Accid. Anal. & Prev. Vol. 1, pp. 159–165. Pergamon Press 1969. Printed in Great Britain.

A PROBABILISTIC FRAMEWORK FOR ACCIDENT DATA ANALYSIS*

WILLIAM K. HALL

Highway Safety Research Institute, The University of Michigan, Ann Arbor, Michigan, U.S.A.

(*Received* 31 *January* 1969; *in revised form* 6 *April* 1969)

IN RECENT years researchers have become increasingly involved in the analysis of accident data in an attempt to make inferences on the processes of accident "causation." As a result of this interest, many statistical hypotheses have been formulated and examined using such data, and many elaborate projects have been undertaken to develop information collection and retrieval schemes based upon these data.

In many cases this increased activity has proceeded in an *ad hoc* manner. In part this is because of the complexity of this type of research. In addition, however, it is due to the fact that a general framework has not been utilized to organize and guide this research.

The purpose of this paper is to introduce such a general framework by applying the concepts of probability theory and statistics. A probabilistic approach is desirable because of the residual uncertainties which always remain when drawing conclusions from accident data analyses. This is due to the nature of the accident process, which is stochastic and not deterministic. In the first section of this paper concepts of probability theory are utilized to develop the basic analysis. Next, techniques for making inferences on these probability distributions are reviewed. Finally, various problems which arise when implementing this approach are briefly examined.

It should be recognized that many of the concepts developed here have been utilized (either explicitly or implicitly) in applied studies. However, the large number of inappropriate, and in many cases inaccurate, statistical analyses of accident data and the increasing interest in this type of analysis make an examination of the foundations appropriate at this time.

It should also be recognized that this article is introductory. The development of more sophisticated data analysis methodologies, and the development of better applications of such methodologies offer challenging, unsolved research problems.

PROBABILISTIC BASIS

For simplicity, assume that the only possible outcomes of the driving process are the occurrence of an accident (the event A) or the non-occurrence of an accident (the event $\bar{A}$). Associated with each of these events is a vector consisting of realizations of the n variables $(X_1, \ldots, X_n)$. One may then consider the probabilities associated with events defined in the sample space consisting of A and $\bar{A}$ as well as all possible realizations of the variables $(X_1, \ldots, X_n)$.

* This research was undertaken as a part of the Systems Analysis Research Program of the Highway Safety Research Institute.

Denote a vector-valued event defined from the variables $(X_1, \ldots, X_n(1))$ by the lower-case symbols $(x_1, \ldots, x_n)$. For instance, consider the variables X_1 = driver age and X_2 = year of vehicle. Then, *one possible* event defined from these variables is: (operator less than 20-years old and driving a 1967 vehicle).

It is also necessary to introduce the concepts of conditional probability into this discussion. To introduce the notation let $\Pr(B \mid C)$ denote the probability of the event B conditional upon the occurrence of event C. Either B or C may be a vector-valued event. For instance, B might consist of the r events $(b_1, b_2, \ldots, b_r)$.

The relevant probabilities in accident data analysis may then be written as:

$\Pr(A)$ = The probability of an accident (the event A). In most cases this probability will reflect the investigator's subjective and historical information as he starts a study—his prior information.

$\Pr(\bar{A})$ = The probability of the non-occurrence of an accident (the event $\bar{A}$). From probability theory it is evident that $\Pr(A) = 1 - \Pr(\bar{A})$.

$\Pr(x_1, x_2, \ldots, x_n \mid A)$ = The probability of the vector-valued event $(x_1, \ldots, x_n)$ conditional upon the occurrence of an accident. The variables this event is defined from may come from official accident reports, on-the-scene accident investigations, or other sources of accident data.

$\Pr(x_1, \ldots, x_n \mid \bar{A})$ = The probability of the same vector-valued event $(x_1, \ldots, x_n)$ conditional upon the non-occurrence of an accident. Essentially this probability represents the involvement of the event of interest in a non-accident control population.

