DEVELOPMENT OF A GUIDE TO STATISTICS IN MAINTENANCE QUALITY ASSURANCE PROGRAMS IN TRANSPORTATION

Project 06-04
April 2006

Midwest Regional University Transportation Center
College of Engineering
Department of Civil and Environmental Engineering
University of Wisconsin, Madison

Authors:
Robert L. Schmitt, University of Wisconsin, Platteville
Samuel Owusu-Ababio, University of Wisconsin, Platteville
Richard M. Weed, Consultant, formerly New Jersey DOT
Erik V. Nordheim, University of Wisconsin, Madison

Principal Investigator:
Robert L. Schmitt, Associate Professor, Department of Civil and Environmental Engineering, University of Wisconsin, Platteville

DISCLAIMER

This research was funded by the Midwest Regional University Transportation Center. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the Department of Transportation, University Transportation Centers Program, in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The contents do not necessarily reflect the official views of the Midwest Regional University Transportation Center, the University of Wisconsin, the Wisconsin Department of Transportation, or the USDOT's RITA at the time of publication.

The United States Government assumes no liability for its contents or use thereof. This report does not constitute a standard, specification, or regulation.

The United States Government does not endorse products or manufacturers. Trade and manufacturers names appear in this report only because they are considered essential to the object of the document.

1. Report No.	2. Govt. Accession No.	3 Recipient's Catalog No. CFDA 20.701	
4. Title and Subtitle Development of a Guide to Statistics for Maintenance Quality Assurance Programs in Transportation		5. Report Date April 2006	
		6. Performing Organization Code	
7. Authors Schmitt, Robert L., Owusu-Ababio, Sam , Weed, Richard M., and Nordheim, Erik V.		8. Performing Organization Report No. MRUTC 06-04	
9. Performing Organization Name and Address Midwest Regional University Transportation Center University of Wisconsin-Madison 1415 Engineering Drive, Madison, WI 53706		10 Work Unit No.	
		11. Contract or Grant No. DTRS 99-G-0005	
12. Sponsoring Agency Name and Address U.S. Department of Transportation Research and Special Programs Administration 400 7th Street, SW Washington, DC 20590-0001		13. Type of Report and Period Covered Research Report May 2005 to April 2006	
		14. Sponsoring Agency Code	
15. Supplementary Notes Project completed for the Midwest Regional University Transportation Center with support from the Wisconsin Department of Transportation.			
16. Abstract This report provides maintenance managers and practitioners with knowledge of how to apply statistics in MQA programs. Literature were reviewed and MQA manuals from 10 states were synthesized to understand state-of-practice for managing statistics in MQA programs. It was observed among lead states that a wide range of measured elements and threshold definitions exist for roadway, roadside and vegetation, drainage, traffic control, and rest areas. The role of statistics in MQA was described, and key statistical terms were defined. Actual MQA data were collected and analyzed to illustrate how fundamental statistical procedures and applications to describe features of MQA data. Examples were demonstrated for basic summary statistics, confidence intervals, data stratification, analysis of variance, sample size determination, precision, sensitivity analysis, and power level of statistics when evaluating the QA process itself. A new ranking procedure was developed that applies the concepts of percent defective (PD) or percent within limits (PWL), and computes confidence limits for the statewide system or any strata (functional class, division, county, or 10-mile roadway segment), and then ordering them by their lower confidence limits. If those limits do not include the desired level of the particular measure, then that particular section of pavement would be considered noncompliant.			
17. Key Words Maintenance, Quality Assurance, Statistics, Applications		18. Distribution Statement No restrictions. This report is available through the Transportation Research Information Services of the National Transportation Library.	
19. Security Classification (of this report) Unclassified	20. Security Classification (of this page) Unclassified	21. No. of Pages 115	22. Price -0-

Form DOT F 1700.7 (8-72) **Reproduction of form and completed page is authorized.**

EXECUTIVE SUMMARY

In recent years, transportation Maintenance Quality Assurance (MQA) programs have been developed to assure that maintenance quality is being achieved. MQA programs have to be capable of detecting insufficient maintenance efforts, poor material performance, and incorrect procedures when evaluating end-product performance. At the October 2004 Maintenance Quality Assurance Peer Exchange held at Madison, Wisconsin, participants raised questions regarding use of statistics in MQA programs. The purpose of this guidebook is to answer these questions and provide maintenance practitioners with knowledge of how to understand and use statistics in MQA programs.

Literature were reviewed and MQA manuals from 10 states were synthesized to understand state-of-practice for managing statistics in MQA programs. The major maintenance categories with quality assurance programs in-place include roadway, roadside and vegetation, drainage, traffic control, and rest areas. It was observed that among lead states, a wide range of threshold definitions exist for features or characteristics monitored for each maintenance category. In addition, for the same maintenance categories, considerable variation exists among the states in the types of features monitored. Statistics have been applied to MQA programs in four main areas including, sampling for customer surveys, sampling facilities for condition assessment, level of service ratings and analyses, and quality assurance testing for automated data collection and processing.

Fundamental statistical procedures are presented to describe features of MQA data, create confidence intervals, and conduct formal statistical tests. These fundamental procedures provide the foundation for the applications where specific questions facing MQA programs are addressed. Data were collected from four states, then data were analyzed from two select states, North Carolina and Wisconsin, to demonstrate how an agency can understand the statistics associated with their MQA program. Traditional statistical tools and methods are applied to the data. Examples using actual data are demonstrated for basic summary statistics, data stratification, analysis of variance (ANOVA), sample size determination, precision, sensitivity analysis, and power level of statistics when evaluating the QA process itself.

A new ranking procedure was developed during this project to take an alternative approach to traditional methods. This prototype procedure applies the concepts of percent defective (PD) or percent within limits (PWL), and computing confidence limits for the statewide system or any strata (functional class, division, county, or 10-mile roadway segment), and then ordering them by their lower confidence limits. If those limits do not include the desired level of the particular measure, then that particular section of pavement would be judged to be noncompliant.

The electronic web-based guidebook is structured with eight chapters to readily find relevant material of interest. The guidebook, created with the eight-chapter structure, has numerous links to access the material. Figures, tables, and equation captions have been created for the electronic guidebook to comply with handicap-accessible requirement in U.S. Section 508 Code. The electronic guidebook can be found at http://www.mrutc.org.

ACKNOWLEDGMENTS

The authors thank the state highway agencies for their contribution to this project. A special thanks goes to the agencies that provided data and both time and assistance in this effort, including California, Florida, Kansas, New York, North Carolina, Texas, Utah, Virginia, Washington, and Wisconsin. The authors thank the project panel for their guidance and input throughout the project. The authors also thank Mr. David Dochterman and Mr. Ryan Horst, engineering students at UW-Platteville, for their assistance preparing this guidebook.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

1.1 Background and Problem Statement

One of the primary components of any thriving transportation agency is the integration of maintenance management into their processes – through total asset management or some enterprise resource program. Highway agencies spend large sums of money maintaining their facilities and assuring product quality so that uniform levels of service are achieved. In recent years, transportation Maintenance Quality Assurance (MQA) programs have been developed to assure that maintenance quality is being achieved. MQA programs have to be capable of detecting insufficient maintenance efforts, poor material performance, and incorrect procedures when evaluating end-product performance. Data collected to assess whether these capabilities are achieved require a statistically-valid sampling and evaluation process.

At the October 2004 Maintenance Quality Assurance Peer Exchange held at Madison, Wisconsin, participants expressed interest in exploring how statistical tools might be more effectively applied in MQA programs (Adams 2004). Consequently, a research team was put together by the Midwest Regional University Transportation Center (MRUTC) in May 2005. Table 1.1 provides a summary of the statistically-related questions raised at the peer exchange, along with a clarification of these questions from the May 2005 meeting with the research team.

There are several types of data associated with MQA Programs, and specifically for this study, only questions pertaining to condition assessment or condition ratings were addressed. Range of categories investigated included roadway, roadside, traffic control, drainage, and rest areas. Lead states and Canadian provinces identified in this effort include California, Florida, Kansas, New York, North Carolina, Texas, Utah, Virginia, Washington, and Wisconsin.

Table 1.1 Statistical Questions Associated with MQA Programs

Index (1)	Specific Question (2)	Related Questions or Concerns (3)
1	*What decisions must this data should support?*	a) Consider the questions you want to answer before you collect the data. b) Are we going to use the data for funding level decisions? c) Are we oriented towards network-level policy decisions? d) Network-level data to legislature must be easily understood. e) What do you need to worry about before you get going and what to worry about after you have started – also relates to cost.
2	*What language accurately (and clearly) describes what I know and how confident I am?*	a) Have a standard to help lay people understand the statistics. b) There should be a glossary. c) Have an explanation of terms. Perhaps more than just a glossary.
3	*How many samples will I need to get valid information? What sampling techniques to use?*	a) How do different states achieve random sampling? b) Ways to explain to practitioners that random sampling is a prerequisite for valid results c) If I change sampling strategy, how do I compare results?
4	*How far "down" can I go, in terms of geographical subdivision, of highway features?*	a) Can the data go down to the county level? b) What kind of features? c) Horizontal/vertical segregation of data – statistical implications. d) Is there a problem with different data sets in different states? e) Not looking for characterizing problem areas, only trying to find a way to analyze the available data f) How large of a sample size do we need? g) Can data be stratified in a meaningful way? h) Sampling does not occur homogeneously – how do you deal with it?
5	*How confident am I?*	a) Confidence intervals (both managerial and statistical). b) If we compute an LOS of 83% and the threshold is 80%, do we have sufficient confidence (95%, say) that the threshold was successfully achieved?
6	*What kinds of trouble signs should I look for?*	a) Common problems within the data, particularly outliers. b) Feature weighting – combining multiple parameters in a single LOS value. c) Point features – assessing individual parameters. d) Data entry errors – incorrect values. e) Data cleansing. When to throw out data?
7	*How to QA the QA?*	a) How to check and verify that MQA data is being sampled correctly. b) How to compare effectiveness of different data collectors or crews.
8	*What do I do when my data isn't that powerful?*	a) Powerful – more than just a statistical reference. Also can mean reliable, valid and adequate. b) Small sample size – what strategies to use in order to overcome? c) What kinds of statements can you make given a small sample?
9	*What if I have inventory information? What if I don't?*	a) There are inventories of pavements, signs and bridges, however much of the data may not exist in such abundance. Need a methodology to determine how much data is enough; cost will be an issue.
10	*How have other states used and reported this?*	a) How to report out the statistical info? b) Hit target at some specified level of confidence? c) Is this data better than last year? How do you compare the data year to year? d) Summary statistics should make sense to agency, public, and legislature.

1.2 Project Objectives

The objectives of this project were to (1) develop a practical guide for the use of statistics in making sound maintenance management decisions in maintenance quality assurance (MQA) programs, and (2) synthesize case studies of agencies with effective MQA programs (as a possible severable deliverable). The guide has been produced in an electronic and printed format that answer specific questions from the October 2004 Maintenance Quality Assurance Peer Exchange, as well as information or guidelines for other statistical concerns.

The objectives were addressed by reviewing relevant literature and MQA program manuals of lead states, reviewing on-going research on statistical validity of MQA, collecting actual MQA data, performing a rigorous statistical analysis, investigating underlying assumptions of MQA data distributions, performing computer simulations and sensitivity analysis of key parameters, and creating both a hardcopy and Americans with Disability Act (ADA) compatible electronic guidebook that is easily understood by practitioners. In addition, answers to practical questions posed by the partners of this research, including agency representatives from Minnesota, New York, North Carolina, and Wisconsin, and both Federal Highway Administration (FHWA) and the MRUTC, were incorporated into the guide.

The guide has been written with the assumption that the reader should have the knowledge that normally would be gained by a basic statistics course. Additionally, training in statistical software is suggested to readily apply the concepts and procedures enumerated in the guidebook. Figure 1.1 places the guidebook in the context of having basic statistical knowledge before using the guidebook, and receiving formal training in statistical applications and software.

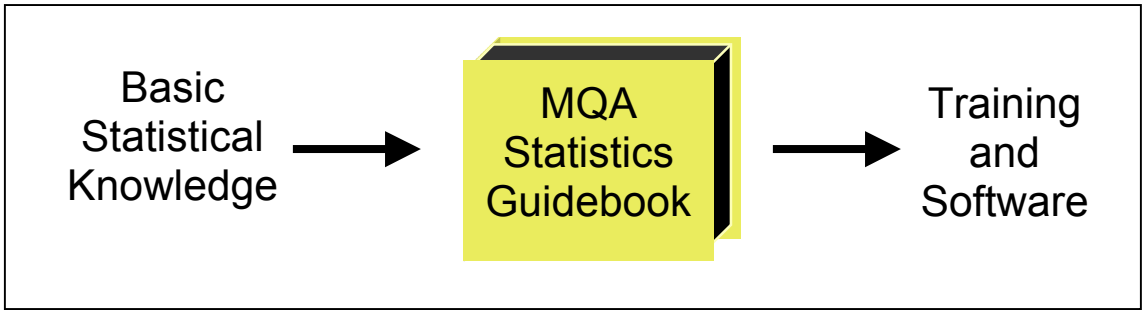


Figure 1.1 Context of Guidebook

1.3 Benefits

Research benefits include:

- A manual guide, both electronic and hardcopy, to serve as a vital resource to apply statistics in MQA programs.
- Improved decision-making and cost savings by providing maintenance practitioners with a deeper understanding of statistics generated during quality assurance efforts.
- Maintenance staff with an improved position to make data-based decisions that acknowledges the sampling approach, inherent risks, power of statistics, and measured quality.
- Fact-based decisions yielding greater accountability, proactive awareness of options or changes to maintenance activities, materials, and methods to improve quality, and overall enhanced knowledge of the process.
- An educational opportunity to University of Wisconsin-Platteville engineering students to translate their enhanced knowledge of maintenance programs to the betterment of transportation agencies in the Midwest.

CHAPTER 2 REVIEW OF CURRENT PRACTICES IN MQA

2.1 Introduction

A literature review was conducted to examine the statistics methods associated with MQA programs. Recent MQA program manuals for lead states were reviewed in detail for requirements when establishing MQA programs, indicators of maintenance quality, and statistical approaches. A literature review was conducted using the TRIS (Transportation Research Information System) database to understand statistical approaches and equations in MQA, and a detailed summary can be found in Appendix A.

Since the first Maintenance Management Workshop in 1968 at Ohio State University, highway maintenance quality continues to gain significant attention due to a wide range of factors. Significant among these factors is the increasing pressure for fiscal austerity on local resources coupled with the expanding awareness of future funding uncertainties. Higher labor, equipment, and material costs associated with continued inflation with no commensurate increase in tax revenues have caused transportation agencies to reduce personnel and cut back expenditures at all levels. The impact of these reductions is compounded by increasing maintenance and traffic management workloads created by an aging infrastructure, dramatic increases in traffic volume, and highway users' growing expectations for higher levels of service. Consequently, some transportation agencies have applied quality assurance as a tool for better maintenance management practice, as well as an effort to increase accountability to the legislature, and satisfy users' expectations for higher levels of service.

MQA applies to a broad scope of agency programs for managing and monitoring the effectiveness on maintenance operations. The scope includes elements such as maintenance condition assessment, level of service, maintenance performance measurements, maintenance accountability (TRB Committee AHD10 2005), and customer expectations and satisfaction of maintenance outcomes.

2.2 Establishing Quality in Maintenance

MQA programs, and related literature, were reviewed to document current practice. This review provided a basis for the statistical procedures and applications provided later in this guide.

According to Miller (1989) quality in maintenance is determined based on specified indicators. The quality indicator is the "evidence that an element of the highway (e.g. roadsides) is being maintained in accordance with the agency's standards". Quality indicators serve the purposes of providing direction to field personnel to ensure uniformity of maintenance effort throughout the agency, providing a tool for scheduling and budgeting, and defining a uniform level of service to which the highway user is entitled. Agencies develop indicators based on factors such as roadway functional classification, facility/road feature type, and customer expectations of quality.

To establish quality in maintenance, Miller (1989) outlined four key steps that an agency needs to undertake. These steps are as follows:

a) Determining the meaning of quality and defining it in terms of standards to be achieved;
b) Instructing field maintenance personnel in the intent, meaning, and use of the standards;
c) Developing and implementing procedures for evaluating the performance of the program to ensure compliance with agency intent; and
d) Having a consistent budgetary base to provide the resources to execute the program.

2.3 Evaluation Levels

The evaluation of quality in maintenance can occur at three levels including activity, project, and network levels. Independent trained quality assessment personnel typically perform the evaluation at all three levels. The *Activity level* evaluation involves an assessment of maintenance work (e.g. pothole patching) performed by field crews on an element of the highway (e.g. travel way) to ensure that field crews adhere to procedural requirements during the maintenance work and that the work product meets a predetermined quality level. At the *Project level*, each maintenance work activity required on a discrete section of the highway system (e.g. a bridge or relatively short section of highway) is evaluated and used to describe the existing condition or level of service (LOS) for the section. The *Network level* evaluation is an extension of the project level evaluation. Sample sections are determined randomly using statistical procedures and assessed to determine an overall level of service for a specific classification of the network or the entire network. The level of service determined is compared to that set by an agency for the network classification or the entire network.

2.4 MQA Program Manuals

Adams (2004) reported on a survey of U.S. states and ten Canadian provinces regarding maintenance quality assurance (MQA) programs. Of the 39 responding agencies (36 states and 3 provinces), 83% indicated having some form of an MQA program used primarily for purposes such as condition assessment, maintenance work planning, maintenance policy analysis, and allocation of maintenance funds. MQA programs have been developed for a wide range of features including culverts, guide rail, rest areas, traffic control devices, highways, bridges, and roadside.

2.4.1 Major Maintenance Categories

Ten states that were solicited and having provided MQA Program guides include: California, Florida, Indiana, New York, North Carolina, Texas, Utah, Virginia, Washington and Wisconsin. Major maintenance categories included roadway, unpaved shoulder, ditches, roadside features, traffic control devices, rest areas, and environmental concerns. Categories varied among states, however, most states had procedures for measuring and assessing roadway, drainage, and traffic control. Features from all ten states can be found in Appendix B.

2.4.2 Sampling Approach

All state MQA programs investigated use some form of simple random sampling. Simple random sampling characteristics described by MQA lead states are summarized in Table 2.1. The table indicates that the majority of states conduct inspections over randomly selected sample units of length equal to 0.1-mile. North Carolina uses a 0.2-mile sample unit, while New York and California use a 1-mile sample unit. Longer sample units may require more inspection crew time and budget compared to shorter sample units. In addition, for a longer sample unit, it may be necessary to subdivide it in the future if rehabilitation projects change the pavement structure; this may create anomalies in condition data analysis.

The length definition of the sample unit enables the total number of possible sample units to be determined for a given size of network. For example, if the total mileage for all interstate routes in a county is 1,800 miles, then for a sample unit of 0.1 mile, the total possible sample units will be equal to 1,800mile/0.1mile/sample unit = 1,800*10 =18,000 sample units from which a sample can be drawn based on the desired level of confidence and precision. Once the total number of samples is determined, random number generator programs such as Microsoft Excel™ can be used to select the candidate sample units for inspection.

Table 2.1 MQA Sampling Plans for States

State (1)	Sampling Unit Length (2)	Confidence Level or Precision (3)	Comments on Condition Assessment (4)
California	1 mile	95% confidence level	LOS2000 evaluation is based on visual inspection of randomly selected samples. An LOS2000 Coordinator determines the total random sample of centerline miles from each district for evaluation. *(Caltrans 1999).* According to McCullouch and Sinha (2003) 10% of the inventory is surveyed.
Florida	0.1 mile	95% confidence level and 3% precision	Sampling is based on sample units within a maintenance section. A random number generator program is used to select the sample units on each facility type contained within a maintenance unit *(Smith et al. 2003).*
Indiana	0.1 mile	90% confidence level	Sample size for inspection of each of three functional classes (interstate, state road, & U.S. Highway) per district is based on Equation A.3 at 90% confidence level. The random generator in Excel is used to select sample units to be inspected *(McCullouch and Sinha 2003).*
New York	1 mile	---	600 random sections are selected from 10 participating regions for review every year by field personnel *(NYSDOT 2004).*
North Carolina	0.2 mile	90-95% confidence level and 6% precision	For each major road system (primary, secondary, urban) approximately 250 (based on Equation A.5) random samples are drawn for each of 14 maintenance divisions. A statewide sample of approximately 250 samples are drawn for the interstate system. Random sample selection for 2006 is be based on variation of data from previous year surveys.
Texas	0.1 mile	---	---
Utah	0.1 mile	---	Inspection is based on 25 randomly-sampled units *((McCullouch and Sinha 2003)*
Virginia	0.1 mile	95% confidence level and 4% precision	Sample size determination is based on Equation A.6 *(Kardian and Woodward 1990)*
Washington	0.1 mile	95% confidence level	Statistical methods are used to identify approximately 2,200 randomly-selected data survey sites around the state *(Washington State DOT 2004, 2005).*
Wisconsin	0.1mile	---	---

2.4.3 Composite Equations for LOS Computing

Some states use a composite equation to compute an overall LOS for a given pavement segment. The calculation of LOS is typically based on observed characteristics relating to several main elements associated with a highway facility, such as the pavement, drainage ditches, and traffic control systems. In North Carolina, for example, six elements are used, as shown in Table 2.2.

Table 2.2 Elements and Characteristics Monitored for LOS (NC DOT 1998)

Element (1)	Monitored Characteristics (2)
Roadway	***Flexible pavement***: alligator cracking, block cracking, reflective cracking, rutting, raveling, bleeding, ride quality, and patching. ***Rigid (shoulder features)***: shoulder lane joint, shoulder drop-off, and shoulder condition. ***Rigid pavement features***: patching, surface wear, pumping, ride, longitudinal cracking, transverse cracking, corner break, spalling, joint seal, and faulting.
Unpaved Shoulders & Ditches	Shoulders, lateral ditches, and lateral ditch erosion.
Drainage	Crossline pipe, driveway pipe, curb & gutter, catch basins & drop inlets, and other drainage features
Roadside	Mowing, brush & tree control, litter and debris, slopes, and guardrail.
Traffic Control Devices	Traffic signs, pavement striping, words and symbols, and pavement markers.
Environmental	Turf condition and miscellaneous vegetation management.

Measurable definitions for each characteristic are provided in the Maintenance Condition Survey Manual (NCDOT 1998). For each characteristic within a 0.2-mile sample unit, a grading scale of A (best condition) through D, and F (worst possible condition) is assigned based on the percent of inventoried characteristic meeting specific defined condition criterion. The grading is translated into a rating scale of 5 (A-grade) to 1 (F-grade). The LOS for any element is calculated as shown in Equation 2.1.

$$LOS_j = 100 * \frac{\sum_{i=1}^{n} R_i W_i}{\sum_{i=1}^{n} W_i} \Big/ R_b \quad \text{.....(2.1)}$$

Where:

LOS_j = Level of service for element j (e.g. drainage);
R_i = Rating for observed characteristic *i* (e.g. catch basin);
W_i = NCDOT assigned weight for observed characteristic *i*;
R_b = Best possible rating = 5;
n = total number of observed characteristics; and
$R_i W_i = S_i$ = score for observed characteristic *i*.

NCDOT weight values assigned to drainage characteristics for Interstate highways, for example, are shown in Table 2.3. Equation 2.1 and Table 2.3 were derived from a database file obtained from the NCDOT Maintenance Unit of the Division of Highways.

Table 2.3 Typical Weights for Drainage Characteristics for Interstates (NCDOT 1998)

Drainage Characteristic (1)	NCDOT Assigned Weight (2)
Crossline pipe	8
Driveway Pipe	0
Curb & Gutter	6
Catch basin	6
Other Drainage features	6

The overall LOS for a sample unit or segment is calculated as shown in Equation 2.2.

$$LOS_s = \frac{\sum_{j=1}^{N} LOS_j W_j}{\sum_{j=1}^{N} W_j} \quad \text{.....(2.2)}$$

Where:

LOS_S = Overall level of service for sample unit or segment;
LOS_j = LOS for element j with some observed characteristics;
W_j = NCDOT assigned weight for element j with observed characteristics for a particular highway functional classification (30% pavement, 14% drainage, etc.); and
N = total number of elements with observed characteristics.

Table 2.4 NCDOT Assigned Weights for Monitored Elements in an Interstate Segment

Element (2)	Assigned Weight for LOS Computation (2)
Roadway/pavement	30
Unpaved Shoulders & Ditches	16
Drainage	14
Roadside	15
Traffic Control Devices	20
Environmental	5

2.5 Summary and Conclusions

A literature review was conducted to examine the application of statistics in maintenance quality assurance (MQA). In addition, elements of MQA programs for lead states were reviewed including purposes, monitored features, and challenges associated with the implementation of MQA programs. Based on the review, the following conclusions are reached:

- Agency MQA programs are focused on asset condition assessment, needs, and customer satisfaction and expectations of maintenance outcomes.
- The major maintenance categories with quality assurance programs in-place include roadway, roadside and vegetation, drainage, traffic control, and rest areas. It was observed that among lead states, a wide range of threshold definitions exist for features or characteristics monitored for each maintenance category. In addition, for the same maintenance categories, considerable variation exists among the states in the types of features monitored.
- Statistics have been applied to MQA programs in four main areas including, sampling for customer surveys, sampling facilities for condition assessment, level of service ratings and analyses, and quality assurance testing for automated data collection and processing.
- MQA programs in lead states use a statistical random sampling approach to determine the number of sample units or sample size to inspect in asset condition assessment. The majority of lead states conduct asset condition assessment over 0.1-mile sample units, with the exception of programs in New York and California, which use 1-mile segments. In addition, lead MQA states indicated using confidence levels of 90-95% with precision in the range of 3-6% in the statistical sampling approach.
- Results from MQA program analyses are used for a variety of purposes including, maintenance work planning and management, maintenance policy decisions, allocation of maintenance funds, description of existing conditions and trends, and demonstration of accountability to all key customers such as the legislature or top decision-makers, and the public at-large.

CHAPTER 3 ROLE OF STATISTICS IN MQA MANAGEMENT

3.1 Introduction

This chapter presents key issues that confront MQA decision-makers at various organizational levels, and how statistics can play an important role across those levels. Specific issues that can use statistics as a tool in the management and decision-making process are identified. Finally, statistical terms and definitions are provided to help communicate features of the data.

3.2 MQA Decision Issues

MQA programs can operate at three levels including the activity, project, and network levels. The basic operating levels of MQA programs and sample activities that occur at each level are listed in Figure 3.1. This figure suggests that the foundation for MQA facility management at any level is data acquisition and processing (database). In addition, criteria to determine acceptable work are essential, as well as analysis of alternatives to determine cost-effective solutions.

The literature, however, indicates that current MQA programs are focused at the network level. Hence, most decisions are made at this level and may involve a network of facilities for a given functional classification system or network within a jurisdiction. Decisions at the network level can result from a series of questions and issues that can be viewed from three basic organizational levels, including technical, administrative, and legislative levels. The questions and issues may differ in focus and scope depending on the agency (i.e. state, county, district, and city) and the management level involved. Knowledge about the pertinent issues provides a basis for identifying specific areas where statistics can be used as a tool in the decision-making process. In addition, such knowledge allows a target audience for MQA reports to be identified, and consequently, the manner in which the reports or information can be communicated to facilitate the understanding of the target audience.

Figure 3.2 lists some of the common decision-making issues associated with the various organizational levels. The administrative level personnel commonly develop maintenance capital spending and programs. They therefore, need to explicitly recognize and respond to legislative level issues based on specific required inputs from the technical level personnel. The administrative and legislative organizational levels tend to emphasize justification for budget requests, while the technical level focuses on the data requirements for decision-making at the various levels.

