

Conducting Statistical Tests of Hypotheses

Five Common Misconceptions Found in Transportation Research

SIMON WASHINGTON

Most statistical methods use hypothesis testing. Analysis of variance, regression, discrete choice models, contingency tables, and other analysis methods commonly used in transportation research share hypothesis testing as the means of making inferences about the population of interest. Despite the fact that hypothesis testing has been a cornerstone of empirical research for many years, various aspects of hypothesis tests commonly are incorrectly applied, misinterpreted, and ignored—by novices and expert researchers alike. On initial glance, hypothesis testing appears straightforward: develop the null and alternative hypotheses, compute the test statistic to compare to a standard distribution, estimate the probability of rejecting the null hypothesis, and then make claims about the importance of the finding. This is an oversimplification of the process of hypothesis testing. Hypothesis testing as applied in empirical research is examined here. The reader is assumed to have a basic knowledge of the role of hypothesis testing in various statistical methods. Through the use of an example, the mechanics of hypothesis testing is first reviewed. Then, five precautions surrounding the use and interpretation of hypothesis tests are developed; examples of each are provided to demonstrate how errors are made, and solutions are identified so similar errors can be avoided. Remedies are provided for common errors, and conclusions are drawn on how to use the results of this paper to improve the conduct of empirical research in transportation.

Hypothesis tests are statistical constructs (tools) that are used to both ask and answer questions about phenomena. The usefulness of hypothesis tests in research depends on sound and informed application of the tests. Hypothesis tests are easily misused and misinterpreted by engineers and scientists. In defense of engineering researchers, it is not all that surprising that statistics are abused and misused (for most of us it is a matter of degree and not a guilty or not-guilty verdict). Elementary textbooks often do not explain all aspects of hypothesis testing, complex textbooks are not read or understood by most engineers, and last but not least most engineers have had but one statistics course in their academic training—often many years ago. Thus, much can be learned by revisiting various aspects of conducting hypothesis tests and how test results are misinterpreted and abused. Because they are a fundamental underpinning of most statistical methods, they are used in an overwhelming amount of empirical research. For example, consider the following common applications:

1. *t*-tests used in regression and discrete choice models are hypothesis tests to estimate the probability that confidence intervals around regression slope coefficients contain zero (or some other prespecified value).

2. *t*-tests in analysis of variance are hypothesis tests to estimate the probability that confidence intervals around two or more group means overlap.

3. *F*-tests used in regression are used to determine the probability that variables in a model provide greater explanatory power than some reduced model form.

4. Chi-squared tests used in contingency tables test the probability that factors are independent.

5. Confidence intervals developed by bootstrap or Monte Carlo methods are an extension of hypothesis tests.

Hypothesis tests are monumentally important for conducting research in that they provide a way for an engineer to feel confident about a decision resulting from the analysis of data. Used properly, hypothesis tests can provide a powerful tool with which decision makers with sufficient information can act appropriately.

As stated previously, however, hypothesis tests are commonly misused. One needs only to peruse published research to find tests being conducted and interpreted on shaky ground. This is disturbing. In a field in which tens of millions of dollars are spent (or not spent) regularly in response to accrued research findings, it is critical for results to be interpreted correctly. Instead of highlighting flaws in the published literature throughout transportation (the reader can confirm that this is easy to do), which neither makes the finger-pointer popular nor better informs the reader, it is left to the initiated reader to develop a more critical eye for past and potential future failings with regard to hypothesis tests.

First the mechanics of hypothesis testing in general are reviewed here. Then, five common misconceptions of hypothesis test applications and interpretations are provided. For each, examples are provided and advice is given about how to avoid these errors in future applications of hypothesis tests.

MECHANICS OF HYPOTHESIS TESTING

The first subject reviewed is the mechanics of hypothesis testing, which is essentially the same regardless of the application (*t*-tests, *F*-tests, chi-squared tests), and so is generalizable to most empirical studies. An example is used to aid in the understanding of the hypothesis testing process.

