Development of a Statistical Decision-Making Framework for the Field Evaluation of Any Automated Pavement Distress Measuring Device

HOSIN LEE, JOE P. MAHONEY, NEWTON C. JACKSON, AND KEITH KAY

A pavement condition survey is one of the most essential elements of any pavement management system. During the last decade, significant progress has been made towards automating pavement distress survey procedures by use of advanced imaging technology without human interference. The procedure of reducing the pavement video image into quantifiable distress data involves a number of steps that are difficult to quantify. The image processing algorithms adopted for the pavements always need to be calibrated using the field data for various field conditions. Currently, there is no simple and objective statistical procedure available for equitably evaluating various automated devices on the basis of resulting data. A basic statistical measurement model for estimating errors of measurement addresses, at least, the accuracy and precision of automated distress measuring devices. Factors that must be considered in any statistical decision-making process include paired versus individual measurements, threshold versus stipulated level of significance concept, and Type I versus Type II errors. Because the statistical measurement model described is general, it can be applied for evaluating any automated pavement distress measuring device.

A pavement distress survey is one of the most essential elements of any pavement management system. Pavement distress surveys usually involve visual examinations of pavement surfaces. The information collected from this type of survey can be used to document the overall condition of a pavement network, evaluate the performance of individual pavements, and help determine appropriate rehabilitation and maintenance strategies.

In the past, pavement distress data were normally recorded on various forms and transferred to a central file for storage and future use. Techniques to automate collection, storage, retrieval, and analysis of these data are being developed and refined. Several devices that use advanced technology such as video imaging have been developed and are being considered for adoption by various highway agencies. The procedure of reducing the video image of pavements into meaningful distress data involves a number of heuristic steps that seem difficult to analyze.

In recent years, several research projects were conducted for evaluating several automated pavement data collection equipments (*1–3*). Some previous studies applied subjective criteria for evaluating the operational aspects of the equipment such as reliability, consistency, repeatability, usefulness, and ease of data processing (*4*). However, they did not provide an objective procedure for evaluating the resulting data. One study has recently demonstrated the difficulties in evaluating the equipment performance and identifying the presence and magnitude of various types of error involved with the automated distress data collection process (*5*). There exists a need for a simple objective procedure that can be used for evaluating any automated device used to conduct pavement distress surveys on an equitable basis.

The problem addressed in this study is to compare a new automated distress measuring device against the existing one. There is no theoretically absolute values against which the new device can be calibrated. The main objective is to present a statistical decision-making framework that can be used for evaluating any instrument for collecting pavement distress data. Statistical problems faced when evaluating new methods and equipment against the existing ones are addressed.

Estimating errors of measurements are emphasized. A simple and objective statistical model is described. The selected automated distress measuring device for this research, Pavedex PAS 1, and its field data collection procedures and statistical analysis results are described by H. Lee in a companion paper in this Record. Statistical considerations necessary for evaluating any automated distress measuring device are discussed.

MOTIVATION

For highway agencies such as Washington State Department of Transportation (WSDOT) that have conducted visual pavement distress surveys for some time, a switch to a new system could increase the cost and may create compatibility problems with the previously collected historical data. Any such state highway agency (SHA) that is satisfied with its currrent distress survey procedure would likely consider a switch only if a substantial increase in accuracy and precision in distress data would result from the new measuring system.

However, the situation could be the opposite, such that a certain SHA is not happy with its current distress survey procedure and is looking for a new method of conducting distress surveys. Any such SHA would consider a switch even if a

H. Lee, Department of Civil and Environmental Engineering, Washington State University, Pullman, Wash. 99164-2910. J. P. Mahoney, Department of Civil Engineering, University of Washington, Seattle, Wash. 98195. N. C. Jackson and K. Kay, Materials Laboratory, Washington State Department of Transportation, Olympia, Wash. 98504.

small increase in accuracy and repeatability in data would result from the new pavement distress measuring system.

Different situations currently existing in each SHA would determine the magnitude of the errors that the agency would be willing to tolerate in the statistical decision-making process. The statistical evaluation procedure for a new distress measuring system should be designed to minimize the overall decision errors with respect to either adopting or rejecting a new system.

Usually, the sample data are collected from the field to aid the highway engineer in any decision-making process. The statistical parameters computed from the sample data are to be used to test any decision hypothesis. The hypothesis to be tested in this research project is simply whether the new automated distress survey system can measure the true distress of the pavement as well or better than the old system.

