NAIL FINDERS, EDIFICES, AND OZ

BY

LEO BREIMAN

TECHNICAL REPORT NO. 32
MAY 1984

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA, BERKELEY

NAIL FINDERS, EDIFICES, AND OZ

LEO BREIMAN

The thesis of this paper is that many, if not most, statisticians in government and industry are poorly trained for their profession with the result that they are poor problem solvers in terms of public policy decisions.

Since this is illustrated by anecdotal material in which I often emerge as the hero, I begin with a situation in which my failure was undeniable.

In the late sixties, I was hired as a consultant to the defense in the famous Ellsberg trial. The defense was interested in bringing a challenge to the court concerning the composition of the empanelled jurors. A Supreme Court ruling had held that empanelled jurors should reflect the characteristics of the population in which the particular court had jurisdiction. In the Federal Court hearing the Ellsberg case the jurisdiction consisted of Los Angeles County and four surrounding counties. The Ellsberg defense was specifically concerned with under-representation of blacks.

The basic data consisted of the questionnaires which were mailed to the people selected for jury duty, filled in by them and returned to the Jury Commissioner. We were given access to the about 6000 questionnaires which had been received over the past year. A large majority of these were from people that subsequently had been excused from jury duty because of some circumstance or hardship.

These 6000 were in boxes, which I took to a commercial card punching company together with a coding for various questions and answers. After the data was put onto the cards I started the tabulations to present to

the court.

There were two fatal errors on my part. The first came out when the Jury Clerk, in her testimony, disputed the numbers in my tabulation. She had gone through the questionnaires and counted the number in a certain category. My count was much higher than hers.

With a sinking feeling, I started checking through the output in detail and finally realized what had happened. The card punchers had entered one box of questionnaires twice! My first failure was that I did not ensure

GOOD QUALITY CONTROL ON THE DATA.

The second error was even more fundamental. The critical question about race on the questionnaire had the comment that answering was optional. In the questionnaires submitted by people who were later empanelled, about 15% of the answers were missing. As I recall, about 5% of those answering the question were black, and there were about 13% blacks in the total population of the court jurisdiction. There was no way in which that data could be used to determine whether blacks were under represented. I had failed to ask the fundamental question

CAN THE RELEVANT QUESTIONS BE ANSWERED BY THE DATA?

In my defense, there were mitigating circumstances. I had just come out of the University and started my consulting career. I learned and did not repeat those early mistakes too often.

My experience in meeting, working with, and reviewing the work of many statisticians in field practice is that they generally suffer from one or the other of these three complexes.

THE FIND-THE-NAIL COMPLEX

Jerome Friedman has a lovely saying (source unknown);

"If all you have is a hammer, then every problem will look like a nail."

As applied to statisticians, this refers to absorption with the technique rather than the problem; to the failure to see the problem whole; to ask, "Does it all make sense?"

THE EDIFICE COMPLEX

This refers to the building of a large, elaborate and many layered statistical analysis often covering up what is simple and obvious.

THE WIZARD OF OZ COMPLEX

The exploitation of the mysteries of statistics to dazzle and mystify the less knowledgeable.

Here are some recent illustrations.

THE ASA AD HOC ADVISORY COMMITTEE ON NUCLEAR REGULATORY RESEARCH

This committee was formed by the ASA in 1980. For background, I include the relevant sections of a June 30, 1980 memo from Fred C. Leone, Executive Director of the ASA.

> The Board of Directors of the American Statistical Association is establishing an Ad Hoc Advisory Committee on Nuclear Regulatory Research. This is a result of negotiation between members of the Nuclear Regulatory Commission and the American Statistical Association, followed by an invitation from the Director of the Office of Nuclear Regulatory Research. This is a major step which the ASA Board has taken and, hence, it is especially important that the advisory committee be very strong and have the necessary balance to be most effective. The terms of reference (charge) of the Advisory Committee are stated in the accompanying sheet.

