the ends of the probability scale, small probabilities are generally overestimated and large probabilities underestimated (Peterson and Beach 1967, Sanders 1973, Schaeffer and Borcherding 1973, Winkler 1967).

There are numerous ways that subjective probabilities can be elicited. Seaver and Stillwell (1983) describe five methods along with the advantages and disadvantages of each method. Paired comparisons and ranking/rating methods derive from their use as psychological scaling techniques (Thurstone 1927, Torgerson 1958, Bock and Jones 1968). Direct and indirect numerical methods, as well as methods based on the use of multiattribute utility theory (Keeney and Raiffa 1976, Edwards 1977), are also discussed. The "best" method to use depends on numerous task characteristics and is often unknown. As a compromise, direct numerical methods are often employed in practice and are considered here. In addition, direct numerical methods are simple, inexpensive, and easy to use.

Subjective numerical judgments can be directly elicited in many ways. The uncertainty in a subjective judgment can be elicited by use of odds, log odds, probability, log probability, fractiles, chances, and other means as well. Seaver, von Winterfeldt, and Edwards (1978) found that subjects responding in probabilities or odds were much better than those responding with fractiles, fractiles expressed in odds, or log odds. Seaver and Stillwell (1983) concluded that odds responses marked on a logarithmically-spaced scale is the best direct elicitation procedure, particularly for relatively unlikely or rare events. For example, an expert on human factors would be asked to mark the logarithmically-spaced scale at the point that represents his or her judgment of the odds (p/1-p) that an operator fails to respond to an annunciated legend light (one light only).

There are several basic problems with the use of odds responses marked on a logarithm scale. First, the odds scale must be preselected of sufficient scope to provide sufficient divisions for the approximate magnitude of the odds. For some events the approximate range and number of divisions for the probability of the event may be unknown. Thus, such a scale may or may not represent a valid choice. Second, the use of such a scale may precondition the expert to make responses within the given range, even though he or she may believe differently. Third, the use of log odds requires that the concept and use of both "odds" and, to a lesser extent, "logarithm" be understood by the expert. Many experts might not feel comfortable with the notion and use of "odds" without sufficient training. Such training may not be feasible in certain modes of non-contact elicitations, such as mailed responses.

An alternative way to directly elicit judgments regarding subjective probabilities of events is by use of chances. Seaver and Staël von Holstein (1975) recommend this approach. For example, in the previous human factors example, the expert would be asked for his or her subjective assessment of the chances (that is, one in how many) that an operator would fail to respond to an annunciated legend light. The desired probability is then simply 1/"how many". The use of chances alleviates some of the difficulties in the use of log odds discussed above. Seaver and Staël von Holstein (1975) also state that the use of chances is particularly appropriate for unlikely or rare events because experts can discriminate more easily between "one in 100" and "one in 1000" than between the absolute numbers 0.01 and 0.001. In addition, the constraint of a closed probability scale, with the attendant human reluctance to avoid the extremes of such a scale, is avoided.

Aggregating Subjective Probabilities

The second major subarea of behavioral decision theory concerns the formation of a group judgment from the individual judgments of the group members. This is the so-called consensus or opinion-pool problem (Stone 1961), for which a large research literature likewise exists. Hogarth (1975), Seaver (1976) (1978), Stillwell, Seaver, and Schwartz (1982), and Mirkin (1979) provide comprehensive bibliographies of various methods for obtaining consensus opinions.

Seaver (1976) concluded that (1) the consensus assessment is, on the average, more accurate than the individual judgments primarily due to a decrease in the error variance around the true value, (2) the improved accuracy of group judgments over individual judgments appears to hold for factual as opposed to value judgments, and (3) a larger diversity of individual experience and opinion among group members leads to relatively more accurate group judgments.

Two general approaches have been extensively employed. The first consists of mathematical or statistical aggregation rules for arriving at a consensus judgment. Such methods range from simple unweighted and weighted averages to more sophisticated schemes based on the use of Markov chain methods and Bayesian models. Generally, small or no differences in the quality of the consensus judgments have been found among the methods (Gough 1975, Rouse, Gustafson, and Ludke 1974, Seaver 1978, Stael von Holstein 1970, Winkler 1971, Winkler and Cummings 1972). In fact, Einhorn and Hogarth (1976) found that equal weighting schemes often resulted in lower expected mean square error than those based on "optimal" least squares weights.

Dalkey (1977) proved a result known as an "impossibility theorem" which states that there is no mathematical aggregation rule for aggregating individual probability assessments that is consistent with the usual rules of probability. However, Bordley and Wolff (1981) claim that not all of the assumptions Dalkey made in proving his theorem are likely to be reasonable in practice. In particular, weighted averages and Bayesian multiplicative product rules (Morris 1977) are not covered by Dalkey's theorem (Bordley and Wolff 1981). Such Bayesian multiplicative aggregation rules are recommended by Seaver and Stillwell (1983) for use in estimating human error probabilities. However, there is a problem in the use of these rules which may make them inappropriate for estimating the probability of simple events (Morris 1983).

The second general approach to the consensus problem has focused on group interaction techniques. In these methods either the individual responses or a group summary response is fedback to each of the group members for use as additional information to be used in subsequent iterated judgments. The constraints placed upon their mode of interaction and the instructions they receive before the interaction constitute the major differences among procedures. In an extreme case, the Delphi procedure (Dalkey 1969b) requires that the individual not interact face-to-face at all, but, instead, they make assessments and are given feedback about what the group as a whole responded, and a new set of judgments is made. If, after some number of iterations, no consensus is reached, mathematical aggregation is often used to produce the final group assessment. The problem with Delphi is the lack of evidence that it gives good answers. Only two studies (Dalkey 1969a, 1969b) support Delphi as superior to even simple face-to-face discussion groups. On the other hand, there is considerable evidence that Delphi yields answers that are no better or

no worse than other procedures (Brockhoff 1975, Van de Ven and Delbecq 1974, Fischer 1981, Seaver 1978).

Another group interaction technique which has received considerable attention is the Nominal Groups Technique (NGT), developed at the University of Wisconsin (Delbecq and Van de Ven 1971). NGT procedures get their name from their main characteristic; namely, that the group does not interact in a normal manner, but rather in a very limited or "nominal" sense. In other words, the group interaction is tightly controlled. As discussed by Delbecq, Van de Ven, and Gustafson (1975), it consists of four steps: (1) silent judgments by individuals in the presence of the group; (2) presentation of all individual assessments to the group without discussion; (3) group discussion of each judgment for clarification and evaluation; and (4) individual reconsideration of judgments and mathematical aggregation. The main drawback of the NGT is the need to assemble the group in one geographic location. Often this is either impossible or impractical. On the other hand, there is no need to physically assemble the group when using either mathematical aggregation or the Delphi method. Winkler (1968) concluded that the use of simple feedback yielded stronger consensus estimates even though there was no group discussion. Overall, the evidence is inconclusive, and even contradictory, regarding the value of group interaction techniques. Some important questions remain: Should group interaction techniques be used to aggregate subjective judgments of uncertain quantities? Should they be used in conjunction with mathematical aggregation rules? If so, what rules should be used?

Los Alamos Experiment

An experiment was designed and conducted at Los Alamos National Laboratory to address these and related questions. Various methods of elicitation and aggregation were considered in a single integrated experiment which had several objectives. The first objective was to investigate the performance of six direct numerical methods for eliciting subjective judgments corresponding to four types of stimuli. Two of the methods involved non-interactive Delphi-type group feedback. The stimuli required opinions regarding magnitude, probability and chance, rate, and percentage. The sampled population consisted of mostly well-educated and well-informed persons, some of which are substantive and/or normative experts for at least some of the stimuli. The second objective was to compare the performance of six mathematical aggregation rules for obtaining consensus point judgments and six rules for obtaining consensus confidence intervals. The final objective was to compare the performance of normative versus non-normative experts in the quality of the judgments produced.

EXPERIMENT

To develop a better understanding of the foregoing issues, an experiment was conducted at Los Alamos National Laboratory during March and April of 1984. The essential approach was to ask questions of "experts" -- individuals with at least some personal knowledge of the subjects in question -- and compare their answers (including uncertainty bounds) with the true answers. Questions were selected so that individuals questioned would have some, but not complete, knowledge; also so that those conducting the experiment could verify the true answers.

Subjects

Subjects in the experiment were 48 Los Alamos employees. All were staff members, supervisors, or higher-level technical workers, and thus represented a relatively homogeneous group in terms of above-average education, a common working environment, and similar exposure to everyday activities. This homogeneity of background made it possible to select questions satisfying the criteria noted above. Although there was no random selection of subjects from any identifiable population, subjects were assigned randomly to groups that were asked questions in particular orders and formats (see below).

Questions

Questions asked of the subjects were divided into four types (henceforth referred to as "stimuli" of the experiment): percentages; magnitudes; rates; and probabilities or chances. Examples of each type are as follows:

> Percentage: As of November 1983, what percentage of Laboratory staff members was constituted by females?
>
> Magnitude: What was the highest annual total snowfall (in inches) ever recorded in Los Alamos?
>
> Rate: At what average monthly rate did Laboratory employees use sick leave during the first half of 1983?
>
> Probability or chance: What are the chances (one in how many) that a 40-year-old person (disregarding sex) will die before the age of 50?

Answers to all questions were in the form of continuous variables, so that subjects could be asked to put uncertainty intervals on their answers.

Response Modes

Questions were each to be answered in one of six response modes, four involving a single response and two involving a small degree of feedback. The four modes are identified as follows:

> A: Subjects gave a certainty interval -- upper and lower bounds within which they were virtually certain that the correct answer would lie. Subsequently, they gave a "best point estimate" for the correct answer. The term "best" was not further defined, nor was "virtually certain."
>
> B: Same as A except that the order was reversed: subjects were asked first to give a best point estimate, then a certainty interval.
>
> C: Subjects were asked to give a 90% confidence interval for the correct answer, i.e., upper and lower bounds such that there is a subjective 5% probability that the correct answer is above the upper bound, and another 5% below the lower bound. Then, subjects were asked to give a median point estimate, i.e., value such that there is an even bet that the correct answer is larger [smaller].

D: Subjects supplied a best point estimate and a symmetric certainty interval, i.e., an additive component (plus or minus x) such that the subject is virtually certain that the correct answer lies within x of the estimate.

E: Those having previously answered a question under response mode A were provided with the answers given by the 11 other subjects answering the same question in the same mode. They were then invited to change either their point or interval estimates. (To preserve anonymity of the participants, the other respondents were not identified.)

F: Similar to E, but iterating the responses in response mode C.

Experimental Design

For the single response portion (modes A-D), the survey response modes were balanced over subjects and questions as shown in the following array. The categories I, II, III, IV refer to four groups of subjects, 12 in each group.

CATEGORY	QUESTION																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
I	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C	D	D	D	D	D
II	D	D	D	D	D	A	A	A	A	A	B	B	B	B	B	C	C	C	C	C
III	C	C	C	C	C	D	D	D	D	D	A	A	A	A	A	B	B	B	B	B
IV	B	B	B	B	B	C	C	C	C	C	D	D	D	D	D	A	A	A	A	A

Considerable randomization was employed. As noted, subjects were randomly assigned to categories I-IV. Next, the order of the response modes A-D was randomized. Finally, the order of questions within a response mode was randomized. For example, a subject in category II might have the response mode order BACD, in which case a possible question order would be 12-15-11-14-13-6-10-9-7-8-18-19-17-20-16-5-4-1-3-2.

RESULTS

As expected in view of human variability, there were extreme outliers present in some of the responses, which would have made interval estimation or point comparisons all but meaningless. Gross outliers were first identified and removed by simply calculating the standardized residual (response-mean)/standard deviation for the point estimates in each question-mode combination. The (somewhat arbitrary) rule used was to eliminate those responses with absolute residuals exceeding 2.5. (See Grubbs, 1950, for a discussion of such criteria.) Using this rule, 43 of the 1440 responses were eliminated. Repeating the procedure with the reduced data set eliminated another 14 outliers; a third iteration eliminated another three; and a fourth eliminated one more. The remaining 1379 estimates revealed no residuals with absolute values exceeding 2.5.

Response Mode and Stimulus Comparisons

Figure 1 compares the average performance (averaged over subjects) for each of the four stimuli and each of the six response modes. The criterion for comparison is the relative absolute error (error of estimation as a fraction of the correct answer). This permitted subsequent aggregation over stimulus types and response modes, which would not have been possible without standardization of some such form. Approximate 99% confidence intervals are shown to illustrate the sizeable effects of subject-to-subject variation. Clearly there are no substantial differences among response modes. However, the errors are much worse (and much more variable) for the probability stimulus than for any of the other three. Figure 2 makes this effect more vivid, by aggregating the results over response mode. Aggregating over stimulus type (Figure 3) illustrates the much lesser effect of response mode. Modes E and F (those incorporating feedback) did give slight improvements compared to the other four, and do appear to reduce variability. (This would be expected, since those subjects giving very extreme answers would be expected to modify their answers after noting how much they differed from those of other respondents.)

Modes E and F in particular gave improvements over modes A and C respectively, in which the same questions had been asked without feedback. A sign-test comparison of E vs. A (used in preference to a paired-t test because of the sizeable number of subjects who did not change their estimates as a result of feedback) indicated significant improvement at the .0001 level, as did a similar comparison of F vs. C.

Subject Performance

As noted in the introduction, earlier studies have found that subjects tend to overstate their confidence in subjective interval estimation -- or equivalently, that they make their intervals too narrow. This effect was verified in our study. Figure 4 illustrates the relative frequency of "surprises" -- outcomes wherein the true answer lies outside the subject's expressed interval -- for each response mode and stimulus type. Again, the response mode appears to have little effect. All modes, though, gave a much higher frequency of surprises than would have been the case had the experts been accurately assessing their own error magnitudes. In Figure 5, which aggregates the results over stimulus type, it is seen that mode C does tend to give slightly more surprises than the others; this would be expected since it called for 90% intervals while modes A, B, and D called for certainty intervals. Also, the feedback modes E and F tended to give fewer surprises, with the 90%-interval mode F giving more than the certainty-interval mode E as would be expected. Even the feedback modes result in intervals that are much too narrow, though.

Winkler (1967) and Staël von Holstein (1971, 1972) argued that subjects with statistical expertise should outperform those lacking such expertise, especially in terms of accurately assessing confidence interval widths. We did not find this to be provable. The average relative absolute errors for statisticians and non-statisticians were .66 $\pm$ .06 and .70 $\pm$ .04, respectively. Containment frequencies (frequencies with which the intervals contained the correct answers) were .57 $\pm$ .03 and .54 $\pm$ .01 respectively. Thus, the statisticians in our sample (including nine subjects with graduate degrees in statistics) showed only slightly better performance than others in the sample.

Combining Judgments

It was suggested by Martz and Bryson (1983) that interval width assessed in the expert's uncertainty judgment might be a suitable weighting factor for the combining of expert opinions, putting more weight on the opinions of those claiming more accurate knowledge. This raises the question: do those putting relatively narrow intervals on their own estimates really tend to come closer to the correct answer? This issue was addressed by comparing absolute errors in point estimates with the subjective confidence interval widths. These two variables showed a significant (.05 level) positive correlation in only six cases, were significantly negatively correlated in three, and showed no significant correlation in 11. Thus, there is no evidence that subjective confidence interval widths tell us anything about the real accuracy of the point estimate.

For each of the 120 question-method groups, the 12 point estimates were combined by several methods:

MEAN: take the unweighted arithmetic mean of the estimates.
MEDIAN: take the .50 fractile of the estimates.
MIDIQR: take the mean of the .25 and .75 fractiles of the estimates.
MIDRANGE: take the mean of the largest and smallest estimates.
WTMEAN1: take a weighted arithmetic mean, with weights proportional to the inverse z-score (standardized residual) associated with each point estimate.
WTMEAN2: take a weighted arithmetic mean with weights proportional to the inverse of the individual's interval width.

Figure 6 compares these methods, using the average relative absolute error as a criterion. There is a slight (though nonsignificant) preference for the median as an aggregate estimate, which when added to its simplicity and robustness seems to indicate that it is a desirable aggregation procedure. Comparison of the median to the individual (unaggregated) estimates showed that, in half of the 120 groups, the median fell in the upper third of the unaggregated estimates. Thus, there is a noticeable gain as a result of aggregating the estimates.

Aggregating interval estimates is more difficult. Five methods were initially considered:

1) An interval from the most extreme lower bound to the most extreme upper bound.

2) Interval from the median of all lower bounds to the median of all upper bounds.

3) Interval from the mean of all lower bounds to the mean of all upper bounds.

4) Interval from the .25 fractile of lower bounds to the .75 fractile of upper bounds.

5) Interval from the .75 fractile of lower bounds to the .25 fractile of upper bounds.

Figure 7 shows the average containment frequencies for each method. It is clear that none of them do particularly well. It is especially interesting to note that methods (1) and (4) gave exactly the same intervals, indicating that there is no information about the correct answer contained in the most extreme 25% of the lower and upper bounds. Basically the intervals produced by (1) and (4) are too wide for practical use while those produced by the other three have such low containment fractions as to be relatively uninteresting.

A different approach recognizes the relatively uninformative nature of the interval estimates and uses strictly the point estimates. Dalkey (1969c) and Martino (1970) indicated that expert point estimates follow roughly a lognormal distribution (a conclusion that was shown to be consistent with our data). For each group, then, we took logarithms of point estimates, calculated their means and standard deviations, and established intervals going k standard errors on either side of the mean. Figure 8 gives the resulting performance (coverage frequency) as a function of k. It appears that an interval width of $\pm$ 5 standard errors gives a reasonable approximation to 90-95% coverage. Using the 5-standard-error rule, Figure 9 shows the coverage results for the four stimulus types, indicating fairly consistent behavior for the rule.

CONCLUSIONS

1) There was a consistent tendency on the part of subjects to understate the amount of uncertainty in their estimates. This was true for all response modes and all forms of questions (stimuli), but was worst for stimuli relating to probability or chance estimates. Relative errors were also worst for those stimuli.

2) Even minimal group feedback (anonymous presentation of other respondents' judgments) gave some improvement both in accuracy and in coverage by probability intervals.

3) Respondents with statistical training showed only minimally better performance than those lacking such training.

4) Aggregation of expert opinions using group medians does give some improvement in accuracy.

5) There is no evident correlation between the width of an expert's self-assessed uncertainty interval and the actual accuracy of the expert's point estimate.

6) Interval estimates do not give useful information in terms of pooling data to get over-all interval estimates. A better approach appears to be one relying on point estimates only, with the pooled interval estimate consisting of the averaged point estimates plus or minus five standard errors. This appears to give, typically, a 90 to 95% coverage probability.

REFERENCES

ALPERT,M. and RAIFFA,H. (1969), "A Progress Report on the Training of Probability Assessors", unpublished report, Harvard University.

ANDERSON,J.R. (1980), Cognitive Psychology and Its Implications. San Francisco: W.H.Freeman.

BARTOS,J. A. (1969), "The Assessment of Probability Distributions for Future Security Prices," unpublished Ph.D. thesis, Indiana University, Bloomington.

BEACH,L. R. and SCOPP,T. S. (1968), "Intuitive Statistical Inferences About Variances," Organizational Behavior and Human Performance, 3, 109-123.

BEACH,L. R. and SWENSON,R. G. (1966), "Intuitive Estimation of Means," Psychonomic Science, 5, 161-162.

BOCK,R. D. and JONES,L. V. (1968), The Measurement and Prediction of Judgment and Choice, San Francisco: Holden-Day.

BORDLEY,R.F. and WOLFF,R.W. (1981), "On the Aggregation of Individual Probability Estimates," Management Science, 27, 959-964.

BROCKOFF,K. (1975), "The Performance of Forecasting Groups in Computer Dialogue and Face-to-Face Discussion," in The Delphi Method: Techniques and Applications, eds. H. Linstone and M. Turoff, Reading, MA: Addison-Wesley.

COHEN,J. (1960), Chance, Skill, and Luck, Harmondsworth, Middlesex, England: Penguin Books.

COHEN,J. and HANSEL,C. E. M. (1955), "The Idea of a Distribution," British Journal of Psychology, 46, 111-121.

DALKEY,N. (1969), "Analysis from a Group Opinion Study, "Futures, 1, 541-551. (a)

DALKEY,N. (1969), "An Experimental Study of Group Opinion: The Delphi Method," Futures, 1, 408-426. (b)

DALKEY,N. (1969), "An Experimental Study of Group Opinion," RAND Corporation Report RM-5888-PR, The RAND Corporation, Santa Monica, CA. (c)

DALKEY,N. (1977), "Group Decision Theory," UCLA Technical Report 7749, University of California, School of Engineering and Applied Science, Los Angeles, CA.

DELBECQ,A. and VAN DE VEN,A. (1971), "A Group Process Model for Problem Identification and Program Planning," Journal of Applied Behavioral Science, 7, 466-492.

DELBECQ,A., VAN DE VEN,A., and GUSTAFSON,D. (1975), Group Techniques for Program Planning, Glenview, IL: Scott, Foresman.

EDWARDS,W. (1977), "How to Use Multiattribute Utility Measurement for Social Decision Making," IEEE Transactions on Systems, Man, and Cybernetics, SMC-7, 326-340.

EINHORN,H. J. and HOGARTH,R. M. (1976), "Unit Weighting Schemes for Decision Making," Organizational Behavior and Human Performance, 18.

EINHORN,H.J. and HOGARTH,R.M. (1981), "Behavioral Decision Theory: Processes of Judgment and Choice," Annual Review of Psychology, 32, 53-88.

FISCHER,G.W. (1981), "When Oracles Fail-A Comparison of Four Procedures for Aggregating Subjective Probability Forecasts," Organizational Behavior and Human Performance, 28, 96-110.

GOUGH,R. (1975), "The Effect of Group Format on Aggregate Subjective Probability Distributions," in Utility, Probability, and Human Decision-Making, eds. D. Wendt and C. Vlek, Dordrecht-Holland: Reidel.

GRUBBS,F.E. (1950), "Sample Criteria for Testing Outlying Observations," Annals of Mathematical Statistics, 21, 27-58.

HAMPTON,J.M., MOORE,P.G., and THOMAS,H. (1973), "Subjective Probability and Its Measurement," Journal of the Royal Statistical Society, Series A, 136, 21-42.

HOFSTÄTTER,P. R. (1939), "Uber die Schätzung von Gruppeneigenschaften," Zeitschrift für Psychologie, 145, 1-44.

HOGARTH,R.M. (1975), "Cognitive Processes and the Assessment of Subjective Probability Distributions," Journal of the American Statistical Association, 70, 271-289.

HOLLOWAY,C.A. (1979), Decision Making under Uncertainty, Models and Choices, New Jersey: Prentice-Hall.

KAHNEMAN,D. and TVERSKY,A. (1972), "Subjective Probability: A Judgment of Representativeness," Cognitive Psychology, 3, 430-454.

KAHNEMAN,D. and TVERSKY,A. (1979), "Intuitive Prediction: Biases and Corrective Procedures," Management Science, 12, 313-327.

KEENEY,R. L. and RAIFFA,H. (1976). Decisions with Multiple Objectives: Preferences and Value Tradeoffs, New York: John Wiley.

LATHROP,R. G. (1967), "Perceived Variability," Journal of Experimental Psychology, 73, 498-502.

LEONARDZ,B. and STAEL VON HOLSTEIN,C. A. S. (1967), "A Comparison between Bayesian and Classical Methods for Estimating Unknown Probabilities," Technical Report No. 3, Division of Applied Mathematics, Brown University, Providence, RI.

LICHTENSTEIN,S., FISCHHOFF,B., and PHILLIPS,L.D. (1977), "Calibration of Probabilities: The State of the Art," in Decision Making and Human Affairs, Dordrecht-Holland: Reidel.

LUDKE,R. L., STRAUSS,F. F., and GUSTAFSON,D. H. (1977), "Comparison of Five Methods for Estimating Subjective Probability Distributions," Organizational Behavior and Human Performance, 19, 162-179.

LUSTED,L. B. (1977), "A Study of the Efficiency of Diagnostic Radiologic Procedures," final report, Efficacy Study Committee of the American College of Radiology, Chicago.

MARTINO,J.P. (1970), "Lognormality of Delphi Estimates," Technological Forecasting, 1, 355-358.

MARTZ,H.F. and BRYSON,M.C. (1984), "A Statistical Model for Combining Biased Expert Opinions," IEEE Transactions on Reliability, R-32.

MIRKIN,B.G. (1979), Group Choice, Washington, DC: Winston.

MORRIS,P.A. (1977), "Combining Expert Judgments: A Bayesian Approach," Management Science, 23, 679-692.

MORRIS,P.A. (1983), "An Axiomatic Approach to Expert Resolution," Management Science, 29, 24-32. Low-Probability/High-Consequence Risk Analysis, New York: Plenum.

MURPHY,A. H. and WINKLER,R. L. (1974), "Credible Interval Temperature Forecasting: Some Experimental Results," Monthly Weather Review, 102, 784-794.

MURPHY,A. H. and WINKLER,R. L. (1977), "Can Weather Forecasters Formulate Reliable Probability Forecasts of Precipitation and Temperatures?", National Weather Digest, 2, 2-9.

PETERSON,C. R. and BEACHH,L. R. (1967), "Man As An Intuitive Statistican," Psychological Bulletin, 68, 29-46.

ROWSE,G., GUSTAFSON,D., and LUDKE,R. (1974), "Comparison of Rules for Aggregating Subjective Likelihood Ratios," Organizational Behavior and Human Performance, 12, 274-285.

SANDERS,F. (1973), "Skill in Forecasting Daily Temperature and Precipitation: Some Experimental Results," Bulletin of the American Meteorological Society, 54, 1171-1179.

SCHAEFFER,R. E. and BORCHERDING,K. (1973), "The Assessment of Subjective Probability Distributions: A Training Experiment," Acta Psychologica, 37, 117-129.

SEAVER,D.A. (1976), "Assessment of Group Preferences and Group Uncertainty for Decision Making," SSRI Research Report 76-4, University of Southern California, Social Science Research Institute, Los Angeles, CA.

SEAVER,D.A. (1978), "Assessing Probability with Multiple Individuals: Group Interaction Versus Mathematical Aggregation," SSRI Research Report 78-3, University of Southern California, Social Science Research Institute, Los Angeles, CA.

SEAVER,D. A. and STAEL VON HOLSTEIN,C. A. S. (1975), "Probability Encoding in Decision Analysis," Management Science, 22, 340-358.

SEAVER,D. A. and STILLWELL,W. G. (1983), "Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations," SNL Contractor Report SAND82-7054(NUREG/CR-2743), Sandia National Laboratories, Albuquerque, NM.

SEAVER,D. A., VON WINTERFELDT,D., and EDWARD,W. (1978), "Eliciting Subjective Probability Distributions on Continuous Variables," Organizational Behavior and Human Performance, 21, 379-391.

SLOVIC,P., FISCHHOFF,B., and LICHTENSTEIN,S. (1977), "Behavioral Decision Theory," Annual Review of Psychology, 28, 1-39.

SLOVIC,P., FISCHHOFF,B., and LICHTENSTEIN,S. (1980), "Fact Versus Fears: Understanding Perceived Risk," in Societal Risk Assessment, New York: Plenum.

SPENCER,J. A. (1963), "A Further Study of Estimating Averages," Ergonomics, 6, 255-265.

SPETZLER,C.S. and STAEL VON HOLSTEIN,C.A.S. (1975), "Probability Encoding in Decision Analysis," Management Science, 22, 340-358.

STAEL VON HOLSTEIN,C. A. S. (1970). "Some Problems in the Practical Application of Bayesian Decision Theory," Behavioral Approaches to Management, Gothenburg: The Graduate School of Economics and Business Administration.

STAEL VON HOLSTEIN,C.A.S. (1971), "The Effect of Learning on the Assessment of Subjective Probability Distributions," Organizational Behavior and Human Performance, 6, 304-315.

STAEL VON HOLSTEIN, C.A.S. (1972), "Probabilistic Forecasting: An Experiment Related to the Stock Market," Organizational Behavior and Human Performance, 8, 139-158.

STILLWELL,W.G., SEAVER,D.A., and SCHWARTZ,J.P. (1982), "Expert Estimation of Human Error Probabilities in Nuclear Power Plant Operations: A Review of Probability Assessment and Scaling," SNL Contractor Report SAND81-7140/NUREG/CR-2255, Sandia National Laboratories, Albuquerque, NM.

STONE,M. (1961), "The Opinion Pool," Annals of Mathematical Statistics, 32, 1339-1342.

THURSTONE,L. L. (1927), "A Law of Comparative Judgment," Psychological Review, 34, 273-286.

TORGERSON,W. S. (1958), Theory and Methods of Scaling, New York: John Wiley.

TVERSKY,A. and KAHNEMAN,D. (1974), "Judgment under Uncertainty: Heuristics and Biases," Science, 185, 1129-1131.

VAN DE VEN,A. and DELBECQ,A. (1974), "The Effectiveness of Nominal, Delphi, and Interacting Group Decision Making Processes," Academy of Management Journal, 17, 605-621.

WAGENAAR,W. A. (1970), "Appreciation of Conditional Probabilities in Binary Sequences," Acta Psychologica, 34, 348-356.

WAGENAAR,W. A. (1972), "Generation of Random Sequences by Human Subjects: A Critical Survey of Literature," Psychological Bulletin, 77, 65-72.

WALLSTEN,T. S. and BUDESCU,D. V. (1980), "Encoding Subjective Probabilities: A Psychological and Psychometric Review," draft report, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, NC.

WINKLER,R.L. (1967), "The Assessment of Prior Distributions in Bayesian Analysis," Journal of the American Statistical Association, 62, 776-800.

WINKLER,R. L. (1968), "The Consensus of Subjective Probability Distributions," Management Science, 1, B-61 -B-75.

WINKLER,R.L. (1971), "Probabilistic Prediction: Some Experimental Results," Journal of the American Statistical Association, 66, 675-685.

WINKLER,R.L. and CUMMINGS,L.L. (1972), "On the Choice of a Consensus Distribution in Bayesian Analysis," Organizational Behavior and Human Performance, 7, 63-76.

WINKLER,R.L. and MURPHY,A.H. (1968), "'Good' Probability Assessors," Journal of Applied Meteorology, 7, 751-758.

VON WINTERFELDT,D. (1980), "Some Sources of Incoherent Judgments in Decision Analysis," Decision Science Consortium, Falls Church, VA.

ZLOTNICK,J. (1972), "Bayes' Theorem for Intelligence Analysis," in Computer Diagnosis and Diagnostic Methods, ed. J. A. Jacquez, Springfield, IL: C.C.Thomas.

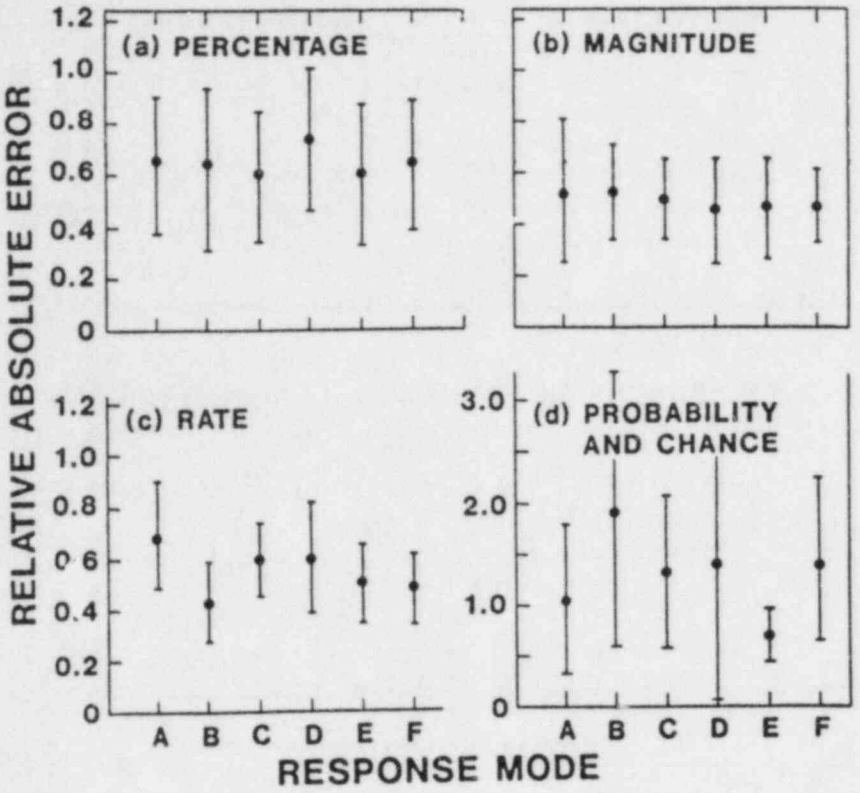

Figure 1. The relative absolute error for each response mode for each type of stimulus.

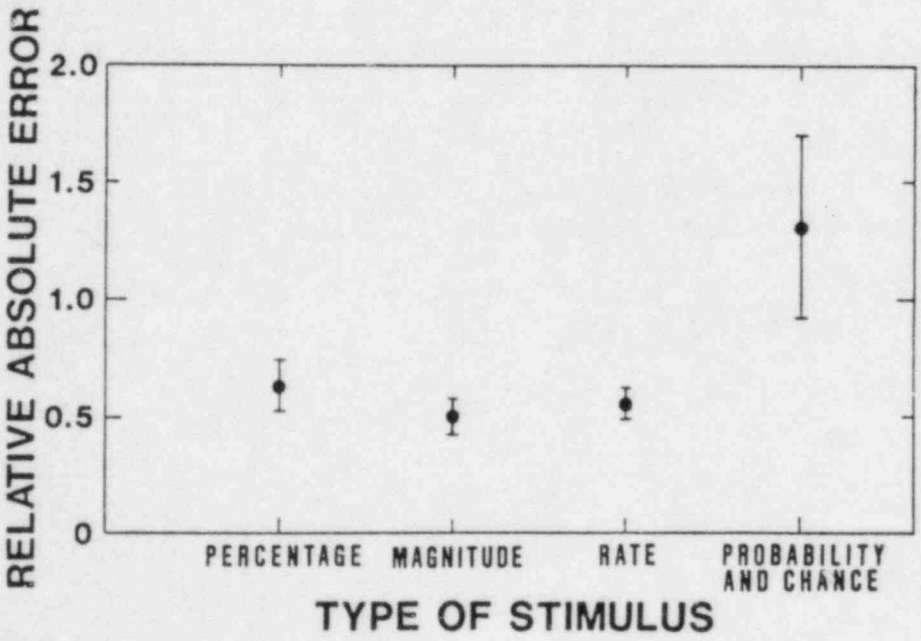

Figure 2. The relative absolute error for each type of stimulus.

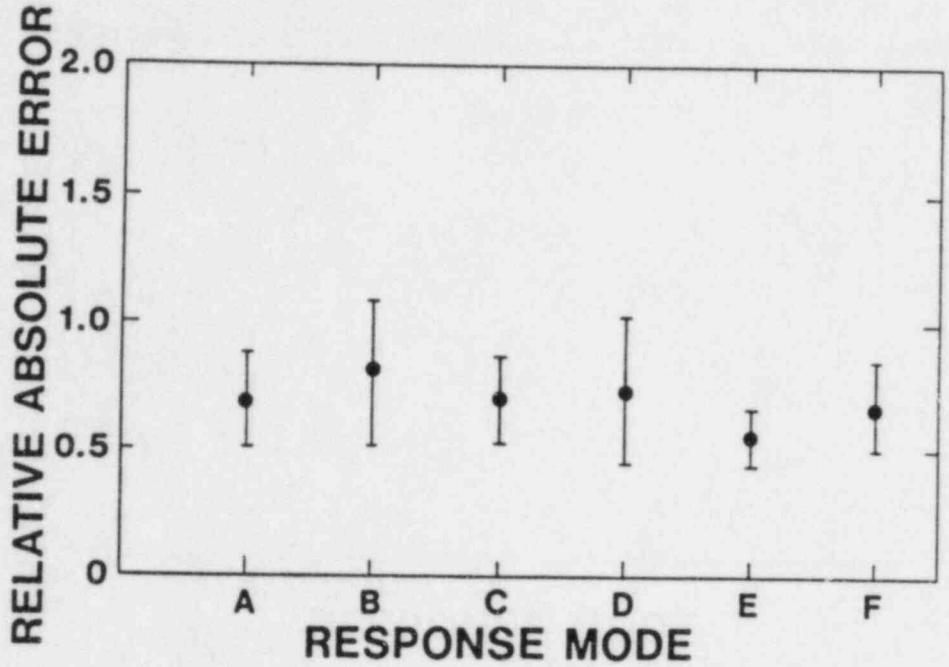

Figure 3. The relative absolute error for each response mode.

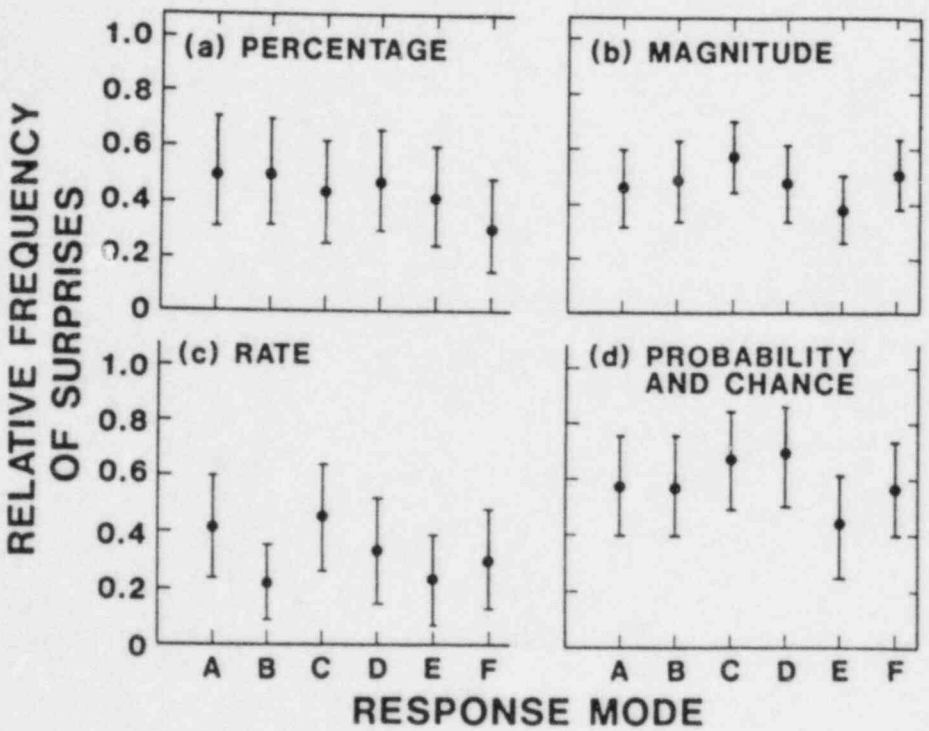

Figure 4. The relative frequency of surprises for each response mode for each type of stimulus.

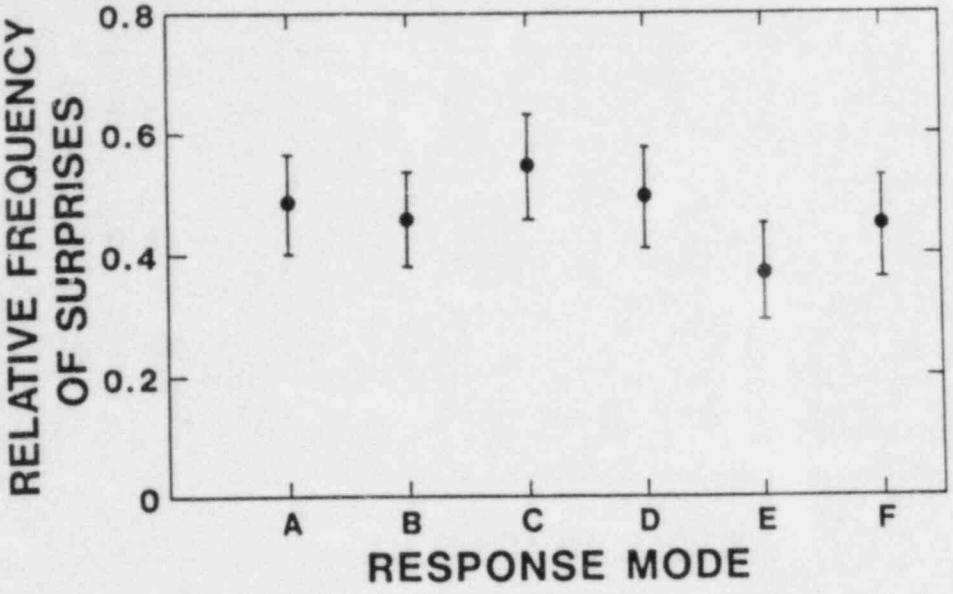

Figure 5. The relative frequency of surprises for each response mode.

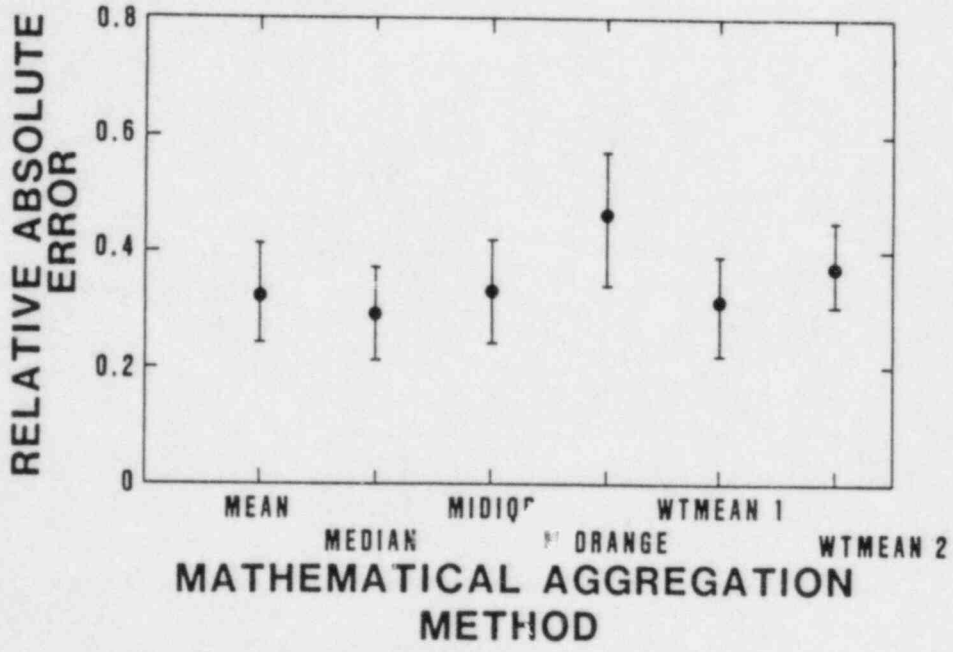

Figure 6. The relative absolute error for each mathematical aggregation method.

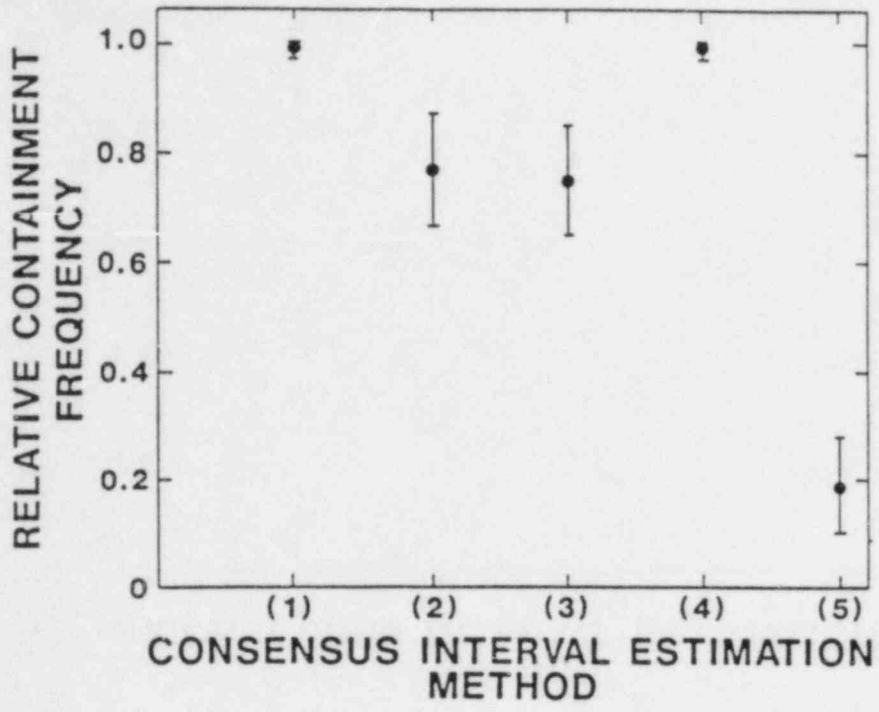

Figure 7. The relative containment frequency for each consensus interval estimation method.

