Effective Experiment Design and Data Analysis in Transportation Research

DETAILS

80 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-25849-4 | DOI 10.17226/22707

CONTRIBUTORS

Richard W. Lyles, M. Abrar Siddiqui, Neeraj Buch, William C. Taylor, Syed Waqar Haider, Dennis C. Gilliland, Bruce W. Pigozzi, and Joseph E. Hummer; National Cooperative Highway Research Program; Transportation Research Board; National Academies of Sciences, Engineering, and Medicine

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

NCHRP REPORT 727

Effective Experiment Design and Data Analysis in Transportation Research

Richard W. Lyles
M. Abrar Siddiqui
Neeraj Buch
William C. Taylor
Syed Waqar Haider
Dennis C. Gilliland
Bruce W. Pigozzi
MICHIGAN STATE UNIVERSITY
East Lansing, Michigan

IN ASSOCIATION WITH

Joseph E. Hummer
NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

Subscriber Category
Research

Research sponsored by the American Association of State Highway and Transportation Officials
in cooperation with the Federal Highway Administration

TRANSPORTATION RESEARCH BOARD

WASHINGTON, D.C.
2012
www.TRB.org

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

Systematic, well-designed research provides the most effective approach to the solution of many problems facing highway administrators and engineers. Often, highway problems are of local interest and can best be studied by highway departments individually or in cooperation with their state universities and others. However, the accelerating growth of highway transportation develops increasingly complex problems of wide interest to highway authorities. These problems are best studied through a coordinated program of cooperative research.

In recognition of these needs, the highway administrators of the American Association of State Highway and Transportation Officials initiated in 1962 an objective national highway research program employing modern scientific techniques. This program is supported on a continuing basis by funds from participating member states of the Association and it receives the full cooperation and support of the Federal Highway Administration, United States Department of Transportation.

The Transportation Research Board of the National Academies was requested by the Association to administer the research program because of the Board's recognized objectivity and understanding of modern research practices. The Board is uniquely suited for this purpose as it maintains an extensive committee structure from which authorities on any highway transportation subject may be drawn; it possesses avenues of communications and cooperation with federal, state and local governmental agencies, universities, and industry; its relationship to the National Research Council is an insurance of objectivity; it maintains a full-time research correlation staff of specialists in highway transportation matters to bring the findings of research directly to those who are in a position to use them.

The program is developed on the basis of research needs identified by chief administrators of the highway and transportation departments and by committees of AASHTO. Each year, specific areas of research needs to be included in the program are proposed to the National Research Council and the Board by the American Association of State Highway and Transportation Officials. Research projects to fulfill these needs are defined by the Board, and qualified research agencies are selected from those that have submitted proposals. Administration and surveillance of research contracts are the responsibilities of the National Research Council and the Transportation Research Board.

The needs for highway research are many, and the National Cooperative Highway Research Program can make significant contributions to the solution of highway transportation problems of mutual concern to many responsible groups. The program, however, is intended to complement rather than to substitute for or duplicate other highway research programs.

NCHRP REPORT 727

Project 20-71
ISSN 0077-5614
ISBN 978-0-309-25849-4
Library of Congress Control Number 2012945556

COPYRIGHT INFORMATION

NOTICE

The project that is the subject of this report was a part of the National Cooperative Highway Research Program, conducted by the Transportation Research Board with the approval of the Governing Board of the National Research Council.

The members of the technical panel selected to monitor this project and to review this report were chosen for their special competencies and with regard for appropriate balance. The report was reviewed by the technical panel and accepted for publication according to procedures established and overseen by the Transportation Research Board and approved by the Governing Board of the National Research Council.

The opinions and conclusions expressed or implied in this report are those of the researchers who performed the research and are not necessarily those of the Transportation Research Board, the National Research Council, or the program sponsors.

The Transportation Research Board of the National Academies, the National Research Council, and the sponsors of the National Cooperative Highway Research Program do not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the object of the report.

Published reports of the

NATIONAL COOPERATIVE HIGHWAY RESEARCH PROGRAM

are available from:

Transportation Research Board
Business Office
500 Fifth Street, NW
Washington, DC 20001

and can be ordered through the Internet at:

http://www.national-academies.org/trb/bookstore

Printed in the United States of America

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. On the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, on its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

The **Transportation Research Board** is one of six major divisions of the National Research Council. The mission of the Transportation Research Board is to provide leadership in transportation innovation and progress through research and information exchange, conducted within a setting that is objective, interdisciplinary, and multimodal. The Board's varied activities annually engage about 7,000 engineers, scientists, and other transportation researchers and practitioners from the public and private sectors and academia, all of whom contribute their expertise in the public interest. The program is supported by state transportation departments, federal agencies including the component administrations of the U.S. Department of Transportation, and other organizations and individuals interested in the development of transportation. **www.TRB.org**

www.national-academies.org

COOPERATIVE RESEARCH PROGRAMS

CRP STAFF FOR NCHRP REPORT 727

Christopher W. Jenks, *Director, Cooperative Research Programs*
Crawford F. Jencks, *Deputy Director, Cooperative Research Programs*
B. Ray Derr, *Senior Program Officer*
Andréa Harrell, *Senior Program Assistant*
Eileen P. Delaney, *Director of Publications*
Sharon Lamberton, *Assistant Editor*

NCHRP PROJECT 20-71 PANEL
Field of Special Projects—Area of Research

Donald L. Dean, *California DOT, Sacramento, CA* (Chair)
Deniz Sandhu, *New York State Education Department, formerly New York State DOT, Albany, NY*
Montasir M. Abbas, *Virginia Polytechnic Institute and State University, Blacksburg, VA*
Tie He, *Nevada DOT (retired), El Dorado Hills, CA*
Gary L. Robson, *West Virginia DOT (retired), Hurricane, WV*
Vincent Van Der Hyde, Jr., *Oregon DOT (retired), Salem, OR*
Zhongjie "Doc" Zhang, *Louisiana DOTD, Baton Rouge, LA*
Peter A. Kopac, *FHWA Liaison*
Richard Pain, *TRB Liaison*

FOREWORD

By B. Ray Derr
Staff Officer
Transportation Research Board

This report describes the factors that should be considered in designing experiments and presents 21 typical transportation examples illustrating the experiment design process, including selection of appropriate statistical tests. The examples encompass a wide range of transportation disciplines and statistical methods. This report will be very beneficial to anyone with limited research experience needing to answer a question based on data (e.g., presenting ozone concentrations in a region, determining whether a contractor's quality assurance/quality control procedures are adequate, estimating the effect of automated enforcement on speeds, monitoring trends in the condition of bridge superstructures, developing a user survey to determine the impact of transit fare changes). The report is a companion to *NCHRP CD-22, Scientific Approaches to Transportation Research,* Volumes 1 and 2, which were developed in NCHRP Project 20-45 and present detailed information on statistical methods. *NCHRP CD-22* is available at http://www.trb.org/Main/Blurbs/152122.aspx

---

Transportation agencies spend millions of dollars conducting research to improve their ability to plan, design, construct, maintain, and operate the transportation system. These research projects cover a broad range of topics and use approaches ranging from fully controlled laboratory experiments to field observational studies. Unfortunately, some research projects use inappropriate experimental designs or data analysis techniques, thereby increasing costs and decreasing the likelihood of success.

There are many excellent university-level texts on experimental design and data analysis, but these are often not well suited to the needs of those involved in state DOT research. Principal investigators and DOT research program and project managers need practical information that focuses on common problems that DOTs face so that they can make better decisions when planning and conducting research. NCHRP Project 20-45, "Scientific Approaches for Transportation Research," produced *NCHRP CD-22* that presents valuable material for transportation researchers. It does not, however, cover experimental design.

In NCHRP Project 20-71, Michigan State University and North Carolina State University determined how state DOTs handle experimental design for contract and in-house research and to what extent the results from NCHRP Project 20-45 are in use. They then described basic principles and approaches that should guide experimental design and data analysis in transportation research. Example cases of experiment design were developed across a broad range of functional areas within transportation. The example cases include: (1) the research question being addressed, (2) the dependent and independent variables, (3) data that will be collected, (4) techniques for analyzing the data, (5) interpretation of results, (6) discussion of the process, and (7) application of the approach used in other areas of transportation.

CONTENTS

Note: Many of the figures and tables in this report have been converted from color to grayscale for printing. The electronic version of the report (posted on the Web at www.trb.org) retains the color versions.

CHAPTER 1

Introduction

Developing Effective Experiment Designs and Data Analysis Plans

Personnel in state departments of transportation (DOTs), county road commissions, and other transportation agencies often engage in evaluations of different products or want to know whether a treatment has resulted in an improvement in their system, such as whether some crack sealer is better than the one that was used in the past or whether changing signs on a number of horizontal curves really led to a crash reduction. Some agencies also engage in more comprehensive and formal studies through ongoing research programs. The questions asked in informal and formal investigations are significantly different, but the common element is that some sort of experiment design should be done that includes specifying the appropriate statistical analysis. In some cases, the required design and analysis are straightforward; in others, complex. The purpose of these guidelines is to help the practitioner ask the right questions and design an analysis that is appropriate to address the research problem.

Throughout *NCHRP Report 727,* the terms *research* and *experiment design* are used fairly loosely. A large, state-funded project undertaken by a local university to evaluate the effectiveness of some statewide highway safety improvement program clearly qualifies as research. In contrast, comparing the effectiveness of two products that are routinely used may not seem like research that requires a rigorous experiment design. For the purposes of this guide, however, such a comparison is classified as research, and determining the approach to take to make that comparison is, more or less, designing the experiment.

What This Guide Is and Is Not

This guide is not intended to turn practitioners into statisticians. For the expert, let alone the uninitiated, designing experiments and understanding and undertaking complex statistical analysis can be truly daunting. It is unnecessary (and would be ill-advised) to try to turn all traffic engineers, pavement designers, and other transportation researchers into qualified statisticians. However, transportation professionals can learn some of the terminology, understand some of the pitfalls in conducting research, learn to ask effective questions, interact more productively with researchers and statisticians, and obtain understandable and valid test results. Practitioners need to be able to understand what needs to be done and what the results really mean with reference to practice.

This guide will help transportation practitioners become better engaged in undertaking any kind of experiment design and analysis, improve interactions with contracted consultants and researchers, know when to engage statistical experts, and discern more useful information from transportation studies.

Organization of the Guide

This report has three chapters. Chapter 2 focuses on how to get organized, ask questions, develop an experiment design, and understand how an appropriate statistical technique/approach is selected. Chapter 3 presents a variety of common examples of experiments in different areas of interest (e.g., traffic engineering, maintenance, and planning). The examples emphasize working through various questions related to experiment design. The questions lead to the selection of an analysis technique that is appropriate for the research question formulated. The examples are generally short, although some are more in-depth to better illustrate the issues that must be addressed.

It should be noted that NCHRP Project 20-45, "Scientific Approaches to Transportation Research," has already resulted in the publication of *NCHRP CD-22,* a CD-ROM containing a two-volume primer on statistical methods. Throughout this report, when more details on specific analysis techniques are required, excerpts from or references to NCHRP 20-45 are included. For readers' convenience, these references also have been summarized in the Resources section at the end of Chapter 3. More information about NCHRP Project 20-45 and *NCHRP CD-22* can be found at: http://www.trb.org.

CHAPTER 2

Some Questions and Answers About Experiment Design

This chapter steps the reader through a question-based approach to developing an experiment design appropriate for what the potential researcher wants to accomplish. What drives the guidelines is the process of designing an experiment or a research project. This chapter follows a question-and-answer format.

What Is the Research Question (What Do You Want To Know)?

The first step in designing an experiment or research project, or just undertaking data analysis, is the definition of the problem statement and the objectives of the research. Formulating the right question is important because varying perspectives can lead to different kinds of analysis. Regardless of the quality of the research approach or the qualifications of the analyst, if the wrong question has been asked, the answer may well prove useless. Unfortunately, this happens far more often than people think!

The basic point to be made in formulating effective research questions is that they are rarely as straightforward as they first seem. In real life, the researcher often may be stuck with a less straightforward or more comprehensive question than he or she had hoped for. Under these circumstances, it is important to at least know the limitations of the study (and the eventual conclusions).

The balance of this chapter examines two typical research questions to show how generating and answering additional, related questions help shape the design of the experiment.

Typical Research Question Involving Comparison of Test Results

Is crack sealant A better than crack sealant B? More generally, is one product better than another? This research question can be further generalized to almost any area of transportation where the results of two (or more) treatments, applications, or products need to be compared.

What Does This Question Imply?

The question about crack sealants implies a comparison of test results for two different treatments leading to a determination of whether sealant A is better than sealant B. The question requires that the word *better* be defined in a specific way. Is a yes-or-no answer sufficient, or does performance have to be improved by some incremental amount? For example, if treatment A costs more than treatment B, is "better" defined as more cost effective? This type of consideration helps define the question and the experiment design. One way to examine what the question

4 Effective Experiment Design and Data Analysis in Transportation Research

implies is to identify and address related questions. Related questions help to define the breadth of the experiment.

- Does the product have to be better in all situations? The research question "Is crack sealant A better than crack sealant B?" implies that one product or treatment will be selected over another in *all* cases, which may not be true. In some situations (e.g., when sealing deeper or more numerous cracks), sealant A may outperform sealant B. Under different conditions, sealant B may outperform sealant A.
- If the definition of "better" includes a result that is either difficult to measure or requires a long observation period, can a surrogate measure be used? (Surrogate measures themselves may have problems. If a surrogate measure is considered, it is important to ask how well the surrogate correlates with the real measure.) A classic example of using a surrogate measure in traffic safety research involves evaluation of crash frequencies and rates, which are the ultimate measures of safety but can require an experiment time frame that covers many years. Some researchers/engineers use traffic conflicts (e.g., near-misses in which one vehicle must unexpectedly give way to another) as a surrogate for crashes. Traffic conflicts provide a reasonably effective surrogate measure for crash frequencies because measuring the near-misses allows researchers to compile a larger sample of data within a shorter period of time.
- How large must the sample be to determine whether one product is better than another? The answer to this question involves confidence limits, the cost of obtaining samples of the various products being evaluated, and the potential dollar savings associated with choosing one product over the other under various experimental conditions.
- Are there other consequences of using one product over the other? These could include, for example, the cost of application, environmental costs, and potential long-term price changes for the product.

Typical Research Question Involving Independent and Dependent Variables

What is the effect of increasing truck traffic on ride quality? More generally, does variation in one variable (e.g., truck volume) affect the outcome or value of some other variable (e.g., ride quality)?

What Does This Question Imply?

Answering this question involves determining correlation or calibrating an equation where a dependent variable is a function of one or more independent variables. Addressing a question like this typically involves obtaining data and then fitting an equation to the data (e.g., using linear or non-linear regression). It is important to be clear whether changes in the independent variable (truck traffic) really have a causal impact on the dependent variable (ride quality).

What Issues Must Be Considered?

- Does the relationship involve cause and effect? One of the most fundamental (and critical) issues that must be considered with research involving variables is whether observed correlations indicate causal relationships. Although establishing correlation is necessary for causation, it is not sufficient. That is, even when correlation can be measured, this may not mean that a cause-and-effect relationship exists. This is a fundamental question to consider when causal models are desired.
- If it can be shown that a causal relationship exists between the independent variable and the dependent variable (e.g., if it can be shown that changes in truck traffic truly cause changes in ride quality), is it sufficient to know that the causal relationship is valid, or can something be done to control the independent variable (e.g., can truck traffic really be regulated)?

- Do other variables also have an impact on ride quality? If so, how can the effects of one variable be separated from the effects of others?
- Is the conclusion true in all cases? In this example, does the relationship between truck volume and ride quality hold for all pavement types over the entire range of truck volumes and for all types of trucks?
- What is the *form* of the relationship between the dependent and independent variables? Statistically, is the form of the relationship linear throughout, non-linear throughout, or does it vary across the observed range of values (e.g., it is linear for low values of the independent variables and non-linear for higher values)?
- Based on the nature of the phenomenon, is it necessary that the fitted regression line pass through the origin (where both the dependent and independent variables equal zero)? If not, what is the meaning of the intercept?
- What statistics are appropriate for evaluating the regression line? What are the effects of outliers in the data set?

What Else Needs to Be Asked?

Anyone undertaking or supervising a study or research project (no matter how big or small) might need to know many other things. It's hard to make a comprehensive list, especially because not everyone is faced with the same situations. That said, some additional questions follow that might be useful to think about.

Why Do a Statistical Analysis?

This question has probably been around as long as researchers have been trying to answer questions about whether some action (or project) is worthwhile, or whether some procedure or product should be changed. Often, research is conducted simply to gain more confidence in decisions that have already been made.

In general, statistical analysis is used to try to explain an observed variation (e.g., variation in pavement roughness, the number of vehicle crashes on curved roadway sections, or the performance of a product like a crack sealant). Statistical analysis can help clarify the relationship between these observations and variations in pavement design, the degree of curvature of the road, or the material differences in sealants. Although these might not be the only variables that should be considered, the idea is to use statistics to examine what it is that explains the variations that have been observed.

How Much Data Must Be Collected?

This age-old question can also be put another way: What size does the sample size need to be? The answer relates to the variation the research seeks to explain. Let's say a researcher wants to estimate a mean or average value of something (x). If all observations of x result in the same value, then there is no variance. The average value is the same as any single observation. If there is no variance, then a single observation—a sample size of one—will establish the "truth."

It would be nice if life were that simple, but in reality, research data almost always involves variation. Simply put, the more variation exists in the data, the more samples (observations) will be needed to get good estimates of statistical measures such as the mean. Likewise, the greater the level of confidence desired (i.e., the more the researcher wants to be sure that the estimates are meaningful and accurate), the more data will be needed. (Sample size and confidence are addressed in detail in NCHRP Project 20-45, Volume 2, Chapter 1, under the heading "Sample Size Determination.")

When determining sample size, it is also important to consider the amount of difference in the results that is important for the decision the experiment is designed to support. The more precision the researcher is looking for in an answer, the more data must be collected. For example, is it important to know whether increased police presence in a work zone reduces the average speed of traffic by 1 mile per hour, or would the researchers recommend using increased enforcement only if it resulted in a reduction of 5 miles per hour or more?

Because the cost of data collection often is a function of the number of observations collected, an understanding of the required accuracy is important.

If a researcher expects estimates or answers to vary according to specific parameters, the researcher needs to consider these variations in the experiment design and the sample size calculations. For example, a researcher might study pavement behavior in a state such as Michigan, which encompasses two distinct freeze-thaw zones. The researcher will need an appropriate sample size in each of the freeze-thaw zones to make conclusions that are applicable statewide. The analysis will also need to consider the amount of the overall variance in pavement behavior that is attributable to differences in freeze-thaw zones.

Does the Type of Data Collected Make a Difference?

The term *type* can be interpreted in several ways, but in statistical analysis the researcher is concerned with what form the data takes. Is the variable that is being measured continuous or discrete? The type of data often determines the kind of analysis that can be done. For example, although exceptions occur, linear regression typically requires that both dependent and independent variables be continuous (or that they can be assumed to be continuous). On the other hand, simple analysis of variance (ANOVA) techniques work when the dependent variable is continuous and the independent variables are discrete/categorical. (Different types of data and the implications for selection of analysis techniques are comprehensively discussed in NCHRP Project 20-45, Volume 2, Chapter 1, "Identification of Empirical Setting.")

What Are Some Typical Problems and Pitfalls?

Even after researchers and their collaborators come to agreement on the basic research question, issues may still be encountered. Understanding the typical problems and pitfalls can help practitioners avoid or address them proactively, before they can jeopardize the usefulness of research results.

Is the Research Question Misleading, Too Simple, or Wrong?

Some questions are either wrong or lead to incomplete or poor experiments. For example, the sample question about the relationship between truck traffic and ride quality is fairly simple. The implicit assumption is that truck traffic is the major (or perhaps the only) determinant of ride quality; however, other factors could easily have an impact on ride quality. It might be interesting to know the extent to which ride quality degradation is related to truck traffic, but determining whether truck traffic can be effectively controlled is problematic. This might be a valid question to ask in relation to local streets and roads (where truck access can be controlled), but it may not be valid for the interstate/freeway system (where there is generally less control).

Misleading, simple, or wrong questions may lead the researcher down the wrong road entirely or lead the researcher to construct a project or experiment that is too narrow or too broad in scope. If the question is too broad, the resulting experiment may be too costly to perform, espe-

cially within a limited budget. The following questions can help transportation professionals determine whether a research question is misleading, too simple, or wrong.

Has the Question Already Been Answered?

In these days of burgeoning information, it is sometimes hard to believe that any questions exist that haven't already been addressed. It is fundamental to determine whether the research project that is envisioned has already been done, and if so whether all the questions have been answered.

If the questions really have been answered, the solution is straightforward—don't do the project. Often, however, researchers find a question has been asked and answered "somewhere else" (i.e., in a different state or region of the country). In such instances, it may still be reasonable to conduct a similar study to make sure that results are consistent in the local environment. For example, in traffic safety it is well known that driving styles differ from place to place and the results of very similar studies are not necessarily consistent. The clear implication here is that the literature should be critically reviewed. The experiment design should be carefully crafted so that results can be compared (to the extent possible) to those from other projects and locations. Indeed, the experiment design used for a past project may be appropriate for a new project with very little change.

Does Real Disagreement Exist About the Question/Issue?

As with the previous issue, sometimes there is simply no good reason to undertake the research. A product or a technique may have been tested so often that everyone knows the answer. In such cases, although there is no need to reinvent the wheel, another level of analysis might be appropriate (i.e., a meta-analysis constructed to consider the results of a variety of similar studies). If a meta-analysis seems warranted, it is advisable to consult with a professional statistician.

If little or no disagreement exists about a question or issue, the most obvious outcome is to question whether the study really needs to be done. In general, research money seems too scarce these days to waste dollars on studies that really don't extend the state of the art.

What Factors Affect Outcomes?

Once the researcher is satisfied that he or she is asking the right research question and is comfortable with its form, the next step is to determine whether other factors may affect or complicate the answer.

There are almost always some other factors that can affect the outcome of a research project. Some factors may be unexpected while others may simply be overlooked. For example, if researchers want to look at the total crashes that occurred at a set of intersections after changing them in some way, the crash frequency might have been affected by an exceptionally hard winter that resulted in many more "icy road" crashes or by the fact that some of the intersections are on a road that served as a detour for another road that was closed for the construction season, thus causing significant changes in traffic volume through the studied site. Another example might be a longer study on the performance of pavements during which the volume of trucks on the road being studied unexpectedly increases (or decreases) because of detours, construction, changes in land use, or any of several other reasons.

In general, researchers have two choices for dealing with such problems:

1. They can control the experiment to eliminate some of the variations that result from other factors; and

8 Effective Experiment Design and Data Analysis in Transportation Research

2. They can explicitly consider the variations due to other factors in the selection of their analysis approach.

At this point, it is sufficient to answer the following questions. The answers to these questions also help determine which data need to be collected.

What Other Factors Can Affect the Answers to Your Research Question?

In the crack sealant example, both the average size of the cracks and the ambient temperature when the sealant was applied might affect the performance of either crack sealant A or crack sealant B in any individual application. This question reveals that there are other effects that the researcher needs to track in addition to measuring the basic performance of the sealant. The experiment design and the analysis approach need to be planned in a way that explicitly considers these other effects.

Can the Experiment Be Designed to Control for the Variations Caused by Other Factors?

For example, a researcher interested in measuring the ride quality of alternative pavement designs over time might consider how to take into account the effects of truck traffic. This question implies that the researcher must either isolate the effects being studied (e.g., by selecting sites with similar truck volumes) or explicitly control for other factors in the analysis (e.g., by using control sites or certain kinds of statistical techniques).

Can Site Selection Affect Results? How Can the Researcher Determine the Suitability of a Site for Collecting Data?

This question implies that researchers must also consider the impact that a site might have on the ultimate analysis and the results of the analysis. The effects of site selection can be handled in a couple of ways. First, researchers can control for the impacts of a site just like any other independent variable. This is typically done to determine whether results obtained from one site are unique or consistent with those obtained at similar sites. For most projects, desirable sites will be similar enough to be effectively interchangeable. This is especially true if the researcher is treating some sites and using others as untreated controls for comparison purposes. To select appropriate sites, the researcher needs to compile the characteristics that may impact the experiment and then select sites that are similar when compared on the basis of those characteristics. For example, for a traffic safety-related project or experiment in which average vehicle speed is an important measure of effectiveness, important site selection criteria would include characteristics of traffic flow such as the speed limit, the truck percentage (or volume), and the grade of the roadway.

What Are Independent Samples? How Might They Affect Results?

For most research projects, it is important that the data collected be both *representative* and *independent.* For example, in a study of pavement characteristics using sample cores in a mile-long section of roadway, the cores will ideally be representative of the entire section. Therefore, the researcher would not take all of the samples in a small area of the pavement, from spots right next to each other. (A useful approach would be to divide up the section of roadway into many small segments, then randomly choose the segments for coring.) Another example might be testing the effectiveness of a traffic control device in slowing down vehicles in advance of a horizontal curve.

In this experiment, researchers might want to collect data about vehicle speeds at, say, the point of curvature (after the motorist has seen and responded to the device). It would be important to collect only the speeds of free-flowing (lead) vehicles, disregarding those vehicles following in a line and vehicles completely separated from other vehicles (e.g., following more than *X* seconds behind a preceding vehicle). The researcher wants to collect the speeds of vehicles whose drivers are affected only by the signs themselves and whose actions are independent of the actions of the drivers in vehicles that precede them.

Summary

The concepts discussed in this chapter are intended to encourage transportation researchers to think about how to define relevant research questions, identify the appropriate dimensions of any research question, recognize the elements of experiment design necessary to answer the question, and anticipate what might go wrong with an experiment design. The goal of the report is not to convert the reader into a qualified statistician but rather to help the reader be more conversant with statistical terms and techniques. The next section presents a collection of examples that illustrate different types and complexities of experiment designs and the appropriate statistical analysis.

CHAPTER 3

Examples of Effective Experiment Design and Data Analysis in Transportation Research

About this Chapter

This chapter provides a wide variety of examples of research questions. The examples demonstrate varying levels of detail with regard to experiment designs and the statistical analyses required. The number and types of examples were selected after consulting with many practitioners. The attempt was made to provide a couple of detailed examples in each of several areas of transportation practice. For each type of problem or analysis, some comments also appear about research topics in other areas that might be addressed using the same approach. Questions that were briefly introduced in Chapter 2 are addressed in considerably more depth in the context of these examples.

All the examples are organized and presented using the outline below. Where applicable, references to the two-volume primer produced under NCHRP Project 20-45 have been provided to encourage the reader to obtain more detail about calculation techniques and more technical discussion of issues.

Basic Outline for Examples

The numbered outline below is the model for the structure of all of the examples that follow.

1. **Research Question/Problem Statement:** A simple statement of the research question is given. For example, in the maintenance category, does crack sealant A perform better than crack sealant B?
2. **Identification and Description of Variables:** The dependent and independent variables are identified and described. The latter includes an indication of whether, for example, the variables are discrete or continuous.
3. **Data Collection:** A hypothetical scenario is presented to describe how, where, and when data should be collected. As appropriate, reference is made to conventions or requirements for some types of data (e.g., if delay times at an intersection are being calculated before and after some treatment, the data collected need to be consistent with the requirements in the *Highway Capacity Manual)*. Typical problems are addressed, such as sample size, the need for control groups, and so forth.
4. **Specification of Analysis Technique and Data Analysis:** The links between successfully framing the research question, fully describing the variables that need to be considered, and the specification of the appropriate analysis technique are highlighted in each example. References to NCHRP Project 20-45 are provided for additional detail. The appropriate types of statistical test(s) are described for the specific example.
5. **Interpreting the Results:** In each example, results that can be expected from the analysis are discussed in terms of what they mean from a statistical perspective (e.g., the *t*-test result from

a comparison of means indicates whether the mean values of two distributions can be considered to be equal with a specified degree of confidence) as well as an operational perspective (e.g., judging whether the difference is large enough to make an operational difference). In each example, the typical results and their limitations are discussed.

6. **Conclusion and Discussion:** This section recaps how the early steps in the process lead directly to the later ones. Comments are made regarding how changes in the early steps can affect not only the results of the analysis but also the appropriateness of the approach.
7. **Applications in Other Areas of Transportation Research:** Each example includes a short list of typical applications in other areas of transportation research for which the approach or analysis technique would be appropriate.

