Accid. Anal. & Prev. Vol. 15, No. 4, pp. 275–285, 1983
Printed in Great Britain.

0001–4575/83 $3.00 + .00

REFLECTIONS ON METHODS OF STATISTICAL INFERENCE IN RESEARCH ON THE EFFECT OF SAFETY COUNTERMEASURES†

EZRA HAUER
Department of Civil Engineering, University of Toronto, Canada M5S 1A4

(*Received* 27 *August* 1982)

Abstract—Sensible management of traffic safety is predicated on having reasonable expectations about the effect of various safety countermeasures. It is the role of evaluative research to derive such intelligence from empirical data. In spite of decades of research and experience, the safety effect of many countermeasures remains unknown. This sorry state of affairs is largely due to the objective difficulty of conducting conclusive experiments. Recognition of this objective difficulty should lead to the realization that in transport safety, knowledge is accumulated gradually from small, noisy and diverse experiments. The statistical tools used to extract knowledge from data should reflect this aspect of reality. One must therefore question the usefulness of classical tests of significance as a device for scientific progress in this field. It is argued that the unquestioning and all-pervasive use of significance testing in evaluative research on transport safety amounts to a self-inflicted learning disability. In contrast, it is shown that classical Point Estimation, Likelihood-Support and Bayesian methods can all make good use of experimental evidence which comes in small doses. In particular, the likelihood function is an efficient device for the accumulation of objective information and a necessary ingredient for Bayesian decision analysis.

1. TRANSPORT SAFETY MANAGEMENT

The management of transport safety is a multifaceted and amorphous public activity. So many are the players in this game and so loosely coordinated appear their game plans to be, that the common aspects of "transport safety delivery" may escape recognition. Yet when one thinks of, say, the air traffic control system, the provision of school crossing guards, the licensing of drivers and highway geometry design as a few selected components of a transport safety management system, one comes to recognize how all-pervasive this public activity is and how large are the resources devoted to it. Legend has it that some 80% of the Transport Canada budget is linked to management of safety. It is therefore both natural and necessary to be concerned about the effectiveness with which resources are allocated and spent in this domain of public expenditure.

Resources are allocated to a variety of safety related tasks, programmes, standards, etc. It is best to lump these under the term: countermeasures. One can accept without qualms the statement that improved knowledge about the safety effect of countermeasures is likely to enhance the rationality of transport safety management. Conversely, no matter how well the outfit responsible for safety management is run and how sophisticated its officials are, if the safety effect of countermeasures is not known, one can not expect efficiency in resource allocation. These dicta are the motivation for engaging in research on the effect of countermeasures on transport safety.

This trite introduction is needed to set the stage for the arguments to come. It is best to state at the outset that evaluative research in transport safety is viewed here not as an element in the quest for the understanding of nature nor as an activity satisfying a primeval human thirst for knowledge. It is viewed within a utilitarian perspective as the kingpin of rational decision-making in the management of safety.

2. A LEARNING DISABILITY

Progress in research on the effect of safety countermeasures has been uneven and in many cases, sluggish. It is hardly unfair to maintain that in spite of many decades of experience with

†The support of Transport Canada (Road Safety) is gratefully acknowledged.

many countermeasures (such as driver education, illumination, speed limit enforcement, traffic signals, vehicle inspection, demerit point systems, etc.) there is no concensus about their safety effect. This is an unusual and perplexing fact which should give pause for thought. There must be reasons which explain why after half a century of doing something as costly and unpopular as, e.g. enforcing speed limits, society has been unable to use such extensive experience in order to learn whether the enforcement activity has some effect on safety.

Some of the formidable objective obstacles to progress in research on safety are well known. Foremost among those is the practical difficulty of conducting large-scale controlled experiments. Yet there is a nagging doubt whether these objective difficulties are sufficient to explain the persistence of ignorance. After all, reality does offer opportunities for study. To examine, e.g. the effect of speed enforcement, one can find speed limit enforcement programs introduced and abandoned, police strikes, differences in levels of enforcement, etc. Researchers are skillful to recognize many such opportunities. Thus, there exists a multitude of research studies, small or extensive, properly conducted as well as deficient in design and execution. Yet, in spite of laborious and costly research efforts spanning several decades, concensus about the effect of many a countermeasure is slow to emerge. Is it not possible, therefore, that in addition to the objective difficulties, there is also some "learning disability" at work? Is there not something in the manner in which we learn from experience and extract information from experimental evidence which acts as an obstacle to the emergence of knowledge about the effect of safety countermeasures?

It seems prudent to admit the possibility that not all fault lies with an adverse "state of nature"; that other factors contribute to slow progress in evaluative research on safety; that some of the impediments to progress are self-inflicted.