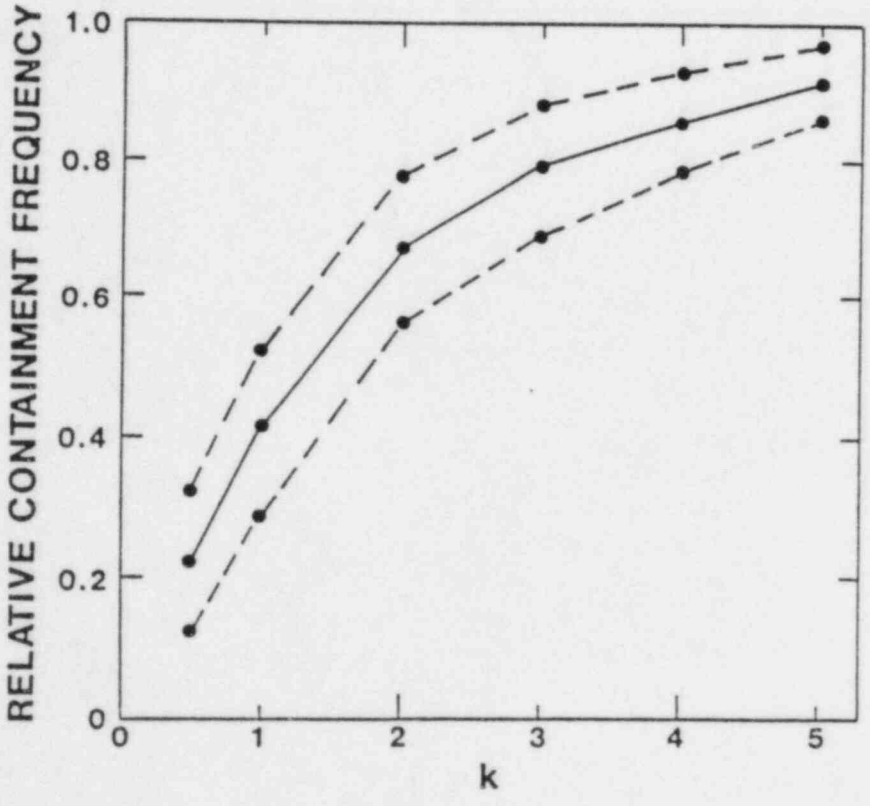

Figure 8. The relative containment frequency plotted as a function of the number of standard errors, k.

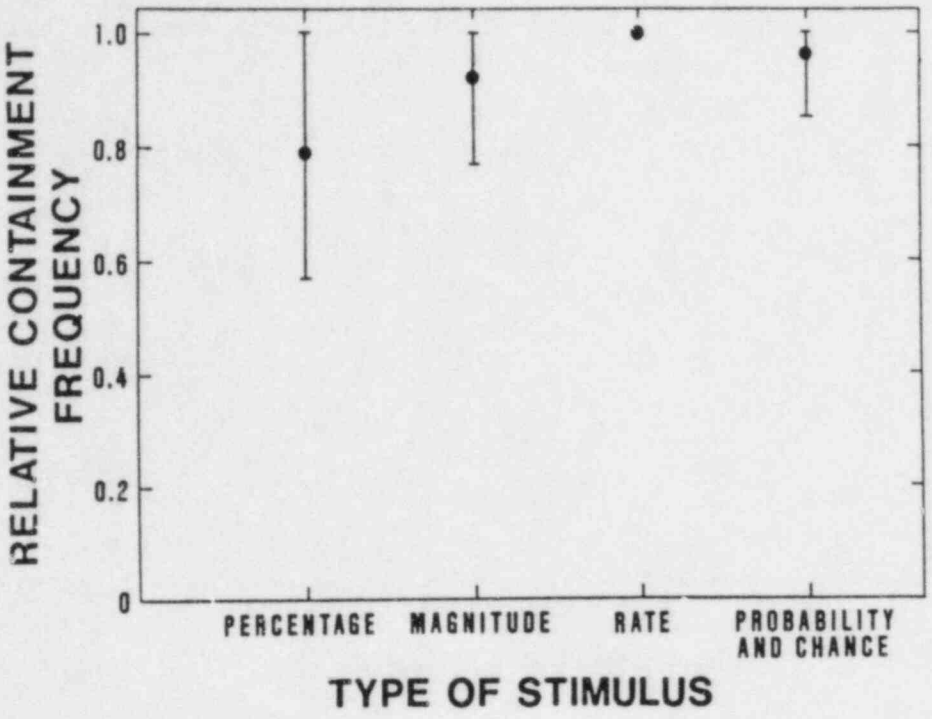

Figure 9. The relative containment frequency for each type of stimulus.

HUMAN FACTORS AFFECTING SUBJECTIVE JUDGMENTS

Mary A. Meyer
Los Alamos National Laboratory

ABSTRACT

Human factors include the ways in which people acquire, process, and convey information. They affect the quality of people's judgements and thus become a concern when these judgments are being elicited for use as data. This paper focuses on five human factors: question phrasing, conservatism, inconsistency, overoptimism, and social pressures. Techniques for detecting and reducing the occurrence of these human factors are given for two methods of eliciting subjective data, the mail survey and the interactive group method. Techniques for structuring the elicitation methods are proposed as the main means for countering the occurrence of human factors.

THE HUMAN FACTORS

Human factors can affect the quality of the subjective data in many ways. Human factors include the ways in which people acquire, remember, process, and present information that inhibit their reaching mathematically optimal decisions. The human acquisition of data is biased because humans selectively learn that which supports, rather than opposes, their views (Mahoney 1976, Hogarth 1980). For example, people are unconsciously drawn to acquire information which supports, rather than refutes, their preconceptions (Mahoney 1976). Then too, people can acquire faulty information because of the role that feedback plays in the learning process. When people receive no feedback, delayed, or only partial feedback, as often occurs, they may draw incorrect conclusions (Hogarth 1980). For example, scientists who often receive only partial confirmation of their hypotheses are likely to consider this sufficient validation or to believe those data points which support their theory and mentally dismiss the others (Mahoney 1976). The information acquired is stored and may be later accessed by the person during an elicitation session.

How easily such information can be accessed from memory also affects peoples' judgments during an elicitation session. Concrete, catastropic, or widely publicized information is more easily accessible and thus more greatly influences a person's judgment than less memorable information (Spetzler and Stael von Holstein 1975, Hogarth 1980). For example, it is thought that the League of Women Voters ranked the nuclear industry as posing the greatest occupational hazards to its employees of any industry because of the disproportionate amount of media coverage this industry had received.

The processing of data in the human mind, such as during an elicitation session, is also subject to human factors. Generally, people have difficulty processing more than seven pieces of information at a time (Miller 1956). Typically, they will select a heuristic for solving a problem in a decision situation which then influences the decision they reach. For example, managers may focus on the major aspects of the problem and ignore the uncertainties and complex interactions of factors to reach a decision (Bender et al., 1981). This simplifying heuristic may point to a different decision than one which had included all the complexities of the problem. In applying these heuristics, people are likely to be inconsistent, thus further complicating the gathering of quality subjective data. For example, the manager may have been forecasting the completion date of a large project by adding together the blocks of time that each major phase was likely to require. He may have forgotten to add in a phase being done by a subcontractor, thus failing to consistently follow his own heuristic.

Additional complications may enter as a result of the mode in which participants are requested to give the judgments. For example, respondents may give different judgments on a survey than they would in an interview situation (Payne 1951). They might give varying judgments to different phrasings of the same question (Payne 1951, Sudman and Bradburn 1982, Gorden 1980). Then too, they might give different judgments if they are giving it in "willingness to gamble" or "probability" schemes (Winkler 1967, Hogarth 1980).

Due to the constraints of time, five human factors were selected for discussion. These five factors are widely prevalent and often interrelated as will be described below. The five human factors include the effects of:
1) Presentation of the decision task and phrasing of the questions or response options;
2) Conservatism;
3) Inconsistency;
4) Overoptimism and;
5) Social pressure.

Evidence of the effect of the presentation of the decision task on the individual's response has been documented by Tversky and Kahnemen (1981). They asked students which alternatives they preferred in gain and loss situations. For example, students chose between: 1) a sure gain of $250; and 2) a 25% chance of gaining $1000 or a 75% chance of gaining nothing. In the set of loss alternatives, they chose between; 1) a sure loss of $750; and 2) a 75% chance to lose $1000 or a 25% chance to lose nothing. The majority preferred the sure gain in the first pair of options and the risky loss in the second pair. Thus, the relative attractiveness of options varies when the same decision is framed in different ways. Furthermore, individuals are generally unaware of the effect of question framing and, if informed of it, uncertain of how to compensate for its effect.

In addition, there is evidence that the response mode, such as probabilities or equivalent gambles, influence peoples' judgment (Winkler 1967, Hogarth 1980). For example, Winkler (1967) recommended that a "willingness to pay" response mode be used because people gave more conservative, hence more realistic, estimates using this response mode than using probabilities. Similarly, the scales used for the responses, such as 1 to 10 or -5 to +5, can influence peoples' judgments.

The effect of question phrasing has been shown most dramatically by Payne (1951) through his use of the split ballot technique in survey questions. The split ballot technique entails giving half of a survey sample one wording of a question or response option and the other, another. For example, one wording of a question might be, "Do you believe that X event will occur by Y time?" The other wording might be, "Do you believe that X event will occur by Y time, or not?" This second option is more balanced because it mentions both possibilities. For this reason it would be likely to receive a higher percentage of "no" responses. Often the difference measured by the split ballot technique is 4-15% even when the rewording has been very slight.

Conservatism, or anchoring bias, involves the individual's tendency to cling to their first judgment instead of adjusting it to reflect new information. Sometimes this tendency is explained in terms of Bayes' Theorum as the failure to adjust a judgment in light of new information as much as it would be according to Bayes' mathematical formula. Spetzler and Stael von Holstein (1972) and Armstrong (1981) describe how people tend to anchor to their initial response, using it as the basis for later responses. For example, the subject may use the last year's sales as a starting point in predicting this year's sales and fail to consider other points on this distribution independently from this starting point. In addition, Ascher (1978) finds this problem to exist in forecasting where panel members tend to anchor to past or present trends in their projection of future trends. Ascher determined that one of the major sources of inaccuracy in forcasting future possibilities, such as markets for utilities, was the extrapolation from old patterns that no longer represented the emerging or future patterns.

Inconsistency occurs when individuals give contradictory judgments. For example, they might give item A a higher rating than B with respect to goal X, B a higher rating than C, and C a higher rating than A. Inconsistency is a common problem because, as mentioned earlier, individuals are generally unable to apply a consistent strategy, or heuristic, to a series of cases (Hogarth 1980). Inconsistency in an individual's judgment can also stem from his remembering or forgetting information during the process of the elicitation session. For example, the individual may remember some of the less spectacular pieces of information and consider these in making judgments later in the session. Or, the individual may forget that particular ratings are only to be given in extreme cases and begin to give them more freely towards the end of a session than at the beginning.

Overoptimism is sometimes referred to as the overestimation of probabilities, overconfidence bias, or the underestimation of uncertainty. Overoptimism is the giving of more optimistic judgments, such as in the form of probabilities, than the person's data warrants. People tend to be overly optimistic of the probability of some event occuring and often underestimate the uncertainty, or the time and resources needed to make this event a reality. Thus, they give too narrow of error bars on these judgments (Capen 1975). Overoptimism can stem from a variety of causes: 1) thinking at too general a level; 2) wishful thinking; and 3) illusion of control. Armstrong (1975) and Hayes-Roth (1980) have shown that people give higher, less realistic, probabilities when they consider decision tasks in general than when they disaggregate them into their component parts. For example, Armstrong (1975) asked straight Almanac questions of one half of his sample. Of the other half, he asked the same Almanac questions but broken into logical parts. For instance, the question "How many

families were living in the U.S. in 1970?" was asked as "What was the population of the U.S. in 1970?" and "How many people were there in the average family then?". The persons answering the disaggregated questions gve significantly more accurate judgments.

Wishful thinking occurs when an estimator's hopes influence his judgment (Hogarth 1980). For example, a project manager in charge of a project may give optimistic probabilities about completing it on schedule because he hopes this will be the case. In general, people exhibit wishful thinking about what they can exhibit in a given amount of time--They overestimate their productivity (Hayes-Roth 1980).

Illusion of control is the tendency to feel greater optimism or greater confidence in some outcome, if one has been involved in its process (Hogarth 1980). People can acquire the impression of having more control over outcomes simply by spending time analyzing a situation as in a elicitation session (Langer 1975). Similarly, people perceive risks as being lower when they feel that they are in control of a process. For example, people perceive less risk when they are driving a car than when they are riding, as a passenger, in a plane (Rowe 1982).

Social pressure induces individuals to slant their responses or to silently acquiese to what they believe will be acceptable to their group, superordinates, institution, or society in general. Zimbardo, a psychologist, explains that it is due to the basic needs of people to be loved, respected, and recognized that they can be induced or choose to behave in a manner which will bring them affirmation (1983). There is abundant sociological evidence of conformity within groups (Weissenberg 1971). Generally, individuals in groups conform to a greater degree if they have a strong desire to remain a member, if they are satisfied with the group, if the group is cohesive, and if they are not a natural leader in the group. Furthermore, the individuals are generally unaware that they have modified their judgment to be in agreement with the group. One mechanism for this unconscious modification of opinion is explained by the theory of cognitive dissonance. Cognitive dissonance occurs when an individual finds a discrepancy between thoughts he holds or between his beliefs and his actions (Festinger 1957). For example, if an individual holds an opinion which is conflict with that of the other group members and he has a high opinion of the other's intelligence, cognitive dissonance will result. Often, the individual's means of resolving the discrepancy is by unconsciously changing his judgment to be in agreement with that of the group (Baron and Byrne 1981).

Irving Janis's study of fiascos in American foreign policy (1972) illustrates how presidential advisors often silently acquiese rather than critically examine what they believe to be the group's opinion. This tendencey has been called "group think", the "bandwagon tendency", or the "follow-the-leader effect."

The effect of social pressure can also be seen in situations where the individual is not in direct contact with others. Payne (1951) has provided evidence that people give socially acceptable answers to survey questions. On surveys, people claim that their educations, salaries, and job titles are better than they are. More people claim subscriptions to socially acceptable magazines and deny it to the lurid ones than subscription records support. Often there is

a 10% difference between what is claimed for "prestige" reasons and what objectively is.

THE METHODS

Methods for eliciting expert opinion vary along several continuums: 1) the number of participants; 2) the degree of interaction among participants and between them and the session leader; 3) the degree of structure imposed on the elicitation process; 4) the degree of participants' expertise; and 5) the degree of "fuzziness" of the data being elicited.

For example, one method, the mail survey, involves many respondents but little interaction among respondents or between them and an interviewer. Interaction is defined as any two-way communication after which the respondent is allowed to change his judgment. When the respondent fills out a survey, there is generally no interaction between him and his peers or between him and an interviewer.

Another possibility, the Delphi method, can include any number of respondents and allow for more interaction between respondents than the traditional mail survey. The respondents' interactions are controlled by the Delphi monitor who sends each respondent the judgments of the others. The respondents are allowed to adjust their judgments in light of this information. The process of allowing respondents to change their judgments can go through any number of iterations even until consensus is reached. RAND corporation developed the Delphi method to overcome some of the problems inherent in an interactive group method, such as social pressures to conformity. For this reason, in the Delphi technique, the respondents do not interact in a face-to-face situation. Instead, the only contact they are supposed to have with one another is via the mail. And then, the names and other identifying features are removed from the judgments before they are circulated so that the origins of these judgments will not unduly affect the recipients.

Another method, the face-to-face interview, usually involves a fewer number of respondents than the mail survey. The respondents are interactive, singly, with the inteviewer during the course of the interview.

Fourthly, there is a interactive group method. In this method, a group of three or more may be convened to give their judgments in the presence of one another. The group sessions are generally monitored and structured by a session leader. For example, the leader may encourage group members to write down their judgments and their reasoning. The leader may require that this information be presented to the group and that a discussion follow. The interactive group method can go through any number of iterations, as in the Delphi method, until consensus, if it is desired, is reached.

For the sake of brevity, this paper will confine its discussion of the detection and reduction of the human factors to two methods, the traditional mail survey and the interactive group method. These two methods were selected because they lie on opposite ends of the continuum with respect to the number of participants and the degree of interaction involved.

The five human factors are manifested in different ways in the various methods so the means by which they can be detected or reduced also vary. For

example, the effect of social pressure is manifested more strongly in the interactive methods such as the face-to-face interview and the interactive group method. Yet, because these methods are interactive, much of the detection of social pressure can be done by a trained observer. This paper's approach to the detection and reduction of human factors in elicitation methods is likely to reflect the orientation of a cognitive or social scientist. The approach is to perform a real time detection or counteraction of the human factors as they occur during a session rather than a later mathematical adjustment of the data.

This paper advocates a structuring of the elicitation methods as a means for reducing the occurrence of human factors. Structuring an elicitation method involves controlling interactions, identifying the parts of the phenomenon on which the respondents are being questioned, defining them and the response options, such as the scale. For example, an unstructured interactive group method would resemble the usual meeting which occurs in the business world. A structured version of the same method would have a program for when each member would present his judgment and rationale to the group, when the floor was open for discussion, and when the next round could begin. In general, the greater the degree of structure imposed on the decision process, the simpler it is to control for the occurence of human factors. Often a method cannot be maximumly structured because each degree of structure imposed slows the process and requires more patience or cooperation on the part of the participants. The client may have deadlines and a fixed budget which limit the amount of structuring which can be done. Thus, the amount of structuring which can be done often involves tradeoffs between the quality of the data and its cost in time and manpower.

The Mail Survey

Detection of Human Factors

In a survey, the occurrence of human factors is not generally detected while the individual is making his judgment but earlier during pilot tests or later when the survey is analysed. Three factors, the effects of question phrasing, social pressure, and inconsistency, can be detected by the use of the split ballot, the sleeper option, and pilot test.

The effects of question wording and sequencing of options can be detected by measuring the differences between the split ballot questions. The split ballot technique is most commonly used for "yes-no" and other multiple choice questions. Use of split ballot techniques in the past (Payne 1951) have shown that people favor generally worded options over those which are highly specific. In addition, they favor options which refer to the status quo over those proposing new alternatives. Split ballot results have also shown that people favor selecting numerical options which are located in the middle of a series whereas they favor nonnumeric options which are located on either end of the series.

Social pressures to give the most acceptable response can also be detected by use of the split ballot technique. One wording on half the surveys can state the options bluntly, the other can contain face saving phrases to encourage people to check the response which is most descriptive of their thoughts or actions. A face-saving option often encourages the respondent to admitt that he does not have X knowledge or Y socially-desirable possession at this time by allowing him to state that he plans to acquire them in the future.

Another common area for the effects of social pressures to emerge is in peoples' unwillingness to admitt ignorance, to check the "I don't know" option. If identification of knowledgable respondents is important, a different technique can be used to get a better indication of people's knowledge than simply totalling those who selected the "Don't know" response. A "sleeper" option that sounds plausible but which does not exist in reality can be inserted into the series of bonafide options. For example, on a survey of public opinion of nuclear reactors a "fast water reactor" might be inserted between a "light water", and a "breeder." The number of people who select the sleeper option can be added to those who marked the "Don't know" option and excluded from the pool of supposedly knowledgable respondents.

Inconsistency in people's responses to surveys is more difficult to detect than the two above mentioned effects. Inconsistency could conceivably be detected by the use of redundant questions but this approach poses problems. If the redundant question is an exact repitition, it can annoy people because they wonder why they are being asked the same question, again. Yet, if the question is asked with a new wording, respondents may give different answers simply because of the difference in phrasing. Inconsistency can occur because the individual has not applied his heuristic consistently, has forgotten instructions or definitions, or has remembered different incidents as he progressed through the survey. An intensive interview type of pilot test can be used to check the survey instrument for these problems. For example, one set of these pilot tests revealed that individuals had forgotten the instructions about half way through the selection of many options. The respondents were supposed to mark their areas of knowledge on a list spanning two pages. Instead by the second page, one fifth of the pilot sample had checked areas in which they would have liked to have had knowledge.

This type of pilot test is the only one, to my knowledge, that can be used to tack peoples' thinking, their consistency, through a survey. I adapted several ethnographic interviewing techniques to create this pilot test method. These techniques gather two types of information: 1) how the respondent progresses through the survey, that is which sections he looks at, in what order, and for how long, his general impressions, and when or why he decides to fill out the survey and to turn it in; and 2) how the respondent specifically interprets each direction, question, and response option.

To obtain the first type of information, the interviewee is asked to handle the survey as he would naturally, if no observer were present. The interviewee is asked to "think outloud" and to mention his impressions. Generally, individuals will skim the cover letter and flip through the rest of the survey. As the individual flips through the survey he might state, "I have problems with this page and I would probably let the survey sit on my desk for several days to decide whether to fill it out. While the interviewee pages through the survey, his pauses and gestures, particularly those indicating confusion or anxiety are noted by the monitor. If the respondent has paused or shown some emotion during his review of a particular section, specific questions will be asked such as, "What was your feeling when you read this?".

To obtain the second type of information, the respondent is asked to paraphrase, in his own words, the meaning of each direction, question, and response option. This information allows the monitor to track the respondent's interpretation of each part of the survey.

Structuring the Method to Reduce the Occurrence of Human Factors

As mentioned earlier, structuring any elicitation method can facilitate the counteraction of many human factors. The following section contains some recommendations on how to set up a mail survey to obtain better quality subjective data by controlling for the intrudence of some human factors.

The first stage in developing the mail survey can have an effect on the amount of inconsistency which shows up later in the respondents' judgments. Often seeming inconsistencies in the respondents' answers arise from their viewing the phenomena in a different manner than the way in which it has been presented on the survey. Because the survey does not generally encourage them to explain the view or assumption which allowed them to make the puzzling responses, their responses are dismissed as inconsistent and unreliable. For this reason, it is recommended that the creator of the survey first talk extensively to a sample of those who will be surveyed to learn what relationships, causes and effects, they believe enter into the problem. For example, respondents from a utility might believe that the future of their utilities market is tied to the nation's gross national product (GNP). If the task is to elicit their projections for a utilities market in year 2000, then the questions should define different levels of GNP. For instance, "Assuming that the GNP is X in the year 2000, what would you predict the market for Y to be?"

Careful composition of the questions can reduce the occurrence of three effects: 1) inconsistencies which arise from the respondents' confusion, 2) phrasing, and 3) social pressure. The use of Basic English is recommended if the survey is targeted for the general public as one means for minimizing misunderstandings. Basic English is a vocabulary of approximately 1000 words that are understood by most people who possess a high school education. Payne (1951) provides a list of these words. He also provides a list of words which have been found to possess different meanings for different people. For example, "this year" means the present fiscal year to some, the present calendar year to others, and this coming year to still others. It is recommended that the use of these problem words or phrases be avoided in the interests of clarity. In addition, it is recommended that question lengths not exceed 25 words because respondents' comprehension has been found to fall off around that point (Payne 1951).

As mentioned earlier, the split ballot techniques can be used to detect or counteract the effect of phrasing and ordinality. For example, response options can be placed first or last in half the surveys and in the middle in the other half to counter the effect of ordinality.

If the pilot test of the survey indicated that prestige was on issue on some questions, then face-saving wordings can be used to obtain a better representation of peoples' opinions. Generally, admission of ignorance involves the loss of prestige, so the "Don't know" option should be carefully worded. "No set opinion at this time" is an example of a face-saving wording.

The presence and placement of definitions is another technique which can be employed to reduce the occurrence of human factors, in this case, inconsistency. Definitions include descriptions of the phenomena, the time frame in which the respondent is to consider these, and the scale in which he is to respond. As an individual progresses through a survey, the definitions becomes blurred in his

mind. He relies on his memory of these definitions and often arrives at a working definition which deviates from the original written one. For this reason, definitions should be incorporated into the question or they should immediately proceed it. For example, "What is the probability that the motor generator will reach a maximum power of X for Y amount of time by calendar year September 1, 1984?" The definition of the phenomena has been mentioned as part of the question. The same treatment can be extended to the response scale. For example, the Sherman Kent scale gives these descriptors, "nearly certain", "highly probably", and "We are convinced", to describe a percent ranging from 90 to 99. Both numbers and verbal descriptors, or definitions, are used in attempt to make people mean approximately the same thing when they give the same rating.

Another structuring technique, hierarchically organizing the survey, is helpful in countering the respondents' tendencies to conservatism and overoptimism (Meyer 1982a). Organizing the survey in a hierarchical manner generally entails beginning with specific questions and progressing to more inclusive questions. The respondent is not asked major questions until his memory has been prodded to remember more than just the easily accessible information. Thus, his judgment is not as likely to be anchored to just the first remembered bits of data. Using the hierarchical structure also involves disaggregating questions, as shown in the Almanac example, to counter peoples' tendency toward overoptimism.

The Interactive Group Method

Detection of Human Factors

The effects of phrasing, conservatism, inconsistency, and social pressure can be detected during elicitation sessions by the trained observer who is monitoring this process (Meyer 1982b). Generally, only the presence of these effects, not their magnitude, can be detected by this means. This mode of detection assumes that the group members have been instructed to "think outloud" in interpreting the questions and giving their judgments. (More details on the group members' verbalization of their thoughts will be given in the next section.)

The respondent's verbal feedback on their interpretations of questions allows misunderstandings to be caught during the sessions. Conservatism can also be detected during the session. If an individual continuously holds to his initial judgment, even though there has been a discussion and an opportunity to revise his judgment, he is a likely candidate for conservatism. Inconsistency can be detected when members rate an element differently than they did a comparable one earlier or when their interpretation of a definition appears to change.

The problem of inconsistency arises from more sources in the interactive group method than in the face-to-face interview or the mail survey. This is because the group meetings are held many times whereas the others tend to be one-time deals. Thus, with the usual group method, there is the chance of the members forgetting information, instruction, and definitions over the course of time. One inconsistency which can emerge is the ease with which a response option is applied. For example, the respondents may select the extremes of the scale with varying frequency through time. In general, fatigue during a session seems to contribute to the occurrence of inconsistencies, perhaps because people

are not thinking as carefully. (Fatigue is indicated by briefer responses and by the degree of the participants' horizontal inclination.)

The degree of inconsistency can be detected by use of Bayesian-based scoring and ranking techniques. The group members' judgments can be entered into a scoring and ranking program, such as that of Saaty's Analytical Hierarchical Process, to obtain a rating of their consistency (Saaty 1980).

Social pressures can also be detected by real-time observations. Generally, if consensus is easily obtained, no difference of opinion is voiced, and the group members appear to defer to another member of the group, group think is a strong possibility. Social pressures can come from the members of the group or from the institution sponsoring the decision session. The institution may favor a particular decision outcome and apply pressure on the group members to this end.

Structuring the Method to Reduce the Occurrence of Human Factors

The first stage of the interactive group method, a free association exercise, can be used to counteract the members' tendency toward conservatism. The free association exercise involves having group members mention any and all elements which might have bearing on the phenomena in question. For example, in considering a problem on which technologies should be exported from the United States, some of the major elements a free association might have produced would be the military, economic, political, and technological significance of the export items. The elements mentioned during a free association are usually recorded for the group members to see. Later, the group members will work from these in developing a model of the decision situation. The purpose of the free association exercise is to start with a wide set of possibilities and to narrow these to the pertinent ones. The free association exercise is to counter the human tendency to anchor narrowly on past or present cases which may not hold in the future.

The next stage, the organization of these elements into a model, has bearing on how much inconsistency will be observed when the members are giving their judgments. Highly inconsistent judgments (as determined by ear and by Bayesian techniques) often indicate a need to restructure the model to better represent the members' view. This stage of the method is the most time consuming because the particpants are not always conscious of how they mentally model the phenomena. Then too, sometimes they are so conscious of some information that they fail to convey it for incorporation into the model.

The elicitation phase can be structured to include various techniques for countering the effects of social pressure, conservatism, and overoptimism. Perhaps, the most critical of all of the structures placed on the elicitation process is the requirement that participants verbalized their judgments and their reasons for giving such judgments. As mentioned earlier, this verbal feedback allows the method to be monitored for the intrusion of many human factors. For example, if group members appeared to exhibit group think, the method can be structured to promote the opposite bias, conservatism. Groups where conformity is likely to be a problem are cohesive groups, groups where the people have worked togeather before, or groups where there is a dominating leader (Janis 1972). By requiring group members to write down and then report on their judgments and rationale, they are more likely to get attached to their

judgments and defend them when the discussion begins. I would recommend having each person record and read his judgments before opening the floor to discusssion and allowing people to modify their judgements. If there is a strong official or even a natural exoffio leader in the group, that individual should be asked to give his judgments last so as not to influence the other group members. In addition, if there is an official leader of the group, he or she should be encouraged to be nondirective during the meetings. An explanation of the group think phenomena usually suffices to convince them that better discussions and data will result from their refraining from "leading."

If on the other hand, group members appear to be too narrow, or anchoring, in their thinking, a series of extreme scenarios can be introduced for their consideration.

If overoptimism has been detected, the group members can be lead to think in greater detail about the elements of the phenomena. This is done in much the way that the Almanac questions were disaggregated for the survey population.

Another technique, the reviewing of definitions, can help reduce respondents' tendency to be inconsistent because of faulty memory. If at the beginning of every session, definitions are verbally reviewed, members will be more consistent in their definitions through time and between themselves. In addition, each time that their judgment is requested, a statement of the question inclusive of definitions, can be given. For example, "What rating would you give to the importance of element X over Y to reaching goal Z?" Their copy of the scale, in this case a Saaty Pairwise Comparison, should include descriptors or definitions of the ratings.

Another technique for reducing inconsistency is to have the group members monitor their own consistency. For this task, they should have copies in front of them of their judgments, and response scale. A matrix structure of the critieria on which the elements are being judged, the elements, and the judgments work well for this task (Meyer 1982b). Often the group members will view an element in a different light than they did earlier and wish to change the earlier judgment to be in line with their current thinking. If their reasoning does not violate the logic of the model or of the definitions, they should be allowed to make the change. Sometimes, consideration of a new element makes them aware that the model and accompanying definitions did not realistically protray this part of the phenomena. Parts of the original model will need to be changed and some of the process of giving judgments will need to be repeated.

REFERENCES

Ascher, William (1978), "Forecasting: An Appraisal for Policymakers and Planners," Baltimore: John Hopkins University Press.

Armstrong, J.S. Denniston, W.B. Jr., and Gordon, M.M., (1975), "Use of the Decomposition Principle in Making Judgments," Organizational Behavior and Human Performance, 14, 257-263.

Armstrong, J.S. (1981), "Long-Range Forecasting: From Crystal Ball to Computer," New York, New York: Wiley-Interscience.

Baron, Robert A. and Byrne, Donn (1981), "Social Psychology: Understanding Human Interaction," Boston: Allyn and Bacon.

Bender, Paul S., Northup, William D., and Shapiro, Jeremy F. (1981), "Practical Modeling for Resource Management," Harvard Business Review, March-April, 163-173.

Capen, E.C. (1975), "The Difficulty of Assessing Uncertainty," Society of Petroleum Engineers AIME 50th annual Fall Technical Conference, Dallas, Texas, September 28-October 1, Paper SPE 5579.

Festinger, Leon (1957), "A Theory of Cognitive Dissonance," Palo Alto, California: Standford University Press.

Gordon, Raymond (1980), "Interviewing: Strategy, Techniques, and Tactics, "Homewood, Illinois: Irwin-Dorsey Limited.

Hayes-Roth, Barbara (1980), "Estimation of Time Requirements During Planning: Interactions Between Motivation and Cognition," Rand report N-1581-ONR, November.

Hogarth, Robin (1980), "Judgment and Choice: The Psychology of Decision," Chicago, Illinois: Wiley-Interscience.

Janis, I.L. (1972), "Victims of Group Think: a Psychological Study of Foreign Policy Decisions and Fiascos," Boston: Houghton Mifflin.

Langer, E.J. (1975), "The Illusion of Control," Journal of Personality and Social Psychology, 32, 311-328.

Mahoney, Michael (1976), "The Scientist as Subject: The Psychological Imperative," Cambridge, Massechusetts: Ballinger Publishing.

Meyer, M.A., Peaslee, A.T., Jr., and Booker, J.M. (1982), "A Data-Gathering Method For Use in Modeling Energy Research, Development, and Demonstration Programs," Energy Programs, Policy, and Economics, Florida: Butterworth Publishers.

Meyer, M.A., Peaslee, A.T., Jr., and Booker, J.M. (1982), "Group Consensus Methods and Results," Los Alamos National Laboratory, report LA-9584-MS.

Miller, George A. (1956), "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," Psychological Review, 63, 81-97.

Payne, Stanley (1951), "The Art of Asking Questions," Princeton, New Jersey: Princeton University Press.

Rowe, William D. (1977), "An Anatomy of Risk," New York: Wiley-Interscience.
Saaty, Thomas L. (1980), The Analytic Hierarchy Process: Planning, Priority Setting, and Resource Allocation," New York: McGraw-Hill.

Spetzler, C.A. and Stael von Holstein, C-A. (1975), "Probability Encoding in Decision Analysis," Journal of Institutional Management Science, 22, 340-58, November.

Sudman, Seymour and Bradburn, Norman M. (1982), "Asking Questions: A Practical Quide to Questionnaire Design," San Francisco: Jossey Bass.

Tversky, Amos and Kahneman, Daniel (1981), "Framing of Decisions and the Psychology of Choice," Science, 211, 453-58.

Weissenberg, Peter (1971), "Introduction to Organizational Behavior: A Behavioral Science Approach to Understanding Organizations," Scranton, Ohio: Intext Educational Publishers.

Winkler, Robert L. (1967), "The Quantification fo Judgment: Some Methodological Suggestions," Journal of the American Statistical Association, 320, 1105-1120.

Zimbardo, Philip G. (1983), "To Control a Mind," Standford Magazine, Winter, 59-64.

Special Topical Session

Health Effects of Energy Technologies

THE HANFORD STUDY -- A REVIEW OF ITS LIMITATIONS AND CONTROVERSIAL CONCLUSIONS

Ethel S. Gilbert
Pacific Northwest Laboratory

ABSTRACT

The Hanford data set has attracted attention primarily because of analyses conducted by Mancuso, Stewart, and Kneale (MSK). These investigators claim that the Hanford data provide evidence that our current estimates of cancer mortality resulting from radiation exposure are too low, and advocate replacing estimates based on populations exposed at relatively high doses (such as the Japanese atomic bomb survivors) with estimates based on the Hanford data. In this paper, it is shown that the only evidence of association of radiation exposure and mortality provided by the Hanford data is a small excess of multiple myeloma, and that this data set is not adequate for reliable risk estimation. It is demonstrated that confidence limits for risk estimates are very wide, and that the data are not adequate to differentiate among models. The more recent MSK analyses, which claim to provide adequate models and risk estimates, are critiqued.

THE HANFORD DATA

The Hanford plant, located in southeastern Washington State, has employed many workers in jobs involving some exposure to radiation. The initial purpose of this plant was the manufacture, chemical separation, and purification of plutonium. In addition research of a diverse nature and, more recently, power generation have been conducted at the facility.

The Hanford data base includes demographic data and employment histories, which have been obtained from employment records, as well as mortality and radiation exposure data. Mortality data, which is obtained primarily from the Social Security Administration, includes date and cause of death, while the exposure data, which is obtained through the use of personal dosimeters, consists of annual estimates of external exposure to radiation. A more detailed description of the Hanford data base is found in Gilbert and Marks (1979).

Exposures of Hanford workers have been deliberately limited as a protection to the worker with the result that exposures are far lower than those received by the Japanese atomic bomb survivors and other populations currently providing the data that form the basis for risk estimation. The fact that exposures are low can be regarded as a strength in that exposures are at the levels of actual interest, thus providing a direct method of assessing risks from low level exposure to radiation for medical, occupational or other reasons. However, the limited exposure is also a weakness in that it severely limits the potential of the study. In fact, if current estimates of radiation effects are correct, it is highly unlikely that statistically detectable effects can be identified in a population such as Hanford, and such data are not adequate for reliable estimation of radiation effects. A principal rationale for studying groups that have been exposed occupationally and environmentally is the investigation of the hypothesis that effects might be many times larger than current estimates would suggest. That is, such studies serve as a rough check on estimates and models that have been developed from other sources.

The Hanford data has been analyzed by several investigators including Mancuso et al. (1978), Hutchison et al. (1979) and Darby and Reissland (1981). In this section we describe analyses that have been conducted at Pacific Northwest Laboratory (PNL). Additional detail regarding these analyses is given in Gilbert and Marks (1979) and Gilbert (1984*).

METHODS FOR ANALYZING THE HANFORD DATA

The procedures that have been used for PNL analyses of the Hanford data can be derived from the proportional hazards model of Cox (1972). Initially PNL analyses emphasized tests of the null hypothesis of no association of radiation exposure and mortality from several different causes, but more recently estimates and confidence limits have also been obtained.

---

*"How Much Can Be Learned From Populations Exposed to Low Levels of Radiation?" To be published in The Statistician 34 (1985).

Both test and estimation procedures are based on comparing doses of workers dying from the cause of interest to the doses of comparable workers alive, and therefore, at risk of dying at that time. "Comparable" is used loosely to mean similar with respect to age, sex, calendar year and other specified variables. For example, suppose that worker i dies at age 56 in 1967 with cumulative dose z_{i56}. We might then compare z_{i56} with the cumulative doses (up to age 56) for workers who are alive and at risk of dying at age 56 in calendar years 1965-1969.

We will call the set of doses of those workers alive at age 56 in a similar calendar year period the risk set for worker i and denote this set by R_i. We will denote the mean of the doses in R_i by μ_i and the variance by σ_i^2. A test of the null hypothesis can be obtained by comparing the observed and expected scores defined, respectively, by $Z = \sum_{i=1}^{n} z_i$ and $MU = \sum_{i=1}^{n} \mu_i$ where i indexes deaths from a particular cause and z_i is a measure of dose for worker i at the time or age of death t_i. Asymptotically, the statistic $Y = (Z - MU)/[\sum_{i=1}^{n} \sigma_i^2]^{1/2}$ will have a standard normal distribution.

The above approach can be regarded as a simple application of Cox's regression model in which the time variable is age, grouped in single year intervals (calendar year or follow-up time could also play this role), in which there is a single time-dependent regression variable, dose, and in which other variables such as calendar year are controlled through stratification. A variety of dose-response relationships can be investigated by varying the definition of dose. For example, to account for a latent period, exposures can be lagged by various intervals; that is, only exposure accumulated up to some specified period prior to t_i would be included both for the worker who dies and for those in his risk set R_i. Doses can also be replaced with scores for exposure categories resulting in analyses similar to those described by Mantel and Haenszel (1958) and by Mantel (1963). A computer program, MOX (Mortality and Occupational eXposure) (Buchanan and Gilbert 1984), is available for performing the needed computations for testing the null hypothesis for several different diseases.

Procedures for obtaining estimates and confidence limits require specification of the hazard or risk function. The model originally described by Cox (1972) was the "log-linear" model, in which the form of the dose response function is exponential, or of the form $\lambda_{kt} \exp(\beta z_{kt})$, where λ_{kt} is the spontaneous hazard for worker k at time t and z_{kt} is the cumulative dose for worker k at time t. However, risks due to radiation are usually expressed in terms of linear (or linear-quadratic) models, and such models cannot be expressed in the exponential form.

A linear form of the proportional hazards model is available and can be written $\lambda_{kt}(1 + \beta z_{kt})$. The partial log likelihood function based on this model can be written as indicated below:

$$\log L(\beta) = \sum_{i=1}^{n} \log \frac{(1 + \beta z_i)}{(1 + \beta \mu_i)}$$

where i indexes deaths and z_i and μ_i are defined as previously. This likelihood is a relatively simple function based only on the observed and expected doses associated with the workers who die of the cause of interest. Maximizing this likelihood thus requires iterating only on a relatively small set of data rather than on the several hundred thousand person-years in the full data set.

Confidence limits based on the likelihood ratio statistic can be obtained by making use of the fact that under the hypothesis that $\beta = \beta_0$, $-2 \log L(\beta_0)/L(\hat{\beta})$ will be asymptotically distributed as chi-square with one degree of freedom. This statistic tends to approach its asymptotic distribution more rapidly than the maximum likelihood estimate $\hat{\beta}$ and thus should provide more accurate confidence limits than the more direct approach using a normal approximation to $\hat{\beta}$. Because of the use of a linear model and the very skewed exposure distribution, the distribution of $\hat{\beta}$ can be expected to be quite skewed. By setting $\beta_0 = 0$, the likelihood ratio statistic can also be used as an alternative to the score statistic Y for assessing the p-values in tests of the null hypothesis of no effect.

RESULTS OF PNL ANALYSES OF THE HANFORD DATA

Until recently, PNL analyses have been aimed primarily at testing for associations of radiation exposure and and mortality from several causes. Results of these analyses have been described in detail in Gilbert and Marks (1979) and Tolley et al. (1983). Here it is noted only that of 17 cancer types tested, only multiple myeloma shows evidence of a significant correlation with radiation exposure. This correlation results from three deaths with relatively large exposures. Since the possibility that the correlation is a false positive finding cannot be ruled out, the appropriate interpretation of the multiple myeloma finding remains uncertain at this time.

Because of the limited magnitude of the exposures received by Hanford workers, estimates based on this data set are not likely to be very meaningful. However, confidence limits provide a useful way of quantifying the uncertainty in this data set. For example, it is possible for the Hanford data to show no evidence of positive correlations of radiation and various cancer types, yet be consistent with effects several times currently accepted estimates; large upper confidence limits will demonstrate this. On the other hand, claims are sometimes made that effects are magnitudes higher than standard estimates would indicate. Upper confidence limits that are less than such extreme claims serve to demonstrate that the Hanford data are inconsistent with such extreme claims.

In Table 1, estimates and 95% confidence limits are presented for several cancer types. With the exception of the estimate for leukemia, the study population upon which these estimates are based consists of 13,632 monitored male workers with at least three dosimeter readings who survived at least 10 years following their initial employment date. The analyses include deaths occurring over the period January 1, 1955 through December 31, 1978. The basic time variable is age (grouped in single year intervals) and the analyses are stratified by calendar year (five-year intervals except for the period 1975-1978 for which single year intervals are used). Exposures have been lagged for 10 years since this is thought to be the minimal latent period for most cancer types other than leukemia. The estimate for leukemia is based on

TABLE 1. Risk estimates and 95% confidence limits for several cancer types.

Cancer Type	Estimate ($\hat{\beta}$)	95% Confidence Limits	Number of Deaths
Leukemia[a] (205-7)	-3.3% per rem	(Negative, 8.7% per rem)	12
All M.N. except[b] leukemia, bone, skin, prostate	-0.0% per rem	(Negative, 3.0% per rem)	433
M.N. of digestive system (150-159)	-0.5% per rem	(Negative, 5.3% per rem)	147
M.N. of stomach (151)	5.3% per rem	(Negative, 37% per rem)	25
M.N. of colon (153)	-7.8% per rem	(Negative, 0.5% per rem)	50
M.N. of pancreas (157)	3.5% per rem	(Negative, 28% per rem)	35
M.N. of lung (162)	0.2% pre rem	(Negative, 6.1% per rem)	153
M.N. of lymphatic and hematopoietic tissue except leukemia (200-202, 209)	4.0% per rem	(Negative, 25% per rem)	38
Multiple Myeloma (203)	96% per rem	(6.4% per rem, ∞)	8

[a]Exposures lagged for two years.

[b]For all cancer types other than leukemia, exposures are lagged for 10 years.

a lag of two years, includes deaths occurring over the period January 1, 1947 through December 31, 1978, and the study population includes monitored male workers with at least three dosimeter readings who survived at least 2 years. Because the estimates are based on a model in which the radiation risk is assumed to be proportional to spontaneous risk, these estimates are expressed as a per cent increase per rem.

The estimate for leukemia is negative, but the upper 95% confidence limit of 8.7% is several times standard estimates. Documents such as BEIR III (1980) do not usually present estimates in this form, but it is possible to determine that the BEIR III linear estimate for leukemia is about 2 to 3% per rem. Thus the upper confidence limit for leukemia is approximately 3 or 4 times the BEIR III linear estimate. The other cancer grouping for which BEIR III presents lifetime risk estimates is all cancer except leukemia, bone, prostate and skin, and here it can be determined that the BEIR III linear estimate is approximately 0.3% per rem. The estimate for this cancer grouping based on the

Hanford data is negative with an upper 95% confidence limit of 3.0%, 10 times the BEIR III estimate.

Estimates for other cancer types are also presented in Table 1 including an estimate for multiple myeloma, the one cancer type showing evidence of a significant correlation with radiation among Hanford workers. The estimate for multiple myeloma is 96% per rem, with an infinite upper 95% confidence limit (the log likelihood function approaches an asymptote). Even the lower confidence limit of 6.4% per rem is larger than the standard linear estimate for leukemia, the cause of death that has been most strongly linked with radiation in other studies. The large estimate and large lower limit reflect in part the fact that this cause of death was selected as the cancer type showing the strongest correlation in the Hanford data. If, for example, one accounted for the fact that 17 cancer types were being considered by calculating a 1 - 0.05/17 = 0.997 level confidence interval, the lower limit would be negative. Also the normal approximation may not be entirely adequate for obtaining confidence limits for an effect that results from three deaths with relatively large doses.

