1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
FHWA/TX-04/0-1861-2		

4. Title and Subtitle	5. Report Date
QUALITY ASSURANCE AND INTERVAL ESTIMATION PROCEDURES FOR ROAD CONDITION DISTRESS SCORES	June 2004 Resubmitted:August 2004
	6. Performing Organization Code

7. Author(s)	8. Performing Organization Report No.
John P. Wikander, Thomas J. Freeman and Clifford H. Spiegelman	Report 0-1861-2

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
Texas Transportation Institute The Texas A&M University System College Station, Texas 77843-3135	11. Contract or Grant No. Project No. 0-1861

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Texas Department of Transportation Research and Technology Implementation Office P. O. Box 5080 Austin, Texas 78763-5080	Technical: September1998 - August 2000
	14. Sponsoring Agency Code

15. Supplementary Notes

Project performed in cooperation with the Texas Department of Transportation and the Federal Highway Administration.

Project Title: Statistical Analysis of PMIS Data Elements

16. Abstract

This report describes the statistical analysis of the Texas Department of Transportation (TxDOT) Pavement Management Information System (PMIS) distress data and audit procedures used to verify satisfactory performance of statewide distress surveys. Several improvements are provided, including a method of determining the sample size that is much more rigorous than the current fixed percentage method. In addition, a review of the confidence interval for the average condition of all pavements in Texas has been conducted.

17. Key Words	18. Distribution Statement
PMIS, Distress, Audit, Statistical Methods, Confidence Interval	No Restrictions. This document is available to the public through NTIS: National Technical Information Service http://www.ntis.gov Springfield, Virginia 22161

19. Security Classif.(of this report)	20. Security Classif.(of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	62	

QUALITY ASSURANCE AND INTERVAL ESTIMATION PROCEDURES FOR ROAD CONDITION DISTRESS SCORES

by

John P. Wikander
Assistant Transportation Researcher
Texas Transportation Institute

Thomas J. Freeman
Engineering Research Associate
Texas Transportation Institute

and

Clifford H. Spiegelman, Ph.D.
Professor and Research Scientist
Texas Transportation Institute

Report 0-1861-2
Project Number 0-1861
Project Title: Statistical Analysis of PMIS Data Elements

Performed in Cooperation with the
Texas Department of Transportation
and the
Federal Highway Administration

June 2004
Resubmitted: August 2004

TEXAS TRANSPORTATION INSTITUTE
The Texas A&M University System
College Station, Texas 77843-3135

DISCLAIMER

The contents of this report reflect the views of the authors who are responsible for the opinions, findings, and conclusions presented herein. The contents do not necessarily reflect the official views or policies of the Federal Highway Administration (FHWA) and/or the Texas Department of Transportation (TxDOT). This report does not constitute a standard, specification, or regulation. Additionally, this report is not intended for construction, bidding, or permit purposes. Thomas J. Freeman was the principal investigator for the project.

ACKNOWLEDGEMENTS

Mr. Bryan Stampley, Construction Division, Materials & Pavements Section served as Project Director. His assistance in providing and explaining the data elements were crucial to the success of this project.

Special thanks are also extended to Craig Cox, Construction Division, Materials & Pavements Section for his help in properly formatting the data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. BACKGROUND AND OBJECTIVES

1.1 OBJECTIVES

The purpose of this project is to develop several statistical tools and measures for PMIS data. The tools include defining appropriate sample sizes for auditing contractor performed distress ratings. The statistical measures to be determined include determining the measurement error for the PMIS Condition Score and to use this error in analyzing recent trends.

1.2 DISTRESS DATA COLLECTION

There are numerous methods that can be used to collect surface distress information. Distress surveys can be conducted and analyzed manually, or equipment can be used to automate some of the steps. In general, methods that are more costly are also more accurate, are more precise, and have the greatest resolution (Smith et al. 1996). Since the terms accuracy, precision, and resolution appear throughout this report, they will be defined so the terms will have the proper meaning. Accuracy is the degree to which the method provides a value that matches an accepted reference value (ASTM 1992). Precision is the closeness of agreement, or repeatability, among multiple measurements obtained under defined conditions (ASTM 1992). Resolution is the smallest increment that can be measured.

The accuracy, precision, and resolution needed depend on the goals of the pavement management system and the funds available to pay for the inspection services. Some methods are more subjective than others. Publications by Hicks and Mahoney (1981), Epps and Monismith (1986), Cable and Marks (1990), and TxDOT (undated) describe in detail many of the data collection methods.

1.3 INTERPRETING DISTRESS INFORMATION

Distress information can be converted into a Condition Score, or information on each distress type and severity can be used individually. The Condition Score combines information from all of the distress ratings into a single number. This number can be used at the network level to define the condition state, to identify when treatments are needed, as a part of ranking/prioritization, and in condition projection. Individual distress type, severity, and

quantity information at the network level is normally restricted to use in a decision tree procedure to identify feasible treatments. For project-level analysis, individual distress information is routinely used in determining the cause of deterioration, identifying feasible treatments, and estimating repair quantities.

From experience with manual systems, it is apparent that even though there is some deviation in distress types and quantities among raters during network-level surveys, the condition indexes may agree reasonably well. Since the condition indexes are used as key management indicators at the network level, they can be used if they are reasonably accurate and precise, even if the collection of individual distress types may not give the accuracy, precision, and resolution desired for individual distresses at the project level.

Different surface types have different distress types that must be addressed in condition indexes. Pavements with hot-mix asphalt concrete (HMAC) surfaces are the predominant surface type in Texas, and the most important distress types must be included. Asphalt concrete overlays on Portland cement concrete have reflective cracks and crack deterioration that would not typically be found on other surface types. Pavements with a slurry seal applied to asphalt concrete (slurry) have fine, relatively uniform surface texture. The bituminous surface treatment (BST) pavements have a coarse surface texture that tends to mask some of the distress types, and they tend to have more distresses caused by pavement layer instability. The Portland cement concrete (PCC) pavements have a completely different set of distress types, but the amount of PCC pavement in Texas is somewhat limited. Although other pavement surface types are present, they can generally be included in one of these groups. For ease of discussion, when the generic term asphalt is used, it includes all of the asphaltic and bituminous surface types (AC, APC, slurry, and BST). PCC will be used to identify pavements with Portland cement concrete surfaces, and the individual names will be used to address the specific surface types.

1.4 PURPOSE OF DISTRESS DATA COLLECTION

Distress surveys are performed to collect data on the entire network (network level), to determine the type and cost of treatment for a specific project (project level), and to collect data for research purposes (research level). These different purposes require different data collection methods and accuracy. The objectives of each are summarized below (FHWA 1995):

Network level:

a. expediency in the field condition survey;

b. reproducibility of survey results should be provided within a reasonable degree of accuracy; and

c. useful information should be provided for identifying potential rehabilitation projects, identifying potential budget needs, and establishing priorities.

Project level:

a. reproducibility of survey results within a high degree of accuracy;

b. useful information should be provided for identifying causes of failures and determining effective maintenance and repair techniques;

c. useful information should be provided for estimating costs of maintenance, repair, and restoration; and

d. expediency in field condition survey (e.g., less than one quarter man day per project).

Research level:

a. accuracy of survey results with a high degree of reproducibility;

b. useful information should be provided for identifying causes of distresses;

c. location information should be provided for locating distresses so that they can be tracked over time; and

d. expediency in field condition survey (e.g., less than one half man day per project).

Network-level inspections are usually a driving type survey. Project-level evaluations can be conducted from a slowly moving vehicle but are also conducted by a walking survey, while research-level inspections are usually a detailed walking survey.

1.5 CURRENT METHOD OF DISTRESS DATA COLLECTION IN TEXAS

Prior to 2001, the frequency of surveys conducted by TxDOT on a particular road in the network was based on the functional class of the highway. Interstate highways (IH) were inspected annually while non-interstate highways were inspected every two years, although some districts did inspect greater percentages.

Traditionally, TxDOT used multiple crews even within a single district. These crews were typically composed of personnel with other duties and were temporarily assigned to the inspection effort.

More recently, in an effort to standardize data collection, reduce variability between districts, reduce the number of raters, complete the work more quickly, and to free up district personnel, TxDOT hired contractors to perform this intensive data collection effort. Prior to using contractors, the large variability of the inspection results made comparisons between districts more difficult since raters in one district might have rated pavements lower (more strict interpretation or slightly different criteria) than a rater in a different district would have. Using contractors for the data collection reduced this variability by using a single crew for surveying multiple districts within a region.

1.6 PMIS DISTRESS TYPES FOR FLEXIBLE PAVEMENTS

Table 1 lists the current distresses used in the TxDOT Pavement Management Information System (PMIS) (TxDOT 2002). Currently, there are no severity levels associated with the PMIS distress survey. The addition of severities would complicate the process of defining distress but would provide added utility to the PMIS process as wider cracks could be separated and accounted for separately from hairline cracks.

Any survey method must accurately, and at least as importantly consistently, quantify and identify the appropriate distress type. Although it is sometimes difficult to obtain consensus, even among qualified inspectors, as to whether a particular distress is still longitudinal cracking or whether it has progressed to the point where it should be considered alligator cracking, once a decision is made, the survey teams must reliably and consistently identify and measure the distress accurately. The following is a brief discussion of each of the distresses. All quotations and references are from the TxDOT PMIS manual (TxDOT 2002).

Table 1. FY2003 PMIS Distress Types for Flexible Pavements.

Distress	Ratings Based On
Rutting, Shallow	Percent of Wheelpath Length (Measured by Profiler)
Rutting, Deep	Percent of Wheelpath Length (Measured by Profiler)
Patching	Percent Lane Area
Failures	Number per Section
Block Cracking	Percent Lane Area
Alligator Cracking	Percent Wheelpath Length
Longitudinal Cracking	Length per 100 feet
Transverse Cracking	Number per 100 feet
Flushing (Optional)	Category
Raveling (Optional)	Category

Rutting

Rutting will continue to be collected by automated equipment.

Patching

"Patches are repairs made to pavement distress. The presence of patching indicates prior maintenance activity, and is thus used as a general measure of maintenance cost."

Problems with recording patches have been a continuing problem in PMIS because the definition of what is to be counted as a defect (patch) is continually being debated and updated. For example, an improved area that is 495 feet long is a patch, but if it were to be 500 feet long, it would be counted as an overlay and have no impact on the Distress Score. One section would be rated as having a score of 74, while the overlay would receive a score of 100. Functionally, these pavements are equivalent. A measure of the complicated nature of patches is that the current PMIS manual contains 11 separate explanations and modifications, called "Special Cases" in the manual.

The identification and determination of quantities for this distress accounts for the largest discrepancy in PMIS score on both the training and audit surveys.

Failures

"A failure is a localized section of pavement where the surface has been severely eroded, badly cracked, depressed, or severely shoved. Failures are important to rate because they identify specific structural deficiencies that may pose safety hazards."

This distress is subjective in nature and is often misidentified during training classes and during audit surveys.

Block Cracking

"Block cracking consists of interconnecting cracks that divide the pavement surface into approximately rectangular pieces, varying in size from 1 foot by 1 foot (0.3 meter by 0.3 meter) up to 10 feet by 10 feet (3 meters by 3 meters). Although similar in appearance to alligator cracking, block cracks are much larger. Block cracking is not load-associated. Instead, it is commonly caused by shrinkage of the asphalt concrete or by shrinkage of cement or lime-stabilized base courses."

The difficulty with this distress is the identification of the pattern associated with block cracking. However, block cracking is merely a lot of longitudinal and transverse cracking, and misidentification may not change the overall distress score.

Alligator Cracking

"Alligator cracking consists of interconnecting cracks which form small, irregularly shaped blocks that resemble the patterns found on an alligator's skin. Blocks formed by alligator cracks are less than 1 foot by 1 foot (0.3 meter by 0.3 meter). Larger blocks are rated as block cracking. Alligator cracks are formed whenever the pavement surface is repeatedly flexed under traffic loads. As a result, alligator cracking may indicate improper design or weak structural layers. Heavily loaded vehicles may also cause alligator cracking."

It is very important to accurately assess this distress since it has a major impact on the Distress Score, maintenance level of service, and maintenance needs. Fortunately, because the PMIS survey does not have severity levels, it should be possible to obtain a high level of accuracy and repeatability. With severity levels, the distress has to be assigned to the different categories based on distress definitions that are not easily quantifiable.

However, during PMIS training classes and in the audit surveys, there is often a large discrepancy in the amount of alligator cracking reported.

Longitudinal Cracking

"Longitudinal cracking consists of cracks or breaks which run approximately parallel to the pavement centerline. Edge cracks, joint or slab cracks, and reflective cracking on composite pavement (i.e., overlaid concrete pavement) may all be rated as longitudinal cracking. Differential movement beneath the surface is the primary cause of longitudinal cracking."

The functional definition of a longitudinal crack is that if it is at least 1/8 inch wide (i.e., generally is visible while seated in the rating vehicle), it should be recorded as a longitudinal crack. As with alligator cracking, the interpretation of longitudinal cracking is both important and relatively straightforward. However, there is also a substantial difference in the values reported during training classes and in PMIS audit surveys.