One of the important traits of observations on system safety is their random nature. This is why safety is explored using the theory of probability and the tools of statistics. The whipping-boy in this paper will be some methods of classical statistics. It will be argued in the sequel that the machinery of classical hypothesis testing is largely irrelevant for the task at hand and is often the reason for conservatism and slow learning. Alternative methods for extracting information from data will be explored later.

3. A BRIEF CRITIQUE OF THE CLASSICAL TESTING OF HYPOTHESES

A typical study about the safety effect of some countermeasure begins by conceiving an experiment and proceeds by collecting the data leading to the analysis thereof. The punch line in most cases is the testing of statistical hypotheses about the safety effect of the countermeasures studied. The conclusion therefore is usually a statement on whether a hypothesis can or cannot be rejected at a chosen level of significance when confronted with some alternative hypotheses.

There is considerable attraction in the apparent rigour of this process. It evokes the image of an idealized "scientific method" in which hypotheses are postulated and supported (or falsified) by experiments. This image of rigour seems to be particularly dear to disciplines not endowed by the facility for purposeful experimentation and deductive reasoning which is so characteristic of the natural sciences.

Be it as it may, the discourse on the method of science is best left to its philosophers. Here we will merely try to substantiate the claim that the persistent practice of interpreting data through classical tests of hypotheses may exert a thwarting influence on progress in research about safety countermeasure effectiveness. This is so, because the "decision" to reject or not to reject a hypothesis is largely arbitrary and cannot be interpreted in any meaningful manner. Moreover, the real-life setting in which research on the effectiveness of safety countermeasures has to be conducted is characterized by relatively small samples and deals with countermeasures the effect of which is typically small. The conventional test of a hypothesis in these circumstances will usually return the answer: the hypothesis "no effect" cannot be rejected. The net results of this built-in conservatism is that most real-life countermeasures are branded as "not shown effective". This in turn leads to perpetuation of the status quo and to stagnation. Thus, it appears important to argue the case against the use of classical tests of hypotheses in research on the effect of safety countermeasures.

The "knocking" of statistics is a popular pastime. It is often done on the basis of some improper use of the statistical method. To any such charge, the statistician would respond by claiming that misuse is not the fault of the method but of those using it incorrectly. (Even the validity of this response can be questioned. It is not common to find statistical methods used properly. Nor is there a good prospect that this will change. Thus, if a method cannot be used properly by well meaning mortals, it may not be fit for general human consumption.) However, this critique is aimed at the case when the classical method is applied as intended–without flaw. We will claim, that in the context of research on countermeasure effectiveness in transport safety, methods of classical hypothesis testing are largely irrelevant and often harmful. This claim will rest on two groups of arguments.

(a) The outcome of a classical test of hypothesis is arbitrary. Whether a hypothesis is "rejected" or "not rejected" depends on considerations that cannot be related in any meaningful fashion to the real circumstances of the issue under scrutiny. Thus, one can "decide" either to reject or to accept any of the hypotheses under consideration.

(b) The "decision" (to reject or not) which is the result of the "test" is difficult to interpret in some straightforward manner. Yet, the obscure concept of "decision" is the fulcrum on which everything hinges. Moreover, the criteria used to make this "decision" cannot be used to weigh the importance of the consequences in a real decision-making context. These two groups of arguments are explored in some detail below.

The first group of arguments leads to the conclusion that the result of a test of hypothesis is arbitrary. The arbitrariness of the procedure stems from three major difficulties.

First, whether the null hypothesis is rejected or not depends entirely on the chosen level of significance. Given the results of an experiment, for some levels of significance you reject the null hypothesis, for others you don't. If the results of a test depend so immediately on the chosen level of significance, one would like to be shown that the custom and convention in this matter have a rational justification. For, if such a basis cannot be found, one is justified in thinking the test to be arbitrary.

Both elementary and advanced texts on statistics either leave it to the discretion of the analyst to choose the appropriate "critical levels" (as if it was evident how to do so) or recommend to follow custom and use the 0.01 or 0.05 levels of significance. But how is one to judge what is "appropriate"? Does adherence to custom make the outcome "arbitrary-by-custom"?

A down-to-earth justification of the conventional critical levels is given by Bross [1971]. His argument is rooted in the practice of biostatistics and consists of two elements. By the first element, progress would be impeded if too many erroneous claims (say, of drug effectiveness) were published. Thus, it is necessary to set some level of significance to keep "false positives" out of the pages of journals. The second element of the argument is that "the 5% level ... is a feasible level at which to do research work"—in biostatistics.

In research on the effect of safety countermeasures, the 5% level is rarely a feasible one. To illustrate, Berg (1980) concludes that in order to detect a 5% safety effect of an advance warning sign for railroad-highway crossing at the 5% level of significance (and with a 50% chance of not finding it), one would require 15000 "treatment" and 15000 "control" sites. This exceeds the total number of level crossings in the USA. Thus, if there is to be a conventional level of significance in safety research, it is not necessarily that which is suitable for biostatistics or agriculture.