The objective of accident data analysis is to examine these four probabilities to predict potential accident occurrence. That is, one wishes to determine $\Pr(A \mid x_1, \ldots, x_n)$, which is the probability of an accident conditional upon the observed vector-valued event $(x_1, \ldots, x_n)$.

This probability may be obtained by applying Bayes Theorem of probability theory (Parzen, 1960). Using this theorem one can calculate:

$$\Pr(A \mid x_1, \ldots, x_n) = \frac{\Pr(x_1, \ldots, x_n \mid A)\Pr(A)}{\Pr(x_1, \ldots, x_n \mid A)\Pr(A) + \Pr(x_1, \ldots, x_n \mid \bar{A})\Pr(\bar{A})}. \tag{1}$$

For example, suppose an investigator is interested in determining the influence of the single event "excessive speed" on traffic accident causation along a stretch of freeway. Denote the variable "speed" by S, and, assuming a 70 mph posted speed limit, define the event excessive speed by "$s>70$". The investigator's prior experiences with this roadway allow him to set the subjective, unconditional probability of an accident along this roadway at 0·03. Analysis of accident data leads him to infer that the probability of speed greater than the 70 mph in the accident population is 0·95. That is, 95 per cent of those in the accident population were traveling at speeds greater than 70 mph. By taking speed measurements along the freeway, the investigator makes the further inference that 25 per cent of the non-accident vehicles traveling the route are exceeding the speed limit.

For this example equation (1) becomes:

$$\Pr(A \mid s>70) = \frac{\Pr(s>70 \mid A)\Pr(A)}{\Pr(s>70 \mid A)\Pr(A) + \Pr(s>70 \mid \bar{A})\Pr(\bar{A})}$$

Substituting the above probabilities one obtains:

$$\Pr(A \mid s>70) = \frac{(0{\cdot}95)\,(0{\cdot}03)}{(0{\cdot}95)\,(0{\cdot}03)+(0{\cdot}25)\,(0{\cdot}97)} = \frac{0{\cdot}0285}{0{\cdot}2710} = 0{\cdot}105.$$

In addition to demonstrating the principles involved, this example points out a common fallacy in data analytic studies of this type. Even though the frequency of excessive speeding in the accident population is very high (0·95), the probability of an accident conditional upon this excessive speeding is still small (0·105). Consequently, analysis of the accident data alone would lead to erroneous inferences.

The large discrepancy between these two probabilities is easily resolved. The small prior probability of an accident (reflecting the fact that accidents are rare events) "deflates" the high probability of speeding in the accident population to yield the much smaller probability of an accident conditional on such excessive speeding.

At this point two questions should be considered:

(1) Why should the subjective probabilities of an accident be incorporated into the analysis when they are, in fact, intangible and highly variable?

(2) How sensitive is $\Pr(A \mid x_1, \ldots, x_n)$ to changes in the four component probabilities?

Both of these questions can be answered by transforming equation (1) to a much simpler and more illuminating form. To do this it is necessary to convert the probabilities of each event to the "odds" in favor of the occurrence of the event. Let

$\Omega(A)$ = Prior subjective odds in favor of accident occurrence. From probability theory these can be related to $\Pr(A)$ by the formula

$$\Omega(A) = \frac{\Pr(A)}{1-\Pr(A)} = \frac{\Pr(A)}{\Pr(\bar{A})}$$

$\Omega(A \mid x_1, \ldots, x_n)$ = Odds in favor of accident occurrence conditional upon the observed event $(x_1, \ldots, x_n)$. This is sometimes called the "posterior" odds and is related to $\Pr(A \mid x_1, \ldots, x_n)$ by the formula

$$\Omega(A \mid x_1, \ldots, x_n) = \frac{\Pr(A \mid x_1, \ldots, x_n)}{1-\Pr(A \mid x_1, \ldots, x_n)}.$$