Techniques Covered in the Examples

The determination of what kinds of statistical techniques to include in the examples was made after consulting with a variety of professionals and examining responses to a survey of research-oriented practitioners. The examples are not exhaustive insofar as not every type of statistical analysis is covered. However, the attempt has been made to cover a representative sample of techniques that the practitioner is most likely to encounter in undertaking or supervising research-oriented projects. The following techniques are introduced in one or more examples:

- Descriptive statistics
- Fitting distributions/goodness of fit (used in one example)
- Simple one- and two-sample comparison of means
- Simple comparisons of multiple means using analysis of variance (ANOVA)
- Factorial designs (also ANOVA)
- Simple comparisons of means before and after some treatment
- Complex before-and-after comparisons involving control groups
- Trend analysis
- Regression
- Logit analysis (used in one example)
- Survey design and analysis
- Simulation
- Non-parametric methods (used in one example)

Although the attempt has been made to make the examples as readable as possible, some technical terms may be unfamiliar to some readers. Detailed definitions for most applicable statistical terms are available in the glossary in NCHRP Project 20-45, Volume 2, Appendix A. Most definitions used here are consistent with those contained in NCHRP Project 20-45, which contains useful information for everyone from the beginning researcher to the most accomplished statistician.

Some variations appear in the notations used in the examples. For example, in statistical analysis an alternate hypothesis may be represented by H_a or by H_1, and readers will find both notations used in this report. The examples were developed by several authors with differing backgrounds, and latitude was deliberately given to the authors to use the notations with which they are most familiar. The variations have been included purposefully to acquaint readers with the fact that the same concepts (e.g., something as simple as a mean value) may be noted in various ways by different authors or analysts.

Finally, the more widely used techniques, such as analysis of variance (ANOVA), are applied in more than one example. Readers interested in ANOVA are encouraged to read all the ANOVA examples as each example presents different aspects of or perspectives on the approach, and computational techniques presented in one example may not be repeated in later examples (although a citation typically is provided).

Areas Covered in the Examples

Transportation research is very broad, encompassing many fields. Based on consultation with many research-oriented professionals and a survey of practitioners, key areas of research were identified. Although these areas have lots of overlap, explicit examples in the following areas are included:

- Construction
- Environment
- Lab testing and instrumentation
- Maintenance
- Materials
- Pavements
- Public transportation
- Structures/bridges
- Traffic operations
- Traffic safety
- Transportation planning
- Work zones

The 21 examples provided on the following pages begin with the most straightforward analytical approaches (i.e., descriptive statistics) and progress to more sophisticated approaches. Table 1 lists the examples along with the area of research and method of analysis for each example.

Example 1: Structures/Bridges; Descriptive Statistics

Area: Structures/bridges

Method of Analysis: Descriptive statistics (exploring and presenting data to describe existing conditions and develop a basis for further analysis)

1. **Research Question/Problem Statement:** An engineer for a state agency wants to determine the functional and structural condition of a select number of highway bridges located across the state. Data are obtained for 100 bridges scheduled for routine inspection. The data will be used to develop bridge rehabilitation and/or replacement programs. The objective of this analysis is to provide an overview of the bridge conditions, and to present various methods to display the data in a concise and meaningful manner.

   Question/Issue

   Use collected data to describe existing conditions and prepare for future analysis. In this case, bridge inspection data from the state are to be studied and summarized.

2. **Identification and Description of Variables:** Bridge inspection generally entails collection of numerous variables that include location information, traffic data, structural elements' type and condition, and functional characteristics. In this example, the variables are: bridge condition ratings of the deck, superstructure, and substructure; and overall condition of the bridge. Based on the severity of deterioration and the extent of spread through a bridge component, a condition rating is assigned on a discrete scale from 0 (failed) to 9 (excellent). These ratings (in addition to several other factors) are used in categorization of a bridge in one of three overall conditions: not deficient; structurally deficient; or functionally obsolete.

Table 1. Examples provided in this report.

Example	Area	Method of Analysis
1	Structures/bridges	Descriptive statistics (exploring and presenting data to describe existing conditions)
2	Public transport	Descriptive statistics (organizing and presenting data to describe a system or component)
3	Environment	Descriptive statistics (organizing and presenting data to explain current conditions)
4	Traffic operations	Goodness of fit (chi-square test; determining if observed/collected data fit a certain distribution)
5	Construction	Simple comparisons to specified values (*t*-test to compare the mean value of a small sample to a standard or other requirement)
6	Maintenance	Simple two-sample comparison (*t*-test for paired comparisons; comparing the mean values of two sets of matched data)
7	Materials	Simple two-sample comparisons (*t*-test for paired comparisons and the *F*-test for comparing variances)
8	Laboratory testing and/or instrumentation	Simple ANOVA (comparing the mean values of more than two samples using the *F*-test)
9	Materials	Simple ANOVA (comparing more than two mean values and the *F*-test for equality of means)
10	Pavements	Simple ANOVA (comparing the mean values of more than two samples using the *F*-test)
11	Pavements	Factorial design (an ANOVA approach exploring the effects of varying more than one independent variable)
12	Work zones	Simple before-and-after comparisons (exploring the effect of some treatment before it is applied versus after it is applied)
13	Traffic safety	Complex before-and-after comparisons using control groups (examining the effect of some treatment or application with consideration of other factors)
14	Work zones	Trend analysis (examining, describing, and modeling how something changes over time)
15	Structures/bridges	Trend analysis (examining a trend over time)
16	Transportation planning	Multiple regression analysis (developing and testing proposed linear models with more than one independent variable)
17	Traffic operations	Regression analysis (developing a model to predict the values that a dependent variable can take as a function of one or more independent variables)
18	Transportation planning	Logit and related analysis (developing predictive models when the dependent variable is dichotomous)
19	Public transit	Survey design and analysis (organizing survey data for statistical analysis)
20	Traffic operations	Simulation (using field data to simulate or model operations or outcomes)
21	Traffic safety	Non-parametric methods (methods to be used when data do not follow assumed or conventional distributions)

14 Effective Experiment Design and Data Analysis in Transportation Research

Table 2. Sample bridge inspection data.

Bridge No.	Owner	Location	Bridge Condition Rating			Overall Condition
			Deck	Superstructure	Substructure	
1	State	Rural	8	8	8	ND*
7	Local agency	Rural	6	6	6	FO*
39	State	Urban	6	6	2	SD*
69	State park	Rural	7	5	5	SD
92	City	Urban	5	6	6	ND

*ND = not deficient; FO: functionally obsolete; SD: structurally deficient.

3. **Data Collection:** Data are collected at 100 scheduled locations by bridge inspectors. It is important to note that the bridge condition rating scale is based on subjective categories, and there may be inherent variability among inspectors in their assignment of ratings to bridge components. A sample of data is compiled to document the bridge condition rating of the three primary structural components and the overall condition by location and ownership (Table 2). Notice that the overall condition of a bridge is not necessarily based only on the condition rating of its components (e.g., they cannot just be added).
4. **Specification of Analysis Technique and Data Analysis:** The two primary variables of interest are bridge condition rating and overall condition. The overall condition of the bridge is a categorical variable with three possible values: not deficient; structurally deficient; and functionally obsolete. The frequencies of these values in the given data set are calculated and displayed in the pie chart below. A pie chart provides a visualization of the relative proportions of bridges falling into each category that is often easier to communicate to the reader than a table showing the same information (Figure 1).

   Another way to look at the overall bridge condition variable is by cross-tabulation of the three condition categories with the two location categories (urban and rural), as shown in Table 3. A cross-tabulation provides the joint distribution of two (or more) variables such that each cell represents the frequency of occurrence of a specific combination of possible values. For example, as seen in Table 3, there are 10 structurally deficient bridges in rural areas, which represent 11.4% of all rural area bridges inspected. The numbers in the parentheses are column percentages and add up to 100%. Table 3 also shows that 88 of the bridges inspected were located in rural areas, whereas 12 were located in urban areas.

   The mean values of the bridge condition rating variable for deck, superstructure, and substructure are shown in Table 4. These have been calculated by taking the sum of all the values and then dividing by the total number of cases (100 in this example). Generally, a condition rating

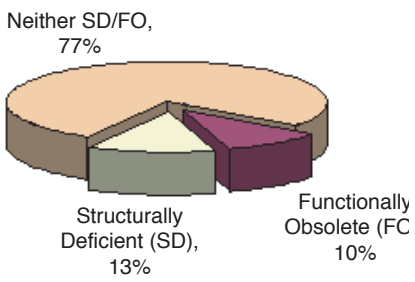

Figure 1. Highway bridge conditions.

Table 3. Cross-tabulation of bridge condition by location.

	Rural	Urban	Total
Structurally deficient	10 (11.4%)	3 (25.0%)	13
Functionally obsolete	6 (6.8%)	4 (33.3%)	10
Not deficient	72 (81.8%)	5 (41.7%)	77
Total	88 (100%)	12 (100%)	100

of 4 or below indicates deficiency in a structural component. For the purpose of comparison, the mean bridge condition rating of the 13 structurally deficient bridges also is provided.

Notice that while the rating scale for the bridge conditions is discrete with values ranging from 0 (failure) to 9 (excellent), the average bridge condition variable is continuous. Therefore, an average score of 6.47 would indicate overall condition of all bridges to be between 6 (satisfactory) and 7 (good). The combined bridge condition rating of deck, superstructure, and substructure is not defined; therefore calculating the mean of the three components' average rating would make no sense. Also, the average bridge condition rating of functionally obsolete bridges is not calculated because other functional characteristics also accounted for this designation.

The distributions of the bridge condition ratings for deck, superstructure, and substructure are shown in Figure 2. Based on the cut-off point of 4, approximately 7% of all bridge decks, 2% of all superstructures, and 5% of all substructures are deficient.

5. **Interpreting the Results:** The results indicate that a majority of bridges (77%) are not structurally or functionally deficient. The inspections were carried out on bridges primarily located in rural areas (88 out of 100). The bridge condition variable may also be cross-tabulated with the ownership variable to determine distribution by jurisdiction. The average condition ratings for the three bridge components for all bridges lies between 6 (satisfactory, some minor problems) and 7 (good, no problems noted).
6. **Conclusion and Discussion:** This example illustrates how to summarize and present quantitative and qualitative data on bridge conditions. It is important to understand the measurement scale of variables in order to interpret the results correctly. Bridge inspection data collected over time may also be analyzed to determine trends in the condition of bridges in a given area. Trend analysis is addressed in Example 15 (structures).
7. **Applications in Other Areas of Transportation Research:** Descriptive statistics could be used to present data in other areas of transportation research, such as:
   - **Transportation Planning—**to assess the distribution of travel times between origin-destination pairs in an urban area. Overall averages could also be calculated.
   - **Traffic Operations—**to analyze the average delay per vehicle at a railroad crossing.

Table 4. Bridge condition ratings.

Rating Category	Mean Value
Overall average bridge condition rating (deck)	6.20
Overall average bridge condition rating (superstructure)	6.47
Overall average bridge condition rating (substructure)	6.08
Average bridge condition rating of structurally deficient bridges (deck)	4.92
Average bridge condition rating of structurally deficient bridges (superstructure)	5.30
Average bridge condition rating of structurally deficient bridges (substructure)	4.54

16 Effective Experiment Design and Data Analysis in Transportation Research

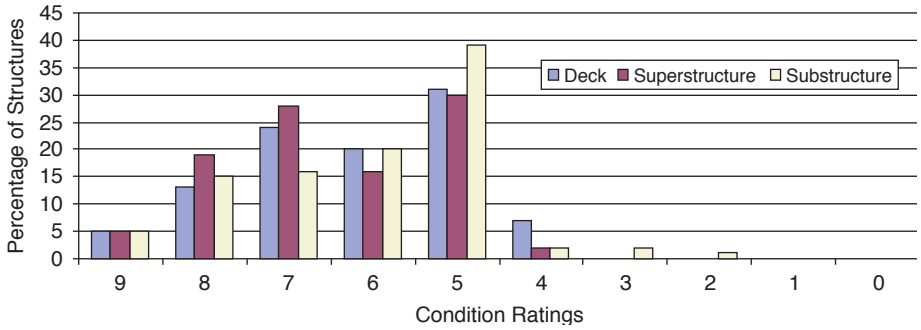

Figure 2. Bridge condition ratings.

- **Traffic Operations/Safety—**to examine the frequency of turning violations at driveways with various turning restrictions.
- **Work Zones, Environment—**to assess the average energy consumption during various stages of construction.

Example 2: Public Transport; Descriptive Statistics

Area: Public transport

Method of Analysis: Descriptive statistics (organizing and presenting data to describe a system or component)

1. **Research Question/Problem Statement:** The manager of a transit agency would like to present information to the board of commissioners on changes in revenue that resulted from a change in the fare. The transit system provides three basic types of service: local bus routes, express bus routes, and demand-responsive bus service. There are 15 local bus routes, 10 express routes, and 1 demand-responsive system.

Question/Issue

Use data to describe some change over time. In this instance, data from 2008 and 2009 are used to describe the change in revenue on each route/part of a transit system when the fare structure was changed from variable (per mile) to fixed fares.

2. **Identification and Description of Variables:** Revenue data are available for each route on the local and express bus system and the demand-responsive system as a whole for the years 2008 and 2009.
3. **Data Collection:** Revenue data were collected on each route for both 2008 and 2009. The annual revenue for the demand-responsive system was also collected. These data are shown in Table 5.
4. **Specification of Analysis Technique and Data Analysis:** The objective of this analysis is to present the impact of changing the fare system in a series of graphs. The presentation is intended to show the impact on each component of the transit system as well as the impact on overall system revenue.

   The impact of the fare change on the overall revenue is best shown with a bar graph (Figure 3).

   The variation in the impact across system components can be illustrated in a similar graph (Figure 4).

   A pie chart also can be used to illustrate the relative impact on each system component (Figure 5).

Table 5. Revenue by route or type of service and year.

Bus Route	2008 Revenue	2009 Revenue
Local Route 1	$350,500	$365,700
Local Route 2	$263,000	$271,500
Local Route 3	$450,800	$460,700
Local Route 4	$294,300	$306,400
Local Route 5	$173,900	$184,600
Local Route 6	$367,800	$375,100
Local Route 7	$415,800	$430,300
Local Route 8	$145,600	$149,100
Local Route 9	$248,200	$260,800
Local Route 10	$310,400	$318,300
Local Route 11	$444,300	$459,200
Local Route 12	$208,400	$205,600
Local Route 13	$407,600	$412,400
Local Route 14	$161,500	$169,300
Local Route 15	$325,100	$340,200
Express Route 1	$85,400	$83,600
Express Route 2	$110,300	$109,200
Express Route 3	$65,800	$66,200
Express Route 4	$125,300	$127,600
Express Route 5	$90,800	$90,400
Express Route 6	$125,800	$123,400
Express Route 7	$87,200	$86,900
Express Route 8	$68.300	$67,200
Express Route 9	$110,100	$112,300
Express Route 10	$73,200	$72,100
Demand-Responsive System	$510,100	$521,300

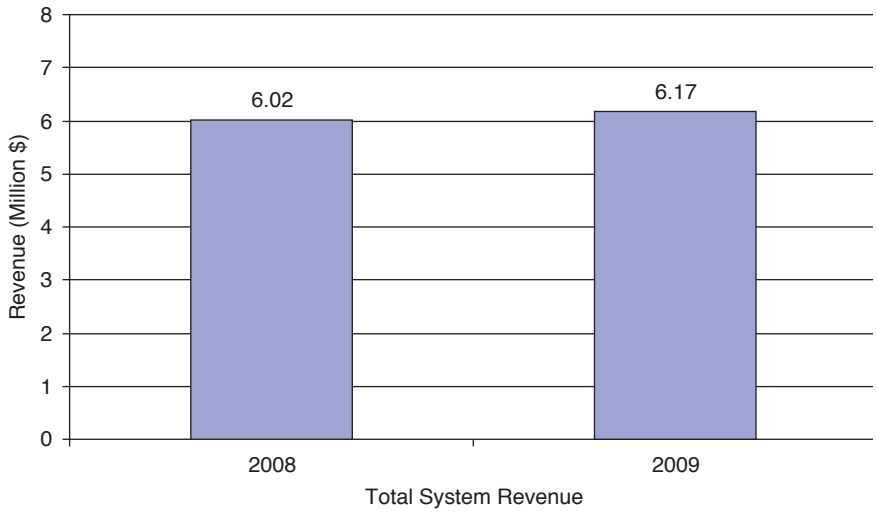

Figure 3. Impact of fare change on overall revenue.

18 Effective Experiment Design and Data Analysis in Transportation Research

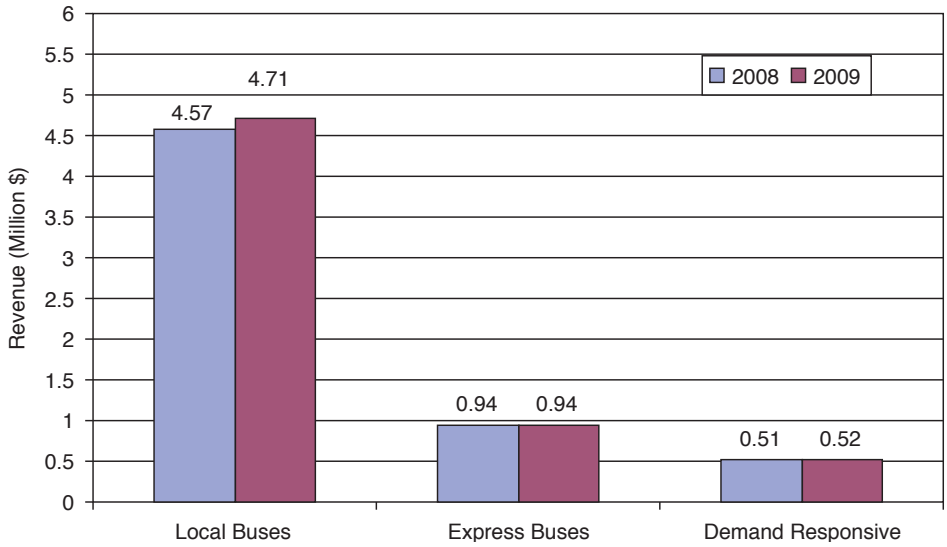

Figure 4. Variation in impact of fare change across system components.

If it is important to display the variability in the impact within the various bus routes in the local bus or express bus operations, this also can be illustrated (Figure 6).

This type of diagram shows the maximum value, minimum value, and mean value of the percent increase in revenue across the 15 local bus routes and the 10 express bus routes.

5. **Interpreting the results:** These results indicate that changing from a variable fare based on trip length (2008) to a fixed fare (2009) on both the local bus routes and the express bus routes had little effect on revenue. On the local bus routes, there was an average increase in revenue of 3.1%. On the express bus routes, there was an average decrease in revenue of 0.4%.

   These changes altered the percentage of the total system revenue attributed to the local bus routes and the express bus routes. The local bus routes generated 76.3% of the revenue in 2009, compared to 75.8% in 2008. The percentage of revenue generated by the express bus routes dropped from 15.7% to 15.2%, and the demand-responsive system generated 8.5% in both 2008 and 2009.

6. **Conclusion and Discussion:** The total revenue increased from $6.02 million to $6.17 million. The cost of operating a variable fare system is greater than that of operating a fixed fare system—hence, net income probably increased even more (more revenue, lower cost for fare collection), and the decision to modify the fare system seems reasonable. Notice that the entire discussion

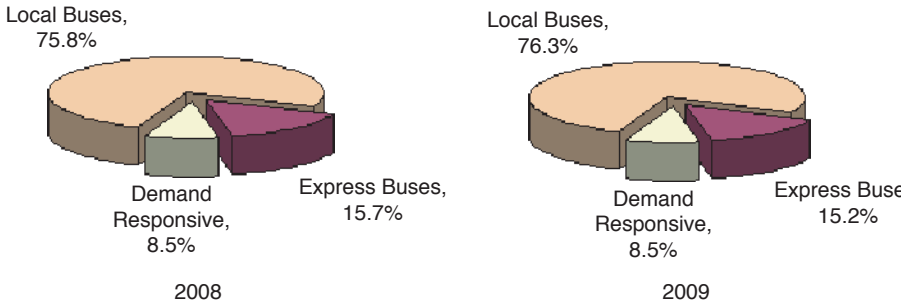

Figure 5. Pie charts illustrating percent of revenue from each component of a transit system.

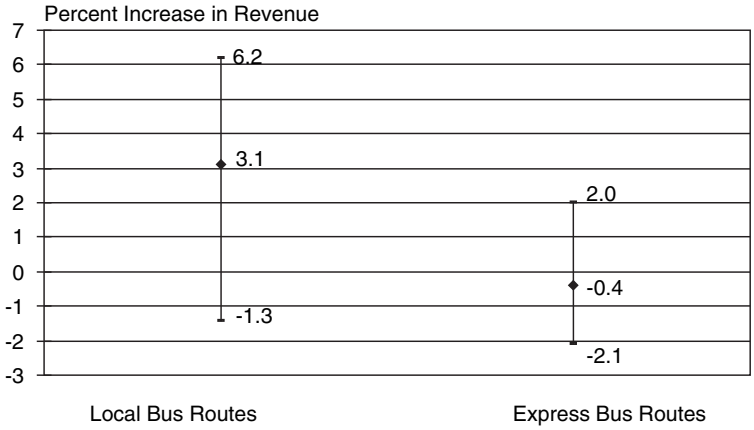

Figure 6. Graph showing variation in revenue increase by type of bus route.

also is based on the assumption that no other factors changed between 2008 and 2009 that might have affected total revenues. One of the implicit assumptions is that the number of riders remained relatively constant from 1 year to the next. If the ridership had changed, the statistics reported would have to be changed. Using the measure revenue/rider, for example, would help control (or normalize) for the variation in ridership.

7. **Applications in Other Areas in Transportation Research:** Descriptive statistics are widely used and can convey a great deal of information to a reader. They also can be used to present data in many areas of transportation research, including:
   - **Transportation Planning—**to display public response frequency or percentage to various alternative designs.
   - **Traffic Operations—**to display the frequency or percentage of crashes by route type or by the type of traffic control devices present at an intersection.
   - **Airport Engineering—**to display the arrival pattern of passengers or flights by hour or other time period.
   - **Public Transit—**to display the average load factor on buses by time of day.

Example 3: Environment; Descriptive Statistics

Area: Environment

Method of Analysis: Descriptive statistics (organizing and presenting data to explain current conditions)

1. **Research Question/Problem Statement:** The planning and programming director in Environmental City wants to determine the current ozone concentration in the city. These data will be compared to data collected after the projects included in the Transportation Improvement Program (TIP) have been completed to determine the effects of these projects on the environment. Because the terrain, the presence of hills or tall buildings, the prevailing wind direction, and the sample station location relative to high volume roads or industrial sites all affect the ozone level, multiple samples are required to determine the ozone concentration level in a city. For this example, air samples are obtained each weekday in the month of July (21 days) at 14 air-sampling stations in the city: 7 in the central city and 7 in the outlying areas of the city. The objective of the analysis is to determine the ozone concentration in the central city, the outlying areas of the city, and the city as a whole.

Question/Issue

Use collected data to describe existing conditions and prepare for future analysis. In this example, air pollution levels in the central city, the outlying areas, and the overall city are to be described.

2. **Identification and Description of Variables:** The variable to be analyzed is the 8-hour average ozone concentration in parts per million (ppm) at each of the 14 air-sampling stations. The 8-hour average concentration is the basis for the EPA standard, and July is selected because ozone levels are temperature sensitive and increase with a rise in the temperature.
3. **Data Collection:** Ozone concentrations in ppm are recorded for each hour of the day at each of the 14 air-sampling stations. The highest average concentration for any 8-hour period during the day is recorded and tabulated. This results in 294 concentration observations (14 stations for 21 days). Table 6 and Table 7 show the data for the seven central city locations and the seven outlying area locations.
4. **Specification of Analysis Technique and Data Analysis:** Much of the data used in analyzing transportation issues has year-to-year, month-to-month, day-to-day, and even hour-to-hour variations. For this reason, making only one observation, or even a few observations, may not accurately describe the phenomenon being observed. Thus, standard practice is to obtain several observations and report the mean value of all observations.

   In this example, the phenomenon being observed is the daily ozone concentration at a series of air-sampling locations. The statistic to be estimated is the mean value of this variable over

Table 6. Central city 8-hour ozone concentration samples (ppm).

Day	Station							
	1	2	3	4	5	6	7	Σ
1	0.079	0.084	0.081	0.083	0.088	0.086	0.089	0.590
2	0.082	0.087	0.088	0.086	0.086	0.087	0.081	0.597
3	0.080	0.081	0.077	0.072	0.084	0.083	0.081	0.558
4	0.083	0.086	0.082	0.079	0.086	0.087	0.089	0.592
5	0.082	0.087	0.080	0.075	0.090	0.089	0.085	0.588
6	0.075	0.084	0.079	0.076	0.080	0.083	0.081	0.558
7	0.078	0.079	0.080	0.074	0.078	0.080	0.075	0.544
8	0.081	0.077	0.082	0.081	0.076	0.079	0.074	0.540
9	0.088	0.084	0.083	0.085	0.083	0.083	0.088	0.594
10	0.085	0.087	0.086	0.089	0.088	0.087	0.090	0.612
11	0.079	0.082	0.082	0.089	0.091	0.089	0.090	0.602
12	0.078	0.080	0.081	0.086	0.088	0.089	0.089	0.591
13	0.081	0.079	0.077	0.083	0.084	0.085	0.087	0.576
14	0.083	0.080	0.079	0.081	0.080	0.082	0.083	0.568
15	0.084	0.083	0.080	0.085	0.082	0.086	0.085	0.585
16	0.086	0.087	0.085	0.087	0.089	0.090	0.089	0.613
17	0.082	0.085	0.083	0.090	0.087	0.088	0.089	0.604
18	0.080	0.081	0.080	0.087	0.085	0.086	0.088	0.587
19	0.080	0.083	0.077	0.083	0.085	0.084	0.087	0.579
20	0.081	0.084	0.079	0.082	0.081	0.083	0.088	0.578
21	0.082	0.084	0.080	0.081	0.082	0.083	0.085	0.577
Σ	1.709	1.744	1.701	1.734	1.773	1.789	1.793	12.243

Table 7. Outlying area 8-hour ozone concentration samples (ppm).

Day	Station							
	8	9	10	11	12	13	14	Σ
1	0.072	0.074	0.073	0.071	0.079	0.070	0.074	0.513
2	0.074	0.075	0.077	0.075	0.081	0.075	0.077	0.534
3	0.070	0.072	0.074	0.074	0.083	0.078	0.080	0.531
4	0.067	0.070	0.071	0.077	0.080	0.077	0.081	0.523
5	0.064	0.067	0.068	0.072	0.079	0.078	0.079	0.507
6	0.069	0.068	0.066	0.070	0.075	0.079	0.082	0.509
7	0.071	0.069	0.070	0.071	0.074	0.071	0.077	0.503
8	0.073	0.072	0.074	0.072	0.076	0.073	0.078	0.518
9	0.072	0.075	0.077	0.074	0.078	0.074	0.080	0.530
10	0.074	0.077	0.079	0.077	0.080	0.076	0.079	0.542
11	0.070	0.072	0.075	0.074	0.079	0.074	0.078	0.522
12	0.068	0.067	0.068	0.070	0.074	0.070	0.075	0.492
13	0.065	0.063	0.067	0.068	0.072	0.067	0.071	0.473
14	0.063	0.062	0.067	0.069	0.073	0.068	0.073	0.475
15	0.064	0.064	0.066	0.067	0.070	0.066	0.070	0.467
16	0.061	0.059	0.062	0.062	0.067	0.064	0.069	0.434
17	0.065	0.061	0.060	0.064	0.069	0.066	0.073	0.458
18	0.067	0.063	0.065	0.068	0.073	0.069	0.076	0.499
19	0.069	0.067	0.068	0.072	0.077	0.071	0.078	0.502
20	0.071	0.069	0.070	0.074	0.080	0.074	0.077	0.515
21	0.070	0.065	0.072	0.076	0.079	0.073	0.079	0.514
Σ	1.439	1.431	1.409	1.497	1.598	1.513	1.606	10.553

the test period selected. The mean value of any data set ($\bar{x}$) equals the sum of all observations in the set divided by the total number of observations in the set (n):

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

The variables of interest stated in the research question are the average ozone concentration for the central city, the outlying areas, and the total city. Thus, there are three data sets: the first table, the second table, and the sum of the two tables. The first data set has a sample size of 147; the second data set also has a sample size of 147, and the third data set contains 294 observations.