The damage done to scientific progress by the publication of an erroneous account of effectiveness must be weighed against the damage done by branding "not proven effective" many a promising countermeasure. Which is the more serious damage is surely a matter of opinion. Thus, even the level-headed argument by Bross leads back to the exercise of judgement. This judgement is far removed from the specific research question at hand or a set of experimental data to be interpreted. It is difficult to think of a test as not arbitrary if its result depends on some metaphysical aspects of scientific progress.

The second major difficulty is the fact that the answer returned by a statistical test of hypothesis depends crucially on yet another custom–the convention by which one selects the "null hypothesis". The outcome of the test is often determined by the decision which of the contending hypotheses to cast in the role of the "null hypothesis".

Consider two simple hypotheses:

X—countermeasure has no effect.

Y—countermeasure is 10% effective.

Given is a small set of experimental results and an appropriate level of significance is agreed upon—somehow. Assume that when X is selected to serve as the null hypothesis, it cannot be rejected. One concludes

(1) The hypothesis that the countermeasure has no effect cannot be rejected.

Assume that if Y were selected to serve as the null hypothesis it also cannot be rejected (at the same level of significance and using the same experimental data). In this case one states:

(2) The hypothesis that the countermeasure is 10% effective cannot be rejected.

Ordinarily, statement (1) would be interpreted as being different in meaning from statement (2). If so, one obtains different answers from a test as a result of a choice about which of the contending hypotheses is to serve in the role of the "null hypothesis". This choice is dictated by arbitrary convention.

One can maintain that statements (1) and (2) should not convey different impressions about the effectiveness of the countermeasure; that they are complementary and illuminate the same facts from two slightly different angles; in fact, that there is a whole range of parameter values, hypotheses about which cannot be rejected. This interpretation runs counter to the concept of a "test" which culminates in a "decision". It runs counter to the manner in which the machinery of hypothesis testing is used.

Finally, even if one swallows convention in both cases and agrees on a level of significance and null hypothesis, the outcome of the test remains arbitrary. For, no matter how small the difference between the "null" and the "alternative" hypotheses, the null hypothesis can be almost always rejected when the experiment is large enough. Thus, not only does the outcome of a "test" depend on two inexplicable conventions, it also depends on the budget available for experimentation. It is difficult to see what meaning to assign to a "test" if its outcome is determined by vague conventions and other extraneous influences.

The second set of objections to the process of classical hypothesis testing centers on the lack of clarity as to what a "decision" is to mean. At the end of a test of a hypothesis, one has to "decide" whether "to reject" or "not to reject" the null hypothesis. What precisely are the consequences or guidance for subsequent action implied by such a decision? One interpretation of what "deciding" means may be, that until such time as the null hypothesis can be rejected, it remains the accepted doctrine. I do not know whether this a workable scheme in the natural sciences. In safety, it is seldom easy to determine what is meant by "accepted doctrine". In any case, this interpretation seems to run counter to the convention of using "no effect" as the hypothesis we are trying to reject. It is difficult to believe that the accepted doctrine about most countermeasures we test is, that they have no effect. The contrary may be closer to the truth. We test them because there is some ground to believe they might be effective.

This leads to another possible interpretation of what "deciding" might mean. The null hypothesis we select (no effect) is really only a strawman; it is placed on the altar only to be destroyed by experimental evidence ("it is merely something set up like a coconut to stand until it is hit" Jeffreys, 1961). When so rejecting the "no-effect" hypothesis we imply that the "yes-effect" hypothesis replaces it. But if we do not manage to unseat the "no-effect" hypothesis (as happens with discouraging but persistent regularity), we are saddled with a reigning hypothesis which we did not believe in from the outset and which was selected, so to speak, out of spite. In this case the most that can be said is that while we do not necessarily believe in the "no effect" hypothesis, we have no sufficient proof to reject it. In, say, the biological sciences, it might be possible to buy more guinea pigs and experiment till the null hypothesis can be rejected. Such an option does not ordinarily exist in research on safety.

The same sentiment is expressed delightfully by Edwards [1972, pp. 179, 180]: "unfortunately any method which invites the contemplation of a "null" hypothesis is open to grave misuse, or even abuse, and this seems particularly so in the social sciences, where high standards of objectivity are especially difficult to attain, and data often of dubious quality. The argument runs as follows: 'I am interested in the effect of A on G (for example, the influence of hereditary factors in the determination of human intelligence, or the effect of increased family

allowances on population growth) and I propose to use approved statistical techniques so that no one can question my methodology. These require me to state a null hypothesis, namely, that A has no effect on B. I now test this null hypothesis against my data. Unfortunately my data are not very extensive, but I have done the Angler–Plumfather two-headed test and found $0.20 \leq P \leq 0.10$. I therefore accept the null hypothesis.' Further sets of data—none of them very extensive—continue to "miss the coconut", and after a time the null hypothesis joins that corpus of hypotheses referred to as "knowledge", on no positive grounds whatever.

