

ACCIDENT RESEARCH MANUAL

Forrest M. Council
Donald W. Reinfurt
B. J. Campbell
Frank L. Roediger
Carol L. Carroll
Amitabh K. Dutt
Janet R. Dunham

University of North Carolina
Highway Safety Research Center
CTP 197A
Chapel Hill, NC 27514

January 1980

UNC/HSRC-80/1/1

Final Report



Document is available to the U.S. public through the
National Technical Information Service,
Springfield, Virginia 22161

Prepared For
U. S. Department of Transportation
Federal Highway Administration
Washington, D.C.

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle ACCIDENT RESEARCH MANUAL		5. Report Date January 1980	
		6. Performing Organization Code	
7. Author(s) Council, F.M., Reinfurt, D.W., Campbell, B.J., Roediger, F.L., Carroll, C.L., Dutt, A.K., & Dunham, J.R.		8. Performing Organization Report No.	
9. Performing Organization Name and Address University of North Carolina Highway Safety Research Center CTP 197A Chapel Hill, NC 27514		10. Work Unit No. (TRAIS)	
		11. Contract or Grant No. DOT-FH-11-9424	
12. Sponsoring Agency Name and Address U.S. Department of Transportation Federal Highway Administration 400 Seventh Street, S.W. Washington, D.C. 20590		13. Type of Report and Period Covered Final Report	
		14. Sponsoring Agency Code	
15. Supplementary Notes			
<p>16. Abstract Included in this manual is a compilation of sound research techniques that can be used by the engineer/analyst to carry out research related to highway accidents. Because highway engineering administrators must daily decide how best to spend limited numbers of safety dollars, they need to have results from properly conducted and clearly presented accident research for inputs in this decision-making process. This manual was prepared to meet the continuing need for upgraded research, both in the area of analysis of relationships between accidents and other variables and in the area of countermeasure evaluation. The manual is designed for use by the engineer/analyst who has some background in statistical analysis.</p> <p>The manual contains material related to 1) the rationale and the need for improving the level of existing research, 2) the underlying issues that researchers must be familiar with, 3) the components and methodologies used in the two basic types of accident research--research aimed at evaluating countermeasures and research aimed at identifying and examining underlying relationships between accidents and other highway factors, 4) the preparation and distribution of research results, and 5) summary guidelines for the engineer/analyst to use in his research. The manual has been developed for use in classroom training, as a reference text, and/or in a self-study program. Review questions, self-study pre- and post-tests, and references to related texts and articles are included.</p>			
17. Key Words Countermeasure Evaluation Statistical Research Highway Accident Research Accident Data Analysis Traffic Records Research Report Preparation		18. Distribution Statement Document is available to the U.S. public through the National Technical Information Service, Springfield, Virginia 22161.	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages iv, 144	22. Price

NOTICE

This document is disseminated under the sponsorship of the Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

TABLE OF CONTENTS

	Page
ACKNOWLEDGMENTS	1v
CHAPTER I THE MANUAL: WHY IT WAS WRITTEN AND WHAT IT IS	
1.1 Introduction to the Problem	1
1.2 Purpose of the Manual	4
1.3 Target Audience	4
1.4 Orientation of the Remainder of the Manual	4
1.5 References	6
SELF STUDY PRE-TEST	7
CHAPTER II ISSUES IN ACCIDENT RESEARCH	
2.1 Introduction	9
2.1.1 What this manual is about	9
2.1.2 What this manual is not about	9
2.1.3 Where accident data come from	10
2.1.4 Problems with the data: A discouraging word	12
2.1.5 An encouraging word	12
2.2 Problems and Issues in Use of Accident Data	12
2.2.1 Problems and issues in accident data collection or accumulation	12
2.2.1a Unreported or inconsistent data	12
2.2.1b Reported but biased data	15
2.2.1c Methods for reducing bias	17
2.2.2 Problems and issues related to the nature of accidents	19
2.2.3 Problems and issues in exposure data	21
2.2.3a What is the need for exposure data?	21
2.2.3b Sources of exposure data	21
2.2.3c Problems in exposure data	22
2.3 Summary	22
2.4 Review Questions	22
2.5 References	23
CHAPTER III EVALUATING COUNTERMEASURES	
3.1 Introduction	25
3.2 Definition of Countermeasures or Modifiable Highway Elements	26
3.3 What is Evaluation	26
3.3.1 Administrative (process) evaluation	26
3.3.2 Effectiveness evaluation	27
3.3.3 Why carry out effectiveness evaluation?	27
3.4 Limits on the Success of Effectiveness Evaluation	28
3.5 Components of Effectiveness Evaluation	29
3.5.1 Attitude	29
3.5.2 The basic idea of cause and effect	29
3.5.3 Determination of what to measure	31
3.5.3a Accidents as the criterion	31
3.5.3b Crash severity as criterion	32
3.5.3c Intermediate measures as criterion	35
3.6 Threats to the Validity of Effectiveness Evaluation	38
3.6.1 History (other causes at the same time)	39
3.6.2 Maturation (trends over time)	39
3.6.3 Regression artifacts	39
3.6.4 Instability	41

TABLE OF CONTENTS (cont.)

	Page
3.7 Common Evaluation Designs Used in Overcoming Threats to Validity	41
3.7.1 Evaluation of single-treatment programs	42
3.7.1a Before/After design	42
3.7.1b Before/After with randomized control groups	45
3.7.1c Before/After with comparison group	48
3.7.1d Interrupted time-series designs	52
3.7.1e Time series with comparison groups	54
3.7.1f Time series with comparison variables	55
3.7.1g Time series with switching replications	55
3.7.1h "Tie-breaking" designs	56
3.7.1i Regression discontinuity design	56
3.7.2 Evaluations involving multiple degrees or types of treatment	57
3.7.3 Evaluation of the "multiple" improvement countermeasures	59
3.8 Statistical Procedures for Evaluating Countermeasures	60
3.8.1 Glossary of terms	60
3.8.2 The importance of Type I and Type II errors	62
3.8.3 Sampling considerations and sample size determination	64
3.8.4 Choice of appropriate statistical test	68
for Poisson frequencies	72
Paired t-test	73
Z-test for proportions	74
Mantel-Haenszel	75
GENCAT	75
ECTA or CONTAB	75
F-Test	76
Kilnogorov-Smirnov Test	77
RIDIT	78
Student's t-test	78
ANOVA	79
Analysis of covariance	79
Median test	80
Mann-Whitney U-Test	80
Bartlett's test	81
3.9 Use of Evaluation Results in Cost/Benefit Analysis	82
3.9.1 Cost/benefit methodology definition	82
3.9.2 Possible algorithms	82
3.10 Review Questions	83
3.11 References	84
 CHAPTER IV IDENTIFYING RELATIONSHIPS AMONG VARIABLES	
4.1 Introduction	87
4.2 Analysis Issues Related to Research Involving Relationships	87
4.2.1 Sampling considerations: Total population versus sample	91
4.2.1a Estimating required sample size	92
4.2.1b Choosing a representative sample	93

TABLE OF CONTENTS (cont.)

	Page
4.2.2 Choice of dependent variable	94
4.2.2a Accidents as the dependent variable	95
4.2.2b Crash severity as the dependent variable	96
4.2.2c Intermediate measures as the dependent criterion	97
4.3 Analysis Techniques	97
4.3.1 Introduction to statistical analysis	98
4.3.2 Variable screening procedures	99
4.3.2a Simple presence of association between two continuous variables	99
4.3.2b Simple presence of association between two categorical variables	103
4.3.3 Relative weight or strengths of relationships--model development	108
4.3.3a Model development when all variables are continuous	108
4.3.3b Model development when all variables are not continuous	115
4.4 Summary	116
4.5 Review Questions	116
4.6 References	117
4.7 Computer Programs	119

CHAPTER V THE FINAL STEP: PREPARATION AND DISTRIBUTION OF RESEARCH RESULTS

5.1 Introduction	121
5.2 Preparation of Reports	122
5.2.1 Report preparation keys	123
5.2.2 Suggested report preparation sequence	123
5.3 Distribution of Results	125
5.3.1 Distribution of results in short article form	125
5.3.2 Presentation of an oral report	126
5.4 Summary	127
5.5 Review Questions	127
5.6 References	127

CHAPTER VI SUMMARY

Guidelines for the Accident Researcher	129
Self Study Aid	133
Closure	133

SELF STUDY POST-TEST	135
--------------------------------	-----

APPENDIX A STANDARD STATISTICAL TABLES

A.1 t-distribution for 1-tail tests	A-2
A.2 t-distribution for 2-tail tests	A-3
A.3 z-distribution for 1 and 2 tail tests	A-5
A.4 χ^2 -distribution for 2-tail tests	A-4
A.5 D for the Kolmogorov-Smirnov 2-sample test (two-tail and one-tail tests)	A-5

APPENDIX B INTRODUCTION TO STATISTICAL TESTING

ACKNOWLEDGMENTS

As is noted in the title page, a number of different HSRC staff members contributed to the manual and are cited as principal authors. However, these individuals are by no means the only people whose contributions were necessary to the preparation of this document.

Primary in this regard were the FHWA personnel who conceived the project, guided the authors through the developmental stage, and provided inputs and revision suggestions that helped make the manual more effective. For these invaluable inputs, the authors wish to express their thanks first and foremost to Mr. David Solomon, Chief of the Environmental Design and Control Division, who was a very conscientious and (thankfully) very understanding Contract Technical Manager. In addition to Mr. Solomon, the other FHWA staff members who contributed to the review and revision process were Harry Lum, Julie Fee, Stan Byington, Charles Horton, and Phyllis Young. The authors also wish to acknowledge the FHWA staff members and the researchers from various organizations and state agencies who participated in the pilot workshop presented as part of this contract; they provided very valuable revision suggestions for the manual and for the proposed training process involving this document.

In addition, the authors are very grateful to the HSRC staff whose efforts made the manual possible. These include Cranaine Brinkhous, Lauren Ogle, and Bill Pope, who helped provide the graphic materials used, and William Hunter, who reviewed the initial draft. Lindsay Griffin of the Texas Transportation Institute also reviewed a draft of the manual and provided helpful suggestions.

Finally, but certainly not least, heartfelt thanks go to Donna Suttles, Peggy James, Teresa Parks, and Martha Apple, the typists who were unfortunately responsible for deciphering poor handwriting and even poorer dictation tapes from many different sources. Without their expertise and continued support and help, the project could not have been completed. For the fact that it is now completed, we are all most grateful.

CHAPTER I

THE MANUAL: WHY IT WAS WRITTEN AND WHAT IT IS

Situation: The traffic engineer of a large eastern state holds a series of staff meetings in which his division engineers from across the state provide documentation concerning numerous highway safety problems and their proposed solutions. In most cases, the analysis and solution are based on engineering judgment following visits to each of the problem sites.

The state engineer's own accident investigation and research unit identifies a number of high-accident locations which need to be corrected and proposes solutions (countermeasures) ranging from edge line delineation to total intersection redesign. The members of the accident investigation and research unit assure him that their proposed solutions are based on the results of before/after accident studies they have conducted in the past.

He has recently been contacted by sales representatives from various companies selling crash cushions, breakaway supports, and innovative traffic control devices. Each salesman has assured him that their respective devices have been tested and shown to reduce accidents and injuries.

The Planning and Research Division, a companion division in the state highway department, calls to say that they have recently completed a laboratory study which proved that the use of larger letters on warning signs is significantly more effective in drawing the attention of the drivers tested in the lab. Thus, they want him to implement the larger letter program on a three-county basis in order to measure the exact effect on crashes.

The FHWA Division Office calls to ask why he has not spent all his categorical safety funds in the areas of railroad grade crossings and edge marking.

In addition, he receives the usual daily quota of calls from legislators, irate parents, and PTA presidents concerning the implementation of potential safety projects at specific locations in their towns.

Finally, his boss, the State Highway Administrator calls to say the overall traffic engineering budget will be reduced by 8 percent in the upcoming fiscal year because a tax reform referendum has made it necessary to cut back on funding to all state government agencies. His Administrator wants to know how much of his safety funds he can give up in excess of 8 percent.

Result: The traffic engineer resigns and joins a private consulting firm at an increased salary.

Main Chapter Topics

- Introduction to the Problem
- Purpose of the Manual
- Target Audience
- Orientation of the Remainder of the Manual

1.1 Introduction to the Problem

Although this hypothetical situation is exaggerated, it may not be too far removed from the current situation that traffic safety administrators must face. Traffic engineers, research engineers, highway program administrators on the federal, state, and local levels, and other administrators, researchers, and implementers involved in the area of program management are daily faced with the task of making the decisions concerning how best to spend limited numbers of safety dollars. Although the decision-making process for such decisions includes various inputs ranging from political consideration to budget constraints, the most important input to the conscientious safety administrator is the relative effectiveness of each available countermeasure in terms of its potential for reducing the frequency or severity of crashes.

or for maintaining the same level of safety while increasing the flow of traffic. The decisions are much more critical today as the demands on our transportation system expand at a greater rate than the resources devoted to insuring the system's safety.

Because of the complexity of accidents, highway administrators are increasingly forced to also consider factors related to the driver and vehicle. For example, the design of guardrails and crash cushions has been complicated by increases in truck size and decreases in average car weight: these devices now need to be strong enough to protect the trucks, yet soft enough to accommodate the lighter cars.

As changes such as this occur in the demands placed on the transportation system, a related change must also occur in administrators' awareness of what designs can transport people and goods safely. It is primarily for this reason--this need to increase knowledge for use in decision-making--that research in the area of highway safety is needed.

Unfortunately, although an impressive number of highway safety research studies have been conducted, many are inadequate because of erroneous conclusions or the absence of conclusive evidence. In 1970, Solomon, Starr, and Weingarten reviewed research and evaluation studies that analyzed 57 highway-oriented countermeasures. The authors felt that they had found "good to excellent" estimates of effectiveness for only eight of the 57 countermeasures. For the remaining 49 countermeasures, effectiveness estimates were "... based either on engineering judgment, involved only fair or poor data, or were little more than guesses."

Since then, the situation has improved somewhat, but recent surveys of research efforts have continued to find deficiencies in countermeasure evaluations. Hunter, et al., (1977) reviewed numerous research reports to develop estimates of the effectiveness of roadside countermeasures such as breakaway sign supports, guardrail placement and modification, bridge or crash attenuation systems and other hardware. Although Hunter, et al., compiled "best guess" estimates of effectiveness (see Table 1.1), they noted that great deficiencies existed in the effectiveness evaluations they reviewed:

"... The fact that the estimates of effectiveness are not more specifically defined is a major roadway safety issue. There is a continuing very serious need for more well-designed effectiveness evaluations of fixed object treatments ... there is a scarcity of good evaluations concerning fixed object improvement programs. Where such evaluations exist, they generally are the before/after type with no control group and thus are subject to accident fluctuations, regression to the mean, and other artifacts."

In addition to problems in the methodology used in many of these studies, part of the existing deficiency can also be related to problems inherent in the primary variable being studied--the traffic accident.

A major problem is that most individual treatments can be realistically expected to reduce only a small proportion of the accidents that occur (i.e., each treatment has a relatively low overall level of effectiveness). The exception to this is the complete redesign of a highway to upgrade it to Interstate standards, a treatment limited in use because of the cost involved. Furthermore, accidents are almost random occurrences that, for the most part, do not occur in large numbers at a given site. Because of these difficulties, attempts have been made to use measures besides accidents to assess treatment effectiveness, but the use of these proxy or surrogate measures has caused a great deal of controversy and created many problems. Although there are times when such substitute measures are appropriate or even necessary for safety evaluations, operational measures such as speed, traffic conflicts, passing maneuvers, etc., must be related to crashes in order to be acceptable to the general public and many decision makers. Accidents are not the only indicator of the operational efficiency of a roadway system, but the "political" situation dictates that the surrogate measures must also be directly related to what is thought of as safety (i.e., to crash frequency or severity) in order to be acceptable substitutes.

Because of this emphasis on accidents as the acceptable measure of interest among decision-makers, accident-oriented research will continue to be of greater interest than research involving surrogate measures. However, the research results currently available cannot always provide administrators with the information they need to make decisions.

Table 1.1 Estimated effectiveness of various roadside countermeasures.

Hazard	Treatment	% Reduction			
		Fatal (%)	Injury (%)	PDO (%)	
1. Utility poles	a. Breakaway	30	-1 ¹	0	
	b. Relocate - 30' from edge of pavement	32	-1.7	0	
	c. Remove	38	-1.5	0	
2. Trees	Remove	50	25	-20	
3. Exposed bridge rail ends	Transition Guardrail	55	20	-50	
4. Substandard bridge rail	Improved rail (thrie beam)	15	5	-3	
5. Underpasses (Bridge piers)	a. Concrete median barrier with end treatment	60	40	-150	
	b. Attenuators				
	1. Water filled cushion	75	60	-300	
	2. Sand filled cell	75	60	-300	
	3. Steel Barrels	75	60	-300	
6. Rigid signs or supports					
	a. Small sign	Breakaway	70	25	-12
	b. Large metal support	Breakaway	60	20	-20
	c. Large metal support	Relocate behind guardrail	55	30	-5
	d. All supports combined	Breakaway	68	24	-14
7. Guardrail ends	a. Breakaway cable terminal	55	25	-15	
	b. Turned down Texas terminal	55	25	-15	
8. Median-involved accidents					
	a. Narrow median	Concrete median barrier	90	10	-10
	b. Wider median	Double faced guardrail	75	2	-28

¹Minus sign indicates an increase in the proportion of accidents.

Source: Hunter, et al. (1977), pp. 14-16

1.2 Purpose of the Manual

This manual has been prepared in an attempt to help overcome this dilemma. The material included in the following chapters presents a detailed discussion of the methods which can and should be used in highway-related accident research and includes: the underlying rationale for these methods; the problems and solutions associated with the implementation of the methods; and the related statistical tools which indicate the strength of a relationship to aid in a final decision concerning the effectiveness of a given countermeasure. This manual, however, is not a statistics text. Statistical analyses are an integral part of accident research, but they are only one part. This manual is aimed at the more general questions involved in 1) specifying a given problem in workable terms, 2) establishing a research design in order to insure that the problem can be answered, 3) implementing the design in terms of collection of data, 4) analyzing the data itself, and 5) presenting and distributing the research results to other individuals in the field.

1.3 Target Audience

The manual is intended primarily for research engineers who are or will be involved in highway accident research. It is further assumed that the primary users of the manual have a high degree of analytic capability frequently associated with a degree in engineering or in a related field and will have completed a comprehensive course in applied statistics. Thus, a basic understanding of statistical terminology and methods is assumed. However, if the manual user does not have this background, the required knowledge may be gained by studying the materials referenced at the end of each chapter. Also, although the manual is primarily aimed at the highway engineering aspect of the safety system and the examples and situations used throughout the manual are very closely tied to this area, the concepts are also valid for accident-related safety research concerning the vehicle or the driver.

The manual has been developed primarily for use (1) in classroom training, (2) as a reference text, and/or (3) in a self-study program. First, the manual can be used as supplemental material for a series of classroom lectures (the manual was field-tested in one such workshop/classroom series). Second, the manual can be used as a reference tool by practicing researchers when certain research problems require a firmer knowledge of underlying principles or a more detailed knowledge of the solution to a specific problem. Finally, the manual can also be used for self-study when classroom lectures are not readily available (self-study is actually the key to the other uses of the manual as well).

To facilitate this self-study use, there are questions at the end of each chapter to measure users' understanding of the material in that chapter. At the end of this first chapter is a short pre-test that surveys the material covered in the entire manual. The reader should take this test as an exercise in self-evaluation so that he can be aware of the areas to which he needs to devote his attention.

1.4 Orientation of the Remainder of the Manual

The remaining five chapters of the manual contain information that will hopefully help fulfill the needs described above (Figure 1.1 presents the general flow of information and topics to be covered in these chapters): Chapter 2 presents various underlying issues which a researcher must be familiar with; Chapters 3 and 4 present the components and methodologies used in the two basic types of accident research--research aimed at evaluating countermeasures and research aimed at identifying and examining underlying relationships between accidents and other highway factors; Chapter 5 presents information about preparing and distributing the results of this research; and Chapter 6 summarizes the key points covered in the manual and provides a self-study post-test whose questions are keyed to the relevant manual pages.

In addition to the self-study questions at the end of each chapter, references are cited throughout the manual both as examples of research and as sources of additional information about a given subject area. For convenience, the references cited in each chapter are listed at the end of that chapter.

This manual was prepared to meet the need for better highway safety research. In reality, however, the key to meeting this need is not the manual, but the user of the manual--the accident researcher.

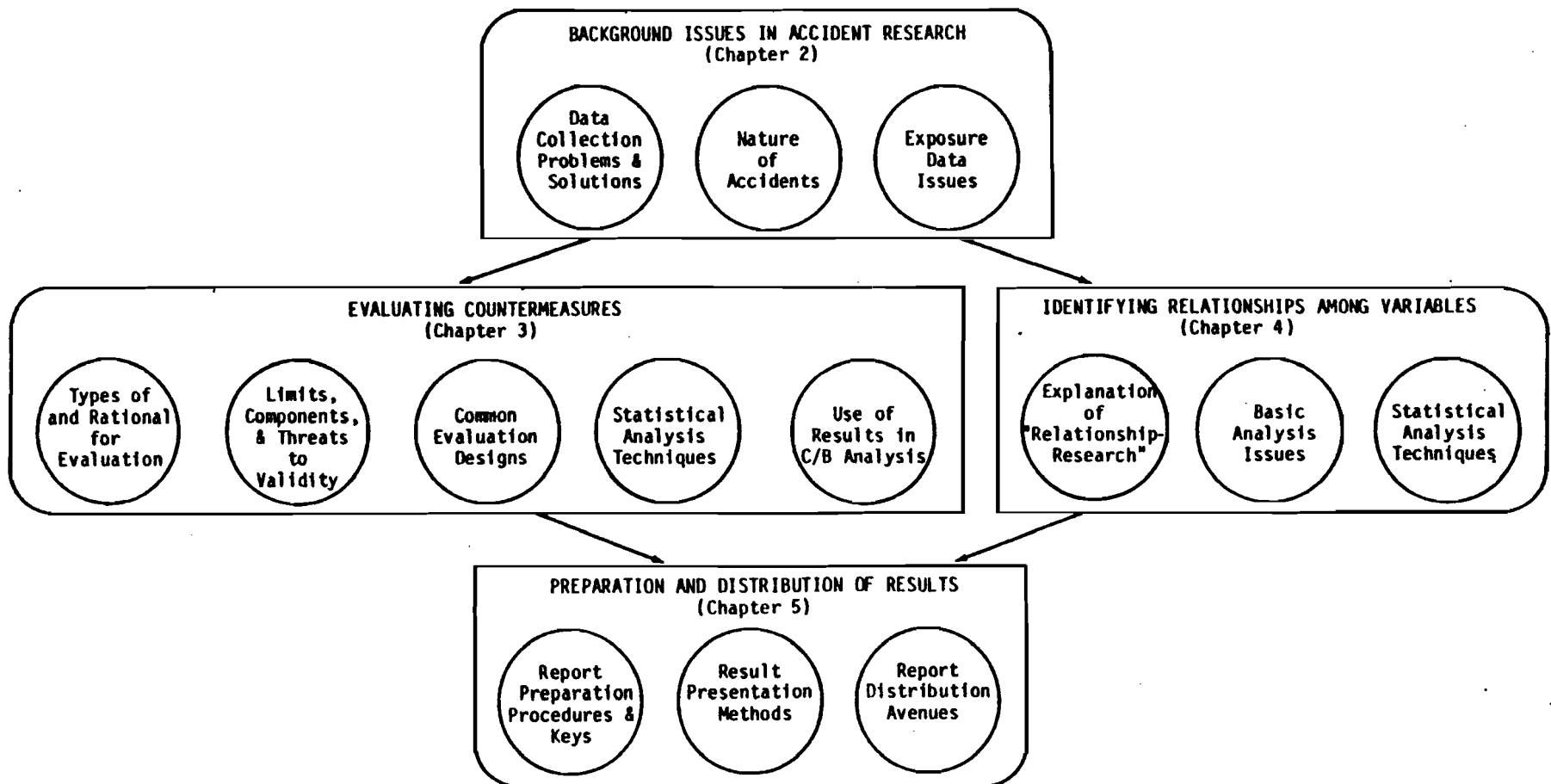


Figure 1.1. Topics included in Chapters 2 - 5 of this manual.

1.5 References

Hunter, W. W., Council, F. M., & Dutt, A. K. Project selection for roadside hazards elimination. Final report. (Vol. 1) Chapel Hill: University of North Carolina Highway Safety Research Center, 1977.

Solomon, D., Starr, S., & Weingarten, H. Quantitative analysis of safety efforts of the Federal Highway Administration. Washington, D.C.: Federal Highway Administration, 1970.

Self Study Pre-Test

1. Describe three causes of potential biases that may be present in a given accident data base of which the researcher should be aware.
2. What is exposure data and why is it so important in accident research? List three existing sources of mileage exposure data.
3. How is a representative sample of a population selected?
4. In some evaluations of countermeasures, a substitute measure (proxy measure) will be used as the criterion in place of accidents. List the two attributes that an acceptable proxy measure must possess.
5. A researcher is interested in ascertaining the relationship between variables which may not be linearly related. What type of analysis should she employ?
6. The people of New Hebrides have decided that lice produce good health since all their healthy tribesmen have lice and none of the sick ones do. The tribe statistician has calculated a high correlation between the number of lice and degree of health. Briefly discuss this correlation in terms of cause-effect.
7. What is the basic question the evaluator should ask in determining what should be measured (i.e., in determining the criterion variable) in an evaluation?
8. A before/after study has indicated that the placement of concrete median barriers has increased accident frequencies on freeways. How can such a treatment still be justified?
9. When a change is detected in any evaluation of a highway countermeasure, there are many possible causes including the treatment itself. List the four main rival explanations for a given change, other than the treatment.
10. There are various types of evaluation designs (e.g. Before/After, control group designs, time series, etc.). What is the basic reason that a researcher would apply a sound design?
11. Which study design would be appropriate to evaluate a law reducing speed limits on all freeways to 55 mph?
12. In budgeting for the coming fiscal year a highway engineering dept. has a set operating improvement budget and the results of the evaluations of three proposed improvements. Which if any of the following improvements should the department make? All cost the same amount.

	<u>α</u>	<u>Calculated values</u>	<u>d.f.</u>	<u>Critical values</u>
Improved Pavement Delineation	.05	$t = .997$	10	$t_c = 1.8$
Breakaway Poles	.05	$\chi^2 = 3.22$	1	$\chi_c^2 = 3.84$
A New Attenuation System	.05	$\chi^2 = 2.49$	1	$\chi_c^2 = 3.84$

13. Due to large increases in Labor Day weekend traffic, police officers in state A decide to report only those crashes that involve personal injury to the occupants of in-state vehicles. How can this practice affect a study of the relationship between accidents and traffic volume?
14. A state traffic engineer is requested by the FHWA to collect accident and highway characteristics data on a sample of sections of Interstate highway. Because the purpose of the study is to predict accident rates based on highway characteristics, the engineer samples those locations which have experienced one or more accidents in the past year. Comment briefly on the adequacy of this sample.

15. A researcher is interested in developing a relationship between some measure of safety and the feet of guardrail per mile, the number of breakaway and non-breakaway telephone poles per mile, and the number of protected bridge piers per mile. What would be an appropriate dependent (predicted) variable to be used in the model?
16. While many statistical tests exist for analyzing data collected in an evaluation, the choice of most appropriate test basically depends on three factors. These are:
 - a. The evaluation design used
 - b.
 - c.
17. (a) To the highway engineer with little money to expend which type of error is more acceptable, Type I or Type II? Why?

(b) What about the researcher attempting to find an effective countermeasure for an important problem area in which no good treatments exist? Explain your reason.
18. A researcher is evaluating the effectiveness of water-filled crash attenuation devices. The devices have been placed in gore areas of arterials which carry heavy commuter traffic involving car-pooling. A comparison group of locations has been chosen from rural freeways experiencing similar ADT's. Would total number of serious occupant injuries or total occupant deaths be appropriate criteria for the evaluation?
19. While many sequences could be followed in the preparation (writing) of a research report, two steps which are often neglected but strongly recommended are
 - a)
 - b)
20. A number of avenues for distribution of highway-related research reports are available to the researcher. Four of these are:
 - 1) Distribution through FHWA
 - 2)
 - 3)
 - 4)

CHAPTER II

ISSUES IN ACCIDENT RESEARCH

***Situation:** A group of engineers are told by their director to attend a one-week workshop concerning accident research. Each of these engineers has some limited statistical training and some familiarity with past accident studies, and each is to be assigned to a newly formed accident research unit which will both conduct internal research and monitor outside research funded by the home agency. The text to be used in the workshop is a new accident research manual. Upon reaching the workshop site, the engineers discover that the instructors are non-engineering researchers from a large university. After introducing the topic of research by pointing out a large series of poor studies (conducted, incidentally, by engineers) the instructors spend the remaining six hours of day one presenting problem after problem which can hinder the chances of conducting a successful research study. Very few solutions are presented.*

***Result:** In the evening, the engineers discuss the day's material among themselves, make a group decision, and call in sick the following morning. (The manual is donated to the local paper recycling program.) [NOTE: THIS IS A PURELY HYPOTHETICAL SITUATION.]*

Main Chapter Topics

- Introduction
- Problems and Issues in Accident Data
 - Data collection or accumulation
 - The nature of accidents
 - Exposure data

2.1 Introduction

2.1.1 What this manual is about.

This manual considers the kind of research that requires the compilation of large numbers of accidents so that statistical analysis techniques can be properly applied to arrive at sound conclusions. In general, the manual does not consider research based on analyses of a limited number of on-scene investigations. However, if adequate samples of such on-scene investigations can be collected or accumulated, both the problems and solutions cited in this manual can pertain.

For example, a typical problem intersection may be identified based on five or more accidents in one year. To improve the situation, the traffic engineer may analyze these accidents and the intersection itself. Such a site-specific accident analysis is not within the scope of this manual.

Accident research of the kind addressed here is usually undertaken based on a specific need resulting from (1) the necessity of identifying and defining the components of a specific safety problem, (2) an accident-reducing countermeasure program or device to be evaluated, or (3) the necessity for studying the interrelationship among a number of variables thought to be relevant to accidents.

Research on actual accidents is desirable because the relationship between various "causal" variables and accidents is not usually clear. Indeed, the history of accident research has shown very often that the relationship between some seemingly valid safety countermeasure or common-sense factor and the end result, accidents, is very difficult to establish. Thus, it is highly desirable to measure actual accidents that occur as the most practical, hard-nosed way of determining whether a program or a countermeasure or a new accident-reducing system is in fact effective and to determine the true form of relationships between accidents and other variables of interest.

2.1.2 What this manual is not about.

This manual will not specifically address research such as test track studies, staged crashes, and mathematical or full scale simulation. Obviously, statistical analysis of actual accidents is not the only valid way to do research related to the safety of the highway; there are even times when it is not necessarily the best way. One instance of safety research in which the use of accident data is not feasible is when accident data may be too crude (it may not be possible to collect sufficient accident data to allow

specific small details of the accident which are really relevant to the question to be extracted). A recent example is a study conducted by Systems Technology, Inc. for the Federal Highway Administration, that examined the aerodynamic effects of large trucks on other vehicles (Weir, et al., in press). This study sought to determine whether allowing larger trucks on the road would detrimentally affect the safety of surrounding vehicles by producing aerodynamic effects which might cause a passing vehicle to deviate from a safe path. One way to study such a question is to allow larger trucks on the roadway, collect accident data on each accident in the immediate vicinity of a large truck, and then ferret out the details of whether or not the aerodynamic effects of the trucks played a part in the accident. Unfortunately, this type of detail is usually not available from the accident report forms that researchers normally use. In many cases, the large truck would not even be at the scene of the accident since it would not be directly involved in the crash. In other cases, even if the accident-involved driver noted that he was "blinded by the spray" or "blown off the road" such statements might not be recorded by the investigating officer or, even if they were recorded, would probably not be computerized for later retrieval.

Thus, for a question in which very specific details are needed to investigate the problem of interest, available accident data may indeed be too crude an indicator to use. In this type of situation, it would be quite relevant to employ a study involving test track and wind tunnel simulation (this was the approach used by Weir, et al.).

A second situation in which non-accident data would be quite appropriate are crash test studies conducted to determine whether or not new developments in roadside hardware increase safety. Such studies represent a large part of the highway research literature in the past decade (Bronstad, et al., 1974; Field and Prysock, 1965; Hayes, et al., 1971; Martinez, 1971).

For example, one way of studying new crash attenuation systems would be to actually install the attenuation systems on existing highways, wait until crashes occur, and then study the results of the crashes in terms of occupant injury. However, if there is some question concerning whether or not the system is indeed safer than what is already on the roadway, it would obviously be better to pretest the system with simulation or full-scale tests in order to determine whether it results in lower collision forces to vehicles. This staged-crash research also will not be covered in the manual.

However, both of the above examples of research are carried out under the assumption that the effects measured are ultimately related to the safety of the roadway. That is, if splash and spray are greater for larger trucks, then this situation should ultimately result in a change in accidents that could be demonstrated if pertinent data from a large enough sample of the proper accidents could be collected. Similarly, changes or decreases in g-forces to the crash test vehicle are also assumed to be related ultimately to occupant injuries in actual crashes. For this reason, particularly in the second example where the crash tests are used, it is always important to follow up such pre-testing with actual on-road accident research.

A third area in which accident data are not used is research involving proxy or surrogate measures as substitutes for accident variables. There are times when, because of a lack of sufficient time to collect an adequate sample size of accident data, or because there is a need for an intermediate measure of effectiveness before the end of the project, it is necessary to conduct a study which involves a surrogate or proxy measure instead of accidents as the outcome criteria. Despite the apparent differences between these two approaches, the problems, the solutions, and the methodologies related to use of proxy measures are similar to those related to the use of accidents. Therefore, this type of research falls within the scope of this manual.

Readers should also be aware that measures not related to accidents are also used extensively in non-accident studies (i.e., decisions concerning lane and shoulder width criteria, bridge clearance, sight distance and passing zone criteria, and many other aspects of roadway design are based on studies of speed, vehicle placement, passing behavior and other non-accident surrogate measures). Obviously, accident research is only one facet of the total research picture which concerns the highway system. However, because this manual is specifically directed toward accident research, the use of surrogate measures in non-accident studies will not be covered.

2.1.3 Where accident data come from.

The raw material of accident research is the accident reports on file in a given jurisdiction. Normally, these are obtained from the standard accident report forms filled out by the police officers in

that jurisdiction (see Figure 2.1). Some of the information recorded on that form is then coded into computerized format. (Regrettably, it has historically been the case that some valuable information from the form is typically not transferred to computer.) Once on computer, large samples of the compiled accidents can be analyzed relatively easily. However, there will continue to be research questions which can only be answered with the raw data, the actual written forms. These cases usually occur when the computerized (coded) data are incomplete. For example, in most states, neither the sketch nor the narrative provided by the investigating officer is computerized. If this is the case, then a question involving the distance from the roadway of the sign support struck in ran-off-road collisions cannot be answered without manual reference to the original forms. Researchers need to understand that information that appears to be unavailable on a coded file may actually be available. Unfortunately, "computer-power" must be replaced with "researcher-power" in these instances. Familiarity with the investigator's basic form is often the key to answering difficult research questions.

Figure 2.1 North Carolina Accident Report Form.

A second major source of accident data, one that is very plentiful in many states, is the driver report. Usually the driver is required to fill out an accident report himself (in addition to any police report). In some states the driver report is a principal source of information since police reports are not necessarily filed for every accident. However, the driver report may be filled out in a self-serving way because the driver may fear being penalized by the state or his insurance company. It is because of this lack of objectivity that the researcher is urged to use the police report data where available.

A third source of accident information, not found as frequently as the above two sources, is accident data collected by the researchers themselves. Although this manual does not specifically apply to on-scene accident investigation per se, if enough on-scene investigations are conducted, the data from them can be compiled into a data base which could be analyzed using the methods described in this manual.

A final source of data, and one which may become increasingly useful and available to researchers studying highway safety problems, is national data compiled by either FHWA or NHTSA. Current examples include the National Accident Sampling System (NASS) and the Fatal Accident Reporting System (FARS). In each of these systems, accident data are collected from a number of states on a common report form and are computerized and made available to government and private researchers, and the general public. In the NASS system, the data are collected by special accident investigation teams located across the nation; the FARS data are coded from accident, vehicle registration, and driver files in a number of different state locations and are merged into the central system at NHTSA.

2.1.4 Problems with the data: a discouraging word.

This whole chapter is designed to list various warnings about using accident data. One or more of the problems related to data collection or accumulation (e.g., the low probability of an accident occurring at a given location or in a given short period of time, the lack of consistency in exposure data, etc.) will be encountered in almost every study. By the end of this chapter the reader may be inclined to throw up his hands and say, "Why even try to use the accident data?"

The answer to that question is that, even with all the inherent problems, accident data remain the most acceptable indicator of whether or not the ultimate goal of safer travel is met. While other measures of safety are continually being developed, advocated, and tested, the rationale for safety funding is the reduction in crash frequency or severity. Because these direct measures are available, and because the ultimate user of all research, the decision maker, is "biased" in favor of these bottom-line measures, accident data should and, in all probability, will continue to be the data of primary interest. Accident research will continue to involve accident data. Countermeasure programs need to be rigorously evaluated to be absolutely sure that society's resources are being expended on programs that really work. The history of the highway safety field is filled with examples of well-intentioned costly programs that seemed like a good idea, but do not actually work. Such programs soak up resources that could be used on effective programs that save lives and reduce injury severity and property damage.

2.1.5 An encouraging word.

The problems discussed in this chapter can perhaps be better characterized as issues of which any good researcher must be aware. Just as any good administrator in any program must be aware of the strengths and weaknesses of his staff, his material, and his product, the researcher needs to be aware of the basic strengths and weaknesses of the data--his basic material--if he is to produce the best product possible. Probably the most common research error is the failure to realize that some unforeseen data characteristic is distorting conclusions by warping analyses in one direction. The researcher must be healthily skeptical of his data, and must be alert for ways in which the data can mislead him.

Although hints and guidelines for overcoming such problems are included in this manual, there is no substitute for a questioning attitude toward the data and making sure that the data really signify what they seem to indicate. All in all, accident data can frequently be used with success with proper planning and knowledgeable, yet skeptical, interpretation.

2.2 Problems and Issues in the Use of Accident Data

The following text will discuss problems and issues which are relevant to the three areas of accident-related highway research: 1) the collection of the data, 2) the basic nature of the accident data, and 3) the collection and use of exposure data, the necessary companion to accident data.

2.2.1 Problems and issues in accident data collection or accumulation.

Perhaps the major issue which the researcher must face is data inadequacies or problems related to the data collected. Barriers to good data collection arise from both planned and actual collection procedures, biases inherent among the data collectors, and continual changes in the collection mechanism.

2.2.1a Unreported or inconsistent data.

(1) Inconsistent data due to reporting thresholds. The accident cases in a given official file do not by any means comprise all the accidents that have occurred in that area: many minor collisions are not reported. In fact, no official attempt is made to collect information on all collisions.

Almost every jurisdiction has a reporting threshold so that accidents are officially reported only if they involve some degree of injury (including death) or, in the absence of injury, a specified amount (in terms of dollars) of property damage (see Table 2.1). It may well be that for every reported accident there are three or four unreported minor mishaps.

Thus, one must consider whether the threshold has changed during the period covered by the research. If a threshold is raised, there may be a downturn in reported accidents immediately thereafter. This indicates nothing more than the threshold change, but could be mistaken as an "improvement" in the accident picture. An example of such a change is an increase in the dollar threshold because of the impact of inflation on auto repair costs.

In an attempt to arrive at an objective threshold, some federally sponsored accident data collection systems are defining the reporting threshold in terms of "towaway" crashes. These are crashes producing vehicle damage severe enough that the vehicle cannot safely leave the scene under its own power. Instead, a tow truck is called. Such a criteria might be thought to be more objective because: (1) the definition of a towaway accident is more objective and less susceptible to inflationary changes than an estimate of dollar damage, (2) if a vehicle is disabled, it will be more likely to remain at the scene long enough for the officer to have an opportunity to thoroughly investigate and completely report the accident, and (3) accidents in which vehicles can easily leave the scene are usually minor and the loss of such cases is less significant than the loss of major ones. (The loss of "low damage" cases can, however, cause problems in the evaluation of crash attenuation systems: if the system works properly, many potential injury accidents will become non-towaways and will go unreported.)

Nevertheless, there is a problem in using the towaway threshold because the likelihood that an accident-involved vehicle will need towing depends on what part of the vehicle is struck. Imagine identical impacts on a series of vehicles starting with the center front (defined as a 12 o'clock impact) and going "around the clock" through the right side, rear, left side, etc. One can readily imagine that identical "strikes" on various parts of the car will have a different likelihood of rendering the vehicle a towaway. A blow on the right front fender that crumples sheet metal onto the tire may render the car inoperable. That same blow on the right passenger door may well leave the vehicle operable, yet that blow might have high injury potential if someone is seated in the right front seat.

Suffice it to say, the researcher should always consider the nature of the reporting threshold and consider whether the threshold rule equally affects all the variables at issue in the study in question and whether the threshold has changed during the study period.

A second threshold concern is the issue of determining when a delayed death is actually a traffic fatality. Traditionally, the definition of a fatal traffic accident has included delayed deaths that occurred within one year of the crash date. There is now a move to change this. The American National Standards Institute has recently approved a 90-day rule. In contrast NHTSA and FHWA have chosen to use a 30-day rule in publishing data on fatalities. The reason for this issue, of course, is the combined factor of late reporting and late death as a factor in the accident toll vs the desire to "close the books" as soon as practical at the end of a year. A certain number of days always pass after the end of a calendar year before the traffic toll is "settled."

A related problem, although not associated with a threshold change per se, is biases which might result from changes in the reporting forms during the course of the evaluation period. Such changes, although they appear innocuous, may result in rather drastic changes in the reporting of a certain data item. A real-world example of this situation occurred when a city reorganized the box on its accident report form concerning driver violations. In that box, six traffic violations were listed and the investigating officer could check the appropriate one. Speeding was listed first. When the form was redesigned, the officials rearranged the order in which the violations were listed and moved speeding to the fourth position in the list. Officials were startled to find that the number of indicated speeding violations in crashes declined sharply. At first they thought they had reduced speeding greatly. However, what had actually happened was that the officers had glanced down the form and checked the first item that seemed logical. In other words, the change in the number of indicated speeding violations did not signify an improvement in the control of speeding. (It is also noted that some other violation probably increased due to being placed first on the list!) Again, the point stressed is that the researcher must be aware of such changes in the data collection forms.

Table 2.1 Accident reporting threshold levels requiring police reports by state.

State	Dollar Amount of Property Damage			
	\$50	\$100	\$200	Other
Alabama	X			
Alaska				\$500
Arizona				\$300
Arkansas		X		
California				Inj.
Colorado				All
Connecticut				\$250
Delaware				\$250
District of Columbia	- No information -			
Florida	X			
Georgia	X			
Hawaii	- No information -			
Idaho	X			
Illinois				All
Indiana			X	
Iowa				\$250
Kansas			X	
Kentucky				Upon Request
Louisiana		X		
Maine			X	
Maryland				All
Massachusetts			X	
Michigan			X	
Minnesota		X		
Mississippi	X			
Missouri				Fatals
Montana		X		
Nebraska				\$250
Nevada				\$250
New Hampshire				\$300
New Jersey			X	
New Mexico		X		
New York				Inj.
North Carolina			X	
North Dakota				\$300
Ohio				All
Oklahoma		X		
Oregon	- No information -			
Pennsylvania				Towaways
Rhode Island	- No information -			
South Carolina	X			
South Dakota				\$250
Tennessee			X	
Texas				Inop. veh.
Utah			X	
Vermont				All
Virginia		X		
Washington				\$300
West Virginia				All
Wisconsin			X	
Wyoming				\$250

Source: Unpublished information provided by Bureau of Operations and Research, International Association of Chiefs of Police, Gaithersburg, MD, 1979.

(2) Inconsistent reporting due to failure to investigate. In some situations, a legally reportable accident is not reported (e.g., when police agencies have heavy criminal investigation duties and are not able to dispatch an officer to the accident scene). While a North Carolina study (House, Waller, and Koch; 1974) indicated that 89 percent of the crashes reported to an insurance company were found on the official Department of Motor Vehicles file, a study of motorcycle crashes in North Dakota (1979) indicated that only 47 percent were on file. Thus, the problem may vary from jurisdiction to jurisdiction and perhaps even according to vehicle types. And this non-reporting can become a significant problem if the occasions of non-reporting are not random in nature.

For example, it is noted that in some major cities, freeway accidents that occur during rush hour are not reported unless they are severe enough to cause injury or the disablement of a vehicle. This policy is followed for the simple reason that during rush hour, disastrous traffic jams can occur if accident-involved vehicles are held at the scene for investigation instead of being quickly removed. Because of this inconsistency in reporting accidents, accident records might inaccurately indicate that freeways are safer in rush hour than at any other time.

(3) Inconsistencies due to cross jurisdictional differences or differences in forms. A researcher must know whether all accident reporting agencies represented in the sample he is studying follow the same reporting rules. An investigator using records made up of city and county jurisdictions that use different report forms or follow different criteria in reporting or storing the data may be in difficulty without ever realizing there is a problem.

Also, there may be certain jurisdictions where the police investigate an accident only if personal injury is involved. If injury is not involved, then the driver's personal report may be the only one available for compilation. Where that situation exists, the combining of injury and non-injury reports may be inappropriate.

2.2.1b Reported but biased data.

In addition to the problems inherent in accident data collection due to the reporting issues described above, accident research data may also be compromised by another aspect of the collection process--the presence of incomplete or biased data. In contrast to the above cases, where the data are not reported or coded, this biased data issue exists even though the data have been reported and coded. Indeed, whereas non-reporting of data or the inconsistencies between reporting mechanisms can sometimes be identified by the researcher through a survey of the formal collection policies, the biases now being discussed are much more subtle and therefore much more difficult to detect. Quite often, they result from the informal "working procedures" used by individual investigating officers rather than from more formal prescribed procedures documented in a manual. The following are examples of problems that can arise by virtue of incomplete data incorrectly reported data, or some kind of statistical bias in the data.

Example 1. A problem existing in many states results from the failure of officers to milepost accidents properly, particularly when a state does not use an accident location system in the field. Frequently officers may merely estimate the distance from an accident site to the nearby mileposted feature. In some cases, officers too frequently round off the distance estimate to convenient distances (e.g., .1 miles, .5 miles, 1.0 miles, 2.0 miles) and the resulting mileposted values are in error (see Figure 2.2).

If distances were always being measured in increments of .1 miles from mileposted benchmarks (such as nearby intersections) in the aggregated statewide data, one would expect a crash to occur 0.5 mile from all benchmarks only one-tenth of the time. In states where the roadway system is a one mile square grid system with an intersection each mile, one would expect a uniform distribution with each tenth equally represented when the entire state is analyzed. In a state where the benchmarks are randomly spaced, the distribution of distances should be somewhat triangular shaped with the .1 and .2 distances outweighing the .5 and 1.0. The fact is that accidents are reported by officers at one-tenth of a mile and one mile from the benchmark many times more often than three tenths, six-tenths, etc. This makes the data appear to say that "one of the most dangerous place is one mile from somewhere." This, of course, indicates that the actual distance in the mileposting data is suspect. While the problem may be minimized by use of a physical mileposting system, it may well continue to exist if actual measurements are not made to the standard benchmark. In this way, the failure to correctly milepost an accident means, for example, that the roadway characteristic computer system cannot accurately associate an accident with the proper location or proper characteristics.