Using the formula just shown, the mean value of the ozone concentration in the central city is calculated as follows:

$$\bar{x} = \sum_{i=1}^{147} \frac{x_i}{147} = \frac{12.243}{147} = 0.083 \text{ ppm}$$

The mean value of the ozone concentration in the outlying areas of the city is:

$$\bar{x} = \sum_{i=1}^{147} \frac{x_i}{147} = \frac{10.553}{147} = 0.072 \text{ ppm}$$

The mean value of the ozone concentration for the entire city is:

$$\bar{x} = \sum_{i=1}^{294} \frac{x_i}{294} = \frac{22.796}{294} = 0.078 \text{ ppm}$$

22 Effective Experiment Design and Data Analysis in Transportation Research

Using the same equation, the mean value for each air-sampling location can be found by summing the value of the ozone concentration in the column representing that location and dividing by the 21 observations at that location. For example, considering Sample Station 1, the mean value of the ozone concentration is 1.709/21 = 0.081 ppm.

Similarly, the mean value of the ozone concentrations for any specific day can be found by summing the ozone concentration values in the row representing that day and dividing by the number of stations. For example, for Day 1, the mean value of the ozone concentration in the central city is 0.590/7=0.084. In the outlying areas of the city, it is 0.513/7=0.073, and for the entire city it is 1.103/14=0.079.

The highest and lowest values of the ozone concentration can be obtained by searching the two tables. The highest ozone concentration (0.091 ppm) is logged as having occurred at Station 5 on Day 11. The lowest ozone concentration (0.059 ppm) occurred at Station 9 on Day 16.

The variation by sample location can be illustrated in the form of a frequency diagram. A graph can be used to show the variation in the average ozone concentration for the seven sample stations in the central city (Figure 7).

Notice that all of these calculations (and more) can be done very easily if all the data are put in a spreadsheet and various statistical functions used. Graphs and other displays also can be made within the spreadsheet.

5. **Interpreting the Results:** In this example, the data are not tested to determine whether they fit a known distribution or whether one average value is significantly higher or lower than another. It can only be reported that, as recorded in July, the mean ozone concentration in the central city was greater than the concentration in the outlying areas of the city. (For testing to see whether the data fit a known distribution or comparing mean values, see Example 4 on fitting distributions and goodness of fit. For comparing mean values, see examples 5 through 7.)

   It is known that ozone concentration varies by day and by location of the air-sampling equipment. If there is some threshold value of importance, such as the ozone concentration level considered acceptable by the EPA, these data could be used to determine the number of days that this level was exceeded, or the number of stations that recorded an ozone concentration above this threshold. This is done by comparing each day or each station with the threshold

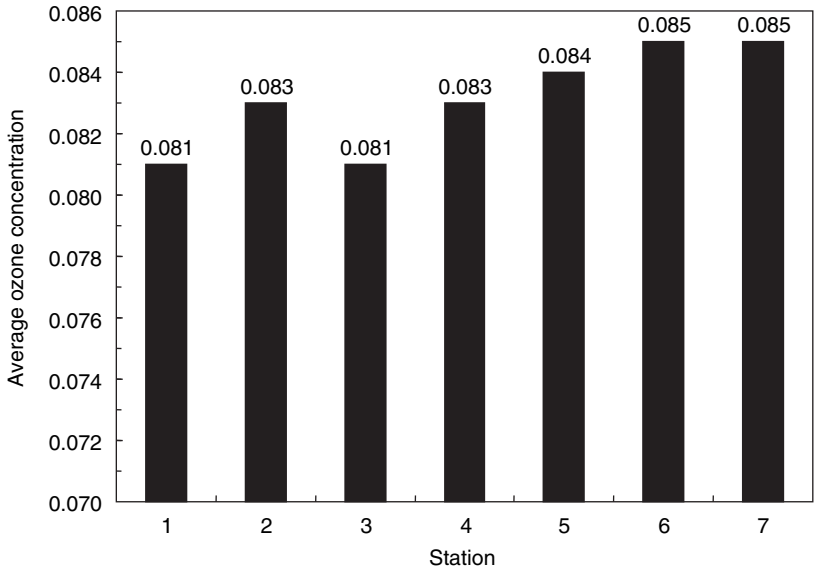

Figure 7. Average ozone concentration for seven central city sampling stations (ppm).

value. It must be noted that, as presented, this example is not a statistical comparison per se (i.e., there has been no significance testing or formal statistical comparison).

6. **Conclusion and Discussion:** This example illustrates how to determine and present quantitative information about a data set containing values of a varying parameter. If a similar set of data were captured each month, the variation in ozone concentration could be analyzed to describe the variation over the year. Similarly, if data were captured at these same locations in July of every year, the trend in ozone concentration over time could be determined.
7. **Applications in Other Areas in Transportation:** These descriptive statistics techniques can be used to present data in other areas of transportation research, such as:
   - **Traffic Operations/Safety and Transportation Planning**
     - to analyze the average speed of vehicles on streets with a speed limit of 45 miles per hour (mph) in residential, commercial, and industrial areas by sampling a number of streets in each of these area types.
     - to examine the average emergency vehicle response time to various areas of the city or county, by analyzing dispatch and arrival times for emergency calls to each area of interest.
   - **Pavement Engineering—**to analyze the average number of potholes per mile on pavement as a function of the age of pavement, by sampling a number of streets where the pavement age falls in discrete categories (0 to 5 years, 5 to 10 years, 10 to 15 years, and greater than 15 years).
   - **Traffic Safety—**to evaluate the average number of crashes per month at intersections with two-way STOP control versus four-way STOP control by sampling a number of intersections in each category over time.

Example 4: Traffic Operations; Goodness of Fit

Area: Traffic operations

Method of Analysis: Goodness of fit (chi-square test; determining if observed distributions of data fit hypothesized standard distributions)

1. **Research Question/Problem Statement:** A research team is developing a model to estimate travel times of various types of personal travel (modes) on a path shared by bicyclists, in-line skaters, and others. One version of the model relies on the assertion that the distribution of speeds for each mode conforms to the *normal distribution.* (For a helpful definition of this and other statistical terms, see the glossary in NCHRP Project 20-45, Volume 2, Appendix A.) Based on a literature review, the researchers are sure that bicycle speeds are normally distributed. However, the shapes of the speed distributions for other users are unknown. Thus, the objective is to determine if skater speeds are normally distributed in this instance.

Question/Issue

Do collected data fit a specific type of probability distribution? In this example, do the speeds of in-line skaters on a shared-use path follow a normal distribution (are they normally distributed)?

2. **Identification and Description of Variables:** The only variable collected is the speed of in-line skaters passing through short sections of the shared-use path.
3. **Data Collection:** The team collects speeds using a video camera placed where most path users would not notice it. The speed of each free-flowing skater (i.e., each skater who is not closely following another path user) is calculated from the times that the skater passes two benchmarks on the path visible in the camera frame. Several days of data collection allow a large sample of 219 skaters to be measured. (An implicit assumption is made that there is no

24 Effective Experiment Design and Data Analysis in Transportation Research

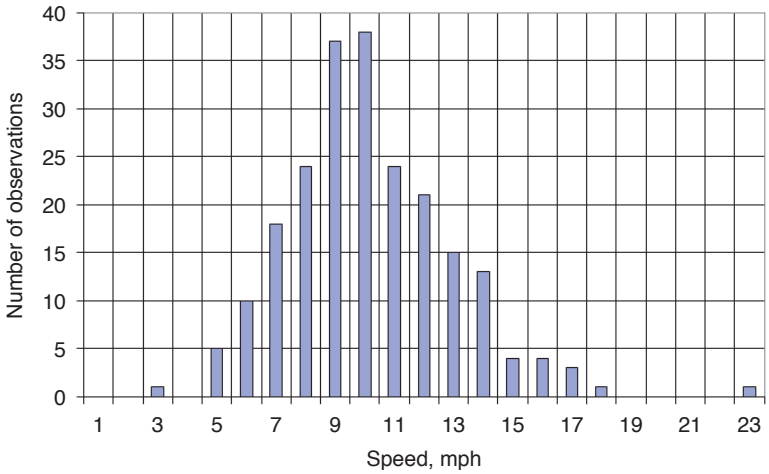

Figure 8. Distribution of observed in-line skater speeds.

variation in the data by day.) The data have a familiar bell shape; that is, when graphed, they look like they are normally distributed (Figure 8). Each bar in the figure shows the number of observations per 1.00-mph-wide speed bin. There are 10 observations between 6.00 mph and 6.99 mph.

4. **Specification of Analysis Technique and Data Analysis:** This analysis involves several preliminary steps followed by two major steps. In the preliminaries, the team calculates the mean and standard deviation from the data sample as 10.17 mph and 2.79 mph, respectively, using standard formulas described in NCHRP Project 20-45, Volume 2, Chapter 6, Section C under the heading "Frequency Distributions, Variance, Standard Deviation, Histograms, and Boxplots." Then the team forms bins of observations of sufficient size to conduct the analysis. For this analysis, the team forms bins containing at least four observations each, which means forming a bin for speeds of 5 mph and lower and a bin for speeds of 17 mph or higher. There is some argument regarding the minimum allowable cell size. Some analysts argue that the minimum is five; others argue that the cell size can be smaller. Smaller numbers of observations in a bin may distort the results. When in doubt, the analysis can be done with different assumptions regarding the cell size. The left two columns in Table 8 show the data ready for analysis.

   The first major step of the analysis is to generate the theoretical normal distribution to compare to the field data. To do this, the team calculates a value of Z, the standard normal variable for each bin i, using the following equation:

$$Z = \frac{x_i - \mu}{\sigma}$$

   where x is the speed in miles per hour (mph) corresponding to the bin, μ is the mean speed, and σ is the standard deviation of all of the observations in the speed sample in mph.

   For example (and with reference to the data in Table 8), for a speed of 5 mph the value of Z will be $(5 - 10.17)/2.79 = -1.85$ and for a speed of 6 mph, the value of Z will be $(6 - 10.17)/2.79 = -1.50$. The team then consults a table of standard normal values (i.e., NCHRP Project 20-45, Volume 2, Appendix C, Table C-1) to convert these Z values into A values representing the area under the standard normal distribution curve. The A value for a Z of -1.85 is 0.468, while the A value for a Z of -1.50 is 0.432. The difference between these two A values, representing the area under the standard normal probability curve corresponding to the speed of 6 mph, is 0.036 (calculated $0.468 - 0.432 = 0.036$). The team multiplies 0.036 by the total sample size (219), to estimate that there should be 7.78 skaters with a speed of 6 mph if the speeds follow the standard normal distribution. The team follows

Table 8. Observations, theoretical predictions, and chi-square values for each bin.

Speed (mph)	Number of Observations	Number Predicted by Normal Distribution	Chi-Square Value
Under 5.99	6	6.98	0.137
6.00 to 6.99	10	7.78	0.637
7.00 to 7.99	18	13.21	1.734
8.00 to 8.99	24	19.78	0.902
9.00 to 9.99	37	26.07	4.585
10.00 to 10.99	38	30.26	1.980
11.00 to 11.99	24	30.93	1.554
12.00 to 12.99	21	27.85	1.685
13.00 to 13.99	15	22.08	2.271
14.00 to 14.99	13	15.42	0.379
15.00 to 15.99	4	9.48	3.169
16.00 to 16.99	4	5.13	0.251
17.00 and over	5	4.03	0.234
Total	**219**	**219**	**19.519**

a similar procedure for all speeds. Notice that the areas under the curve can also be calculated in a simple Excel spreadsheet using the "NORMDIST" function for a given x value and the average speed of 10.17 and standard deviation of 2.79. The values shown in Table 8 have been estimated using the Excel function.

The second major step of the analysis is to use the chi-square test (as described in NCHRP Project 20-45, Volume 2, Chapter 6, Section F) to determine if the theoretical normal distribution is significantly different from the actual data distribution. The team computes a chi-square value for each bin i using the formula:

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the number of actual observations in bin i and E_i is the expected number of observations in bin i estimated by using the theoretical distribution.

For the bin of 6 mph speeds, $O = 10$ (from the table), $E = 7.78$ (calculated), and the χ_i^2 contribution for that cell is 0.637. The sum of the χ_i^2 values for all bins is 19.519. The degrees of freedom (df) used for this application of the chi-square test are the number of bins minus 1 minus the number of variables in the distribution of interest. Given that the normal distribution has two variables (*see* May, *Traffic Flow Fundamentals,* 1990, p. 40), in this example the degrees of freedom equal 9 (calculated $12 - 1 - 2 = 9$).

From a standard table of chi-square values (NCHRP Project 20-45, Volume 2, Appendix C, Table C-2), the team finds that the critical value at the 95% confidence level for this case (with df = 9) is 16.9. The calculated value of the statistic is ~19.5, more than the tabular value.

The results of all of these observations and calculations are shown in Table 8.

5. **Interpreting the Results:** The calculated chi-square value of ~19.5 is greater than the critical chi-square value of 16.9. The team concludes, therefore, that the normal distribution is significantly different from the distribution of the speed sample at the 95% level (i.e., that the in-line skater speed data do not appear to be normally distributed). Larger variations between the observed and expected distributions lead to higher values of the statistic and would be interpreted as it being less likely that the data are distributed according to the

26 Effective Experiment Design and Data Analysis in Transportation Research

hypothesized distribution. Conversely, smaller variations between observed and expected distributions result in lower values of the statistic, which would suggest that it is more likely that the data are normally distributed because the observed values would fit better with the expected values.

6. **Conclusion and Discussion:** In this case, the results suggest that the normal distribution is not a good fit to free-flow speeds of in-line skaters on shared-use paths. Interestingly, if the 23 mph observation is considered to be an outlier and discarded, the results of the analysis yield a different conclusion (that the data are normally distributed). Some researchers use a simple rule that an outlier exists if the observation is more than three standard deviations from the mean value. (In this example, the 23 mph observation is, indeed, more than three standard deviations from the mean.) If there is concern with discarding the observation as an outlier, it would be easy enough in this example to repeat the data collection exercise.

   Looking at the data plotted above, it is reasonably apparent that the well-known normal distribution should be a good fit (at least without the value of 23). However, the results from the statistical test could not confirm the suspicion. In other cases, the type of distribution may not be so obvious, the distributions in question may be obscure, or some distribution parameters may need to be calibrated for a good fit. In these cases, the statistical test is much more valuable.

   The chi-square test also can be used simply to compare two observed distributions to see if they are the same, independent of any underlying probability distribution. For example, if it is desired to know if the distribution of traffic volume by vehicle type (e.g., automobiles, light trucks, and so on) is the same at two different freeway locations, the two distributions can be compared to see if they are similar.

   The consequences of an error in the procedure outlined here can be severe. This is because the distributions chosen as a result of the procedure often become the heart of predictive models used by many other engineers and planners. A poorly-chosen distribution will often provide erroneous predictions for many years to come.

7. **Applications in Other Areas of Transportation Research:** Fitting distributions to data samples is important in several areas of transportation research, such as:
   - **Traffic Operations—**to analyze shapes of vehicle headway distributions, which are of great interest, especially as a precursor to calibrating and using simulation models.
   - **Traffic Safety—**to analyze collision frequency data. Analysts often assume that the Poisson distribution is a good fit for collision frequency data and must use the method described here to validate the claim.
   - **Pavement Engineering—**to form models of pavement wear or otherwise compare results obtained using different designs, as it is often required to check the distributions of the parameters used (e.g., roughness).

Example 5: Construction; Simple Comparisons to Specified Values

Area: Construction

Method of Analysis: Simple comparisons to specified values—using Student's *t*-test to compare the mean value of a small sample to a standard or other requirement (i.e., to a population with a known mean and unknown standard deviation or variance)

1. **Research Question/Problem Statement:** A contractor wants to determine if a specified soil compaction can be achieved on a segment of the road under construction by using an on-site roller or if a new roller must be brought in.

The cost of obtaining samples for many construction materials and practices is quite high. As a result, decisions often must be made based on a small number of samples. The appropriate statistical technique for comparing the mean value of a small sample with a standard or requirement is Student's *t*-test.

Formally, the working, or null, hypothesis (H_o) and the alternative hypothesis (H_a) can be stated as follows:

H_o: The soil compaction achieved using the on-site roller (C_A) is less than a specified value (C_S); that is, ($C_A < C_S$).

H_a: The soil compaction achieved using the on-site roller (C_A) is greater than or equal to the specified value (C_S); that is, ($C_A \geq C_S$).

Question/Issue

Determine whether a sample mean exceeds a specified value. Alternatively, determine the probability of obtaining a sample mean ($\bar{x}$) from a sample of size *n,* if the universe being sampled has a true mean less than or equal to a population mean with an unknown variance. In this example, is an observed mean of soil compaction samples equal to or greater than a specified value?

2. **Identification and Description of Variables:** The variable to be used is the soil density results of nuclear densometer tests. These values will be used to determine whether the use of the on-site roller is adequate to meet the contract-specified soil density obtained in the laboratory (Proctor density) of 95%.
3. **Data Collection:** A 125-foot section of road is constructed and compacted with the on-site roller, and four samples of the soil density are obtained (25 feet, 50 feet, 75 feet, and 100 feet from the beginning of the test section).
4. **Specification of Analysis Technique and Data Analysis:** For small samples ($n < 30$) where the population mean is known but the population standard deviation is unknown, it is not appropriate to describe the distribution of the sample mean with a normal distribution. The appropriate distribution is called Student's distribution (*t*-distribution or *t*-statistic).

   The equation for Student's *t*-statistic is:

$$t = \frac{\bar{x} - \bar{x}'}{\frac{S}{\sqrt{n}}}$$

where $\bar{x}$ is the sample mean,
$\bar{x}'$ is the population mean (or specified standard),
S is the sample standard deviation, and
n is the sample size.

The four nuclear densometer readings were 98%, 97%, 93% and 99%. Then, showing some simple sample calculations,

$$\bar{X} = \sum_{i=1}^{4} \frac{X_i}{4} = \frac{98 + 97 + 93 + 99}{4} = \frac{387}{4} = 96.75\%$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

$$S = \sqrt{\frac{20.74}{3}} = 2.63\%$$

and using the equation for t above,

$$t = \frac{96.75 - 95.00}{\frac{2.63}{2}} = \frac{1.75}{1.32} = 1.33$$

The calculated value of the t-statistic (1.33) is most typically compared to the tabularized values of the t-statistic (e.g., NCHRP Project 20-45, Volume 2, Appendix C, Table C-4) for a given significance level (typically called *t critical* or t_{crit}). For a sample size of $n = 4$ having 3 $(n - 1)$ degrees of freedom (df), the values for t_{crit} are: 1.638 for $\alpha = 0.10$ and 2.353 for $\alpha = 0.05$ (two common values of α for testing, the latter being most common).

Important: The specification of the significance level (α level) for testing should be done before actual testing and interpretation of results are done. In many instances, the appropriate level is defined by the agency doing the testing, a specified testing standard, or simply common practice. Generally speaking, selection of a smaller value for α (e.g., $\alpha = 0.05$ versus $\alpha = 0.10$) sets a more stringent standard.

In this example, because the calculated value of t (1.33) is less than the critical value (2.353, given $\alpha = 0.05$), the null hypothesis is accepted. That is, the engineer cannot be confident that the mean value from the densometer tests (96.75%) is greater than the required specification (95%). If a lower confidence level is chosen (e.g., $\alpha = 0.15$), the value for t_{crit} would change to 1.250, which means the null hypothesis would be rejected. A lower confidence level can have serious implications. For example, there is an approximately 15% chance that the standard will not be met. That level of risk may or may not be acceptable to the contractor or the agency. Notice that in many standards the required significance level is stated (typically $\alpha = 0.05$).

It should be emphasized that the confidence level should be chosen before calculations and testing are done. It is not generally permissible to change the confidence level after calculations have been performed. Doing this would be akin to arguing that standards can be relaxed if a test gives an answer that the analyst doesn't like.

The results of small sample tests often are sensitive to the number of samples that can be obtained at a reasonable cost. (The mean value may change considerably as more data are added.) In this example, if it were possible to obtain nine independent samples (as opposed to four) and the mean value and sample standard deviation were the same as with the four samples, the calculation of the t-statistic would be:

$$t = \frac{96.75 - 95.00}{\frac{2.63}{3}} = 1.99$$

Comparing the value of t (with a larger sample size) to the appropriate t_{crit} (for $n - 1 = 8$ df and $\alpha = 0.05$) of 1.860 changes the outcome. That is, the calculated value of the t-statistic is now larger than the tabularized value of t_{crit}, and the null hypothesis is rejected. Thus, it is accepted that the mean of the densometer readings meets or exceeds the standard. It should be noted, however, that the inclusion of additional tests may yield a different mean value and standard deviation, in which case the results could be different.

5. **Interpreting the Results:** By themselves, the results of the statistical analysis are insufficient to answer the question as to whether a new roller should be brought to the project site. These results only provide information the contractor can use to make this decision. The ultimate decision should be based on these probabilities and knowledge of the cost of each option. What is the cost of bringing in a new roller now? What is the cost of starting the project and then determining the current roller is not adequate and then bringing in a new roller? Will this decision result in a delay in project completion—and does the contract include an incentive for early completion and/or a penalty for missing the completion date? If it is possible to conduct additional independent densometer tests, what is the cost of conducting them?

If there is a severe penalty for missing the deadline (or a significant reward for finishing early), the contractor may be willing to incur the cost of bringing in a new roller rather than accepting a 15% probability of being delayed.

6. **Conclusion and Discussion:** In some cases the decision about which alternative is preferable can be expressed in the form of a probability (or level of confidence) required to make a decision. The decision criterion is then expressed in a hypothesis and the probability of rejecting that hypothesis. In this example, if the hypothesis to be tested is "Using the on-site roller will provide an average soil density of 95% or higher" and the level of confidence is set at 95%, given a sample of four tests the decision will be to bring in a new roller. However, if nine independent tests could be conducted, the results in this example would lead to a decision to use the on-site roller.
7. **Applications in Other Areas in Transportation Research:** Simple comparisons to specified values can be used in a variety of areas of transportation research. Some examples include:
   - **Traffic Operations—**to compare the average annual number of crashes at intersections with roundabouts with the average annual number of crashes at signalized intersections.
   - **Pavement Engineering—**to test the comprehensive strength of concrete slabs.
   - **Maintenance—**to test the results of a proposed new deicer compound.

Example 6: Maintenance; Simple Two-Sample Comparisons

Area: Maintenance

Method of Analysis: Simple two-sample comparisons (*t*-test for paired comparisons; comparing the mean values of two sets of matched data)

1. **Research Question/Problem Statement:** As a part of a quality control and quality assurance (QC/QA) program for highway maintenance and construction, an agency engineer wants to compare and identify discrepancies in the contractor's testing procedures or equipment in making measurements on materials being used. Specifically, compacted air voids in asphalt mixtures are being measured. In this instance, the agency's test results need to be compared, one-to-one, with the contractor's test results. Samples are drawn or made and then literally split and tested—one by the contractor, one by the agency. Then the pairs of measurements are analyzed. A paired *t*-test will be used to make the comparison. (For another type of two-sample comparison, see Example 7.)

Question/Issue

Use collected data to test if two sets of results are similar. Specifically, do two testing procedures to determine air voids produce the same results?

Stated in formal terms, the null and alternative hypotheses are:

H_o: There is no mean difference in air voids between agency and contractor test results:

$H_o : \bar{X}_d = 0$

H_a: There is a mean difference in air voids between agency and contractor test results:

$H_a : \bar{X}_d \neq 0$

(For definitions and more discussion about the formulation of formal hypotheses for testing, see NCHRP Project 20-45, Volume 2, Appendix A and Volume 1, Chapter 2, "Hypothesis.")

2. **Identification and Description of Variables:** The testing procedure for laboratory-compacted air voids in the asphalt mixture needs to be verified. The split-sample test results for laboratory-

30 Effective Experiment Design and Data Analysis in Transportation Research

Table 9. Laboratory-compacted air voids in split samples.

Sample	Air Voids (%)		Difference
	Contractor	Agency	
1	4.37	4.15	0.21
2	3.76	5.39	-1.63
3	4.10	4.47	-0.37
4	4.39	4.52	-0.13
5	4.06	5.36	-1.29
6	4.14	5.01	-0.87
7	3.92	5.23	-1.30
8	3.38	4.97	-1.60
9	4.12	4.37	-0.25
10	3.68	5.29	-1.61
$\bar{X}$	3.99	4.88	$\bar{X}_d$ = -0.88
S	0.31	0.46	s_d = 0.70

compacted air voids are shown in Table 9. Twenty samples are prepared using the same asphalt mixture. Half of the samples are prepared in the agency's laboratory and the other half in the contractor's laboratory. Given this arrangement, there are basically two variables of concern: who did the testing and the air void determination.

3. **Data Collection:** A sufficient quantity of asphalt mix to make 10 lots is produced in an asphalt plant located on a highway project. Each of the 10 lots is collected, split into two samples, and labeled. A sample from each lot, 4 inches in diameter and 2 inches in height, is prepared in the contractor's laboratory to determine the air voids in the compacted samples. A matched set of samples is prepared in the agency's laboratory and a similar volumetric procedure is used to determine the agency's lab-compacted air voids. The lab-compacted air void contents in the asphalt mixture for both the contractor and agency are shown in Table 9.
4. **Specification of Analysis Technique and Data Analysis:** A paired (two-sided) *t*-test will be used to determine whether a difference exists between the contractor and agency results. As noted above, in a paired *t*-test the null hypothesis is that the mean of the differences between each pair of two tests is 0 (there is no difference between the means). The null hypothesis can be expressed as follows:

$$H_o : \bar{X}_d = 0$$

The alternate hypothesis, that the two means are not equal, can be expressed as follows:

$$H_a : \bar{X}_d \neq 0$$

The *t*-statistic for the paired measurements (i.e., the difference between the split-sample test results) is calculated using the following equation:

$$t = \frac{|\bar{X}_d - 0|}{\frac{s_d}{\sqrt{n}}}$$

Using the actual data, the value of the *t*-statistic is calculated as follows:

$$t = \frac{|0.88 - 0|}{\frac{0.7}{\sqrt{10}}} = 4$$

For $n - 1$ $(10 - 1 = 9)$ degrees of freedom and $\alpha = 0.05$, the t_{crit} value can be looked up using a *t*-table (e.g., NCHRP Project 20-45, Volume 2, Appendix C, Table C-4):

$t_{0.025, 9} = 2.262$

For a more detailed description of the *t*-statistic, see the glossary in NCHRP Project 20-45, Volume 2, Appendix A.

5. **Interpreting the Results:** Given that $t = 4 > t_{0.025, 9} = 2.685$, the engineer would reject the null hypothesis and conclude that the results of the paired tests are different. This means that the contractor and agency test results from paired measurements indicate that the test method, technicians, and/or test equipment are not providing similar results. Notice that the engineer cannot conclude anything about the material or production variation or what has caused the differences to occur.
6. **Conclusion and Discussion:** The results of the test indicate that a statistically significant difference exists between the test results from the two groups. When making such comparisons, it is important that random sampling be used when obtaining the samples. Also, because sources of variability influence the population parameters, the two sets of test results must have been sampled over the same time period, and the same sampling and testing procedures must have been used. It is best if one sample is drawn and then literally split in two, then another sample drawn, and so on. The identification of a difference is just that: notice that a difference exists. The reason for the difference must still be determined.

   A common misinterpretation is that the result of the *t*-test provides the probability of the null hypothesis being true. Another way to look at the *t*-test result in this example is to conclude that some alternative hypothesis provides a better description of the data. The result does not, however, indicate that the alternative hypothesis is true.

   To ensure practical significance, it is necessary to assess the magnitude of the difference being tested. This can be done by computing confidence intervals, which are used to quantify the range of effect size and are often more useful than simple hypothesis testing.