Terms of Reference (Charge) of ASA Ad Hoc Advisory Committee on Nuclear Regulatory Research

This Committee willprovide advice and peer review with respect to programs of the Office of Nuclear Regulatory Research of the U.S. Nuclear Regulatory Commission which involve or require statistical and probabilistic techniques and approaches. In particular, it will

1. take a broad responsibility for the review of the statistical and probabilistic technique proposed for the Numerical Risk Criteria Project and assessment of statistical contributions to risk assessment procedures and applications.
2. Review and comment as requested on statistical and probabilistic approaches or techniques proposed together with other nuclear regulatory research and development programs.
3. Define (a) monographs and guideline documents on existing statistical and probabilistic topics and techniques and (b) areas of statistical and probabilistic research and development, that are needed to further the effective use of statistics in connection with nuclear regulatory procedures and programs.

The committee first met on August 1, 1980. It consisted of 18 members of whom 17 were statisticians, the majority academics. It interfaced to the Office of Nuclear Regulatory Research (NRR).

The NRR was committed to the probabilistic risk assessment (PRA) methodology. This methodology was initiated in the Rasmussen Report (WASH-1400, 1975). An NRC requirement as of 1978-79 was that each new nuclear power plant conduct a PRA prior to operation.

A PRA starts with each basic component in the plant, i.e. pipes, valves, diesel generators, etc. and estimates a failure rate for each such component. Then it attempts to construct all possible sequence of events leading to severe core damage or meltdown. In some way, a probability is assigned to each sequence. Then these probabilities are combined to give an overall probability per year of severe care damage.

A PRA is a serious and extensive undertaking. The original volumes of the WASH-1400 report form a stack almost a foot high. The cost of a PRA for

an individual power plant is several million dollars. The superstructure of a PRA is extremely large, elaborate, and constructed using a variety of tenuous assumptions. After a long climb it finally emerges at the top with "the bottom-line number" i.e. the probability per year of severe core damage. This is the number reported to the press and bandied about in the NRC licensing decisions.

After the original Rasmussen report came under fire, the Lewis Committee was appointed to review it. The Lewis Committee report approved the basic methodology, but had a number of criticisms including lack of adequate peer review of statistical methodology. Against this background the ASA committee was formed with travel and other expenses funded by the NRR.

Here are two examples of what happened at committee meetings.

COMMITTEE REPORT, MAY 15, 1981

2. Modeling component failure and reactor error sequences. In general, members of the Committee felt that significantly more attention should be paid to the use of methodology from stochastic processes in this area. In particular, it was felt that the methodology of renewal processes and non-homogeneous Poisson processes could be useful in producing mathematical models which would describe certain observed phenomena more closely than models currently in use. To be more specific, a more flexible but more complicated model for the binomial failure rate common cause model which could account for the differing lifetimes of different components could be constructed using non-homogeneities in both the Poisson process and the binomial probabilities. For some types of equipment, the use of renewal processes may prove fruitful in modeling such events as non-catastrophic breakdowns.

MINUTES, APRIL 25-27, 1982 MEETING

Suggestions and issues raised in general discussion included:

a. A discussion of the positive and negative aspects of Bayesian methodology when applied in risk estimation,

b. Questions and responses on the meaning and implications of the phrase "uncertainty propogation" and on the communications problems engendered in its use.

c. The suggestion that a more thorough survey of the literature and greater methodological adaptation be attempted in addressing NRC RES's statistical problems. Such a review and comparison of methods and problems should yield areas for further statistical research by individuals in academic or other research institutions.

The excerpts above are, admittedly, not a random selection. But they illustrate what was often happening. In the course of its meetings many committee members were operating in the find-the-nail mode, and trying to find technical pieces of the problem that could be dealt with using known statistical methods.

The fundamental issue which we should have been addressing from the start was;

DOES THE WHOLE IDEA OF PRS's MAKE SENSE?

My conclusion is absolutely not, at least in terms of producing a believable estimate of risk. In mid-1982 I submitted suggested recommendations to the ASA committee for submission to the NRR. These are contained in the appendix to this paper. To quote;

> "The opinion of the ASA Ad Hoc Committee is that 'bottom line' estimates of severe core damage are misleading and inaccurate. The continued focus on them is harmful to the goal of nuclear reactor safety."
>
> "...that overall risk assessments of severe core damage be based on the analysis of past nuclear reactor operating experience."