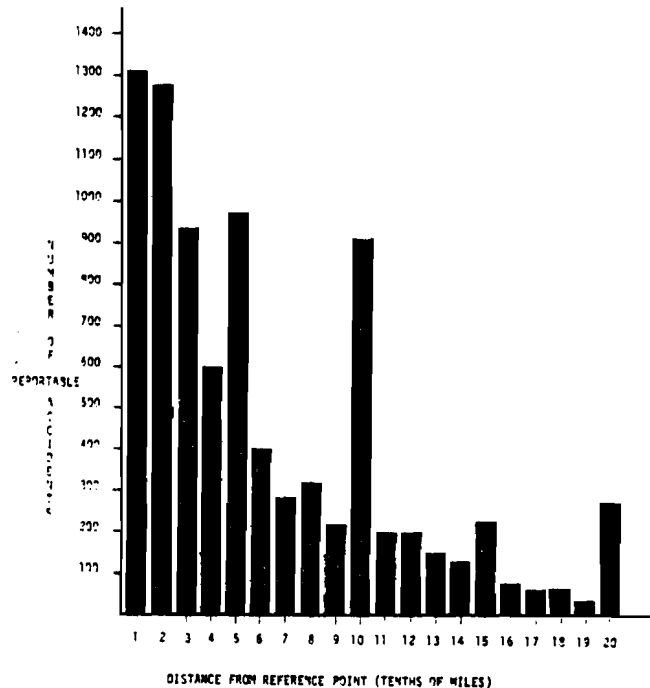


Figure 2.2. Number of reportable Interstate accidents in North Carolina (1972-1977) according to the reported distance from the nearest milepost.

Example 2. There can be biases in certain reported variables based on an officer's judgment about the situation. For example, if there is an occupant fatality, an officer arriving at the scene after occupants have been moved may assume the seat belt was not worn simply because the occupant is dead. He implicitly assumes the person would not be dead had the belt been worn. Thus, he may not pursue his investigation to find out who removed the victim from the car, and whether the deceased was actually belted: therefore, he may miss those important instances in which the person is deceased despite wearing a seat belt.

An example more pertinent to the highway area is that the officer may make a lower estimate of the impact speed of a car that hits a crash cushion than that of a car that has crashed into a standard barrier because he is "fooled" by the reduced amount of deformation that occurs when a car hits a crash cushion.

Example 3. There may be biases due to shortcomings in the accident report form itself or poor definition of the reporting variables. For example, in studying new no-passing zone markings, the study of crashes related to passing maneuvers might be hampered by a lack of data if the report form is structured to describe the first crash event and not the precipitating event. Thus, the precipitating event might be an illegal passing maneuver, but the actual first crash event might be running off the road. Consequently, the form might report the crash as a "ran-off-road" accident and might make no reference to the illegal passing maneuver.

Another bias may be due to the very fact that the study is being done. That is, if a study is underway to evaluate a countermeasure, and if attention is drawn to the fact that the study is being done between the

before and after reporting periods, this could conceivably cause changes in the reporting practices so that the "before" and "after" data are not comparable.

For example, if new pavement edge markings are installed, and reporting police agencies know about the edge marking, they might become more sensitive to (and more likely to report) ran-off-road accidents in the area, even though the official reporting guidelines are unchanged. Such a bias could cause the countermeasure to appear counter-productive rather than helpful.

To summarize: missing data are not likely to be random. To offset this, the investigator needs to deduce whether there is reason to think that the data's incompleteness or bias could work against study accuracy. Study accuracy will be compromised if such biases affect some study variables more strongly than others, or if the bias affects the "before" data more strongly than the "after" data, or if it affects the experimental data more strongly than the control data.

It should be noted that many of the problems with which researchers must contend result from the report form itself. Usually accident report forms are not created for research purposes, so very little attention is given to the needs of research. Instead, the form's content is based primarily on administrative and legal requirements. As discussed below, this does not have to continue to be the case.

2.2.1c Methods for reducing bias.

Any set of accident data is going to have significant shortcomings when it is first used for research. Nevertheless, accident researchers should not passively accept these shortcomings; rather they should get themselves "into the loop" to improve the system. It is unrealistic to hope that a data system not designed to serve research purposes should serve that purpose adequately, but there are several ways to improve the situation over a period of time (see Figure 2.3)

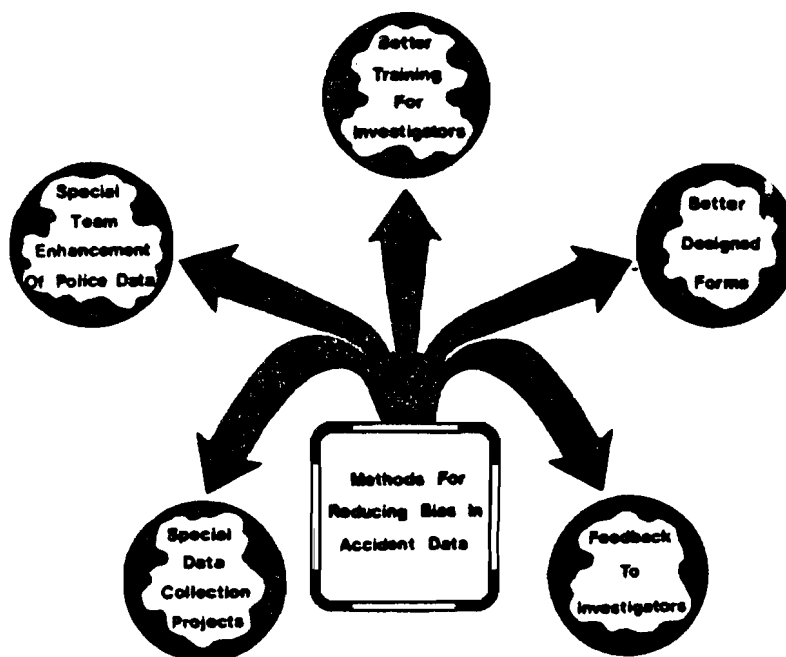


Figure 2.3. Ways to improve the data system.

(1) Better training for investigators. The research agency could seek to provide input into the training program for accident investigators and to alert police students to the needs of research, the importance of certain variables, and the applications of the data. This will tend to give the police some added appreciation of the importance of what they are doing.

(2) Better designed forms. The research interests and the highway engineering staff should participate in the process by which the accident report form is revised from time to time, and should seek ways of introducing onto the form some of the key variables necessary for research. Without such inputs on the part of the research/engineering community, the form will remain primarily a driver-oriented enforcement document. Engineers/researchers who become involved in such a redesign should refer to the AAMVA (1978) and the National Safety Council's Committee on Motor Vehicle Traffic Accident Classification (1976) publications, which provide detailed information on definitions and classifications of many accident variables.

(3) Feedback to investigating officers. A most important aspect of improving the system is to give the police officers a sense of participation. The police officer may feel that he is completing the form in vain; he may have a suspicion that the form is never really used and just goes into a file cabinet somewhere. Or he may feel that he is really being forced to fill out the form for some commercial interest. (It's not unusual to hear a police officer say, "We're filling out this form for the insurance companies anyway.")

A way to combat this is to adopt some way of getting feedback to the investigating officers, such as traffic records workshops throughout the state (see Figure 2.4) or some sort of accident reporting newsletter (see Figure 2.5). Through these devices it is possible to emphasize to the officers that their "detective" skills at the crash scene are critical to the process of saving lives. Such feedback programs should include examples of how the accident data are actually used to further the cause of traffic safety.

Figure 2.4 Possible topics to be covered in Traffic Records Workshop.

<p>ACCIDENT REPORT FORM MODIFICATIONS</p> <ul style="list-style-type: none">● Changes Resulting from Recent Legislation● Changes Made (Proposed) by State Agencies● Changes Proposed by Investigation Officers <p>REPORTING PROBLEMS & EMPHASIS AREAS</p> <ul style="list-style-type: none">● Variables on Form Where Errors Have Been Detected by DMV<ul style="list-style-type: none">- Driver variables (age, inaccurate injury data, etc.)- Vehicle variables (VIN, vehicle type, estimated speeds, etc.)- Roadway variables (poor location information, object struck, gore area descriptions, etc.)● New Emphasis Areas<ul style="list-style-type: none">- Mileposting to .01 miles, reporting crashes into new types of attenuators, distance from pavement to object struck, new points in sketches, etc. <p>SPECIAL DATA COLLECTION EFFORTS</p> <ul style="list-style-type: none">● Supplementary Forms and Procedures for Special Collection Effort● Data Collection by Special Teams● The NASS and FARS Systems <p>USES OF DATA</p> <ul style="list-style-type: none">● Recently Completed Research Projects--How Data was Used & Results<ul style="list-style-type: none">- Driver- Vehicle- Roadway● Problem Identification Usage● New safety problems detected (mopeds, speed zoning)
--

The Accident Reporter

Published for North Carolina Highway Safety Research Center by University of North Carolina

Jan., 1980

Top
of the
month.

Although there is overwhelming evidence that the use of crash-tested child restraint devices dramatically reduces the chance of death and serious injury from auto accidents, very few parents properly restrain their children while they are riding in automobiles.

In an effort to encourage the use of child restraint devices, the University of North Carolina Highway Safety Research Center is engaged in a comprehensive program aimed at reaching a wide variety of civic and health professional groups that impact on parents and their young children.

On a statewide level, HSRC is participating with the North Carolina Department of Human Resources in a series of regional workshops to promote understanding of the role of child restraint devices in reducing death and injury and to build public support for local programs that encourage the use of the devices.

To date, workshops have been held in Winston-Salem and Greenville, and others are scheduled on a monthly basis.

The regional workshops provide instruction in conducting community action programs that involve civic clubs and health professionals in the dissemination of information regarding the benefits of properly used child restraint devices. The workshops also focus on approaches to making the devices available to the public.

In the Chapel Hill area, the Highway Safety Research Center is conducting a program that might well serve as a model for local communities throughout the state.

HSRC is asking all health professionals in North Carolina involved with the parents of young children to

HSRC Expands Program for Child Restraints



distribute information that points out the safety advantages of child restraint devices and to recommend to parents that they use the devices.

In Chapel Hill, HSRC is asking local civic clubs not only to assist in promoting the use of the devices, but also to assist in financing child restraints that would be loaned to parents who could not afford them.

HSRC Researcher Lorraine Lohman says response to the appeal has been encouraging. "Almost everyone has expressed an interest in distributing educational materials," she says.

To further encourage use of child restraint devices, HSRC is providing free installation to parents who purchase them. A special installation

team has been established to carry out this task at the Highway Safety Research Center.

Finally, a special investigation team has been organized to investigate motor vehicle accidents involving children in the vicinity of Chapel Hill. The team will gather and compile data which will be available to the General Assembly in the event legislation regarding child restraint devices is considered.

With so much evidence indicating the benefits of child restraint devices, the dilemma has been how to convince parents to use the devices. The comprehensive effort now being undertaken is one step in resolving that dilemma.

Figure 2.5. Example of an accident reporting newsletter - The Accident Reporter.

(4) Special police data collection projects. When the existing form simply does not contain the needed information, arrangements can be made for supplemental reporting by officers on a sampling basis: after the necessary additional information is specified, supplementary forms are designed, and a sampling scheme is introduced by which officers report the supplementary data either statewide or within a given sampling district for a specified time. After the sample has been collected, the special reporting provisions are discontinued.

(5) Special team enhancement of police data. It might also be possible to have a small cadre of accident investigation professionals who periodically follow up police investigations to enhance the information developed by the police. This concept is being used by NHTSA in the National Crash Severity Study, and will be used in NASS (National Accident Sampling System).

These solutions to data-related problems are only a few illustrations of approaches that can be initiated by the researcher/engineer to improve his basic material over a period of time. Ultimately, however, the most important "solutions" to the data-related bias is the researcher's own knowledge of the intricacies of data formats and definitions, and the data collection and storage processes in his jurisdiction. Such knowledge can only be acquired through "hands on" work with the raw data and continuous contact with the real world of the data collector, the investigating officer.

2.2.2 Problems and issues related to the nature of accidents.

The above issues notwithstanding, the most basic problem the uninitiated highway accident researcher must face is the very nature of accidents: because crash rates have been greatly reduced over the past 30-40 years, it is sometimes difficult to study the small number that remain at a given spot or in a small geographic area or time period, and it is becoming increasingly difficult to design treatments which have a major impact on the remainder of the problem.

Accidents are very low probability events per time period, location, or driver (this is especially true for fatal accidents). Because of this fact, often only a very small number of accidents occur in any given time period or geographical area. This is illustrated in Table 2.2, which presents real-world accident and fatality rates and the predicted accident and fatality frequencies per mile of highway for various highway types and ADT levels. As the table indicates, a large sample (438 expected accidents per mile of highway) would be available for studying high-volume city streets, but a very small sample (less than one expected accident per mile of highway) could be expected for a low-volume two-lane rural highway. And predicted fatalities are even lower: the highest predicted count is three per mile.

Table 2.2. Typical Numbers of Fatalities and Accidents
Per mile of Highway Per Year.

Highway Type	ADT	Assumed Rates*		Per Mile of Highway Per Year	
		Accidents	Fatalities	Accidents	Fatalities
Urban Freeways	10,000	100	1	4	0.04
	100,000	100	1	36	0.4
Rural 2-Lane Highways	100	200	5	0.07	0.002
	1,000	200	5	0.7	0.02
	10,000	200	5	7.3	0.2
Multi-Lane, Uncontrolled Access, Rural Arterials	1,000	400	8	1.5	0.03
	10,000	400	8	14.6	0.3
	100,000	400	8	146.0	3.0
City Streets	1,000	1200	3	4.4	0.01
	10,000	1200	3	43.8	0.1
	100,000	1200	3	438.0	1.0

*Assumed rates are numbers of accidents or fatalities per 100 million vehicle-miles of travel.

The lack of numerical stability inherent in such small numbers can be overcome by aggregating numbers over a greater time or greater space. For example, although accidents on low-volume (100-1000 ADT) highways are generally concentrated at intersections, each intersection typically has only one or no accidents per year. In such cases, a large number of intersections would be needed to provide reliable comparisons. Conversely, for high-volume (20,000-50,000 ADT) city streets, an intersection may typically have between 20 and 100 accidents per year, and fewer intersections are needed to provide a reliable accident sample. The problem is not one of raw numbers of accidents: although accident rates have decreased sharply, approximately 18 million total accidents, two million injury accidents, and 50,000 fatal accidents still occur each year in the U.S. However, very few (if any) researchers will have access to national accident data. Most are restricted to smaller subsets of accidents in their own state or locality. Also, the researcher's usable data base is often further restricted to accidents occurring at specific locations, for specific treatments, and thus to specific types of crashes. These necessary restrictions often result in low sample sizes for the accidents of interest.

Related to the low probability of accidents is individual treatment's modest benefit in terms of overall accident reduction. Most new countermeasure activities can be expected to reduce accidents in any given area or location by only 15-20 percent, and often even less. In fact, many of our programs may well have benefits below the ten percent effectiveness level. However, even these modest benefits may well be worth funding, because even such low benefit levels may result in payoffs that exceed program costs.

The two major exceptions to the low effectiveness levels in the highway area are treatments aimed at "cushioning the crash" to reduce crash severity, and complete redesign of the highway to Interstate (freeway) standards. In the former case, decreases in fatalities due to crash cushions and other well-designed highway hardware have been measured at levels of 50-75 percent. In the latter case, reductions in total predicted frequencies of crashes due to the development of the Interstate system also range from 50 to 75 percent (Fee, et al., 1970). However, this total redesign is actually a combination of many individual treatments (e.g., access control, clear roadsides, wider medians, etc.), each of which contributes some percentage of the overall reduction. Because the development of new freeway-type sections is limited by cost factors, most current research does not examine this pervasive type of treatment.

The generally modest effect from a given treatment program, coupled with the low probability of accidents occurring on a given set of roadways means that the researcher will be attempting to decipher extremely small changes in accident patterns. This can be done, but only with proper research planning.

2.2.3 Problems and issues in exposure data.

In order for accident data to be meaningful, they must be compared with the experience of the non-accident population (often called the population at risk). Data about the population at risk are also called "exposure" data. (Exposure data are sometimes called denominator data: in calculating accident rates, the number of accidents is the numerator, and some measure of the population exposure is the denominator--i.e., total vehicle miles of travel is the denominator for calculating the number of accidents per million vehicle miles.)

2.2.3a What is the need for exposure data?

Exposure data are important because they are crucial to calculating the actual likelihood of an accident. That is, the researchers need to have some estimate of what might be called the "accident-opportunity" level--the number of chances or opportunities that could result in an accident. As a simple example, consider a hypothetical intersection at which ten accidents have occurred over a one-year period. These 10 accidents per year can mean one thing if the annual vehicle volume (one measure of the "accident opportunity" level) totals 10 million vehicles and something quite different if it is only 10,000. In like fashion, knowing that 50 percent of all drivers killed in accidents have been drinking (Perrine, et al., 1971; Solomon, 1970) cannot be correctly interpreted without having some measure of the proportion of drivers on the road who are drinking. If 75 percent of the drivers on the road are drinking and only 50 percent of the drivers involved in fatal crashes are drinking, then the non-drinking drivers are in more than their share of fatal crashes. (In reality, less than two percent of the drivers on the road are drinking, indicating that drinking drivers are greatly overrepresented in fatal crashes [Perrine, et al., 1971; Borkenstein, et al., 1964; Hurst, 1970].) A third example is the daytime vs. nighttime accidents. A simple tally of accident frequency indicates that daytime accidents are much more frequent. However, when driving mileage during the two periods is collected as exposure data, the indication is reversed, and nighttime accidents become about twice as likely to occur as daytime crashes (Solomon, 1964, p. 13). Exposure data not only clarify the relationship, but can even alter what the accident data signify.

2.2.3b Sources of exposure data.

There are a variety of places from which exposure data can be obtained. In some cases the State Highway Planning Department can provide an excellent source of exposure data through their systematic origin and destination studies. On a more general basis, summaries based on gasoline tax revenues or on the traffic surveys generally accumulated annually in each state may be useful. In addition, the Motor Vehicle Manufacturers Association publishes an annual estimate of state-by-state rural and urban vehicle mileage in "Motor Vehicle Facts and Figures." The best source of general vehicle mileage data is provided by the Federal Highway Administration in an annual report entitled "Highway Statistics." Here, statewide data on vehicle miles is categorized by urban/rural, highway system, and, in some cases, roadway configuration (2-lane, 4-lane, etc.).

Additional exposure data can be extracted in states with periodic motor vehicle inspection programs, but these may require some additional effort. Since each vehicle is inspected at regular intervals vehicle mileage data can be collected and special surveys can be conducted.

2.2.3c Problems in exposure data.

Just as with accident data, problems arise in the collection and analysis of exposure information. In general, to get the most out of exposure data, the researcher needs to have data on the same variables for the population at risk and for the accident population. (For example, if wet weather accidents at certain locations are being studied, the researcher needs to know the number of "opportunities" for wet weather accidents to occur, i.e., the number of vehicles that travel on these sections during wet weather. This is a very stiff requirement which few accident research studies have been able to meet. (A notable exception was the study by Blackburn, et al. (1978)).

However, readily available exposure data are usually very general. For example, if vehicle exposure measures are based on tax-related estimates of millions of vehicle miles, on traffic count surveys, or even on statewide highway statistics categorized by road type, etc., such exposure data unfortunately cannot be categorized by day or night driving, wet or dry weather, or even more pertinently, by location on the highway. In most cases this nonspecificity of data results in large problems for the researcher.

Second, even where exposure data exist, they are sometimes biased by the way they are collected. Consequently, the researcher needs to be aware of how exposure data have been collected so he can understand the biases they contain. In many cases, the data are not collected on a random, year-round basis, but by "samples of convenience": in many states, traffic count data are collected during summer months when temporary help is readily available. Similarly, spot checks of the percentage of trucks in a given traffic stream at a given location may be collected only during normal working hours and not on a 24-hour basis. Because accident data are collected around the clock, the researcher who combines accident data with such non-24 hour exposure counts implicitly assumes that the percentage of trucks observed during the 8 a.m. to 5 p.m. collection period is representative of the entire 24-hour period. (Such problems, however, can sometimes be overcome by using sophisticated counting equipment.)

There are times, however, when the exposure data may be superior to the corresponding accident data. For example, Solomon (1964), who studied the relationships between accidents, speed, and other driver and vehicle variables, and collected exposure information with interviews, noted:

. . . "On a unit basis, accident data were more difficult and expensive to obtain than interview and speed data. Accordingly, a much larger volume of the latter type of data were obtained--on the average, nearly 30 times as much. This permitted the statistical reliability of the involvement rate to be based on the number of accidents alone because the number of accidents nearly always was much smaller than the number of interviews or speed observations and therefore governed the reliability of the computed rates."

2.3 Summary

The problems inherent in accident data may be discouraging for researchers, but they are not insoluble. Like any other administrative problem, these matters must be anticipated, studied, and taken into account in planning and implementing the research. After all, the goal of accident-oriented research is to try to get the best possible information from a set of data. To do this, researchers need only keep clearly in mind the specific relationship they are studying, consider what set of observations will give the fairest and most objective chance for defining that relationship, and anticipate "tricks" the data may play due to factors like exposure, data biases, etc.

2.4 Review Questions

1. Name two instances in which the use of accident data is not feasible for conducting highway safety research.
2. Describe three potential biases that may be present in a given accident data base of which the researcher should be aware.
3. Indicate three methods available for reducing these biases in the accident data.

4. What are exposure data and why are they so important in accident research? List four existing sources of mileage exposure data.
5. Define a towaway accident and give the principal advantage of using this criterion as a reporting threshold.
6. Due to large increases in Labor Day weekend traffic, police officers in state A decide to report only those crashes that involve personal injury to the occupants of in-state vehicles. How can this practice affect a study of the relationship between accidents and traffic volume?
7. What type of data would be available if a researcher were interested in the effect of a new aggregate type on tire wear (and thus on accidents)?
8. A researcher is interested in studying the relationship between "estimated crash speed" and resulting injury. The police agency collecting the accident data has a training policy in which all accidents involving property damage or minor injury are investigated by rookie policemen. When a serious injury or fatality occurs, a supervisor (expert investigator) is called to the scene to investigate. Would this reporting practice affect the researcher's study? How?

2.5 References

- American Association of Motor Vehicle Administrators. Data element dictionary. (final ed.) Washington, D.C.: Author, 1978.
- Blackburn, R. R., Harwood, D. W., St. John, A. D., & Sharp, M. C. Effectiveness of alternative skid reduction measures. Kansas City, MO: Midwest Research Institute, 1978.
- Borkenstein, R. F., Crowther, R. F., Shumate, R. P., Zeil, W. B., & Zylman, R. The role of the drinking driver in traffic accidents. Bloomington: Indiana University Department of Police Administration, 1964.
- Bronstad, M.E., Mickie, J.D., Viner, J.V., & Behm, W.E. Crash test evaluation of thrie beam traffic barriers. Transportation Research Record, No. 488, 1974, 34-45.
- Committee on Motor Vehicle Traffic Accident Classification. Manual on classification of motor vehicle traffic accidents. (3rd ed.) Chicago: National Safety Council, 1976.
- Fee, J. A., Beatty, R. L., Deitz, S. K., Kaufman, S. F., & Yates, J. G. Interstate system accident research study. (3 Vols.) Washington, D.C.: U. S. Government Printing Office, 1970.
- Field, R. N., & Prysock, R. H. Dynamic full scale impact tests of double blocked-out metal beam barriers and metal beam guardrailing series X. Sacramento: California Department of Public Works, Structural Materials Section, 1965.
- Hayes, G.G., Ivey, D.L., & Hirsch, T.J. Performance of the hy-dro cushion cell barrier vehicle-impact attenuator. Highway Research Record, No. 343, 1971, 93-99.
- Helmets cut crash odds for North Dakota cyclists. The Highway Loss Reduction Status Report, 1979, 14(1), 9.
- House, E. G., Waller, P. F., & Koch, G. G. How complete are driver records? An analysis based on insurance claim crashes. Chapel Hill: University of North Carolina Highway Safety Research Center, 1974.
- Hurst, P. M. Estimating the effectiveness of blood alcohol limits. Behavioral Research in Highway Safety, 1970, 1 (Summer), 87-89.
- Martinez, J.E., Olsen, R.M., & Post, E.R. Impact response of overhead sign bridges mounted on breakaway supports. Highway Research Record, No. 346, 1971, 11-18.

- Perrine, M. W., Waller, J. A., & Harris, L. S. Alcohol and highway safety: Behavioral and medical aspects. Burlington: University of Vermont Department of Psychology, 1971.
- Solomon, D. Accidents on main rural highways, related to speed, driver, and vehicle. Washington, D.C.: U.S. Department of Commerce, 1964.
- Solomon, D. Highway safety myths. In P. F. Waller (Ed.), Proceedings of the North Carolina Symposium on Highway Safety. Vol. 2. Highway and traffic safety: A problem of definition. Chapel Hill: University of North Carolina Highway Safety Research Center, 1970.
- Weir, D. H., Strange, J., & Heffley, R. K. Reduction of adverse aerodynamic effects of large trucks. Washington, D.C.: Federal Highway Administration, 1979.

EVALUATING COUNTERMEASURES

Situation: The evaluator in a state traffic engineering department is asked to evaluate a program in which signs reading "DANGEROUS LOCATION, A CITIZEN DIED HERE" are placed at all high accident locations where a fatality has occurred in the past five years. The signs have been in place for two years prior to the beginning of the evaluation. Having been told of the need for a comparable "no-treatment" control group in a good evaluation, the evaluator develops a computer listing of all other locations which have had at least one fatality in the five year period and attempts to match these untreated locations with the high-accident locations on the basis of the frequency of crashes in the year before the treatment was implemented. He then uses the subset of high-accident locations which he could match as his treatment group and their respective matches as his control group.

Result: The analysis of the data indicate that, while the matched controls have a declining accident rate in the two years following treatment, the matched high accident group experiences an increase in frequency. Since the procedure he used guaranteed that all other factors were equal, the evaluator is forced to conclude that the signs are causing accidents, probably because of drivers being distracted or becoming emotionally erratic. He reports his very intriguing findings to the local newspaper. His boss then informs him that the signs were suggested by the current governor. The evaluator is transferred to an outlying highway district as an "engineer-in-training."

Main Chapter Topics

Introduction
 Definition of a Countermeasure
 What Is Evaluation?
 Limits to the Success of Effectiveness Evaluations
 Components of Effectiveness Evaluations
 Threats to the Validity of Effectiveness Evaluations
 Common Evaluation Designs Used in Overcoming These Threats
 Statistical Procedures
 Use of Evaluation Results in Cost/Benefits Analyses

3.1 Introduction

A review of research reports will often lead the uninitiated reader to conclude that the subject is very complex (as indeed it is). However, despite this complexity, all have a common underlying goal: to identify possible causal relationships between subsequent accidents and other factors of interest while accounting for all other factors which may contaminate or confuse the results. Thus, research can be characterized as the building of a mountain of evidence to help draw sound conclusions by disproving (or controlling for) all other possible explanations of an effect. For example, a researcher may wish to examine the differences between the effects of various curvature/superelevation combinations on subsequent accident frequencies and severity. That is, he may wish to determine which combination of curvature and superelevation appears to result in more accidents and which appears to result in fewer accidents. In doing this, however, the researcher has to account for many other possible contaminating or confusing factors such as vehicle speed, traffic volume, vehicle mix, pavement condition, weather, etc., since these factors themselves can result in changes in accident frequencies. This has been documented quite clearly, for example, in studies which have shown that changes in volumes and changes in speed variance both affect accident frequencies (Kihlberg & Tharp, 1968; Solomon, 1964). In a similar manner, the researcher may wish to study a new pavement edge marking scheme at a series of dangerous curves to see whether or not accidents that occur at the curves will be reduced. Again, in order to determine whether or not the variable of interest--the pavement marking scheme--is really affecting the outcome variable (subsequent accidents), the researcher must somehow account for, control for, or parcel out the effects of all other factors which could also affect accidents at these locations.

This control can sometimes be gained through actual manipulation of other factors in a planned experimental design. On other occasions, when extraneous variables are either not controllable or were not controlled ahead of time, the researcher must resort to statistical procedures in an attempt to gain such control. In this chapter and in Chapter 4, these two strategies will be discussed in detail. Strategies that apply to the first situation--planned experimental design--are covered in this chapter; those that apply to the second--extraneous variables that must be controlled with statistical procedures--are discussed in Chapter 4: materials that pertain to both are presented in this chapter.

3.2 Definition of Countermeasures or Modifiable Highway Elements

In this chapter we refer to the situation in which the highway department or another agency recognizes an undesirable safety situation and therefore intervenes through the implementation of some countermeasure designed to reduce the danger. Following intervention, it is necessary to determine whether the dangerous situation has improved. There are many countermeasures which meet this definition. An example is detection of a deterioration in wet weather performance of a section of roadway surface, and intervention in the form of grooving the pavement to improve drainage and traction. A measurement and evaluation of the success of this countermeasure could be undertaken to detect a reduction of skidding accidents.

For the sake of consistency, we will be using the term countermeasure to define any of a number of "treatments" or "fixes" which the engineer has at his disposal. Such countermeasures may range from the placement of a single treatment at a single location (e.g., the installation of a warning sign at a hazardous curve or crash attenuator system at an unprotected bridge pier) to a series of treatments at multiple locations (e.g., the complete redesign of a series of intersections along a given highway). In general, the term countermeasure will not be used to indicate the construction of a totally new facility to replace an older one (although evaluation of such a situation would follow the same general rules). For the sake of discussion, countermeasures are defined to be treatments implemented at a highway location (segment) that can be modified by the engineer.

3.3 What is Evaluation?

Evaluation is a fashionable word today, at least in part, due to federal highway safety standards that require evaluation as a part of each program. Each of us uses some form of evaluation of relevant factors in our everyday consumer decisions, usually without being precisely aware of how we did the evaluation and usually without feedback with which to measure our success. But what is evaluation in the context of this manual?

It will clarify the definition to some degree to discuss two forms of evaluation present in highway safety programs: Administrative (Process) Evaluation and Effectiveness (Outcome) Evaluation. While the two forms differ drastically, both are an integral part of the overall evaluation of a countermeasure program.

3.3.1 Administrative (Process) Evaluation.

Administrative Evaluation involves determining whether the implementation of a countermeasure or countermeasure program was performed according to plan (i.e., "To what extent was the planned process actually carried out?"). For example, if a countermeasure program involved the installation of raised centerline markers on 100 miles of 2-lane rural highway in a given county or township, the administrative evaluation would be aimed at determining how many miles had actually been marked correctly with the raised markers. While appearing extremely straightforward and thus, perhaps, unnecessary, it is important that such process evaluation be conducted each time an effectiveness evaluation is attempted in order to correctly determine what treatment is being studied. Specifically, an evaluation of the effects of a new type of guardrail on crash severity will be severely hindered if the evaluator doesn't study those crashes occurring at locations where the new guardrail was actually installed rather than in crashes where it was planned to be installed.

We will further broaden the definition of administrative evaluation to include those studies of programs (countermeasures) which either (1) do not directly impinge on accidents or (2) do so indirectly via the action of several other variables. In such instances the process form of evaluation may be all that is available. An example of this is seen when the goal of a highway safety activity is the improvement of an accident records system (a support activity). Even though the record system is being improved to detect dangerous locations so that improvements can be made to reduce accidents, it is obvious that the process of

improving the records system itself cannot be evaluated in terms of accidents prevented. Rather it must be evaluated in terms of the goals, objectives, or criteria of a better records system.

Although administrative evaluation is important and no doubt needs to be improved and expanded in its application to highway safety, nevertheless the more neglected and more important topic is effectiveness evaluation as defined below.

3.3.2 Effectiveness Evaluation.

This is a formal process, following definite procedures for determining whether and how effectively a highway safety activity has brought about the desired result. For example, if a pavement grooving program is introduced to prevent skidding, then we assume that program success means that vehicles have fewer skidding accidents than on similar ungrooved sections. Effectiveness evaluation is the set of procedures by which one formally determines whether such assumptions are correct.

Although administrative evaluation is an integral part of the overall evaluation process, effectiveness evaluation is more frequently lacking in the field of highway research: therefore, the remainder of this chapter will concentrate solely on this latter type.

In the context of this working definition of effectiveness evaluation, it is necessary to provide limited background material concerning issues in evaluation and detailed discussions of appropriate criteria, evaluation design, and analysis techniques.

3.3.3 Why carry out effectiveness evaluations?

Even though the manual is written on the assumption that the user is already committed to conducting research, it appears worthwhile at this point to present a limited discussion of the underlying bases for countermeasure evaluation for two reasons. First, in order to carry through the necessary but rather involved and tedious steps of a sound evaluation, the engineer-turned-evaluator must himself be totally convinced of the necessity. Second, even though the evaluator may be convinced of the need, he may have to convince the program manager of (1) the need to evaluate, and (2) the need to allow the evaluator some control (or input) over the ultimate treatment implementation scheme so that the evaluation can be meaningful.

Not every administrator is convinced of the need for all the trouble and expense of formal evaluation. However, with proper planning, the "perceived" trouble and expense of evaluating a countermeasure can be reduced to a very feasible level. In many cases the cost of poor evaluation may be greater than the cost of a sound evaluation which optimizes the use of available data and circumstances.

Another obstacle can be administrators who feel that it is self-evident that most highway programs are successful: after all, haven't automobile death rates plummeted five-fold since the 1920's? Such persons believe that "common sense" is a sufficient indicator of what works, are faintly amused or perhaps irritated with "proving the obvious" (when their beliefs are confirmed by evaluation), and are puzzled or angered at the "negativism" of the evaluator (when their belief in a program is not confirmed by a formal evaluation). Evaluation is also sometimes seen as a rather heartless way of judging programs because it dismisses the appealing philosophy that "if it saves just one life the program is worth any cost."

In rebuttal to these objections:

First, because only limited financial resources are available for highway safety programs, it is tritely but truthfully a matter of life and death that these monies be directed toward the programs that have the most direct impact in reducing highway death and injury. Thus, to sustain programs because "it may save a life, never mind the cost" can actually cost lives if other, more effective programs are not supported (or in some cases are not discovered). It is bad policy to finance marginal (or perhaps ineffective) programs while allowing other possibly effective programs to be inadequately funded or go unfunded altogether. Only with information from rigorous evaluations can sound administrative decisions be made.

Second, the success of a program is not "self-evident," even to individuals with an inordinate amount of common sense. The actual effectiveness of many highway safety programs is modest: because this modest

improvement affects only a portion of the comparatively few highway segments, locations, or drivers actually treated by the program, the impact is scarcely measureable at all in the total accident picture. For example, if a given highway safety program reaches only a few percent of crashes (i.e., motor vehicle inspection can favorably affect only the small percent of crashes caused by mechanical failure), and if the program itself succeeds in bringing about only a ten percent reduction in such crashes, then we have a situation in which a ten percent change is brought about in a quantity that is itself (let us say) only ten percent of the problem. This net benefit would be a scarcely-detectable one percent improvement overall. Since many programs including those in the highway area do indeed have only this kind of modest effect when successful, then evaluation becomes imperative in order to find out and document whether programs do in fact have an effect or not.

There is no quarrel here with modest successes. A program that makes only a one percent difference overall may very well be quite a good safety bargain. Indeed, progress in this field usually comes in small bits. The only point is that to detect these modest but important successes, careful evaluation is necessary.

Third, we need evaluation because in real life we rarely see a simple cause and effect relationship operating in a vacuum. Usually, many factors that can influence accidents are operating simultaneously--changes in traffic volume, population size, etc. Furthermore, countermeasure programs themselves are in effect concurrently and can augment or obscure each other's effects. In such a situation, only a formal evaluation that rigorously follows prescribed rules, can provide information about the effectiveness of the particular program under examination.

To sum up, highway safety programs are too important--too many lives depend on their outcome--to allow guesswork to guide program decisions. Because of the complicated mix of factors influencing the setting in which any highway safety program operates, it is imperative that formal evaluation procedures be used to measure actual program results. Nothing could be more practical than hard-headed assessment of actual program effectiveness. It is much more "ivory tower" to by-pass evaluation on the grounds of theory, hope, or optimism.

3.4 Limits on the Success of Effectiveness Evaluation

It is by no means always possible to perform an objective and defensible effectiveness evaluation. There are considerable barriers to this, and we need to consider them. The nature of accidents themselves places limits on evaluation; because accidents are rare events, the numbers used in an evaluation can be too small to be stable, and because accidents in the aggregate are produced by a host of causative factors; there is always the risk that the accident changes that do occur in the sampling area are produced by a hidden variable instead of by the countermeasure (see Chapter 2 for amplification of these issues).

Other practical problems are imposed by the necessity of carrying out the countermeasure evaluation in the context of a governmental situation where other considerations often interfere with evaluation (e.g., there may be times when program administrators, in their planning, omit any provision for evaluation).

In other instances, program administrators may be so sincerely persuaded of the program's worth that they feel there is no necessity to expend effort or funds in evaluating the effect of the program. They may even be hostile toward any suggestion that the program should be evaluated (and even more hostile to any suggestion that the program is not effective.) In still other instances, the administrator may desire that no evaluation be made because he fears it may be politically untenable to indicate that the program is not as effective as desired.

D. T. Campbell (1975) summarizes the basic reason why decision makers in politically sensitive positions, cannot advocate rigorous evaluation: "... specific reforms are advocated as though they were certain to be successful. For this reason, knowing outcomes has immediate political implications..." and thus, failure is intolerable. As an answer to this dilemma, Campbell says that the administrator should "shift from the advocacy of specific reform to the advocacy of the seriousness of the problem, and hence to the advocacy of persistence in alternative reform efforts should the first one fail. The political stance would become: 'This is a serious problem. We propose to initiate Policy A on an experimental basis. If after five years there has been no significant improvement, we will shift to Policy B.'" While the average highway administrator is perhaps less political and therefore more able to advocate "knowing the truth,"

such a posture, especially if expressed publicly (or to his boss), might help eliminate future resistance to evaluation efforts.

To insure that this philosophy is heard, the evaluation unit needs very much to become a part of the overall planning team. From this position, evaluators can participate in the advance planning, and acquaint officials with the realities of evaluation (i.e., taking some of the mystery out of the concept of evaluation).

3.5. Components of Effectiveness Evaluation

A number of factors are necessary for carrying out an objective evaluation of program effectiveness and can effect the chance of getting the correct answer. These factors include proper experimental attitude, knowledge of the basis of a cause-effect relationship, and knowledge of proper choice of criteria to be used.

3.5.1 Attitude

It seems almost idle to point out that the researcher should have an objective attitude when undertaking an evaluation. He should not consider himself a part of any advocacy position so that he feels compelled to show the benefits or disprove the merits of a program. Rather, he needs to take the view that the facts of the program effectiveness are not known, and that it is in the best public interest to know the truth. It is the researcher's task to set up the fairest, most impartial mechanism by which any effectiveness of a program can show up.

Sadly, there are pressures opposing such an objective attitude. Government agencies tend to thrive on successful programs and therefore may be rather cool toward disclosure of the fact that programs do not work. Sincere government officials believe in their programs or they would not be dedicating their careers to implementing them. Therefore, understandably, they may find it difficult to be objective when the evaluation and the possible future of their program may be at stake. (Cynical researchers may believe that objective evidence, even if it indicates that a program is not effective, rarely has an impact on program support.) Because it is incumbent upon the researcher to assure the objectivity of the evaluation, he needs to be able to recognize the subtle as well as obvious factors that can diminish that objectivity.

3.5.2 The Basic Idea of Cause and Effect

In everyday life in a thousand simple ways we exercise the basic model around which evaluation is based. That model is:

1. A situation exists
2. We want to change it, so we intervene
3. The situation changes

Following are four examples, each representing a successively more complex setting, in which we try to illustrate this model.

A. Toward evening it becomes too dark to read, you intervene by turning on the light, and continue to read in a better lighted situation.

B. The lawn is not doing too well, you intervene by fertilizing, and see an improvement.

C. Your child has a bacterial throat infection, you intervene with penicillin, and see a rapid uneventful recovery.

D. Your business profit situation is worsening, you institute what you think are some cost saving procedures and you hope to see better profits.

The first example is the simplest cause and effect situation. Repeated experience has shown us that turning the switch almost always bring light. Furthermore, we are familiar with the underlying process, and this gives us even more reason to believe that the simple cause and effect relationship is true.

In the case of the lawn, the situation is a bit more complex in that other factors such as rainfall, soil acidity, etc., also influence the end result, but still the process usually operates with reasonable reliability.

In the case of the child's sore throat, we normally expect recovery even without medication and know that many other factors can operate to influence the course of the disease and the recovery. Here we are not as confident of a simple cause-effect relationship. We even are aware that sometimes penicillin may not help or may even be followed by an adverse reaction. Thus, the assumed cause-effect relationship is less clear and the underlying process less well understood.

Finally, the example of the business's profits, the whole situation becomes very fuzzy because of the many other factors also operating, and the uncertain end-results of the cost cutting intervention.

The cause and effect relationship which could be assumed to underlie all these examples follows the sequence:

Situation ———> Intervention ———> Changed situation

But the examples show varying degrees of clarity and complexity of relationship.

This same general relationship holds true with highway safety programs and their evaluation:

1. A dangerous highway safety situation exists;
2. Intervention takes place in the form of some kind of highway safety countermeasure;
3. A change for the better either does or does not occur

An underlying cause and effect relationship is assumed to exist and evaluation is (1) the process of monitoring the events to see if the situation did in fact change, and (2) the process of judging whether in fact any change that has occurred can actually be attributed to the intervention. If the situation were this simple, a text on the subject would hardly be necessary. Unfortunately, there are a number of factors that complicate the situation.

First, however, a limited discussion of a related topic appears to be necessary--the concept of statistical correlation. Correlation coefficients are used to measure the degree of underlying relationships between two or more variables. The problem that exists in some published evaluations of social programs stems from the fact that certain authors have used correlational type analyses in an attempt to define a cause/effect relationship. As pointed out in numerous studies (Cook & Campbell, 1976; Huff, 1954; Campbell & Stanley, 1963), there are basic underlying flaws with this use of such correlational analysis. Perhaps the best way to illustrate these problems is with the following example provided by Huff (1954).

"It is rather like the conviction among the people of the New Hebrides that body lice produce good health. Observation over the centuries has taught them that people in good health usually had lice and sick people very often did not. The observation itself was accurate and sound, as observations made informally over the years surprisingly often are. Not so much can be said for the conclusion to which these primitive people came from their evidence: Lice make a healthy man. Everyone should have them."

No doubt, if the New Hebrides had had a tribe statistician and if he had studied the relationship between the number of lice and the health of an individual, he would have found a strong correlation between the two. However, the question remains, "Are the lice really causing the good health?"

Of course, the answer would be "no". The situation is complicated by the fact that there is an underlying third variable which is related both to the presence of lice on a person and to the person's health. In reality, almost everyone in the New Hebrides Islands had lice most of the time. It was a normal condition. However, when anyone took a fever (which possibly was caused by the same lice) and his body became too hot for comfortable habitation, the lice left. Therefore, sick people rarely had lice while healthy people did. Here, as stated by Huff, "You have cause and effect altogether confusingly distorted, reversed, and intermingled."

A second problem with the use of correlational analyses as evaluation studies in which a cause/effect relationship is being ferreted out lies in the basic deviance from the above stated model:

Situation → Intervention → Changed Situation

In correlational studies, the intervention is not being made by the evaluator. If an intervention really exists, is it made by some unknown natural force? As Cook and Campbell (1976) state:

"another and perhaps more compelling reason for relegating these [correlational] models to a low place among quasi-experimental designs is their passivity. Essential to the idea of an experiment is a deliberate, arbitrary human intervention--a planned intrusion or disruption of things as usual. Probably the psychological roots of the concept of cause are similar. Causes are preeminently things we can manipulate deliberately to change other things. Evidence of cause best comes as a result of such manipulation."

Thus, as the reader will note in the following discussion of evaluation designs to be used, correlational designs are not included. Such analyses certainly have their place in the study of underlying relationships. However, there are definite questions concerning their place in the study of cause/ effect determinations.

3.5.3 Determination of what to measure.

Let us now turn our attention from the more philosophical material to the first issue which must be resolved in carrying out an evaluation--the determination of the criterion to be measured.

3.5.3a Accidents as the criterion.

Although choosing the criterion variable (the variable to be studied) appears to be a quite simple process, review of past accident research studies has indicated that it is not as simple as it seems. The proper choice of criterion is based primarily on the purpose of the research being done, the program being evaluated, or the element under study. Thus, while the dependent variable in accident research is often the frequency or rate of total crashes, this is not always the case. For example, in an evaluation of speed warning signs at hazardous curves, the most appropriate criterion variable would probably not be all the accidents that occur on the curves, but a specific subcategory of those accidents. A likely candidate would be ran-off-road accidents, because these are most likely to be related to the elements of interest.

A simple means of beginning to determine the proper criterion variable is to ask, "What is the countermeasure program intended to do?" or, more specifically, "Which accidents is such a countermeasure intended to affect?" The researcher can then limit the data used to those which are most likely to be related to the criterion variable.

For example, let us assume that a jurisdiction is upgrading protection at high-volume railroad grade crossings by installing train-activated flashing signals in place of the existing crossbucks. What criterion should be studied?

Obviously, the use of these flashing signals cannot be expected to affect accidents at all locations on all highways throughout the state. The treatment is designed to reduce car-train collisions at the upgrade railroad crossings. Thus, the evaluator should not study accidents on all roadways or even at all grade crossings. His primary criterion variable should be accidents at the crossings of interest. In fact, the sample should be reduced even further. Past research has shown that only approximately one-third of the accidents at grade crossings involve trains (Schoppert & Hoyt, 1968). The remaining two-thirds are single car and car-to-car crashes. Thus, the evaluator should not even consider all railroad grade crossing accidents. He must limit the sample to car-train collisions.

There is one note of caution which must be expressed related to this use of subsets of crashes. The researcher must be aware that there are times when even a well-designed treatment aimed at eliminating one type of accident may "cause" an increase in a second type. For example, studies of the signalization of intersections have, as would be expected, indicated decreases in right angle collisions, most probably as the result of decreases in opportunities for conflict between crossing vehicle flows (Clyde, 1964; Conner, 1960; Solomon, 1959; Vey, 1933). However, the same studies have consistently shown increases in

rear-end collisions, and, in some cases, in total crashes. Thus, evaluators who study certain subsets of accidents must be alert to other possible undesirable relationships between the program of interest and other types of crashes. Knowledge of only limited "positive" relationships without attention to these associated undesirable trends may well result in improper based funding or design policies. In the example of the railroad grade crossing signalization, the evaluator might also examine the single-vehicle fixed object crashes for possible increases, especially if the treatment involved the installation of larger "fixed objects" such as guardrails around the flashing signals. However, this problem would not be expected to be a major one.

Continuing with a discussion of those instances in which accidents, or some subset of accidents, are used as the criterion, a final issue concerns whether accident frequencies or accident rates should be used. Should the researcher analyze accidents per driver, accidents per hundred million vehicle miles, accidents per location, per vehicle, or should he simply use the number of accidents? A very strong argument can be made for including some measure of "crash opportunity"--exposure--in any research involving accidents. Without such a measure of exposure, the interpretation of the results of evaluation studies and of research involving relationships is very difficult and at times almost impossible. The primary measure of exposure used in accident research is some measure of the average traffic flow or volume through a location or over a highway segment. The units of such a measure could include vehicle miles, vehicles per year, vehicles per day, passenger miles, and ton-miles-- measures which have been shown, at least by some studies, to be associated with accident frequencies (Kihlberg & Tharp, 1968; Fee, et al., 1970; Raff, 1953). Indeed, ADT is more strongly related to crashes than almost any other variable. Thus, in meaningful evaluations, some measure of exposure must be accounted for.