Transverse Cracking

"Transverse cracking consists of cracks or breaks which travel at right angles to the pavement centerline. Joint cracks and reflective cracks may also be rated as transverse cracking. Differential movement beneath the pavement surface usually causes transverse cracks. They may also be caused by surface shrinkage due to extreme temperature variations."

Transverse cracks may be the easiest distress to catalog, though narrow cracks (1/8 inch wide) will be more difficult as the speed of travel increases. However, for flexible pavements, they have little impact unless there are many cracks (5 cracks per 100 feet result in a Distress Score of 91-94).

Flushing and Raveling

Flushing (sometimes called bleeding) and raveling are optional distresses that do not affect the Distress Score. While these optional distresses may be useful in identifying pavements needing routine or preventive maintenance, comprehensive skid testing rather than the network-level approach of one skid test per half-mile provides a more quantifiable measure of pavements that need attention.

1.7 PMIS DISTRESS TYPES FOR CONTINUOUSLY REINFORCED CONCRETE PAVEMENTS (CRCP)

Table 2 lists the current distresses used in the TxDOT Pavement Management Information System (TxDOT 2002) for CRCP pavements. All quotations and references are from the TxDOT PMIS manual (TxDOT 2002).

Table 2. FY2003 PMIS Distress Types for Continuously Reinforced Concrete Pavements.

Distress	Ratings Based On
Spalled Cracks	Number per Section
Punchouts	Number per Section
Asphalt Patches	Number per Section
Concrete Patches	Number per Section
Average (Transverse) Crack Spacing	Distance

Spalled Cracks

"A spalled crack is a crack that shows signs of chipping on either side, along some or all of its width."

The distress definition also includes a width definition (spall is greater than 3 inches), but width is difficult to estimate while in a moving vehicle. The spalled crack distress is often a source of substantial variation in the rater class ratings.

Punchouts

"A typical punchout is a full depth block of pavement formed when one longitudinal crack crosses two transverse cracks. Although usually rectangular in shape, some punchouts may appear in other shapes."

Punchouts can be difficult to accurately and consistently recognize during a field survey and are often a source of substantial variation in the rater class ratings.

Asphalt Patches

"An asphalt patch is a localized area of asphalt concrete which has been placed to the full depth of the surrounding concrete slab, as a temporary method of correcting surface or structural defects."

This distress is easy to identify. Few problems occur in identifying this distress.

Concrete Patches

"A concrete patch (a 'longer lasting' repair) is a localized area of newer concrete which has been placed to the full depth of the existing slab as a method of correcting surface or structural defects."

Concrete patches are occasionally a source of differences during the field surveys, because of the similarity of the color and the difficulty in estimating the length of the patch (patches greater than 10 feet are recorded as one patch for every 10 feet) on a busy highway while driving along the shoulder.

Average Crack Spacing

"Average crack spacing is not, in itself, a pavement distress type. It is rated as a method of obtaining the percentage of transverse cracks that are spalled. However, average crack spacing is valuable as a measure of whether or not the CRCP slab is behaving as designed. A CRCP section with a small average crack spacing may deteriorate rapidly into a series of small punchouts if the proper corrective procedures are not applied."

This distress is relatively easy to measure; however, at higher speeds the narrow cracks may be a problem. Fortunately, average crack spacing has no impact on the Distress Score.

1.8 PMIS DISTRESS TYPES FOR JOINTED CONCRETE PAVEMENTS (JCP)

Table 3 lists the current distresses used in the TxDOT Pavement Management Information System (TxDOT 2002) for JCP pavements. All quotations and references are from the TxDOT PMIS manual (TxDOT 2002).

Table 3. FY2003 PMIS Distress Types for Jointed Concrete Pavements.

Distress	Ratings Based On
Failed Joints and Cracks	Number per Section
Failures	Number per Section
Shattered (Failed) Slabs	Number per Section
Slabs with Longitudinal Cracks	Number of Slabs per Section
Concrete Patches	Number per Section
Average Joint Spacing	Distance

Failed Joints and Cracks

"The distress type 'failed joints and cracks' covers two major items: spalled joints and transverse cracks, and asphalt patches of spalled joints and transverse cracks."

Although the definitions for this distress are straightforward and well defined, considerable variability in results was typical for the rating classes.

Failures

"Failures are localized areas in which traffic loads do not appear to be transferred across the reinforcing bars. Failures are typically areas of surface distortion or disintegration."

Failures in JCP were usually identified fairly consistently in the training classes. This distress is a major component of the PMIS Distress Score, but the distress is fairly rare.

Shattered Slabs

"A shattered slab is a slab that is so badly cracked that it warrants complete replacement."

This distress was also identified fairly consistently and accurately.

Slabs with Longitudinal Cracks

"A longitudinal crack is a crack that roughly parallels the roadbed centerline."

Although this distress was sometimes misidentified, most of the inspectors were accurate. When it was misidentified, the results were dramatically different. Since the measurements are

fairly well defined, more training is needed to ensure that this distress is recorded accurately. However, jointed concrete pavement is somewhat rare, and most slabs are not distressed.

Concrete Patches

"A concrete patch (a 'longer lasting' repair) is a localized area of newer concrete which has been placed to the full depth of the existing slab as a method of correcting surface or structural defects."

Concrete patches are usually easy to identify and record.

Apparent Joint Spacing

"Some transverse cracks may become so wide (long) that they look and act like joints. The crack must be greater than ½ inch (13 mm) wide (long) across the complete width of the lane. These 'apparent' joints are important because they serve to divide the original slab into smaller units."

There are usually no problems identifying this distress.

1.9 RATER CERTIFICATION AND TRAINING METHODOLOGY

Currently, the method used to conduct network-level distress data collection for pavements in Texas involves having crews drive at approximately 15 miles per hour (mph) (24 km/hr) along the shoulder, or in the lane, and estimate the quantity of each distress type for a given length. For all asphalt and CRCP surfaces the unit of measurement is 40 feet (12 m), which in the field is the distance from the start of one lane stripe to the start of the next lane stripe. For JRCP, most of the distresses are estimated for each slab or joint. Except for the under-construction or otherwise identified miles, all of the mileage is inspected. For roads that have undivided roadbeds, only one lane (the lane that appears to be in the worst condition) is inspected. For divided roads, one lane in each direction is inspected. TxDOT hires contractors to perform this service and has TxDOT and personnel from the Texas Transportation Institute (TTI) check the results by performing an audit survey of a small percentage of the pavements. Part of the research was to determine whether the audit percentage was adequate.

Prior to the start of the inspection season, contractor, TxDOT, and TTI personnel are required to attend training classes to become certified inspectors for the calendar year. Changes to the manual, including clarifications and interpretations, and inspections of selected sections are used in order to reduce the rater-to-rater variability between inspection teams.

The training classes begin with the TxDOT representative (usually Doug Chalman) explaining the reference marker (RM) and lane designation system to any raters that had not attended previous certifications. The next step involves explaining each of the distresses including the definition, rating procedure, acceptable values, and a detailed description of the special cases involved with the distress. The class then conducted field trips to illustrate the distresses and data collection procedures.

Later in the week, experienced raters arrive and must attend the distress data collection review. These are called refresher classes. As with the inexperienced raters, each distress was reviewed and a brief field trip was taken to refresh the raters on field measurements. After that, the raters were separated into teams containing both experienced and inexperienced raters.

Several field test sections were inspected, and the results were reviewed. Teams whose results were far from the average of the group were chastised, reminded of the distress definition, and counseled to bring their ratings in-line. The following table (Table 4) from a flexible training class in Austin lists the values for the mean, standard deviation (SDev), and coefficient of variation (CVar = Mean/SDev).

Finally, prior to receiving their certification, raters are required to pass an open book test developed by TxDOT that covers the material discussed in the class. A score of 70 is required to pass. It is very rare that anyone fails this test.

Table 4. Statistical Values for PMIS Flexible Distresses.

Sect.	Patch			Failures			Alligator			Longitudinal		
	Mean	SDev	CVar	Mean	SDev	CVar	Mean	SDev	CVar	Mean	SDev	CVar
7	24.8	8.5	34%				1.0	2.1	207%	14.6	7.4	51%
10	5.1	4.7	92%	2.0	1.9	93%				90.3	11.6	13%
14	13.4	4.8	36%							20.8	4.6	22%
15	56.0	11.9	21%							18.1	9.0	50%
18	13.3	7.1	53%	0.4	0.5	138%	13.4	5.6	42%	70.4	24.9	35%
20							13.0	5.0	43%	134.0	32.0	24%
Ave.		7.4	47%		1.2	115%		4.2	97%		14.9	32%

1.10 AUDIT PROCEDURE

In order to verify that the contractors are performing inspections accurately, TxDOT has enlisted the support of TTI to provide certified auditors that travel with TxDOT district personnel and rate a portion of the pavements. TTI raters attend the certification classes alongside the raters provided by the contractors.

TxDOT provides a list of audit sections to the districts based on providing sections that have representative pavement types in approximately the same percentage as that found in the district, pavements that were in good, fair, and poor condition as of the last inspection, and with the constraint that it includes at least three consecutive PMIS sections.

TxDOT or TTI personnel create maps of the counties and district to aid in scheduling inspections, and the audit is coordinated with contractor personnel to ensure that the audit and annual inspections are conducted within two weeks of each other. Reducing the time between annual and audit inspections reduces the chance that pavement condition will change between the inspections. Occasionally, seal coat crews will still be working in the area, and there have been pavements rated by one team where the road was then seal coated before the second team could perform the inspection. In order to document this and other problems, such as having a patching crew repair the road, TTI inspectors include notes in the inspection file that document these conditions and provide other indications on potential problems with the inspections. In addition, audit crews typically note whether the sections should be easy, medium, or hard to rate. A section is easy when there is little distress or only one distress that is easy to count. Hard sections have substantial distress types and quantities and may have high traffic volumes or other factors that make the inspection difficult. After the annual and audit surveys are completed, these notes can help explain discrepancies between the surveys.

1.11 PROBLEMS WITH CURRENT METHOD

In spite of the best efforts of TxDOT and all parties involved, there are significant problems with the current inspection method:

- One problem is certainly the cost. The annual cost of these inspections is approximately $1,500,000. When the current contract expires and is renegotiated, it is expected that the costs will increase.

- Vehicles driving very slowly along the shoulder or within the lane are a potential safety hazard for both the inspectors and the traveling public.

- There is considerable variability in the inspections. Normally, inspections conducted during the class would be expected to be the most consistent and have a smaller standard deviation than there would be during production inspection because:
  - all parties know they are being observed;
  - there should be no "burn out" since only a few pavements are inspected during class;
  - there is no incentive, time limit, or rush to finish the inspection;
  - definitions of distresses to be used are fresh in the minds of the inspectors; and
  - there are three or four inspectors in each vehicle, so no distresses should be missed.

- Inspectors in different regions may still be rating differently. Since there is no overlap of inspections on the same pavements, this hypothesis is never tested.

1.12 IMPROVEMENTS TO AUDIT METHODOLOGY

As part of the review of the audit procedures, researchers suggest the following improvements. Most of these suggestions were immediately implemented (1, 2, 4, and 5). Those that have not been implemented (3 and 6) are identified as such.

1. Audit should be from RM to RM. Originally, three segments, usually each 0.5 mile in length were selected for the audit survey. The rationale was that with at least three segments, the auditors would spend less time simply driving from one area to another and more time rating. Since the direction of travel on a road is dictated by the location of the sun (distresses can be seen ***much*** better when looking into the sun), auditors often located a reference marker and then skipped the first 0.5 mile segment before starting the audit. Changing the selection process to have the audit section start at an RM and run continuously to the next RM increased accuracy (because the location where the audit was to start was more closely identified), reduced the driving time (fewer locations but more segments in an area), and simplified the paperwork.

2. Increase Audit Percentage but Reduce Counties. As the statistical analysis will demonstrate, inspecting a higher percentage within a county reduces the errors and improves the confidence level that poor inspection procedures will be identified. However, since the statistical analysis was not yet completed, the primary reason that this was adopted was because it ***greatly*** reduced the nonproductive travel time. Under the original plan, it was common for auditors to drive two hours, inspect just a few sections, and then drive two or more hours to get to the next area.

3. Use Variable Audit Percentage. Later in this report, a method of selecting the statistically appropriate percentage is identified. This method would increase the accuracy of the audit process and make the re-rating of suspect counties statistically defensible. Counties with less than about 300 segments should not be surveyed. This recommendation has not been implemented.

4. Eliminate 0.1 Mile Segments. At the start of a highway, the reference marker is usually offset from the beginning of the road so that the RM can be placed on a highway sign. These offsets are usually identified as being 0.1 mile in length but are often as short as 50 feet and as long as 750 feet. In these intersections, atypical distresses are often found and the traffic is usually higher and can be turning onto the road with little prior notice. Because these segments are difficult to inspect, they make poor locations to check the work of others.