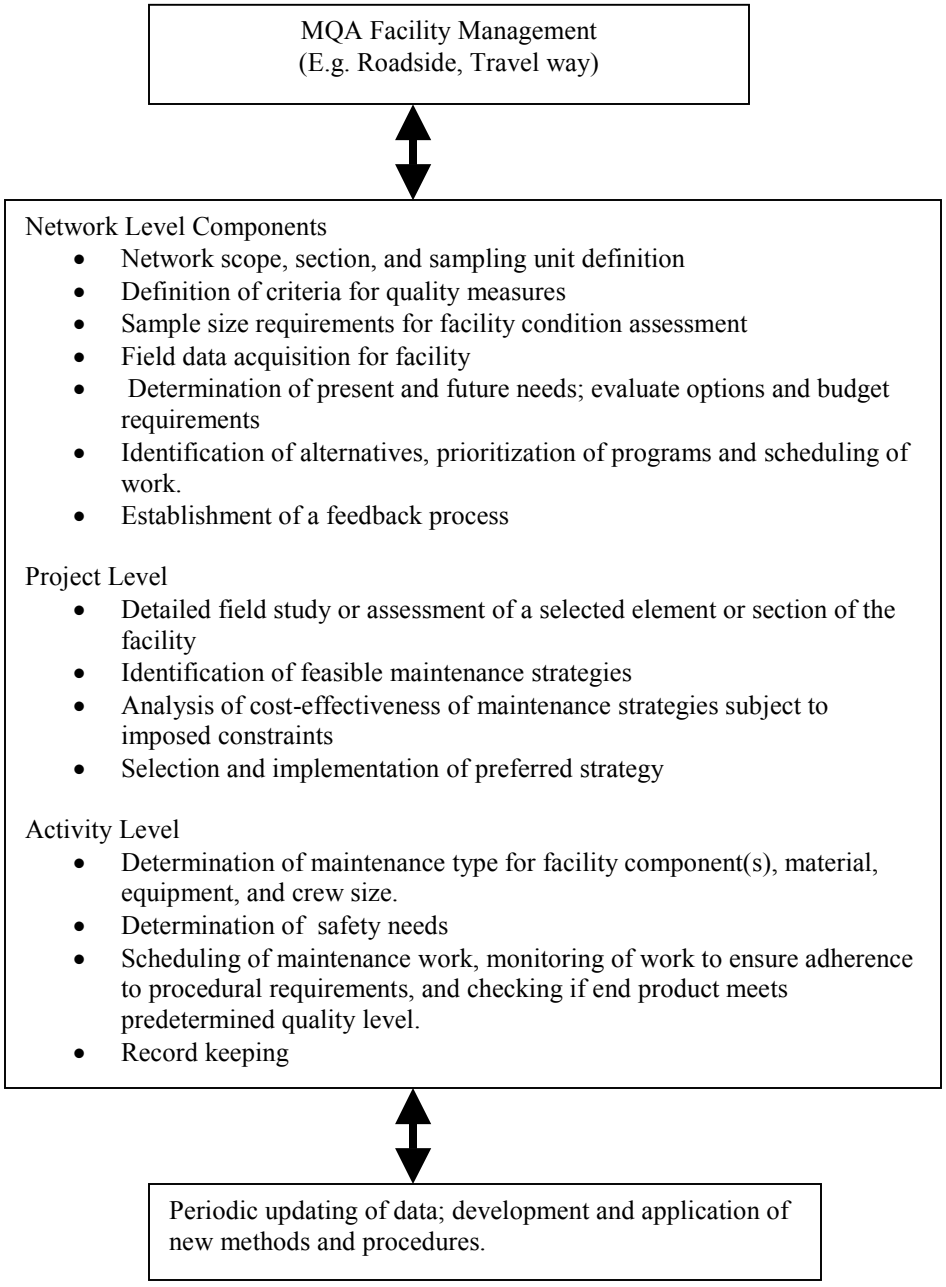


Figure 3.1 Basic Operating Levels of MQA Programs

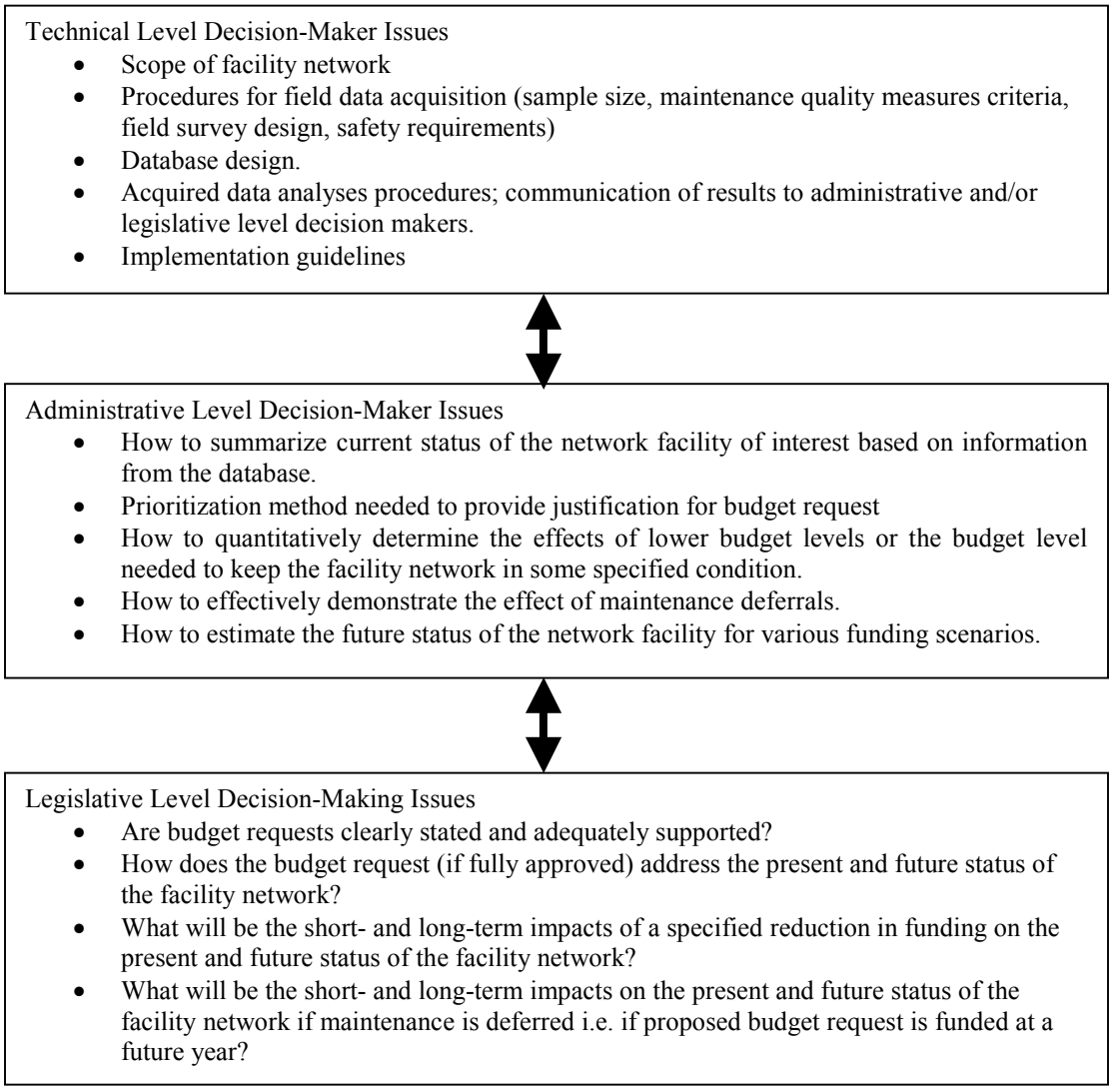


Figure 3.2 Network level Organizational Decision-Making Issues for MQA Programs

3.3 Adequacy of Statistics in MQA Decision-Making

As described earlier, the common decisions made at the network level of a MQA include determining the extent of the facilities to be inventoried, maintenance needs from facility condition evaluation surveys, and funds needed to address the identified needs. In addition, feasible funding options and strategies, as well as their impacts on the conditions of the facilities under consideration and on the using public are determined. Decisions that require the use of statistics at the network level may include but are not limited to the following:

a) Determining the sample size for condition evaluation survey. It is economically not practical to conduct an inventory of the network facility as a whole. Hence, some statistical sampling is required. Chapter 5 discusses issues involving sampling design and sample size determination.
b) Determining the proportions of facility characteristics (whether statewide, countywide, or district-wide) that meet agency target value for a specific performance measure at a given confidence level. It is common practice for agencies to set target maintenance performance values for their assets, and periodically, determine the proportions that meet, exceed or fall below the set target in order to plan and program work load. This may require constructing a confidence interval relative to the target.
c) Determining whether significant differences exist in facility maintenance performance for the different functional classification systems, geographical regions or jurisdictions. This may require the use of analysis of variance tests.
d) Determining whether significant changes exist in facility maintenance performance from one year to the next. This may require the use of comparison of means tests. If significant differences exist, agency intent will be to determine the role that funding level played in the variation and use the information for future funding requests.
e) Determining how to create some awareness among the public and policy makers regarding facility condition over time. This may require the use of simple plots/trend analysis of condition. This information can be related to consequences of delayed or inadequate funding on facility.

A review of the MQA decision issues, as related to the three organizational levels outlined in the previous subsections suggests that the use of statistics is more appropriate to addressing issues at the technical level. The other two levels rely heavily on the technical level data for their decision-making. In the following sections, specific questions or issues that apply statistics as a tool to facilitate the decision-making process are presented.

3.4 Characterizing Data

Statistics play a fundamental role in characterizing data, and in terms of this guide, characterizing maintenance condition. Maintenance characteristics can be described by a variety of elementary statistics, and several well-established procedures are available by which these elementary statistics can be used to make rational and objective decisions concerning the effectiveness of different maintenance materials or procedures, or to make other similar comparisons.

After deciding the objective of an analysis, the next step is to determine how much data, and of what type, will be necessary to compute the elementary statistics and perform the appropriate analyses. It is usually impractical to sample all, or even a large percentage, of a particular stretch of highway. Therefore, a key element of statistical analysis is the use of a limited number of random samples to represent the entire area of interest. The

validity of the information gained in this manner is strongly linked to the size of the samples and the manner in which they are obtained.

Since the outcomes of the analyses will often result in the decision to spend substantial sums of money for reconstruction or repair, or to adopt some method or material as standard practice, it is important to know and control the level of confidence with which these decisions are made.

3.5 Elementary Statistical Definitions

In an effort to communicate statistics in various MQA decisions, key terms and definitions are available. The following is a list of common terms and definitions, in order of detail, that pertain to various decision levels and organizational levels. Most of these terms are in such common use that no references are cited with the definitions, and definitions of some of the most common terms have intentionally been omitted. For further clarification, refer to almost any standard statistical text. Another particularly useful reference for terms related to quality assurance is the *Glossary of Highway Quality Assurance Terms* published by the Transportation Research Board (TRB 2005).

Population – All items or portions of a quantity of similar material, construction, or product – such as all paving done on a specific highway on a given day, etc. In quality-assurance parlance, sometimes used interchangeably with *lot*.

Types of Populations – A population may be either *continuous* (measured, such as pavement thickness) or *discrete* (counted, i.e., made up of elements that either have or do not have some characteristic). Continuous populations may be further subdivided into two categories, depending on their frequency *histograms. Normal* populations are approximately bell shaped, whereas *non-normal* populations depart significantly from the bell shape in any of several possible ways. The type of population often determines which statistical procedures are appropriate.

Sample – A portion or subset of the population.

Random Sample – A sample selected in such a manner as to allow all members of the population an equal likelihood of appearing in the sample.

Stratified Random Sample – A sample selected by first subdividing the population into a specified number of equal subsections, and then selecting a single random sample from each. The purpose of stratified random sampling is to spread the sample throughout the population and to prevent individual samples from being clustered too closely together.

Attributes Sampling – Sampling from a population of discrete values. A distinct advantage of this approach is that no distributional assumptions are required. A disadvantage is that it has less discriminating power than variables procedures.

Variables Sampling – Sampling from a continuous (or essentially continuous) normal population followed by the calculation of various statistical parameters such as the mean, standard deviation, etc.

Quality Index (Q Statistic) – A statistic used with variables sampling and appropriate tables to estimate the *percent defective* or the *percent within limits* of a population.

Parameter – A measure of a population that takes into account every member of the population. The mean, when applied to a population, is an example of a parameter.

Statistic – A measure computed from a sample. Sometimes called *sample statistic*.

Estimate – A statistic computed from a sample in order to determine the population parameter. An estimate may be a *point estimate* (single value assumed to represent the population parameter), or an *interval estimate* (two values between which the population parameter is believed to lie).

Confidence Interval Estimate – An interval estimate within which the true population parameter is expected to lie with a specified level of probability.

Confidence Limit(s) – Lower and/or upper bounds of the confidence interval estimate. Confidence intervals may be *two-sided*, in which case there are two confidence limits, or *one-sided*, in which case one of the limits is plus or minus infinity.

Mean – A measure of *central tendency*. For most applications, the mean is taken to be the arithmetic average of a group of numbers. The mean may be calculated from either a sample or a population.

Standard Deviation – A measure of *dispersion* of a set of data. It may be calculated from either a sample or a population, although the computational formulas are slightly different.

Variance - A measure of the average squared distance between the mean and each item in the population. The square root of the variance is the *standard deviation*.

Hypothesis – In statistics, a claim or conjecture about a population.

Hypothesis Test – A statistical procedure by which a hypothesis is tested.

Null Hypothesis – The hypothesis that is tested, often the opposite of the conjecture that motivated the hypothesis test.

Alternative Hypothesis – The hypothesis to be accepted if the null hypothesis is rejected. Technically, neither the null hypothesis nor the alternative hypothesis can be proven true in this manner, but if the null hypothesis is rejected at a specified low risk of error, it is conventional to accept the alternative hypothesis as true. Conversely, if the null hypothesis is not rejected by the statistical test, it is conventional to regard it as true.

Type I Error - Rejecting a null hypothesis when it is true. The Greek letter alpha (α) is the probability of Type I error

Type II Error - Accepting a null hypothesis when it is false. The Greek letter beta (β) is the probability of a Type II error

Power of the Hypothesis Test - It is defined as the probability of correctly rejecting the null hypothesis when it is false. The power of a test is computed as 1-Beta (β). The test is more powerful if beta is kept reasonable small. A powerful test can have values of 0.80, 0.90, or 0.95.

Significance Level – The probability of falsely rejecting the null hypothesis when it is true. Usually represented by the Greek letter alpha (α). Commonly used values are 0.01, 0.05, and 0.10.

Confidence Level – The probability that statisticians associate with an interval estimate of a population parameter indicating how confident they are that the interval estimate will include the population parameter. The quantity 1.0 minus the significance level. Commonly used values are 0.90, 0.95, and 0.99.

Normal Distribution – A data distribution commonly occurring in nature that is approximately bell shaped. This distribution is a reasonably accurate representation of many construction and maintenance characteristics, and often occurs when there is a process with a consistent central tendency that has no nearby lower or upper physical constraints.

Student-t Distribution – A statistical distribution used for a variety of comparisons and *hypothesis tests*. It is bell shaped similar to the normal distribution, but generally broader than the normal distribution, and typically applies to small samples of $n < 30$.

F Distribution – Another statistical distribution used for a variety of comparisons and *hypothesis tests*. The F statistic is the ratio of two variances and, unlike the normal and Student t-distributions, the F distribution is skewed.

Analysis of Variance (ANOVA) – A procedure to simultaneously test whether the sample means of several populations are equal or not. Variation is partitioned both within and between the populations, then ratio of between variance and within variance are considered in the determination.

F Ratio - It is the ratio used in to determine the area, or probability, under the skewed F distribution. It is usually used with the ANOVA, among other tests, to compare the magnitude of two or more estimates of the population variance to determine it the two or more estimates are approximately equal.

Non-Parametric - It is when the parameters of the variable of interest in the population are not known. Non-parametric methods do not rely on the estimation of parameters, such as the mean or standard deviation, to describe the variable of interest. These methods are sometimes called *parameter-free* method or *distribution-free* methods.

CHAPTER 4 STATISTICAL PROCEDURES

4.1 Introduction

This chapter presents fundamental statistical procedures that help describe features of MQA data, create confidence intervals, and conduct formal statistical tests. These fundamental procedures provide the foundation for the applications in the following chapter where specific questions facing MQA programs are addressed. Examples are provided for these statistical procedures using sample data.

- Normality Tests
- Confidence Interval for a Proportion (Discrete Population)
- Confidence Interval for the Mean of a Normal Population
- Confidence Interval for Percent Defective (Normal Population, Single Limit)
- Confidence Interval for Percent Defective (Normal Population, Double Limits)
- Hypothesis Test for Mean of a Normal Population
- Hypothesis Test for Difference between Means of Two Normal Populations
- Hypothesis Test for Dependent Normal Means (Paired-t Test)
- Hypothesis Test for Proportion of a Discrete Population
- Hypothesis Test for Difference between Proportions of Discrete Populations
- Hypothesis Test for Difference between Means of Several Normal Populations
- Sample Size Determination
- Random Sampling

4.2 Normality Tests

The validity of several statistical procedures depends upon the assumption of sampling from a population that is normally distributed, or at least nearly so. Figure 4.1 provides a hypothetical case showing an approximately normal histogram and three possible examples of distinctly non-normal histograms. Provided a reasonable amount of data values are available, visual inspection of the histogram may be sufficient to determine whether or not the distribution is sufficiently bell shaped to safely assume it is normal. Based on past experience of the authors, $N > 50$ is desirable when assessing normality with histograms since it is robust to the number of histogram bars, and the individual bar width. Traditional normality tests traditionally require a minimum sample size of $N > 30$ (Thompson 1992).

There are a variety of statistics available for testing normality, including the Anderson-Darling test, Shapiro-Wilks W test, and Kolmogorov-Smirnov test. A simplified, and more visually appealing test, is the Normal Probability Plot. Refer to a range of statistical textbooks for detailed descriptions of each test procedure.

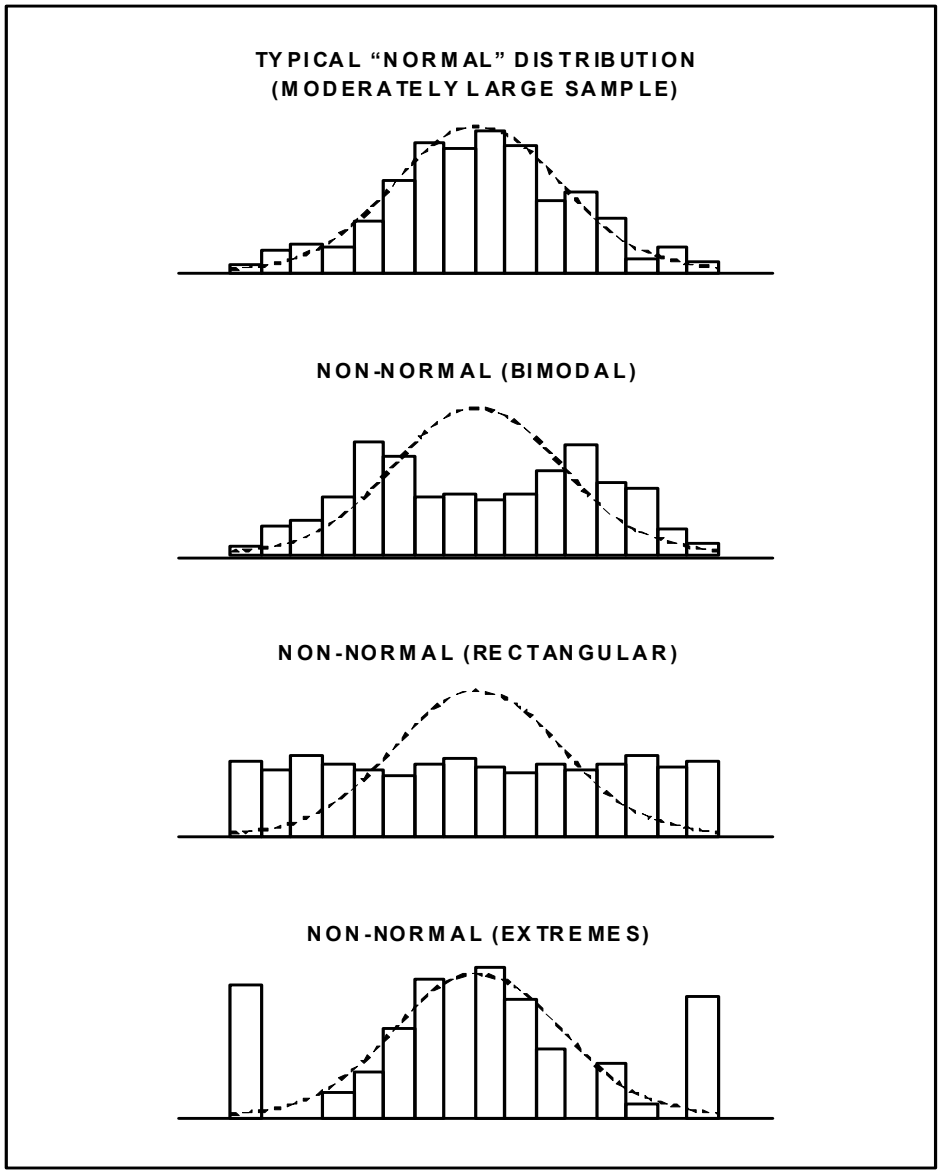


Figure 4.1 Typical Data Distributions

If there is doubt about normality, and a more formal decision process is desired, a variety of tests are readily available in standard statistical texts, such as the Anderson-Darling Test. These tests use significance testing to determine whether the sample is normal or not. Typically, a 0.05 or 0.10 probability value (p-value) is used to determine significance.

Several statistical computer packages can readily determine normality, such as Minitab™ or StatGraph™. An example of a normal distribution test using the Minitab™ software and North Carolina 2000 level of service (LOS) data is shown in Figure 4.2. In Figure 4.2, the weighted level of service (LOS) ratings for n = 151 segments in Division 1 are provided. The histogram shows strong central tendency and approximation of a normal "bell-shaped" distribution, with $\bar{x}$ = 71.6 and s = 11.2. With a stated *null hypothesis* that the shape is normal, with 5% probability of rejection, the Anderson-Darling Test concludes a normal distribution since the p-value > 0.05. From this distribution, 95%confidence intervals are constructed for the mean; an interval where the true mean can be found with 95% probability.

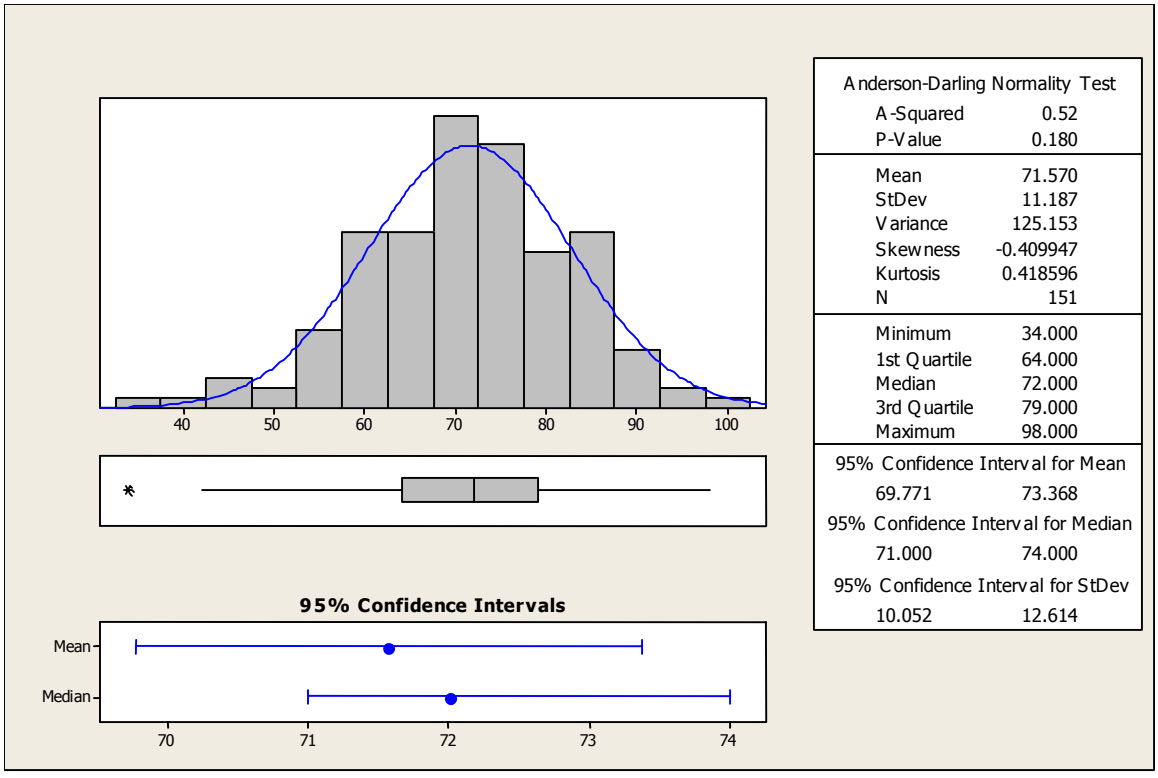


Figure 4.2 North Carolina 2000 Weighted LOS for Division 1

Quartiles are also provided in Figure 4.2, where the data are grouped into quarters. The 1st Quartile defines data points in the lower quarter, while the 3rd Quartile defines data points in the upper quarter. Additional statistics are listed, such as the skewness and kurtosis, that describe the shape of the distribution. (Refer to any statistical textbook for

definitions and examples for skewness and kurtosis. These are advanced statistical parameters that can be investigated and help better characterize a distribution. Based on a review of current MQA practice, they are not recommended without advanced training.)

A fundamental question that may arise is "what happens if the data are not normal"? In Figure 4.3, the weighted LOS ratings for n = 175 segments in Division 3 are provided. The histogram shows strong central tendency and approximation of a normal "bell-shaped" distribution, with $\bar{x}$ = 77.9 and s = 9.3. However, at a set 5% value for the null hypothesis, the Anderson-Darling Test concludes that the distribution is not normal, since the p-value = 0.038, which is lower than the 0.05 threshold. In this case, the agency may assume that a normal distribution would be met if additional segments are sampled, or a *non-parametric* approach can be pursued.

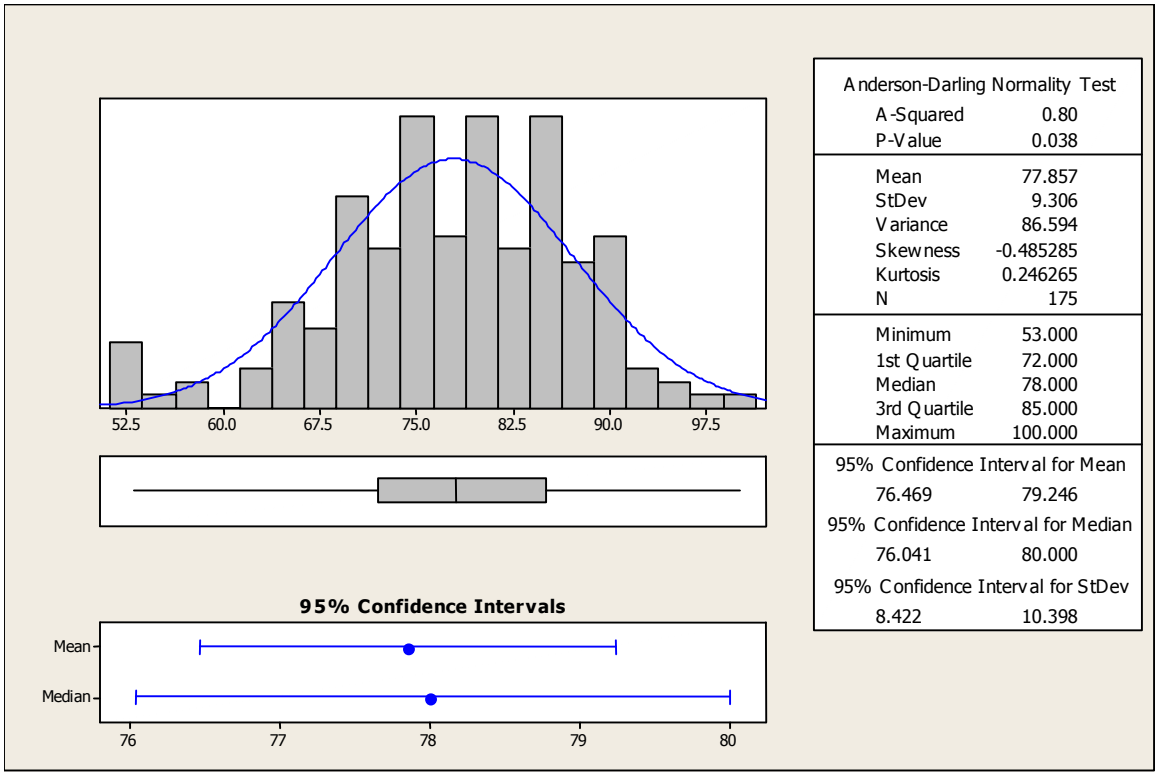


Figure 4.3 North Carolina 2000 Weighted LOS for Division 3

Non-parametric methods are a simplified approach to characterize the data, such as the median, mean, quartile(s), and range. These statistics do not require that a specific distributional assumption be met, such as the normal distribution. However, non-parametric methods cannot construct confidence intervals or estimate precision for a stated probability level, since the data lack the underlying distributional parameters. Thus, in the previous example, an agency can compute the mean, standard deviation, or mean; however, a 95 percent confidence interval of the mean would not be possible since the data were determined not normally distributed.

4.3 Confidence Interval for the Mean of a Normal Population

Example: Confidence Interval for Number of Obstructed Drains

This procedure is often used in lieu of a point estimate, or with a point estimate to provide an indication of the precision of the estimate, stated at some specified confidence level. For example, suppose that the number of drains in County #1 are counted for having either (1) the outlet, endwalls, or end protection closed or crushed, or (2) the water flow or end protection is obstructed. Data determine that a point estimate of 3.0 drains are obstructed, or stated as an interval estimate of 2.5 - 3.5 drains with a confidence level of 95%. The correct interpretation is that, if the complete estimation procedure were repeated many times using a confidence level of 95% each time, then approximately 95 percent of the intervals estimated in this manner would contain the true population mean.

The top diagram in Figure 4.4 illustrates the conventional manner in which the confidence interval for a population mean would be computed from the sample mean. For example, suppose a sample of size n = 10 in County #1 had produced a mean value, ($\bar{x}$) for a count of 3.5 drains and a standard deviation (s) of 0.29 drains. Since the sample size is less than 30, the Student-t distribution is used. Sample sizes with a distinct normal distribution (n > 30) could use the standard normal tables with Z statistic), if desired. In this example, degrees of freedom, df = n – 1 = 10 – 1 = 9, and a significance level of $\alpha/2$ = 0.025 (giving a total tail area of 0.05 for a two-tailed interval at a confidence level of 0.95), the appropriate t-statistic is $t_{df=9,\ \alpha/2=0.025} = 2.262$. The lower and upper confidence limits are then computed as follows:

LOWER: $$L = \bar{x} - \frac{t(s)}{\sqrt{n}} \qquad (4.1)$$

$$L = 3.5 - \frac{2.262(0.29)}{\sqrt{10}}$$

$$L = 3.29$$

UPPER: $$U = \bar{x} + \frac{t(s)}{\sqrt{n}} \qquad (4.2)$$

$$U = 3.5 + \frac{2.262(0.29)}{\sqrt{10}}$$

$$U = 3.71$$

Therefore, for this example it is conventional to state that, at a confidence level of 95%, the true population mean might be as small as 3.29 drains or as large as 3.71 drains. If a

higher level of confidence had been chosen (such as 99%), the price for this greater degree of assurance would be a still wider interval. The t-value in this case would have been 3.250, and the corresponding intervals would have been L = 3.20 and U = 3.80 drains.