The analyses described above demonstrate that estimates based on the Hanford data are very unstable. Additional analyses, which are not presented here, indicate that such estimates may also be highly dependent on the model upon which the estimates are based. Furthermore, we cannot hope to address such issues as the shape of the dose-response function, the effect of variables such as age at exposure, or the manner in which radiation risks are related to spontaneous risks. Thus we must continue to place a strong degree of reliance on estimates and models derived from populations exposed at relatively high levels. However, data on populations exposed at low levels can serve as a valuable check on such estimates and models.

RECENT ANALYSES BY MANCUSO, STEWART, AND KNEALE (MSK)

Early analyses of the Hanford data by Mancuso, Stewart, and Kneale (1977) resulted in claims that Hanford workers were experiencing cancer risks due to radiation that were far greater than would be predicted based on estimates

such as found in BEIR III (1980). These analyses have been criticized by many scientists including Hutchison et al. (1979), Anderson (1978), Reissland (1978) and Gilbert and Marks (1979). Several problems with the early Mancuso, Stewart, and Kneale (MSK) analyses have been identified. Analyses of the Hanford data by other investigators (Hutchison et al. 1979, Gilbert and Marks 1979, and Darby and Reissland 1981) have failed to confirm the conclusions of MSK.

In the more recent MSK analyses, which are described in Kneale et al. (1981, 1984), the methodology has been revised considerably. Although these recent analyses avoid most of the problems for which the early papers were criticized, a number a new problems have been introduced. These recent analyses have not attracted the attention of the earlier ones, but risk estimates based on the 1981 paper have been used in workman's compensation hearings. Although there is not space here for a complete critique of these analyses, an attempt is made below to describe some of the difficulties.

We will start with a discussion of Kneale et al. (1981), since this is the paper upon which the estimates used in recent hearings have been based. The analyses in this paper are based on the Cox model with initial efforts directed toward testing the null hypothesis of no effect. Conclusions based on these analyses differ from our own (which are also based on the Cox model) primarily because of differences in the cancer categories chosen for analysis, and differences in the control variables included.

The only disease groupings selected for analysis are all causes of death combined, a group of cancers identified as "radiosensitive" cancers (pharynx, most digestive cancers, breast, lung, thyroid, lymphatic and hematopoietic), and a group including all remaining cancer types (which will be referred to here as "non-radiosensitive" cancers). MSK claim that this grouping of cancers was obtained independently of any analyses of the Hanford data since it is similar to that found in ICRP 14 (1969). However, many of the cancers that are excluded from the "radiosensitive" group are those that show negative associations with radiation exposure in the Hanford population. Thus analyses based on the "radiosensitive" group show stronger evidence of association with radiation exposure and also lead to higher estimates than would an analysis based on all cancers.

Although the group of cancers chosen does not seem entirely unreasonable, it is not one that would be universally accepted by all scientists as appropriate. Also since MSK had analyzed the Hanford data before arriving at this choice, the possibility that results of these early analyses may have subtly influenced this choice cannot be ruled out. Finally, Darby, Nakashima, and Kato (1984) have used data on the Japanese A-bomb survivors to investigate whether the "radiosensitive" cancers showed stronger evidence of association with radiation exposure than the "non-radiosensitive" cancers. They found that the relative risk estimate based on the "non-radiosensitive" cancers was actually slightly higher (although not significantly so) than the estimate based on the "radiosensitive" cancers.

We turn now to a discussion of the control variables used in the recent MSK analyses. The control variables used in the initial analyses presented in Kneale et al. (1981) are a group of variables described as "obvious factors", and which include sex, hire age, hire date, and duration of employment, with follow-up time serving as the time variable. No evidence of a significant positive correlation of radiation exposure and death from the "radiosensitive" cancers was identified, but a significant negative correlation was identified for all causes of death combined. Based on this negative correlation, MSK argue that there is evidence of selective bias in the Hanford population, and that it is therefore necessary to introduce an additional control variable to eliminate this bias.

Their choice for this task is the level of internal monitoring. In addition to being monitored for external radiation exposure, Hanford workers are also monitored for internal exposure through urinalysis (bioassay) and whole body counts. MSK define four levels of monitoring for internal depositions as follows: 1) no record of bioassays or whole body counts, 2) records of these tests but all with negative findings, 3) no record of whole body counts or internal deposition but at least one of the bioassays recorded some radioactivity (positive bioassays), and 4) either definite evidence of internal deposition or a combination of positive bioassays and whole body counts.

There are several points to be made with regard to the use of this variable. First, there is always the possibility of bias in an epidemiological study, and the general idea of trying to minimize such bias is a laudable one. However, it does not follow, that just because the inclusion of a particular control variable results in a flat dose response curve for all causes of death combined, it is therefore appropriate to include the variable. (One could for example control for survival itself and achieve an absolutely flat response curve.) A part of the explanation for the negative correlation observed in the Hanford data is that workers who die will frequently be ill for some period preceding their deaths, and thus will not be reporting to work and having their dosimeters read. We have found that the relatively straightforward procedure of lagging exposures for ten years greatly reduces the negative correlation observed for all causes of death. Also the use of age and calendar year as control variables (as well as sex and length of employment in analyses including females and short term workers) seems to result in less negative correlation than the use of the "obvious" factors identified by MSK.

Perhaps the most important point with regard to the use of the bioassay variable is that it is used inappropriately in that workers are classified as being in their final category throughout the follow-up period. It is evident that when workers initiate employment at Hanford, they will be in category 1 (never bioassayed). After some period of time, they may progress to category 2 (bioassayed but with no positive results), and so forth. In a correct application of the Cox model, they should not be considered to be in the higher level categories until they actually get there. This incorrect application by MSK leads to an artificial bias toward a positive correlation of radiation exposure and mortality.

MSK justify classifying workers according to the highest level of internal monitoring throughout the analysis by stating that "...although a worker might take some time to reach this level, he could easily be doing dangerous work for several years before personally reaching the level for the job" (Kneale et al. 1981). The problem with this argument is that workers who die relatively early in their follow-up will never have the opportunity to be appropriately classified, and this is where the bias comes in. Suppose for

the moment we accept the idea that at least for workers who stay at Hanford long enough, the final bioassay state is an appropriate measure of the "dangerousness" of their work. Also, for simplicity, suppose that there are two types of workers, those doing "dangerous" work and destinied for the higher bioassay levels (Group I), and those not doing "dangerous" work and who will never attain the higher bioassay levels (Group II). Because potential for internal exposure is correlated with external exposure, workers in Group I will tend to have higher levels of external exposure than those in Group II. Bias results because workers in Group I who die early are classified according to the low bioassay categories associated with Group II, and thus their relatively high external doses are inappropriately compared with the lower doses in Group II, resulting in a spurious positive correlation. This artifact explains at least in part why including control for level of internal monitoring removes the negative correlation initially observed for all causes of death combined, and pushes the correlation for "radiosensitive" cancers in the positive direction. In an analysis presented by MSK in which workers are allowed to progress through the four bioassay levels as they are followed through time, the negative correlation for all causes was not removed.

To summarize, MSK obtain a significant correlation of radiation exposure and cancer mortality only be restricting the analysis to "radiosensitive" cancers and by including the final level of internal monitoring as a control variable. Once MSK have obtained significant results, they then go on to fit a model upon which risk estimates can be based. They determine that the shape of the dose response function is non-linear and is best described by the square root function. It is claimed that linearity can be rejected. In PNL attempts to investigate the shape of the dose response function, we have found that linear and square root functions were basically indistinguishable although the square root function did fit slightly better than the linear one. MSK estimate that the doubling dose is 15 rad, or that the dose response function is of the form $\lambda(1 + \sqrt{z/15})$ where λ is the spontaneous risk. This estimate of 15 rad was obtained from an analysis of "radiosensitive" cancers with control for the final internal monitoring level so that biases resulting from these choices will affect the estimate obtained.

MSK also estimate parameters describing the latency relationship and the effect of age at exposure. We will not comment on the latency parameter since it is not that unreasonable (although the Hanford data is hardly strong enough to effectively investigate latency), but we will comment on the results for age exposure. In the model determined by MSK, the relative risk of a worker with a given exposure is to be calculated by multiplying each annual radiation exposure by an exponential function of the age at which the exposure occurs. According to this model, an exposure received at age 60 is estimated to be about 12 times as effective at producing cancer as an exposure received at age 40, and nearly 150 times as effective as an exposure received at age 20.

First, we note that the age at exposure effect determined by MSK contradicts findings based on the Japanese A-bomb survivors (Kato and Schull 1982) and other populations (Boice et al. 1977) that suggest that relative risks decrease rather than increase with increasing age at exposure. Second, it is noted that the application of the age at exposure factor in workman's compensation hearings involving workers exposed relatively late in life can result in very large risk estimates. Finally, the main reason that MSK obtain their result for age at exposure is that analyses are controlled for age only in fairly broad categories (<25, 25-34, 35-34, 35-44, 45-54, and 55+). Spontaneous rates for cancer mortality increase markedly with age, and to a large extent MSK are simply picking up this increase within the broad ten-year intervals, and calling it an effect of age at exposure.

Even if none of the specific problems discussed above were present, there would be tremendous statistical uncertainty in risk estimates based on the Hanford data. In the first part of this paper, it was demonstrated that even under the assumption of a fairly simple model with only one parameter to be estimated, confidence limits were very broad. In the MSK model, several parameters are estimated. It is difficult to assess the full uncertainty resulting from simultaneous estimation of the parameters defined by MSK, but it is clear that the risk estimates obtained would have extremely large confidence regions. In summary, risk estimates based exclusively on the Hanford data are far too unreliable to provide an adequate basis for setting radiation protection standards and for estimating the "probability of

causation" in court cases involving persons exposed to radiation. At most the role of the Hanford data must be to supplement information obtained through extrapolation from populations exposed at high levels.

Early this year, Kneale et al. (1984) updated their analyses of the Hanford data to include recent deaths, and introduced a new job hazard variable. These recent papers have been critiqued by Gilbert and Petersen (1985*). Here we note only that it is evident from some of the findings presented in this most recent analysis that the updating has substantially reduced the correlation of radiation exposure and mortality from "radiosensitive" cancers. No estimates are presented in this recent analysis, but it is clear that they would be quite different from those presented in the 1981 paper.

ACKNOWLEDGEMENTS

This work was performed for the U.S. Department of Energy under Contract DE-AC06-76RLO 1830.

REFERENCES

ANDERSON, T.W. (1978), "Radiation Exposures of Hanford Workers: A Critique of the Mancuso, Stewart and Kneale Report, Health Physics, 35, 743-750.

BEIR REPORT, (1980), The Effects on Populations of Exposure to Low Levels of Ionizing Radiation, Report of the Advisory Committee on the Biological Effects of Ionizing Radiations, Division of Medical Sciences, National Academy of Sciences--National Research Council, Washington, DC.

BOICE, J.D. Jr., LAND, C.E., SHORE, R.E., NORMAN, J.E. and TOKANAGA, M. (1979), "Risk of Breast Cancer Following Low-dose Exposure," Radiology, 131,589-597.

BUCHANAN, J.A. and GILBERT, E.S., (1984), "MOX, A User's Guide", PNL-5023, Pacific Northwest Laboratory, Richland, WA.

---

*Gilbert, E.S. and G.R. Petersen. "A Note on 'Job Related Mortality Risks of Hanford Workers and Their Relation to Cancer Effects of Measured Doses of External Radiation." To be published in Br. J. of Ind. Med.

COX, D.R. (1972), "Regression Models and Life Tables," Journal of the Royal Statistical Society, Series B, 34, 187-220.

DARBY, S.C., NAKASHIMA, E., and KATO, H. (1984), "A Parallel Analysis of Cancer Mortality AMong Atomic Bomb Survivors and Patients with Ankylosing Spondylitis Given X-Ray Therapy, RERF TR 4-84, Radiation Effects Research Foundation, Hiroshima, Japan.

DARBY, S. and REISSLAND, J. (1981), "Low-levels of Ionising Radiation and Cancer--Are We Underestimating the Risk?" Journal of the Royal Statistical Society, Series A, 144, 298-231.

GILBERT, E.S. and MARKS, S. (1979), "An Analysis of the Mortality of Workers in a Nuclear Facility," Radiation Research, 79, 122-148.

HUTCHISON, G.B., MACMAHON, B., JABLON, S. and LAND, C.E. (1979), "Review of a Report by Mancuso, Stewart and Kneale of Radiation Exposure of Hanford Workers," Health Physics, 37, 207-220.

INTERNATIONAL COMMISSION ON RADIOLOGICAL PROTECTION (ICRP) (1969), Recommendations of the International Commission on Radiological Protection. ICRP Publication 14, New York: Pergamon Press.

KATO, H. and SCHULL, W.J. (1982), "Studies of the Mortality of A-Bomb Survivors 7. Mortality, 1950-1978: Part I. Cancer Mortality." Radiation Research, 90, 395-432.

KNEALE, G.W., MANCUSO, T.F. and STEWART, A.M. (1984), "Job Related Mortality Risks of Hanford Workers and Their Relation to Cancer Effects of Measured Doses of External Radiation." British Journal of Industrial Medicine, 41, 9-14.

KNEALE, G.W., MANCUSO, T.F., and STEWART, A.M. (1981). "Hanford Radiation Study III: A Cohort Study of the Cancer Risks from Radiation to Workers at Hanford (1944-77 deaths) by the Method of Regression Models in Life-Tables," British Journal of Industrial Medicine, 38, 156-66.

MANCUSO, T.F., STEWART, A. and KNEALE, G, (1977), "Radiation Exposures of Hanford Workers Dying from Cancer and Other Causes," Health Physics, 33, 369-385.

MANTEL, N. (1963), "Chi-square Tests with One Degree of Freedom. Extensions of the Mantel-Haenszel Procedure," Journal of the American Statistical Association, 58, 690-700.

MANTEL, N. and HAENSZEL, W. (1958), "Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease," Journal of the National Cancer Institute 22, 719-748.

REISSLAND, J.A. (1978), "An Assessment of the Mancuso Study," in National Radioecological Protection Board Document, NRPB-R79, Springfield: National Technical Information Service.

TOLLEY, H.D., MARKS, S, BUCHANAN, J.A., and GILBERT, E.S. (1983), "A Further Update of the Analysis of Mortality of Workers in a Nuclear Facility," Radiation Research, 95, 211-213.

MELANOMA AMONG LAWRENCE LIVERMORE NATIONAL LABORATORY EMPLOYEES: AN EPIDEMIOLOGIC PUZZLE

Dan Moore
Deborah Bennett
Mortimer Mendelsohn

Biomedical Sciences Division
Lawrence Livermore National Laboratory
Livermore, California 94550

Prepared for presentation at the 1984 Statistics Symposium on National Energy Issues, October 16-18, Seattle, Washington.

Work performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under contract number W-7405-ENG-48.

MELANOMA AMONG LAWRENCE LIVERMORE NATIONAL LABORATORY EMPLOYEES: AN EPIDEMIOLOGIC PUZZLE

Dan Moore, Deborah Bennett and Mortimer Mendelsohn
Biomedical Sciences Division
Lawrence Livermore National Laboratory
Livermore California 94550

ABSTRACT

Since 1972 the Lawrence Livermore National Laboratory has experienced malignant melanoma diagnosis among its 7000-8000 employees at a rate three to four times that for the surrounding community. A brief history including how the increase was detected, what analysis was used to measure the size of the increase and what correlations have been investigated to understand the possible causes of this increase is presented. Results of a survey of incidence among former employees and a separate survey of malignant melanoma (MM) among Los Alamos National Laboratory employees suggest that the increased incidence is unique to current LLNL workers. A case-control study, conducted by the California Resource for Cancer Epidemiology, found that work exposure to radioactive materials increased risk of MM while a study based on badge dosimetry found no association with ionizing radiation exposure. A mortality study, covering the time period 1964-1979, found six MM deaths, significantly fewer than the number expected based on the observed high incidence. These and other findings are compared with results from other epidemiologic studies of MM to illustrate the puzzling nature of this poorly understood disease which is increasing rapidly in fair-skinned populations throughout the world.

INTRODUCTION

This paper describes an unexplained recent excess incidence of malignant melanoma (MM) among employees of the Lawrence Livermore National Laboratory. We begin with a brief description of how the Laboratory became aware of the problem and then describe some of the studies we have undertaken to try to understand the problem. Finally, we outline current plans for further studies.

First, we begin with some background information on cancer in the United States. This year approximately 870,000 Americans will get cancer; about 450,000 will die from cancer. Slightly more women than men will get

cancer, while more men than women will die from cancer (Silverberg, 1984). Figure 1 shows cancer incidence and mortality rates for white males in the U.S. over the time period 1969 to 1976. The all cancer incidence and mortality rates have been rising very slowly, about 1% per year, over this 7-year period. Lung cancer, the most common cancer among males accounting for 22% of the incidence and 35% of the mortality, has also been increasing slowly, 1.4% per year for incidence and 2.6% per year for mortality. In contrast, melanoma, which accounts for less than 2% of the total, is the most rapidly increasing form of cancer among white males. Incidence increased at a 6.8% annual rate while mortality rose at 4.0% per year over this period.

Cancer incidence among employees of the Lawrence Livermore National Laboratory for the time period 1969-1980 is shown in Figure 2. These numbers were obtained by comparing yearly rosters of employees with California Tumor Registry files. The tumor registry covers five San Francisco Bay Area counties and includes the residences of over 90% of Laboratory employees. In contrast to statistics for the U.S., malignant melanoma (MM) is the most frequently diagnosed cancer among Laboratory employees. Melanoma stands out as the only cancer with significantly more observed cases (in both sexes) than the number expected based on U.S.

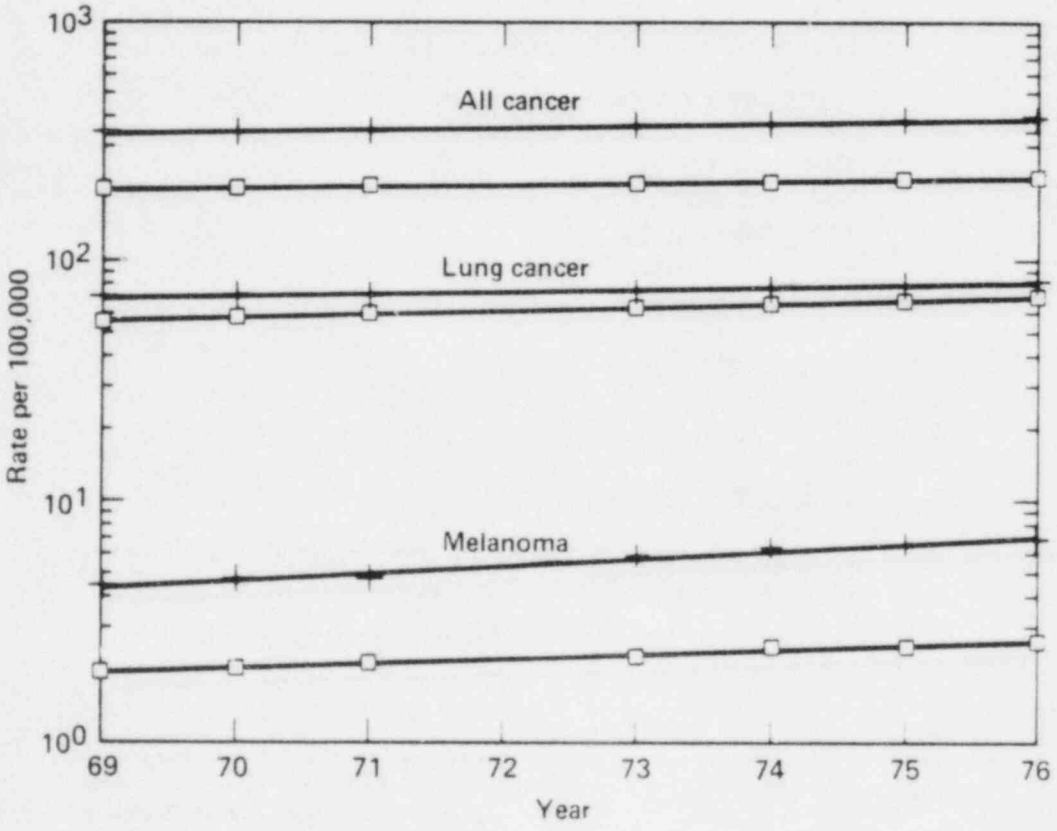

Figure 1. US White Male Cancer Incidence (+) and Mortality (□) Rates 1969-1976.

rates. During 1969-1980 30 cases of MM were diagnosed among employees compared to 8.3 expected. Lung cancer, the most frequently expected cancer with 42.1 expected cases, was significantly under-represented among LLNL employees with 24 cases. The total number of diagnosed cancers was 178 compared to 173 expected.

HISTORY OF MELANOMA INCIDENCE AT THE LABORATORY

Figure 3 shows the year-by-year incidence of MM among current Laboratory employees. From the beginning of the Laboratory in 1952 through 1959 there were no diagnosed cases of MM (according to Laboratory medical files). In 1960 the first case appeared but there were no further cases until 1963. In 1964 there was a third case but no more cases until 1968. Starting in 1968 there was one case per year until 1972 when a cluster of four cases occurred. This was the first year that the cumulative Laboratory rate (based on accumulating cases and person-years of employment over a 6-year period) diverged from the rate for Alameda County, which houses 80% of LLNL employees. The rise in the Laboratory rate continued through the mid 1970's so that by the end of 1976 a total of 21 cases had

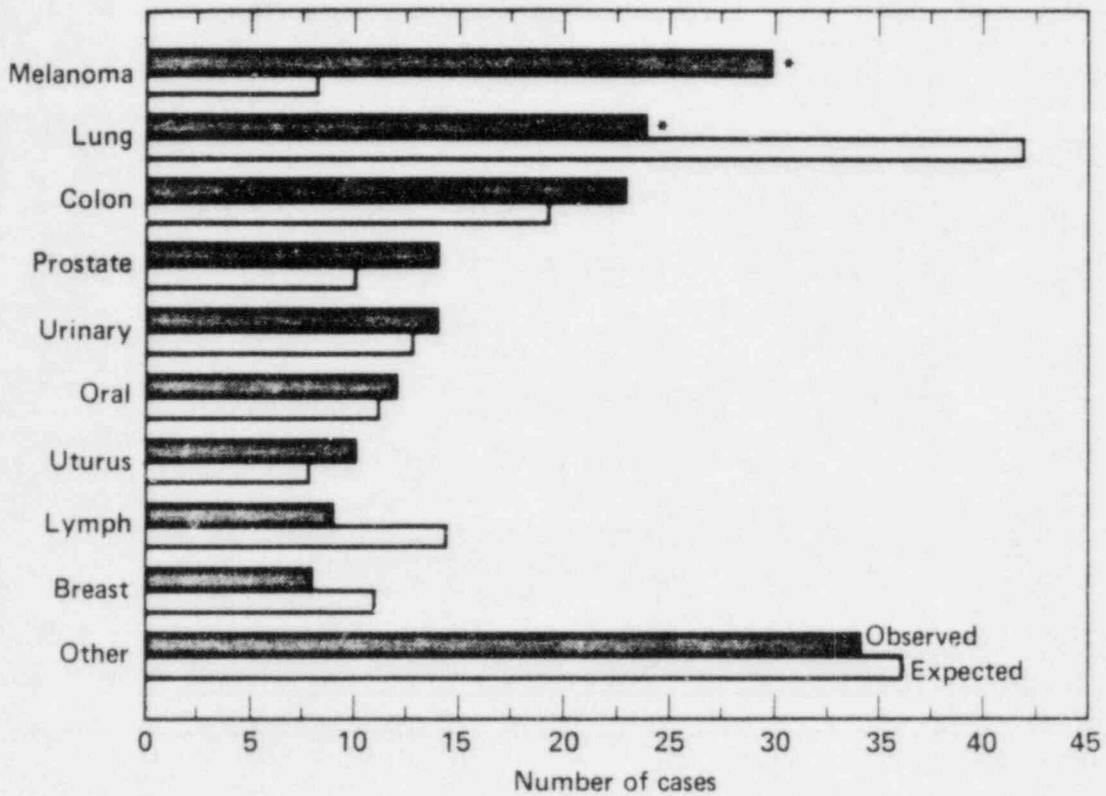

Figure 2. Cancer incidence at LLNL during 1969-1980. An asterisk (*) denotes a significant ($p < .01$) difference between observed and expected (based on cancer incidence rates for the San Francisco-Oakland Standard Metropolitan Statistical Area) numbers of cases.

occurred. It was at this point that the medical services department became alarmed, although at the time the medical services department was aware of only 15 cases. (There is no requirement that an employee inform the Laboratory of a diagnosis of any form of cancer.) A comparison was made between the number of cases known at that time and the number expected based on rates for the Bay Area counties; it appeared that roughly twice as many melanomas had occurred as would be expected. In early 1977 the medical services department requested help from the Resource for Cancer Epidemiology (RCE) under Dr. Donald Austin, the keepers of the tumor registry, to determine whether or not the Laboratory was experiencing abnormally high MM diagnosis. This request was accompanied by headlines in the local press which may have influenced additional case-finding in 1977 when six new cases were discovered. In 1978 only one new case was diagnosed, suggesting a harvesting of cases the previous year. The study by the RCE was completed in 1980 confirming the high incidence of MM among employees (Austin, 1980). This announcement was attended by exposure to the mass media which dramatically affected case finding for 1980 when an additional eight cases were diagnosed. (In addition to the eight cases shown in Figure 3, there were three "atypical melanocytic hyperplasias", very early lesions not considered melanomas, diagnosed among employees.)

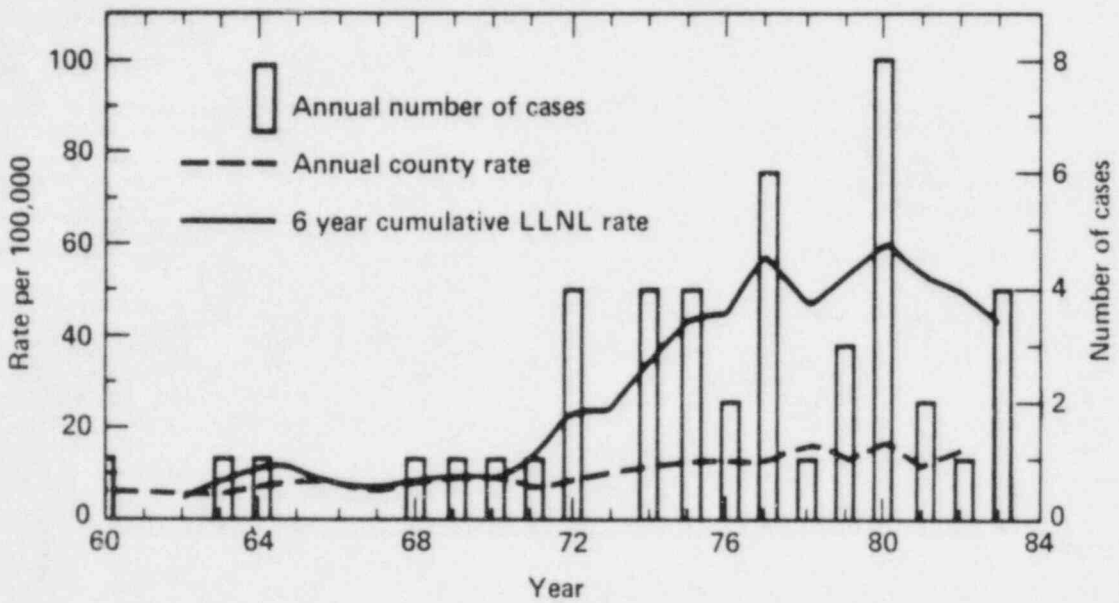

Figure 3. Melanoma at LLNL. Numbers of cases (both sexes) per year are shown as histogram bars. The LLNL incidence rate, based on six-year accumulations of cases and employee-years at risk and plotted at the ends of the six-year period, is shown as a solid line (the upper line). Incidence rates for Alameda County white males are shown as a dashed line (the lower line). The Laboratory is located in Alameda County and the employees are 80% white male.

Finally, starting in 1981 there was a return to the "normal" (high) Laboratory level of incidence. Two additional cases have been diagnosed in 1984 (as of October 1) and are not shown in the figure.

POST 1980 STUDIES OF MELANOMA AMONG LABORATORY EMPLOYEES

The confirmation of the suspected high incidence rate of MM among LLNL employees by the Resource for Cancer Epidemiology (RCE) led to the initiation of several studies by the Laboratory. These studies are of three types:

(1) In-house studies, including
   (a) a study of mortality among LLNL employees,
   (b) a study of incidence among former employees who continue to reside in the five bay area counties, and
   (c) studies utilizing Laboratory records to determine possible associations between melanoma cases and occupational factors;

(2) A case-control study conducted by the Resource for Cancer Epidemiology;

(3) A comparison of medical records of LLNL Kaiser Health Plan members with those of non-LLNL plan members. (Approximately 50% of LLNL employees are members of the Kaiser Health Plan.)

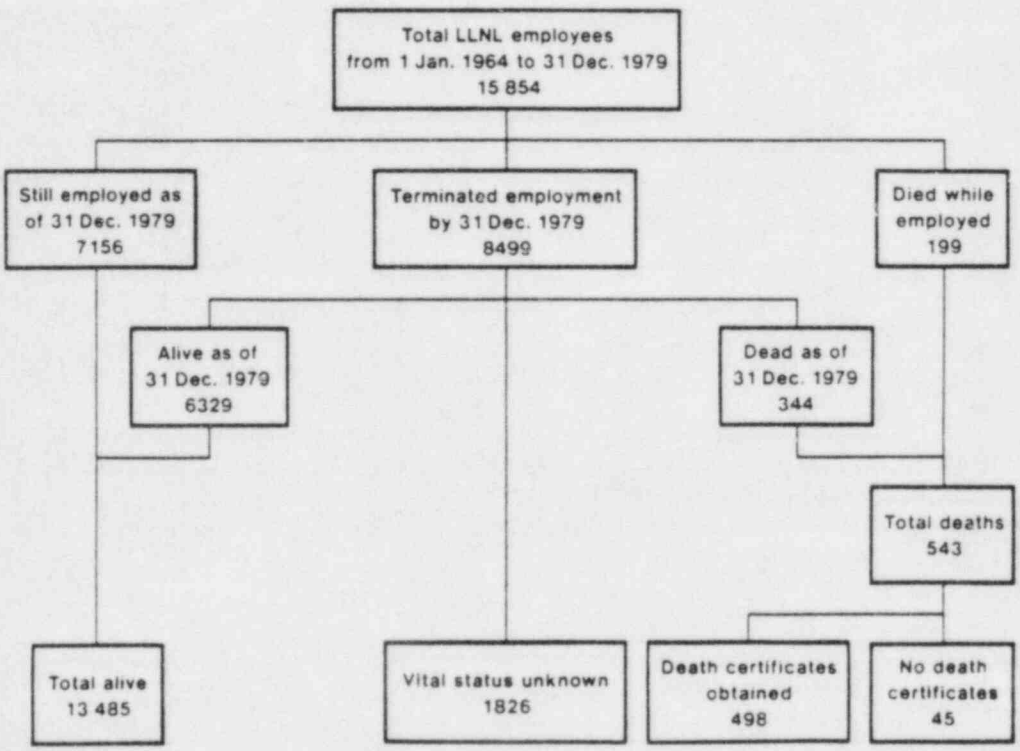

Figure 4. Vital status of the LLNL cohort from January 1964 to December 1979. At the end of the study period, 85.1% of the former employees were still alive, 3.4% had died, and the vital status of the rest (11.5%) was unknown.

RESULTS FROM MORTALITY STUDY AMONG LLNL EMPLOYEES

The mortality study covers all persons who ever worked at the Laboratory between 1964 and 1979. The 1964 date was selected because it marked the beginning of computer files that contain the names of all former LLNL employees and the ending date is the latest for which the Social Security Administration (SSA) had complete records of deaths among its members. The vital status of the 15,854 persons included in this study is shown in Figure 4. The age distribution at time of hire is shown in Table 1 and the number of person-years lived by this cohort is shown in Table 2. The Occupational Cohort Mortality Analysis Program (OCMAP) (Marsh, 1980) was used to analyse the mortality data. The data were assembled on computer tape by the Death Certificate Retrieval Office (DCRO) at Oak Ridge Associated Universities. A nosologist from the DCRO coded cause of death from death certificates for those who were determined by the SSA as dead.

TABLE 1. CHARACTERISTICS OF LLNL STUDY POPULATION.

Age at Hire	White		Nonwhite		Totals
	Males	Females	Males	Females	
< 20	317	319	232	150	1018
20-24	2638	715	454	221	4028
25-29	2615	519	286	101	3521
30-34	1843	400	189	83	2515
35-39	1276	346	161	44	1827
40-44	967	257	123	22	1369
45-49	576	176	66	13	831
50-54	330	67	43	7	447
55-59	174	29	27	1	231
60-64	45	5	5	2	57
65 +	0	4	0	0	10
Totals	10787	2837	1586	644	15854

TABLE 2. PERSON-YEARS EXPERIENCED BY STUDY POPULATION.

Age at Hire	White		Nonwhite		All Totals
	Males	Females	Males	Females	
< 20	347	358	326	199	1230
20-24	5840	2654	1937	1107	11538
25-29	13745	4216	2460	1142	21563
30-34	17903	4167	1867	655	24592
35-39	17879	3111	1528	389	22907
40-44	16701	2568	1347	278	20894
45-49	14874	2322	978	170	18344
50-54	11877	2008	706	114	14705
55-59	8196	1336	483	41	10056
60-64	4466	659	307	18	5450
65 +	2872	350	218	8	3448
Totals	114700	23749	12157	4121	154727

The results of this mortality study are summarized in Table 3 where the observed numbers of deaths are compared with those expected based on U.S. vital statistics death rates. The Standardized Mortality Ratio (SMR), equal to the observed divided by the expected number of deaths multiplied by 100 to convert to a percent, is the basis for comparison and is shown in the fourth column of the table. An SMR of 100 means that the observed number of deaths is equal to that expected based on rates for the entire U.S. population, adjusted for age, sex, race and calendar year. SMRs less than 100 mean that fewer than the expected number of deaths were observed, while those greater than 100 mean that more deaths were observed than expected. The overall mortality was about 59% that expected ($p < .01$). The major contributors to this low SMR were the paucity of deaths due to circulatory system diseases and accidents and suicides. Laboratory

TABLE 3. MORTALITY AMONG ALL LLNL EMPLOYEES FROM 1964 TO 1979 COMPARED WITH MORTALITY EXPECTED ON THE BASIS OF U.S. VITAL STATISTICS.

Cause of Death	Observed	Expected	Standardized Mortality Ratio, %	95% Confidence Limit	
				Lower	Upper
All causes of death	543	920	59[a]	54	64
All cancers	132	188	70[a]	59	83
Buccal cavity and pharynx	2	6	32	4	115
Esophagus	4	5	84	23	214
Stomach	4	8	53	15	137
Large intestine	11	15	75	37	134
Rectum	4	5	87	24	222
Liver	4	3	140	38	358
Pancreas	9	10	95	43	180
Respiratory system	40	65	62[a]	44	84
Bone	0	1	0	–	–
Skin	6	4	139	51	302
Breast	5	6	87	28	202
Female genital system	1	4	26	1	145
Prostate	5	5	93	30	217
Testis	1	2	61	2	343
Bladder	4	3	117	32	299
Kidney	4	5	85	23	217
Eye	0	0	0	–	–
Brain and central nervous system	8	7	108	47	213
Thyroid	0	0	0	–	–
All lymphopoietic cancer	8	20	40[a]	17	79
Benign neoplasms	1	3	36	1	201
All circulatory system diseases	223	383	58[a]	51	66
All respiratory diseases	17	44	39[a]	23	63
Allergy, endocrine, metabolic, and nutritional diseases	9	16	55	25	105
All digestive system diseases	14	57	25[a]	13	41
Cirrhosis of the liver	9	37	25[a]	11	46
Accidents and suicides	90	167	54[a]	43	66
All other causes	3	25	12[b]	3	35

[a]Significant at the 1% level.
[b]Significant at the 5% level.

employees also experienced significantly lower cancer mortality (SMR=70, $p < .01$). The major contributor to this low ratio was the small number of lung cancer deaths, probably attributable to a low percentage of smokers among LLNL employees. Four cancer sites had SMRs greater than 100: liver, skin (mortality due entirely to melanoma among LLNL employees), bladder and brain and central nervous system. None of these SMRs are statistically significant, however. This is evidenced by the inclusion of 100 in the 95% confidence limits shown in the last two columns of the table.

Table 4. COMPARISON OF STANDARDIZED MORTALITY RATIOS AMONG WHITE MALES FROM FIVE WORKER POPULATIONS.

Cause of Death	LLNL 1964-79	Hanford[a] 1945-67	Uranium Workers[b] 1943-73	Petrochemical Workers[c] 1941-77	DuPont Chemists[d] 1964-77
All causes of death	63	75	93	83	47
All cancers	77	85	85	86	50
Buccal cavity and pharynx	38	106	79	–	–[e]
Esophagus	112	93	64	–	–
Stomach	67	69	73	–	55
Large intestine	83	101	50	78	174
Rectum	77	68	32	–	85
Liver	183	51	57	161	78
Pancreas	88	100	96	108	43
Respiratory system	65	77	106	85	29
Bone	0	65	90	–	–
Skin	157	84	95	–	0
Prostate	109	90	81	32	70
Testis	64	–	55	–	–
Bladder	130	67	80	51	0
Kidney	96	104	75	107	83
Eye	0	–	90	–	–
Brain and central nervous system	111	102	95	162	29
Thyroid	0	0	0	–	–
All lymphopoietic cancer	48	54	77	116	99
Benign neoplasms	46	–	92	235	–
All circulatory system diseases	61	76	85	82	48
All respiratory diseases	42	–	110	58	33
Allergy, endocrine, metabolic, and nutritional diseases	57	–	65	–	–
All digestive system diseases	27	–	77	39	41
Cirrhosis of the liver	24	–	–	–	29
Accidents and suicides	57	75	109	103	41
All other causes	19	65	–	–	–
Total number of deaths	468	2089	5394	765	198

[a]Tables II and III (Gilbert and Marks 1979).
[b]Table 3 (Polednak and Frome, 1981).
[c]Table 2 (Austin and Schnatter, 1983).
[d]Table 7 (Hoar and Pell, 1981).
[e]Not reported.

The low SMRs evident in the table are at least partially explained by the so-called "healthy worker effect". This refers to the fact that only the healthy are able to work whereas many terminally or seriously ill persons are included in the U.S. population. The healthy worker effect is usually not as strong for cancer SMRs since cancer is primarily a disease of the old and is seldom detectable at hiring age. Table 4 compares Laboratory SMRs, this time for white males only, with those for four other sets of workers. The SMR for all causes is significantly lower than those for three of the four comparison populations. The same is true for all cancers. Each population has a few cancer SMRs greater than 100 but there is no consistent pattern and the LLNL pattern does not stand out as being unusual. These results suggest that, at least as of the end of 1979, LLNL was not experiencing a significant increase in melanoma mortality.

It may be argued that the mortality study is too small to detect an increase in melanoma mortality but calculations based on the Poisson distribution suggest that the power was at least 90% of detecting a three-fold increase (corresponding to the observed increased incidence) and 50% of detecting a doubling. Another limitation is that the study period may not be long enough to couple incidence to mortality since melanoma has a relative long (for cancer) expected survival time and the mass of cases did not really occur until 1977 and after. We are currently extending the study to cover 1982 (a date for which SSA records are now complete) which should provide a better indication of whether or not mortality is rising with incidence.

MELANOMA INCIDENCE AMONG FORMER EMPLOYEES

This study was designed to determine whether former LLNL employees experience the same MM incidence rate as current employees. Questionnaires were sent to a sample of 1000 former employees (out of 11,284 who terminated employment between 1963 and 1981, inclusive) to determine their length of residence during the years 1972-1981 in the five Bay Area counties covered by the tumor registry. From the 270 returned questionnaires we determined that on average each respondent spent 3.04 years in the tumor registry collection area. This number was then multiplied by the total number of former employees (11,284) to obtain person-years at risk. Published melanoma incidence rates (Austin, 1981) can then be applied to person-years, subdivided by age and sex, to obtain an expected number of cases among former employees. The entire roster of terminated employees was then compared with the tumor registry to determine the number of MM cases. Eight cases among former employees were found by the tumor registry. Two of these cases had less than one year at LLNL prior to diagnosis (3 weeks for one and 7 months for the other). The remaining six cases are not significantly greater than the 4.71 expected based on current rates for and length of time spent by former employees in the registry area. The observed number is also significantly below the number expected under the assumption that former employees have the same rates as current employees (i.e. three to four times the tumor registry rates).

IN-HOUSE STUDIES

Radiation Dosimetry

Twenty MM cases occurring between 1974 (the first year for which computerized dosimetry records are available) and 1979 were matched with all LLNL employees who were the same age (within 1 year), same sex and who had been employed the same amount of time (within one year). Cases occurring after 1979 were omitted from the analysis because it was felt that the underlying etiology of the disease may have been different for many of the 1980 cases who were alerted to their lesions by the publicity following the Austin report. The cumulative radiation dose, as measured by badge reading, for each case was then ranked among the peers. The results, summarized in Table 5, show that most MM cases (13 of 20) received less radiation than their peers. A statistical test of the distribution of the ranks showed no significant departure from randomness and confirms a lack of association between ionizing radiation and MM.

Job Category

Table 6 compares the observed numbers of MM cases in each job category with those expected assuming a homogeneous MM rate over all job

TABLE 5. DOSIMETRY RANK FOR CASES AND MATCHED CONTROLS 1974-1979.

Case Number	Age at Dx	Years at Lab	Case Rank	Total Matches	Probability[a]
1	44	19	15	90	0.17
2	46	9	8	12	0.67
3	42	18	42.5	59	0.72
4	65	5	3.5	4	0.88
5	34	4	57.5	82	0.70
6	47	17	48	88	0.55
7	46	24	8	13	0.62
8	43	11	29	82	0.35
9	42	17	55	95	0.58
10	42	8	5	33	0.15
11	30	28	2	11	0.18
12	55	18	67	72	0.93
13	57	21	55.5	69	0.80
14	47	17	71	90	0.79
15	49	19	41	97	0.42
16	30	1	118	203	0.58
17	43	12	9	12	0.75
18	29	2	126	216	0.58
19	55	20	3	11	0.27
20	48	7	14.5	19	0.76
Mean			38.93	68	0.57

[a]Probability that case rank is distributed at random.

categories. This table is based on data for the years 1967-1979. (Data for numbers of employees by age, sex and job class for years prior to 1967 are not available in computer format. Again cases occurring after 1980 were omitted for the same reasons as outlined previously.) The table shows that scientists, and in particular chemists, have higher rates of MM than the rest of the Laboratory. This finding agrees with results from the case-control study (reported below). A death certificate survey of US chemists conducted by the American Chemical Society revealed an excess of pancreatic cancer, but no excess of melanoma as causes of death (Li et al., 1969). Two more recent studies, however, found elevated MM rates among chemists (Hoar and Pell, 1981; Wright et al., 1983). The five LLNL cases among chemists explain less than one-third of the excess of 18 cases occurring among LLNL employees in 1967-1979 and more importantly, of the subsequent 18 cases of skin melanoma since 1979, only 2 have been chemists.

Building Study

Table 7 shows the distribution of cases occurring between 1960 and 1979 according to building location at time of diagnosis. An expected number of MM cases can be calculated for each building by multiplying the overall MM incidence rate for the Laboratory by the number of person-years occupancy (equal to the number of persons times the number of years of building occupancy). The probability of observing a given number of cases in a building by chance can then be determined from the Poisson distribution. Low probabilities suggest clustering of cases. However, the values reported as probabilities in Table 7 do not take into account the large number of buildings in which clustering, real (i.e., due to a presumed actual but unidentified cause) or spurious (i.e., due to random scattering of cases), could have occurred. In June 1980 there were 100 permanent buildings and 181 trailers on site of which 55 buildings and 66 trailers each housed 11 or more employees. The probability that there would be one or more clusters with probabilities as small as 0.0063 (the

TABLE 6. MELANOMA RATES BY JOB CATEGORY 1967-1979

Category	Number of Cases		Ratio
	Observed	Expected	
Scientist	15	9.12	1.64
Chemist	5	1.42	3.52*
Physicist	4	2.82	1.42
Engineer	3	3.52	0.85
Other	3	1.36	2.21
Administrator	1	1.79	0.56
Supervisor	4	2.86	1.40
Technician	4	5.52	0.72
Clerical	2	2.84	0.70
Craft	0	2.38	0.00
Other	1	2.49	0.40
Totals	27	27	1.00

*Significantly elevated ($p < 0.05$).

corrected value for building 111) among these 121 buildings and trailers is 0.53. Thus the observed clustering of cases in building 111 could well be due to chance. There have been no new cases in this building since March 1977. The argument about chance clustering can also be applied to the five other buildings with individual probabilities below 0.05. None of the buildings has more than two cases. We conclude from this that there is no strong evidence for a building effect.