"The dangers are obvious. In the first place, the problem is usually one of estimation (to use the conventional work) rather than hypothesis-testing; in the second place, the chosen null hypothesis is often such that no rational man could seriously entertain it: who doubts that hereditary factors have some influence on human intelligence, or that increased family allowances have some influence on population growth? And in the third place, not only is each test itself devoid of justification, but sequential rather than concentrated assaults on the null hypothesis are practically powerless in difficult cases: it is like trying to sink a battleship by firing lead shot at it for a long time. What used to be called judgement is now called prejudice, and what used to be called prejudice is now called a null hypothesis. In the social sciences, particularly, it is dangerous nonsense (dressed up as 'the scientific method'), and will cause much trouble before it is widely appreciated as such".

Edwards' diagnosis is near prophetic in view of the following comment by Meehl [1978]:

"I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happened in the history of psychology."

The last escape from this morass I can think of is, that the consequences of deciding to reject or not to reject vary from case to case and therefore each situation requires separate analysis. This might be a sensible escape route. Unfortunately it is blocked. To weigh the importance of "rejection" or "non-rejection" in a specific case, we need probabilities of erring. The classical test of a hypothesis does not yield estimates of the probability of making an error one way or another. It returns conditional probabilities. That is, the probability of erring *if* one (or the other) hypothesis is true. But since it remains unknown *whether* one or the other hypothesis is true, *we are not in a position to weigh the importance of deciding to reject or not to reject.*

In summary, the classical test leads to a "decision" in some contorted sense of the word and uses for that purpose conditional probabilities which cannot be related to the importance of the consequences of error—the costs and benefits associated with any real decision-making.

One cannot fault classical hypothesis testing for being entirely devoid of any consideration based on cost or benefits. The method has not been conceived to be used as a decision making tool. (Perhaps the terminology which uses the word "decision" in the context of testing hypotheses has been an unfortunate choice.) It is possible that when experimentation is cheap, the classical method is a good safeguard against unproven hypotheses and its consistent use promises to keep the body of knowledge unpolluted and the ship of progress on even keel.

However, research on the effect of safety countermeasures *is* in support of deliberations about costs and benefits. It is therefore incongruous to think that "decisions" can be made without reference to resource implications. For this reason alone, classical hypothesis testing should be rejected as a legitimate tool for summarizing experimental results in research on safety countermeasure effectiveness.

As used, classical hypothesis testing is usually a one-shot affair. It is the punch-line of a study. An experiment is conducted, outcomes are tabulated and then subjected to a test. When, as often happens, the null-hypothesis is not rejected, the information contained in the outcome of the experiment loses much of its value. The data is seldom used in the next study of the same issue. What one finds more often is a tally of asterisks—how many previous studies found significant differences, at what level, and how many failed to do so. One then relies on scientists or decision-makers to fuse the nose-count of asterisks and their own opinions by some undefined mental process. This is possibly the most unfortunate feature of classical significance testing—as used. We operate in an environment in which knowledge accumulates through many experiences. Failure to make use of all accumulated information in an explicit and purposeful learning process virtually assures stagnations.

4. HOW TO LEARN FROM DATA

The central question before us is: "how to extract information from results of experiments and how to learn from it".

In discussing this question here, the primacy of two aspects of the prevailing situation has to be recognized. First, that the purpose of research on the effect of safety countermeasures is to facilitate sensible decisions about the allocation of resources in safety management. Second, that a single experiment by itself is likely to be too small and noisy to yield authoritative knowledge. Therefore, in most cases, one cannot settle a question once and for all.

Thus, research on countermeasure effectiveness is a process. The need to engage in it arises when one suspects that the present state of knowledge on some countermeasures is insufficient to make good decisions about its implementation. That, in fact, if we knew more, we might make better use of resources. It follows, that the purpose of research is to improve, update and revise present knowledge. This revised knowledge will serve for any decisions that need to be made here and now. But if one still has ground to believe that better knowledge might save resources, there is room for more research and yet another revision of present knowledge, and so on ad infinitum.

To elucidate the role played by research on the effect of safety countermeasures we start by describing a simple-minded but general framework for making decision. This will be followed by an illustrative example.