The middle diagram in Figure 4.4 illustrates a slightly different way this procedure could be formulated. As depicted in this diagram, it is desired to find how low the true population mean (L in this diagram) could be such that there would be a probability of exactly 0.025 of obtaining the observed sample mean of $\bar{x}$ = 3.5 drains. Similarly, it is desired to find how high the true population mean could be such that there would be a probability of exactly 0.025 of obtaining that same observed value of the sample mean. It can be demonstrated that, for this example, precisely the same calculations would be made as shown in Equations 4.1 and 4.2. In this case the procedures are exactly equivalent because the "L" and "U" sampling distributions shown in the middle diagram are identical to the distribution shown in the top diagram.

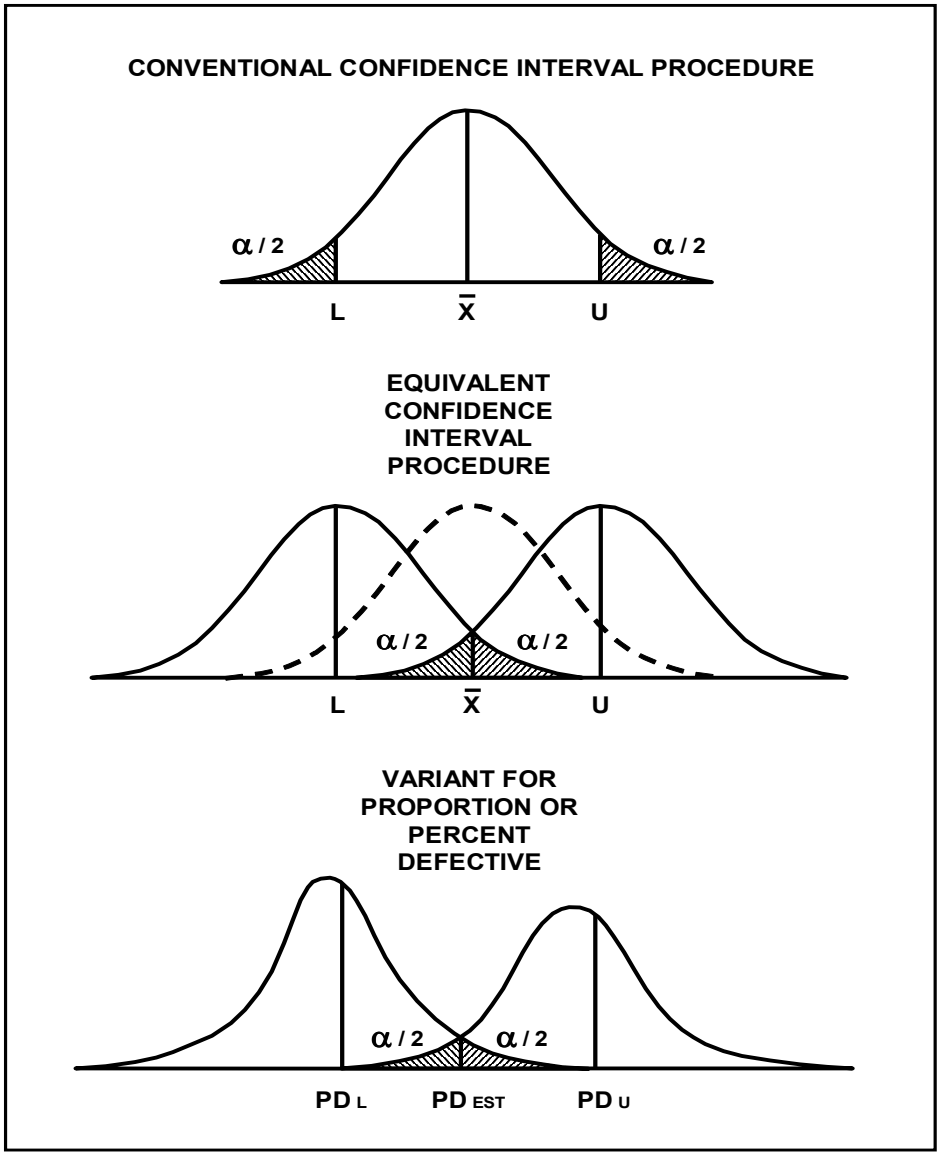


Figure 4.4 Confidence Interval Concepts

4.4 Confidence Interval for a Proportion (Discrete Population)

Example: Proportion of Vegetation Obstruction

When the statistic of interest is a proportion, it might be stated as a decimal or, alternatively, as a percent defective (PD). In both cases, the conceptual approach shown in the lower diagram in Figure 4.4 applies, which is simply a variant of the middle diagram.

If the proportion has been obtained by sampling from a discrete population, and counting the number of sample items that have some particular characteristic, a procedure based on the binomial distribution will be used as described below. For example, in a given state in County #1, n = 20 randomly selected equal-length sections of roadway were sampled for vegetation obstructions, and k = 3 segments were found to be defective by having sufficient obstruction to the driver, thus, the point estimate of the proportion defective would be p = 3 / 20 = 0.15 (which could also be represented as PD = 15). This value is certainly of interest, but it would be still better to know if that value of 0.15 is precise to plus or minus 0.05, for example, or some much larger tolerance.

Although there is no simple equation by which the confidence limits can be computed in this case, they are nonetheless easy to obtain. Convenient tables exist for a wide range of possible sample values, such as those published by the Chemical Rubber Company (CRC 1968). Available graphs cover an even wider range of data but are not quite as accurate or easy to use, such as those published by Dixon and Massey (1969), and several on-line calculation tools also exist. (It is recommended to cross-check any on-line tools with well established tables before adopting them as standard practice because several minor discrepancies have been found). Still another method, used by the authors to double-check the aforementioned sources, is to create a computer algorithm that performs binomial calculations using an efficient search routine to obtain the desired results.

For the example just presented, in which p = 3 / 20 = 0.15, and a two-tailed 95% confidence interval is desired, the CRC table (CRC 1968) gives the lower and upper confidence limits as $p_L = 0.032$ and $p_U = 0.379$. As a check, the graph for a 95% confidence level was consulted in Dixon and Massey (1969) where it is seen that the curves for n = 20 intersect the p = 0.15 grid line at about $p_L = 0.03$ and $p_U = 0.38$. Similar tables in other texts should produce essentially the same results. As a further test to demonstrate the accuracy of these results, a direct computation was made using the cumulative binomial expression for the lower confidence limit given by Equation 4.3.

$$\sum_{k}^{n}\left[\frac{n!}{x!(n-x)!}\right]p^{x}(1-p)^{(n-x)} = \alpha/2 \qquad (4.3)$$

where

k	=	number of failed sections;
n	=	sample size;
x	=	summation variable, ranges from k to n;
p	=	population proportion defective (p = PD / 100); and
α	=	statistical significance level (1 – confidence level).

For a two-sided confidence interval of 0.95, the significance level is $\alpha = 1 - 0.95 = 0.05$ and the right hand side of Equation 4.3 is then $\alpha/2 = 0.025$. For this particular case, instead of performing 18 computations to sum from $k = 3$ to $k = n = 20$ to confirm that a probability value of $\alpha/2 = 0.025$ is obtained, it will be far simpler to sum from $k = 0$ to $k = 2$ to check for a probability value of $1 - 0.025 = 0.975$. No trial-and-error process is needed for this check because the table and the graph have already indicated that $p_L = 0.032$. Therefore, this value is used in Equation 4.3 to obtain the following results:

$k = 0$: Prob[$x = k$] = 0.52180
$k = 1$: Prob[$x = k$] = 0.34499
$k = 2$: Prob[$x = k$] = 0.10835
$\sum$ = 0.97514 (EXPECTED 0.975)

These calculations confirm that the values obtained from both the table and graph are accurate. This particular case can also be used as a quick check of any on-line calculation tool under consideration.

As a final note, it can be seen that for a sample size of n = 20 segments for vegetation obstruction, the estimate of p (or PD) is not obtained with as high a degree of precision ($0.032 \leq p \leq 0.379$) as might be desired for a decision of any consequence to be made. This emphasizes that, for some applications, larger samples may be desirable. However, it will be demonstrated in the next section that when conditions permit a variables procedure to be used, significantly narrower confidence intervals can be obtained.

The procedure for a discrete population is summarized as follows:

Sample n = 20;
Sample k = 3;
Compute p = 3 / 20 = 0.15 (PD = 15.0);
Select confidence level (0.95 in this example);
Compute $\alpha/2 = (1 - 0.95) / 2 = 0.025$;
Obtain $p_L = 0.032$ from table or graph;

Obtain $p_U = 0.379$ from table or graph;
$PD_L = 100\ (0.032) = 3.2$; and
$PD_U = 100\ (0.379) = 37.9$.

POINT ESTIMATE: PD = 15.0

INTERVAL ESTIMATE: $3.2 \leq PD \leq 37.9$ (0.95 Confidence)

4.5 Confidence Interval for Percent Defective (Normal Population, Single Limit)

Example: Percentage of Adequately Mowed Area

CAUTION: The term "single limit" in the heading of this section refers not to a single-sided confidence interval, but to those applications for which percent defective in one tail or the other (but not both) of the population is of concern.

When it is known that the sampled population is at least approximately normally distributed, a variables sampling procedure can be used to estimate the percent defective (PD) or percent within limits (PWL) of a population. (Note that these two measures are interchangeable since PD + PWL = 100.) This method requires that a sample of size $n \geq 3$ be taken, and the sample mean and standard deviation be computed. The next step is to compute the quality index for lower or upper limit, as follows:

$$\text{LOWER:} \qquad Q_L = \frac{\bar{x} - L_L}{s} \tag{4.4}$$

$$\text{UPPER:} \qquad Q_U = \frac{L_U - \bar{x}}{s} \tag{4.5}$$

where

Q_L = quality index computed for lower limit;
Q_U = quality index computed for upper limit;
L_L = lower limit;
L_U = upper limit;
$\bar{x}$ = sample mean; and
s = sample standard deviation.

Quality indexes computed in this manner are then referred to appropriate tables, specific for the sample size, to obtain the PD or PWL estimates. Alternatively, special software can be used. Sets of tables can usually be found with literature on statistical quality assurance, and one such set of tables for PD and PWL is provided in Appendix C and D, respectively.

In order to make a direct comparison with the previous example, suppose that the agency desires to measure the percentage of area that has been adequately mowed to alleviate vegetation obstructions in randomly-chosen 0.1-mile roadway segments. A sample of size n = 20 segments has been obtained, and the mean and standard deviation are are $\bar{x}$ = 86.0% and *s* = 5.79% mowed adequately. For this example, assume the lower limit is L_L = 80.0%, so that the quality index is calculated using Equation 4.4 as follows:

$$Q_L = \frac{(86.0 - 80.0)}{5.79} = 1.036$$

Interpolating for Q_L = 1.036 in a table for n = 20, the point estimate for percent defective is found to be PD = 15.01, essentially the same as obtained with the same sample size in the previous example. For the variables case, however, the procedure for determining the confidence limits is entirely different, and Equations 4.6 and 4.7 are used to obtain lower and/or upper confidence limits on the PD estimate (Weingarten 1982).

$$Z_L = -Q + Z_{\alpha/2}\sqrt{\frac{1}{n} + \frac{Q^2}{2n}} \tag{4.6}$$

$$Z_U = -Q - Z_{\alpha/2}\sqrt{\frac{1}{n} + \frac{Q^2}{2n}} \tag{4.7}$$

where

Z_L	=	standard normal variate associated with the lower proportion defective ($p_L = PD_L / 100$);
Z_U	=	standard normal variate associated with the upper proportion defective ($p_U = PD_U / 100$);
$Z_{\alpha/2}$	=	standard normal variate associated with the confidence level, i.e., for 0.95 confidence, $\alpha/2 = 0.025$ and $Z_{\alpha/2} = -1.96$;
n	=	sample size; and
Q	=	Q statistic computed by Equation 4.4 or 4.5.

Since a lower confidence limit is sought in this case, Equation 4.6 is used, with a value of $\alpha/2 = 0.025$ and $Z_{\alpha/2} = -1.96$ to be consistent with the previous example, as follows:

$$Z_L = -1.036 + (-1.96)\sqrt{\frac{1}{20} + \frac{1.036^2}{2(20)}} = -1.579$$

Then, using a table of areas under the standard normal distribution, Z_L = - 1.579 is found to correspond to an area of 0.058, which is equivalent to PD = 5.8. Therefore, with a sample size of n = 20 and a single-sided confidence level of 0.975 ($\alpha/2 = 0.025$), the

variables procedure has produced a point estimate of PD = 15 with a lower confidence limit of $PD_L = 5.8$. Under comparable conditions, the attributes procedure described in the previous section produced a lower confidence limit of $PD_L = 3.2$, a one-sided interval that is almost 30 percent less precise (broader).

In a similar manner, Equation 4.7 can be used to compute an upper confidence limit for this example of $PD_U = 31.1$. The comparable upper confidence limit in the example in the section, *Confidence Interval for a Proportion (Discrete Population)*, is $PD_U = 37.9$ and is less precise by about 32 percent. This example clearly demonstrates that the variables procedure is capable of producing more precise interval estimates when the data are sufficiently normally distributed to allow its use. Thus, a more accurate depiction of vegetation obstruction is possible.

The procedure for a normal population with a single limit is summarized as follows:

Sample n = 20;
Lower limit $L_L = 80.0$;
Calculate $\bar{x} = 86.0$;
Calculate $s = 0.579$;
Calculate $Q_L = 1.036$ using Equation 4.4;
Obtain PD = 15.01 from table;
Select confidence level (0.95 in this example);
Compute $\alpha/2 = (1 - 0.95) / 2 = 0.025$;
Compute $Z_L = -1.579$ using Equation 4.6;
Compute $Z_U = -0.493$ using Equation 4.7;
Obtain AREA(Z_L) = 0.058 from standard normal table;
Obtain AREA(Z_U) = 0.311 from standard normal table;
Compute $PD_L = 100\ (0.058) = 5.8$; and
Compute $PD_U = 100\ (0.310) = 31.1$.

POINT ESTIMATE: PD = 15.01

INTERVAL ESTIMATE: $5.8 \leq PD \leq 31.1$ (0.95 Confidence)

4.6 Confidence Interval for Percent Defective (Normal Population, Double Limits)

Example: Spacing of Erosion Fence

CAUTION: The term "double limits" in the heading of this section refers not to a double-sided confidence interval, but to those applications for which percent defective in either or both tails of the population is of concern.

To illustrate this procedure, an entirely different example will be worked. Suppose an agency desired to place temporary erosion fence every 50 feet in a drainage ditch having severe erosion from a heavy rainfall. It is desired to have the spacing of every 50 feet, but not less than 40 feet or greater than 60 feet to maintain a uniform roadside

appearance. Assume a sample size of n = 30 is obtained, the sample mean and standard deviation have been computed to be $\bar{x}$ = 52.0 and s = 8.0, and the lower and upper limits are L_L = 40 and L_U = 60.

Using Equations 4.4 and 4.5, the Q values are computed to be as follows:

$$Q_L = \frac{(52.0 - 40.0)}{8.0} = 1.50$$

$$Q_U = \frac{(60.0 - 52.0)}{8.0} = 1.00$$

To avoid confusion with terminology, the PD values from the lower and upper tails of the population will be designated PD_1 and PD_2, respectively. Using the above two Q values, and a table for n = 30, the two PD values are obtained as follows:

LOWER TAIL: $PD_1 = 6.46$

UPPER TAIL: $PD_2 = 15.88$

TOTAL PD = $PD_1 + PD_2 + 6.46 + 15.88 = 22.34$

For the double-sided procedure, the same table is then used in reverse to obtain a Q value equivalent to a single-tailed value of PD = 22.34, producing Q = 0.764.

CAUTION: It is not appropriate to simply add the two individual Q values when using this procedure. (Note that this would produce Q = 1.50 + 1.00 = 2.50 which is drastically different from the correct composite value of Q = 0.764.)

Once the composite Q value has been obtained in the correct manner, the method is identical to the single-sided procedure using Equations 4.6 and 4.7. As in the previous examples, $\alpha/2 = 0.025$ and $Z_{\alpha/2} = -1.96$.

$$Z_L = -0.764 + (-1.96)\sqrt{\frac{1}{30} + \frac{0.764^2}{2(30)}} = -1.171$$

$$Z_U = -0.764 - (-1.96)\sqrt{\frac{1}{30} + \frac{0.764^2}{2(30)}} = -0.357$$

Again using a table of areas under the standard normal distribution:

AREA(Z_L) = 0.121, producing a lower confidence limit of PD_L = 12.1

AREA(Z_U) = 0.361, producing an upper confidence limit of PD_U = 36.1

The procedure for a normal population with a double limit is summarized as follows:

Sample n = 30;
Lower limit $L_L = 40$;
Upper limit $L_U = 60$
Calculate $\bar{x} = 52.0$;
Calculate $s = 8.0$;
Calculate $Q_L = 1.50$;
Calculate $Q_U = 1.00$;
Obtain $PD_1 = 6.46$ from table;
Obtain $PD_2 = 15.88$ from table;
Compute total PD = 6.46 + 15.88 = 22.34;
Obtain corresponding Q = 0.764 using table in reverse;
Select confidence level (0.95 in this example);
Compute $\alpha/2 = (1 - 0.95) / 2 = 0.025$;
Compute $Z_L = - 1.171$ using Equation 4.6;
Compute $Z_U = - 0.357$ using Equation 4.7;
Obtain AREA(Z_L) = 0.121 from standard normal table;
Obtain AREA(Z_U) = 0.361 from standard normal table;
$PD_L = 100\ (0.121) = 12.1$; and
$PD_U = 100\ (0.361) = 36.1$.

POINT ESTIMATE: PD = 22.34

INTERVAL ESTIMATE: $12.1 \leq PD \leq 36.1$ (0.95 Confidence)

Thus, the percent defective of temporary erosion fence (PD) outside the 40- to 60-foot uniformity interval, at 0.95 confidence, is 12.1% to 36.1%. These PD ranges could be compared against an established interval to warranty remedial action, if necessary. Note how the sample mean in this example (52 feet) was very close to the desired 50-foot interval, but the relatively high standard deviation (8 feet) produced approximately 22% PD.

4.7 Hypothesis Test for Mean of a Normal Population

Example: Width of Mowing

For this test, it is hypothesized that the population mean is equal to some specific value, μ (i.e., the null hypothesis). A random sample is obtained and then checked statistically to see if the values are sufficiently far from the hypothesized mean to conclude that the null hypothesis is false.

For purposes of this example, assume that an agency desires mowing to be a uniform distance of 10 feet from the edge of the shoulder. The null and alternative hypotheses can be stated as follows:

Null hypothesis: mean = $\mu = 10.0$ feet
Alternative hypothesis: $\mu \neq 10.0$ feet

The statistical procedure consists of calculating the sample mean ($\bar{x}$) and standard deviation (s), and then computing the t statistic given by Equation 4.8.

$$t = \frac{(\bar{x} - \mu)}{\left(\frac{s}{\sqrt{n}}\right)} \tag{4.8}$$

These statements describe a two-sided hypothesis test because the null hypothesis will be rejected if a sample mean is obtained that is either too small or too large to have occurred just due to random chance. A sample size of n = 15 was obtained in County #1 and the following calculations have been made:

$\bar{x}$ = 9.7 feet
s = 0.45 feet

Using Equation 4.8, the calculations is as follows:

$$t = \frac{(9.7 - 10)}{\left(\frac{0.45}{\sqrt{15}}\right)} = -2.582$$

The next step is to determine if the computed statistic is sufficiently far out in the tail of the t distribution (in either a positive or negative direction for a two-tailed test) that it warrants rejection of the null hypothesis. To determine the critical value of t to make this decision, the following information must be compiled before consulting a table of the t distribution:

n = 15 (sample size)
df = n – 1 = 14 (degrees of freedom for t statistic)
Level of significance of test = $\alpha = 0.01$ (for this example)
For two-tailed test, $\alpha/2 = 0.005$

Using this information to enter the table, $t_{df=14,\ \alpha/2=0.005} = 2.977$ is found to be the critical value. Since the absolute value (for a two-tailed test) of $|t| = 2.582$ computed with Equation 4.8 does not exceed the critical value obtained from the table, there is not sufficient basis to reject the null hypothesis. Therefore, at a significance level of $\alpha = 0.01$, there is no reason to believe that the true population mean is other than $\mu = 10.0$ feet of mowing width.

A significance level of $\alpha = 0.01$ is a fairly stringent test, however, and it will be interesting to see what would have happened if a more moderate significance level of $\alpha =$

0.05 had been chosen, in which case $\alpha/2 = 0.025$ would be used to determine the critical value of $t_{df=14,\ \alpha/2=0.025} = 2.145$. Since the absolute value of the computed t statistic of 2.582 exceeds the critical value in this case, the null hypothesis would be rejected and the alternate hypothesis accepted that the true population mean is something different from 10.0 feet of mowing width.

It should be noted that there is no "right" or "wrong" choice in choosing the level of significance to conduct tests such as this; it is more a matter of tradition and practical convenience. If the significance level is set at a low value (e.g., $\alpha = 0.01$), then the test will occasionally fail to detect a mean that differs from the assumed value by only a moderate amount. On the other hand, if a rather large significance level is chosen (e.g., $\alpha = 0.10$), then it may frequently be concluded that a difference exists when, in fact, there really was no difference. Since it is inevitable that errors of either type will occasionally be made, the choice of significance level is largely influenced by which type of error has the greater adverse consequences. As a practical matter, a significance level of $\alpha = 0.05$ is often chosen as a reasonable compromise.

The other application of this method to be considered is the single-tailed test. This is used when there is a concern only if the true population mean deviates in one particular direction, and little concern when it deviates in the other direction. In the previous example, suppose there were concern only if the population mean were lower than the assumed null value of $\mu = 10.0$ feet mowing width. In this case, exactly the same calculations would be made; only the determination of the critical t value would be different.

Since there is concern only if the sample mean is lower than the assumed population mean of $\mu = 10.0$ feet, the hypothesis test need not be applied unless the sample mean is less than 10.0 feet in this case. For sample means less than 10.0 feet, the t statistic computed with Equation 4.8 will be negative, and it will be considered statistically significant if its absolute value exceeds the absolute value of the critical t value in the lower half of the t distribution. In this case, the entire significance level (α) is put in the lower tail of the t distribution, and for $\alpha = 0.05$, for example, the critical value is $t_{df=14,\ \alpha=0.05} = 1.761$. (Technically, this is –1.761, but it is conventional just to list it as 1.761.) Finally, since the absolute value of the computed $t = 2.582$ is substantially larger than the critical value, the null hypothesis is clearly rejected by this single-sided test.

The hypothesis test for mean of a normal population for mowing width from edge of shoulder is summarized as follows:

Null hypothesis: mean = $\mu = 10.0$ feet.
Alternative hypothesis: $\mu \neq 10.0$ (two-sided test)
n = 15 (sample size)
$\bar{x} = 9.7$
$s = 0.45$

$$t = \frac{(9.7 - 10)}{\left(\frac{0.45}{\sqrt{15}}\right)} = -2.582$$

n = 15 (sample size)
df = n – 1 = 14 (degrees of freedom for *t* statistic)
Level of significance of test = $\alpha = 0.01$
For two-tailed test, $\alpha/2 = 0.005$
$t_{df=14,\ \alpha/2=0.005} = 2.977$ (obtained from *t* table)
$|-2.582| = 2.582 < 2.977$

Null hypothesis that $\mu = 10.0$ feet is not rejected at $\alpha/2 = 0.005$ significance level.

Therefore, it can be concluded that mowing from the unpaved shoulder edge is at a uniform width of 10.0 feet, based on the n = 15 segments representing the population.

4.8 Hypothesis Test for Difference between Means of Two Normal Populations

Example: Comparing Mowing between Two Districts

For this test, it is hypothesized that the means of two independent populations are equal to each other. Random samples are obtained from each population and then a statistical test is performed to determine if the sample means are sufficiently different to be judged to have come from two different populations.

In this case, the null and alternative hypotheses are stated as follows:

Null hypothesis: $\mu_1 = \mu_2$
Alternative hypothesis: $\mu_1 \neq \mu_2$

The statistical procedure consists of calculating the sample mean ($\bar{x}$) and standard deviation (s) for each of the two samples, and then computing the *t* statistic given by Equation 4.9.

$$t = \frac{\overline{x_1} - \overline{x_2}}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{4.9}$$

where S_P is the "pooled" standard deviation given by,

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{4.10}$$

with degrees of freedom of,

$$df = n_1 + n_2 - 2 \qquad (4.11)$$

This approach is valid providing the standard deviations of the two populations can be assumed to be at least approximately equal. If this assumption cannot confidently be met, the procedure becomes considerably more complex, and the reader is referred to a standard statistical text on hypothesis testing.

Like the test for a single mean, this can be either a one-sided or a two-sided test. If it is a two-sided test, the MQA manager is not concerned about which mean is larger than the other, only that they are significantly different. Once the sample *t* statistic has been calculated with Equation 4.9, the procedure is identical to that of the previous section.

For this example, suppose an MQA manager wants to compare the percentage of adequately mowed areas between two transportation districts within the state. Values around 80% of total area are common. The following statistics have been estimated from each district:

District #1	District #2
$n = 10$	$n = 15$
$\bar{x} = 80.0$	$\bar{x} = 85.0$
$s = 3.87$	$s = 3.16$

$$S_p = \sqrt{\frac{(10-1)3.87^2 + (15-1)4.46^2}{10+15-2}} = 4.239 \qquad \text{(from Equation 4.10)}$$

$$t = \frac{80.0 - 85.0}{4.239\sqrt{\frac{1}{10} + \frac{1}{15}}} = 2.889 \qquad \text{(from Equation 4.9)}$$

which, combined with a two-sided significance level of

$$\alpha/2 = 0.025$$

produces a critical *t* value of

$$t_{df=23,\alpha=0.025} = 2.069 \quad \text{(from } t \text{ table)}$$

DECISION: 2.889 > 2.069, so reject null hypothesis. Since the *t* statistic computed from the sample exceeds the critical value from the table, the means of the two populations are judged to be significantly different. Conclusion is that the percentage of areas with adequate mowing are different between the two districts.

4.9 Hypothesis Test for Dependent Normal Means (Paired-t Test)

Another extremely useful hypothesis test that involves the use of the *t* distribution is the comparison of dependent means, as might occur when it is desired to make a more efficient comparison of some specific test condition. For example, if it were desired to compare two different repair techniques, it is effective to design the experiment so that the same maintenance crew applies both techniques on the same day. The objective is to eliminate as many known differences as possible in order to minimize any random variability that might tend to obscure any real differences between the two methods under test.

In this case, the data are treated differently, as shown in Table 4.1. The last column in this table produces a total of n individual differences for which the null hypothesis is $\mu_D = 0.0$. If the computed *t* statistic exceeds the critical value obtained from the table, the null hypothesis is rejected and there is sufficient evidence that the two methods are not equally effective. Otherwise, it would be concluded that there is no significant difference between the two methods.

Table 4.1 Preparation of Data for Paired-*t* Test

Method #1 (1)	Method #2 (2)	Difference, D (3)
X_1	X_1	D_1
X_2	X_2	D_2
--	--	--
--	--	--
X_n	X_n	D_n

A double-sided test is performed if it is only desired to detect a difference between the two methods without regard to which method is superior. If the motivating hypothesis is that one treatment is better than the other, then a single-sided test would be applied. No example is given in this case because the actual application of the test is identical to the example in the section, *Hypothesis Test for Mean of a Normal Population*. A detailed example, and further discussion of this method is provided in the *Statistical Applications* chapter.

4.10 Hypothesis Test for Proportion of a Discrete Population

Example: Proportion of Complying Items in a Rest Area

Two different procedures are available to perform this test, and both are presented for instructional purposes. The proportion of sample items from a discrete population found to have some particular characteristic of interest is distributed as a binomial variable, and

the sampling distribution is described by an equation very similar to Equation 4.3, thus permitting direct probability calculations to be made. However, because such calculations can be quite tedious, most statistical tests recommend an alternate approach that involves the use of the standard normal distribution to approximate the binomial distribution under certain conditions that will be explained. Since the direct calculation approach has already been discussed in the section *Confidence Interval for a Proportion (Discrete Population)*, that method is presented first.

The section *Confidence Interval for a Proportion (Discrete Population)* covers the calculation of the confidence interval for a proportion estimated from a sample from a discrete population. In effect, the calculation of a confidence interval provides an alternative way to perform a hypothesis test for this proportion. In this case, the null hypothesis is that the true population proportion is equal to some specific level p_0. The sample estimate is obtained and the procedure of the section *Confidence Interval for a Proportion (Discrete Population)* is used to calculate the limits within which the true population value is expected to lie with some specified level of confidence $(1 - \alpha)$. If those limits do not include the hypothesized level of the proportion (p_0), then the null hypothesis is rejected and it is concluded that the true population proportion is something different from p_0.

(This approach, with the use of either the tables or the graphs described in the section *Confidence Interval for a Proportion (Discrete Population)*, will be found not only to be a convenient procedure in its own right, it will also be seen to be useful for a new ranking method presented in Chapter 6.)

The second method for performing a hypothesis test for a proportion from a discrete population involves approximating the sampling distribution with a standard normal distribution having $\mu = 0$ and $\Phi = 1$, i.e., a standard table of areas under the normal distribution. The procedure is as follows:

Null hypothesis H_0: True population proportion $p = p_0$;
Alternative hypothesis H_A: $p \neq p_0$;
Take sample of size n and count x, the number of items having the characteristic of interest;
Compute the sample proportion $p = x / n$; and
Compute the standard deviation of the sampling distribution for p:

$$\Phi = \sqrt{\frac{p_0(1-p_0)}{n}} \tag{4.12}$$

Compute the test statistic:

$$Z = \frac{p - p_0}{\Phi} \tag{4.13}$$

Select level of significance of test = α ($\alpha/2$ for two-tailed test)

Determine critical Z value from a table of the standard normal distribution
If |Computed Z| > |Critical Z|, reject H_0, otherwise accept.