However, exposure is inherently accounted for in the stronger evaluation designs (i.e., designs involving randomly assigned controlled groups and designs involving very similar comparison groups). Other less strong designs can help account for exposure differences to some extent, but not completely. In the poorer design (i.e., before/after) little control is afforded over exposure or any other contaminating variable. The purpose of designing an evaluation correctly is to be sure that other variables such as exposure do not lead the evaluator to the wrong conclusion. Thus, to guard against this, the best guideline to follow is to use rates as the criterion variable when a poorer design is being used. When randomly assigned control groups are used or when similar comparison groups are used, the use of rates is not as important because differences in exposure will be accounted for by the design itself.

3.5.3b Crash severity as criterion.

In many cases, the criterion variable will be related either to total accident frequencies or to the rates of special subsets of crashes. However, there are other times when the frequency of accidents is not the most appropriate measure. Again, the user must mentally refer to the basic question--"What can the countermeasure being evaluated be expected to affect?" In some cases, the answer to this question will involve crash severity rather than crash frequency. While this difference may appear subtle at first glance, the erroneous choice of criterion can completely disguise any effect that is present.

In reality, when countermeasure programs are being evaluated, the use of accident severity as a criterion will often be the case. Let us assume that on a section of four-lane undivided highway a certain number of crashes occur because vehicles evidently go out of control, cross over into an opposing lane, and crash into vehicles travelling in the opposite direction. Naturally, other crashes also occur, such as rear-end collisions, same direction sideswipes, single vehicle crashes, etc.

Because of the frequency and severity of these cross-over crashes, a concrete median barrier (New Jersey design) is installed to guard against such events. What accident should be measured as an indicator of the benefits of the median guardrail?

In this instance, it is again quite obvious that one would not record all crashes in the state, nor even all the ones in the county. Since the median barrier cannot possibly influence any crashes except those where the rail is installed, it is logical to study crashes on that particular stretch of highway.

The next question is whether the median barrier, even if successful, can act to reduce all types of crashes on the particular section in question. Again, the answer has to be that the guardrail can only

influence one type of event--the cross-median crash configuration. And it is that one type of accident that should be used as the indication of benefits.

But now the criterion question becomes more subtle. When there is no median barrier, two cross median situations can exist:

1. Car crosses median and hits another car or object (crash);
2. Enters median (or crosses it) but hits nothing and recovers (no crash).

Thus, sometimes cars can enter or cross over the median without a crash resulting. However, because the barrier's installation reduces the available recovery area by 50 percent, all former complete cross-over events and many of the partial cross-overs become crashes with the barrier. Because only hits are recorded in an accident data system, the researcher cannot know how many actual crossovers or partial crossovers actually occur.

Given this situation, the number of median crashes might be expected to increase. An example of just such results is found in a 1969 Arizona study (Olivarez, 1969) involving concrete median barriers (CMB) and metal beam guardrail installed on 15 miles of basically 6-lane urban freeway with a 12-foot median. Table 3.1 presents the results for the 10.2 mile section in which CMB was installed.

Table 3.1 Before/after accident analysis for Concrete Median Barriers

	Before (18 months) <u>No Barrier</u>	After (18 months) <u>CMB</u>
Miles	16	10.2
ADT	33323	41775
MVM	266	223
Total Accidents	355	424
Median-involved accidents	58	79

Here, while the million vehicle miles of travel (MVM) is less for the shorter segment with the CMB, the total accidents and median-involved accidents have increased. Does this mean that the CMB treatment actually caused harm? It is impossible to answer that question from these data.

Let us reexamine the purpose of the treatment. The median barrier is designed to eliminate head-on cross-over crashes. But we also know (from these data and from analytical thought about reduced recovery area) that the barrier may cause an increase in median-involved crashes. But, wouldn't we expect these resulting crashes to be less severe than the head-ons we eliminate? Thus, the criterion of interest has quite subtly shifted from accident-based to severity-based. The proper criterion for study would be the injury distributions of all median-involved crashes (including cross-over head-ons and barrier hits). We would anticipate that, while more total driver injuries might result (from the increased number of crashes), fatalities and serious injury would be reduced. (It should be noted that the authors of the above study did present data on "injury or fatal accidents." However, they studied the total "injury and fatal accidents" within the segment instead of only those involving the median.)

If examination of the basic question concerning the purpose of the countermeasure indicates that crash severity is the appropriate criterion, the researcher must then decide which of a number of different severity-related measures he should use. Choices could include the number of total injuries, the number of injuries per vehicle, the number of fatal injuries per vehicle, the number of serious plus fatal injuries per vehicle, the number of vehicles experiencing damage above a certain level, or measures related to shifts in injury distributions. No single choice is the most appropriate one in all cases, but there are some issues to take into account in selecting the severity related criterion.

Measuring vehicle damage commonly involves one of these variables from accident report forms: 1) total damage costs to the vehicle, 2) the Vehicle Damage Index (VDI) Scale used primarily by in-depth accident investigators and by some police agencies (see Figure 3.1), and 3) the Traffic Accident Damage

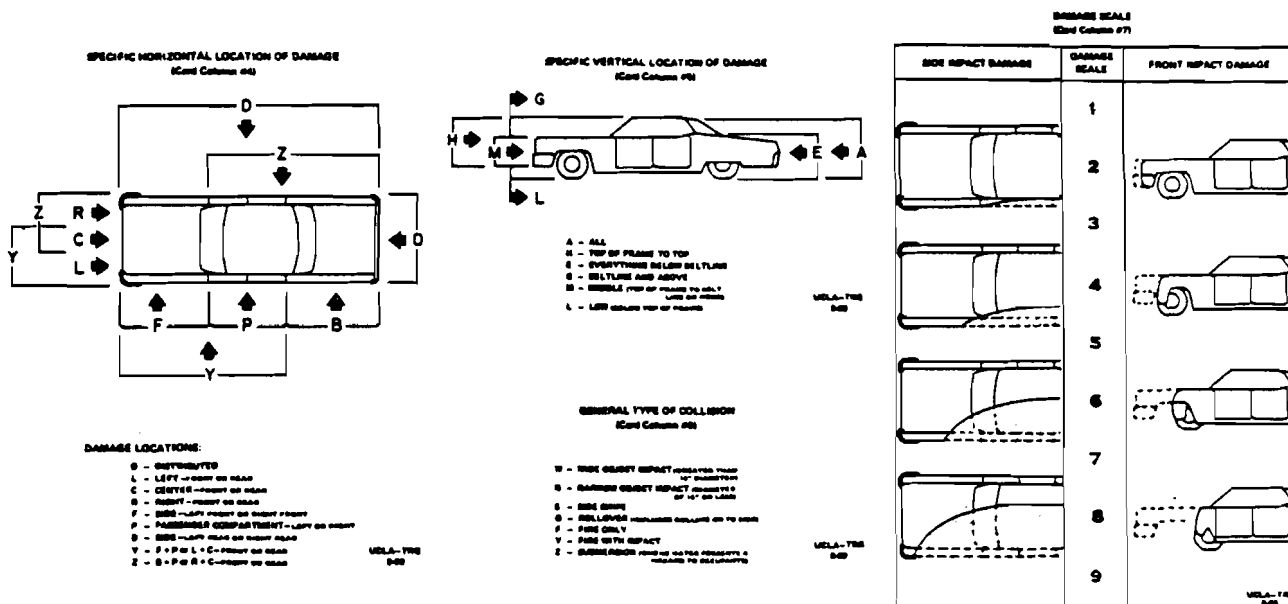
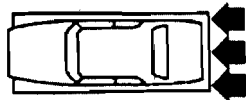


Figure 3.1. Damage coding index for VDI Scale.

SEVERITY SCALE

BD

Rear End Damage
Distributed Impact



This scale is applicable to damage to rear of subject vehicle resulting from full contact of rear end of subject vehicle with other vehicle or object.

Damage Rating

BD-1

BD-2

BD-3

BD-4

BD-5

BD-6

BD-7

BD

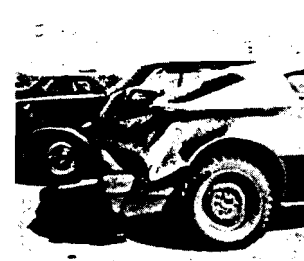


Figure 3.2. Sample page from the TAD Scale Manual.

(TAD) Scale used by some police departments (see Figure 3.2). The latter two scales are composed of coded letters designating the part of the car damaged followed by a numerical value designating the extent of the damage. Here, some combination of the numerical values could be used as the dependent variable, i.e., a numerical value of greater than or less than a certain amount (e.g., a TAD rating greater than 3 or a VDI rating greater than 2), or the researcher could most appropriately examine shifts in the overall damage distributions.

While the definition of such a value is relatively simple, there are problems associated with using damage as a measure of severity. First, the data provided represent estimates by the reporting officer rather than by assessments of competent mechanics. Second, biases can occur because damage to different types and years of vehicle may be judged as higher or lower even though the actual severity of the crash is the same (e.g., an older vehicle striking a guardrail at 20 mph may be more extensively damaged than a newer vehicle striking the same guardrail at the same speed because of deterioration due to aging). Thus, if the roadway segments under investigation for some reason have differing aged vehicles, one type of guardrail may incorrectly appear "better".

Traditionally, the primary measure of accident severity has been injury severity for vehicle occupants. Again choices exist concerning whether to use total number of injuries, injuries per car, fatalities per car, total number of serious injuries, shifts in injury distribution, etc. The following issue should be considered. If a simple count of total injuries (fatal + serious + moderate + minor) is used, then there must be control over the number of vehicles on the road and thus the total number of crashes. The number of injuries is obviously not only a function of the safety device but also of the number of crashes that occur and the number of vehicle occupants. In addition, roadside safety devices are not designed to completely eliminate injury but to reduce its severity--to shift fatalities to injuries and serious injuries to minor injuries. The use of a count of total injuries would not measure such a shift. If the most severe injury in a given vehicle is used, it will be affected to some extent by the occupancy rates for the vehicles: the more occupants, the greater the chance of a more serious injury to one of them. Finally, because there is often a great deal of interest in fatalities, the number of fatalities is often proposed as a dependent variable. However, the researcher should keep in mind that fatalities occur only in a very small proportion of crashes (usually less than one-half of one percent), and thus, the use of fatalities as a meaningful dependent variable requires exceedingly large sample sizes of data.

One severity-related dependent variable which appears to overcome at least some of these problems would be shifts or differences in driver injury distribution. By using driver injuries, the problem stemming from differential occupancy rates is overcome (almost all vehicles at least have a driver present). In addition, by examining changes in the overall distributions rather than just in a given injury class, the researcher is more likely to detect the subtle shifts within the distributions which might not be detected by analysis of injury counts within classes. Again, although counts of total injuries should be avoided, a less satisfactory alternative would be to use moderate and severe injury rates per driver.

3.5.3c Intermediate measures as the criterion.

In most evaluations, the researcher will be using either some measure of crash frequency or crash severity as the dependent variable. However, there are other cases in which other measures should be considered. Because past advances in highway design have resulted in a relatively low level of accidents per mile or per year for a given section of roadway, very long time periods will often be required for the accumulation of adequate accident data samples. This problem can be overcome to some degree, of course, if the jurisdiction records being analyzed are computerized. However even when such a system exists, when specific subsets of accidents are needed, the problem will remain, and unreasonable periods of time may be necessary for the accumulation of an adequate sample.

In addition, even when accident data from a long time period are being used or accumulated, there is often a need for some intermediate indication of relationship for use in decision-making, particularly in the evaluation of treatments. The decision-maker is often faced with the problem of needing information on countermeasure effectiveness for input into a current decision when inadequate accident data exist. Thus, the researcher may find that he may have to develop other criterion measures rather than accident frequency or severity.

There is much controversy today concerning the use of these so-called proxy or surrogate measures among accident researchers. One issue concerns defining what constitutes such measures. Two primary terms used today for referring to these substitute criteria are "proxy measure" and "surrogate measure". While the

following definitions may not be accepted by the entire research community, they will be used in all subsequent discussion in this manual for the sake of clarity.

Surrogate measures are a large group of (hopefully) adequate substitutes for accidents and other measures of inadequacies in highway system operation. The important term here is "system operation". The proposed surrogate measures have a known relationship with total highway operation (i.e., they affect or are related to traffic volume, delay, movement of goods and services, etc.), but they do not have a known relationship with subsequent accidents. Thus, "surrogate measures" is used in the remainder of the manual to refer to operational measures which usually do not have a known relationship with crash frequency or severity, although such a relationship might be hypothesized on the basis of "common sense" or "engineering judgment".

Proxy measures, on the other hand, are adequate intermediate substitutes for crash frequency or crash severity (i.e., they can realistically be substituted for actual crashes or crash severity). To be an acceptable proxy, the measure must have two attributes. First, it must be measurable, i.e., the researcher must be able to count or measure the frequency with which the measure occurs. Some proposed proxy measures fail this rather basic criterion. In the field of driver safety, for example, the use of driver attitudes is sometimes proposed as a proxy measure. This criterion is an unacceptable proxy because attitude is very difficult to define, and changes in attitude are hard to quantify. However, the use of a speed distribution (i.e., the number of vehicles within established speed intervals) can be measured and therefore is a good proxy measure.

The second attribute of all proxy measures is that they must have a known relationship with accidents. This requirement greatly reduces the current number of acceptable proxy measures. Although many substitute criteria can be hypothesized based on engineering judgment, very few have been shown to have a known relationship to crashes. Perhaps the best current example of this controversy is traffic conflicts, erratic vehicle maneuvers which occur at a given location and can be measured by the application of brake lights, vehicles crossing of the center line, vehicles swerving into other lanes, etc. Unfortunately, data currently do not indicate a relationship between conflicts and subsequent accidents.

On the other hand, past research has shown that speed variance or difference from mean speed on a given section of roadway is related to the probability of a crash (Solomon, 1964; Cirillo, 1968; Research Triangle Institute, 1970). Thus, this is a criterion which is both measurable and has a known relationship with accidents--this would be an acceptable proxy. An example of its use might be in a study of the relationship between the presence of advisory speed signs at high accident locations and subsequent crashes. Here, speed variance might be an acceptable proxy measure (criterion variable) to relate to other factors, including the presence or absence of the signs.

As an additional example, let us assume that, because police reports indicate that excessive speed is contributing to accidents on certain segments of two-lane rural roadways, the traffic engineering department has decided to reduce the speed limit and to install special speed signing which include flashing lights to draw attention to the new limit. What criterion variable should be measured to evaluate the treatment?

Here determining the proper criterion is more difficult than in the preceding cases. Although it may initially appear that the evaluator might look at either total accidents or at accidents occurring above the speed limit, it may well be the case that the effect of this relatively low cost treatment could realistically be so small (and yet worthwhile) that it would not be apparent in the examination of either of these criteria. In addition, the second criterion, accidents occurring above the speed limit, involves the judgment of the reporting traffic officers, who are, most likely, aware that their earlier accident reports were the reason the treatment was installed. Thus, there may be bias in police reports of speed after the treatment.

The main problem, however, is that the effect of such a treatment may indeed be so small that it would be almost impossible to measure in a reasonable period of time. In cases like this, appropriate proxy measures can be useful. When we analyze the purpose of the treatment, we find that the treatment is intended to affect those accidents that are related to speed. Research (e.g., Solomon, 1964) has shown that speed variance or difference from mean speed on a given section of roadway is related to the probability of a crash. Thus, this proxy criterion, which is both measurable and has a known relationship with crashes, could be an acceptable criterion measure in the evaluation of this treatment. For example, if analysis of

subsequent speed data were to indicate that the speed variance had decreased, there would be a basis for assuming that a decrease in accidents would occur.

Two other intermediate criteria which have a known relationship with accidents involve precipitating behaviors in pedestrian accidents and traffic behavior measures which are related to inadequate gap times (following-too-closely). In the first instance Snyder (1972) analyzed 2100 pedestrian accidents in 13 cities to identify common pedestrian behaviors which result in accidents. These included pedestrians "darting out" from street side locations (often from between parked cars), pedestrians dashing across intersections, driver attention conflicts, and child pedestrian/vehicle movement in the vicinity of ice cream vendors. Based on these known relationships to accidents, it would appear that counts of such behaviors at specific locations could be used as an intermediate substitute for pedestrian crashes. (This was actually done in a follow-up study by Knoblauch (1975) in which various treatments were evaluated on the basis of changes in these behaviors.)

In like fashion, a 1976 study by Lohman, et al., indicated that another proxy measure which might be appropriate is the number of vehicles or rate of vehicles which are following-too-closely, vehicles which have not allowed a sufficient gap time between themselves and the vehicles ahead. Analysis of accident and exposure data indicated that not only is this maneuver associated with a large number of crashes, but that when the occurrence rate is counted in non-crash situations (i.e., in the normal population at risk), the insufficient gap time maneuver is seen approximately 21 times more often in crashes than would be expected from what occur in normal driving. Therefore, because following-too-closely is both measurable and associated with accidents, it can be a proxy measure for evaluating certain countermeasures.

This discussion is not intended to be a complete listing of appropriate proxy measures. Indeed, there are many more with which the individual researcher might be familiar. Remember, however, that any proxy measure used in research must be measurable and have some known relationship with subsequent crashes or crash severity.

On the basis of this discussion of surrogate and proxy measures, it should be clear why this manual recommends that only the latter be used in accident research. In order to be acceptable to the general public and to many decision makers, operational measures (e.g., speed, passing maneuvers, traffic conflicts, etc.) must be related to crashes. Although accidents are not the only indicators of the operational efficiency of a highway system, existing funding criteria and the "political" situation dictates that other operational measures must directly relate to what is thought of as safety (i.e., to crash frequency or severity) in order to be acceptable substitute measures. Specifically, in the safety improvement area, the administrator is often involved with what are known as categorical safety funds, monies which must be spent to improve the safety level of a given highway component (e.g., railroad grade crossing, pavement delineation, etc.) rather than to improve the entire system's operational efficiency. Although the two are undoubtedly related, the criterion which the researcher is expected to use to measure this increase in safety level is also expected by the engineering community to have some relationship with what is considered to be the ultimate measure of system safety, the accident configuration.

In addition, the funds available for improving the operational efficiency of a system or highway, the funds available to improve the ability of the roadway to carry higher levels of traffic more efficiently, are, in general, much larger funding pools than the ones that are available to improve the "safety" of the system. Because there are enough safety problems to absorb all the currently available safety dollars, safety administrators are disposed to relying on "bottom line" measures (i.e., accident frequency and severity). They are not satisfied in spending their limited funds for changes that do not seem to be directly associated with safety (i.e., the reduction of crashes or crash severity). Because of this real-world "bias," accident rates (or accident severity) are the measure of interest among decision-makers, and accident research is and probably will continue to be considered more important than research involving other operational measures. Because of this bias, this manual will concentrate on proxy measures as the only adequate substitutes to the "bottom line" criteria of accident rates or accident frequency. In defining the proper evaluation criterion, it is important to alter the criterion only if the alteration gives a clearer indication of the countermeasure's effectiveness. This does not necessarily mean using the criterion that produces the most favorable results. For example, it is not justifiable to change in midstream from the best, most relevant measure to an alternate measure simply because preliminary analysis of data indicate no effect on the best criterion but an apparent effect on the alternate one. The guiding principle for establishing evaluation criteria should be to assess as accurately as possible a program's effectiveness and

not to fabricate criteria that produce figures that support some pre-conceived notion of what the countermeasure should achieve.

It is important to remember that most safety countermeasures are designed to address a specific problem and therefore affect only a small subset of accidents. Consequently, the best way to detect program benefits is to measure its impact on the affectable accident subset. An integral part of this process is specifying the goals of the countermeasure in sufficiently precise terms to be able to define a suitable measure of effectiveness. If one can state only in very general terms what a countermeasure is intended to accomplish, then obviously it will be difficult to specify a clear-cut criterion or measure of success.

3.6 Threats to the Validity of Effectiveness Evaluations

After determining the proper criterion for assessing the impact of a specific countermeasure, the next step is to determine the evaluation design to be used. By design of the evaluation, we mean the specific method to be used in collecting the data -- primarily the choice of which locations or segments to study and the specification of the observation time periods. While there are numerous alternative methods or designs which can be used, the choice of the proper design is based on one key principle -- the chosen design should insure that any change observed in the measured criterion has been caused by the treatment implemented and not by anything else. Unfortunately, because this "insurance" cannot usually be perfect, the evaluator's goal is to choose the design which will best discount as many other causes as possible. If other causes are allowed to contribute to a given change, they become threats to the researcher's ability to draw sound conclusions concerning his countermeasure program.

When such other causes are discussed in more technical literature, they are usually described as "threats to the internal validity of experiments." As noted by Campbell (1975), "if a change or difference occurs, there are rival explanations that could be used to explain away [this] effect and thus to deny that in this specific experiment any genuine effect of the experimental treatment has been demonstrated." These are the rival hypotheses or explanations which the researcher must consider and hopefully be able to discount through his evaluation design.

A second principle involved in choice of designs is that the chosen design should insure that the experimental results obtained from the sample studied can be interpreted, and can be generalized (extrapolated) to the population in question. Threats to this generalization are termed "threats to external validity" (e.g., external threats would arise when a researcher attempts to extrapolate from the evaluation of one edge-marking scheme on one segment of two-lane rural roadway to all possible locations where the treatment could be applied). Obviously, external threats are related to some degree to the representativeness of the sample chosen. While these threats to external validity are both ever-present and important, because they are present even in the best designed (most carefully controlled) studies, and because they are less affected by the choice of design, the remainder of this section will concentrate on the internal threats.

Campbell and his colleagues have, in a number of papers (Campbell, 1975; Cook and Campbell, 1976; Campbell and Stanley, 1963), enumerated and discussed up to thirteen general classes of such rival explanations. While all are possible threats to the internal validity of highway accident evaluations, the nature of available data and the nature of treatment programs usually under study in this particular area reduces the importance of many of these candidate threats. For a specific example, in evaluations of safety countermeasures related to drivers, threats exist which are a direct or indirect result of the subjects being able to detect their being tested (evaluated). For example, the threat of "testing" refers to the fact that people (drivers) may perform better on a posttest simply because they have practiced on a pretest. Another rival hypothesis would be "Hawthorne effect" in which groups that are experimented with will change regardless of the treatment simply because they are part of an experiment.

In research involving highways, this is not the case since the specific locations (our subjects) should not have the ability to "learn" or "react" to the fact that they are being evaluated. The highway program evaluator is fortunate. For this reason, the following discussion is limited to but four of the thirteen classes of threats -- history, maturation, regression artifacts, and instability.

NOTE: The manual user should not consider this discussion as "research philosophy" which can be viewed lightly. The ultimate decision of whether a proposed evaluation design is valid depends on overcoming these

threats. The choice of a proper design rests completely on the evaluator's understanding of these rival explanations.

[Manual users who evaluate driver-oriented countermeasures need to be aware of the complete list of threats; we strongly suggest that they refer to Campbell's lengthy but highly informative treatment of this area (Campbell and Stanley, 1963; Cook and Campbell, 1976).]

3.6.1 History (other causes at the same time).

The first threat, history, is the possibility that specific causes other than the treatment we are investigating resulted in all or part of any observed difference. The evaluator's goal is to ascertain the effectiveness of the treatment itself and discount the other potential causes of a given change.

For example, many of the recent evaluations of the 55 mph speed limit involved examination of accidents before and after the speed limit was enforced. In a very simplistic sense, a researcher (or administrator) interested in the question of the effectiveness of the speed limit as a traffic safety countermeasure might look only at fatalities per year before the imposition of the speed limit and compare them to fatalities per year after the imposition of the lower speed limit.

In almost every state such an evaluation would have shown a decrease in the number of fatalities that occurred. Thus, the researcher might conclude that the 55 mph limit was the cause of this drop in fatalities. However, it is quite clear that there were other causes which were also operating at the same time and which were related to this drop in fatalities, principal among them being the decrease in mileage driven due to the energy shortage. In addition to lower miles driven, other possible causes of the decrease in fatalities could have included such mechanisms as (1) change in driver mix--because of lower availability of fuel, teen-age drivers may not have been allowed to use the family cars as much, thus increasing the proportion of older, safer drivers in the driving population; and (2) changes in the time of day that exposure was accumulated--if few miles were driven at high-risk times (e.g., on early weekend mornings), fewer fatalities would have resulted; etc. Thus, there are a series of "causes" of the decrease in fatalities which can be grouped under the threat "history", and each is a potential rival explanation of the decrease. Again, the use of the proper evaluation design could help discount these rival explanations.

3.6.2 Maturation (trends over time).

The second threat to the validity of an effectiveness evaluation is maturation, the natural aging of the data being used. In terms of highway research, the most obvious examples of this phenomenon are accident trends over time. For example, if an evaluation of a specific countermeasure shows a change in accident rate between Time A and Time B, it is possible that this change was due to the treatment applied. However, an alternative explanation might be that this decrease in accident rate was simply the extension of a continuing decreasing trend which had been occurring for years. Specifically, accidents per million vehicle miles (and particularly fatal accidents) for the entire U.S. have been decreasing for several years. (See Figure 3.3.) If a researcher did not realize this and was in the process of evaluating a change in roadway standards in his state, he might conclude that the observed decrease in accidents per million vehicle miles from one time period to the next was simply due to the change in design standards. Although this could be the case, another alternative cause of this decrease could simply be the continuing decrease in accident rates per million miles resulting from the combination of many other factors.

3.6.3 Regression artifacts.

Perhaps the most important cause of erroneous conclusions in highway-related evaluations, however, is the threat of regression or, as it is more commonly known, regression to the mean. Regression is a phenomenon which operates to the greatest degree when potential sites are chosen because of their extreme rate in a given time period. That is, from among the potential sites for treatment, the ones selected are those with the very worst recent accident histories. Statistically, this regression effect occurs anytime when measurement is made on two variables which are not perfectly related (such as accidents in two time periods at a given location). In simple language, "the highest (best) get lower and the lowest (worst) get higher automatically."

This phenomenon can be best demonstrated with an accident-related example. First, let us assume that the points in Figure 3.4 represent the number of accidents occurring at a certain location (say, an intersection) in each of the previous ten years. Although the average number of accidents per year is 20, the individual frequencies range from 8 to 32. Now, observe each point which greatly deviates from the average (e.g., 1971, 1972, 1975, 1977), and study the situation in the next year. In each case, the deviant

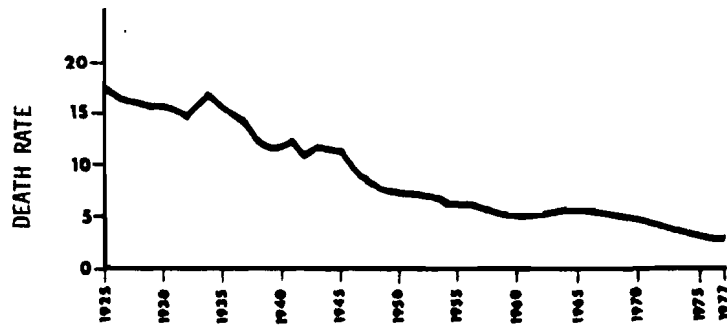


Figure 3.3. Deaths per 100,000,000 vehicle miles.

Source: Accident Facts, 1978, p. 40

points have "regressed toward the overall trend mean" without any treatment having been applied. Let us further assume that the year is early 1973 (and thus we don't have the benefit of the 1973 or future pattern). Because the accident experience has been so bad in the past year (1972), we decide to treat the intersection by new signalization. At the end of 1973, we study the effects of our program, and find that crashes have been reduced by 28 percent. And if we analyze the data for the two following years (1973-1974), we find that the crashes have already been reduced by half. But has the change in signalization reduced the accidents? With our knowledge of the entire 1969-1978 trend, we realize that the decrease was simply the natural result of the regression phenomenon and not caused by the treatment.

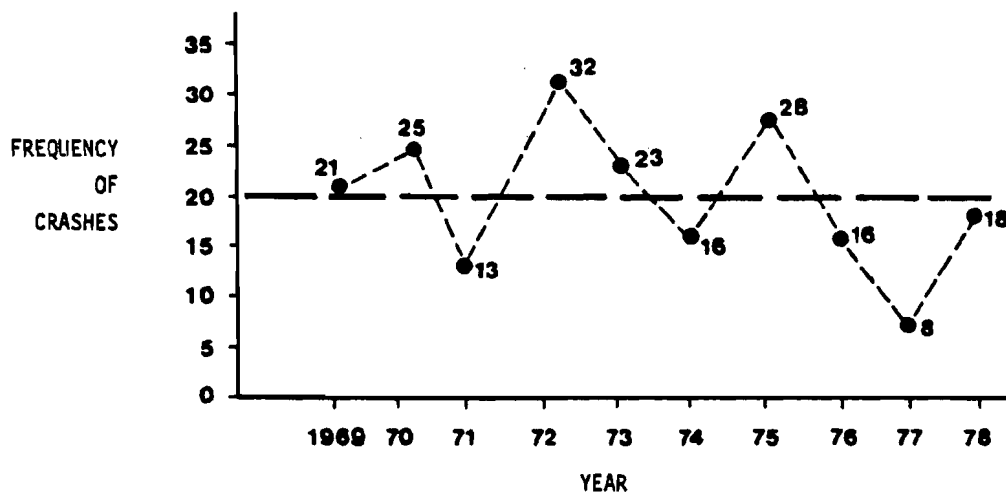


Figure 3.4. Frequency of accidents at intersection X.

The phenomenon can also be illustrated in a simple probability demonstration involving a series of drawings from a box of red and white marbles with the red marbles being defined as accidents in a given

year's time. If the true percentage of the red marbles in the box is say, 30 percent (i.e., an average of three accidents per year), repeated draws of ten marbles will sooner or later result in one draw in which eight or nine red marbles are drawn. We can see the effects of the regression phenomenon on the next draw. This next draw will, with a high probability, contain many fewer red marbles (fewer accidents per time period). Because nothing has been done to the box of marbles to change the actual overall proportion of red marbles (it remains 30 percent), the "improvement" we see from the high marble draw to the low marble draw is the regression phenomenon (natural fluctuation) and not really the result of a treatment.

In real life, the situation is slightly different. Our treatment probably will cause some improvement. Unfortunately, if treatment locations are chosen because of their recent deviant history, it will be difficult to determine how much of the change is caused by the treatment and how much is the result of the rival regression-to-the-mean explanation.

This is not to say, however, that we should not continue to treat high-accident locations. However, the above discussion does clearly point out the need to identify locations that are truly deviant in the long run (e.g., have true averages of 30, or 40, instead of a true average of 20) rather than locations which are not really high-accident locations but only appear to be due to a normal short-term chance fluctuation. (Thus, the engineer who must identify such locations should study as long a history as possible for each potential high-accident location). Even with these precautions, however, the fact that the evaluator must often study programs implemented because of recent deviant accident histories means that this "cause" will often be present to rival the explanation of changes due to treatment.

Understanding the regression phenomenon is important because of the highway safety field's normal preoccupation with "spending money where the problem is" (i.e., with treating high accident times, drivers, or locations). The regression phenomenon has invalidated or confused the results of a greater number of past evaluations than any other threat. Unfortunately, upon close scrutiny, the huge benefits claimed for many of our problem driver or problem location treatments have simply been the result of the regression phenomenon.

3.6.4 Instability.

The final threat to internal validity which will be discussed is instability. As defined by Campbell (1975), this alternative explanation of effect refers to the "unreliability of measures, fluctuations in sampling persons or components, autonomous instability in repeated or equivalent measures" -- in short, the chance or random fluctuations of the data. Accident data over time or over locations or over other groups will not remain consistent but will vary. Thus, the threat is that what might be interpreted as a treatment effect is, in actuality, only a random fluctuation of the observed data.

Of interest is the fact that, unlike the above described threats (and unlike the remaining threats not discussed here), instability is the only threat that can be overcome through the use of statistical techniques, rather than through the use of the proper evaluation design.

[This point may be viewed as rather surprising by the user given the great amount of emphasis normally placed on statistical techniques and the limited amount of information on experimental design in an engineer's education.]

To further reemphasize this point, Cook and Campbell (1976) view statistics as "fallible gate keepers" in that they can, with a degree of certainty, help determine whether an observed change or difference is "real" or only a chance occurrence, but they cannot determine the true underlying cause of the change. Statistical tests will accept as real all changes that can result from the rival explanations described above.

In summary, this section has described the major categories of rival explanations to help choose a proper evaluation design. Only through the use of appropriate designs (rather than appropriate statistics) can these rival explanations be discounted so that the change measured can be assumed to be due to the treatment implemented.

3.7 Common Evaluation Designs Used in Overcoming Threats to Validity

After taking these threats into account, it is possible to select an evaluation design. The designs presented in this manual can be categorized into (1) those evaluating a single countermeasure treatment, and

(2) those involving more than one "degree" of treatment in the same study. The first group will usually be aimed at a difference between a treatment condition versus a no-treatment condition and would, for example, include the replacement of all rigid sign supports with breakaway supports. The second category usually involves determining the difference between the effects of multiple levels or types of the same treatment and would include, for example; the evaluation of three different types of pavement grooving and a no-treatment condition on curves on two-lane roadways.

3.7.1 Evaluation of single-treatment programs.

Each of the single-treatment designs discussed below is first described and then analyzed in terms of the internal validity threats that it helps control.

3.7.1a Before/after design.

The first design to be discussed is the simple Before/After design. The model is as shown below:

Measurement Before ———→ Treatment ———→ Measurement After

This design is discussed for two reasons: first, it has traditionally been the most widely used design in the evaluation of highway countermeasures; second, it provides a prime example of a design which does not control for the important threats to internal validity and thus is very vulnerable to yielding the wrong answer.

Because this design is found so often in the highway research literature, referenced examples will not be given. Instead, let us turn our attention to the inability of the design to account for each of the rival hypotheses.

History. With a Before/After design in which only one location or group of locations is studied, causes other than the treatment can occur at the same time the treatment is implemented. The rival explanations cited earlier to the 55 mph's effect on fatalities is a classic example of the loopholes in the Before/After study design. Likewise, in studies of pavement grooving, rainfall amounts (and therefore the frequency of potential skidding accidents) could have been different in the before and after measurement periods. In a Before/After evaluation of a countermeasure involving modifying telephone poles to make them breakaway, an observed decrease in driver injury severity may be the result of the treatment, but at least part of the shift may also be attributable to changes in car designs which make the striking vehicle safer, to increases in the use of restraint devices, or to more days of adverse weather which might lower the speeds of traffic in general and therefore reduce the speed of collisions. While these rival causes are, to some extent, obvious, there are others that researchers can overlook. As in all evaluations, it is incumbent on the researcher to demonstrate that these alternative causes have not affected the situation being examined. In this case, the design does not provide help in ruling out these rival explanations.

Maturation. This threat can occur when the evaluator is unaware of trends because he only measures at two points in time. For example, assume that the situation depicted in Figure 3.5 exists. Here the researcher has measured accident frequency both before (B) and after (A) the treatment implementation. There is no question that a decrease in accidents equal to $Acc_B - Acc_A$ has occurred. (For present discussion, let us assume that this difference is not simply a random fluctuation of the data. This possibility is discussed as instability.) In the absence of other information, the evaluator concludes that the total difference, $Acc_B - Acc_A$, is the result of the treatment. However, this conclusion implicitly assumes that the before measure, B_B , is representative of the expected accident level in the future if no treatment were introduced (i.e., the underlying accident trend in Figure 3.6 is assumed).

This is an appropriate point to reemphasize an underlying component of all evaluation. We are attempting to measure the difference between what we observe (after treatment) and what we would have predicted to occur in the absence of treatment. (While the term "expected" could be used here, "expected" has other connotations in statistical analysis techniques, particularly in use of the χ^2 statistic. For clarity, we will use the term "predicted.") Thus, with the simple Before/After design, because we are limited to one observation point before treatment, we are forced to assume that the observed value also represents the predicted value -- what would have been observed in later time periods if the treatment were not implemented.

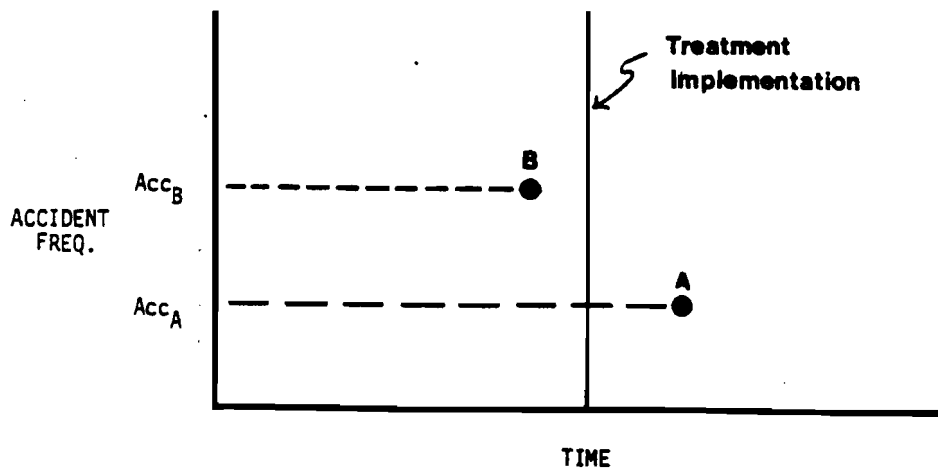


Figure 3.5

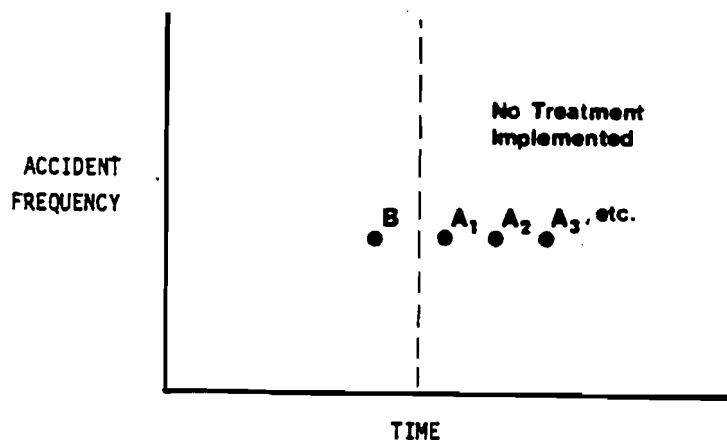


Figure 3.6

But what if the true situation (which is unknown to the Before/After user) is the one illustrated in Figure 3.7? Here, a decreasing accident frequency in the past would certainly lead us to a different estimate of predicted level at Point A than we assumed previously. We would predict that, even without treatment, Measurement A should fall somewhere close to the extension of the dotted line. Thus, a rival explanation for all or part of the decrease from B to A is maturation; the simple Before/After design cannot discount this threat.

Regression artifacts. Assume now that the treatment example used above is further complicated by the fact that the location selected for treatment was chosen on the basis that Year B was a high accident year. Given the initial situation in Figure 3.5, it would still be necessary to conclude that the change in accidents was caused by the treatment. However, assume that B was collected in 1972 and A in 1973, and that the underlying distribution is the same as the one shown in Figure 3.4. By superimposing Figure 3.4 on Figure 3.5, we get Figure 3.8. Recall that, in the earlier discussion, we noted that the decrease between B and A was the result of the regression phenomenon. Thus, even though a treatment was introduced, any

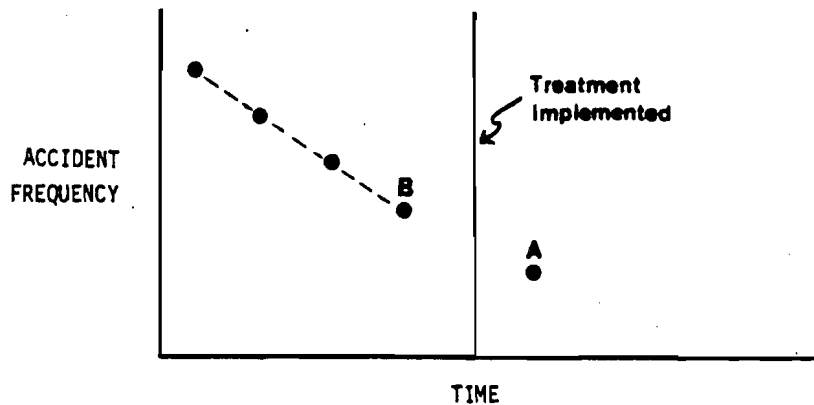


Figure 3.7

difference between the two points would, to some unknown degree, have to be attributed not to the treatment but to the regression explanation. (In this particular example, given that we "know" the underlying pattern, the total decrease is due to this threat and to random variation.) Again, the simple Before/After design does not help eliminate this rival explanation of the change.

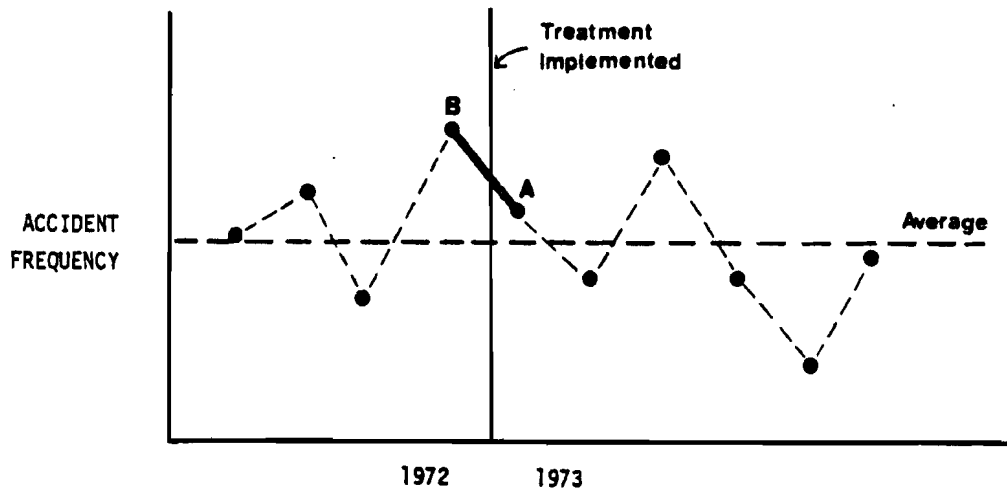


Figure 3.8

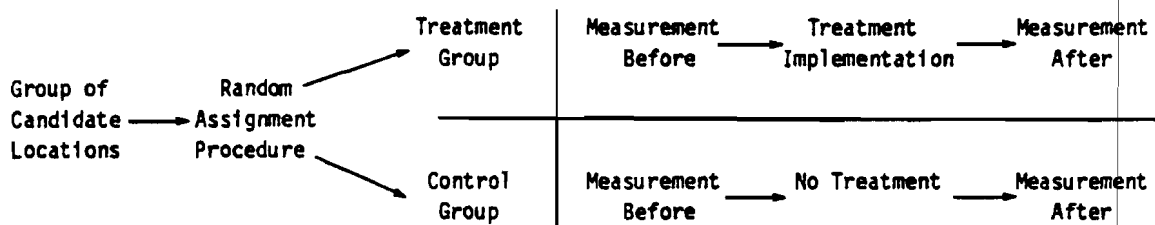
Instability. Here, at last, is a threat to internal validity that does not plague the Before/After design (at least not more than it does other designs). Given sufficient samples of accidents at B and A and given that certain assumptions concerning the underlying distributions are met, the choice and use of the proper statistical technique can help eliminate the threat that the instability or randomness of the accident data was the cause of the change. However, the point of importance is that even if a "statistically significant" difference is shown, we will continue to be ignorant of the true cause(s) of this difference. While these statistical "gate keepers" can keep out differences so small that they might be caused by chance, they will allow a large difference in, regardless of whether the cause of the large difference is the treatment or some other rival explanation.

In summary, the Before/After design, although quite simple and almost always available, is a very poor design. Consequently, highway administrators and evaluators should carefully consider the idea of delaying or eliminating evaluation if this is the only design available. Far fewer erroneous conclusions would be generated and the money saved could be applied to more worthwhile studies.

We will now turn our attention to much stronger designs, designs which can help control rival explanations and which, with proper planning, can be implemented in the real world of highway programs.

3.7.1b Before/After with randomized control groups.

We will now move from a very weak design to a design which is one of the strongest available to the evaluator, the Before/After Design with a Randomized Control Group. This design is similar to the simple Before/After design in that measurement is made before and after a treatment is implemented. However, it is quite different both in terms of the mechanism of comparison between the observed and predicted levels of the criterion and in the fact that it must be planned for before the treatment is implemented. The evaluation model is depicted below:



Here a group of locations which are candidates for a given treatment are first randomly assigned to either a treatment group or a control group. The mechanism used for this assignment could be the flip of a coin or the use of a table of random numbers. The underlying requirement is to give all locations the same chance of receiving the treatment. The purpose of this random assignment is to attempt to make the control and treatment groups equal on all factors except for the execution of the treatment. (The two groups do not need to be of equal size. The size of the smallest group will be determined by sample size requirements discussed in Section 3.8.3.)

After this random assignment, the before measurement is made (in truth, in highway accident evaluations, this before measurement may be legitimately made on data collected even before randomization). Next, the treatment is implemented at locations in only one of the groups, and then the after measurement is made in both groups. Although the data collection intervals can be different for the before and the after periods, data must be collected identically in both periods for the treatment and the control groups.

With this design, unlike in the case of the simple Before/After design, the "predicted" level of the after measure is based on the experience of the control group. That is to say, because all other factors have been "made comparable" through randomization, we predict that, without intervention, the treatment group will behave exactly like the control group. (The reader will recall that this was not the case with the simple Before/After design, where the evaluator was forced to predict the future based only on the before measurement.) Thus, any changes that occur in the treatment group, besides the ones related to the intervention, (i.e., rival explanations of the effect) can also be expected to be reflected in the post-treatment measurements of the control group. Thus, the Before/After Design with a Randomized Control Group controls for the threat of history, or other causes occurring at the same time.

In similar fashion, any changes that underlie the normal accident trends (i.e., the threat of maturation) should equally affect each group because both are representative of the same original larger group. Effects due to maturation should be reflected in the post-treatment measurements of both groups.

The same holds true with the regression threat. Even if the original larger group was composed of "high-accident locations", and therefore even though we would expect each location's accident frequencies to regress toward its true mean, there is no reason to expect this regression to occur differentially in the control and treatment groups. Even though regression may affect the measurements of the treatment group after the intervention, its effect can be expected to equally affect the control group.

An example of the interpretational difference between the simple Before/After and the Before/After with Randomized Control Group is illustrated with Table 3.2. Assume this table presents the ran-off-road accident frequencies observed in a study of hazardous curves on two-lane rural roads where raised delineators (reflectorized paddles) were installed on the shoulder of the curves.

Table 3.2. Ran off road accident frequencies on hazardous curves on two-lane rural roads.

		Period	
		Before	After
Group	Treatment	100	30
	Control	110	40

First, assume the evaluator employed a Before/After design. In this case, only the top line of the table could be developed. Based on these data, he would conclude that the treatment had reduced ran-off road accidents by 70 percent (i.e., $(100-30)/100$).