Eye and Hair Color

Eye and hair color data were obtained for each current (September 1982) employee from files in the LLNL badge office. The results were then compared to the 29 MM cases who were employed at that time. As summarized in Table 8, the data suggest an excess of blond hair among the melanoma cases (9 cases vs. 3.67 expected based on the percent of blond hair in the workforce). This excess is statistically significant. It would be interesting to determine whether or not there are proportionately more blonds in the LLNL workforce than in the surrounding community, but thus far we have been unsuccessful in obtaining data on hair color composition for the surrounding community.

TABLE 7. MELANOMA RATES FOR ON SITE BUILDINGS WITH CASES 1960-1979.

Building	Activity	Observed Cases	Expected Cases[a]	Probability[b]	Job Classes[c]
111	Theoretical Physics	5	0.89	0.0022[d]	Admin, Chem, EE, Prog, Phys
113	Computations	2	0.92	0.23	E Tech, Clerk
116	Computations	1	0.13	0.12	E Tech
121	Experimental Physics	2	0.65	0.14	Phys, E Tech
131	Engineering	4	3.28	0.41	ME, E Tech, Chem, Prntr
174	Laser Research	1	0.04	0.04	EE
212	Accelerator	1	0.15	0.14	Drfmn
222	Chemistry	2	0.63	0.13	Chem, Chem
253	Hazards Control	1	0.32	0.27	H & S Tech
315	MFE Physics	1	0.49	0.38	Phys
321	Metals Fabrication	1	0.80	0.55	Assm Tech
361	Biomedical	2	0.23	0.02	Biophys, Biophys
381	Lasers	1	0.35	0.29	Phys
416	Police	2	0.25	0.03	Guard, Guard
511	Crafts	1	1.77	0.83	Laborer
612	Dry Waste	1	0.07	0.02	H & S Tech
T114	Computations	1	0.07	0.07	Phys
T1403	Energy	1	0.05	0.05	Chem
All Others		0	18.97	1.00	

[a]Expected cases = Overall Lab rate × Person Years of building occupancy.
[b]Probability of observed based on poisson distribution with expected cases.
[c]Job Class abbreviations: Admin = Administrator, Chem = chemist, EE = electrical engineer, Prog = programmer, Phys = physicist, E Tech = electronic technician, ME = mechanical engineer, Prntr = printer, Drfmn = draftsman, H & S Tech = health & safety technician, Assm Tech = assembly technician, Biophys = biophysicist.
[d]Probability = 0.0063 when adjusted for age, sex, and job class distribution in building.

There is a slight excess of blue and hazel eyes among the melanoma cases, but the excess is not statistically significant. A large hospital study in New York City, found 9% blonds and 26% blue eyes among 1938 non-MM white patients (Gellin et al., 1969). In this loose comparison, the LLNL percentages are 50% higher but this is not enough to explain the three-fold Laboratory excess number of cases.

Medical Description of MM Tumors

We have obtained information on anatomical location of the MM tumor

TABLE 8. HAIR AND EYE COLOR FOR LLNL EMPLOYEES.

Hair Color	All Workers		Melanoma Cases	
	Number	Percent	Number	Expected
Blonde	965	13.09	9	3.67
Red	207	2.81	0	0.79
Brown	5317	72.12	18	20.19
Black	883	11.98	1	3.35
Totals	7372	100	28	28

χ^2 = 10.44 (Probability that number of melanoma cases is no different from expected = 0.015).

Eye Color	All Workers		Melanoma Cases	
	Number	Percent	Number	Expected
Blue	2782	36.38	12	10.55
Green/Grey	740	9.66	2	2.80
Hazel	1053	13.74	7	3.99
Brown/Black	3081	40.22	8	11.66
Totals	7661	100	29	29

χ^2 = 3.86 (Probability that number of melanoma cases is no different from expected = 0.28).

TABLE 9. ANATOMICAL DISTRIBUTION OF MALIGNANT MELANOMA IN LLNL AND QUEENSLAND.[a]

Body Location	Males			Females		
	LLNL		Queensland	LLNL		Queensland
	Number	Percent	Percent	Number	Percent	Percent
Trunk	22	59	47	6	60	22
Arm	6	16	17	2	20	23
Leg	6	16	13	2	20	36
Head	6	8	23	0	0	19
Totals	37			10		

[a]Queensland data based on 341 males and 349 females (Little *et al., Med. J. Aust.* 1: 66-69, 1980). LLNL data are for 1960-1983.

for 47 of the cases studied. These distributions are shown in Table 9 where they are compared with anatomical distributions for MM in Queensland, Australia, where overall incidence rates are comparable to those for the Laboratory. Our distributions do not differ significantly from those in the Queensland study, although there seem to be fewer lesions on the head among both sexes at LLNL.

We have also obtained information on tumor thickness for 36 LLNL employees who were referred to the UCSF melanoma clinic. Mean thickness by calendar year for all clinic patients was also obtained. Figure 5 shows that thickness has decreased with calendar time for the LLNL cases while no such decline is apparent in the clinic. The figure shows that awareness and medical surveillance can have a dramatic effect on the stage at which melanoma is discovered. In Queensland, where MM rates are comparable to ours, it has been noted that increased awareness results in earlier diagnosis and smaller lesions (Little et al., 1980).

HEALTH PLAN COMPARISON STUDY

The results of this study conducted by Robert Hiatt and Bruce Fireman of the Kaiser Foundation Research Institute have been recently published (Hiatt and Fireman, 1984). The purpose of this study was to determine whether the pattern of health plan usage (specifically usage of a

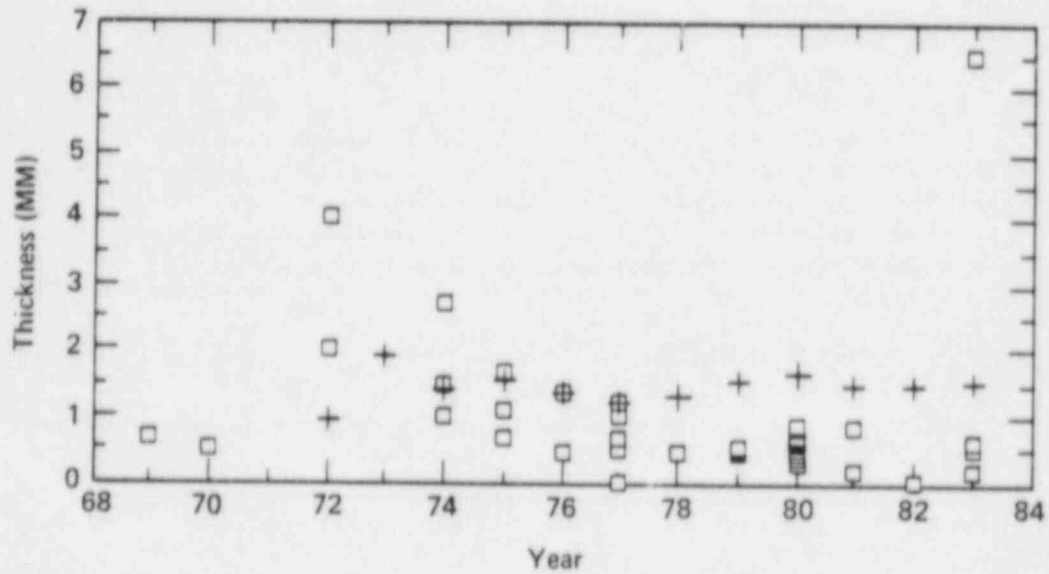

Figure 5. MM lesion thickness by year of diagnosis. Thirty-six (of 42) MM lesions (shown as □ in the figure) occurring at LLNL since 1969 have been examined by a single dermopathologist (Dr. Richard Sagebiel, UCSF Melanoma Clinic). Except for a single thick lesion in 1983, there is a striking reduction in tumor thickness with time at LLNL. In contrast, average tumor thickness for clinic patients, 1972-1983, (shown as + in the figure) shows no such reduction.

dermatology clinic at one of the plan's hospitals) was different for LLNL employees compared to other non-LLNL health plan members. The main findings of this study can be summarized as follows:

- LLNL members are three times as likely to be diagnosed with MM as non-LLNL members. (This confirmed the findings of the initial RCE study.)
- LLNL members are three times as likely to have cutaneous biopsies as non-LLNL members.
- LLNL members do not have elevated rates of other skin cancers or of actinic (solar) changes of the skin.
- LLNL members without melanoma are no more likely to visit the dermatology clinic than non-LLNL members.
- There is a suggestion that the MM incidence rate among LLNL family members is also elevated three-fold. However this finding is based on only six cases and is not statistically significant.

CASE-CONTROL STUDY

The purpose of this study, conducted by Donald Austin and Peggy Reynolds of the RCE (1984), was to identify work-related factors which may contribute to the increased incidence of MM among LLNL employees. The study design matched four LLNL employees to each of 31 LLNL MM cases occuring between 1969 amd mid-1980. The matching variables were age, sex and concurrent employment at time of MM diagnoses in each case. Over 180 factors were investigated by means of an extensive questionnaire followed by an in-depth interview. Several constitutional risk factors, all of which have been found to be related to MM in other studies, were identified. These include: the presence of numerous large moles (6 or more larger than 1/2 cm in diameter), a parental history of skin cancer, a previous (non-melanoma) skin cancer, a proclivity to burn rather than tan, and the acquisition of an advance educational degree.

In addition to these constitutional factors five occupational factors were reported as significant in influencing the risk of being diagnosed with MM. They are: exposure to radioactive materials, exposure to volatile photographic chemicals, work at Site 300 (a non-nuclear testing grounds), presence at the Pacific Test Site at the time of a nuclear event and duties as a chemist. These findings are summarized in Table 10. Austin and Reynolds identify these factors as "causal" and "independent contributors of risk" based on the results of fitting a multiple logistic regression model to the data.

We have obtained a copy of the data, except for the identity of the individuals, and have begun a reanalysis. We are able to reproduce the results reported by Austin and Reynolds. However, we have found that by entering the variables in a stepwise manner, into the same model as used by Austin and Reynolds, that only the first two occupational factors are significantly and independently related to MM incidence. In our analysis we first found a subset of constitutional factors which are independent and significantly related to melanoma and then searched for significant

occupational factors. The final model contains four constitutional factors (six or more large moles, previous non-MM skin cancer, advanced educational degree and fewer than 105 days per year spent outdoors after age 21) and two occupational factors (exposure to radioactive materials and exposure to volatile photographic chemicals). We do not agree with Austin and Reynolds' conclusion that the occupational factors are causal. First, the design of the study is such that it is highly likely that several factors will appear to be "significant" by chance alone. Over 180 statistical tests for significance are reported in Austin and Reynolds' tables; chance alone can account for nine significant results (5% of 180). Austin found 2 significant occupational exposures among 43 items tested: exposure to radioactive materials and exposure to volatile photographic chemicals. The probability that these occurred by chance when 43 are tested is 0.64 (based on the binomial distribution with p=0.05). The "most significant" exposure - with a calculated p-value of 0.0015 - was to radioactive materials. The probability of this occurring by chance is 0.0625. Thus it is reasonably likely that the reported significant associations between occupational exposures and melanoma arose by chance rather than by cause and effect. A second reason for doubting Austin and Reynolds' conclusion of a cause and effect relationship is the lack of previous evidence linking these occupational exposures with melanoma. While it is possible that the link between photographic chemicals and MM was previously undiscovered because the question had never been asked, this explanation cannot apply to ionizing radiation. There is a voluminous literature on radiation effects that is consistently negative with regard to an association between ionizing radiation and MM. There is also some internal inconsistency regarding these "causes". Exposure by contact to photographic chemicals was not higher among cases; neither was photography as a hobby. Similarly, working with glove boxes, wearing protective clothing, or using breathing apparatuses did not appear more frequently among cases than controls, although these do correlate with occupational exposure to radioactive materials.

TABLE 10. SUMMARY OF FINDINGS FROM CASE-CONTROL STUDY: UNIVARIATE RESULTS.

	Cases	Controls	Odds Ratio	*P*-value
Constitutional Factors				
6 or more moles	11/30	5/110	12.2	0.00002
Advanced ed. degree	13/31	19/109	3.4	0.006
Parental skin cancer	8/30	10/110	3.6	0.017
Burn w/o tanning	11/30	21/110	2.4	0.040
Previous non-MM skin cancer	4/31	3/110	5.3	0.042
Occupational Factors				
Radioactive materials	20/31	36/110	3.7	0.0015
Volatile photo chem	11/31	17/110	3.0	0.016
Chemist duties	4/31	2/110	8.0	0.021
Work at Site 300	18/31	42/110	2.2	0.039
At Pacific Test	4/30	4/110	4.1	0.065

DISCUSSION

The incidence of MM continues to be high among current employees. At this time it is not known whether employees also have elevated MM mortality rates, although the mortality study covering 1964-1979 suggests that mortality will fall nearer to the normal rates than to the three-fold higher incidence rates experienced by the Laboratory. The size of recent lesions (Figure 5) suggests early detection of this disease which should lead to improved survival. The RCE case-control study found the usual associations of MM with multiple large moles, advanced education and tendency to burn rather than tan. The associations of MM with parental non-melanoma skin cancer and with previous non-melanoma skin cancer among cases appear to be new, but plausible, findings. Our record study also confirmed previous reports of an association between MM and lightly colored hair.

The case-control study also reported significant associations of MM with exposure to radioactive materials and to volatile photographic chemicals, chemist duties, assignment to a non-nuclear explosive test site and presence at an atmospheric nuclear test in the Pacific. Previous studies have shown increased MM rates among chemists, but only 4 of 31 LLNL cases are chemists. When the occupational factors are entered step-wise into the analysis (based on a multiple logistic linear regression model) only two exposure factors emerge as significant. We believe it is likely that these two arose by chance. Neither badge dosimetry analysis nor reports of other groups exposed to ionizing radiation confirm the association between MM and exposure to radioactive materials.

A recent study of MM at Los Alamos National Laboratory (LANL) has profound impact on our findings (Acquavella et al., 1982). This study, based on the New Mexico cancer registry, found no significant increase in MM incidence among LANL employees. This makes it much more difficult to explain the LLNL increase as a result of job-related activities, since the two laboratories have nearly identical research activities and their employees are potentially exposed to similar chemicals and radiations. A second study at LANL found that the most significant risk factor for MM was level of education (Acquavella et al., 1983). Those who had college degrees had two-fold risk while those with graduate degrees experienced three-fold risk compared to those with no college degree. A similar finding was also made in Western Australia where "professionals" had much higher MM incidence rates than "laborers" (Holman et al., 1980).

We have assembled a number of hypotheses that might explain the high incidence of MM at LLNL.

The Statistical Hypothesis: that the LLNL rates are a chance association. The initial high incidence of malignant melanoma among LLNL employees during 1972-1977 was calculated by Austin to have a probability of 10^{-8} of occurring by chance (Austin et al., 1981). However, this calculation did not take into account the prior selection of the population for study based on the strong suspicion that LLNL employees were experiencing an

unusual melanoma incidence. Under these circumstances, an posteriori probability calculated from much of the very same data must be interpreted with caution. However, the observation has now been replicated on new data by finding the same continuing high rates for a second, subsequent six-year period. Overall, the probability that chance caused the sustained 12-year elevated incidence of malignant melanoma at LLNL is infinitesimally small, making this hypothesis essentially untenable.

The Infectious Hypothesis: that an infectious agent is the cause. Under this hypothesis an unknown virus (or some other undefined organism) capable of causing melanoma infected LLNL employees sometime before 1972. To explain the sustained high incidence of melanoma, the agent must be either propagating slowly among susceptible employees or displaying long, variable latency in its effect. This hypothesis cannot be ruled out, but is rendered unlikely by the failure of any previous study, on humans or animals, to find such an agent. Another counterargument might be the normal rates found by Austin for the city of Livermore which contains thousands of LLNL family members who should also have been susceptible to the agent. However, the more recent study by Hiatt and Fireman uncovered a suggestion that family rates may be showing the same roughly three-fold elevation that is found in employees.

The Hypersusceptibility Hypothesis: that the LLNL population includes an unusually large pool of melanoma-sensitive persons. Examples of this hypothesis might be the presence of an overabundance of blond, blue-eyed, hazel-eyed or easily sunburned people at the Laboratory. The LLNL population would have to differ significantly from the surrounding population in the area, and from the otherwise closely similar professional population at Los Alamos. Neither comparison has been tested directly, but both seem unlikely. (See discussion of Table 8.) Major differences from Los Alamos are also unlikely in view of the continued high rate of interchange of scientists between the two laboratories and the fact that both organizations are hiring from the same professional pool. In any case, the Austin case-control study showed the typical presence and frequency of predisposing factors found in other studies of melanoma. Livermore cases and their pathology materials have been examined by Dr. Richard Sagebiel of the Melanoma Clinic at University of California, San Francisco, and apart from the thinness of the lesions, there is nothing unusual about them. In particular, there is no evidence of the inherited nevus syndrome, a well known, but unusual genetic predisposer for malignant melanoma. The hypothesis of hypersusceptibility cannot be ruled out but seems unlikely to be anything more than a minor contributor to LLNL's excess of cases.

The Occupational Hypothesis: that an occupational agent is causing melanoma at LLNL. This hypothesis focuses on radiation and chemicals, and is reinforced by the varied, extremely technical and often exotic nuclear and energy-related activities at the Laboratory. There is no known connection in the literature between melanoma of the skin and ionizing radiation or specific chemicals. The wide variety of jobs performed by the melanoma cases at LLNL implies that the putative causative agent must be

ubiquitous. Thus one would have to explain the presence at LLNL of a widely distributed agent which is absent, or present to much lesser degree, at Los Alamos. Attempts to identify candidate agents for this role have so far been unrewarding, but are continuing.

The Solar Hypothesis: that LLNL employees are uniquely exposed or poorly adapted to sunlight. Sunlight is the only generally accepted etiologic agent for human melanoma. Melanoma rates in the U.S. and Australia correlate with proximity to the equator, however for any particular location, outdoor workers have lower melanoma rates than indoor workers (Armstrong, 1984). This apparent paradox is attributed in the literature to differences in adaptation to sunlight. The indoor worker is exposed intermittantly, primarily during leisure activities, and is subject to sunburn, while the outdoor worker is consistently exposed, is protected by clothing, tanning or other internal and external factors, and rarely sunburns. Livermore is a sunny location, and the employees, while generally working indoors, are prone to outdoor leisure activities. The Livermore valley has more sunshine than much of the Bay Area, including the two counties used as controls for Austin's study of melanoma incidence. However, that study carefully matched controls to employees by census tract of residence. Also, the Hiatt study compared LLNL melanoma rates with those of the University of California Davis, and again showed the roughly three-fold difference. Davis, if anything, is sunnier than Livermore. Los Alamos is at lower latitude and a 2000m higher altitude than LLNL, and is sunny; all of which points to a higher potential for solar exposure at Los Alamos than at Livermore. Patterns of dress and leisure activities at the two laboratories are superficially similar, but have not been studied in detail. These comparisons between Livermore and its surround or Livermore and Los Alamos argue strongly against solar intensity or solar adaptation as an explanation for the high melanoma incidence among LLNL employees, although one cannot rule out the small possibility that a subtle aspect of LLNL exposure patterns is somehow having a major effect.

The Surveillance Hypothesis: that the high rate of melanoma at LLNL is primarily due to the aggressive detection of very early lesions. Early detection of LLNL cases has clearly been a consequence of increased medical surveillance and employee awareness following the observed and widely discussed initial cluster of melanoma cases in the mid 1970s. For any progressive disease, early detection increases incidence transiently, with the duration of the transient being roughly the average time between early detection and ordinary detection of the disease. It is unlikely for melanoma that this interval is as long as 9 years which is the minimal duration of the transient at LLNL assuming that early diagnosis began in 1975. However, the surveillance hypothesis can be further elaborated by arguing that aggressive detection of very early lesions carries with it the probability of labelling as melanoma a variety of lesions that are evanescent, nonprogressive, or otherwise unlikely to ever be clinically significant. The literature contains evidence that the growth of human melanoma can be erratic, and that well-established lesions often show sectored areas of spontaneous regression (Ariel, 1981). Little is known of the progression of very early lesions because when identified they are

routinely and promptly excised. According to the elaborated hypothesis, the institution of early detection and removal would cause a spike of cases followed by a stable increased level of melanoma incidence. The height of the new, stable level is determined by the ratio of nonprogressive to progressive lesions. For example, a three-fold increase in melanoma incidence is explained by the detection of three nonprogressive lesions for every progressive lesion. Early detection and excision should reduce mortality because of the early removal of the fraction of lesions that are progressive. There is no way at present to distinguish progressive from nonprogressive lesions, hence the physician has no choice but to remove all lesions. According to this hypothesis the crucial difference between LLNL and LANL is the triggering event which happened in the mid 70's at LLNL and has not happened at LANL. This hypothesis could also explain an absence of increased melanoma incidence in former employees if they escape heightened surveillance.

There is presently insufficient evidence to establish any one hypothesis, although one, the statistical hypothesis, can be discarded, and several others, the infectious hypothesis and the solar hypothesis, seem unlikely. Our personal priority would be to place the surveillance hypothesis first, the occupational hypothesis second, and the remaining three viable hypotheses a distant third. (Additional evidence in support of the surveillance hypothesis is provided by the results of a recent telephone survey of a random sample of current LLNL employees. This survey found that 23% of respondents had a skin biopsy.) Various combinations of the hypotheses are also a possibility.

The hypotheses are useful for suggesting strategies of research. To test the surveillance hypothesis we would recommend that the National Cancer Institute (or some corresponding body) select a white population with normal melanoma rates and deliberately subject them to heightened surveillance. The hypothesis predicts that melanoma rates would increase. We see no good way to test this hypothesis on LLNL employees, although it may be possible for the Laboratory to find an outside group willing to act as controls by submitting itself to the Laboratory's level of surveillance. The occupational hypothesis can be tested in a variety of ways: by case-control examination of occupational factors in LLNL and other high incidence populations, and by a detailed comparison of chemical inventories at Livermore and Los Alamos, looking for the set of candidate chemicals that might be responsible for the difference in melanoma rates. Similarly for the susceptibility hypothesis, the two laboratory populations can easily be compared for hair and eye color, for sunburn susceptibility, and for any other candidate mechanisms for heightened susceptibility. Livermore could also be compared to the surrounding community. For the infectious hypothesis and for general reassurance, it is important to keep track of family incidence as well as incidence in the communities immediately surrounding the Laboratory. Parenthetically, if the surveillance hypothesis is correct, family and surrounding rates should increase as awareness of melanoma pervades the community. The solar hypotheses might be approached by comparing patterns of outdoor leisure activities, tanning, and sunburn in the two Laboratories.

ACKNOWLEDGMENTS

This study could not have been carried out without the help of many people. Special appreciation for their contributions go to Ora Lowe for patiently transferring information from employee badges to computer files and for tracking down former employees, to Bob Patton for efficiently gathering information from employee records, to Rodger Johnson for dosimetry information, and to members of the melanoma task group under the leadership of Dr. R. Lowry Dobson for constant advice and unfailing support.

REFERENCES

ACQUAVELLA, J.F., TIETJEN, G.L., WILKINSON, G.S., KEY, C.R., and VOETZ, G.L. (1982), Malignant melanoma incidence at the Los Alamos National Laboratory. Lancet i, 883-884.

ACQUAVELLA, J.F., WILKINSON, G.S., TIETJEN, G.L., KEY, C.R., STEBBURGS, J.M., and VOELZ, G.L. (1983), A melanoma care-control study at the Los Alamos National Laboratory. Health Physics 45: 587-592.

ARIEL, R.M. (1981), Malignant Melanoma Appleton-Century-Crofts, New York.

ARMSTRONG, B.K. (1984), Melanoma of the skin. Bt. Med. Bull. 40: 346-350.

AUSTIN, D.F. (1980), A study of cancer incidence in Lawrence Livermore National Laboratory employees. Report #1, Malignant Melanoma.

AUSTIN, D.F. (1981), Cancer incidence and mortality in San Francisco-Oakland SMSA, 1973-77. In Surveillance, Epidemiology, and End Results. NCI Monograph 57. National Cancer Inst., Bethesda, MD.

AUSTIN, D.F., REYNOLDS, P.J., SNYDER, M.A., BIGGS, M.W., and STUBBS, H.A. (1981), Malignant melanoma among employees of the Lawrence Livermore National Laboratory. Lancet ii, 712-716.

AUSTIN, D.F., and REYNOLDS, P. (1984), A case-control study of malignant melanoma among Lawrence Livermore National Laboratory employees. Report #3.

AUSTIN, S.F. and SCHNATTER (1983), A cohort mortality of petrochemical workers. J. Occup. Med. 25: 304-312.

GELLIN, G.A., KOPF, A.W., GARFINKEL, L. (1969), Malignant melanoma a controlled study of possibly associated factors. Arch. Derm. 99: 43-48.

GILBERT, E.S. and MARKS, S. (1979), An analysis of the mortality of workers in a nuclear facility. Rad. Res. 79: 122-148.

HIATT, R.A. and FIREMAN, B. Malignant melanoma in the Kaiser Foundation Health Plan of Northern California: A comparison of incidence and measures of health care utilization between Lawrence Livermore National Laboratory and surrounding health plan groups. Dept. of Med. Methods Res. Kaiser Foundation Research Inst. Oakland, California.

HOAR, S.K. and PELL, S. (1981), A retrospective cohort study of mortality and cancer incidence among chemists. J. Occup. Med. 23: 485-494.

HOLMAN, C.D.J., MULRONEY, C.D., and ARMSTRONG, B.K. (1980), Epidemiology of pre-invasive and invasive malignant melanoma in Western Australia. Int. J. Cancer 25: 317-323.

LI, F.P., FRAUMENI, J.F., MANTEL, N., and MILLER, R.W. (1969), Cancer mortality among chemists. J. Nat. Cancer Inst. 43: 1159-1164

LITTLE, J.H., HOLT, J., and DAVIS, N. (1980), Changing epidemiology of malignant melanoma in Queensland. Med. J. of Australia 1: 66-69.
MARSH, G.M. (1980), OCMAP: A user-oriented occupational cohort mortality analysis program. Amer. Statis. 34: 245.
POLEDNAK, A.P. and FROME, E.L. (1981), Mortality among men employed between 1943 and 1947 at a uranium processing plant. J. Occup. Med. 23: 169-178.
SILVERBERG, E. (1984), Cancer Statistics 1985. Ca 34: 7-23.
WRIGHT, W.E., PETERS, J.M., and MACK, T.M. (1983), Organic chemicals and malignant melanoma. Amer. J. Ind. Med 4: 577-581.

Contributed Papers

HAZARD-FUNCTION MODELING OF EARLY EFFECTS MORTALITY RISKS ASSOCIATED WITH LIGHT WATER NUCLEAR REACTOR ACCIDENTS

B. R. Scott, F. F. Hahn, R. G. Cuddihy,
B. B. Boecker, and F. A. Seiler
Inhalation Toxicology Research Institute
Lovelace Biomedical and Environmental Research Institute
P. O. Box 5890
Albuquerque, NM 87185

ABSTRACT

A hazard-function modeling technique is used to derive risk estimators for mortality from specific early and continuing effects of exposure to radiations that could result from an accident at a light water nuclear power plant. The risk estimators allow for the accommodation of dose rate effects. Two modes of exposure are considered: (1) brief external exposure, mainly to cloud-shine and ground-shine gamma rays and (2) protracted internal exposure, mainly to beta radiation from inhaled and ingested radionuclides. Critical organs considered are the bone marrow, gastrointestinal tract, and lungs. The procedure used to develop the risk estimators is generic and, with additional parameters, could accommodate other types of accidents, including those associated with plutonium or thorium fuel cycles.

INTRODUCTION

The U. S. Nuclear Regulatory Commission (NRC) in 1975 issued the Reactor Safety Study (WASH 1400, 1975) which provided quantitative estimates of the health and economic impacts of a light water nuclear power plant accident. In a recent critique of that report (Cooper et al., 1983), additional research needs were pointed out including the need for additional work in developing early mortality models that allow for dose protraction effects.

In this paper, a hazard-function modeling technique is used to develop risk estimators for mortality from specific early and continuing effects of exposure to radiations that could result from a major light water nuclear power plant accident. Two modes of exposure are considered: (1) brief exposure mainly to external cloud-shine and ground-shine gamma rays and (2) protracted internal exposure, mainly to beta and gamma radiations. The beta and gamma radiations are treated as being equally effective for similar dose rates.

The critical organs for lethality considerations are the bone marrow, intestine, and lungs. Dose to these organs are evaluated over various time periods of dose delivery to accommodate effects of dose protraction. Uncertainties are addressed by the use of upper and lower bounds.

HAZARD-FUNCTION MODELING APPROACH

The hazard-function modeling approach to be described is based on two assumptions regarding quantal (all-or-none) effects of combined exposure to different radiations: (a) Each radiation involved produces initial damage called critical damage that could lead to the radiobiological effect of interest. (b) Doses of different radiations that lead to the same level of cumulative hazard (or the associated risk) can be viewed as producing the same amount of critical damage and being indistinguishable as far as effects of subsequently administered radiation are concerned.

Based on assumptions (a) and (b), one can define, for a given quantal effect of combined exposure to different ionizing radiations, a global (or overall hazard) h(t) given as a function of radiation-specific doses $d_j(t)$ and instantaneous dose rates $c_j(t)$, for $j = 1, 2, \ldots, n$, at exposure time t as (Scott, 1984)

$$h = \sum_j \int c_j \dot{h}_j(d_j^*)dt. \qquad (1)$$

The integral is evaluated over the exposure time period of concern and for simplification of notation, the limits of integration as well as the exposure-time dependence of the dose rates, doses, and cumulative hazards have been omitted.

The function $\dot{h}_j(d_j^*)$ is the dose derivative of $h_j(d_j)$ evaluated at isoeffect dose d_j^*, where d_j^* is the solution to the equation

$$h(d_j^*) = h. \qquad (2)$$

Conditional on the global hazard h just before dose accumulation in the small exposure time interval (t, t + dt), the product $c_j \dot{h}_j(d_j^*)dt$ insures that an individual exposed to the dose increment $c_j dt$ in (t, t + dt) is treated as responding in the same way as if brought to the same level of hazard h by exposure to a dose d_j^* of the jth radiation; this follows from assumption (b).

Exposure time t appears in Equation 1 for the purpose of evaluating dose increments over the pattern of radiation exposure and should not be viewed as a temporal description of changes in risk. To determine the global hazard, one traverses an n-dimensional dose space (d_1, d_2, ..., d_n). The path taken is determined by the temporal pattern of dose accumulation and may influence the outcome. For example, use of this approach for modeling the combined cell killing effects of sequential exposure to neutrons and X rays has led to the prediction that exposure first to neutrons followed by exposure to X rays would be more effective than reversing the order of the exposures (Scott, 1983). This prediction is supported by experimental data (Masuda, 1960; Scott, 1983; Scott, 1984A) both qualitatively and quantitatively.

Whether analytical solutions to Equation 1 can be obtained depends on the types of radiation-specific cumulative hazards used (Scott, 1984A). Some cases where analytical solutions have been found are summarized in Table 1. In each case radiation-specific cumulative hazards were assumed to be of the same type. Also given in Table 1 are the radiation-specific risk functions r_j and global risk function r which are related to $h_j(d_j)$ and h by

$$r_j = 1 - \exp[-h_j(d_j)] \tag{3}$$

and

$$r = 1 - \exp(-h). \tag{4}$$

Reactor Accident Early Mortality Assessment

We considered the case of a major light water nuclear power plant accident which could lead to brief exposure mainly from external cloud-shine and ground-shine gamma rays followed by protracted exposure mainly from inhaled and ingested beta- and gamma-emitting radionuclides. We develop estimators of cumulative hazard for three possible causes of death from early effects: (i) from injury to the bone marrow; (ii) from injury to the intestine; and (iii) from injury to the lungs.

The integration in Equation 1 is carried out over specific successive exposures: A brief period b followed by various time periods of dose protraction p1, p2 ..., pm, where the m + 1 time periods may differ for each of the critical organs considered. The length and number of time periods can be selected based on available information on the effectiveness of low-LET radiations when delivered over various time periods. Parameters of the cumulative hazard functions may differ for different time periods, but remain constant within a given period. Dose rates will be relatively high during the brief exposure period, and will

TABLE 1. GLOBAL EXCESS RISK FUNCTION r(D) FOR SIMULTANEOUS EXPOSURE TO n DIFFERENT RADIATIONS[a]

	Radiation specific, j = 1, 2, ..., n		
Models	Risk function r_j	Hazard function h_j	Global risk function r
1. Linear	$a_j d_j$	$-\ln(1 - a_j d_j)$	AD
2. Linear quadratic 1	$a_j d_j + b(a_j d_j)^2$	$-\ln[1 - a_j d_j - b(a_j d_j)^2]$	$AD + b(AD)^2$
3. Linear quadratic 2	$1 - \exp[-(a_j d_j + b(a_j d_j)^2)]$	$a_j d_j + b(a_j d_j)^2$	$1 - \exp[-(AD + b(BD)^2)]$
4. One hit	$1 - \exp(-a_j d_j)$	$a_j d_j$	$1 - \exp(-AD)$
5. Multitarget 1	$[1 - \exp(-a_j d_j)]^m$	$-\ln[1 - [1 - \exp(-a_j d_j)]^m]$	$[1 - \exp(-AD)]^m$
6. Power function 1	$(b_j d_j)^m$	$-\ln[1 - (b_j d_j)^m]$	$(BD)^m$
7. Weibull 1	$1 - \exp[-(b_j d_j)^m]$	$(b_j d_j)^m$	$1 - \exp[-(BD)^m]$

a. Products AD and BD are inner vector products of vectors A, B, and D where A and B are row vectors given by $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$ and D is the column vector of the respective doses, from Scott (1984A).

decrease for each of the successive m periods of protraction of the internal dose (WASH 1400, 1975).

With the separation of the integration into various periods and with the assumption that beta and gamma rays are of (approximately) equal effectiveness, Equation 1 reduces to

$$h = \sum_k \int c_{bg} \dot{h}_{bg}(d_{bg}^*)dt = \sum_k \Delta h_k, \qquad (5)$$

where Δh_k is the increment in the global hazard due to the dose delivered in the time period k, c_{bg} is the instantaneous beta-gamma dose rate, d_{bg}^* is the beta-gamma isoeffect dose which would lead to the same level of hazard as h just before dose buildup in (t, t + dt), and $h_{bg}(d_{bg})$ is the low-LET cumulative lethality hazard for a specified critical organ.

Radiation-Specific Cumulative Hazard

For lethality from early and continuing effects of exposure to ionizing radiations, certain conclusions have been made based mainly on studies with laboratory animals (Mole, 1984; Jones, 1981; Scott and Hahn, 1980; NCRP, 1974):

1. The dose-effect relationships are sigmodial.
2. Large doses are required, suggesting the existence of an absolute or effective threshold.
3. The dose-effect relationships are quite steep so that doses which are just sufficient to cause a few deaths do not differ much from the smallest dose which leads to the death of all exposed individuals.

Because Weibull-type risk functions (and associated cumulative hazards) are flexible enough to accommodate each of the above criteria,

provide for a systematic characterization of both mortality and morbidity (Scott and Seiler, 1984), and lead to a convenient way of predicting the combined effects of different radiations when the hazard-function modeling approach (Equation 1) is used, we have adopted the Weibull-type cumulative hazards of the form

$$h_{bg} = \ln(2)(d_{bg}/d50)^{v}. \qquad (6)$$

The parameter d50 represents the median lethal dose and its value determines the location of the dose-effect curve along the dose axis; for this reason we call it the location parameter. The parameter v determines the steepness or shape of the dose-effect relationship and is called the shape parameter. For lethal effects of low-LET radiations, v is generally greater than or equal to 4.

That the Weibull function is quite flexible is demonstrated in Figures 1 and 2, where the hypothetical risk of lethality is given as a function of critical organ dose and varying values of d50 and v. For values of v greater than about 3, an effective threshold exists as indicated in Figure 2.

Estimates of Shape Parameters

Available data based on exposure of humans are too limited for estimating the shape parameters v. However, there is much evidence that the shape of the dose-effect curve for lethality from early effects of bone marrow irradiation is similar for different mammalian species (Mole, 1984; Jones, 1981; NCRP, 1974). We assume the shape of the dose-effect curves for lethality from injury to the bone marrow, intestine, or lungs to be the same for different mammalian species. Data based on exposure of laboratory animals can, therefore, be used to arrive at an estimate of the shape parameters v for lethality from injury to the bone marrow,

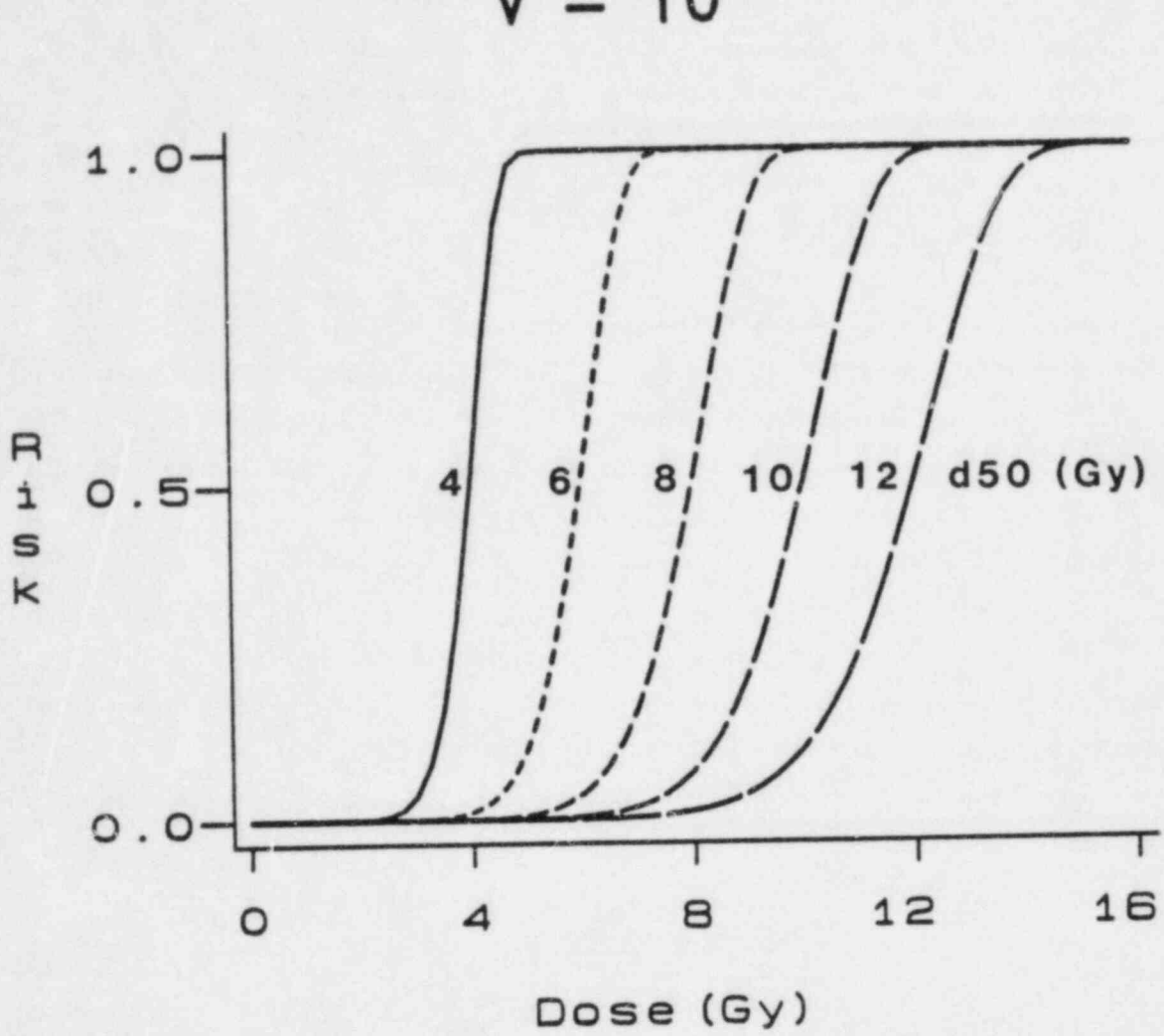

Figure 1. Family of hypothetical risk vs dose relationships based on two-parameter Weibull-type risk estimators with shape parameter v and location parameter d50 in Gy. Parameter v = 10 with d50 taking specified values.

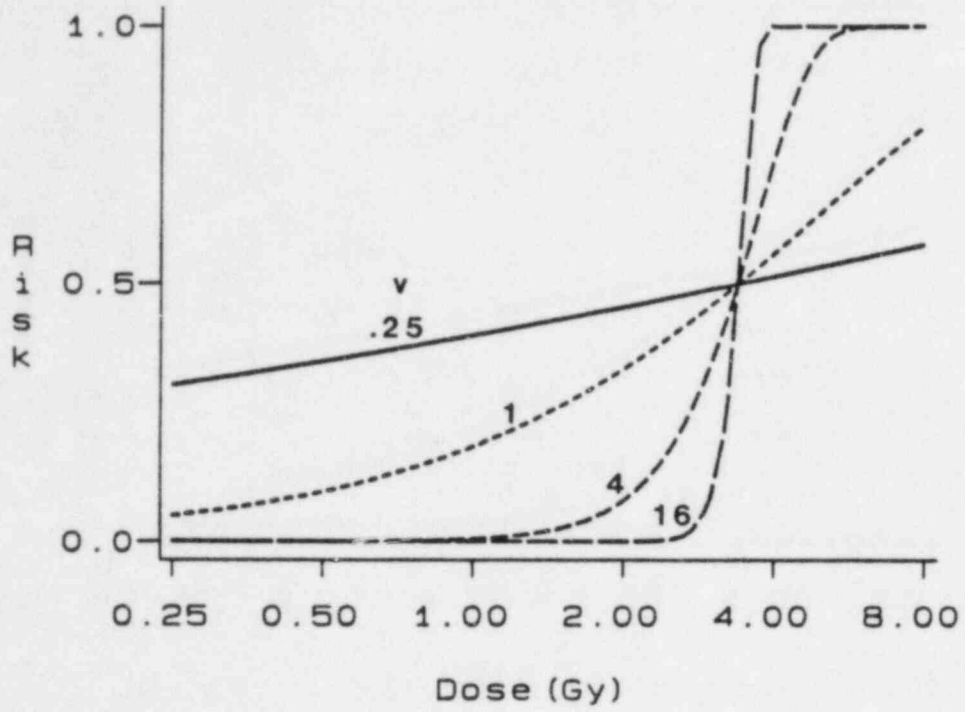

Figure 2. Family of hypothetical risk vs dose relationships based on two-parameter Weibull model with location parameter d50 = 3.4 Gy (current best estimate of d50 for humans after brief total-body exposure to photon radiation) and with v taking on specified values.

intestines, or lungs. Shape parameter estimates arrived at in this way are summarized in Table 2. Also given in Table 2 are d50 estimates for brief exposure.

Dose Protraction Effects

Dose rate influences the d50 in such a way that it increases by a factor proportional to the exposure time to the one-third power

TABLE 2. SHAPE PARAMETER AND d50 ESTIMATES FOR LETHALITY FROM INJURY TO THE BONE MARROW, INTESTINES, OR LUNGS

Critical Organ	Shape Parameter Estimate[a]	d50 (Gy)[b]
Bone marrow	10	3.4[c]
Intestine	10	15[d]
Lungs	4	9.5[e]

a. Estimate for bone marrow based on 60-day lethality data after bilateral photon irradiation of dogs (Michaelson et al., 1968; Hansen et al., 1961). Estimate for intestine is based on data for exteriorized exposure of rat intestine to photon radiation (Sullivan et al., 1959). The estimate for the lung is the mean of several estimates for various patterns of dose protraction to the lungs of dogs after inhalation exposure to beta-emitting radionuclides (Scott and Seiler, 1984; McClellan et al., 1982) and for thoracic exposure of rats (Dunjic et al., 1960).

b. Values are for brief exposure to low-LET radiation.

c. Based on exposure of humans (WASH 1400, 1975).

d. Based on exposure of rats (Sullivan et al., 1959).

e. Based on dose-effect relationship for radiation pneumonitis in humans after brief exposure of the thorax to photon radiation (Van Dyk et al., 1981).

