IDENTIFICATION OF THE PROBLEM DRIVER FROM DRIVER RECORDS

A PRELIMINARY ANALYSIS

By WILLIAM L. CARLSON
Systems Analysis

Report SyA-1
January 1968

HIGHWAY SAFETY RESEARCH INSTITUTE • THE UNIVERSITY OF MICHIGAN

IDENTIFICATION OF THE PROBLEM DRIVER FROM DRIVER RECORDS

A PRELIMINARY ANALYSIS

INTRODUCTION

The characteristics that uniquely describe the problem driver (i.e., the driver who has a large number of accidents) must be determined before effective countermeasures can be instituted. Driving records are maintained by all states, and despite certain shortcomings beyond the control of the record-keepers, these records should not be overlooked as sources of vital predictive information. Data analysis, combining statistics, powerful computing hardware and software, and qualified scientific judgment can begin to determine what these large bodies of data have to tell us.

The objective of the overall research program--the first step of which is reported here--is to create a mathematical model by which groups of problem drivers can be identified from their driving records. Because of the complexities of characterizing human behavior, it is unlikely that the model will be able to predict individual problem drivers. However, if problem groups can be identified, then countermeasures can be concentrated where they should have the greatest effect.

PRELIMINARY FINDINGS

1. The average number of accidents per driver for all drivers over the 6.5-year period (January 1, 1961 - June 1, 1967) is 0.42. Interestingly, a study (2) conducted in California of 150,000 drivers during the years 1961 through 1963 reported an average number of accidents of 0.20 for the three-year period.

2. Groups of problem drivers who have a significantly greater number of accidents than the average accident rate for the entire group of drivers can be identified. This identification has been accomplished with only that data currently available in official driver files.

3. The most significant single identifier of these problem drivers is the total number of convictions. Those drivers who did not incur any convictions over the six and one-half year period of this study had an accident rate which was almost one-third that of the average driver in the State of Michigan.

4. There are significant interactions between the total number of convictions and other factors as reported by the Michigan Driver Record-Keeping System. For example, of the drivers with convictions, (a) male drivers have a higher average number of accidents than female drivers, (b) drivers in the age group 20-25 have more accidents, and, (c) chauffeurs or commercial drivers have a higher number of accidents than do persons having operator licenses. For the drivers who did not have any convictions, these differences did not occur.

5. This study should be extended to provide sharper tools for identifying groups of problem drivers. Two major types of expansion are needed:

a. An increase in the number of factors considered. This would allow us to refine our identification of groups of problem drivers. For example, we are confident that the inclusion of additional alcohol-involvement data would significantly improve the predictive capability of the model.

b. An increase in the number of drivers considered. This will allow a better cancelling of random factors not included in the study. It will also provide a means for both testing those hypotheses based on the present small sample and for generating new hypotheses.

DISCUSSION

The dependent factor chosen in this study was total number of accidents in the 6.5-year period from January 1, 1961 through June 1, 1967; this is the measure of driver quality used in this study. The Michigan State Police daily submit accident data to the Department of State. The data are taken from accident reports which are submitted to the state police by all of the reporting agencies in the state. Usually, the data are available within a few days to a month following the accident.

The independent factors from the Department of State driver files that were retained, following considerable preselection, are given below. Unfortunately, several types of convictions which we believe may be highly predictive, such as Reckless Driving and D.U.I.L., had to be discarded because of insufficient sample size.

1. Number of restrictions on driver licenses. The major categories under this grouping are those drivers having no restrictions and those drivers wearing glasses. Since this factor was subsequently found not to be a significant predictor, it can be assumed that the wearing of glasses corrected for any problems these drivers might have.
2. Sex
3. Age group, divided as follows:

Age Group 1	drivers under 20 years of age.
Age Group 2	drivers in the age range 20-25.
Age Group 3	" " " " " 26-35.
Age Group 4	" " " " " 36-45.
Age Group 5	" " " " " 46-55.
Age Group 6	" " " " " 56-65.
Age Group 7	" " " " " 66-75.
Age Group 8	" " " " " above 75.

4. Type of License. This could be divided into major groups: persons having operator licenses and persons having chauffeur licenses (i.e., commercial drivers).
5. Origin of License. This can be divided into two major categories, new licenses and renewal licenses. Those persons having new licenses did not show up as having a higher average number of accidents. However, since our measure of number of accidents extended over a six and one-half year period, it would be expected that persons having new licenses would not have as many accidents because they would be only represented for a maximum of three years; the other drivers would have been represented for the entire six and one-half years.