(LIMITATION OF THIS PROCEDURE: $np_0 > 5$ and $n(1 - p_0) > 5$)

To illustrate both methods, suppose that an agency specified that rest area condition be at least 90 percent compliant (building interior/exterior, grounds, parking lot, etc.). In this case, the null hypothesis is that the proportion of complying items in the rest area population is $p_0 = 0.90$, and the alternative hypothesis is that this proportion is $p_0 < 0.90$. Since there will be concern only if the true population proportion is lower than the hypothesized value, this constitutes a one-sided test.

Further suppose that a sample of size n = 100 is available, and that a total of x = 85 items have been found to be in compliance. The first step is to check that the limitations of the procedure have been satisfied, i.e., that both $np_0 > 5$ and $n(1 - p_0) > 5$. For this example,

$$np_0 = 100(0.9) = 90$$

$$n(1 - p_0) = 100(1 – 0.9) = 10$$

Thus, this requirement is easily met. The next step is to compute the sample estimate of the population proportion as:

$$p = 85 / 100 = 0.85$$

which enables the remaining calculations to be made using Equations 4.12 and 4.13, as follows:

$$\Phi = \sqrt{\frac{0.85(1-0.85)}{100}}$$

$$Z = \frac{0.85 - 0.90}{0.0357} = -1.401$$

Assuming that a one-sided significance level of $\alpha = 0.05$ is chosen for this test, the corresponding critical value from the table of the standard normal distribution is found to be

$$Z_{crit} = – 1.645$$

Since the absolute value of the computed sample statistic of |Z| = 1.410 does not exceed the critical value of $|Z_{crit}| = 1.645$, there is no basis to reject the null hypothesis. Therefore, it is concluded on the basis of this test that there is no evidence to reject the claim that the true proportion conforming is 0.90. In other words, the rest area condition is in compliance.

For comparison purposes, it will be interesting to see what the calculation method based on Equation 4.3 will produce. In this case, for k = 85 and n = 100, 1 - α = 0.95 confidence limits on the true population proportion can be obtained (with interpolation) from a table as:

$p_L = 0.763$

$p_U = 0.913$

Since this interval includes the hypothesized value of $p_0 = 0.90$, there is no basis to reject the null hypothesis, and this procedure confirms the findings of the method using the normal-distribution approximation.

Note that it would virtually always be possible to construct an example in which one method rejected the null hypothesis by a very small amount while the other failed to do so by a correspondingly small amount. If the lower and upper confidence limits obtained with the procedure based on the binomial distribution were obtained without the need for interpolation, then it would be logical to assume that it is the more correct method since the procedure using the normal distribution is acknowledged to be an approximation. However, for practical purposes, it is believed that either procedure may be considered reliable provided the stated limitations of the normal approximation of $np_0 > 5$ and $n(1 - p_0) > 5$ are satisfied.

4.11 Hypothesis Test for Difference between Proportions of Discrete Populations

Example: Operability of Vending Machines between Two Highway Corridors

In the section *Hypothesis Test for Proportion of a Discrete Population*, two methods were presented by which a proportion obtained from a single population could be tested. One involved the summing of terms of a binomial expansion similar to Equation 4.3, and the other made use of a normal approximation of the binomial distribution. For the case in which it is desired to compare the proportions of two different discrete populations, the procedure using the normal approximation is generally recommended.

For this example, suppose an agency wanted to compare the operability of rest area vending machines between two interstate corridors, I-85 and I-95. Different vendors supply and maintain the machines on each corridor, and an agency would like to compare the proportion of operating vending machines on each corridor. The following population data have been collected, in other words, data from all vending machines:

I-85 (Population #1)	I-95 (Population #2)
$n_1 = 20$	$n_2 = 40$
$p_1 = 15 / 20 = 0.75$	$p_2 = 34 / 40 = 0.85$

Further assume that it is only desired to test if there is a difference between the two corridor populations. Therefore, this will be a two-sided test and the hypotheses will be stated as follows:

H_0: $p_1 - p_2 = 0$

H_A: $p_1 - p_2 \neq 0$

It should be noted that this procedure is subject to similar limitations as the procedure presented in the section *Hypothesis Test for Proportion of a Discrete Population*, namely that:

$n_1 p_1 > 5$
$n_2 p_2 > 5$
$n_1(1 - p_1) > 5$
$n_2(1 - p_2) > 5$

Again assume a significance level of $\alpha = 0.05$ is desired so that, since this is a two-tailed test, $\alpha / 2 = 0.025$ will be used, and the critical value of the Z statistic is:

$Z_{crit} = \pm 1.96$

The appropriate statistic and terms to be used to test the difference between two proportions are given by equations 4.14 through 4.16.

$$p = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2} \tag{4.14}$$

$$\Phi_{p_1 - p_2} = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{4.15}$$

$$Z = \frac{(p_1 - p_2)}{\Phi_{p_1 - p_2}} \tag{4.16}$$

The first step is to check that the necessary requirements are met:

$n_1 p_1$	=	20 (0.75)	=	15	
$n_2 p_2$	=	40 (0.85)	=	34	
$n_1(1 - p_1)$	=	20 (1 – 0.75)	=	5	(judged barely satisfactory)
$n_2(1 - p_2)$	=	40 (1 – 0.85)	=	6	

After this check is complete, the following calculations are made:

$$p = \frac{(20(0.75) + 40(0.85))}{(20 + 40)} = 0.817$$

$$\Phi_{p_1 - p_2} = \sqrt{0.817(1 - 0.817)\left(\frac{1}{20} + \frac{1}{40}\right)} = 0.106$$

$$Z = \frac{(0.75 \ - 0.85 \ )}{0.106} = -0.943$$

Since the absolute value of the computed sample statistic of $|Z| = 0.943$ does not exceed the absolute value of the critical Z statistic of $|Z_{crit}| = 1.96$, the null hypothesis is not rejected and there is insufficient evidence to consider these two population proportions to be different. In other words, the operability of vending machines on the two corridors is not considered different.

4.12 Hypothesis Test for Difference between Means of Several Normal Populations

Analysis of variance (ANOVA) methods can determine if the mean level among several normal populations are different or not, such as roadway class, district, county, or any other meaningful category to an agency. ANOVA is a very powerful statistical tool, and can readily determine if the features are different at specified risk levels. There are three fundamental types of ANOVA available, including the Fixed Effects Model (Model I ANOVA), the Random Effects Model (Model II ANOVA), and the Mixed Model, that uses a combination of fixed and random effects. At this time, the Fixed Effects Model is most appropriate for MQA data, since deliberate fixed effects are of interest (roadway functional class, geographic location, district, county, etc.). The Random Effects Model is appropriate where there is a single classification of data and no specific treatments are applied to any of the data. The objectives of the Random Effects Model are different from those of the Model I ANOVA. The Model I ANOVA makes a deliberate comparison among specific "treatments", or features of interest.

A hypothesis test determines if mean level between any number of features is significantly different or not. The null hypothesis, H_O, assumes they are not different, while the alternative hypothesis, H_A, hypothesizes they are different:

H_O: Features are not different (mean difference = 0).
H_A: Features are different (mean difference $\neq$ 0).

Two standard statistics are calculated and used to determine significance: (1) F-value and (2) p-value. The F-value calculates the ratio of mean variances "between" features and "within" features, and is then plotted on the F-distribution to determine a probability level of significance, or p-value. High F ratios yield p-values equal to or less than a traditional value of 5% would indicate the null hypothesis should be rejected. Equation 4.17 shows how the F-value is calculated using the ratio of mean squares (MS) both "between" and

"within" features. (The calculation for the mean squares several intermediate steps that are normally performed with a computer software program, and the reader is referred to detailed descriptions found in most statistical textbooks.)

$$F_{Feature} = \frac{MS\,(Between\ Feature)}{MS\,(Within\ Feature)} \qquad (4.17)$$

From a statistical definition, stratifying data occurs when the population is partitioned into regions or strata, and a sample is selected by some design within each stratum (Thompson 1992). In practical terms of MQA data, a stratum can be defined as a region, district, county, or some other meaningful grouping, provided that the sampling design is valid. From the review of current MQA practices, all states are using some form of random sampling. This is a valid sampling design when stratifying data, since all roadway segments in the state highway system (i.e., population) have been given an equal opportunity to be included in the collected sample. The overall purpose of stratification is to determine whether significant differences exist in facility maintenance performance for the different functional classification systems, geographical regions or jurisdictions. A specific application of this method using actual MQA data is provided in the following chapter.

4.13 Sample Size Determination

Example: Samples Needed to Estimate Lineal Feet of Shoulder Drop-Off or Build-Up

The precision of a statistical estimate, or the power of a statistical test to detect a real difference of some sort, is strongly influenced by the amount of data available. Because the square root of the sample size (n) appears in the denominator in Equations 4.1 and 4.2, for example, the confidence interval can be made as narrow as desired by taking increasingly larger samples. In general, the price to be paid for a more precise estimate or more discriminating power is the cost of obtaining a larger sample, and there will usually be a point of diminishing returns beyond which it will be impractical to further increase the sample size.

To estimate the sample size needed, it is first necessary to specify the level of risk one is willing to take. In the case of a confidence interval on the mean discussed in the section *Confidence Interval for the Mean of a Normal Population*, this involves the selection of the "alpha" risk that, in turn, influences the *t* statistic that appears in Equations 4.1 and 4.2. If the plus-or-minus precision with which the population mean is to be estimated is designated by PREC, then either of these equations can be transposed to produce

$$PREC = \frac{t(s)}{\sqrt{n}} \qquad (4.18)$$

such that

$$n = \left(\frac{t(s)}{PREC} \right)^2 \qquad (4.19)$$

However, two problems with this approach become immediately apparent. First, it is necessary to have an estimate of the standard deviation (*s*) in order to solve Equation 4.19 for the sample size n, and second, the value of *t* in this equation is itself dependent on n, thus requiring an iterative trial-and-error solution.

This first problem can usually be overcome by substituting a realistic estimate of the standard deviation obtained from prior knowledge or historical data. The second problem is the more difficult one, and a common solution is to assume that the standard deviation is sufficiently well known that the *z* statistic (standard normal distribution) can be used in place of the *t* statistic in Equation 4.19, producing Equation 4.20 and making it possible to solve for n directly. Alternatively, many statistical texts include tables that will provide the necessary sample size for situations in which the standard deviation is not known precisely.

$$n = \left(\frac{z(\sigma)}{PREC} \right)^2 \qquad (4.20)$$

where

n	=	required sample size;
z	=	standard normal variates associated with significance level α for a one-sided test or $\alpha/2$ for a two-sided test;
σ	=	assumed known (or reasonably well known) standard deviation;
PREC	=	desired level of precision.

For example, suppose that lineal feet of drop-off or build-up for an unpaved shoulder, 0.1-mile segment of roadway usually ranges from 50 - 100 lineal feet with a typical standard deviation of about 5 lineal feet, and that it were desired to estimate it within plus or minus 2 lineal feet. If a two-sided risk level of $\alpha/2 = 0.025$ is chosen, then $z_{\alpha/2=0.025} = 1.96$ (from a standard normal table) and Equation 4.20 becomes,

$$n = \left(\frac{1.96(5)}{2} \right)^2 = 24.01$$

which in this case would probably be rounded to n = 24. (Normally, it is appropriate to round up to the next larger integer.) Thus, the agency should randomly sample n = 24 segments for lineal feet of shoulder drop-off or build-up to estimate length within ±2 lineal feet.

Note that this procedure has provided an estimate of the sample size needed to make a confidence interval statement of ±2 about the mean obtained from the sample, based on an assumed known value for the standard deviation. It is possible that when the sample is

actually taken, and the confidence interval is calculated using Equations 4.1 and 4.2 and the standard deviation calculated from the sample, that the interval could be somewhat broader than desired. If this should occur, and the width of the interval is not acceptable, then the only recourse is to obtain a larger sample in order to obtain a more precise estimate.

Alternatively, many standard statistical texts contain tables especially designed for this purpose, such as *Table A-11, Number of Observations for t Test of Mean*, found in the textbook by Rickmers and Todd (1967). Using the same assumed values in the foregoing example, a value of n = 26 was obtained, indicating that the lack of knowledge of the standard deviation exacts a price in terms of required sample size.

For the case involving a proportion of a discrete distribution discussed in Section 4.3.3, a different procedure must be used. One source (Natrella 1966) indicates that an exact solution for sample size could be obtained from tables of the binomial distribution to the extent that such tables are available, but the frequent need to interpolate in two directions, or often extrapolate, greatly limits the practicality of this approach. Instead, it is preferable to use standard tables constructed specifically for this purpose. Tables A-25 and A-26 in the Natrella reference are one example of suitable tables for this application.

For the determination of confidence intervals for percent defective (PD) or percent within limits (PWL), presented for single-sided and double-sided limits in the sections *Confidence Interval for Percent Defective (Normal Population, Single Limit)* or *Confidence Interval for Percent Defective (Normal Population, Double Limits)*, respectively, no suitable reference was found for determining the necessary sample size. However, an approximate method can be proposed. Conceptually, a sample could be obtained using a trial sample size, and then confidence limits could be calculated for either the single-limit or double-limit case as previously described. If the confidence interval were found to be too broad, additional samples could be obtained and the calculations repeated with the larger (combined) sample in order to obtain the desired level of precision. This approach can be approximated if historical data are available prior to sampling that could be used to provide likely values for the average and standard deviation that must be entered into Equations 4.4 and/or 4.5. In this case, the calculations would proceed with an assumed sample size (n) to determine the likely outcome, thus enabling the user to gauge the suitability of the choice of sample size and modify it, if it seems necessary. This method is admittedly crude, but may nonetheless be effective from a practical standpoint.

For the hypothesis test for the mean of a normal population discussed in the section *Hypothesis Test for Mean of a Normal Population*, there is a great similarity to the procedure for computing the confidence interval about the sample mean described in the section *Confidence Interval for the Mean of a Normal Population*. If the confidence interval contains the hypothesized value of the mean, then there is no statistical basis to reject the null hypothesis. This reasoning leads again to Equation 4.19 so that, if a reasonably accurate value of the standard deviation is known from historical data, it is possible to solve directly for *n*. Otherwise, a table such as Table A-11 in the Rickmers and Todd reference may be used (Rickmers and Todd 1967).

For the case of a hypothesis test concerning the difference between two population means described in the section *Hypothesis Test for Difference between Means of Two Normal Populations*, a similar but different table is used to determine the required sample size. One such table is *Table A-12, Number of Observations for t Test of Difference Between Two Means*, in the reference by Rickmers and Todd (1967).

Since the hypothesis test for the difference between two dependent means (paired *t* test) presented in the section *Hypothesis Test for Dependent Normal Means (Paired-t Test)*, is similar in structure to the hypothesis test for the mean described in the section *Hypothesis Test for Mean of a Normal Population*, essentially the same procedure can be used to determine the appropriate sample size. In theory, Equation 4.20 would again apply, except that the appropriate value of the standard deviation to be substituted into Equation 4.20 would be the standard deviation of the differences listed in the last column of Table 4.1. As before, for this approach to be practical it would be necessary to have an estimate of the standard deviation of these differences from historical data to justify the use of the standard normal *z* value in Equation 4.20. Alternatively, Table A-11 in the Rickmers and Todd reference may be used (1967).

Like the hypothesis test for the mean of a normal population, the hypothesis test for a proportion estimated from a sample from a discrete population presented in Section 4.3.9 is very similar to the procedure for computing the confidence interval about the sample proportion described in the section *Confidence Interval for a Proportion (Discrete Population)*. Consequently, the same method can be used to determine the appropriate sample size for this procedure, and Tables A-25 and A-26 in the Natrella reference are recommended (Natrella 1966).

According to the Natrella reference (1966), the determination of the required sample size for a hypothesis test for the difference between two proportions described in the section *Hypothesis Test for Difference between Proportions of Discrete Populations* is complicated by the fact that the sample size depends on the true but unknown values of the proportions involved. However, it is usually possible to make sufficiently realistic assumptions in order to obtain a reasonably accurate estimate of the sample size. The procedure is somewhat involved, and the reader is referred to Chapter 8 of the Natrella handbook (1966), or to other similar publications.

To determine the required sample sizes for the analysis of variance (ANOVA) procedure described in the section *Hypothesis Test for Difference between Means of Several Normal Populations,* any number of samples can be used, but $n > 10$ per population is recommended to ensure a conclusive result. That size will lessen the effect that an individual data point has upon the F-ratio calculation. The ANOVA procedure accommodates a range of sample sizes in the determination, and the p-values associated with the chosen sample size can be found in most textbooks.

4.14 Random Sampling

Sampling consists of selecting some part of a population to observe so that one may estimate something about the whole population (Thompson 1992). A data acquisition or sampling plan can be adequately designed to obtain any feature of interest. Several sampling designs are available to understand and measure the feature of interest, such as simple random sampling, sampling with replacement, sampling without replacement, stratified random sampling, cluster and systematic sampling, multistage designs, double sampling, and network sampling. More sophisticated designs also exist, such as capture-recapture sampling, line-intercept sampling, Latin Squares, and stratified adaptive cluster sampling

It is recommended that the sampling process be simple and clearly understood to be effective. For MQA Programs, simple random sampling or stratified random sampling are the most effective and easily understood methods to collect a sample, based upon the commonly accepted practices disclosed in the MQA manual review.

Randomization is extremely important to the sampling process. It allows each part of the population an equal chance of being selected and it protects against unsuspected sources of bias. Violation of the randomization principle can produce biased samples that will inaccurately reflect true characteristics of the population, in particular, "trying to be fair" by knowingly and purposely sampling a certain segment. A random number generator is the most effective and recommended approach to ensure randomization, where segments are drawn based on a random number assignment.

The following guidelines apply to the MQA sampling process:

1. Length among sample segments must be equal. This allows an equal chance of encountering the feature of interest (both attributes and variables).

2. Sample length can be any length (0.1 mile, 0.2 mile, 1 mile, etc.) as long as the sample length is equal among segments. This allows an equivalent region to be measured among samples.

3. If a specific strata is of interest, such as interstate highways within the state or a given district, overall county highway system, etc., subdivide the strata into equal segments and randomly sample the segments. Do not discard any segments, since this will introduce bias in the statistical parameters.

4. If a specific feature is desired during the sampling process, such as culvert pipe blockage or drainage ditches, and the feature is not encountered in the randomly-sampled segment, a stratified random sampling approach is recommended. Divide the population segment into equal stratified segments having the feature, then randomly choose a predetermined number of stratified segments. For example, for a 1-mile population segment, subdivide into ten 0.1-mile segments, identify which segments contain the specific feature, then randomly select from

those 0.1-mile segments. If a certain sample size is desired, use the earlier section for guidance, *Sample Size Determination*. Caution – this approach is strictly limited to the feature of interest, and not recommended for surveying all roadway elements within the chosen segment (pavement, ditches, traffic, etc.), since some segments had been inadvertently discarded by not having the original feature of interest.

CHAPTER 5 STATISTICAL APPLICATIONS

5.1 Introduction

The previous chapter presented basic statistical procedures for use in a range of MQA data analysis situations. This chapter provides statistical applications to more specific questions or concerns expressed at both the Maintenance Quality Assurance Peer Exchange held in October 2004 at Madison, Wisconsin, and May 2005 kick-off meeting for this project. Table 5.1 presents the specific application and the statistical tool(s) or method(s) recommended for the application. These applications are meant to provide a deeper understanding of the practicality, and relative simplicity, that statistics can provide for a range of situations in MQA practice.

The statistical application began by choosing two lead states in MQA programs that had readily available data for analysis, including North Carolina and Wisconsin. Data were transferred via email attachment or mailed disc. MQA manuals were reviewed from the select states to understand how the MQA data were collected and reported. During this review, potential problems and concerns in understanding the data were noted, and are presented throughout this chapter. A term known as "data cleansing" was conducted to yield valid statistics. For example, alpha characters were converted to numerical values, and blank cells were identified and marked. The "data cleansing" term will be revisited during discussion of trouble areas or statistical pitfalls later in this chapter.

The data were analyzed using commercially available software, including Microsoft Excel™, SAS™, and StatGraph™. These packages were used due to familiarity with programming code and efficiency. It is not the intent of this study to advocate any software program; each agency must choose a package appropriate for their needs. FHWA has developed several software packages specifically for state highway agencies and, at the time of this writing, it is believed that contracts for additional software are under consideration.

Table 5.1 MQA Application and Statistical Approach

Index (1)	Application (2)	Statistical Approach (3)	Equation (4)
1	*Determining number of samples to yield valid information*	Statistical distributions and statistical parameters can assist in determining sample size. Any number of samples is valid, provided there were randomly chosen. The t-distribution is used for small sample sizes when the population is not known (n<30), while the normal distribution (z statistic) is used for larger sample sizes when the population statistics are known (or reasonably well known). Precision and confidence interval estimation equations can provide an interrelationship of sample size, standard deviation, and confidence level.	Population parameters unknown and only estimated $n = \left(\frac{t(s)}{PREC} \right)^2$ n<30 Population parameters known $n = \left(\frac{z(\sigma)}{PREC} \right)^2$ n>30
2	*Developing confidence in an estimate*	Confidence Intervals can be constructed around the average using the chosen level of confidence (i.e., 95%), underlying variability, and sample size.	$C.I. = \overline{LOS_s} \pm \left(Z * \frac{s}{\sqrt{n}} \right)$
3	*Stratifying data in terms of geographical or highway features*	Analysis of Variance can detect whether there is a difference between features of data (region, crews, etc.), while incorporating the variability into the determination. Some states may have significant differences between regions that can be detected using Analysis of Variance.	$F_{Feature} = \frac{MS\,(Between\,Feature)}{MS\,(Within\,Feature)}$
4	*Comparing results of MQA data collectors*	T-tests can detect the mean difference between data sets generated by two different MQA data collectors. Then, with the assistance of Power Curves , the "true" mean difference between them can be measured and controlled. Standard statistical tests are available with which to compare data sets obtained by different MQA technicians or teams.	$n = \sigma_D^2 \frac{(Z_{\alpha/2} + Z_{\beta})^2}{d^2}$
5	*Are years different or not*	F-tests and t-tests can provide a statistical comparison of means. Paired-sample t-tests are used when data are collected from the same roadway segment from year-to-year, while a two-sample t-test is used when the roadway segments are independent of each other.	$\left\lvert \bar{X}_1 - \bar{X}_2 \right\rvert < t_{\alpha,\nu} \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)}$

Table 5.1 Cont.

Index (1)	Application (2)	Statistical Approach (3)	Equation (4)
6	*Looking for trouble signs*	Outliers, or data points that are abnormal from a distribution, can detect trouble signs. Several standard tests for outliers exist, and the chi-square or other goodness-of-fit tests can be used to check normality.	$Z_i = \frac{X_i - \bar{X}}{S}$
7	*Reporting Data*	Beginning with simple fundamental statistical measures is always the best start (plot the data, calculate the average and standard deviation, etc.). The sampling design largely drives if/how a statistically-valid analysis can proceed, so effort must be placed on sampling design at the beginning.	Line graph, bar chart. pie chart, table, and box-and-whisker plots.

5.2 Sample Size and Confidence Limits

Sample size and the confidence of population estimates are a fundamental concern in any MQA program. Equations 2.6, 2.7, 2.8, 4.19, and 4.20 are appropriate to determine the number of sample segments for individual MQA measures or LOS at statewide, functional class, district, county level, or any other meaningful strata. There are four inter-related components to these equations that can be evaluated, including: (1) precision or confidence limits of the sample, (2) t-statistic or Z-statistic for the probability level, (3) estimated standard deviation, and (4) number of samples. As Equations 2.8, 4.19, and 4.20 imply, there is a lesser number of samples required when the precision is reduced, when the probability level is reduced, or when the standard deviation is small.

Reasonable estimates for the standard deviation can be found from collected data within each MQA program, however, consideration must be given to the strata of interest. Earlier examples of determining sample size using precision were provided in Chapter 4 (see section *Sample Size Determination*). These examples would apply to any MQA measure using variables data.

From the review of MQA manuals and relevant literature, LOS is a common measure of facility condition. North Carolina data from 2000, 2002, and 2004 were collected and analyzed to apply Equations 2.6, 2.7, and 2.8 to determine sample size and confidence limits of mean statewide LOS. North Carolina DOT weights values of the roadway elements to compute the overall LOS for sample segment, as shown earlier by Equation 2.2.

Summary statistics and a histogram plot for NC 2000 statewide LOS values are provided in Figure 5.1. The histogram shows strong central tendency and approximation of a normal "bell-shaped" distribution, with a $\bar{x}$ = 75.3 and s = 10.8. Because of the large sample size (n > 30) and normal distribution, these sample statistics could be considered population parameters for mean (μ) and standard deviation (σ) of LOS in 2000. These

statistics were derived from 2,436 randomly chosen pavement sections. In 2002, a total of 3,462 segments yielded LOS statewide statistics of $\bar{x}$ = 76.4 and s = 12.2. In 2004, a total of 10,932 segments produced LOS statewide statistics of $\bar{x}$ = 79.5 and s = 8.7. Thus, statewide LOS from 2000 to 2004 had the mean trending upwards from 75.3 to 79.5, while the standard deviation fluctuated from 10.8 to 12.2, then down to 8.7. It is apparent that LOS increased in this time period, but an application provided later in this chapter (see *Comparison of Means*) can be conducted to state this conclusion with confidence.

Based on these summary statistics, confidence limits were developed for the mean LOS using Equation A.8. Table 5.2 provides the relationship of sample size and confidence limits at a 95% probability level for standard deviation values of 8.7 and 12.2, respectively. Sample size was incrementally increased by n = 20 to illustrate the effect on the confidence limits.

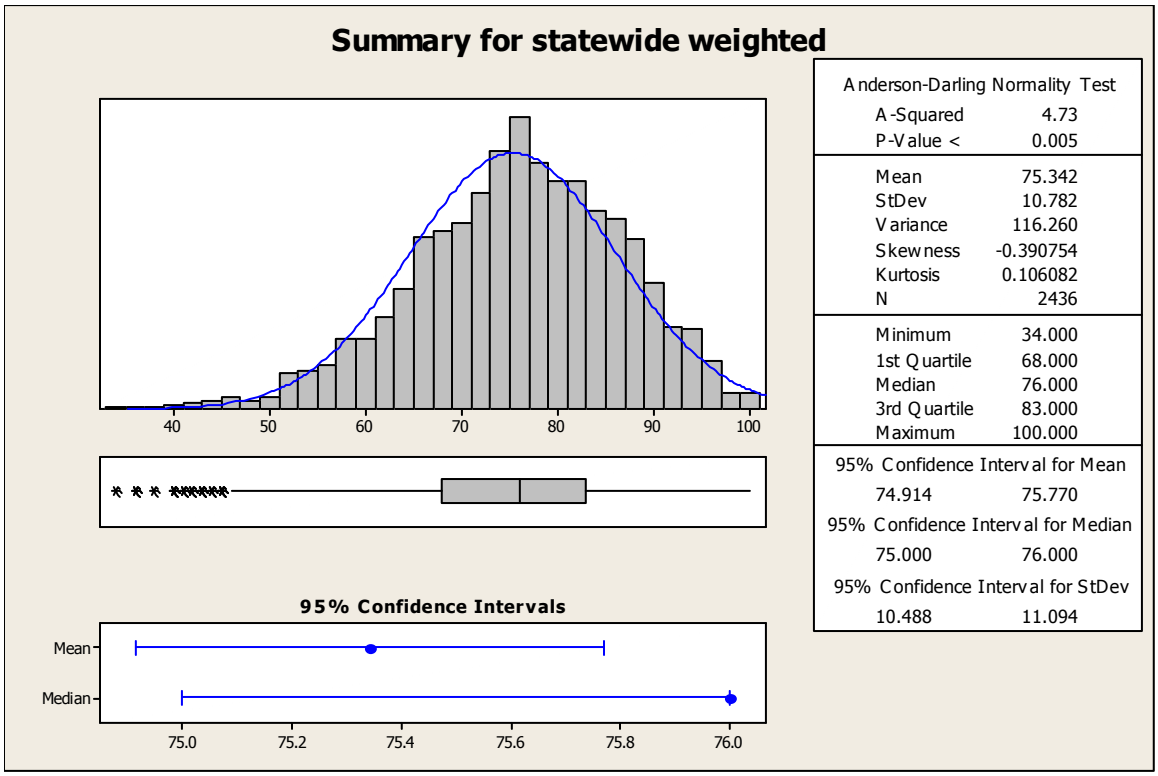


Figure 5.1 North Carolina Statewide Weighted LOS (2000)

For s = 8.7, nearly 300 samples are necessary to achieve a confidence limit of +/- 1 LOS. Over 500 samples are needed to achieve the same level when s = 12.2. Clearly, a substantial number of segments are needed to develop confidence limits for +/- 1 LOS, based on the variability and specified probability level. In fact, the relationship is non-linear due to the nature of Equation A.8, with an increasing number of samples needed to proportionally reduce the confidence limits.

This diminishing return on investment (more samples) is a concern to any agency, since sampling consumes limited resources. A simple way to reduce the number of samples is wider confidence limits or a decrease in probability level, however, this exposes the agency to greater risk of underestimating or overestimating the true mean LOS. After an examination of these statistics, a consensus value can be chosen for sample size determination. If an agency cannot comfortably choose a value, sensitivity analysis is recommended by holding all equation variables constant and adjusting the single variable in question. This application can be used across a range of agency levels, such as statewide, division or district, or county level.

Table 5.2 Confidence Limit Estimate at 95% Probability Level

s = 8.7 LOS		*s* = 12.2 LOS	
Sample Size (1)	C.I. +/- (2)	Sample Size (3)	C.I. +/- (4)
20	3.81	20	5.35
40	2.70	40	3.78
60	2.20	60	3.09
80	1.91	80	2.67
100	1.71	100	2.39
120	1.56	120	2.18
140	1.44	140	2.02
160	1.35	160	1.89
180	1.27	180	1.78
200	1.21	200	1.69
220	1.15	220	1.61
240	1.10	240	1.54
260	1.06	260	1.48
280	1.02	280	1.43
300	0.98	300	1.38
320	0.95	320	1.34
340	0.92	340	1.30
360	0.90	360	1.26
380	0.87	380	1.23
400	0.85	400	1.20
420	0.83	420	1.17
440	0.81	440	1.14
460	0.80	460	1.11
480	0.78	480	1.09
500	0.76	500	1.07

5.3 Data Stratification and ANOVA

The following two sections provide applications of stratifying data for Wisconsin and North Carolina.