4.1 *A naive framework for decision analysis*

Implementation of a countermeasure is considered. Two of its attributes are of interest: the cost of the countermeasure and its ability to improve safety. Let then C denote the annual cost of the countermeasure and θ be an index measuring the effectiveness of the device. When the device is implemented on a system on which the expected annual number of accidents is M, the expected annual number of accidents after implementation will be $M\theta$. (For the sake of simplicity, it is assumed that the effect of countermeasure is to change the expected number of accidents. Safety is not measured only in terms of quantity of accidents but also their severity. This simplification has no effect on the essence of the argument to follow.) If $\theta = 0$, the device is totally effective. If $\theta = 1$, the countermeasure does not reduce expected annual number of accidents. The magnitude of θ is never known exactly. Informative statements about θ must always be couched in terms of odds and probabilities. The aim of research on the effectiveness of safety countermeasures is to obtain a good estimate of θ.

The decision to implement the countermeasure has positive and negative repercussions. On the credit side of the ledger is the reduction in the expected annual number of accidents $(1 - \theta)$. On the debit side is the expenditure of resources C.

Whether the implementation of the countermeasure is a worthwhile investment can be judged only if it is known whether the same resources cannot result in a larger safety improvement if invested in a different countermeasure. (The problem of project divisibility is neglected here. It is assumed the investment in countermeasures is finely divisible. Nor do we propose to deal with uses of money in fields other than management of transport safety.) Let then A denote the reduction in the annual number of accidents obtainable by investing one unit of money per annum in the best alternative countermeasure. With this meagre notational arsenal, a sensible decision rule might be:
Implement device if:

$$M(1 - \theta) > CA. \tag{1}$$

The rule incorporated in the inequality urges one to implement the accident reducing device if (with equal expenditure of money) it is expected to save more accidents than the best alternative countermeasure. (This decision rule is unrealistically stringent. Even if the new device only exceeds the cost effectiveness of the *worst* presently financed countermeasure it should be implemented, provided that resources can be diverted from one to the other.)

Estimation of three out of the four variables in inequality 1 is relatively straightforward. It makes common sense that when considering the investment of some money, one should know what the cost of the project (C) is, what the magnitude of the problem (M) is and what could be done with the money elsewhere (A). Normally, however, the index of countermeasure effectiveness (θ) is

known with little certainty and decisions must be made on the basis of the best available evidence.

So far, θ was assumed to have some unique value, estimates of which are obtainable from experimental data. However, the same countermeasure when applied to different systems is likely likely to be different in effectiveness. Thus, it is legitimate to think of θ as a random variable with a probability distribution function and an expected value. It will be argued in the sequel that the decision-analytic framework is well served by a Bayesian approach to extracting information from data. This approach is based on a recurring revision of the probability distribution function of θ. Thus, to keep all options open, we will replace θ in inequality (1) by $E\{\theta\}$.

Within this framework, the role of research on the effect of safety countermeasures is to revise current estimates of $E\{\theta\}$ on the basis of new empirical evidence. The value of evaluative research is then measurable from the (expected) increase in the effectiveness with which resources are allocated.

4.2 *Illustrative example*

The principal issues can be illustrated by a hypothetical numerical example. Wide implementation of a novel pedestrian crossing device (PCD) is contemplated. Naturally, little is known about its safety performance. In this state of ignorance, the decision makers might wish to ask for expert opinion. Of the 10 experts polled, three opinion groups can be identified:

1 thinks the PCD will increase accidents by 10% (Group I)

6 think it will have no effect (Group II) and

3 think it will bring about a 10% reduction in accidents (Group III).

Installing the device at sites which have (jointly) some 100 fatal pedestrian accidents per annum would cost $200,000 per annum.

Depending on which group of experts is right, the following are the repercussions of implementing the PCD (Table 1).

Table 1.

	Expected Savings (fat./year)	Costs ($/year)
Group I is right	-10	200,000
Group II is right	0	200,000
Group III is right	10	200,000

If the decision makers value equally the opinion of every expert, and no other information about the effectiveness of the new PCD is available, the best presently available estimate of $(1 - E\{\theta\}) = 0.02$. $[0.1 \times (-0.10) + 0.6 \times (0.00) + 0.3 \times (0.10) = 0.02]$. Accordingly, the best estimate of $M(1 - E\{\theta\})$ is $100 \times 0.02 = 2$ fatalities/year. This is the estimate of the expected reduction in pedestrian fatalities obtained by the investment of $200,000 per year.

Assuming that one could hire for the same amount of money some 10 crossing guards and expect to save thereby 2.5 fatalities per year, it would be wise to decide against implementing the novel PCD.

If, however, the experts in Group III happen to be right, an incorrect decision has been made (to save 2.5 lives instead of 10). It may therefore be worthwhile to do some research about the effectiveness of the PCD so as to improve our ability to make good decision and thereby to enhance the efficiency with which money is used in safety management.

