Inferences from Alarming Events

John W. Pratt
Richard J. Zeckhauser

Abstract *An extreme event, such as a nuclear accident, an earthquake, a cluster of adverse reactions to a particular drug, or excessive breakdowns of some class of equipment, frequently focuses attention for the first time on an important issue. By then, however, data on the incidence and magnitudes of relevant past events may be unavailable or too costly to reconstruct. Using a simple probability model, we derive methods for drawing statistical inferences based only on the magnitude of the first event noticed and the amount of exposure before this event occurred. We assume that an event is noticed only when its magnitude exceeds some threshold, and we develop methods of inference that are valid even when this threshold is unknown. One tempting but incorrect approach is to treat the magnitude of the observed event as if it were the threshold, forgetting that smaller magnitudes might have been noticed as well. The biases that arise when this mistake is made turn out to be substantial; risks can easily be overstated by a factor of 3.*

THE PROBLEM

Modern industrial societies employ a proliferating array of substances and technologies that appear when introduced to involve low-probability risks of possibly high consequence, but of low probability. It is neither physically possible nor economically sensible to monitor fully the performance of every conceivable hazard in each of the settings in which it is found. The screening strategy that society follows, at least implicitly, is to permit the use of these substances and technologies, requiring a test that "proves safety" in some cases, but to remain alert for surprising or extreme outcomes. Three Mile Island and the heightened incidence of leukemia among vinyl chloride workers are classic cases of such outcomes. In the same way, many social processes, such as the operation of businesses or government agencies, are assumed to be

Journal of Policy Analysis and Management, Vol. 1, No. 3 371–385 (1982)
© 1982 by the Association for Public Policy Analysis and Management
Published by John Wiley & Sons, Inc CCC 0276-8739/82/030371-15$02.50

acceptable as long as no significantly untoward consequences come to light.

Because economic development has tended to increase the physical concentration of the population as well as its reliance on large-scale interdependent technologies, society has become increasingly vulnerable to low-probability natural disasters as well. A major quake along the San Andreas Fault would provide a dramatic illustration. Yet California remains densely populated in many regions near the fault; significant tremors might sound an alarm and lead to changes in the location and construction of residential and industrial buildings.

Society's strategy for dealing with low-probability risk relies, in effect, on a watchdog that is aroused only occasionally. Whatever the merits of this approach—the quintessence of crisis management—once the dog does bark, appropriate inferences need to be drawn to predict future levels of risk. The problem is subtle because the level of stimulus required to arouse the dog, a critical consideration here, may not even be known. Inappropriate inferences are all too easy to make and, as we shall show, may lead to large systematic errors, creating the potential for important biases in policy.

To predict future levels of risk, a statistically attractive possibility would be to go back and collect all data on relevant past experience and then draw appropriate conclusions based on the complete time series. For example, after an assassination attempt that wounds a leader, his security staff may assess how frequently they have received threats or warnings, intercepted bombs or weapons, or had other indications of danger. Often, unfortunately, such reconstruction is not possible. In some instances, no data at all have been collected. Thus, an auto manufacturer might not know how often a particular steering mechanism in its cars was involved in accidents. Consistent reconstruction of past data is especially likely to be infeasible (or excessively expensive) in areas not covered by the normal recording procedures of our society. This may be the case with activities outside the law, such as illicit drugs or illegal immigrants, or with currently unidentified risks such as those associated with industrial wastes or with food substances that are generally regarded as safe (GRAS).

In other instances, the past data may have been collected in incompatible forms, perhaps from a diversity of sources. In the case of the Three Mile Island nuclear accident,[1] the different accounts of past histories diverge dramatically. One group of commentators tells of 500 reactor years without a significant accident. Another recounts several serious situations and near misses. Although it may be possible to reconstruct our experiences with nuclear power on a consistent and agreed-upon basis, a significant commitment of resources and time will be required, and not a little debate. An interim evaluation may be useful.