PAIRED VERSUS INDIVIDUAL MEASUREMENTS

Two sets of pavement distress data were collected from the same pavement sections located around the city of Spokane, Washington, by both the Pavedex PAS 1 device and the hand-mapping procedure for this research (H. Lee in a companion paper in this Record). The hypothesis-testing procedure adopted for the research used the differences between paired measurements instead of individual measurements. The two groups of measurements, one collected by the Pavedex PAS 1 device and the other by hand-mapping procedures, could not be treated as independent because they were collected from identical sections. Therefore, paired-comparison methodology seemed to be a logical choice for this study.

Furthermore, pairing would control the effects of the extraneous factors that could cause significant differences in means and variances. Pairing will reduce the variance introduced by the extraneous factors and thereby increase the power of the statistical test. By using matched pairs, variability of distress amount among a number of sample pavement sections can be made as small as possible (*6*). Another advantage of pairing includes placing of fewer restrictions on the data sets such that the samples need not be independent and two separate populations need not be assumed to have equal variance (*7,8*).

STATISTICAL MEASUREMENT MODEL

Any measurements are composed of two parts; one is the true value and the other is the measurement error. In general, it is difficult to separate the measurement error from the true value (*9*). For example, if the amount of longitudinal cracking that is measured by Pavedex PAS 1 device is, say, 15 ft, then it is not known what part of 15 ft is the error of measurement. With such a single measurement, it is impossible to separate an error of measurement from the true value. Precision can be defined as a measure of the variation in the errors of measurement (*10*). The automated distress measuring device would possess the utmost in precision if all the errors of measurement were zero (or equal).

A problem in which two instruments are used simultaneously to make measurements has been extensively discussed by several authors (*9,11,12*). A technique for separating measurement error from the true value has been presented and significance tests for these estimators were provided (*13,14*). Such a technique was applied to evaluating an automated pavement distress measuring device.

Assume that two measurements made on each pavement section by two different methods are denoted by X_i and Y_i, where the subscript i stands for the ith pavement section in a random sample. Let b_x and b_y represent the average bias errors for distress measuring devices X (new automated device) and Y (hand-mapping procedure), respectively. The t_i value is the true but unknown distress value of the ith pavement section and e_{xi} and e_{yi} are measurement errors and are assumed to be independent and normally distributed with mean zero. The basic statistical model can be written (*11*) as

$$X_i = t_i + b_x + e_{xi}$$

$$Y_i = t_i + b_y + e_{yi}$$

In comparing these two distress-measuring procedures, two tests are of interest; one concerns the magnitude of the biases b_x and b_y, and the other is random measurement errors e_{xi} and e_{yi}. The unbiased estimate of the difference $(b_x - b_y)$ is the difference between the average of all the measurements by procedures X and Y. If the average bias of $(b_x - b_y)$ is equal to zero, then the errors of measurements average out to zero and the average observed value would be accurate; but on the other hand, the measurements may vary considerably from one measurement to another. In order to determine the precision of the measurement, the amount of variation in the quantities of e_{xi} and e_{yi} should be examined (*9*).

In order to estimate the accuracy of an automated distress measuring device the paired t test can be performed on the differences between measurements by two different distress measuring procedures. The theoretical sampling distribution of the differences may be assumed to be a t distribution with a mean of zero and standard deviation that is the estimated standard error of the difference (*7*). For hypothesis test of bias errors, the paired t statistic can be used to determine the probability of Type I errors by finding the area in the tail of the t distribution.

By making repeated measurements from the same pavement section many times, it can be assumed that there is no bias in the measurements. This repeated sampling practice would allow estimating standard errors directly, which is defined as the precision of the automated distress measuring device. The problems of significance tests for judging whether $e_{xi} - e_{yi}$ values differ significantly from zero need to be further investigated. For this research, a coefficient of variation that was based on the repeated measurements by the automated distress measuring device was adopted for precision definition. The results from the analysis of variance using various sets of distress measurement data are presented by H. Lee in a companion paper in this Record. The following sections discuss the basic statistical decision-making concepts that must be considered (but are often avoided) in most statistical decision-making procedures.

THRESHOLD VERSUS STIPULATED LEVEL OF SIGNIFICANCE

A hypothesis can be either accepted or rejected depending on the level of significance (α) selected. The question is how

close the sample average should be to be considered as the same as the true value. The acceptance region for this true hypothesis will be determined by the selected level of significance. Therefore, any hypothesis test should be properly designed through consideration of different types of and a tolerable amount of errors based on different levels of significance.

A natural question to ask at this point is, what level of significance should be used? Common values that are most often used in practice are 0.01, 0.05, and 0.10, but the rationale for the selection of a particular level of significance is not usually discussed. For example, the level of significance of 0.01 was used for evaluating nuclear density gauges (*15*). What if the level of significance was set at 0.05 instead of 0.01? There is no one universal answer to the question of what level of significance should be used. Indeed, different individuals including statisticians will select different levels of significance for the same hypothesis test (*16*).