Assume then that a research project is funded and the PCD is installed at some representative sites. During two years prior to PCD implementation, these sites recorded 10 fatal pedestrian accidents; during two years following PCD implementation, 5 fatalities occurred.

The question is, how to make good use of this meagre (but costly) empirical evidence.

Were one to subject this result to a classical test of hypothesis, the inescapable conclusion is that the hypothesis, "PCD is not effective" cannot be rejected. What effect such a conclusion would (or should) have on the decision about the future of the PCD is unclear. Most persons with normal instincts would conclude that the result lends some support to the hypothesis that the PCD

is effective which is in stark contrast to the result of the test. Thus, having conducted a test of hypothesis, one is hardly further ahead. It is therefore important to try and suggest better ways for squeezing information out of data. Three important options present themselves:

(1) Classical "point estimation".

(2) The "Likelihood" method, and

(3) "Bayesian" estimation.

The three methods are interwined. Thus, e.g. one of the more popular techniques for obtaining a point estimate is to search for the maximum of the likelihood function. Also, the likelihood function plays a central role in Bayesian estimation.

All three options are viable devices for the accumulation of empirical evidence, as will be shown below.

Were one to adopt classical "point estimation", the data indicate $\hat{\theta} = 5/10 = 0.50$. Following Cox and Lewis [1966, pp. 223–225], the 95% confidence interval for $\hat{\theta}$ is 0.14–1.70.

These are constructive and informative statements. On the basis of the empirical evidence available, they convey the correct impression about what the effect of the PCD seems to be and what the associated uncertainty is.

One could even use the estimate $\hat{\theta}$ in inequality 1 and conclude that the implementation of the PCD is very attractive. Such a conclusion, however, takes no cognizance of the large uncertainty surrounding θ and disregards completely the body of expert opinion. These shortcomings cannot be easily remedied. Within the classical frame of thought it is not legitimate to ask: what is the probability, say, of $\theta \geq 1$? "Nor is there a satisfactory way to reconcile expert opinion with empirical findings. It must be left to the "decision maker" to do the best with the information presented."

On the other hand, the accumulation of evidence from several studies appears to be feasible. Should results of a subsequent study on the PCD be, say, 4 fatalities"without" and 3 fatalities "with", it is easy to revise the earlier estimate. Now, $\hat{\theta} = (5+3)/(10+4) = 0.57$ and the 95% confidence interval is 0.21–1.44.

Thus, it is not only possible, but in some cases very simple, to ensure that the estimate of θ is revised when new data becomes available; the current estimate is to reflect all evidence accumulated untill then.

As the body of experimental evidence grows, the confidence interval becomes narrower. In this sense, the confidence interval is a useful index of the degree of uncertainty in $\hat{\theta}$. A word of caution is in order. It is a common misconception to think of θ as a random variable which is contained with probability of 0.95 in the 95% confidence interval calculated from some data. This is nonsense. The value θ either is or is not within a specific range and to speak of probability in this context has no meaning. However, this misconception seems to be much less damaging than the acceptance-rejection malaise. Therefore, no more will be said about it.

The second of the aforementioned three options for extracting knowledge from data is to make use of the notion of "likelihood". If an experimental result has a greater probability of occurring if θ_1 is true than if θ_2 prevails, θ_1 is said to be more "likely" than θ_2. Thus, likelihood measures the strength of support which data lend to alternative values of an unknown parameter. The concept has first been introduced by Fisher [1925]; its logical foundations are explored by Hacking [1965] and one of its convincing proponents is Edwards [1972].

In Fig. 1, two likelihood functions are shown. The ordinate of the dashed curve is the likelihood of θ when only information about 10 "before" accidents and 5 "after" accidents is available. Thus, e.g., $\theta = 0.9$ has twice the likelihood of $\theta = 1.2$. In this case, the best supported (most likely) value of θ is 0.5, in accord with common sense and the "point estimate" obtained earlier. When new evidence in the form of 4 accidents "without" and 3 accidents "with" becomes available, the dashed curve is modified and the solid likelihood function obtains. Now the best supported value of θ is 0.57. An increase in information will, in general, result in a narrower likelihood function. This merely indicates that outlying values of θ become less supported by the data in comparison with the most likely value.

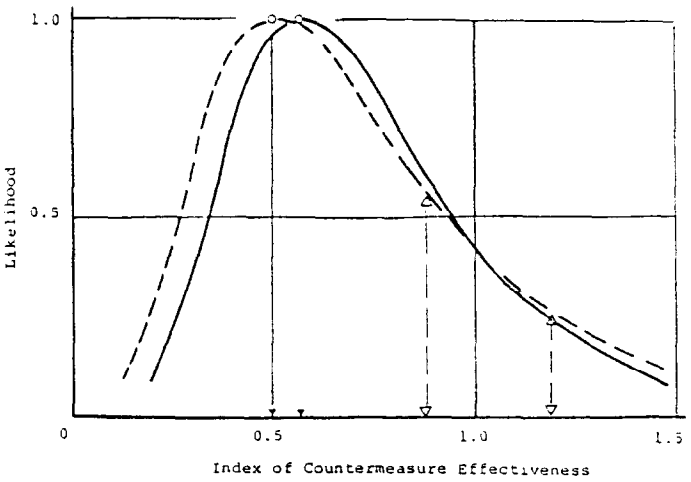

Fig. 1. Two likelihood functions.