To keep the analysis tractable, we shall invoke a strong simplifying assumption: The only known facts are the magnitude of the extreme event that activated the alarm and the number of trials or total amount of exposure before it occurred. Many fields offer examples that could fit our models with no more than the level of elaboration traditionally required to pass from simple abstractions to real world phenomena. An adverse symptom may be noticed in a cluster of individuals taking a particular drug; corruption may be discovered within a government agency; some class of equipment may be found to break down excessively; three tenured professors may quit a department in a single year. The inference problem we consider is particularly germane to the control of potentially toxic substances. The monitoring and regulation of thousands of such substances is the purpose of the widely debated but indisputably consequential Toxic Substances Control Act of 1976.[2]

For purposes of illustration, we shall use a simple, noncontroversial example based on an actual legal case in which one of the authors was asked but chose not to participate. It involves a rivet gun. A rivet has ricocheted and blinded an individual in one eye. We know the number of rivets that have been sold for this brand of gun, most of which have presumably been fired. To settle a product liability suit, or to decide what regulatory action should be taken toward the gun, it must be determined how dangerous it is. Surely any previous injury as severe as a blinding would have been reported. If less severe injuries had occurred, however, some of them would presumably also have been reported. What inferences should be drawn from the fact that no previous injuries have been reported? As another investigator demonstrated, it is a subtle matter to draw inferences when a watchdog doesn't bark.

> "Is there any other point to which you would wish to draw my attention?"
> "To the curious incident of the dog in the night-time."
> "The dog did nothing in the night-time."
> "That was the curious incident," remarked Sherlock Holmes.[3]

With the rivet gun, the question for attention is: How severe an injury would have led to a report? Our tale concerns a not-so-elementary follow-up question: How can we draw inferences properly from the magnitude of and time until an alarming event, when the threshold for notice or report is not known?

We have two analytic objectives in this article. First, we wish to examine the bias inherent in some intuitively appealing but incorrect approaches to this problem; this is a subject of importance to anyone concerned with the formulation of public policy towards risk. To determine the importance of the bias, we shall examine it numerically as well as algebraically. Second, we seek to illustrate the use of statistical methodologies appropriate for situations that only become a focus of attention once some extreme event occurs; this is a subject that will be germane for only those readers who might actually conduct an analysis. For ease of exposition, we emphasize classical statistical methods, and hypothesis testing in particular, though our personal philosophy of inference is, in principle, Bayesian.[4]

Our broader purpose is to heighten consciousness of the importance of drawing inferences logically and clearly when, by changing the perception of a process, extreme outcomes motivate discussion or policy action. We examine only those biases that arise when formal statistical methods are brought into play and are used inappropriately. Thus we present at best a metaphor for the processes of informal reasoning that are employed most frequently in real world decisions, particularly crisis decisions. The decision processes actually used are unlikely to match the logic and clarity of formal statistical analysis. Consequently, our estimates of biases are likely to prove conservative as compared with real world decisions, and our warnings abcut pitfalls are likely to be understated.

THE UNDERLYING PROBABILITY MODEL

To provide a specific analysis of situations like those described above, picture the following process. At random intervals of time, events occur (for example, a rivet hits someone). Time is measured in units appropriate to the situation (for example, rivets fired or reactor years). Each event has a magnitude, also random, that indicates its seriousness. How seriousness is measured and how the measure is distributed turn out to matter little for our main results.

We assume that an event will be noticed if and only if its magnitude exceeds an unchanging threshold a.[5] (A more general formulation would make the probability of noticing an event increase as a function of its magnitude over some range, rather than jump abruptly from 0 to 1 at a single point a. This would allow, for example, some classes of injuries that would be reported sometimes but not always.) For the rivet gun, the threshold might be an injury that requires hospitalization. A central aspect of the problem is that the value of a is not necessarily known. Let the first event noticed have magnitude b and occur at time T. In this notation, our concern is drawing inferences from the limited observed information T and b, and whatever information (possibly subjective) exists about the value of the threshold a.

The probability model we employ is natural and standard. Its parameters are an unknown constant λ, denoting the frequency or rate at which events occur on average, and a function $G(y)$, denoting the probability that an event, whenever it occurs, will have magnitude y or greater. The average frequency with which events of magnitude y or greater occur therefore is $\lambda G(y)$. Technically, the model is equivalent to an assumption that the number of events of magnitude at least y occurring before time t has the Poisson probability distribution with mean $t\lambda G(y)$.[6] The model treats time as continuous. For discrete time, the effects we are concerned with are stronger, but only slightly, and the analysis is more complex.[7]