5. Eliminate "Bad" Sections. To ensure an adequate mix of sections in good, fair, and poor condition, the original plan used the previous year's rating to determine these sections. However, the practical significance of this was that those sections that were "bad" last year were usually on the seal coat or rehabilitation list for this year and were often in perfect condition. The recommendation is to eliminate these. Although not directly implemented, the section list generated a few years ago has continued to be used and those sections are beginning to age. Likewise, those sections that were once in "fair" condition are now becoming "poor."

6. Require Raters to Rate to Receive Certification. This is the most controversial of the recommendations and has not been implemented. In order to be certified, a rater need only to be able to see far enough to read a newspaper and have enough intelligence or experience to find the answers to the test in the manual. Some raters have even used tests from previous years to assist them on the test. While this level of expertise may be good enough for the person who records the data, the person actually identifying the distress types and quantities should be held to a higher standard. Additionally, although the teams that are far from the mark are cautioned to get their ratings more in line with the class, the teams often feel that they are the ones who are correct and the rest of the class was in error. To improve this situation, raters should be ***required*** to actually rate some pavement properly in order to become certified. Test sections should be identified and surveyed by a TxDOT instructor and at least one other trusted rater. These sections should be surveyed and a "ground truth" established. Raters wishing to be certified should be required to inspect these sections and if the results of their PMIS Distress Scores are not all within 10 points of the "expert" ratings, they are not certified. Another option would be to require all distresses to be within a certain variance, but as noted before the identification of distresses can be problematic.

CHAPTER 2. STATISTICAL ANALYSIS AND PROCEDURES

2.1 INTRODUCTION

The need to reduce survey error is easily understood although the methods to do so are not usually so obvious. One method TxDOT has already implemented is to use automated equipment to measure rutting. The automated rutting measurement has proven to be effective and accurate and reduces the exposure of the rating team to traffic. There has been considerable interest by TxDOT, other states, FHWA, municipalities, and other agencies in the use of an automated survey vehicle to measure distress.

An important part of reducing the error is to define the quantity of measurement error of PMIS surveys. Figure 1 illustrates the problem with not knowing the measurement error. In Figure 1, there are three data points representing three years of data. The question to be answered from this graph is whether there is a trend in the data. For our example the question would be "Is the Condition Score improving?"

The data points seem to show a general increase in Condition Score in that the 1996 and 1997 values are greater than the 1995 value and the regression line drawn through the points has a positive slope. However, if the measurement error is as little as three points, the plot shown in Figure 2 is also possible.

Based on the previous example, with a known measurement error, the appropriate answer to the question of a trend would be to say that there is no trend.

There are a variety of methods that can be used to determine the error.

Statistical Comparison of Data Sets

If the distress data collected by different persons using the same method, or different persons using different methods, on the same section of pavement are exactly the same, the distress types, severities, quantities, and Distress Score calculated from each set will be the same. However, even if the same person inspects the same section of pavement on the same day, there will generally be some difference between the results. This difference is called the measurement error.

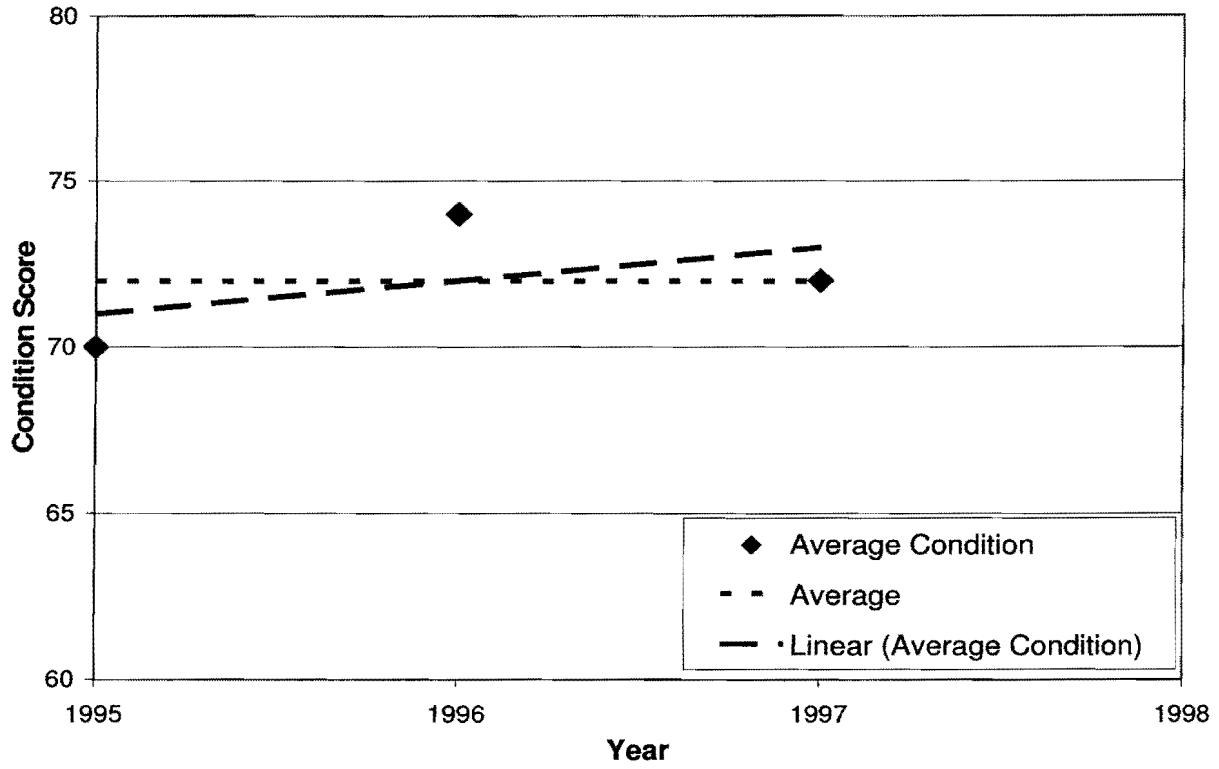

Figure 1. Example of Condition Score Analysis.

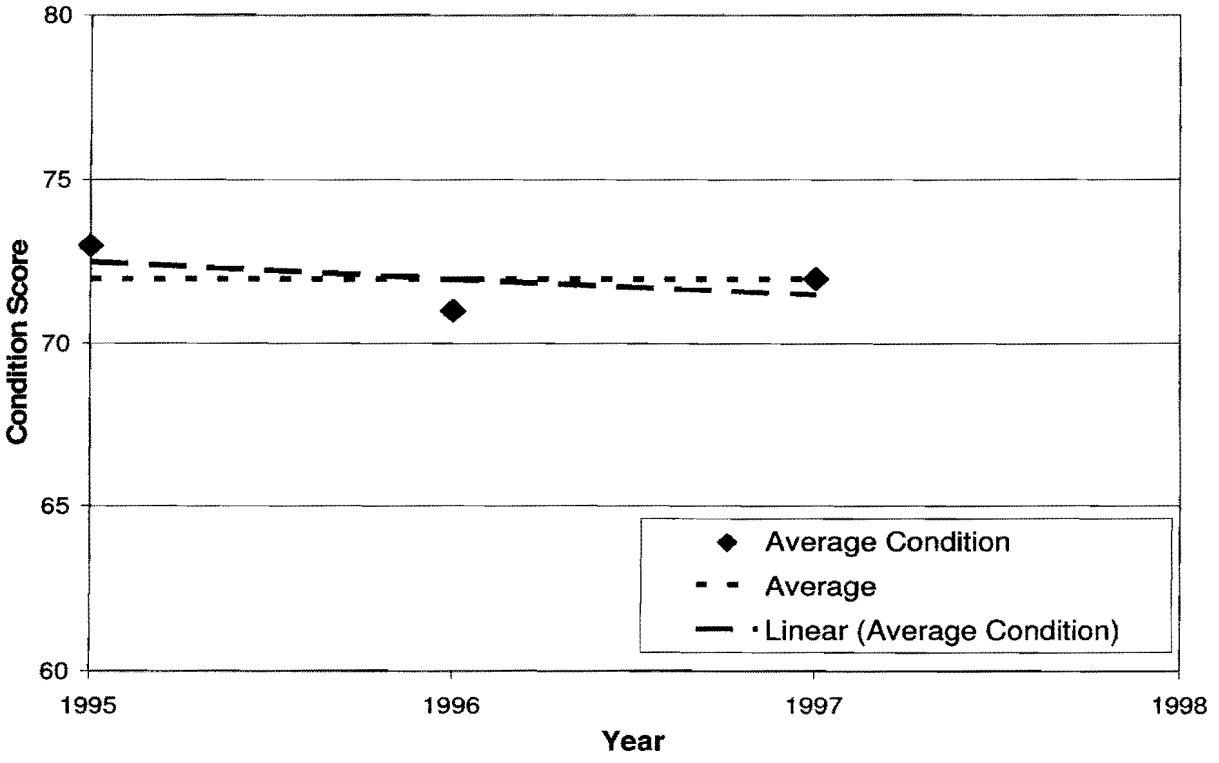

Figure 2. Example of Condition Score with Measurement Error.

Part of the concern about manual surveys is that there is considerable subjectivity in how inspectors define and interpret the distress types and severities, which leads to some of this error. When the same inspector inspects the same section of pavement several times, and when different inspectors inspect the same pavement at the same time, they may interpret the distress present as a different type of distress or as the same distress but with a different quantity or severity. These differences that occur when inspectors use the same set of definitions indicate the error that is expected either within repeat inspections by the same inspector or the error that is expected when more than one inspector is used in inspecting pavements. Because the PMIS Distress Score is calculated based on type and quantity of distress, differences in distress data will generally also affect the Distress Score. When different methods of defining the distresses are used, the results generally cannot be directly compared because of the differences in the definitions.

Even if all of the subjectivity due to manual definitions were removed, there is still a difference in the appearance of the road surface depending on the direction and angle of the sunlight striking the road surface, the amount of moisture on the pavement surface, and the direction from which the inspector views the pavement surface. When the temperature is high, some cracks may be pressed closely together, while they may be separated when the temperature is low. If the pavement surface is evaluated on different days, there may be an increase in the distress present on the surface, or maintenance personnel may apply maintenance treatments that will change the severity. These factors cause some of the error, and they cannot be changed by improving definitions.

Contractors provide teams of raters to rate the condition of the roads. It is important to ensure that these ratings are as accurate as possible. To accomplish this, inspection crews from TTI and TxDOT independently check the ratings performed by the contractors. TxDOT mandates that the contractor must redo the inspections in a county if more than 10 percent of the distress scores in that county differ by more than 15 points from the distress scores reported by the TTI/TxDOT audit team. Originally, a 6 percent sample of road segments was selected to be audited.

One purpose of this research was to examine the PMIS audit procedure data from a statistical approach in order to revise the current audit procedures.

When considering a sampling-based verification system, the possibility of decision-making errors should be considered. We examine in detail the performance of the current quality assurance procedure in the context of statistical decision theory, identify problem areas in the present specification, and propose possible improvements

2.2 STATISTICAL CONSIDERATIONS FOR DECISION RULES

The following definitions will be used throughout this report.

Noncompliant Score - A contractor Distress Score that differs by more than 15 from the corresponding auditor's score for a particular road segment.

Noncompliant Vendor - A contractor who has more than 10 percent noncompliant scores in a county.

Type I Error - An error that occurs when we incorrectly conclude a vendor is noncompliant (false positive).

Type II Error - An error that occurs when we incorrectly conclude a vendor is compliant; i.e., we fail to detect that a vendor is noncompliant (false negative).

In a sampling situation, we can never know for certain whether we have made a Type I or Type II error; rather, we speak in terms of the ***probabilities*** of these errors occurring, given certain assumptions about the behavior of our data. The probability of making a Type I error is denoted by α (the probability we will ***erroneously*** require the contractor to resurvey a county), while the probability associated with making a Type II error is denoted by β (the probability we will ***not*** require the contractor to resurvey when we should have). We can alternatively view α (Type I) and β (Type II) as the false positive and false negative rates, respectively. A measure related to Type II error is ***power***, defined as $(1-\beta)$. Power can be viewed as the probability of ***correctly*** concluding a vendor is noncompliant and measures our ability to detect noncompliance.

The number of samples (sample size) is very important in any sampling scheme. One can view any statistically based decision as an educated guess that depends on the amount of information at hand. The required sample size can be viewed as the amount of information necessary for a sufficiently good guess, where the criteria for a good guess depends on the error probabilities α and β we are willing to tolerate. All three criteria are mutually antagonistic in

that we can only optimize our decision procedure with respect to one criterion at the expense of the other criteria. Trade-offs must be made and we must seek a compromise among the above three criteria by constructing an appropriate decision procedure. If we feel that the consequences of incorrectly rejecting a rater as noncompliant are more severe than failing to identify a noncompliant rater, we should seek, subject to limitations on sample size, a decision procedure which minimizes α at the expense of β. If, however, we are more concerned with the possible consequences of failing to identify noncompliant raters, we should seek to minimize β at the expense of α.