Now, assume that the evaluator planned ahead of treatment implementation (since he was privy to the next year's budget, he knew in advance that such treatments would be introduced), doubled his group of potential locations and randomly assigned individual locations to the treatment or control groups. This would enable him to build the second line of the table. Here, the predicted experience for the treatment group should be based on the experience of the control group. In this case, the control group experienced a 64 percent decrease in ran-off-road crashes from some combination of unknown causes. Using this, we would predict a similar decrease in the treatment group if no treatment were introduced. The predicted after frequency would be $100 - (.64)(100) = 36$ crashes. The observed number of crashes was 30. Thus, (without regard to statistical significance at this point) we do see an apparent effect of the delineators. But the effect is approximately 17 percent (i.e., $(36-30)/36$) rather than the 70 percent indicated earlier by the Before/After design.

Conversely, there will be instances in which the treatment group rates stay the same (or increase slightly) while the control group rates increase greatly due to other factors. In this case, what would be interpreted in a Before/After study to be "no effect" or "very limited" effect could well be shown to be a larger significant effect by a Before/After with Control Study. This occurred in a study (Foody and Taylor, 1966) in which a simple Before/After analysis would have indicated a 6 percent reduction, but the better design indicated a real reduction of 15 percent, a large difference in terms of crashes reduced. The point is that the decrease in the control group could have resulted from any number of other factors, including maturation, regression, or history. These rival explanations, however, can be accounted for by incorporating the randomized control group into the basic Before/After design.

Appropriate statistical tests for this design. In terms of appropriate statistical tests, two approaches are possible. In both cases, the evaluator should first compare the before measurements (averages, rates, proportions, or distributions) to see whether the treatment and control groups are indeed equal. Although randomized assignment will cause the control and treatment groups to be equivalent if the original sample of locations is large enough, the fewer the number of locations in the overall potential treatment group, the higher the probability that the assignment procedure will not "equalize" all other factors. (The user should refer to Table 3.3, Section 3.8.4 for proper statistical procedures.) However, if the before groups are equivalent, the evaluation can compare the after measurements to see if significant differences exist using the statistical procedures cited in the same table.

Discussion of control groups. The above design, although one of the most powerful ones available to the evaluator of highway related countermeasures, is very seldom used. Frequently, the objections to using it are:

- 1) My treatment has to be implemented jurisdiction-wide, or
- 2) It is (morally, ethically) impossible to identify control locations and to leave them untreated.

First, if the treatment must truly be implemented on a jurisdiction-wide basis, then this design is indeed not possible (in this case, the user should refer to the discussion of Interrupted Time Series designs). However, as discussed below, even when a treatment is to be implemented on a jurisdiction-wide basis, there are times that the treatment cannot be implemented at the exact same point in time due to construction scheduling, equipment purchase, etc. In these cases, suitable control groups can be identified.

The second point is even more important because it refers directly to the underlying rationale for evaluation. This argument is true only if we know the treatment works (in which case there is no reason to evaluate it!) Control locations can be identified just as treatment segments can be identified. It requires that the evaluator be totally committed to making his evaluation work, and thus, willing to do the extra work involved in carefully identifying the "extra" sites. It also requires that the evaluator have some input into the selection of the treatment implementation scheme. He both should know what is planned and should have the authority to suggest implementation sequences. This of course means that the administrator must also be totally committed to the idea of evaluation ("educating" the administrator to espouse the correct philosophy may be the single most important task the evaluator faces).

While the problem is not an easy one, there are two underlying "givens" in any highway-related program implementation process which the evaluator can use in his search for controls. Again, it is stressed that these will not be useful unless the evaluator makes the effort to become part of the improvement planning team. The first "given" is that there are never enough safety dollars to improve all candidate locations. The second "given" is that, except for changes in laws or regulations, improvements at all sites cannot be made at exactly the same point in time.

Budgetary constraints may, at times, be the savior of proper evaluation. In almost all cases, a situation will exist where the careful engineer/evaluator, using the best tool for choosing potential treatment sites, will identify more locations than his departmental budget can possibly treat (not necessarily twice the number, but at least more than he can treat). Given that too many locations exist, and given that we do not know whether a certain treatment works, there is no "ethically fairer" method for deciding which sites receive treatment than random assignment. Every site has an equal chance of getting the limited dollars; no location is unfairly discriminated against. Indeed, knowing that he will be correcting 50 sites with one treatment (e.g., signalization), 40 sites with a second (active grade crossing protection) and 100 with a third (replacement of non-breakaway sign supports), the evaluator could identify too many sites in each category, assuring himself of potential pools of treatment and control sites.

The second aid to the designation and protection of control sites mentioned above is the time always necessary to completely implement a wide spread improvement -- the staging sequence. This time lag can aid the careful evaluator even in cases when enough funds exist to treat all locations because, at any point in time, there will be some candidate locations which have not yet been treated and are accumulating accidents at the same time as the treatment group. These locations can serve as controls. This is particularly useful for special projects in which the entire job will have to be done by the same work crew, say from the central office. With planning, it can also be useful in cases where different crews, say in each of several state highway divisions, will be implementing the treatment. It should be noted that the use of staging also helps overcome any legal commitment to treating all "needy" locations. In this procedure, all are treated, fulfilling any liability requirements, but are treated in such a sequence that useful information can be gained, fulfilling a research requirement.

Consider the case of replacing non-breakaway sign supports on all four-lane divided highways in the state. Such a large scale project could, undoubtedly, take one to three years (or more) to complete. First, all new supports will not be delivered at the same time. Second the work will be competing with, and have to be scheduled into, other work schedules. Rather than allowing each division to implement the treatment as they see fit (and hope for a "natural" random scheduling), the careful evaluator could randomly choose the time of implementation for whole divisions or for highway segments within division (the latter is the better strategy). Using central control of supply distribution and headquarter's control over work schedules, such a scheme might be possible. Note, however, that the evaluator would have to collect careful records of when and where the treatment was implemented to determine when a given segment has been shifted from the control to the treatment group.

A special case arises with respect to high-accident locations. Here (unfortunately from an evaluation standpoint), the inquiring office will not only have chosen such locations, but have, through some mechanism, "ranked" them according to "need." In these cases, the administrator is less likely to agree to the random assignment of treatments, citing ethical (or legal) grounds for treating the most needy first.

Two points are raised with regard to this situation. First, the best methods we have to date for ranking high accident locations are just that. They are the best to date. They will continue to be modified based on information gained from studying relationships. Thus, our rankings are far from perfect and should not be considered definitive yard sticks of "most needy." This lack of a perfect hazardness indicator is clearly shown in a study by Carlson (unpublished) reported by Taylor and Thompson (1977). Carlson obtained information on the hazard index formulas used in various states and combined similar ones into a total of 13 formulae. He then collected accident and other data on 15 sites in Pennsylvania and applied each of the 13 formulae to rank each site with each formula. While some degree of agreement between formulae existed, there were major differences in the rankings calculated for identical sites. For example, for three different sites, ranks ranged from one to 13. Thus, the hazardness formulae used by different states obviously are not consistent.

The second point concerning this situation is that, if we "know" that a location is "most needy," and if we "know" that our treatment will improve the situation, evaluation is not needed in the first place since the only goal of evaluation is to determine if and to what degree a treatment works.

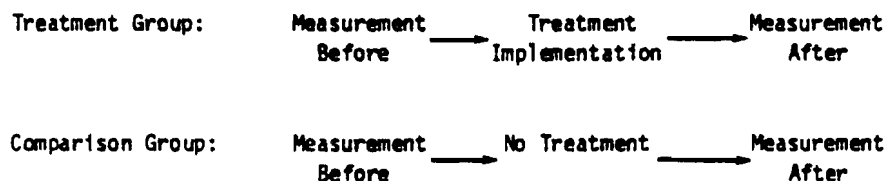
If we, as an engineering community, examine our knowledge, however, we realize that we very seldom "know" what works. We are having to guess. If we must guess, we can fall back on the philosophy of "the problem is serious. We do not know what will improve it, but we plan to try Treatment A on some locations, Treatment B on some others, and Treatment C (a no-treatment or low-level treatment) on the remainder."

However, it would be naive to assume that all administrators (or evaluators) will embrace this philosophy. The issue is not a simple one, particularly in that the legal system is increasingly judging the engineering accident-related processes. For this reason, two of the designs which might be of use with high-accident locations are described in later sections. (Sections 3.7.1h, 3.7.1i).

Finally, to conclude this discussion of Randomized Control Groups and to provide evidence that such designs can be and have been implemented in the field, readers should consult two studies of pavement edge marking conducted in the early 1960's (Musick, 1960; Basile, 1961). In the second study, Kansas was moving toward a statewide edgemarking program. Two earlier Before/After studies had indicated a 21 percent reduction in accidents and a 59 percent reduction in fatalities. However to better control extraneous factors, in 1959, twenty-nine pairs of equivalent sections of rural highway (384 miles) were chosen for further experimentation. The pairs were then randomly divided into treatment and control sections, the treatment sections were marked, and accident data were accumulated for one year. In contrast to the findings of the earlier Before/After study, the treatment group study indicated a non-significant 1 percent increase in accidents on the treated sections over what would have been expected from the experience of the control group. (This indicates that the most crude Before/After study fell victim to one or more of the aforementioned threats to validity and resulted in the wrong answer.) Basile did, however, find a significant 46 percent reduction in crashes at intersections and driveways. No significant change was noted between access points. Thus, through careful planning, the establishment of a Randomized Control Group was plausible. (The only possible criticism of the study might be in the choice of criterion. It is left to the reader to decide if certain classes of accidents might have been more appropriate.)

3.7.1c Before/After with comparison group.

A variation of the Before/After with Control Group design is the Before/After Design with a Non-random Comparison group. A diagrammatic representation of this design is presented below. The only difference



between this design and the previous one is that the groups are not assigned on a random basis. This design appears very appealing since, with good historical records, it would be possible to choose a comparison group even after implementation. However, the lone difference (the lack of random assignment) causes major differences in the relative strength of the two designs. This is usually the case because even careful

choice of comparison locations will not completely assure that the two groups were entirely equal before treatment. However, this design is much stronger than the simple Before/After design. The strength of the design is directly proportional to how similar the treatment and control groups are. Thus, in using this design, the evaluator should always carefully compare the measures for the two groups in the before period.

Even the similarity of the groups in the before period, however, cannot completely insure that the treatment group, without the intervention, could have been expected to act exactly like the control group. Other unidentified factors can still result in differences. Partly responsible for this problem is the fact that in most of the instances in which this design is employed, the treatment group will have been chosen because of "demonstrated need" (e.g., high-accident locations). Thus, by definition, if our ranking device is accurate, any comparison groups which remain after the treatment group is chosen are automatically "lower in need" than the treatment group, and the factors which make it "lower in need" could also cause it to respond differently across time than the treatment group. Cook and Campbell (1976) present a series of alternative outcomes to an evaluation conducted with this design and discuss the threats to validity which are more bothersome in each outcome. The four outcomes most likely to occur in the evaluations of highway countermeasures are shown in Figure 3.9.

First, each of these outcomes could be interpreted by the evaluator as a treatment effect. That is, since in B, C, and D, the treatment group improves more (accidents decrease more) than does the comparison group, this difference must have been caused by the treatment. In A, a treatment benefit could also be hypothesized in that the treatment more effectively kept the "normal growth rate of accidents" under control. However, these outcomes could also be explained by the rival explanations of "local" history, differential maturation, and regression. Thus, they represent interactions between the threats and the possible treatment effects. For example in A, while we would not expect regression to be a plausible explanation because the "high risk" treatment group does not regress downward, we would have to agree that the pattern could be caused by "local" history (i.e., these are some factors that affect one group but not the other one). For example, in a study of pavement grooving, if the comparison location experienced more wet weather, this pattern could emerge without any treatment effect. Second, the two groups could have underlying trends over time which differed. If the pavement at the comparison locations was a different composition and thus was becoming more slippery at a faster rate, this pattern could result. Patterns B and C are more likely to be observed, especially if the treatment is given on the basis of "need." With both of these patterns, in addition to the threats of local history and differential maturation, we must also now suspect regression artifacts (since the high accident treatment group is indeed regressing toward the "normal" comparison group mean).

Pattern (D) is interpretable, but only if the treatment group's measurement before the intervention is significantly higher than the comparison group's and the treatment group's measurement after the intervention is significantly lower than the comparison group's. This is because the rival explanations appear less likely to explain the pattern. For example, while other factors (local history) might cause the treatment group's measurements to drop, unless these causes are very strong (and do not affect the comparison group at all), they would not be expected to cause the treatment group's after measurements to cross over and be significantly lower than the control group's measurements. In like manner, while regression might explain a decrease toward the "normal" mean, it would not explain a decrease to a level significantly lower than the normal mean. Finally, study of maturation effects historically indicates that accident data trends, even though different, would not be expected to "crossover" significantly.

Even though this outcome can be interpreted, two problems remain. First, this outcome is very unlikely to occur in real-world accident studies; A, B, or C is much more likely. Second, there is no way of estimating how much of the decrease is due to the treatment and how much to the combination of rival explanations. The careful reader will have noted, however, that the above criticisms of the design are all keyed to the basic issue of comparability of groups. If groups can be found that appear comparable under very close scrutiny, this design becomes a very strong one.

One excellent example in which this comparability can be found by the careful evaluator is cited in a study by Foody and Taylor (1966). Working for the Ohio Department of Highways, the authors were examining the question of whether Ohio's policy of placing raised markers (delineators) on the outside shoulders of horizontal curves resulted in accident reductions and was therefore cost effective.

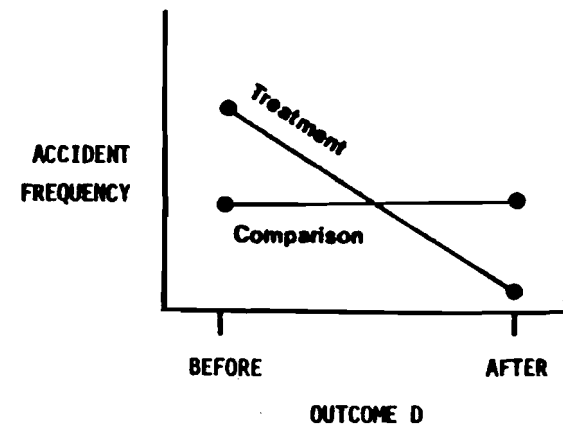
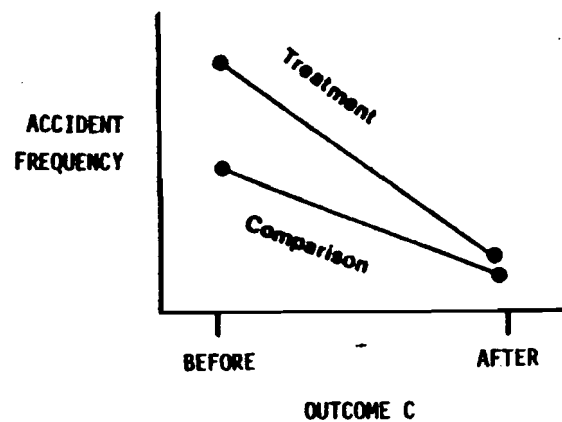
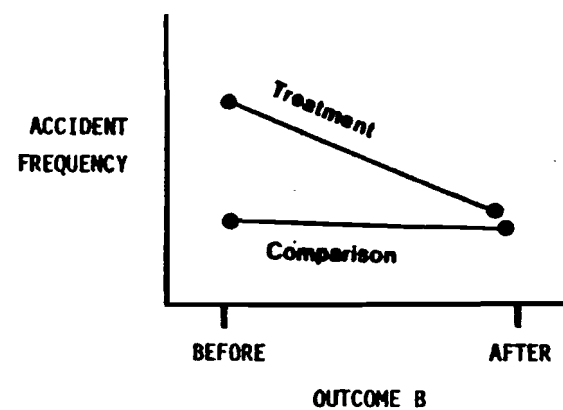
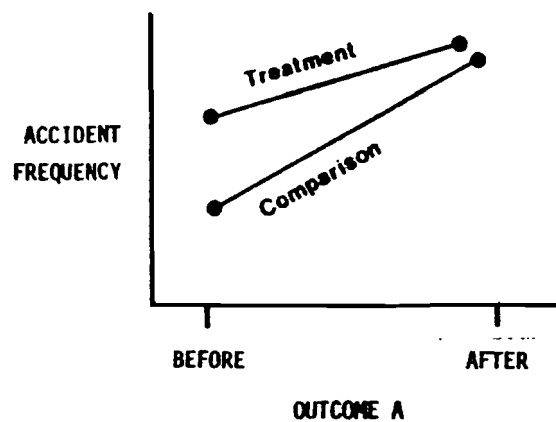


Figure 3.9. Possible evaluation outcomes in a Before/After with Comparison Group Design.

Four years' data were available for analysis. Of the 914 curves which met the criteria for delineation and were originally chosen to be delineated in the program, only 557 had been delineated during the first two years of the data sequence. Thus, because of real world implementation schedules, the authors were able to identify 557 test curves which had been delineated and 357 control curves which had not been delineated but which met the same criterion as the test sections did. Even though the study was not planned ahead of time, the authors found it possible to identify a situation in which both treatment and comparable control sections did exist. Changes in accident frequencies at the treated test curves were compared with changes in the control sections and the study indicated a 15.5 percent reduction in the frequency of all accidents on the treated curves when compared with predicted values based on the experience of the control curves. Because this design was basically a Before/After with Comparison Group design, the confidence placed in the results of the study were much greater than if the design had simply been a Before/After design. By choosing this design, the authors helped guard against the threats of (1) other causes occurring at the same time, (2) trends in the data, (3) regression artifacts, and (4) selection bias: because all the curves were chosen using the same criteria, any changes (other than those caused by the treatment) should have equally affected both the treatment and the control sections.

Matching. In many cases in which an evaluator is attempting to build a comparison group after-the-fact, he will be tempted to resort to some form of matching, in which a location in the treatment group is "matched" to a location from the remaining population on the basis of other factors (e.g., prior accidents, ADT, number of lanes, pavement width, etc.). Only the matched locations are then placed into the treatment and control categories and analyzed. This is done in an attempt to "equalize" the two groups. Unfortunately, although it appears logical, this practice is erroneous. While Campbell (1963, 1975) states that in research involving social programs, matching on the before measure always undercorrects and thus would lead to erroneously thinking the treatment has an effect, the reverse could be true in highway research.

For example, assume that the treatment group is a number of high-accident locations. (In contrast to the "pseudo" high-accident locations discussed in earlier sections, these locations have a long history of a high level of crashes.) The evaluator searches his computerized file to find the untreated locations which are similar to the treated ones, especially on the basis of similar accident frequencies in the one-year before treatment. Observe what is occurring: if it is assumed that the high-accident locations are correctly chosen (i.e., they actually have had a higher average frequency or rate), the situation for the before period is the one depicted in Figure 3.10.

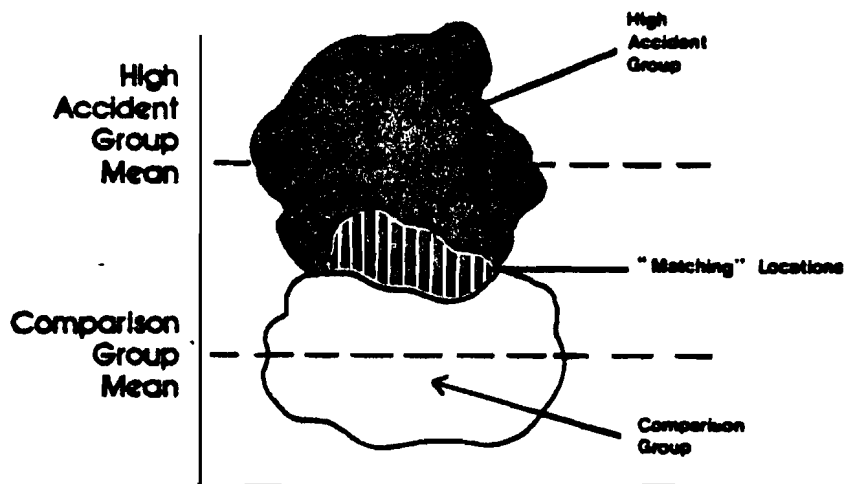


Figure 3.10. Before period crashes.

The locations that would match on the basis of accidents in the one-year before period are those high-accident locations whose frequency deviated downward from their group mean in that year and those comparison group locations whose frequency deviated upwards from their true group mean in that year. But these are deviant points in the two groups. Remembering the previous discussion of regression artifacts, what is likely to happen to the accident frequencies for these locations in the after measurement period even without treatment? They are likely to regress toward their individual group means. As shown in Figure 3.11, the matched high accident locations would get worse and the matched comparison locations would get better because of regression.

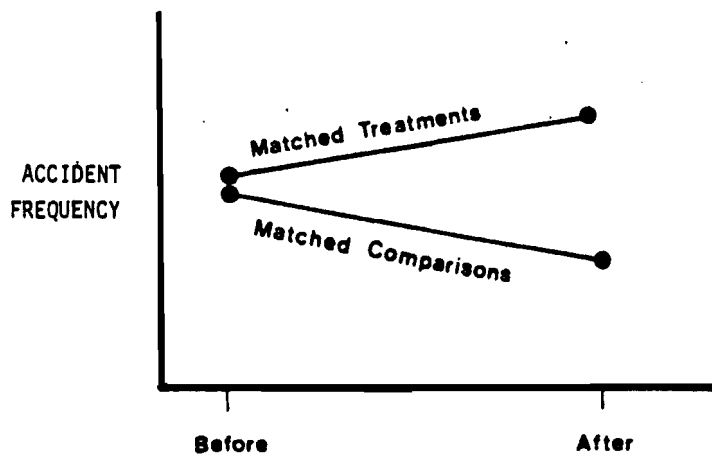


Figure 3.11

Thus, if a treatment had no effect or even a slight positive effect, the evaluator might conclude on the basis of the observed pattern that the treatment causes accidents (as in the opening "situation"). This could be most damaging in that while treatments which show little or no effect might be tried again, a treatment that appears to have a negative effect would probably be permanently abandoned.

Even if factors other than the Before accidents are used to match, not all of the inherent differences are likely to be controlled for. At least some part of an observed difference could still be due to the rival explanations. As a last choice, however, matching other factors can at least help overcome some of the differences.

In all fairness, it must be noted that if the above matching procedure was used, and the treatment group "overcame" the natural regression tendencies and was significantly better than the control group in the after period, we would probably conclude that the treatment was effective. However, just as with outcome D discussed earlier, this is very unlikely to happen. When it does, the level of treatment effect will be almost impossible to define.

[NOTE: There is a great difference between this matching procedure and one in which locations are matched and then randomly assigned to treatment and control groups before the intervention. This latter procedure can help strengthen even the Before/After with Randomized Control Group design. The key, however, is the random assignment after matching.]

In summary, the Before/After with Comparison Group design can be a weak design or a very strong design, depending entirely on the evaluator's ability to isolate a truly comparable comparison group. In every case, it is better than a simple Before/After design. In some cases, such as in the example provided earlier, it can be nearly as sound as the Randomized Control Group design.

3.7.1d Interrupted time series designs.

A relatively powerful group of designs that involve multiple observations of the criterion both before and after an abruptly introduced treatment is the Time Series (or Interrupted Time Series) designs. The format is presented below. Here, "X" represents the implementation of the treatment while each "M" represents a point in time when a measurement of the criterion variable is made.

. . . M M M M M X M M M M M . . .

Time Series designs are most useful to the engineer/evaluator when he must evaluate a treatment (such as a law change) which is implemented at one point in time over his entire area. In such a situation, the possibility of defining control or comparison sites does not exist.

The design can also be very useful when comparison sites do exist and should be utilized much more often in the highway research area than it has been. A major point in favor of using this design is the

fact that, unlike in other research areas (e.g., education, social aid program) and even in the driver safety area, the evaluator/engineer will quite often have at his disposal a long crash history for a given location prior to treatment.

An example series illustrates the advantages of this design in terms of control over internal threats. Assume that the time series illustrated in Figure 3.12 represents the monthly crash frequency of all four-lane rural roads in a given state. Further assume that the speed limit on each of these roads has been 65 mph for the four years, 1973-1976.

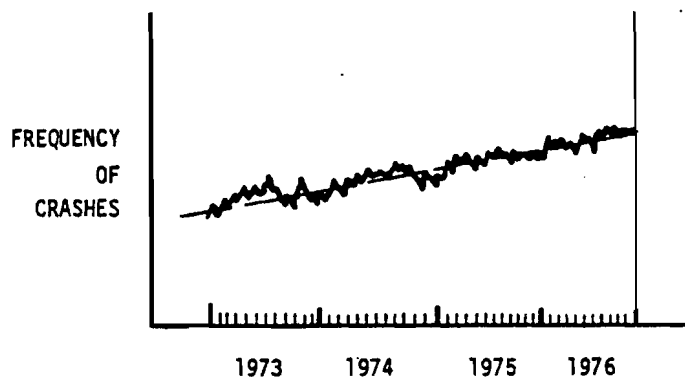


Figure 3.12

As can be seen, while the series has random shifts, and even some cycles (i.e., lower during the winter months), it is fairly consistent in its gradual increase and could be represented by the broken regression line. Further, assume that the state legislature passed a law which reduced the speed limit on all four-lane roads to 55 mph as of January 1, 1977 (the treatment or intervention). The evaluator now plots the monthly frequencies for 1977 and 1978 to determine whether the treatment has been effective.

First, it's necessary to determine what changes in the series indicate a treatment effect. Basically, the resulting series could (1) remain the same, (2) shift downward (or upward) from the original series without a change in slope, (3) experience a change in slope but no shift, (4) experience both a shift and a change of slope, or (5) experience a later shift and/or change in slope. These alternatives are illustrated in Figure 3.13.

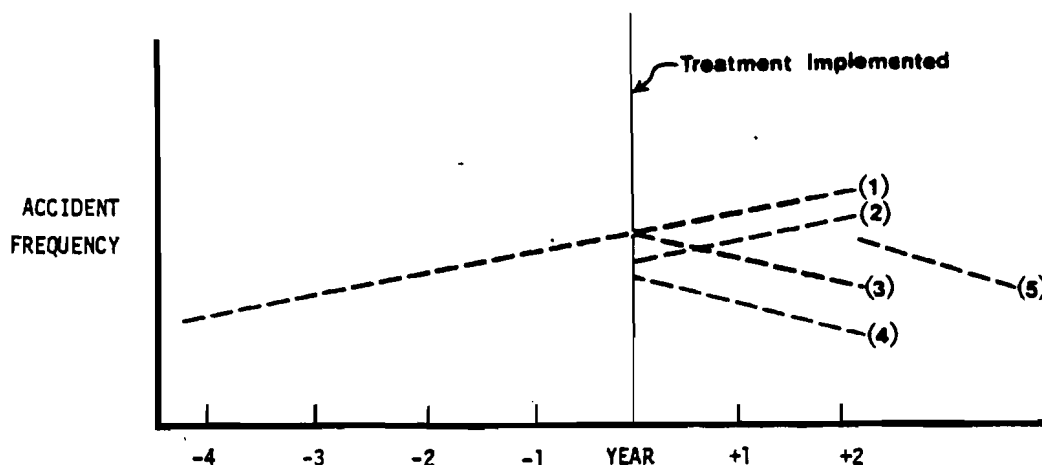


Figure 3.13. Possible evaluation outcomes with a Time Series design.

Pattern (2), (3), or (4) signifies a change; because there is no reason before hand to think that the change in limit would cause a delayed effect, (5) does not indicate a treatment effect.

Having now decided how to interpret the series, let us examine the advantages and disadvantages of the design. Of the threats we have discussed, only one, history, appears to be a plausible rival explanation of effect. That is, other causes which occur at the same time as the limit change (such as a fuel shortage) cannot be discounted by the design. All the other threats can. Maturation is ruled out by not accepting (1) as evidence of change. We now have knowledge of the underlying trend and can control for it. In like manner, regression is ruled out since the interpretation of an effect is based on shifts from a long term "average" trend rather than on a shift from one (possible deviant) measurement.

This design is a strong one. There are certain suggestions which the evaluator should follow when using it, both to insure that his interpretation is not hindered by normal cycles in the accident experience and to help overcome the threat of history.

- (1) Use a long time series to insure that all natural data cycles are accounted for. For example consider an evaluation in which a treatment was implemented in mid January. If weekly accident frequencies gradually decrease in the winter and increase as spring arrives, an October-March time series might indicate a treatment effect which in reality is due only to the calendar cycle of the accident frequencies.
- (2) Use measurement points that are as close in time as possible to help minimize the chance that another cause occurred at the same time. Monthly observations are recommended. For example, while many historical causes could occur over the course of a year, few have the opportunity of occurring in a given month.
- (3) Do not use this design when the treatment is not abrupt. It will not be meaningful. For example, if an evaluator is studying edge marking, and it will take a year or more for all his sample to be marked, he should instead use a Before/After Design with Controls gained from the staging. The effect of the treatment, even if present, will be delayed and thus easily confused with effects from other causes.

Statistical analysis techniques. For years, there have been problems in the statistical interpretation of Time Series: because the adjacent points are more highly correlated than remote ones, the standard least-squares regression analysis is not appropriate. Recently, however, new techniques have been developed to overcome this problem (see Table 3.3 Section 3.8.4).

3.7.1e Time series with comparison groups.

As noted in the introduction to the basic Time Series design, the design can also be used in cases where the treatment is not jurisdiction-wide. Just as with previous designs, this design can be further strengthened by including comparison groups, as illustrated below.

Treatment Group: . . . M M M M M X M M M M M . . .

Comparison Group: . M M M M M M M M M M .

Actually, this is the Before/After with Comparison Group Design with multiple before and after measures. By using these multiple measures (which are often available), the evaluator is able to "buy" protection against two of the rival hypotheses noted in the discussion of the Comparison Group design (see Section 3.7.1c). Both differential maturation and regression artifacts are controlled for. First, although there may be different underlying accident trends, this design indicates to the evaluator whether these trends exist and need to be accounted for. Second, as explained in the above section, regression is not a threat to any Time Series design since interpretation is not based on one (possibly deviant) before measurement.

The use of comparison groups can help control for the single remaining threat, history, if the comparison groups are carefully chosen. The discussion presented in the earlier section remains appropriate. Again, how well the threat of history (or local history) is accounted for is directly proportional to the equality of the groups. In contrast to the Before/After Comparison Group Design, this design can provide even more control over history by using short measurement intervals. In terms of practicality, this design is essentially possible every time a Before/After with Comparison Group is

possible. Because it is a much more powerful design, it is strongly recommended for use by the evaluator/engineer.

Statistical analysis techniques. Campbell and Stanley (1963) suggested that when this design is used, the differences between the pairs of measurements (Treatment versus Comparison at the same point in time) be calculated, and these differences (in crashes) be analyzed as a Time Series. The more appropriate technique may be to analyze the series separately and compare parameter shifts (see Table 3.3 Section 3.8.4).

3.7.1f Time series with comparison variables.

A less powerful modification of the above described design is one in which the Time Series for the treatment locations is compared to one or more series for the other independent factors or variables that may be associated with a history threat. Here, the comparison series is not other locations, but is instead any other factor which might be a possible alternative hypothesis. The design would be the same one illustrated on the previous page.

While numerous factors might be other possible causes in evaluations related to highway treatments, the main group of comparison factors might well be exposure measures such as ADT, million vehicle miles of travel, or the number of vehicles entering an intersection. For example, since a rival explanation for the decrease in fatalities after the imposition of the 55 mph limit was a decrease in mileage driven, an alternate Time Series of MVM could be plotted and compared to the accident series. Similarly, a Time Series related to the frequency of skidding accidents (where the treatment was the simultaneous implementation of pavement grooving at a series of locations) could be compared with a series of wet days per measurement period.

The major problem with this design variation is that the success of overruling the threat of other historical causes rests on the evaluator's ability to identify and develop data for all possible rival explanations, a major undertaking to say the least. At times, this can be done to a satisfying degree. The reader interested in an example of such a study should refer to Ross, et al. (1970).

3.7.1g Time series with switching replications.

A final refinement of the Time Series with Comparison Group design (Section 4.7.1f) would be to replicate the treatment phase of the evaluation by implementing the same treatment in the comparison group at a later time. This design is illustrated below.

Treatment Group: . . . M M M M X M M M M M M M . . .

Control Group: . . . M M M M M M M M X M M M M . . .

Here, the outcome would be a shift or slope change in the treatment group after treatment implementation with no change in the comparison locations, followed by a shift or slope change in the comparison group after the second implementation of the treatment. Theoretically, the second shift restores the groups to their original relationship. The additional strength of this design over the previous Comparison Group Time Series would be in making the threat of history less plausible. This is true because, while other causes may occur at the same time as the first treatment, the probability of a second "dose" of alternative cause occurring at exactly the same time as the second treatment only in the comparison group is quite low.

Although it seems complicated, this design can be practical for almost all the situations that lend themselves to Time Series with Comparison Groups or Before/After with Comparison Groups. The reason for this is that the second treatment would be imposed in only those cases where the first treatment appears effective. If the first treatment does appear effective, there would be little objection to treating the comparison group. Also, such a design could easily be built into these situations in which staggered implementation of the treatment made comparison (control) groups available. The only additional work would be the continuation of data collection, a very low price to pay for the additional power.

Statistical analysis techniques. While no specific statistical technique has been developed for this design, it appears that the design could be analyzed as two separate pairs of Time Series. The Time Series for the treatment and control groups up to the point of second treatment implementation could be compared,

followed by comparison of the pair of series extending from the first implementation to the end of the series.

In summary, because of the normal availability of long histories of data and the possibility of identifying comparison groups, the various Time Series designs should be strongly considered for evaluating highway safety programs. The additional power gain from use of these designs without a great amount of extra work or difficult randomization requirements places them in a position of great importance in the evaluation of highway-related treatments.

Two single treatment designs that can be useful in evaluating treatments applied to high-accident locations are the "Tie-breaking" design and the Regression Discontinuity design. While each has limited utility because of certain restrictions, both are included here for use by any evaluators who can surmount those restrictions.

3.7.1h "Tie-breaking" designs.

Actually, this "Tie-breaking" design is not a new design at all. Instead, it is a method that strengthens evaluations of high-accident location treatments by providing a means for randomizing the assignment of the locations to the treatment and control groups. The method was derived to accommodate similar situations in educational research, where there is a need to evaluate the effect that rewards (e.g., scholarships, Deans List designations, etc.) have on subsequent performance.

As discussed earlier, the current methods used in choosing and ranking high-accident locations are far from perfect; it is highly probable that many of the locations identified as high-accident sites are not actually high-accident locations, and that many of the ones that are not designated should be. This is particularly true for those locations which are near the cut-off point (the score or rank above which a location is designated "high-accident"). Thus, for example, if locations are ranked by "hazardness," and the top 200 are to be treated, there is a group of locations with ranks around the level of the 200th location (e.g., ranks 170-230) which are probably equally worthy of treatment. These locations could be considered "tied" at the cut-off point.

The "Tie-breaking" method simply admits the limitations of the rankings, spends 60-80 percent of the available budget to treat the top ranked locations, and earmarks the remaining 20-40 percent of the budget for those in the "tied" group. Because there is no incontrovertible way to determine the relative need of the remaining sites, it is possible to assume that they all deserve the remaining funds.

For the reader who studied the discussion of "automatic control groups," the next step is obvious: how else would a good evaluator/administrator break a series of ties except by random assignment? By grouping the locations by proposed treatments before breaking the ties, the evaluator may be able to generate randomly assigned treatment and control groups for at least some of the sites. (See Section 3.7.1b for the mechanics of the analysis.)

3.7.1i Regression discontinuity design.

The Regression Discontinuity Design is an attempt to exploit the cut-off point that differentiates between the locations that do and do not receive treatment. A regression line is constituted and examined to determine if a discontinuity (shift), which would be predicted by effective treatment, exists at the cut-off point. More specific to the case at hand, when identifying high accident locations, there are a series of locations which are ranked on past accident experience from best to worse. Only, those that fall above a particular cut-off point are treated. (This strategy assumes that the administrator is not willing to admit the fallability of his identification method and thus will not allow ties. Otherwise, the evaluator could use the "Tie-breaking" method described in the preceding section, which is both more powerful and much more practical.)

In this after-the-fact analysis, the evaluator collects two pieces of data for each location--the accident frequency or rate used in the ranking (the Before data) and the corresponding frequency or rate for the After period. Each location is then plotted by the two measures, as shown in Figure 3.14.

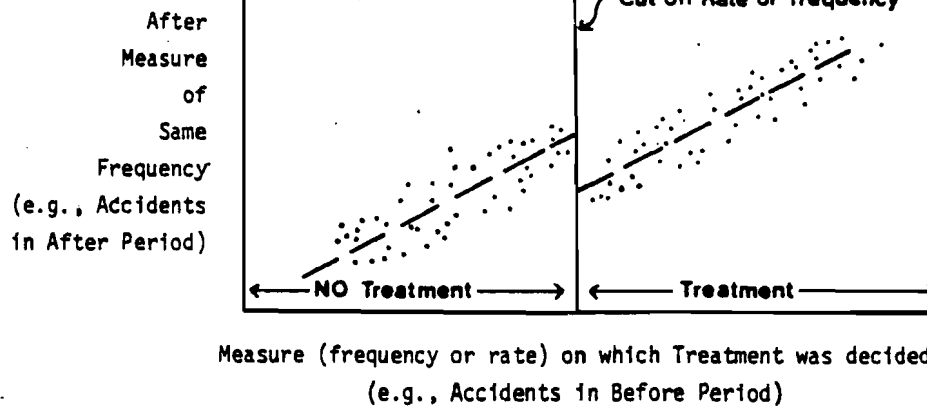


Figure 3.14

Thus, each point represents one location plotted by its original experience and later experience. If the treatment is effective, we would expect the treated group, falling to the right of the cut-off point, to have relatively fewer accidents than would be expected in the later time period. Thus, the above result indicates just such an effective treatment. We would expect just such a shift downward at the cut-off point. (We would hope not to get a change in slope, because this could indicate a confounding explanation--an underlying curvilinear relationship without a treatment effect.) Without an effect, the left line should have continued unbroken. Thus for the design to work, when the evaluator fits separate regression lines (see Chapter 4) to the "no treatment" and "treatment" locations, he expects that a treatment effect would result in a discontinuity at the cutting point. A full discussion of this design is presented in Campbell and Stanley (1963) and Cook and Campbell (1976).

In summary, the following restrictions are noted:

- (1) Both of the regression lines must be linear and parallel. Differences in slope could indicate an underlying curvilinear relationship rather than a treatment effect.
- (2) Multiple locations experiencing the same or similar treatments are required. Without adequate data points, especially in the treated, high-accident sites, little confidence can be placed in the fitted regression lines, and thus little confidence can be placed in any indication of a shift.

If the treatments used at the high-accident locations differ, it is possible to conclude only that the treatments as a group were effective, a relatively useless finding. This requirement for large samples of data points for the same treatment will, in all likelihood, require the evaluator to go back into his files to pull treatment-location data from previous years.

- (3) The cut-off point cannot be too high. If it is, there will not be a large enough horizontal spread of the points in the treatment group to allow valid fittings of the regression line.

As noted, this design has its limitations. However, in certain cases, with some effort it may provide useful information.

Statistical analysis techniques. Campbell and Stanley (1963) suggest that the most efficient test would be a covariance analysis, in which the cut-off point frequency or rate is the covariable for subsequent accident experience (see Table 3.3, Section 3.8.4).

3.7.2 Evaluations involving multiple degrees or types of treatment

The preceding section has presented detailed discussion of evaluation designs suitable for use with one treatment (i.e., treatment versus no treatment). The following narrative will present a more limited discussion of one design which can be used if multiple levels or types of the same treatment are to be evaluated (e.g., the various types of pavement grooving or various levels of railroad grade crossing protection), all of which are to be tested simultaneously. In discussion of this design, let us assume that

an engineer/evaluator wishes to determine the effectiveness of pavement grooving on accident reduction in rural locations. However, unlike in the preceding sections, rather than simply determining the effectiveness of one type of pavement grooving, the evaluator now wishes to simultaneously compare the effects of three types of pavement grooving (i.e., three different grooving patterns). In addition, the evaluator wishes to compare each of these three different treatments to a no-treatment condition, so that there is a total of four treatment conditions. If possible, the evaluator would also like to know the differential effects of each of the grooving types on curves vs tangent sections and on freeways vs two-lane roadways.

In attempting to find an efficient yet scientifically sound method for meeting his goals, the evaluator turns to one of many available texts on experimental designs (e.g., Cochran & Cox, 1950; Steel & Torrie, 1960). The first design which might come to his attention is known as the Latin Square design. In such a design, the treatments are assigned randomly to each of the various combinations in such a manner that other factors are controlled for. For example, in the situation above, the four types of locations can be divided among four time periods to form a four-by-four square. The four treatments (the three grooving patterns and the no-treatment condition, T_1 , T_2 , T_3 , and N, respectively) can then be assigned randomly to the first row, and then assigned to the second, third, and fourth rows so that the same treatment does not appear in a given row (or column) more than once. One of the possible assignments is shown below.

	Freeway		2-Lane	
	Curve	Tangent	Curve	Tangent
Time 1	T_3	T_2	N	T_1
Time 2	T_1	N	T_2	T_3
Time 3	T_2	T_3	T_1	N
Time 4	N	T_1	T_3	T_2

This is a "Latin Square." By assigning the treatments in this fashion, it is possible to discount the influence of other factors by using analysis of variance techniques. The advantage of this design is that it is one of the more economical and efficient designs available when multiple treatments are being studied. Only four location types were required (although better control will result from having more than one location in each cell).

However, there are problems with using this design for highway accident research. Foremost is the fact that each treatment must be applied to each location. In the example, this means that each location receives each of the four different types of treatments at different time intervals. While this can be done in some highway countermeasure studies, in most cases this is not feasible. For example, in the example under consideration, this would mean that a certain grooving pattern would be implemented at one of the location types, accident data would be collected, the grooving pattern would be physically removed and replaced by a new grooving pattern, accident data would again be collected, etc. Obviously, because of the earlier described problems in the collection of sufficient samples of accident data, it would be necessary that the individual time periods be rather long, resulting in a very long total time period for covering all possible treatments. This long time period can allow other roadway system changes to take place so that rival explanations for the observed effect may develop.

A second alternative design used in tests of multiple treatments would be a Factorial design. In fact, such a design is the natural outgrowth of the single-treatment information presented earlier. Without any knowledge of multiple treatment designs, the evaluator familiar with single-treatment designs might decide that one way of conducting the pavement grooving experiment would be to identify potential treatment locations on freeway curves and tangents. Then for example, within the group of freeway tangent sections, the researcher could randomly assign each of the types of pavement grooving to the locations and leave a set of locations untreated as a control group. The same procedure could be conducted on the potential freeway curve locations, and four-lane tangent locations, two-lane curve locations, etc. Comparisons then could be made of the accident experience of each of the treatments vs the accident experience of the control group within each roadway type/road character to determine whether any effect exists. In addition, to determine

the differences in levels (degree) of effect within each location type, the accident experience of each of the treatments could be compared to the remaining treatments.

The above process indeed describes a Factorial design. A way of visualizing such a design is shown below. Here the potential group of Freeway Curves (designated above by FC # 1,2,3,4) are randomly assigned to the treatments. While the Latin Square design required a minimum of only four locations, this design requires a minimum of four of each type for a total of 16 locations. In fact, the design is much stronger if there is more than one location of a given type in each cell. (This is termed "replication of the factorial" in statistical texts.) For example, if 12 potential locations on freeway curves could be itemized, three would be randomly assigned to each cell.

	Location Type			
	Freeway		2-Lane	
	Curve	Tangent	Curve	Tangent
Treatment 1	FC2	FT3	2-LC3	2-LT1
Treatment 2	FC3	FT2	2-LC1	2-LT2
Treatment 3	FC1	FT4	2-LC4	2-LT4
Treatment 4	FC4	FT1	2-LC2	2-LT3

Statistical analysis techniques. The analysis technique most appropriate for this design is analysis of variance (see Table 3.3). There will also be instances in which some uncontrolled factor may be felt to possibly hinder the analysis. In the above example, an unexpected large number of wet days at the locations in only one cell might result in a given treatment being interpreted as having no effect. In such cases, the use of such factors as the covariant in an analysis of covariance (see Table 4.2) appears warranted. These analyses will reveal whether or not there is an overall significant difference among the treatments. However, to determine whether the difference between pairs of treatments is significant (e.g., Treatment 2 vs Control, or Treatment 3 vs Treatment 1), the evaluator should use the Duncan's multiple range test or a comparable statistic (Cochran & Cox, 1950; Steel & Torrie, 1960).

In closing, the preceding discussion of evaluation designs has included those which appear to be most appropriate for use in the evaluation of highway-related treatments. There are undoubtedly numerous other designs and numerous variations to the designs presented which could be used. The choice is limited only by the knowledge and imagination of the evaluator. The point that has been continually stressed, however, is that the appropriate choice of design is based on reducing the plausibility of rival explanations for an observed effect.

3.7.3 Evaluation of the "multiple" improvement countermeasure.

The preceding sections concerning evaluation designs have been related to evaluations of single treatments at one or more locations and evaluations of multiple degrees or types of treatments at a number of different locations. However, there is another situation that is often found in highway accident research. This is a case in which the researcher is required to evaluate the effects of a "do all we can" treatment. For example, a high-accident intersection will be identified using existing procedures. The intersection will then be examined by a traffic engineer who will make an engineering judgment concerning the combination of improvements which appear to be most appropriate. These improvements may include signalization, channelization, repainting, addition of lanes, etc. The researcher is then asked to determine the relative effectiveness of each of the improvements.

Although quite usual, this situation poses a most difficult research task. While the overall effectiveness of the total package of treatments can be ascertained using the earlier cited designs e.g., Before/After with Comparison Group, Time Series design, Before/After with Control Group, etc.), the resulting findings can only be stated in terms of the overall effectiveness of the total package. The

parcelling out of individual effectiveness levels is virtually impossible, particularly if only a small number of intersections are treated.

If a large number of intersections with similar characteristics are treated with different combinations of the treatments, the researcher can attempt to parcel out the relative effectiveness of each of the treatments by using a form of regression-type analysis. The individual treatments would be used as independent or predictor variables for accident frequencies or rates, and tests of the relative difference in the coefficients would give some indication of their relative importance. However, use of such a technique is far from optimal and, in fact, is discouraged by some statisticians working in the accident research area. Regression analysis has problems in defining causal relationships due to the inherent errors in measuring the exact extent to which each treatment is involved and in defining and including all other possible predictor variables which could affect accident rates. Unless these other predictor variables are included, observed differences cannot be attributed to the treatments.

Thus, in the "do all we can" situation (a situation which the researcher should attempt to prevent if he is interested in measuring the true effectiveness of the individual countermeasures), the best that can be done is to determine the overall effectiveness of the package of improvements. Again, little benefit can be obtained from such a measure since little information can be parceled out concerning individual treatments.

3.8 Statistical Procedures for Evaluating Countermeasures

The preceding sections related to the evaluation of countermeasure treatments have explored the threats to the validity of an evaluation and the strengths and weaknesses of alternative designs to overcome these threats. In the remainder of this chapter an overview of statistical techniques useful in such evaluations of countermeasures will be presented. It is again stressed that even though this material is included as a major section of the manual, the strongest statistical technique can only overcome one of the threats to validity, the threat of instability. Only proper experimental design can overcome the threats of history, regression, maturation, or related problems. Thus, the following discussion will assume that the researcher has established and carried out a strong design or that he is familiar with the threats and has accounted for them in his interpretation. The statistical techniques to be presented here only determine whether a measured difference is sufficiently large to be considered statistically significant.