The likelihood method has many attractive features. First, it presents the available empirical evidence in a manner which is closely tied to intuition and common sense. In particular, it does away with the obfuscation which goes with "levels of significance". Second, the method virtually invites accumulation of experimental results because the revision of the likelihood function when new data is acquired is usually very simple. All that is needed is to multiply the ordinate of the old likelihood function by the corresponding ordinate of the likelihood function based on the new data only. Third the maximum likelihood estimate extracts all pertinent information from the data (i.e. is a sufficient statistic). Thus, as a device for the accumulation and presentation of empirical evidence, our preference is to use the likelihood method instead of the more popular point and interval estimation.

Neither "point estimation" nor the "likelihood" take cognizance of what knowledgeable people believe the PCD can do to enhance safety. It may be inappropriate to disregard their experience in reaching a decision, particularly in view of the paucity of experimental data. Nor are any of the aforementioned methods capable of answering questions such as: "what are the odds that the PCD is harmful?" In contrast, the Bayesian approach is ideally suited for the task of making decisions as is illustrated below. In the present context, the Bayesian method is used to further three aims: to make both expert opinion and empirical evidence count as information, to ensure that every bit of new empirical evidence exerts its appropriate influence on current knowledge, and to express the current knowledge in a comprehensive manner which is directly usable for making decisions.

The kingpin of the method is the revision machine embodied in:

$$\left[\begin{array}{c}\text{Probability that the index of countermeasure effectiveness} = \theta \\ \text{when new information is available}\end{array}\right]$$

is proportional to

[The Likelihood Function which contains all empirical evidence]

multiplied by

$$\left[\begin{array}{c}\text{Probability that the index of countermeasure effectiveness} = \theta \\ \text{before the new information became available.}\end{array}\right]$$

To illustrate the workings of the machine, we continue with the same numerical example. Consider first the point in time before any empirical evidence became available. Relying on expert opinion only (Table 1), the probability (as degree of belief) that the PCD will, say, increase accidents by 10% ($\theta = 1{\cdot}10$) is 0.10. These probabilities are listed in column 3 of Table 2.

Table 2.

1	2	3	4	5
Group	Index of Effectiveness θ	Probability estimate based on expert opinion only $\pi(\theta)$	Revised probability estimate based on expert opinion and first data set $\pi(\theta)$	Revised probability estimate based on expert opinion and both data sets $\pi(\theta)$
I	1.10	0.10	0.07	0.07
II	1.00	0.60	0.57	0.56
III	0.90	0.30	0.36	0.37

When later the first data set became available (10 fatalities "before" and 5 "after"), the likelihood function shown by the dashed curve can be calculated. Its ordinates at $\theta = 0.9$, 1.0 and 1.1 are 0.545, 0.428 and 0.331, respectively. Using these and the revision machine, the new probability probability estimates in column 4 are obtained. If the decision maker adopted earlier expert opinion to to base his assessment of the probabilities on, now he must revise his views in the face of the newly acquired experimental evidence to what is given in column 4. As one should expect, comparison of columns 3 and 4 reveals that the data support the views of the experts in Group III and shift some of the probability mass to that row. The magnitude of the shift reflects (partly) the amount of information contained in the data. When at a later time, results of another study become available (4 fatalities "without" and 3 "with" PCD), the probabilities are revised again. This time the ordinates of the likelihood function, shown by the solid curve in Fig. 1, are used. The results are listed in column 5 of Table 2. Now the estimate of $E(\theta)$ is 0.971.

Accordingly, the current estimate of the expected reduction in pedestrian fatalities obtainable by implementing the PCD is $100 \times 0.029 = 2.9$ fatalities per year. This appears now to be more attractive than the hiring of crossing guards. In fact, by spending $200,000 per annum on crossing guards, we expected to save 2.5 fatalities; however, one needs to spend only $170,000 per annum on the PCD's to expect to save 2.5 fatalities. The expected saving of $30,000 per annum is attributable to the information obtained through research which led to a better decision. It is possible that more research on this countermeasure could be justified. Results of any such research would be used to revise the estimate of $\pi(\theta)$ in the rightmost column of Table 2.