One way to answer this question is to compute the threshold level of significance that will just barely cause the hypothesis to be rejected. Normally, in the past, the hypothesis tests were conducted in three steps:

1. Determine the level of significance (say, 0.1),
2. Compute the sample statistics, and
3. Reject (or accept) the hypothesis.

This procedure is simple and straightforward. But, what if the level of significance was set at 0.05 instead of 0.1? Then, the procedure must be repeated for any different level of significance. For example, suppose the level of significance was set at 0.1, the test was performed, and the hypothesis was rejected. Suppose the threshold level of significance that would just barely cause the hypothesis to be rejected is then computed as 0.02. This means that even if the level of significance had been set at 0.05 instead of 0.1, the hypothesis would still have been rejected. However, if the level of significance had been set at 0.01, then the hypothesis should have been accepted.

Instead of determining whether the hypothesis should be rejected for the stipulated level of significance, the threshold level of significance was computed and used throughout this research. The threshold level concept seems more complicated than the stipulated level of significance concept. But, this approach allows the highway engineer to make the final decision through easily asking what-if questions with regard to different levels of significance. Furthermore, this computed threshold level of significance will provide the highway engineer with the additional insight into how strong a conviction should be held with regard to his decision (*16*).

TYPE I VERSUS TYPE II ERRORS

As previously discussed, the hypothesis testing method is to choose a decision procedure so that there would be a minimum probability of making an incorrect decision (*17*). There are two ways that errors can be made in evaluating the new automated device. For example, WSDOT could conclude that the new device (Pavedex) does not meet the accuracy requirements when, in fact, it does. In some cases, WSDOT can conclude the Pavedex device meets the accuracy requirements when, in fact, it does not. The former is the Pavedex's risk (called Type I error), and the latter is WSDOT's risk (called Type II error). Because making an incorrect decision is always possible, it becomes desirable to compute the probabilities associated with those errors.

The probability of rejecting the null hypothesis when it is true (Pavedex's risk) is denoted by α, and the probability of accepting the hypothesis when it is false (WSDOT's risk) is denoted by β. In other words, α is the probability of Type I error occurrence, and β is the probability of Type II error occurrence (*17*); α is often called the level of significance.

The Type I and Type II errors are typically shown in Figure 1. Further, the distribution of both the null hypothesis and the alternate hypothesis are schematically shown in Figure 2. The area with horizontal lines in Figure 2 represents the probability of Type I error, and the area with vertical lines represents the probability of Type II error. As shown in Figure 2, Type I error can be made only when the null hypothesis is true as determined from the probability distribution of the null hypothesis.

The Type II error can be made only when the null hypothesis is false, and requires the probability distribution of the alternate hypothesis. The probability of Type II error is difficult to determine because it only occurs when the true mean is different from the mean of the null hypothesis. Therefore, the probability of Type II error varies with the magnitude of that difference. As shown in Figure 2, the probabilities of Type I and Type II errors are in general inversely related, i.e., the larger the probability of Type I error (α), the smaller the probability of Type II error (β) (*17*).

Assume that the hypothesis that there is no significant difference between two sets of measurements of longitudinal cracking can be rejected with the probability of Type I error

		Decision: Accept Null Hypothesis	Decision: Reject Null Hypothesis
Truth: Null Hypothesis		correct decision	Type I error
Truth: Alternate Hypothesis		Type II error	correct decision

FIGURE 1 Two types of errors in hypothesis testing.

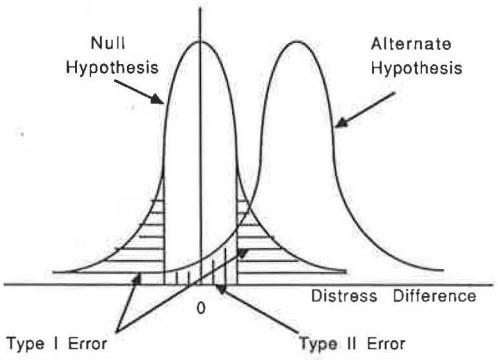

FIGURE 2 Distribution of the null and alternate hypotheses.

of 0.13. In other words, on the basis of the sample data, the longitudinal cracking data measured by the new automated device are different from the data measured by the hand-mapping procedure and a Type I error could be made with approximately 13 percent probability (i.e., this conclusion will be wrong 13 times out of 100). If this probability of Type I error seems too high, then accept the hypothesis that the Pavedex device is as good as the hand-mapping procedure with respect to longitudinal cracking.