Bayes Theorem (1) can then be converted to odds by straight-forward algebraic manipulation. The final result is:

$$\Omega(A \mid x_1, \ldots, x_n) = \frac{\Pr(x_1, \ldots, x_n \mid A)}{\Pr(x_1, \ldots, x_n \mid \bar{A})}\,\Omega(A). \tag{2}$$

The ratio $R = \dfrac{\Pr(x_1, \ldots, x_n \mid A)}{\Pr(x_1, \ldots, x_n \mid \bar{A})}$ is termed the probability ratio or overrepresentation

ratio of the event $(x_1, \ldots, x_n)$. Equation (2) states that the posterior odds in favor of accident occurrence, given the observed event $(x_1, \ldots, x_n)$, are equal to the product of the probability ratio and the prior odds in favor of accident occurrance. A probability ratio which is greater than one means that the prior odds are increased by the occurrance of the event.

In analyzing accident data using relation (2) above, the investigator is not necessarily forced to specify $\Omega(A)$, but only to examine the probability ratio. The sensitivity of $\Omega(A \mid x_1, \ldots, x_n)$ to changes in either the probability ratio or the prior odds is also obvious.

Returning to the earlier example one may write:

$$\Omega(A \mid s>70) = \frac{\Pr(s>70 \mid A)}{\Pr(s>70 \mid \bar{A})}\,\Omega(A) = \frac{0{\cdot}95}{0{\cdot}25}\,\Omega(A) = 3{\cdot}8\,\Omega(A).$$

Hence in this artificial example one may conclude that the event "excessive speeding" multiplies the prior odds in favor of an accident by a factor of 3·8.

STATISTICAL BASIS

In practice, the investigator rarely has complete information on the two conditional probabilities which comprise the probability ratio. Instead, he has samples from the accident and non-accident populations and wishes to base his conclusions on these. That is, he estimates the probability ratio by utilizing

$$\hat{R} = \frac{\hat{\Pr}(x_1, \ldots, x_n \mid A)}{\hat{\Pr}(x_1, \ldots, x_n \mid A)}$$

where the "^" notation denotes an estimator of a quantity. Inference procedures for developing such estimators are examined in this section.

The most common approach to developing inferences on probabilities of the above form assumes that the investigator has enough knowledge to specify the probability distribution up to a set of p unknown parameters $(\theta_1, \ldots, \theta_p)$. He then draws a sample of size n from this distribution and utilizes some estimation technique to obtain $(\hat{\theta}_1, \ldots, \hat{\theta}_p)$.

In the accident data analysis situation it is necessary to estimate both $\Pr(x_1, \ldots, x_n \mid A)$ and $\Pr(x_1, \ldots, x_n \mid \bar{A})$. For simplicity of notation, let $\mathbf{x} = (x_1, \ldots, x_n)$. If both relevant distributions have $\theta = (\theta_1, \ldots, \theta_p)$ as the set of unknown parameters, then the above probabilities may be written as $\Pr(\mathbf{x} \mid \theta, A)$ and $\Pr(\mathbf{x} \mid \theta, \bar{A})$. That is, the probabilities are also conditional on the values assumed by the unknown parameters.

Suppose an investigator draws a sample $(\mathbf{x}_1, \ldots, \mathbf{x}_\lambda)$* of size λ from the accident population and a sample $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ of size m from the non-accident population. He then estimates the parameters of the accident distribution by $\hat{\theta}_A = F(\mathbf{x}_1, \ldots, \mathbf{x}_\lambda)$ and estimates the parameters of the non-accident distribution by $\hat{\theta}_{\bar{A}} = F(\mathbf{x}_1, \ldots, \mathbf{x}_m)$. In both cases the function F is dependent upon the chosen estimation technique. Finally

* Recall that in general each member of the sample is itself a vector-valued event. That is $\mathbf{x}_i = (x_1, \ldots, x_n)_i$.