5.3.1 Wisconsin, Hazardous Debris Data

An example of basic statistics for Hazardous Debris for 2004 Wisconsin MQA data are provided in Table 5.3. Frequency of hazardous debris in randomly-chosen 0.1-mile segments were recorded for approximately 240 segments in eight districts, and the mean and standard deviation for segments were calculated. To address the question, "How far 'down' in terms of geographical subdivision and features can I go", the Wisconsin data were stratified into (1) roadway classification (divided and undivided), (2) 8 transportation districts, and (3) 72 counties. Only the first 5 counties in the data set are shown to simplify interpretation of the data.

Table 5.3 Basic Summary Statistics for Wisconsin 2004 Hazardous Debris Data

Geographic or Attribute Division (1)	Sample Size, n (2)	Mean, count (3)	Standard Deviation, count (4)
Statewide	1892	0.251	1.041
Divided Highway	345	0.530	1.587
Undivided Highway	1547	0.178	0.777
District 1	231	1.026	2.214
District 2	239	0.138	0.460
District 3	239	0.138	0.478
District 4	235	0.068	0.386
District 5	236	0.144	0.595
District 6	237	0.063	0.391
District 7	239	0.201	1.160
District 8	239	0.251	0.872
County 1	14	0.214	0.426
County 2	19	0.105	0.315
County 3	29	0.000	0.000
County 4	22	0.227	0.612
County 5	33	0.152	0.508

The statewide average for Wisconsin in 2004 indicates a mean of 0.251 pieces of debris per 0.1-mile segment. When comparing roadway class (divided versus undivided), the divided highways had a greater frequency of hazardous debris, and higher variability, as indicated by the standard deviation. When the data are stratified into districts, District 1 had a higher frequency and variability. Districts 4 and 6 had a much lower frequency of

hazardous debris. For counties, the mean level varied, with no hazardous debris found in County 3.

This simple breakdown of the data into strata summary statistics, however, did not have the capability of determining whether the mean hazardous debris level among roadway classes, districts, or counties were different from each other. One could argue that they appear different, but a formal determination is necessary that is defensible and statistically valid.

Table 5.4 provides the ANOVA results (using Equation 4.17 as a basis) for each variance component, degrees of freedom (pieces of data used in determination minus one for the mean itself), and expression for the Mean Square (MS). Results of the ANOVA conclude that the three methods for subdividing hazardous debris are significantly different (roadway class, district, and county). In practical terms, the level of hazardous debris is higher on divided highways, and different among districts and counties. This finding is important, since it indicates an inconsistent mean level for this maintenance quality indicator for roadway class, geographic subdivision by district, or county. Action, as the agency deems necessary, can then be focused on the higher mean level segments, or where there is a high degree of variability, as measured by the standard deviation.

Table 5.4 Results of ANOVA for Wisconsin 2004 Hazardous Debris Data

Geographic or Attribute Division (1)	Degrees of Freedom, n (2)	Sum of Squares, count^2 (3)	Mean Square, count^2 (4)	F-value, ratio (5)	p-value, %/100 (6)	Significantly Different, Yes/No (7)
(a) Roadway Class (Divided and Undivided)						
Between Features	1	34.701	34.701	36.44	0.0001	Yes
Within Features	1890	1799.973	0.952	---	---	
Total	1891	1834.674	---	---	---	
(b) District						
Between Features	7	164.344	23.478	23.46	0.0001	Yes
Within Features	1887	1888.091	1.001	---	---	
Total	1894	2052.435	---	---	---	
(c) County						
Between Features	71	671.715	9.461	12.49	0.0001	Yes
Within Features	1823	1380.719	0.757	---	---	
Total	1894	2052.435	---	---	---	

5.3.2 North Carolina, Blocked Ditch Data

A stratification and analysis of blocked ditches in North Carolina was performed for 2000, 2002, and 2004 data. A feature of North Carolina MQA data was that random 0.2-mile segments are sampled within each mile of roadway. For the blocked ditch measure, total lateral ditch length parallel to the roadway was recorded (North Carolina 1998). For example, if a typical two-lane, two-way roadway is being inspected, the total length will be 2,112 feet (0.2 mi. x 5,280 ft. x 2 ditches). For the assessment of a divided roadway that has a single median longitudinal ditch, the total ditch length will be 3,168 feet (0.2 mi. x 5,280 ft. x 3 ditches). Then, within the 0.2-mile ditch segment, the longitudinal length with 50% or more blockage was measured and recorded. Thus, the maximum length of at least 50% blockage for any given 0.2-mile segment of ditch would be 1,056 feet.

The data were stratified by (1) roadway classification (interstate, primary, secondary, and urban), (2) 14 transportation districts, and (3) 100 counties. Table 5.5 shows the basic summary statistics for North Carolina blocked ditch data for the three years. Some roadway segments had no ditches, and those segments were excluded from the analysis. For example, there were 2,681 roadway segments sampled in 2000, but only 2,300 segments had longitudinal drainage ditches parallel to the traveled roadway. Only the first four divisions and four counties in the data set are shown to simplify interpretation.

One visible feature in the summary statistics is the drop in mean statewide level, and variability, from 2000 to 2004. From randomly chosen 0.2-mile roadway segments, the average length of blocked ditches dropped from 124 to 44 lineal feet.

When the data were stratified by functional class, the mean level was higher for interstate and urban sections. From years 2000 to 2004, the mean level and standard deviation dropped within primary, secondary, and urban classes. Interstate roads had fluctuations in mean level and standard deviation.

Stratification by division showed inconsistent yearly mean levels. When comparing blockage from 2000 to 2004 within each division, Divisions 1 and 3 had a reduction in mean lineal feet, while Division 2 and 4 had fluctuations. Stratification by county also measured inconsistent mean levels from year to year, and when comparing any two counties for a given year. Standard deviation values also varied.

This analysis highlighted a potential concern for practitioners in the manner in which these data were analyzed. Some 0.2-mile roadway segments had 2 ditches, while others had 3 or 4 ditches. Thus, some roadway segments had a greater lineal length "sample" of blocked ditches than others, creating a possible bias in the data, particularly for divided highways. In the LOS calculation for drainage, North Carolina does account for the difference in number of ditches.

Table 5.5 Summary Statistics for North Carolina Blocked Ditch Data

Geographic or Attribute Division (1)	Year (2)	Sample Size, n (3)	Mean, L.F. (4)	Standard Deviation, L.F. (5)
Statewide	2000	2300	123.5	280.3
	2002	3314	72.5	217.0
	2004	9736	44.2	156.4
Interstate	2000	253	149.4	351.7
	2002	246	99.0	329.5
	2004	264	132.5	362.0
Primary	2000	621	122.7	283.4
	2002	1040	58.9	198.1
	2004	1659	29.7	118.6
Secondary	2000	1058	105.7	248.4
	2002	1658	74.7	213.0
	2004	2607	40.6	131.0
Urban	2000	368	158.5	302.0
	2002	370	83.0	188.1
	2004	5206	46.2	160.0
Division 1	2000	142	246.5	386.0
	2002	211	190.8	385.5
	2004	594	55.5	179.5
Division 2	2000	135	42.4	81.7
	2002	206	69.2	199.9
	2004	644	38.9	225.5
Division 3	2000	167	220.5	417.9
	2002	224	117.6	383.3
	2004	679	52.0	158.4
Division 4	2000	182	210.2	369.7
	2002	236	40.1	165.0
	2004	742	191.5	303.1
County 0*	2000	31	65.5	142.9
	2002	49	2.0	14.2
	2004	---	---	---
County 1	2000	15	14.6	20.3
	2002	25	28.9	74.3
	2004	130	9.0	67.7
County 2	2000	16	55.8	110.6
	2002	15	97.9	201.9
	2004	58	11.4	29.2
County 3	2000	28	141.2	235.3
	2002	54	14.0	46.7
	2004	61	18.2	135.2
* In 2004, County numbering changed from 0-99, to 1-100.				

To address a potential imbalance that may occur in practice, a calculation was made for the average blocked lineal feet of 0.2-mile long *ditches*, not the blocked ditch length for a 0.2-mile *roadway segment*. An average blocked length was computed for each segment by dividing the total measured blocked length by number of ditches. For example,

Interstate 95 in Division #4 (County #63) between Mileposts 140.43 and 140.63 had 3,168 lineal feet of ditch inventory, and a total of 100 L.F. of blocked ditch. The 3,168 L.F. total was divided by 1,056 L.F. to yield 3 ditches in this roadway segment. Then, the 100 L.F. total blocked ditch length was divided by 3 ditches to yield an average of 33.33 L.F. blocked per 0.2-mile ditch.

Table 5.6 provides a comparison of year 2000 statistics derived from 0.2-mile roadway segments, and the average blocked length for 0.2-mile ditch in the same 0.2-mile segment. (Only 2000 data are provided to simplify interpretation). A significant result of the investigation was the reported mean levels for functional class. When roadway segments are considered, interstate and urban roadways had a higher level of blocked ditches, however, when the average blocked ditch length was calculated, interstates had the lowest level. Thus, it could be concluded that interstate roadways had the highest level of service with respect to individual blocked ditches.

Table 5.6 Summary Statistics for North Carolina 2000 Blocked Ditches

Geographic or Attribute Division (1)	0.2-mile Roadway Segment			Average of 0.2-mile Ditches in Segment		
	Sample Size, n (2)	Mean, L.F. (3)	Standard Deviation, L.F. (4)	Sample Size, n (5)	Mean, L.F. (6)	Standard Deviation, L.F. (7)
Statewide	2300	123.5	280.3	2300	69.6	151.4
Interstate	253	149.4	351.7	253	48.8	114.9
Primary	621	122.7	283.4	621	74.6	158.4
Secondary	1058	105.7	248.4	1058	60.4	134.6
Urban	368	158.5	302.0	368	101.5	195.5
Division 1	142	246.5	386.0	142	128.8	198.7
Division 2	135	42.4	81.7	135	28.4	60.8
Division 3	167	220.5	417.9	167	107.5	199.2
Division 4	182	210.2	369.7	182	95.0	175.6
County 0	31	65.5	142.9	31	47.5	107.2
County 1	15	14.6	20.3	15	11.0	16.1
County 2	16	55.8	110.6	16	52.4	110.6
County 3	28	141.2	235.3	28	70.6	117.6

Another result of this comparison was the reduction in variability. Clearly, a less volatile measure is provided for individual ditches, rather than the combination of ditches in a segment. For example, a roadway with two ditches has a range of 0 to 2,112 L.F. available to be blocked, while an individual ditch has 0 to 1,056 L.F.

When comparing divisions, the blocked lineal feet dropped by a factor of two from roadway segments to the ditch average, suggesting that most roadways in districts have an average of two ditches. A “canceling” effect also occurs between roadways with a single ditch and 3 ditches, and no ditches and 4 ditches. Comparison of county data

showed a slight decrease in the mean and variability for Counties #0, #1, and #2, while County #3 estimates were cut in half.

To determine whether the mean levels of the average blocked ditch segments were similar or different for functional class, division, or county, an analysis of variance was performed and results reported in Table 5.7. The conclusion for the three stratification groupings was the same; there was a different mean level.

Table 5.7 Results of ANOVA for North Carolina 2000 Blocked Ditch Data

Geographic or Other Attribute (1)	Degrees of Freedom, n (2)	Sum of Squares, L.F.^{2} (3)	Mean Square, L.F.^{2} (4)	F-value, ratio (5)	p-value, %/100 (6)	Significantly Different, Yes/No (7)
(a) Functional Class						
Between Feature	3	589245	196415	8.66	0.0001	Yes
Within Feature	2296	52103701	22693	---	---	
Total	2299	52692946	---	---	---	
(b) Division						
Between Feature	13	4545499	349654	16.60	0.0001	Yes
Within Feature	2286	48147447	21062	---	---	
Total	2299	52692946	---	---	---	
(c) County						
Between Feature	99	8772554	88612	4.44	0.0001	Yes
Within Feature	2200	43920393	19964	---	---	
Total	2299	52692947	---	---	---	

Results of the analysis of data stratification yield the following observations and guidelines:

- Stratification is an important tool in understanding MQA data. An agency is better able to understand features of the data when the samples are partitioned into groupings of interest, such as functional class, district/division, or county. Then, maintenance can be better managed since more useful statistics can be provided at the level of interest.

- An analysis of variance provides the best tool to determine whether there is a difference between features within any stratum of interest. With factual-based knowledge of a difference, necessary action can then be taken to correct higher (or lower) mean levels, or try to minimize variability to yield a consistent level of service.

- Stratification can provide useful statistics on which to base a sampling plan. For example, if an agency wants to develop a division-based maintenance program, naturally, statistics from a division-based stratum are most useful. If a county-based maintenance program is desired, county-based stratum statistics are preferred.

- Asking the question of what the data should answer is vital when developing an effective sampling and reporting plan. An investigation of NC blocked-ditch data provided two different perceptions of the service levels on interstate roadways. Defining what measure is most useful and important in managing maintenance, and/or providing a specified level of service to the public, should drive the sampling and reporting processes.

- ANOVA is possible with small sample sizes within each strata. It is recommended that a minimum of n = 10 samples be collected for each population of interest to minimize the impact of any single data point on mean or standard deviation. The F-ratio and p-value calculations to determine significance can accommodate a range of populations and sample sizes.

5.4 Comparing Results of MQA Data Collectors

From time to time, an agency may want to verify that QA field surveys are being conducted in a correct manner. In other words, the agency may want to "QA the QA." This practice is common in highway construction where an agency may independently verify (Independent Assurance) that the QA functions are properly fulfilled. Such a comparison requires a comparison of means, and statistical tools are readily available for this purpose.

Equations 2.10 and 2.11 are recommended for a comparison of means, or comparison of QA measurements. These equations state that the number of tests used in a comparison of means is a function of: (1) variability, (2) true difference between means, (3) probability that the mean difference is contained within a defined acceptance region (H_o), and (4) probability that the mean difference is outside the defined acceptance region (H_A). As noted in Chapter 2, Equations 2.10 and 2.11 apply to the split-sample comparison within the same sample segment.

An example of $\mu_{D,\ HO}$ would be a difference of 0.0 pieces of hazardous debris between two QA field survey staff, while $\mu_{D,\ HA}$ would be a difference of 0.5 pieces. In other words, two staff members would be considered very different if they counted an average difference of 0.5 pieces of debris on a statewide, district, or county level. Obviously, it is desired to have two staff members derive the same field values for hazardous debris, however, this may not be practical. Thus, the agency may need a "tolerance value", or specified mean difference, that would determine if QA field surveys are statistically different among staff members, in order to identify corrective action.

Equation A.12, the one-sided test, was used to determine mean differences for specified risk values and sample sizes. In a practical sense, how far can two QA field surveyors be apart from each other as measured between the normal null value, zero, and an alternative value where the probability of experiencing a true mean difference becomes very large (say, 80% to 95%). At the time of this study, standard deviation values for the difference between field staff were not available, so values of 0.1, 0.2, and 0.3 for hazardous debris will be assumed for example purposes. Table 5.8 relates the number of comparison tests

with several variables, including (1) a specified mean difference, (2) estimated values for the population standard deviation for pieces of debris per 0.1-mile segment, (3) "two-sided" $Z_{\alpha/2} = 1.96$ (95 percent probability), and (4) $Z_{\beta} = 0.842$ (80 percent power) were used in the calculation.

Table 5.8 Allowable Difference between QA Hazardous Debris Counts

Standard Deviation, count (1)	Allowable Difference for Sample Segments, count			
	1 (2)	2 (3)	3 (4)	4 (5)
0.10	0.3	0.2	0.2	0.1
0.20	0.6	0.4	0.3	0.3
0.30	0.8	0.6	0.5	0.4

If a comparison is desired between split-sample QA measurements for hazardous debris (standard deviation = 0.1), there would be an 80 percent chance of detecting a true difference of 0.3 pieces for a single segment. If the difference is less than 0.3 pieces, the probability of detecting a true difference would be much less (less than 80 percent). Figure 5.2 illustrates the power concept. The ability to discriminate true mean differences is lost as the tests move closer together. In other words, a strong statement can only be made for true differences when the means are further apart. Another option is to compare debris counts at four 0.1-mile segments. Then, if there were a mean difference of 0.1 pieces, the probability of detecting a true mean difference would be 80 percent, much higher than the single-segment probability. Clearly, as the number of tests was increased for the comparison, the true mean difference between tests is decreased.

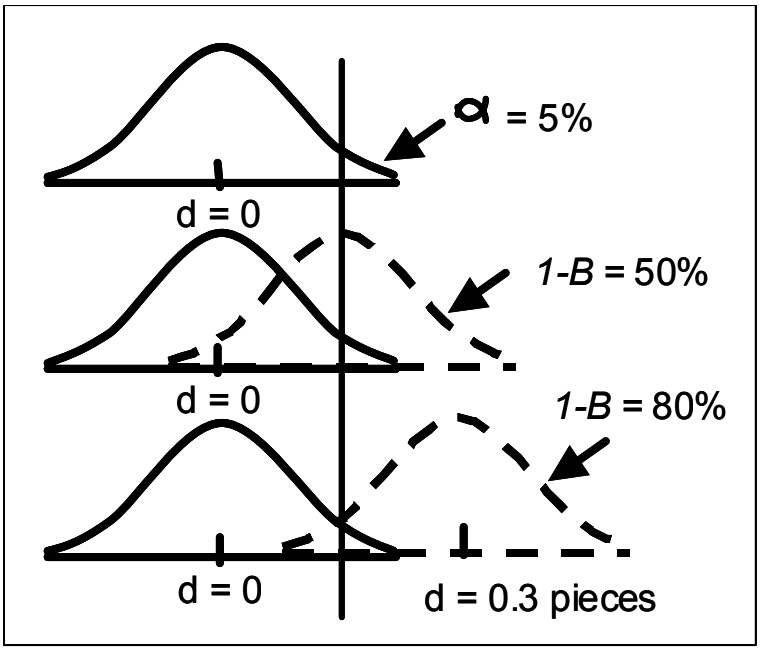


Figure 5.2 Illustration of Power Concept for Detecting True Mean Difference

5.5 Comparing Years (Comparison of Means)

Significant changes in facility maintenance condition from one survey period to the next can shed light, for example, on the role that funding levels impact facility condition. Such information can be translated to impact (e.g. vehicle operating cost) on the using public and may form a basis for future funding requests. To ascertain whether significant differences exist between facility conditions measured for any two particular time periods, a statistical approach dealing with comparison of means test can be applied. The test focuses on the sampling distribution of the difference between sample means. According to Greenshields and Weida (1978), the difference in the population means can be tested by Equation 5.1. This equation is similar to the statistical procedure described in Chapter 4, the section titled, *Hypothesis Test for Difference between Means of Two Normal Populations*. Equations 5.1 and 4.9 are similar, with only a rearrangement of terms.

$$\left| \bar{X}_1 - \bar{X}_2 \right| < t_{\alpha,\nu} \sqrt{\left(\frac{S_1^{\,2}}{n_1} + \frac{S_2^{\,2}}{n_2} \right)} \qquad \text{.....(5.1)}$$

Where:

t = the value of t for the significance level α and the degrees of freedom v;
$v = (n_1 + n_2 - 2)$; and
n_i and S_i^2 = represent the respective size and variance for sample i.

If the inequality in Equation 5.1 holds, then the means represented by the two samples being compared may be considered equal with a level of confidence (1- α), otherwise the hypothesis that the means are equal is rejected. The squared-root term in Equation 5.1 represents the standard error of the difference between the two means, or the pooled standard deviation as described earlier in Chapter 4.

The test is demonstrated using the 2002 and 2004 Interstate LOS data for selected elements in the North Carolina MQA program. The basic summary statistics of the Interstate element LOS are shown in Table 5.9, while the summary analysis results are shown in Table 5.10. Table 5.10 suggests that there is no significant difference between the 2002 Interstate system-wide LOS and that for 2004, while the individual elements, with the exception of "Roadside", had significant difference in LOS between the two time periods. This observation highlights the need for data stratification, in this case by roadway element, so that each can be understood from one year to the next. The overall system LOS had no significant change (from a considerable amount of variability with respect to the mean difference), while some elements had significant increases in element LOS. This approach allows a deeper understanding of movement or shifts in elements that comprise the overall statewide LOS, and where appropriate maintenance resources can be allocated to achieve a desired LOS within elements, and the overall composite LOS.

Table 5.9 LOS Summary Statistics for Interstate Elements in North Carolina

Interstate Element (1)	2002				2004			
	Mean (2)	Standard Deviation (3)	Count (4)	Range (5)	Mean (6)	Standard Deviation (7)	Count (8)	Range (9)
Shoulder ditch	79.9	17.2	250	0-100	87.1	15.7	265	38.7-100
Drainage	77.7	24.8	250	0-100	69.5	28.1	265	0-100
Roadside	74.4	14.9	250	25.4-100	75.7	13.9	265	32.2-100
Traffic	75.8	23.0	250	0-100	82.2	20.0	265	20-100
Entire System	78.4	11.3	250	38.5-97.7	79.8	9.3	265	45.2-93.7

Table 5.10 Test of Means for North Carolina LOS Data

Interstate Element (1)	$\lvert Mean_{2002} - Mean_{2004} \rvert$ (2)	Degrees of freedom, v (3)	T-value for 95% confidence and v = 513 (4)	$\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$ (5)	Column (4) x Column (5) (6)	Are the means significantly different based on Columns (1) and (6)? (7)
Shoulder ditch	7.2	250 + 265-2 = 513	1.97	1.45	2.86	Yes
Drainage	8.2	513	1.97	2.33	4.60	Yes
Roadside	1.3	513	1.97	1.27	2.51	NO
Traffic	6.4	513	1.97	1.90	3.75	Yes
Entire System	1.4	513	1.97	0.92	1.80	NO

5.6 Looking for Trouble Spots (Detecting Outliers in Data)

A potential problem in any data set is the presence of data entry errors, outliers or any other anomaly. Each case must be treated on an individual basis, with an appropriate course of action dependent upon the severity or error of the individual data point. Identification of outliers is a useful tool for checking the validity of data. Among the several methods used for outlier detection are Z-scores and Box-and-Whisker-plots. A brief presentation of each follows.

5.6.1 Z-Score

The Z-score for any data value can be represented as Equation 5.2

$$Z_i = \frac{X_i - \bar{X}}{S} \quad \text{.....(5.2)}$$

Where:

Z_i = Z score;
X_i = data value I;
$\overline{X}$ = the sample mean; and
S = the sample standard deviation.

Data values are classified as outliers when the Z-score is greater than 3 or less than –3 (Anderson et al., 1990). This implies that the data point is an outlier if it exceeds three standard deviations from the mean, or approximately beyond the 99% probability region. Data values that fall in these ranges can be reviewed for accuracy and a decision can be made whether they belong in the data set or not.

As an example, in Figure 5.1 several LOS values in the North Carolina 2002 data were less than 50, causing a slight skew in the histogram to the left, towards the lower LOS values. The Z score for LOS = 50 is computed as follows:

$$Z_i = \frac{50 - \bar{75}.4}{10.8} = -2.35$$

Since the Z score (-2.35) is less than –3, it can be concluded that the LOS = 50 value is not an outlier. A computation using Equation 5.2, and solving for X_i, suggests that LOS $\leq$ 43 are considered outliers. Although this conclusion was reached formally, it may warrant an investigation to determine if it is valid. During analysis of the MQA data from North Carolina and Wisconsin, data entry errors were noted, such as negative LOS values for a particular element and alpha characters inadvertently entered for a numeric characters.

5.6.2 Box-and-Whisker Plots

This procedure allows mild and extreme outliers in a data set to be identified. It involves the computation of the first quartile (Q1) or the 25th percentile, the median, the third quartile (Q3) or the 75th percentile, and the interquartile range (IQR = Q3-Q1). Data falling between 1.5 IQR and 3.0 IQR from the first quartile or third quartile are classified as mild outliers, while those falling beyond 3.0IQR are considered extreme outliers (Anderson et al., 1990). With the availability of statistical software such as STATGRAPH™, such analyses can be easily carried to assess the validity of the data. Figures 5.3 through 5.6 show box-and-whisker plots for the 2002 LOS values associated with selected elements in the North Carolina Interstate system. The system-wide (Figure 5.3) and roadside (Figure 5.5) LOS data are characterized by a number of mild outliers, compared to the shoulder/ditch data (Figure 5.4), which has one extreme outlier. The drainage data show no outliers (Figure 5.6).

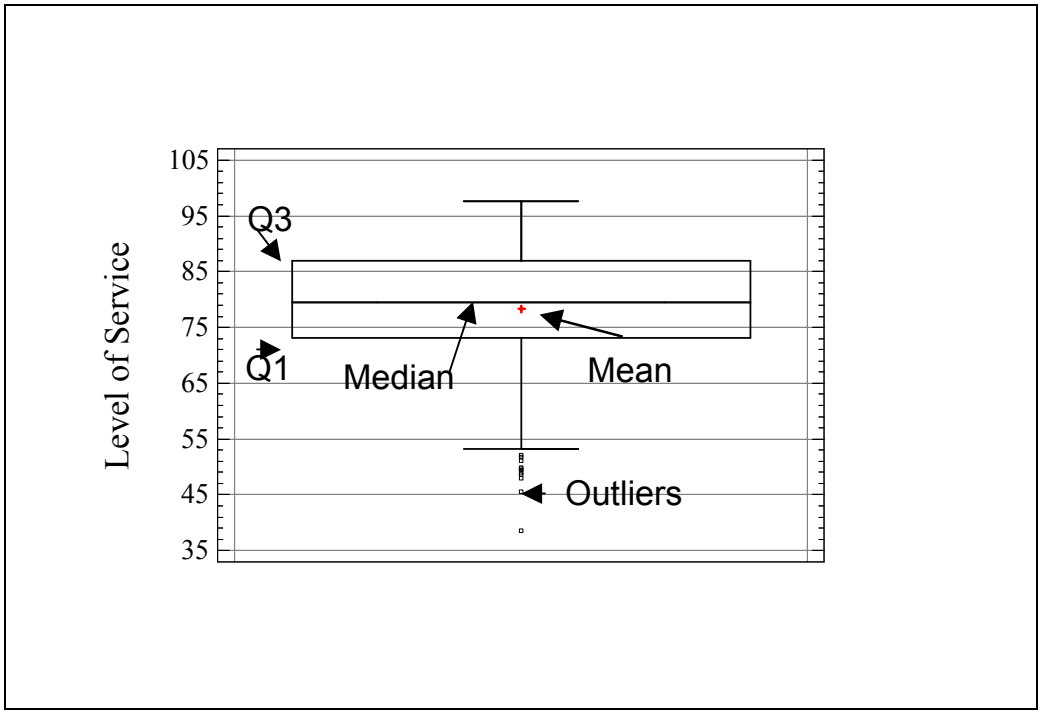


Figure 5.3 Box-and-Whisker Plot for Interstate LOS (NC 2002)

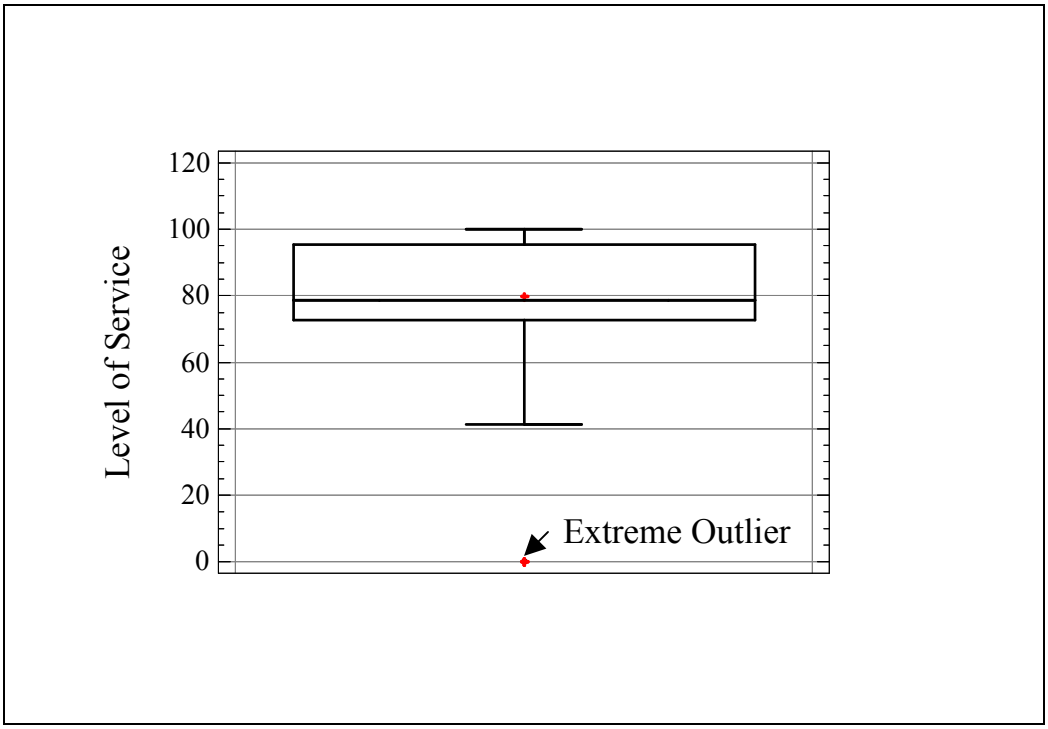


Figure 5.4 Box-and-Whisker Plot for Shoulders/Ditches LOS (NC 2002)