   Failure to reject a hypothesis also provides important information. Possible explanations include: occurrence of a type-II error (erroneous acceptance of the null hypothesis); small sample size; difference too small to detect; expected difference did not occur in data; there is no difference/effect. Proper experiment design and data collection can minimize the impact of some of these issues. (For a more comprehensive discussion of this topic, see NCHRP Project 20-45, Volume 2, Chapter 1.)
7. **Applications in Other Areas of Transportation Research:** The application of the *t*-test to compare two mean values in other areas of transportation research may include:
   - **Traffic Operations**—to evaluate average delay in bus arrivals at various bus stops.
   - **Traffic Operations/Safety**—to determine the effect of two enforcement methods on reduction in a particular traffic violation.
   - **Pavement Engineering**—to investigate average performance of two pavement sections.
   - **Environment**—to compare average vehicular emissions at two locations in a city.

Example 7: Materials; Simple Two-Sample Comparisons

Area: Materials

Method of Analysis: Simple two-sample comparisons (using the *t*-test to compare the mean values of two samples and the *F*-test for comparing variances)

1. **Research Question/Problem Statement:** As a part of dispute resolution during quality control and quality assurance, a highway agency engineer wants to validate a contractor's test results concerning asphalt content. In this example, the engineer wants to compare the results

of two sets of tests: one from the contractor and one from the agency. Formally, the (null) hypothesis to be tested, H_o, is that the contractor's tests and the agency's tests are from the same population. In other words, the null hypothesis is that the means of the two data sets will be equal, as will the standard deviations. Notice that in the latter instance the variances are actually being compared.

Test results were also compared in Example 6. In that example, the comparison was based on split samples. The same test specimens were tested by two different analysts using different equipment to see if the same results could be obtained by both. The major difference between Example 6 and Example 7 is that, in this example, the two samples are randomly selected from the same pavement section.

Question/Issue

Use collected data to test if two measured mean values are the same. In this instance, are two mean values of asphalt content the same? Stated in formal terms, the null and alternative hypotheses can be expressed as follows:

H_o: There is no difference in asphalt content between agency and contractor test results:

$(H_o: m_c - m_a = 0)$

H_a: There is a difference in asphalt content between agency and contractor test results:

$(H_a: m_c - m_a \neq 0)$

2. **Identification and Description of Variables:** The contractor runs 12 asphalt content tests and the agency engineer runs 6 asphalt content tests over the same period of time, using the same random sampling and testing procedures. The question is whether it is likely that the tests have come from the same population based on their variability.
3. **Data Collection:** If the agency's objective is simply to identify discrepancies in the testing procedures or equipment, then verification testing should be done on split samples (as in Example 6). Using split samples, the difference in the measured variable can more easily be attributed to testing procedures. A paired *t*-test should be used. (For more information, see NCHRP Project 20-45, Volume 2, Chapter 4, Section A, "Analysis of Variance Methodology.") A split sample occurs when a physical sample (of whatever is being tested) is drawn and then literally split into two testable samples.

   On the other hand, if the agency's objective is to identify discrepancies in the overall material, process, sampling, and testing processes, then validation testing should be done on independent samples. Notice the use of these terms. It is important to distinguish between testing to verify only the testing process (verification) versus testing to compare the overall production, sampling, and testing processes (validation). If independent samples are used, the agency test results still can be compared with contractor test results (using a simple *t*-test for comparing two means). If the test results are consistent, then the agency and contractor tests can be combined for contract compliance determination.
4. **Specification of Analysis Technique and Data Analysis:** When comparing the two data sets, it is important to compare both the means and the variances because the assumption when using the *t*-test requires equal variances for each of the two groups. A different test is used in each instance. The *F*-test provides a method for comparing the *variances* (the standard deviation squared) of two sets of data. Differences in means are assessed by the *t*-test. Generally, construction processes and material properties are assumed to follow a normal distribution.

In this example, a normal distribution is assumed. (The assumption of normality also can be tested, as in Example 4.) The ratios of variances follow an *F*-distribution, while the means of relatively small samples follow a *t*-distribution. Using these distributions, hypothesis tests can be conducted using the same concepts that have been discussed in prior examples. (For more information about the *F*-test and the *t*-distribution, see NCHRP Project 20-45, Volume 2, Chapter 4, Section A, "Compute the F-ratio Test Statistic." For more information about the *t*-distribution, see NCHRP Project 20-45, Volume 2, Chapter 4, Section A.)

For samples from the same normal population, the statistic *F* (the ratio of the two-sample variances) has a sampling distribution called the *F*-distribution. For validation and verification testing, the *F*-test is based on the ratio of the sample variance of the contractor's test results (s_c^2) and the sample variance of the agency's test results (s_a^2). Similarly, the *t*-test can be used to test whether the sample mean of the contractor's tests, $\bar{X}_c$, and the agency's tests, $\bar{X}_a$, came from populations with the same mean.

Consider the asphalt content test results from the contractor samples and agency samples (Table 10).

In this instance, the *F*-test is used to determine whether the variance observed for the contractor's tests differs from the variance observed for the agency's tests.

Using the *F*-test

Step 1. Compute the variance (s^2), for each set of tests: $s_c^2 = 0.064$ and $s_a^2 = 0.092$. As an example, s_c^2 can be calculated as:

$$s_c^2 = \frac{\sum_i \left(x_i - \bar{X}_c\right)^2}{n-1} = \frac{(6.4-6.1)^2}{11} + \frac{(6.2-6.1)^2}{11} + \cdots + \frac{(6-6.1)^2}{11} + \frac{(5.7-6.1)^2}{11} = 0.0645$$

Step 2. Compute $F_{calc} = \frac{s_a^2}{s_c^2} = \frac{0.092}{0.064} = 1.43$.

Table 10. Asphalt content test results from independent samples.

Contractor Samples		Agency Samples	
1	6.4	1	5.4
2	6.2	2	5.8
3	6.0	3	6.2
4	6.6	4	5.4
5	6.1	5	5.6
6	6.0	6	5.8
7	6.3		
8	6.1		
9	5.9		
10	5.8		
11	6.0		
12	5.7		
Descriptive Statistics	$\bar{X}_c = 6.1$ $s_c^2 = 0.064$ $s_c = 0.25$ $n_c = 12$	Descriptive Statistics	$\bar{X}_a = 5.7$ $s_a^2 = 0.092$ $s_a = 0.30$ $n_a = 6$

34 Effective Experiment Design and Data Analysis in Transportation Research

Step 3. Determine F_{crit} from the *F*-distribution table, making sure to use the correct degrees of freedom (df) for the numerator (the number of observations minus 1, or $n_a - 1 = 6 - 1 = 5$) and the denominator ($n_c - 1 = 12 - 1 = 11$). For $\alpha = 0.01$, $F_{crit} = 5.32$. The critical *F*-value can be found from tables (see NCHRP Project 20-45, Volume 2, Appendix C, Table C-5). Read the *F*-value for $1 - \alpha = 0.99$, numerator and denominator degrees of freedom 5 and 11, respectively. Interpolation can be used if exact degrees of freedom are not available in the table. Alternatively, a statistical function in Microsoft Excel™ can be used to determine the *F*-value.

Step 4. Compare the two values to determine if $F_{calc} < F_{crit}$. If $F_{calc} < F_{crit}$ is true, then the variances are equal; if not, they are unequal. In this example, F_{calc} (1.43) is, in fact, less than F_{crit} (5.32) and, thus, there is no evidence of unequal variances.

Given this result, the *t*-test for the case of equal variances is used to determine whether to declare that the mean of the contractor's tests differs from the mean of the agency's tests.

Using the *t*-test

Step 1. Compute the sample means ($\bar{X}$) for each set of tests: $\bar{X}_c = 6.1$ and $\bar{X}_a = 5.7$.

Step 2. Compute the pooled variance s_p^2 from the individual sample variances:

$$s_p^2 = \frac{s_c^2(n_c - 1) + s_a^2(n_a - 1)}{n_c + n_a - 2} = \frac{0.064(12 - 1) + 0.092(6 - 1)}{12 + 6 - 2} = 0.0731$$

Step 3. Compute the *t*-statistic using the following equation for equal variance:

$$t = \frac{|\bar{X}_c - \bar{X}_a|}{\sqrt{\frac{s_p^2}{n_c} + \frac{s_p^2}{n_a}}} = \frac{|6.1 - 5.7|}{\sqrt{\frac{0.0731}{12} + \frac{0.0731}{6}}} = 2.9$$

$$t_{0.005,16} = 2.921$$

(For more information, see NCHRP Project 20-45, Volume 2, Appendix C, Table C-4 for $A = 1 - \frac{\alpha}{2}$ and $\nu = 16$.)

5. **Interpreting the Results:** Given that $F < F_{crit}$ (i.e., 1.43 < 5.32), there is no reason to believe that the two sets of data have different variances. That is, they could have come from the same population. Therefore, the *t*-test can be used to compare the means using equal variance. Because $t < t_{crit}$ (i.e., 2.9 < 2.921), the engineer does not reject the null hypothesis and, thus, assumes that the sample means are equal. The final conclusion is that it is likely that the contractor and agency test results represent the same process. In other words, with a 99% confidence level, it can be said that the agency's test results are not different from the contractor's and therefore validate the contractor tests.
6. **Conclusion and Discussion:** The simple *t*-test can be used to validate the contractor's test results by conducting independent sampling from the same pavement at the same time. Before conducting a formal *t*-test to compare the sample means, the assumption of equal variances needs to be evaluated. This can be accomplished by comparing sample variances using the *F*-test. The interpretation of results will be misleading if the equal variance assumption is not validated. If the variances of two populations being compared for their means are different, the mean comparison will reflect the difference between two separate populations. Finally, based on the comparison of means, one can conclude that the construction materials have consistent properties as validated by two independent sources (contractor and agency).

   This sort of comparison is developed further in Example 8, which illustrates tests for the equality of more than two mean values.

7. **Applications in Other Areas of Transportation Research:** The simple *t*-test can be used to compare means of two independent samples. Applications for this method in other areas of transportation research may include:
   - **Traffic Operations**
     - to compare average speeds at two locations along a route.
     - to evaluate average delay times at two intersections in an urban area.
   - **Pavement Engineering—**to investigate the difference in average performance of two pavement sections.
   - **Maintenance—**to determine the effects of two maintenance treatments on average life extension of two pavement sections.

Example 8: Laboratory Testing/Instrumentation; Simple Analysis of Variance (ANOVA)

Area: Laboratory testing and/or instrumentation

Method of Analysis: Simple analysis of variance (ANOVA) comparing the mean values of more than two samples and using the *F*-test

1. **Research Question/Problem Statement:** An engineer wants to test and compare the compressive strength of five different concrete mix designs that vary in coarse aggregate type, gradation, and water/cement ratio. An experiment is conducted in a laboratory where five different concrete mixes are produced based on given specifications, and tested for compressive strength using the ASTM International standard procedures. In this example, the comparison involves inference on parameters from more than two populations. The purpose of the analysis, in other words, is to test whether all mix designs are similar to each other in mean compressive strength or whether some differences actually exist. ANOVA is the statistical procedure used to test the basic hypothesis illustrated in this example.

Question/Issue

Compare the means of more than two samples. In this instance, compare the compressive strengths of five concrete mix designs with different combinations of aggregates, gradation, and water/cement ratio. More formally, test the following hypotheses:

H_o: There is no difference in mean compressive strength for the various (five) concrete mix types.
H_a: At least one of the concrete mix types has a different compressive strength.

2. **Identification and Description of Variables:** In this experiment, the factor of interest (independent variable) is the concrete mix design, which has five levels based on different coarse aggregate types, gradation, and water/cement ratios (denoted by *t* and labeled A through E in Table 11). Compressive strength is a continuous response (dependent) variable, measured in pounds per square inch (psi) for each specimen. Because only one factor is of interest in this experiment, the statistical method illustrated is often called a one-way ANOVA or simple ANOVA.
3. **Data Collection:** For each of the five mix designs, three replicates each of cylinders 4 inches in diameter and 8 inches in height are made and cured for 28 days. After 28 days, all 15 specimens are tested for compressive strength using the standard ASTM International test. The compressive strength data and summary statistics are provided for each mix design in Table 11. In this example, resource constraints have limited the number of replicates for each mix design to

36 Effective Experiment Design and Data Analysis in Transportation Research

Table 11. Concrete compressive strength (psi) after 28 days.

Replicate	Mix Design				
	A	B	C	D	E
1	y_{11} = 5416	y_{21} = 5292	y_{31} = 4097	y_{41} = 5056	y_{51} = 4165
2	y_{12} = 5125	y_{22} = 4779	y_{32} = 3695	y_{42} = 5216	y_{52} = 3849
3	y_{13} = 4847	y_{23} = 4824	y_{33} = 4109	y_{43} = 5235	y_{53} = 4089
Mean	$\bar{y}_{1.}$ = 5129	$\bar{y}_{2.}$ = 4965	$\bar{y}_{3.}$ = 3967	$\bar{y}_{4.}$ = 5169	$\bar{y}_{5.}$ = 4034
Standard deviation	s_1 = 284.52	s_2 = 284.08	s_3 = 235.64	s_4 = 98.32	s_5 = 164.94
Overall mean			$\bar{y}_{..}$ = 4653		

three. (For a discussion on sample size determination based on statistical power requirements, see NCHRP Project 20-45, Volume 2, Chapter 1, "Sample Size Determination.")

4. **Specification of Analysis Technique and Data Analysis:** To perform a one-way ANOVA, preliminary calculations are carried out to compute the overall mean ($\bar{y}_{..}$), the sample means ($\bar{y}_{i.}$), and the sample variances (s_i^2) given the total sample size ($n_T = 15$) as shown in Table 11. The basic strategy for ANOVA is to compare the variance between levels or groups—specifically, the variation between sample means—to the variance within levels. This comparison is used to determine if the levels explain a significant portion of the variance. (Details for performing a one-way ANOVA are given in NCHRP Project 20-45, Volume 2, Chapter 4, Section A, "Analysis of Variance Methodology.")

   ANOVA is based on partitioning of the total sum of squares (TSS, a measure of overall variability) into within-level and between-levels components. The TSS is defined as the sum of the squares of the differences of each observation (y_{ij}) from the overall mean ($\bar{y}_{..}$). The TSS, between-levels sum of squares (SSB), and within-level sum of squares (SSE) are computed as follows.

$$TSS = \sum_{i,j} \left(y_{ij} - \bar{y}_{..} \right)^2 = 4839620.90$$

$$SSB = \sum_{i,j} \left(\bar{y}_{i.} - \bar{y}_{..} \right)^2 = 4331513.60$$

$$SSE = \sum_{i,j} \left(y_{ij} - \bar{y}_{i.} \right)^2 = 508107.30$$

   The next step is to compute the between-levels mean square (MSB) and within-levels mean square (MSE) based on respective degrees of freedom (df). The total degrees of freedom (df_T), between-levels degrees of freedom (df_B), and within-levels degrees of freedom (df_E) for one-way ANOVA are computed as follows:

$$df_T = n_T - 1 = 15 - 1 = 14$$

$$df_B = t - 1 = 5 - 1 = 4$$

$$df_E = n_T - t = 15 - 5 = 10$$

where n_T = the total sample size and
t = the total number of levels or groups.

   The next step of the ANOVA procedure is to compute the *F*-statistic. The *F*-statistic is the ratio of two variances: the variance due to interaction between the levels, and the variance due to differences within the levels. Under the null hypothesis, the between-levels mean square (MSB) and within-levels mean square (MSE) provide two independent estimates of the variance. If the means for different levels of mix design are truly different from each other, the MSB will tend

to be larger than the MSE, such that it will be more likely to reject the null hypothesis. For this example, the calculations for MSB, MSE, and *F* are as follows:

$$MSB = \frac{SSB}{df_B} = 1082878.40$$

$$MSE = \frac{SSE}{df_E} = 50810.70$$

$$F = \frac{MSB}{MSE} = 21.31$$

If there are no effects due to level, the *F*-statistic will tend to be smaller. If there are effects due to level, the *F*-statistic will tend to be larger, as is the case in this example. ANOVA computations usually are summarized in the form of a table. Table 12 summarizes the computations for this example.

The final step is to determine F_{crit} from the *F*-distribution table (e.g., NCHRP Project 20-45, Volume 2, Appendix C, Table C-5) with $t - 1$ $(5 - 1 = 4)$ degrees of freedom for the numerator and $n_T - t$ $(15 - 5 = 10)$ degrees of freedom for the denominator. For a significance level of $\alpha = 0.01$, F_{crit} is found (in Table C-5) to be 5.99. Given that $F > F_{crit}$ $(21.31 > 5.99)$, the null hypothesis that all mix designs have equal compressive strength is rejected, supporting the conclusion that at least two mix designs are different from each other in their mean effect. Table 12 also shows the *p*-value calculated using a computer program. The *p*-value is the probability that a sample would result in the given statistic value if the null hypothesis were true. The *p*-value of 0.0000698408 is well below the chosen significance level of 0.01.

5. **Interpreting the Results:** The ANOVA results in rejection of the null hypothesis at $\alpha = 0.01$. That is, the mean values are judged to be statistically different. However, the ANOVA result does not indicate where the difference lies. For example, does the compressive strength of mix design A differ from that of mix design C or D? To carry out such multiple mean comparisons, the analyst must control the experiment-wise error rate (EER) by employing more conservative methods such as Tukey's test, Bonferroni's test, or Scheffe's test, as appropriate. (Details for ANOVA are given in NCHRP Project 20-45, Volume 2, Chapter 4, Section A, "Analysis of Variance Methodology.")

   The coefficient of determination (R^2) provides a rough indication of how well the statistical model fits the data. For this example, R^2 is calculated as follows:

$$R^2 = \frac{SSB}{TSS} = \frac{4331513.60}{4839620.90} = 0.90$$

   For this example, R^2 indicates that the one-way ANOVA classification model accounts for 90% of the total variation in the data. In the controlled laboratory experiment demonstrated in this example, $R^2 = 0.90$ indicates a fairly acceptable fit of the statistical model to the data.
6. **Conclusion and Discussion:** This example illustrates a simple one-way ANOVA where inference regarding parameters (mean values) from more than two populations or treatments was

Table 12. ANOVA results.

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	*F*	Probability > *F* (Significance)
Between	4331513.60	4	1082878.40	21.31	0.0000698408
Within	508107.30	10	50810.70		
Total	4839620.90	14			

desired. The focus of computations was the construction of the ANOVA table. Before proceeding with ANOVA, however, an analyst must verify that the assumptions of common variance and data normality are satisfied within each group/level. The results do not establish the cause of difference in compressive strength between mix designs in any way. The experimental setup and analytical procedure shown in this example may be used to test other properties of mix designs such as flexure strength. If another factor (for example, water/cement ratio with levels low or high) is added to the analysis, the classification will become a two-way ANOVA. (In this report, two-way ANOVA is demonstrated in Example 11.) Notice that the equations shown in Example 8 may only be used for one-way ANOVA for balanced designs, meaning that in this experiment there are equal numbers of replicates for each level within a factor. (For a discussion of computations on unbalanced designs and multifactor designs, see NCHRP Project 20-45.)

7. **Applications in Other Areas of Transportation Research:** Examples of applications of one-way ANOVA in other areas of transportation research include:
   - **Traffic Operations—**to determine the effect of various traffic calming devices on average speeds in residential areas.
   - **Traffic Operations/Safety—**to study the effect of weather conditions on accidents in a given time period.
   - **Work Zones—**to compare the effect of different placements of work zone signs on reduction in highway speeds at some downstream point.
   - **Materials—**to investigate the effect of recycled aggregates on compressive and flexural strength of concrete.

Example 9: Materials; Simple Analysis of Variance (ANOVA)

Area: Materials

Method of Analysis: Simple analysis of variance (ANOVA) comparing more than two mean values and using the *F*-test for equality of means

1. **Research Question/Problem Statement:** To illustrate how increasingly detailed analysis may be appropriate, Example 9 is an extension of the two-sample comparison presented in Example 7. As a part of dispute resolution during quality control and quality assurance, let's say the highway agency engineer from Example 7 decides to reconfirm the contractor's test results for asphalt content. The agency hires an independent consultant to verify both the contractor- and agency-measured asphalt contents. It now becomes necessary to compare more than two mean values. A simple one-way analysis of variance (ANOVA) can be used to analyze the asphalt contents measured by three different parties.

Question/Issue

Extend a comparison of two mean values to compare three (or more) mean values. Specifically, use data collected by several (>2) different parties to see if the results (mean values) are the same. Formally, test the following null (H_o) and alternative (H_a) hypotheses, which can be stated as follows:

H_o: There is no difference in asphalt content among three different parties:

$\left(H_o: m_{contractor} = m_{agency} = m_{consultant}\right)$

H_a: At least one of the parties has a different measured asphalt content.

Table 13. Asphalt content data summary.

Party Type	Asphalt Content Percent
Contractor	$\bar{X}_1 = 6.1$ $s_1 = 0.254$ $n_1 = 12$
Agency	$\bar{X}_2 = 5.7$ $s_2 = 0.303$ $n_2 = 6$
Consultant	$\bar{X}_3 = 5.12$ $s_3 = 0.186$ $n_3 = 12$

2. **Identification and Description of Variables:** The independent consultant runs 12 additional asphalt content tests by taking independent samples from the same pavement section as the agency and contractor. The question is whether it is likely that the tests came from the same population, based on their variability.
3. **Data Collection:** The descriptive statistics (mean, standard deviation, and sample size) for the asphalt content data collected by the three parties are shown in Table 13. Notice that 12 measurements each have been taken by the contractor and the independent consultant, while the agency has only taken six measurements.

   The data for the contractor and the agency are the same as presented in Example 7. For brevity, the consultant's raw observations are not repeated here. The mean value and standard deviation for the consultant's data are calculated using the same formulas and equations that were used in Example 7.
4. **Specification of Analysis Technique and Data Analysis:** The agency engineer can use one-way ANOVA to resolve this question. (Details for one-way ANOVA are available in NCHRP Project 20-45, Volume 2, Chapter 4, Section A, "Analysis of Variance Methodology.") The objective of the ANOVA is to determine whether the variance observed in the dependent variable (in this case, asphalt content) is due to the differences among the samples (different from one party to another) or due to the differences within the samples. ANOVA is basically an extension of two-sample comparisons to cases when three or more samples are being compared. More formally, the technician is testing to see whether the between-sample variability is large relative to the within-sample variability, as stated in the formal hypothesis. This type of comparison also may be referred to as between-groups versus within-groups variance.

   Rejection of the null hypothesis (that the mean values are the same) gives the engineer some information concerning differences among the population means; however, it does not indicate which means actually differ from each other. Rejection of the null hypothesis tells the engineer that differences exist, but it does not specify that $\bar{X}_1$ differs from $\bar{X}_2$ or from $\bar{X}_3$.

   To control the experiment-wise error rate (EER) for multiple mean comparisons, a conservative test—Tukey's procedure for unplanned comparisons—can be used for unplanned comparisons. (Information about Tukey's procedure can be found in almost any good statistics textbook, such as those by Freund and Wilson [2003] and Kutner et al. [2005].) The *F*-statistic calculated for determining the effect of who (agency, contractor, or consultant) measured

40 Effective Experiment Design and Data Analysis in Transportation Research

Table 14. ANOVA results.

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	*F*	Significance
Between groups	5.6	2	2.8	49.1	0.000
Within groups	1.5	27	0.06		
Total	7.2	29			

the asphalt content is given in Table 14. (See Example 8 for a more detailed discussion of the calculations necessary to create Table 14.)

Although the ANOVA results reveal whether there are overall differences, it is always good practice to visually examine the data. For example, Figure 9 shows the mean and associated 95% confidence intervals (CI) of the mean asphalt content measured by each of the three parties involved in the testing.

5. **Interpreting the Results:** A simple one-way ANOVA is conducted to determine whether there is a difference in mean asphalt content as measured by the three different parties. The analysis shows that the *F*-statistic is significant (p-value < 0.05), meaning that at least two of the means are significantly different from each other. The engineer can use Tukey's procedure for comparisons of multiple means, or he or she can observe the plotted 95% confidence intervals to figure out which means are actually (and significantly) different from each other (see Figure 9). Because the confidence intervals overlap, the results show that the asphalt content measured by the contractor and the agency are somewhat different. (These same conclusions were obtained in Example 7.) However, the mean asphalt content obtained by the consultant is significantly different from (and lower than) that obtained by both of the other parties. This is evident because the confidence interval for the consultant doesn't overlap with the confidence interval of either of the other two parties.

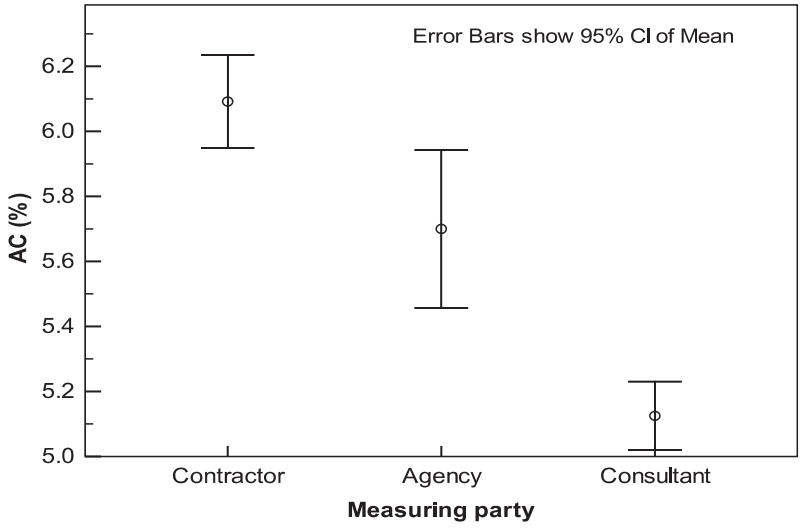

Figure 9. Mean and confidence intervals for asphalt content data.

6. **Conclusion and Discussion:** This example uses a simple one-way ANOVA to compare the mean values of three sets of results using data drawn from the same test section. The error bar plots for data from the three different parties visually illustrate the statistical differences in the multiple means. However, the *F*-test for multiple means should be used to formally test the hypothesis of the equality of means. The interpretation of results will be misleading if the variances of populations being compared for their mean difference are not equal. Based on the comparison of the three means, it can be concluded that the construction material in this example may not have consistent properties, as indicated by the results from the independent consultant.
7. **Applications in Other Areas of Transportation Research:** Simple one-way ANOVA is often used when more than two means must be compared. Examples of applications in other areas of transportation research include:
   - **Traffic Safety/Operations—**to evaluate the effect of intersection type on the average number of accidents per month. Three or more types of intersections (e.g., signalized, non-signalized, and rotary) could be selected for study in an urban area having similar traffic volumes and vehicle mix.
   - **Pavement Engineering**
     – to investigate the effect of hot-mix asphalt (HMA) layer thickness on fatigue cracking after 20 years of service life. Three HMA layer thicknesses (5 inches, 6 inches, and 7 inches) are to be involved in this study, and other factors (i.e., traffic, climate, and subbase/base thicknesses and subgrade types) need to be similar.
     – to determine the effect of climatic conditions on rutting performance of flexible pavements. Three or more climatic conditions (e.g., wet-freeze, wet-no-freeze, dry-freeze, and dry-no-freeze) need to be considered while other factors (i.e., traffic, HMA, and subbase/base thicknesses and subgrade types) need to be similar.

Example 10: Pavements; Simple Analysis of Variance (ANOVA)

Area: Pavements

Method of Analysis: Simple analysis of variance (ANOVA) comparing the mean values of more than two samples and using the *F*-test

1. **Research Question/Problem Statement:** The aggregate coefficient of thermal expansion (CTE) in Portland cement concrete (PCC) is a critical factor affecting thermal behavior of PCC slabs in concrete pavements. In addition, the interaction between slab curling (caused by the thermal gradient) and axle loads is assumed to be a critical factor for concrete pavement performance in terms of cracking. To verify the effect of aggregate CTE on slab cracking, a pavement engineer wants to conduct a simple observational study by collecting field pavement performance data on three different types of pavement. For this example, three types of aggregate (limestone, dolomite, and gravel) are being used in concrete pavement construction and yield the following CTEs:
   - 4 in./in. per °F
   - 5 in./in. per °F
   - 6.5 in./in. per °F

   It is necessary to compare more than two mean values. A simple one-way ANOVA is used to analyze the observed slab cracking performance by the three different concrete mixes with different aggregate types based on geology (limestone, dolomite, and gravel). All other factors that might cause variation in cracking are assumed to be held constant.