For the rivet gun, with the usual approximation of discrete trials by continuous time, each shot has probability λ of hitting someone and probability $\lambda G(y)$ of inflicting an injury of severity of at least y;

the number of such injuries inflicted by the first t shots has a Poisson distribution with mean $t\lambda G(y)$. If one shot in 2500 hits someone, then $\lambda = 0.0004$. If y denotes blinding, and such a severe injury occurs once in 1000 hits, then $G(y) = 0.001$. In 1,000,000 shots, the number of injuries at least as severe as blinding will then have a Poisson distribution with mean $t\lambda G(y) = 1{,}000{,}000 \times 0.0004 \times 0.001 = 0.4$. According to this distribution, the probability of no such injury in 1,000,000 firings is 0.67, the probability of 1 is 0.27, the probability of 2 is 0.05, and the probability of 3 or more is 0.01.

One implication of our model willl be especially useful. Clearly T, the time until the occurrence of the first event exceeding the threshold a, exceeds a given time t if and only if no such event occurs up to time t. The probability of this is given by the Poisson distribution as

$$\text{Prob}(T > t) = e^{-t\lambda G(a)} \tag{1}$$

The expression on the right is the exponential function evaluated at $-t\lambda G(a)$. (The distribution of T is accordingly called "exponential.")

Once an event has been noticed, how should inferences be drawn? In addressing this question, there is a tendency to forget that smaller magnitudes might have been noticed as well. This natural mistake in effect treats the magnitude b of the first event noticed as if it were the threshold a. In this article we (1) identify correct methods of inference when a is known, (2) provide a quantitative appraisal of the bias arising when b is mistakenly treated as the threshold a, and (3) define correct methods to be used when a is known. (This explains the mongrelized subtitle we used in the discussion paper version of this article, "False Tails and True When Finally the Watchdog Barks.")

We carry out this analysis for classical hypothesis testing and apply the results to the rivet gun. Then with a bow to our upbringing, we present a straightforward "Bayesian" probability analysis. Using the wrong threshold produces similar biases in other forms of inference.[7] Throughout, technicalities are relegated to the Notes section where possible.

HYPOTHESIS TESTING

Society employs two basic approaches to the control of low-probability risks. In the framework of classical hypothesis testing,[8,9] these approaches give rise to two different types of null hypothesis. A null hypothesis representing danger would be in the spirit of the current U.S. regulatory procedure for the introduction of new drugs: A drug is considered unacceptable for use until it is proven both safe and efficacious.[10] Our approach to most consumer products, on the other hand, is to assume they are safe until contrary evidence appears.[11] A null hypothesis of safety would formalize this assumption. When extreme events excite notice, the approach assuming safety is ordinarily the one that has been followed. Hence we shall concentrate on it, but each of our tables includes an additional row for the case where the null hypothesis represents danger.

Null Hypothesis Representing Safety

We consider a null hypothesis chosen to represent conditions of low risk or safe operation. Rejecting it should sound an alarm; accepting it should provide reassurance. Specifically, adopting the model described earlier, let an event frequency λ_0 and magnitude distribution G_0 be chosen to represent safe conditions. The statistical null hypothesis is that extreme events are no more likely than under λ_0 and G_0. Let the test statistic be the amount of time T until an extreme event is first noticed. By eq. (1), the distribution of time until notice depends only on the product $\lambda G(a)$, where a is the threshold for notice. We may therefore take as the null hypothesis

$$H_0\colon \lambda G(a) \leqslant \lambda_0 G_0(a) \tag{2}$$

This says specifically that the frequency with which events sufficiently extreme for notice occur on average is at most $\lambda_0 G_0(a)$, a chosen safety level.

Correct Inference When the Threshold is Known

Assume now that the threshold for notice, a, is known. The null hypothesis of safety is rejected if a noticeably extreme event occurs too soon, that is, if T is too small. Under the null hypothesis (2), by eq. (1), the probability that an event as extreme as a would have occurred by time T was at most

$$P_a(T) = 1 - e^{-T\lambda_0 G_0(a)} \tag{3}$$

A traditional classical statistical test at significance level α is to reject H_0 if and only if this probability is less than or equal to the significance level, i.e., if $P_a(T) \leqslant \alpha$. The value of $P_a(T)$ is sometimes called the P value. P values are often interpreted as measuring the strength of the evidence against the null hypothesis.