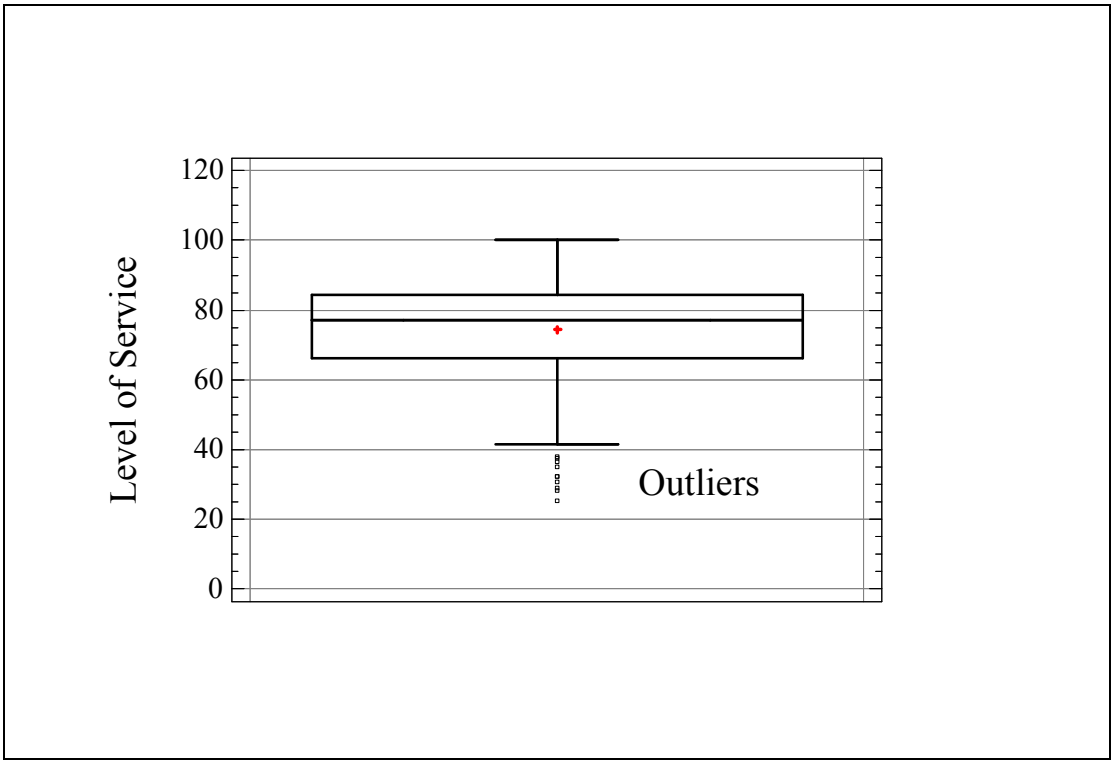


Figure 5.5 Box-and-Whisker Plot for Roadside LOS (NC 2002)

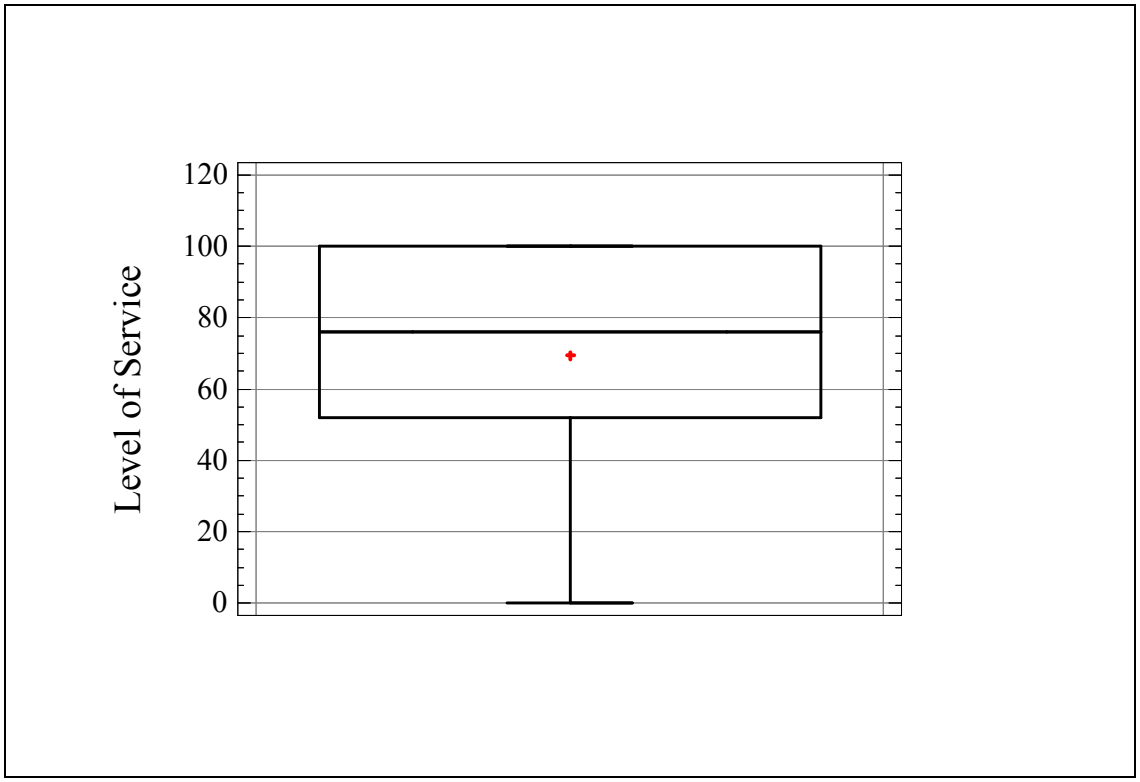


Figure 5.6 Box-and-Whisker Plot for Drainage LOS (NC 2002)

5.7 Reporting Data

Effective presentations are prepared to address specific audience including technical and non-technical people. Hobbs (1987), however, noted that engineers have a tendency to rely on technical vocabulary, complex analysis models, and multiple tables of data. Consequently, the presentation of funding needs and suggested changes in policies to non-technical engineers have suffered from several problems (Braun 1987). Mouaket and Lucy (1987) reported that well-designed graphics are much more effective communication tools than words or tables of numbers, particularly, when the information is being presented to a non-technical audience such as the legislature, transportation commissioners, upper level management, and the public. It is recommended that the level of details presented to those reviewing funding requests should decrease as the level of decision makers increase.

A wide range of formats can be used to present the results of MQA statistical analysis including tables, line graphs, bar charts, pie charts, box-and-whisker plots, and summary reports. The format will depend on factors such as the amount and type of information to be presented, as well as the target audience. Table 5.11 provides guidance on possible presentation formats for combinations of information type and target audience. Since each audience and data type has its own unique combination of circumstances, this table recommends a basis from which to start.

Table 5.11 Recommended Formats for MQA Statistical Data Presentation

Desired Information (1)	Target Audience (2)	Presentation Formats (3)
1. Create awareness regarding average facility network condition over a selected time period.	Public, Administrative, and legislative personnel	Line graph, bar chart.
2. Basic condition summary statistics for network or classification within the network	Administrative, technical	Table
3. Distribution of facility conditions (e.g. good, fair, poor)	Administrative, legislative, technical	Bar chart, pie chart
4. Outliers in data	Technical	Box-and-Whisker plots
5. Funding allocation among maintenance categories	Legislative, administrative, technical	Bar chart, pie chart
6. Impact of funding scenarios (e.g. full, partial, and deferred funding) on facility condition	Administrative, legislative, technical	Line graph, bar chart
7. Indicate whether differences exist in facility condition for different classifications e.g. climatic regions, functional class, districts etc)	Administrative, technical	Table

CHAPTER 6 STATISTICAL RANKING PROCEDURE

6.1 Introduction

The previous two chapters provided statistical procedures and applications for understanding MQA data. Traditional statistical equations from the literature review, as well as standard equations documented in the literature review and most statistical textbooks, were applied to MQA data.

It this chapter, a new procedure is described that ranks roadway sections based on an individual MQA measure (hazardous debris, blocked ditches, etc.) or the individual element LOS, or overall composite LOS. Since a highway maintenance management system has many facets, with the primary purpose being to discover and apply the safest and most cost-effective repair strategies, a ranking procedure can help prioritize which components of the system are in greatest need of treatment. In the case of the pavement, for example, the challenge is to first survey the system to determine the range of condition, usually in terms of suitable measures such as level of service or individual measure, and then create an ordered ranking of individual segments of the system by need for repair based on some established set of rules.

This new procedure was developed during this project to take an alternative approach to the methods described earlier, such as precision and confidence interval construction. This prototype procedure makes use of well-known statistical measures and procedures in an innovative way, and has the added advantage of providing an effective graphical display of any individual MQA element, or the traditional level of service (LOS) measure. Before this ranking procedure is presented, several other established ranking methods are described to put the proposed ranking procedure in context.

6.2 Ranking Methods

Various approaches for ranking pavement sections have been presented by AASHTO (2001). The following approaches will be discussed in the following paragraphs, including, use of damage measures, performance function, usage weighted performance function, composite criteria, first cost, and least life-cycle cost.

The damage measure approach ranks sections based on the amount of some specific distress type (e.g. average rut depth) occurring on candidate sections. This approach addresses the worst conditions first, and as such has the potential to minimize user costs and user complaints. It however, fails to account for the effect of facility usage. In addition, Peshkin et al. (1999) reported that it is sometimes possible to fix five sections in better condition for the price of fixing one in worst condition. Thus, if the objective is to preserve the investment made in the facility network, then the worst first approach is not the best ranking method to use.

The performance function ranking procedure is similar to the damage measure approach except that distress types are combined to yield a single index such as pavement condition index (PCI) or present serviceability index (PSI) to represent the overall condition of each candidate section. It has similar weaknesses as the damage measure approach.

The usage weighted performance function ranking approach is a modified version of the performance function, in that the performance function is divided by a usage variable such as vehicle miles traveled. The main weakness is that funding allocation tends to favor those candidate sections with high traffic even in situations where the cause of deterioration may not be traffic related.

Ranking by composite criteria combines condition measures with other data such as safety and traffic to develop a priority score. Each variable to be included in the priority score computation is placed on a common scale and assigned a weight based on its perceived importance. The score is calculated as the sum of the product of the individual weights and corresponding measured indices. If the highest score represents the best condition, then the lowest numbers represent the candidate sections with the highest priority. The main weakness is that the weights are difficult to interpret and if not properly developed, it could be subject to similar weaknesses associated with the worst first approach.

The first cost approach ranks sections based on least cost for their repair. Unit costs are used to normalize for section size. This avoids small sections being favored over large sections if total cost is used.

The least life-cycle cost procedure assesses the economics of all anticipated costs over the life of the facility, and is used to identify sections that will provide the desired performance at the minimum cost over a given analysis period. The challenges in this approach include being able to identify and quantify all present and future costs, selecting an appropriate discount rate, and the length of an analysis period. Once these variables are known, a discounted cash-flow method such as present value or annualized cost can be used to compare the sections.

6.3 Approach to New Ranking Procedure

The new ranking procedure applies the concepts of percent defective (PD) or percent within limits (PWL). PD and PWL which are simply different representations of the same measure, are currently widely used in the quality assurance field because they address both mean level and variability in a statistically efficient manner. Since both mean level and variability are similarly important in evaluating MQA elements, and since the emphasis in maintenance management is focused on identifying unsatisfactorily performing areas or sections, it was decided to use PD for the initial development. Another distinct advantage of this statistical measure is that it is well suited for both attributes (counted) and variables (continuous) data, and applies equally well for

characteristics with either single or double (both lower and upper) limits, all of which are encountered in maintenance management applications.

Once the appropriate measure of condition (LOS or individual element) and a suitable statistical measure (PD) have been selected, the next step is to define exactly what is or is not acceptable in terms of the statistical measure. For example, in highway quality assurance applications, $PD \leq 10$ is often chosen as the acceptable quality level (AQL), meaning that construction that exhibits a larger PD value would be judged either rejectable or subject to some degree of pay reduction. For purposes of this prototype development, it will be assumed that a similar requirement would be appropriate for LOS when evaluated in terms of PD.

Once these fundamental decisions have been made, a formal procedure must be established by which the evaluation process is to be applied. Conventionally, this takes the form of a hypothesis test by which some null condition is assumed (i.e., the null hypothesis), data are obtained, and an analysis is performed. Based on the results of that analysis, the null hypothesis is either accepted or rejected. To put this in the context of maintenance management, the null hypothesis might be that some specific section of highway meets the LOS requirement. Then after data are collected and an analysis is performed, if the data are sufficiently extreme that there is a very low probability (significance level, α) that they could have come from the population assumed by the null hypothesis, then it would be rejected and it would be concluded that this particular section does not meet the desired level of LOS, or the threshold for any individual MQA measure (centerline striping, delineators, etc.).

6.4 Confidence Interval Procedure as Hypothesis Test

Another standard statistical procedure, the calculation of confidence limits about some point estimate, provides an alternative way to perform exactly this same hypothesis test. In this case, the same sample estimate is obtained and limits are calculated within which the true population value is expected to lie with some specified level of confidence ($1 - \alpha$). If those limits do not include the desired level of the particular measure, then that particular section of pavement would be judged to be noncompliant.

Thus, the procedure employing confidence limits about a point estimate is functionally equivalent to a hypothesis test of that same estimate. Whereas both of these approaches are commonly applied to point estimates such as the mean, they can apply to any point estimate, including PD, and the necessary statistical methodology is well developed for both variables (Weingarten 1982) and attributes (CRC 1968) methods of measurement.

Figure 6.1 is repeated from Chapter 4, and the top diagram in this figure illustrates the conventional procedure used to compute lower and upper confidence limits for the population mean based on sample statistics. An application of the confidence interval procedure was provided in Chapters 4 and 5. (See the sections *Confidence Interval for the Mean of a Normal Population* in Chapter 4, and *Sample Size and Confidence Limits* in Chapter 5). The middle diagram in this figure shows a more intuitive way to

conceptualize exactly this same procedure by considering how low the population mean might be such that there would be an α/2 probability that a sample mean as high as the one observed would be obtained, with the same reasoning applied to the computation of the upper confidence limit.

.

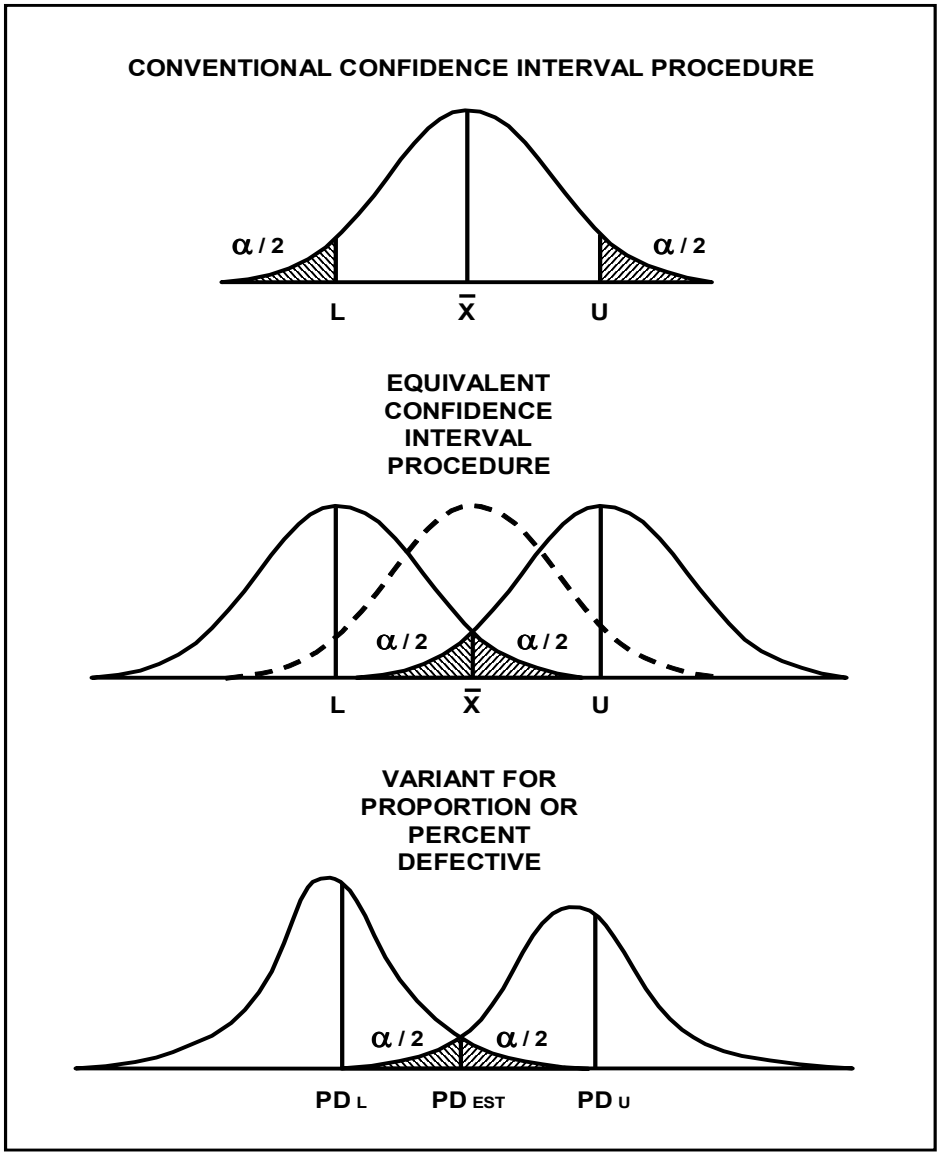


Figure 6.1 Confidence Interval Concepts

For confidence limits on a proportion (or PD) using the binomial sampling distribution associated with attributes data, or confidence limits on PD obtained by variables sampling, only the procedure shown in the bottom diagram in Figure 6.1 applies since the sampling distributions are no longer symmetrical.

6.5 Graphical Display of Ranking Procedure

The application of the ranking procedure based on these methods consists of computing confidence limits for the statewide system or any strata (functional class, division, or county), and then ordering them by their lower confidence limits, as illustrated in Figure 6.2. In this illustration, the length of each individual bar in this figure represents the degree of precision (width of confidence interval) with which the level of service of that roadway section has been estimated. For those cases in which the sample sizes are larger, there will be a tendency for the bars to be shorter, indicating a greater degree of precision. The bars have been ordered from top to bottom by ascending values of the lower confidence limit on LOS, expressed in terms of PD. The darkly shaded bars at the bottom of the figure indicate six sections of pavement that have failed the null hypothesis that they are within the required range of LOS. Thus, the figure not only clearly shows the results of a series of standard hypothesis tests, it also presents a convenient snapshot of all sections in a logical order of LOS based on those tests.

6.6 Data Requirements

The proposed ranking procedure can be applied to virtually any individual measure (hazardous debris, blocked ditches, mowing width, etc.) or the LOS measure under the following conditions:

1. A suitably short, discrete length of roadway (e.g., 0.1-mile, 0.2-mile) that can appropriately be rated as either satisfactory or unsatisfactory (attributes procedure), or given a single quantitative MQA measure or LOS value (variables procedure).

2. A meaningful acceptable level can be defined for these discrete segments; a level that describes exactly what constitutes a satisfactory or unsatisfactory rating, or describes precisely how a numerical measure of LOS or individual element quality is to be obtained.

3. An acceptable percentage (PD) can be defined for nonconforming segments within any given sampling area (e.g., PD $\leq$ 10, meaning no more than 10 percent of the discrete pavement segments are allowed to be nonconforming).

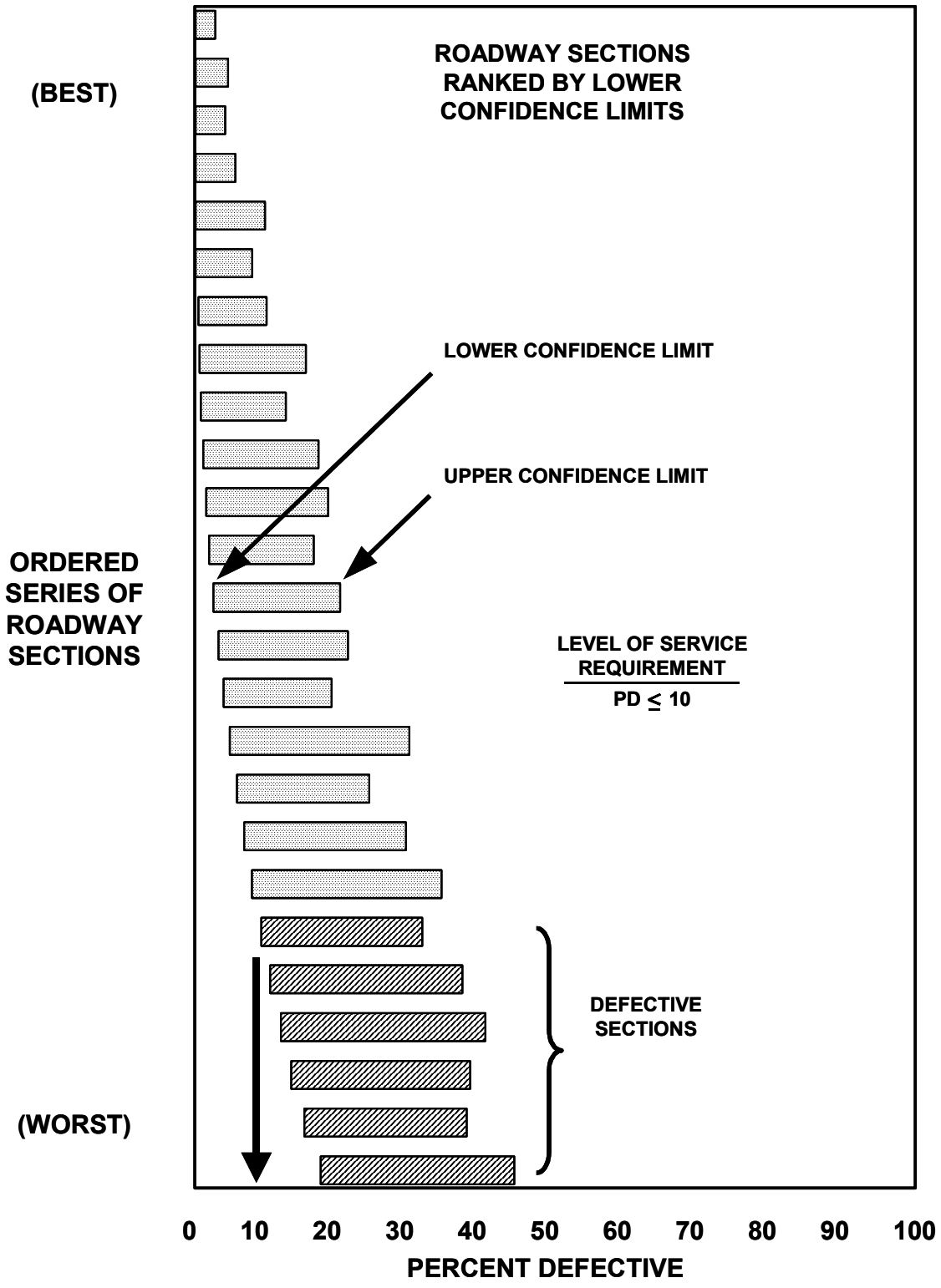


Figure 6.2 Typical Display of Ranking Procedure

6.7 Developing the Attributes Procedure

The attributes procedure is the more general of the two methods because, unlike the variables procedure that requires at least an approximately normally distributed population, it can be applied to data of any distributional form. The "attribute" being referred to in this particular application is the "yes/no" or "pass/fail" result regarding the LOS requirement defined by the highway agency for each discrete 0.1-mile, 0.2-mile, or other short section of roadway.

The data obtained in this manner is "count" data which, to put it in a form suitable for analysis by the binomial distribution, might be expressed as k = 6 failing segments out of a total of n = 30 randomly selected segments of pavement, for example. The percent defective estimate for this particular data grouping or strata (functional class, division, or county) is then given by Equation 6.1.

$$PD = 100\left(\frac{k}{n}\right) = 100\left(\frac{6}{30}\right) = 20 \tag{6.1}$$

Assuming a typical critical limit of PD ≤ 10 has been selected by the agency, the next question to be answered is whether or not the sample estimate of PD = 20 is sufficiently larger than PD = 10 to be confident that it should be rejected, and this is where the statistical hypothesis test comes into play. Recalling the lower diagram in Figure 6.1, it is the lower confidence limit that is primarily of interest which, by definition, is the population mean (in terms of PD) that would produce precisely an α/2 probability of observing k = 6 rejects out of a total of n = 30 trials. In terms of the cumulative binomial distribution, this can be expressed as

$$\left(\sum_{k}^{n}\right)\left(\frac{n!}{x!(n-x)!}\right)p^{x}(1-p)^{(n-x)} = \alpha/2 \tag{6.2}$$

where

k	=	number of failed sections
n	=	sample size
x	=	summation variable, ranges from k to n
p	=	population proportion defective (p = PD / 100), and
α	=	statistical significance level.

The objective is to solve for p. In the absence of a closed-form solution, a computer program was written to choose a logical starting value and then use a searching algorithm to obtain p by trial and error. The algorithm is sufficiently fast that it requires very little computational time to obtain p to the third decimal place. The desired lower confidence limit is then obtained as $PD_L = 100\ p$. The upper confidence limit can be obtained in a similar manner.

To complete this example, assume that a confidence interval of $1 - \alpha = 0.95$ were desired. In this case, $\alpha/2$ in Equation 6.2 would be 0.025, and the value of p would be found to be 0.077. From this, $PD_L = 100\ p = 7.7$. What this means is that the population PD could be as low as 7.7 percent and there would still be an 0.025 probability that, with a sample size of n = 30, a value as high as PD = 20 could be observed by random chance. If the critical limit chosen by the agency had been $PD \leq 10$ in this case, $7.7 < 10$ so the null hypothesis would not be rejected and this particular section of highway would be considered to be in compliance.

An alternate way to do this example would be to use an appropriate handbook of statistical tables (CRC 1968). In this case, a table for a confidence interval of $1 - \alpha = 0.95$ would be chosen, and then entered with the values k = 6 and k = 30 to obtain lower and upper confidence limits of $PD_L = 7.7$ and $PD_U = 38.6$. This is a simple matter when the values for k and n happen to correspond to those in the headings of the table, but when that is not the case, it could be necessary to interpolate in two directions. Therefore, it usually is more practical to use computer software for this step.

6.8 Developing the Variables Procedure (Single-Limit Case)

The variables procedure is somewhat less applicable than the attributes procedure because it requires that the population being sampled be at least approximately normally distributed. However, it is worth using when it is appropriate because it provides more precise estimates of PD and narrower confidence intervals, as will be demonstrated in a subsequent section.

The data obtained for the variables procedure are continuous (or nearly continuous) measurements from which the mean ($\bar{x}$) and standard deviation (s) are calculated from the sample. For this example, assume the agency has defined a composite measure of LOS that ranges over a scale of 0 - 100, and that the requirement for each discrete 0.1-mile segment of highway is $LOS \geq 85$. To make this example comparable to the preceding attributes example, assume the sample size is once again n = 30 and that the values listed in Table 6.1 have been obtained.

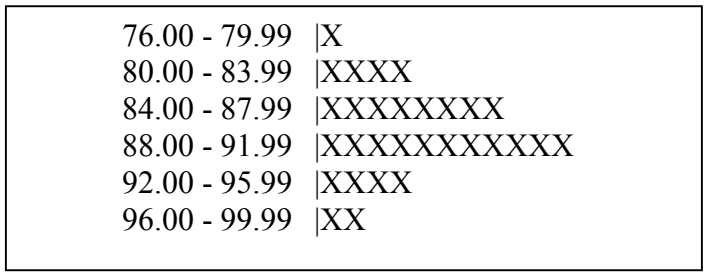
Table 6.1 LOS Data for Variables Example

95.10	95.25	86.43
87.82	89.60	88.08
89.51	86.31	88.47
96.75	82.53	90.30
84.11	88.66	82.04
97.05	87.37	90.03
85.71	95.79	86.01
90.78	79.39	90.78
85.53	90.28	87.15
94.05	91.67	82.27

$\bar{x} = 88.83$, $s = 4.54$

In most cases, the agency would know from past experience if the data were sufficiently normally distributed to allow the variables procedure to be used. However, if this were the first time such an application was being considered, it would be wise to plot a simple histogram, such as that shown in Figure 6.3. More formal tests for normality may also be used (see Chapter 4 section *Normality Tests*). Although a sample size of only n = 30 is seldom enough to make a reliable determination about distributional form, in this case an approximate bell shape is seen so it is judged appropriate to proceed.

Figure 6.3 Histogram of Data from Table 6.1



The next step is to compute the mean and standard deviation which have already been shown at the bottom of Table 6.1. Since the lower limit in LOS units has been set at L = 85 for this example, the "Q-statistic" which will be used to obtain the PD estimate is calculated as follows:

$$Q = \frac{\bar{x} - L}{s} = \frac{88.83 - 85.00}{4.54} = 0.844 \tag{6.3}$$

It should be noted for completeness that, had the application involved an upper limit (U) instead of a lower limit, the formula for Q is given by Equation 6.4. If there were both lower and upper limits, both equations would be used and the total PD estimate would be the sum of the individual estimates.

$$Q = \frac{U - \bar{x}}{s} \tag{6.4}$$

Since there is only a single lower limit in this case, only the calculated value of Q = 0.844 is required. The corresponding estimate of PD can be found either in an individual table created specifically for a sample size of n = 30, or in a consolidated table under the n = 30 column. The consolidated table has the Q values in the body of the table, which usually requires interpolation to obtain the PD value, and therefore tends to be less convenient. A portion of the individual table is reproduced in Table 6.2 to illustrate its use. Ordinarily, a Q value to two decimal places is sufficient for most applications, but three places are used in this example to show that it equates exactly to the previous attributes example.

Table 6.2 Portion of Table Used To Obtain PD Estimates (n = 30)

Q	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	50.00	49.60	49.21	48.81	48.42	48.02	47.63	47.24	46.84	46.45
0.1	46.05	------	------	------	------	------	------	------	------	------
0.2	42.15	------	------	------	------	------	------	------	------	------
-										
-										
-										
0.8	21.27	------	------	------	20.12	19.84	------	------	------	------

It can be seen in Table 6.2 that Q = 0.844 corresponds to a PD value that lies between 19.84 and 20.12. By interpolation, the value is found to be PD = 20.0, the same result obtained in the previous (attributes) example.

