

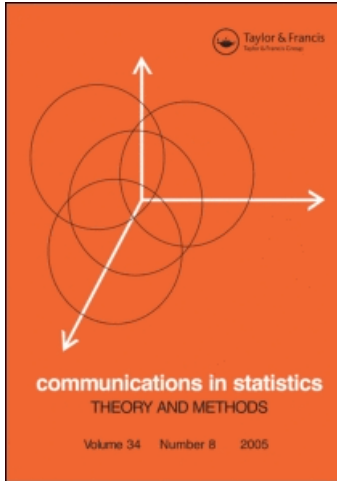
This article was downloaded by:

On: 7 January 2010

Access details: *Access Details: Free Access*

Publisher *Taylor & Francis*

Informa Ltd Registered in England and Wales Registered Number: 1072954 Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Communications in Statistics - Theory and Methods

Publication details, including instructions for authors and subscription information:

<http://www.informaworld.com/smpp/title~content=t713597238>

### Balancing type I and type II error probabilities: further comments on proof of safety vs proof of hazard

Burt Holland <sup>a</sup>; Nasser K. Ordoukhani <sup>b</sup>

<sup>a</sup> Department of Statistics, Temple University, Philadelphia, Pennsylvania <sup>b</sup> Department of Mathematics, East Carolina University, Greenville, North Carolina

**To cite this Article** Holland, Burt and Ordoukhani, Nasser K.(1990) 'Balancing type I and type II error probabilities: further comments on proof of safety vs proof of hazard', *Communications in Statistics - Theory and Methods*, 19: 10, 3557 – 3570

**To link to this Article: DOI:** 10.1080/03610929008830397

**URL:** <http://dx.doi.org/10.1080/03610929008830397>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article may be used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

**BALANCING TYPE I AND TYPE II ERROR  
PROBABILITIES: FURTHER COMMENTS ON PROOF OF  
SAFETY VS PROOF OF HAZARD**

**Burt Holland**  
Department of Statistics  
Temple University  
Philadelphia, Pennsylvania 19122

**Nasser K. Ordoukhani**  
Department of Mathematics  
East Carolina University  
Greenville, North Carolina 27858

**KEY WORDS AND PHRASES:** *environmental testing, joint error probability, power of test, sample size determination*

**ABSTRACT**

This paper elaborates on earlier contributions of Bross (1985) and Millard (1987) who point out that when conducting conventional hypothesis tests in order to "prove" environmental hazard or environmental safety, unrealistically large sample sizes are required to achieve acceptable power with customarily-used values of Type I error probability. These authors also note that "proof of safety" typically requires much larger sample sizes than "proof of hazard". When the sample has yet to be selected and it is feared that the sample size will be insufficient to conduct a reasonable

“proof of safety”, we recommend that consideration be given to equating Type I and Type II error probabilities at somewhat higher than the habitually-used levels in the “proof of hazard” test in order to reduce sample size requirements to attainable levels. To assist in the implementation of this approach, several charts that display the possible tradeoffs among the various decision quantities are given. The ideas and suggestions presented here apply as well in many areas of application aside from environmental health.

## 1. INTRODUCTION

Bross (1985) and Millard (1987a) indicate that when monitoring for environmental safety, hypothesis tests undertaken to “prove” either safety or hazard require unrealistically large sample sizes to achieve acceptable power of tests at customarily-used levels of  $\alpha$ . They also point out that “proof of safety” requires far larger sample sizes than does “proof of hazard”. Unfortunately, in the context of environmental testing, small sample sizes are the rule rather than the exception, and often insufficient attention is given to the power of tests (Bross (1985), Millard (1987b)).

In this paper, we provide a brief discussion of the reason for the asymmetry between “proof of safety” and “proof of hazard”. Then we consider the situation where the sample has not yet been selected and there is concern whether it will be possible to obtain a sample of adequate size to conduct a defensible “proof of safety”. Here we propose that serious consideration routinely be given to the option to balance the competing interests involved with moderate-sized samples by equating Type I and

Type II error probabilities at values greater than the habitual 0.05 level. Formulas which display the relationships among the decision quantities involved are supplied, and charts are provided in order to display explicitly the available tradeoffs.

Definitions and notation will be presented in Section 2. Some discussion of Type I and Type II errors in terms of safety and hazard in the present setting is presented in Section 3. Details of our suggested approach for determining sample size follow in Section 4. As in the earlier papers, our presentation will be couched in terms of testing for environmental safety, but we note in Section 5 that similar problems and remedies occur in many other important areas of application.