2.3 DECISION PROCEDURE

Keeping in mind the previously discussed tradeoffs, we commence interpreting the TxDOT criterion in the context of statistical decision theory. Recall that the TxDOT criterion defines a rater to be noncompliant if more than 10 percent of the Distress Scores for road segments in a county differ by more than 15 points from the corresponding audit Distress Scores. We would like our decision procedure to have a low Type I error probability when the proportion of noncompliant vendor scores, p, is below 0.10 and a low Type II error probability when p is above 0.10. In fact, we will specify the maximum allowable value for α to be 0.05 when $p = 0.10$; our optimum decision procedure will be one which minimizes β at all values of p greater than 0.10, subject to the restriction on α.

If we denote the proportion of noncompliant Distress Scores observed in a sample of road segments by $\hat{p}$, then our decision procedure for determining whether a rater is compliant will be the following:

If $\hat{p} > c$, conclude that the rater is noncompliant.

If $\hat{p} \leq c$, conclude that the rater is compliant.

The number c, $0 \leq c \leq 1$, is a cutoff value for our decision procedure. It specifies the maximum proportion of noncompliant Distress Scores we are willing to tolerate in a particular sample of road segments. This is currently set at 0.10.

Statistics involving proportions usually assume that the sample from which the statistics are calculated come from an infinitely large population. Under this assumption, the individual

observations in the sample can be considered to be independent of one another, and the sample proportion has, by definition, a binomial distribution (Montgomery 1991). In the case of the road condition data, there are only a finite number of road segments (finite population) which can be sampled. If the sample is large relative to the population, the sample proportion will not change as much between different samples as it would for an infinite population. For example, if the entire population is included in the sample, the sample proportion will always be the same no matter how many samples are taken (assuming the population remains the same). In other words, observations in a sample from a finite population are not independent and the degree of correlation between successive samples depends on the sample size relative to the total population. This type of distribution is called hypergeometric. The hypergeometric distribution of the sample proportion will not have as much variation as for the infinite population case (binomial distribution), so test statistics involving the sample proportion will have smaller cutoff or critical values. Since we are dealing with proportions arising from hypergeometric random variables, we may apply a normal approximation with a finite population correction factor to obtain the following formula for the value of c:

$$c = p + 1.645\sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}, \tag{1}$$

where

p is the maximum allowable proportion of noncompliant distress scores for a county, specified to be 0.10;

N is the total number of road segments in a county, and

n is the number of road segments in the audit sample.

We make the important distinction here between the total number of rater scores in a county, termed the ***population*** of scores, and the ***sample*** of scores we use to evaluate raters. Our decision procedure makes an inference about the population of scores based on the information present in our sample. With this distinction made, one must then bear in mind that ***the proportion of noncompliant scores in a sample, while being representative of the county-wide noncompliance proportion, is not the same as the latter.*** To illustrate, suppose a rater obtains 100 scores for a county; of these, 10 scores are sampled for evaluation. Let us further presume

that the total proportion of noncompliant scores by that rater for the entire county was 0.10; that is, if we were to audit all 100 scores, we would find that 10 scores were noncompliant (the proportion noncompliant was 0.10). Finally, suppose that 5 of the noncompliant scores by that rater just happen to be included in the sample of 10 scores. Our rater will then be judged to have 50 percent of his scores noncompliant, even though his county-wide proportion of noncompliant scores is a much lower 0.10. Obviously, in this example, our rater has been extraordinarily unlucky in that so many of his noncompliant scores were included in the sample. Had a different sample been chosen, perhaps no noncompliant scores would have been included. Our decision procedure has been designed to guard against exactly this type of problem: the variability of results from sample to sample.

Unlike the current, fixed criterion specified by TxDOT where the maximum allowable proportion of noncompliant scores is 0.10, our cutoff will be a variable proportion of noncompliant scores depending on the number of segments in a county. Examining Equation 1, we see that the maximum allowable proportion of noncompliant scores increases as sample size decreases. This occurs because the uncertainty in a decision procedure is greater for smaller samples, where we have less information on which to base our decision, than for larger samples. In order to ensure that raters having 10 percent noncompliant scores are not rejected as noncompliant more than 5 percent of the time, we need to increase the maximum allowable proportion of noncompliant scores in a sample from the currently fixed proportion of 0.10 to a higher value to include a reasonable margin of error to allow for uncertainty in the rating and audit procedures.

The power of our decision procedure to detect a noncompliant rater may also be determined. Let $p_A > p = 0.10$ be a rater's proportion of noncompliant distress scores within the entire county; then the probability of identifying this rater as noncompliant is

$$P\left[Z > 1.645\sqrt{\frac{p(1-p)}{p_A(1-p_A)}} - \frac{p_A - p}{\sqrt{p_A(1-p_A)}}\sqrt{\frac{n(N-1)}{N-n}}\right], \qquad (2)$$

where Z is a random variable having a standard normal distribution.

2.4 APPLICATION TO ATLANTA AND LUFKIN DISTRICTS

Sample data from two representative highway districts where contractors had been collecting distress data for two years and which had a wide range of number of segments in a county were used to develop and test the statistical procedures. The Atlanta and Lufkin districts constitute 895 audit segment ratings, covering 15 Texas counties. These are listed by county number in Table 5. The number of audited segments in each county ranged from a minimum of 27 to a maximum of 89. The proportion of noncompliant scores observed for the audited road segments is also shown for each county.

Table 5. Data for Atlanta and Lufkin Districts.

County Number	Total Number of Segments (N)	Audited Segments (n)	Percent Audited	Proportion Noncompliant ($\hat{p}$)
3	872	71	8.1	0.2817
32	257	48	18.7	0.0833
103	1209	86	7.1	0.1395
114	842	71	8.4	0.2958
172	325	27	8.3	0.1111
174	873	72	8.2	0.2639
183	708	28	4.0	0.2500
187	838	89	10.6	0.1460
202	475	56	11.8	0.2500
203	538	81	15.1	0.0123
204	505	56	11.1	0.1071
210	788	64	8.1	0.1250
225	541	38	7.0	0.1842
228	439	44	10.0	0.3181
230	707	64	9.1	0.0781

For each county, we found the cutoff value for the decision procedure yielding minimum Type II error (false negative), for a Type I (false positive) error not to exceed 0.05, at $p = 0.10$. The cutoff values are shown in Table 6, along with the decision outcome under both the unmodified TxDOT criterion and our procedure. The sample fractions, f, shown in Table 6 are simply the proportion of audited road segments in a county. In examining the table, one sees that the necessary proportion of noncompliant scores in a sample for rejecting a rater as noncompliant is, as expected, larger than 0.10. Note that the cutoff value is inversely proportional to sample size, with larger samples requiring lower cutoff values. We also see from the table that our

decision procedure rejects raters less often than decisions adhering to a strict 10 percent criterion; of the 15 counties examined, 7 are deemed noncompliant under our statistical decision procedure, while 12 are identified as noncompliant under the 10 percent rule. Note also that in marginal cases, where the proportion noncompliant is slightly greater than 0.10, our decision procedure, with its margin of error, does not reject a rater as noncompliant. For example, County 204, with a proportion of noncompliant scores in the sample of 0.1071, is rejected as noncompliant under the 10 percent rule, while our decision procedure more reasonably concludes the proportion is within the margin of error. Another principle advantage of our decision procedure is that, properly constructed, it is very fair to raters, since its margin of error provides a formal mechanism for giving "benefit of the doubt." Decisions made using such a procedure are easier to justify and interpret.

Table 7 investigates the power of our decision procedure to detect noncompliant raters for each of the 15 counties. The probabilities of rejecting a rater depend on the county-wide proportion of noncompliant scores, p. While we do not know this proportion, we do know the quantitative relationship between it and the power of our procedures. We may then calculate power for a wide range of p and generate what are known as ***power curves*** to describe the performance of our decision procedures. Two such power curves, corresponding to Counties 103 and 225, are shown in Figure 3. Looking more closely at Table 7, one sees that the power to detect noncompliant raters increases as the county-wide proportion of noncompliant scores becomes much larger than 0.10. The rate of this increase is more rapid for larger sample sizes, as can be observed by comparing the power for counties 103 and 225 in Table 7 and Figure 3.

Note that the probability of detecting a noncompliant rater, given that their proportion of noncompliant scores across a county is 0.15, is less than 0.5 for all but one of the counties. In other words, if a rater's county-wide noncompliance rate is 0.15, on average, more than half the time the county will not be identified. Power improves for $p = 0.20$, with our decision procedure being able to detect noncompliant raters over 90 percent of the time, on average. We see, however, that the ability to reliably detect noncompliance only emerges when a raters' county-wide proportion of noncompliant scores is at least twice the currently allowable rate of 10 percent.

Table 6. Optimum Decision Rules.

County Number	Audit Sample Size (n)	Audit Sample Fraction (f)	Proportion Non-compliant ($\hat{p}$)	Cutoff (c)	Reject Rater If $\hat{p} > 0.10$?	$\hat{p} > c$?
3	71	0.0814	0.2817	0.1562	Yes	Yes
32	48	0.1868	0.0833	0.1644	No	No
103	86	0.0711	0.1395	0.1513	Yes	No
114	71	0.0843	0.2958	0.1561	Yes	Yes
172	27	0.0831	0.1111	0.1911	Yes	No
174	72	0.0825	0.2639	0.1557	Yes	Yes
183	28	0.0395	0.2500	0.1915	Yes	Yes
187	89	0.1062	0.1460	0.1495	Yes	No
202	56	0.1179	0.2500	0.1620	Yes	Yes
203	81	0.1506	0.0123	0.1506	No	No
204	56	0.1109	0.1071	0.1622	Yes	No
210	64	0.0812	0.1250	0.1592	Yes	No
225	38	0.0702	0.1842	0.1773	Yes	Yes
228	44	0.1002	0.3181	0.1706	Yes	Yes
230	64	0.0905	0.0781	0.1589	No	No

Table 7. Power of the Decision Procedures.

County Number	Audit Sample Size (n)	Audit Sample Fraction (f)	Probability of Rejecting a Rater If $p = 0.15$	$p = 0.20$	$p = 0.25$	$p = 0.30$
3	71	0.0814	0.4398	0.8561	0.9768	0.9977
32	48	0.1868	0.3789	0.7799	0.9446	0.9904
103	86	0.0711	0.4860	0.8996	0.9889	0.9993
114	71	0.0843	0.4405	0.8569	0.9771	0.9978
172	27	0.0831	0.2666	0.5741	0.7866	0.9100
174	72	0.0825	0.4434	0.8600	0.9781	0.9979
183	28	0.0395	0.2655	0.5718	0.7842	0.9083
187	89	0.1062	0.5058	0.9149	0.9921	0.9996
202	56	0.1179	0.3946	0.8017	0.9553	0.9933
203	81	0.1506	0.4937	0.9058	0.9903	0.9995
204	56	0.1109	0.3929	0.7995	0.9542	0.9930
210	64	0.0812	0.4153	0.8282	0.9667	0.9959
225	38	0.0702	0.3129	0.6696	0.8727	0.9618
228	44	0.1002	0.3432	0.7239	0.9118	0.9792
230	64	0.0905	0.4176	0.8310	0.9678	0.9961

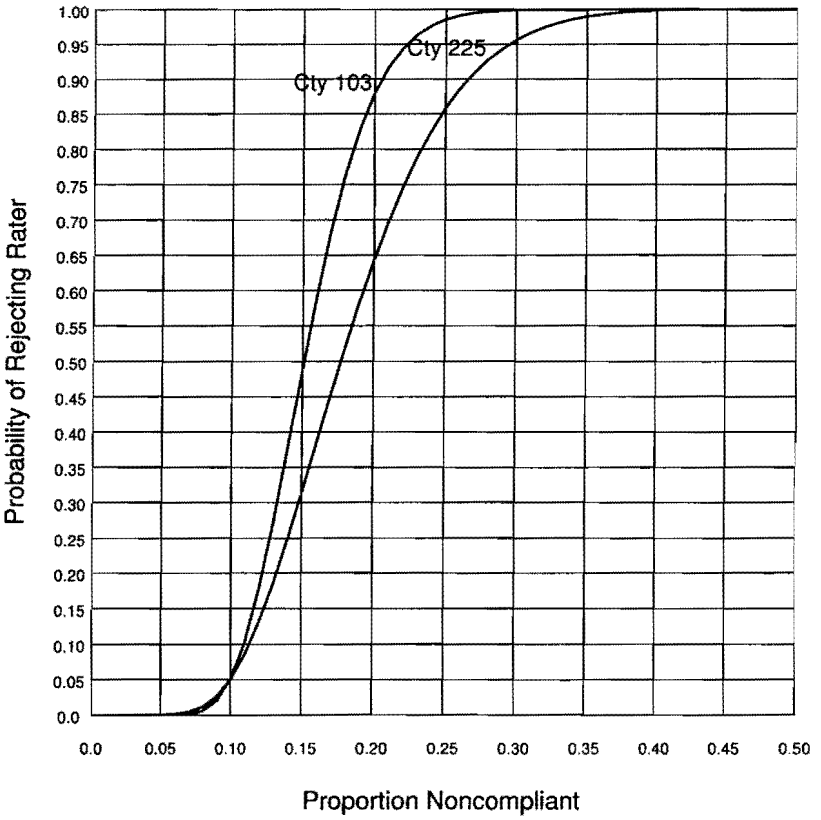

Figure 3. Power Comparison of Decision Procedures for Two Counties.