6. Driver Education. An indication of whether or not the person had had driver education.
7. Total convictions since 1961.
8. Speeding convictions since 1961.

Convictions are defined as follows: The convicting court in the State of Michigan sends a record of motor vehicle offense convictions to the Department of State. Any arrests that do not result in convictions are not recorded in the driver records files. This study used total convictions over the entire period of January 1, 1961, to June 1, 1967. Total speeding convictions were added in an attempt to differentiate between the more serious aggressive types of offenses and more routine convictions such as driving without a license, faulty equipment, and so forth.

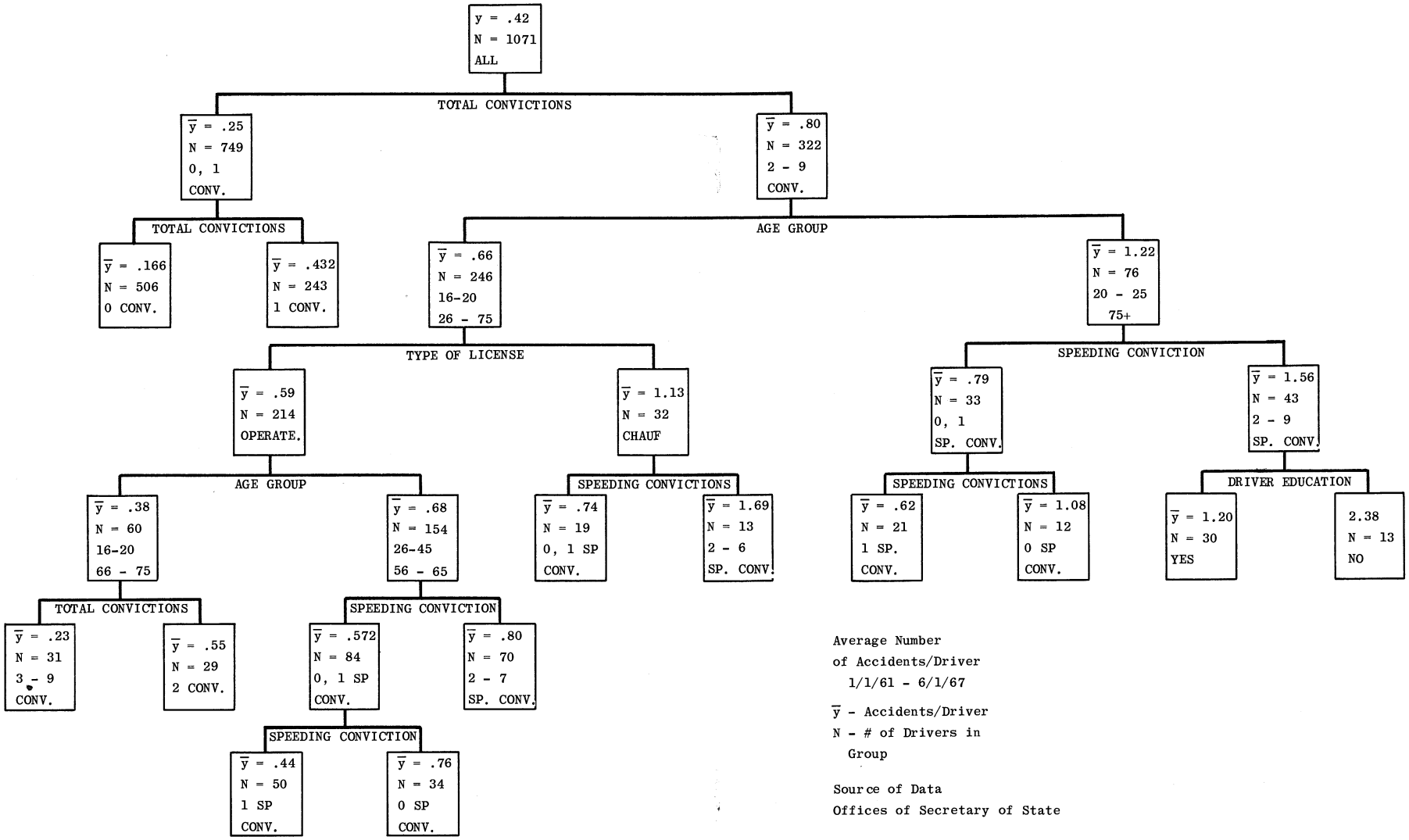
ANALYTICAL PROCEDURE

The first task was to identify the factors which best predict drivers who have accidents. Our initial approach was to divide the data into a number of subgroups based on the significant predictor factors. There are a number of different ways in which the data could have been split depending upon the levels of these different factors. For example, taking the factor sex, we could split the data into male and female groups. We could then further subdivide each of these groups into groups based upon, for example, age group or number of speeding convictions. Obviously, by following this procedure to its ultimate and using all eight independent factors, we would end up with a large number of different groups to compare, one against the other. In the simplest case, if we make a two-way split on each of the eight independent factors, we would end up with 256 different groups. It seems reasonable to assume that a large number of these groups would not be significantly different from the total population. Therefore, our real task is to sort out those groups that are truly different from the other groups. In other words, determine which combinations of these independent factor levels identify groups of people who are problem drivers.

The Institute for Social Research at The University of Michigan has developed a computer program titled the Automatic Interaction Detector (AID) which identifies and differentiates according to average dependent factor response (3). This large and powerful program automatically tests all possible groups which can be generated on the basis of the various combinations of the independent factor levels. The program selects from this vast number those groups which are truly different from all of the other groups; this indicates it will allow us to sort out groups of problem drivers. The algorithm splits the total sample into successive subgroups, using first the most important predictor. It then further splits these subgroups using other predictors in order of their relative importance. This technique was used to gain knowledge from the data concerning the characteristics of problem drivers.

RESULTS

The results of this phase of the analysis are shown in Figure 1. The most significant factor found to explain number of accidents was



Average Number
of Accidents/Driver
1/1/61 - 6/1/67

$\bar{y}$ - Accidents/Driver
N - # of Drivers in
Group

Source of Data
Offices of Secretary of State

FIGURE 1.