Throughout this paper, we put quotes around the word "proof" to emphasize that the results of a hypothesis test constitute *evidence* in favor of one hypothesis or its competitor, as opposed to proof in any formal sense.

## 2. NOTATION AND DEFINITIONS

We retain the notation and problem setting used in Bross (1985) and Millard (1987a) where for a stable population at risk, interest lies in comparing the true unknown probability  $p_1$  of death at a site before some "event" with the probability  $p_2$  of death at the site following the event, assuming equal time intervals before and after. If  $p_2$  exceeds  $p_1$ , the site is said to be hazardous; otherwise, the site is referred to as safe. However, inference proceeds in terms of the parameter  $\theta = (p_2 - p_1)/p_1$ ,

the relative increase in deaths following the event. As usual,  $\alpha$  will denote the maximum probability of committing a Type I error, while  $\beta$ , which depends on  $\theta$ , is the probability of committing a Type II error.

Bross (1985) takes as his definition of "site is safe" the condition  $\theta \leq A$  and conversely defines "site is hazardous" to be the contrary statement  $\theta > A$ . Here  $A$  is selected to be some small fractional allowance; Bross uses  $A = 0.1$  and we will also examine below the choice  $A = 0$ .

Let  $x$  be the number of deaths before the event,  $y$  the number of deaths after the event, and  $z = x + y$ . Bross's test statistic is

$$\zeta = 2\sqrt{z} \left[ \frac{x}{z} - \frac{1}{A+2} \right],$$

which he shows to be approximately standard normal when  $\theta = A$ .

### 3. TYPE I AND TYPE II ERRORS

There are two ways to undertake to decide between "site is safe" and "site is hazardous". The first procedure, labeled "proof of hazard," constructs the null hypothesis  $H_0 : \theta \leq A$  (i.e., the site is safe) *vs* the alternative hypothesis  $H_a : \theta > A$  (i.e., the site is hazardous). Then one rejects "site is safe" in favor of "site is hazardous" at level of significance  $\alpha$  if  $\zeta \leq -z_\alpha$ , where  $z_\alpha$  is the 100(1 -  $\alpha$ )th percentage point of the standard normal distribution. With this test, a Type I error is to state that the site is hazardous when in fact it is safe and a Type II error is to state that the site is safe when in fact it is hazardous.

The second procedure, labeled "proof of safety," sets up the null hypothesis  $H_0 : \theta > A$  (i.e., the site is hazardous) vs the alternative hypothesis  $H_a : \theta \leq A$  (i.e., the site is safe). Here one rejects "site is hazardous" in favor of "site is safe" at level  $\alpha$  if  $\zeta \geq z_\alpha$ . Now, a Type I error is to declare that the site is safe when it actually is hazardous, while a Type II error is to declare that the site is hazardous when in fact it is safe.

The inherent imbalance between "proof of safety" and "proof of hazard" arises from the fact that conventional hypothesis testing procedures are chosen to minimize Type II error probability subject to controlling Type I error probability (under a null hypothesis) at an arbitrarily designated tolerance  $\alpha$ , and as a result, Type I error is implicitly being regarded as more serious and more worthy of control than Type II error. However, from the point of view of interested parties who would view Type II error as more serious than Type I error, the test as conducted would seem unfair and misleading. [Other statistical decision procedures (e.g., Bayesian) often can alternatively be used, but these have their own arbitrary aspects.] The essential point here is that those who are inclined to view the declaration that the site is hazardous when it is actually safe as a more serious error than saying the site is safe when it actually is hazardous (e.g., alleged polluters) would naturally prefer to test via "proof of hazard", while those who would regard the determination that the site is safe when in fact it is hazardous as a more serious error than claiming the site is hazardous when in fact it is safe (e.g., persons residing close to a site of alleged pollution) would tend to prefer testing with "proof of safety".

Now suppose that a researcher who is truly concerned about controlling both errors plans to collect sample data but fears that it will not be possible to select a sample of adequate size to undertake a "proof of safety". Is the researcher stuck, or can a "minimally acceptable" sample size nevertheless be determined? We propose that the researcher undertake one of the testing procedures, say "proof of hazard", in such a way that, apart from the allowance factor  $A$ , "proof of safety" and "proof of hazard" are made equally difficult. This proposal, introduced in the next Section, gives an uncommon amount of consideration to the control of Type II errors. In addition, some thought should be given as to what constitutes a reasonable choice for the allowance factor,  $A$ .