42 Effective Experiment Design and Data Analysis in Transportation Research

Question/Issue

Compare the means of more than two samples. Specifically, is the cracking performance of concrete pavements designed using more than two different types of aggregates the same? Stated a bit differently, is the performance of three different types of concrete pavement statistically different (are the mean performance measures different)?

2. **Identification and Description of Variables:** The engineer identifies 1-mile sections of uniform pavement within the state highway network with similar attributes (aggregate type, slab thickness, joint spacing, traffic, and climate). Field performance, in terms of the observed percentage of slab cracked ("% slab cracked," i.e., how cracked is each slab) for each pavement section after about 20 years of service, is considered in the analysis. The available pavement data are grouped (stratified) based on the aggregate type (CTE value). The % slab cracked after 20 years is the dependent variable, while CTE of aggregates is the independent variable. The question is whether pavement sections having different types of aggregate (CTE values) exhibit similar performance based on their variability.
3. **Data Collection:** From the data stratified by CTE, the engineer randomly selects nine pavement sections within each CTE category (i.e., 4, 5, and 6.5 in./in. per °F). The sample size is based on the statistical power (1-β) requirements. (For a discussion on sample size determination based on statistical power requirements, see NCHRP Project 20-45, Volume 2, Chapter 1, "Sample Size Determination.") The descriptive statistics for the data, organized by three CTE categories, are shown in Table 15. The engineer considers pavement performance data for 9 pavement sections in each CTE category.
4. **Specification of Analysis Technique and Data Analysis:** Because the engineer is concerned with the comparison of more than two mean values, the easiest way to make the statistical comparison is to perform a one-way ANOVA (see NCHRP Project 20-45, Volume 2, Chapter 4). The comparison will help to determine whether the between-section variability is large relative to the within-section variability. More formally, the following hypotheses are tested:

   H_O: All mean values are equal (i.e., $m_1 = m_2 = m_3$).

   H_A: At least one of the means is different from the rest.

   Although rejection of the null hypothesis gives the engineer some information concerning difference among the population means, it doesn't tell the engineer anything about how the means differ from each other. For example, does m_1 differ from m_2 or m_3? To control the experiment-wise error rate (EER) for multiple mean comparisons, a conservative test—Tukey's procedure for unplanned comparisons—can be used. (Information about Tukey's procedure can be found in almost any good statistics textbook, such as those by Freund and Wilson [2003] and Kutner et al. [2005].)The *F*-statistic calculated for determining the effect of CTE on % slab cracked after 20 years is shown in Table 16.

Table 15. Pavement performance data.

CTE (in./in. per °F)	% Slab Cracked After 20 Years
4	$\bar{X}_1 = 37,\ s_1 = 4.8,\ n_1 = 9$
5	$\bar{X}_2 = 53.7,\ s_2 = 6.1,\ n_2 = 9$
6.5	$\bar{X}_3 = 72.5,\ s_3 = 6.3,\ n_3 = 9$

Table 16. ANOVA results.

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	*F*	Significance
Between groups	5652.7	2	2826.3	84.1	0.000
Within groups	806.9	24	33.6		
Total	6459.6	26			

The data in Table 16 have been produced by considering the original data and following the procedures presented in earlier examples. The emphasis in this example is on understanding what the table of results provides the researcher. Also in this example, the test for homogeneity of variances (Levene test) shows no significant difference among the standard deviations of % slab cracked for different CTE values. Figure 10 presents the mean and associated 95% confidence intervals of the average % slab cracked (also called the mean and error bars) measured for the three CTE categories considered.

5. **Interpreting the Results:** A simple one-way ANOVA is conducted to determine if there is a difference among the mean values for % slab cracked for different CTE values. The analysis shows that the *F*-statistic is significant (p-value < 0.05), meaning that at least two of the means are statistically significantly different from each other. To gain more insight, the engineer can use Tukey's procedure to specifically compare the mean values, or the engineer may simply observe the plotted 95% confidence intervals to ascertain which means are significantly different from each other (see Figure 10). The plotted results show that the mean % slab cracked varies significantly for different CTE values—there is no overlap between the different mean/error bars. Figure 10 also shows that the mean % slab cracked is significantly higher for pavement sections having a higher CTE value. (For more information about Tukey's procedure, see NCHRP Project 20-45, Volume 2, Chapter 4.)
6. **Conclusion and Discussion:** In this example, simple one-way ANOVA is used to assess the effect of CTE on cracking performance of rigid pavements. The *F*-test for multiple means is used to formally test the (null) hypothesis of mean equality. The confidence interval plots for data from pavements having three different CTE values visually illustrate the statistical differences in the three means. The interpretation of results will be misleading if the variances of

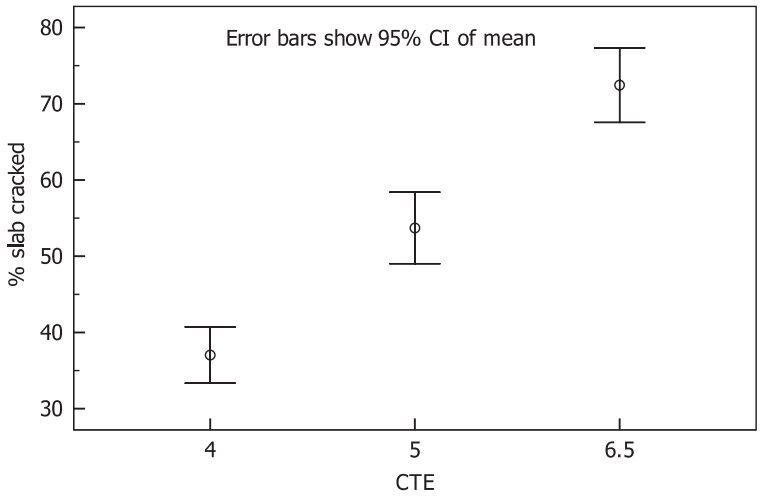

Figure 10. Error bars for % slab cracked with different CTE.

populations being compared for their mean difference are not equal or if a proper multiple mean comparisons procedure is not adopted. Based on the comparison of the three means in this example, the engineer can conclude that the pavement slabs having aggregates with a higher CTE value will exhibit more cracking than those with lower CTE values, given that all other variables (e.g., climate effects) remain constant.

7. **Applications in Other Areas of Transportation Research:** Simple one-way ANOVA is widely used and can be employed whenever multiple means within a factor are to be compared with one another. Potential applications in other areas of transportation research include:
   - **Traffic Operations—**to evaluate the effect of commuting time on level of service (LOS) of an urban highway. Mean travel times for three periods (e.g., morning, afternoon, and evening) could be selected for specified highway sections to collect the traffic volume and headway data in all lanes.
   - **Traffic Safety—**to determine the effect of shoulder width on accident rates on rural highways. More than two shoulder widths (e.g., 0 feet, 6 feet, 9 feet, and 12 feet) should be selected in this study.
   - **Pavement Engineering—**to investigate the impact of air void content on flexible pavement fatigue performance. Pavement sections having three or more air void contents (e.g., 3%, 5%, and 7%) in the surface HMA layer could be selected to compare their average fatigue cracking performance after the same period of service (e.g., 15 years).
   - **Materials—**to study the effect of aggregate gradation on the rutting performance of flexible pavements. Three types of aggregate gradations (fine, intermediate, and coarse) could be adopted in the laboratory to make different HMA mix samples. Performance testing could be conducted in the laboratory to measure rut depths for a given number of load cycles.

Example 11: Pavements; Factorial Design (ANOVA Approach)

Area: Pavements

Method of Analysis: Factorial design (an ANOVA approach used to explore the effects of varying more than one independent variable)

1. **Research Question/Problem Statement:** Extending the information from Example 10 (a simple ANOVA example for pavements), the pavement engineer has verified that the coefficient of thermal expansion (CTE) in Portland cement concrete (PCC) is a critical factor affecting thermal behavior of PCC slabs in concrete pavements and significantly affects concrete pavement performance in terms of cracking. The engineer now wants to investigate the effects of another factor, joint spacing (JS), in addition to CTE. To study the combined effects of PCC CTE and JS on slab cracking, the engineer needs to conduct a factorial design study by collecting field pavement performance data. As before, three CTEs will be considered:
   - 4 in./in. per °F,
   - 5 in./in. per °F, and
   - 6.5 in./in. per °F.

   Now, three different joint spacings (12 ft, 16 ft, and 20 ft) also will be considered. For this example, it is necessary to compare multiple means within each factor (main effects) and the interaction between the two factors (interactive effects). The statistical technique involved is called a multifactorial two-way ANOVA.
2. **Identification and Description of Variables:** The engineer identifies uniform 1-mile pavement sections within the state highway network with similar attributes (e.g., slab thickness, traffic, and climate). The field performance, in terms of observed percentage of each slab cracked (% slab cracked) after about 20 years of service for each pavement section, is considered the

dependent (or response) variable in the analysis. The available pavement data are stratified based on CTE and JS. CTE and JS are considered the independent variables. The question is whether pavement sections having different CTE and JS exhibit similar performance based on their variability.

Question/Issue

Use collected data to determine the effects of varying more than one independent variable on some measured outcome. In this example, compare the cracking performance of concrete pavements considering two independent variables: (1) coefficients of thermal expansion (CTE) as measured using more than two types of aggregate and (2) differing joint spacing (JS). More formally, the hypotheses can be stated as follows:

$H_o : a_i = 0$, No difference in % slabs cracked for different CTE values.
$H_o : g_j = 0$, No difference in % slabs cracked for different JS values.
$H_o : (ag)_{ij} = 0$, for all *i* and *j,* No difference in % slabs cracked for different CTE and JS combinations.

3. **Data Collection:** The descriptive statistics for % slab cracked data by three CTE and three JS categories are shown in Table 17. From the data stratified by CTE and JS, the engineer has randomly selected three pavement sections within each of nine combinations of CTE values. (In other words, for each of the nine pavement sections from Example 10, the engineer has selected three JS.)
4. **Specification of Analysis Technique and Data Analysis:** The engineer can use two-way ANOVA test statistics to determine whether the between-section variability is large relative to the within-section variability for each factor to test the following null hypotheses:
   - $H_o : a_i = 0$
   - $H_o : g_j = 0$
   - $H_o : (ag)_{ij} = 0$

   As mentioned before, although rejection of the null hypothesis does give the engineer some information concerning differences among the population means (i.e., there are differences among them), it does not clarify which means differ from each other. For example, does μ_1 differ from μ_2 or μ_3? To control the experiment-wise error rate (EER) for the comparison of multiple means, a conservative test—Tukey's procedure for an unplanned comparison—can be used. (Information about two-way ANOVA is available in NCHRP Project 20-45, Volume 2,

Table 17. Summary of cracking data.

		CTE (in/in per °F)			Marginal μ & σ
		4	5	6.5	
Joint spacing (ft)	12	$\bar{x}_{1,1} = 32.4$ $s_{1,1} = 0.1$	$\bar{x}_{1,2} = 46.8$ $s_{1,2} = 1.8$	$\bar{x}_{1,3} = 65.3$ $s_{1,3} = 3.2$	$\bar{x}_{1,.} = 48.2$ $s_{1,.} = 14.4$
	16	$\bar{x}_{2,1} = 36.0$ $s_{2,1} = 2.4$	$\bar{x}_{2,2} = 54$ $s_{2,2} = 2.9$	$\bar{x}_{2,3} = 73$ $s_{2,3} = 1.1$	$\bar{x}_{2,.} = 54.3$ $s_{2,.} = 16.1$
	20	$\bar{x}_{3,1} = 42.7$ $s_{3,1} = 2.4$	$\bar{x}_{3,2} = 60.3$ $s_{3,2} = 0.5$	$\bar{x}_{3,3} = 79.1$ $s_{3,3} = 2.0$	$\bar{x}_{3,.} = 60.7$ $s_{3,.} = 15.9$
Marginal μ & σ		$\bar{x}_{.,1} = 37.0$ $s_{.,1} = 4.8$	$\bar{x}_{.,2} = 53.7$ $s_{.,2} = 6.1$	$\bar{x}_{.,3} = 72.5$ $s_{.,3} = 6.3$	$\bar{x}_{.,.} = 54.4$ $s_{.,.} = 15.8$

Note: $n = 3$ in each cell; values are cell means and standard deviations.

46 Effective Experiment Design and Data Analysis in Transportation Research

Table 18. ANOVA results.

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)	*F*	Significance
CTE	5677.74	2	2838.87	657.16	0.000
JS	703.26	2	351.63	81.40	0.000
CTE × JS	0.12	4	0.03	0.007	0.999
Residual/error	77.76	18	4.32		
Total	6458.88	26			

Chapter 4. Information about Tukey's procedure can be found in almost any good statistics textbook, such as those by Freund and Wilson [2003] and Kutner et al. [2005].)

The results of the two-way ANOVA are shown in Table 18. From the first line it can be seen that both of the main effects, CTE and JS, are significant in explaining cracking behavior (i.e., both p-values < 0.05). However, the interaction (CTE × JS) is not significant (i.e., the p-value is 0.999, much greater than 0.05). Also, the test for homogeneity of variances (Levene statistic) shows that there is no significant difference among the standard deviations of % slab cracked for different CTE and JS values. Figure 11 illustrates the main and interactive effects of CTE and JS on % slabs cracked.

5. **Interpreting the Results:** A two-way (multifactorial) ANOVA is conducted to determine if difference exists among the mean values for "% slab cracked" for different CTE and JS values. The analysis shows that the main effects of both CTE and JS are significant, while the interaction effect is insignificant (p-value > 0.05). These results show that when CTE and JS are considered jointly, they significantly impact the slab cracking separately. Given these results, the conclusions from the results will be based on the main effects alone without considering interaction effects. In fact, if the interaction effect had been significant, the conclusions would be based on them. To gain more insight, the engineer can use Tukey's procedure to compare specific multiple means within each factor, or the engineer can simply observe the plotted means in Figure 11 to ascertain which means are significantly different from each other. The plotted results show that the mean % slab cracked varies significantly for different CTE and JS values; that is, the CTE seems to be more influential than JS. All lines are almost parallel to

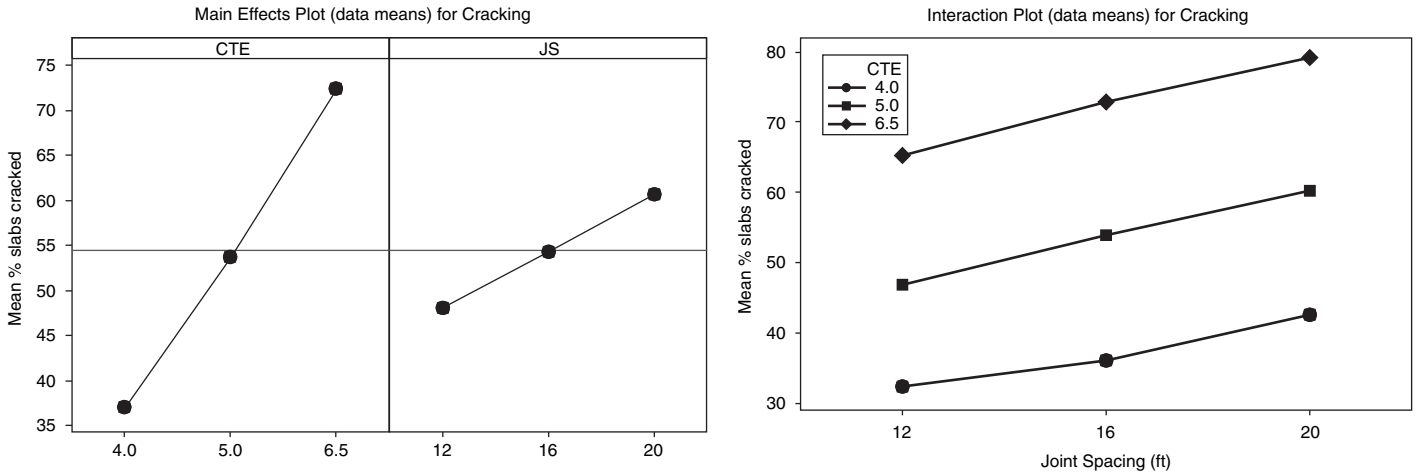

Figure 11. Main and interaction effects of CTE and JS on slab cracking.

each other when plotted for both factors together, showing no interactive effects between the levels of two factors.

6. **Conclusion and Discussion:** The two-way ANOVA can be used to verify the combined effects of CTE and JS on cracking performance of rigid pavements. The marginal mean plot for cracking having three different CTE and JS levels visually illustrates the differences in the multiple means. The plot of cell means for cracking within the levels of each factor can indicate the presence of interactive effect between two factors (in this example, CTE and JS). However, the *F*-test for multiple means should be used to formally test the hypothesis of mean equality. Finally, based on the comparison of three means within each factor (CTE and JS), the engineer can conclude that the pavement slabs having aggregates with higher CTE and JS values will exhibit more cracking than those with lower CTE and JS values. In this example, the effect of CTE on concrete pavement cracking seems to be more critical than that of JS.
7. **Applications in Other Areas of Transportation Research:** Multifactorial designs can be used when more than one factor is considered in a study. Possible applications of these methods can extend to all transportation-related areas, including:
   - **Pavement Engineering**
     - to determine the effects of base type and base thickness on pavement performance of flexible pavements. Two or more levels can be considered within each factor; for example, two base types (aggregate and asphalt-treated bases) and three base thicknesses (8 inches, 12 inches, and 18 inches).
     - to investigate the impact of pavement surface conditions and vehicle type on fuel consumption. The researcher can select pavement sections with three levels of ride quality (smooth, rough, and very rough) and three types of vehicles (cars, vans, and trucks). The fuel consumptions can be measured for each vehicle type on all surface conditions to determine their impact.
   - **Materials**
     - to study the effects of aggregate gradation and surface on tensile strength of hot-mix asphalt (HMA). The engineer can evaluate two levels of gradation (fine and coarse) and two types of aggregate surfaces (smooth and rough). The samples can be prepared for all the combinations of aggregate gradations and surfaces for determination of tensile strength in the laboratory.
     - to compare the impact of curing and cement types on the compressive strength of concrete mixture. The engineer can design concrete mixes in laboratory utilizing two cement types (Type I & Type III). The concrete samples can be cured in three different ways for 24 hours and 7 days (normal curing, water bath, and room temperature).

Example 12: Work Zones; Simple Before-and-After Comparisons

Area: Work zones

Method of Analysis: Simple before-and-after comparisons (exploring the effect of some treatment before it is applied versus after it is applied)

1. **Research Question/Problem Statement:** The crash rate in work zones has been found to be higher than the crash rate on the same roads when a work zone is not present. For this reason, the speed limit in construction zones often is set lower than the prevailing non-work-zone speed limit. The state DOT decides to implement photo-radar speed enforcement in a work zone to determine if this speed-enforcement technique reduces the average speed of free-flowing vehicles in the traffic stream. They measure the speeds of a sample of free-flowing vehicles prior to installing the photo-radar speed-enforcement equipment in a work zone and

48 Effective Experiment Design and Data Analysis in Transportation Research

then measure the speeds of free-flowing vehicles at the same location after implementing the photo-radar system.

Question/Issue

Use collected data to determine whether a difference exists between results before and after some treatment is applied. For this example, does a photo-radar speed-enforcement system reduce the speed of free-flowing vehicles in a work zone, and, if so, is the reduction statistically significant?

2. **Identification and Description of Variables:** The variable to be analyzed is the mean speed of vehicles before and after the implementation of a photo-radar speed-enforcement system in a work zone.
3. **Data Collection:** The speeds of individual free-flowing vehicles are recorded for 30 minutes on a Tuesday between 10:00 a.m. and 10:30 a.m. before installing the photo-radar system. After the system is installed, the speeds of individual free-flowing vehicles are recorded for 30 minutes on a Tuesday between 10:00 a.m. and 10:30 a.m. The before sample contains 120 observations and the after sample contains 100 observations.
4. **Specification of Analysis Technique and Data Analysis:** A test of the significance of the difference between two means requires a statement of the hypothesis to be tested (H_o) and a statement of the alternate hypothesis (H_1). In this example, these hypotheses can be stated as follows:

   H_o: There is no difference in the mean speed of free-flowing vehicles before and after the photo-radar speed-enforcement system is displayed.

   H_1: There is a difference in the mean speed of free-flowing vehicles before and after the photo-radar speed-enforcement system is displayed.

   Because these two samples are independent, a simple *t*-test is appropriate to test the stated hypotheses. This test requires the following procedure:

   Step 1. Compute the mean speed ($\bar{x}$) for the before sample ($\bar{x}_b$) and the after sample ($\bar{x}_a$) using the following equation:

$$\bar{x}_i = \frac{\sum_{i=1}^{n} x_i}{n_i}; \; n_b = 120 \text{ and } n_a = 100$$

   Results: $\bar{x}_b = 53.1$ mph and $\bar{x}_a = 50.5$ mph.

   Step 2. Compute the variance (S^2) for each sample using the following equation:

$$S^2 = \sum_{i-1}^{n} \frac{(x_i - \bar{x}_i)^2}{n-1}$$

   where $n_a = 100$; $\bar{x}_a = 50.5$ mph; $n_b = 120$; and $\bar{x}_b = 53.1$ mph

   Results: $S_b^2 = \sum \frac{(x_b - \bar{x}_b)^2}{n_b - 1} = 12.06$ and $S_a^2 = \sum \frac{(x_a - \bar{x}_a)^2}{n_a - 1} = 12.97.$

   Step 3. Compute the pooled variance of the two samples using the following equation:

$$S_p^2 = \frac{\sum (x_a - \bar{x}_a)^2 + \sum (x_b - \bar{x}_b)^2}{n_b + n_a - 2}$$

   Results: $S_p^2 = 12.472$ and $S_p = 3.532$.

Step 4. Compute the *t*-statistic using the following equation:

$$t = \frac{\bar{x}_b - \bar{x}_a}{S_p}\sqrt{\frac{n_a n_b}{n_a + n_b}}$$

Result: $t = \frac{53.1 - 50.5}{3.532}\sqrt{\frac{(100)(120)}{100+120}} = 5.43.$

5. **Interpreting the Results:** The results of the sample *t*-test are obtained by comparing the value of the calculated *t*-statistic (5.43 in this example) with the value of the *t*-statistic for the level of confidence desired. For a level of confidence of 95%, the *t*-statistic must be greater than 1.96 to reject the null hypotheses (H_o) that the use of a photo-radar speed-enforcement system does not change the speed of free-flowing vehicles. (For more information, see NCHRP Project 20-45, Volume 2, Appendix C, Table C-4.)
6. **Conclusion and Discussion:** The sample problem illustrates the use of a statistical test to determine whether the difference in the value of the variable of interest between the before conditions and the after conditions is statistically significant. The before condition is without photo-radar speed enforcement; and the after condition is with photo-radar speed enforcement. In this sample problem, the computed *t*-statistic (5.43) is greater than the critical *t*-statistic (1.96), so the null hypothesis is rejected. This means the change in the speed of free-flowing vehicles when the photo-radar speed-enforcement system is used is statistically significant. The assumption is made that all other factors that would affect the speed of free-flowing vehicles (e.g., traffic mix, weather, or construction activity) are the same in the before-and-after conditions. This test is robust if the normality assumption does not hold completely; however, it should be checked using box plots. For significant departures from normality and variance equality assumptions, non-parametric tests must be conducted. (For more information, see NCHRP Project 20-45, Volume 2, Chapter 6, Section C and also Example 21).

   The reliability of the results in this example could be improved by using a control group. As the example has been constructed, there is an assumption that the only thing that changed at this site was the use of photo-radar speed enforcement; that is, it is assumed that all observed differences are attributable to the use of the photo-radar. If other factors—even something as simple as a general decrease in vehicle speeds in the area—might have impacted speed changes, the effect of the photo-radar speed enforcement would have to be adjusted for those other factors. Measurements taken at a control site (ideally identical to the experiment site) during the same time periods could be used to detect background changes and then to adjust the photo-radar effects. Such a situation is explored in Example 13.
7. **Applications in Other Areas in Transportation:** The before-and-after comparison can be used whenever two independent samples of data are (or can be assumed to be) normally distributed with equal variance. Applications of before-and-after comparison in other areas of transportation research may include:
   - **Traffic Operations**
     - to compare the average delay to vehicles approaching a signalized intersection when a fixed time signal is changed to an actuated signal or a traffic-adaptive signal.
     - to compare the average number of vehicles entering and leaving a driveway when access is changed from full access to right-in, right-out only.
   - **Traffic Safety**
     - to compare the average number of crashes on a section of road before and after the road is resurfaced.
     - to compare the average number of speeding citations issued per day when a stationary operation is changed to a mobile operation.
   - **Maintenance**—to compare the average number of citizen complaints per day when a change is made in the snow plowing policy.

50 Effective Experiment Design and Data Analysis in Transportation Research

Example 13: Traffic Safety; Complex Before-and-After Comparisons and Controls

Area: Traffic safety

Method of Analysis: Complex before-and-after comparisons using control groups (examining the effect of some treatment or application with consideration of other factors that may also have an effect)

1. **Research Question/Problem Statement:** A state safety engineer wants to estimate the effectiveness of fluorescent orange warning signs as compared to standard orange signs in work zones on freeways and other multilane highways. Drivers can see fluorescent signs from a longer distance than standard signs, especially in low-visibility conditions, and the extra cost of the fluorescent material is not too high. Work-zone safety is a perennial concern, especially on freeways and multilane highways where speeds and traffic volumes are high.

Question/Issue

How can background effects be separated from the effects of a treatment or application? Compared to standard orange signs, do fluorescent orange warning signs increase safety in work zones on freeways and multilane highways?

2. **Identification and Description of Variables:** The engineer quickly concludes that there is a need to collect and analyze safety surrogate measures (e.g., traffic conflicts and late lane changes) rather than collision data. It would take a long time and require experimentation at many work zones before a large sample of collision data could be ready for analysis on this question. Surrogate measures relate to collisions, but they are much more numerous and it is easier to collect a large sample of them in a short time. For a study of traffic safety, surrogate measures might include near-collisions (traffic conflicts), vehicle speeds, or locations of lane changes.

   In this example, the engineer chooses to use the location of the lane-change maneuver made by drivers in a lane to be closed entering a work zone. This particular surrogate safety measure is a measure of effectiveness (MOE). The hypothesis is that the farther downstream at which a driver makes a lane change out of a lane to be closed—when the highway is still below capacity—the safer the work zone.
3. **Data Collection:** The engineer establishes site selection criteria and begins examining all active work zones on freeways and multilane highways in the state for possible inclusion in the study. The site selection criteria include items such as an active work zone, a cooperative contractor, no interchanges within the approach area, and the desired lane geometry. Seven work zones meet the criteria and are included in the study.

   The engineer decides to use a before-and-after (sometimes designated B/A or b/a) experiment design with randomly selected control sites. The latter are sites in the same population as the treatment sites; that is, they meet the same selection criteria but are untreated (i.e., standard warning signs are employed, not the fluorescent orange signs). This is a strong experiment design because it minimizes three common types of bias in experiments: history, maturation, and regression to the mean.

   History bias exists when changes (e.g., new laws or large weather events) happen at about the same time as the treatment in an experiment, so that the engineer or analyst cannot separate the effect of the treatment from the effects of the other events. Maturation bias exists when gradual changes occur throughout an extended experiment period and cannot be separated from the effects of the treatment. Examples of maturation bias might involve changes like the aging of driver populations or new vehicles with more air bags. History and maturation biases are referred to as specification errors and are described in more detail in NCHRP Project 20-45, Volume 2,

Table 19. Lane-change data for before-and-after comparison using controls.

Time Period	Number of Vehicles Observed in Lane to be Closed at Midpoint		
	Control	Treatment	Total
Before	1262	2099	3361
After	1476	1877	3353
Total	2738	3976	6714

Chapter 1, in the section "Quasi-Experiments." Regression-to-the-mean bias exists when sites with the highest MOE levels in the before time period are treated. If the MOE level falls in the after period, the analyst can never be sure how much of the fall was due to the treatment and how much was due to natural fluctuations in the values of the MOE back toward its usual mean value. A before-and-after study with randomly selected control sites minimizes these biases because their effects are expected to apply just as much to the treatment sites as to the control sites.