(Lushbaugh, 1982). Available data presented in Figure 3 for lethal effects of irradiation of the lungs suggest that it is reasonable, for reactor risk assessment purposes, to assume that v is independent of the time periods over which the low-LET doses are delivered. Average exposure times ranged from less than one day for upper-body exposure to external photon radiation to greater than one year after inhalation exposure to an insoluble aerosol containing the beta emitter ^{144}Ce. An

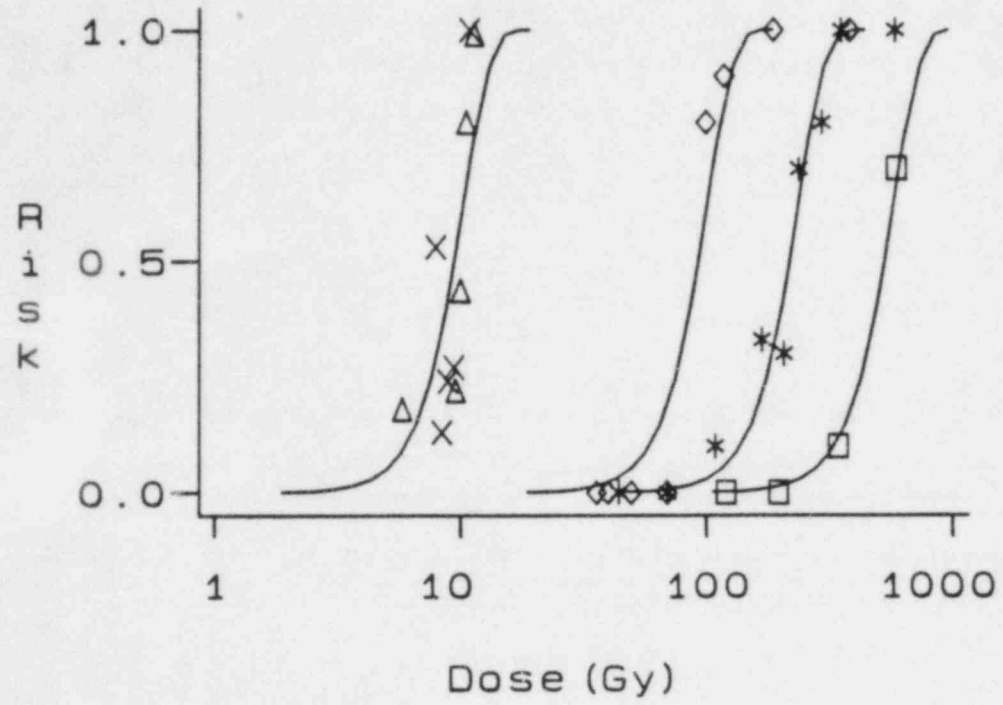

Figure 3. Dose-effect relationships for lethality from injury to the lungs of mammals after brief exposure of the thorax of rats to external photon radiation (triangles), based on data from Dunjic et al. (1960); after inhalation exposure of dogs to the beta emitters Y-90 (diamonds), Y-91 (stars), or Ce-144 (squares) each inhaled in an insoluble aerosol. The inhalation exposure data are from McClellan et al. (1982) as reanalyzed by Scott and Seiler (1984). Also shown are data for the incidence of radiation pneumonitis in humans (x) after brief exposure of the thorax to external photon radiation (Van Dyk et al. 1981).

average value of v = 4 derived from the animal data was used to fit all the curves in this figure. Also shown are data (Van Dyk et al., 1981) for frequency of radiation pneumonitis observed in humans after brief exposure of the lungs to external photon radiation; most individuals eventually died from early effects of lung irradiation. Assuming v in Equation 6 to be constant leads to a very useful solution to Equation 5 given by

$$h = \ln(2)\left(\sum_k \Delta_k/d50_k\right)^v, \tag{7}$$

where Δ_k is the increment in the critical organ dose over time period k, where k = b, p1, p2, ..., pm. Note that the summation in Equation 7 represents the addition of dimensionless doses in units of d50. If we represent this dimensionless d50 dose for each time period k as X_k, then one can define an overall dose in these units as

$$X = \sum_k X_k = X_b + X_p \tag{8}$$

and

$$X_p = \sum_j X_j, \qquad j = p1, p2, \ldots, pm. \tag{9}$$

Equation 8 is a very important result as it allows the reduction of the hypersurface (for n > 2) given by Equation 7 to a function of a single variable X, or as a function of X_b and X_p instead of a function of n variables. Futhermore, any dose combination such that X_b plus X_p equals one would be expected to represent a median-lethal dose for the overall exposure.

Risk surface estimates are shown in Figure 4 for lethality from injury to the hematopoietic tissue of the bone marrow as a function of the brief (high dose rate) external gamma ray dose X_b and overall protracted internal beta and gamma dose X_p in dimensionless d50 units.

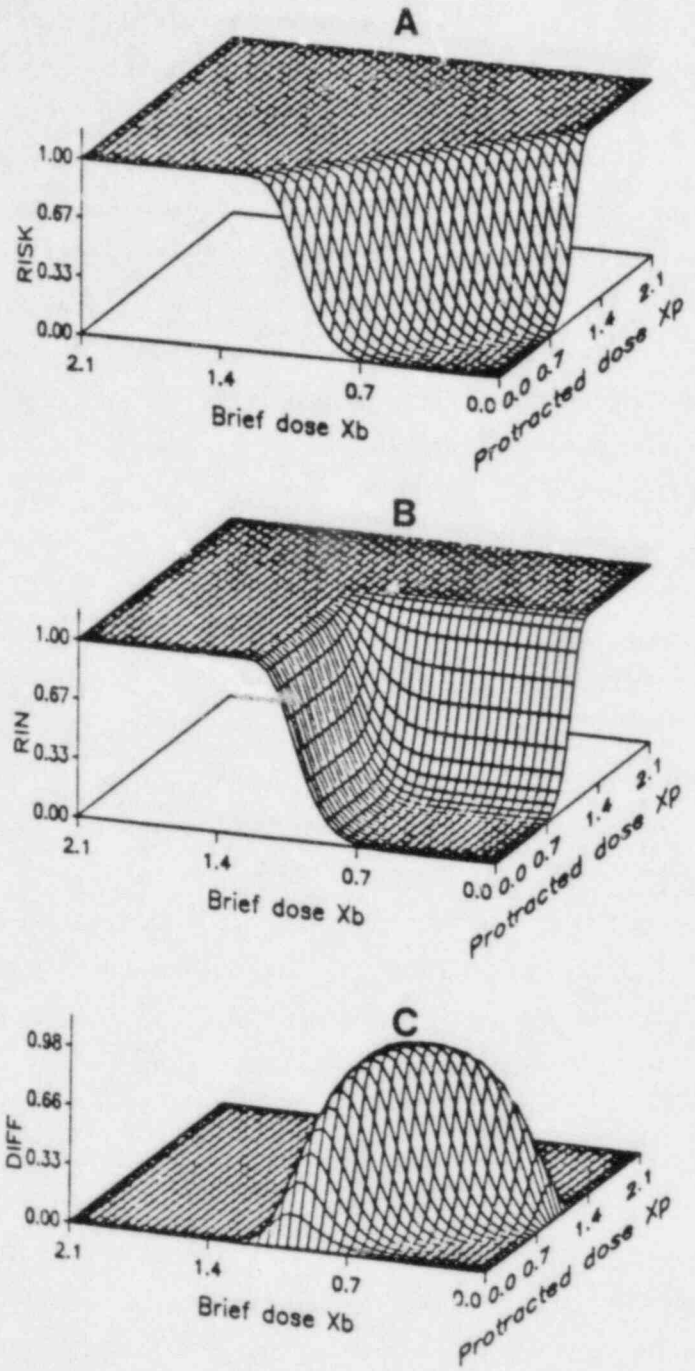

Figure 4. Risk surface estimates for lethality from injury to the bone marrow after total-body exposure as a function of brief Xb and protracted X_p doses in dimensionless d50 units. (A) Dependent effects model based on hazard-function modeling approach discussed in the text. (B) Independent (RIN) effects model. (C) Difference between surfaces in (A) and (B).

The surface in Figure 4A is based on the hazard-function modeling approach. The surface in Figure 4B is based on the independent-effects assumption. The surface (DIFF) in Figure 4C is the difference between those in Figure 4A and 4B and is a measure of the expected enhancement in the effectiveness of the internal protracted dose because of the prior brief high-dose-rate exposure. Such an enhancement is of major concern to regulators and is one of the reasons why the hazard-function modeling approach is becoming more popular. This method of modeling also provides a convenient way of accommodating alpha radiation and therefore could be used for assessing health risk associated with plutonium or thorium fuel cycles (Scott, 1984B).

Multiple Injuries

Adding cumulative hazards for each cause of death leads to an overall hazard for lethality. While this may be viewed as implying the assumption of independent effects of different critical organs, some clarification is necessary. The cumulative hazard for lethality from injury to the bone marrow is based on data for total-body exposure and therefore accommodate nonindependent effects. Cumulative hazards for lethality from injury to the lungs or intestines are based on data for which only a single organ was significantly irradiated. Further research is needed to clarify the importance of interorgan interaction effects.

Risk Estimators

Risk vs dimensionless d50 dose curves are plotted in Figure 5 for death from injury to the bone marrow, intestine, or lungs. Note that when the dimensionless dose X is used, the marrow and intestinal syndrome lethality curves superimpose; this is because the shape parameter estimates are the same for lethality from injury to the bone marrow and for lethality from injury to the intestines. The apparent greater variance for the lung curve may be due to systematic errors associated

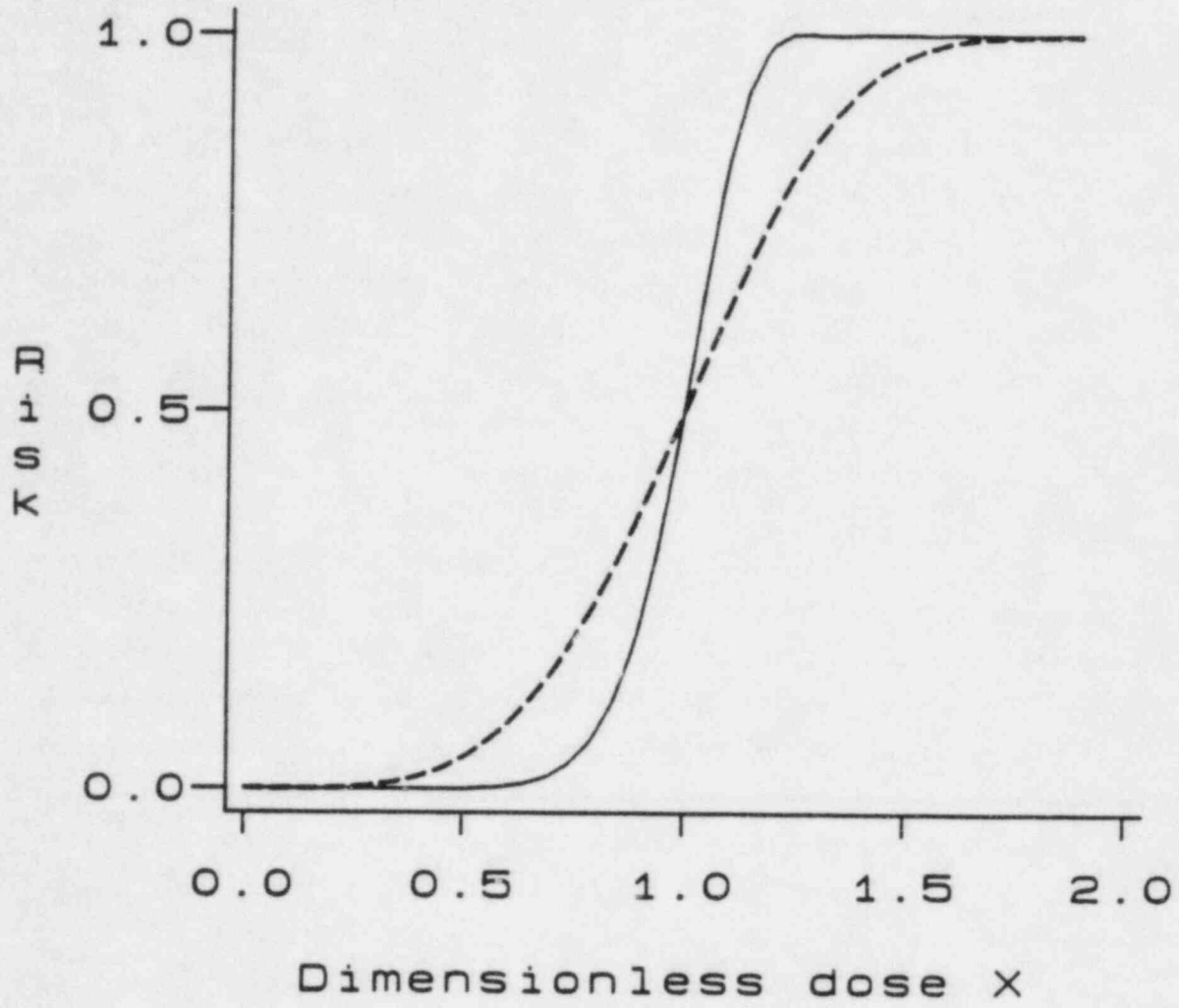

Figure 5. Central risk estimates for lethality from injury to the bone marrow, intestine, or lungs of humans based on information discussed in the text. The smooth curve is for injury to the bone marrow or intestine. The dashed curve is for injury to the lungs.

with grouping of individuals that had different doses in order to arrive at dose-effect relationships after inhalation exposures which resulted in different doses to each individual. With the doses used to arrive at the curves for the bone marrow and intestines, individuals in the same dose group received essentially the same dose of external photon radiation.

Uncertainties

Several uncertainties have been identified in addition to the statistical uncertainties associated with model parameters:

1. Systematic errors which could be associated with the use of Weibull-type risk estimators.
2. Systematic errors which could be associated with method used to account for dose protraction effects.
3. Uncertainties associated with cross-species extrapolation.
4. Uncertainties associated with possible existence of sensitive subgroups within a species.

To account for both statistical and systematic errors, we choose to use upper and lower bounds as indicators of degree of uncertainty. We are concerned about the relatively small number of individuals expected to receive doses large enough to produce fatalities from early effects; therefore, the choice of model is not as critical as it would be for cancer or genetic effect where the large number of individuals that could be exposed to low radiation doses are of concern. Because of the steep threshold type dose-effect relationships, we would expect about the same number of deaths from early effects from all plausible sigmoidal dose-effect models. This is because the probability of receiving a dose in the small region where model predictions differ is small. Recall that for threshold-type relationships, one has to multiply the probability of receiving a given dose times the risk at that dose level and summing these products over the dose distribution for the population to arrive at the average risk. The expected deaths can then be obtained by multiplying the risk set size times the average risk. For threshold-type dose-effect relationships, one should not use person-Sv (or person-rem) population doses.

Because the shape of the early mortality dose-effect relationship is quite similar for different mammalian species, the major contribution to

the uncertainty associated with cross-species extrapolation is uncertainty in the d50. For the most likely accident scenarios, injury to the bone marrow is the most important because lethal doses to the bone marrow are relatively small in comparison to lethal doses to the intestine or lungs (Table 1). For this reason, the uncertainties in the d50 values for lethality from injury to the intestine or lungs will not contribute much to the overall uncertainty in the estimated deaths from early effects. The major contribution will be due to errors in the d50 for lethality from injury to the hematopoietic system.

Estimates of the d50 for lethality from injury to the hematopoietic system range from 2.8 Gy to the bone marrow for the very ill (Lushbaugh, 1967) to 4.5 for the case of protracted exposure (Smith, 1983). For central risk estimates for healthy individuals, a d50 of 3.4 Gy is suggested (WASH 1400, 1975).

There is evidence that the d50 increases by a factor approximately proportional to the exposure time to the one-third power (Lushbaugh 1982). In the absence of data to the contrary, we assume that the uncertainty associated with predicting the effects of dose protraction are relatively small and can be accounted for by the upper and lower bounds discussed below.

Conservative bounds which account for sensitive subgroups can be arrived at and should also account for the other cited uncertainties. We use data for very sick patients with inoperable cancer or terminal leukemia that received brief total-body exposure to photon radiation and which demonstrated a median lethal dose of about 2.8 Gy to the bone marrow (Lushbaugh, 1967). The cumulative hazard for these sick individuals was about 6 times larger than that estimated for healthy individuals, assuming the shape parameter ($v = 10$) to be the same for both groups and using a d50 of 3.4 Gy to the bone marrow for healthy individuals (WASH 1400, 1975). Assuming that the cumulative hazards for

lethality from injury to other critical organs (lungs, intestine) differs by no more than this factor of 6 for sensitive subgroups provides a crude means of quantifying uncertainty for the overall lethality hazard. However, because these bounds are likely to be overly conservative, further research is needed on better methods of accounting for both systematic and statistical errors associated with radiation-effects risk assessment.

REFERENCES

DUNJIC, A.C., MAISIN, J., MALDAGUE, P., and MAISIN, H.A. (1960), "Incidence of Mortality and Dose-Response Relationship Following Partial Body X-Irradiation of the Rat," Radiat. Res., 12, 155-166.

JONES, T.D. (1981), "Hematologic Syndrome in Man Modelled From Mammalian Lethality," Health Phys., 41, 83-103.

LUSHBAUGH, C.C. (1967), "Clinical Studies on the Radiation Effects in Man," Radiat. Res. Suppl., 7, 398-412.

_____, HUBNER, K.F., and FRY, S.A. (1982), "The Impact of Estimates of Human Radiation Tolerance Upon Radiation Emergency Management, in the Control of Exposure of the Public to Ionizing Radiation in the Event of Accident or Attact," in Proceedings of a Symposium held 27-29 April, 1981. National Council on Radiation Protection and Measurements, pp. 46-57.

MCCLELLAN, R.O., BOECKER, B.B., CUDDIHY, R.G., GRIFFITH, W.C., HAHN, F.F., MUGGENBURG, B.A., SCOTT, B.R., and SEILER, F.A. (1982), "Health Effects From Internally Deposited Radionuclides Released in Nuclear Disasters," in The Control of the Exposure of the Public to Ionizing Radiation in Event of Accident or Attack, Proceedings of a Symposium Sponsored by the National Council on Radiation Protection and Measurements, Reston, Virginia, April 17-29, 1981, pp. 28-39.

MOLE, R.H. (1984), "The LD50 or Uniform Low-LET Irradiation of Man," British Journal of Radiology, 57, 355-369.

NCRP 42 (1974), Radiological Factors Affecting Decision Making in Nuclear Attack, National Council of Radiation Protection and Measurements, Washington, D.C.

SCOTT, B.R., and HAHN, F.F. (1980), "A Model That Leads to the Weibull Distribution Function to Characterize Early Radiation Response Probabilities," Health Phys., 39, 521-530.

———(1983), "Theoretical Models for Estimating Dose-Effect Relationships After Combined Exposure to Cytotoxicants," Bull. Math. Biol., 45, 323-345.

———(1984A), "Methodologies for Predicting the Expected Combined Stochastic Radiobiological Effects of Different Ionizing Radiations and Some Applications," Radiat. Res., 98, 182-197.

———(1984B) HAHN, F.F., GUILMETTE, R.A., MUGGENBURG, B.A., SNIPES, M.B., BOECKER, B.B., and MCCLELLAN, R.O. (1984B), "Use of Studies with Laboratory Animals to Assess the Potential Early Health Effects of Combined Internal Alpha and Beta Irradiation," in Proceedings of the 22nd Hanford Life Science Symposium: Life-Span Radiation Effects Studies in Animals; What Can They Tell Us?, Richland, Washington, Sept. 27-29, 1983 (in press).

_____, and SEILER, F.A. (1984), "Mortality and Morbidity Risk Estimators for Quantal Effects of Large Radiation Doses," Paper Presented at 32nd Annual Scientific Meeting of Radiation Research Society, Orlando, Florida. Copies can be obtained from authors.

SMITH, H. (1983), "Dose-Effect Relationships for Early Response to Total Body Irradiation," NRPB-R139, Chilton, Didcot, Oxon.

SULLIVAN, M.F., MARKS, S., HACKETT, P.L., and THOMPSON, R.C. (1959), "X-Irradiation of the Exteriorized of In Situ Intestines of the Rat," Radiat. Res., 11, 652-636.

VAN DYK, J., KEANE, T.J., KAN, S., RIDER, W., and FRYER, C.J.H. (1981), "Radiation Pneumonitis Following Large Single-Dose Irradiation: A Re-evaluation Based on Absolute Dose to Lung," Int. J. Radiat. Oncol. Biol. Phys., 7, 461-467.

WASH 1400 (1975), Reactor Safety Study, An Assessment of Accident Risk in U.S. Commercial Nuclear Power Plants, U.S. Nuclear Regulatory Commission.

ACKNOWLEDGMENTS

We are grateful for Dr. R. O. McClellan and members of the ITRI staff for their strong support of both the experimental and theoretical aspects of this research and to Dr. J. H. Diel for review of the manuscript. This research was performed under U. S. Department of Energy Contract DE-AC04-76EV01013 and supported by the Nuclear Regulatory Commission under an Interagency Agreement.

BAYESIAN METHODS USING INFORMED OPINION

R.V. Canfield, K.T. Chen
Utah State University

ABSTRACT

Bayesian methods provide an intuitively attractive approach to the solution of many engineering problems which require destructive testing of expensive experimental units and for which there exists a history of related experience. These applications of Bayesian methods presume the use of informative priors. The primary difficulty associated applying the method is specification of the prior. A great deal of literature is devoted to the use of noninformative priors and to the use of convenient families (e.g., conjugate families) of priors, but little seems to deal with complete specification.

A method of selecting the prior from within a predetermined family of distributions is developed in this paper. Selection is based upon the concept of information contained in the prior relative to information in the sample. A univariate measure of information is defined for both prior and sample distributions with multiple parameters. The influence of the prior in an analysis is controlled by specifying the ratio of prior information to sample information and the location (e.g., mean value of each parameter) of prior information.

INTRODUCTION

The theory of Bayesian statistics has great intuitive appeal as potential solution to many engineering problems. These problems involve the situation in which limited data is available and at the same time there exists prior information in the form of prior experience. An example is the estimation of soil profile characteristics in geotechnical engineering.

Unfortunately, this intuitive appeal has not lead to overwhelming acceptance of Bayesian methods. In order to use the method it is required that the prior information be transformed in to a prior distribution. The main obstacle in applications is the lack of an effective method for quantifying this prior information. Even if the prior information exists as quantified data from prior tests under similar conditions, there exists the problem of controlling the influence of this data relative to the influence of the present data in the estimation or decision process. The purpose of this paper is to develop a method for quantifying prior information.

The procedures presently published for selecting a prior distribution may be classified into two categories: (1) procedures which result in informative priors, and (2) procedures which result in noninformative priors. Noninformative priors are defined as those which exert negligible influence on the final parameter estimate. The selection of informative priors is addressed in this paper.

The notation to be used is given in the next section. Selection of informative priors follows with the final sections devoted to a generalization of a method of choosing informative priors.

NOTATION

$\underline{X}$	an n-dimensional vector random variable
$\underline{x}$	A sample value of $\underline{X}$
$f_{\underline{X}}(\underline{x},\theta)$	Joint density of $\underline{X}$, indexed by the vector parameter of dimension K.
$\underline{\Theta}$	The vector parameter considered as a random variable for Bayes estimation
$g_{\underline{\Theta}}(\underline{\theta},\underline{\delta})$	Prior density function of Θ, indexed by the vector parameter $\underline{\delta}$ of dimension m.
$\underline{\theta}_p$	Prior point estimate of $E(\underline{\Theta})$.
$h_{\underline{\Theta}}(\theta\underline{x},\underline{\delta})$	Posterior density function of $\underline{\Theta}$ given $\underline{x}$ and $\underline{\delta}$.
$I_{\underline{X}}(\underline{\Theta})$	'Information' in $\underline{X}$ relative to $\underline{\theta}$ (defined subsequently)
$BI_{\underline{\delta}}(\underline{\theta})$	'Bayes Information' in the prior (defined subsequently)
γ	$BI_{\underline{\delta}}(\underline{\Theta})/I_{\underline{X}}(\underline{\theta})$

INFORMATIVE PRIOR DISTRIBUTIONS

Although Bayesian methods have been available for some time, there seems to be very little information of a general nature for use in selecting an informative prior. Most selections are based on ad hoc considerations peculiar to the particular setting in which application is made (e.g.) The major problem seems to be that although the concept of prior information is very intuitive, the structure of prior information can be very nebulous. Thus there is need for a starting point in order to quantify it.

The principle of maximum entropy has been used as an approach to finding the prior. The entropy of a random variable Θ is defined as:

$$-\int_{-\infty}^{\infty} g(\theta,\underline{\delta}) \ln g(\theta,\underline{\delta}) dx$$

where $g(\theta, \underline{\delta})$ is the density of Θ. It may be considered as a measure of the uncertainty or randomness associated with Θ. The distribution on the parameter space with maximum entropy subject to appropriate constraints is considered the least informative prior. This method is also invariant to one to one transformations of $\underline{\delta}$.

One method of obtaining a prior is applicable when the prior information is data from a population with known distribution. This distribution must have as one of its parameters, the parameter to be estimated using the present data. Using the maximum entropy or invariance principle, derive noninformative priors for the parameters of the prior data. Theposterior distribution obtained using the non-informative prior and the prior data becomes the appropriate prior for Bayesian estimation using the present data.

This technique is good when the prior information satisfies the given requirements. Too often, however, the prior data has some deficiencies. For example, the prior information may be data arising from a similar experiment conducted under conditions different from those of the present experiment or historical averages only may be available. In these cases the prior data may not carry as much information as when conditions are identical. In other situations the prior information may be in the form of experience only.

An information theoretic approach provides a framework within which the concepts needed to quantify the prior information may be simply stated. It can be regarded as a maximum entropy (with constraints) estimator, however in most applications entropy will not be a consideration. It is intended that the information required to specify the prior can be understood and supplied by the decision makers, not necessarily the analysts.

This method will be generalized in the next section. In this section the philosophical basis is described. It is assumed that a parametric family of distributions can be found which adequately represents all

possible quantitative realization of the prior experience. Conjugate families may be used here.

A simple quantitative description of the characteristics of prior information is necessary. Two essential characteristics are needed in this approach. First, location of prior information, i.e., point estimates of the value of the parameters is required. The point estimates may arise from data of previous experiments or they can be expert opinion. The second aspect of prior information which is important is the reliability or strength of the information. If it is thought to be weak, it is necessary to limit its influence in determining the posterior distribution. However it is difficult to limit its influence without some measure of its strength.

An absolute measure of the strength of prior information is difficult to quantify. However, strength or reliability measured relative to that of the present data is much more intuitive. A single positive number (ν) is used to express this strength and is interpreted as the overall influence of the prior information compared with the present data in forming the posterior distribution.

GENERALIZED INFORMATION MEASURE

For one parameter Fisher's information is defined:

$$I_{\underline{X}}(\theta)=E\left[\frac{\partial \ln f_{\underline{X}}(\underline{x},\theta)}{\partial\theta}\right]^2 \qquad (1)$$

A heuristic interpretation of (1) is useful in order to develop a compatible measure of information for the prior. Fisher's measure (1) is the average square of the slope (i.e., steepness) of $\ln f_{\underline{X}}(\underline{x},\theta)$ in the parameter space. The steeper $f_{\underline{X}}(\underline{x},\theta)$, the more information provided. In

order to be compatible, information in the prior must also be derived in the parameter space of $\underline{X}$. Thus Bayes information in the prior is defined:

$$BI_{\underline{\delta}}(\theta)=E\left[\frac{\partial \ln g_{\Theta}(\theta,\underline{\delta})}{\partial\theta}\right]^2$$

Note that the derivative is with respect to θ so that Baye's information is defined in the parameter space of $\underline{K}$. This definition leads to an important prope' ty discussed later.

Consider the following multiparameter generalization. For this case bec mes a vector $\underline{\theta}$ and the slope (tangent line) becomes a tangent plane or hyperplane. The slope of any tangent line in the tangent plane which passes through the point of tangency constitutes the family of directional derivatives at that point. Denote these directional derivatives as

$$\frac{\partial \ln f_{\underline{X}}(\underline{x},\underline{\theta})}{\partial \underline{S}} \quad \text{or} \quad \frac{\ln g_{\Theta}(\underline{\theta},\underline{\delta})}{\partial \underline{S}}$$

where $\underline{S}$ is a function of $\underline{\theta}$. The maximum absolute value of the slope is used here in the measurement of information and is denoted:

$$\max_{\underline{S}}\left|\frac{\partial \ln f_{\underline{X}}(\underline{x},\underline{\theta})}{\partial \underline{S}}\right| \quad \text{or} \quad \max_{\underline{S}}\left|\frac{\ln g_{\Theta}(\underline{\theta},\underline{\delta})}{\partial \underline{S}}\right|$$

Thus, the generalized Fisher's information and Bayes' information is given by

$$I_{\underline{X}}(\underline{\theta})=E\left[\max_{\underline{S}}\left|\frac{\ln f_{\underline{X}}(\underline{x},\underline{\theta})}{\partial \underline{S}}\right|\right]^2 \quad \text{and} \quad BI_{\underline{\delta}}(\underline{\theta})=E\left[\max_{\underline{S}}\left|\frac{\ln g_{\Theta}(\underline{\theta},\delta)}{\partial S}\right|\right]^2$$

As in the one parameter case, the relative strength or influence of the prior is γ where:

$$\gamma=BI_{\underline{\delta}}(\underline{\theta})/\ I_{\underline{X}}(\underline{\theta})$$

Given the value of γ, the parameters $\underline{\delta}$ of the prior must be chosen to satisfy (3). It is also desirable to choose $\underline{\delta}$ such that the prior distribution has strength or influence γ with respect to each component of

the vector $\underline{\theta}$. The following lemma provides the basis for establishing this property.

Lemma.

$$I_{\underline{X}}(\underline{\theta}) = \sum_{i=1}^{K} I_X(\theta_i) \qquad \text{and} \qquad BI_{\underline{\delta}}(\underline{\theta}) = \sum_{i=1}^{K} BI_{\underline{\delta}}(\theta_i) \tag{4}$$

Where θ_i is the ith component of $\underline{\theta}$.

Proof

The proof follows immediately from a property of the directional derivative i.e. for any differentiable function h(y)

$$\max_{S} \left| \frac{\partial h(\underline{y})}{\partial S} \right| = \left[\sum_{i=1}^{K} \left(\frac{\partial h(\underline{y})}{\partial y_i} \right)^2 \right]^{\frac{1}{2}}$$

where y_i, i=1,2,...,K is the ith component of $\underline{y}$.

Thus the total information can be partitioned into nonoverlapping parts, each part being the univariate Fisher's information or Bayes' information as the case may be. This property reduces the multiparameter problem to several (K) univariate problems since

$$BI_{\underline{\delta}}(\theta_i) / I_{\underline{X}}(\theta_i) = \gamma \qquad i=1,2,\ldots,K. \tag{5}$$

implies (3). Equations (5) insure that the prior distribution exerts γ relative strength or influence for each θ_i.

ESTIMATION OF INFORMATION

Since $\underline{\theta}$ is unknown, the value $I_{X(\underline{\theta})}$ is also unknown and reliance must be placed on estimates of $\underline{\theta}$. Since $I_{\underline{X}}(\underline{\theta})$ represents information from present data, an estimate of $\underline{\theta}$ from present data will be used. This seems unfortunate since $I_{\underline{X}}(\underline{\theta})$ will then be subject to sampling variation. However on careful scrutiny, the procedure has an error correcting feature which makes estimated information value seem more desirable than exact information.

If the exact value of $I_{X(\underline{\theta})}$ is also unknown and reliance must be reliance must be placed on estimates of $\underline{\delta}$. Since $I_X(\underline{\theta})$ will then be subject to sampling variation. However on careful scrutiny, the procedure has an error correcting feature which makes estimated information value seem more desirable than exact information.

If the exact value of $I_X(\underline{\theta})$ were known and used to determine the value of $\underline{\theta}$ to satisfy (3) then on the average (i.e., over may applications of the method) the relative strength of the prior is γ. However, in most instances one is interested in γ for the particular problem at hand. For this case the estimated information more closely reflects the influence which the present data will have in the posterior distribution. (The posterior doesn't know what $\underline{\theta}$ is either.) Thus, if an estimate $(\underline{\hat{\theta}})$ of $\underline{\theta}$ is such that $I_{X(\underline{\hat{\theta}})}$ is high, then $\underline{\delta}$ chosen so that (3) holds results in a higher value for the Bayes information $BI_{\underline{\delta}}(\underline{\theta})$. Thus, the relative strength or influence remains constant at each application of the method. This feature is illustrated in a later example.

The most intuitively natural method is to substitute for $\underline{\theta}$ its maximum likelihood estimate $\underline{\hat{\theta}}$ into $I_X(\underline{\theta})$. $I_X(\underline{\hat{\theta}})$ may be termed the "expected" Fisher's information (generalized) consistent with its use in one parameter families. However, there may be reason to consider use of the "observed" Fisher's information (generalized). At this point there is no reason to favor one form over the other.

SELECTION OF PRIOR

The condition which the prior distribution must satisfy given γ and the prior point estimates $\underline{\theta}_p$ are

$$BI_{\underline{\delta}}(\underline{\theta}_i)=\gamma I_{\underline{X}}(\underline{\hat{\theta}}_i) \qquad i=1,2,\ldots,K \tag{6}$$

and

$$E(\underline{\theta})=\underline{\theta}_p \tag{7}$$

Thus, the appropriate is a prior in the family $q_{\underline{\theta}}(\underline{\theta}<\underline{\delta})$ whose parameter $\underline{\delta}$ is a solution to equations (6) and (7).

Since the dimensions of $\underline{\theta}$ and $\underline{\delta}$ are K and m respectively, there are 2K equations in m unknowns. In any practical application it is possible to control the prior family so that m = 2K and a unique solution is obtained. If however, m > 2K, it seems reasonable to employ the maximum entropy principle to obtain a unique solution.

DISCUSSION

The case considered in this paper corresponds to the situation in which each component of $\underline{\theta}_p$ has the same strength relative to the present data. The formulation can easily be generalized to the case where the strengths of the components of $\underline{\theta}_p$ differ. Then equations (6) are modified so that

$$BI_{\underline{\delta}}(\theta_i)=\gamma_i I_{\underline{X}}(\hat{\theta}_i) \qquad i=1,2,\ldots,K \qquad (8)$$

Where γ_i is chosen for each i to reflect the strength of prior information with respect to θ_i, i=1,2,...,K. The overall influence of the prior is $BI_{\underline{\delta}}(\underline{\theta})/I_X(\underline{\theta})$.

The generality of this procedure lies in the diversity of prior families which may be considered. For example, suppose the mean (μ) and variance ($\sigma 2$) of the particle size distribution of soil is to be estimated. Suppose also that there have been several previous investigations of this distribution in the region. The overall mean and variance for the region has been estimated and in addition it is found that there is a correlation between the mean and variance. A reasonable choice for the prior family in this case is the bivariate normal-lognormal with correlation. The normal is associated with μ and the log normal is associated with $1/\sigma$.

It is instructive to carry out an estimation example which may be worked using either classical or Bayes methods. The Bayes method is then provided some check on its validity if the Bayes estimates agree with those of the classical method. The following example illustrates this agreement.

Example:

Estimation of the mean (μ) and variance (σ^2) of a normal population from two independent samples on the same population.

Data:

Sample #1 $n_1 = 15, \sum X1 = 4, S1 = 7.5$

Sample #2 $n2 = 15, \sum X2 = 3, S2 = 6.5$

The classical estimates of μ and σ^2 are obtained in this case by pooling the data. The pooled estimates of μ and σ are $\hat{\mu}^* = 3.50$ and $\hat{\sigma}^* = 7.02$.

Consider the Bayes procedure. Assume that the first sample represents prior information. The choice of prior which simplifies computation is a bivariate independent normal-gamma. The prior on μ is normal and the prior on $1/\sigma^2$ is gamma distributed. In order to be useful however, the Bayes estimates should not be sensitive to a reasonable choice of the prior family. To illustrate this insensitivity, an independent normal (μ), log normal ($1/\sigma$) is used.

For this choice of prior, let $\underline{\theta}' = (\theta_1, \theta_2)$ where $\theta_1 = \mu$ and $\theta_2 = 1/\sigma$. Then:

$$g_{\Theta}(\underline{\theta}, \underline{\delta}) = \frac{1}{2\pi\theta_2\tau\beta} \; e^{-\frac{1}{2}\left[\left(\frac{\theta_1-\eta}{\tau}\right) + \left(\frac{\ln\theta_2-\alpha}{\beta}\right)^2\right]}$$

where $\underline{\delta}' = (\eta, \tau, \alpha, \beta)$.

The components of $I_{\underline{X}}(\underline{\theta})$ and $BI_{\delta}(\underline{\theta})$ are

$$BI_{\underline{\delta}}(\theta_1) = \frac{1}{\tau^2} \qquad \text{and} \qquad BI_{\underline{\delta}}(\theta_2) = e^{-2\alpha+2\beta^2}(1+1/\beta^2) \tag{9}$$

$$I_{\underline{X}}(\hat{\theta}_1) = n/S_1^2 = .355 \qquad \text{and} \qquad I_{\underline{X}}(\hat{\theta}_2) = 2n/S_1^2 = .710 \tag{10}$$

Since the sample sizes are equal, $\gamma = 1$ is used. The prior estimate of $\underline{\theta}$ is:

$$\underline{\theta}'_p = (\bar{x}, S_1) = (4, 7.5) \tag{11}$$

By using equations (9) and (10) in (6) and (7), the solution for the parameter vector $\underline{\delta}$ is

$$\underline{\delta}' = (\eta, \tau, a, \beta) = (4, 1.6783, 2.001, .16725)$$

The posterior density becomes:

$$h_{\Theta}(\underline{\theta}\ \underline{x}, \underline{\delta}) = Cg_{\underline{\Theta}}(\underline{\theta}, \underline{\delta}) \frac{1}{2\pi\ \theta 2} e^{-\frac{1}{2}\Sigma\left(\frac{x_i - \theta_1}{\theta_2}\right)^2}$$

where C is such that

$$\int_{\infty}^{\infty} h_{\Theta}(\underline{\theta}\ \underline{x}, \underline{\delta}) d\underline{\theta} = 1$$

Using the mean of the posterior as the Bayes estimate of $\underline{\theta}$, the following value were obtained using numerical integration to evaluate expected values.

$$\mu_B = E(\theta_1) = 3.54 \qquad \text{and} \qquad \sigma_B = E(\theta_2) = 7.04$$

The close agreements between $\hat{\theta}^*$ and μ_B indicates that the Bayes procedure performs as should be expected for the mean and it is seen that the Bayes procedure is equivalent to pooling variance information with no assumptionconcerning population means. This is of course preferred since equality of means is not assured.

REFERENCES

Canfield, R.V., and Teed, J.C. (1977), Selecting the prior distribution in Bayesian Statistics, IEEE Transactions on Reliability R-26 No. 4, Pg. 283-285.

Cox, D.R. and Hinckley, D.V. (1974), Theoretical Statistics, Chapman and Hall, London.

Crellin, C.L. (1972), The philosophy and mathematics of Bayes' equation, IEEE Transactions on Reliability, R-21, pg. 131-135.

Kendall, M.G. and Stuart, A. (1961), The Advanced Theory of Statistics, Vol. II, Charles Grittin & Co. Limited, London.

Raiffa, H. and Schlaifer, R. (1961), Applied Decision Theory, Harvard University Press, Cambridge, Massachusetts.

Sokolnikoff, I.S. (1939), Advanced Calculus, McGraw-Hill, New York.

Soland, R.M., (1968), Bayesian Analysis of the Weibull process with unknown scale parameter and its applications to acceptance sampling, IEEE Transactions on Reliability, R-17, Pg. 84-90.

Vanmarcke, E.H. (1977), Probabilistic modeling of soil profiles, "Journal of the Geotechnical Engineering Division," ASCE, Vol. 03, No GT11, Proc. Paper 13364, Pp. 1227-1246.

Yakowitz, S., Duckstein, L. and Kistel, C. (1974), Decision analysis of a gamma hydrologic variable, Water Resources Research, Vol. 10, No. 4.

RISK EVALUATION IN HIGH-ALTITUDE LEVEL FLIGHT

James A. Lechner

National Bureau of Standards

ABSTRACT

Consideration is being given to reducing the minimum planned vertical separation between aircraft flying above 29000 ft over CONUS, from the present 2000 ft (pressure altitude) to 1000 ft. (The allowed separation is already 1000 ft for aircraft flying below 29000 ft altitude). Special Committee 150 of the Radio Technical Commission for Aeronautics is charged with producing draft performance specifications which will ensure safety at 1000 ft separation. Considerations to be discussed include:

- Delimiting the problem;
- Choice of appropriate measure(s) of safety;
- Choice of criteria for these measures;
- Modeling of the system and its stochastic behavior;
- "Calibration" of the model to reality by estimation of the model parameters;
- Determination of performance specifications to achieve safety in a cost-effective manner, considering the variety of users and desiderata.

The main technical result relates to peaks in a distributed-risk situation. Peaks taken over too-small pieces of the system may be meaningless. A characterization of the risk thoughout the system leads to a determination of the appropriate size pieces to consider.

BACKGROUND

At present, aircraft under control of the Air Traffic Control System over the conterminous United States (CONUS) are assigned to "flight levels" corresponding to multiples of 1000 feet pressure altitude.* Up to FL290 (29,000 feet pressure altitude) every multiple of 1000 ft is used; above FL290, only every other one is used (the odd multiples). Thus the minimum vertical separation betweem passing or overtaking aircraft on a given route is 1000 ft if the aircraft are below FL300, and 2000 ft if above FL290. In addition, over most of the airspace, the "cardinal rule" is applied: the odd multiples up to FL290, and every other odd multiple above FL290, are reserved for eastbound traffic, while the remaining flight levels are reserved for westbound traffic. Under these circumstances, the minimum separation between overtaking (i.e., "same-direction") traffic is 2000 ft below FL290, and 4000 ft above FL290. The larger separation at high altitudes was instituted when jet aircraft began to fly that high, because altimetry systems were judged not to be sufficiently accurate to permit safe operation with 1000 ft separation.

Several attempts have been made to reduce the high-altitude separation, without success. Another, considerably more elaborate, effort is now underway, involving Special Committee 150 of the Radio Technical Commission for Aeronautics, formed in March of 1982 for this purpose. Two major reasons for the desire for decreased separation are the added flexibility it allows to flight controllers (thus easing congestion and presumably increasing safety and/or capacity), and the economic benefits to operators. The economic benefits are connected with the characteristics of modern turbojet aircraft, for which the optimum cruise altitude is generally close to the maximum cruise altitude. At present, with 4000 ft separation between available altitudes in a given direction, fully-loaded aircraft must spend long periods at a given altitude before being able to climb to the next level. If the intermediate altitudes were available, the aircraft could climb 2000 ft as soon as it was able to do so, rather than having to wait until it could climb 4000 ft. It has been estimated that the improved efficiency would save on the order of $100M per year in fuel costs for US operators. In addition, aircraft could be assigned to altitudes closer to their optimal altitudes if overcrowding becomes a problem, because there would be twice as many altitudes available.

*Worldwide, air traffic uses "feet" (and other English units) rather than metric units. We follow this usage for ease of communication within the community.

The mandate to SC150, when it was formed in March 1982, was to "develop minimum system performance standards necessary to safely reduce vertical separation to 1000 ft above flight level 290...identify any needed equipment performance improvements...identify any recommended procedural changes." Considering that about 15 percent of the traffic above FL290 is military, and 15 percent is General Aviation, most of which is not scheduled as predictably as the scheduled air carriers; and that an increasing percentage of the traffic is being given clearance to fly "direct" or "great circle" rather than on the established routes; and that reliable height-keeping data are scarce; the task is nontrivial.

THE PROBLEM

Considering again the mandate, one could detail the task as follows:

1) Decide what is meant by "safely". We need to define, first, what measures of safety to use; and second, what criterion value(s) to adopt.

2) Develop a method for determining whether a system implementation meets the criteria.

3) Choose an implementation that is "acceptable", considering the varied interests of the interested parties.

4) Develop a means of verifying that the implementation has been performed.

5) Show that the implementation actually does achieve the desired safety level.

6) Allow for future changes, at least by monitoring.

Each of these will be expanded in later sections.

THE MEANING OF "SAFELY"

The statistical literature abounds with discussions of risk, its measurement, and its comparison. I will not attempt to review these discussions. The Special Committee agreed, early on, to consider 1000' separation "safe" if the expected number of collisions between

aircraft in level cruise above FL290 over CONUS is sufficiently small. (The probability of one or more collisions is smaller, of course.) It has further been agreed that the risk will be expressed as "expected number of collisions per something", with the "something" often being 10^7 system hours. A system hour is accrued for each flight hour of each aircraft in the system -i.e., in high-altitude level cruise.

From an overall systems point of view, an obvious measure is what I call "total risk": the expected number of collisions anywhere in the high-altitude level-flight portion of the CONUS airspace. The reason is that any collision anywhere reflects on the safety of the whole system; any collision is a "disaster", especially if it involves a passenger aircraft, and more than 90 percent of the passings in this airspace involve at least one commercial aircraft.

Choice of a criterion level, or Target Level of Safety (TLS), is not a technical issue, and therefore is beyond the responsibility of this Committee. However, we cannot come up with the performance standards unless we have a criterion to meet, so we have chosen to work for now with the value used in the North Atlantic study, namely one expected collision per 10^8 system hours. Considerable discussion has centered around this criterion level, comparing it with risk from different activities and from different portions of a flight. However, in the end it remains an arbitrary choice.