#### 4. EQUATING TYPE I AND TYPE II ERROR PROBABILITIES IN THE "PROOF OF HAZARD"

Millard (1987a) shows that for the above "proof of hazard,"

$$\sqrt{z} = \frac{(z_\alpha + z_\beta)(\theta + 2)(A + 2)}{2(\theta - A)}, \quad (4.1)$$

where  $\beta$  is the probability of Type II error corresponding to  $\theta$ , for  $\theta > A$ .

Suppose we equate  $\alpha$  and  $\beta$ , and henceforth refer to this common quantity as the *joint error probability*. In doing so, we do not claim that the errors are literally equally serious. Instead, we feel that both need to be under control, and the sample size resulting from application of our proposal will give more needed control over Type II error than would a "proof of hazard" using the automatic  $\alpha = 0.05$ .

From  $\alpha = \beta$  we have  $z_\alpha = z_\beta$ , and (4.1) may be restated in various ways:

$$z = \left[ \frac{z_\alpha(\theta + 2)(A + 2)}{(\theta - A)} \right]^2, \quad (4.2)$$

or

$$\theta = \frac{A\sqrt{z} + 2z_\alpha(A + 2)}{\sqrt{z} - z_\alpha(A + 2)}, \quad (4.3)$$

or

$$\alpha = 1 - \Phi \left[ \frac{\sqrt{z}(\theta - A)}{(\theta + 2)(A + 2)} \right], \quad (4.4)$$

with  $\Phi(\cdot)$  the standard normal cumulative distribution function.

Equations (4.2)–(4.4) are alternate representations of the menu of choices involving the sample size  $z$ , the joint error probability  $\alpha$ , the allowance  $A$ , and the true value of parameter  $\theta$  under  $H_a$ . We propose to use them in order to assess whether a suggested sample size and desired joint error probability are compatible.

The use of a nonzero value of  $A$  in our formulation allows the investigator to regard as unequally serious the Type I and Type II errors in the “proof of hazard” discussed in Section 3 by making it possible to declare a site is safe when there are actually 100A% excess deaths. Thus the selection of a value for  $A$  may be a political matter reflecting one’s attitude toward these two errors. Bross (1985) states that “there is general agreement that  $A$  should be a small fraction” and suggests the choice  $A = 0.1$ ; i.e., an excess of deaths following the event of only 10% or less is not suggestive of hazard. We will examine this value of  $A$  and also the

choice  $A = 0$ , which errs on the side of overcautiousness in claiming that any excess of deaths is suggestive of hazard. Other values of  $A$  could also be considered, but since even 10% excess deaths seems to us to provide plenty of margin for error to advocates of proof of hazard, we will not consider  $A > 0.1$ .

We do not envision the use of (4.2)–(4.4) with a single pair of values  $(\alpha, \theta)$  because researchers are interested in how well errors are controlled for a variety of possible values of relative increase in deaths,  $\theta$ . Sample size determination should proceed from examination of what is possible for a few or several selections of  $(\alpha, \theta)$ . To facilitate this, we present, in Figures 1 and 2 for  $A = 0.1$  and  $A = 0$  respectively, two-dimensional depictions of the interrelationships among  $z$ ,  $\alpha$ , and  $\theta$ . The researcher can determine from Figure 1a or 2a which is the northeast–most curve that is consistent with his or her views concerning various  $\alpha$  and  $\theta$ . If the researcher wishes to concentrate on a particular fixed  $\theta$ , Figures 1b or 2b may be more useful.

For  $A = 0.1$ , Figure 1a presents contours of constant  $z \leq 150$  for  $0 < \alpha < .20$  and  $0.3 < \theta < 2.3$ , and Figure 1b contains contours of constant  $\theta$  for  $0 < \alpha < .20$  and  $z \leq 200$ . Conclusions such as the following emerge from these charts. We can control  $\alpha = \beta = 0.10$  at  $\theta \geq 0.90$  with  $z \leq 100$ . But if we are willing to accept  $\alpha = \beta = 0.20$ , this can be attained at  $\theta \geq 0.60$  with  $z \leq 100$ ; that is, if we can increase the joint error probability to 0.20, this can be achieved with a sample of only 100 cases when there are truly 60% excess deaths. In order to equate  $\alpha$  and  $\beta$  when there are actually 100% excess deaths (i.e.,  $\theta = 1.0$ ), we can test with only 80 cases and get  $\alpha = \beta = 0.10$  or under 40 cases if  $\alpha = \beta = 0.20$ .