For any threshold a and significance level α, there will be a time such that if an event as extreme as a occurs before then, the null hypothesis of safety is rejected. If it takes longer, then we will be reassured. Accordingly, we shall call this time the reassurance time. (It is also called the "critical value" of the test statistic T.) It is obtained by setting eq. (3) equal to α and solving for T. By a bit of algebraic manipulation, the reassurance time, denoted $T_a(\alpha)$, can be expressed in terms of natural logarithms as

$$T_a(\alpha) = \frac{-\ln(1-\alpha)}{\lambda_0 \mathrm{G}_0(a)} \tag{4}$$

In summary, when the threshold a is known, a classical statistical test of the null hypothesis of safety can be based on the time T to first notice. A safety level $\lambda_0 G_0(a)$ must be chosen for the frequency of noticeable events. Either formula (3) for the P value or formula (4) for the reassurance time can be evaluated. The null hypothesis is rejected if the P value is less than some chosen significance level α or, what amounts to the same thing, if notice occurs before the reassurance time.

Mistaking the Observed Magnitude for the Threshold

When the threshold a is unknown or ignored, what is the effect of mistakenly treating the first magnitude noticed, b, as if it were the threshold? That is, if the P value (3) or the reassurance time (4) is erroneously calculated with b in place of a, how serious is the error?

To a strict hypothesis tester, the relationship between the true and asserted significance levels of the erroneous test probably provides the most definitive measure of error. Remember that G_0 stands for the distribution of the magnitude of the events chosen to represent borderline safety. The true significance level corresponding to a given asserted significance level depends, perhaps surprisingly, on neither the true threshold nor G_0. As long as G_0 is continuous, which we assume henceforth, Table 1 reveals in pure form the considerable overreaction implied by treating the observed magnitude as if it were the threshold. For the null hypothesis of safety, an erroneous test supposedly at the widely used 5-percent level, for example, has a true significance level of 17.5 percent. That is, even if the procedure is safe by our own standards, we face a 17.5-percent chance that we will conclude that it is not. True and asserted P values satisfy the same relationship as significance levels. Thus, Table 1, though limited in coverage, applies also to P values: an asserted P value of 2.5 percent, for instance, is truly 10.4 percent.

The bottom row of Table 1 gives the parallel results if we start with a null hypothesis of danger. The overstatement of danger persists. For example, an erroneous test thought to be conducted at

Table 1. True significance level of erroneous tests.

Supposed or nominal significance level α	0.005	0.010	0.025	0.050	0.100
True significance level when null hypothesis represents safety	0.029	0.051	0.104	0.175	0.287
True significance level when null hypothesis represents danger	0.00071	0.0016	0.0046	0.011	0.025

the 1-percent level has actually at most a 0.16-percent chance of providing false reassurance by rejecting this null hypothesis if true.

There are also significant differences between erroneous and correct reassurance times. Table 2 gives the factor by which the erroneous ıeassurance time $T_b(\alpha)$ exceeds the correct reassurance time of a test based on the time of first notice, assuming that the true threshold is unknown. At the 1-percent level, for example, the erroneous reassurance time is 7 times too large if the null hypothesis represents safety, 1.5 times too large if it represents danger.

It is instructive as well to compare the erroneous inferences with those that would be correct for a known threshold a. Comparison is easiest for the reassurance times: by eq. (4), the erroneous reassurance time is too large by the factor

$$\frac{T_b(\alpha)}{T_a(\alpha)} = \frac{1}{R} \tag{5}$$

where

$$R = \frac{G_0(b)}{G_0(a)} \tag{6}$$

The central question then becomes how big the error factor R is likely to be, i.e., how it is distributed. The answer is simple in the most important case: R is distributed uniformly between 0 and 1 when the magnitudes have the borderline null distribution G_0.[12] Moreover, R is distributed independently of the time until first notice.[13] This implies that the probability is ½ that the reassurance time will be overstated by a factor of 2 or more; overstatement by a factor of 3 or more has probability ⅓, etc.

Note, furthermore, that because of the overstatement, the expected value of the erroneous reassurance time is infinite. In contrast, the actual time of operation without a noticed event has

Table 2. Ratio of erroneous to correct reassurance times.

Significance level α	0.005	0.010	0.025	0.050	0.100
Ratio when null hypothesis represents safety	7.79	7.01	5.96	5.17	4.39
Ratio when null hypothesis represents danger	1.46	1.51	1.59	1.68	1.81

finite mean, namely, $1/\lambda G(a)$. This infinite versus finite discrepancy may sound rather puzzling. There are times when mathematics is like that.