PEAK RISK

The concept of total risk does not necessarily capture all the concerns of the Committee. Suppose there were some portion of the system, say a route from point A to point B, for which the risk (per flight hour) were a thousand times as high as the system average. Is this fair to the traveler on that route? We have not attempted to answer this (value-laden) question. However, we are concerned with determining whether there are such relatively high-risk parts of the system. Initially, we looked for segments with a high density of passings, since the calculated risk is a sum of terms proportional to the numbers of passings (of various types). And we found them. But even if they are real, and not artifacts of the data-collection, are they meaningful? Not necessarily. Consider the following simple example.

Suppose there is a route with two hours of high-altitude level flight, which only has two flights a day, in opposite directions, which pass enroute. And suppose that the scheduling is perfect and the winds almost constant, so that in fact these two airplanes always pass each other directly over a certain house. Now if one were to look at this route over very small segments -say 200 foot increments -then there is one passing in that 200 ft. segment, for every flight. And since it only takes 0.00007 hours to traverse that segment, the number of passings per flight hour for that segment is about 15,000! The number of passings per flight hour for this small segment appears entirely intolerable, compared to the passing frequency for the system as a whole, which is less than one per flight hour. And yet this is a safe route, by any common-sense consideration. Something is amiss: it cannot be proper to use such small segments for calculating peak risk. So we need to consider what kind of segment makes sense. The following paragraphs address this question.

CHARACTERIZATION OF RISK-MEASURES

This section is devoted to determining what a normative or prescriptive approach can tell us about risk measures. In other words, we ask the question, what must a measure of risk look like, to be sensible? Usually this approach does not produce a complete specification of the risk measure; rather, it "narrows the field" by eliminating some measures from consideration, and also enables us to obtain conclusions which should hold whatever risk measure is eventually chosen for use.

In order to simplify the stochastic process approach, consider a hypothetical system where a "collision" is counted, but does not interrept the flight. This hypothetical system is identical to the actual system, up to the time of the first (if any) collision. Since we are talking about very small risks of collision, it seems reasonable to talk about this hypothetical system instead; the actual system might well change immediately upon the occurrence of a collision, anyway.

Now consider an individual aircraft on the system. Its behavior consists of a succession of takeoff-flight-landing sequences. Since we are concerned only with the high-altitude level cruise portions, let us use the term "leg" to denote the entire high-altitude level cruise portion of a flight between one takeoff and the next landing.

Finally, let us introduce some notation. Let $m(t)$ be the expected number of collisions suffered by this aircraft up to time t, where $t=0$ at takeoff. Let $m=m(t_o)$, where t_o is the time of landing. Let m/T denote the "leg risk", i.e., the expected number of collisions for this aircraft on this leg divided by the corresponding flight time T. Finally, let $\lambda(t)=(d/dt)m(t)$ denote the intensity function for the random process $N(t)$ whose value is the actual number of collisions suffered by the aircraft up to time t. (Under reasonable assumptions, $N(t)$ can be considered to be a generalized Poisson process, which has the value 0 as long as no collisions have occurred.)

Now suppose one were to perturb the system somewhat, for example by modifying the schedules or the altitude or route assignments, but in such a way that all the leg risks remained unchanged. (The $\lambda(t)$ functions might change drastically, but their integrals over legs would not change.) Then there is no reason (based on safety considerations) for any passenger, pilot, or airline to prefer either version of the system to the other. They are "risk-equivalent", since the risk of passage is the same for any passenger, pilot, etc. But then it follows that <u>any</u> reasonable risk-measure should assign the same risk to each. Thus risk-measures should depend only on the leg risks, and not at all on the peak heights of the $\lambda(t)$ functions.

If we know $\lambda(t)$, we can calculate m. Now the <u>average</u> risk is just m/T, where T is the total time spent in high-altitude level flight. And since we are assuming that T is known, m is sufficient to determine the risk. By simple extension, knowing m/T for every leg is tantamount to knowing the collision risk for every leg (and thus for every passenger, every flight, and the whole system), so we shouldn't have to look further than m/T.

In the real system, an aircraft may spend part of a leg on a very lightly-traveled route segment with a very small $\lambda(t)$, and another part on a heavily-traveled segment with a large $\lambda(t)$. In that case, it might be easier to calculate its λ by determining an average λ for each segment and calculating a weighted average of the two. It would <u>not</u> be appropriate to look at the larger λ to define the peak risk, unless there were another flight for which one leg consisted only of that higher λ; because only in this latter case could a passenger experience an average risk corresponding to the higher λ, and the average risk over the leg is what determines the probability of safe arrival. (Note again the example in the previous section: it's not the 15,000 passings per flight hour on that 200-ft segment, but the <u>one</u> passing in two flight hours, that determines the risk.)

Taking a cue from the parenthetical comment just above, consider another way of looking at this question. Note that risk comes from passings: no passings, no risk. In fact, the total risk is proportional to the number of passings. More precisely, the risk consists of two main terms, one related to same-direction and one to opposite-direction passings; each is proportional to the corresponding number of passings.*

(We have ignored passings at more than the minimum spacing here, but the terms corresponding to them are each again proportional to the number of such passings.) Now since the number of passings is the determining factor, it can't matter where on the leg the passings occur, and again only the average number of passings per hour matters.

This conclusion seems to make some people uncomfortable. Thus we should seriously consider why, for at least two reasons: first, gut feelings often lead to insights about the relation between the model and reality, sometimes leading to a better model; and second, even if the gut feelings don't stand up to inspection, they are real, and need to be addressed in the real (political and psychological) world. Several reasons for possible discomfort come to mind; perhaps there are more.

1) There seems to be a difference between <u>actual</u> and <u>perceived</u> risk. If for example passings occur frequently at a given point, then perhaps they are more likely to be noticed; in turn, then, the impression might well occur that the risk is higher than if the same number of passings were distributed more broadly across the flight. No doubt, perceived risk is important - but the subject of <u>this</u> paper is actual risk. The problem of differences between fact and perception is common; good communication should help.

*The proportionality discussed here holds precisely for the Collision Risk Model treatment of on-route traffic, where traffic passes in parallel. Traffic cleared to fly "direct" (off-route) passes traffic <u>on</u> a route at other angles. For such passings, the angle matters, too. These passings constitute a growing fraction of total passings, and need to be treated (by an extension of the CRM?), unless it can be shown in general that moving traffic off route can only decrease the risk.

2) Even though risk depends on average exposure, there are other reasons for looking at "hot spots". For example, if one is looking for ways to improve the saftey of a system (as contrasted to measuring the safety), places where passings occur with high frequency are obvious places to look, since they hold
out the hope of achieving greater return per effort invested. This is a legitimate concern, but it does not contradict the thesis that risk assessment should involve whole-leg averages.

3) Perhaps the model is wrong, in some important way. Consider: If a pilot performs an emergency measure to avoid a collision, he can only maneuver into a different aircraft if there is another plane there (besides the one he is attempting to avoid); that says that passings are less dangerous if they are widely separated, and indeed that must be true. Our (Collision Risk) model takes no account of evasive maneuvers nor does it include controller intervention. Thus our model tends to overcompensate for this possible problem, even though it treats all passings as equal; ignoring the above-mentioned effect of bunching on evasive maneuvers is more than compensated for by ignoring the possible increase in safety due to controller intervention. And further, there are other effects of bunching: perhaps it makes pilots and/or controllers more attentive, thus contributing to safety?

4) There is a concern for risk to people/property on the ground. This concern may well be justified, but as in 2), it's not part of this problem.

BOUNDS FOR PEAK RISK; SEGMENT RISK.

We have argued that any measure of system risk should depend on the average over whole legs. By a pathological (extreme) example, we have illustrated that averages over shorter segments may give a misleading impression. Naturally, it is difficult to measure parameters like density, number of passings, etc., over a whole leg: they are usually estimated from flight strips which refer to movement within an air traffic control sector, and (with the present system) can only be processed and tabulated manually. Of course, if there is a leg whose high-altitude level flight portion coincides with a particular segment, then the segment risk for that segment is a leg risk. Perhaps this is true for most (or rather, for the peak-risk) segments. If so, then the peak segment risk is the peak leg risk. The important consideration is to avoid averaging over such a short segment that the average is unrepresentative of any leg-average.

How might we use segment risks to get at peak leg risks? Segments with high density lead to upper bounds for the leg risk. We could systematically look at the highest of these, expanding the size of the segments to get better approximations to the leg risk, until the segment with highest risk is a whole leg; we have then found the highest "peak risk" in the system. Another possibility is to expand the sizes till all legs have been shown to have risk below a certain threshold. This might be possible without going nearly all the way to whole legs. Note that the peak whole-leg risk cannot be larger than the peak obtained by breaking the entire system into segments smaller than a whole leg. Therefore, any breakup of the system into smaller segments necessarily results in an upper bound for the peak risk, as long as the whole system is included.

WHICH PEAK SHOULD WE BOUND?

Is it the absolute highest peak we should put a limit on? Even if it is one really short segment which experiences a peak on one day of the year, at one particular time, say when there is a fly-in? Or should we limit the peak of "normal" flights? And should it be a peak of the daily averages, or individual flights? How can we calculate any of these? Or should it be the peak (risk • total flight hours), a rough indicator of the contribution of a given route to the total, when risk is measured per flight hour? What kind of weather should we assume?

Finally, since it is not feasible to go out and measure density on every route at all times, we cannot really estimate system risk by measuring every leg (but see the later Section entitled Simulation). We can, however, make good guesses as to which are peak-risk routes, from judgment based on the experience accumulated in the Air Traffic Control Centers. So we have hopes of getting reasonable estimates of "peak risk". Such an estimate is, of course, an upper bound on average (or total) risk; but surely an extremely conservative one. Is there some way that we could find a factor, say K, such that bounding the peak risk to be less than KR will bound the average risk to be less than R?

SETTING SPECIFICATIONS.

This Committee is charged with developing specifications which will allow 1000ft separation above FL290. The realities of life are:

a) We don't know how to go directly from a specified safety level to a (minimum) performance specification which will ensure that safety level.

b) We can (via the Collision Risk Model, with various assumptions applied) go from performance characteristics to an estimate of the achieved risk (even though the estimate is thought to be a very conservative number, for given performance levels).

c) We need to use things like altitude-utilization models, occupancy models (or extensive data), flight schedules, etc. to obtain the estimates of risk.

d) We can iterate: i.e., come up with a specification, calculate the risk, check it against a target or targets, then modify the specification and repeat the cycle.

e) Since the specification is inherently multi-dimensional, one could easily spend a lot of time going round this loop; one needs to use good engineering judgment to come up with sensible specifications which will do the job at reasonable "cost".

f) Performance specifications are not the only option. We might want to (and are charged to) state any operational changes or restrictions that are necessary in addition to the performance specifications. Options include, of course, separating a high-density route into two routes; making bidirectional routes into unidirectional routes; and possibly putting restrictions on how (certain classes of) aircraft use the airspace. The shift toward more direct-clear flights will need to be considered, too.

g) There is a difference between assessing the safety of a given mix of aircraft in a given situation, and evaluating a specification. In the former case, one takes account of the different performance levels of the different aircraft. In the latter, one is wise not to assume anything beyond the legal requirement (witness the recent appearance of inertial navigation systems which do not far exceed the performance requirements, but do cost less). It is simpler to assume only a given level of performance. One does not need to know the details of performance for the specific aircraft on a given segment, in order to calculate a risk. In other words, the performance and the traffic pattern have been decoupled. Also, the overall performance is simpler to represent, no longer being a mixture of several different distributions. But on the other hand, one cannot take advantage of the fact that many aircraft far exceed the average performance.

h) Ought we to consider recommending specifications which differ for different aircraft types? Consider the following scenario. Suppose we count on no better than the minimum performance specification, and determine a specification which meets the desired safety level. And suppose that achieving that height-keeping performance turns out to be quite unreasonable for (say) high-performance aircraft. And finally, suppose there are very many aircraft (for instance, all large commercial jets) which already far exceed that performance. Perhaps there is a higher standard of performance, which is already being met by this latter class, and a lower standard, which can reasonably be met by the other aircraft, which together will ensure that the desired safety level is met. If so, why not go that way?

SIMULATION

Aircraft passings are (relatively) infrequent events, and there is considerable apparent randomness in aircraft movements and schedules. Therefore it is hard to get good estimates of passings from actual observation. There is another approach. Suppose we had reasonably good estimates of total number of flights (by origin and destination), schedules, and flight planning procedures. Then by using a simulation program, including randomization, one might well come up with better estimates of overlap frequencies. And of course, it would be possible to do sensitivity studies, to pinpoint the weak spots of such an approach; one thus finds out which kinds of data are insufficient to support this activity. Furthermore one can implement system changes and evaluate the effect; this includes (but is not limited to) the distribution of preferred altitude in a 1000-ft environment.

VERIFICATION

The Committee is still deliberating the proper form of verification and monitoring. Certainly a sizeable data-collection effort will be mounted (and is already being planned), using precision radars to measure actual height and/or height differences. This will be applied first to aircraft which do not meet the proposed performance specifications, to verify that the specifications do what they are intended to do. It will then be applied to all aircraft, to verify that they all do at least that well, and to ensure that there are no surprises. Finally, after implementations, the same sort of technique will be applied in some fashion yet to be decided, to ensure the continued safe operation of the high-altitude system.

ACKNOWLEDGEMENT

The work of the author has been partially funded by the Federal Aviation Agency, US Department of Transportation. This paper is an input to the Committee's deliberations, not solely the work of the author, but presented through the author's eyes; the responsibility for the views expressed therein is thus solely the author's.

REFERENCES

1. Gillsinn, J. F. and Shier, D. R., Mathematical Approaches to Evaluating Aircraft Vertical Separation Standards, NBSIR 76-1067, National Bureau of Standards, May 1976.

2. Busch, A. C., Colamosca, B., and VanderVeer, J. R., Collision Risk and Economic Benefit Analysis od Composite Separation for the Central East Pacific Track System, FAA-EM-775, Federal Aviation Administration, June 1977.

3. Preliminary Estimate of System Performance Requirements Necessary to Support a 1,000-Foot Vertical Separation Standard At and Above Flight Level 290. Technical Note DOT/FAA/CT-TN83/41, Feb. 1984.

LAPLACE'S LAW OF SUCCESSION AND PREDICTION INTERVALS*

David Rubinstein
U.S. Nuclear Regulatory Commission

ABSTRACT

Several conservative prediction intervals for the time of occurrence of future (untoward) events of a Poisson process are derived as functions of the number of events observed to the present. The form of the prediction intervals considered here provides a limit such that the future event of interest will fall beyond the stated limit with prescribed confidence. In case of zero observed events to the present, at least one of the prediction intervals resembles LaPlace's law of succession; e.g. the sun will fail to rise on the next day with probability 1/(n+1) if it rose during the n preceding days. In fact one can consider the prediction intervals as a rigorous formulation of LaPlace's law of succession. Prediction intervals provide a simple and appealing alternative to quantifying risk by means of confidence limits for occurrence rates. Several practical applications are suggested.

INTRODUCTION

One of the early efforts at numerical risk assessment was LaPlace's determination of the probability that the sun would rise on the next day given his postulation that it had risen for the previous consecutive n = 1,826,213 days. His betting odds were n/(n+1) (Feller 1950, pp. 83-85). In modern times his formulation of the law of succession, which addresses the sun rise problem, is regarded with considerable skepticism. Uspensky (1937, pp. 68-71) derives the law of succession by Bayes' formula and the assumption of a uniform prior distribution for the probability of sun rise. Feller derives it on the basis of a "somewhat artificial" urn scheme. Both authors argue that the law of succession does not represent a persuasive basis for inference. When addressing a practical modern problem, I obtained a result that had a resemblance to the law of succession with a formulation that was free from the postulation of a prior distribution or artificial schemes. The problem posed by a statistical layman was a request for "the correct way to state the statistical significance.... of 400 reactor years of operation without major accident." One of the standard ways of dealing with this problem is to calculate confidence limits for the occurrence rate of major accidents by well established methods based on the

---

*This paper was prepared by an employee of the United States Nuclear Regulatory Commission. It expresses opinions that do not necessarily represent a staff position of the NRC. The report has been neither approved nor disapproved.

Poisson distribution of the number of observed events. I opted for a prediction interval for the time occurrence of the next (and this case the first) major accident. Parenthetically, I like to observe that all of this happened just prior to the Three Mile Island accident when no major accidents had occurred yet.

There are noteworthy differences between the confidence intervals and prediction intervals and, in fact, there are also differences between focusing on rates and occurrence times. These differences may be particularly important in the perceptions of lay statisticians.

a. The time of occurrence of an event is more concrete than its occurrence rate. The choice of formulation may have different psychological impact. If one learns that a given catastrophe occurs at the rate of one in a 100 years, the psychological reaction may be that this is remote from me and my children. If one hears, there is a 10% chance that a catastrophe will occur within the next 10 years, one might find the threat worrisome.

b. The confidence interval involves two probabilities one related to the rate of occurrence, the other to the confidence level. The prediction interval essentially only involves confidence. Thus, the prediction interval provides a simpler statement. Interpretation of a prediction interval by a layman is likely to come closer to the mark than that of a confidence interval. In fact it may as well have more meaning to the statistician.

c. A confidence limit for an occurrence rate may be used for computing probabilities for the time of occurrences yielding essentially a tolerance limit. Prediction intervals of comparable assurance or confidence are likely to provide tighter bounds in many cases. Except for citing specific examples, this is difficult to formalize because prediction intervals and tolerance limits are not directly comparable.

To sum up, prediction intervals deal with the risk problem in concrete fashion, are more in line with the reasoning of the layman, and provide statistically tighter bounds in many applications.

PREDICTION INTERVALS

In simple formulation, a prediction interval is a probability statement about two random variables: one which has already been observed and one which will materialize in the future. From this formulation one can provide bounds with probabilistic or confidence qualification for the future random variable to be observed.

One of the simplest prediction intervals is that for the next observation X of a normal random variable with mean μ and variance σ^2 after n samples of the same random variable have been observed. If $\bar{x}$ is the mean of the past observations, then $X-\bar{x}$ is normally distributed with mean 0 and variance $(1+1/n)\sigma^2$.

Accordingly,

$$P\left[X-\bar{x} \leq z_{\gamma}\sqrt{(1+\tfrac{1}{n})\,\sigma^2}\right] = \gamma, \tag{1}$$

where z_{γ} is the γ'th quantile of the standard normal distribution. The inequality in the bracket of (1) translates to the inequality

$$X \leq \bar{x} + z_{\gamma}\sqrt{(1+\tfrac{1}{n})\,\sigma^2}, \tag{2}$$

to which we attach probability γ. After observation of $\bar{x}$ it may be more rigorous to think of γ as a confidence coefficient for one sided interval for X defined by (2)-analogous to confidence intervals of parameters.

PREDICTION INTERVALS FOR THE OCCURRENCE OF THE NEXT EVENT

Model

Consider Poisson process X(t) with parameter λ and arrival time T_j for the j'th occurrence. Let $t_o > 0$ be the time during which X(t) has been observed; therefore, $X(t_o)$ is the number of occurrences in the interval $[0, t_o]$. Let $t_1 \geq t_o$ be an arbitrary point in time not before t_o. Let Y be the waiting time to the first occurrence after t_1, and measured from t_1; i.e.,

$$Y = T_{X(t_1)+1} - t_1. \tag{3}$$

Because X(t) is Poisson process and $t_o \leq t_1$, Y is independent of $X(t_o)$ and T_j has a gamma distribution with scale parameter λ and shape parameter j.

Derivations

It is our objective to find prediction intervals for Y as a function of $X(t_o)$. Two prediction intervals are derived for Y one of which easily generalizes easily to the waiting time Y_{β} to the β'th occurrence after t_1. The first prediction interval is easily derived; it is reasonably effective if the expected number of occurrences in the interval $[0, t_o]$ is small, a condition appropriate to LaPlace's sun rise problem or the major nuclear accident problem. The second prediction interval is of a more general nature.

Prediction Interval A and its Relationship to LaPlace's Law of Succession

Let $T^* = \min[t_o, T_1]$. The following theorem holds.

Theorem 1: For $k > 0$,

$$P(Y > kT^*) > 1/(k+1). \tag{4}$$

Proof: Because Y is independent of T^*,

$$P[Y>kT^*] = \int_0^{t_o} \exp(-\lambda kz)\exp(-\lambda z)\lambda dz + \exp(-\lambda kt_o)\exp(-\lambda t_o)$$

$$= [1/(k+1)\{1 - \exp[-\lambda t_o(k+1)]\} + \exp[-\lambda t_o(k+1)]$$

$$= 1/(k+1) + [k/(k+)] \exp[-\lambda t_o(k+1)]\} > 1/(k+1). \quad (5)$$

One may treat $1/(k+1)$ as a confidence coefficient for a conservative one sided prediction interval for Y.

Setting $k = 1/n$, equating t_o with n, and observing that no failure in n trials (or units of time) is equivalent to $T^* = t_o$ one obtains an expression resemling LaPlace's law of succession.

$$P[Y>1] > 1/[(1/n)+1]=n/(n+1). \quad (6)$$

The formulation of (6) is not correct, the correct statements are:

$$P[Y>T^*/n] > n/(n+1). \quad (7)$$

$$P(Y>1) = \exp(-\lambda), \quad (8)$$

if Y is exponentially distributed.

$$P(Y>1) = 1-p \quad (9)$$

in the corresponding discrete formulation, where p is probability of the event occurring in a single trial.

For the general prediction interval with confidence coefficient γ one sets $k = \gamma^{-1} - 1$ and obtains

$$P[Y>(\gamma^{-1} - 1)T^*]>\gamma. \quad (10)$$

Prediction Interval B

A conservative prediction which takes account of all occurrences during observation time t_o, is implied by Theorem 2 given below. It mimics the conventional prediction interval in which the predictor variable is also a waiting time.

Let Z_1 and Z_2 have gamma distributions with common scale parameter and respective shape parameters $\alpha_1 = 1$ and $\alpha_2 = \alpha$. The ratio $Z_1/(Z_2/\alpha)$ has an F distribution with 2 and 2α degrees of freedom, and

$$P(Z_1>kZ_2/\alpha_2) = P[Z_1/(Z_2/\alpha_2)>k] = [1+(k/\alpha)]^{-\alpha}. \quad (11)$$

Equating $[1+(k/\alpha)]^{-\alpha}$ with γ one obtains that $k=\alpha(\gamma^{-1/\alpha}-1)$.

This leads immediately to

$$P[Z_1>(\gamma^{-1/\alpha}-1)Z_2] = \gamma. \tag{12}$$

We have

Theorem 2:

$$P\{Y>[\gamma^{-1/[X(t_o)+1]}-1]t_o\} > \gamma. \tag{13}$$

Proof: Let Z be an exponential random variable with parameter λ independent of $X(t_o)$ and $T_{X(t_o)+1}$. Let $X = X(t_o)$ and $k = \gamma^{-1/(X+1)}-1$. Then

$$P(Y>kt_o) = P(Z>kt_o) \geq P(Z>kT_{X+1}) = E_X[P(Z>kT_{X+1})]=\gamma. \tag{14}$$

Table 1 demonstrates the conservatism of equation (13). It lists the differences

$$D = P\{Y>[\gamma^{-1/[X(t_o)+1]}-1]t_o\} - \gamma \tag{15}$$

for selected values of λ and γ; without loss of generality we choose $t_o = 1$; λ is then the occurrence rate per t_o.

TABLE 1. CONSERVATISMS IN FORMULA (13) GIVEN IN VALUES OF D

	λ				
γ	1/2	1	2	4	8
.01	.01077	.00651	.00344	.00167	.00071
.10	.06205	.03938	.01861	.00797	.00286
.25	.13757	.06038	.02937	.01105	.00344
.50	.19411	.08749	.02991	.00936	.00233
.75	.13008	.06992	.02304	.00460	.00079
.90	.05760	.03348	.01173	.00180	.00016
.99	.00604	.00365	.00134	.00018	.00000

Prediction Interval C

Theorem 2 generalizes easily to the waiting time of any future occurrence. Let Y_β be the waiting time to the β'th occurrence after t_1, then

$$P\{Y_\beta > \beta t_o F(2\alpha, 2[X(t_o)+1]; \gamma)/[X(t_o)+1]\} > \gamma, \tag{16}$$

where $F(\alpha_1, \alpha_2; q)$ is the (1-q) quantile of the F-distributions with α_1 and α_2 degrees of freedom.

The generalization of Theorem 2 is contemplated for specifying the interval between inspections of snubbers (shock absorbing devices). It has been the tradition to set the interval between inspections of snubbers on the basis of the number of snubbers failed during the last interval between inspection. An alternative to past methods is to take the interval between inspection as the lower limit of the prediction interval to the m'th failure; m-1 is a number of snubber failures that can be readily tolerated. This provides probabilistic control over the number of failed snubbers until the next inspection if the failure rate remains unchanged.

References

Feiler, (1950), An Introduction to Probability and its Applications; Vol. 1, New York, John Wiley.

Uspensky, J. V. (1937), Introduction to Mathematical Probability, New York, McGraw-Hill Book Company.

COMPARISON OF IN-SITU AND LABORATORY MEASUREMENT METHODS FOR RA-226

P. R. Engelder, H. L. Fleischhauer, and S. J. Marutzky
Bendix Field Engineering Corporation
Grand Junction, Colorado
and
R. O. Gilbert
Battelle/Pacific Northwest Laboratories
Richland, Washington

ABSTRACT

A number of methods for measuring Ra-226 in surface soil are used to determine compliance with EPA standards in support of DOE remedial action programs. A study is being conducted to estimate the relationship between laboratory and in-situ measurements, and to evaluate the consistency of results obtained using five different field instruments.

An experiment was designed to estimate the mean Ra-226 concentration within a 100-m^2 area. Using a comparison-of-means test and analysis of variance, the laboratory and in-situ methods were evaluated to determine whether results are significantly different. Another experiment was designed to evaluate the linear relationship between field and laboratory measurements of Ra-226. These linear relationships were estimated via regression analysis of data collected over a wide range of concentrations. Their stability was examined by testing the equality of results of two independent regression experiments.

Preliminary results indicate that the field methods yield consistently higher estimates of Ra-226 than does laboratory analysis, suggesting a potential problem with calibration and/or conversion of the raw data to concentrations. Furthermore, for each field instrument, the regression equations that describe the relationship between in-situ and laboratory measurements differ significantly from one sample suite to the other.

INTRODUCTION

A large number of organizations are engaged in remedial action activities associated with the four programs conducted under the auspices of the U.S. Department of Energy (DOE) Division of Remedial Action Projects. These programs are the Surplus Facilities Management Program (SFMP), Formerly Utilized Sites Remedial Action Program (FUSRAP), Uranium Mill

Tailings Remedial Action Project (UMTRAP), and Grand Junction Remedial Action Program (GJRAP). Remedial action comprises three main phases: site characterization, site cleanup, and verification of site cleanup. Each phase is conducted with a view to meeting U.S. Environmental Protection Agency (EPA) standards for cleanup of sites contaminated by uranium mill tailings (U.S. EPA 1983). In the surface soil layer (0 to 15 cm), Ra-226 concentration averaged over a 100-m^2 area is not to exceed 5 pCi/g above background. The EPA standards, however, do not specify how this compliance determination is to be accomplished, resulting in the development of various techniques, using several different kinds of instruments. The question arises: Do these various measurement methods yield comparable results? The DOE Technical Measurements Center (TMC), which was established to recommend methods and instrumentation for use in remedial action programs, was asked to address this question.

OBJECTIVES

Two objectives in the form of questions were defined to accomplish this task, the first being: Do laboratory soil-sample analysis and the five in-situ methods yield the same mean response when measuring Ra-226 in a small field plot characterized by low Ra-226 spatial variability? Objective 1 can be restated as the following null and alternative hypotheses:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_{soil}$
H_1: One or More of the Means Are Not Equal

where μ_i is the true mean Ra-226 response for the ith type of in-situ instrument (i=1,...,5) and μ_{soil} is the true mean Ra-226 concentration derived from laboratory analysis of soil samples collected at the in-situ measurement points.

Objective 2 involved answering the following question: Is the regression relationship between in-situ and laboratory measurements the same when the experiment is repeated? The null and alternative hypotheses are

H_0: $\alpha_1 = \alpha_2$ and $\beta_1 = \beta_2$
H_1: $\alpha_1 \neq \alpha_2$ and/or $\beta_1 \neq \beta_2$

where α_i and β_i are the intercept and slope parameters, respectively, of the simple linear regression between in-situ and laboratory measurements for data set i.

EXPERIMENTAL DESIGN

Five different scintillation-type instruments/configurations were used in this study: geoMetrics GR-410 Spectrometer (one shielded and

one unshielded), Scintrex GAD-6 Spectrometer, Bendix Field Engineering Corporation EL-0018 Delta Counter, and Mount Sopris PS-872 Delta Counter. These instruments are or have been used by remedial action contractors; their inclusion in this study does not signify a preference over other available instrumentation. A description of scintillation-type instruments and associated calibration methods can be found in Marutzky et al. (1984).

Design for Objective 1

The first objective of this study was addressed using an experimental design based on conditions imposed by the EPA standards for residual Ra-226 in soil. These standards are stated in terms of average concentration over a 100-m^2 area. A randomized block design with replications was selected as the most appropriate technique for assessing equality of mean concentrations. This type of design is very sensitive to small differences in the mean (Griffiths 1967), and permits isolation of both spatial variation and experimental error. The latter may include counting error, operator error, and error due to such temporal effects as ambient temperature and radon in the air.

The study was conducted at a remedial action site for which Ra-226 distribution had been previously characterized. Two measurement localities were identified, one where the Ra-226 concentration was expected to be near 5 pCi/g and another where it was expected to be between 50 and 100 pCi/g. Both were scanned with instruments to determine whether radioactivity was reasonably homogeneous. At each locality, a 10-m-by-10-m area was surveyed. Within the area, 36 cells, hereinafter referred to as blocks, were laid out in a square grid pattern, each block measuring 1.67 m by 1.67 m. The center of each block was flagged so that each instrument could be placed at the same point. A measurement was made using each of the five instruments at each of the 36 flagged points. The measurement sequence was randomized for all instruments. This randomness was examined for each instrument across all points, as well as for the order of instruments at each point, using the runs test as described by Dixon and Massey (1969). Replicate measurements were made by repeating the same random sequence for each instrument. After all measurements were completed, a 1500-cc soil sample for laboratory analysis was collected at each flagged point, using a stainless-steel sampler inserted to a depth of 15 cm. Soil-sampling procedures are described in Fleischhauer (1984).

Design for Objective 2

For purposes of the regression study, field measurement points were selected such that they would be uniformly distributed over the range of Ra-226 concentrations and soil types at the site. A total of 403 in-situ Ra-226 measurements had been made on a grid basis during the characterization of the 80-acre site (Abramiuk et al. 1983). For this regression study, these grid data were ranked from lowest to highest and subdivided

into 13 equal groups, each of which represented a small slice of the distribution. Within each group, one primary and one alternate point were randomly selected, without replacement, for each of the two sample suites. One exception, the highest group (range of 78 to 304 pCi/g), involved selection of three primary and three alternate points to avoid large gaps in the sample-suite data. Thus, a total of 15 primary locations were selected for each sample suite, with estimated concentrations ranging from 2 to 150 pCi/g. Alternate points were to be used only when conditions prohibited either vehicular access to the primary point or collection of a representative soil sample at the primary point.

Identical sets of instruments operated by different teams were used so that the two data sets could be collected simultaneously. Measurement locations were flagged so that each instrument could be placed at the same point. Measurements were made at each point by placing the instruments on the point in a fixed sequence, which was the same for each of the 15 points. Replicate measurements were made by repeating the measurement sequence. In other words, each instrument was placed at a point, then removed and replaced only after the entire sequence of instruments had been placed on the same point. This process required approximately 2 hours per point. After all measurements were completed at a point, a 1500-cc soil sample was collected to a depth of 15 cm, using the procedure described by Fleischhauer (1984).

RESULTS

For the two grid studies, the average of the two replicate measurements at each point was used in the statistical analyses. An examination of the descriptive statistics for each of the six treatments (laboratory assay and five instrument measurements) on the low-concentration grid reveals that many of the distributions are slightly skewed toward high values, and several of these exhibit statistically significant kurtosis (Table 1). Statistics for the instrument data on the high-concentration grid indicate that only two distributions are slightly skewed, but most exhibit statistically significant kurtosis. Log transformation of these data does not enhance the normality in many cases (Table 1).

A comparison of the six box and whisker plots (Tukey 1977) for the low-concentration grid data (Figure 1) indicates that the medians for all six treatments are very similar. The range of the laboratory data is larger than the range for each of the five in-situ instruments. The low-concentration data appear to comprise at least two groups of treatments, (1) the spectrometers and the laboratory analysis, and (2) the delta counters.

Comparing the box and whisker plots for the high-concentration grid data (Figure 2) shows that the medians and ranges for all five instruments are very similar. The means lie between approximately 120 and 137 pCi/g. Conversely, the median and mean of the laboratory-analysis results on the

TABLE 1. DESCRIPTIVE STATISTICS OF ANOVA DATA

Measurement Type	Mean	Variance	Skewness	Kurtosis
	Low-Concentration Grid (n=36)			
Mount Sopris Delta	5.78	1.01	0.34	-0.19
Log Transformation	0.75	0.01	-0.11	-0.34
Bendix Delta Counter	5.46	1.77	0.68[a]	1.64[a]
Log Transformation	0.72	0.01	-0.39	0.84[b]
GAD-6 Spectrometer	4.65	0.88	-0.11	0.05
Log Transformation	0.66	0.01	-0.74[b]	0.39
GR-410 Spectrometer	3.66	0.86	0.26	0.52
Log Transformation	0.55	0.02	-0.68[a]	0.76[a]
Shielded GR-410	4.13	1.23	1.03[b]	2.10[b]
Log Transformation	0.60	0.01	0.00	0.47
Laboratory Assay	3.65	2.91	2.53[b]	8.39[b]
Log Transformation	0.53	0.03	0.75[b]	1.58[b]
	High-Concentration Grid (n=36)			
Mount Sopris Delta	120.15	1161.07	0.45	-0.88[b]
Log Transformation	2.06	0.02	0.04	-1.07[b]
Bendix Delta Counter	137.29	1604.24	0.59[a]	-0.62[a]
Log Transformation	2.12	0.02	0.15	-0.99[b]
GAD-6 Spectrometer	127.50	1006.41	0.26	-0.65[a]
Log Transformation	2.09	0.01	-0.19	-0.89[b]
GR-410 Spectrometer	129.93	1029.25	0.23	-0.72[a]
Log Transformation	2.10	0.01	-0.20	-0.93[b]
Shielded GR-410	124.70	1167.01	0.31	-1.17[b]
Log Transformation	2.08	0.01	-0.03	-1.17[b]
Laboratory Assay	95.20	1151.44	0.74[b]	0.35
Log Transformation	1.95	0.03	-0.41	0.62[a]

a. Value is significant at the 0.05 level.

b. Value is significant at the 0.01 level.

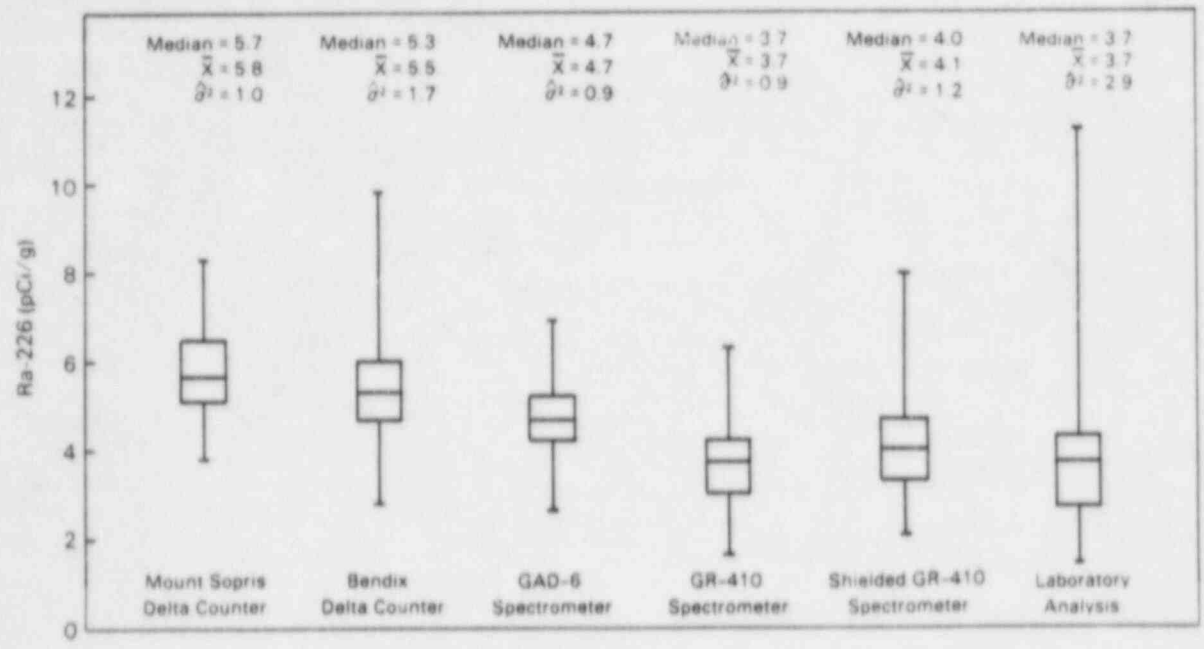

Figure 1. Box and whisker plots of low-concentration grid data (n=36).

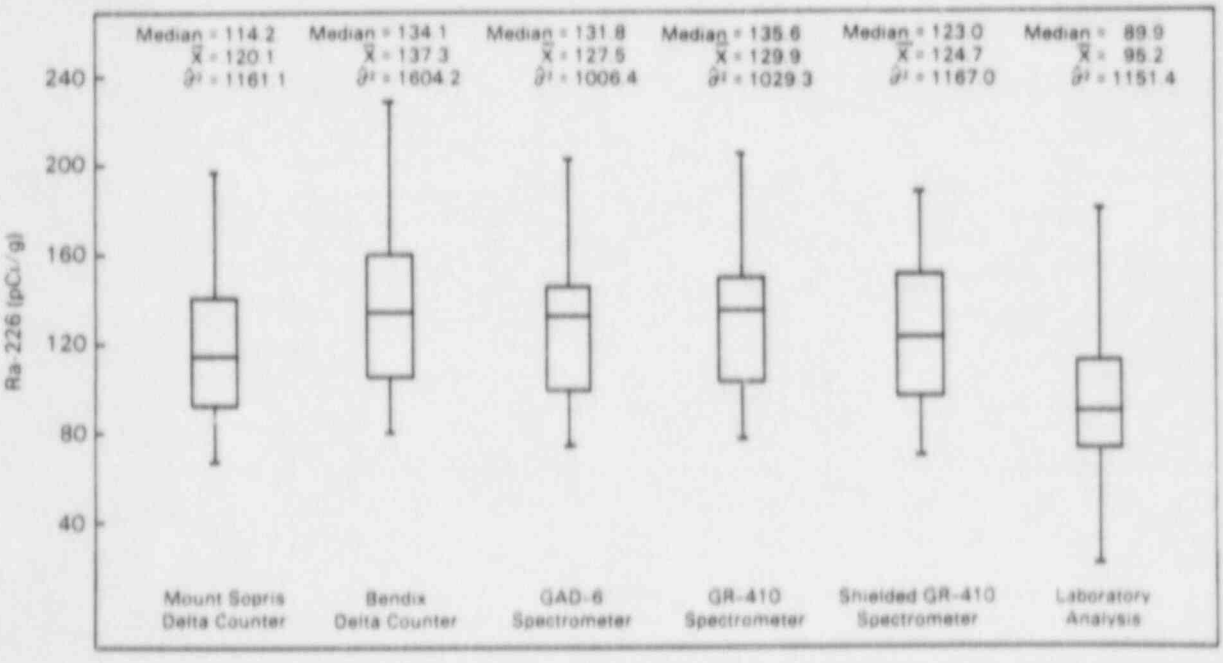

Figure 2. Box and whisker plots of high-concentration grid data (n=36).

soil samples appear to be much lower than those of the in-situ measurements. In addition, the range of the laboratory data is larger. This dissimilarity suggests that the high-concentration data also comprise two distinct groups of treatments, the in-situ measurements and the laboratory assays.

Figure 3 presents a diagram of a generalized random block design. The appropriate analysis-of-variance model described by Winer (1971) is

$$X_{ij} = \mu + \beta_i + \tau_j + \varepsilon_{ij}$$

where

X_{ij} = each observation
$i = 1, \ldots, n$ (blocks)
$j = 1, \ldots, k$ (treatments)
μ = the overall mean
β_i = the block effect
τ_j = the treatment effect
ε_{ij} = all other sources of unassigned error.

This model is a mixed-effects model as defined by Scheffé (1959). The blocks are a random effect, and the treatments are a fixed effect. This model was applied to the data using BMDP2V (Dixon and Brown 1979).

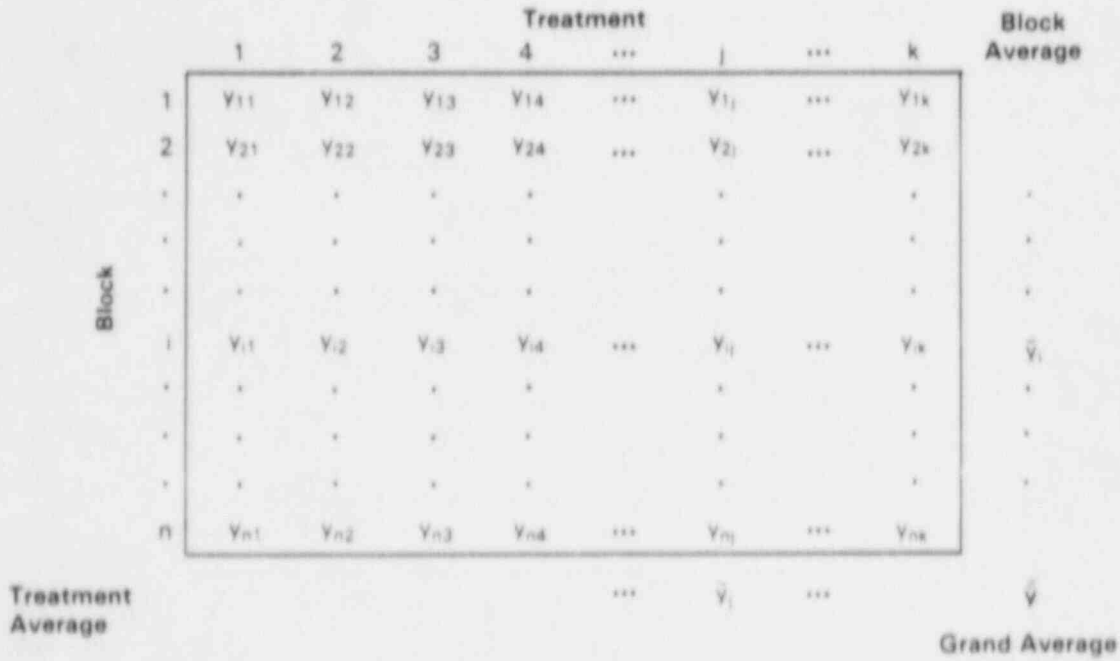

Figure 3. Generalized random block design.

Table 2 presents the analysis-of-variance results for the low-concentration grid data. The results are similar for the high-concentration data. In both cases, the F statistic for the treatment comparison is highly significant, indicating that one or more of the treatment means are significantly different from the other treatment means at the 5 percent level. Diagnostic checking of the model through an examination of the residuals, as recommended by Box et al. (1978), indicates that the model is appropriate for the data. Furthermore, the low- and high-concentration data were also examined using Friedman's Nonparametric Test (Conover 1971, p. 266), results of which confirm that at least one of the treatment means is different from the other means for both experiments.

A multiple comparison of all possible pairs of treatment means, using both sets of concentration data, is shown in Table 3. The Bonferroni t-tables developed by Bailey (1977) were used for this analysis. By examining the results of the various pairs of treatment means, several groupings are revealed. In the low-concentration grid, two groups of treatments are characterized by means that are not statistically different. These are (1) the two delta counters, and (2) the laboratory analysis of soil samples and the GR-410 spectrometers, both shielded and unshielded. The GAD-6 spectrometer appears to be a transitional treatment between the groups, because it is not different from one member of each of the other two groups. Using the Bonferroni t-tables for analysis of the high-concentration grid data, two groups of treatments yield statistically different results. These are (1) all of the field instruments, and (2) the laboratory analysis of soil samples.