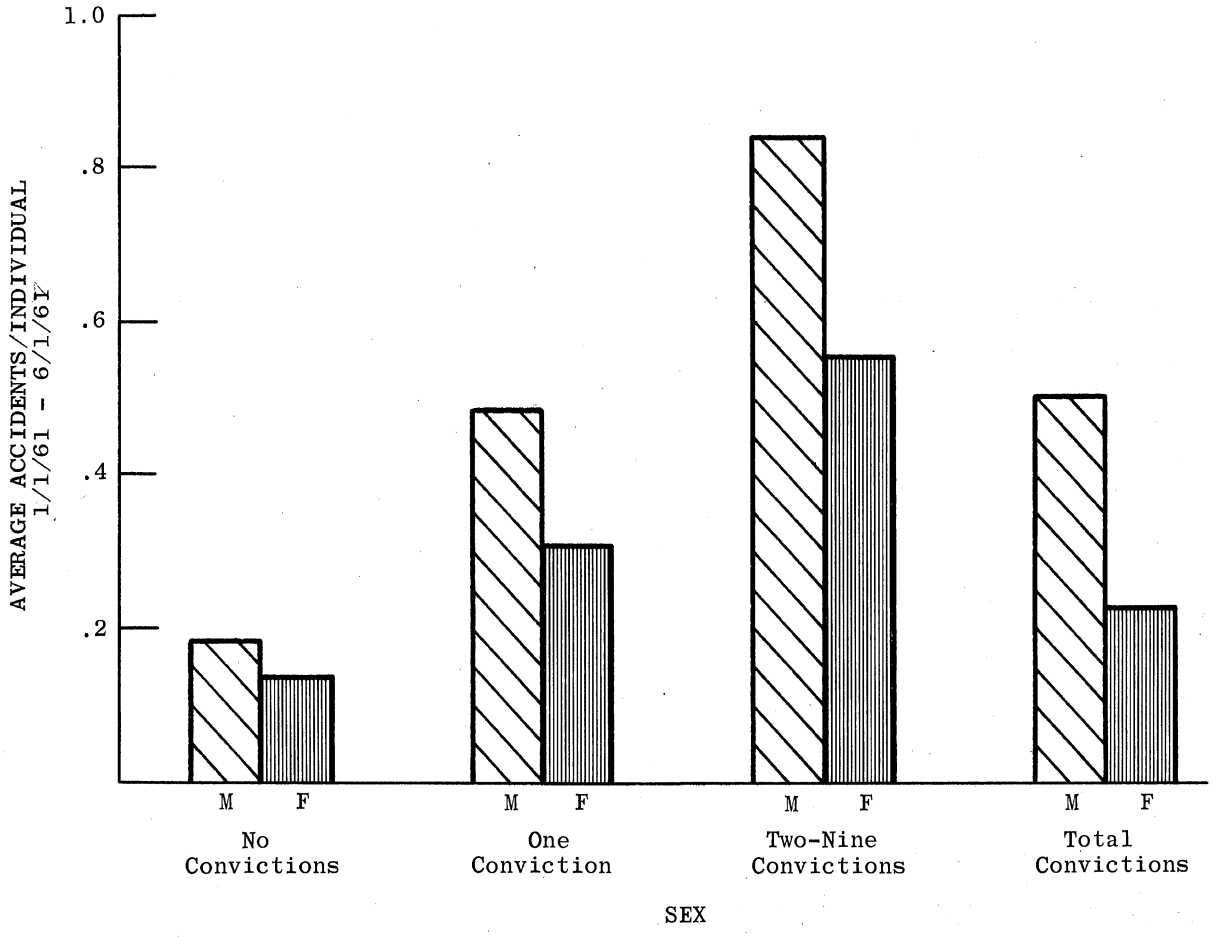


FIGURE 2. ACCIDENTS OF MEN VS WOMEN

total convictions. As shown, the total sample of 1071, having an average of 0.42 accidents per driver, was divided into:

1. A group of 749 drivers with 0 or 1 convictions and an average of 0.25 accidents per driver, and,
2. A group of 322 drivers with 2 or more convictions and an average of 0.80 accidents per driver, a value over three times that of the first group.

Each of these groups was further split by convictions and by age group, as shown in Figure 1. The former showed the further effect of even one conviction for predicting accidents. In addition, of the drivers having more than one conviction, those in age group 20-25 had more accidents than those in other age groups. Further study of Figure 1 indicates the splitting of other groups having a high number of accidents. The fact that most of the splitting occurred in the high conviction group indicates a definite interaction between convictions and other factors such as license type and age group.

Figures 2, 3, and 4 further depict this phenomenon. Each of these graphs shows average accidents per individual on the ordinate. Figure 2 indicates how the average number of accidents differs for men and for women. For the group with no convictions, both men and women have fewer accidents than does the population as a whole. In addition, there is little difference between the average number of accidents for men and for women. However, as convictions increase, the differences between men and women widen.

This same tendency exists with regard to age group (Figure 3) and type of license (Figure 4). Drivers with no convictions show a low number of accidents regardless of their age group. The difference in number of accidents between operators and chauffeurs, as shown in Figure 4, does not occur if we look at only those drivers with 0 convictions.