Finally, we compare the erroneous P values and significance levels with the correct ones for a known threshold a. The erroneous P value $P_b(T)$ may be expressed in terms of the correct P value $P_a(T)$ as

$$P_b(T) = 1 - \left[1 - P_a(T)\right]^R \tag{7}$$

where R has the same meaning as before. The relationship is the same for significance levels. Two implications of this formula may provide some insight. First, if the correct P value (or significance level) is small, then the ratio of the incorrect to the correct P value is approximately R. Second, the expectation of this ratio in the borderline case is between 0.5 and 0.53 if the correct P value is 0.3 or less, and is always at least 0.5.[14]

Correct Inference When the Threshold is Unknown

Correct P values for an unknown threshold are given by Table 1 and formulas in Note 13. Comparing the correct P value with α then gives a correct test. This amounts to viewing in reverse the relationship discussed above between the supposed and true significance levels of the erroneous test procedure that treats the first noticed magnitude b as if it were the threshold. Specifically, define H_0 as requiring that eq. (2) hold for all a and some given, continuous G_0. Then any desired true significance level can be obtained by employing the erroneous test at the corresponding nominal significance level in Table 1 (or a more complete table of the same sort). Equivalently, for the true significance levels commonly used, Table 3 provides the critical values of the test statistic $T\lambda_0 G_0(b)$: The null hypothesis is rejected if the time T of first notice is less than this critical value divided by $\lambda_0 G_0(b)$.[15] Alternatively, dividing the entry in Table 3 by T gives the strictest safety level

Table 3. Critical values of $T\lambda_0 G_0(b)$ when threshold is unknown.

Significance level α	0.005	0.010	0.025	0.050	0.100
Critical value when null hypothesis represents safety	0.000643	0.00143	0.00425	0.00994	0.0240
Critical value when null hypothesis represents danger	3.62	3.05	2.32	1.78	1.27

$\lambda_0 G_0(b)$ we could adopt and still accept the null hypothesis; this is a confidence limit. (For the null hypothesis of danger, "accept" and "reject" are reversed.)

Example: The Rivet Gun In the case of the rivet gun, suppose that the blinding occurred after 220,000 firings. Suppose that the null hypothesis is that the gun is "safe," and is defined so that the probability on each firing of an injury at least as severe as a blinding in one eye is no greater than 10^{-7}, with milder or severer injuries correspondingly more or less likely. (Choosing a null hypothesis after the data are known risks biasing the choice to make a case. Though perhaps falling prey to this bias, we are not investigating it here.) Then borderline safety corresponds to $\lambda_0 G_0(b) = 10^{-7}$.

If the true threshold a were known, a test of a null hypothesis (2) stated in terms of a could be based on the P value (3) or the reassurance time (4). Unfortunately a is unknown.

If blinding (which is b, the first observed magnitude) is treated as the threshold, then for $\alpha = 0.05$ eq. (4) gives the erroneous critical value or reassurance time

$$\frac{-\ln(1 - \alpha)}{\lambda_0 G_0(b)} = 10^7 \ln(0.95) = 512,933 \tag{8}$$

This is too large by the factor

$$\frac{1}{R} = \frac{\text{fraction of injuries that would have been reported}}{\text{fraction of injuries at least as severe as blinding}} \tag{9}$$

The erroneous P value (3) is

$$1 - e^{-220,000 \times 10^{-7}} = 0.022 \tag{10}$$

which is too small by approximately the same factor. The problem is, of course, that the factor R is unknown. Suppose that severity of injury has a continuous distribution for injuries that would have been reported. (This may be approximately though not exactly true.) Then by Table 1 the true significance level of the erroneous test, the chance of false alarm, is 17.5 percent: If the null hypothesis is just satisfied, there is more than 1 chance in 6 that it will be rejected, in contrast to the 1 chance in 20 we were aiming for. Table 1 also shows that the true P value is a little less than 0.1, rather than 0.022.