The NRR faced with budget problems, cut out the funding for the ASA committee in late 1982 and meetings stopped before any careful consideration of these recommendations. However, the NRR had already started a precursor study, which got an estimate of risk by looking at all records of serious power plant incidents, and, applying the PRA methodology to get estimates that the

incident might have led to severe core damage. The advantage is that here one starts far along in the sequence of events. This study resulted in a risk estimate two orders of magnitude higher than that given in the Rasmussen report.

If it had looked at the problem whole instead of finding nails, the ASA committee might have realized that PRA's are an outrageous misuse of statistics, probability, and common sense.

THE EPA CRITERION DOCUMENT ON TSP

TSP stands for total suspended particulates, that is, the particles in air that are small enough to remain suspended. The larger of these are kept from entering the lungs by the body's defense system. The smaller, microscopic particles can get through, lodge in the lung tissue and cause some damage. Because of this, EPA requires TSP monitoring, which is done for 24 hours every 6 days at about 4000 sites in the U.S.

The measurement is done by a Hi-Vol sampler. This instrument sucks a measured volume of air through a filter for 24 hours. The filter is weighed before and after. The weight difference (in micrograms) is divided by the volume of air (in $(\text{meters})^3$) to give the TSP reading.

The current standards are:

annual geometric mean less than 75
2^{nd} highest 24 hour reading in the year less than 260

By the provisions of the Clean Air Act, the EPA periodically reviews all research relevant to the standards and based on its review, decides whether to change the standard.

The procedure is that the EPA puts together a criterion document which summarizes all available relevant information and serves as the basis for

its decision. This document is then submitted for review and written public comments before it becomes final.

Among others, I reviewed the draft criterion document, the written public comments, and a review of both by an EPA contractor. The fundamental issue is:

WHAT EVIDENCE OF HEALTH EFFECT EXIST THAT IS RELEVANT TO CURRENT STANDARDS?

The criterion document discussed a number of epidemiological studies. The most important of these are in England, because of the concern there over high smoke levels. Most of the urban particulate matter in England consists of fine carbon particles generated by the use of coal in house heating and in factories. Because of this, they use the BS method of measurement, which is based on optical reflectivity of the filter instead of the total mass of the deposited particles.

The British have had extremely high smoke episodes in some of their cities. Over 20 years ago, London experienced an episode which was estimated to cause several thousand deaths. The data from such episodes and from persistently high smoke areas give the most clear cut evidence of particulate health effects.

The written public comments consisted of over 200 pages. Well over half were filled with discussions of statistical techniques. The correct way to model using multiple time series, transformation of variables, lags, standard errors, confidence intervals, etc. were subjects written about by a number of very eminent statisticians. It constitutes a very nice example of nail finding and edifice building. Here are some excerpts from my review:

> "The statistician's first question when faced with data must be 'is this data capable of answering the questions I am interested in?' No amount of fancy statistical footwork can make up for unsuitable data. All of the statisticians who are bemoaning the fact that some data sets have not been analyzed using high-powered time series analysis are missing the fundamental point--to wit YOU CAN'T MAKE A SILK PURSE OUT OF A SOW'S EAR. After reading over some of the fundamental epidemiological papers and many descriptions and criticisms of other studies, the clear fact emerges that there is very little or no data suitable for setting TP or TSP standards in reference to health hazards...The big problem is not lack of appropriate statistical technique, but lack of good data. Give me a well thought out and carefully executed experiment resulting in good data and I will tell you the names of at least a dozen statisticians around the country who will do a very credible job in the analysis of the data."

In brief, the problem was that the British studies were virtually useless for two reasons. First, the health effects observed took place at particulate levels that by any standard of comparison, were much higher than our current EPA standards. Second, because there is no site independent method to reliably convert BS measurements to TSP. The few U.S. studies had other flaws.