The performance of the decision procedures for Counties 103 and 225 in Figure 3 is especially relevant when we consider the sizes of these counties. The sample fractions for both are approximately 0.07 (as seen in Tables 6 and 7), but the total number of road segments in county 103 is 1209, over twice the number for county 225 (see Table 5). Clearly, there is a wide difference in performance between the decision procedures for the two counties; the power of the decision procedure for County 225 is always much poorer than that for County 103. It seems that using the same sample fraction across counties of differing size yields different performance results.

Tables 8 through 10, along with Figures 4 and 5, examine the relationship of sample size to power in more detail. Table 8 shows the decision procedures calculated for 10 counties with varying total numbers of road segments. All the procedures were obtained using 6 percent sample sizes and a fixed Type I (false positive, rater is rejected unnecessarily) error of 0.05. We notice in Table 9 that, for various proportions of noncompliant scores, the probability of rejecting

a rater as noncompliant is very poor for smaller audit samples and county-wide proportions of noncompliant scores near 10 percent, but improves as both the audit sample size and the proportion of noncompliant scores increase. This may be more easily seen in Figure 4. One may also observe from Table 10 and Figure 4 that as sample size decreases, the decision procedures will require large county-wide proportions of noncompliant segments to detect rater noncompliance with any high level of power.

The behavior we observe in Tables 9 and 10 and Figures 4 and 5 is explained by the fact that the performance of statistical decision procedures depends on the amount of data at hand. Specifying a fixed sample fraction gives inconsistent results in terms of the ability of the decision procedure to detect noncompliance; in this case, we are allowing sample size to dictate the performance of our decision procedure, instead of having the decision procedure dictate the sample size. Given a variable sample size, the error probabilities of our decision procedure will also be variable and hence uncontrolled. Consistent performance may only be achieved by ensuring that sample sizes are adequate for the degree of performance required.

The primary impact of this analysis is that under the current audit procedures, we are unjustified, statistically speaking, to reject a contractor's rating of a county unless the percentage of noncompliant scores is much, much higher than the current 10 percent criterion.

Table 8. Decision Procedures for Various County Sizes (6 Percent Sampling).

Segments per County (N)	Audit Sample Size (n)	Cutoff (c)
100	6	0.5000
200	12	0.4167
300	18	0.3333
400	24	0.2917
500	30	0.3000
600	36	0.2778
700	42	0.2619
800	48	0.2500
900	54	0.2407
1000	60	0.2500

Table 9. Probabilities of Rejecting a Rater with 6 Percent Sampling.

Segments per County (N)	Proportion Noncompliant $p = 0.15$	$p = 0.20$	$p = 0.25$	$p = 0.30$
100	0.1515	0.2848	0.4059	0.5189
200	0.1886	0.3839	0.5538	0.6968
300	0.2206	0.4657	0.6633	0.8088
400	0.2497	0.5360	0.7462	0.8801
500	0.2770	0.5970	0.8092	0.9252
600	0.3029	0.6502	0.8570	0.9537
700	0.3276	0.6967	0.8931	0.9714
800	0.3512	0.7373	0.9204	0.9825
900	0.3740	0.7727	0.9408	0.9893
1000	0.396	0.8036	0.9561	0.9935

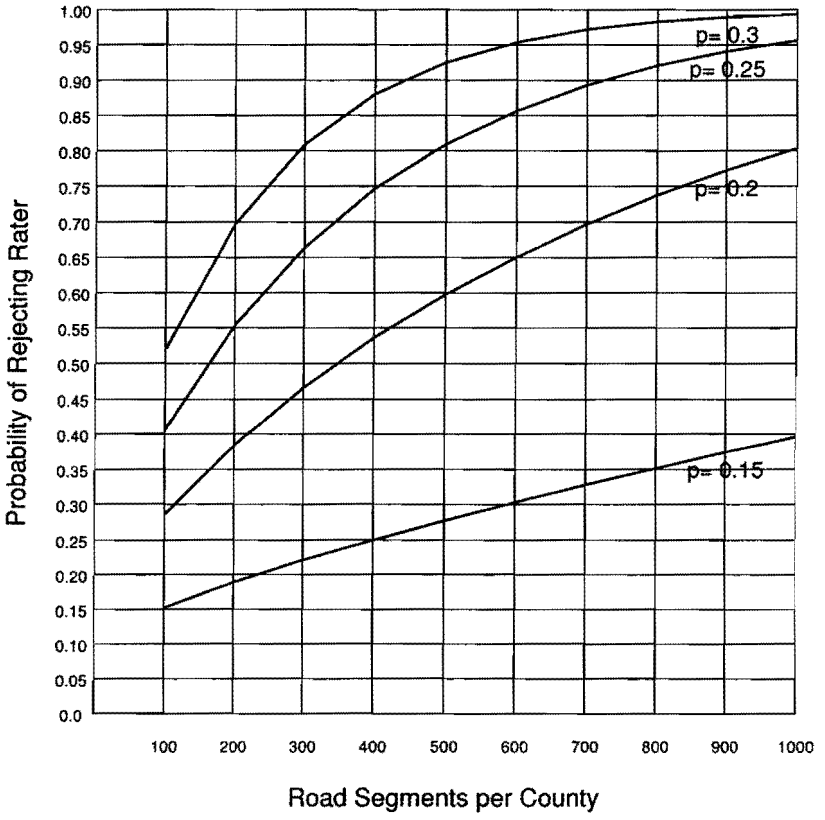

Figure 4. Probabilities of Rejecting a Rater under 6 Percent Sampling.

Table 10. Proportion of Noncompliant Scores Required before Rater Is Rejected (6 Percent Sampling).

Segments per County (N)	$\Pr(\text{rej. rater}) = 0.9$	$\Pr(\text{rej. rater}) = 0.75$	$\Pr(\text{rej. rater}) = 0.5$
100	0.550	0.430	0.300
200	0.420	0.330	0.240
300	0.355	0.285	0.215
400	0.320	0.260	0.200
500	0.295	0.240	0.190
600	0.275	0.230	0.180
700	0.260	0.220	0.175
800	0.250	0.210	0.170
900	0.240	0.205	0.170
1000	0.230	0.200	0.165

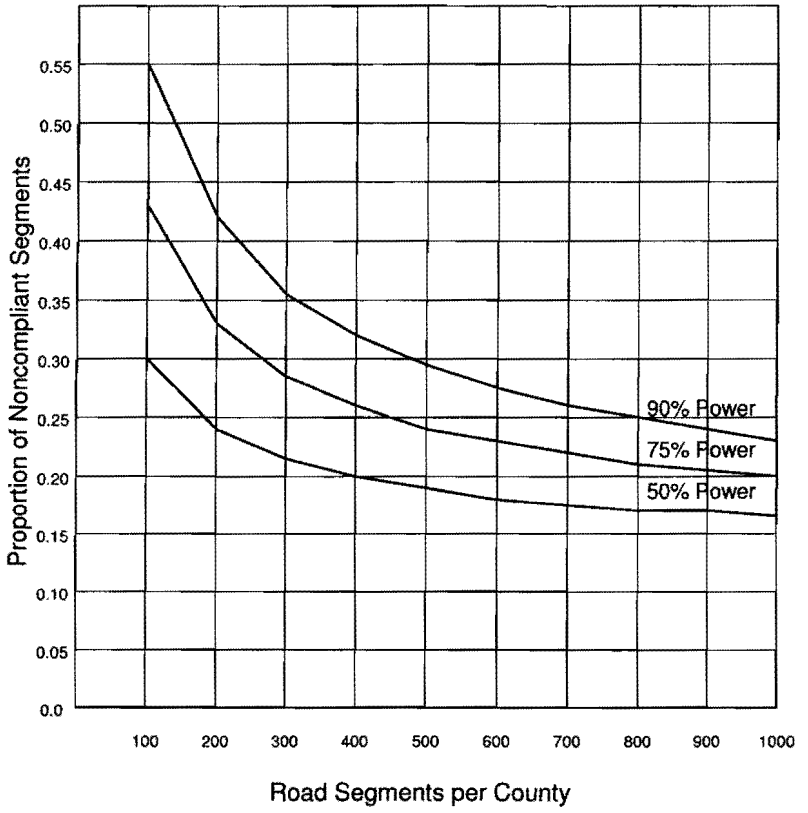

Figure 5. Proportion of Noncompliant Scores Required before Rater Is Rejected (6 Percent Sampling).

2.5 AN ALTERNATIVE TO FIXED PERCENTAGE SAMPLES

We propose an alternative criterion for determining the appropriate sample sizes, based on considerations of Type I error and power. For this method, we choose the sample size n to be the smallest integer such that

$$n \geq \frac{K^2 N}{K^2 + N + 1}, \tag{3}$$

where

$$K = \left(z_\alpha \sqrt{\frac{p(1-p)}{p_A(1-p_A)}} + z_{(1-\beta)} \right) \frac{\sqrt{p_A(1-p_A)}}{p_A - p}. \tag{4}$$

Here, z_α and $z_{(1-\beta)}$ are the appropriate quantiles from a standard normal distribution, corresponding to the desired Type I error (α) and power ($1-\beta$), respectively. For example, suppose we wish to construct a decision procedure which specifies $\alpha = 0.05$ at $p = 0.10$, and $(1-\beta) = 0.9$ at $p_A = 0.20$. The corresponding values for z_α and $z_{(1-\beta)}$ would then be 1.645 and 1.28, respectively.

Table 11 and Figure 6 show the sample sizes required for 90 percent power at four assumed county-wide noncompliance proportions p_A, for county sizes of 100 to 1000 road segments. All sample sizes are calculated using Equations 3 and 4. Table 12 and Figure 7 show the corresponding required audit fractions; note that the required audit fractions for smaller counties are much larger than those required under a fixed 6 percent sampling scheme. The required sample fractions for large counties are less than they would be under a strict 6 percent sampling scheme only if we are satisfied with achieving 90 percent power at a county-wide proportion of noncompliant scores greater than 0.225. Figure 7 indicates that only large sample fractions can ensure adequate power to detect noncompliant raters when the audited counties have less than 300 highway segments.

2.6 APPLICATION TO INSPECTION PROCEDURES

The application of the preceding statistical analysis, when applied to the actual distribution of pavement conditions in Texas, illustrates that as currently implemented a contractor could submit distresses that would result in a Distress Score of 85 for all pavement sections throughout the state ***without inspecting any pavement and could not be penalized.*** This

is because of the high average distress score for Texas pavements and the statistical analysis of the auditing procedures presented in this report. Table 13 lists the percentages of pavement segments within a district that have a Distress Score less than 70.

Under the current procedures, some counties would be identified as being noncompliant. In that case, the contractor could simply re-inspect, or in this case inspect, those counties.

To solve this potential problem, the researchers propose that the procedures identified in the previous section, along with Table 11, be implemented.

Table 11. Number of Audit Segments Required for 90 Percent Power at Specified p_A ($\alpha = 0.05$).

Segments per County (N)	Proportion Noncompliant (p_A)			
	0.200	0.225	0.250	0.300
100	51	41	34	23
200	68	51	40	26
300	76	56	43	27
400	81	59	44	28
500	85	60	45	28
600	87	61	46	28
700	89	62	46	29
800	90	63	47	29
900	92	64	47	29
1000	92	64	47	29

Table 12. Sample Fractions Required for 90 Percent Power at Specified p_A ($\alpha = 0.05$).