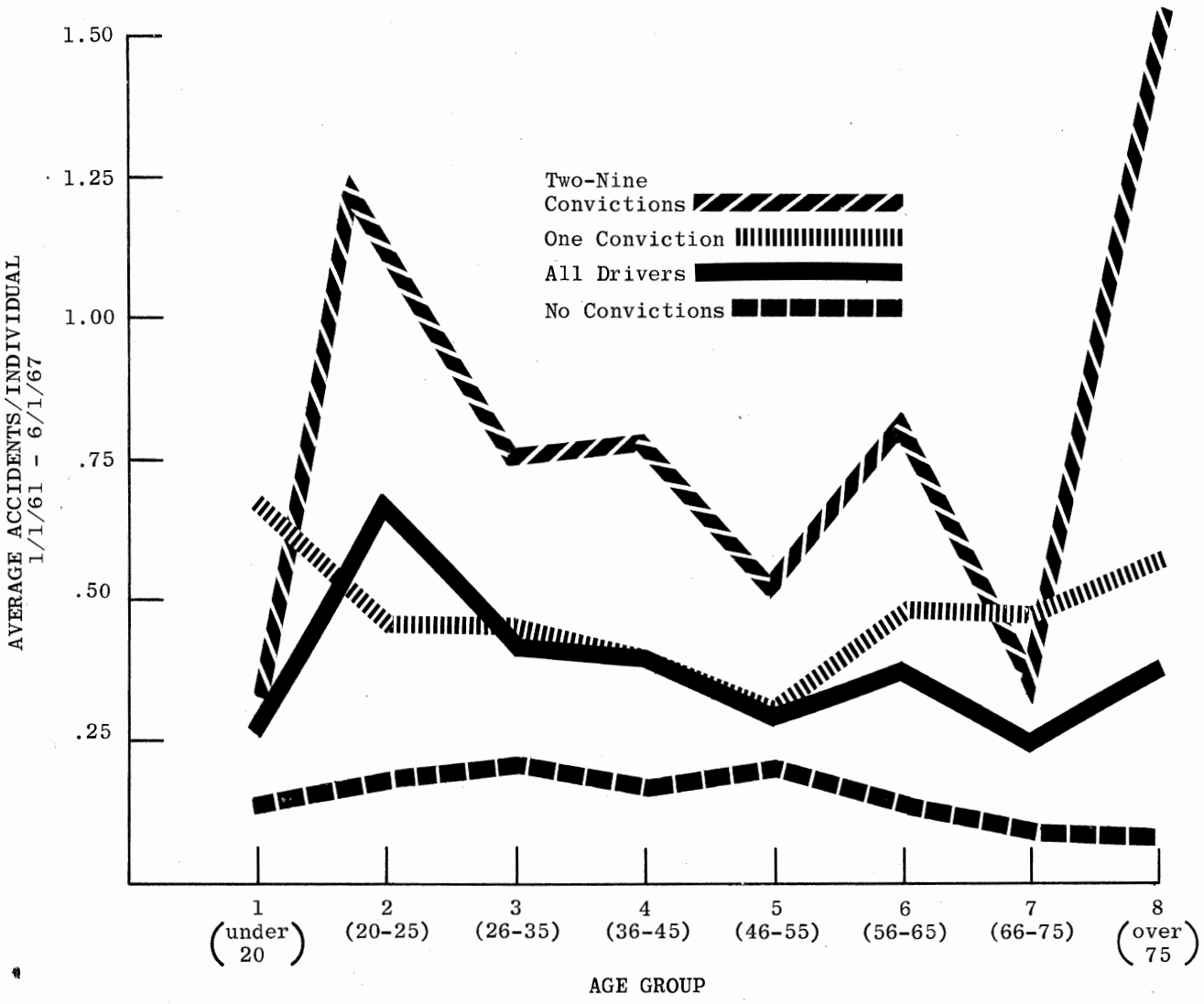


FIGURE 3. EFFECT OF AGE GROUP ON ACCIDENTS

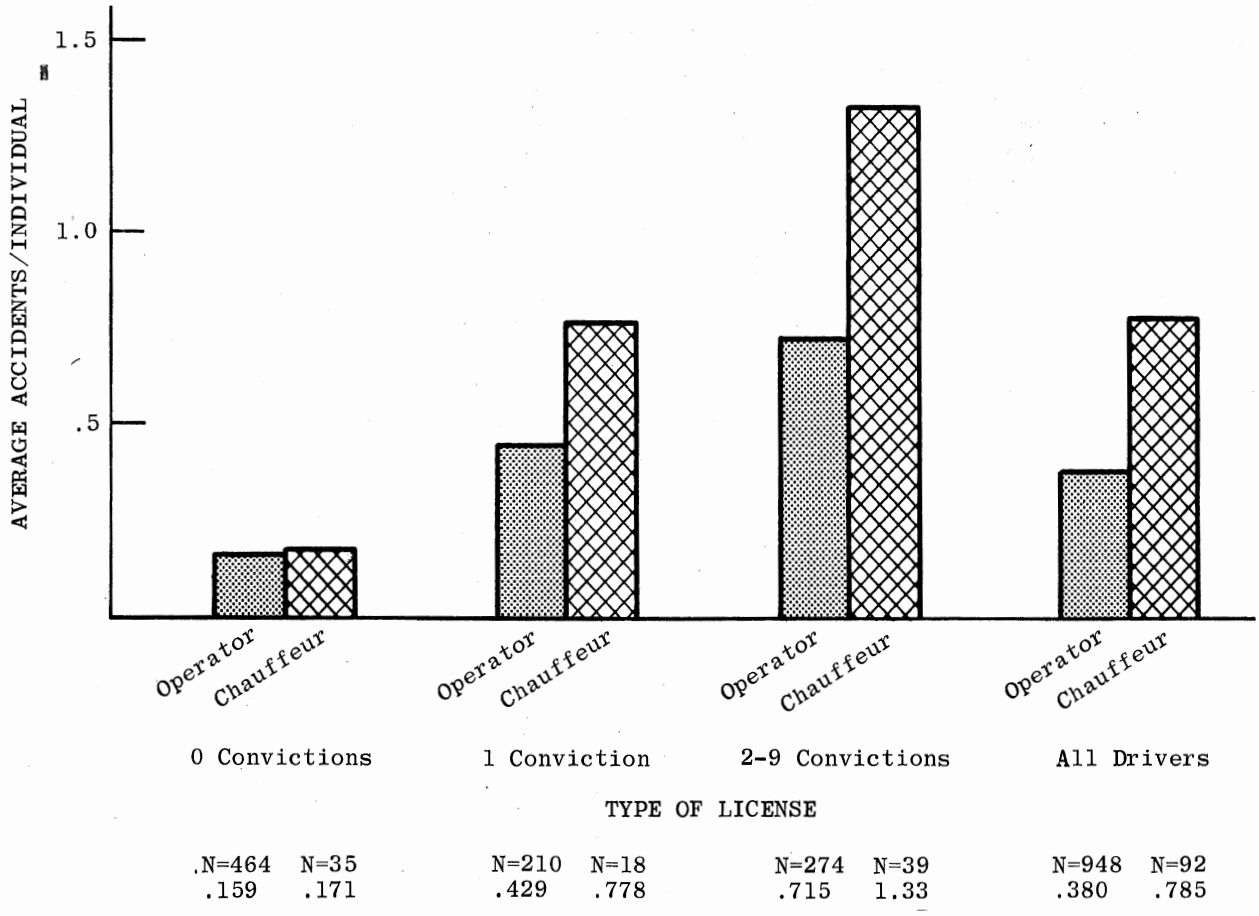


FIGURE 4. EFFECT OF LICENSE TYPE ON ACCIDENTS

This would suggest the existence of a group of "good" drivers (i.e., no convictions) who are much less involved in accidents than is the driver population as a whole. Of the remaining drivers, certain groups have many more accidents than does the population as a whole. It would seem reasonable that our efforts should thus be concentrated on identifying the "problem" drivers among those drivers having convictions. If these groups can be made small enough, individual attention can be given to these drivers.

MODEL EXTENSION AND VALIDATION

As a result of the initial analysis of the data, we propose to construct a mathematical model which will predict the average number of accidents for drivers falling into various groups. Using this mathematical model, we should then be able to rank drivers based upon the expectation of their having any given average number of accidents. These drivers should then be investigated further.

Once this model has been constructed, it will be necessary to validate it. A first measure of effectiveness of the model can be based upon such statistical criteria as the multiple correlation coefficient, T tests of the individual factor coefficients, and so forth. However, it is much more important to ask "Is the model truly doing the job that it is supposed to be doing?" In this case, since the model is supposed to single out groups of bad drivers, we wish to determine how well the model actually predicts groups of bad drivers. The first step in the validation process will be to apply the model to the group of data which was used to develop it, and thus, to determine how well the model predicts average number of accidents over the range of the various factors contained in it. The results of this analysis may suggest necessary modifications of our model. Once this procedure has been completed, model validation must be extended to independent sample groups. It would seem obvious that the model will make better predictions for the group of drivers from which it was developed than for the independent sample groups; however, if the model is to have general use, it must be able to predict problem drivers from any groups of drivers to which it is applied. In other words, it must literally predict future events as well as "predicting" historically observed results.

REFERENCES

1. J. W. Little, Highway Safety Research Institute, The University of Michigan, "The Michigan Driver Profile," Unpublished Report.

2. R. S. Coppin, The 1964 California Driver Record Study, Part 2, page 19, Department of Motor Vehicles, State of California.

3. J. A. Sonquist and J. N. Morgan, The Detection of Interaction Effects, Monograph #35, Survey Research Center, Institute for Social Research, The University of Michigan, Ann Arbor, 1964.