Further, additional checks can be made with regard to the hypothesis test. If the null hypothesis is to be accepted, then the Type II error needs to be determined because this error is the risk the WSDOT takes when accepting a false hypothesis. However, the probability of a Type II error is difficult to determine because the probabilities vary with the mean value of the alternate hypothesis.

However, if the mean of the alternate hypothesis is known, the probability of Type II error can be obtained by computing the area under the distribution of the alternate hypothesis for the acceptance region as previously determined on the basis of the stipulated level of significance. As an example, for longitudinal cracking measurements, assume that the hypothesis was accepted at the stipulated significance level of 0.10 (say, 10 percent probability of making Type I error allowed by WSDOT). The mean value of the alternate hypothesis is assumed to be the 10-ft difference in measuring longitudinal cracking. The Type II error can be computed by finding the area under the t distribution function of the alternate hypothesis between the acceptance boundaries for the 0.10 level of significance.

As shown in Figure 3, the computed Type II error is 0.32. This Type II error seems too high; thus the decision to accept the null hypothesis at the level of significance of 0.10 should be cautioned. However, this Type II error can be reduced by

1. Assuming the mean value of the alternate hypothesis would be more different from that of the original hypothesis (e.g., assuming 15 ft of difference instead of 10 ft);
2. Increasing the stipulated level of significance (e.g., increasing the level of significance up to the threshold α of 0.13 instead of 0.10); and
3. Increasing the number of samples.

For example, if the mean value of alternate hypothesis is assumed to be 15 ft instead of 10 ft, then the probability of Type II error is significantly reduced to 0.06 from 0.32, as shown in Figure 4. In general, it is difficult to accurately estimate the probability of Type II error, and it could be reduced to an acceptable level as needed. Therefore, only Type I error is considered in this study, like most other statistical decision-making procedures used in practice.

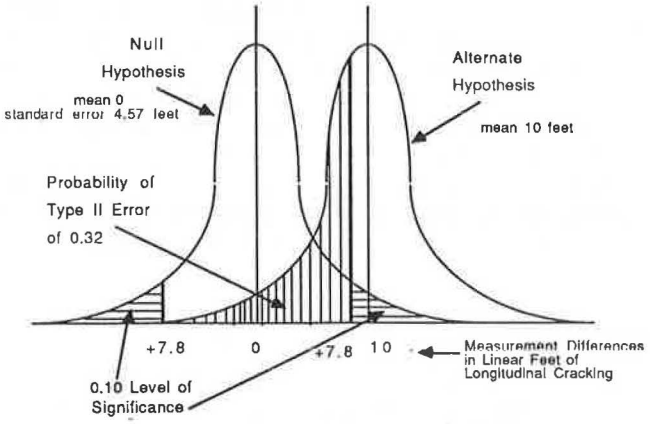

FIGURE 3 Probability of Type II error of 0.32 for longitudinal cracking with mean of 10 ft for alternate hypothesis.

SUMMARY AND CONCLUSIONS

The purpose is not to evaluate a specific pavement distress measuring device about its operating characteristics but to present a simple and objective statistical decision-making framework for evaluating the relative accuracy and precision of an automated distress-measuring device. The presented statistical model is general, and can be applied for evaluating any automated distress-measuring device. The rationale for selecting the most appropriate statistical model for the problem has been discussed.

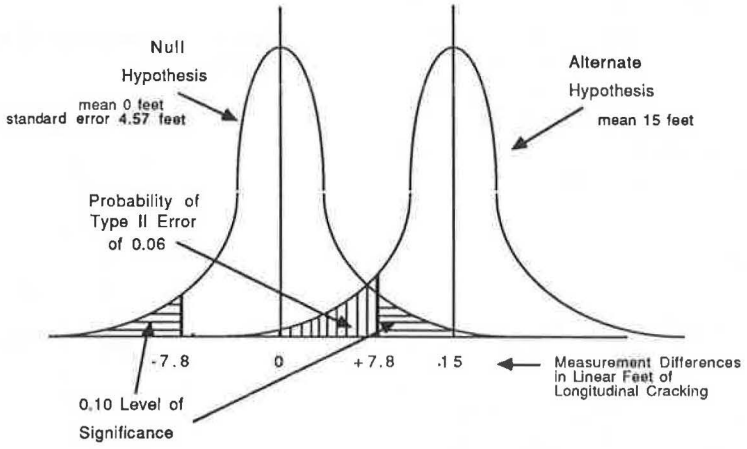

FIGURE 4 Probability of Type II error of 0.06 for longitudinal cracking with mean of 15 ft for alternate hypothesis.