In this example, the engineer randomly selects four of the seven work zones to receive fluorescent orange signs. The other three randomly selected work zones received standard orange signs and are the control sites. After the signs have been in place for a few weeks (a common tactic in before-and-after studies to allow regular drivers to get used to the change), the engineer collects data at all seven sites. The location of each vehicle's lane-change maneuver out of the lane to be closed is measured from video tape recorded for several hours at each site. Table 19 shows the lane-change data at the midpoint between the first warning sign and beginning of the taper. Notice that the same number of vehicles is observed in the before-and-after periods for each type of site.

4. **Specification of Analysis Technique and Data Analysis:** Depending on their format, data from a before-and-after experiment with control sites may be analyzed several ways. The data in the table lend themselves to analysis with a chi-square test to see whether the distributions between the before-and-after conditions are the same at both the treatment and control sites. (For more information about chi-square testing, see NCHRP Project 20-45, Volume 2, Chapter 6, Section E, "Chi-Square Test for Independence.") To perform the chi-square test on the data for Example 13, the engineer first computes the expected value in each cell. For the cell corresponding to the before time period for control sites, this value is computed as the row total (3361) times the column total (2738) divided by the grand total (6714):

$$3361 * 2738/6714 = 1371 \text{ vehicles}$$

The engineer next computes the chi-square value for each cell using the following equation:

$$\chi_i^2 = \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the number of actual observations in cell i and E_i is the expected number of observations in cell i.

For example, the chi-square value in the cell corresponding to the before time period for control sites is $(1262 - 1371)^2 / 1371 = 8.6$. The engineer then sums the chi-square values from all four cells to get 29.1. That sum is then compared to the critical chi-square value for the significance level of 0.025 with 1 degree of freedom (degrees of freedom = number of rows − 1 * number of columns − 1), which is shown on a standard chi-square distribution table to be 5.02 (see NCHRP Project 20-45, Volume 2, Appendix C, Table C-2.) A significance level of 0.025 is not uncommon in such experiments (although 0.05 is a general default value), but it is a standard that is difficult but not impossible to meet.

52 Effective Experiment Design and Data Analysis in Transportation Research

5. **Interpreting the Results:** Because the calculated chi-square value is greater than the critical chi-square value, the engineer concludes that there is a statistically significant difference in the number of vehicles in the lane to be closed at the midpoint between the before-and-after time periods for the treatment sites relative to what would be expected based on the control sites. In other words, there is a difference that is due to the treatment.
6. **Conclusion and Discussion:** The experiment results show that fluorescent orange signs in work zone approaches like those tested would likely have a safety benefit. Although the engineer cannot reasonably estimate the number of collisions that would be avoided by using this treatment, the before-and-after study with control using a safety surrogate measure makes it clear that some collisions will be avoided. The strength of the experiment design with randomly selected control sites means that agencies can have confidence in the results.

   The consequences of an error in an analysis like this that results in the wrong conclusion can be devastating. If the error leads an agency to use a safety measure more than it should, precious safety funds will be wasted that could be put to better use. If the error leads an agency to use the safety measure less often than it should, money will be spent on measures that do not prevent as many collisions. With safety funds in such short supply, solid analyses that lead to effective decisions on countermeasure deployment are of great importance.

   A before-and-after experiment with control is difficult to arrange in practice. Such an experiment is practically impossible using collision data, because that would mean leaving some higher collision sites untreated during the experiment. Such experiments are more plausible using surrogate measures like the one described in this example.
7. **Applications in Other Areas of Transportation Research:** Before-and-after experiments with randomly selected control sites are difficult to arrange in transportation safety and other areas of transportation research. The instinct to apply treatments to the worst sites, rather than randomly—as this method requires—is difficult to overcome. Despite the difficulties, such experiments are sometimes performed in:
   - **Traffic Operations—**to test traffic control strategies at a number of different intersections.
   - **Pavement Engineering—**to compare new pavement designs and maintenance processes to current designs and practice.
   - **Materials—**to compare new materials, mixes, or processes to standard mixtures or processes.

Example 14: Work Zones; Trend Analysis

Area: Work zones

Method of Analysis: Trend analysis (examining, describing, and modeling how something changes over time)

1. **Research Question/Problem Statement:** Measurements conducted over time often reveal patterns of change called trends. A model may be used to predict some future measurement, or the relative success of a different treatment or policy may be assessed. For example, work/construction zone safety has been a concern for highway officials, engineers, and planners for many years. Is there a pattern of change?

Question/Issue

Can a linear model represent change over time? In this particular example, is there a trend over time for motor vehicle crashes in work zones? The problem is to predict values of crash frequency at specific points in time. Although the question is simple, the statistical modeling becomes sophisticated very quickly.

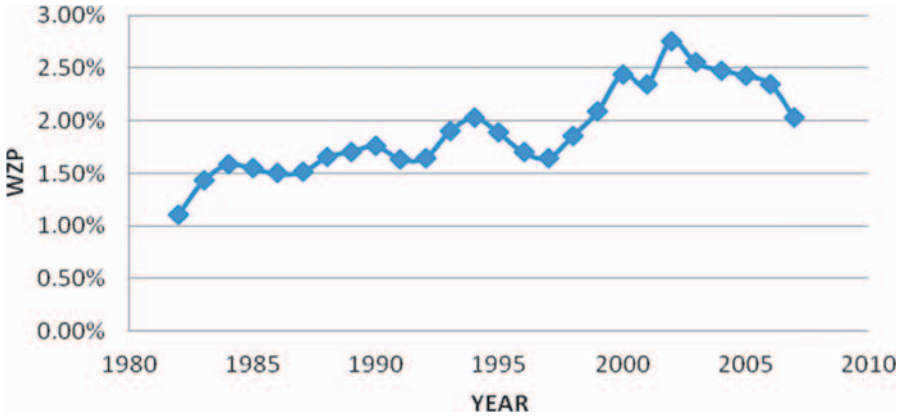

Figure 12. Percentage of all motor vehicle fatalities occurring in work zones.

2. **Identification and Description of Variables:** Highway safety, rather the lack of it, is revealed by the total number of fatalities due to motor vehicle crashes. The percentage of those deaths occurring in work zones reveals a pattern over time (Figure 12).

   The data points for the graph are calculated using the following equation:

   $$WZP = a + b\ YEAR + u$$

   where WZP = work zone percentage of total fatalities,
   $YEAR$ = calendar year, and
   u = an error term, as used here.

3. **Data Collection:** The base data are obtained from the Fatality Analysis Reporting System maintained by the National Highway Traffic Safety Administration (NHTSA), as reported at www.workzonesafety.org. The data are state specific as well as for the country as a whole, and cover a period of 26 years from 1982 through 2007. The numbers of fatalities from motor vehicle crashes in and not in construction/maintenance zones (work zones) are used to compute the percentage of fatalities in work zones for each of the 26 years.
4. **Specification of Analysis Techniques and Data Analysis:** Ordinary least squares (OLS) regression is used to develop the general model specified above. The discussion in this example focuses on the resulting model and the related statistics. (See also examples 15, 16, and 17 for details on calculations. For more information about OLS regression, see NCHRP Project 20-45, Volume 2, Chapter 4, Section B, "Linear Regression.")

   Looking at the data in Figure 12 another way,

WZP = −91.523 +	0.047(YEAR)	R = 0.867	$R^2 = 0.751$
(−8.34)	(8.51)	*t*-values	
(0.000)	(0.000)	*p*-values	

   The trend is significant: the line (trend) shows an increase of 0.047% each year. Generally, this trend shows that work-zone fatalities are increasing as a percentage of total fatalities.
5. **Interpreting the Results:** This experiment is a good fit and generally shows that work-zone fatalities were an increasing problem over the period 1982 through 2007. This is a trend that highway officials, engineers, and planners would like to change. The analyst is therefore interested in anticipating the trajectory of the trend. Here the trend suggests that things are getting worse.

54 Effective Experiment Design and Data Analysis in Transportation Research

How far might authorities let things go—5%? 10%? 25%? Caution must be exercised when interpreting a trend beyond the limits of the available data. Technically the slope, or *b*-coefficient, is the trend of the relationship. The *a*-term from the regression, also called the intercept, is the value of WZP when the independent variable equals zero. The intercept for the trend in this example would technically indicate that the percentage of motor vehicle fatalities in work zones in the year zero would be –91.5%. This is absurd on many levels. There could be no motor vehicles in year zero, and what is a negative percentage of the total? The absurdity of the intercept in this example reveals that trends are limited concepts, limited to a relevant time frame.

Figure 12 also suggests that the trend, while valid for the 26 years in aggregate, doesn't work very well for the last 5 years, during which the percentages are consistently falling, not rising. Something seems to have changed around 2002; perhaps the highway officials, engineers, and planners took action to change the trend, in which case, the trend reversal would be considered a policy success.

Finally, some underlying assumptions must be considered. For example, there is an implicit assumption that the types of roads with construction zones are similar from year to year. If this assumption is not correct (e.g., if a greater number of high speed roads, where fatalities may be more likely, are worked on in some years than in others), then interpreting the trend may not make much sense.

6. **Conclusion and Discussion:** The computation of this dependent variable (the percent of motor-vehicle fatalities occurring in work zones, or MZP) is influenced by changes in the number of work-zone fatalities and the number of non-work-zone fatalities. To some extent, both of these are random variables. Accordingly, it is difficult to distinguish a trend or trend reversal from a short series of possibly random movements in the same direction. Statistically, more observations permit greater confidence in non-randomness.

   It is also possible that a data series might be recorded that contains regular, non-random movements that are unrelated to a trend. Consider the dependent variable above (MZP), but measured using monthly data instead of annual data. Further, imagine looking at such data for a state in the upper Midwest instead of for the nation as a whole. In this new situation, the WZP might fall off or halt altogether each winter (when construction and maintenance work are minimized), only to rise again in the spring (reflecting renewed work-zone activity). This change is not a trend per se, nor is it random. Rather, it is cyclical.
7. **Applications in Other Areas of Transportation Research:** Applications of trend analysis models in other areas of transportation research include:
   - **Transportation Safety—**to identify trends in traffic crashes (e.g., motor vehicle/deer) over time on some part of the roadway system (e.g., freeways).
   - **Public Transportation—**to determine the trend in rail passenger trips over time (e.g., in response to increasing gas prices).
   - **Pavement Engineering—**to monitor the number of miles of pavement that is below some service-life threshold over time.
   - **Environment—**to monitor the hours of truck idling time in rest areas over time.

Example 15: Structures/Bridges; Trend Analysis

Area: Structures/bridges

Method of Analysis: Trend analysis (examining a trend over time)

1. **Research Question/Problem Statement:** A state agency wants to monitor trends in the condition of bridge superstructures in order to perform long-term needs assessment for bridge rehabilitation or replacement. Bridge condition rating data will be analyzed for bridge

superstructures that have been inspected over a period of 15 years. The objective of this study is to examine the overall pattern of change in the indicator variable over time.

Question/Issue

Use collected data to determine if the values that some variables have taken show an increasing trend or a decreasing trend over time. In this example, determine if levels of structural deficiency in bridge superstructures have been increasing or decreasing over time, and determine how rapidly the increase or decrease has occurred.

2. **Identification and Description of Variables:** Bridge inspection generally entails collection of numerous variables including location information, traffic data, structural elements (type and condition), and functional characteristics. Based on the severity of deterioration and the extent of spread through a bridge component, a condition rating is assigned on a discrete scale from 0 (failed) to 9 (excellent). Generally a condition rating of 4 or below indicates deficiency in a structural component.

   The state agency inspects approximately 300 bridges every year (denominator). The number of superstructures that receive a rating of 4 or below each year (number of events, numerator) also is recorded. The agency is concerned with the change in overall rate (calculated per 100) of structurally deficient bridge superstructures. This rate, which is simply the ratio of the numerator to the denominator, is the indicator (dependent variable) to be examined for trend over a time period of 15 years. Notice that the unit of analysis is the time period and not the individual bridge superstructures.
3. **Data Collection:** Data are collected for bridges scheduled for inspection each year. It is important to note that the bridge condition rating scale is based on subjective categories, and therefore there may be inherent variability among inspectors in their assignments of rates to bridge superstructures. Also, it is assumed that during the time period for which the trend analysis is conducted, no major changes are introduced in the bridge inspection methods. Sample data provided in Table 20 show the rate (per 100), number of bridges per year that received a score of four or below, and total number of bridges inspected per year.
4. **Specification of Analysis Technique and Data Analysis:** The data set consists of 15 observations, one for each year. Figure 13 shows a scatter plot of the rate (dependent variable) versus time in years. The scatter plot does not indicate the presence of any outliers. The scatter plot shows a seemingly increasing linear trend in the rate of deficient superstructures over time. No need for data transformation or smoothing is apparent from the examination of the scatter plot in Figure 13.

   To determine whether the apparent linear trend is statistically significant in this data, ordinary least squares (OLS) regression can be employed.

Table 20. Sample bridge inspection data.

No.	Year	Rate (per 100)	Number of Events (Numerator)	Number of Bridges Inspected (Denominator)
1	1990	8.33	25	300
2	1991	8.70	26	299
5	1994	10.54	31	294
11	2000	13.55	42	310
15	2004	14.61	45	308

56 Effective Experiment Design and Data Analysis in Transportation Research

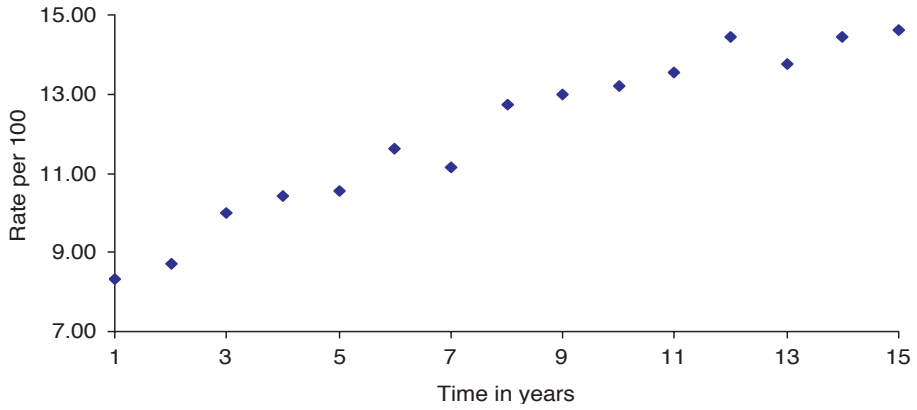

Figure 13. Scatter plot of time versus rate.

The linear regression model takes the following form:

$$y_i = \beta_o + \beta_1 x_i + e_i$$

where $i = 1, 2, \ldots, n$ ($n = 15$ in this example),
y = dependent variable (rate of structurally deficient bridge superstructures),
x = independent variable (time),
β_o = y-intercept (only provides reference point),
β_1 = slope (change in unit y for a change in unit x), and
e_i = residual error.

The first step is to estimate the β_o and β_1 in the regression function. The residual errors (e) are assumed to be independently and identically distributed (i.e., they are mutually independent and have the same probability distribution). β_1 and β_o can be computed using the following equations:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 0.454$$

$$\hat{\beta}_o = \bar{y} - \beta_1 \bar{x} = 8.396$$

where $\bar{y}$ is the overall mean of the dependent variable and $\bar{x}$ is the overall mean of the independent variable.

The prediction equation for rate of structurally deficient bridge superstructures over time can be written using the following equation:

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x = 8.396 + 0.454x$$

That is, as time increases by a year, the rate of structurally deficient bridge superstructures increases by 0.454 per 100 bridges. The plot of the regression line is shown in Figure 14.

Figure 14 indicates some small variability about the regression line. To conduct hypothesis testing for the regression relationship (H_o: $\beta_1 = 0$), assessment of this variability and the assumption of normality would be required. (For a discussion on assumptions for residual errors, see NCHRP Project 20-45, Volume 2, Chapter 4.)

Like analysis of variance (ANOVA, described in examples 8, 9, and 10), statistical inference is initiated by partitioning the total sum of squares (TSS) into the error sum of squares (SSE)

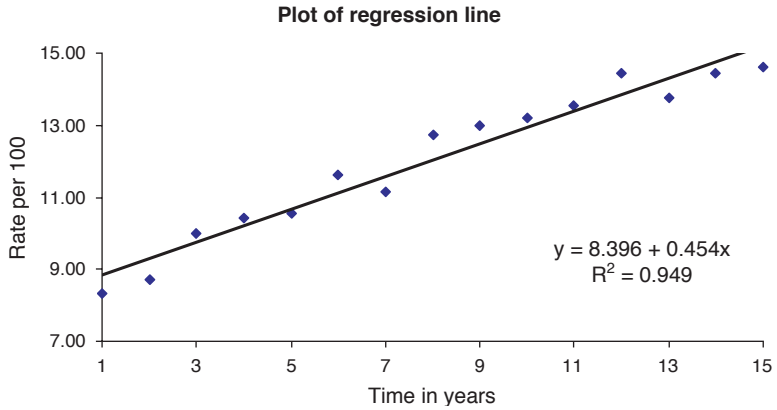

Figure 14. Plot of regression line.

and the model sum of squares (SSR). That is, TSS = SSE + SSR. The TSS is defined as the sum of the squares of the difference of each observation from the overall mean. In other words, deviation of observation from overall mean (TSS) = deviation of observation from prediction (SSE) + deviation of prediction from overall mean (SSR). For our example,

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 60.892$$

$$SSR = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 = 57.790$$

$$SSE = TSS - SSR = 3.102$$

Regression analysis computations are usually summarized in a table (see Table 21).

The mean squared errors (MSR, MSE) are computed by dividing the sums of squares by corresponding model and error degrees of freedom. For the null hypothesis (H_o: $\beta_1 = 0$) to be true, the expected value of MSR is equal to the expected value of MSE such that F = MSR/MSE should be a random draw from an F-distribution with 1, $n - 2$ degrees of freedom.

From the regression shown in Table 21, F is computed to be 242.143, and the probability of getting a value larger than the F computed is extremely small. Therefore, the null hypothesis is rejected; that is, the slope is significantly different from zero, and the linearly increasing trend is found to be statistically significant. Notice that a slope of zero implies that knowing a value of the independent variable provides no insight on the value of the dependent variable.

5. **Interpreting the Results:** The linear regression model does not imply any cause-and-effect relationship between the independent and dependent variables. The y-intercept only provides a reference point, and the relationship need not be linear outside the data range. The 95% confidence interval for β_1 is computed as [0.391, 0.517]; that is, the analyst is 95% confident that the true mean increase in the rate of structurally deficient bridge superstructures is between

Table 21. Analysis of regression table.

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square	*F*	Significance
Regression	57.790	1	57.790 (MSR)	242.143	8.769e-10
Error	3.102	13	0.239 (MSE)		
Total	60.892	14			

0.391% and 0.517% per year. (For a discussion on computing confidence intervals, see NCHRP Project 20-45, Volume 2, Chapter 4.)

The coefficient of determination (R^2) provides an indication of the model fit. For this example, R^2 is calculated using the following equation:

$$R^2 = \frac{SSE}{TSS} = 0.949$$

The R^2 indicates that the regression model accounts for 94.9% of the total variation in the (hypothetical) data. It should be noted that such a high value of R^2 is almost impossible to attain from analysis of real observational data collected over a long time. Also, distributional assumptions must be checked before proceeding with linear regression, as serious violations may indicate the need for data transformation, use of non-linear regression or non-parametric methods, and so on.

6. **Conclusion and Discussion:** In this example, simple linear regression has been used to determine the trend in the rate of structurally deficient bridge superstructures in a geographic area. In addition to assessing the overall patterns of change, trend analysis may be performed to:
   - study the levels of indicators of change (or dependent variables) in different time periods to evaluate the impact of technical advances or policy changes;
   - compare different geographic areas or different populations with perhaps varying degrees of exposure in absolute and relative terms; and
   - make projections to monitor progress toward an objective.

   However, given the dynamic nature of trend data, many of these applications require more sophisticated techniques than simple linear regression.

   An important aspect of examining trends over time is the accuracy of numerator and denominator data. For example, bridge structures may be examined more than once during the analysis time period, and retrofit measures may be taken at some deficient bridges. Also, the age of structures is not accounted for in this analysis. For the purpose of this example, it is assumed that these (and other similar) effects are negligible and do not confound the data. In real-life application, however, if the analysis time period is very long, it becomes extremely important to account for changes in factors that may have affected the dependent variable(s) and their measurement. An example of the latter could be changes in the volume of heavy trucks using the bridge, changes in maintenance policies, or changes in plowing and salting regimes.
7. **Applications in Other Areas of Transportation Research:** Trend analysis is carried out in many areas of transportation research, such as:
   - **Transportation Planning/Traffic Operations—**to determine the need for capital improvements by examining traffic growth over time.
   - **Traffic Safety—**to study the trends in overall, fatal, and/or injury crash rates over time in a geographic area.
   - **Pavement Engineering—**to assess the long-term performance of pavements under varying loads.
   - **Environment—**to monitor the emission levels from commercial traffic over time with growth of industrial areas.

Example 16: Transportation Planning; Multiple Regression Analysis

Area: Transportation planning

Method of Analysis: Multiple regression analysis (testing proposed linear models with more than one independent variable when all variables are continuous)

1. **Research Question/Problem Statement:** Transportation planners and engineers often work on variations of the classic four-step transportation planning process for estimating travel demand. The first step, trip generation, generally involves developing a model that can be used to predict the number of trips originating or ending in a zone, which is a geographical subdivision of a corridor, city, or region (also referred to as a traffic analysis zone or TAZ). The objective is to develop a statistical relationship (a model) that can be used to explain the variation in a dependent variable based on the variation of one or more independent variables. In this example, ordinary least squares (OLS) regression is used to develop a model between trips generated (the dependent variable) and demographic, socio-economic, and employment variables (independent variables) at the household level.

Question/Issue

Can a linear relationship (model) be developed between a dependent variable and one or more independent variables? In this application, the dependent variable is the number of trips produced by households. Independent variables include persons, workers, and vehicles in a household, household income, and average age of persons in the household.

The basic question is whether the relationship between the dependent (Y) and independent (X) variables can be represented by a linear model using two coefficients (**a** and **b**), expressed as follows:

$$Y = \mathbf{a} + \mathbf{b} \cdot X$$

where **a** = the intercept and **b** = the slope of the line.

If the relationship being examined involves more than one independent variable, the equation will simply have more terms. In addition, in a more formal presentation, the equation will also include an error term, ε, added at the end.

2. **Identification and Description of Variables:** Data for four-step modeling of travel demand or for calibration of any specific model (e.g., trip generation or trip origins) come from a variety of sources, ranging from the U.S. Census to mail or telephone surveys. The data that are collected will depend, in part, on the specific purpose of the modeling effort. Data appropriate for a trip-generation model typically are collected from some sort of household survey. For the dependent variable in a trip-generation model, data must be collected on trip-making characteristics. These characteristics could include something as simple as the total trips made by a household in a day or involve more complicated breakdowns by trip purpose (e.g., work-related trips versus shopping trips) and time of day (e.g., trips made during peak and non-peak hours). The basic issue that must be addressed is to determine the purpose of the proposed model: What is to be estimated or predicted? Weekdays and work trips normally are associated with peak congestion and are often the focus of these models.

   For the independent variable(s), the analyst must first give some thought to what would be the likely causes for household trips to vary. For example, it makes sense intuitively that household size might be pertinent (i.e., it seems reasonable that more persons in the household would lead to a higher number of household trips). Household members could be divided into workers and non-workers, two variables instead of one. Likewise, other socio-economic characteristics, such as income-related variables, might also make sense as candidate variables for the model. Data are collected on a range of candidate variables, and

60 Effective Experiment Design and Data Analysis in Transportation Research

the analysis process is used to sort through these variables to determine which combination leads to the best model.

To be used in ordinary regression modeling, variables need to be continuous; that is, measured ratio or interval scale variables. Nominal data may be incorporated through the use of indicator (dummy) variables. (For more information on continuous variables, see NCHRP Project 20-45, Volume 2, Chapter 1; for more information on dummy variables, see NCHRP Project 20-45, Volume 2, Chapter 4).

3. **Data Collection:** As noted, data for modeling travel demand often come from surveys designed especially for the modeling effort. Data also may be available from centralized sources such as a state DOT or local metropolitan planning organization (MPO).
4. **Specification of Analysis Techniques and Data Analysis:** In this example, data for 178 households in a small city in the Midwest have been provided by the state DOT. The data are obtained from surveys of about 15,000 households all across the state. This example uses only a tiny portion of the data set (see Table 22). Based on the data, a fairly obvious relationship is initially hypothesized: more persons in a household (PERS) should produce more person-trips (TRIPS).

   In its simplest form, the regression model has one dependent variable and one independent variable. The underlying assumption is that variation in the independent variable causes the variation in the dependent variable. For example, the dependent variable might be $TRIPS_i$ (the count of total trips made on a typical weekday), and the independent variable might be PERS (the total number of persons, or occupants, in the household). Expressing the relationship between TRIPS and PERS for the *i*th household in a sample of households results in the following hypothesized model:

$$TRIPS_i = \mathbf{a} + \mathbf{b} \cdot PERS_i + \varepsilon_i$$

where **a** and **b** are coefficients to be determined by ordinary least squares (OLS) regression analysis and ε_i is the error term.

The difference between the value of TRIPS for any household predicted using the developed equation and the actual observed value of TRIPS for that same household is called the residual.

The resulting model is an equation for the best fit straight line (for the given data) where **a** is the intercept and **b** is the slope of the line. (For more information about fitted regression and measures of fit see NCHRP Project 20-45, Volume 2, Chapter 4).

In Table 22, R is the multiple R, the correlation coefficient in the case of the simplest linear regression involving one variable (also called univariate regression). The R^2 (coefficient of determination) may be interpreted as the proportion of the variance of the dependent variable explained by the fitted regression model. The adjusted R^2 corrects for the number of independent variables in the equation. A "perfect" R^2 of 1.0 could be obtained if one included enough independent variables (e.g., one for each observation), but doing so would hardly be useful.

Table 22. Regression model statistics.

Coefficients	*t*-values (statistics)	*p*-values	Measures of Fit
a = 3.347	4.626	0.000	R = 0.510
b = 2.001	7.515	0.000	R^2 = 0.260
			Adjusted R^2 = 0.255

Restating the now-calibrated model,

$$\text{TRIPS} = 4.626 + 7.515 \cdot \text{PERS}$$

The statistical significance of each coefficient estimate is evaluated with the *p*-values of calculated *t*-statistics, provided the errors are normally distributed. The *p*-values (also known as probability values) generally indicate whether the coefficients are significantly different from zero (which they need to be in order for the model to be useful). More formally stated, a *p*-value is the probability of a Type I error.

In this example, the *t*- and *p*-values shown in Table 22 indicate that both **a** and **b** are significantly different from zero at a level of significance greater than the 99.9% confidence level. *P*-values are generally offered as two-tail (two-sided hypothesis testing) test values in results from most computer packages; one-tail (one-sided) values may sometimes be obtained by dividing the printed *p*-values by two. (For more information about one-sided versus two-sided hypothesis testing, see NCHRP Project 20-45, Volume 2, Chapter 4.)

The R^2 may be tested with an F-statistic; in this example, the F was calculated as 56.469 (degrees of freedom = 2, 176) (See NCHRP Project 20-45, Volume 2, Chapter 4). This means that the model explains a significant amount of the variation in the dependent variable. A plot of the estimated model (line) and the actual data are shown in Figure 15.

A strict interpretation of this model suggests that a household with zero occupants (PERS=0) will produce 3.347 trips per day. Clearly, this is not feasible because there can't be a household of zero persons, which illustrates the kind of problem encountered when a model is extrapolated beyond the range of the data used for the calibration. In other words, a formal test of the intercept (the **a**) is not always meaningful or appropriate.

Extension of the Model to Multivariate Regression: When the list of potential independent variables is considered, the researcher or analyst might determine that more than one cause for variation in the dependent variable may exist. In the current example, the question of whether there is more than one cause for variation in the number of trips can be considered.