A further warning: there is a difference between statistical significance and practical significance, particularly to the administrator. A statistically significant difference is only a valid "difference" when it is large enough to be meaningful. When statistical tests are applied to particularly large samples of data (as is sometimes the case in accident research), any reasonable statistical test will be able to detect statistically significant differences due to the resulting increased precision (or, for the statistician, decreased estimate of variance). Thus, while these differences will be labeled "statistically significant," the administrator will still have to decide whether the corresponding degree of difference (e.g., reduction in accidents) is large enough to warrant the funds required to implement a countermeasure. This implies that some level of cost-effectiveness analysis should also be considered (see Section 3.9). Another warning concerns the choice of proper criterion, particularly where small samples are involved. Because there are many potential causes of accidents, even many effective treatments can reduce overall accident frequencies by only a small amount. Because statistical significance is considered so important and is based on the magnitude of change as related to the number of accidents being studied, it is crucial that only the subset of accidents that are affectable be studied. This helps guarantee that if a real difference exists, it will be shown to be statistically significant by the appropriate test.

The remainder of this chapter presents (1) a glossary of common statistical terms, (2) additional points of emphasis related to Type I and Type II errors; (3) a discussion of sampling guidelines, (4) a discussion of one and two-tailed tests, and (5) the basic criteria for choosing the appropriate statistical test. Finally, it presents, with examples, the details of the more important statistical tests. In addition, Appendix A contains a limited set of standard statistical tables and Appendix B presents a very basic introduction to statistical testing procedures as a review for the user with a very limited statistical background.

3.8.1. Glossary of terms.

Provided below is a rather limited list of words and phrases used in the more technical sections of Chapters 3 and 4. This glossary is designed, not to be extensive, but to include those items which are

either used repeatedly or which might confuse the user who has a limited statistical background. The user with a more extensive statistical background may notice that some of the definitions differ somewhat from the more theoretical definitions usually found in statistical texts. Hopefully, these modifications will result in a more readable manual.

1. Alpha (α) level = p level = "level of statistical significance": Represents the probability that a difference of a given size or a relationship of a given strength could result from chance alone. Also represents the probability that the researcher has made a "Type I" error by deciding (on the basis of a statistical test) that an effect exists when it really doesn't.
2. Beta (β) level: Represents the probability that the researcher has made a "Type II" error by deciding that an effect does not exist when it really does.
3. Power of the test (1- β): Represents the probability that a certain test used by the researcher will correctly detect an effect which really exists.
4. Critical value (Z_c , t_c , χ^2_c , etc.): The value of the statistical test (Z , t , χ^2) used, which is the breakpoint between significant and nonsignificant differences. This critical value is established by the choice of the alpha level and is extracted from a statistics table. The calculated value of the test statistic (i.e., Z , t , χ^2) is compared to this critical value.
5. Continuous data: Data which can assume a range (or continuum) of numerical values (e.g., pavement width, percent grade, speed).
6. Categorical data: Data which are not continuous but rather fall naturally into categories. Categorical data can be either scalar, ordinal, or nominal (see below).
7. Scalar data: Categorical data in which the labels of the categories are known distances apart (e.g., no. of lanes).
8. Ordinal data: Categorical data in which the labels of the categories are known and are ordered but are not necessarily a known distance apart (e.g., injury scales, any data graded "poor, fair, good," etc.).
9. Nominal data: Categorical data in which the labels of the categories are neither ordered nor known distances apart but are simply names of categories (eg., sex of driver, race, pavement type, urban/rural location).
10. Independent variable: Variables in a regression equation which predict the outcome variable--predictor variables.
11. Dependent variable: The outcome variable in a regression equation or model which is predicted by the other variables.
12. Parametric tests or procedures: Statistical tests or procedures which should be used only when the data being analyzed can be assumed to follow a known underlying distribution (i.e., normal, binomial, Poisson, etc.)
13. Nonparametric tests or procedures: Statistical tests or procedures which are appropriate for use in analyzing variables where assumptions cannot be made about the underlying distributions or where the distribution is known, but the parameters are unknown.
14. Main effect: In a regression or analysis of variance situation, the contribution to the variation in the outcome variable accounted for by a specific independent variable (e.g., effect of ADT on accidents at intersections).
15. Interaction: In a regression or analysis of variance situation, the contribution to the variation in the outcome variable accounted for by a combination (>2) of independent variables (e.g., effect of ADT and number of lanes simultaneously on accidents at intersections).

3.8.2 The importance of Type I and Type II errors

This chapter is concerned with hypothesis testing--did the countermeasure have the desired effect? For example, does the installation of truck escape ramps on downgrades significantly reduce brake-related truck accidents? Is one type of guardrail design more effective in reducing injuries than another? Which of a variety of curve delineation configurations is most effective in a given situation?

The researcher is required to use the sample of accident data collected in his evaluation design to draw inferences concerning whether or not the countermeasure would have an effect on the total population. In making such inferences, the researcher may be led to either of two errors: one error (Type I) is to claim that a particular countermeasure has an effect when indeed it does not; the other (Type II) is to conclude that the countermeasure does not have the desired effect when in fact it does (see Figure 3.15).

		True Situation	
		It Works	It Does Not Work
Statistical Decision	It Works	Correct	Type I Error (α)
	It Does Not Work	Type II Error (β)	Correct

Figure 3.15. True situations and statistical decisions.

For a simple example, suppose we hypothesize that a particular countermeasure will reduce accidents, i.e., that the average number of accidents will be lower after the treatment (say, pavement grooving) than prior to the treatment or when compared to a control group. If μ_c represents the expected number of accidents without grooving and μ_t is the number that occur with grooving, we are interested in testing the null hypothesis, $H_0: \mu_t = \mu_c$ versus the alternative, $H_A: \mu_t \neq \mu_c$. In the general case, this is done by calculating averages, \bar{x}_t , \bar{x}_c , computing a test statistic, say, t , and seeing if $|t| > t_c$, where t_c is the critical value for the significance level (α) selected. In such a case, the test statistic (t) will actually concern the differences in the two means ($\bar{x}_t - \bar{x}_c$). If the treatment has had no effect (the null hypothesis), the difference is zero. If an effect is present, the absolute value of the difference will be greater than zero.

What is actually being analyzed by the statistical test is the distribution of this difference. Under certain assumptions, this distribution of differences (standardized by dividing through by an estimate of the variance of this difference) follows the t distribution (as in Figure 3.16) with the distribution centered around the difference = 0 in the case of no effect.

Note that the distribution is composed of repeated samples of $\bar{x}_t - \bar{x}_c$. In actuality, the researcher has only one such sample and does not know the value of the true mean, $\mu_{\text{difference}} (\mu_D)$. That is, not knowing what the underlying distribution is, he does not know whether μ_D is as shown in Figure 3.16a (where the true effect is zero) or whether μ_D is as shown in Figure 3.16b, where a difference actually exists. The purpose of the testing is to estimate the value of the μ_D with a single sample.

To conduct such a test, a critical value of t , equal to t_c , is chosen based on the desired α level under the assumption of no difference (the null hypothesis). If the actual difference, $\bar{x}_t - \bar{x}_c$, is large enough to produce a calculated t which is less than t_c (to the left of t_c in Figure 3.16a), the null hypothesis of $\mu_D = 0$ is rejected and a difference is assumed. If the difference, $\bar{x}_t - \bar{x}_c$,

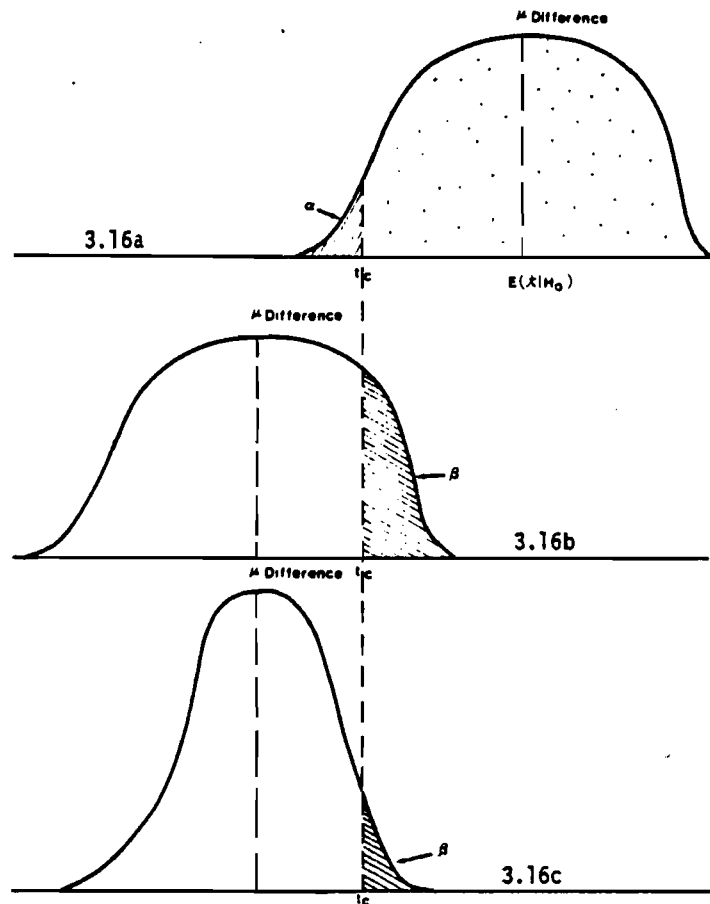


Figure 3.16. Graphical illustration of Type I and Type II errors.

produces a calculated t which is greater than (to the right of) t_c , the null hypothesis is not rejected and a lack of a real difference is assumed.

But observe Figure 3.16a. Obviously, even though the underlying distribution is unknown, and even if this difference, μ_D , was equal to zero, there will be some cases in which the single sample drawn would be drawn from the tail of the distribution to the left of t_c . This would cause the researcher to reject the "no-difference" null hypothesis and assume a real difference even though no such difference actually exists (i.e., even though $\mu_D = 0$). This is a Type I error (erroneously thinking there is a difference). The probability of such an error occurring is the area under the curve to the left of t_c and is equal to α , our preset value. Obviously, by choosing a very small α -level (moving t_c further to the left), the probability of a Type I error can be reduced.

However, observe Figure 3.16b. Here, again unknown to the researcher, the true difference, μ_D is actually less than zero, i.e., the treatment has had an effect ($\mu_t - \mu_c$ is negative). Again, t_c has been established by choosing α (under the null hypothesis of no effect). Here, if our single sample, $\bar{x}_t - \bar{x}_c$, were to fall to the left of t_c , we would correctly reject the null hypothesis and assume a difference exists.

But there are also cases where one sample might fall to the right of t_c even though the true difference, μ_D has the distribution shown in Figure 3.16b. If this occurred, we would not reject the null hypothesis of no difference (since we do not know whether we are drawing from 3.16a or 3.16b) and we would conclude that the treatment was not effective even though, in truth, it was. This is a Type II error, the probability of which is β , the cross-hatched area in Figure 3.16b.

The probability of committing a Type I error can be reduced by moving t_c to the left (choosing a smaller α level). But observe what happens to β , the Type II error, if a true difference exists (as in 3.16b): as α is reduced, β is increased. Thus, reducing the probability of erroneously claiming an effect increases the chances of missing a real effect. Obviously the choice of α is critical.

One more factor is also related to α and β . This factor is N , the sample size. As shown in Figure 3.16c if sample size is increased while holding t_c at the same level, the distribution becomes more narrow (i.e., the variance decreases) and the probability of a Type II error (and likewise a Type I error) is reduced. As will be further explored in the following section concerning the determination of sample size, if α and β are fixed, N can be calculated.

An issue that the researcher must resolve before making any significance tests is to decide on values of α and β (usually just α , since N is given, and thus β is determined by α). For obvious reasons, this issue cannot be resolved unequivocally for all situations. The values of α and β should depend on the consequences of making Type I and Type II errors, respectively. For example if the researcher is considering the installation of a very expensive countermeasure, he may hypothesize that the countermeasure is effective. If this hypothesis is true but he rejects it, the consequences are less economically severe than they would be if the hypothesis is false but he accepts it. For such a case, α should be set relatively small and β should be quite large.

On the other hand, the purpose of the evaluation is to help identify countermeasures which reduce accidents. Because there is not an unlimited supply of such proven treatments in existence, it is important not to reject one that is effective, especially since rejection may mean that the treatment is not tried again. Thus, it may be important to reduce the chance of a Type II error (small β), even though doing so increases the chances of a Type I error (large α). A more detailed discussion of this point is presented in Campbell (1972). As a guideline, α -levels of .10, or even .15, are sometimes considered appropriate in evaluation studies. While the $\alpha = .05$ level will be used in most of the examples which follow, this may not always be appropriate.

One-tailed versus two-tailed test? It is possible to conduct either one-tailed or two-tailed statistical tests. With a two-tailed test (the more usual test found in research of all types), the null hypothesis is that one treatment is no different from another treatment. The alternative hypothesis is that one of the treatments is either better or worse than the other treatment. Significance is indicated if the second treatment yields either higher or lower outcome measurements.

On the other hand, in using the one-tailed test the researcher must specify ahead of time the direction of expected change. For example, in most countermeasure evaluations, the researcher will be comparing a treatment group to an untreated group or an after period to a before period. In both of these cases, the expected direction of effect is for the treatment to positively affect accidents -- to reduce them. It is inherent in the mathematical formulation of statistical tests that statistical significance is more readily shown with a one-tailed test than with a two-tailed test. While it is important that the researcher be aware of the fact that his treatment may actually cause harm, the fact that the direction of effect is known in advance leads to the advocacy of the use of one-tailed statistical tests in these analyses. Perhaps the most reasonable procedure is to test using a one-tailed test, and if no difference is noted and the treatment appears to have an adverse effect, to then apply a two-tailed test.

3.8.3. Sampling considerations and sample size determination.

An issue that needs to be addressed before the details of the different statistical tests are introduced is the determination of the sample size required for evaluating the effectiveness of a given countermeasure. As can be seen in current literature, this relatively simple step can be one of the more important steps in evaluation design and analysis. However, it is also the step which is most often neglected by the researcher. Establishing the required sample size is important because, even if an evaluation is carefully planned, implemented, and analyzed, even meaningful differences will appear statistically insignificant (a Type II error) unless samples are of adequate size. If this occurs, limited evaluation dollars will have been wasted and a beneficial program misclassified. The basic problem stems from most countermeasure programs being limited to effecting modest changes in accidents. The smaller the change, the larger sample is needed to detect statistical significance. Because there are many locations where a countermeasure will be implemented where very few accidents will occur in a one-year period, the evaluator must often employ longer time periods or more than one location to evaluate this countermeasure. The question then becomes, "How does the evaluator best select this sample of accidents or locations and how large a sample is needed?"

In specifying guidelines concerning how to best select a sample, it is important that the researcher be familiar with some of the basic considerations involved in sampling techniques. While these considerations are included here because they are involved in evaluations, they are also pertinent in the later discussion of research involving relationships.

In most cases, even though he might attempt to obtain a total population of accidents, the researcher is faced with the fact that he usually has only a sample available. This is due to the various problems discussed in Chapter 2 which result in data biases and underreporting of accidents. But does the use of a sample instead of the total population of accidents at a given location or series of locations really affect the results of our evaluation? The answer, of course, depends on the amount of bias found in the data. The importance of the sample versus total population issue stems from the fact that the basic rationale for conducting the research is to be able to draw conclusions that can be used in the future concerning the total population. That is, whether he realizes it or not, the researcher is extrapolating from the sample to the total population. He is assuming that whatever effects (or relationships) are identified in the sample would also be found if the researcher were somehow able to examine the total population of accidents involving all drivers at all times. The more biased the sample, the less faith the researcher should have in his conclusions or inferences concerning the total population.

For example, if a study involving a particular guardrail design indicated that this particular design adequately redirected vehicles, the researcher would not conclude that all guardrail designs were of adequate strength to safely redirect vehicles. While in this situation, the lack of generality is obvious, in many cases, the issues are much more subtle. (An example of the subtlety of sampling plan errors in research involving relationships is included in Chapter 4.)

In summary, in almost every instance involving evaluation of countermeasures, the researcher is forced to draw conclusions concerning the total population of locations from his sample of either one location or a small number of locations. With the careful preplanning required to design and implement one of the stronger designs discussed earlier, there may be times when the evaluator will have the opportunity to select a location or series of locations prior to implementation. In such instances, he should employ a methodology such as random or stratified random sampling. However, because these instances are rare in highway evaluations, the detailed discussion of these methods is deferred until Chapter 4. Instead the remainder of this section deals with determining the sample size required to produce statistical significance. Again the importance of determining required sample sizes cannot be overemphasized, since meaningful differences or effects will not be indicated as statistically significant unless the sample size is adequate.

The evaluator must specify three quantities before he can calculate an appropriate sample size. He needs to specify (1) the degree or level of difference (effect) that is important for him to detect, e.g., whether the countermeasure must decrease accidents by 5 percent, 10 percent, 40 percent; (2) the probability of missing a real effect that he is willing to accept (β), and thus the power, or probability of detecting a real effect ($1-\beta$); and (3) the significance level (α).

The previous section discussed factors that should be considered while setting levels for α and β . Once the β -level is established, $1-\beta$ is determined (e.g., for $\beta = .20$, $1-\beta = .80$). In specifying the degree of difference that is important to detect, the investigator has to rely upon his own judgement. He can either arbitrarily decide that a given countermeasure must reduce accidents by some level or, in a more systematic fashion, he may conduct what could be termed a reverse economic analysis. While the details of such an analysis will be presented in an example which follows, the goal is to combine the countermeasure costs and the possible accident savings in order to determine the level of effectiveness which would make the countermeasure break even economically.

There are a number of different types of criteria which will be used in evaluations including frequencies, rates, proportions, variances and shifts in distribution (see related material in Section 3.8.4). While there are no simplified sample size determination formulas for all of these, there are formulas for three types of these measures: (1) when the researcher is comparing two proportions, p_1 and p_2 , e.g., proportions of vehicles involved in accidents before or after implementation of a countermeasure or proportions of accident involved vehicles at a control location compared to proportion at a treatment location, (2) when the researcher is comparing the means for two groups given a known variance, and (3) when the variance is unknown, but Poisson distribution assumptions are appropriate. The formulas for these three cases are presented below.

Sample size for differences in two proportions. (See Fleiss, 1973, p.30.) If the researcher is interested in comparing two proportions, p_1 and p_2 at the significance level (α) with power $(1-\beta)$, he should:

1. Calculate

$$N' = \frac{(z_{\alpha/2}\sqrt{2\bar{p}\bar{q}} - z_{1-\beta}\sqrt{p_1q_1 + p_2q_2})^2}{(p_2 - p_1)^2}$$

where

p_1 = estimated proportion in group 1

p_2 = estimated proportion in group 2

$q_1 = 1-p_1$

$q_2 = 1-p_2$

$$\bar{p} = \frac{p_1 + p_2}{2}$$

$$\bar{q} = 1-\bar{p}$$

$z_{\alpha/2}$ = critical value of z which leaves $\alpha/2$ in the upper tail of the standard normal distribution. Extracted from a table of the normal distribution.

$z_{1-\beta}$ = critical value of z which leaves $1-\beta$ in the upper tail of the standard normal distribution. Extracted from a table of the standard normal distribution.

Note: For $1-\beta$ greater than .50 (usual case), this critical value of z will be negative.

2. Applying a continuity correction, the required sample size is given by

$$N = \frac{N'}{4} \left[1 + \sqrt{1 + \frac{8}{N' |p_2 - p_1|}} \right]^2$$

Thus N represents the sample size for each group, either control and treatment or before and after.

Sample size for differences in two means. If the researcher is interested in comparing two means (or rates) at significance level (α) and power $(1-\beta)$, he also needs additional information. If the two samples are to be of equal size, and if the variance for the groups (σ^2) is known, or can be calculated from past data, the sample size N for each group can be calculated by:

$$N = \frac{(z_{\alpha} + z_{\beta})^2 \sigma_D^2}{\delta^2}$$

where

δ = difference in means which is important

σ_D^2 = variance of this difference

z_α = critical value of z which leaves α in the upper tail of the standard normal distribution. (Extracted from appropriate table)

z_β = critical value of z which leaves β in the upper tail of the standard normal distribution. (Extracted from appropriate table)

However, while there may be times when it is possible to calculate the variance σ^2 for the control group or before group from past data, or other times when σ^2 can be estimated from other information or knowledge, for most real-world situations the variance is unknown. Here the t-test replaces the normal deviate test but the formula for N is very complex and involves integral equations. The reader is referred to "The Design and Analysis of Industrial Experiments" by Owen Davies (1956) for details in such cases.

There is, however, a case in which an alternative to this complex interval equation solution exists. If it can be assumed that the number of accidents at the location(s) to be studied can be considered to have a Poisson distribution, (see Section 4.3.2.a), and if the evaluator has an estimate of the mean accident rate or frequency in the control group or before period, an approximate sample size can be estimated due to the fact that the mean of a Poisson variable is equal to the variance. If these assumptions and requirements can be met, then the evaluator can calculate N for each group by:

1. Determining or estimating the average accident frequency or rate without treatment = λ_0 .
2. Specifying the percentage reduction (change) in mean thought to be important = c , where c is expressed as a proportion, 0 to 1.0 (e.g., a 20 percent change would define $c = .20$).

3. Calculating
$$N = \frac{2(z_\alpha + z_\beta)^2}{c^2 \lambda_0}$$

where z_α , z_β are as defined above.

Example of sample size determination with "reverse economic analysis." Let us now consider a hypothetical situation which will illustrate a procedure to determine the difference that the researcher needs to detect and the calculation of the required sample size. In this situation the researcher is attempting to design an evaluation for a countermeasure which will reduce injuries (including fatalities) but will not reduce the number of accidents (e.g., a series of attenuators at elevated gore areas). The following is known:

Probability of a serious injury given an accident (No countermeasure)	= $p_1 = 0.70$
Probability of an accident at a gore area (accidents per encroachment)	= 0.03
Exposure to accidents (number of encroachments per year)	= 500
Cost of serious injury	= \$30,000
Cost of countermeasure	= \$215,000 per year (amortized)

Hence, the expected number of accidents without treatment = $(500)(0.03) = 15$. Therefore, the number of expected injuries = $(15)(0.70) = 10.5$, and the cost of injuries without the countermeasure = $(10.5)(30,000) = \$315,000$

For the countermeasure to break even economically, the number of serious injuries which must be reduced must have a value equal to the cost of the treatment, (i.e., \$215,000).

Thus,

$$\frac{\$215,000}{\$30,000} = 7.17 \text{ serious injuries must be reduced}$$

The number which can continue to occur = $10.5 - 7.17 = 3.33$.

Thus, the probability of injury/accident with the treatment should equal:

$$p_2 = \frac{3.33}{15} = 0.2215$$

Hence, the difference to be detected is

$$p_1 - p_2 = (0.70 - 0.22) = 0.48.$$

Let $\alpha = 0.10$ and power = $1 - \beta = 0.90$.

Thus $z_{\alpha/2} = 1.96$ and $z_{1-\beta} = -1.282$

$$\bar{p} = \frac{0.70 + 0.22}{2} = 0.46$$

$$\bar{q} = 1.00 - 0.46 = 0.54$$

$$\text{and } N' = \left[\frac{(1.960) \sqrt{(2)(.46)(.54)} + 1.282 \sqrt{(.70)(.30) + (.22)(.78)}}{(.70 - .22)} \right]^2 = 20.55$$

$$\text{and } N = \frac{20.55}{4} \left[1 + \sqrt{1 + \frac{8}{(20.55)(.48)}} \right]^2$$

$$= 28.27$$

Hence, we need 29 accidents in each group to confirm the effectiveness of the countermeasure. If the sites have 15 accidents per year, the researcher would need to study two years of data without the countermeasure and two years with the countermeasure in place. In such a case, a "normal" one year before and after period would not detect an effect even if it existed. The researcher interested in additional details of sample size determination for various cases ranging from differences in means or proportions to analysis of variance cases should refer to Statistical Power Analysis for the Behavioral Sciences by Cohen (1969). This text contains both formulas and tables for sample size determination in related analysis of the power of a given test.

3.8.4. Choice of appropriate statistical test.

The proper choice of a statistical test for a given situation basically depends on:

1. The evaluation design used;
2. The nature of the criterion variable studied; and
3. The type of data (continuous, categorical) involved.

The first key is related to the type of evaluation design (Before/After, Control Group designs, Time Series, etc.) actually employed in the research study. Given the evaluation design the second key concerns the nature of the criterion variable -- the type of data to be studied. In general, in highway accident research, the criterion will usually be of the following types:

1. Frequencies
2. Rates
3. Proportions

4. Variances

5. Shifts in distribution

For example, a Before/After design might involve numbers of accidents or it might involve shifts in the injury distribution from the before to after period. For most of the designs discussed, this total list of criterion types are possible. Examples of each of these types include:

1. Frequencies--number of accidents, number of injuries, number of locations experiencing accidents, number of serious injuries, number of fatalities, number of fatal accidents;
2. Rates--accident rates per million vehicle miles (mvm), accident rates per year, accident rates per entering vehicle, total injury rates per entering vehicle, fatality rates per crash;
3. Proportions--proportion of locations experiencing accidents, proportion of locations experiencing more than two accidents in a given time period, proportion of locations experiencing fatal accidents, proportion of entering vehicles involved in accidents;
4. Variances--changes in speed variances between before/after situations or between comparison and control groups;
5. Shifts in distribution--shifts in injury distribution following the imposition of a severity-reducing countermeasure.

For the convenience of the reader, an attempt has been made to denote which of the available tests will be appropriate for each design and each type of criterion (Table 3.3). The first column denotes the type of evaluation design used, the second indicates the type of the criterion variable, the third indicates the appropriate test or series of tests, the fourth indicates the page number of Chapter 3 in which the test is explained, and the fifth and final column presents appropriate references for further study. For some of the statistical tests presented, a detailed example has been provided for the user's convenience and study. These cases are designated by an "(Exp.)" in column 4. In other cases, where no example is given, a description of the test, its limitations, and its underlying assumptions and extensions are presented along with a statistical reference. For example, the analysis techniques used with data collected in time series designs have become so complex that it is not possible to present a single test which is useful. For this reason, a reference is given for the reader to examine for applications.

Table 3.3 Guide to Statistical Tests

Evaluation Design	Nature (type) of Criterion	Test(s) or Procedures	Page in Manual	Reference
1. Before/After	frequencies	a. χ^2 for Poisson Freq. b. Paired t-test (if normality assumed)	72 (exp.) ¹ 73	Snedecor & Cochran (1967) p.92-100
	rates	a. Paired t-test	73 (exp.)	"
	proportions	a. z-test for prop. If statistical control of others factors is attempted: b. Modified Mantel-Haenszel c. GENCAT d. ECTA e. CONTAB	74 (exp.) 75 75 75 75	Ostle (1969) p. 115-117 Campbell (1970) Landis, Stannish, Freeman, & Koch (1978). Goodman & Fay (1974) Gokhale & Kullback (1976)
	variances	a. F-test	76 (exp.)	Snedecor & Cochran (1967) p.116-117
	distribution shifts	a. RIDIT b. Kolmogorov-Smirnov	78 77 (exp.)	Hochberg (1975) Siegel (1956) p.127-136
2. Before/After with Randomized controls and 3. Before/After with comparison groups	frequencies	a. χ^2 for Poisson Freq. b. Paired t-test for B to A within group c. t-test for group vs. group d. Analysis of Covariance e. Median test (categorical data) f. Mann-Whitney U (categorical data)	72 73 78 79 80 80	Snedecor & Cochran (1967) p.92-100 Snedecor & Cochran (1967) p.100-106 Snedecor & Cochran (1967) Chap. 14 Siegel (1956) p.111-116; Conover (1971) Siegel (1956) p.116-127; Conover (1971)
	proportions	a. z-test for prop. between groups If statistical control of other factors is attempted: b. Modified Mantel-Haenszel c. Analysis of Covariance d. GENCAT e. ECTA f. CONTAB	74 (exp.) 75 79 75 75 75	Ostle (1969) p.115-117 Campbell (1970) Snedecor & Cochran (1967) Chap. 14 Landis, Stannish, Freeman, & Koch (1978) Goodman & Fay (1974) Gokhale & Kullback (1976)
	rates	a. Paired t-test for B to A within group b. t-test for group vs. group c. Analysis of Covariance	73 78 79	Snedecor & Cochran (1967) p.92-100 Snedecor & Cochran (1967) p.100-106 Snedecor & Cochran (1967) Chap. 14

¹Example problem included with test.

	variances distribution shifts	a. F-test a. Kolmogorov-Smirnov b. RIDIT (two sample)	76 (exp.) 77 78	Snedecor & Cochran (1967) p.116-117 Siegel (1956) p.127-136; Conover (1971) Hochberg (1975)
4. Interrupted Time Series	frequencies or rates	a. Fitting T-S model and testing of parameters		Glass, Wilson and Gottman (1975)
5. Time Series with Comparison Groups and 6. Time Series with Comparison Variables and 7. Time Series with Switching Replications	frequencies or rates (usually freq.)	a. Fitting T-S models to both groups and testing for diff. in parameters " "		" " "
8. Tie Breaking Designs	all	Same as for Before/After with Randomized Controls		See #2 above
9. Regression Discontinuity Analysis	frequencies or rates	Analysis of Covariance.	79	Snedecor & Cochran (1967) Chap. 14
10. Latin Square Design	frequencies or rates	a. ANOVA, followed by Scheffe's test Tukey procedure Duncan's procedure b. Analysis of Covariance c. Kruskal-Wallis (ordinal data)	79 79	Snedecor & Cochran (1967) Chaps.10 & 11 Kleinbaum & Kupper (1978) p.271-276 Kleinbaum & Kupper (1978) p.268-271 Sarhan & Greenberg (1962) pp. 147-148 Snedecor & Cochran (1967) Chap.14 Siegel (1956) p. 184-193
11. Factorial Design	proportions	a. χ^2 for proportions b. ANOVA (after transformation to make mean and var. independent) If statistical control of other factors is attempted: c. ECTA d. CONTAB	79 75 75	Fleiss (1973) p. 92-96 Snedecor & Cochran (1967) Chaps 10 & 11 Goodman & Fay (1974) Gokhale & Kullback (1976)
	variances distribution shifts	a. Bartlett's test a. Analyze means and variances as in #2 above	81	Neter & Wasserman (1974) p.509

χ^2 FOR POISSON FREQUENCIES

Analyses Question: Are the frequencies for one group significantly different from that of another?

Type of Data: Discrete (e.g., accident counts)

Underlying Assumptions: Data follow a Poisson process.

Statistic:

$$\chi^2 = \sum_{j=1}^k \frac{(N_{Aj} - \hat{N}_{Aj})^2}{\text{Var}(N_{Aj})}$$

where

$$\hat{N}_{Aj} = (N_{Bj} + N_{Aj}) \left[\frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right]$$

$$\text{Var}(N_{Aj}) = \hat{N}_{Aj} \left[1 - \frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right]$$

since under the null hypothesis,

$$N_{Aj} | N_{Bj} + N_{Aj} \sim B_i \left(N_{Bj} + N_{Aj}, \frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right)$$

$$\text{Var}(N_{Aj}) = (N_{Bj} + N_{Aj}) \left(\frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right) \left(1 - \frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right)$$

$$= \hat{N}_{Aj} \left(1 - \frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right)$$

where

t_{Aj} = length of the j -th time period for the after (A) sample;
likewise for t_{Bj} .

N_{Aj} = number of accidents in the j -th time period for the
after (A) sample; likewise for N_{Bj} .

$$N = N_{Aj} + N_{Bj}$$

k = number of locations.

Interpretation: If $\chi^2 > \chi^2_C$ with k degrees of freedom (see "Modifications"), reject null hypothesis of no difference.

Modifications: If a location has zero accidents in BOTH the before and after periods (resulting in a $\text{Var}(N_{Aj})=0$) then drop this term from the calculation. The degrees of freedom would then be reduced by 1.

EXAMPLE

1. Purpose: To test for a difference in the number of accidents based on a two-year before and one-year after period.

2. Data (hypothetical):

Group	Location (j)					
	1	2	3	4	5	6
Before	10(2)	10(2)	12(2)	14(2)	18(2)	12(2)
After	10(1)	8(1)	6(1)	6(1)	9(1)	6(1)

3. Calculate:

$$= \sum_{j=1}^6 [(N_{Aj} - \hat{N}_{Aj})^2 / \text{Var}(N_{Aj})]$$

where

$$\hat{N}_{Aj} = (N_{Bj} + N_{Aj}) \left(\frac{t_{Aj}}{t_{Bj} + t_{Aj}} \right)$$

For example,

$$\hat{N}_{A1} = (10 + 10) \left(\frac{1}{1 + 2} \right) = 6.7$$

$$\text{Var}(N_{A1}) = 6.67 \left[1 - \frac{1}{3} \right] = 4.45$$

Thus

$$\chi^2 = \left(\frac{10-6.7}{4.45} \right)^2 + \left(\frac{8-6.0}{4} \right)^2 + \left(\frac{6-6.0}{4} \right)^2 + \left(\frac{6-6.7}{4.45} \right)^2 + \left(\frac{9-9.0}{6} \right)^2 + \left(\frac{6-6.0}{4} \right)^2$$

$$\chi^2 = 3.56$$

4. Conclusion: Comparing $\chi^2 = 3.56$ with the tabular value χ^2_C of 12.59 with an $\alpha = 0.05$ and 6 d.f., there is no significant difference in the number of accidents at the 6 locations before and after the introduction of the countermeasure.

PAIRED T-TEST

Analysis Question: Is the before mean for a group of locations significantly different from the after mean for the same locations.

Type of Data: Continuous

Underlying Assumptions: Underlying distributions are approximately normal with means μ_B , μ_A and variances σ_B^2 , σ_A^2 , respectively.

Statistic:

$$t = \frac{\bar{x}_B - \bar{x}_A}{s_D / \sqrt{N}}$$

where \bar{x}_B = Before sample mean.

\bar{x}_A = After sample mean.

and

$$s_D^2 = s_B^2 + s_A^2 - 2 \left[\frac{1}{N-1} \sum_{i=1}^N (x_{Bi} - \bar{x}_B)(x_{Ai} - \bar{x}_A) \right]$$

N = number of locations.

Interpretation: If $t > t_c$, difference in means is statistically significant where degrees of freedom is equal to the number of locations - 1.

Modifications: None

Example

1. **Purpose:** To evaluate the effectiveness of improved pavement delineation on the number of annual accidents per 1,000 miles for one set of locations where before and after data is collected.

2. **Data:** No. of Accidents Per 1,000 Miles/Year

	Location						Total	Avg.
	1	2	3	4	5	6		
Before New Delineation	7.0	14.1	19.0	20.6	30.2	41.1	132.0	22.0
After New Delineation	7.3	8.5	14.2	17.5	18.5	30.1	96.1	16.02

3. **Assume:** The underlying distributions are approximately normal with means μ_B , μ_A and variances σ_B^2 , σ_A^2 , respectively.

4. **Compute:**

$$t = \frac{\bar{x}_B - \bar{x}_A}{s_D / \sqrt{N}}$$

where: $\bar{x}_B = 22.0$

$\bar{x}_A = 16.02$

$s_B^2 = 146.08$

$s_A^2 = 68.50$

N = 6

d.f. = 5

then

$$\begin{aligned} s_D^2 &= s_B^2 + s_A^2 - 2 \left[\frac{1}{N-1} \sum_{i=1}^N (x_{Bi} - \bar{x}_B)(x_{Ai} - \bar{x}_A) \right] \\ &= 214.58 - 2(96.48) = 21.62 \\ s_D &= 4.65 \end{aligned}$$

and

$$t = \frac{22.0 - 16.02}{4.65 / \sqrt{6}} = 3.15$$

5. **Conclusion:** Comparing $t = 3.15$ with $t_c = 2.01$ for $\alpha = .05$ and 5 d.f. (one-sided test), we reject the null hypothesis and conclude that the new delineation is effective.

Z-TEST FOR PROPORTIONS

Analysis Question: Is the proportion of occurrences in one group significantly different from the proportion in a second group.

Type of Data: Continuous (proportions)

Underlying Assumptions:

1. Underlying distribution is binomial (observation is either success or failure -- no other level)
2. Observations are independent.
3. Large samples are collected in each group ($N > 30$).

Statistic:

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)\left(\frac{1}{N_1} + \frac{1}{N_2}\right)}}$$

where

$$p_1 = \frac{x_1}{N_1}$$

$$p_2 = \frac{x_2}{N_2}$$

$$p = \frac{x_1 + x_2}{N_1 + N_2} = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

x_1 = number of occurrences in group 1 (e.g., serious injuries); likewise for x_2 .

N_1 = number of possible occurrences or trials (e.g., number of drivers); likewise for N_2 .

Interpretation: If $z > z_c$, the difference in proportions is statistically significant.

Modifications: If $N < 30$, refer to Ostle, 1969, p. 116.

Example

1. **Purpose:** To test the relative effectiveness of two different attenuation systems in reducing the proportion of drivers seriously injured in gore area-type crashes. Note: The two systems are placed at various Interstate locations which are as comparable as possible.

2. **Data:**

	Proportion of drivers with serious injuries	Total N
Crash Cushion A	.40	210
Crash Cushion B	.47	309

3. **Compute:**

$$z = \frac{p_A - p_B}{\sqrt{p(1-p)\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}}$$

where: p_A = proportion of seriously injured drivers striking crash cushion A.

N_A = number of drivers striking crash cushion A

p = estimate of overall proportion

$$\begin{aligned}
 &= \frac{N_A p_A + N_B p_B}{N_A + N_B} \\
 &= \frac{210(.40) + 309(.47)}{210 + 309} = .44
 \end{aligned}$$

Therefore

$$z = 1.58$$

4. **Conclusion:** For $\alpha = .10$, $z_c = 1.28$. Since the z does exceed 1.28 it appears that there is a difference between types of attenuation systems in preventing serious driver injury.

MANTEL-HAENSZEL (M-H) (Especially Appropriate for Low Frequency Situations)

Analysis Question: Is the proportion of a comparison group different from that of a reference group, taking into consideration the levels of other variables?

Type of Data: Categorical data.

Underlying Assumptions: None.

Statistic:

I = Calculated index for the ratio of observed to expected frequencies summed over the various strata

Having calculated I , the χ^2 is used for testing significance (see Campbell, 1970)

Interpretation: Index provides indication of treatment worth. If $\chi^2 > \chi^2_c$, then the comparison group is significantly different from the reference group in, say, the proportion of serious driver injuries ($A + K$).

Modifications: None

GENCAT (Generalized Categorical Data Analysis Using Weighted Regression Procedures)

Analysis Question: Is the proportion of the outcome measure different in the after period from that of the before period?

Type of Data: Categorical data.

Underlying Assumptions: None

Statistic: Regression-type models are fitted to the data using weighted regression procedures and appropriate tests of functions of the parameters are carried out (see computer package references)

Interpretation: See Landis, et al., (1978)

Modifications: See Landis, et al., (1978)

ECTA OR CONTAB FOR LOG LINEAR MODELS

Analyses Question: Is the proportion of occurrence in one group significantly different from the proportion in a second group when other variables (some categorical) are controlled for statistically.

Type of Data: Categorical

Underlying Assumption: Cell proportions are asymptotically normal.

Statistic: Regression-type models are fitted to the data and appropriate tests of the parameters are conducted (see computer package references).

Interpretation: See Goodman & Fay (1974) and Gokhale & Kullback (1976)

Modifications: See Goodman & Fay (1974) and Gokhale & Kullback (1976)

F-TEST

Analysis Question: Is there a significant difference between the variances of two populations?

Type of Data: Continuous

Underlying Assumptions:

1. Independent random samples.
2. Underlying distributions are normal.

Statistic:

$$F = \frac{S_A^2}{S_B^2}$$

where

$$S_A^2 = \sum \frac{(x_{A1} - \bar{x}_A)^2}{N_A - 1}$$

S_B^2 likewise

Interpretation: If $F > F_C$ where d.f. = $((N_A - 1), (N_B - 1))$ then the variances are significantly different.

Modifications: None

Example

1. **Purpose:** To evaluate the effectiveness of an electronic speed warning signal in reducing speed variability.

2. **Data:**

	N	Mean Speed	Speed Variance
No Speed Warning Sign (A)	280	55.4	49.0
Speed Warning Sign (B)	401	55.5	38.8

3. **Compute:** Using the formulae

$$S_A^2 = \sum \frac{(x_{A1} - \bar{x}_A)^2}{N_A - 1}$$

$$S_B^2 = \sum \frac{(x_{B1} - \bar{x}_B)^2}{N_B - 1}$$

speed variances were obtained for both groups. Then, using these variances

$$F = \frac{S_A^2}{S_B^2} = \frac{49.0}{38.8} = 1.26$$

4. **Conclusion:** For sample sizes $N = 280$, $N = 401$, the critical value for F with $\alpha = .05$ is 1.11. Since $F = 1.26$, it may be concluded that the data do provide sufficient evidence to indicate that speed variability is reduced with this electronic speed warning device.

KOLMOGOROV-SMIRNOV TEST

Analysis Question: Has there been a shift in the distribution from before to after or group to group.

NOTE: The Kolmogorov-Smirnov test will detect changes (or differences) in the shape of the distributions (e.g., skewed left to skewed right, bell-shaped to skewed, etc.) as well as shifts in central tendency without shifts in shape. Riddit analysis procedures, an alternative test, will primarily detect shifts in central tendency only.

Type of Data: Ordinal (may be applied to small samples).

Underlying Assumptions: Underlying continuous distribution.

Statistic:

$$D = \max |S_{N_A}(x) - S_{N_B}(x)|$$

where

$S_{N_A}(x)$ = the observed cumulative step function for the sample corresponding to after.

Interpretation: If $D > D_c$, then the distribution of one group is significantly different from the distribution of a second group.

Modifications: None

Example

1. Purpose: To compare the driver injury distribution before imposition of the 55 mile per hour speed limit with the corresponding distribution after the speed limit was imposed.

2. Data: (Fleiss, 1973, p. 104) Frequencies of driver injury severities before and after 55 mph speed limit.

(1)	(2)	(3)	(4)	(5)	(6)
Injury Severity	Frequency Before 55 mph Speed Limit N_B	Frequency After 55 mph Speed Limit N_A	Cumulative Frequency Distribution Before 55 mph Speed Limit $S_{N_B}(x)$	Cumulative Frequency Distribution After 55 mph Speed Limit $S_{N_A}(x)$	Col(4)-Col(5) $ S_{N_A}(x) - S_{N_B}(x) $
None	17	5	0.095	0.100	0.005
Minor	54	10	0.397*	0.300	0.097
Moderate	60	16	0.732	0.620	0.112
Severe	19	5	0.838	0.720	0.118
Serious	9	3	0.888	0.780	0.108
Critical	6	6	0.922	0.900	0.022
Fatal	14	5	1.000	1.000	0.000
Total	179	50			

*For example, $0.397 = 0.095 + \frac{54}{179}$

3. Compute: Columns 2 and 3 in the table represent the injury frequencies for various severity levels before and after the 55 mph speed limit, respectively. Columns 4 and 5 represent the respective cumulative frequency distributions. Col.(6) equals the absolute value of (4)-(5).

From column 6, D = maximum difference in column (6)
= 0.118

From Siegel, 1956 (p. 279) for $\alpha = 0.05$, $D_c = 1.36 \sqrt{\frac{N_B + N_A}{N_A N_B}} = 0.217$

4. Conclusion: Since $D = 0.118 < 0.2175$, it may be concluded that the two samples do not arise from two different distributions; i.e., that the injury severity distributions are the same before and after imposition of the 55 mile-per-hour speed limit.

RIDIT

Analyses Question: Has there been a shift in the distribution from before to after or group to group.

NOTE: RIDIT is an alternative to the Kolmogorov-Smirnov test. Unlike the K-S test, it will primarily detect shifts in central tendency rather than changes or differences in the shape of the distribution.

Type of Data: Ordinal data

Underlying Assumption: Underlying continuous distribution.

Statistic: Using one group as a baseline, a RIDIT (r) for the other group is calculated using procedures noted in Hochberg (1975). Using a z-test, is the r significantly different from 0.5?

Interpretation: If $z > z_c$, the distributions of the two groups are significantly different.

Modification: See Hochberg (1975)

STUDENT'S T-TEST

Analysis Question: Is the mean of one group significantly different from the mean of another group?

Type of Data: Continuous.

Underlying Assumptions:

1. Underlying distribution approximately normal with means μ_A and μ_B and common variance σ^2 (see modif. #3).
2. Observations must be independent (see modif. #2).

Statistic:

$$t = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{s_p^2 \left(\frac{1}{N_B} + \frac{1}{N_A} \right)}}$$

where

$$s_p^2 = \frac{(N_B - 1)s_B^2 + (N_A - 1)s_A^2}{N_B + N_A - 2}$$

Interpretation: If $t > t_c$, the difference in means is statistically significant.

Modifications:

1. If data is ordinal, use Mann-Whitney U test or Kolmogorov Smirnov test.
2. When the observations are not independent (e.g., same locations) use paired t-test.
3. If variances are unequal, use Satterthwaite's procedure (see Dixon and Massey, 1957, Chapter 9, and Cochran and Cox, 1950).

ANOVA

Analysis Question: Are the means of two or more treatments all equal or are any significantly different from the others.

Type of Data: Continuous

Underlying Assumptions: Normally distributed residuals with mean residual = 0 and common variance.

Statistic: Using appropriate procedures, an ANOVA table is calculated producing various Sums of Squares (SS).

Then

$$F = \frac{\frac{SS_{\text{TREATMENTS}}}{d.f.}}{\frac{SS_{\text{ERROR}}}{d.f.}}$$

Interpretation: If $F > F_c$, some of the means are significantly different.

Modifications: For cases of multiple treatments with significant F, the determination of which treatment(s) is significantly different from the others is done using Duncan's, Scheffe's, or Tukey's procedures (see Table 3.3 for references or consult statistician).

ANALYSIS OF COVARIANCE

Analysis Question: Are the means of two or more treatments all equal or are they significantly different while controlling for an additional variable(s).

Type of Data: Continuous

Underlying Assumptions: Normally distributed residuals with mean residual = 0 and common variance.

Statistic: Using appropriate procedures, an analysis of covariance table is calculated producing various Sums of Squares.
Then

$$F = \frac{\frac{SS_{\text{TREATMENTS (adjusted)}}}{d.f.}}{\frac{SS_{\text{ERROR (adjusted)}}}{d.f.}}$$

Interpretation: If $F > F_c$, the means are significantly different.

Modifications: None

MEDIAN TEST

Analysis Question: Is the median of one group significantly different from the median of another?

NOTE: Procedure readily extends to more than two groups.

Type of Data: Ordinal (categorical)

Underlying Assumptions:

1. Use of χ^2 requires a two-sided test.
2. Ties require special treatment.

Statistic:

$$\chi^2 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{(N_{ij} - M_{ij})^2}{M_{ij}}$$

where the data are arranged in the following table:

	Above Median	Below Median	
Group 1	N_{11}	N_{12}	$N_{1\cdot}$
Group 2	N_{21}	N_{22}	$N_{2\cdot}$
	$N_{\cdot 1}$	$N_{\cdot 2}$	N

where N_{11} = number of observations in Group 1 below the combined median; N_{12} , N_{21} , N_{22} likewise.

$$M_{11} = \frac{N_{1\cdot} \cdot N_{\cdot 1}}{N}; M_{12}, M_{21}, M_{22} \text{ likewise.}$$

Interpretation: When $\chi^2 > \chi^2_{\alpha}$ for d.f. = 1 and given α the difference in means is significantly different.

Modifications:

1. When $N_1 + N_2 > 40$, use χ^2 corrected for continuity.
2. If the smallest expected frequency is less than 5, use the Fisher exact test (see Siegel (1956), pp. 96-104.)
3. If ties occur, see Hays and Winkler, 1970, p. 227.