Segments per County (N)	Proportion Noncompliant (p_A)			
	0.200	0.225	0.250	0.300
100	0.5100	0.4100	0.3400	0.2300
200	0.3400	0.2550	0.2000	0.1300
300	0.2533	0.1867	0.1433	0.0900
400	0.2025	0.1475	0.1100	0.0700
500	0.1700	0.1200	0.0900	0.0560
600	0.1450	0.1017	0.0767	0.0467
700	0.1271	0.0886	0.0657	0.0414
800	0.1125	0.0788	0.0588	0.0362
900	0.1022	0.0711	0.0522	0.0322
1000	0.0920	0.0640	0.0470	0.0290

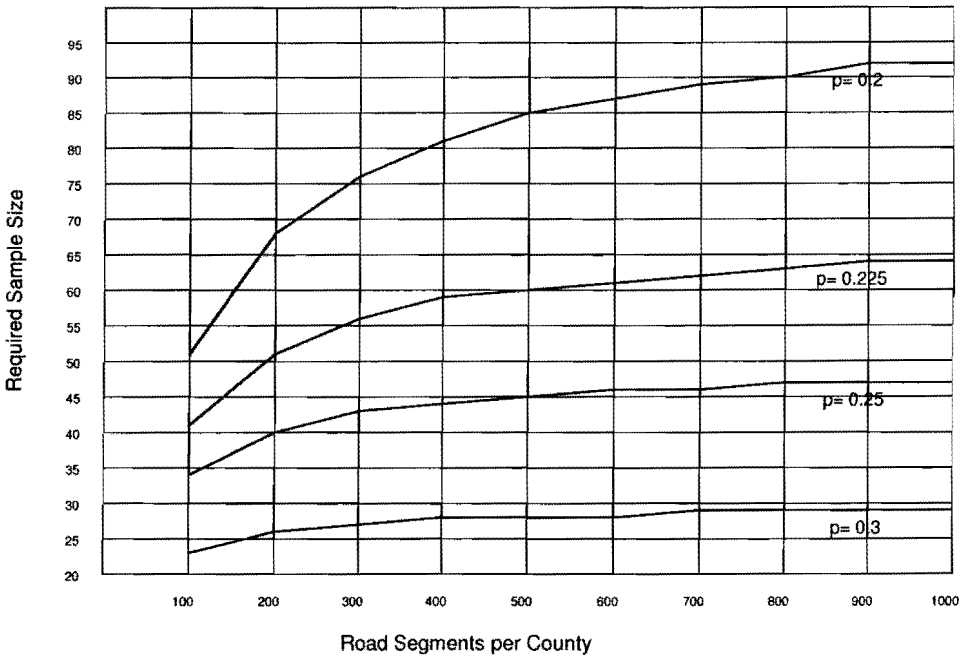

Figure 6. Required Sample Sizes for 90 Percent Power at Specified p_A ($\alpha = 0.05$).

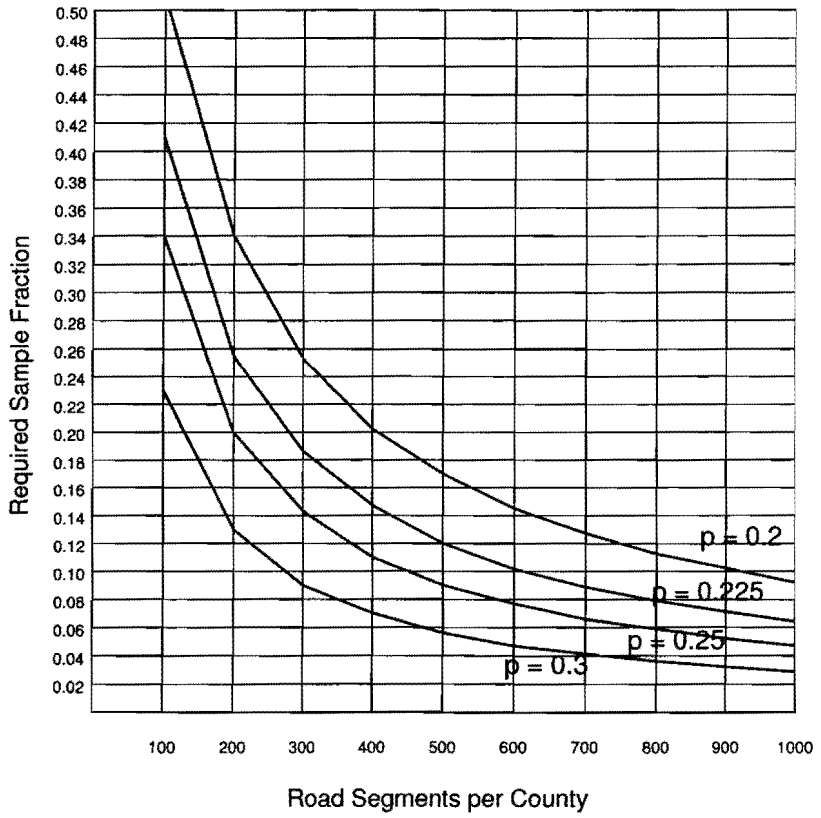

Figure 7. Required Sample Fractions for 90 Percent Power at Specified p_A ($\alpha = 0.05$).

Table 13. Percentage of Segments with Distress Score Less than 70.

District	Fiscal Year			
	2001	2002	2003	2004
Paris	10.2	14.3	12.1	10.8
Fort Worth	7.3	7.7	7.7	7.5
Wichita Falls	5.4	10.8	7.6	7.0
Amarillo	17.4	13.4	17.5	12.9
Lubbock	12.9	14.4	12.9	10.6
Odessa	3.5	3.7	2.7	3.7
San Angelo	6.5	6.6	4.7	3.6
Abilene	7.0	5.9	6.4	7.0
Waco	6.9	9.3	9.8	6.6
Tyler	7.6	8.5	12.0	9.0
Lufkin	12.8	11.8	10.0	8.7
Houston	15.8	20.0	19.9	20.3
Yoakum	13.0	12.1	11.0	9.5
Austin	10.6	15.8	10.6	9.4
San Antonio	11.0	10.9	10.2	12.6
Corpus Christi	14.5	15.3	13.3	13.9
Bryan	12.0	11.6	8.4	11.2
Dallas	26.7	25.1	18.3	14.7
Atlanta	3.5	6.4	4.5	4.9
Beaumont	14.7	16.2	18.2	9.1
Pharr	4.6	6.6	6.6	7.4
Laredo	6.9	11.4	14.5	12.8
Brownwood	5.4	7.3	4.9	3.5
El Paso	9.2	7.0	6.1	7.4
Childress	7.2	6.2	8.6	8.8

CHAPTER 3. CONFIDENCE INTERVALS FOR PAVEMENT CONDITION SCORES

3.1 INTRODUCTION

Each year the Texas Department of Transportation inspects over 100,000 segments of roadway across the state. The results of these inspections are important for both ascertaining the condition of Texas roads and determining the proper allocation of funding for road maintenance. One principle concern related to these measurements is whether the fluctuations observed in average road condition from year to year represent real increases or decreases in this statistic, as opposed to being the result of random measurement errors. A goal of this project is to characterize the variability of mean pavement Distress Scores and develop a statistical framework for evaluating their year-to-year differences.

The most familiar method of reporting measurement uncertainty is the confidence interval, where a statistic is combined with information about its variability to produce an interval estimate. The method of constructing this interval is expected, in repeated application, to include the value of the parameter being estimated by the statistic a certain proportion of the time, termed the confidence level. Such intervals or their counterparts, statistical hypothesis tests, may then be used to determine statistically significant differences between measurements.

We have pursued two approaches to the problem of interval estimation of pavement Distress Scores:

1. parametric confidence intervals, which involve assumptions about the probability distribution of the scores; and

2. bootstrap (or nonparametric) confidence intervals, which are a computational approach and not reliant on distributional assumptions.

We compare the performance of both approaches by applying them to recent road condition distress scores, and demonstrate the methods used for statistical comparison of differences among scores.

3.2 LARGE SAMPLE PARAMETRIC CONFIDENCE INTERVALS FOR MEANS

The confidence intervals discussed here all involve the application of the law of large numbers, which states: given a large enough sample size from which to compute our statistic, the probability distribution of that statistic will converge to that of a normally distributed random variable (Casella and Berger 1990). Thus, for a sample mean $\hat{\mu}$ computed from a sample of size n, a $100(1-\alpha)$ percent confidence interval for the population mean μ is

$$\hat{\mu} \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \tag{5}$$

where

$\hat{\sigma}$ is the estimated standard deviation of the sample and

$z_{\alpha/2}$ is the upper $\alpha/2th$ percentile of a standard normal distribution.

A brief example is useful for understanding the above nomenclature. Suppose we desire a 95 percent confidence interval for μ. This implies that $100(1-\alpha)=95$; solving this equation for α yields $\alpha=0.05$. The value of $z_{\alpha/2}$ will be the percentile of a standard normal distribution for which $P(Z > z_{\alpha/2}) = \alpha/2 = 0.025$, which is 1.96.

As another example, on the Stanford-Binet Intelligence Test, a score of 100 means that 50 percent of all individuals taking the test score below this value. Thus, 100 is the quantile corresponding to $P(X > 100) = 0.5$; alternatively, 100 is the 50^{th} percentile of the distribution of scores on the test. The distinction between quantiles and percentiles is contextual, in that the term quantile is used when probabilities are being discussed (decimal number), and percentiles are used when percents are being discussed (a percentage is the probability multiplied by 100).

We may also readily construct confidence intervals to compare two populations. Let $\hat{\mu}_1$ and $\hat{\mu}_2$ be the sample means obtained from samples of size n_1 and n_2, respectively, and assume that these samples were taken from two independent populations having respective means μ_1 and μ_2. Then a $100(1-\alpha)$ percent confidence interval for the difference in the population means $\mu_1 - \mu_2$ is

$$(\hat{\mu}_1 - \hat{\mu}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}, \tag{6}$$

where the quantities $\hat{\sigma}_1$ and $\hat{\sigma}_2$ are the standard deviations obtained from the two samples so that $\sqrt{\hat{\sigma}_1/n_1 + \hat{\sigma}_2/n_2}$ represents the standard deviation of the estimator $\hat{\mu}_1 - \hat{\mu}_2$.

3.3 RELATION BETWEEN CONFIDENCE INTERVALS AND HYPOTHESIS TESTING

Any $100(1-\alpha)$ percent confidence interval has a corresponding hypothesis test at a significance level (maximum Type I error rate) of α. For determining whether two means are different from each other at significance level α, the appropriate set of hypotheses are

$$H_0: \ |\mu_1 - \mu_2| = 0,$$
$$H_A: \ |\mu_1 - \mu_2| > 0.$$

Under the null hypothesis H_0 (which assumes the means are equal), and provided that n_1 and n_2 are sufficiently large, the statistic

$$Z = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}} \tag{7}$$

is a random variable having a standard normal distribution. Now assume that, for two given samples, we obtain a value for Z, say z^*. The decision rule for choosing between hypotheses will depend on the likelihood of observing, under the null hypothesis, a value of the test statistic Z whose magnitude is as least as large as z^*. Specifically, we will base our decision on the probability $P(|Z > z^*|)$ of observing this, which is referred to as the ***p-value*** of the test statistic. The decision rule for the above hypotheses may then be stated:

> If $P(|Z > z^*|) > \alpha$, reject H_0 and conclude that $|\mu_1 - \mu_2|$ is larger than 0 at significance level α. The means are statistically different.

If $P(|Z > z^*|) \leq \alpha$, fail to reject H_0 and conclude that there is insufficient evidence to disprove the assumption that $|\mu_1 - \mu_2| = 0$ at significance level α. The difference between the means is not statistically different.

The *p*-value may be interpreted as the strength of evidence against the null hypothesis, with *p*-values much smaller than α emphatically disproving H_0 and there is a difference in the means from year to year. Similarly, large *p*-values indicate the lack of evidence against the null hypothesis and there is no trend.

The power of the test, or the probability of rejecting the null hypothesis when H_A is true, may be obtained by

$$P\left(Z > z_{\alpha/2} - \frac{\mu_A}{\hat{\sigma}}\right) + P\left(Z < -z_{\alpha/2} - \frac{\mu_A}{\hat{\sigma}}\right), \tag{8}$$

where

Z is a random variable having a standard normal distribution,

$z_{\alpha/2}$ is the upper $\alpha/2th$ percentile of a standard normal distribution,

$\mu_A = \mu_1 - \mu_2$, $\mu_A \neq 0$ is the mean difference between the two populations, and

$\hat{\sigma} = \sqrt{\frac{\hat{\sigma}_1}{n_1} + \frac{\hat{\sigma}_2}{n_2}}$ is the standard deviation of the estimate for $\mu_1 - \mu_2$.

The operation of a smoke alarm illustrates the nature of Type I and Type II error. The alarm needs to be sensitive enough (have a low enough Type II error rate) such that it will be able to reliably detect a fire. Imagine if the sensitivity of the detector is such that it correctly detects the presence of a fire half the time. Such a detector will probably have a low false alarm (Type I error) rate; i.e., it will not sound when frying food in the kitchen, however, it is clearly unacceptable to have a smoke detector which only "works" half the time. If the sensitivity of the smoke detector is increased so that it can correctly detect a fire 99 percent of the time, it will have a much higher false alarm (Type I error) rate and the fried chicken dinner may be interrupted.

In the case of evaluating raters, a similar trade-off must be made. One must decide whether it is more important to protect raters from the consequences of false alarms, or to ensure that noncompliant raters are reliably identified. Based on this decision, appropriate Type I and Type II error rates can be specified for the tests.

3.4 BOOTSTRAP CONFIDENCE INTERVALS FOR MEANS

Even if the observations are not normally distributed, the distribution of the mean or the proportion will be normally distributed if the sample size is very large (Casella and Berger 1990). This is known as the law of large numbers. If the average distress score is computed from thousands of scores, the distribution of the average distress score can safely be assumed to have a normal distribution. For nonparametric confidence intervals, we can use a method known as bootstrapping to simulate the distribution of our statistic from the observed sample data (Davison and Hinkley 1997). Bootstrap methods are often used for analysis of experiments when sample sizes are small and the data are not normally distributed.