No illustrative example can be convincing in all detail. Nevertheless, the overall impression is, I hope, of common sense and rationality. The meaning and context of decision making is clear and fits the task of managing safety; new information created by research contributes to better decisions by revising previously held views. There is no need for obscurantism about imaginary decisions, pseudo-scientific hypotheses, arbitrary levels of significance and convoluted definitions of "errors".

The Bayesian frame of thought seems to solve several problems: even small experiments contribute to knowledge; the method ensures that results of all previous experiments influence present knowledge; the output is in a form which can be directly related to cost effectiveness exercises; it is possible to assess the importance of research beforehand to identify research designs which will not yield information.

4.3 *Discussion*

Three options for learning from experimental evidence have been outlined. Each can be used to update and revise estimates of countermeasure effectiveness in an accumulative manner. Thus, each option is a promising cure to the learning disability brought about by the affliction of hypothesis testing.

Our preference is for a combination of the likelihood and the Bayesian method. The likelihood function is to retain and accumulate all "objective" evidence. Decisions are to be taken using probabilities provided by the Bayesian machine.

The starting point of the Bayesian procedure in the illustrative example is subjective "expert" opinion. This fact is commonly viewed as a weakness which renders the method "unscientific". After all, arguments are convincing when based on facts, not opinion. Also, as is well known, people are poor assessors of probability and expertise is no insurance against bias.

Proper exploration of this issue leads to the realm of the modern philosophy of probability which is beyond the scope of this paper and the competence of its author. However, for a realistic perspective and balanced view a few points deserve mention. For one, the role of expert opinion in the process of revising the estimates of $\pi(\theta)$ is limited. It is only the first step, setting into motion a self-rectifying process, which in principle, goes on indefinitely. The set of probabilities which is based on expert opinion is, as it were, the initial guess in a convergent iterative solution algorithm. A poor initial guess implies many iterations until a sufficiently accurate solution is obtained, whereas with a successful starting guess, convergence is obtained in a few cycles. However, nobody will brand the result of an iterative solution as unscientific because it began from a subjective guess. In the same manner, it is hardly just to stigmatize the Bayesian procedure because its starting point is expert opinion.

Second it has been argued earlier that the rigor of classical statistical methods begins only after some arbitrary specification of, say, the level of significance. This, in some sense, is worse than expert opinion since any honest attempt of such specification is unaided by intuition and does not derive from any human experience. Thus, one ordinarily resorts to the illusory comfort of convention.

The "Bayesian Controversy" has raged for decades. Numerous books and articles by eminent scholars have been devoted to this subject. It is not the aim of this paper to add new insights into the argument. Its purpose is to introduce the Bayesian controversy into the realm of evaluative research on transport safety. Since the proof of the pudding is in the eating, results of an application to a specific countermeasure are described in detail in a companion paper [Hauer, 1983].

6. BRIEF SUMMARY

A message which runs counter to much of present practice is bound to raise discussion. To keep discussion on the central issues, it may be useful to restate the main points made.

Experimental data about the effect of many safety countermeasures comes often in small doses. Accordingly, the process of learning about the safety effect of countermeasures must allow for gradual accumulation of evidence. The classical hypothesis testing frame of thought is ill-suited for this task. In contrast, point estimation, the method of likelihood and the Bayesian probability revision machine are promising candidates. The decision-making frame of reference is well served by the Bayesian machine. To keep data and opinion separate, it is useful to adhere to a reporting practice explicitly identifying the likelihood function and the original "prior".

REFERENCES

Berg W. D., Fuchs C. and Coleman J., Evaluating the safety benefits of railroad advance-warning signs. *Transpn Res. Rec.* 773. National Academy of Sciences, Washington, D.C., 1980.

Bross I. D. J. Critical levels, statistical language and scientific inference. In *Foundations of Statistical Inference*, (Edited by Godambe and Sprott). Holt, Rinehart and Winston of Canada., 1971.

Cox D. R. and Lewis P. A. W. *The Statistical Analysis of a Series of Events*. Methuen & Company Limited. London, 1966.

Edwards A. W. F., *Likelihood*. Cambridge University Press., 1972.

Fisher R. A. *Statistical Methods for Research Workers*. Oliver and Boyd., Edinburgh: 1925.

Hacking I. *Logic of Statistical Inference*, Cambridge, University Press., 1965.

Hauer E. An application of the Bayesian approach to the estimation of safety countermeasure effectiveness, *Accid. Anal. & Prev.* 15, 287–298, 1983.

Jeffreys H. *Theory of Probability*, 3rd Edn. (p. 377). Oxford Clarendon Press, 1961.

Meehl P. E. Theoretical risk and tabular asterisks; Sir Karl, Sir Ronald and the slow progress of soft psychology. *J. Consulting and Clinical Psychology*, **46**, 806–834., 1978.