Hypothesis testing is premised on asking theoretical questions about engineering phenomena and uses the rigor of statistical theory to help to assess the evidence. In hypothesis testing the answer provided is binary: yes or no, good or bad. To begin, the engineer poses two competing hypotheses: a null hypothesis (the hypothesis to be nullified) and an alternative. Generally the engineer intends to demonstrate that the data are supported by an alternative to the null hypothesis.

To illustrate, suppose an engineer has off-ramp volume data 3 months before and after installation of a variable message sign

with route-diversion information. By comparing downstream off-ramp use data during the time periods before and after provision of route diversion information, the engineer hopes to quantify the effect of the variable message sign on route diversion. For the purposes of this paper, it is assumed that potential confounding variables have been controlled, that nonhypothesis testing aspects of the study have been conducted soundly, and that the effect of time on sign effectiveness is studied separately. The engineer sets up two hypotheses:

- Null hypothesis (H_0). There has been no change in downstream mean off-ramp traffic volumes since installation of the variable message sign.
- Alternative hypothesis (H_A). There has been a significant increase in mean traffic volumes on the off-ramp since installation of the variable message sign.

Because the effectiveness of the variable message sign is supported by the alternative hypothesis, the engineer assesses the evidence against the null hypothesis.

Testing of these hypotheses is probabilistic and not deterministic. Thus, by using knowledge of statistical distributions, engineers make judgments about the likelihood that an observed increase in traffic volumes was the result of random chance acting alone (the null hypothesis). If it is not likely that random chance alone produced such observed data, then there is more support for the alternative hypothesis.

Figure 1 depicts the general hypothesis-testing scenario for this example. The statistical distribution of traffic volumes on the off-ramp assumed to occur under the null hypothesis is shown on the left. This distribution of traffic volumes, those that occurred naturally before installation of the variable message sign, would not be expected to change in the after period if the variable message sign were ineffective. The alternative hypothesis—that off-ramp traffic volumes increased because motorists were diverted by improved route information—will result in a shifting of the entire distribution of traffic volumes to the right as shown in the figure. It should be noted that the magnitude of the shift in traffic volumes, reflected by a shifting of the means μ_{null} and $\mu_{\text{alternative}}$, is unknown. Thus the test is set up to test whether any (nonzero) positive shift in the mean of traffic volumes occurs. The alternative hypothesis encompasses many possible outcomes. Shown in the figure is one possible outcome for the alternative hypothesis—in actuality the distribution of

traffic volumes in the after period could be farther to the left or right of the one shown in the figure and still be larger than μ_{null} . This fact is important for later interpretations of hypothesis test results.

The decision to be made by the engineer is which of the two hypotheses, *the null or an alternative*, provides a better description or explanation of the observed data. Empirical evidence against a chance explanation (the null hypothesis) supports an alternative explanation for the observed data. Lack of evidence to reject a chance explanation generally raises some doubt about the validity of an engineering claim and often leads to further data collection and scrutiny.

These statistical decisions are made by using notions of probability or long-run frequency of occurrence. To understand how these decisions are made, the engineer must necessarily understand type I and type II errors. Table 1 shows the differences between type I and type II errors and their associated probabilities. From Figure 1 and Table 1 it can be observed that four outcomes of a statistical test are possible: two incorrect outcomes (type I and type II errors) and two favorable outcomes.

The significance level α , shown in Figure 1, represents the probability that off-ramp traffic volumes are greater than V_{critical} under the assumption that traffic volumes did not change when advanced warning information was provided to motorists (the null hypothesis). α is the area under the null hypothesis distribution curve (probability) to the right of V_{critical} . In the hypothesis testing framework, the significance level α , commonly known as a type I error rate, is the probability that the null hypothesis explanation of the data is rejected when in fact the null hypothesis is true. In the spirit of the example, traffic volumes greater than V_{critical} will occur naturally α percent of the time. If traffic volumes greater than V_{critical} are observed in the after period, then the null hypothesis will wrongly be rejected α percent of the time.