The correct reassurance time at level $\alpha = 0.05$ is $0.00994/10^{-7}$ = 99,400 firings, the relevant entry in Table 3 divided by the borderline safety level for blinding. Thus 220,000 firings before a reported injury, if the injury is blinding, is more than twice the correct reassurance time, though less than half the incorrect one. The incorrect reassurance time is 5.17 times the correct one, by Table 2. The null hypothesis of safety would be rejected only if it required the probability of blinding on each firing to be below $0.00994/220,000 = 0.452 \times 10^{-7}$, or about 1 in 22 million.

BAYESIAN METHODS

Hypothesis testing is a very limited form of inference. To go beyond it on the basis of a single observed event and a history of nonevents, however, requires strengthening our assumptions. In many real world situations, the next step would be to obtain further data, but our interest here is in developing more powerful inferences without further data. The simplest plausible assumption for this purpose appears to be that our earlier model will hold with a known distribution of magnitude. Accordingly, we assume that G is known.

In a more technical background paper,[7] besides hypothesis testing we treat confidence limits and several likelihood methods with known and unknown thresholds. Here we consider only a type of probability analysis called Bayesian[4] when both the frequency of occurrence, λ, and the true threshold for notice, a, are unknown.

The object of the Bayesian approach is to obtain a probability distribution for the quantity that is of interest, given the data at hand. This is called the posterior distribution. To obtain it, it is necessary to start with a probability distribution representing the opinion one would have held in the absence of data. This is called the prior distribution. Here we need a prior probability density for the frequence λ and the true threshold a. For any given λ and a, the probability density of the data, the observed magnitude b and time T, is[16]

$$p(T,b\,|\,\lambda,a) = \lambda g(b) e^{-T\lambda G(a)} \quad \text{for } b \geqslant a, \text{ and otherwise } 0 \qquad (11)$$

where g is the probability density function of the magnitude distribution. According to an elementary law of probability theory known as "Bayes' rule," the posterior density of λ and a is simply the prior density times the quantity in eq. (11) times a constant. Integrating over a then gives the posterior of λ, the quantity of interest. Erroneously treating b as the threshold replaces the distribution of a by the single value b. Since $a > b$ is impossible once b has been observed, this error increases the relative weight on large values of λ in expression (11) and in the posterior density of λ, in much the spirit of the exaggerations discussed previously.[17]

Now we apply the approach to the case of the rivet gun. On the 220,000th firing it was noticed: it blinded someone. We now wish to draw inferences about the probability of different values of a and λ. In particular, we wish to compute the probability density for the hit frequency λ given the data, the posterior density of λ. To do so, we must specify a prior distribution for the two unknowns. Assume that the prior distribution has λ and $G(a)$ independently uniformly distributed on intervals $[0, L]$ and $[C',1]$, respectively, where L is essentially infinite and $C' < G(b)$. The posterior density of λ for this illustrative prior distribution is $(220/0.999)\,(e^{-220\lambda} - e^{-220{,}000\lambda})$.[18] The erroneous posterior density of λ, obtained by treating blinding b as the threshold, with the same prior distribution of λ, is $(220)^2\lambda e^{-220\lambda}$. Graphs of the correct and erroneous posterior densities λ appear in Figure 1. The overweighting of large values of λ caused by the erroneous assumption that no accident less than blinding would have been noticed is eminently clear.

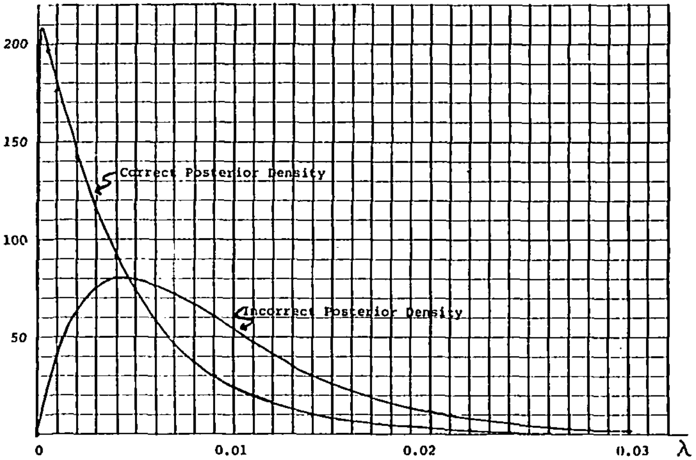

Figure 1. Correct and incorrect posterior densities for frequency of hit in the rivet gun example.