TABLE 2. ANALYSIS OF VARIANCE USING THE RANDOMIZED BLOCK DESIGN (low-concentration grid data)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squared Error	F Value	Probability
Mean	4485.4	1	4485.4	8759.8	0.00
Treatment	147.3	5	29.5	57.5	0.00
Block	214.3	35	6.1	11.9	0.00
Error	89.6	175	0.5		

TABLE 3. PAIRWISE COMPARISON OF TREATMENT MEANS USING STUDENT'S T TEST[a]

Measurement Type	Mount Sopris Delta Counter	Bendix Delta Counter	GAD-6 Spec.	GR-410 Spec.	Shielded GR-410
		Low-Concentration Grid (n=36)			
Bendix Delta Counter	1.15				
GAD-6 Spectrometer	4.88[b]	2.94			
GR-410 Spectrometer	9.24[b]	6.61[b]	4.52[b]		
Shielded GR-410	6.59[b]	4.59[b]	2.19	-1.93	
Laboratory Analysis	6.41[b]	4.98[b]	3.08[c]	0.01	1.38
		High-Concentration Grid (n=36)			
Bendix Delta Counter	-1.96				
GAD-6 Spectrometer	-0.95	1.15			
GR-410 Spectrometer	-1.25	0.86	-0.32		
Shielded GR-410	-0.56	1.44	0.36	0.67	
Laboratory Analysis	3.11[c]	4.81[b]	4.17[b]	4.46[b]	3.68[b]

a. Numbers in table are t values resulting from a test of the difference between two means. The critical value at the $\alpha = 0.05$ level is 3.08 with $k = 15$ and $\nu = 70$ using Bonferroni t tables given by Bailey (1977).

b. Value is significant at the 0.01 level.

c. Value is significant at the 0.05 level.

Simple linear regression of in-situ data on laboratory data was performed using BMDP1R (Dixon and Brown 1979). For each instrument, regression lines were calculated for the two sample suites combined, as well as for each sample suite separately. Results of the regressions were used to construct an analysis-of-variance test for Objective 2, based on the method described by Neter and Wasserman (1974).

The initial run using all data indicated significant differences between the two sample suites for each instrument. Sample Suite 1 yielded equations with higher slopes and intercepts than Sample Suite 2. A 95 percent confidence interval on the difference between the two slopes does not contain zero. Therefore, for each instrument, the slopes are significantly different.

Scatter plots reveal several outliers in Sample Suite 1. Figures 4 and 5 present a plot of the observed in-situ measurements versus the laboratory assay and a plot of the residual value versus predicted value, respectively, for all data combined using the GR-410 spectrometer. The locations in question have high in-situ values associated with low laboratory assays. Three are located on top of stabilized tailings piles, suggesting a problem with source geometry and disequilibrium between radium and its daughters.

In-situ detectors measure gamma radiation emitted by daughter products of Ra-226. Accurate assays using field instruments depend on conditions of secular equilibrium between the daughters and the parent. [Wasure (1977) presents a complete discussion of secular equilibrium versus disequilibrium.] Ra-226 decays to Rn-222, a gas that can migrate through porous media such as soil and thereby induce disequilibrium.

For the locations mentioned above, Rn-222 probably migrated from the tailings into the thin soil cover where it decayed to yield the higher apparent concentrations detected by the field instruments. In the soil sample, these radon daughters started to decay into nonradioactive daughters, resulting in the lower concentration determined by laboratory analysis. When the field disequilibrium data are combined with the data presented in Figure 5, one of these locations is confirmed as an outlier, and is excluded from subsequent regression analysis. Other potential outliers have been retained in the analysis pending further examination.

Table 4 presents results of the regression analysis using the reduced data set. Scatter plots of regression residuals versus predicted values and regression residuals versus the independent variable, as recommended by Draper and Smith (1981), do not suggest any need to transform the data. Under ideal circumstances, slopes should be equal to one and intercepts should be equal to zero. In this analysis, all slopes differ significantly from one at the 0.05 level. With the exception of the intercept for the Mount Sopris delta counter on soil in Sample Suite 1, none of the intercepts differs from zero. The F-statistic indicates that equations for Sample Suite 1 and Sample Suite 2 differ significantly for each instrument. The 95 percent confidence interval for $(\beta_1 - \beta_2)$ does not

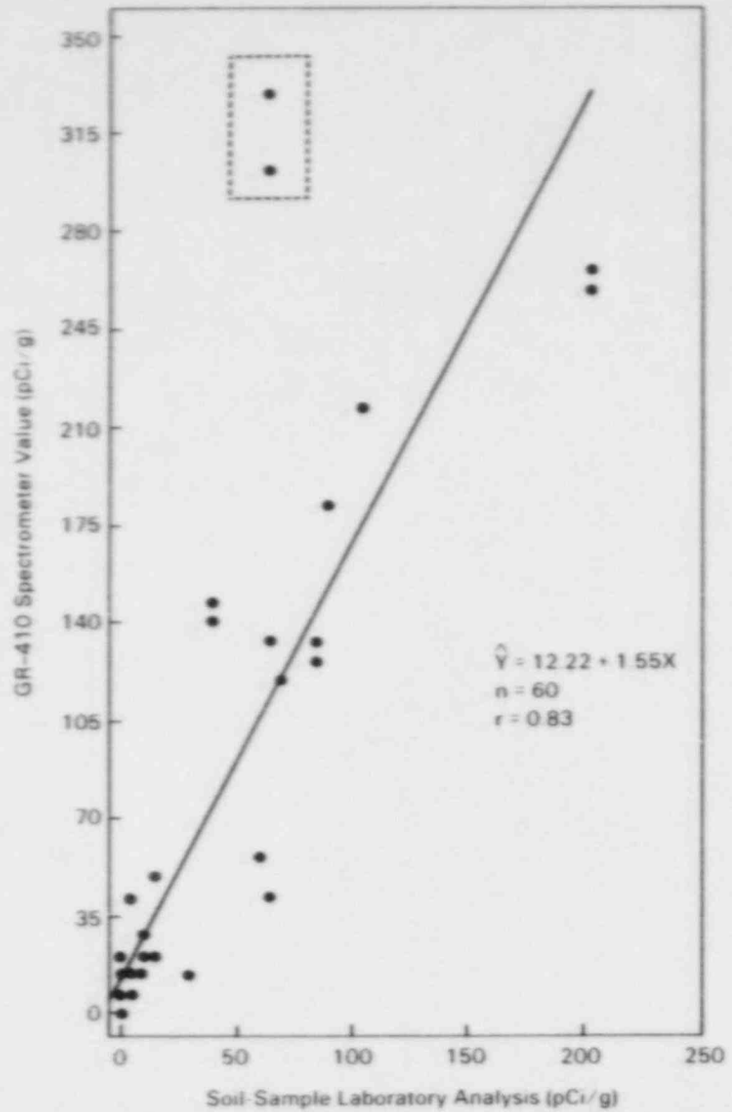

Figure 4. Regression plot of all data combined for GR-410 spectrometer.

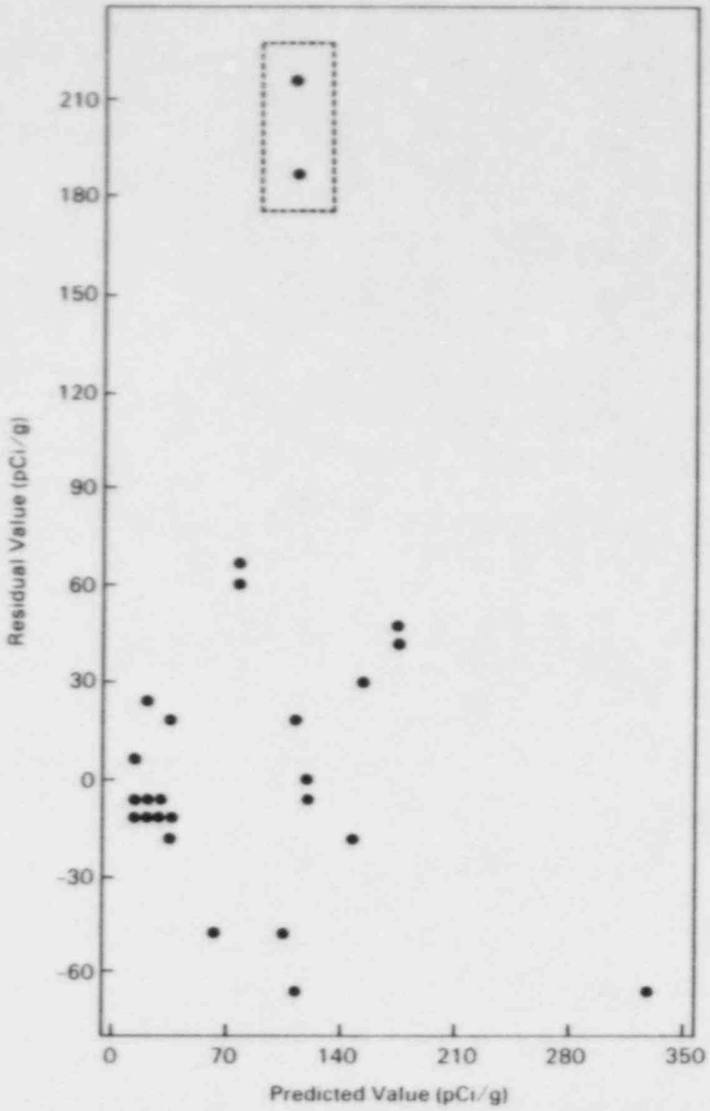

Figure 5. Plot of predicted vs. residual values for all data combined for GR-410 spectrometer.

TABLE 4. COMPARISON OF TREATMENT REGRESSION LINES WITH OUTLIERS REMOVED

Instrument	Data Set	N	Slope	Intercept	R[a]	Standard Error[b]	F Test of Regression	Equality of Lines[c]
Mount Sopris Delta	All Data	58	1.36	14.98	0.87	35.92	166.44	
	Suite 1	28	1.89	19.77	0.86	38.22	71.06	
	Suite 2	30	1.28	2.48	0.96	18.94	373.91	13.63
Bendix Delta Counter	All Data	58	1.39	10.76	0.87	35.24	180.13	
	Suite 1	28	2.15	9.79	0.92	30.99	139.98	
	Suite 2	30	1.22	2.11	0.96	19.39	325.81	25.88
GAD-6 Spec.	All Data	58	1.43	9.94	0.91	28.91	281.68	
	Suite 1	28	2.07	7.16	0.94	24.09	214.54	
	Suite 2	30	1.27	4.94	0.96	18.74	377.70	23.69
GR-410 Spec.	All Data	58	1.44	9.07	0.92	27.79	308.82	
	Suite 1	28	2.01	6.89	0.94	23.77	209.11	
	Suite 2	30	1.30	4.18	0.96	19.30	371.03	19.48
Shielded GR-410 Spec.	All Data	58	1.42	10.03	0.89	33.98	202.68	
	Suite 1	28	2.25	7.26	0.94	27.87	190.29	
	Suite 2	30	1.23	2.59	0.97	15.27	530.20	38.33

a. Correlation coefficient.

b. Standard error of regression line, i.e., square root of mean squared error.

c. These are F values. All of the F values in this table are significant at the 0.01 level.

include zero for any instrument. Thus, the slopes of each pair of equations appear to differ beyond that which is expected by chance. Figures 6 and 7 are scatter plots of GR-410 spectrometer data versus laboratory assay for Sample Suite 1 and Sample Suite 2, respectively.

DISCUSSION

Results of preliminary analysis of the study data indicate that various measurement methods may yield different estimates of mean Ra-226 concentration in small areas. At low concentrations, differences between methods seem most pronounced. At high concentrations, in-situ measurements are consistent, but differ from the laboratory assay. Regression

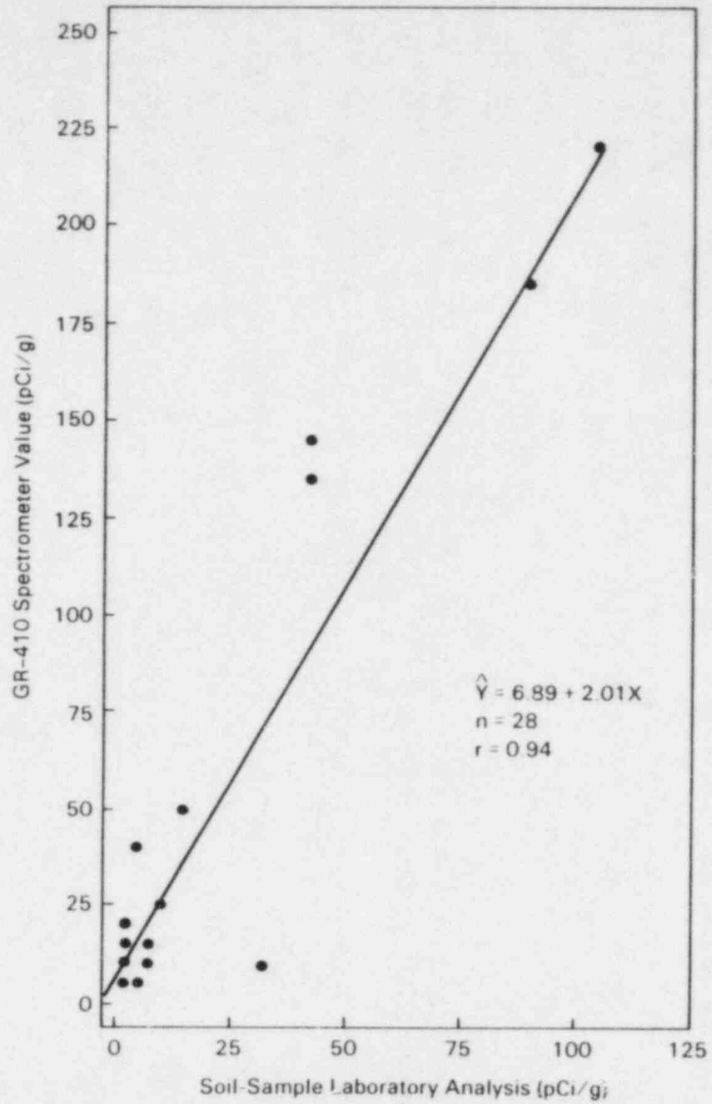

Figure 6. Plot of Sample Suite 1 data for GR-410 spectrometer (outliers removed).

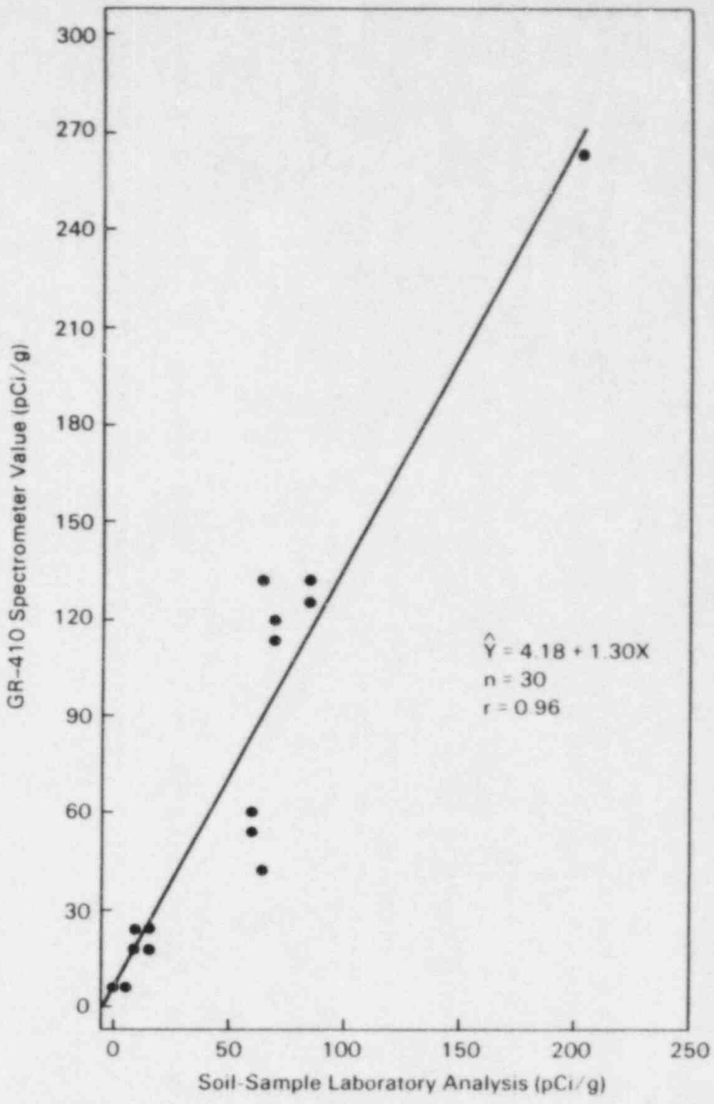

Figure 7. Plot of Sample Suite 2 data for GR-410 spectrometer (outliers removed).

equations have slopes greater than one, confirming the differences between the field methods and laboratory analysis.

Since the field methods systematically yield higher estimates of mean Ra-226 concentration than does laboratory analysis, the calibration and data-reduction procedures used for converting the in-situ data are suspect. The problem may lie in one or more areas, including inappropriate application of moisture and/or disequilibrium corrections to both calibration and raw field data. Investigations are under way to identify and resolve the problems.

The regression study, on the other hand, suggests that correcting these problems may not eliminate all discrepancies. Regression equations for in-situ data on laboratory measurements differ significantly between the two sample suites for each instrument. Effects of operator error, instrument bias, and ambient weather conditions, however, cannot be used to explain the differences between suites. Operator error, due to inconsistent positioning of the instrument or misreading the digital output, is minimized by flagging the measurement locations and by repeating measurements. Instrument bias is not a factor, because the regression equations within a sample suite are similar. Also, one suite yields uniformly higher slopes and intercepts than the other. Finally, since both suites were sampled simultaneously, weather conditions would have been the same.

The differences between the suites appear to be a function of source geometry. In general, there are three main types of source geometry: layers of tailings and ore with soil; homogeneous mixtures of soil with tailings and/or ore particles; and inhomogeneous mixtures of soil with tailings and/or ore particles. Various combinations of these three types are distributed throughout a typical remedial action site. Therefore, any randomly chosen suite of samples will probably yield a unique relationship between in-situ and laboratory measurements.

REFERENCES

ABRAMIUK, I. N., BLANCHFIELD, L. A., COTTER, E. T., FLEISCHHAUER, H. L., GOODKNIGHT, C. S., JOHNSON, V. G., KARP, K. E., KEARL, P. M., KORTE, N. E., REDOLFI, C. A., ROQUEMORE, R. R., SCHAER, D. W., and SEWELL, J. M. (1983), Monticello Remedial Action Project, Site Analysis Report (draft), GJ-10(83), Grand Junction: Bendix Field Engineering Corporation.

BAILEY, B. J. R. (1977), 'Tables of the Bonferroni t Statistic,' Journal of the American Statistical Association, 72, 469-478.

BOX, G. E. P., HUNTER, W. G., and HUNTER, J. S. (1978), Statistics for Experimenters - An Introduction to Design, Data Analysis, and Model Building, New York: John Wiley and Sons, Inc., 653 p.

CONOVER, W. J. (1971), Practical Nonparametric Statistics, New York: John Wiley and Sons, Inc., 462 p.

DIXON, W. J., and BROWN, M. B. (1979), BMDP Biomedical Computer Programs P-Series 1979, Berkeley: University of California Press, 880 p.

DIXON, W. J., and MASSEY, F. J, Jr. (1969), Introduction to Statistical Analysis, New York: McGraw-Hill Book Co., 638 p.

DRAPER, N. R., and SMITH, H. (1981), Applied Regression Analysis, New York: John Wiley and Sons, Inc., 709 p.

FAURE, G. (1977), Principles of Isotope Geology, New York: John Wiley and Sons, Inc., 464 p.

FLEISCHHAUER, H. L. (1984), Procedures For Sampling Radium-Contaminated Soils (draft), U.S. Department of Energy Technical Measurements Center Report GJ/TMC-13, Grand Junction: Bendix Field Engineering Corporation.

GRIFFITHS, J. C. (1967), Scientific Methods in Analysis of Sediments, New York: McGraw-Hill Book Co., 508 p.

MARUTZKY, S. J., STEELE, W. G., KEY, B. N., and KOSANKE, K. (1984), Surface Gamma-Ray Measurement Protocol, U.S. Department of Energy Technical Measurements Center Report GJ/TMC-06, Grand Junction: Bendix Field Engineering Corporation.

NETER, J., and WASSERMAN, W. (1974), Applied Linear Models, Homewood, Illinois: Richard D. Irwin, Inc., 842 p.

SCHEFFÉ, H. (1959), The Analysis of Variance, New York: John Wiley and Sons, Inc., 477 p.

TUKEY, J. W. (1977), Exploratory Data Analysis, Reading, Massachusetts: Addison-Wesley Publishing Co., 688 p.

U.S. ENVIRONMENTAL PROTECTION AGENCY (1983), 'Standards for Remedial Actions at Inactive Uranium Processing Sites,' in Federal Register, v. 48, no. 3, January 5, Washington: U.S. Government Printing Office, 590-606.

WINER, B. J. (1971), Statistical Principles in Experimental Design, Second Edition, New York: McGraw-Hill Book Co., 907 p.

THE VARIANCE OF MEASUREMENTS FROM A CALIBRATION FUNCTION DERIVED FROM DATA WHICH EXHIBIT RUN-TO-RUN DIFFERENCES

A. M. Liebetrau
Pacific Northwest Laboratory

ABSTRACT

The volume of liquid in a nuclear process tank is determined from a calibration equation which expresses volume as a function of liquid level. Successive calibration "runs" are made to obtain data from which to estimate either the calibration function or its inverse. For tanks equipped with high-precision measurement systems to determine liquid level, it frequently happens that run-to-run differences due to uncontrolled or uncontrollable ambient conditions are large relative to within-run measurement errors. In the strict sense, a calibration function cannot be developed from data which exhibit significant run-to-run differences. In practice, run-to-run differences are ignored when they are small relative to the accuracy required for measurements of the tank's contents. The use of standard statistical techniques in this situation can result in variance estimates which severely underestimate the actual uncertainty in volume measurements. This paper gives a method whereby reasonable estimates of the calibration uncertainty in volume determinations can be obtained in the presence of statistically significant run-to-run variability.

INTRODUCTION

Accurate volume measurements are an essential component of any system to control and account for nuclear materials. The volume of liquid in a process tank is determined from a calibration equation which expresses the relationship between volume and liquid level. The calibration equation is estimated from data obtained during one or more calibration "runs." During each run, carefully measured increments of a liquid are added to the tank. These volume measurements, together with observations of the corresponding liquid levels (or observations of some surrogate for liquid level, such as pressure), constitute the basic data from which the calibration equation is derived.

Liquid level and volume measurements can be made with great precision, so it is often the case that run-to-run differences due to uncontrolled or uncontrollable ambient conditions are significantly greater than within-run measurement errors. In a strict sense, data which exhibit significant run-to-run differences cannot be used to develop a calibration function. In practice, these differences are ignored when they are small relative to the accuracy required for measurements of the tank's contents. In this case, the use of standard statistical methods to estimate the variances of volume measurements can lead to serious underestimates of measurement uncertainty. The remainder of this paper is devoted to the development of a method, necessarily ad hoc, which yields operationally reasonable estimates of calibration uncertainty for volume measurements in the situation described above.

In practice, it is often necessary to adjust raw liquid level and volume measurements to a standard set of conditions before a calibration function can be developed. If temperature differences are observed during or between runs for example, it may be necessary to standardize volume measurements. Moreover, the data frequently require "alignment" (see Goldman and Liebetrau 1984) to compensate for differences in the initial volume of liquid in the tank at the start of each calibration run. For purposes of the present discussion, it is assumed that all necessary data normalization has been accomplished. It is also assumed that the data have passed all the screening tests used to verify their validity and internal consistency. A recent paper by Jones (1984) contains a detailed discussion of data normalization techniques for tanks equipped with pressure measurement systems. Methods used to screen data for internal consistency at the time of acquisition are discussed by Liebetrau (1984). Empirical methods for aligning and comparing the data from several calibration runs are presented by Goldman and Liebetrau (1984).

MODEL FITTING AND ESTIMATION

The standardized calibration data consist of a series of volume measurements together with observations of the corresponding liquid levels. These data are obtained during several calibration runs. Let X_{ij} be the total

of the first i calibrated volume increments added to the tank during the jth calibration run, and let Y_{ij} be the corresponding liquid level. For simplicity and clarity of presentation, it is assumed that the number of volume increments is the same for all runs, so that i = 1, ..., N for each j, j = 1, ..., r.

The calibration function, i.e., the relationship between the standardized liquid level (Y) and the standardized cumulative volume (X) measurements, is often estimated by fitting an equation of the form Y = f(X) to the observed data. A typical calibration function consists of several segments, each of which is fit with a first or second degree polynomial. The data available for fitting a model to the kth segment consist of the pairs $(Y_{i+i_k,j}, X_{i+i_k,j})$, i = 1, ..., n_k, j = 1, ..., r, where $n_0 = 0$ and $i_k = \sum_{\ell=0}^{k-1} n_\ell$ is an integer such that $0 \leq i_k < N$. For present purposes, it is not necessary to identify particular segments of the data, so all references to segment number will be suppressed in order to simplify the notation.

It is also assumed that classical least-squares regression techniques are appropriate for estimating the function f, where liquid level (Y) is taken as the response variable and volume (X) is taken as the control variable. Further, normality of liquid level measurement errors is assumed as necessary for testing hypotheses. Although these assumptions should be verified for each set of calibration data, they are plausible for calibration data with which the author is familiar. With suitable modification, the methods presented here can be used to produce realistic estimates of volume measurement variability when other estimation methods are used, as well as when the number of volume increments per run is not constant.

For the segment in question, it is initially necessary to determine the functional form of the model, after which several model-fitting steps are carried out. Attention is restricted to linear and quadratic functions in this paper. The linear function is used for illustrative purposes, in which case the necessary sums of squares and parameter estimates can be obtained from a standard analysis of of covariance. Three model-fitting steps are described below.

Step 1. Once the functional form of the model is determined, fit this function individually to the data of each calibration run and compute the

corresponding total pooled sum of squared residuals, say SSE(1). The corresponding degrees of freedom is df(1) = r(n - p), where p-1 is the degree of the polynomial being fitted.

The quantity MSE(1) = SSE(1)/df(1) is the best available estimate of random measurement error (plus, of course, any lack of fit due to inadequacy of the model).

For a linear function, this step yields

$$SSE(1) = \sum_{j=1}^{r} \sum_{i=1}^{n} (Y_{ij} - \hat{\alpha}_j - \hat{\beta}_j X_{ij})^2 \quad , \qquad (1)$$

where $\hat{\alpha}_j$ and $\hat{\beta}_j$ are the ordinary least squares estimates (with $\overline{X}_{.j} = \frac{1}{n} \sum_{i=1}^{n} X_{ij}$, etc.)

$$\hat{\alpha}_j = \overline{Y}_{.j} - \hat{\beta}_j \overline{X}_{.j}$$

and $$\hat{\beta}_j = (\sum_i X_{ij}Y_{ij} - n\overline{X}_{.j}\overline{Y}_{.j})/(\sum_i X_{ij}^2 - n\overline{X}_{.j}^2) \quad .$$

The degrees of freedom associated with SSE(1) is df(1) = r(n-2).

Step 2. Fit functions individually to the data from each run as in Step 1, but in this case, estimate all coefficients except the intercepts from the pooled data from all runs, so that the fitted models are "parallel." The corresponding total pooled sum of squared residuals, say SSE(2), has degrees of freedom df(2) = (n - 1)r - (p - 1) = nr - (r + p - 1), where p-1 is the degree of the polynomial being fitted.

The difference D_1 = [SSE(2) - SSE(1)] is a measure of the increased lack-of-fit, relative to that for individual functions, when parallel functions (all coefficients, except intercepts, are estimated from pooled data) are fit to the data. The statistic

$$T_1 = \left[\frac{SSE(2) - SSE(1)}{df(2) - df(1)}\right] / \left[\frac{SSE(1)}{df(1)}\right] \qquad (2)$$

can be used to test the hypothesis that the data are adequately fit by parallel functions. Under the assumption of normality, T_1 has an F-distribution with parameters $df(2) - df(1) = (r-1)(p-1)$ and $df(1) = r(n-p)$.

In the linear case, this step involves fitting lines, one for each run, which have a common slope but differing intercepts. The least-squares estimates are $\hat{\alpha}'_j = \overline{Y}_{.j} - \hat{\beta}\overline{X}_{.j}$, $j = 1, 2, \ldots, r$, and

$$\hat{\beta} = \sum_j \sum_i (X_{ij}Y_{ij} - nr\,\overline{X}\,\overline{Y}) / \left(\sum_j \sum_i X_{ij}^2 - nr\overline{X}^2\right) , \qquad (3)$$

where $\overline{X} = \overline{X}_{..} = \sum_i \sum_j X_{ij}/(nr)$, etc. For these estimates, SSE(2) has the form

$$SSE(2) = \sum_{j=1}^{r} \sum_{i=1}^{n} (Y_{ij} - \hat{\alpha}'_j - \hat{\beta}X_{ij})^2 \qquad (4)$$

and degrees of freedom $df(2) = r(n-1) - 1$. For a linear function, (2) is the statistic used to test the hypothesis that the regression slopes for all groups are equal. A convenient computational form for [SSE(2) - SSE(1)] in this case is

$$D_1 = \sum_j (\hat{\beta}_j - \hat{\beta})^2 \sum_i (X_{ij} - \overline{X}_{.j})^2.$$

Step 3. Fit a single function to the pooled data from all runs and compute the residual sum of squares, SSE(3). The degrees of freedom associated with SSE(3) is $df(3) = nr - p$, where $p-1$ is the degree of the polynomial fitted.

By the reasoning of Step 2, the statistic

$$T_2 = \left[\frac{SSE(3) - SSE(1)}{df(3) - df(1)}\right] / \left[\frac{SSE(1)}{df(1)}\right] \qquad (5)$$

can be used to test the hypothesis that a single function fits the data as well as individual functions. Under the assumption of normality, T_2 has an F-distribution with parameters df(3) - df(1) = p(r-1) and df(1) = r(n-p).

The difference D_3 = SSE(3) - SSE(2) is a measure of the increased lack-of-fit, relative to that for parallel functions, when a single function is fit to the pooled data. The hypothesis that the data are adequately fit by a function with a single intercept, relative to the fit of parallel models, can be tested with the statistic

$$T_3 = \left[\frac{SSE(3) - SSE(2)}{df(3) - df(2)}\right] / \left[\frac{SSE(2)}{df(2)}\right] \quad , \tag{6}$$

which, under the assumption of normality, has an F-distribution with parameters df(3) - df(2) = r-1 and df(2) = rn - (r+p-1).

For a linear function, the least-squares estimate of the intercept is $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}$, the least squares estimate of the slope is given by (3), the residual sum of squares is

$$SSE(3) = \sum_j \sum_i (Y_{ij} - \hat{\alpha} - \hat{\beta}X_{ij})^2 \quad , \tag{7}$$

and df(3) = nr - 2. In this case, a computational form for [SSE(3) - SSE(2)] is

$$D_3 = n\left[\hat{\beta}^2 \sum_j (\overline{X}_{.j} - \overline{X})^2 + \sum_j (\overline{Y}_{.j} - \overline{Y})^2\right] \quad .$$

If the hypothesis that a single function fits the data is rejected, i.e., if T_2 is too large, then the data exhibit significant run-to-run differences. These differences may be due to lack of parallelism or lack of a common intercept, or both. Various scenarios unfold depending upon the nature of run-to-run differences. Estimates of measurement variance are proposed for each in the next section.

VARIANCE ESTIMATES FOR INDIVIDUAL MEASUREMENTS

To compute an estimate of the variance of a predicted value of Y, given X_0, the following basic formula is used:

$$S^2(Y_0|X_0) = S_M^2(Y_0|X_0) + S_R^2(Y_0|X_0) \qquad (8)$$

In (8), $Y_0 = \hat{f}(X_0)$ is the predicted mean value of Y, given $X = X_0$, and $S_M^2(Y_0|X_0)$ is the estimated variance of Y_0 due to uncertainty in the predictive model. The term $S_R^2(Y_0|X_0)$ is an estimate of run-to-run variability. The predictive model $Y = \hat{f}(X)$ is obtained by fitting a single function to the pooled data from all runs. The strategy is to find suitable values for the terms on the right-hand side of (8). Depending upon the relationships among SSE(1), SSE(2) and SSE(3), there are four cases to consider.

Case 1. A single function fits the pooled data. In this case, SSE(3) is not significantly greater than SSE(1); equivalently, run-to-run variability is not significantly greater than random measurement variability. Consequently,

$$S^2 = SSE(3)/df(3) = SSE(3)/(rn - p)$$

is an appropriate estimate of random measurement variability. Because run-to-run variability is not significant, $S_R^2(Y_0|X_0)$ is taken to be zero in this case.

For a linear function, (8) yields the familiar formula

$$S^2(Y_0|X_0) = S^2\left(\frac{1}{nr} + \frac{(X_0 - \bar{X})^2}{\sum_i \sum_j (X_{ij} - \bar{X})^2}\right)$$

$$= \frac{SSE(3)}{df(3)}\left(\frac{1}{nr} + \frac{(X_0 - \bar{X})^2}{\sum_i \sum_j (X_{ij} - \bar{X})^2}\right) \qquad (9)$$

for the variance of a new observation of Y, given $X = X_0$. The estimator (9) has df(3) = rn - 2 degrees of freedom.

With an appropriate change in the right-hand side of (9), an estimate of $S^2(Y_0|X_0)$ can also be obtained for a quadratic function.

Case 2. A single function does not fit the data, but the lack of fit is due to differences in the intercepts and not to a lack of parallelism among the individual models. In this case, SSE(2) is not significantly greater than SSE(1), but SSE(3) is significantly greater than SSE(1). An appropriate estimate of random measurement variability is

$$S^2 = SSE(2)/df(2) = SSE(2)/[rn - (r + p - 1)] \quad . \qquad (10)$$

Run-to-run variability, on the other hand, is estimated by

$$S_R^2(Y_0|X_0) = \frac{SSE(3) - SSE(2)}{n(r - 1)} = \frac{D_3}{n(r - 1)} \quad . \qquad (11)$$

The estimator (11) has df(3) - df(2) = r - 1 degrees of freedom.

For a linear function, substitution of (10) and (11) into (8) yields

$$S^2(Y_0|X_0) = \frac{SSE(2)}{df(2)}\left(\frac{1}{nr} + \frac{(X_0 - \overline{X})^2}{\sum_i \sum_j (X_{ij} - \overline{X})^2}\right) + \frac{D_3}{n(r - 1)} \quad . \qquad (12)$$

The first term on the right-hand side of (12) has df(2) = nr - (r+1) degrees of freedom, and the second has df(3) - df(2) = r - 1 degrees of freedom. Since SSE(3) is significantly greater than SSE(2), the (approximate) degrees of freedom for (12) can be obtained by means of Satterthwaite's approximation formula (see Brownlee 1965):

$$df = \frac{(A + B)^2}{A^2/df(A) + B^2/df(B)} \quad , \qquad (13)$$

where A denotes the first term on the right-hand side of (12), B is the second, and df(A) and df(B) are their respective degrees of freedom.*

Case 3. A single function does not fit the data, but the lack of fit is due to both differences in the intercepts and a lack of parallelism among the individual functions. This is the case in which SSE(2) is significantly greater than SSE(1), but SSE(3) is not significantly greater than SSE(2). The appropriate estimate of random measurement variability in this case is

$$S^2 = SSE(1)/df(1) = SSE(1)/[r(n - p)]. \qquad (14)$$

Lack of parallelism in the individual functions is a significant contributor to run-to-run variability in this case (and the next). When the individual functions are not parallel, the estimator $S_R^2(Y_0|X_0)$ may depend upon Y_0 in a complicated fashion. In these cases, the estimated run-to-run variability at $Y = Y_0$ can often be computed from that at $Y = \overline{Y}$ by means of the formula

$$S_R^2(Y_0|X_0) = C^2 S_R^2(\overline{Y}|\overline{X}) \equiv C^2 S_R^2(\overline{Y}) \quad , \qquad (15)$$

where C is an appropriate scale factor. The choice of a suitable scale factor may involve some trial-and-error. If $S_R(Y_0|X_0)$ is proportional to Y_0, for example, then the correct choice for C is $C = Y_0/\overline{Y}$. Another choice of C which sometimes works is $C = (Y_0/\overline{Y})^{1/2}$.

A suitable estimator of $S_R^2(\overline{Y})$ in this case is

$$S_R^2(\overline{Y}) = \frac{SSE(3) - SSE(1)}{n(r - 1)} \quad . \qquad (16)$$

The estimator has degrees of freedom $df(3) - df(1) = p(r - 1)$.

---

*Strictly speaking, (13) should be computed for each value of X_0. In practice, changes in df are small when the term involving $S_R^2(-|-)$ dominates.

For a linear function, (8) becomes

$$s^2(Y_0|X_0) = s^2\left(\frac{1}{nr} + \frac{(X_0 - \bar{X})^2}{\sum_i \sum_j (X_{ij} - \bar{X})^2}\right) + c^2 s_R^2(\bar{Y}) \quad , \qquad (17)$$

where s^2 is given by (14) and $s_R^2(\bar{Y})$ by (16). The first term on the right-hand side of (17) has r(n-2) degrees of freedom and the second term has 2(r-1) degrees of freedom. The estimators (14) and (16) are significantly different in this case, so the approximate degrees of freedom for (17) can be computed from (13), where A and B are taken to be the first and second terms of (17), respectively, and df(A) = r(n-2) and df(B) = 2(r-1).

As in previous cases, a formula analogous to (17) can also be obtained for a quadratic function.

Case 4. Parallel functions do not fit the data, but the lack of fit for a single function is significantly greater than the lack of fit for parallel functions. In this case SSE(2) is significantly greater than SSE(1) and SSE(3) is significantly greater than SSE(2). As in Case 3, (14) is used to estimate random measurement variability and (16) is used to estimate run-to-run variability. However, the differences D_3 = SSE(3) - SSE(2) and D_1 = SSE(2) - SSE(1) are significantly different in this case. Consequently, the degrees of freedom associated with the estimator (16) is computed from Satterthwaite's formula (13) with A = $D_3/[n(r-1)]$, B = $D_1/[n(r-1)]$, df(A) = df(3) - df(2) = r - 1 and df(3) = df(2) = df(1) = (r-1)(p-1), where p-1 is the degree of the polynomial function being fitted. From this point, estimation proceeds as in the previous case: The estimates (14) and (16) are substituted into the appropriate form of (17), depending upon whether the function is linear or quadratic, and the degrees of freedom is computed by means of (13).

EXAMPLE

Figure 1 shows a profile plot of an actual set of aligned calibration data. A profile plot is simply a plot of the residuals (versus X, and connected for each run) resulting from a linear least-squares regression fit to the entire data set. Thus, a profile plot shows clearly the nonlinearity of the calibration function. It is apparent from Figure 1 that the residuals from Run 1 are smaller than those from the other three runs. The discrepancy between Run 1 and the others is the major source of measurement uncertainty in this case.

For comparison, Figure 2 shows a plot of the data for increments 27-31 (105-122 kg, approximately) with the vertical scale expressed in the original units. The total height of the tank is approximately 415 cm (164 in.) and the total volume is approximately 210 liters.

After examining the residuals more closely, it was decided that the data could be adequately fit in three major segments. The three segments consist of increments 8-28 (30-110 kg, approximately), increments 29-38 (114-150 kg) and increments 39-53 (150-208 kg). The first two segments were fit with a linear function, but the third required a quadratic polynomial. Residuals from these fits are shown in Figure 3, along with those from fits for the smaller segments (not discussed in this paper). As in Figure 1, it is evident that the residuals from the first run deviate from those for other runs. Also shown in Figure 3 are confidence limits derived by the methods proposed in the preceding section; computational details for the first segment are given below. The V-shaped envelope indicates measurement accuracy criteria which were established for this tank.

For the segment consisting of increments 8-28, Table 1 shows the numerical values of the sum of squares identified in the second section. For these data, the values of (2), (5) and (6) are $T_1 = 101.33$, $T_2 = 1067.42$ and $T_3 = 422.78$. All are significantly large, so the discussion for Case 4 of the preceding section applies. For illustrative purposes, variances are estimated at the mean $(\overline{Y}, \overline{X})$. The estimate of random measurement variance, from (14), is

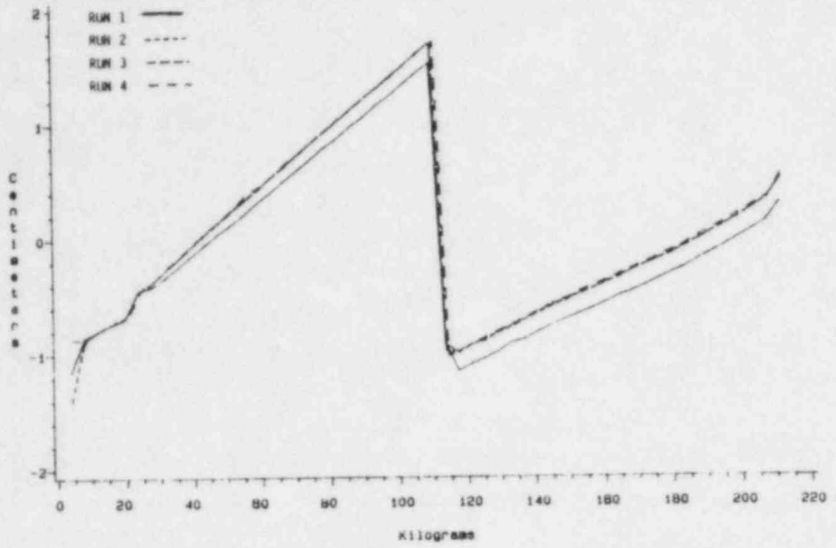

FIGURE 1. Profile Plot of Long Tube Pressures

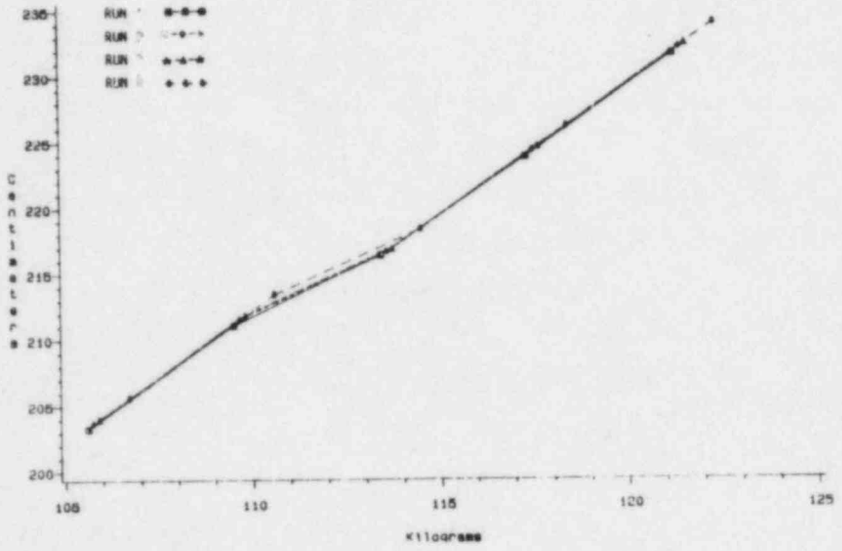

FIGURE 2. Long Tube Pressure vs. Weight, Increments 27-31

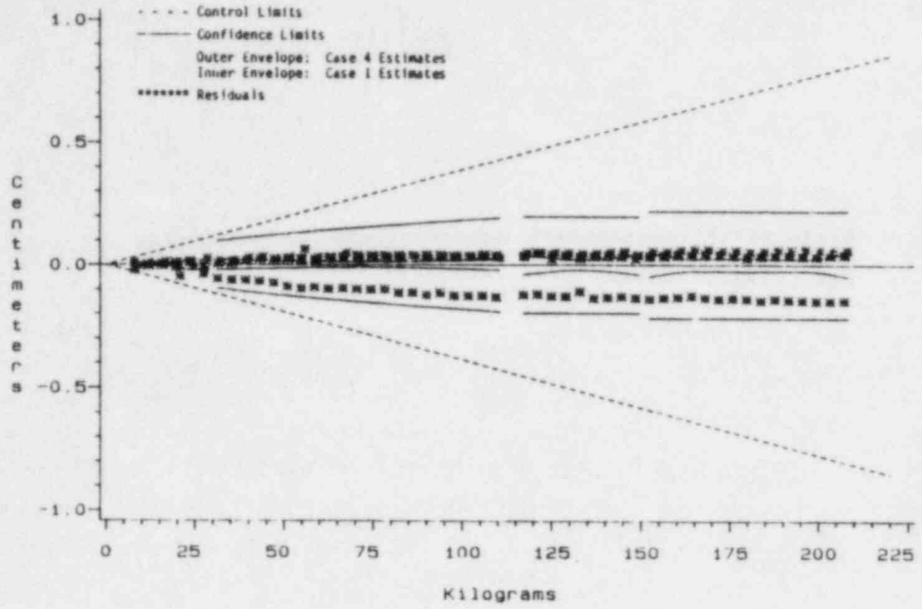

FIGURE 3. Control Limits, Confidence Limits and Residuals from Fitted Calibration Function

$$S^2 = 0.0033303/76 = 0.0000438 \quad . \tag{18}$$

From (16), the estimate of run-to-run variance is

$$S_R^2(\overline{Y}|\overline{X}) = \frac{0.2839689 - 0.0033303}{21 * 3}$$

$$= 0.0044546 \quad . \tag{19}$$

From (13), the degrees of freedom for $S_R^2(\overline{Y}|\overline{X})$ is 3.30. Substitution of (18) and (19) into (17) yields

$$S^2(\overline{Y}|\overline{X}) = 0.0000438 \left(\frac{1}{84}\right) + 0.0044546$$

$$= 0.0044551 \quad . \tag{20}$$

TABLE 1. Sums of Squares and Degrees of Freedom for the Segment Consisting of Increments 8-28

Model	SSE	df
Individual Functions, Each Run	SSE(1) = 0.0033303	df(1) = 76
Parallel Functions, Each Run	SSE(2) = 0.0166503	df(2) = 79
Single Function, All Runs	SSE(3) = 0.2839689	df(3) = 82

From (13), the degrees of freedom for $S^2(\overline{Y}|\overline{X})$ is computed to be 3.36. The 90% confidence limits computed from (17) using (19), (20), and $C = Y_0/\overline{Y}$ yield the outer envelope shown in Figure 3.