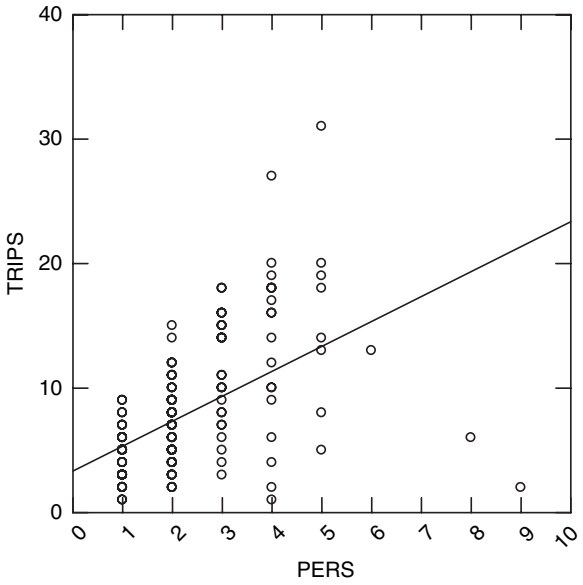

Figure 15. Plot of the line for the estimated model.

62 Effective Experiment Design and Data Analysis in Transportation Research

Table 23. Results from fitting the multivariate model.

Coefficients	*t*-values (statistics)	*p*-values	Measures of Fit
a = 8.564	6.274	3.57E-09*	R = 0.589
b = 0.899	2.832	0.005	R^2 = 0.347
c = 1.067	3.360	0.001	adjusted R^2 = 0.330
d = 1.907E-05*	1.927	0.056	
e = -0.098	-4.808	3.68E-06	

*See note about scientific notation in Section 5, Interpreting the Results.

The model just discussed for evaluating the effect of one independent variable is called a univariate model. Should the final model for this example be multivariate? Before determining the final model, the analyst may want to consider whether a variable or variables exist that further clarify what has already been modeled (e.g., more persons cause more trips). The variable **PERS** is a crude measure, made up of workers and non-workers. Most households have one or two workers. It can be shown that a measure of the non-workers in the household is more effective in explaining trips than is total persons; so a new variable, persons minus workers (**DEP**), is calculated.

Next, variables may exist that address entirely different causal relationships. It might be hypothesized that as the number of registered motor vehicles available in the household (**VEH**) increases, the number of trips will increase. It may also be argued that as household income (**INC,** measured in thousands of dollars) increases, the number of trips will increase. Finally, it may be argued that as the average age of household occupants (**AVEAGE**) increases, the number of trips will decrease because retired people generally make fewer trips. Each of these statements is based upon a logical argument (hypothesis).

Given these arguments, the hypothesized multivariate model takes the following form:

$$\mathbf{TRIPS} = a + b \cdot \mathbf{DEP} + c \cdot \mathbf{VEH} + d \cdot \mathbf{INC} + e \cdot \mathbf{AVEAGE} + \varepsilon$$

The results from fitting the multivariate model are given in Table 23.

Results of the analysis of variance (ANOVA) for the overall model are shown in Table 24.

5. **Interpreting the Results:** It is common for regression packages to provide some values in scientific notation as shown for the *p*-values in Table 23. The coefficient **d**, showing the relationship of TRIPS with INC, is read 1.907 E-05, which in turn is read as $1.907 \cdot 10^{-5}$ or 0.000001907.

All coefficients are of the expected sign and significantly different from 0 (at the 0.05 level) except for **d**. However, testing the intercept makes little sense. (The intercept value would be the number of trips for a household with 0 vehicles, 0 income, 0 average age, and 0 dependents, a most unlikely household.) The overall model is significant as shown by the F-ratio and its *p*-value, meaning that the model explains a significant amount of the variation in

Table 24. ANOVA results for the overall model.

ANOVA	Sum of Squares (SS)	Degrees of Freedom (df)	*F*-ratio	*p*-value
Regression	1487.5	4	19.952	3.4E-13
Residual	2795.7	150		

the dependent variable. This model should reliably explain 33% of the variance of household trip generation. Caution should be exercised when interpreting the significance of the R^2 and the overall model because it is not uncommon to have a significant F-statistic when some of the coefficients in the equation are not significant. The analyst may want to consider recalibrating the model without the income variable because the coefficient **d** was insignificant.

6. **Conclusion and Discussion:** Regression, particularly OLS regression, relies on several assumptions about the data, the nature of the relationships, and the results. Data are assumed to be interval or ratio scale. Independent variables generally are assumed to be measured without error, so all error is attributed to the model fit. Furthermore, independent variables should be independent of one another. This is a serious concern because the presence in the model of related independent variables, called multicollinearity, compromises the *t*-tests and confuses the interpretation of coefficients. Tests of this problem are available in most statistical software packages that include regression. Look for Variance-Inflation Factor (VIF) and/or Tolerance tests; most packages will have one or the other, and some will have both.

   In the example above where PERS is divided into DEP and workers, knowing any two variables allows the calculation of the third. Including all three variables in the model would be a case of extreme multicollinearity and, logically, would make no sense. In this instance, because one variable is a linear combination of the other two, the calculations required (within the analysis program) to calibrate the model would actually fail. If the independent variables are simply highly correlated, the regression coefficients (at a minimum) may not have intuitive meaning. In general, equations or models with highly correlated independent variables are to be avoided; alternative models that examine one variable or the other, but not both, should be analyzed.

   It is also important to analyze the error distributions. Several assumptions relate to the errors and their distributions (normality, constant variance, uncorrelated, etc.) In transportation planning, spatial variables and associations might become important; they require more elaborate constructs and often different estimation processes (e.g., Bayesian, Maximum Likelihood). (For more information about errors and error distributions, see NCHRP Project 20-45, Volume 2, Chapter 4.)

   Other logical considerations also exist. For example, for the measurement units of the different variables, does the magnitude of the result of multiplying the coefficient and the measured variable make sense and/or have a reasonable effect on the predicted magnitude of the dependent variable? Perhaps more importantly, do the independent variables make sense? In this example, does it make sense that changes in the number of vehicles in the household would cause an increase or decrease in the number of trips? These are measures of operational significance that go beyond consideration of statistical significance, but are no less important.

7. **Applications in Other Areas of Transportation Research:** Regression is a very important technique across many areas of transportation research, including:
   - **Transportation Planning**
     - to include the other half of trip generation, e.g., predicting trip destinations as a function of employment levels by various types (factory, commercial), square footage of shopping center space, and so forth.
     - to investigate the trip distribution stage of the 4-step model (log transformation of the gravity model).
   - **Public Transportation—**to predict loss/liability on subsidized freight rail lines (function of segment ton-miles, maintenance budgets and/or standards, operating speeds, etc.) for self-insurance computations.
   - **Pavement Engineering—**to model pavement deterioration (or performance) as a function of easily monitored predictor variables.

Example 17: Traffic Operations; Regression Analysis

Area: Traffic operations

Method of Analysis: Regression analysis (developing a model to predict the values that some variable can take as a function of one or more other variables, when not all variables are assumed to be continuous)

1. **Research Question/Problem Statement:** An engineer is concerned about false capacity at intersections being designed in a specified district. False capacity occurs where a lane is dropped just beyond a signalized intersection. Drivers approaching the intersection and knowing that the lane is going to be dropped shortly afterward avoid the lane. However, engineers estimating the capacity and level of service of the intersection during design have no reliable way to estimate the percentage of traffic that will avoid the lane (the lane distribution).

Question/Issue

Develop a model that can be used to predict the values that a dependent variable can take as a function of changes in the values of the independent variables. In this particular instance, how can engineers make a good estimate of the lane distribution of traffic volume in the case of a lane drop just beyond an intersection? Can a linear model be developed that can be used to predict this distribution based on other variables?

The basic question is whether a linear relationship exists between the dependent variable (Y; in this case, the lane distribution percentage) and some independent variable(s) (X). The relationship can be expressed using the following equation:

$$Y = \mathbf{a} + \mathbf{b} \cdot X$$

where **a** is the intercept and **b** is the slope of the line (see NCHRP Project 20-45, Volume 2, Chapter 4, Section B).

2. **Identification and Description of Variables:** The dependent variable of interest in this example is the volume of traffic in each lane on the approach to a signalized intersection with a lane drop just beyond. The traffic volumes by lane are converted into lane utilization factors (f_{LU}), to be consistent with standard highway capacity techniques. The *Highway Capacity Manual* defines f_{LU} using the following equation:

$$f_{LU} = \frac{v_g}{v_{g1}(N)}$$

where V_g is the flow rate in a lane group in vehicles per hour, V_{g1} is the flow rate in the lane with the highest flow rate of any in the group in vehicles per hour, and N is the number of lanes in the lane group.

The engineer thinks that lane utilization might be explained by one or more of 15 different factors, including the type of lane drop, the distance from the intersection to the lane drop, the taper length, and the heavy vehicle percentage. All of the variables are continuous except the type of lane drop. The type of lane drop is used to categorize the sites.

3. **Data Collection:** The engineer locates 46 lane-drop sites in the area and collects data at these sites by means of video recording. The engineer tapes for up to 3 hours at each site. The data are summarized in 15-minute periods, again to be consistent with standard highway capacity practice. For one type of lane-drop geometry, with two through lanes and an exclusive right-turn lane on the approach to the signalized intersection, the engineer ends up with 88 valid

data points (some sites have provided more than one data point), covering 15 minutes each, to use in equation (model) development.

4. **Specification of Analysis Technique and Data Analysis:** Multiple (or multivariate) regression is a standard statistical technique to develop predictive equations. (More information on this topic is given in NCHRP Project 20-45, Volume 2, Chapter 4, Section B). The engineer performs five steps to develop the predictive equation.

Step 1. The engineer examines plots of each of the 15 candidate variables versus f_{LU} to see if there is a relationship and to see what forms the relationships might take.

Step 2. The engineer screens all 15 candidate variables for multicollinearity. (Multicollinearity occurs when two variables are related to each other and essentially contribute the same information to the prediction.) Multicollinearity can lead to models with poor predicting power and other problems. The engineer examines the variables for multicollinearity by

- looking at plots of each of the 15 candidate variables against every other candidate variable;
- calculating the correlation coefficient for each of the 15 candidate independent variables against every other candidate variable; and
- using more sophisticated tests (such as the variance influence factor) that are available in statistical software.

Step 3. The engineer reduces the set of candidate variables to eight. Next, the engineer uses statistical software to select variables and estimate the coefficients for each selected variable, assuming that the regression equation has a linear form. To select variables, the engineer employs forward selection (adding variables one at a time until the equation fit ceases to improve significantly) and backward elimination (starting with all candidate variables in the equation and removing them one by one until the equation fit starts to deteriorate). The equation fit is measured by R^2 (for more information, see NCHRP Project 20-45, Volume 2, Chapter 4, Section B, under the heading, "Descriptive Measures of Association Between X and Y"), which shows how well the equation fits the data on a scale from 0 to 1, and other factors provided by statistical software. In this case, forward selection and backward elimination result in an equation with five variables:

- Drop: Lane drop type, a 0 or 1 depending on the type;
- Left: Left turn status, a 0 or 1 depending on the types of left turns allowed;
- Length: The distance from the intersection to the lane drop, in feet ÷ 1000;
- Volume: The average lane volume, in vehicles per hour per lane ÷ 1000; and
- Sign: The number of signs warning of the lane drop.

Notice that the first two variables are discrete variables and had to assume a zero-or-one format to work within the regression model. Each of the five variables has a coefficient that is significantly different from zero at the 95% confidence level, as measured by a *t*-test. (For more information, see NCHRP Project 20-45, Volume 2, Chapter 4, Section B, "How Are *t*-statistics Interpreted?")

Step 4. Once an initial model has been developed, the engineer plots the residuals for the tentative equation to see whether the assumed linear form is correct. A residual is the difference, for each observation, between the prediction the equation makes for f_{LU} and the actual value of f_{LU}.

In this example, a plot of the predicted value versus the residual for each of the 88 data points shows a fan-like shape, which indicates that the linear form is not appropriate. (NCHRP Project 20-45, Volume 2, Chapter 4, Section B, Figure 6 provides examples of residual plots that are and are not desirable.) The engineer experiments with several other model forms, including non-linear equations that involve transformations of variables, before settling on a lognormal form that provides a good R^2 value of 0.73 and a desirable shape for the residual plot.

Step 5. Finally, the engineer examines the candidate equation for logic and practicality, asking whether the variables make sense, whether the signs of the variables make sense, and whether the variables can be collected easily by design engineers. Satisfied that the answers to these questions are "yes," the final equation (model) can be expressed as follows:

$$f_{LU} = \exp(-0.539 - 0.218 \cdot \text{Drop} + 0.148 \cdot \text{Left} + 0.178 \cdot \text{Length} + 0.627 \cdot \text{Volume} - 0.105 \cdot \text{Sign})$$

5. **Interpreting the Results:** The process described in this example results in a useful equation for estimating the lane utilization in a lane to be dropped, thereby avoiding the estimation of false capacity. The equation has five terms and is non-linear, which will make its use a bit challenging. However, the database is large, the equation fits the data well, and the equation is logical, which should boost the confidence of potential users. If potential users apply the equation within the ranges of the data used for the calibration, the equation should provide good predictions. Applying any model outside the range of the data on which it was calibrated increases the likelihood of an inaccurate prediction.
6. **Conclusion and Discussion:** Regression is a powerful statistical technique that provides models engineers can use to make predictions in the absence of direct observation. Engineers tempted to use regression techniques should notice from this and other examples that the effort is substantial. Engineers using regression techniques should not skip any of the steps described above, as doing so may result in equations that provide poor predictions to users.

   Analysts considering developing a regression model to help make needed predictions should not be intimidated by the process. Although there are many pitfalls in developing a regression model, analysts considering making the effort should also consider the alternative: how the prediction will be made in the absence of a model. In the absence of a model, predictions of important factors like lane utilization would be made using tradition, opinion, or simple heuristics. With guidance from NCHRP Project 20-45 and other texts, and with good software available to make the calculations, credible regression models often can be developed that perform better than the traditional prediction methods.

   Because regression models developed by transportation engineers are often reused in later studies by others, the stakes are high. The consequences of a model that makes poor predictions can be severe in terms of suboptimal decisions. Lane utilization models often are employed in traffic studies conducted to analyze new development proposals. A model that under-predicts utilization in a lane to be dropped may mean that the development is turned down due to the anticipated traffic impacts or that the developer has to pay for additional and unnecessary traffic mitigation measures. On the other hand, a model that over-predicts utilization in a lane to be dropped may mean that the development is approved with insufficient traffic mitigation measures in place, resulting in traffic delays, collisions, and the need for later intervention by a public agency.
7. **Applications in Other Areas of Transportation Research:** Regression is used in almost all areas of transportation research, including:
   - **Transportation Planning—**to create equations to predict trip generation and mode split.
   - **Traffic Safety—**to create equations to predict the number of collisions expected on a particular section of road.
   - **Pavement Engineering/Materials—**to predict long-term wear and condition of pavements.

Example 18: Transportation Planning; Logit and Related Analysis

Area: Transportation planning

Method of Analysis: Logit and related analysis (developing predictive models when the dependent variable is dichotomous—e.g., 0 or 1)

1. **Research Question/Problem Statement:** Transportation planners often utilize variations of the classic four-step transportation planning process for predicting travel demand. Trip generation, trip distribution, mode split, and trip assignment are used to predict traffic flows under a variety of forecasted changes in networks, population, land use, and controls. Mode split, deciding which mode of transportation a traveler will take, requires predicting mutually exclusive outcomes. For example, will a traveler utilize public transit or drive his or her own car?

Question/Issue

Can a linear model be developed that can be used to predict the probability that one of two choices will be made? In this example, the question is whether a household will use public transit (or not). Rather than being continuous (as in linear regression), the dependent variable is reduced to two categories, a dichotomous variable (e.g., yes or no, 0 or 1).

Although the question is simple, the statistical modeling becomes sophisticated very quickly.

2. **Identification and Description of Variables:** Considering a typical, traditional urban area in the United States, it is reasonable to argue that the likelihood of taking public transit to work (Y) will be a function of income (X). Generally, more income means less likelihood of taking public transit. This can be modeled using the following equation:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where X_i = family income,
$Y = 0$ if the family uses public transit, and
$Y = 1$ if the family doesn't use public transit.

3. **Data Collection:** These data normally are obtained from travel surveys conducted at the local level (e.g., by a metropolitan area or specific city), although the agency that collects the data often is a state DOT.
4. **Specification of Analysis Techniques and Data Analysis:** In this example the dependent variable is dichotomous and is a linear function of an explanatory variable.

Consider the equation $E(Y_i | X_i) = \beta_1 + \beta_2 X_i$.

Notice that if P_i = probability that Y = 1 (household utilizes transit), then $(1 - P_i)$ = probability that Y = 0 (doesn't utilize transit). This has been called a linear probability model. Note that within this expression, "*i*" refers to a household.

Thus, Y has the distribution shown in Table 25.

Any attempt to estimate this relationship with standard (OLS) regression is saddled with many problems (e.g., non-normality of errors, heteroscedasticity, and the possibility that the predicted Y will be outside the range 0 to 1, to say nothing of pretty terrible R^2 values).

Table 25. Distribution of Y.

Values that Y Takes	Probability	Meaning/Interpretation
1	P_i	Household uses transit
0	$1 - P_i$	Household does not use transit
	1.0	Total

68 Effective Experiment Design and Data Analysis in Transportation Research

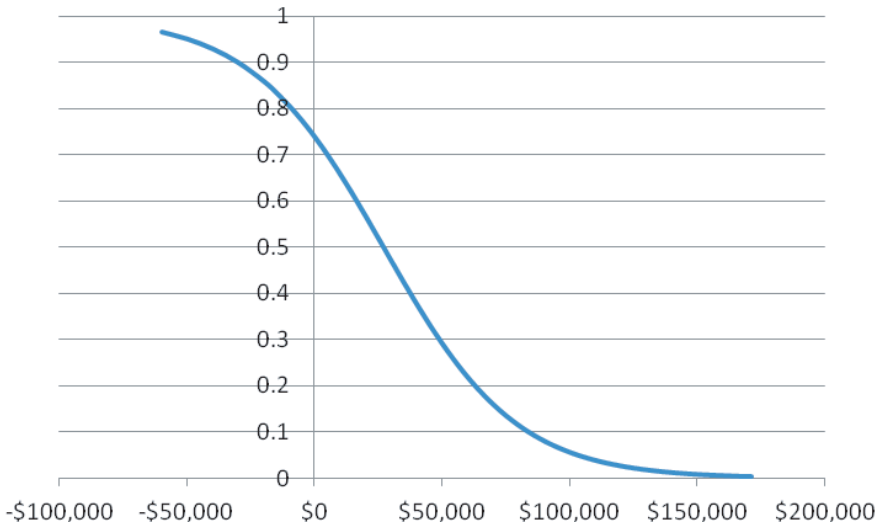

Figure 16. Plot of cumulative logistic distribution showing a lazy Z-curve.

An alternative formulation for estimating P_i, the cumulative logistic distribution, is expressed by the following equation:

$$P_i = \frac{1}{1+\varepsilon^{-(\beta_1+\beta_2 X_i)}}$$

This function can be plotted as a lazy *Z*-curve where on the left, with low values of X (low household income), the probability starts near 1 and ends at 0 (Figure 16). Notice that, even at 0 income, not all households use transit. The curve is said to be asymptotic to 1 and 0.

The value of P_i varies between 1 and 0 in relation to income, X. Manipulating the definition of the cumulative logistic distribution from above,

$$\left(1+\varepsilon^{-(\beta_1+\beta_2 X_i)}\right)P_i = 1$$

$$\left(P_i + P_i\varepsilon^{-(\beta_1+\beta_2 X_i)}\right) = 1$$

$$P_i\varepsilon^{-(\beta_1+\beta_2 X_i)} = 1 - P_i$$

$$\varepsilon^{-(\beta_1+\beta_2 X_i)} = \frac{1-P_i}{P_i}$$

and

$$\varepsilon^{(\beta_1+\beta_2 X_i)} = \frac{P_i}{1-P_i}$$

The final expression is the ratio of the probability of utilizing public transit divided by the probability of not utilizing public transit. It is called the odds ratio.

Next, taking the natural log of both sides (and reversing) results in the following equation:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = \beta_1 + \beta_2 X_i$$

L is called the logit, and this is called a logit model. The left side is the natural log of the odds ratio.

Unfortunately, this odds ratio is meaningless for individual households where the probability is either 0 or 1 (utilize or not utilize). If the analyst uses standard OLS regression on this

Table 26. Data examined by groups of households.

X_j ($)	N_j (Households)	n_j (Utilizing Transit)	P_j (Defined Above)
$6,000	40	30	0.750
$8,000	55	39	0.709
$10,000	65	43	0.662
$13,000	88	58	0.659
$15,000	118	69	0.585
$20,000	81	44	0.543
$25,000	70	33	0.471
$30,000	62	25	0.403
$35,000	40	16	0.400
$40,000	30	11	0.367
$50,000	22	6	0.273
$60,000	18	4	0.222
$75,000	12	2	0.167
Total:	701	380	

equation, with data for individual households, there is a problem because when P_i happens to equal either 0 or 1 (which is all the time!), the odds ratio will, as a result, equal either 0 or infinity (and the logarithm will be undefined) for all observations.

However, by using groups of households the problem can be mitigated. Table 26 presents data based on a survey of 701 households, more than half of which use transit (380). The income data are recorded for intervals; here, interval mid-points (X_j) are shown. The number of households in each income category is tallied (N_j), as is the number of households in each income category that utilizes public transit (n_j). It is important to note that while there are more than 700 households (i), the number of observations (categories, j) is only 13.

Using these data, for each income bracket, the probability of taking transit can be estimated as follows:

$$\hat{P}_j = \frac{n_j}{N_j}$$

This equation is an expression of relative frequency (i.e., it expresses the proportion in income bracket "j" using transit).

An examination of Table 26 shows clearly that there is progression of these relative frequencies, with higher income brackets showing lower relative frequencies, just as was hypothesized.

We can calculate the odds ratio for each income bracket listed in Table 26 and estimate the following logit function with OLS regression:

$$L_j = \ln\left(\frac{\frac{n_j}{N_j}}{1-\frac{n_j}{N_j}}\right) = \beta_1 + \beta_2 X_j$$

The results of this regression are shown in Table 27.

The results also can be expressed as an equation:

$$\text{LogOddsRatio} = 1.037 - 0.00003863 * \text{X}$$

5. **Interpreting the Results:** This model provides a very good fit. The estimates of the coefficients can be inserted in the original cumulative logistic function to directly estimate the probability of using transit for any given X (income level). Indeed, the logistic graph in Figure 16 is produced with the estimated function.

70 Effective Experiment Design and Data Analysis in Transportation Research

Table 27. Results of OLS regression.

Coefficients	*t*-values (statistics)	*p*-values	Measures of "Fit"
$\beta_1 = 1.037$	12.156	0.000	$R = 0.980$
$\beta_2 = -0.00003863$	-16.407	0.000	$R^2 = 0.961$
			adjusted $R^2 = 0.957$

6. **Conclusion and Discussion:** This approach to estimation is not without further problems. For example, the *N* within each income bracket needs to be sufficiently large that the relative frequency (and therefore the resulting odds ratio) is accurately estimated. Many statisticians would say that a minimum of 25 is reasonable. This approach also is limited by the fact that only one independent variable is used (income). Common sense suggests that the right-hand side of the function could logically be expanded to include more than one predictor variable (more Xs). For example, it could be argued that educational level might act, along with income, to account for the probability of using transit. However, combining predictor variables severely impinges on the categories (the *j*) used in this OLS regression formulation. To illustrate, assume that five educational categories are used in addition to the 13 income brackets (e.g., Grade 8 or less, high school graduate to Grade 9, some college, BA or BS degree, and graduate degree). For such an OLS regression analysis to work, data would be needed for 5×13, or 65 categories.

   Ideally, other travel modes should also be considered. In the example developed here, only transit and not-transit are considered. In some locations it is entirely reasonable to examine private auto versus bus versus bicycle versus subway versus light rail (involving five modes, not just two).

   This notion of a polychotomous logistic regression is possible. However, five modes cannot be estimated with the OLS regression technique employed above. The logit above is a variant of the binomial distribution and the polychotomous logistic model is a variant of the multinomial distribution (see NCHRP Project 20-45, Volume 2, Chapter 5). Estimation of these more advanced models requires maximum likelihood methods (as described in NCHRP Project 20-45, Volume 2, Chapter 5).

   Other model variants are based upon other cumulative probability distributions. For example, there is the probit model, in which the normal cumulative density function is used. The probit model is very similar to the logit model, but it is more difficult to estimate.
7. **Applications in Other Areas of Transportation Research:** Applications of logit and related models abound within transportation studies. In any situation in which human behavior is relegated to discrete choices, the category of models may be applied. Examples in other areas of transportation research include:
   - **Transportation Planning—**to model any "choice" issue, such as shopping destination choices.
   - **Traffic Safety—**to model dichotomous responses (e.g., did a motorist slow down or not) in response to traffic control devices.
   - **Highway Design—**to model public reactions to proposed design solutions (e.g., support or not support proposed road diets, installation of roundabouts, or use of traffic calming techniques).

Example 19: Public Transit; Survey Design and Analysis

Area: Public transit

Method of Analysis: Survey design and analysis (organizing survey data for statistical analysis)

1. **Research Question/Problem Statement:** The transit director is considering changes to the fare structure and the service characteristics of the transit system. To assist in determining which changes would be most effective or efficient, a survey of the current transit riders is developed.

Question/Issue

Use and analysis of data collected in a survey. Results from a survey of transit users are used to estimate the change in ridership that would result from a change in the service or fare.

2. **Identification and Description of Variables:** Two types of variables are needed for this analysis. The first is data on the characteristics of the riders, such as gender, age, and access to an automobile. These data are discrete variables. The second is data on the riders' stated responses to proposed changes in the fare or service characteristics. These data also are treated as discrete variables. Although some, like the fare, could theoretically be continuous, they are normally expressed in discrete increments (e.g., $1.00, $1.25, $1.50).
3. **Data Collection:** These data are normally collected by agencies conducting a survey of the transit users. The initial step in the experiment design is to choose the variables to be collected for each of these two data sets. The second step is to determine how to categorize the data. Both steps are generally based on past experience and common sense.

   Some of the variables used to describe the characteristics of the transit user are dichotomous, such as gender (male or female) and access to an automobile (yes or no). Other variables, such as age, are grouped into discrete categories within which the transit riding characteristics are similar. For example, one would not expect there to be a difference between the transit trip needs of a 14-year-old student and a 15-year-old student. Thus, the survey responses of these two age groups would be assigned to the same age category. However, experience (and common sense) leads one to differentiate a 19-year-old transit user from a 65-year-old transit user, because their purposes for taking trips and their perspectives on the relative value of the fare and the service components are both likely to be different.

   Obtaining user responses to changes in the fare or service is generally done in one of two ways. The first is to make a statement and ask the responder to mark one of several choices: strongly agree, agree, neither agree nor disagree, disagree, and strongly disagree. The number of statements used in the survey depends on how many parameter changes are being contemplated. Typical statements include:

   1. I would increase the number of trips I make each month if the fare were reduced by $0.xx.
   2. I would increase the number of trips I make each month if I could purchase a monthly pass.
   3. I would increase the number of trips I make each month if the waiting time at the stop were reduced by 10 minutes.
   4. I would increase the number of trips I make each month if express services were available from my origin to my destination.

   The second format is to propose a change and provide multiple choices for the responder. Typical questions for this format are:

   1. If the fare were increased by $0.xx per trip I would:
      a) not change the number of trips per month
      b) reduce the non-commute trips
      c) reduce both the commute and non-commute trips
      d) switch modes
   2. If express service were offered for an additional $0.xx per trip I would:
      a) not change the number of trips per month on this local service
      b) make additional trips each month
      c) shift from the local service to the express service

72 Effective Experiment Design and Data Analysis in Transportation Research

Table 28. Table of responses to sample statement, "I would increase the number of trips I make each month if the fare were reduced by $0.xx."

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Total responses	450	600	300	400	100

These surveys generally are administered by handing a survey form to people as they enter the transit vehicle and collecting them as people depart the transit vehicle. The surveys also can be administered by mail, telephone, or in a face-to-face interview.

In constructing the questions, care should be taken to use terms with which the respondents will be familiar. For example, if the system does not currently offer "express" service, this term will need to be defined in the survey. Other technical terms should be avoided. Similarly, the word "mode" is often used by transportation professionals but is not commonly used by the public at large.