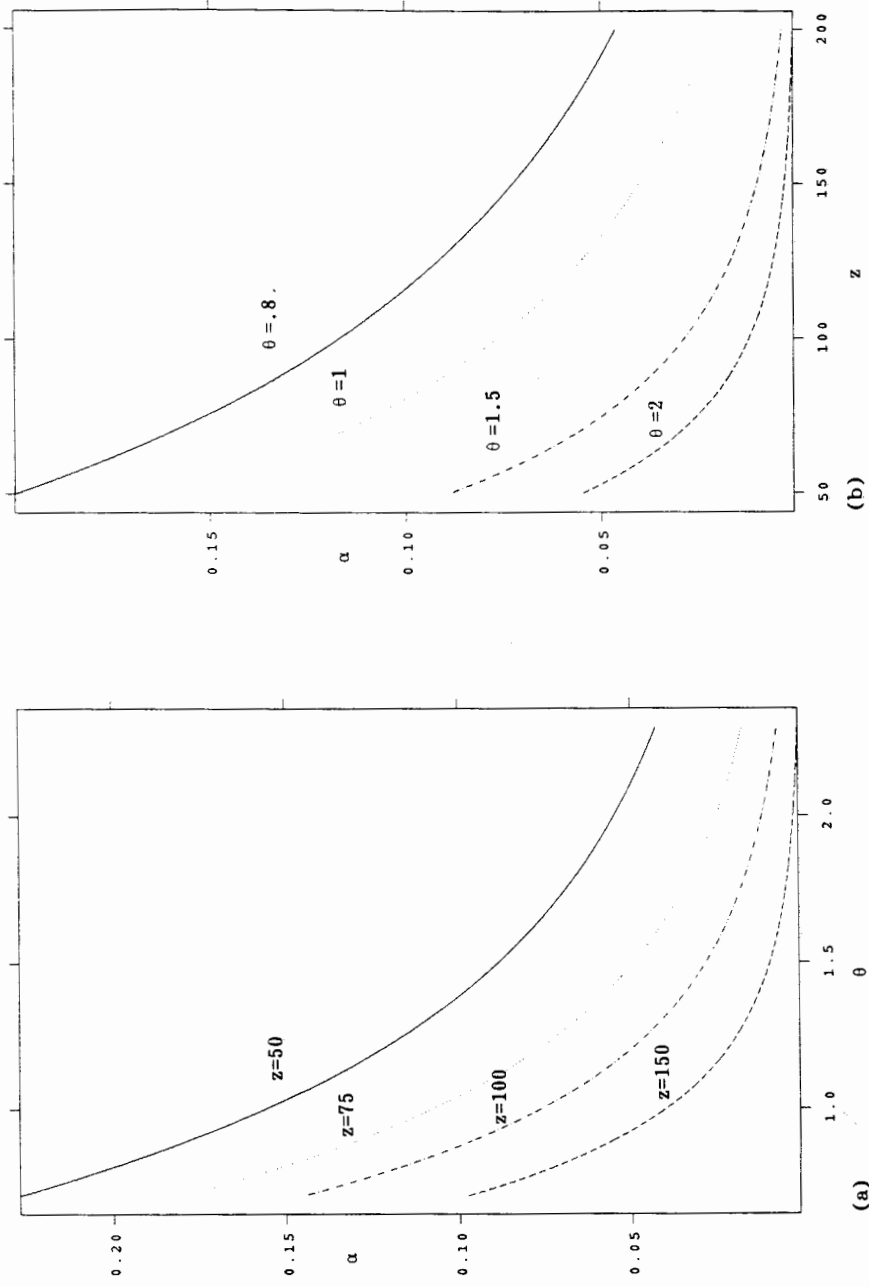


Figure 1. The interrelationship among joint error probability ( $\alpha$ ), sample size ( $z$ ) and true relative increase in death rate ( $\theta$ ) for allowance  $A=0.1$ .