THE ETHYL CORPORATION MMT APPLICATION

Ethyl Corporation manufactures a lead-based additive to increase the octane rating of gasoline. With all of the new cars using unleaded gas, Ethyl devised a new additive MMT based on manganese (Mn). Use of an additive is a serious affair. Quoting from Ethyl Corporation's Reapplication for MMT Waiver (May 22, 1981), "use of a 1/64 g.Mn./gal in unleaded...indicate a savings of 35,000 B. oil per day and $350 million in processing facilities."

The process that any proposed additive must go through to get a waiver (EPA permission to use the additive) is that it must present proof that use of the additive will not increase automotive exhaust pipe levels of nitrous oxides (NOX), carbon monoxide (CO) and hydrocarbons (HC).

Ethyl Corporation funded a study by the Coordinating Research Council. It used 63 cars of 7 different types, i.e. Ford, GM, Chrysler and two foreign. The 9 cars of each type were divided into 3 groups of 3 each. The first group used unleaded gas with 0 MMT, the second group used gas with 1/32 g./gal. of MMT and the third used 1/16 g./gal. of the additive. The cars were all driven 50,000 miles and the NOX, CO, and HC levels checked at .0, .3, 5, 10, 15, 22.5, 30, 37.5, 45 and 50 thousand miles.

The 1979 report of the Coordinating Research Council states:

> "The results of this study indicate that the use of MMT at either test concentration increases both engine and tailpipe hydrocarbon emissions compared to clear fuel."

The EPA disallowed the waiver. But on May 22, 1981, Ethyl Corporation reapplied for a MMT waiver for 1/64 g./gal. Their application states:

> "Interpolating the available emission data for clear fuel and 1/32 and 1/64 g./gal. Mn to 1/64 g./gal. Mn shows no significant effect of MMT on emissions at this low concentration."

Over half of the reapplication document was devoted to a summary (50 pages) of a statistical analysis which "proved" the above statement. The analysis is a marvelous exercise in edifice building and statistical wizardry.

To begin with, the analysis dropped 3 car types from the analysis (the reasons were interesting), leaving a total of 360 recorded HC values. A total of 105 regression equations were fit to these 360 data points. There were various discussions of significance tests for rejecting outliers, for linearity, of degrees of freedom, of reduced variance, etc.

At the end of this long and complex analysis, they produced prediction equations for HC at any mileage for any given amount of MMT additive. An outcome was that their equations predicted lower HC emissions at 1/64 g./gal. MMT than for clear gas for all 4 automobile types tested.

I was asked to review their analysis and requested the original data. Perhaps the most telling point emerged when I used their equations to predict actual data values. Averaged over the 4 types, at 30,000 miles, here are the results:

HC Emissions

	0 MMT	1/32 MMT	Increase
Ethyl Prediction	.410	.422	3%
Actual Data	.400	.455	14%

The statistical jargon and complexity of the analysis make it hard to penetrate. But this magic edifice had the effect of using the actual data with a 14% increase at the bottom and produced a predicted 3% increase at the top.

There are many other illustrations that could be given, but the three preceding are, I think, enough to get me to my major point:

BECAUSE OF THEIR FAILURE TO TREAT A PROBLEM WHOLE, MANY STATISTICIANS ARE POOR SERVANTS OF PUBLIC POLICY.

In practice, the primary issues are:

1. Problem formulation - What are the right questions?
2. Data
   (a) How to gather data capable of answering the relevant questions,
   (b) Assessing whether the data at hand is capable of answering the questions,
   (c) Understanding the measurement methods producing the data,
   (d) Data quality.

3. Analysis - interpretation

   (a) An analysis <u>appropriate</u> to the data,

   (b) Sensible interpretation of results.

But succeeding in (1) and (2) are three-quarters of the battle. Yet these issues are rarely addressed in formal statistical training.

> I KNOW OF NO FIELD IN WHICH THERE IS SUCH A LARGE DIVERGENCE BETWEEN WHAT IS NEEDED IN PRACTICE AND THE TEACHING AND RESEARCH OF THE UNIVERSITIES.

We do not encourage

> CAREFUL THINKING
> INTELLIGENT FORMULATION
> COMMON SENSE

Instead, statisticians are equipped with a narrow and often inapplicable methodology that produces

> LIMITED VISION
> WIZARD-OF-OZISM
> EDIFICE BUILDING

That the impact of statisticians on public policy has not been larger and statisticians distrusted is due to a good extent, not to our stars, dear statisticians, but to ourselves.