By sampling with replacement from our observed data, whereby we generate a new sample from our data, we can generate a "new" set of data, called a bootstrap replicate or bootstrap sample, and obtain a different value for our statistic. If we repeat this process for a large number B of bootstrap replications, we get a distribution of many different values for our statistic. It is by ordering these values and selecting the appropriate percentiles that we create bootstrap confidence intervals. As an example, if we had 20 data points, we could take the 20 numbers and put them in a hat. A new bootstrap sample would be generated by drawing a number out of the hat (sampling), writing down the number, and then returning that number to the hat (replacement). The hat is re-shuffled and a number drawn again (it could be the same number). This process is continued until a total of 20 numbers are drawn.

The obvious advantage of bootstrap confidence intervals is that they may be obtained readily in many circumstances when a parametric approach is difficult or impossible to implement. Bootstrap intervals may also serve to corroborate parametric results in cases where both methods are applicable, by affording an examination of how closely the assumed and empirical probability distributions of the test statistic match. The disadvantage of the bootstrap

approach lies in the computational difficulties associated with resampling extremely large sample sizes, especially when there are 10^5 observations, such as in the TxDOT PMIS database.

Bootstrap methods may be viewed as an approximation to another nonparametric approach known as permutation methods, which require the calculation of every possible permutation of the sample. For those who may feel uncomfortable with the use of nonparametric methods such as these, note that, historically, permutation methods were the preferred choice. Parametric probability distributions eventually achieved prominence for the primary reasons that they were computationally less demanding than their permutation counterparts and served as excellent approximations for permutation distributions. With the phenomenal improvement in computational resources in recent decades, nonparametric methods are regaining the recognition they enjoyed for the majority of the history of statistics.

3.5 APPLICATION TO DATA FROM ATLANTA AND LUFKIN DISTRICTS

The data shown in Table 14 comprise the pavement Distress Scores for 15 counties audited in both 1999 and 2000. Figure 8 shows the smoothed estimates for the probability densities of each year's Distress Scores. Note that the figure indicates pronounced skewness, with many scores clustering near the maximum value of 100. The densities also exhibit minor modes centered around 70, most likely due to heterogeneity in road conditions across counties. That is, within a county many scores seem to be either excellent or fair. Fortunately, the sample sizes are quite large so that we can be reasonably optimistic that the law of large numbers still holds.

Parametric and bootstrap 95 percent confidence intervals were constructed for the mean distress score for each of the two years, denoted by μ_{1999} and μ_{2000}, as well as their difference, with the results shown in Table 15. All bootstrap intervals were obtained by using $B = 5000$ bootstrap replications. One may see immediately the excellent agreement between the bootstrap and parametric results; most values differ only by a few hundredths of a point. Examining the confidence interval (CI) for the difference in means, we see that the interval includes the value of 0 for both methods. Drawing on the relationship between confidence intervals and hypothesis tests, we would conclude that the difference in mean Distress Scores across the two years is not significant at $\alpha = 0.05$ (null hypothesis is true).

Table 14. Data for Atlanta and Lufkin Districts.

County Number	Total Number of Segments	1999 Audited Segments	2000 Audited Segments
3	872	71	None
32	257	48	9
103	1209	86	62
114	842	71	34
172	325	27	16
174	873	72	49
183	708	28	34
187	838	89	23
202	475	56	24
203	538	81	27
204	505	56	20
210	788	64	29
225	541	38	26
228	439	44	27
230	707	64	33
Total Audited Segments		895	413

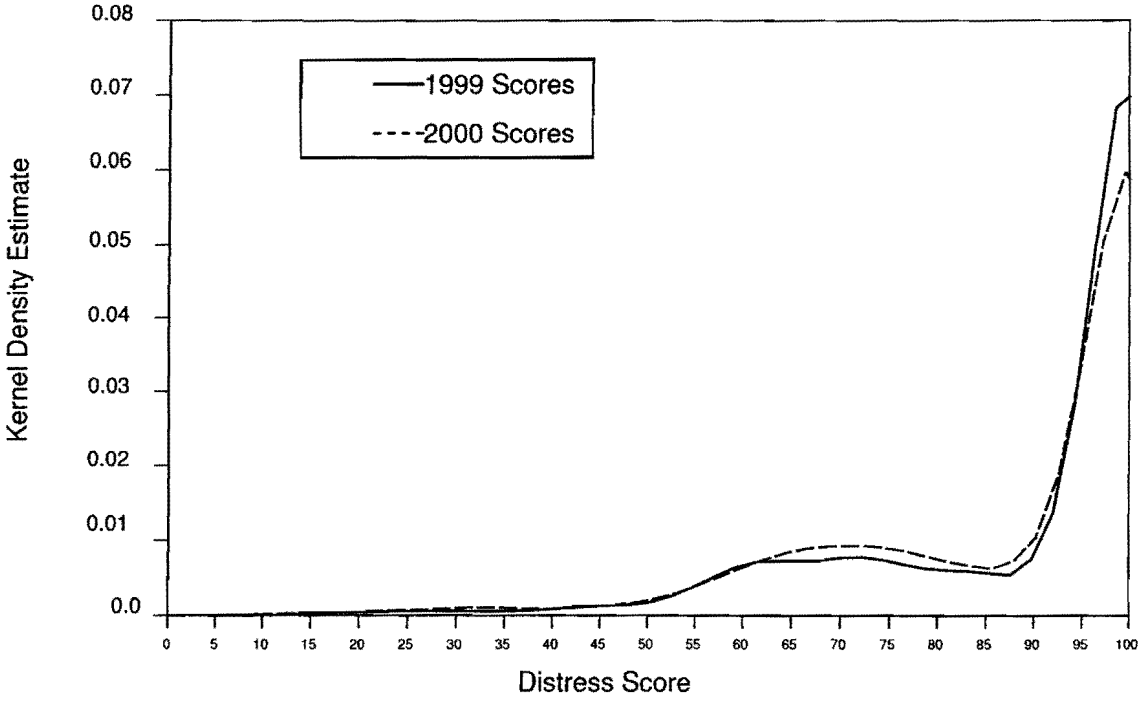

Figure 8. Estimated Probability Density for 1999 and 2000 Distress Scores.

We may better understand the close correspondence between the results of the two methods by considering the plots of the bootstrap density estimates for mean distress scores shown in Figure 9, and the bootstrap density estimate for the difference in mean distress scores shown in Figure 10. The two figures indicate that all three statistics are normally distributed; the distribution for the 2000 data is "fatter" as a result of the smaller sample size. This confirms our initial conjecture that the law of large numbers would assure convergence of the distributions of our statistics to normality.

Hypothesis tests were also conducted at level $\alpha = 0.05$ to determine if the difference in mean distress scores $\mu_1 - \mu_2$ was nonzero, using both parametric and nonparametric methods. Test results are shown in Table 16; we see again that both methods yield very similar values. Both p-values are greater than 0.05, indicating insufficient evidence to reject the null hypothesis of zero difference, and the mean values of the Distress Scores for all pavements in the two districts are not statistically different from one year to the other. The power of the parametric test is displayed graphically in Figure 11, where we see that the test's ability to detect a nonzero difference between mean distress scores is relatively poor for any difference less than ± 2. However, if the difference is greater than 3, we would be able to detect a statistical difference 85 percent of the time. Coincidentally, the calculated standard deviation of the test statistic $\hat{\mu}_1 - \hat{\mu}_2$ is 1.0051, which means that the scale of the x-axis in the figure represents standardized units. Hence, Figure 11 may also be interpreted as the general power curve for tests of differences in means.

Table 15. Parametric and Bootstrap Confidence Intervals for Lufkin and Atlanta Districts.

Parameter	Estimate	95 Percent Parametric CI		95 Percent Bootstrap CI	
		Lower	Upper	Lower	Upper
μ_{1999}	90.9452	89.8818	92.0088	89.86	91.99
μ_{2000}	89.7070	88.0489	91.3652	88.07	91.37
$\mu_{1999} - \mu_{2000}$	1.23823	-0.7317	3.2081	-0.7485	3.233

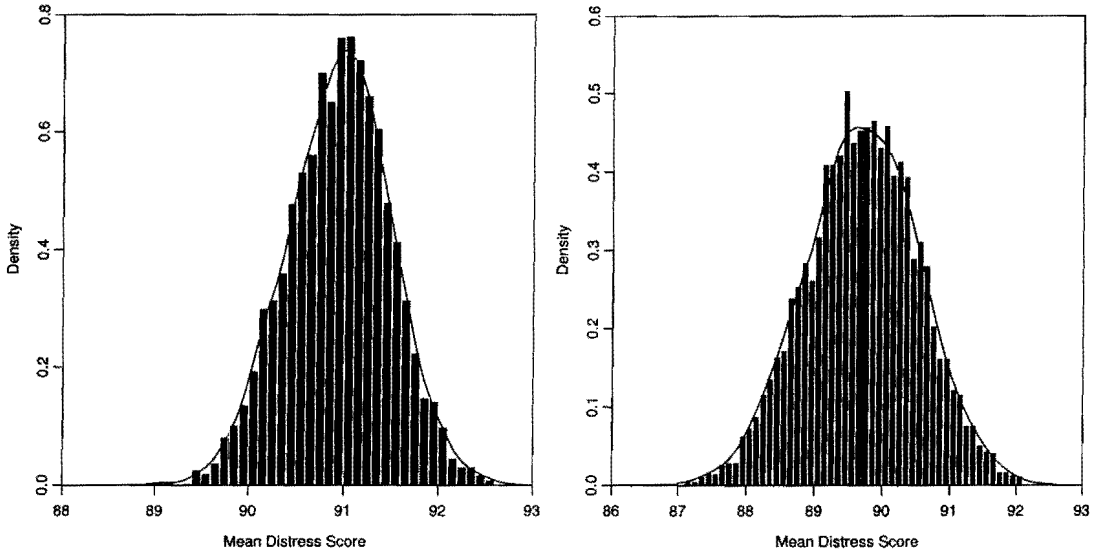

Figure 9. Bootstrap Density Estimates for 1999 (Left Pane) and 2000 (Right Pane) Mean Distress Scores.

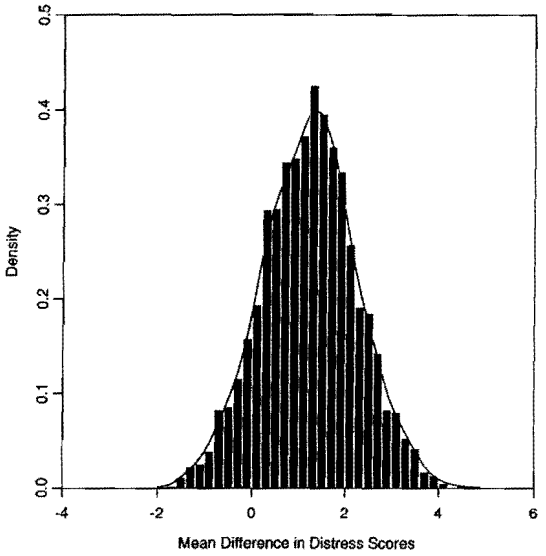

Figure 10. Bootstrap Density Estimate for the Difference between 1999 and 2000 Mean Distress Scores.

Table 16. Hypothesis Test Results for Difference in Mean Distress Scores.

Method	Test Statistic Value	*p*-value
Parametric	1.2320	0.2180
Bootstrap	1.2820	0.1996

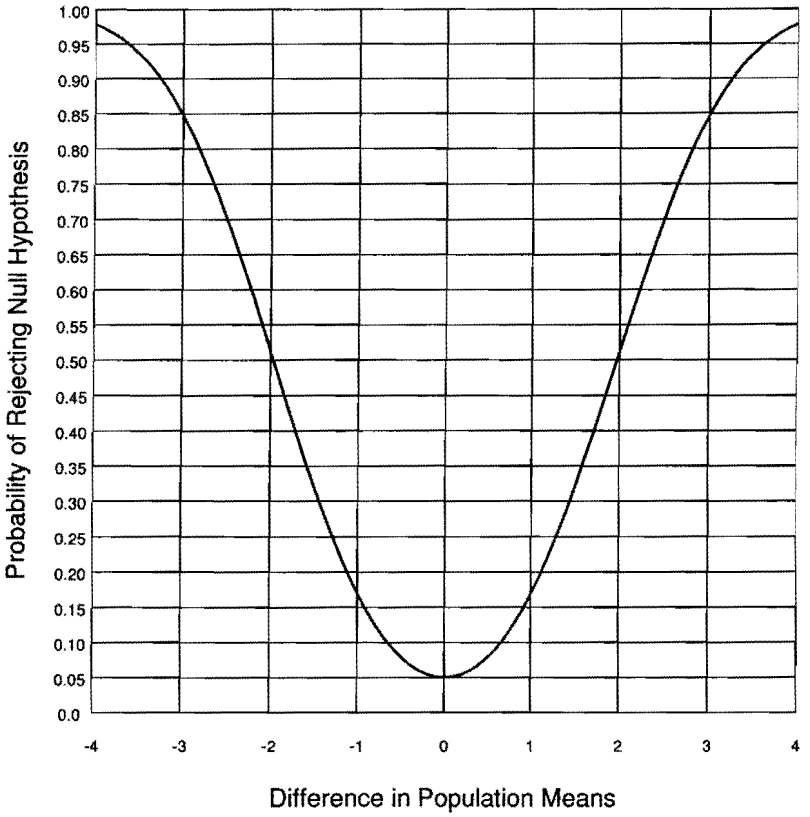

Figure 11. Power of the Test for Differences in Mean Distress Scores.