the desired probability estimators are derived as:

$$\hat{\Pr}(\mathbf{x} \mid A) = \Pr(\mathbf{x} \mid \hat{\theta}_A, A)$$
$$\hat{\Pr}(\mathbf{x}) \mid \bar{A}(1) = \Pr(\mathbf{x} \mid \hat{\theta}_{\bar{A}}, \bar{A}).$$

Alternately, one might adopt a Bayesian approach to this estimation problem. The investigator, when using this approach, thinks of the unknown parameters as subjective random variables. He then develops subjective probability distributions $g(\theta \mid A)$ over possible values of the parameters in the accident population and $g(\theta \mid \bar{A})$ over possible values in the non-accident population. The estimators then become:

$$\hat{\Pr}(\mathbf{x} \mid A) = \int \Pr(\mathbf{x} \mid \theta, A)\, g(\theta \mid A)\, d\theta$$
$$\hat{\Pr}(\mathbf{x} \mid \bar{A}) = \int \Pr(\mathbf{x} \mid \theta, \bar{A})\, g(\theta \mid \bar{A})\, d\theta.$$

Although this estimation procedure has many desirable properties, it will not be considered further in this paper.

Two examples may be useful in showing the applicability of the first estimation technique. Consider again the problem of predicting the effect of the event "excessive speed". Suppose the investigator has reason to believe that the distribution of speed in the accident population can be satisfactorily approximated by a normal random variable* with two unknown parameters, the mean μ_A and the variance σ^2_A. The density of this random variable can then be written:

$$f(s \mid \mu_A, \sigma^2_A, A) = \frac{1}{\sqrt{(2\pi)}\sigma_A} \exp \left[-\frac{1}{2\sigma^2_A} (s - \mu_A)^2 \right] \quad -\infty < s < \infty$$

and

$$\Pr(s > 70 \mid \mu_A, \sigma^2_A, A) = \int_{70}^{\infty} f(s \mid \mu_A, \sigma^2_A, A)\, ds.$$

Similarly the investigator assumes that speed in the non-accident population is also normally distributed with unknown mean $\mu_{\bar{A}}$ and variance $\sigma^2_{\bar{A}}$.

The investigator now randomly samples m vehicles in each of the populations to estimate the unknown parameters. Although there are many possible estimators for these parameters, the following are generally utilized because of their desirable properties:

$$\hat{\mu} = \frac{1}{m} \sum_{j=1}^{m} s_j$$

$$\hat{\sigma}^2 = \frac{1}{m-1} \sum_{j=1}^{m} (s_j - \hat{\mu})^2$$

where s_j is the jth sampled speed.

By substituting these estimators in the appropriate probability expressions, the investigator can then estimate the probability ratio:

* Clearly the choice of this probability model introduces some unrealistic assumptions, for a normally distributed random variable takes on all values over the real line. Nevertheless the popularity and flexibility of the normal model make it of interest in this example.

$$\hat{R} = \frac{\int_{70}^{\infty} f(s \mid \hat{\mu}_A, \hat{\sigma}^2_A, A)\, ds}{\int_{70}^{\infty} f(s \mid \hat{\mu}_{\bar{A}}, \hat{\sigma}^2_{\bar{A}}, \bar{A})\, ds}.$$

To illustrate the calculations assume the following estimates are obtained: $\hat{\mu}_A = 65$, $\hat{\sigma}^2_A = 16$, $\hat{\mu}_{\bar{A}} = 64$, $\hat{\sigma}^2_{\bar{A}} = 4$. Observe that the sample means are similar in the two populations, but that the sample variance in the accident population exceeds that in the non-accident population by a factor of four. From tables of the normal distribution one can quickly compute

$$\hat{R} = \frac{1-\Phi\left[\dfrac{s-\hat{\mu}_A}{\hat{\sigma}_A}\right]}{1-\Phi\left[\dfrac{s-\hat{\mu}_{\bar{A}}}{\hat{\sigma}_{\bar{A}}}\right]} = \frac{1-\Phi\left[\dfrac{70-65}{4}\right]}{1-\Phi\left[\dfrac{70-64}{2}\right]} = 10{\cdot}6$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} w^2\right) dw \text{ and } z = \frac{s-\hat{\mu}}{\sigma}.$$

Thus, in this example the prior odds in favor of an accident are multiplied by a factor of 10·6 when excessive speeding is observed.