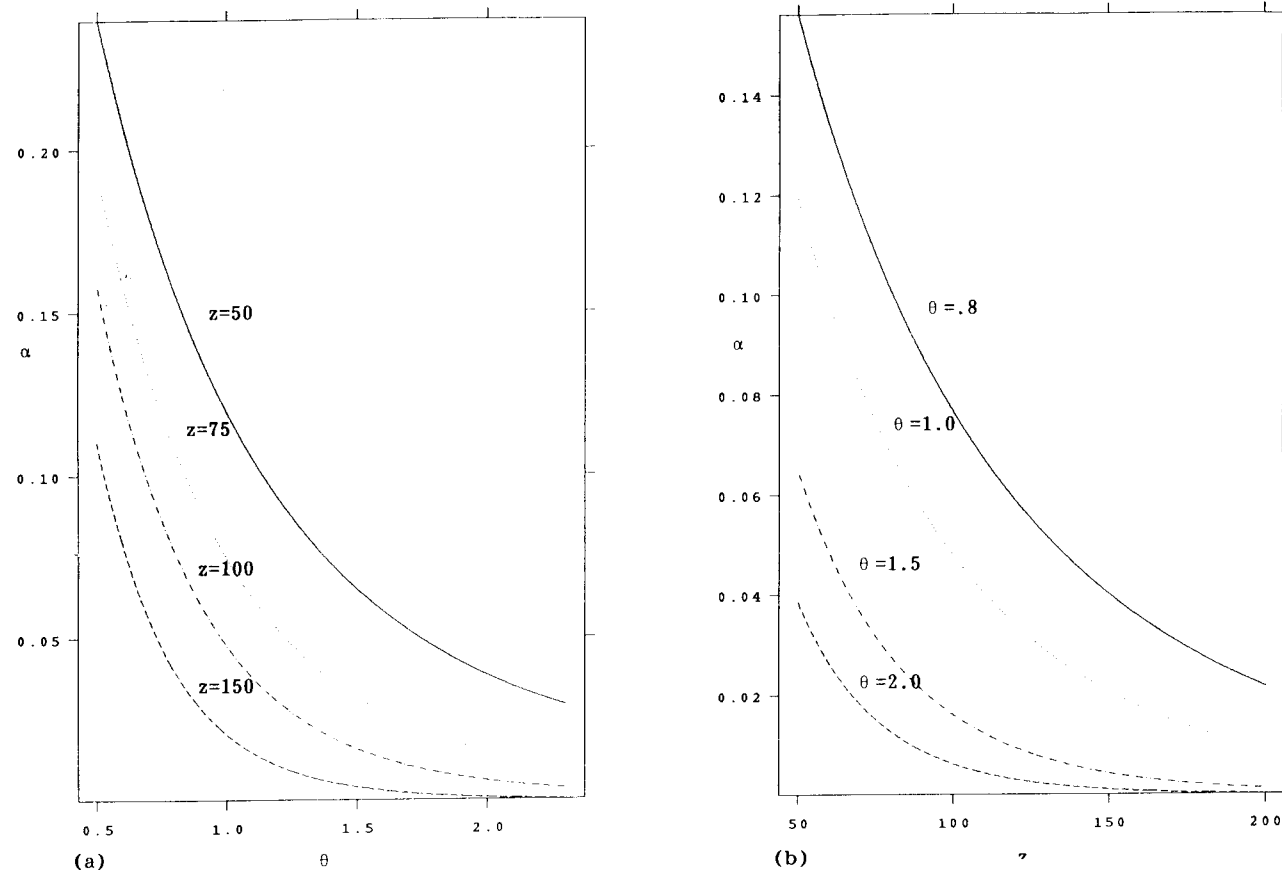


Figure 2. The interrelationship among joint error probability ( $\alpha$ ), sample size ( $z$ ) and true relative increase in death rate ( $\theta$ ) for allowance  $A=0$ .

We now consider the choice  $A = 0$ . As compared with  $A = 0.1$ , for a given value of  $\alpha$  we can reduce somewhat either the sample size  $z$  or the true value of the relative change in deaths  $\theta$ , or both. The following are illustrations of the type of information conveyed by figures 2a and 2b for  $A = 0$ . We can have  $\alpha = \beta = 0.10$  at  $\theta \geq 0.70$  with  $z \leq 100$ , or  $\alpha = \beta = 0.20$  at  $\theta \geq 0.50$  with  $z$  at most only 75. To equate  $\alpha$  and  $\beta$  when there are actually 100% excess deaths ( $\theta = 1.0$ ), we can have  $\alpha = 0.10$  with only 60 cases or  $\alpha = 0.20$  with only about 25 cases.