However, although the PD estimate is the same, the method for calculating the confidence interval is quite different. As derived in the paper by Weingarten 1982) the formulas for obtaining the lower and upper confidence limits on the PD estimate are given by Equations 6.5 and 6.6.

$$Z_L = -Q + Z_{\alpha/2}\sqrt{\frac{1}{n} + \frac{Q^2}{2n}} \tag{6.5}$$

$$Z_U = -Q - Z_{\alpha/2}\sqrt{\frac{1}{n} + \frac{Q^2}{2n}} \tag{6.6}$$

where

Z_L = standard normal variate associated with the lower proportion defective (p_L)

Z_U = standard normal variate associated with the upper proportion defective (p_U)

$Z_{\alpha/2}$ = standard normal variate associated with the confidence level, i.e., for 0.95 confidence, $\alpha/2 = 0.025$ and $Z_{\alpha/2} = -1.96$

n = sample size

Q = Q statistic computed by Equation 6.3 or 6.4

The lower confidence interval is computed with Equation 6.5 using Q = 0.844, $Z_{\alpha/2}$ = -1.96, and n = 30 as follows:

$$Z_L = -0.844 - 1.96\sqrt{\frac{1}{30} + \frac{0.844^2}{2(30)}} = -1.26 \tag{6.7}$$

Then, using a table of the standard normal distribution, it is found that Z_L = -1.26 corresponds to $p_L = 0.1038$. From this:

$$PD_L = 100p_L = 100(0.1038) = 10.4 \tag{6.8}$$

To compute the upper limit for this example, it is noted that Equation 6.6 is identical to Equation 6.5 except for the minus sign before the second term. In this case, Z_U = -0.43 and $p_U = 0.3336$ so that $PD_U = 100\ (0.3336) = 33.4$.

The necessary tables to convert a Q statistic to a PD estimate are readily available in Appendix C, while PWL estimates are provided in Appendix D. Another handy source for the individual tables is the software manual that was furnished with FHWA Demonstration Project 89 on Quality Management (Weed 1996). However, for actual implementation of this procedure, it is recommended that a computer algorithm specifically designed for this purpose be used.

6.9 Developing the Variables Procedure (Double-Limit Case)

According to the source document (Weingarten 1982) the double-limit procedure is not a formally derived procedure but is an approximation. One of the difficulties in deriving such a procedure is that, for a double-limit requirement, there is an infinite number of ways the total PD could be split between the two tails of the normal population. To confirm that the procedure for the double-limit case is reasonably accurate, the results of several computer simulation tests will be presented in the next section.

The procedure is very similar to the single-limit procedure except for the manner in which the final Q value is obtained. First, Equations 6.3 and 6.4 are used with the individual lower and upper limits (L and U) to obtain two Q values (Q_L and Q_U). Note that it is not appropriate to add the two Q values to obtain the final value. Instead, two PD values are obtained in the usual manner (PD_L and PD_U), and then these two values are added to obtain the total PD estimate. The remaining step is to determine the equivalent Q value associated with the total PD estimate, and that is obtained by working backwards in the same statistical table used previously, this time entering the table with a PD value (and sample size, n) in order to obtain the corresponding Q value. (This is the one case for which the consolidated table is preferable but, like the single-limit case, it will still be more practical to use specialized computer software.) Once the final Q value has been

obtained, it is used with Equations 6.5 and 6.6, exactly as was done for the single-limit procedure.

6.10 Checking the Procedures

To confirm that the foregoing procedures are correct, several tests were conducted. The most straightforward way to check the procedure to estimate PD when sampling by attributes is to demonstrate that it precisely duplicates the values contained in standard statistical tables. As part of this check, a computer algorithm was written to conveniently handle all cases without needing to interpolate within published tables, and Table 6.3 demonstrates that essentially perfect agreement was obtained.

Table 6.3 Demonstration of Confidence Interval Procedure for Attributes Sampling

			0.95 CONFIDENCE INTERVALS IN UNITS OF PD	
SAMP SIZE	DEFECT COUNT	EST PD	FIGURE 6.1 ALGORITHM	STANDARD TABLE [a]
10	1	10	0.3 - 44.5	0.3 - 44.5
20	2	10	1.2 - 31.7	1.2 - 31.7
50	10	20	10.0 - 33.7	10.0 - 33.8
100	40	40	30.3 - 50.3	30.3 - 50.3
200	100	50	42.9 - 57.1	42.9 - 57.1

CRC Handbook Tables (CRC 1968)

As a further check of the attributes method, several computer simulation tests were run to apply the procedure literally thousands of times to data sampled randomly from populations with various known levels of PD. The frequencies with which the computed confidence intervals contained the true population values were generally accurate within the expected statistical margin of error.

For the variables procedure, no tables were found similar to those for the attributes method, and only graphical methods appear to be available. Since graphs are difficult to read accurately, it was decided to test the variables method entirely by computer simulation. The source document (Weingarten 1982) notes that the procedure is derived for the single-limit case and, accordingly, this case was used to generate the random results shown in Table 6.4. These tests demonstrate that the procedure produces consistently accurate confidence intervals, at least for the single-sided case.

Table 6.4 Demonstration of Confidence Interval Procedure for Variables Sampling with Single Limits

TRUE PD	SAMPLE SIZE	NUMBER OF INTERVALS[a] CONTAINING TRUE VALUE	NUMBER OF REPLICATIONS	SUCCESS RATE (EXPECT 0.95)
10	5	952	1000	0.952
10	10	943	1000	0.943
10	50	960	1000	0.960
20	10	946	1000	0.946
20	100	945	1000	0.945

Source: Weingarten / Figure 6.1 Procedure (Weingarten 1982)

Since the variables procedure was derived specifically for the single-limit case, and no information was given concerning its accuracy for the double-limit case, a second set of simulation tests was run to test this application. The results are reported in Table 6.5 where it is seen that the double-limit application, also, appears to be quite accurate, at least for the ranges of PD and sample size tested. Based on this limited series of tests, it seems reasonable to conclude that these procedures will be accurate enough for practical engineering applications.

Table 6.5 Demonstration of Confidence Interval Procedure for Variables Sampling with Double Limits

TRUE PD	SAMPLE SIZE	NUMBER OF INTERVALS CONTAINING TRUE VALUE	NUMBER OF REPLICATIONS	SUCCESS RATE (EXPECT 0.95)
10	5	974	1000	0.974
10	10	947	1000	0.947
10	50	943	1000	0.943
20	10	962	1000	0.962
20	100	953	1000	0.953

Source: Weingarten / Figure 6.1 Procedure (Weingarten 1982)

6.11 Application of Ranking Procedure

There are two distinctly different situations in which the ranking procedure might be applied: 100-percent sampling and fractional sampling. For 100-percent sampling, no statistical estimation process is required, and the resultant PD values would be obtained

by noting the fraction of 0.1-mile (for example) pavement segments that failed the defined LOS requirement. In this case, the PD values plotted in Figure 6.2 would be represented by points rather than bars, and all those failing the PD ≤ 10 requirement would be judged nonconforming.

By far the more likely scenario will be the one in which some form of fractional random sampling has been applied. In this case, the PD values obtained must be regarded as estimates of the condition of the roadway, so it is appropriate to proceed with computing the confidence intervals, represented by the bars in Figure 6.2.

Depending on how the data have been obtained, there are two methods by which the confidence limits are computed. For either "count" data, or continuous data that cannot confidently be assumed to be normally distributed, the attributes procedure represented by Equations 6.1 and 6.2 is used. If the data represent continuous (or nearly continuous) measurements that can be assumed to be at least approximately normally distributed, then it is permissible to use the variables approach represented by Equations 6.3 through 6.8, along with the appropriate tables. Therefore, if the LOS values for the short, discrete segments of pavement described under data requirements are essentially normally distributed, then the variables procedure may be used. The advantage of this method, when appropriate, is that it generally leads to a more precise estimate with narrower confidence limits, as demonstrated in the next section.

To assist an agency in making this decision, a variety of normality tests are readily available in standard statistical texts or computer packages. In practice, however, it may be sufficient just to view the histogram of the data provided the sample size is sufficiently large, as outlined in the *Normality Tests* section in Chapter 4.

6.12 Comparison of Attributes and Variables Methods

As a final example, it will be useful to demonstrate the increase in precision that can be obtained when the data are sufficiently normally distributed to justify the use of the variables procedure. Several examples are presented in Table 6.6 where it can be seen that the width of the interval computed by the variables method is consistently narrower than that for the attributes method, demonstrating the improved efficiency of this procedure. The next-to-last row in this table illustrates a case in which, if the requirement had been PD ≤ 10, the pavement would have been found in conformance by the attributes procedure but clearly out of conformance by the variables procedure.

Table 6.6 Comparison of Precision of Attributes and Variables Methods

		0.95 CONFIDENCE INTERVALS IN PD UNITS			
SAMP	EST	ATTRIBUTES METHOD		VARIABLES METHOD	
SIZE	PD	LIMITS	WIDTH	LIMITS	WIDTH
10	10	0.3 - 44.5	44.3	1.8 - 33.4	31.6
50	10	3.3 - 21.8	18.5	4.9 - 18.3	13.4
10	20	2.5 - 55.6	53.1	5.7 - 44.8	39.1
50	20	10.0 - 33.7	23.7	12.2 - 30.1	17.9
100	20	12.7 - 29.2	16.5	14.2 - 27.0	12.8

6.13 Combining Multiple Measures of LOS

The discussion up to this point has made no mention of precisely what constitutes an LOS value. Most realistic applications will involve multiple measures of roadway condition, such as potholes, cracking, skid resistance, drainage, etc. By far the simplest way to deal with this is for the agency to define a composite LOS value that takes into account the contribution of each individual LOS value. In this case, the procedure for computing confidence intervals for composite LOS values would be identical to the method just described.

For those agencies contemplating the development of a composite measure of LOS, there are certain basic approaches to be considered. An intuitive approach would be to simply average the individual LOS values to obtain a composite LOS value. If certain individual measures were considered more important than others, a more sophisticated approach would be to use a weighted average. It is unlikely that there is a single "right" way to do this, and any rational method an agency chooses to use should be sufficient.

CHAPTER 7 IMPLEMENTATION

7.1 Introduction

This chapter presents implementation issues to consider for the MQA program itself, as well as understanding and applying statistics in the program. First, general challenges associated with the implementation of the MQA program are addressed, then statistically-related issues are enumerated.

7.2 General Implementation Issues

The main barriers for implementing an MQA program have been widely discussed by Stivers et al (NCHRP 422). Lack of commitment by all levels of management was cited as the most critical factor. The authors indicated that this can be addressed by making all managers fully aware of the principles involved in the MQA program, how work priorities will be accomplished under reduced funding levels, and the impact of reduced funding on the level of service that can be provided. Other factors cited included funding limitations, employee attitudes, special interest, privatization and unions.

It was pointed out that funding shortages during the fiscal year might be attributed to a number of factors including:

(a) Emergencies from flooding, wildfires, earthquakes, heavier-than-normal snowfalls, etc. If funding does not increase in these situations, priorities developed under funding shortage scenario should be implemented and must be strictly adhered to.
(b) Unwillingness within the agency to request adequate funding to provide the desired LOS. In this situation, it was recommended that modifications to either the LOS to be provided on each activity or the methods and cost estimates of producing the output should help address upper management concerns.
(c) Failure of the final approval authority to fund agency request. This is unlikely to occur, if the agency adequately explains the principles of its MQA program, the desired LOS resulting from the proposed funding limitation, and the method of implementing priorities. However, should this happen, the priorities scenario developed for this situation must be implemented.

The attitude of employees was reported to be critical to the success of an MQA program. The implementation of a new program may require changes in employee work habits and familiar techniques. There is therefore, the tendency to resist the proposed system because it represents change with the added potential to threaten employee job security. To overcome this problem it was recommended management must create a clear statement of purpose or the picture of what the end result should be. In addition, it must involve labor unions and special interest groups in the MQA process, making them familiar with the issues and providing an opportunity for feedback. This will help reduce the potential for resistance to the proposed changes.

7.3 Statistical Implementation Issues

Understanding and comprehending statistics can be a challenge from both a personal perspective, as well as an organizational setting, such as a maintenance management program. There are numerous textbooks in statistics, ranging from basic statistical methods to more advanced experimental and computational procedures. In either case, a recommended approach to using statistics has been offered by Box et al. (1978):

(a) *Define Objectives*. Clearly define the objectives of what is wanted from the data. Be sure that all parties agree upon the objectives. For example, the stated goal may be a uniform mean and standard deviation for a specific MQA measure (hazardous debris, delineators, etc.) among districts or counties.
(b) *Do Not Forget About Non-statistical Knowledge.* There may be an inclination to be over zealous in the use of some particular statistical tool or methodology. Do not forget about what you know about your subject-matter field. Statistical techniques are most effective when combined with appropriate subject-matter knowledge, in this case, maintenance management and maintenance quality assurance. The statistical methods are an important adjunct to, not a replacement for, the basic knowledge and experience from your area of expertise.
(c) *Learn from Each Other.* Good statistical work seems to result from a genuine interest in practical problems, and sharing and learning from each other. Interplay between the statistics and the practical perception of the problem can not underestimated.

In addition to the overarching themes provided by Box et al. (1978), several practical issues arise:

(a) *Training*. Education and training are key to understanding statistics. This report and electronic guidebook offer detailed explanations of statistical procedures and applications. The agency may want to supplement the provided information with additional training via seminars or on-line short courses. It is recommended that a formal training course be developed for this guidebook. Two references to begin development of the course include the National Highway Initiative (FHWA 2006) or NCHRP Project 20-45 (Washington et al. 2001).
(b) *Software*. There are numerous software packages available to conduct statistical analysis. Packages referenced in the guide are the preference of authors, and it is not the intent to advocate one package over another. Selection should be made using personal preference, ability to be easily understood, cost, and/or graphical output. (In some cases, it may be desirable to develop new software for certain specialized applications.)

CHAPTER 8 CONCLUSIONS AND RECOMMENDATIONS

A guidebook was developed to understand and use statistics in MQA programs. To reach this goal, literature and MQA manuals were reviewed, actual MQA data were collected and analyzed, and statistical procedures and applications were described in detail to address specific questions presented by lead states in MQA. The following are specific conclusions and recommendations reached from these primary subject matters.

8.1 Literature Review

The literature review described and critiqued specific equations for use in MQA programs, and synthesized MQA manuals from 10 states. Based on this review, certain equations were recommended for MQA statistical procedures and applications. The major maintenance categories with quality assurance programs in-place include roadway, roadside and vegetation, drainage, traffic control, and rest areas. It was observed that among lead states, a wide range of threshold definitions exist for features or characteristics monitored for each maintenance category. In addition, for the same maintenance categories, considerable variation exists among the states in the types of features monitored.

Statistics have been applied to MQA programs in four main areas including, sampling for customer surveys, sampling facilities for condition assessment, level of service ratings and analyses, and quality assurance testing for automated data collection and processing. Only the statistics associated with condition assessment were described in this guide. A programs in lead states use a statistical random sampling approach to determine the number of sample units or sample size to inspect in asset condition assessment. The majority of lead states conduct asset condition assessment over 0.1-mile sample units, with the exception of programs in New York and California, which use 1-mile segments. In addition, lead MQA states indicated using confidence levels of 90-95% with precision in the range of 3-6% in the statistical sampling approach.

8.2 Role of Statistics in MQA

Specific issues that can use statistics as a tool in the management and decision-making process were identified for the activity, project, and network levels. Based on the needs of this guidebook, statistical tools and definitions were described for the network level. Decisions that require the use of statistics at the network level, and that were addressed in the statistical procedures and applications, included determining

(a) Sample size for condition evaluation,
(b) Proportions of facility characteristics (whether statewide, countywide, or district-wide) that meet agency target value for a specific performance measure at a given confidence level.

(c) Whether significant differences exist in facility maintenance performance for the different functional classification systems, geographical regions or jurisdictions.
(d) Whether significant changes exist in facility maintenance performance from one year to the next.
(e) How to create some awareness among the public and policy makers regarding facility condition over time.

8.3 Statistical Procedures

Based on the network level determinations requested by the project panel, as well as issues enumerated in the literature, fundamental statistical procedures were described in detail. Fundamental procedures included normality tests, confidence intervals, hypothesis tests, sample size determination, and random sampling.

For normality, it was recommended that $N > 50$ is desirable to perform a visual inspection or formal determination. Based on past experience of the authors, $N > 50$ is desirable when assessing normality with histograms since it is robust to the number of histogram bars, and the individual bar width. More formal normality tests traditionally require a minimum sample size of $N > 30$. Methods for constructing confidence intervals were recommended for proportions, means, and percent defective. Hypothesis test recommendations were described for the mean, differences among independent or dependent populations, and proportions. Recommendations for determining sample size using inputs for precision, variability, and confidence level were provided. Guidelines were provided for random sampling, in particular, length of sample segments must be equal, stratification is valid if the strata are subdivided into equal segment lengths, and when it is desired to sample a specific feature, the random sampling approach is limited to the feature and cannot include all roadway elements within the chosen segment.

Specific examples were illustrated for:
(a) Normality of LOS data;
(b) Confidence interval for number of obstructed drains;
(c) Proportion of vegetation obstruction;
(d) Percentage of adequately mowed area;
(e) Spacing of erosion fence;
(f) Width of mowing;
(g) Comparing mowing between two districts;
(h) Proportion of complying items in a rest area;
(i) Operability of vending machines between two highway corridors; and
(j) Samples needed to estimate lineal feet of shoulder drop-off or build-up

8.4 Statistical Applications

In addition to the recommended statistical procedures, specific statistical applications were provided based on questions expressed by the project panel. These specific

applications were illustrated using actual MQA data from North Carolina and Wisconsin, and included:

(a) Determining number of samples to yield valid information;
(b) Developing confidence in an estimate;
(c) Stratifying data in terms of geographical or highway features;
(d) Comparing results of MQA data collectors;
(e) Are years different or not; and
(f) Looking for trouble signs.

These applications provide maintenance managers and practitioners with the tools and procedures to make objective and reliable decisions using their MQA data.

Recommendations for formatting and reporting data were described, such as tables, line graphs, bar charts, pie charts, box-and-whisker plots, and summary reports. The format will depend on factors such as the amount and type of information to be presented, as well as the target audience. Since each audience and data type has its own unique combination of circumstances, a summary table recommends a basis from which to start.

8.5 New Statistical Ranking Procedure

As work progressed in this guide, it became apparent that approaches from other highway quality assurance fields could be efficiently adapted to maintenance management. The new procedure for ranking roadway sections can be based on an individual MQA measure (hazardous debris, blocked ditches, etc.) or the individual element LOS, or overall composite LOS. The procedure applies the concepts of percent defective (PD) or percent within limits (PWL), and computes confidence limits for the statewide system or any strata (functional class, division, county), and then orders them by their lower confidence limits. If those limits do not include the desired level of the particular measure, then that particular section of pavement would be considered noncompliant. The new method has the benefit of graphically illustrating the length of each individual bar to represent the degree of precision (width of confidence interval) with which the level of service of that roadway section has been estimated.

8.6 Implementation

General challenges associated with implementing an MQA program, along with statistically-related issues were enumerated. With respect to using statistics, it is recommended that the agency clearly d*efine the objectives, do not become so immersed in the statistics that you forget about non-statistical knowledge, and to learn from each other.* Since understanding and comprehending statistics can be a challenge from both a personal and an organizational perspective, training is recommended, along with appropriate software.

REFERENCES

AASHTO Joint Task Force on Pavements (2001). "Pavement Management Guide", AASHTO, Washington, D.C.

Adams, T. (2004). "Maintenance Quality Assurance Peer Exchange", Midwest Regional University Transportation Center, http://www.mrutc.org/research/0404/index.htm, Madison, WI.

Adams, T. (2004). "Pre Workshop Survey on MQA Programs", Proceedings of the Maintenance Quality Assurance Peer Exchange Workshop, Madison, WI., October 2004

Box, George E.P., Hunter, William G., and Hunter, J. Stuart (1978). Statistics for Experimenters. John Wiley and Sons, Inc., New York, NY.

Braun, R.P. (1987). "Justifying Required Funds-A review of success Stories", Vol. 3, Proceedings, Second North American Conference on Managing Pavements, Ontario Ministry of Transportation, Toronto, Canada.

California Department of Transportation (1999), "Evaluator's Reference Module and Field Evaluation Guide", 1999.

CRC (1968). CRC Handbook of Tables for Probability and Statistics. 2nd Edition, William H. Beyer, Editor, The Chemical Rubber Company, Cleveland, OH.

Dixon, Wilfred J., and Massey, Frank J., Jr. (1969). Introduction to Statistical Analysis. Third Edition, McGraw-Hill, Inc., Boston, MA.

Federal Highway Administration (2006). National Highway Initiative Training Programs, <http://www.nhi.fhwa.dot.gov/training.asp>, updated 2006.

Good, Phillip I., and Hardin, James W. (2003), " Common Errors in Statistics (and How to Avoid Them)", John Wiley & Sons, Inc., Hoboken, N.J.

Graff, J. S. (2004). "The Use of Condition Assessment to Improve maintenance Level of Service in Texas", Proceedings of the Maintenance Quality Assurance Peer Exchange Workshop, Madison, WI, October 2004.

Greenshields, B.D. and Weida, F.M. (1978). "Statistics with Applications to Highway Traffic Analyses", ENO Foundation for Transportation, Inc., Westport, CT.

Hobbs, D.G. (1987). " A Decision-Maker's Perspective on Managing Pavements", Proceedings, Second North American Conference on Managing Pavements, Ontario Ministry of Transportation, Toronto, Canada.

Kardian, R.D., and Woodward, W.W. Jr. (1990). "Virginia Department of Transportation's Maintenance Quality Evaluation Program", Transportation Research Record 1276, TRB, National Research Council, Washington, D.C..

Kopac, P.A. (1991). "Q&A: How to Conduct Questionnaire Surveys." *Public Roads*, Volume 55, No. 1, pp.8-15.

Lebwohl, A.S. (2003). "Compass-- Case Study of How Wisconsin Adapted NCHRP Report 422 to Work at Home", Transportation Research Circular Number E-C052, TRB, Washington, D.C.

Levin, Richard I. and Rubin, David S. (1991). "Statistics for Management", 5th Edition, Prentice Hall, Englewood Cliffs, N.J.

McCullouch, Bob and Sinha, Kumares (2003). "Maintenance Quality Assurance Program", Final Report, Joint Transportation Research Program, Project No. C-36-78F, Purdue University, West Lafayette, IN.

Miller, Charles (1989), "Indicators of Quality in Maintenance", NCHRP Synthesis of Highway Practice # 148, TRB, National Research Council, Washington, D.C.

Mouaket, I.M., and Lucy, J.F. (1987). "Communicating for Results in Managing Pavements in Ontario", Proceedings, Second North American Conference on Managing Pavements, Ontario Ministry of Transportation, Toronto, Canada.

Natrella, Mary G. (1966). Experimental Statistics, Handbook 91. U. S. Department of Commerce, National Bureau of Standards, Washington, D.C.

New York State Department of Transportation (2004). "Roadside and Traffic Quality Assurance Condition Assessment Score Sheet", NYSDOT Transportation Maintenance Division, Albany, NY.

New York State Department of Transportation (2001). "Paving Project Quality Assurance Assessment Rating Guidelines", NYSDOT Transportation Maintenance Division, Albany, NY.

North Carolina Department of Transportation (1998). "Maintenance Condition Survey Manual", Operations Division of Highways.

Peshkin D.G., Smith K.D., Zimmerman K.A, and Geoffroy D.N. (1999). Pavement Preventive Maintenance, Reference Manual. FHWA Contract DTFH61-T-70003. Washington, D.C.

Rickmers, Albert D., and Todd, Hollis N. (1967). Statistics: An Introduction. McGraw Hill Book Company, Boston, MA.

Schmitt, R.L., Hanna, A.S., Russell, J.S., and Nordheim, E.V. (2001). "Analysis of Bias in HMA Field Split-Sample Testing," *Journal of the Association of Asphalt Paving Technologists,* Vol. 70, pp. 273-300.

Selezneva, O., Mladenovic, G., Speir, R., Amenta J., and Kennedy, J. (2004). "National Park Service Road Inventory Program-Quality Assurance Sampling Considerations for Automated Collection and Processing of Distress Data", *In Transportation Research Record No. 1889*, TRB, Washington, D.C.

Shahin, M.Y. (1994). "Pavement Management for Airports, Roads, and Parking Lots" Chapman & Hall, New York, NY.

Smith K.L., Stivers, M.L., Hoerner T.E., and Romine A.R. (1997). "Highway Maintenance Quality Assurance", NCHRP Web Document 8 (project 14-12), ERES Consultants, Champaign, IL, 1997

Smith, K.L., Beckemeyer, C.K., Bourdon, R., and Myzie, D. (2003). "Development and Application of the Expanded Version of the Florida Maintenance Rating Program", Transportation Research Circular Number E-C052, TRB, Washington, D.C.

Templeton, C.J., and Lytton, R.L. (1984). "Estimating Pavement Condition and Rehabilitation Costs Using Statistical Sampling Techniques", Research Report 239-5, College Station,Texas Transportation Institute.

Thompson, S.K. (1992). Sampling. John Wiley and Sons, New York, NY.

Transportation Research Board (2005). "Transportation Research Circular E-C074: Glossary of Highway Quality Assurance Terms", Third Update, Transportation Research Board, National Research Council, Washington, D.C.

TRB Committee AHD10 (2005). "Quality Assurance and Condition Assessments for Asset Management in Maintenance—Call for Papers", Annual Transportation Research Board Meeting, Washington, D.C.

Utah Department of Transportation (2005). "Maintenance Management Quality Assurance Plus (MMQA+) Inspection Manual", Utah Department of Transportation, Salt Lake City, UT.

Washington State Department of Transportation (2004). "Maintenance Accountability Process Manual", Maintenance and Operations Division, Maintenance Office, Olympia, WA.

Washington State Department of Transportation (2005). "Maintenance Accountability Process—Field Data Collection Manual volume 1", Maintenance and Operations Division, Maintenance Office, Olympia, WA.

Washington, S., Leonard, J., Manning, D.G., Roberts, C., Williams, B., Bacchus, A.R., Devanhalli, A., Ogle, J., and Melcher, D. (2001). Scientific Approaches to Transportation Research, National Cooperative Highway Research Program Project 20-45, http://trb.org/publications/nchrp/cd-22/start.htm, November 7, 2001.

Weed, R. M. (1996). "Quality Assurance Software for the Personal Computer", Publication No. FHWA-SA-96-026, Federal Highway Administration, Washington, D.C.

Weingarten, H. (1982). "Confidence Intervals for the Percent Nonconforming Based on Variables Data," *Journal of Quality Technology*, American Society for Quality, Volume 14, Number 4, page 207, Milwaukee, WI, October 1982.

APPENDIX A – Statistical Literature Review

A.1 Statistics in Condition Assessment

The use of statistics in pavement engineering is not new. Highway agencies and the construction industry have used statistics in quality assurance for more than two decades. In this case, quality assurance implies separate contractor quality control and agency acceptance. The main aspects of a QA program in the construction industry are testing, inspection, and certification. Quality assurance in maintenance, however, has generally focused on asset/facility conditions, asset needs, and customer satisfaction and expectations of maintenance outcomes. Statistical sampling and analysis have been applied in these areas to facilitate the quality assurance process.

Condition assessment involves determining the inventory of roadway facilities to be monitored and the establishment of the roadway facility population. The sample size to draw from the population of interest uses one of two approaches: (1) a specified percentage of the population, and (2) a statistically determined sample based on a specified confidence level and desired precision. The following sections describe these methods along with related methods for assessing facility condition.

A.1.1 Sample Size Determination

A.1.1.1 Percentage of Population

Traditionally, some agencies target a certain percentage of the population in the sample size determination. The specified percentage appears to be based on the overall use of the data. As reported in the AASHTO Pavement Management Guide (2001), studies conducted by the Texas Department of Transportation found that a sample size of 2 to 5%, depending on the size of the network being sampled, would be adequate to determine average condition of the network of roads. A sample size of 10 to 15% is recommended if the goal is to predict the distribution of condition so that the percent of the network below some selected score can be identified. According to Templeton and Lytton (1984), a sample size of 30 to 35% is needed if the goal is to predict the cost to repair selected segments below some selected value.

Shahin (1994) also presents Table A.1 regarding network level sampling criteria used by some agencies. The number of sample units to inspect depends on the total number of sample units within the section of interest.

Table A.1 Example Network Level Sampling Criteria (Shahin 1994)

No. of Sample Units in Section, N (1)	No. of Units to be Inspected, n (2)
1-5	1
6-10	2
11-15	3
16-40	4
>40	10%

A.1.1.2 Precision

Another sample size determination is based on simple random sampling of the total roadway segment population and desired precision. Random sampling assures each individual segment in the population the same chance of being chosen for field inspection. The entire population of roadway segments is numbered in order to generate random numbers for the segments. The sample size required to achieve the desired precision is computed using Equation A.1 (Stivers et al. 1999).

Where stratification is desired (e.g., by functional class, district, maintenance unit, county), Equation A.1 can still be used on every stratum of interest. The only exception is that the total number of strata is limited to 10 (Stivers et al. 1999).

$$n = \frac{Z^2 * S^2}{d^2} \quad \text{.....} (A.1)$$

Where:

n = required sample size;
S = standard deviation of the ratings from a pilot study;
d = precision (e.g., for precision of ± 5 on a 1-100 scale, use 5.0); and
Z = z-statistic, standard normal variate associated with a particular confidence coefficient (e.g., for 95% confidence, z = 1.96).

[Author's Note – Traditionally, the z-statistic is associated with population parameters such as the population mean (μ) and population standard deviation (σ), and t-statistic is associated with sample statistical parameters, such as the sample mean ($\overline{X}$) and sample standard deviation (S). In the Equation A.1, the population z-statistic has been used to compute the probability level using a sample standard deviation, S.]