3.6 APPLICATION TO MEAN DISTRESS SCORES FOR ALL DISTRICTS

We next examined Distress Scores from 1999 for all 25 districts, totaling 149,504 rated road segments. The distributions of the distress scores varied across districts, as can be seen from the smoothed density estimates in Figure 12. The three districts shown in the figure are representative of the types of distributions of distress scores observed. Distress Scores for the Houston District, for example, have a highly skewed, unimodal distribution, indicating that most road segments across all counties in the district had high distress scores. The elongated left tail of the distribution indicates that isolated road segments in poorer condition could be found scattered across the district as well. The distribution for the Wichita Falls District, however, shows that while the majority of distress scores were 90 or above, there was a significant minority of scores in the range of 70-80. This would seem to indicate that there are identifiable subgroups of highway segments that differ in their mean distress scores, perhaps due to recent road reconstruction or different pavement types. The San Angelo District presents the most

extreme example of multimodality; here we see that highway segments are divided between those with distress scores above 90 and those whose distress scores are centered around 82. There is also a substantial minority of segments having a Distress Score in the 50-70 range.

95 percent confidence intervals were constructed for mean pavement Distress Score in each district, with the results tabulated in Table 17. All bootstrap intervals were again constructed from $B = 5000$ replicates. Despite the variety of distributions of distress scores, we could be assured of the appropriateness of employing large-sample parametric confidence intervals because the data for each district contained at least 604 road segments. Indeed, as in the case for the data from the Lufkin and Atlanta Districts in Section 3.5, we observe close agreement between the parametric and bootstrap results. The large sample sizes ensure reasonably precise interval estimates, with most intervals having widths less than one.

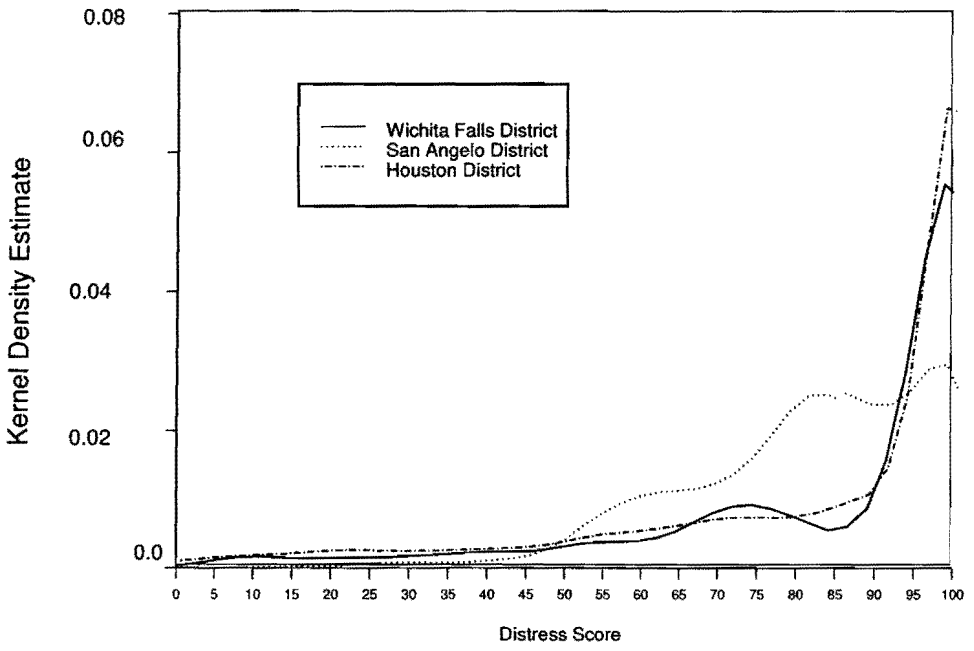

Figure 12. Estimated Probability Densities for Distress Scores in Selected Districts.

We may also observe a rough correspondence between sample size n and the widths of the confidence intervals in Table 17. Recall that from Equation 5 the width of an interval is inversely proportional to the square root of the sample size. For example, the ratio of widths of the parametric intervals for the Wichita Falls and Lubbock Districts is approximately 6 (that is

the width of the interval for Wichita Falls is 6 times the size as that for Lubbock), while the ratio of the square roots of sample sizes for the two districts is 4.2.

Table 17. Parametric and Bootstrap Confidence Intervals for Mean Distress Scores.

District	Segments Rated (n)	Mean ($\hat{\mu}$)	95 percent Parametric CI		95 percent Bootstrap CI	
			Lower	Upper	Lower	Upper
Paris	6725	92.3526	91.9947	92.7104	92.0101	92.7151
Fort Worth	3614	93.4427	92.9183	93.9671	92.8953	93.9715
Wichita Falls	604	86.5861	84.8099	88.3623	84.7875	88.3282
Amarillo	9170	86.9662	86.5810	87.3513	86.6031	87.3923
Lubbock	10651	90.8091	90.5123	91.1060	90.5184	91.0987
Odessa	6444	96.5663	96.3409	96.7917	96.3278	96.7851
San Angelo	6183	81.3962	81.0020	81.7905	81.0104	81.7964
Abilene	8350	94.2783	94.0217	94.5350	94.0204	94.5311
Waco	6582	94.3177	94.0231	94.6123	93.9869	94.5976
Tyler	1030	91.9573	91.0970	92.8176	91.0340	92.7806
Lufkin	3386	92.4259	91.9473	92.9044	91.9079	92.8686
Houston	6780	87.5618	87.0640	88.0596	87.0737	88.0732
Yoakum	7341	88.0672	87.6704	88.4639	87.6698	88.4398
Austin	6199	87.2376	86.8373	87.6380	86.8166	87.6395
San Antonio	9928	91.0626	90.7544	91.3707	90.7381	91.3606
Corpus Christi	6401	91.0987	90.6912	91.5063	90.6904	91.5019
Bryan	6692	91.2862	90.8712	91.7011	90.8742	91.7010
Dallas	8812	80.3740	79.8098	80.9383	79.8175	80.9537
Atlanta	5747	95.5634	95.3358	95.7911	95.3134	95.7728
Beaumont	3697	88.4266	87.7458	89.1073	87.7278	89.0756
Pharr	4990	95.8060	95.5187	96.0934	95.4989	96.0775
Laredo	4667	93.4131	93.0609	93.7653	93.0352	93.7680
Brownwood	5652	95.1515	94.8796	95.4233	94.8939	95.4303
El Paso	4494	93.0120	92.5941	93.4300	92.5691	93.4058
Childress	5365	92.1940	91.8775	92.5105	91.8571	92.5055

We also constructed a 95 percent parametric confidence interval for the overall mean Distress Score, which is shown in Table 18. The extremely large number of observations precluded the generation of bootstrap intervals although this difficulty could be ameliorated by performing computations on a computer with a fast processor and at least 512 megabytes of

RAM. Note that the very large sample size has resulted in a very precise interval estimate, with the width of the confidence interval being less than 0.2. That is, if the mean Distress Score varies from year to year by more than 0.35, this difference is statistically significant.

Table 18. Parametric Confidence Interval for Overall Mean Distress Score.

Segments Rated (n)	Mean ($\hat{\mu}$)	95 Percent Parametric CI Lower	Upper
149,504	90.5986	90.5141	90.6831

From this we can see that there is no statistical reason for the scores to vary back and forth from year to year. Therefore, these differences represent either a true physical difference where the actual condition of the pavement is changing from year to year or some discrepancy in the data collected.

An analysis was conducted of statewide data from 1999 through 2004 which showed that year to year changes were all statistically significant at an experiment wise significance level of 0.005. This is due to the very large number of observations. While all results were statistically significant, the range of values was 91.309 (2002) to 92.734 (2004) which was a range of 1.425.

CHAPTER 4. SUMMARY AND CONCLUSIONS

Each year the Texas Department of Transportation inspects over 100,000 segments of roadway across the state. Except for mileage under construction, or otherwise identified, all of the mileage is inspected. The results of these inspections are important for both ascertaining the condition of Texas roads and determining the proper allocation of funding for road maintenance.

Prior to the start of the inspection season, contractor, TxDOT, and TTI personnel are required to attend training classes to become certified inspectors for the calendar year. Changes to the manual, including clarifications and interpretations, and inspections of selected sections are used in order to reduce the rater-to-rater variability between inspection teams.

The current method of certifying qualified raters for PMIS inspections can be improved by adopting the recommendations to:

- audit from one RM to the next RM,
- eliminating the inclusion of short sections,
- not including pavements that were in poor condition during the last inspection,
- increasing the percentage of sections that are audited to a variable percentage based on the number of sections in a county, and
- requiring raters to rate pavement as part of the certification process.

Most of these suggestions were immediately implemented, but others are still under consideration.

In an effort to standardize data collection, reduce variability between districts, reduce the number of raters, complete the work more quickly, and to free-up district personnel, TxDOT hired contractors to perform this intensive data collection effort and has TxDOT and personnel from TTI check the results by performing an audit survey of 6 percent of the pavements. The original TxDOT criterion defines the rating of a county to be noncompliant if more than 10 percent of the Distress Scores for road segments in that county differ by more than 15 points from the corresponding audit Distress Scores.

In our analysis of two districts, the probability of detecting a noncompliant rater, given that their proportion of noncompliant scores across a county is 0.15, is less than 50 percent for all but one of the counties. However, the ability to reliably detect noncompliance only emerges

when a raters' county-wide proportion of noncompliant scores is at least twice the currently allowable rate of 10 percent.

The primary impact of this analysis is that under the current audit procedures, we are unjustified, statistically speaking, to reject a contractor's rating of a county unless the percentage of noncompliant scores is much, much higher than the current 10 percent criteria. The application of the preceding statistical analysis, when applied to the actual distribution of pavement conditions in Texas, illustrates that as currently implemented a contractor could submit distresses that would result in a Distress Score of 85 for all pavement sections throughout the state ***without inspecting any pavement and could not be penalized.*** This is because of the high average distress score for Texas pavements and the statistical analysis of the auditing procedures presented in this report.

We propose an alternative criterion for determining the appropriate sample sizes, based on statistical considerations. The analysis shows that the required audit fractions for smaller counties are much larger than those required under a fixed 6 percent sampling scheme. The required sample fractions for large counties are less than they would be under a strict 6 percent sampling scheme only if we are satisfied with achieving 90 percent power at a county-wide proportion of noncompliant scores greater than 22.5 percent.

Another principle concern of this research was to determine whether the fluctuations observed in average road condition from year to year represent real increases or decreases in this statistic, as opposed to being the result of random measurement errors.

The very large sample size has resulted in a very precise interval estimate, with the width of the confidence interval being less than 0.2. That is, if the mean Distress Score varies from year to year by more than 0.35, this difference is statistically significant.

REFERENCES

ASTM, *Standard Practice for Use of the Terms Precision and Bias in ASTM Test Methods*, E 177-90a, ASTM, Philadelphia, PA, 1992.

Cable, J. K., and V. J. Marks, *Automated Pavement Distress Data Collection Equipment Seminar*, FHWA-TS-90-053, Iowa DOT, Ames, IA, and Federal Highway Administration, Washington, D.C., 1990.

Casella, G., and R. L. Berger, *Statistical Inference,* Duxbury Press, Belmont, CA, 1990.

Davison, A. C., and D. V. Hinkley, *Bootstrap Methods and Their Application*, Cambridge University Press, 1997.

Epps, J. A., and C. L. Monismith, *Equipment for Obtaining Pavement Condition and Traffic Loading Data*, NCHRP Synthesis 126, Transportation Research Board, Washington, D.C., 1986.

FHWA, *Pavement and Road Surface Management for Local Agencies*, Participant's Manual, Federal Highway Administration, Washington, D.C., 1995.

Hicks, R. G., and J. P. Mahoney, *Collection and Use of Pavement Condition Data*, NCHRP Synthesis 76, Transportation Research Board, Washington, D.C., 1981.

Montgomery, D. C., *Introduction to Statistical Quality Control*, 2nd ed., John Wiley, NY, 1991.

Smith, R. E., T. J. Freeman, and O. J. Pendelton, *Evaluation of Automated Pavement Distress Data Collection Procedures for Local Agency Pavement Management*, Texas Transportation Institute, Texas A&M University, College Station, TX, 1996.

TxDOT, *Test and Evaluation Project No. 21, Automated Pavement Distress Survey Equipment, November 15-19, 1993, Austin, Texas*, Draft Final Report, Texas Department of Transportation and Federal Highway Administration, Austin, TX, undated.

TxDOT, *Pavement Management Information System Rater's Manual, Fiscal Year 2003*, Texas Department of Transportation, Austin, TX, 2002.