CONCLUSION

When drawing inferences from the occurrence of a single event whose magnitude exceeds an unknown threshold, the natural mistake of treating the size of the event itself as if it were the threshold for notice introduces a substantial bias toward the exaggeration of risk. (Overstatement by a factor of at least 3 was typical in our examples.) Our formal investigation was limited to models in which observations are temporally stationary and independent, thresholds for notice are sharp, and the only knowledge available from before the alert is how much time had already passed. We believed that the significant magnitude of the bias, as well as its qualitative nature, would persist not only through other forms of inference,[7] but also through a variety of alternative formulations, which would be a worthy subject for future research.

At least for the simple models employed here as paradigms, tractable methodologies are available for correctly drawing inferences after the occurrence of an extreme event. The existence of appropriate methods for assessing low-probability risks is, however, no guarantee that they will be employed once such a risk eventuates. Extreme events tend to generate extreme passions and interpretations. Indeed, when such events occur, however plentiful the data, statistical analyses are likely to receive low priority. Even correctly drawn statistical inferences are likely to be abused or ignored. Society needs mechanisms that make it more likely that correct statistical methods will be employed—and used appropriately—when policy decisions are undertaken in response to alarming events.

The authors thank Albert Carnesale, Nancy Jackson, Robert Klitgaard, William Krasker, Howard Raiffa, and referees for helpful comments, and Robert Schlaifer for computation. Pratt's research was supported by the Associates of the Harvard Business School and the S.S. Huebner Foundation. Zeckhauser's research was supported by NSF Grant SOC(77-16602), the Harvard Faculty Project on Regulation, and the S.S. Huebner Foundation.

JOHN W. PRATT is Professor of Business Administration at the Harvard Business School.

RICHARD ZECKHAUSER is Professor of Political Economy, Harvard University.

NOTES

1. *Report of the President's Commission on the Accident at Three Mile Island,* John G. Kemeny, Chairman (Washington, DC: U.S. GPO, 1979).
2. *Toxic Substances Control Act,* Public Law 94-969, 1976.
3. Doyle, Sir Arthur Conan, "Silver Blaze," *The Annotated Sherlock Holmes, Vol. II,* William S. Baring-Gould, Ed. (New York: Clarkson N. Potter, Inc., 1967).
4. Savage, Leonard J., *The Foundations of Statistics* (New York: Wiley, 1954).
5. We avoid complication by assuming that the threshold is stationary over time. This assumption does not always hold. In the discussion of regulatory reform in recent years, considerable concern has been expressed about the disparity between our increasingly sensitive tests for nonsafety and regulatory standards that were formulated for a world of much cruder science. Consider, for example, the much-debated Delaney Clause [*Federal Food, Drug and Cosmetic Act,* Section 409(c) (3) (A), 1958], which prohibits the use as a human food additive of any substance known to be carcinogenic in any concentration in animal or man. Given the advances in science in the past 20-odd years, it employs what turns out to be an increasingly hair-trigger alarm system. Moreover, on a short-term basis, thresholds may change dramatically, most likely downward, when an alarm has been sounded. Once a corrupt official at City Hall has made the news, reporters will be scurrying for other stories on the same beat. After Three Mile Island, at least for a while, the media and the public became much more sensitive to minor malfunctions at nuclear plants.
6. In the terminology of probability theory, the occurrence times follow a Poisson process with parameter λ, and given the occurrence times, the magnitudes are independently, identically distributed with cumulative right-tail probability function G.
7. Pratt, John W., and Zeckhauser, Richard J., "Retrospective inferences from an extreme event: False tails and true when finally the watchdog barks," Harvard University Graduate School of Business Administration, Working Paper 78-48.
8. Fisher, Sir Ronald A., *Statistical Methods and Scientific Inference,* 1956, 3rd ed. (New York: Hafner, 1973).
9. Neyman, J., *First Course in Probability and Statistics* (New York: Henry Holt, 1950).
10. *Kefauver–Harris Drug Amendments to the Food, Drug, and Cosmetics Act,* Public Law 87-781, 1962.

11. *Consumer Product Safety Act*, Public Law 92-573, 1972.

12. This follows directly from eq. (6). R is the conditional probability that a magnitude reaching the threshold will be as large as that observed if G_0 governs the probability distribution of the magnitudes, i.e., if the null hypothesis of safety is just barely satisfied. Technically, R is the right-tail "probability integral transformation" of b, conditional on $b \geqslant a$, when G_0 applies.