MANN-WHITNEY U-TEST

Analyses Question: Is the mean of one group significantly different from the mean of another?

Type of Data: Ordinal (categorical)

Underlying Assumptions: Independent samples.

Statistic: Using specified procedure, the observations are ranked and two statistics (U_1 and U_2) are calculated (See Siegel, 1956, p. 116-127). Using tables, the minimum U is converted to a probability value.

Interpretation: If the calculated p value is less than α , the means are significantly different.

Modification: Treatment of ties: If ties occur between 2 or more observations involving both groups, the value of U is affected. See Hays and Winkler (1970), p. 234.

Note: This is a useful substitute for parametric t-test when researcher wants to avoid t-test assumptions or when measurement in the research is weaker than interval scaling. Mann-Whitney U exhibits greater power than the median test (i.e., it is more likely to detect a difference when indeed there is a difference).

BARTLETT'S TEST

Analyses Question: Are the variances resulting from multiple treatments significantly different.

Type of Data: Continuous data.

Underlying Assumptions: Data approximately normally distributed for each treatment.

Statistic:

$$B = \frac{2.3026}{C} \left[(N_T - r) \log_{10} (MSE) - \sum_{j=1}^r (N_j - 1) \log_{10} (s_j^2) \right]$$

where

r = no. of treatments

$N_T = \sum N_j$, and N_j = no. of observations in treatment j .

s_j^2 = sample variance for treatment j

$$MSE = \frac{1}{N_T - r} \sum_{j=1}^r (n_j - 1) s_j^2$$

$$C = 1 + \frac{1}{3(r-1)} \left[\sum_{j=1}^r \frac{1}{N_j - 1} - \frac{1}{N_T - r} \right]$$

Interpretation: If $B > \chi^2(\alpha; r-1)$, then not all the variances are equal.

Modifications: If N_j are all equal, this procedure can be simplified. Test statistic is $H = \max(s_j^2) / \min(s_j^2)$. If $H > H_c(\alpha; r, N_j)$ then not all the variances are equal. For critical values of H , refer to table on p. 830, Netter & Wasserman (1974), or Pearson & Hartley (1954). Also see Snedecor & Cochran (1967, p. 298).

3.9 Use of Evaluation Results in Cost/Benefit Analysis

The preceding sections of this chapter have been related to planning and implementing evaluations aimed to determine whether a given countermeasure is effective in changing the criterion of interest and, if so, what the level of effectiveness is. The evaluation is being conducted for one purpose--to provide the administrator with information with which to make decisions. In fact, program decisions could be made simply on the basis of the accident reductions determined in the above described evaluations. For example, in determining where to spend a limited amount of money, the program administrator could decide to use those treatments which have been indicated to have the highest level of effectiveness. This scheme, however, would fail to take into account the number of crashes that could be affected by a given program. It is quite possible that, because of the nature of problems in a given state, a countermeasure of a lower level of effectiveness may save more lives simply because it can be implemented at more locations or at locations which carry more traffic and experience more accidents.

However, there is a third refinement which would produce a more rational approach than either of the two above. This would be to combine the accident benefits calculated above with the related program cost, especially when there is more than one alternative program competing for funds. This combination is the essence of cost/benefit analysis and related budget optimization.

3.9.1 Cost-benefit methodology definition.

Various cost/benefit or cost-effective methodologies have been discussed in numerous reports in the literature, particularly in the past five years. While these presentations can appear to be quite complex, in truth, the methodology itself is simply another tool which is used by the administrator in establishing his or her program priorities. At its simplest, it is an algorithm or formula for combining accident--related benefits with program costs in a scientifically meaningful way.

3.9.2 Possible algorithms.

There are a number of different algorithms or formulas for combining these accident-related benefits and program costs. Each has its strengths and weaknesses and the choice of the most appropriate one will depend on many factors. Example algorithms include benefit-cost analysis, the use of net discounted present worth, dynamic programming, integer programming, incremental benefit-cost analysis, and others.

Because this manual is primarily concerned with accident research, a discussion of the strengths and weaknesses of these different algorithms will not be presented. However, the reader interested in such a discussion should refer to a recent report prepared for FHWA by McFarland, et al. (1978). This report contains information concerning better methods of determining accident cost, statistical procedures for calculating countermeasure effectiveness, internally consistent systems for evaluating accident cost or combining accident cost and countermeasure effectiveness, and improved incremental cost/benefit algorithms for ranking safety projects. In addition to developing these improved techniques, the report also reviews selected accident countermeasure studies and provides a critique of current procedures for evaluating safety programs. Finally, a specific algorithm for use in allocating an optimum safety budget is recommended.

In addition to this recent report, a number of past reports have also included computerized methods for combining accident costs and program benefits (see Council, et al., 1977). In general, the computer programs involved are usually simple enough to be modified for individual highway departments for their own use.

In summary, however, the important thing for the manual user to realize is that any cost/benefit methodology is but a further extension of the tools needed in the decision-making process discussed throughout this manual. It is the methodology which is the next step to the process described earlier in that it takes the results of the evaluations conducted and combines them with program costs in an attempt to optimize the expenditure of safety funds.

3.10 Review Questions

1. What is the difference between administrative (process) evaluation and effectiveness evaluation? Why is administrative evaluation an integral part of effectiveness evaluation?
2. D. T. Campbell has suggested a philosophy for administrators to publicly state when questioned about their safety programs. What is the basis of this philosophy?
3. The people of New Hebrides have decided that lice produce good health since all their healthy tribesmen have lice and none of the sick ones do. Does this mean lice should be imported to the U.S. as a health aid? Why or why not?
4. What is the basic question the evaluator should ask in determining the criterion to be used?
5. What does "threat to internal validity of an evaluation" imply in terms of the cause of observed effect?
6. List the four main potential threats to internal validity and note which of the four most often causes problems in research involving high-accident locations.
7. What is the basic purpose of the evaluation design?
8. Which design would be appropriate to evaluate a law reducing speed limits on all freeways to 55 mph?
9. A highway engineering department in budgeting for the coming fiscal year has a set operating improvement budget and the results of the evaluations of three proposed improvements. Which if any of the following improvements should the department make? All cost the same amount.

	<u>α</u>	<u>Calculated Values</u>	<u>d.f.</u>	<u>Critical Values</u>
Improved Pavement Delineation	.05	$t = .997$	10	$t_c = 1.8$
Breakaway Poles	.05	$\chi^2 = 3.22$	1	$\chi_c^2 = 3.84$
A New Attenuation System	.05	$\chi^2 = 2.49$	1	$\chi_c^2 = 3.84$

10. In order to evaluate a particular countermeasure, namely the effectiveness of 4 different types of grooving in reducing accidents in rainy weather, what would be the most appropriate test to employ?
11. To examine differences in distributions of ordinal data, it is appropriate to use which test?
12. To the highway engineer with little money to expend which type of error is more acceptable Type I or Type II? Why?

What about the researcher attempting to find an effective countermeasure for an important problem area in which no proven treatments exist?
13. A researcher is evaluating the effectiveness of water-filled crash attenuation devices. The devices have been placed in gore areas of arterials which carry heavy commuter traffic involving car-pooling. A comparison group of locations has been chosen from rural freeways experiencing similar ADT's. Would total number of serious occupant injuries or total occupant deaths be appropriate criteria for the evaluation?
14. While many statistical tests exist for analyzing data collected in an evaluation, the choice of most appropriate test basically depends on three factors. These are:
 - a. The evaluation design used
 - b.
 - c.

3.11. References

- Anderson, T. W. The statistical analysis of time series. New York: John Wiley & Sons, 1971.
- Basile, A. J. The effect of pavement marking on traffic accidents in Kansas. Highway Research Board Proceedings, 1961.
- Box, G. E. P., & Jenkins, G. M. Time series analysis forecasting and control. San Francisco: Holden Day, 1970.
- Box, G. E. P., & Tiao, G. C. A change in level of a non-stationary time series. Biometrika, 1965, 52, pp. 181-192.
- Campbell, B. J. Driver injury in automobile accidents involving certain car models. Chapel Hill: University of North Carolina Highway Safety Research Center, 1970.
- Campbell, B. J. Comments on David Solomon's paper, "Highway Safety Myths." In P. F. Waller (Ed.), Proceedings of the North Carolina Symposium on Highway Safety, Vol. 2, Highway and traffic safety: A problem of definition. Chapel Hill: University of North Carolina Highway Safety Research Center, 1972.
- Campbell, D. T. Reforms as experiments. In E. L. Struening & M. Guttentag (Eds.), Handbook of evaluation research. Beverly Hills, California: Sage Publications, 1975.
- Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research and teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.
- Cochran, W. G., & Cox, G. M. Experimental designs. New York: John Wiley & Sons, 1950.
- Conover, W. J. Practical nonparametric statistics. New York: John Wiley & Sons, 1971.
- Cook, T. D., & Campbell, D. T. The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunnette (Ed.), Handbook of industrial and organizational psychology. Chicago: Rand McNally, 1976.
- Council, F. M., Dutt, A. K., Hunter, W. W., Leung, A. Y., & Woody, N. C. Project selection for roadside hazard elimination. Vol. 2. User manual for roadside hazard correction ranking program. Chapel Hill: University of North Carolina Highway Safety Research Center, 1977.
- Davies, O. L. The design and analysis of industrial experiments. (2nd ed.) Edinburgh, Scotland: Oliver and Boyd, 1956.
- Dixon, W. J., & Massey, F. J., Jr. Introduction to statistical analysis. New York: McGraw-Hill, 1957.
- Fleiss, J. L. Statistical methods for rates and proportions. New York: John Wiley & Sons, 1973.
- Footy, T. J., & Taylor, W. C. Curve delineation and accidents. Columbus: Ohio Department of Highways, Bureau of Traffic, 1966.
- Glass, G. V., Willson, V. L., & Gottman, J. M. Design and analysis of time-series experiments. Boulder: Colorado Associated University Press, 1975.
- Gokhale, & Kullback, S. The information in contingency tables. Mimeographed notes. George Washington University, 1976.
- Goodman, L., & Fay, R. A computer program for fitting log-linear models to contingency tables. Chicago: University of Chicago Department of Statistics, 1974.

- Hayes, W. L., & Winkler, R. L. Statistics: Probability, inference, and decision. Vol. 2. New York: Holt, Rinehart and Winston, 1970.
- Hedlund, J. E. The severity of truck accidents. Technical note, Mathematical Analysis Division, National Highway Traffic Safety Administration, April 1975.
- Hicks, C. R. Fundamental concepts in the design of experiments. (2nd ed.) New York: Holt, Rinehart and Winston, 1973.
- Hochberg, Y. On the variance estimates of a Mann-Whitney statistic for ordered group data. Chapel Hill: University of North Carolina Highway Safety Research Center, 1975.
- Hunter, W. W., Council, F. M., & Dutt, A. K. Project selection for roadside hazards elimination. Final Report. (Vol. 1) Chapel Hill: University of North Carolina Highway Safety Research Center, 1977.
- Huff, D. How to lie with statistics. New York: W. W. Norton, 1954.
- Kleinbaum, D. G., & Kupper, L. L. Applied regression analysis and other multivariable methods. North Scituate, Massachusetts: Duxbury Press, 1978.
- Landis, J., Stannish, W., Freeman, J., & Koch, G. G. A computer program for the generalized chi-square analysis of categorical data using weighted least squares. (Rev. ed.) Chapel Hill: University of North Carolina Department of Biostatistics, 1978. [Biostatistics Report No. 8].
- McFarland, W. F., Griffin, L. I., Rollins, J. B., Stockton, W. R., Phillips, D. T., & Dudek, C. L. Assessment of techniques for cost-effectiveness of highway accident countermeasures. (4 Vols.) College Station: Texas Transportation Institute, 1978.
- Musick, J. V. Effect of pavement edge marking on two-lane rural highways in Ohio. Highway Research Board Bulletin, No. 266, 1960, pp. 1-8.
- Neter, J., & Wasserman, W. Applied linear statistical models: Regression, analysis of variance and experimental designs. Homewood, IL: Richard D. Irwin, Inc. 1974.
- Ostle, B. Statistics in research: Basic concepts and techniques for research workers. (2nd ed.) Ames: Iowa State University Press, 1963.
- Olivarez, D. R. Safety experiences with concrete and metal beam barriers. Phoenix: Arizona Highway Department, 1969.
- Ross, H. L., Campbell, D. T., & Glass, G. V. Determining the social effects of a legal reform: The British "breathalyser" crackdown of 1967. American Behavioral Scientist, 1970, 13, 493-509.
- Sarhan, A. E., & Greenberg, B. G. Contributions to order statistics. New York: John Wiley & Sons, 1962.
- Schoppert, D. W., & Hoyt, D. M. Factors influencing safety at highway-rail grade crossings. National Cooperative Highway Research Program Report, No. 50, 1968.
- Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.
- Solomon, D. Accidents on main rural highways, related to speed, driver, and vehicle. Washington, D.C.: U.S. Department of Commerce, 1964.
- Steel, R. G. D., & Torrie, J. H. Principles and procedures of statistics with special references to the biological sciences. New York: McGraw-Hill, 1960.

CHAPTER IV

IDENTIFYING RELATIONSHIPS AMONG VARIABLES

Situation: A young graduate from engineering school is hired by the accident research unit of the Federal Highway Administration, and as part of the orientation, his supervisor gives him a set of data collected in a number of states concerning accidents, descriptive vehicle and driver variables, and exposure (mileage) information. Because of a recent multi-fatality accident in a Far West state involving a heavy truck, a request for an analysis of accidents involving large trucks has come down from the administrator of the department. The supervisor, busy with a similar request concerning increased fixed-object crashes involving small vehicles, suggests that the young engineer call the resident statistician. The statistician suggests that the engineer should first examine the relationship between accident rate and size of vehicle using Spearman's rho and then should employ an analysis of covariance with vehicle mileage as the covariant. He would like to help but is also over-committed to other analysis efforts. The engineer, not understanding (or trusting) the statistician, has the computerized data printed out in simple tables in which the number of accident-involved vehicles is presented by each of a number of vehicle types. Based on the fact that the heaviest trucks are involved in more accidents than the lighter trucks and that the number of fatalities per crash is seven times larger in crashes involving heavy trucks (and knowing personally of the problems with passing the larger vehicles, especially in bad weather with their splash and spray), the engineer recommends a reduction in allowable trailer length.

Result: A local consumer group against larger trucks (TRUCKSTOP) hears of his recommendation and selects him "Researcher of the Month." The national headquarters of TMA (Trucks Move America) asks for a meeting with the Secretary of Transportation. The young engineer requests an immediate transfer to the roadway design department.

Main Chapter Topics

- Introduction
- Analysis Issues Related to Research Involving Relationships
 - Sampling considerations
 - Choice of dependent variable
- Analysis Techniques
 - Introduction to techniques
 - Variable screening techniques
 - Model development procedures

4.1 Introduction

Despite their complexities, the basic goal of all research and evaluation studies is to identify possible causal relationships between subsequent accidents and other factors of interest while accounting for all other factors which may contaminate or confuse the results. Sometimes, contaminating factors can be controlled by implementing a countermeasure with a planned experimental design (see Chapter 3 for a complete discussion of this strategy). In other instances, however, when contaminating factors cannot be manipulated or were not controlled for in the implementation of the experimental design, the researchers must seek to discount their effort with statistical procedures. This chapter discusses the techniques involved in this second strategy.

4.2 Analysis Issues Related to Research Involving Relationships

In analyzing relationships, researchers generally follow one of two strategies. They either conduct what can be called a descriptive or comparative study or they attempt to develop an equation or mathematical formula (usually of a predictive nature). Although the remainder of this chapter concentrates on the development of mathematical formulas or equations, this should not be construed as an indication that descriptive studies are not a highly valuable tool for examining underlying relationships.

For a descriptive study, researchers generally have at their disposal a set of data in which a number of single accidents are each accompanied by information concerning a number of other variables (i.e., for each accident there is included information on driver age, driver sex, lane width, temperature, speed limit, collision speed, average daily traffic, etc.). In a descriptive study, the accidents are subdivided into categories in a series of tables. The tables are then examined (not necessarily statistically) for underlying relationships. For example, accidents might be classified according to the day/night variable, the type-of-vehicle variable or by a combination of the two. The researcher then compares the trends among vehicle types as they differ by day/night.

One example of such a descriptive study which has provided a large amount of information to subsequent roadway decision was the study by Solomon (1964) in which he analyzed accidents that had occurred on main rural highways. The purpose was to establish relationships between accidents and speed, driver, and vehicle types. The data base consisted of accident data, exposure data, speed measurements taken at the locations of interest, and driver interview data.

The author examined numerous questions involving relationships between accident involvement and various vehicle and driver characteristics, but perhaps the most important findings were related to speed. The key to the analysis of speed was the use of involvement rates. As pointed out by Solomon:

"It is not enough, however, to know that certain number of drivers involved in accidents were travelling at a particular speed; it is also essential to determine how much driving was done at that same speed. Then by relating the travel speed of accident-involved drivers and of all drivers, it is possible to determine the hazard associated with specific driving speeds--the accident involvement rate."

For example, in his analysis, the author developed the data shown in Table 4.1 concerning the number of daytime vehicle miles of travel and the corresponding number of vehicle accident involvements for each speed category. From these two numbers a rate was calculated. These daytime rates were then plotted in the graph shown in Figure 4.1.

Table 4.1 Accident involvements, vehicle miles, and involvement rates by travel speed for daytime periods.

Travel Speed	Vehicle Involvements	Vehicle Miles	Rate per 100 MVM
Standing	493	--	--
22 or less	1,183	2,736,000	43,238
23-32	331	28,850,000	1,147
33-37	355	64,497,000	550
38-42	558	250,142,000	223
43-47	698	395,097,000	177
48-52	911	714,925,000	127
53-57	700	513,552,000	136
58-62	441	462,238,000	95
63-72	259	307,786,000	84
73 or more	54	38,841,000	139
Total	5,983	2,778,664,000	215

Source: Table 5, Solomon (1964), p. 12.

Study of the data indicated that involvement rates were highest for the very low speed drivers and reached a low point at approximately 60-65 mph. When these same rates were further analyzed based on the average speed of the highway, the previously mentioned indications of substantially increased involvement rates for large deviations from the mean speed of travel were shown.

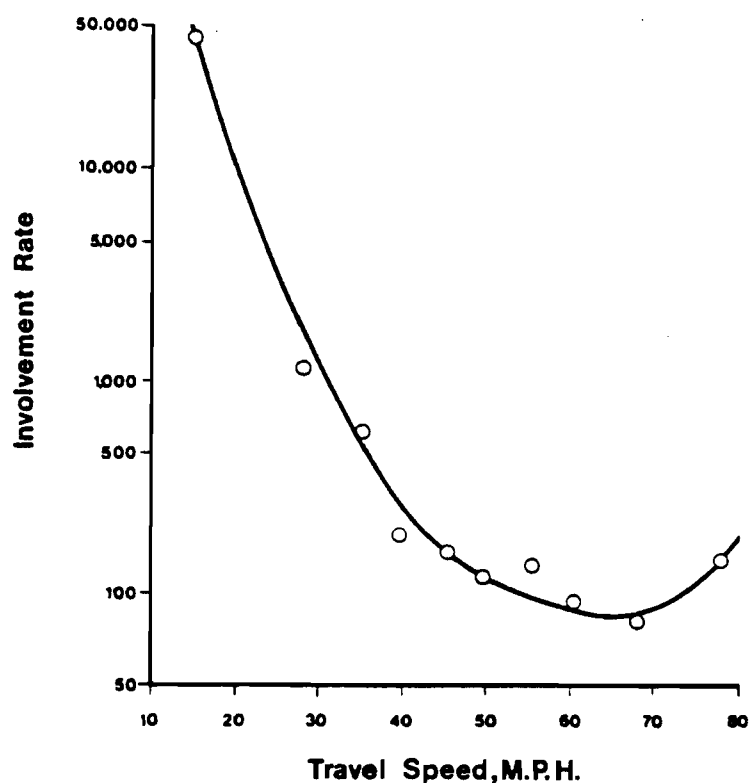


Figure 4.1. Vehicular involvement rate by travel speed on rural highways.
Source: Solomon, 1964, p. 10.

A second indication of the importance of using exposure or "crash opportunity" data in accident analyses is noted by this study in the examination of accident involvements by type of vehicle. As shown in Table 4.2, when the simple number of accident involvements was categorized by vehicle type and the day/night variable, examination of the accident involvements for trucks indicated that trucks with six or more tires appear to be more hazardous than trucks with four tires. However, when vehicle miles is included in the

Table 4.2 Number of involvements by type of vehicle, day and night.

Type of Vehicle	Daytime Accident Involvements	Nighttime Accident Involvements
Passenger car	4,534	3,074
Truck, 4 tires	562	239
Truck, 6 or more tires	780	482
Bus	46	10
Other and not known	61	28
Total	5,983	3,833

analysis and a rate is calculated, as shown in Table 4.3, the trend is reversed. Here, the rates for the trucks with six or more tires are substantially lower than the rates for trucks with four tires in both day and night accidents. Again, this type of descriptive or comparative study clearly points out the need for sound exposure data in order to use rates in drawing conclusions concerning relationships.

Table 4.3 Vehicle-miles, number of involvements, and involvement rate by type of vehicle, day and night.

Type of Vehicle	Day			Night		
	Veh.-miles	Accident Involvements		Veh.-miles	Accident Involvements	
		Number	Rate		Number	Rate
Passenger car	2,186,262,000	4,534	207	530,425,000	3,074	580
Truck, 4 tires	199,765,000	562	281	59,992,000	239	398
Truck, 6 or more tires	374,552,000	780	208	293,198,000	482	164
Bus	17,273,000	46	266	8,437,000	10	119
Other and not known	812,000	61	(1)	460,000	28	(1)
Total	2,778,664,000	5,983	215	892,512,000	3,833	429

¹Rate calculations not meaningful.

Source: Table 25, Solomon (1964), p. 24.

Thus, in summary, the first type of research involving relationships, a study type which has been very important in the highway field when done correctly, is the comparative or descriptive study. The keys to this study are grouping the accidents into various categories of interest for study and combining accidents with some measure of exposure so that rates can be examined within the categories.

Although such descriptive studies have and will continue to be an important part of accident research, the second basic type of research involving relationships is analysis in which the researcher is attempting to develop an equation of a predictive nature which will provide information concerning how a change in a factor or variable of interest (e.g., a change in ADT or lane width) affects the safety measure of interest (e.g., the frequency of accidents). The equation is usually of the form:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 \dots \quad (\text{equation 4.1})$$

where

\hat{y} = the measure of the safety-related variable (e.g., frequency of accident)

x_1, x_2, x_3 = the variables which affect crashes (e.g., lane width, ADT, etc.)

In terms of nomenclature, the factor on the left of the equality sign (usually accidents) is called the "dependent variable" or "outcome variable," while the factor or factors on the right side of the equation are designated as the "independent" or "predictor variables." [The term "independent" has various statistical meanings, some of which will be discussed later, but in the present sense, it refers to variables which do not depend on (are independent of) what is on the other side of the equation. The dependent variable, (y) on the other hand, is being predicted by (is dependent on) the independent variables.]

While there can be more than one dependent variable being predicted simultaneously in a statistical sense, the usual case is to have only one dependent variable in highway accident research. On the other hand, there are always numerous independent variables in any relationship since there are numerous factors which can have an effect on accidents. While "a" in equation (4.1) represents to some extent a baseline value for the dependent variable, the b_1 , b_2 , b_3 , etc. represent the coefficients associated with the independent variables of interest. They are developed statistically when the equation or model is built, and each defines the amount of change in the dependent variable (crashes) due to a one unit change in the independent variable with which the coefficient is associated. For example, in a study of access control on multi-lane rural and urban highways (Cribbins, et al., 1967), the following equation was developed:

$$y = -28.3419 + 0.00011x_1 + 3.28169x_2 + 0.34218x_3 + 0.0005x_4 + 7.34777x_5$$

where y = (predicted) number of injury accidents per mile
 x_1 = access-point index
 x_2 = number of signalized openings per mile
 x_3 = speed limit (mph)
 x_4 = ADT (vehicles per day)
 x_5 = level of service index

Here, b_1 = +0.00011,
 b_2 = +3.28169,
 b_3 = +0.34218, etc.

In this specific example, the coefficient b_2 , associated with x_2 (the number of signalized openings per mile), indicates that as the number of openings increases by one (say, from 1 to 2 or 2 to 3), the average number of injury accidents per mile increases by 3.28169 (everything else holding constant).

Before moving to the specific statistical techniques used in developing and testing such models, the following sections will expand some of the general issues introduced earlier as they relate to research involving relationships.

4.2.1 Sampling considerations: total population versus sample.

When carrying out any type of accident research study, the researcher is almost always forced to use less than the total population of accidents because of problems concerning underreported or biased data (see Chapter 2). Because the basic rationale for conducting the research is to be able to draw conclusions that can be used in the future concerning the total population, the researcher is assuming that whatever relationships or causal factors are identified in the sample would also be found if the researcher were somehow able to examine the total population of accidents involving all drivers at all times. The more biased the sample, the less faith the researcher will have in his conclusions or inferences concerning the total population. As pointed out in the discussion of countermeasure evaluation, while the lack of generality of the sample is often obvious, in some cases the issues are much more subtle.

A specific example of a case in which a non-representative sample was drawn (or at least planned) for use in making inferences concerning the total population is found in a study conducted by Fee, et al. (1970). Of interest here is the fact that, while the following discussion will indicate the proposed sampling plan's obvious bias, this bias was far from obvious to the engineers conducting the data collection phase. (Again, hindsight is far more powerful than foresight.) Fee was attempting to analyze various underlying relationships associated with the level of safety on Interstate roadways. Data were to be collected from the majority of states with Interstate mileage or who were planning to build such mileage. The actual accident, traffic, and roadway characteristics data were to be collected by the state highway departments and submitted to FHWA for analysis. Thus, the actual sampling was carried out by the states, although instructions had been provided by FHWA.

The specific problem being discussed arose in one state collecting the data. Because the local state engineers did not wish to accumulate data on all segments, they had planned an alternative sampling scheme for collecting data. In this scheme, the engineers decided to use only roadway sections in which one or more accidents had occurred. Perhaps without planning to, they were, in reality, drawing a sample from the total population, a sample consisting of those sections where one or more accidents occurred. Indeed, this would be a rather limited sample since most highway segments that are in the total population would not experience an accident in the time period analyzed.

With thought, it becomes apparent that while conclusions drawn from this sample could be generalized to the limited sample of segments in the total population which have experienced previous accidents, they could not be generalized to all highway segments. By failing to include a sample of the non-accident sections, the sampling design had eliminated an entire segment of information needed to define the true relationships involving the highway system. The biased sample would indeed provide relationships, but the relationships would be very limited in their application.

The problem that results from this study (sample) limitation stems from the fact that the results of every research study are implicitly assumed to pertain to all roadways. After all, without prior knowledge of which highway segments would experience future accidents (which doesn't exist), the administrator must apply the research conclusion to all parts of his system. As shown in this study, the engineers had defined a sample (probably without specifically realizing it), were drawing the sample, and the researchers would have analyzed it and formulated conclusions which the decision maker would apply to the entire roadway system. Because of the failure to note that "zero is a good number" in terms of accident research, the sample was severely biased and the resulting conclusions would have been, at best, of limited utility. (Fortunately, the biased sample scheme came to light in discussions with FHWA researchers early in the sample period. The sampling plan was immediately modified to include sections with zero accidents.)

In summary, there are many instances in which the researcher can be forced to draw conclusions concerning the total population of accidents when (either known or unknown to him) the data he is working with may indeed not be the total population or a representative sample. Thus, the underlying importance of the sample vs total population issue lies in the fact that the researcher will indeed be forced to draw conclusions which he assumes will be valid for the total population from whatever sample he has, and his knowledge of the sample and the population will make him a much better judge of whether or not such inferences are indeed valid.

4.2.1a Estimating required sample size.

In the earlier discussed establishment of an experimental design, the estimation of a required sample size was noted to be extremely important because one of the main determinants of a successful analysis of a countermeasure is the advanced planning in which the researcher determines how large a sample will be needed in order to be able to detect the required level of effect. While a parallel issue exists in accident research involving relationships, in truth the issue is far less important. Most research involving relationships will depend upon the use of large computerized files of accidents and characteristics. Because the data are already collected, the researcher will use as much data as are available in order to insure that his sample is representative of the population his inferences will concern. In addition, because the statistical techniques to be discussed in this chapter are also to a large extent computerized, there is seldom a need to draw a sample from the computerized file. Thus, the prime consideration in the sample size used in a study involving relationships is simply economics--the researcher will collect and use all the data he can afford.

Unfortunately, there will be some cases in which the data are not computerized and thus the researcher is forced to use hand tallies of information or to convert file cabinets of hard copy information into a coded computer format. Unlike the situation depicted in Chapter 3 in which there are established rules and mechanisms (and indeed books) which give the researcher definite guidelines for determining sample size under given assumptions, there are no such guidelines in the area of research involving relationships. Careful review of statistical textbooks bear this out. In general, in order to define a minimum sample size for a study involving relationships, the researcher needs to have some information about the basic underlying relationship between each of the independent variables and the dependent variable. Specifically, if the variables are of a continuous nature (e.g., lane width) the researcher would need to know something concerning the correlation of the independent variable with accidents. If the independent variable is more categorical in nature, (e.g., weather) the researcher will again need to have some indication of how different weather conditions are related to accidents. If such information is available, the researcher could follow the procedures specified in Chapter 3 to calculate a sample size required to show statistical significance at a given level for each independent variable and choose the maximum sample size calculated. However, in research involving relationships, such knowledge does not exist. Indeed the reason for carrying out such research is to determine or define such knowledge. Without guidelines, the researcher is simply faced with a situation when he or she should use all the data that are available.

However, if the researcher finds himself in a situation which requires sample size in advance (e.g., whether to draw a sample of 50, 500, or 50,000 accident reports), one method which can help generate crude estimates involves choosing the critical independent variable (or contingency table variable) -- the single variable of interest which will "naturally" have the least data in the real world of accidents. For example,

in a descriptive accident study involving truck size by highway type, one might expect to find the least number of large truck accidents on secondary roads. Thus the critical accident subsample would be on secondary roads. If the researcher can specify the degree of difference in the accident rates for different sized trucks on these secondary roads that is meaningful to him, the techniques described in section 3.8.3 can be used to determine sample size for this particular subsample. Then, if the researcher knows (or can determine from a small "pre-sample") the proportion of total accident reports which concern this type accident (i.e., heavy trucks on secondary roads), an estimate of total sample size can be found by dividing the calculated subsample size by this proportion. Again, because such a process requires more advance knowledge than is usually available, the basic guideline is to use all the data economically available.

4.2.1b Choosing a representative sample.

Because there may be special situations in which the researcher wishes to (or is forced to) choose a sample when examining relationships, and evaluations in which a choice of location/accidents is possible, some discussion of proper sampling techniques is warranted. This general discussion will not in any sense present detailed information concerning sampling. Entire textbooks have been written on the subject, and the reader who is interested in increasing his knowledge in the area of sampling techniques should refer to the appropriate references which have been included at the end of this chapter (Cochran, 1977; Deming, 1963).

The major ensuing requirement in choosing any sample is to have the sample be representative of the population from which it is drawn. Thus, the basis for most sampling techniques is to assure that each observational unit (e.g., each record containing accident or characteristics information, or each section of roadway) has the same chance of being included in the sample as any other observational unit or data point. If this basic requirement is met, then such data collections as "samples of convenience" (choosing data points which are easiest to acquire) are not allowed because the data points that are not as easy to acquire do not have an equal chance of being included in the sample.

(Technically, the above "equal-chance" statement is true only in "simple random sampling". Other more complex schemes weight certain data subgroups more heavily, meaning that all units did not have equal chances of being included. Some discussion of stratified sampling is included later in this same section. However, in general, the purpose is again to have the sample be representative of its population.)

The best way to guarantee that each data point or each observation has an equal chance of appearing in the sample is through what is called random sampling. The term simply signifies that, through some mechanism, each data point is provided an equal chance of being drawn. One method of accomplishing this end would be to place each observational unit in a hat, blindfold a sampler, and have him draw out the number of units that need to appear in the sample. In this way each of the units presumably has the same chance of being drawn. Mechanically, this is a very difficult and time consuming procedure if large sets of data are being sampled.

A second method of random sampling involves the use of published random number tables found in most statistical texts. These are simply columns of numbers which have been generated to guarantee that they fall in random order. In this procedure, each unit that is in the population (or sampling frame) is first assigned a unique number. The researcher then refers to the table of random numbers, starts at some random point on the page, and then reads down the columns of numbers, thus determining the numbers of the specific units to be chosen.

In truth, however, when working in relationship research, the normal situation is to have very large numbers of accidents and related characteristics on a computerized system. Using random number tables to generate random samples and then going into the computerized files to pull out these specific cases can be a very time consuming and expensive operation. An alternative procedure which can be much more efficiently accomplished with computerized data is to draw a systematic sample with a random start. In this procedure, the researcher divides the required sample size (from above) by the total population available to determine the proportion of the population that will have to be drawn. For example, if a sample size of 10,000 highway segments was required and there were 100,000 segments on the file, then the researcher would obviously need to draw one segment from each ten on the file. To draw the systematic sample with the random start, the researcher would go to a random number table and randomly find a number between one and ten (most easily done by closing eyes and pointing to the page while thinking random thoughts). This number, say "6," would give the starting point in the first ten units of the computerized file. Thus, the first unit to be drawn into the sample would be highway segment or accident number 6. From that point on, the computer would simply count in units of 10 and extract each tenth unit into the sample. In this particular example, units

numbered 6, 16, 26, 36, etc. would be included in the random sample. While such samples are not quite as technically sound as a totally random sample, they represent what is considered to be a very suitable substitute for a random sample, provided there are no biases in the way the file is structured, and they are often much less expensive to acquire.

A further refinement to the above described random sampling procedure is stratified random sampling. As detailed in Cochran (1977) and Deming (1963), stratified random sampling is particularly appropriate in cases where there is a need to obtain an accurate estimate of the variable (e.g., mean accident rate) under study but where, for economic reasons, the total number of sampling units which can be chosen is small.

The basic difference between a stratified random sample and the simple random sample described is that a simple random sampling draws units from the total population, but a stratified random sampling first stratifies or subdivides the total population into meaningful subcategories and then draws a random sample from within each subcategory. The categories are formed by grouping data which have similar (homogeneous) characteristics. Specifically, in the current context, past highway accident research would indicate that it might be appropriate to stratify accidents according to such variables as highway type, urban-rural location, and speed limit. While the referenced textbooks present numerous types of and methods for drawing stratified samples, one appropriate technique to follow in the study of highway relationships is described below. In the example used, it is assumed that the researcher wishes to stratify accidents based on urban/rural location and three highway types (Interstate, other primary, and secondary). In this case there would be six strata possible (e.g., urban Interstate, rural Interstate, urban primary, . . .).

Step 1. The accident population to be sampled from should be categorized into the various strata.

Step 2. Within each stratum the accident units are numbered and the number of total units (accidents) within each stratum is counted.

Step 3. The total sample to be used is drawn by randomly drawing the same proportion of units from each of the strata. The random drawing procedure is the same as described above. For example, a "10 percent sample" would mean that 10 percent of the accidents in each stratum is drawn. This will, course, mean that the larger strata (e.g., rural primary) will contribute many more accidents to the final sample than do the smaller strata (e.g., urban Interstate). However, the proportion included in the final sample will be representative of the proportion of accidents in the total population that occur on each roadway type.

Step 4. After the accidents are randomly drawn from each of the strata, they are merged into the data set to be analyzed.

This procedure results in a total sample in which each of the subcategories that are felt to be important are "forced" to have representatives in the final sample. Although the subsample sizes are different from stratum to stratum, each stratum will have some representation in the final group of accidents to be studied. This is not necessarily guaranteed in the earlier described simple random sample. For example, in drawing a simple random sample it is possible to "miss" all accidents occurring on urban secondary roadways simply because a very small number of accidents occur on such roadways. Stratification before sampling eliminates this possibility. (NOTE: There are methods other than this "equal proportion" technique for drawing samples from each strata, particularly in cases in which the researcher wishes to "force" a larger sample size in a given strata for improved accuracy in making estimates for that stratum. These will not be discussed, but the interested reader should refer to the previously cited references.)

After the drawing of the sample, the researcher continues with the analysis of the data and the development of underlying relationships. It is noted here that the reasoning for stratified sampling in this particular context (the development of relationships) is somewhat circular in that the researcher must make use of information concerning the underlying relationships between accidents and other variables in forming the strata--the very information that is to result from the predictive model to be developed. Thus, knowledge of prior research is important in this particular procedure.

4.2.2 The choice of dependent variable.

As noted earlier, this chapter concerns research questions in which the researcher is attempting to examine relationships between accidents and other factors or variables. Generally, these factors are used to develop an equation (usually of a predictive nature) of the form shown in (4.1) (see section 4.2). As noted, the researcher must first define a dependent variable and a series of independent variables.

4.2.2a Accidents as the dependent variable.

Just as in evaluations, the proper choice of dependent variable is based primarily on the purpose of the research being done, the program being evaluated, or the element under study.

As noted in the preceeding chapter, a simple means of beginning to determine the proper criterion in an evaluation is to ask the question, "What is the countermeasure intended to do?" In like manner, the researcher should ask a similar question in research involving relationships, i.e., "Which accidents should be related to the predictor variables of interest?" He should then limit the data used as the dependent variable to those which are most likely to be related.

If accidents, or some subset of accidents, are used as the criterion, should accident frequencies or accident rates be used? Should the researcher attempt to define the relationships between various independent variables and accidents per driver, accidents per hundred million vehicle miles, accidents per location, per vehicle, or should he simply use the number of accidents?

From the discussion in Chapter 3, some measure of "crash opportunity" -- exposure--should be included in any research involving accidents. Without such a measure of exposure, the interpretation of the results of the research is very difficult and at times almost impossible.

The question, however, concerns which side of the equation exposure should be entered in. First, exposure could be accounted for on the left or predicted side of the equation if the dependent variable used is an accident rate per million vehicle miles, per hundred vehicles, etc. As an alternative, the measure of exposure could be entered as one of the independent or predictor variables on the right side of the equation. At first glance it would appear that because the use of rates gives more stability to accident data, it would be more logical and more appropriate to include the exposure measure as part of the dependent variable on the left by using a rate-based dependent variable. In addition to appearing more logical and providing some apparent stability, research has shown that average daily traffic is so heavily related to accidents that there might be some question concerning whether it would not cover up or mask the effects of other independent variables which are really of interest if used as an independent predictor.

In most cases, independent variables other than ADT are the ones that the researcher is most interested in since they are the ones he has some control over. For example, take a situation in which the researcher is trying to define the relationship between accidents and such independent variables as number of lanes, pavement width, presence or absence of paved shoulders, curvature, and superelevation. In a sense, he would have control over these factors in the future design of highways and thus would be more interested in their effects. He would, on the other hand, have less control over the average daily traffic that use these facilities since ADT is generally a function of the user demand which, in turn, is based on societal economics, living patterns, shifts in population centers, and many other factors over which the engineer has very little control. Thus, it might appear more desirable to delete exposure from the list of independent variables.

In contrast, however, there appear to be some rather valid arguments from a statistical point which would indicate that exposure might much more appropriately be included as an independent variable on the predictor side of the equation. The arguments are as follows:

First, the evaluation of rates is usually accomplished by simply dividing the number of accidents by the exposure measure, ADT. However, when this is done, an implied assumption is being made that these two variables are linearly related. That is to say, a unit increase in ADT will be accomplished by a unit increase in accidents throughout the entire range of ADT. If the relationship is not linear, if, at some point in the ADT range, a unit increase in exposure results in say, a two unit increase in accidents, the resulting rate used as the dependent variable will be somewhat inconsistent in terms of its ability to be predicted by other independent variables. Past studies have shown that ADT is not precisely linearly related to accident rate. As shown in Figure 4.2, a very pertinent study was conducted by Kihlberg and Tharp (1968) in which the authors found an exponential relationship between average daily traffic and accident rate. The relationship changes slightly when freeway accident rates and ADT are compared, but the fact remains that, when accidents per mile are compared to average daily traffic, the relationship is slightly nonlinear.

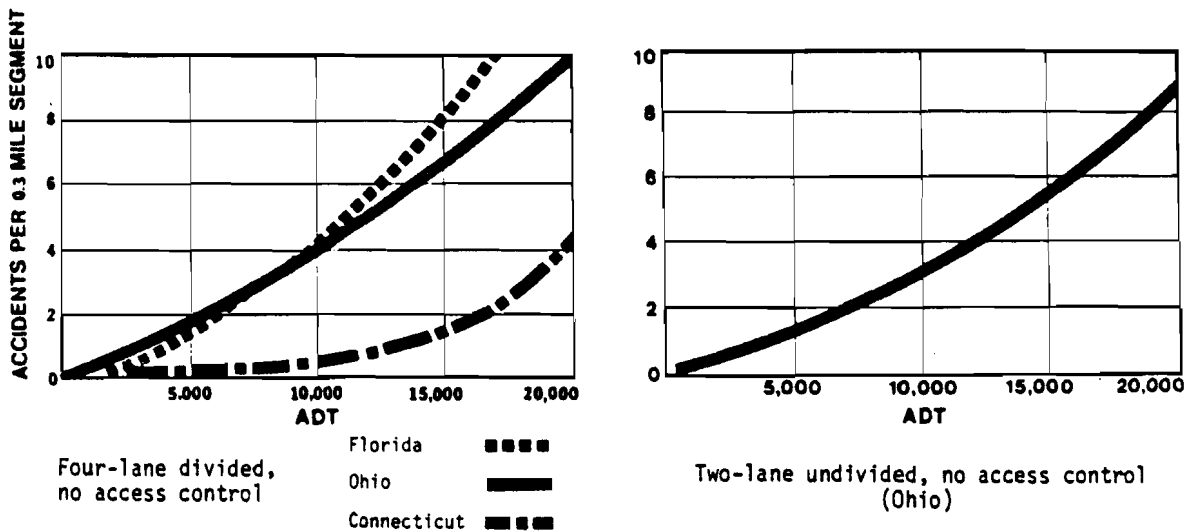


Figure 4.2. Accidents and ADT on four-lane and two-lane rural roadways.
Source: Kihlberg & Tharp, 1968, pp. 63, 65, 70, 74.

Second, and of a more subtle nature, if the exposure measure is used as a denominator in a rate on the left or dependent variable side of the equation, there may be times when the predictive nature of the equation developed may be more dependent on the exposure measure than on accidents. That is to say, when the model is being built, the researcher is attempting to define a relationship related to accidents. However, if the independent variables being used to predict accidents are more highly related (highly correlated) to average daily traffic than to accidents, then what might appear to be a good model for predicting changes in accidents may really be predicting changes in ADT.

A final point concerns the fact that ADT may be highly related to other independent variables, producing what are discussed later as interactions or interactive effects. For example, the relationship of accidents to number of lanes may be dependent on the average daily traffic on the different numbers of lanes. Accident rates on two-lane roads may differ widely if the ADT on the two-lane roads differs from low volume to high volume roadways. If such relationships exist, these can be accounted for in a model by using interaction terms. However, in order to include an interaction term, there is a need to include both the main factors which interact (ADT and number of lanes) on the right or predictor side of the equation.

Thus, in summary, while these arguments may appear to be somewhat theoretical in nature, there does at least appear to be some statistical evidence indicating that if there is a choice, the researcher should use the exposure measure as an independent variable on the right side of the equation rather than using it as the dependent variable to develop accident rates.

4.2.2b Crash severity as the dependent variable.

Just as with countermeasure evaluations, the answer to the basic question--"what can the independent factors being examined be expected to affect?" will sometimes involve crash severity rather than crash frequency. For example, if a researcher were attempting to examine the relationships between accidents and the presence or absence of various roadside safety devices, such as guardrails, crash attenuation systems, breakaway signs supports, etc., he might be examining a function of the following form:

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \dots$$

where

x_1 = the number of feet of guardrail in a given
mile of roadway

- x_2 = the number of breakaway sign supports
in a given mile of roadway
- x_3 = the number of crash attenuation systems
in a given mile of roadway

Here the researcher is attempting to determine the relationship between the roadside hardware described above and some dependent variable associated with the level of safety. But what is the proper dependent variable? Should " \hat{y} " in the equation be accident frequency or should something else be used? With thought, the researcher would conclude that, while the number of feet of guardrail, the number of breakaway sign supports, and the number of crash attenuation systems may slightly affect the number of accidents (or might even result in an increase if such devices are placed in what otherwise would be clear roadside recovery area), these devices are each designed only to reduce the severity of a crash. Thus, the dependent variable of principal interest would most logically be a measure of accident severity. (Because these devices may increase the incidence of accidents, both frequency and severity measures may be desirable in certain studies.)

If crash severity is to be studied, what is to be the character of this dependent variable? When accident frequencies were used in the preceding section, it was obvious that the dependent variable would be a count of the accidents themselves. However, when severity measures are used, the nature of severity measures themselves results in a choice which is not so obvious. In the earlier discussion of countermeasure evaluations where severity was a criterion, emphasis was given to examining shifts in driver injury distribution--a shift from a more serious to a less serious distribution. However, in the type of equation that is being examined here, examining shifts in distributions is not quite as simple. Rather than a shift in distribution, a variable must be defined which is capable of being counted. Choices could include the number of total injuries, the number of injuries per vehicle, the number of fatal injuries per vehicle, the number of serious plus fatal injuries per vehicle, the number of vehicles experiencing damage above a certain level, or some other measure. (The same issues are discussed in Chapter 3; the reader should refer to section 3.5.3b for details.) Consideration of those issues indicated that one possible severity-related dependent variable which appears to overcome at least some of the problems discussed would be the number of injuries experienced by the driver per crash or the proportion of drivers experiencing these injuries. By using driver injuries, the problem stemming from differential occupancy rates is overcome (almost all vehicles at least have a driver present). The issue of not detecting shifts within the injury distribution can be overcome to some extent by using only moderate and severe injuries (including fatalities). By using this measure of moderate or severe driver injuries per vehicle, or the corresponding proportion of vehicles in which the driver experiences a moderate or severe injury, some control is also gained over the number of vehicles that are on the highway and the related crash opportunity level. Of course, in using driver injury, injury-related factors other than those associated with the highway would also have to be included in the equation (e.g., restraint use, age, etc.).

4.2.2c Intermediate measures as the dependent criterion.

In almost all analyses of relationships, the researcher will be using either some measure of crash frequency or crash severity as the dependent variable. Very little relationship research involving proxy measures is found in past literature. However, if such instances arise due to the lack of sufficient accident data coupled with the availability of some surrogate or proxy data, the issues discussed in section 3.5.3c should be reviewed. It is noted, however, that the designation of a sound proxy measure can only be accomplished through the type of research being discussed. The establishment of the link between accidents and a possible proxy measure is the result of this type study.

4.3 Analysis Techniques

Having now discussed some of the general issues which the researcher must face in the development of relationships in accident research, let us turn to the more technical issues concerning the statistical techniques appropriate for use in this work. The following sections may appear rather lengthy, but the material included represents an overview of what would normally be covered in an entire series of statistical courses, each with its own text. To accomplish this in as few pages as possible, decisions were made based on the following points. First, the user will be assumed to have some basic knowledge of statistical terms and processes. Second, the coverage provided to individual techniques will be limited, and the user may wish to refer to the additional references provided for further explanations. Third, underlying theory will be kept to a minimum, and in places, for simplicity and clarity, some liberties will be taken with statistical notation normally followed.