There always exist some measurement errors in any distress-measuring device. The overall decision as to whether the measurement error is significant or not depends on factors such as needs in the agency for the new survey method, the type and amount of errors allowed, and the validity of sampling and data collection procedures. A valid statistical decision-making framework that can be used for evaluating the accuracy and precision of any pavement distress measuring device has been demonstrated.

Several devices that use advanced technology such as video imaging have been developed and are being considered for adoption by many highway agencies. However, the procedure of reducing the video image data into meaningful distress data involves a number of steps that are difficult to quantify. The heuristic image-processing algorithms adopted for the pavements must be calibrated using the field data for various field conditions. Therefore, the statistical decision-making framework described is timely and appropriate, and addresses an important problem facing highway engineers who must evaluate alternate procedures for equitably evaluating various types of automated distress-measuring device.

ACKNOWLEDGMENTS

The authors are pleased to acknowledge the combined efforts and support of the Washington State Transportation Center at University of Washington and Washington State University, and the Washington State Department of Transportation in cooperation with FHWA.

REFERENCES

1. D. G. Richardson. *Evaluation of the Automatic Road Analyzer "ARAN" for Measuring Roughness and Rut Depth*. FHWA-DP-88-072-007. FHWA, U.S. Department of Transportation, May 1988.
2. J. Stone. *Evaluation of the Laser Road Surface Tester for Measuring Pavement Roughness and Rut Depth*. FHWA-DP-88-072-008. FHWA, U.S. Department of Transportation, May 1988.
3. K. Jeyapalan, J. K. Cable, and R. Welper. Automated Pavement Data Collection Equipment. In *Iowa DOT Evaluation of the PASCO Road Survey System*, FHWA-DP-72-2, FHWA, U.S. Department of Transportation, April 1987.
4. K. R. Benson, G. E. Elkins, W. Uddin, and W. R. Hudson. Comparison of Methods and Equipment to Conduct Pavement Distress Survey. In *Transportation Research Record 1196*, TRB, National Research Council, Washington, D.C., 1988.
5. F. Humplick and S. McNeil. Evaluation Errors in Pavement Surface Distress Data Acquisition Using Optical Techniques. *Proc., 1st International Conference on the Application of Advanced Technologies to Transportation Engineering*, San Diego, Calif., Feb. 1989.
6. H. F. Smith. Estimating Precision of Measuring Instruments. *Journal of the American Statistical Association*, Vol. 45, Sept. 1950, pp. 447–451.
7. C. T. Clark and L. L. Schkade. *Statistical Analysis for Administrative Decisions*, 3rd ed., South-Western Publishing Co., Cincinnati, Ohio, 1979.
8. B. Ostle and L. C. Malone. *Statistics in Research: Basic Concepts and Techniques for Research Workers*, 4th ed., Iowa State University Press, Ames, 1988.
9. F. E. Grubbs. On Estimating Precision of Measuring Instruments and Product Variability. *Journal of the American Statistical Association*, Vol. 43, June 1948, pp. 243–264.
10. W. G. Cochran. Errors of Measurement in Statistics. *Technometrics*, Vol. 10, No. 4, Nov. 1968, pp. 637–666.
11. F. E. Grubbs. Errors of Measurement, Precision, Accuracy and the Statistical Comparison of Measuring Instruments. *Technometrics*, Vol. 15, Feb. 1973, pp. 53–66.
12. G. J. Hahn and W. Nelson. A Problem in the Statistical Comparison of Measuring Devices. *Technometrics*, Vol. 12, Feb. 1970, pp. 95–102.
13. J. L. Jaech. Further Tests of Significance for Grubbs' Estimators. *Biometrics*, Vol. 27, Dec. 1971, pp. 1097–1101.
14. C. J. Maloney and S. C. Rastogi. Significance Test for Grubbs' Estimators. *Biometrics*, Vol. 26, Dec. 1970, pp. 671–676.
15. M. Stroup-Gardiner and D. Newcomb. Statistical Evaluation of Nuclear Density Gauges Under Field Conditions. In *Transportation Research Record 1178*, TRB, National Research Council, Washington, D.C., 1988.
16. J. W. Barnes. *Statistical Analysis for Engineers: A Computer-Based Approach*. Prentice-Hall, Englewood Cliffs, N.J., 1988.
17. A. H. Bowker and G. J. Lieberman. *Engineering Statistics*, 2nd ed., Prentice-Hall, Englewood Cliffs, N.J., 1972.

Publication of this paper sponsored by Committee on Pavement Management Systems.