Had run-to-run differences been ignored, and had (9) been used to compute $S^2(\overline{Y}|\overline{X})$ on the basis of SSE(3), the result comparable to (20) is

$$S^2(\overline{Y}|\overline{X}) = \frac{0.2839689}{82}\left(\frac{1}{84}\right) = 0.00004123 \quad , \qquad (21)$$

with corresponding degrees of freedom 82. Thus, it is clear that ignoring run-to-run differences results in overestimation of the first term in (8) and underestimation of the second. When the second term dominates, measurement variability is underestimated. As this example shows, the error can be extremely large. Corresponding confidence intervals are even more disparate because of the widely differing degrees of freedom. The discrepancy may be seen in Figure 3 by comparing the outer envelope of confidence intervals (Case 4) with the inner envelope, the latter envelope having been obtained by substitution of (21) into (9).

Also, by way of comparison, the first run was omitted and the calculations were redone. The new value of (2) is

$$T_1 = \frac{0.0035470 - 0.0027614}{61 - 57} \div \frac{0.0027614}{57} = 4.05 \quad .$$

Although this value is still significant, no serious distortion of estimated variances occurs (relative to control limits) if run-to-run differences are ignored. The Case 1 estimate of (9) is

$$s^2(\overline{Y}|\overline{X}) = \frac{0.0035470}{61}\left(\frac{1}{63}\right) = 0.923 \times 10^{-7}$$

with corresponding degrees of freedom 61.

ACKNOWLEDGEMENTS

This work was supported by the U.S. Department of Energy under Contact DE-AC06-76RLO 1830. The author wishes to thank D. J. Bates for assistance with the calculations, S. M. Popp for her efficient preparation of this manuscript, and D. D. Scott for providing the financial support to prepare this paper.

REFERENCES

BROWNLEE, K. A. (1965), Statistical Theory and Methodology in Science and Engineering (2nd Ed.), New York: John Wiley and Sons, Inc.

GOLDMAN, A. S. and LIEBETRAU, A. M. (1984), "The Analysis of Tank Calibration Data from Several Runs," paper presented at the 25th Annual INMM Meeting, Columbus, Ohio, July 15-18, 1984.

JONES, F. E. (1984), "A Tank Volume Calibration Algorithm," paper presented at the 25th Annual INMM Meeting, Columbus, Ohio, July 15-18, 1984.

LIEBETRAU, A. M. (1984), "A Program for the Automated Acquisition of Tank Calibration Data," in Proceedings of the Conference on Safeguards Technology: The Process-Safeguards Interface, ed. E.A. Hakkila, M.H. Campbell, and N.M. Trahey, pp. 96-105, CONF-831106, U.S. Department of Energy, New Brunswick Laboratory, Argonne, IL.

SAMPLING INSPECTION OF NUCLEAR POWER PLANTS

Julius Goodman
Bechtel Power Corporation

ABSTRACT

Advantages of the sampling method are discussed. A classification of different sampling procedures and methods of statistical evaluations is provided. The method delivering the shortest confidence interval of the parameter, under the evaluation given evidence, is developed. New results in the sampling by attributes, kinds and variables are obtained. In cases when it is impossible to select a random sample because of the large size of the population or limited accessibility the cluster sampling procedure with a likelihood density function evaluation method is proposed. A method allowing credit for human error during inspection is also developed.

INTRODUCTION

Sampling inspection of nuclear power plants is used for quality assurance, reinspection and reevaluation, and verification of compliance with different regulations. There are four principal advantages of sampling over a complete inspection of a population. These advantages will be discussed below closely following Cochran (1977).

Reduced Cost

The cost of inspection is proportional to the size of the sample. If we inspect only a small fraction of the population, we can significantly reduce the cost. Because the precision of the evaluation is inversely proportional to the sample size, the correct choice of sample size is a typical tradeoff between accuracy and cost.

Greater Speed

For the same reason, the data can be collected and evaluated quickly using a sample rather than a complete population. This is a vital consideration when the information is urgently needed.

Greater Scope

In many cases highly trained personnel or specialized equipment, limited in availability, must be used to obtain the data. A complete inspection is technically impossible: the choice lies between obtaining the information by sampling or not at all. Therefore, under these conditions sampling has a broader scope than an ill-conducted inspection of the entire population.

Greater Accuracy

Because personnel of higher quality can be used and given intensive training and careful supervision in the field, better inspection and processing of the results becomes feasible when the volume of work is reduced. As a result, a sample may produce a more accurate solution than the complete inspection of the population.

The sampling procedure can also reduce paper work by verifying QC and QA inspection effectiveness and credibility of the inspector's decision that the product is acceptable. However, for valid statistical predictions, major statistical rules have to be followed. The purpose of this paper is to discuss the state-of-the-art in the sampling inspection.

CLASSIFICATION OF SAMPLING PROCEDURES AND METHODS

Sampling procedures and methods can be classified by several signs: number of phases, complexity, types and mathematical methods.

According to the number-of-phases classification we can distinguish a single sampling plan, a double sampling plan and multiple sampling plans. The purpose of multiple sampling is to minimize the cost of an inspection when there is a significant risk that the result of a small sample will be indecisive. In this case we increase the size of the sample, step by step, looking for a chance that a sampling procedure will be terminated before we begin the largest sample.

The three kinds of sampling are classified by various degrees of complexity: simple, stratified and cluster sampling. In simple sampling we draw a random sample from the entire population. In stratified and cluster sampling the population is first divided into subpopulations (strata or clusters) and a random sampling is drawn from the subpopulations separately. There are several reasons to depart from simple sampling: large unaccountable population, need of more precise knowledge of different strata, existence of clusters with different properties.

The classification by types includes a sampling inspection by attributes, kinds and variables. An inspection by attributes uses "pass-nonpass" criteria; however, to reduce a sample size or resolve ambiguity with marginally acceptable items an inspection by kinds and variables is also used. A detailed discussion of these types of sampling inspections will follow later.

The classification of sampling by applicable mathematical methods is given in Fig. 1. A hypothesis testing approach gives the probability that a lot of quality p (p is a fraction of defective items or nonconformances) given acceptance and rejection criteria will be accepted or rejected if submitted. It does not provide for the probability that a lot in question has a quality p.

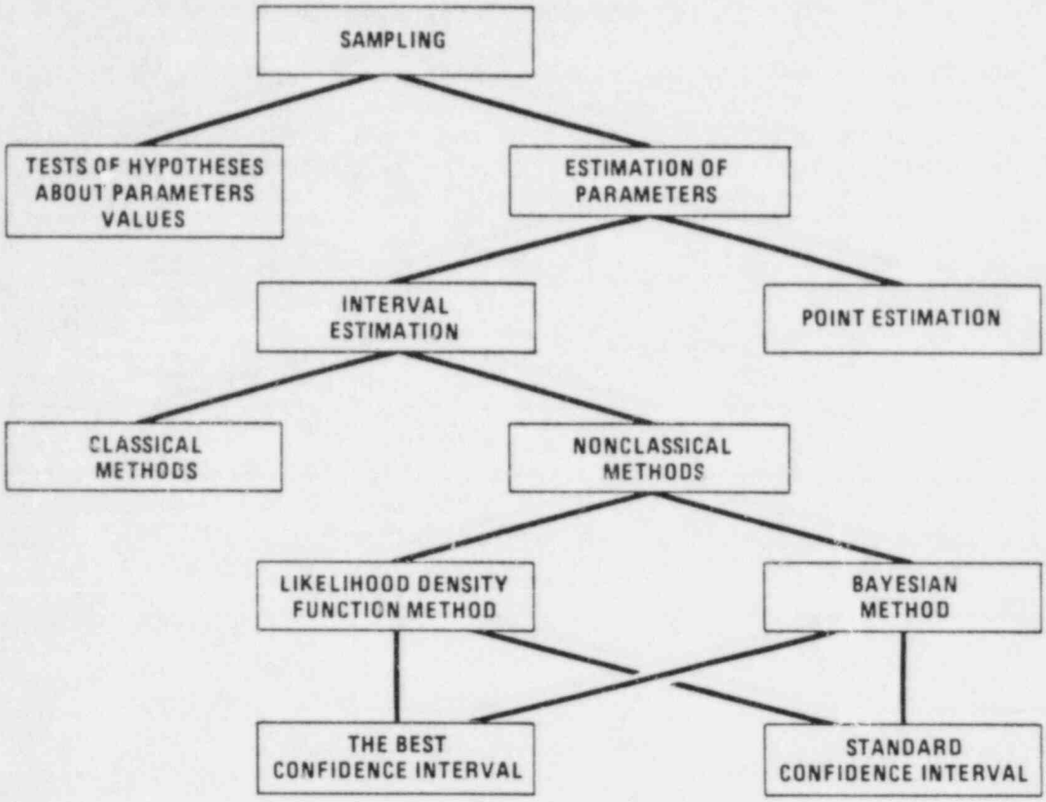

Figure 1. FLOW-CHART OF SAMPLING EVALUATION METHODS

To the contrary, the estimation approach is trying to make a direct or indirect estimation of the lot quality p. The point estimation provides a number and the interval estimation provides a confidence interval for the parameter under evaluation. One of the popular methods of the point estimation is a maximum likelihood method.

The interval estimation can be divided into two approaches: classical and nonclassical. Classical methods provide the confidence interval for a sample estimate and the nonclassical methods provide an unknown population parameter. Specially the classical methods provide the probability of a given observation assuming different values of the unknown population parameter; the nonclassical methods provide a direct evaluation of the probability of the population parameter with the given evidence. The nonclassical methods use two techniques: Bayes' theorem or the likelihood density function method. The result of the interval estimation can be presented in the form of standard or the best confidence interval (Goodman, 1984a).

Hypothesis testing and nonclassical methods are widely described in literature (see, for example, Winkler and Hays, 1975). Therefore, we will concentrate our attention on nonclassical methods.

NONCLASSICAL METHODS OF INTERVAL ESTIMATION

The Bayesian method is known as 1763 (Bayes, 1763). A recent application to the nuclear industry can be found in (Kaplan, 1984). If the sample size n is taken from the population of size N and k nonconformances are observed, then a Bayesian probability $f(m|k,n,N)$ of m nonconformances in the population can be obtained from the Bayes' theorem:

$$f(m|k,n,N) = \frac{\pi(m)L(k,n,N|m)}{\sum_{m=k}^{N-n+k} \pi(m)L(k,n,N|m)} \qquad (1)$$

where $\pi(m)$ is a prior probability and $L(k,n,N|m)$ is a likelihood function.

In the case of sampling from the finite population the likelihood function is a hypergeometrical distribution:

$$L(k,n,N|m) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad (2)$$

Determining the prior probability $\pi(m)$ is beyond the scope of the Bayesian approach. Usually a previous experience or knowledge, or expert opinion is used.

If the population size $N \to \infty$ while n=constant and

$$\lim_{N\to\infty} \frac{m}{N} = p \qquad (3)$$

where p is a fraction of defects or nonconformances then the hypergeometrical distribution turns into the binomial distribution.

$$L(k,n|p) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (4)$$

and Bayesian theorem takes the form:

$$f(p|k,n) = \frac{\pi(p)L(k,n|p)}{\int_0^1 \pi(p)L(k,n|p)dp} \qquad (5)$$

Practically, we can use the binomial distribution if $N \geq 10{,}000$ and $N \leq 0.1N$.

Bayesian probability $f(m|k,n,N)$ or density function $f(p|k,n)$ can be upgraded with new information. In this case, the result of the previous evaluation should be considered as prior distribution and the likelihood function can be constructed on the results of the new sampling (Kaplan, 1984).

Another nonclassical method is a likelihood density function method. This method was used in the past (see, for example, Wilks, 1941, and Goodman, 1969) without recognition as a separate method. According to this method we define the likelihood density function as a normalized likelihood function:

$$f(m|k,n,N) = \frac{L(k,n,N|m)}{\sum_{m=k}^{N-n+k} L(k,n,N|m)} \tag{6}$$

or

$$f(p|k,n) = \frac{L(k,n|p)}{\int_0^1 L(k,n|p)dp} \tag{7}$$

Expressions (6) and (7) can be readily obtained from the Bayesian formulas (1) and (5) if the prior distribution is assumed uniform. This means that we ignore any prior knowledge. However, it makes this method objective because any controversy in selecting the prior distribution - the target of 200 years of criticism - is removed.

Data upgrading or multiple sample plan can be easily handled with the likelihood density function method. It can be shown that during any phase of the Binomial and hypergeometrical distributions we can use formulas (6) and (7) with accumulated numbers of defects $k=k_1+k_2+\ldots$ and combined sample size $n=n_1+n_2+\ldots$ It simplifies the calculations because formulas (6) and (7) can be rewritten in the close form:

$$f(m|k,n,N) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N+1}{n+1}} \tag{8}$$

$$f(p|k,n) = \frac{(n+1)!}{k!(n-k)!}\, p^k\, (1-p)^{n-k} \tag{9}$$

Nonclassical methods allow an interval estimate of unknown parameters (m or p) to be made directly. The standard confidence interval with confidence γ can be defined as:

$$m_{low} \leq m \leq m_{up} \qquad (10)$$

$$p_{low} \leq p < p_{up}$$

where p_{low} and p_{up} are solutions of the equations:

$$\int_0^{p_{low}} f(p|k,n)dp = \frac{1-\gamma}{2} \qquad (11)$$

$$\int_0^{p_{up}} f(p|k,n)dp = \frac{1+\gamma}{2} \qquad (12)$$

and m_{low} and m_{up} are approximate solutions to the equations:

$$\sum_{m=k}^{m_{low}} f(m|k,n,N) = \frac{1-\gamma}{2} \qquad (13)$$

$$\sum_{m=k}^{m_{up}} f(m|k,n,N) = \frac{1+\gamma}{2} \qquad (14)$$

because m_{low} and M_{up} should be integer numbers.

The binomial distribution of all three methods (hypothesis testing classical and nonclassical estimations) will give the same result even though the meaning of these estimates are different. The hypergeometrical distribution does not produce the same results: the nonclassical estimate produces a direct evaluation and a shorter interval. However, the shortest interval, or the best statistical evaluation, can be achieved with a technique proposed by the author (Goodman, 1984a). The comparison of the standard and the best confidence intervals are illustrated in Table 1.

All methods are implemented into two computer codes: BINOM, based on the binomial likelihood function, and HYPER, based on the hypergeometrical likelihood function. As an example, the multiple sample plan based on 95/95 criterion (95 percent confidence and 95 percent reliability, i.e., 5 or less percent for defects) is illustrated in Table 2. This plan provides the minimal sample size for acceptance of the population given the number of observed defects. However, the risk of rejection to a population with a proportion of defects less than 5 percent is very high during the first several phases. Therefore, a cost-benefit analysis is required to determine the size of the beginning sample.

TABLE 1. COMPARISON OF THE BEST AND STANDARD 90 PERCENT CONFIDENCE INTERVALS

Sample Size	Number of Defects	Standard 90 percent confidence interval				The best 90 percent confidence interval				Ratio of the best interval to the standard interval
n	k	5th percentile	Median	95th percentile	Range	Lower limit	Best estimate	Upper limit	Range	
50	0	0.001	0.014	0.058	0.057	0.000	0.000	0.045	0.045	0.789
50	1	0.007	0.033	0.092	0.085	0.002	0.020	0.075	0.073	0.859
100	1	0.004	0.017	0.047	0.043	0.001	0.010	0.039	0.038	0.884
150	1	0.002	0.011	0.031	0.029	0.001	0.007	0.026	0.025	0.862
150	2	0.005	0.018	0.041	0.036	0.003	0.013	0.036	0.033	0.917
200	2	0.004	0.013	0.031	0.027	0.002	0.010	0.027	0.025	0.926

TABLE 2. MULTIPLE SAMPLE PLAN BASED ON 95/95 CRITERION

Observed number of defects, k, for:		Minimal sample size, n, for acceptance if the population N is				Risk of rejection of the population if the real proportion of defects p is			
Acceptance	Rejection	N=500	N=1000	N=5,000	N>10,000	p=0.005	p=0.01	p=0.025	p=0.05
0	5	53	56	57	59	0.256	0.447	0.775	0.952
1	7	85	88	91	93	7.94×10^{-2}	0.238	0.679	0.950
2	9	110	117	122	124	2.48×10^{-2}	0.128	0.602	0.950
3	11	135	144	150	153	7.61×10^{-3}	6.85×10^{-2}	0.534	0.951
4	13	160	170	177	181	2.32×10^{-3}	3.63×10^{-2}	0.474	0.951
5	15	185	195	204	208	6.95×10^{-4}	1.90×10^{-2}	0.420	0.951
6	17	206	218	230	234	2.06×10^{-4}	9.81×10^{-3}	0.369	0.950
7	18	225	240	255	260	6.28×10^{-5}	5.07×10^{-3}	0.326	0.950
8	20	248	265	280	286	2.12×10^{-5}	2.63×10^{-3}	0.289	0.951
9	21	265	285	305	311	8.88×10^{-6}	1.33×10^{-3}	0.254	0.950
10	23	285	310	330	336	5.66×10^{-6}	6.76×10^{-4}	0.223	0.950

SAMPLING BY KINDS AND VARIABLES

There are two major reasons for using an inspection by kinds and variables:

1) Reduce sample size;

2) Reduce human error during inspection.

In a sample by kinds we divide the items into several categories: acceptable, different degree of marginally acceptable, and unacceptable. Therefore, the results of a misclassification are less of a danger when we have several categories rather than just two. The distribution by categories, or kinds, provides us more information than just the percent defective. Hence, a reduced sample size is needed to provide the same accurate prediction.

Lets define the j deficiency categories. We will have j + 1 of different kinds: j - 1 marginally acceptable (i = 1, 2, ... j - 1), one unacceptable (i = j), and one acceptable (i = 0). The likelihood functions $L(k_1, k_2, \ldots k_j, n, N \mid m_1, m_2, \ldots m_j)$ for finite population and $L(k_1, k_2, \ldots k_j, n \mid p_1, p_2, \ldots p_j)$ for infinite population are hypergeometrical and multinomial distributions correspondly:

$$L(k_1,k_2 \ldots k_j,n,N \mid m_1,m_2,\ldots m_j) = \frac{\binom{m_1}{k_1}\binom{m_2}{k_2}\cdots\binom{m_j}{k_j}\binom{N-m}{n-k}}{\binom{N}{n}} \qquad (15)$$

where

$$m = \sum_{i=1}^{j} m_i \qquad (16)$$

$$k = \sum_{i=1}^{j} k_i \qquad (17)$$

and

$$L(k_1,k_2,\ldots k_j,n \mid p_1,p_2,\ldots p_j) = \frac{n!}{k_1!k_2!\ldots k_j!(n-k)}\, p_1^{k_1} p_2^{k_2} \cdots p_j^{k_j} p^{n-k} \qquad (18)$$

where

$$p = \sum_{i=1}^{j} p_i \qquad (19)$$

Probability function $f(m_1,m_2,\ldots m_j \mid k_1,k_2,\ldots k_j,n,N)$ and probability density function $f(p_1,p_2,\ldots p_j \mid k_1, k_2,\ldots k_j,n)$ can be calculated according to formulas similar to (6) and (7).

The tolerance limit for each kind of deficiency is different. The simpliest approach is to assign an independent tolerance limit for every kind of defect, and adopt an acceptance criterion requiring that these limits will not be exceeded. In this case, inspection by kinds will be reduced to multiple inspection by attributes (see Kaplan, 1984, as an example). However, this approach does not meet safety requirements. Even if every kind of defect does not exceed its tolerance limit, the cumulative effects of all kinds of defects could be dangerous. This situation could be corrected by a new acceptance criterion given in the form:

$$\sum_{i=1}^{j} \frac{m_i}{M_i} \leq 1 \qquad (20)$$

or

$$\sum_{i=1}^{j} \frac{P_i}{P_i} \leq 1 \qquad (21)$$

where M_i is a maximum tolerable number of defects of ith kind and P_i is a maximum tolerable probabilities of defects of ith kind in the population assuming zero defects of all other kinds in the population.

Using functions $f(m_1,m_2,\ldots m_j \mid k_1,k_2,\ldots k_j,n,N)$ or $f(P_1,P_2,\ldots P_j \mid k_j,k_2\ldots k_j,n)$ and Monte Carlo or analytical, or the Latin Hypercube method, we can calculate the confidence in the acceptance criteria (20) or (21) for the given evidence (number of deficiencies $k_1,k_2,\ldots k_j$ in the sample of the size n).

In the case of sample by variables we measure some as continuous variable and establish some tolerance limits for it. This methodology for a normal distribution of the variable was developed by Wilks (1941), and Wald and Wolfowitz (1946). Let m and s be a sample mean and a sample standard deviation. Therefore with every sample size n, a tolerance factor K(n) can be established so that the probability is γ that at least a proportion $1-\alpha$ of the population will be included between limits $m\pm K(n)s$ (or will be less than $m+K_1(n)s$, or greater than $m-K_1(u)s$ for one-sided tolerance factor K_1). The tables for one-sided and two-sided tolerance factors are illustrated, for example, by Bowker and Lieberman (1972).

However, this methodology has some limitations. The major ones are listed below:

1) Tolerance factors are calculated with asymptotic formula and are not accurate for small samples (for example, for n = 3, $\gamma = 0.95$ and $\alpha = 0.10$ the estimated by Bowker and Lieberman (1972) tolerance factor K is less than the exact value by 17 percent)

2) The methodology cannot be applied to distributions other than normal;

3) Acceptance criteria assumes that the conditional probability of acceptance jumps from zero to one at the tolerance limit rather than a smooth transition from zero to one in some boundary area near tolerance limits.

These limitations were removed in the approach developed by the author (Goodman, 1984b). The likelihood density, function can be developed for any type of distribution. The methodology discussed in (Goodman, 1984b) specially addressed the Edgeworth-Kapteyn distribution. This includes normal, lognormal and modified lognormal distributions as a partial case. However, we can expand the logarithm of the likelihood function near the best estimate point into Taylor's series and fit it with an appropriate Edgeworth-Kapteyn distribution.

Another refinement is the introduction of the function a(x) (see Figure 2 for one-sided tolerance limit). Now, we have three domains of the variable x:

$a(x) = 1$ acceptance domain (D_a) (22)

$o<a(x)<1$ doubtful domain (D_d) (23)

$a(x) = 0$ rejection domain (D_r) (24)

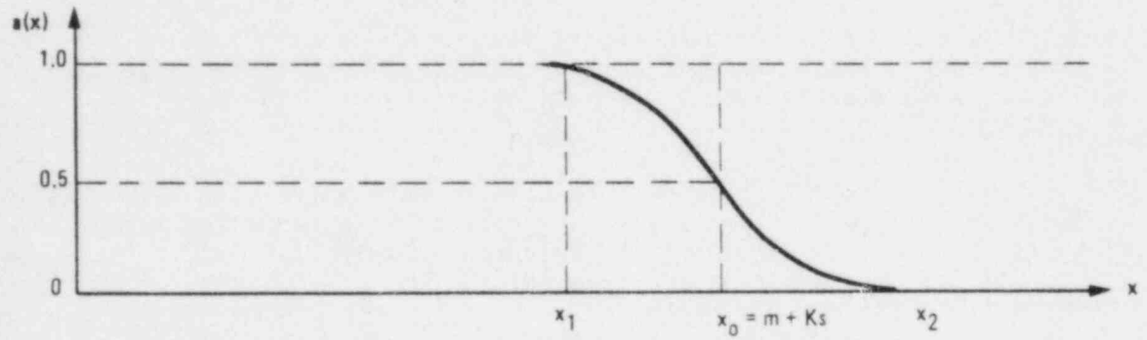

Figure 2. ACCEPTANCE FUNCTION a(x)

x_o is a tolerance limit; $x<x_1$, is an acceptance domain; $x_1 \leq x \leq x_2$ is a doubtful domain; $x > x_2$ is a rejection domain.

Previous methodology uses the acceptance function as:

$$a(x) = \begin{cases} 1, & x < x_o \\ 0, & x > x_o \end{cases} \tag{25}$$

where $x_o = m + Ks$ for an upper one-sided tolerance limit.

The methodology of calculation for a one-sided tolerance limit, for example, follows:

(1) Generate randomly parameters $\vec{\beta}(\beta_1, \beta_2, \ldots \beta_p)$ of the likelihood density function $f(x;\vec{\beta})$;

(2) Calculate integral $P = \int_{-\infty}^{+\infty} f(x; \vec{\beta}) a(x) dx$;

(3) Compare integral P with $1-\alpha$; if $P \geq 1-\alpha$ then SUM = SUM+1;

(4) Repeat steps (1)-(3) N times and, then, calculate γ = SUM/N

As a result we will calculate $\gamma = \gamma(K)$. Inverting this function we can obtain a tolerance factor K as a function of α and γ.

CLUSTER SAMPLING

If the population is too large to be accountable or part of the population is not accessable then we cannot generate a random, unbiased sample. Therefore, we need a methodology to make prediction using biased samples.

A possible approach is the cluster sample. Assume that the population is divided into N clusters. If the subpopulation of the ith cluster is designated as n_i and frequency of defects in the subpopulation as P_i then the frequency of defects in the population is:

$$P = \frac{1}{n} \sum_{i=1}^{N} n_i p_i \tag{26}$$

where:

$$n = \sum_{i=1}^{N} n_i \tag{27}$$

In our analysis we adopt two assumptions:

(1) The subpopulation size n_i for any randomly selected cluster can be accountable;

(2) Numbers n_i and p_i for all clusters provides smooth distributions.

If we randomly select N_s clusters and within each cluster we count n_i and p_i (using a 100 percent inspection or a simple sampling), then a cluster estimate of the frequency of defects is:

$$p_s = \frac{1}{n_{(s)}} \sum_{i=1}^{N_s} n_i p_i \tag{28}$$

where

$$n_{(s)} = \sum_{i=1}^{N_s} n_i \tag{29}$$

Using sample n_i and p_i we can develop a distribution for them. Then the population frequency of defects can be estimated as:

$$p = \frac{1}{n}\left[\sum_{i=1}^{N_s} n_i p_i + \sum_{i=N_s+1}^{N} n_i p_i\right] \tag{30}$$

where:

$$n = \sum_{i=1}^{N_s} n_i + \sum_{i=N_s+1}^{N} n_i \tag{31}$$

The first series in equations (30) and (31) are based on the sample estimate and the second series are based on Monte Carlo simulation. The formula (30) can be presented in the form:

$$p = \frac{n_{(s)}}{n} p_s + \frac{1}{n} \sum_{i=N_s+1}^{N} n_i p_i \tag{32}$$

The methodology described in (Goodman, 1984b) is used to develop distributions of n_i and p_i ($N_s + 1 \leq i \leq N$). The distributions of p_i ($i \leq N_s$) when a simple sample is used to assess them are binominal or hypergeometrical likelihood density functions.

The uncertainty of evaluation P according to formula (32) depends on the number of randomly selected clusters. The fewer number of clusters the more uncertainty there is. The confidence interval for P is larger than for simple sampling and this is a price of using a biased sample.

HUMAN ERROR IN SAMPLING PROCEDURE

All methods of sampling evaluation are based on the assumption that any sample procedure excludes a human error. However, it is not so. There are many sources of human error. Some of them are objective like poor instrumentation, or poor training, or poor instructions. Some of them are subjective like tireness, rush, mood, etc. Therefore, our evaluation of the frequency of defects should incorporate the uncertainty of human error during inspection.

Consider the case when we have QC and QA inspections and QA can catch the items that QC misses. For simplicity we assume that QC inspectors can commit only one-sided error: qualify a defective item as nondefective.

Let n_{QC} and n_{QA} be sample sizes, k_{QC} and k_{QA} be the number of observed defects, p_{QC} and p_{QA} be the probabilities of error for QC and QA inspection correspondly. Then the probability of QC error p_{QC} and the frequency of defects in the population can be estimated by formulas:

$$p_o = \frac{\left(\frac{k_{QC}}{n_{QC}}\right)}{1-P_{QC}} \tag{33}$$

$$p_{QC} = \frac{1 - p_{QC}}{1 - p_{QA}} \cdot \frac{\left(\frac{k_{QA}}{n_{QA}}\right)}{\left(\frac{k_{QC}}{n_{QC}}\right)} \cdot \left(1 - \frac{k_{QC}}{n_{QC}}\right) \tag{34}$$

The probabilities p_o and p_{QC} are expressed through observed data k_{QC}, k_{QA}, n_{QC}, n_{QA} and the probability of QA error.

If

$$p_{QA} << 1 \tag{35}$$

then the expression (34) is not sensitive to the particular value of p_{QA}. Because the qualification of QA inspectors is better than QC

inspectors and conditions of inspection is favorable we can assume that

$$p_{QA} \leq p_{QC} \tag{36}$$

We developed a computer code QCQA which evaluates the uncertainty distributions of p_o and p_{QC}. In this code the uniform distribution of p_{QA} is incorporated assuming that mean values of p_{QC} and p_{QA} coincides.

Lets consider a numerical example. Say, $n_{QC} = 2047$, $k_{QC} = 742$, $n_{QA} = 99$, $k_{QA} = 7$. Then the result of calculation is shown in Table 3.

TABLE 3. FREQUENCY OF DEFECTS P_o WITH HUMAN ERROR, P WITHOUT HUMAN ERROR, AND PROBABILITY OF QC INSPECTORS' ERROR P_{QC}

	Lower Limit (5th percentile)	Median (50th percentile)	Upper Limit (95th percentile)
P_o	0.386	0.414	0.454
p	0.345	0.363	0.380
P_{QC}	0.075	0.122	0.195

We can see that without taking into account of QC inspectors' error the estimate of the frequency of defects is significantly lower. The real frequency p_o is greater than apparent frequency p by 14 percent for medians and by 19.5 percent for upper limits. Therefore, the incorporation of human error correction into sampling procedure is very important.

CONCLUSION

The statistical sampling methodology discussed in this paper can be implemented by quality assurance groups for audit and surveillance of the quality control programs at nuclear power plants to provide a high level of confidence in the quality of workmanship and adherence to NRC requirements.

REFERENCES

Bayes, T. (1763), "An Essay Toward Solving a Problem in the Doctrine of Chances," *Philosophical Transactions of the Royal Society of London.*

Bowker, A. H. and Lieberman, G. J. (1972), *Engineering Statistics* (2nd edition), Englewood Cliffs, New Jersey: Prentice-Hall

Cochran, W. G. (1977), *Sampling Technique* (3rd edition), New York: John Wiley and Sons

Goodman, J. (1969), "Test as an Instrument of Estimating Students' Knowledge," Fifth Conference on Programming Education and Technical Means of Education, Symposium 7, 29-33, Moscow

_____ (1984a), "On the Definition of the "Best" Confidence Interval," *Reliability Engineering,* 7, 213-228

_____ (1984b), "Estimating Fragility Curves Using Few Experimental Data," Proceedings of the Symposium on Advances in Probabilistic Structural Mechanics at the 1984 Pressure Vessel and Piping Conference and Exhibition, PVP - Vol. 93, 41-52, San Antonio, Texas, June 17-21

Kaplan, S. (1984), "A Bayesian Framework for the Sampling Activity in the Zimmer PVQC," Pickard, Lowe and Garrick, Inc., PLG-0329, Prepared for Cincinnati Gas and Electric Company, Cincinati, Ohio, January

Wald, A. and Wolfowitz, J. (1946), "Tolerance Limits for a Normal Distribution," *Annals of Mathematical Statistics,* 17, 208-215

Wilks, S. S. (1941), "Determination of Sample Sizes for Setting Tolerance Limits," *Annals of Mathematical Statistics,* 12, 91-96

Winkler, R. L. and Hays, W. L. (1975), *Statistics: Probability Inference, and Decision* (2nd edition), New York: Holt, Rinehart and Winston.

1984 STATISTICS SYMPOSIUM ON NATIONAL ENERGY ISSUES

REGISTRANTS

Lee Abramson
Nuclear Regulatory Commission
Washington, DC 20555
301-443-7628

Richard J. Beckman
Los Alamos National Laboratory
P.O. Box 1663, Mail Stop F600
Los Alamos, NM 87545
505-667-3308

Carl A. Bennett
Battelle-HARC
4000 N. E. 41st Street
Seattle, WA 98105
206-525-3130

Deborah E. Bennett
Lawrence Livermore National Laboratory
P.O. Box 808, L-316
Livermore, CA 94550
415-423-2056

Jane M. Booker
Los Alamos National Laboratory
P.O. Box 1663, Mail Stop F600
Los Alamos, NM 87545
505-667-3308

Lawrence A. Bruckner
Los Alamos National Laboratory
M/S F600
Los Alamos, NM 87545
505-667-6246

Gary R. Burdick
U.S. Nuclear Regulatory Commission
M/S 1130-SS
Washington, D. C. 20555

Ronald V. Canfield
Utah State University
UMC 24
Logan, UT 84322
801-750-2434

G. M. Christensen
Rockwell Hanford Operations
P.O. Box 800
Richland, WA 99352
509-373-2827

Pamela G. Doctor
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352
509-376-4326

Mark J. Durst
Lawrence Livermore National Laboratory
P.O. Box 808, L-316
Livermore, CA 94550
415-422-4272

Richard Engelder
Bendix Field Engineering
P.O. Box 1569
Grand Junction, CO 81502
303-242-8621

Ethel Gilbert
Battelle-Northwest
P.O. Box 999
Richland, WA 99352
509-376-4424

Richard O. Gilbert
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352
509-376-4218

Ronald E. Glaser
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94550
415-423-0681

Julius Goodman
Bechtel Power Corporation
12400 E. Imperial Hwy
Norwalk, CA 90650
213-807-4061

Robert Jacobs
Rockwell Hanford Operations
P.O. Box 300
Richland, WA 99352
509-373-2073

Samuel C. Kao
Brookhaven National Laboratory
Applied Math Dept., 515, BCL
Upton, NY 11973
516-282-4138

Robert R. Kinnison, Ph.D.
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352
509-376-4760

Les Lancaster
U.S. Nuclear Regulatory Commission
MS 5650-NL
Washington, DC 20555
301-443-7617

James Lechner
NBS
Admin. A337
Gaithersburg, MD 20899
301-921-3651

A. M. Liebetrau
Pacific Northwest Labortory
P.O. Box 999
Richland, WA 99352
509-376-4761

David S. Margolies
Lawrence Livermore National Laboratory
P.O. Box 808, L-316
Livermore, CA 94550
415-422-0591

Timothy Margulies
U.S. Nuclear Regulatory Commission
MS NL-5650
Washington, DC 20555
301-443-7626

Harry F. Martz
Los Alamos National Laboratory
P.O. Box 1663, Mail Stop F600
Los Alamos, NM 87545
505-667-3308

Teresa Meachum
EG&G Idaho
1520 Sawtelle
Idaho Falls, ID 83401
208-526-9029

William Q. Meeker
Iowa State University
Department of Statistics
Ames, Iowa 50010

Richard W. Mensing
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94550
415-422-4269

Dan Moore
Lawrence Livermore National Laboratory
P. O. Box 808
Livermore, CA 94550

Roger H. Moore
BPA - PNE
Box 3621
Portland, OR 97208
503-230-3697

L. Craig Murray
Southern California Edison Company
P.O. Box 800
Rosemead, CA 91770
818-302-2697

David L. Nelson
Boeing Computer Services
P.O. Box 24346
Seattle, WA 98124
206-575-5070

Neal Oden
Brookhaven National Laboratory
Bldg. 475
Upton, N.Y. 11790
516-282-2060

A. R. Olsen
Pacific Northwest Laboratory
Richland, Washington 99352
509-376-4265

Dale M. Rasmuson
U. S. Nuclear Regulatory Commission
NL 5650
Washington, D. C. 20555

David Rose
Boeing Computer Service
4919 E. Mercer Way
Mercer Island, WA 98040
206-236-2711

David Rubinstein
U.S. Nuclear Regulatory Commission
Washington, D. C. 20555
301-492-4723

Paul D. Sampson
Dept. of Statistics
University of Washington
Seattle, WA 98195
206-545-2664

Bobby Scott
Lovelace Research ITRI
P.O. Box 5890 Bldg., 9200 Area Y
KAFB Albuquerque, NM 87185
505-844-8970

Jeanne C. Simpson
Pacific Northwest Laboratory
P.O. Box 999
Richland, WA 99352
509-376-4327

Nora G. Smiriga
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94550
415-422-4281

Floyd W. Spencer
Sandia National Laboratories
Kirkland AFB
Albuquerque, New Mexico 87185
505-844-5647

Judy Steverson
EG&G Idaho Inc.
1520 Sawtelle
Idaho Falls, ID 83442
208-526-1241

Gary Tietjen
Los Alamos National Laboratory
Box 1663, MS F600
Los Alamos, NM 87545
505-667-3308

Steve Verrill
Lawrence Livermore National Laboratory
P.O. Box 808, L-316
Livermore, CA 94550
415-422-4014

Ray A. Waller
Los Alamos National Laboratory
525 Navajo
Los Alamos, NM 87544
505-667-4567

DISTRIBUTION

No. of Copies

OFFSITE

2 U.S. Nuclear Regulatory Commission
Division of Technical Information and Document Control
7920 Norfolk Avenue
Bethesda, MD 20014

Dale M. Rasmuson
U.S. Nuclear Regulatory Commission
Mail Stop NL-5650
Washington, DC 20555

Lee Abramson
U.S. Nuclear Regulatory Commission
Washington, DC 20555

Richard J. Beckman
Los Alamos National Laboratory
P.O. Box 1663, Mail Stop F600
Los Alamos, NM 87545

Carl A. Bennett
Battelle-HARC
4000 N.E. 41st Street
Seattle, WA 98105

Deborah E. Bennett
Lawrence Livermore National Laboratory
P.O. Box 808, L-316
Livermore, CA 94550

Jane M. Booker
Los Alamos National Laboratory
P.O. Box 1663, Mail Stop F600
Los Alamos, NM 87545

No. of Copies

Lawrence A. Bruckner
Los Alamos National Laboratory
Mail Stop F600
Los Alamos, NM 87545

Gary R. Burdick
U.S. Nuclear Regulatory Commission
Mail Stop 1130-SS
Washington, DC 20555

Ronald V. Canfield
Utah State University
UMC 24
Logan, UT 84322

G. M. Christensen
Rockwell Hanford Operations
P.O. Box 800
Richland, WA 99352

Mark J. Durst
Lawrence Livermore National Laboratory
P.O. Box 808, L-316
Livermore, CA 94550

Richard Engelder
Bendix Field Engineering
P.O. Box 1569
Grand Junction, CO 81502

Ronald E. Glaser
Lawrence Livermore National Laboratory
P.O. Box 808
Livermore, CA 94550

Julius Goodman
Bechtel Power Corporation
12400 E. Imperial Hwy
Norwalk, CA 90650

No. of
Copies

Robert Jacobs
Rockwell Hanford Operations
P.O. Box 800
Richland, WA 99352

Samuel C. Kao
Brookhaven National Laboratory
Applied Math Dept., 515, BCL
Upton, NY 11973

Les Lancaster
U.S. Nuclear Regulatory
Commission
Mail Stop NL-5650
Washington, DC 20555

James Lechner
NBS
Admin. A337
Gaithersburg, MD 20899

David S. Margolies
Lawrence Livermore National
Laboratory
P.O. Box 808, L-316
Livermore, CA 94550

Timothy Margulies
U.S. Nuclear Regulatory
Commission
Mail Stop NL-5650
Washington, DC 20555

Harry F. Martz
Los Alamos National Laboratory
P.O. Box 1663, Mail Stop F600
Los Alamos, NM 87545

Teresa Meachum
EG&G Idaho
1520 Sawtelle
Idaho Falls, ID 83401

William Q. Meeker
Iowa State University
Department of Statistics
Ames, IA 50010

No. of
Copies

Richard W. Mensing
Lawrence Livermore National
Laboratory
P.O. Box 808
Livermore, CA 94550

Dan Moore
Lawrence Livermore National
Laboratory
P.O. Box 808
Livermore, CA 94550

Roger Moore
BPA - PNE
Box 3621
Portland, OR 97208

L. Craig Murray
Southern California Edison
Company
P. O. Box 800
Rosemead, CA 91770

David L. Nelson
Boeing Computer Services
P.O. Box 24346
Seattle, WA 98124

Neal Oden
Brookhaven National Laboratory
Bldg. 475
Upton, NY 11790

David Rose
Boeing Computer Service
4919 E. Mercer Way
Mercer Island, WA 98040

David Rubinstein
U.S. Nuclear Regulatory
Commission
Washington, DC 20555

Paul D. Sampson
Dept. of Statistics
University of Washington
Seattle, WA 98195

No. of
Copies

Bobby Scott
Lovelace Research ITRI
P.O. Box 5890 Bldg.,
9200 Area Y
KAFB Albuquerque, NM 87185

Nora G. Smiriga
Lawrence Livermore National
Laboratory
P.O. Box 808
Livermore, CA 94550

Floyd W. Spencer
Sandia National Laboratories
Kirkland AFB
Albuquerque, NM 87185

Judy Steverson
EG&G Idaho Inc.
1520 Sawtelle
Idaho Falls, ID 83442

Gary Tietjen
Los Alamos National Laboratory
Box 1663, Mail Stop F600
Los Alamos, NM 87545

No. of
Copies

Steve Verrill
Lawrence Livermore National
Laboratory
P.O. Box 808, L-316
Livermore, CA 94550

Ray A. Waller
Los Alamos National Laboratory
525 Navajo
Los Alamos, NM 87544

ONSITE

16 Pacific Northwest Laboratory

P. G. Doctor (3)
E. S. Gilbert
R. O. Gilbert
R. R. Kinnison, Ph.D.
A. M. Liebetrau
A. R. Olsen
J. C. Simpson
Publishing Coordination (2)
Technical Information (5)

NRC FORM 335 (2 84) NRCM 1102, 3201, 3202

U.S. NUCLEAR REGULATORY COMMISSION

BIBLIOGRAPHIC DATA SHEET

SEE INSTRUCTIONS ON THE REVERSE

1. REPORT NUMBER (Assigned by TIDC add Vol. No., if any)
NUREG/CP-0063

2. TITLE AND SUBTITLE
Proceedings of the 1984 Statistical Symposium on National Energy Issues

3. LEAVE BLANK

4. DATE REPORT COMPLETED
MONTH: March YEAR: 1985

5. AUTHOR(S)
Robert Kinnison
Pamela Doctor

6. DATE REPORT ISSUED
MONTH: July YEAR: 1985

7. PERFORMING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
Pacific Northwest Laboratory
Richland, Washington 99352

8. PROJECT/TASK/WORK UNIT NUMBER

9. FIN OR GRANT NUMBER
B2386

10. SPONSORING ORGANIZATION NAME AND MAILING ADDRESS (Include Zip Code)
Division of Risk Analysis and Operations
Office of Nuclear Regulatory Research
U.S. Nuclear Regulatory Commission
Washington, DC 20555

11a. TYPE OF REPORT
Conference Proceedings

b. PERIOD COVERED (Inclusive dates)

12. SUPPLEMENTARY NOTES

13. ABSTRACT (200 words or less)

The 1984 Statistical Symposium on National Energy Issues was the tenth in a series of annual symposia bringing together statisticians and other interested parties who are actively engaged in the pursuit of solving the nation's energy problems. Initially the symposium was sponsored by U.S. Department of Energy (DOE) and named the DOE Statisticial Symposium. The symposium is organized by a steering committee made up of representatives from the national laboratories.

The 1984 symposium was hosted by Pacific Northwest Laboratory, and it was organized around four special topical sessions: (1) Assessing and Assuring High Reliability, (2) Spatial Statistical, (3) Quantificaiton of Informed Opinion, and (4) Health Effects of Energy Technologies. These were chosen by the steering committee as topics currently of high importance in energy research and data analysis. Several contributed papers were also presented.

14. DOCUMENT ANALYSIS - a. KEYWORDS/DESCRIPTORS
Statistics, Statistical Symposium
Spacial statistics, informed opinion,
High reliability, health effects

b. IDENTIFIERS/OPEN ENDED TERMS

15. AVAILABILITY STATEMENT

16. SECURITY CLASSIFICATION
(This page) Unclassified
(This report)

17. NUMBER OF PAGES

18. PRICE

UNITED STATES
NUCLEAR REGULATORY COMMISSION
WASHINGTON, D.C. 20555

120555078877 1 1AN
US NRC
ADM-DIV OF TIDC
POLICY & PUB MGT BR-PDR NUREG
W-501
WASHINGTON DC 20555