APPENDIX

Suggested Committee Recommendation on PRAs

L. Breiman

Probabilistic Risk Assessment (PRA) has two major functions:

First: It forces an extensive engineering analysis of the system, starting at component level and working its way up. By isolating higher probability paths, it focuses attention on the critical parts of the system and can lead to corrective action.

Second: It produces "bottom line" estimates of the probability of severe core damage. These estimates appear prominently in the summary. They are widely circulated to the public and used by the Commission in their licensing decisions.

The opinion of the ASA Ad Hoc committee is that the "bottom line" estimates of severe core damage are misleading and inaccurate. The continued focus on them is harmful to the goal of nuclear reactor safety. The reasons for this opinion will be expanded below.

Use of PRAs in their first function, as an engineering systems analysis tool, does provide valuable information concerning the failure modes of the system. Therefore, we recommend that

RECOMMENDATION: That PRAs make no attempts to estimate overall probabilities of severe core damage. Instead of numerical assignments, paths to failure should be ranked as High, Medium and Low Probability. That licensing decisions not be based on numerical estimates of core damage, but instead on whether the best current safety standards have been met by the plant.

We note, to begin, that neither the committee nor any of the technical staff of the NRC with whom the committee has been in contact have any belief in the scientific merit or accuracy of the "bottom line" estimates.

The major reason for this disbelief is inherent in the structure of the fault/event tree analysis. At each stage in the tree construction, questionable estimates or questionable methods of combining previous estimates are introduced. Errors are compounded and propagate upward. The final estimates have many sources of error, some of which are difficult. if not impossible, to quantify.

Two particularly weak places in the analysis are:

I. The impossibility of quantifying human error probabilities to within several orders of magnitude. This problem has been seen repeatedly in the various precursor events involving surprising and unanticipated modes of human failure.

II. The similarly difficult problem of assessing probabilities of common mode failures. The estimated probability for the simultaneous occurrence of two events can differ by several orders of magnitude depending on whether the events are assumed independent or have a common cause origin.

In addition, there are numbers of other quite questionable assumptions used in PRAs to arrive at the final estimate.

It is sound and accepted statistical practice to always compute error bounds for any estimate. In view of the methodological obstacles mentioned above, realistic error bounds on estimated probability of core damage would be so wide as to make the estimates useless for decision making. For instance, we do not consider it unlikely that error bounds on a 10^{-6} estimate might be a lower bound of 10^{-2} and an upper bound of 10^{-10}.

Continued use and emphasis on these "bottom line" estimates has some harmful effects.

First: Since many of the NRC's own technical staff and much of the outside scientific community do not place any credibility in these numbers, an atmosphere of cynicism and frustration is created. For the sake of public relations, suspension of sound judgment is required. Not only is this harmful to internal morale, but it also exposes the NRC to justifiable external criticism.

Second: Because of the focus on the overall estimates, otherwise important engineering information may be distorted. We find it hard to believe that a PRA analysis carried out by a consulting firm hired by the utility will produce unacceptably high overall risk estimates. The emphasis is not only distorting in this way, but also it diverts technical time and funding away from the more important engineering systems analysis aspects, both in terms of NRC technical staff and of the direction of research carried out by subcontractors. If the emphasis were removed from the overall risk assessment and PRAs viewed instead as an engineering analysis tool, this might open the way to significant technical improvements; a much more realistic set of goals would be set; and attention and research directed at those goals.

If overall risk assessments are needed, then a much sounder approach is the statistical analysis of the precursor events generated over the history of many hundreds of reactor-years of operating experience. The committee commends the NRC for moving in this direction and recommends

RECOMMENDATION: That overall risk assessments of severe core damage be based on the analysis of past nuclear operating experience. Furthermore,

that the accuracy of past PRAs in locating high probability paths be retrospectively assessed in terms of the history of precursor events.

Adoption of the recommendations will help in re-establishing the credibility of the NRC risk assessment program and place it on a more honest and realistic statistical footing.