These examples are consistent with the contentions of Bross and Millard that the sample size requirements are somewhat insensitive to the choice of  $A$ .

Many advocates of "proof of safety" would find any  $A > 0$  unpalatable. Another way to regard the two errors asymmetrically would be to set  $\beta - \alpha = \delta$  for some given  $\delta \neq 0$  (rather than  $\delta = 0$  as in Section 4). Then replacing  $\beta$  with  $\alpha + \delta$  in (4.1), we have

$$\sqrt{z} = \frac{(z_\alpha + z_{\alpha-\delta})(\theta + 2)(A + 2)}{2(\theta - A)}. \quad (4.5)$$

Clearly, this relationship could be displayed in the manner of two-dimensional Figures 1 and 2 except that now *two* of the five quantities  $z, \alpha, \delta, \theta$ , and  $A$  need to be fixed in each figure. Explicit solution of (4.5) for  $\alpha$  appears to be difficult, but numerical solution would present no problem.

## 5. DISCUSSION

In some instances it may happen that the use of these or similar charts will suggest an impossibly large sample size unless the investigator is only willing to protect against an unrealistically large  $\theta$ . In such cases, our approach would be unhelpful, and one can only bemoan the impossibility of a larger sample. Nevertheless, there will be other circumstances where by raising Type I error tolerance, researchers will be able to have adequate protection against Type II errors for modest true alternatives with attainably small sample sizes. In any event, it is hoped that when performing hypothesis tests concerning environmental safety (and in other applications as well), the uncritical use of the  $\alpha = 0.05$  standard will cease to be automatic. Other investigators have also advocated consideration of  $\alpha > 0.05$  in environmental testing contexts; see, for example, the use of  $\alpha = 0.10$  in Bross (1985), Millard (1987b), and Stoline and Cook (1986). We feel that on rare occasions, even  $\alpha > 0.10$  may be worth considering.

Inadequate attention to Type II error is not limited to questions of safety or hazard. It occurs any time the outcome of a hypothesis test is of interest to each of two organizations or groups of people with opposing views (such as regulators *vs* regulated entity, management *vs* labor, candidates for public office, competing firms, etc.), when sample sizes are small and there is uncritical insistence on  $\alpha = 0.05$  or  $0.01$ . As a result, there typically is excessive control of the error of declaring that an alternative hypothesis statement such as "the innovation (insurgent interest) (change) is better" when in fact this is not the case, at the cost of little or insufficient control of the error of supporting a null hypothesis statement such

as "the standard (vested interest) (status quo) is better" when truthfully that is not so. This tendency stifles innovation.

Whenever it is deemed desirable to balance the chances of incorrectly finding for the null hypothesis statement with the chances of incorrectly supporting the alternative hypothesis statement, unless the investigator is in the unusual luxurious situation of an overabundance of sample observations, serious consideration should be given to the abandonment of the  $\alpha = 0.05$  standard. Increasing this tolerance to 0.10 or possibly somewhat higher will always reduce  $\beta$  for all values of the alternative, and it seems to us that this is appropriate and defensible in many applications with small samples. The precise relationships between sample size, error probabilities and true parameter values will vary with the application, but equations and charts analogous to those presented in the previous Section can be easily developed.

#### ACKNOWLEDGEMENTS

The authors are grateful to the editors and referees for many useful suggestions, and to Barnes Johnson for several helpful comments.

#### REFERENCES

- Bross, I. (1985): Why proof of safety is much more difficult than proof of hazard. *Biometrics*, 41, pp. 785-793.
- Millard, S. P. (1987a): Proof of safety vs proof of hazard. *Biometrics*, 43, pp. 719-725.

Millard, S. P. (1987b): Environmental Monitoring, Statistics and the Law: Room for Improvement. *The American Statistician*, 41, pp. 249-253.

Stoline, M. R. and Cook, R. J. (1986): A Study of Statistical Aspects of the Love Canal Environmental Monitoring Study. *The American Statistician*, 40, pp. 172-177.

Received July 1989; Revised August 1990.

Recommended by S. J. Schwager, Cornell University, Ithaca, NY.