The length of a survey is almost always an issue as well. To avoid asking too many questions, each question needs to be reviewed to see if it is really necessary and will produce useful data (as opposed to just being something that would be nice to know).

4. **Specification of Analysis Technique and Data Analysis:** The results of these surveys often are displayed in tables or in frequency distribution diagrams (see also Example 1 and Example 2). Table 28 lists responses to a sample question posed in the form of a statement.

Figure 17 shows the frequency diagram for these data.

Similar presentations can be made for any of the groupings included in the first type of variables discussed above. For example, if gender is included as a Type 1 question, the results might appear as shown in Table 29 and Figure 18.

Figure 18 shows the frequency diagram for these data.

Presentations of the data can be made for any combination of the discrete variable groups included in the survey. For example, to display responses of female users over 65 years old,

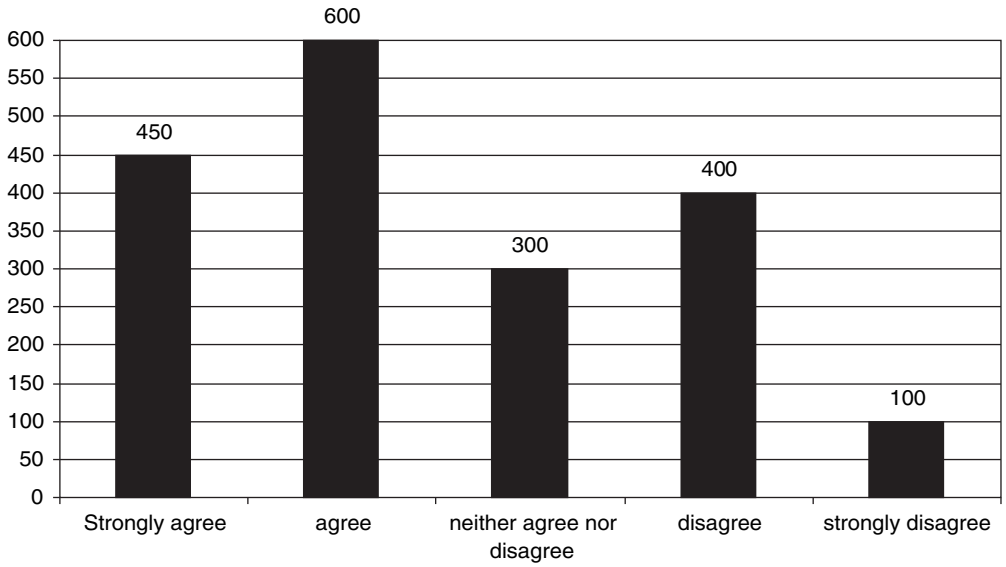

Figure 17. Frequency diagram for total responses to sample statement.

Table 29. Contingency table showing responses by gender to sample statement, "I would increase the number of trips I make each month if the fare were reduced by $0.xx."

	Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree
Male	200	275	200	200	70
Female	250	325	100	200	30
Total responses	450	600	300	400	100

all of the survey forms on which these two characteristics (female and over 65 years old) are checked could be extracted and recorded in a table and shown in a frequency diagram.

5. **Interpreting the Results:** Survey data can be used to compare the responses to fare or service changes of different groups of transit users. This flexibility can be important in determining which changes would impact various segments of transit users. The information can be used to evaluate various fare and service options being considered and allows the transit agency to design promotions to obtain the greatest increase in ridership. For example, by creating frequency diagrams to display the responses to statements 2, 3, and 4 listed in Section 3, the engineer can compare the impact of changing the fare versus changing the headway or providing express services in the corridor.

   Organizing response data according to different characteristics of the user produces contingency tables like the one illustrated for males and females. This table format can be used to conduct chi-square analysis to determine if there is any statistically significant difference among the various groups. (Chi-square analysis is described in more detail in Example 4.)
6. **Conclusions and Discussion:** This example illustrates how to obtain and present quantitative information using surveys. Although survey results provide reasonably good estimates of the relative importance users place on different transit attributes (fare, waiting time, hours of service, etc.), when determining how often they would use the system, the magnitude of users' responses often is overstated. Experience shows that what users say they would do (their stated preference) generally is different than what they actually do (their revealed preference).

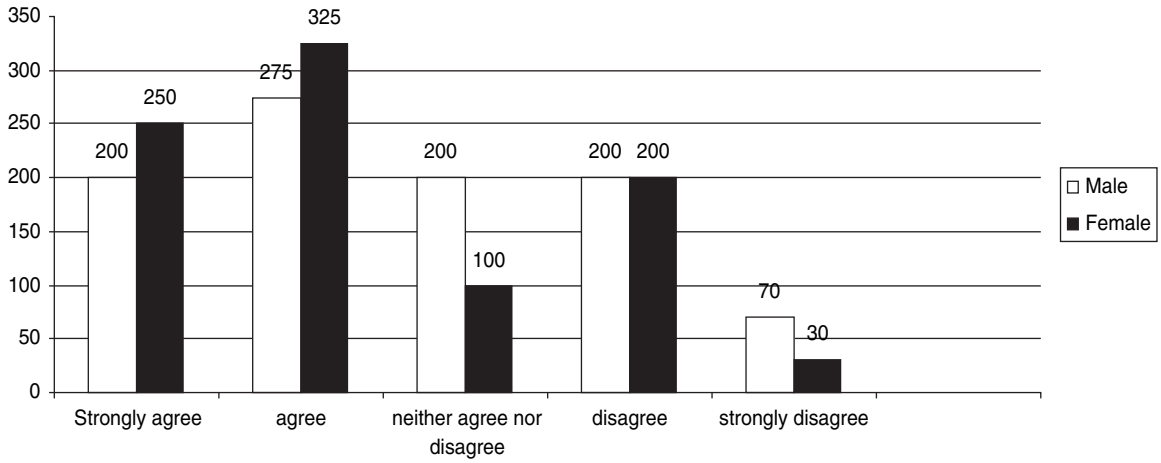

Figure 18. Frequency diagram showing responses by gender to sample statement.

In this example, 1,050 of the 1,850 respondents (57%) have responded that they would use the bus service more frequently if the fare were decreased by $0.xx. Five hundred respondents (27%) have indicated that they would not use the bus service more frequently, and 300 respondents (16%) have indicated that they are not sure if they would change their bus use frequency. These percentages show the stated preferences of the users. The engineer does not yet know the revealed preferences of the users, but experience suggests that it is unlikely that 57% of the riders would actually increase the number of trips they make.

7. **Applications in Other Area in Transportation:** Survey design and analysis techniques can be used to collect and present data in many areas of transportation research, including:
   - **Transportation Planning—**to assess public response to a proposal to enact a local motor fuel tax to improve road maintenance in a city or county.
   - **Traffic Operations—**to assess public response to implementing road diets (e.g., 4-lane to 3-lane conversions) on different corridors in a city.
   - **Highway Design—**to assess public response to proposed alternative cross-section designs, such as a boulevard design versus an undivided multilane design in a corridor.

Example 20: Traffic Operations; Simulation

Area: Traffic operations

Method of Analysis: Simulation (using field data to simulate, or model, operations or outcomes)

1. **Research Question/Problem Statement:** A team of engineers wants to determine whether one or more unconventional intersection designs will produce lower travel times than a conventional design at typical intersections for a given number of lanes. There is no way to collect field data to compare alternative intersection designs at a particular site. Macroscopic traffic operations models like those in the *Highway Capacity Manual* do a good job of estimating delay at specific points but are unable to provide travel time estimates for unconventional designs that consist of several smaller intersections and road segments. Microscopic simulation models measure the behaviors of individual vehicles as they traverse the highway network. Such simulation models are therefore very flexible in the types of networks and measures that can be examined. The team in this example turns to a simulation model to determine how other intersection designs might work.

Question/Issue

Developing and using a computer simulation model to examine operations in a computer environment. In this example, a traffic operations simulation model is used to show whether one or more unconventional intersection designs will produce lower travel times than a conventional design at typical intersections for a given number of lanes.

2. **Identification and Description of Variables:** The engineering team simulates seven different intersections to provide the needed scope for their findings. At each intersection, the team examines three different sets of traffic volumes: volumes from the evening (p.m.) peak hour, a typical midday off-peak hour, and a volume that is 15% greater than the p.m. peak hour to represent future conditions. At each intersection, the team models the current conventional intersection geometry and seven unconventional designs: the quadrant roadway, median U-turn, superstreet, bowtie, jughandle, split intersection, and continuous flow intersection.

   Traffic simulation models break the roadway network into nodes (intersections) and links (segments between intersections). Therefore, the engineering team has to design each of the

alternatives at each test site in terms of numbers of lanes, lane lengths, and such, and then faithfully translate that geometry into links and nodes that the simulation model can use. For each combination of traffic volume and intersection design, the team uses software to find the optimum signal timing and uses that during the simulation. To avoid bias, the team keeps all other factors (e.g., network size, numbers of lanes, turn lane lengths, truck percentages, average vehicle speeds) constant in all simulation runs.

3. **Data Collection:** The field data collection necessary in this effort consists of noting the current intersection geometries at the seven test intersections and counting the turning movements in the time periods described above.

   In many simulation efforts, it is also necessary to collect field data to calibrate and validate the simulation model. Calibration is the process by which simulation output is compared to actual measurements for some key measure(s) such as travel time. If a difference is found between the simulation output and the actual measurement, the simulation inputs are changed until the difference disappears. Validation is a test of the calibrated simulation model, comparing simulation output to a previously unused sample of actual field measurements. In this example, however, the team determines that it is unnecessary to collect calibration and validation data because a recent project has successfully calibrated and validated very similar models of most of these same unconventional designs.

   The engineer team uses the CORSIM traffic operations simulation model. Well known and widely used, CORSIM models the movement of each vehicle through a specified network in small time increments. CORSIM is a good choice for this example because it was originally designed for problems of this type, has produced appropriate results, has excellent animation and other debugging features, runs quickly in these kinds of cases, and is well-supported by the software developers.

   The team makes two CORSIM runs with different random number seeds for each combination of volume and design at each intersection, or 48 runs for each intersection altogether. It is necessary to make more than one run (or replication) of each simulation combination with different random number seeds because of the randomness built into simulation models. The experiment design in this case allows the team to reduce the number of replications to two; typical practice in simulations when one is making simple comparisons between two variables is to make at least 5 to 10 replications. Each run lasts 30 simulated minutes.

   Table 30 shows the simulation data for one of the seven intersections. The lowest travel time produced in each case is bolded. Notice that Table 30 does not show data for the bowtie design. That design became congested (gridlocked) and produced essentially infinite travel times for this intersection. Handling overly congested networks is a difficult problem in many efforts and with several different simulation software packages. The best current advice is for analysts to not push their networks too hard and to scan often for gridlock.

4. **Specification of Analysis Technique and Data Analysis:** The experiment assembled in this example uses a factorial design. (Factorial design also is discussed in Example 11.) The team analyzes the data from this factorial experiment using analysis of variance (ANOVA). Because

Table 30. Simulation results for different designs and time of day.

Time of Day	Total Travel Time, Vehicle-hours, Average of Two Simulation Runs						
	Conventional	Quadrant	Median U	Superstreet	Jughandle	Split	Continuous
Midday	67	64	61	74	63	**59***	75
P.M. peak	121	**95**	119	179	139	114	106
Peak + 15%	170	**135**	145	245	164	180	142

*Lowest total travel time.

the experimenter has complete control in a simulation, it is common to use efficient designs like factorials and efficient analysis methods like ANOVA to squeeze all possible information out of the effort. Statistical tests comparing the individual mean values of key results by factor are common ways to follow up on ANOVA results. Although ANOVA will reveal which factors make a significant contribution to the overall variance in the dependent variable, means tests will show which levels of a significant factor differ from the other levels. In this example, the team uses Tukey's means test, which is available as part of the battery of standard tests accompanying ANOVA in statistical software. (For more information about ANOVA, see NCHRP Project 20-45, Volume 2, Chapter 4, Section A.)

5. **Interpreting the Results:** For the data shown in Table 30, the ANOVA reveals that the volume and design factors are statistically significant at the 99.99% confidence level. Furthermore, the interaction between the volume and design factors also is statistically significant at the 99.99% level. The means tests on the design factors show that the quadrant roadway is significantly different from (has a lower overall travel time than) the other designs at the 95% level. The next-best designs overall are the median U-turn and the continuous flow intersection; these are not statistically different from each other at the 95% level. The third tier of designs consists of the conventional and the split, which are statistically different from all others at the 95% level but not from each other. Finally, the jughandle and the superstreet designs are statistically different from each other and from all other designs at the 95% level according to the means test.

   Through the simulation, the team learns that several designs appear to be more efficient than the conventional design, especially at higher volume levels. From the results at all seven intersections, the team sees that the quadrant roadway and median U-turn designs generally lead to the lowest travel times, especially with the higher volume levels.

6. **Conclusion and Discussion:** Simulation is an effective tool to analyze traffic operations, as at the seven intersections of interest in this example. No other tool would allow such a robust comparison of many different designs and provide the results for travel times in a larger network rather than delays at a single spot. The simulation conducted in this example also allows the team to conduct an efficient factorial design, which maximizes the information provided from the effort.

   Simulation is a useful tool in research for traffic operations because it
   - affords the ability to conduct randomized experiments,
   - allows the examination of details that other methods cannot provide, and
   - allows the analysis of large and complex networks.

   In practice, simulation also is popular because of the vivid and realistic animation output provided by common software packages. The superb animations allow analysts to spot and treat flaws in the design or model and provide agencies an effective tool by which to share designs with politicians and the public.

   Although simulation results can sometimes be surprising, more often they confirm what the analysts already suspect based on simpler analyses. In the example described here, the analysts suspected that the quadrant roadway and median U-turn designs would perform well because these designs had performed well in prior *Highway Capacity Manual* calculations. In many studies, simulations provide rich detail and vivid animation but no big surprises.

7. **Applications in Other Areas of Transportation Research:** Simulations are critical analysis methods in several areas of transportation research. Besides traffic operations, simulations are used in research related to:
   - **Maintenance—**to model the lifetime performance of traffic signs.
   - **Traffic Safety**
     - to examine vehicle performance and driver behaviors or performance.
     - to predict the number of collisions from a new roadway design (potentially, given the recent development of the FHWA SSAM program).

Example 21: Traffic Safety; Non-parametric Methods

Area: Traffic safety

Method of Analysis: Non-parametric methods (methods used when data do not follow assumed or conventional distributions, such as when comparing median values)

1. **Research Question/Problem Statement:** A city traffic engineer has been receiving many citizen complaints about the perceived lack of safety at unsignalized midblock crosswalks. Apparently, some motorists seem surprised by pedestrians in the crosswalks and do not yield to the pedestrians. The engineer believes that larger and brighter warning signs may be an inexpensive way to enhance safety at these locations.

Question/Issue

Determine whether some treatment has an effect when data to be tested do not follow known distributions. In this example, a nonparametric method is used to determine whether larger and brighter warning signs improve pedestrian safety at unsignalized midblock crosswalks. The null hypothesis and alternative hypothesis are stated as follows:

H_o: There is no difference in the median values of the number of conflicts before and after a treatment.
H_a: There is a difference in the median values.

2. **Identification and Description of Variables:** The engineer would like to collect collision data at crosswalks with improved signs, but it would take a long time at a large sample of crosswalks to collect a reasonable sample size of collisions to answer the question. Instead, the engineer collects data for conflicts, which are near-collisions when one or both of the involved entities brakes or swerves within 2 seconds of a collision to avoid the collision. Research literature has shown that conflicts are related to collisions, and because conflicts are much more numerous than collisions, it is much quicker to collect a good sample size. Conflict data are not nearly as widely used as collision data, however, and the underlying distribution of conflict data is not clear. Thus, the use of non-parametric methods seems appropriate.
3. **Data Collection:** The engineer identifies seven test crosswalks in the city based on large pedestrian volumes and the presence of convenient vantage points for observing conflicts. The engineering staff collects data on traffic conflicts for 2 full days at each of the seven crosswalks with standard warning signs. The engineer then has larger and brighter warning signs installed at the seven sites. After waiting at least 1 month at each site after sign installation, the staff again collects traffic conflicts for 2 full days, making sure that weather, light, and as many other conditions as possible are similar between the before-and-after data collection periods at each site.
4. **Specification of Analysis Technique and Data Analysis:** A nonparametric statistical test is an efficient way to analyze data when the underlying distribution is unclear (as in this example using conflict data) and when the sample size is small (as in this example with its small number of sites). Several such tests, such as the sign test and the Wilcoxon signed-rank (Wilcoxon rank-sum) test are plausible in this example. (For more information about nonparametric tests, see NCHRP Project 20-45, Volume 2, Chapter 6, Section D, "Hypothesis About Population Medians for Independent Samples.") The decision is made to use the Wilcoxon signed-rank test because it is a more powerful test for paired numerical measurements than other tests, and this example uses paired (before-and-after) measurements. The sign test is a popular nonparametric test for paired data but loses information contained in numerical measurements by reducing the data to a series of positive or negative signs.

Table 31. Number of conflicts recorded during each (equal) time period at each site.

	Site 1	Site 2	Site 3	Site 4	Site 5	Site 6	Site 7
Standard signs	170	39	35	32	32	19	45
Larger and brighter signs	155	26	33	29	25	31	61
Difference	15	7	2	3	7	-12	-16
Rank of absolute difference	6	3.5	1	2	3.5	5	7

Having decided on the Wilcoxon signed-rank test, the engineer arranges the data (see Table 31). The third row of the table is the difference between the frequencies of the two conflict measurements at each site. The last row shows the rank order of the sites from lowest to highest based on the absolute value of the difference. Site 3 has the least difference ($35 - 33 = 2$) while Site 7 has the greatest difference ($54 - 61 = -16$).

The Wilcoxon signed-rank test ranks the differences from low to high in terms of absolute values. In this case, that would be 2, 3, 7, 7, 12, 15, and 16. The test statistic, x, is the sum of the ranks that have positive differences. In this example, $x = 1 + 2 + 3.5 + 3.5 + 6 = 16$. Notice that all but the sixth and seventh ranked sites had positive differences. Notice also that the tied differences were assigned ranks equal to the average of the ranks they would have received if they were just slightly different from each other. The engineer then consults a table for the Wilcoxon signed-rank test to get a critical value against which to compare. (Such a table appears in NCHRP Project 20-45, Volume 2, Appendix C, Table C-8.) The standard table for a sample size of seven shows that the critical value for a one-tailed test (testing whether there is an improvement) with a confidence level of 95% is $x = 24$.

5. **Interpreting the Results:** Because the calculated value ($x = 16$) is less than the critical value ($x = 24$), the engineer concludes that there is not a statistically significant difference between the number of conflicts recorded with standard signs and the number of conflicts recorded with larger and brighter signs.
6. **Conclusion and Discussion:** Nonparametric tests do not require the engineer to make restrictive assumptions about an underlying distribution and are therefore good choices in cases like this, in which the sample size is small and the data collected do not have a familiar underlying distribution. Many nonparametric tests are available, so analysts should do some reading and searching before settling on the best one for any particular case. Once a nonparametric test is determined, it is usually easy to apply.

   This example also illustrates one of the potential pitfalls of statistical testing. The engineer's conclusion is that there is not a statistically significant difference between the number of conflicts recorded with standard signs and the number of conflicts recorded with larger and brighter signs. That conclusion does not necessarily mean that larger and brighter signs are a bad idea at sites similar to those tested. Notice that in this experiment, larger and brighter signs produced lower conflict frequencies at five of the seven sites, and the average number of conflicts per site was lower with the larger and brighter signs. Given that signs are relatively inexpensive, they may be a good idea at sites like those tested. A statistical test can provide useful information, especially about the quality of the experiment, but analysts must be careful not to interpret the results of a statistical test too strictly.

   In this example, the greatest danger to the validity of the test result lies not in the statistical test but in the underlying before-and-after test setup. For the results to be valid, it is necessary that the only important change that affects conflicts at the test sites during data collection be

the new signs. The engineer has kept the duration short between the before-and-after data collection periods, which helps minimize the chances of other important changes. However, if there is any reason to suspect other important changes, these test results should be viewed skeptically and a more sophisticated test strategy should be employed.

7. **Applications in Other Areas of Transportation Research:** Nonparametric tests are helpful when researchers are working with small sample sizes or sample data wherein the underlying distribution is unknown. Examples of other areas of transportation research in which nonparametric tests may be applied include:
   - **Transportation Planning, Public Transportation—**to analyze data from surveys and questionnaires when the scale of the response calls into question the underlying distribution. Such data are often analyzed in transportation planning and public transportation.
   - **Traffic Operations—**to analyze small samples of speed or volume data.
   - **Structures, Pavements—**to analyze quality ratings of pavements, bridges, and other transportation assets. Such ratings also use scales.

Resources

The examples used in this report have included references to the following resources. Researchers are encouraged to consult these resources for more information about statistical procedures.

Freund, R. J. and W. J. Wilson (2003). *Statistical Methods.* 2d ed. Burlington, MA: Academic Press. *See* page 256 for a discussion of Tukey's procedure.

Kutner, M. et al. (2005). *Applied Linear Statistical Models.* 5th ed. Boston: McGraw-Hill. *See* page 746 for a discussion of Tukey's procedure.

NCHRP CD-22: Scientific Approaches to Transportation Research, Vol. 1 and 2. 2002. Transportation Research Board of the National Academies, Washington, D.C. This two-volume electronic manual developed under NCHRP Project 20-45 provides a comprehensive source of information on the conduct of research. The manual includes state-of-the-art techniques for problem statement development; literature searching; development of the research work plan; execution of the experiment; data collection, management, quality control, and reporting of results; and evaluation of the effectiveness of the research, as well as the requirements for the systematic, professional, and ethical conduct of transportation research. For readers' convenience, the references to NCHRP Project 20-45 from the various examples contained in this report are summarized here by topic and location in *NCHRP CD-22.* More information about *NCHRP CD-22* is available at http://www.trb.org/Main/Blurbs/152122.aspx.

- Analysis of Variance (one-way ANOVA and two-way ANOVA): *See* Volume 2, Chapter 4, Section A, Analysis of Variance Methodology (pp. 113, 119–31).
- Assumptions for residual errors: *See* Volume 2, Chapter 4.
- Box plots; Q-Q plots: *See* Volume 2, Chapter 6, Section C.
- Chi-square test: *See* Volume 2, Chapter 6, Sections E (Chi-Square Test for Independence) and F.
- Chi-square values: *See* Volume 2, Appendix C, Table C-2.
- Computations on unbalanced designs and multi-factorial designs: *See* Volume 2, Chapter 4, Section A, Analysis of Variance Methodology (pp. 119–31).
- Confidence intervals: *See* Volume 2, Chapter 4.
- Correlation coefficient: *See* Volume 2, Appendix A, Glossary, Correlation Coefficient.
- Critical *F*-value: *See* Volume 2, Appendix C, Table C-5.
- Desirable and undesirable residual plots (scatter plots): *See* Volume 2, Chapter 4, Section B, Figure 6.

- Equation fit: *See* Volume 2, Chapter 4, Glossary, Descriptive Measures of Association Between X and Y.
- Error distributions (normality, constant variance, uncorrelated, etc.): *See* Volume 2, Chapter 4 (pp. 146–55).
- Experiment design and data collection: *See* Volume 2, Chapter 1.
- F_{crit} and *F*-distribution table: *See* Volume 2, Appendix C, Table C-5.
- *F*-test (or F-test): *See* Volume 2, Chapter 4, Section A, Compute the F-ratio Test Statistic (p. 124).
- Formulation of formal hypotheses for testing: *See* Volume 1, Chapter 2, Hypothesis; Volume 2, Appendix A, Glossary.
- History and maturation biases (specification errors): *See* Volume 2, Chapter 1, Quasi-Experiments.
- Indicator (dummy) variables: *See* Volume 2, Chapter 4 (pp. 142–45).
- Intercept and slope: *See* Volume 2, Chapter 4 (pp. 140–42).
- Maximum likelihood methods: *See* Volume 2, Chapter 5 (pp. 208–11).
- Mean and standard deviation formulas: *See* Volume 2, Chapter 6, Table C, Frequency Distributions, Variance, Standard Deviation, Histograms, and Boxplots.
- Measured ratio or interval scale: *See* Volume 2, Chapter 1 (p. 83).
- Multinomial distribution and polychotomous logistical model: *See* Volume 2, Chapter 5 (pp. 211–18).
- Multiple (multivariate) regression: *See* Volume 2, Chapter 4, Section B.
- Non-parametric tests: *See* Volume 2, Chapter 6, Section D.
- Normal distribution: *See* Volume 2, Appendix A, Glossary, Normal Distribution.
- One- and two-sided hypothesis testing (one- and two-tail test values): *See* Volume 2, Chapter 4 (pp. 161 and 164–5).
- Ordinary least squares (OLS) regression: *See* Volume 2, Chapter 4, Section B, Linear Regression.
- Sample size and confidence: *See* Volume 2, Chapter 1, Sample Size Determination.
- Sample size determination based on statistical power requirements: *See* Volume 2, Chapter 1, Sample Size Determination (p. 94).
- Sign test and the Wilcoxon signed-rank (Wilcoxon rank-sum) test: *See* Volume 2, Chapter 6, Section D, and Appendix C, Table C-8, Hypothesis About Population Medians for Independent Samples.
- Split samples: *See* Volume 2, Chapter 4, Section A, Analysis of Variance Methodology (pp. 119–31).
- Standard chi-square distribution table: *See* Volume 2, Appendix C, Table C-2.
- Standard normal values: *See* Volume 2, Appendix C, Table C-1.
- t_{crit} values: *See* Volume 2, Appendix C, Table C-4.
- *t*-statistic: *See* Volume 2, Appendix A, Glossary.
- *t*-statistic using equation for equal variance: *See* Volume 2, Appendix C, Table C-4.
- *t*-test: *See* Volume 2, Chapter 4, Section B, How are *t*-statistics Interpreted?
- Tabularized values of *t*-statistic: *See* Volume 2, Appendix C, Table C-4.
- Tukey's test, Bonferroni's test, Scheffe's test: *See* Volume 2, Chapter 4, Section A, Analysis of Variance Methodology (pp. 119–31).
- Types of data and implications for selection of analysis techniques: *See* Volume 2, Chapter 1, Identification of Empirical Setting.

Abbreviations and acronyms used without definitions in TRB publications:

AAAE	American Association of Airport Executives
AASHO	American Association of State Highway Officials
AASHTO	American Association of State Highway and Transportation Officials
ACI–NA	Airports Council International–North America
ACRP	Airport Cooperative Research Program
ADA	Americans with Disabilities Act
APTA	American Public Transportation Association
ASCE	American Society of Civil Engineers
ASME	American Society of Mechanical Engineers
ASTM	American Society for Testing and Materials
ATA	American Trucking Associations
CTAA	Community Transportation Association of America
CTBSSP	Commercial Truck and Bus Safety Synthesis Program
DHS	Department of Homeland Security
DOE	Department of Energy
EPA	Environmental Protection Agency
FAA	Federal Aviation Administration
FHWA	Federal Highway Administration
FMCSA	Federal Motor Carrier Safety Administration
FRA	Federal Railroad Administration
FTA	Federal Transit Administration
HMCRP	Hazardous Materials Cooperative Research Program
IEEE	Institute of Electrical and Electronics Engineers
ISTEA	Intermodal Surface Transportation Efficiency Act of 1991
ITE	Institute of Transportation Engineers
NASA	National Aeronautics and Space Administration
NASAO	National Association of State Aviation Officials
NCFRP	National Cooperative Freight Research Program
NCHRP	National Cooperative Highway Research Program
NHTSA	National Highway Traffic Safety Administration
NTSB	National Transportation Safety Board
PHMSA	Pipeline and Hazardous Materials Safety Administration
RITA	Research and Innovative Technology Administration
SAE	Society of Automotive Engineers
SAFETEA-LU	Safe, Accountable, Flexible, Efficient Transportation Equity Act: A Legacy for Users (2005)
TCRP	Transit Cooperative Research Program
TEA-21	Transportation Equity Act for the 21st Century (1998)
TRB	Transportation Research Board
TSA	Transportation Security Administration
U.S.DOT	United States Department of Transportation