As a second example suppose the investigator is unwilling to assume that speed is normally distributed in the two populations. Instead he wishes to estimate $\Omega(A \mid s>70)$ directly without specifying an underlying distribution of vehicle speed. In this case a binomial probability model is appropriate. Let

$Y_j = 1$ if $s_j > 70$ mph,

$Y_j = 0$ if $s_j \leq 70$ mph,

where again s_j is the jth sampled speed.

In this situation estimators of the desired probabilities can be derived directly by using the proportion of sample speeds over 70 mph. This estimator may be written as $\frac{1}{m} \sum_{j=1}^{m} Y_j$. To illustrate the calculations assume a random sample of size $m=500$ is taken in each population. The results show 70 vehicles exceeding 70 mph in the accident population and only 20 in the non-accident population. The estimate of $\hat{R}$ is then

$$\hat{R} = \frac{70/500}{20/500} = 3{\cdot}5.$$

It was stated earlier that a probability ratio greater than one means that the odds in favor of accident occurrence are increased given observations on the specified event. When the probability ratio is estimated by $\hat{R}$, it is then necessary to test the hypothesis that the true R is greater than one. Unfortunately the development of a statistical test

of this hypothesis is difficult, since the sampling distributions of these quantities are in general complicated. Consequently, informal and approximate tests will probably be utilized in assessing the likelihood that the probability ratio is actually greater than one.

SOME PROBLEMS IN IMPLEMENTATION

The investigator is confronted with four interrelated problems in applying these theoretical techniques to the analysis of real data:

(1) Selection of the specific variables $(X_1, \ldots, X_n)$ to include in the analysis. This selection must be based on the objectives of the analysis, the extent of data available from the two populations, and possible biases or errors in the data.

(2) Selection of the probabilistic models to utilize in the analysis. This selection is difficult because of the complex multivariate and interactive nature of the relevant variables and events defined from them. It will probably be necessary for the investigator to sacrifice some realism in his specification in order to obtain distributions which are tractable.

(3) Selection of a proper sample size. This decision must be based on the trade-off between increasing sampling cost and decreasing sampling error as the sample is enlarged. In general the complexity of this trade-off may make informal analysis necessary.

(4) Selection of the populations to sample. This problem is perhaps the most serious of the four mentioned, for it is essential to sample from accident and non-accident populations which are "comparable". Ideally the investigator desires populations which have the same "exposure to risk", but this concept is not yet well-defined or understood in accident research.* Consequently, approximations and judgment must be employed in population selection.

These problems must be solved before the techniques developed in this paper can be applied. Nevertheless, by developing certain approximate procedures it is possible to apply these methods to the analysis of real problems. This has been done in at least one case (Little and Hall, 1968). Furthermore, these probabilistic foundations are useful in conceptualizing and organizing research problems as well as in providing a means for logically and consistently analyzing accident data.

Acknowledgements—The author would like to thank Mr. Lyle Filkins and Mr. James O'Day of the Highway Safety Research Institute for their helpful discussions on this topic.

REFERENCES

JACOBS H. H. (1961). Conceptual and methodological problems in accident research. *Behavioral approaches to accident research*. Association for the Aid of Crippled Children, New York.

LITTLE J. W. and HALL W. K. (1968). *The association of accident frequency and severity with vehicle age*. Highway Safety Research Institute, Ann Arbor, Michigan.

PARZEN E. (1960). *Modern probability theory and its applications*. Wiley, New York.

RAPOPORT A. (1967). The alcoholic driver: A critique. *The prevention of highway injury*. Highway Safety Research Institute, Ann Arbor, Michigan.

* The interested reader might refer to Jacobs (1961) or Rapoport (1967) for a discussion of exposure.