13. We can now derive the formula used to calculate Table 1. By eq. (3), $1 - P_b(T) = [1 - P_a(T)]^R$. The erroneous test rejects for $P_b(T) \leqslant \alpha$ and hence for $P_a(T) \leqslant 1 - (1 - \alpha)^{1/R}$. The P value $P_a(T)$ is distributed uniformly between 0 and 1 in the "least favorable" case λ_0, G_0. The true significance level of the erroneous test is therefore $1 - (1 - \alpha)^{1/R}$, conditionally on b, when a is known. (This is of some interest in itself.) The unconditional significance level is the expectation of this conditional level in the least favorable case, namely $1 - \int_0^1 (1 - \alpha)^{1/r}\, dr$. Table 1 was calculated from this formula. By simple manipulation, the unconditional significance level can be expressed in terms of standard functions as $1 - E_2[-\ln(1 - \alpha)]$, where E_2 is the exponential integral of order 2, defined by $E_2(x) = \int_1^\infty t^{-2} e^{-xt} dt = \int_0^1 e^{-x/r}\, dr$. Substituting $P_b(T)$ for α gives the unconditional P value, which is therefore $1 - E_2[T\lambda_0 G_0(b)]$.

14. Expression (7) follows directly from the second sentence of Note 13. The expectation of the ratio for $P_a(T) = P$ is $\int_0^1 [1 - (1 - P)^r] dr/P = (1/P) + 1/\ln(1 - P)$.

15. The critical value of $T\lambda_0 G_0(b)$ appearing in the second line of Table 3 is $E_2^{-1}(1 - \alpha)$, obtained by setting the unconditional P value given in the last sentence of Note 13 equal to α. It follows from eq. (4) that the ratio $T_b(\alpha)$ to the critical value of T is $-\ln(1 - \alpha)/E_2^{-1}(1 - \alpha)$. This appears in the second line of Table 2.

16. As a function of λ and a, this quantity, or any constant multiple of it, is called the likelihood. The famous method of maximum likelihood, which chooses estimates to maximize the likelihood and is often excellent, appears to do badly in this model. The maximum likelihood estimate of the threshold a is the observed magnitude b. This is a peculiarly extreme estimate, since it is the largest possible value of the threshold given the data. Furthermore, the maximum likelihood estimate of the frequency λ is $\lambda = 1/TG(b)$. If the threshold a were known, an appropriate (and maximum likelihood) estimate would have $G(a)$ in place of $G(b)$. Therefore λ is "too large" by the factor $G(a)/G(b) = 1/R$ discussed after eq. (6). This factor, the estimate λ, and the bias in λ all have infinite mean. Thus, the extreme estimate of the threshold chosen by maximum likelihood has unpleasant repercussions for the frequency as well.

17. Thus, if the prior density of λ is $p(\lambda)$ and that of a given λ is $q(a|\lambda)$, then the posterior density of λ is proportional to $(\propto)$ $p(\lambda)L(\lambda|T, b)$ where

$$L(\lambda|T, b) \propto \int_0^b \lambda e^{-T\lambda G(y)} q(y|\lambda) dy \tag{12}$$

$L(\lambda|T, b)$ is sometimes called the "marginal likelihood" of λ. In particular, if a is *a priori* independent of λ with a distribution such that $G(a)$ is distributed uniformly on an interval $[C', C]$, then $C > G(b)$ and

$$L(\lambda|T, b) \propto \int_B^C \lambda e^{-T\lambda x} dx \propto e^{-TB\lambda} - e^{-TC\lambda} \tag{13}$$

where $B = \max\{C', G(b)\}$. The integration is also easy to carry out in closed form if $x = G(a)$ has density $(c + dx)e^{bx}$ on some interval, or if its density is a linear combination of such terms.

18. Note 17 applies. Hence eq. (13) holds with $B = G(b) = 0.001$ and $C = 1$ (and $T = 220{,}000$), giving the marginal likelihood $e^{-220\lambda} - e^{-220{,}000\lambda}$. Since λ is distributed uniformly *a priori*, its posterior density is simply the marginal likelihood times a normalizing constant, as given. Similarly, treating b as the threshold gives a posterior density of λ proportional to the erroneous likelihood $\lambda e^{-220\lambda}$.