The material provided is organized into two sections. First, an introductory section provides an overview of the procedures to be covered including keys to the choice of proper procedure. A guide diagram is provided to aid the researcher in choosing the most appropriate procedure for his situation. Second, descriptions of each technique including examples and assumptions are presented. (The reader should also refer to the glossary of terms presented in Section 3.8.1.)

4.3.1 Introduction to Statistical Analysis Techniques to be Presented

Again, this manual is not intended to be a statistical text. Various well-written and easily obtainable texts including those referenced in this chapter present detailed discussions of statistical techniques aimed at identifying relationships among variables. Indeed the techniques abound in such numbers that the sheer magnitude of available tests can be both confusing and disturbing to the engineer/researcher. In the following discussion, an attempt has been made to reduce this large array of techniques to a more manageable number and to present a limited number of details and examples concerning the techniques which are particularly appropriate for use with accident-related research.

As a basic guideline, the manual user should remember that the overall goal of the methodology being discussed in this chapter is simply to examine, determine and quantify relationships between accidents and other variables. In meeting this goal, statistical procedures are used. All of these procedures have three basic purposes.

First, certain procedures are intended to aid in preliminary screening of available variables to see if these variables are related to our criterion of interest (e.g., accidents or injuries). That is, the researcher is trying to determine which of the many independent or predictor variables that are available to him have any "real" (not chance-related) relationship with accidents. Through this screening process, the researcher is often able to eliminate variables which will be of little use to him in subsequent analysis and therefore to reduce his analysis workload in the steps which follow.

The second purpose of statistical procedures, and one which is most pertinent to the research described in this chapter, is to help determine which specific variables are strongly enough related to be included in a model and to help determine the degree or extent of the relationship in order to precisely define the components of the model. As presented in Section 4.2, the mathematical model to be developed is usually of the form

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 \dots \quad (4.1)$$

where

\hat{y} = the measure of the safety related variable (e.g., frequency of accidents)

x_1, x_2, x_3 = the independent variables which affect crashes (e.g., lane width, ADT, etc.)

b_1, b_2, b_3 = coefficients associated with the independent variables which define the amount of resultant change in the dependent variable related to a unit change in the specific independent variable (all else held constant)

Thus, the second purpose of the various statistical procedures is simply to define the model--to calculate the values of the coefficients.

Finally, having developed a potential model, a third group of statistical techniques will be used to examine the model to see first, how well the overall model predicts y (e.g., accidents) and second, whether each of the individual relationships depicted in the model are real rather than chance related. In a statistical sense, the second goal, that of examining the individual relationships, is simply a statistical test to determine whether or not each of the coefficients in the model is significantly different from zero. If a coefficient is not significantly different from zero, that particular variable can be deleted from the model.

Thus, while reading the following sections, the manual user should remember that each of the techniques described is simply a tool to help the researcher:

1. Conduct preliminary screening of variables to detect presence of relationships,
2. Develop a specific model, or
3. Test the model to measure its predictive accuracy.

In keeping with this goal orientation, the techniques to be discussed will be categorized according to which of these three goals it helps meet. Each of the resulting three major categories will be further divided into two parts, with the parts defined by the nature of the data to be analyzed: Are the data continuous or categorical in nature? Second, if continuous, are the data distributed approximately according to a normal distribution? If the data base is categorical in nature, is it nominal, ordinal or scalar (see Glossary, 3.8.1)?

To aid the reader in finding the proper test to be used in a specific instance, a guide diagram is presented in Figure 4.3. By following the branches of the tree, the reader should be able to find the appropriate test and the appropriate section to refer to in the following pages.

4.3.2 Variable Screening Procedures

The procedures discussed in this section are statistical methods which aid the researcher in the earlier described preliminary screening process. Here the researcher is attempting to determine whether or not an association or relationship between two variables is present. As will be noted, many of the procedures that aid in this determination of the presence of association do so by outputting a number whose magnitude (without regard to positive or negative sign) defines the strength of the association. This number ranges from -1.0 to +1.0, and the closer to -1 or +1, the stronger the association. A resulting number close to zero indicates a lack of association between the two variables (also stated as "the two variables are independent of each other" in statistical terms). The sign of the resulting number defines the direction of the relationship.

For example, if a test of the relationship between speed deviation and accidents resulted in a number of +.99, the researcher would conclude that a strong direct relationship is present and that, as one of the variables increases, the second also increases. If the resulting number had been -.95, the researcher would conclude the presence of a strong inverse relationship, i.e., as one variable increases, the second variable decreases.

In the engineering and physical sciences, when a researcher is investigating the degree or strength of association between two variables, a measure of 0.8 or higher is common. In the field of accident investigation, however, such high levels of association are rare. The researcher must be willing to accept far lower values, starting from around ± 0.4 -0.5. The low levels of association in accident studies may be attributed to the very nature of the relationship. In the physical sciences the relationships are most likely to be simple, direct ones, but in accident research, the relationships are often very complex, and changes in accidents are often the result of the interplay of many factors. Thus, any single factor will not usually be highly related to accidents. The use of these low levels is further justified in screening procedures since the strengths of the relationships will be tested again in the model development process.

Let us now turn to the individual tests to be used in this determination of the presence of an association. Presented first will be those procedures suitable for use when the two variables to be examined are both continuous. The second group of tests is appropriate for those variables which are categorical (either nominal or ordinal) in nature.

4.3.2a Simple Presence of Association Between Two Continuous Variables

The statistical procedure most commonly used to detect the presence of a relationship between two variables (e.g., accidents and ADT) is Pearson's product moment correlation. In the strictest statistical sense, the procedure requires that both variables be normally distributed (as in Figure 4.4). However, the procedures appear valid when the distributions of the variables are non-normal. For example, the procedure appears valid for accidents per location at certain locations even though these are distributed according to a Poisson distribution (see Figure 4.5). For example, the procedure would be appropriate in the latter two distributions shown where the mean number of accidents per location (λ) is equal to five or more.

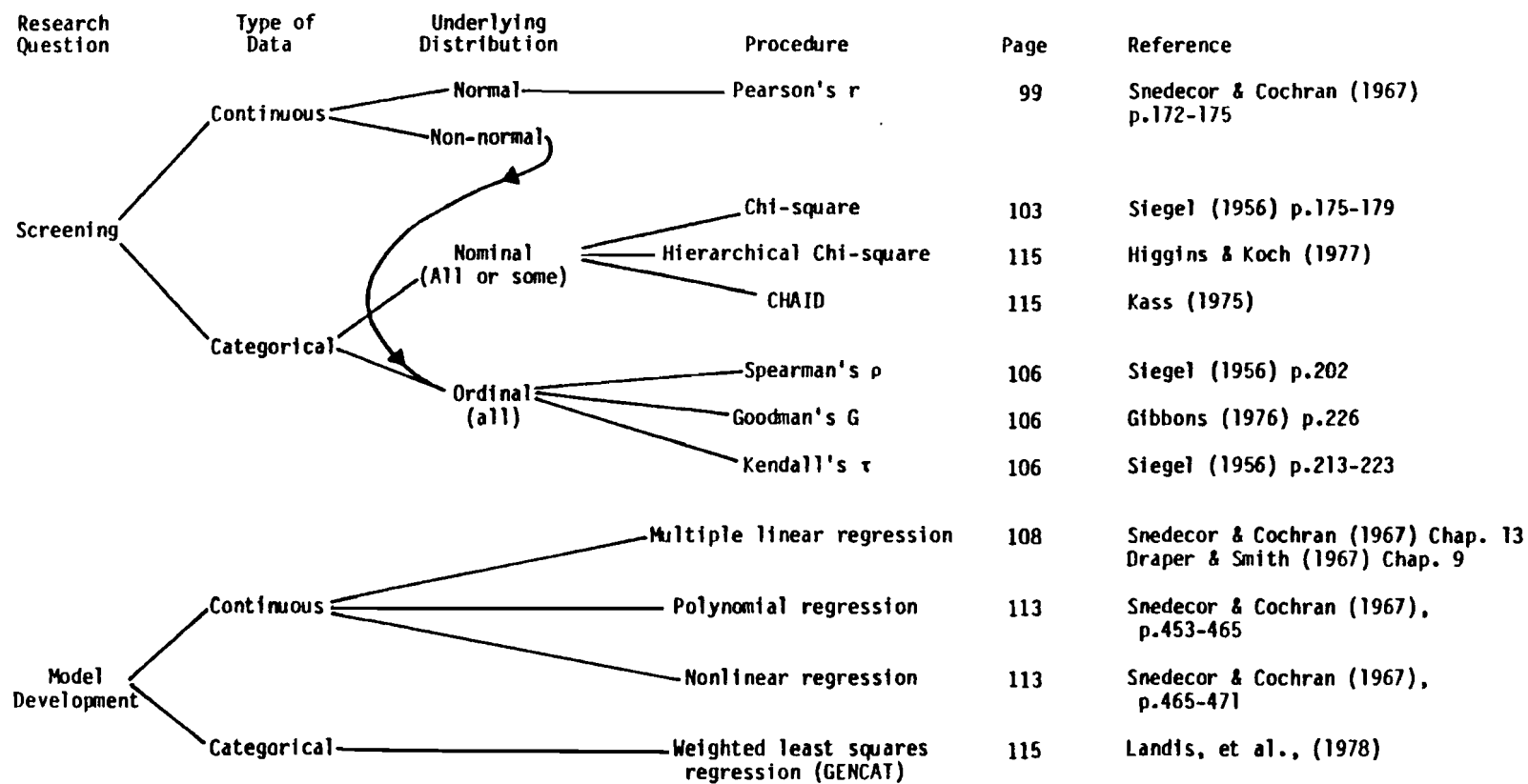


Figure 4.3 Guide to appropriate statistical procedures.

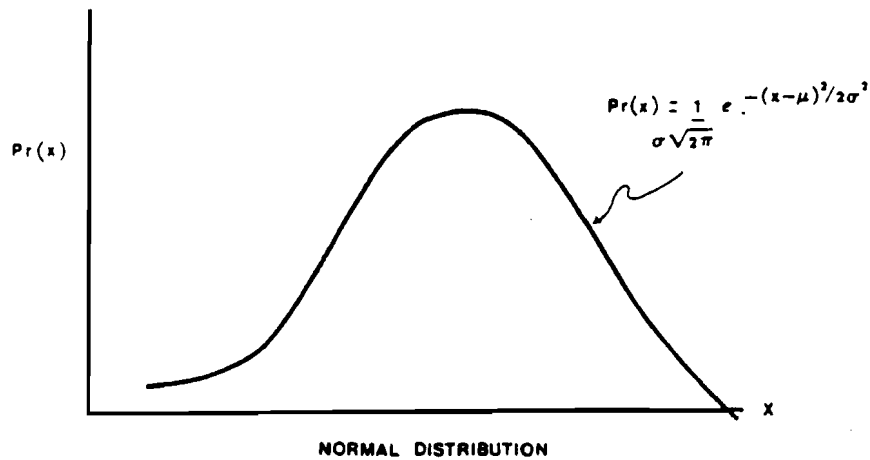


Figure 4.4

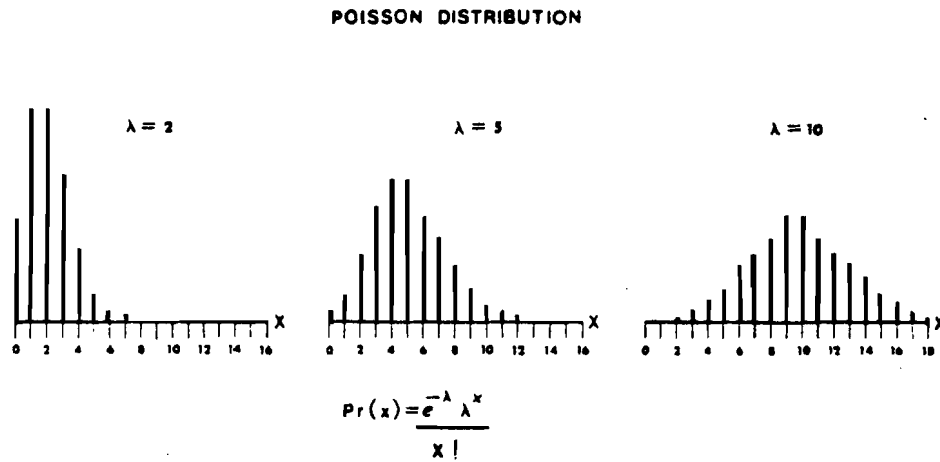


Figure 4.5

If, however, the distribution of one of the variables being studied is known to have a distribution very dissimilar to the bell-shaped normal distribution (e.g., $\lambda = 2$), the researcher should subdivide his data into categories and use one of the tests described in the next section. The formula for Pearson's r is as follows:

Statistic

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{[\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2]^{1/2}}$$

where

y = one variable of interest (e.g., accidents)

x = second variable of interest

$$\bar{x} = \frac{1}{N} \sum_i x_i \quad i = 1, 2, \dots, N; \text{ likewise for } \bar{y}$$

Example

1. Purpose: To examine the relationship between accident frequency and speed standard deviations, and between accident frequency and median width.

2. Data:

Location	Number of Accidents y	Speed Deviation x_1	ADT (in 10,000's) x_2	Median Width x_3
1	0	6.10	0.770	36
2	0	6.30	0.664	44
3	0	6.10	0.366	54
4	1	6.60	2.073	41
5	1	6.70	1.368	32
6	1	6.80	1.673	37
7	2	9.60	1.987	33
8	2	7.00	2.245	23
9	3	7.60	0.847	27
10	3	7.90	2.739	29
11	3	8.10	2.245	23
12	4	8.30	2.421	24
13	4	8.55	3.034	23
14	5	8.20	2.733	24
15	6	9.10	2.854	21
Total	35	112.95	28.019	471

Source: Interstate System Accident Research, U.S. Department of Commerce, Bureau of Public Roads (1961). For this example some of the numbers have been modified.

3. Compute:

$$r_{xy} = \frac{\sum_i (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\left[\sum_i (x_{1i} - \bar{x}_1)^2 \sum_i (y_i - \bar{y})^2 \right]^{1/2}}$$

where x_{1i} = i -th observation of x_1
 \bar{x}_1 = sample mean for $x_1 = 7.53$
 y_i = i -th observation of y

\bar{y} = sample mean for y
 $= 2.33$

and, for example with $i = 9$,

$$(x_{1i} - \bar{x}_1)(y_i - \bar{y}) = (7.60 - 7.53)(3 - 2.33) = 0.0469$$

$$(x_{1i} - \bar{x}_1)^2 = (7.60 - 7.53)^2 = 0.0049$$

4. Conclusion: For this data,

$$r_{yx_1} = 0.81$$

$$r_{yx_3} = 0.84$$

5. Interpretation: Here the Pearson's r for accidents and speed deviation ($r_{yx_1} = .81$) indicates that a strong positive relationship exists. The corresponding Pearson's r for accidents and median width ($r_{yx_3} = .84$) indicates a slightly stronger relationship.

Extensions of This Procedure

Further extensions of this procedure involve situations in which the researcher wishes to examine the relationship between two variables while controlling for, or taking into account, the effect that one or more other variables has on the two variables being studied (e.g., the presence of association between accidents and speed deviation while controlling for ADT which could affect both accidents and speed deviation). In such a case, the researcher could use the Pearson's r for partial association (see Snedecor and Cochran; 1967, p.400).

The researcher must also note that the two Pearson's procedures described above are measures of the degree of linear relationship only. Thus, two variables may be closely associated by a curvilinear relationship and yet their measure of association can be zero. Thus it is recommended that the researchers plot a scattergram of the two variables and examine the trend of the relationship. If a linear relationship is indicated, the interpretation of the Pearson's r is fairly unequivocal. If, however, the scattergram indicates a non-linear or curvilinear relationship, then a low level of association should not be construed as a lack of relationship between the two variables. The variables should be retained for use in the later model building process, and the researcher should consider a polynomial or non-linear regression procedure in building the model (see Snedecor & Cochran, 1967, p. 453).

4.3.2b. Simple Presence of Association Between Two Categorical Variables

Just as in the preceding section, the researcher is again attempting to detect the presence of a relationship between two variables. However, in the previous section, the researcher was working with continuous variables and could assume that both the variables were distributed according to a normal (or approximately normal) distribution. The procedures presented in this section are used when one or both of these assumptions is not true--when at least one of the variables is continuous but known to be non-normal or when a variable of interest is categorical in nature.

In such cases, the categorical data are usually either nominal, where the categories are described by name only (e.g., light condition: dawn, daylight, dusk, darkness) or the data are ordinal, where there is an order implied by the levels of the variable (e.g., degree of injury: none, slight, moderate, serious, fatal; or condition of pavement: poor, fair, good).

If either one of the variables of interest is nominal (not ordinal) in nature, the most appropriate procedure is to calculate Pearson's Chi-square statistic for the contingency table formed with the values of one variable along the side (rows) and the values of the second variable across the top (columns).

Statistic:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(N_{ij} - M_{ij})^2}{M_{ij}}$$

where

N_{ij} = observed number of fatalities on road class i with light condition j

M_{ij} = expected number of fatalities in this cell under the assumption of independence (or no association)

$$= \frac{(\sum_i N_{ij})(\sum_j N_{ij})}{\sum_i \sum_j N_{ij}}$$

To test for a statistically significant association, one compares the calculated with the critical value of Chi-square for a given significance level (say, $\alpha = .05$) with degrees of freedom (d.f.) equal to [(number of rows - 1)(number of columns - 1)].

The researcher should note that Pearson's Chi-square is strongly affected by sample size in that if large samples of accidents are being studied (which is usually the case in studies of relationships), the χ^2 will prove to be statistically significant even for relatively weak associations. To make this procedure meaningful for the large sample cases, the contingency coefficient

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

should be calculated after the χ^2 is calculated and proves to be significant. The contingency coefficient can also vary from 0 to nearly 1, with values close to 1 indicating a very strong relationship. Again values of C greater than 0.5 could be assumed to indicate the presence of meaningful association.

Example

1. Purpose: To test for an association between number of accidents and road condition. (Cell entries are number of locations.)

2. Data:

Number of Accidents	Road Condition				Row Total
	Dry	Wet	Muddy Oily	Snow Ice	
0	98	14	12	10	134
1	14	85	7	8	114
2	25	13	43	6	87
3 or more	10	19	24	24	77
Column Total	147	131	86	48	412

3. Compute:

$$\chi^2 = \sum_{i=1}^4 \sum_{j=1}^4 \frac{(N_{ij} - M_{ij})^2}{M_{ij}} = 266.1$$

where for example, for column 1, row 1,

$$N_{11} = 98$$

$$M_{11} = \frac{(134)(147)}{412} = 47.81$$

4. Conclusion: For $\alpha = 0.05$ with $(4-1)(4-1) = 9$ d.f., $\chi^2_c = 16.9$. Since the calculated χ^2 exceeds 16.9 there is a significant association between road condition and number of accidents.

Note: As the Chi-square statistic is affected by sample size, the contingency coefficient is determined:

$$C = \sqrt{\frac{266.1}{266.1 + 412}} = 0.63.$$

This is a rather high contingency coefficient and indicates that the significance of the relationship is not due to sample size.

A further use of the Chi-square statistic, and indeed the one more often seen, is in the examination of existing relationships between two variables (one or both of which are nominal) in terms of a third variable. While this usage--contingency table analysis--is not directly related to a screening procedure, per se, it is an important categorical data analysis technique which is often found in the accident research literature.

For example, the researcher might wish to examine the presence of a relationship between two variables (e.g., road type and light condition) in terms of accident frequencies. This could also be expressed as examining accidents within road type to see if the frequencies differ by light condition (i.e., does the light condition present during accidents on Interstates differ from the light condition during accidents on U.S. and N.C. routes, Rural Paved routes, etc.?). Here, a contingency table is set up the same as in the previous two-variable case except that the entries in the cells are now the frequencies of accidents rather than the number of locations experiencing 0, 1, etc. accidents.

Example

1. Purpose: To test for an association between road class and light condition with respect to accidents.

2. Data:

Road Type	Light Condition				Total
	Daylight	Dusk	Dawn	Dark	
Interstate	276	9	18	139	442
US & NC	4,442	173	117	1,939	6,671
Rural Paved	2,950	144	74	1,954	5,122
Rural Unpaved	362	23	5	144	534
City Street	7,915	364	121	2,438	10,838
Total	15,945	713	335	6,614	23,607

3. Compute:

$$\chi^2 = \sum_{i=1}^5 \sum_{j=1}^4 \frac{(N_{ij} - M_{ij})^2}{M_{ij}} = 484.6$$

where, for example, for column 1, row 1,

$$N_{11} = 276$$

$$M_{11} = \frac{(442)(15,945)}{23,607} = 298.5$$

4. Conclusions: For $\alpha = .05$ with $(5-1)(4-1) = 12$ d.f., $\chi^2_c = 21.0$. Since the calculated χ^2 exceeds 21.0, it appears that there is a strong association between road class and light condition with respect to accidents.

Note: As the calculated Chi-square statistic is affected by sample size, the contingency coefficient is calculated:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} = .14$$

This is a rather small contingency coefficient and indicates, that although the χ^2 is very large, much of the apparent association is due to the size of the sample.

While the preceding procedures are most appropriate for cases in which at least one of the variables is nominal in nature, more appropriate tests exist if both categorical variables are ordinal. While Goodman's coefficient (G) of regular association (see Gibbons, 1976, p. 226) and Kendall's tau (τ) (see Siegel, 1956, p. 213-223) are appropriate, a procedure which is also appropriate and computationally less complex is Spearman's rho (ρ) (see Siegel, 1956, p.202).

Procedure and Statistic.

The observed values for each of the two variables being studied (e.g., number of accidents per site and pavement condition) are, in reality, pairs of values associated with the same accident location. That is, each accident frequency is associated with a specific pavement condition. Here, let y_i denote the accident frequency for location (i) and x_i denote the individual pavement grade for the same location (i). Thus, the resulting pairs of observed values can be denoted (y_i, x_i) .

Step 1. For each location calculate the ranks of both the accident frequency and the pavement condition. The ranks for both variables will extend from 1, for the lowest value of the variable, to rank = n for the highest rank in the sample. For example, the lowest number of accidents would be assigned the rank of 1. In accident research, however, there will usually be a series of ties in the lower ranks. For example, many of the locations will have experienced zero accidents in the preceding time period. Since zero accidents would normally receive the lowest possible rank, and since all of the zero observations should receive the same rank, the rank which is to be applied to each of these tied observations is the average rank for the set of ties. For example, if there are three observations that are tied and the three observations are the lowest observations, then the three would account for ranks 1, 2 and 3. The average of these ranks is equal to $(1+2+3)/3 = 2$. Thus, each of these three observations would be assigned a rank of 2. If there was one location with zero accidents and four locations at which one accident occurred, the zero location would be assigned rank 1 and the four locations which follow would be assigned the average of the ranks 2, 3, 4 and 5 (i.e., 3.5). Thus, observations 2, 3, 4 and 5, each of which had one accident, would each be assigned the rank 3.5. The same procedure for ranking ties would be used in determining ranks for tied observations in the x variable (e.g., pavement condition equals good, fair or poor).

Step 2. Following the ranking procedure for both the y variable and the x variable, calculate the mean rank for both variables. (These two means should be equal.)

Step 3. For each location, calculate the difference between the rank for y_i and the mean rank for all y's. The same would be done for the x variable.

Step 4. Obtain the product of these differences for each location.

Step 5. Apply the following statistic:

$$\rho = \frac{\sum_{i=1}^N (r_{y_i} - \bar{r}_y) (r_{x_i} - \bar{r}_x)}{\sqrt{\sum_{i=1}^N (r_{y_i} - \bar{r}_y)^2 \sum_{i=1}^N (r_{x_i} - \bar{r}_x)^2}}$$

Step 6. To test for significance of the calculated value of ρ where the total number of locations (i.e., the total sample size) is greater than 4 and less than or equal to 30, the calculated ρ would be compared to the value presented in the table in Siegel, 1956 (p. 284). If the sample size is greater than 30, the fact that the distribution approaches normality could be used to calculate a t-value in which

$$t = \rho \sqrt{\frac{N-2}{1-\rho^2}}$$

This calculated t would then be compared with the tabular value for the student's t distribution given in most statistics books, and with the degrees of freedom equal to the number of locations minus 2 (i.e., N-2).

Example

1. Purpose: To test for association between number of accidents and pavement condition at ten locations.

2. Data:

Location	Number of Accidents y_i	Pavement Condition x_i
1	0	good
2	1	good
3	7	poor
4	0	fair
5	0	good
6	1	fair
7	3	fair
8	2	poor
9	1	fair
10	0	good

3. Ranking:

	(1)	(2)	(3)	(4)	(5)	(6)	(7)=(6)(5)
Location	y	x	r_y Rank y_i	r_x Rank x_i	$(r_y - \bar{r}_y)$	$(r_x - \bar{r}_x)$	
1	0	good	2.5	8.5	-3.0	+3.0	-9.0
2	1	good	6	8.5	0.5	+3.0	+1.5
3	7	poor	10	1.5	4.5	-4.0	-18.0
4	0	fair	2.5	4.5	-3.0	-1.0	+3.0
5	0	good	2.5	8.5	-3.0	3.0	-9.0
6	1	fair	6	4.5	0.5	-1.0	-0.5
7	3	fair	9	4.5	3.5	-1.0	-3.5
8	2	poor	8	1.5	2.5	-4.0	-10.0
9	1	fair	6	4.5	0.5	-1.0	-0.5
10	0	good	2.5	8.5	-3.0	3.0	-9.0
			$\bar{r}_y = 5.5$	$\bar{r}_x = 5.5$			-55.0

4. Compute:

$$\rho = \frac{\sum_{i=1}^N (r_{y_i} - \bar{r}_y)(r_{x_i} - \bar{r}_x)}{\sqrt{\sum_{i=1}^N (r_{y_i} - \bar{r}_y)^2 \sum_{i=1}^N (r_{x_i} - \bar{r}_x)^2}}$$

$$= \frac{-55}{\sqrt{(75.5)(72)}} = -0.746$$

5. Conclusions: Using the table of critical values of ρ with $N = 10$, the association is significant at the $p < .01$ level. Thus, a strong relationship exists between accidents and pavement condition.

4.3.3 Relative Weight or Strengths of Relationships--Model Development

Having now covered some of the statistical procedures used in the preliminary variable screening step described in section 4.3.1, let us now turn to the second, more pertinent question -- defining the relative weights of the relationships between a series of independent variables and one dependent variable (often accidents). Thus, the statistical procedures to be described are those used in developing a specific model and testing the model to measure its predictive accuracy. As in the preceding sections, the techniques discussed will be categorized according to whether the variables of interest are continuous or categorical.

4.3.3a Model development when all variables are continuous.

The primary technique for building models and testing their strengths is regression analysis. Depending on the number of independent variables studied and the hypothesized underlying relationship, this analysis may take the form of either multiple linear regression, polynomial regression, or nonlinear regression.¹

¹The researcher with a more advanced knowledge of statistics should also note (for his own benefit) that this chapter utilizes a family of related techniques, namely analysis of variance (one-way, two-way, multi-way with fixed effects, random effects, or mixed effects), multivariate analysis of variance, analysis of covariance--virtually all of which (including regression) are special cases of the multivariate general linear model (MGLM). In most applications, there will be a single outcome or dependent variable and hence a univariate general linear model (GLM). That is to say, regression and analyses of variance and covariance applications, while usually covered in separate texts or sections of texts, are special cases of the general linear model.

In the most often used type of regression analysis, general linear regression, the researcher is attempting to estimate the coefficients of the following model using the available data.

$$\hat{y} = a + b_1x_1 + b_2x_2 + \dots (4.1)$$

where, as before,

\hat{y} = predicted measure of the safety related variable

x_1, x_2 = the (independent) variables which affect crashes

b_1, b_2, b_3 = coefficients associated with the independent variables which define the amount of resultant changes in the dependent variable related to a unit change in the specific independent variable (all else held constant)

(Polynomial and nonlinear analyses require a different basic model, but the procedures are similar.) In general, the procedure for estimating the coefficients (b_1, b_2 , etc.) is to use mathematical formulas to fit the straight line to the data which minimizes the absolute differences (in actuality to minimize the sum of the squared differences) between the predicted values of y and the actual observed values of y found in the data.

For example, in the most simplistic case where the model being developed has only one independent variable (e.g., ADT), the formulas which are used define a line which minimizes the deviations between predicted and observed accidents/MVM at various levels of ADT. (These deviations are depicted by the lengths of the dotted lines in Figure 4.6. The object is to minimize the sum of the squares of the lengths of these dotted lines.)

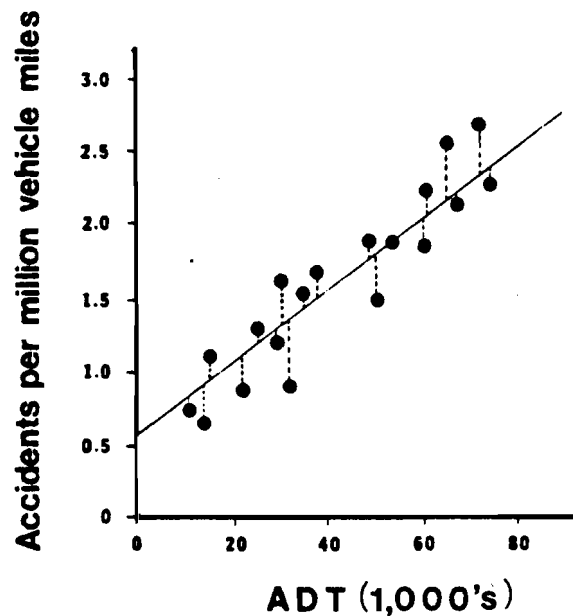


Figure 4.6. Regression line and deviations between predicted and observed accidents/MVM for different ADT values.

Source: Lundy, 1965.

The underlying procedure is the same when the model has more than one independent variable, but the plot is depicted in more than one dimension. The procedure for minimizing these deviations and thus calculating the b's is called the "least squares procedure." The actual mathematical formulas used in these procedures are presented in many texts (Snedecor & Cochran (1967), Draper & Smith (1966), Freund (1971)).

In conducting a general linear regression analysis, the researcher will carry out the following basic steps.

1. Determine the regression coefficients through use of the least squares procedures;
2. Test the statistical significance of each regression coefficient (a , b_1 , b_2 , etc.) to see if it should remain in the model; and
3. Determine the overall predictive accuracy of the final model to determine how well it predicts the dependent variable of interest.

Finally, if the model developed is to be used as a predictive tool, the model should be applied to a new set of data (whenever one can be accumulated) to determine its true predictive ability in the real world (i.e., the true unbiased estimate of R). Because of the inherent mathematical conditions, the estimate of predictive ability given when using the original test data used in model development will be larger than the true ability as determined by applying the final model to new samples of data. (This procedure is known as "cross-validation" of the model.)

Step 3 above, the initial determination of the predictive ability of the model, is an extension of the previously discussed procedure for determining the correlation (Pearson's r) for two variables (section 4.3.2a). In this extension, the multiple correlation coefficient (denoted as R) measures the strength of the overall relationship between the dependent variable and the linear combination of the independent variables (i.e., the relationship between accidents and the entire right hand side of the equation developed). By squaring the coefficient (i.e., calculating R^2), the researcher is able to determine the amount of the variation in the dependent variable accounted for by the model. For example, an $R^2 = .6$ would indicate that 60 percent of the variation in accidents is accounted for by the current model.

While this concept can be somewhat confusing, it can perhaps be best understood by noting that the model being developed is built from N different data sets where N is equal to the number of locations we are studying. At each location, there is an observed number of accidents and observed values for different "descriptor" variables (say ADT, pavement width, speed deviation, percent trucks, etc.). The number of accidents will differ from one location to the other in some but not all cases. For example, as noted earlier, many locations will have experienced zero accidents. The descriptor variables will also differ between locations, but only some cases. Even in locations where the accidents are equal, the descriptors may differ. In like manner, there will be cases in which the descriptors are equal (or nearly equal) at two or more locations, but the accident frequencies will differ. Thus, there is a built-in variation in accidents in the total sample of N locations partly due to the changes in the identified factors, but partly due to some other "causes." (If this were not the case, every location with the same descriptors, or predictor variables, would experience equal numbers of accidents.) No matter how many independent predictor variables we identify to include in our model, there will always be other unidentified factors which affect accidents. Thus, there will always be inherent variation in the accidents in our sample. The R^2 measures how much of this variation the developed model accounts for. The goal of any model development procedure is to maximize the predictability (i.e., maximize R^2) with as few independent variables as possible.

CAUTION: The user of regression equations must remember one important restriction. When used in a prediction sense (the normal case), the model must not be used to predict values when the input values for the independent variables are outside the range of the corresponding independent variable values used in the development procedure. For example, if the values of ADT ranged from 100 to 10,000 vpd in the development phase, it would be improper to use the model to predict accident frequencies for a case when ADT is equal to 50 or to 30,000. It is improper to extrapolate outside the initial ranges.

For the sake of brevity and clarity, let us now turn to a simple example. Assume that the researcher has applied the previously discussed screening process to all of the continuous variables at his disposal and has decided that the only two variables which have shown a strong association with accident frequency

are speed deviations and ADT. This same situation might arise if these were the only two variables available (or of interest) to the researcher.

Example

1. Purpose: To establish a linear function to predict number of accidents from speed deviation and ADT using least squares procedures.
2. Data: See Example data in Section 4.3.2a.
3. Compute: The linear model whose parameters are to be estimated is of the form:

$$\hat{y} = a + b_1x_1 + b_2x_2$$

where

\hat{y} = predicted number of accidents

x_1 = speed deviations

x_2 = ADT in ten thousand vehicles

a, b_1, b_2 are regression coefficients.

Using the least squares procedure, estimates for b_1 and b_2 are given by the following formulas:

$$b_1 = \frac{(\sum_i (x_{2i}')^2)(\sum_i x_{1i}'y_i') - (\sum_i x_{1i}'x_{2i}')(\sum_i x_{2i}'y_i')}{D}$$

$$b_2 = \frac{(\sum_i (x_{1i}')^2)(\sum_i x_{2i}'y_i') - (\sum_i x_{2i}'x_{1i}')(\sum_i x_{1i}'y_i')}{D}$$

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

where

$$D = (\sum_i (x_{1i}')^2)(\sum_i (x_{2i}')^2) - (\sum_i x_{1i}'x_{2i}')^2$$

$$y_i' = y_i - \bar{y}$$

$$x_{1i}' = x_{1i} - \bar{x}_1$$

$$x_{2i}' = x_{2i} - \bar{x}_2$$

Here, using these formulas

$$b_1 = 0.821$$

$$b_2 = 0.955$$

$$a = -5.630$$

Thus the model developed is

$$\hat{y} = -5.630 + 0.821x_1 + 0.955x_2$$

4. Testing the coefficients: To test for the significance of b_1 and b_2 , it is assumed that $\frac{b_1 - \beta_1}{s_{b_1}}$ is distributed as t with $(N-k)$ degrees of freedom

where s_{b_1}

s_{b_1} = standard error of b_1

$$= \sqrt{\frac{\left(\sum_{i=1}^{15} (y_i - \hat{y}_i)^2\right) \left(\sum_{i=1}^{15} (x_{2i}')^2\right)}{N - k \quad D}}$$

$$= 0.355$$

since

$N = 15$ = number of observations

$k = 3$ = number of coefficients in the model

$$x'_{2i} = x_{2i} - \bar{x}_2$$

D defined previously.

Similarly

$$s_{b_2} = 0.451$$

To test, for example, whether b_1 is significantly different from 0 and thus to test whether it should be retained in the model:

$$t_1 = \frac{b_1 - 0}{s_{b_1}} = \frac{0.821}{0.355} = 2.31$$

Since $t = 2.31 > 2.131 = t_{13, .05}$ (2-sided), we conclude that $\beta_1 \neq 0$. In similar manner the test of b_2 results in $t = 2.14$, again indicating that the variable should be retained.

5. Compute R^2 for the model:

$$R^2 = \frac{\left[\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right]^2}{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_i - \bar{\hat{y}})^2}$$

where

y_i = observed number of accidents at i-th location

\bar{y} = overall mean number of accidents at all locations

\hat{y}_i = predicted number of accidents at i-th location

$\hat{\bar{y}}$ = overall mean number of predicted accidents

In this example, for the final model

$$R^2 = 0.750$$

Conclusion: Both coefficients are significant, and thus both ADT and speed deviation should be retained in the model. The overall model accounts for 75 percent of the total variation in accidents at the different locations.

The above example, used to depict the key points of regression analysis, was rather simple in nature. In most situations in which accident researchers will be using these procedures, the situation will be more complex. First there will often be large samples of data with many possible predictor variables. In these cases, the researcher will usually require access to a computerized analysis system since, first, manual calculations could be quite cumbersome, and second, the initial screening procedure for numerous potential predictor variables would be quite tedious. There exist techniques which can be carried out both manually and with a computer which help determine which of many available continuous variables should be included in the "best" regression equation and used to build the model, both in one step. These techniques include (1) all possible regressions; (2) backwards elimination; (3) forward selection; (4) stepwise regression; (5) variations of the previous method, and (6) stagewise regression. Because this manual is not intended to be a statistics text, the details of these techniques will not be presented. However, the reader is referred to Draper and Smith (1966, pp. 163-167) for a description of the pro's and con's of each of these techniques. Again it is noted that while many of these procedures have been computerized and are thus suitable for large data sets, some statistically oriented accident researchers continue to emphasize that each of these techniques has its inherent disadvantages. For the researcher with considerable training and experience with regression analysis, the most preferable technique may continue to be the development of a correlation matrix in which the strength of each individual relationship is examined and a judgment is made by the researcher concerning which variable should be included in the initial model.

The regression procedure is made slightly more complex in certain situations where the best linear model developed will need to include interaction terms -- cases in which two of the independent variables are related to one another. For example, it may be the case that the speed deviation will differ depending on the ADT level, requiring the use of interactive terms on the predictor side of the model. Discussion of the terms and of the related least squares procedure for determining the coefficients is included in Draper and Smith (1966).

Again, however, although appearing more complicated, the basic procedure continues to be the same. The interaction term is but another independent predictor variable which has a separate coefficient which must be determined and tested. The problem with including interactive terms in a predictive model is that they are often quite difficult to interpret. Thus, the researcher should attempt to build a meaningful model without including them if possible.

Finally, there will also be the case when a simple linear model is not sufficient. The underlying relationship may not be depicted by a straight line, but instead may require a curvilinear function. In these cases, the researcher should rely on polynomial regression or nonlinear regression. Both of these types of regression are important as it is indicated by the fact that the underlying relationship between, for example, ADT and accidents is not linear through the entire range of ADT (see Figure 4.7). However, the scope of this manual will not allow detailed discussion of these two techniques. Instead the reader is referred to Snedecor & Cochran (1967, p. 453). It is also noted that in certain cases an alternative to

polynomial or nonlinear regression is to transform the predictor variables into some arithmetic or logarithmic function before using simple linear regression techniques. See Draper & Smith (1966, p. 131).

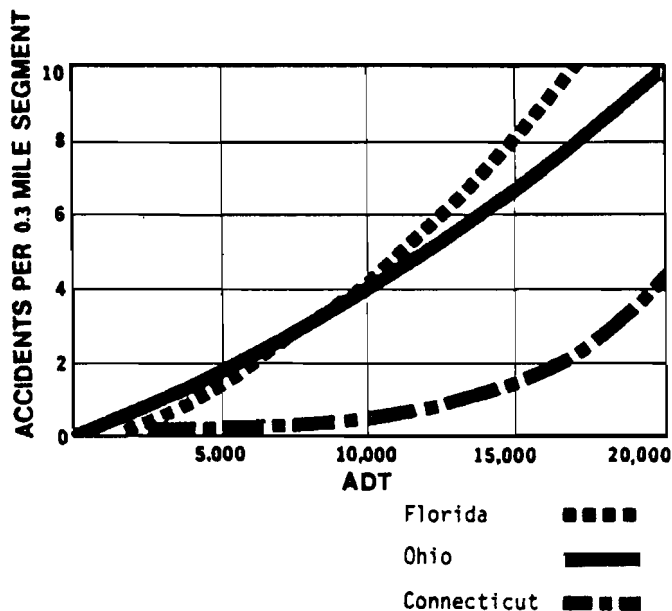


Figure 4.7. Accidents and ADT Four Lane Divided, No Access Control.
Source: Kihlberg & Tharp, 1968, pp. 65, 70, 74.

In summary, while this section could not attempt to provide the detailed discussion of regression analysis presented in the statistical texts referenced, a limited listing of key points which accident researchers should be familiar with follows:

1. Because of the non-normal nature of accident frequencies and because of the fact that accidents are the consequence of a multi-causal system, the resulting predictive nature of the developed models will in many cases be low (as depicted by relatively low R^2 's). The researcher using the results of such models in making future program decisions must do so with care. It is strongly advocated that when underlying relationships are identified, they be further tested using the experimental techniques as described in this chapter before they are used as major tools in decision-making.
2. The researcher should always cross-validate his model on a new sample of data before making a final judgment concerning its real world level of predictability.
3. When inferring casual relationships from a regression analysis (which one is always attempting to do) the researcher must carefully study the direction of the casual chain. One always assumes that the independent or predictor variables "cause" the dependent variable. However, there may be cases when this is not true. A recent example of such a situation occurred in a driver-related study in which regression analysis was used to analyze data from various states related to the number of young drivers taking driver education and the subsequent number of young drivers being licensed and becoming involved in accidents (Robertson & Zador, 1978). The authors concluded that the offering of driver education courses caused more young drivers to be licensed than would normally be the case and thus caused more accidents to occur. As was pointed out by critics of this study (Seaver, et al., in press), in this case the causal chain could have very easily been in the opposite direction. Rather than driver education "causing" young drivers to be licensed, the demand for licensing among the young drivers could have "caused" a state to offer driver education. While such a "reversed" casual chain is less likely to be found in highway-related studies, the researcher must always be aware that such a relationship could be confusing the issue.

4.3.3b Model development when all variables are not continuous.

Having now covered regression analysis, the most appropriate technique when the variables under consideration are continuous in nature, let us now turn to the case where the variables in question are categorical in nature. (Note that continuous variables are often subdivided into categories.) While the specific statistical procedures used in the examination of underlying relationships for categorical data differ somewhat from general linear regression, the underlying procedure is basically the same. Again, an attempt is being made to build a model by determining coefficients associated with important independent variables, to test the significance of the coefficients, and to determine the actual level of predictiveness of the overall model.

Unlike the continuous variable case in which computerized procedures are available to both screen variables and build the model in a single step, model building for categorical data remains a two-step process, with Step 1 being variable selection for inclusion, and Step 2 being the actual development of the model itself.

The reader should note that the model development procedures for categorical data are neither as simple nor as familiar as are the corresponding regression procedures for continuous data. This results from the fact that these procedures are relatively new. However, even though they are quite new, somewhat complex, and unfamiliar to some statisticians in the field of accident research, they are very important techniques since much of the data with which the highway accident researcher must work is categorical in nature. Although there are techniques for including categorical data in regression models, these new categorical techniques produce models which are stronger and more meaningful than are models developed by using less appropriate regression techniques. Because these models are more appropriate in the accident research field, an overview is presented here. Due to the complexity of the procedures, details will not be presented. However, references, including references to existing computerized packages, are provided, and the researcher is urged to contact a knowledgeable statistician in determining first, whether or not to use these procedures, and second, how to actually use them.

Step 1. Variable selection procedures. In selecting the variables to be included in the model, one option would be to use the screening techniques described earlier for categorical data (i.e., the simple χ^2 and the Spearman's rho). However, there are more appropriate selection techniques which have been developed. These include first, a computerized CHAID program (Kass, 1975), and second, a non-computerized procedure involving what is known as hierarchical Chi-square screening.

CHAID (Chi-square Automatic Interaction Detection) is a computerized program which determines relationships through a branching process. Use of the program has indicated that it is heavily dependent upon sample sizes. Thus, important variables which have smaller sample sizes are less likely to be selected for the final model. The principal advantage of the CHAID procedure lies in its computerization.

The hierarchical Chi-square screening procedure, considered more appropriate by some statisticians, involves a step-by-step procedure in which the variables to be included are chosen based on their relationship with the dependent variable (e.g., accidents). (For the advanced statistician, the selection algorithm used proceeds in the same spirit as the algorithm used in forward stepwise regression analysis.) While the reader interested in the exact procedure should refer to Higgins and Koch (1977), this basic screening procedure involves the following steps.

1. The initial independent variable to be included is the single independent variable of all those available which has the strongest relationship with the dependent variable (i.e., accidents). The strength of the relationship is determined by the Pearson Chi-square statistic divided by the degrees of freedom. The first independent variable selected is the one having the largest Chi-square per degree of freedom with respect to accidents.
2. The second independent variable chosen is not simply the independent variable that has the second highest Chi-square per degree of freedom as related to accidents. Instead, it is the variable which has the highest Chi-square per degree of freedom as related to accidents within the categories of the first independent variable chosen. Thus, after the first variable is chosen, all remaining variables are individually placed into tables in which one dimension is accidents and the second dimension is all levels of the combinations of the first variable with levels of the second variable.

The remaining variables follow the same procedure. Because this procedure is somewhat complex, neither an example nor further details are presented in the manual. Again, the reader must note that while both the CHAID and this hierarchical Chi-square procedure are somewhat complex, the basic underlying goal is to reduce the number of possible independent variables to a usable list which includes the most relevant variables to be used in the model. Let us now turn to the model fitting step.

Step 2. Model fitting. Just as with continuous variables, the procedure used for developing the model for categorical variables is analogous to the regression procedures followed earlier. Perhaps the most appropriate procedure for the categorical data situation is weighted least squares regression. The data may be nominal, ordinal, and/or continuous with the latter types of data grouped into categories.

Just as with regression, the underlying model to be developed is of the form

$$F = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots$$

In this case, however, rather than the dependent variable being a continuous variable usually related to accidents (e.g., accident rates), the dependent variable is either the proportion of total accidents at a given location (multiplied by a constant) or the log of the proportion, depending on the nature of the data. The choice is made by a qualified statistician.

Again, similar to regression analysis, the coefficients are calculated by a (weighted) least squares procedure. The procedure is computerized in a program called GENCAT (see computer program references) for ease of handling. While no details will be provided here, the reader is referred to a paper by Freeman, et al. (1975) for details of a rather complex example of accident research. An additional paper by Grizzle, et al. (1969) provides other less complex examples.

4.4. Summary

Chapter 4 has provided an overview of methodologies involved in research aimed at identifying and quantifying underlying relationships between accidents and other factors. The procedure is basically a two-step process.

1. Screening potential predictor variables to select those strongly related to the variable of interest (e.g., accidents).
2. Developing models in which the relationships are quantified and tested.

While a limited number of basic statistical procedures have been presented, and others have been referenced, the choice of procedure is always dependent on:

1. The nature of the question (screening or model development).
2. The nature of the data (continuous or categorical).

4.5 Review Questions

1. How is a representative sample of a population selected?
2. A state traffic engineer is requested by the FHWA to collect accident and highway characteristics data on a sample of sections of Interstate highway. Because the purpose of the study is to predict accident rates based on highway characteristics, the engineer samples those locations which have experienced one or more accidents in the past year. Comment briefly on the adequacy of this sample.
3. Rather than using a table of random numbers, the highway researcher may employ other sampling techniques. What are they?
4. Based on prior research, ADT is known to affect accidents in that accident rates vary greatly for locations experiencing ADT's of less than 100 vpd, between 100 and 2000 vpd, and greater than 2000 vpd. The researcher is attempting to draw a 10% sample from all the locations in his jurisdiction to develop a model predicting accidents. How might he draw such a sample?