A.1.1.3 Probability

In the development of an MQA program for the state of Indiana, McCullouch and Sinha (2003) used a probabilistic method, shown in Equation A.2, for a pilot study to determine the number of road segments to inspect from a 1,624-mile network of roads in the

Laporte District, Indiana. The total population of units to sample from (N represented in Equation A.1), was determined based on 0.1-mile sample units. Thus, the possible number of samples in the network (i.e. the population) was estimated to be 1,624*10 = 16,240. A confidence level of 80% was used resulting in a total sample size of 132, which were selected using the random generator in Excel. Based on the pilot study, a statewide implementation was carried out during the 2002-2003 winter. There was however, a modification to the sample size estimation approach. The sample size was calculated using Equation A.2.

$$n = \frac{P_Y * P_n}{(S.E)^2} \quad \text{.....} (A.2)$$

Where:

n = required sample size for road category in district (Interstate, U.S. Highway, State Road);

S.E. = standard error = (% uncertainty) / (Z-statistic for desired confidence level);

P_Y = probability of yes or pass = 75% (From drainage section, worst case); and

P_n = probability of no or fail = 25%.

Based on a 90% confidence level (Z=1.645), an $n = \frac{0.75 * 0.25}{\left(0.1 / 1.645\right)^2} = 50$ samples were estimated for each road category in the district and selected randomly for inspection.

A.1.1.4 Pavement Condition Index

For project-level inspection for pavement condition index (PCI), Shahin (1994) proposed Equation A.3 for estimating the number of sample units to be surveyed:

$$n = \left[\frac{N * S^2}{e^2 / 4(N-1) + S^2} \right] \quad \text{.....} (A.3)$$

Where:

n = number of sample units to survey;

N = total number of sample units in the pavement section;

e = allowable error in the estimate of the section's PCI (a value of 5 PCI points is recommended); and

S = standard deviation of the PCI between sample units in the section.

Based on data obtained from field surveys, Shahin (1994) recommends S-values of 10 and 15, respectively, for asphalt concrete and Portland cement concrete surfaces when initial inspection is being performed. For subsequent inspections, the actual PCI standard deviations are recommended. Once the sample size is determined, a systematic random

sampling based on a sampling interval is used to select candidate sample units for inspection. The sampling interval is calculated based on Equation A.4:

$$i = N/n \quad \text{.....}(A.4)$$

Where

N = total number of sample units in the section;
n = number of sample units to inspect; and
i = sampling interval rounded off to the smaller whole number.

The first sample unit to be inspected is selected at random from the group of sample units numbered between 1 and *i*. The sequential number of that sample unit is designated *x*. The sample units to be surveyed are therefore, identified as *x, x+i, x+2i, x+3i, +...,* until *x+ji* is equal to or greater than *N* (Shahin 1994).

At the network level, Shahin (1994) proposed the use of Equation A.3 with $e = 5$ and $S = 5$ to determine the number of units to be inspected. Based on this assumption, Table A.2 is suggested. Shahin (1994), however, admits that there is no scientific basis for the assumptions made for the standard deviation, *S*, equal to the allowable error, e, of 5, but it provides a consistent basis for selecting the number of sample units to inspect for different size conditions.

Table A.2 Network Level Sampling Based on Equation A.3 (e = 5, *S* = 5)

No. of Sample Units in Section, N (1)	No. of Units to Be Inspected, n (2)
1	1
2-4	2
5-20	3
>20	4

A.1.1.5 Proportion

Kardian and Woodward (1990) reported Equation A.5 to determine the sample size for inspecting Virginia' roadway systems (Interstate, primary, and secondary):

$$n = \frac{Z^2 * N * p(1-p)}{(A^2 * N) + Z^2 * p(1-p)} \quad \text{.....}(A.5)$$

Where:

n = sample size;
N = population size (centerline Mileage*No. of 0.1 mi. sites/mile);
p = Expected proportion that does not meet desired LOS (30%, 23%, and 25% respectively for Interstate, primary, and secondary systems; this is based on a pilot study of 50 randomly-selected sites);
A = desired precision (precision % / 100%); and

Z = z-statistic for desired confidence level.

A.1.2 Level of Service

The concept of level of service (LOS) is often used to indicate maintenance quality of a highway element (e.g., shoulder or drainage ditch). It is derived from clear and measurable definitions regarding the conditions that can be allowed to exist before specific maintenance features no longer meet agency expectations. According to Miller (1989) LOS serves three purposes, including: (1) provide direction to field personnel to ensure uniformity of maintenance effort throughout the agency, (2) provide a tool for scheduling and budgeting, and (3) define uniform LOS to which the highway user is entitled.

Smith et al. (1997) discussed two statistical quality measurement approaches used in (LOS) field inspections including, method of attributes and method of variables. The method of attributes requires an observer to determine whether a certain standard has been achieved, and then recording the results in a yes-no or pass-fail fashion. It is therefore, more qualitative than quantitative but appears to be the popular method among highway maintenance agencies. The method of variables on the other hand, involves measuring and recording the numerical value according to a specified scale of measurement. It is more complicated than the attributes approach in that it requires both an approximately normal population and a greater number of calculations, but it is worth consideration when appropriate because it provides more precise estimates and narrower confidence intervals.

Statistics have also been applied in LOS ratings and analyses. To determine the LOS of a highway facility, each individual rating for each sample unit is calculated and the statistical mean and variance are calculated for the overall LOS. According to Smith et al., (1997) the mean and standard deviation for LOS can be calculated based on Equations A.6 and A.7, while the confidence interval is based on Equation A.8.

$$\overline{LOS_s} = \frac{\sum LOS_{si}}{n} \quad (A.6)$$

$$s = \sqrt{\frac{\sum \left(LOS_{si} - \overline{LOS_s}\right)^2}{n-1}} \quad (A.7)$$

$$C.I. = \overline{LOS_s} \pm \left(Z * \frac{s}{\sqrt{n}} \right) \quad (A.8)$$

Where:

n = number of sample units;
s = standard deviation of LOS ratings;
Z = z-statistic for desired level of confidence (e.g. Z =1.96 for 95% confidence);
$\overline{LOS_s}$ = mean segment LOS;
LOS_{si} = individual segment LOS values for n sample segments; and
C.I. = confidence interval.

[Author's Note – Comments following Equation A.2 apply to Equations A.6 through A.8.]

A.2 Statistics in Quality Assurance Testing

Selezneva et al. (2004) investigated approaches to quality assurance (QA) sampling, including selection of statistical procedures for quality assurance testing and required sample sizes that will enable conclusions based on QA sample testing to be extrapolated to the whole data set with a certain level of confidence. One statistical testing method investigated was QA testing based on comparison of mean values. This comparison involved the evaluation of the differences between the mean values of field QA sample survey of selected roadway sections associated with specific parks, and data collected from the same sections in the original survey of the park road network by a contractor using an automated road analyzer (ARAN). The QA comparison was based on the statistical *t*-tests. This application is similar in concept to the application requested for MQA.

A.2.1 Test for Mean Differences

A fundamental issue is whether the field staff use the *same* (split-sample) test sites, or *independent* (independent-sample) test sites. If the same test site is used, a "paired" comparison is made, where the effects of sampling different areas of the mat are blocked out. If a data collector chooses to sample a different location, they will be exposed to additional field variability through independent sampling. Agencies are free to choose split sampling or independent sampling in their assurance procedure, however, the additional variability of independent sampling will usually double the sample size comparison for an equal risk and precision level. This feature is further described in the following sections.

A.2.1.1 Split Sample Comparison

For sample size estimation when testing the mean differences of a split sample, Equations A.9 and A.10 were proposed by Selezneva et al. (2004); both equations take into account target values for the probabilities of Type I and Type II errors. Type I error (alpha) in

this case was defined as the probability of concluding, on the basis of the test results, that a significant difference exists between the QA sample and the original survey data while no true difference is present. The consequence of a high probability of a Type I error is rejecting original survey data when indeed the data are acceptable. Type II error (beta) on the other hand, was defined as the probability of failure to identify, based on test results that a significant difference exists between the QA survey and the original survey results.

For a two-sided t-test, sample size can be estimated as:

$$n = \left[\frac{(Z_{\alpha/2=0.025} + Z_{\beta=0.1}) * s}{d} \right]^2 \quad \text{.....}(A.9)$$

Where:

n = sample size;
Z = standard normal variable for specified test significance and test power ($Z_{0.025}$ =1.96 for 95% confidence and $Z_{0.1}$ = 1.28 for 90% power);
s = estimated standard deviation; and
d = difference (tolerance).

For a one-sided t-test, sample size can be estimated as:

$$n = \left[\frac{(Z_{\alpha=0.025} + Z_{\beta=0.1}) * s}{d} \right]^2 \quad \text{.....}(A.10)$$

Where:

Z = standard normal variable for specified test significance and test power ($Z_{0.025}$ =1.64 for 95% confidence and $Z_{0.1}$ = 1.28 for 90% power)

[Author's Note – Comments following Equation A.2 apply to Equations A.9 and A.10.]

Selezneva et al. (2004) applied the above statistical concepts to data collected for the Lake Mead National Recreational Area road network and concluded that, QA testing based on comparison of mean values is appropriate for QA surveys if the purpose of the data is for network level pavement management decision-making. Its use at the project level is discouraged since some individual differences can exceed the acceptable range at the project level due to limitations and evolving nature of the automated data collection and processing technology.

In construction QA split-sample comparisons, Schmitt et al. (2001) recommended Equation A.11 to determine sample sizes for specified risk levels, *d* was defined as the difference between means, or in other words, μ_{HO} (mean at null hypothesis) and μ_{HA} (mean at alternative hypothesis).

$$n = \sigma_D^2 \frac{(Z_{\alpha/2} + Z_{\beta})^2}{d^2} \quad \text{.....}\ldots\ldots\ldots\ldots (A.11)$$

where, n = number of tests.

σ_D^2 = variance between split samples (std. dev.)^2;

$Z_{\alpha/2}$ = standardized statistic of null hypothesis in acceptance region (at 95 percent, $Z_{\alpha/2} = 1.96$);

Z_β = standardized statistic of null hypothesis in rejection region (at 80 percent, $Z_\beta = 0.842$); and

d = difference between means.

Variability plays an important role in determining the true mean difference among QA field measurements taken between staff members. The quantitative measure for the variance in split sampling, σ_D^2, is found through actual data (Schmitt et al. 2001). It is not appropriate to use a segment-specific sample for σ_D^2. An estimate of the population value can be derived through an analysis of data from different strata (roadway functional class, district, county, etc.). If there is uncertainty in the population value, it is recommended that an agency calculate "n" and "d" for a range of possible σ_D^2 values.

A limitation of Equations A.9 and A.10, and Equation A.11, is that the comparison of means between QA field staff must be made within the *same* 0.1-mile segments, not *independent* 0.1-mile segments. The standard deviation in those equations pertains to a "paired" comparison, where the effects of sampling different 0.1-mile roadway segments are blocked from the analysis. In essence, a split-sample of the same 0.1-mile segment is taken between two field staff.

A.2.1.2 Independent Sample Comparison

If an agency chooses to sample a different segment of the roadway, they will be exposed to additional field variability through an independent sampling of the MQA feature. In that case, the pooled variance between data collectors is doubled to acknowledge exposure to the added variability inherent in independent sampling. An agency is free to choose split sampling or independent sampling in their verification procedure, however, the additional variability of independent sampling may double the sample size comparison for an equal risk and precision level.

Schmitt et al. (2001) recommended Equation A.12 to determine the sample size for mean differences of independent-sample data, for specified risk levels and sample sizes, and where *d* is defined as the difference between μ_{HO} (mean at null hypothesis) and μ_{HA} (mean at alternative hypothesis). In a practical sense, how far can two measurements be apart from each other as measured between the normal null value, zero, and an alternative value where the probability of experiencing a true mean difference becomes very large (say, 80% to 95%).

$$n = 2\sigma^2 \frac{(Z_{\alpha/2} + Z_\beta)^2}{d^2} \quad \text{..... (A.12)}$$

where, n = number of tests;

$2\sigma^2$ = twice the pooled variance of Data Collector #1 and Data Collector #2;

$Z_{\alpha/2}$ = standardized statistic of null hypothesis in acceptance region (at 95 percent, $Z_{\alpha/2} = 1.96$);

Z_{β} = standardized statistic of null hypothesis in rejection region (at 80 percent, $Z_{\beta} = 0.842$); and

d = difference between means.

A.2.1.3 Risk and Power

Earlier, Type I and II errors were described. Another way to describe these risks is through the concept of "power." Power is the level of probability that true mean differences between two data sets will be correctly detected at a specified risk level. Power is a function of several parameters, including the size of the departure from Ho ($\mu_{HO} - \mu_{HA}$), sample size, variability, significance level, and the type of test (one-sided or two-sided). Power is increased as the sample size increases, granted all other variables remain the same. Power is computed only from population values and generally does not apply when the standard deviation is unknown (Schmitt et al. 2001). Sampling data are not used in power computations. A common question is what is a good value to have for power. Based on risk tolerance, values of 80%, 90%, or 95% are preferred so there is reasonable assurance that a true mean difference is detected.

A.2.2 Probabilistic Method

Selezneva et al (2004) also investigated quality assurance testing based on individual measurement rating. This approach was based on selecting a representative sample from a data set, rating each individual observation within this sample using established pass-fail criteria for minimum-acceptable quality, and concluding whether the whole data set satisfies criteria for minimum-acceptable quality based on the number of "failed" observations in the sample. The statistical procedure adopted was based on the hypergeometric distribution, which suggests that if a total data set of *N* observations has an actual fraction of defective items equal to *p*, then the probability *g(p)* of accepting the total data set N could be determined using Equation A.13.

$$g(p) = \sum_{x=0}^{r} \frac{\binom{Np}{x}\binom{N(1-p)}{n-x}}{\binom{N}{n}} \quad \text{.....(A.13)}$$

If n is small relative to N, Equation A.13 could be approximated by using a binomial distribution represented as Equation A.14:

$$g(p) \approx \sum_{x=0}^{r} \binom{n}{x} p^x (1-p)^{n-x} \quad \text{.....(A.14)}$$

With Equations A.13 and A.14, a maximum number of unacceptable observations *r* could be determined for a given sample size *n*, or a sample size could be determined for a known targeted maximum number of unacceptable observations in the sample.

A.2.3 Pass-Fail Criteria

Pass-fail criteria have been recommended for QA comparative testing, in a study that assessed park road condition (Selezneva et al. 2004). Regarding sample size selection for testing individual measurements, sample size estimation was considered for two cases: a) large parks with at least 15 miles of surveyed roads and b) small parks with less than 15 miles of surveyed roads (Selezneva et al. 2004). For large parks, with a number of observations *n* in a sample significantly less than the total number of observations for the park (10% or less), the sample size with minimum acceptable number of bad observations *r* could be determined on the basis of a binomial approximation as follows:

(a) The size of the random sample of n_0 observations that must pass the minimum quality acceptance criteria without a single unacceptable observation failing the QA pass-fail criteria could be found from Equation A.15, assuming r = 0, as follows:

$$n_0 = \frac{\ln(1-C)}{\ln(R)} \quad \text{.....(A.15)}$$

(b) The size of the random sample of n observations that must pass the minimum quality acceptance criteria with no more than one unacceptable observation failing the QA pass-fail criteria could be found from Equation A.16, assuming r = 1, as follows:

$$(R)^n + n(1-R)(R)^{n-1} = 1-C \quad \text{.....(A.16)}$$

Where:

R = reliability of the data collection or data-processing procedures expressed as a fraction; and

C= confidence level expressed as a fraction (e.g. for a 95% confidence, C= 0.95).

The reliability value *R* implies that the acceptable percentage of the observations not passing the QA pass-fail criteria because of limitations of the data collection or data-processing methodology is equal to 100*(1-R).

For small parks with less than 15 miles of surveyed roads, Equation A.15 was recommended to determine the sample size when the maximum acceptable number of unacceptable observations *r* is required from the sample n. The input parameters required to compute sample size in this case include the size of the data set, N; reliability level, R=1-p; and confidence level, C= 1-g(p).

APPENDIX B – Major Maintenance Categories

Ten states that were solicited and having provided MQA Program guides include: California, Florida, Indiana, New York, North Carolina, Texas, Utah, Virginia, Washington and Wisconsin. Major maintenance categories included roadway, unpaved shoulder, ditches, roadside features, traffic control devices, rest areas, and environmental concerns. Table B.1 identifies major maintenance categories and corresponding elements monitored in MQA programs. Each feature/characteristic in Column 3 is generally associated with condition standards that must be evaluated in the field. As indicated in Table B.1, categories (Column 2) vary in number. In addition, for the same categories, considerable variation exists between states in the feature types (Column 3) that are monitored.

Table B.1 Maintenance Categories and Characteristics in Agency MQA Programs

State (1)	Maintenance/Operations Category (2)	Monitored Feature/Characteristics (3)	Other Pertinent Characteristics of MQA Programs (4)
California	Travelway	***Flexible***: rideability, cracks, alligator cracking, potholes, wheel rutting, coarse raveling, bleeding, pavement edge, paved shoulders, unpaved shoulders, ramps ***Rigid***: joint separation, slab failure, cracks, spalls, paved shoulders, unpaved shoulders, ramps	--Referred to as the LOS2000 Caltrans Maintenance Program. --LOS2000 evaluation is based on visual inspection of randomly selected sample of one-mile segments. --An LOS2000 Coordinator determines the total random sample of centerline miles from each district for evaluation. Pass/Fail threshold criteria have been established for each of the maintenance attributes listed in Column 3. Pass/Fail results are combined for each sample segment using weighted values for each attribute to derive a Maintenance LOS Rating. LOS ratings from all segments are then combined to generate the overall (district-wide, statewide) level of service. *(Caltrans 1999)*
	Slopes, Drainage and Ditches	Surface drains, cross drains, ditches, slope, ramps	
	Roadside	Roadside vegetation, fences, tree/bush encroachment, roadside litter/debris, graffiti, landscaping, ramps	
	Traffic Guidance	Striping, pavement marking, raised markers, guide markers, signs, guardrail, barriers, attenuators, ramps	
Florida	Roadway	***Flexible Pavement***: Potholes, edge raveling, shoving, depression/bump, paved shoulder ***Rigid Pavement***: Pothole, depression/bump, joints/cracks, paved shoulder	-- Referred to as the Maintenance Rating Program (MRP). --Originally developed and implemented in 1985; --Uses a quantifiable process to determine LOS of maintenance activities performed on various highway classes; --Sampling is based on 0.1-mile sample units within a maintenance section; sample size is determined based on a 95% confidence level and accuracy within 3%. *(Smith et al. 2003)*
	Roadside	Unpaved shoulder, front slope, slope pavement, sidewalk, fence	
	Traffic	Raised pavement Markers, striping, pavement symbols, guardrail, impact attenuators, signs $\leq 30ft^2$, signs$>30ft^2$, object markers and delineators, and lighting.	
	Drainage	Side/cross drain, roadside/median ditch, outfall ditch, inlets, miscellaneous drainage structures, and roadway sweeping,	
	Vegetation and aesthetics	Roadside mowing, slope mowing, landscaping, tree trimming, curb or sidewalk edge, litter removal, and turf condition.	

Table B.1 Cont.

State (1)	Maintenance/Operations Category (2)	Monitored Feature/Characteristics (3)	Other Pertinent Characteristics of MQA Programs (4)
Indiana	Roadway	***Flexible Pavement***: smoothness, depressions/bumps, potholes, shoving, cracking, raveling (including edge)/striping, rutting. ***Rigid Pavement***: smoothness, joints, depressions/bumps, potholes, spalls, cracking.	--Implemented during winter 2002-2003; --Sample size for inspection of each of three functional classes per district (interstate, state road, & U.S. Highway) is based on Equation A.3 at a confidence level of 90%. -- Inspection scores for features are recorded as 1 (pass), 2 (failed), 3 (not applicable); scores are used in conjunction with weighting factors to calculate LOS. *(McCullouch and Sinha 2003)*
	Paved Shoulders	Mainline drop off, potholes, miscellaneous categories, drainage	
	Drainage	Ditches, culverts/pipes, catch basins/drop inlets, curb/gutter, subsurface/underdrains.	
	Traffic control	Striping, signs, attenuators, guardrails, raised pavement markers, barrier wall, pavement markings.	
	Roadside	Unpaved shoulder, mowing/grass, litter/debris, fence, tree/bush control, landscaping (plantings).	
New York	Roadside	Centerline markings, edgeline markings, regulatory signs, warning signs guide signs, cable guide rail, corrugated guide rail, box beam guide rail, edge of pavement drainage, drainage ditches, cross culverts, driveway culverts, drainage structures, closed drainage systems, turf height, turf height under guide rail, clear zone, shoulder drop-off, litter and debris.	--Referred to as Roadside and Traffic Quality Assurance Condition Assessment; --600 random 1-mile sections are selected from 10 participating regions for review every year by field personnel --features are rated on a scale of 4 (best possible condition) to 0 (worst possible condition) --average and standard deviation score values are calculated for features by region and for all regions combined. *(NYSDOT 2004)*
	Rest Areas	***Building (Exterior)***: Roof, walls, doors, wall maps. ***Building (Interior)***: sinks, soap/soap dispenser, paper towels/hand dryers, toilets, partitions, urinals, diaper changing tables, trash, environment/odor, lighting, windows, floors and walls, mirrors, temperature, vending machines, pay phones, and water fountains. ***Grounds***: lawn, picnic tables, trash and litter, trees/shrubs/plantings, fencing and walls, and pet walking area. ***Parking lot***: pavement condition, sidewalk condition, drainage, signing, striping, lighting, commercial vehicle parking, lighting, snow and ice control. *Disabled Persons Accommodation*: parking spaces, sidewalk/curb and building access ramp, exterior door, water fountain, stall, sink and mirrors, and grates.	--Referred to as the Rest Area Quality Assurance Assessment; it assesses the consequences of maintenance efforts rather the technical evaluation of rest area maintenance. --Letter grades A (best) through F (worst) are used to describe condition for characteristics of the rest area. *(NYSDOT 2004)*
	Maintenance paving	***Travel lanes, ramp turning lanes, , shoulders & bike lanes***: Surface texture, longitudinal joints, ride, cross slopes, start/end/bridge and turning lane transitions, shoulder back-up & rounding. ***Intersections & Driveways***: Surface texture, radii, and transitions. ***Drainage***: adjustments to drainage and/or utility structures, curb reveal, and shoulder transition to gutters. ***Striping***: centerline/lane & shoulder stripes, and special markings.	-Referred to as Paving Project Quality Assurance Assessment; --it assesses quality from a user's perception of consequences of the paving project. It is not intended to be a technical evaluation of the project; --letter grades A (best possible quality) through F (worst quality) are used in rating finished project condition. *(NYSDOT 2001)*

Table B.1 Cont.

State (1)	Maintenance/Operations Category (2)	Monitored Feature/Characteristics (3)	Other Pertinent Characteristics of MQA Programs
North Carolina	Roadway	***Flexible pavement***: alligator cracking, block cracking, reflective cracking, rutting, raveling, bleeding, ride quality, and patching. ***Rigid (shoulder features)***: shoulder lane joint, shoulder drop-off, and shoulder condition. ***Rigid pavement features***: patching, surface wear, pumping, ride, longitudinal cracking, transverse cracking, corner break, spalling, joint seal, and faulting.	--Forms part of the Maintenance Assessment and Funding Needs Program. --sampling for condition assessment is based on a 90-95% confidence level and 6% precision. Observations are made on 0.2-mile sample units. --Threshold criteria have been established to describe LOS conditions for features associated with Interstate, Primary, Secondary, and Urban roadway systems. Five LOS values are used A (best condition) through D (poor condition), and F (worst possible condition). Results from LOS analysis are used to determine the investment needed to achieve the current service levels and investment thresholds for achieving other LOS scenarios. *(NCDOT 1998)*
	Unpaved Shoulders & Ditches	Low shoulder, high shoulder, lateral ditches, and lateral ditch erosion.	
	Drainage	Crossline pipe, driveway pipe, curb & gutter, catch basins & drop inlets, and other drainage features	
	Roadside	Mowing, brush & tree control, litter and debris, slopes, and guardrail.	
	Traffic Control Devices	Traffic signs, pavement striping, words and symbols, and pavement markers.	
	Environmental	Turf condition and miscellaneous vegetation management.	
Texas	Pavement	Rutting, cracking, failures, ride quality, edge drop offs, shoulders	--The Maintenance Assessment Program (MAP) evaluates 21 elements in three components. Condition assessment of the elements is based on a rating scale of 5 ("new or like new") to 1 ("failed"). The rating scores are used to determine maintenance levels of service, which in turn are used to effectively communicate maintenance planning and performance expectations, as well as determine the desirable, acceptable, and tolerable levels of funding needed. *(Graff 2004)*
	Traffic	Raised pavement markers, large signs, small signs, striping, attenuators, delineators	
	Roadside	Mowing/herbicide, drainage, litter, sweeping, trees/brush, encroachments, guardrails, mailboxes, public rating,.	
Utah	Non-Hard Surface	Shoulder work: edge rut, vertical drop-off	--Originally referred to as the Maintenance Management Quality Assurance (MMQA) program when implemented in 1997. The MMQA was modified to MMQA^{+} in 2003. MMQA+ has three measurement frequencies regarding assessment of facilities for maintenance. Signs, delineators, sweeping and rest areas are assessed on a monthly basis; road conditions are also assessed 1 hr after precipitation in the form of snow and ice has occurred. All other features listed in Column 3 are measured on semi-annual basis. Generated reports of level of maintenance (LOM) are used at the district level to manage maintenance programs. At the central level reports are used to observe maintenance trends, project budgets, and communication with all key customers. *(Utah DOT 2005)*
	Roadside	Litter and fence	
	Vegetation	Weed, vegetation obstructions, mowing	
	Drainage and Slope Repair	Grade and ditches, inlets, erosion	
	Traffic Services	Pavement striping, pavement markings, signs, delineation, guardrail, sweeping, curb/gutter and islands.	
	Rest Area	***Janitorial services***: rest room smell, walls, floor, trash containers, soap and paper supplies. ***Building & utilities***: mechanical & electrical systems, doors, dispensers, hand dryers, partitions. ***Site***: landscape plants, lawns, sidewalks, picnic areas. ***Operations***: open hours, and presence of an attendant.	

Table B.1 Cont.

State (1)	Maintenance/Operations Category (2)	Monitored Feature/Characteristics (3)	Other Pertinent Characteristics of MQA Programs
Virginia	Traveled Way	***Flexible***: cracking, raveling, rutting, shoving/pushing, bleeding/flushing, ride quality, potholes, distress ***Rigid***: cracking, spalling, faulting, joint material, ride quality, potholes ***Stabilized***: dust control, rutting, potholes, corrugations	--Referred to as the Maintenance Quality Evaluation (MQE) Program; --implemented in 1989 --the MQE program objectives include monitoring the overall quality of highway maintenance, identifying areas of inconsistent performance, and providing a more formal process for assuring that consistent LOS are provided statewide; --sample size determination is based on Equation A.6 using 0.1-mi sample units for a desired confidence level of 95% and precision rate of 4%. *(Kardian and Woodward 1990)*
	Shoulders	***Hard-surfaced***: drainage, distortion, joint separation, failure ***Non-hard surfaced***: drainage, distortion, rutting	
	Drainage	Ditches, culverts, catch basins/drop inlets, curbs and gutters	
	Traffic Control & Safety	Signs, pavement markings, signals, luminaries, barriers, delineators, object markers	
	Roadside	Mowing, litter & debris, tree/bush control, landscaping, sidewalks	
	Structures	Deck distress, expansion joints, parapets, bearing devices	
Washington	Roadway	***Maintenance***: Pavement patching & repair, crack sealing, shoulder, sweeping and cleaning ***Operations***: safety patrol	--Forms an integral part of the Maintenance Accountability Process (MAP); --implemented in 1996 --statistical methods are used to identify approximately 2200 randomly selected data survey sites around the state. Each site length is a 0.10-mi section. All highway features falling within the sample section are evaluated based on specific threshold values. The threshold values correspond to LOS scale A (best possible condition) through F (worst possible condition). The MAP manual indicates that the data collected provides a level of precision of over 95%. An investment model is used to identify the investment needed to achieve the current service level, and estimates the investment threshold for achieving LOS A through F scenarios. *(Washington State DOT 2004, 2005)*
	Drainage	Ditches, culverts, catch basins & inlets, detention/retention basins, slope repair	
	Roadside & Vegetation	Litter pickup, noxious weed control, nuisance vegetation control, vegetation obstructions, landscape	
	Bridge and Urban Tunnel	***Maintenance***: Bridge deck repair, structural bridge repair, bridge cleaning. ***Operations***: movable & floating bridge, ferries, urban tunnel systems	
	Snow and Ice	Snow & Ice control operations	
	Traffic Control	***Maintenance***: pavement striping, raised/depressed pavement marker, pavement marking, regulatory sign, guide sign, guidepost, guardrail. ***Operations***: traffic signal systems, highway lighting systems, intelligent traffic systems, permits	
	Rest Areas	Rest area operations	
	Training and testing	Employee technical & safety training, support and testing	
	3rd-party damages and disaster operations	3rd party damages, disaster operations	
Wisconsin	Shoulders	***Paved Shoulder***: Hazardous debris, drop off/build-up, cross slope, cracking, potholes/raveling ***Unpaved Shoulder***: Hazardous debris, drop off/build-up, cross slope, erosion	--Referred to as the COMPASS system; --implemented statewide in 2002; --uses a random sampling approach to select the number of 0.1-mi sample units for inspection. *(Lebwohl 2003).*
	Drainage	Ditches, culverts, under/edge drains, flumes, curb & gutter, storm sewer.	
	Roadside	Litter, graffiti, mowing, noxious weeds, woody vegetation, fences, landscaping, barriers	
	Traffic control & safety	Centerline markings, edgeline markings, special pavement markings, raised pavement parkers, regulatory/warning signs, other signs, delineators, protective barriers.	

APPENDIX C – Percent Defective Tables

APPENDIX D – Percent Within Limits Tables