5. What are some potential biases that occur when using vehicle damage as a measure of severity?
6. What are the two attributes that a proxy measure must have to be acceptable?
7. A researcher is interested in developing a relationship between some measure of safety and the feet of guardrail per mile, the number of breakaway and non-breakaway telephone poles per mile, and the number of protected bridge piers per mile. What would be an appropriate dependent (predicted) variable to be used in the model?
8. The choice of the most appropriate statistical procedure for use in building models is primarily based on one factor. What is it?
9. What would be the most appropriate statistic for examining the association between weather condition and injury severity?
10. In what context should hierarchical Chi-square screening be utilized and basically what does it accomplish?
11. A researcher is interested in ascertaining the relationship between variables which may not be linearly related. What type of analysis should she employ?
12. The researcher is attempting to determine whether there is an association between accident frequency and intersection pavement condition (i.e., poor, fair, good). Which statistical procedure would be appropriate for use?

4.6 References

- Clyde, M. N. Michigan study indicates signals increase accidents. Traffic Engineering, 1964, 35(2), 32, ff.
- Cochran, W. G. Sampling techniques. (3rd ed.) New York: John Wiley & Sons, 1977.
- Conner, R. E. Traffic signals and accidents on rural state highways in Ohio. Columbus: Ohio Department of Highways, 1960.
- Cribbins, P. D., Arey, J. M., & Donaldson, J. K. Effects of selected roadway and operational characteristics on accidents on multilane highways. Highway Research Record, No. 188, 1967, pp. 140-157.
- Crosstown Associates. Correlation of accident rates and highway geometric configurations. Preliminary draft, 1968.
Cited in Peter A. Mayer, Ed., Traffic control & roadway elements--Their relationship to highway safety. (rev. ed.) Chapter 12. Alignment. Washington, D.C.: Highway Users Federation for Safety and Mobility, 1971.
- Deming, W. E. Sample design in business research. New York: John Wiley & Sons, 1963.
- Draper, N., & Smith, H. Applied regression analysis. New York: John Wiley & Sons, 1966.
- Fee, J. A., Beatty, R. L., Dietz, S. K., Kaufman, S. F., & Yates, J. G. Interstate system accident research study. (3 Vols.) Washington, D.C.: U.S. Government Printing Office, 1970.
- Freeman, J. L., Koch, G. G., Hunter, W. W., & Lacey, J. H. Shoulder harness usage in the population of drivers at risk in North Carolina. Chapel Hill: University of North Carolina Highway Safety Research Center, 1975.
- Freund, J. E. Mathematical statistics. (2nd ed.) Englewood Cliffs, NJ: Prentice-Hall, 1971.

- Gibbons, J. D. Nonparametric methods for quantitative analysis. New York: Holt, Rinehart, and Winston, 1976.
- Grizzle, J. E., Starmer, C. F., & Koch, G. G. Analysis of categorical data by linear models. Biometrics, 1969, 25(3), 489-504.
- Head, J. A. Predicting traffic accidents from roadway elements on urban extensions of state highways. Highway Safety Research Board Bulletin, No. 208, 1959, pp. 45-63.
- Higgins, J. E., & Koch, G. G. Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey. International Statistical Review, 1977, 45, 51-62.
- Kass, G. V. Significance testing in, and some extensions of, automatic interaction detection. Unpublished doctoral dissertation, University of Witwatersrand, South Africa, 1975.
- Kihlberg, J. K., & Tharp, K. J. Accident rates as related to design elements of rural highways. National Cooperative Highway Research Program Report, No. 47, 1968.
- Kleinbaum, D. G., & Kupper, L. L. Applied regression analysis and other multi-variable methods. North Scituate, Massachusetts: Duxbury Press, 1978.
- Klimko, L. A., & Friedman, K. Statistical analysis of crash conditions and their relationship to injuries. Final report. Washington, D.C.: National Highway Traffic Safety Administration, in press.
- Knoblauch, R. L. Urban pedestrian accident countemeasures experimental evaluation. 2 Vols. Falls Church, Virginia: BioTechnology, Inc., 1975.
- Leisch, J. E., Pfefer, R. C., & Moran, P. J. Effect of control devices on traffic operations. National Cooperative Highway Research Program Report, No. 41, 1967.
- Lohman, L. S., Campbell, B. J., Leggett, E. C., & Stewart, J. R. Identification of unsafe driving actions and related countermeasures. Chapel Hill: University of North Carolina Highway Safety Research Center, 1976.
- Lundy, R. A. Effect of traffic volume and number of lanes on freeway accident rate. Highway Research Board Record, No. 99, 1965.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. SPSS: Statistical Package for the Social Sciences. (2nd ed.) New York: McGraw-Hill, 1970.
- Noether, G. E. Elements of non-parametric statistics. New York: John Wiley & Sons, 1967.
- Quade, D. Nonparametric partial correlation. In M. M. Blalock (Ed.), Measurements in the Social Sciences. Chicago: Aldine Publishing Company, 1974.
- Raff, M. S. Interstate highway-accident study. Highway Research Board Bulletin, No. 74, 1953, pp. 18-45.
- Research Triangle Institute. Speed and accidents. Vol. 1. Research Triangle Park, North Carolina: Author, 1970.
- Robertson, L. S., & Zador, P. L. Driver education and fatal crash involvement of teenaged drivers. American Journal of Public Health, 1978, 68(10), 959-965.
- Seaver, W. B., Nichols, J. L., Carlson, W. L., & Voas, R.B. Driver education and the licensing of 16 and 17 year olds. Washington, D.C.: National Highway Traffic Safety Administration, in press.
- Siegel, S. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, 1956.

Slatterly, G. T., Jr., & Cleveland, D. E. Traffic volume. Chapter 2 in Traffic control and roadway elements--Their relationship to highway safety. (Rev. ed.) Washington, D. C.: Automotive Safety Foundation, 1969.

Snedecor, G. W., & Cochran, W. G. Statistical methods. (6th ed.) Ames: Iowa State University Press, 1967.

Snyder, M. B. Traffic engineering for pedestrians' safety: Some new data and solutions. Highway Research Record, No. 406, 1972, pp. 21-27.

Solomon, D. Traffic signals and accidents in Michigan. Public Roads, 1959, 30 (10), 234-237.

Solomon, D. Accidents on main rural highways related to speed, driver, and vehicle. Washington, D.C.: U.S. Government Printing Office, 1964.

Stewart, J. R., & Stutts, J. C. A categorical analysis of the relationship between vehicle weight and driver injury in automobile accidents. Chapel Hill: University of North Carolina Highway Safety Research Center, 1978.

U.S. Department of Commerce, Bureau of Public Roads, Traffic Operations Division. Interstate system accident research: Revised procedure manual. Washington, D.C.: Author, 1961.

Vey, A. H. Effect of signalization on motor vehicle accident experience. In Proceedings of the Institute of Traffic Engineers, 1933.

4.7. Computer Programs

CHAID

Kass, G. V. Significance testing in, and some extensions of, automatic interaction detection. Unpublished doctoral dissertation, University of Witwatersand, South Africa, 1975.

CONTAB

Gokhale, and Kullback, S. The information in contingency tables, New York: Marcel Dekker, 1978.

ECTA

Goodman, L., & Fay, R. A computer program for fitting log-linear models to contingency tables. Chicago: University of Chicago Department of Statistics, 1974.

GENCAT

Landis, J., Stannish, W., Freeman, J., & Koch, G. G. A computer program for the generalized chi-square analysis of categorical data using weighted least squares. (Rev. ed.) Chapel Hill: University of North Carolina Department of Biostatistics, 1978. [Biostatistics Report No. 8].

CHAPTER V

THE FINAL STEP: PREPARATION & DISTRIBUTION OF RESEARCH RESULTS

Situation: The State Highway Department, in an effort to upgrade its research unit, hires a recent non-engineering Ph.D. graduate from the state university statistics department and assigns him the job of assessing the new pavement grooving treatment which has been applied at a number of pilot locations. The researcher draws all the data together, carries out an after-the-fact study (the only type possible), and issues a 300 page report which describes the study and its results. The final paragraph in the Executive Summary reads as follows:

"In summary, examination of the independent variables indicated a lack of homogeneity and normality in the data. Because of this a logarithmic variance stabilizing transformation of the categorical data was carried out. The data was first examined in hierarchical chi-square analyses followed by model development using maximum likelihood estimates. In addition, an analysis of covariance was conducted. Analysis of the null hypothesis indicated no significant effect on the dependent variable at the .01 alpha level. A significant change was indicated at the .05 level but only when annual rainfall was controlled for as a covariant."

Result: Knowing that even with all his engineering expertise he cannot control annual rainfall, the administrator files the report (and the treatment) in his circular file.

Main Chapter Topics

Introduction
Preparation of Reports
Distribution of Results

5.1 Introduction

The preceding chapters of the manual have provided information on why and how to conduct accident research. A researcher with a firm grasp of that material should now be in the position to plan and carry out a meaningful accident research project. However, even with careful planning, such research can benefit a given jurisdiction only if the following conditions exist:

1. The researcher reports the research results to the administrator.
2. The administrator understands the significance of the research results.
3. The administrator acts, based upon the research results.

The action (or inaction) of the decision-maker occurs regardless of the research results, but the initial two steps must take place if the research is even to be considered. Unfortunately, the proper reporting and distributing of research results, a seemingly simple task, is often not carried out adequately, perhaps because researchers believe that the most important phases of the sequence are the actual planning and implementation of the research project. Getting the research done is their job.

However, research is only one of many steps in the complex decision-making process. Because decisions are made whether or not there is well-conducted accident research available, the researcher is not completely doing his job unless the results of his research are taken into account. In this respect, it is not the implementation of the research but rather the presentation of usable research results to the user--the administrator--which is the most important step.

Thus, the researcher must first do an acceptable job of conducting the research, and then also do an acceptable job of communicating the results to potential users. One important element in this communication process is the graphic presentation of information. Report authors should remember that groups of numbers are more accessible to readers if they are presented in tables instead of embedded in text, and that quantitative relationships are more evident if they are depicted in figures instead of described in words. (Winfrey (1962) thoroughly covers this and other aspects of technical report writing.) It is also important that the report be organized according to recognized conventions so that readers can know where to find the various components that they expect the report to contain. (The conventions described in the Department of Transportation's (1975) manual are required for technical reports prepared for DOT and its member agencies, but they are a good set of standards to follow for other audiences as well.)

It is also important that the language of the research report be appropriate to the intended audience. The situation described at the beginning of the chapter, although perhaps somewhat exaggerated, illustrates the need for the researcher to report his results clearly and in terms that are understandable to his readers. Unfortunately, there is an inherent conflict between the realm of researchers and that of administrators: administrators try to operate by assertion (e.g., this program is having an effect), but researchers and statisticians operate by negation (e.g., there is no indication that this program is not having an effect). The reason for this difference is a fundamental principle of statistics which holds that it is not possible to actually prove the existence of a real difference (the alternate hypothesis). Instead, it is necessary to hypothesize that there is no difference (the null hypothesis) and then attempt to reject this hypothesis by determining the probability that it is correct. If the probability that it is correct can be shown to be small enough, then the null hypothesis can be rejected with a certain level of confidence. Because the statistician's basic philosophy is a reflection of this principle, the interpretation of research results in a meaningful way is indeed a new experience to many researchers, especially those who have no engineering background and who have little experience working with decision makers in the highway area.

Second, even when the results are presented clearly, they must not be allowed to die on the shelf, or, as is more common, to stop moving either vertically or horizontally in the existing information channels. In an earlier study (Council and Hunter, 1975) involving interviews with state traffic engineering and highway design personnel concerning problems with research, a problem cited frequently by the working professionals involved their failure to receive available information related to projects and techniques tested in other states or reported by research organizations. While at times this information flow "short circuit" was due to internal problems in the vertical information channels (e.g., the technical report stopping on an upper level administrator's desk rather than being passed down to subordinates who could actually use the information in their work), quite often the nature of the problem resulted from the fact that, while the state experimented with modified designs and studied their effectiveness, such findings were never published for use by other states.

In particular, the head of the Highway Department in a state which is one of the more innovative in terms of highway research produced several internal technical reports concerning appurtenance design and testing as examples of the research work conducted in-house. He also noted that none of these reports was being sent to other states within his region nor were they being presented at national meetings. A second agency head in another state indicated that he received very little information from neighboring states concerning their research work except by word of mouth at professional meetings, and often, when he did receive such information, it was because he was duplicating a research project that had been conducted earlier.

Research information first needs to be prepared so that it is usable to both the decision maker and to other researchers in the field, and then it needs to be disseminated as widely as possible. The specific steps for properly preparing and disseminating research reports are discussed in more detail below.

5.2 Preparation of Reports

Research reports may be prepared in a number of formats, but the researcher must realize that the manner in which information is presented affects whether or not it is properly interpreted by other researchers and by decision makers, and, ultimately, whether it is used. The researcher may be intimately aware of the problems and biases that can affect the interpretation of results, but unless he specifies them in the report, the reader will not be aware of them.

This chapter does not attempt to discuss all the details that are important in the preparation of a well-written research report. The reader is referred to Winfrey (1962) for a comprehensive treatment of report preparation details ranging from style to reproduction methods. However, there are two basic keys which appear to make some reports more useful than others.

5.2.1 Report preparation keys.

The first key is the knowledge of the audience the report is written for--whether the primary users of a report will be other researchers in the field or decision makers. Quite often this decision is based on both the original rationale for the research (i.e., whether the research was to evaluate a given countermeasure or program activity or whether it was to add knowledge to the state-of-the-art by examining underlying relationships) and on the actual results of the research itself. For example, if a study that was originally planned to evaluate a given countermeasure failed to actually provide a conclusive evaluation but did identify new relationships that had not been previously noted in the field, the primary user would properly be other researchers instead of decision makers. The writing style would therefore shift from administrative jargon to a more technical research language. Most of the time, however, the user of the research results will be the decision maker. Therefore, the research report needs to be written so that it is usable by the administrators who must incorporate the findings in their later decisions as well as by other researchers who will be building future research based on this information.

A second key to the preparation of a well-designed report appears to be an emphasis on the interpretation of results. While most statisticians are trained to hedge to some extent when explaining results (i.e., "failure to reject the null hypothesis" rather than "prove there is a real difference" is stressed in all statistical courses), it is necessary for the researcher to interpret his findings in real-world terms if the user is to understand and subsequently incorporate them into his decisions. It is not enough for the researcher to present his findings in statistical terms and leave it up to the administrator to decide what the results actually mean.

This more traditional manner of reporting scientific research implicitly assumes that the reader--the decision maker--is very familiar with the statistical techniques used in the research, and thus has a strong basis for deciding whether the results are sound and how they can be implemented in his program. This assumption is not valid. Indeed, it is not even the responsibility of the administrator to acquire such knowledge. Instead, it is the responsibility of the researcher to present his results so that they can be used by the reader, just as it is the responsibility of the traffic engineer to present his marking and signing configurations so that they are understood by the principal user, the driver. While this means that the researcher is often forced to extrapolate or infer from his results further than he might in the context of pure research, this appears to be a definite necessity if research results are to be used in the decision making process. Otherwise, the results of even very good research are often lost. Guidelines concerning better methods for presenting and interpreting statistical results can be found in Turkey (1977).

After establishing the report's target audience (and the level of interpretation that is necessary), the researcher can begin the actual preparation of the report itself.

5.2.2 Suggested report preparation sequence.

As noted by Winfrey (1962), many report preparation sequences and report formats can be used to adequately provide information in a usable form. The following sequence is commonly followed:

Step 1. Prepare outline. The first report preparation step, which takes place after the statistical analysis has been completed, is preparing a detailed outline for the research report. The purpose of the outline is to help the researcher prepare to present his information in a coherent fashion. The outline in Figure 5.1 presents the type of information that is commonly included in each section of a conventional eight-part report format.

By preparing an outline, the researcher can specify what will be included in his complete report. Often, preparing a detailed outline can be the most difficult part of the entire report preparation sequence, but it can also be the most important in terms of ultimate report usefulness.

Step 2. Prepare initial draft. Following preparation of the outline, the researcher prepares the initial draft of the report, which should contain the abstract, the executive summary, the interpretation of results, and the final conclusions and recommendations. (The initial draft usually does not contain such details as table of contents or appendices.) The author should not consider the initial draft inviolate because he will need to make numerous changes in the text as it is scrutinized during the review process.

Figure 5.1 Suggested report outline.

I. Abstract

The initial report segment will usually be brief (one to two pages) and will provide a description of the project including the goal or goals, the methods employed, problems found, and a brief summary of the major results and recommendations. It is important to remember in preparation of such an abstract that this is often the only part of the report read by decision-makers.

II. Executive Summary

The abstract is often expanded in the initial section of the text of a report through use of an "Executive Summary," a longer (5-10 page) expansion of the same material. While traditionally found only in longer, more technical reports, this summary is now being required more often by contractors as part of the product they are funding. This is based on the knowledge that while a decision-maker may not have the desire nor the time to review an entire report, he will quite often review a 5-10 page summary.

III. Introduction

The initial part of the detailed text narrative should introduce the reader to the problem or area of need. It will provide narrative concerning why the study is being done, what information past research has provided concerning the issue in question (either in this Introduction or in a separate Review of the Literature section), and a preview of the remainder of the report.

IV. Methodology

The methodology section should contain a description and discussion of the data collection procedures, the study design, and the statistical analysis procedures. It is in this part of the paper that the researcher is primarily providing information for other researchers to aid in their assessment of the soundness of the results. This section is usually quite detailed.

V. Results

This section presents the results of the analyses and will in most cases include tables of data, statistical test results, and limited discussion of the results from such tests.

VI. Discussion and Recommendations

The discussion section is the part of the paper which should provide the interpretation of the results. In this respect it is the most important section to be written. It is in this part that the researcher is able to take the results from the statistical tests and interpret them so that they are meaningful to the report user. In addition, this section should discuss the implementation of these results--what do the results mean to your program and to programs in other states? Do the results suggest other evaluations which are needed? What limitations existed in this study, either in the data or the research design? What conclusions can be drawn?

Finally, based on the findings, a list of recommendations related to both the researcher's specific program (i.e., recommendations for one's own state or jurisdiction), and recommendations related to national areas of concern should be included.

VII. References

Detailed citations of any other studies used in the research report or other studies which might be useful to other researchers in the area carrying on related research should be listed. It is suggested that a standard reference format be used.

VIII. Appendices

Most research reports will document the data used by including a series of appendices containing such information as data formats, definitions used, special analyses formulae, questionnaires or data collection forms used, and, in some cases, tables of raw data. This display of raw data has not traditionally been done. However, it is now quite often requested by other researchers and contractors. The more raw data that are presented, the more additional analyses other researchers can conduct. In many FHWA research efforts, these raw data are included in separate volumes. Indeed, in most current contract research, the researcher will be required to retain the data base or to present it to the contractor for future use.

Step 3. Review of initial draft by colleagues. This step is perhaps the one which is carried out least often by researchers. It can be a very important step if the proper persons conduct the review. The researcher is so intimately involved in the actual implementation of the research that he may not be able to anticipate problems which can arise for readers who do not share his familiarity with the research. Consequently, he may fail to include information that is necessary for the reader to completely understand the research effort. Careful review by others before distribution will help specify areas where modification or expansion is needed.

If possible, the report should be reviewed first by another analyst or researcher who can assess the methodology used and the interpretation of results. Second, an engineer (not necessarily an analyst) should review the report for style, clarity in writing, and interpretation of results (i.e., Are the results usable by the decision makers in the field?). Finally, if possible, the report should be reviewed by a non-engineer, non-analyst editor whose primary purpose is to make necessary editorial changes and also provide important inputs concerning the ability of a non-analyst, non-engineer (similar to many top level decision makers) to understand the results as presented.

Step 4. Revision of the draft. Based on the inputs from the review of colleagues, the author should revise the draft. He should then check with the reviewers to make sure that the revisions are adequate. Again, the use of well-designed tables, figures, and illustrative photographs presents important material better than large amounts of text, and makes the report more interesting to the reader.

Step 5. Review by sponsor/user. In most research studies funded by a sponsoring agency such as FHWA, the researcher will be required to provide copies of the report to the sponsor for review before distribution. Although most sponsoring agencies designate a liaison person (Contract Technical Manager) to monitor ongoing project implementation, they will also review the reports to assure that the desired objectives of the contract are met and that the results are presented in a manner suitable for use by the sponsors and other concerned parties.

Step 6. Final revision of the report. Finally, following review by the sponsor, the researcher should incorporate suggested revisions he considers appropriate and make the final changes in the report.

The report preparation sequence described above is but one example of a number of similar sequences which should produce a usable research report. Regardless of what sequence is followed, preparation of a detailed outline and reviews of both technical content and clarity of presentation are strongly recommended.

5.3 Distribution of Results

The final major step in the dissemination of usable research information is the distribution of the well-written report. The distribution avenues which exist, and thus the actual degree of dissemination, are often determined by the nature of the sponsoring agency. For example, while the Federal Highway Administration virtually guarantees distribution of research it funds through a standardized information distribution program, distribution of research funded by state, local, or private agencies will often be at the discretion of the researcher.

For example, numerous studies are conducted either "in-house" or are funded by state or local agencies. Since there is no guaranteed distribution scheme for such studies, it is the researcher's responsibility to see that the results are brought to the attention of those who can use them. In many cases, direct mailings may be the only means possible. If this is done (and it is strongly recommended along with the other avenues discussed in the following sections), the mailing should at least be sent to FHWA, other researchers (especially those whose studies have been referenced, if they are still active in research), other state highway divisions, and, where appropriate, the Governor's Highway Safety Program in each state.

5.3.1 Distribution of results in short article form.

In addition to FHWA's information distribution scheme and the suggested direct mailing to other state agencies, it is also possible for the researcher to distribute his results in the form of short, technical articles which will summarize the full-scale technical report. The condensation process usually involves shortening all sections of the report, but typically, the detailed Introduction, Review of the Literature,

and the Methodology section are condensed most. While the Results and the Discussion and Recommendations sections will also be condensed, they will need to continue to contain all pertinent information.

There are a number of journals to which these short articles can be submitted for publication. These include the following:

1. The ITE Journal (a publication of the Institute of Transportation Engineers);
2. The Transportation Research Board Record (a publication of the Transportation Research Board);
3. The Transportation Research News (an additional publication of TRB which usually includes condensed summaries of technical studies);
4. Public Roads (a publication of the Federal Highway Administration);
5. The Journal of Safety Research (a publication of the NSC);
6. Accident Analysis and Prevention (British journal published by Pergamon Press);
7. Traffic Safety (a publication of the National Safety Council); and
8. Traffic Quarterly (a publication of the ENO Foundation for Transportation)

While there are other magazines such, as Public Works and Civil Engineering, which will accept engineering-type articles related to accident research, the ones cited above are the journals which appear to be most used by researchers and administrators in the field. It is for this reason that it is recommended that they be used in the distribution process.

5.3.2 Presentation of an oral report.

The most effective way to present any kind of research information, however, may be to present it orally at various annual meetings. While many different professional engineering organizations meet regularly, the four organizational meetings which are perhaps the best forums for oral presentations of accident research information are:

1. Transportation Research Board Annual Meeting. The TRB Annual meeting, held each year in the latter part of January in Washington, DC, is perhaps the most diverse engineering-oriented research meeting that exists. The meeting consists of various technical paper sessions and committee sessions. In the mechanism for oral reporting, a paper is submitted to the Transportation Research Board for review by various committee members, and if accepted, the authors are invited to present an oral report. If the paper is not accepted for presentation at a full session, there is always the opportunity to present it at the appropriate committee session. Indeed, it has been the authors' experience that committee sessions may be the best place to receive up-to-date information concerning latest developments in the field of accident research.
2. The National Association of Governor's Highway Safety Representatives Annual Meeting, usually held in early fall, is a meeting of the Governor's Highway Safety Representatives from each state. While the meeting is much less technical in nature than is the TRB meeting, there are usually limited sessions dedicated to reports on recent research. This meeting is a particularly appropriate forum for research results which are relevant to the non-engineering Governor's Highway Safety Program side of DOT.
3. The American Association of State Highway and Transportation Officials Annual and Regional Meetings. AASHTO, composed of most state highway administrators and department heads as well as local engineers, also holds an annual meeting in which technical papers are presented. In addition, regional meetings following the same format are held annually in each of the regions of the U.S.

4. The Institute of Transportation Engineers Annual Meeting. Finally, but certainly not least, the Institute of Transportation Engineers holds an annual meeting each year at which various technical presentations are made. In addition, and perhaps even more appropriate for the researcher wishing to get research information out to people in his own state, there are regional or state ITE meetings that are often held on a monthly basis.

Finally, while the presentation of an informative and interesting oral presentation is, to some extent, dependent on both the presenter's knowledge of the audience and his prior experience, knowledge of the material, prior practice of the presentation, and good visual aids are also critical. The reader is again referred to Winfrey (1962; Chapter 15) for further discussion.

5.4 Summary

The purpose of this chapter has been to provide the researcher with guidelines to the preparation and distribution of his research results. Although there are various formats or preparation schemes for presenting the text, and although there are numerous forums and distribution networks for disseminating the information, the important point is that the information be distributed and that it be distributed to both other researchers and especially to the decision maker. Unless the information is prepared so that it can be easily interpreted and used, and unless the findings are disseminated, the effects of all the other research tasks are lost or at best, greatly minimized.

5.5 Review Questions

1. What appear to be the two basic keys to the preparation of a usable research report?
2. While many preparation sequences could be followed, two steps which are often neglected but strongly recommended are:
3. Four avenues for distribution of research reports are presented in this chapter. They are:
 - 1) Distribution through FHWA
 - 2)
 - 3)
 - 4)

5.6 References

- Council, F.M., & Hunter, W.W. Implementation of proven technology in making the highway environment safe. Chapel Hill: University of North Carolina Highway Safety Research Center, 1975.
- Standards for the preparation and publication of DOT scientific and technical reports. Washington, D.C.: U.S. Department of Transportation, 1975.
- Turkey, J. W. Exploratory data analysis, Reading Massachusetts: Addison-Wesley Publishing Co., 1977.
- Winfrey, R. Technical and business report preparation, (3rd ed.) Ames: Iowa State University Press, 1962.

CHAPTER VI

SUMMARY

Although there has been a relatively long history of research in the highway area, many of the studies of highway-related treatments have not produced reliable results. Consequently, the highway administrator is often forced to make decisions without the benefit of sound information. This gap in the knowledge has been caused by both poor study methodology and inadequate preparation and distribution of reports. Because even effective highway-related treatments can realistically be expected to reduce only a small proportion of the total accidents that occur, it has become increasingly important that evaluators and researchers in the area utilize the most powerful research techniques available.

The manual has attempted to help meet this need by compiling material from a number of sources including existing studies of highway treatments, existing studies of experimental methodology (particularly from other social science areas), existing statistical texts, and finally, from the combined knowledge of FHWA and the writers. In this regard, the manual represents a condensation and combination of the work of others, rather than the development and description of new methodologies. The rationale for this approach is justified because the problem has not been caused by a lack of appropriate methodologies which can be used in highway safety research, but by the failure to use existing methodologies.

Because of the amount of information which has been included in the preceding sections, it is difficult to summarize the important aspects without repeating details. However, in the following section, an attempt has been made to provide the researcher with guidelines that emphasize some (but not all) of the key ideas in each chapter.

Guidelines for the Accident Researcher

Chapter I.

1. The researcher should always remember the rationale for research is to provide inputs to decision makers.
2. While much has been done, a large number of gaps remain in our documented knowledge. These gaps can only be filled by sound research conducted by competent researchers.

Chapter II.

1. Accident research is by no means the only type of safety research. However, when other methods are used (e.g., test track studies, crash tests, surrogate measure studies), they should ultimately be followed up with well-designed accident research if the cost of the treatments are to be weighed against direct safety benefits.
2. Although computerized police reports are the most common source of data, other potential sources should also be considered (e.g., the hard copy of the police report, objective driver reports, the reports from special on-scene investigation teams, and national data bases).
3. Be aware of possible reporting threshold differences across jurisdictional boundaries and changes within the study period.
4. Consider if and how the nature of the reporting threshold could influence the outcome of the study.
5. Obtain information on "real world" reporting practices among the police investigators and determine how these practices could affect accident data (e.g., failure to report minor accidents during rush hour).
6. Study the mileposting practices of investigators for possible erroneous location (and thus erroneous characteristics) data.

7. Study the basic accident report form carefully to detect possible problems related to definition of variables (e.g., "first crash event," "crash speed").
8. The researcher must involve him/herself in the traffic records system in order to input and help facilitate changes which could upgrade the data. Such active participation includes:
 - a. Providing better training for investigators
 - b. Providing inputs to better designed forms
 - c. Providing feedback to investigating officers
 - d. Designing and implementing special supplementary police data collection procedures
 - e. Enhancing basic data by special team investigations.
9. The researcher should anticipate relatively low sample sizes of accidents at a given set of locations and relatively modest treatment benefits and thus must carefully plan and design his or her research.
10. Exposure data, while often not collected, are fundamental to the prediction of the likelihood of an accident. Collect these data if at all feasible.
11. Although many sources of exposure data exist, the one source the researcher can control and use most often is the origin and destination study.
12. Beware of collection biases in exposure data, such as collection during only certain hours of the day or parts of the year.

Chapter III.

1. If there is a choice between a designed evaluation and a regression type analysis of a countermeasure, always choose the evaluation approach to increase control over extraneous factors.
2. Administrative (process) evaluation, while appropriate for system support activities, is only one part of the necessary evaluation of countermeasures. Effectiveness evaluation is the primary goal.
3. The keys to selling well-designed effectiveness evaluations to administrators include:
 - a. Limited safety funds requiring knowledge of which programs work.
 - b. The necessity to measure what are expected to be modest benefits for most individual treatments.
 - c. The advocacy of the "experimental basis" approach to problem solving rather than the advocacy of a specific treatment.
4. The determination of the most appropriate criterion to be measured directly affects the possibility of detecting a true benefit. The criterion should be determined by "what the countermeasure is intended to do."
5. Appropriate criterion include accident frequencies or rates, accident severity, or appropriate proxy measures.
6. Appropriate proxy measures should be measurable and have a known relationship to accidents or accident severity.
7. The researcher must always attempt to establish an evaluation design which helps insure that any change observed in the measured criterion is due to the treatment implemented and not due to any other causes, and that the results obtained can be generalized to the population in question.

8. In highway related studies, the major threats to evaluation validity are:
 - a. History
 - b. Maturation
 - c. Regression artifacts
 - d. Instability
9. Always avoid a simple Before/After design. In the absence of other possibilities, the researcher should at least attempt to expand the data into a time series design.
10. Attempt to plan for a Before/After (or time series) with a randomly assigned control group.
11. Search carefully for "automatic" control groups due to funding limitations or to implementation staging schedules.
12. Become part of the project planning team to insure that the strong designs can be implemented.
13. Embrace "matching" (of locations) prior to randomization. Avoid matching after treatment implementation, particularly in studies involving high accident locations.
14. If high accident locations must be studied in the absence of randomly assigned controls, consider the tie breaking and regression discontinuity designs.
15. For any test, it is important to consider "practical" significance along with statistical significance. Often results which are not significant in the practical sense will be statistically significant because of large sample size.
16. Avoid the "do all we can" situation. It is almost impossible to determine effectiveness levels except for the entire treatment package.
17. Remember that statistical procedures overcome only one threat to evaluation validity (i.e., instability). The remaining threats can only be attacked by designing the evaluation correctly.
18. In testing, give special attention to alpha, the probability of making a Type 1 error (i.e., concluding that an effective program is not effective), and beta, the probability of making a Type 2 error (i.e., failing to conclude that a program is effective when indeed it is) along with the consequences of these errors.
19. Always use the most appropriate statistical test, and consider higher alpha-levels (e.g., 0.1 or 0.2) to help reduce the chance of a Type 2 error (i.e., failing to detect a true difference).
20. The researcher should always attempt to calculate the sample size necessary to detect a "real" difference before the evaluation and treatment begin. (Sample size is established by the level of alpha and beta. Consult a statistician when in doubt.)
21. In general, one-tailed statistical tests appear to be appropriate where they can be carried out. Even with significance, check to be sure that it is in the expected direction.
22. The choice of statistical tests should always be based on:
 - a. The design used
 - b. The nature of the criterion (frequencies, rates, proportions, variances, shifts in distribution)
 - c. The type of data (continuous, categorical)
23. Consult Table 3.3 for a listing of the appropriate statistical tests for a given evaluation design, criterion nature, and data type. (References to various statistical texts are listed to supplement the material and examples given in the manual.)

24. Attempt to combine evaluation results with program costs as the best decision-making tool.

Chapter IV.

1. In the study of relationships, the sample size will usually be established by available data. There are no general guidelines to sample size requirements in this type study.
2. The sample chosen should be representative of the total population. The guarantee is to choose a random sample.
3. If a random sample is too expensive, consider a systematic sample with a random start.
4. In developing models, consider including ADT as an independent variable (or use on both sides of the equation).
5. Appropriate dependent variables include accident frequencies or rates, accident severity, or appropriate proxy measures.
6. Consider using moderate or severe driver injury as a severity related criterion.
7. In examining data to identify and quantify relationships, the researcher should:
 - a. Conduct preliminary screening of variables
 - b. Develop models
 - c. Test the models for predictive accuracy
8. The variable screening procedures will depend on the nature of the data:
 - a. Continuous: Pearson product moment correlations
 - b. Categorical:
 - 1) Nominal: Chi-square
CHAID
 - 2) Ordinal: Spearman's rho
Kendall's tau
Goodman's G
9. For model building, the choice of statistical procedure will also depend on the type of data being used.
 - a. Continuous: regression (multiple linear, polynomial, non-linear)
 - b. Categorical: weighted least squares regression (GENCAT)
10. For prediction using regression models, the researcher SHOULD NOT extrapolate outside the range of the independent variables used in the model building.
11. The goal of regression procedures is to build a model that accounts for the maximum amount of the variation in the dependent variable with the minimum number of independent variables possible.
12. In accident research, relatively low levels of association and R^2 's should be expected since most individual variables will have only a modest effect on the outcome variable. Accidents are complex events with a host of simultaneously contributing factors.
13. Since much accident-related data are categorical in nature, the researcher should consider (with the aid of a statistician) the use of the new nonparametric procedures. They are more appropriate than the traditional techniques as they better fulfill the assumptions required.

Chapter V.

1. The researcher should remember that a very important determinant of whether the research is utilized is proper reporting to the administrators and to others in the field.
2. Be aware of the basic keys to report preparation:
 - a. Knowledge of the intended reader
 - b. Emphasis on interpretation of results
3. While many report formats and preparation sequences are possible always prepare a detailed outline and build in review of both technical content and clarity of presentation.
4. Always assure that the research results are distributed to users, administrators, and other researchers through:
 - a. Written reports
 - b. Short articles
 - c. Oral presentations

Self Study Aid

One of the requirements for the manual was that it be suitable for self-study by engineers or evaluators who do not have access to classroom instruction. A pre-test in Chapter 1 and review questions at the end of each chapter provide the researcher with some measure of his current level of knowledge. As a final aid to the user, a post-test has been provided at the end of this final chapter. Unlike the preceding tests or review questions, the questions in the post-test have been keyed to the appropriate sections in the manual where the reader can refer for the solution.

Closure

The writers feel that one final point should be made here. For this manual to remain a usable document, it would be anticipated that there will need to be periodic updates of the material contained herein. In the spirit of peer review so strongly advocated earlier in Chapter 5, both the writers and FHWA would greatly appreciate comments from users concerning better ways of presenting the material contained herein and comments concerning other material which should be included.

The manual has been developed as an aid to the highway accident researcher in his attempts to assure that the limited evaluation dollars at his disposal are well-spent--that his study methods are sound and that the results he presents are usable. As pointed out earlier, however, people rather than books provide the final solutions to problems. While this text can hopefully be an aid in overcoming the needs cited above, the researcher himself is the real key in the effort aimed at increasing the amount of sound research findings used in real world decision making.

Self Study Post-Test

1. Describe three causes of potential biases that may be present in a given accident data base of which the researcher should be aware. (Section 2.2.1)
2. What is exposure data and why is it so important in accident research? List three existing sources of mileage exposure data. (Section 2.2.3a,b)
3. How is a representative sample of a population selected? (Section 3.8.3, 4.2.1b)
4. In some evaluations of countermeasures, a substitute measure (proxy measure) will be used as the criterion in place of accidents. List the two attributes that an acceptable proxy measure must possess. (Section 3.5.3c)
5. A researcher is interested in ascertaining the relationship between variables which may not be linearly related. What type of analysis should she employ? (Section 4.3.2a)
6. The people of New Hebrides have decided that lice produce good health since all their healthy tribesmen have lice and none of the sick ones do. The tribe statistician has calculated a high correlation between the number of lice and degree of health. Briefly discuss this correlation in terms of cause-effect. (Section 3.5.2)
7. What is the basic question the evaluator should ask in determining what should be measured (i.e., in determining the criterion variable) in an evaluation? (Section 3.5.3)
8. A before/after study has indicated that the placement of concrete median barriers has increased accident frequencies on freeways. How can such a treatment still be justified? (Section 3.5.3b)
9. When a change is detected in any evaluation of a highway countermeasure, there are many possible causes including the treatment itself. List the four main rival explanations for a given change, other than the treatment. (Section 3.6)
10. There are various types of evaluation designs (e.g. Before/After, control group designs, time series, etc.). What is the basic reason that a researcher would apply a sound design? (Section 3.6)
11. Which study design would be appropriate to evaluate a law reducing speed limits on all freeways to 55 mph? (Section 3.7.1d,e,f)
12. In budgeting for the coming fiscal year a highway engineering dept. has a set operating improvement budget and the results of the evaluations of three proposed improvements. Which if any of the following improvements should the department make? All cost the same amount. (Section 3.8.1, 3.8.2, Appendix B)

	<u>α</u>	<u>Calculated values</u>	<u>d.f.</u>	<u>Critical values</u>
Improved Pavement Delineation	.05	$t = .997$	10	$t_c = 1.80$
Breakaway Poles	.05	$\chi^2 = 3.22$	1	$\chi_c^2 = 3.84$
A New Attenuation System	.05	$\chi^2 = 2.49$	1	$\chi_c^2 = 3.84$

13. Due to large increases in Labor Day weekend traffic, police officers in state A decide to report only those crashes that involve personal injury to the occupants of in-state vehicles. How can this practice affect a study of the relationship between accidents and traffic volume? (Section 2.2.1a)
14. A state traffic engineer is requested by the FHWA to collect accident and highway characteristics data on a sample of sections of Interstate highway. Because the purpose of the study is to predict accident rates based on highway characteristics, the engineer samples those locations which have experienced one or more accidents in the past year. Comment briefly on the adequacy of this sample. (Section 4.2.1)

15. A research is interested in developing a relationship between some measure of safety and the feet of guardrail per mile, the number of breakaway and non-breakaway telephone poles per mile, and the number of protected bridge piers per mile. What would be an appropriate dependent (predicted) variable to be used in the model? (Section 4.2.2b)
16. While many statistical tests exist for analyzing data collected in an evaluation, the choice of most appropriate test basically depends on three factors. These are: (Section 3.8.4)
 - a. The evaluation design used
 - b.
 - c.
17. (a) To the highway engineer with little money to expend which type of error is more acceptable, Type I or Type II? Why? (Section 3.8.2)

(b) What about the researcher attempting to find an effective countermeasure for an important problem area in which no good treatments exist? Explain your reason. (Section 3.8.2)
18. A researcher is evaluating the effectiveness of water-filled crash attenuation devices. The devices have been placed in gore areas of arterials which carry heavy commuter traffic involving car-pooling. A comparison group of locations has been chosen from rural freeways experiencing similar ADT's. Would total number of serious occupant injuries or total occupant deaths be appropriate criteria for the evaluation? (Section 3.5.3b)
19. While many sequences could be followed in the preparation (writing) of a research report, two steps which are often neglected but strongly recommended are (Section 5.2.1)
 - a)
 - b)
20. A number of avenues for distribution of highway-related research reports are available to the researcher. Four of these are: (Section 5.3)
 - 1) Distribution through FHWA
 - 2)
 - 3)
 - 4)

APPENDIX A

Standard Statistical Tables:

- A.1 t-distribution for 1-tail tests
- A.2 t-distribution for 2-tail tests
- A.3 z-distribution for 1 and 2 tail tests
- A.4 χ^2 -distribution for 2-tail tests
- A.5 D for the Kolmogorov Smirnov 2-sample test
(two-tail and one-tail tests)

Sources: Data in Tables A.1 and A.2 extracted from tables produced by the University of North Carolina Department of Biostatistics. Data used with permission of the Department of Biostatistics.

Data in Tables A.3 and A.4 extracted from Introduction to Statistical Analysis by W. J. Dixon and F. J. Massey, Jr., pp. 382-383, 386-387, respectively. (Copyright by McGraw - Hill Book Company, Inc., 1957.) All data used with permission of McGraw - Hill.

Data in Table A.5 extracted from "Table for Estimating the Goodness of Fit of Empirical Distributions" by N. Smirnov, Annals of Mathematical Statistics, Vol. 19 (1948) pp. 280-281. Data were used with permission of the Institute of Mathematical Statistics.

Table A.1 The t distribution for 1-tail test. (Values of t_c where α equals the area under the t-distribution to the right of t_c .)

Degrees of Freedom	<u>α-level</u>			
	0.20	0.10	0.05	0.01
1	1.376	3.078	6.314	31.821
2	1.061	1.886	2.920	6.965
3	0.978	1.638	2.353	4.541
4	0.941	1.533	2.132	3.747
5	0.920	1.476	2.015	3.365
6	0.906	1.440	1.943	3.143
7	0.896	1.415	1.895	2.998
8	0.889	1.397	1.860	2.896
9	0.883	1.383	1.833	2.821
10	0.879	1.372	1.812	2.764
11	0.876	1.363	1.796	2.718
12	0.873	1.356	1.782	2.681
13	0.870	1.350	1.771	2.650
14	0.868	1.345	1.761	2.624
15	0.866	1.341	1.753	2.602
16	0.866	1.337	1.746	2.583
17	0.863	1.333	1.740	2.567
18	0.862	1.330	1.734	2.552
19	0.861	1.328	1.729	2.539
20	0.860	1.325	1.725	2.528
21	0.859	1.323	1.721	2.518
22	0.858	1.321	1.717	2.508
23	0.858	1.319	1.714	2.500
24	0.857	1.318	1.711	2.492
25	0.856	1.316	1.708	2.485
26	0.856	1.315	1.706	2.479
27	0.855	1.314	1.703	2.473
28	0.855	1.313	1.701	2.467
29	0.854	1.311	1.699	2.462
30	0.854	1.310	1.697	2.457
40	0.851	1.303	1.684	2.423
60	0.848	1.296	1.671	2.390
120	0.845	1.289	1.658	2.358
∞	0.842	1.282	1.645	2.326

Table A.2 The t-distribution for 2-tail tests. (Values of t_c where α equals the sum of the area under the t distribution to the right of t_c and to the left of $-t_c$.)

Degrees of Freedom	<u>α-level</u>			
	0.20	0.10	0.05	0.01
1	3.078	6.314	12.706	63.657
2	1.886	2.920	4.303	9.925
3	1.638	2.353	3.182	5.841
4	1.533	2.132	2.776	4.604
5	1.476	2.015	2.571	4.032
6	1.440	1.943	2.447	3.707
7	1.415	1.895	2.365	3.499
8	1.397	1.860	2.306	3.355
9	1.383	1.833	2.262	3.250
10	1.372	1.812	2.228	3.169
11	1.363	1.796	2.201	3.106
12	1.356	1.782	2.179	3.055
13	1.350	1.771	2.160	3.012
14	1.345	1.761	2.145	2.977
15	1.341	1.753	2.131	2.947
16	1.337	1.746	2.120	2.921
17	1.333	1.740	2.110	2.898
18	1.330	1.734	2.101	2.878
19	1.328	1.729	2.093	2.861
20	1.325	1.725	2.086	2.845
21	1.323	1.721	2.080	2.831
22	1.321	1.717	2.074	2.819
23	1.319	1.714	2.069	2.807
24	1.318	1.711	2.064	2.797
25	1.316	1.708	2.060	2.787
26	1.315	1.706	2.056	2.779
27	1.314	1.703	2.052	2.771
28	1.313	1.701	2.048	2.763
29	1.311	1.699	2.045	2.756
30	1.310	1.697	2.042	2.750
40	1.303	1.684	2.021	2.704
60	1.296	1.671	2.000	2.660
120	1.289	1.658	1.980	2.617
∞	1.282	1.645	1.960	2.576

Table A.4 The χ^2 distribution for 2-tail test. (Values of χ^2_c where α equals the area under the χ^2 distribution to the right of χ^2_c).

Degrees of Freedom	<u>α-level</u>			
	0.20	0.10	0.05	0.01
1	1.642	2.706	3.841	6.635
2	3.219	4.605	5.991	9.210
3	4.642	6.251	7.815	11.345
4	5.989	7.779	9.488	13.277
5	7.289	9.236	11.070	15.086
6	8.558	10.645	12.592	16.812
7	9.803	12.017	14.067	18.475
8	11.030	13.362	15.507	20.090
9	12.242	14.684	16.919	21.666
10	13.442	15.987	18.307	23.209
11	14.631	17.275	19.675	24.725
12	15.812	18.549	21.026	26.217
13	16.985	19.812	22.362	27.688
14	18.151	21.064	23.685	29.141
15	19.311	22.307	24.996	30.578
16	20.465	23.542	26.296	32.000
17	21.615	24.769	27.587	33.409
18	22.760	25.989	28.869	34.805
19	23.900	27.204	30.144	36.191
20	25.038	28.412	31.410	37.566
21	26.171	29.615	32.671	38.932
22	27.301	30.813	33.924	40.289
23	28.429	32.007	35.172	41.638
24	29.553	33.196	36.415	42.980
25	30.675	34.382	37.652	44.314
26	31.795	35.563	38.885	45.642
27	32.912	36.741	40.113	46.963
28	34.027	37.916	41.337	48.278
29	35.139	39.087	42.537	49.588
30	36.250	40.256	43.773	50.892
35	41.778	46.059	49.802	57.342
40	47.269	51.805	55.758	63.691
45	52.729	57.505	61.656	69.957
50	58.164	63.167	67.505	76.154
60	68.972	74.397	79.082	88.379
70	79.715	85.527	90.531	100.425
80	90.405	96.578	101.879	112.329
90	101.054	107.565	113.145	124.116
100	111.667	118.498	124.342	135.806
120	132.806	140.233	146.567	158.950
140	153.854	161.827	168.613	181.840
160	174.828	183.311	190.516	204.530
180	195.743	204.704	212.304	227.056
200	216.609	226.021	233.994	249.445

Table A.3 The z-distribution for 1 and 2-tail tests. (Values of z_c where α equals the area in the tail(s) of the distribution.)

	<u>α-level</u>				
	0.20	0.15	0.10	0.05	0.01
1-tailed	0.84	1.04	1.28	1.64	2.33
2-tailed	1.28	1.44	1.64	1.96	2.58

Table A.5 Table of critical values of D_c in the Kolmogorov-Smirnov 2-tail test* for 2 samples. (n_1 and n_2 are sample sizes)

<u>α</u>	<u>D_c</u>
.10	$1.22 \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$
.05	$1.36 \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$
.01	$1.63 \sqrt{\frac{n_1 + n_2}{n_1 n_2}}$

*For one tail test, convert D to a χ^2 with 2 degrees of freedom using:

$$\chi^2 = 4D^2 \left[\frac{n_1 n_2}{n_1 + n_2} \right]$$

Then compare this χ^2 to the critical value of χ_c^2 found in Table A.4.