The other error made during hypothesis testing is known as a type II error. Figure 1 and Table 1 depict this error as well. If the alternative hypothesis were true and traffic volumes did increase on the off-ramp in the after period, then the distribution of traffic volumes would look more like the distribution with mean $\mu_{\text{alternative}}$ (or some similar distribution shifted to the left or right). The area to the left of V_{critical} under the alternative hypothesis distribution is β . It represents the probability that the null hypothesis is concluded to be true even though it is false (because the observation resulted under the alternative hypothesis).

Several important relationships emerge by inspecting the relationship between type I and type II errors. First, the smaller is α , and the larger is β . Thus, if α were made to be really small, the cost would

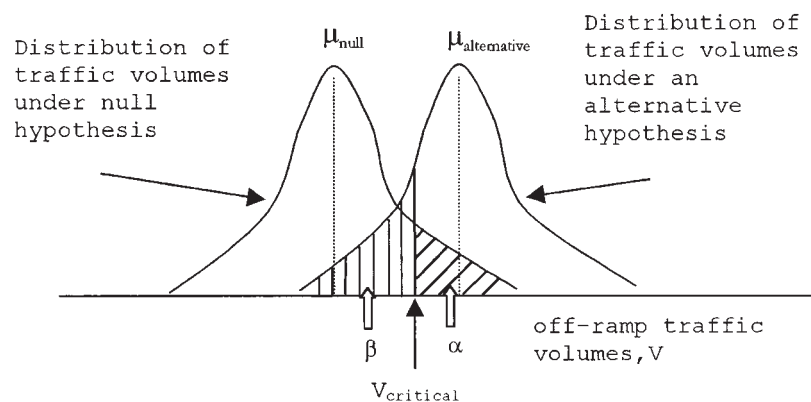


FIGURE 1 Diagram of hypothesis test of traffic volumes under null and alternative hypotheses.

TABLE 1 Outcomes of Statistical Hypothesis Testing and Associated Probabilities

Test Result	Reality	
	Null Hypothesis is correct	Null Hypothesis is incorrect
Reject H_0	Error: Type I Probability = α	Correct Decision Probability = $1-\beta$
Do not reject H_0	Correct Decision Probability = $1-\alpha$	Error: Type II Probability = β

be a higher probability of a type II error. Second, the larger the effect size, or difference between the null and alternative hypothesis means ($\mu_{\text{alternative}} - \mu_{\text{null}}$), the lesser the probability of making a type II error. In practical terms, larger effects (differences between the means of the distributions under the null and alternative hypotheses) will be easier to discern than smaller ones. Finally, the variance of the distributions affects decision-making errors. The larger the variance, the greater the probability of making a type II error.

Often the probability of making a type I error, α , is set to 0.10 or 0.05 (5 and 10 percent error rates, respectively). The selection of α often dictates the level of β for a given hypothesis test. The computation of β of under different hypothesis testing scenarios is beyond the scope of this paper, so the reader is referred to more complete references on the subject such as Glenberg (1).

There is nothing magical about a 5 percent α level. As stated previously, it is the probability of incorrectly rejecting the null hypothesis. The selection of an appropriate α level should be based on the consequences of making a type I error. For instance, if human lives are at stake when an error is made (which may be the case in accident investigations and medical studies), then an α value of 0.01 or 0.005 may be appropriate. On the other hand, if the result merely determines where monies are spent for improvements (congestion relief, transit level of service improvements), then perhaps a less stringent α value is more appropriate.

After the α and β values are selected, the engineer can conduct hypothesis tests, compute the results, and interpret and report the results. But as was described previously, there are many pitfalls to watch out for in this process. The following section describes five of these pitfalls, or misconceptions, and provides examples and recommended solutions.

FIVE MISCONCEPTIONS OF HYPOTHESIS TESTING

There are many opportunities for foibles and blunders before hypothesis testing. They can occur as early on as sampling of the data (nonrandom sampling methods), selection of error rates (too liberal or conservative), unchecked assumptions of a statistical test or distribution (choosing a continuous as opposed to discrete distributions), and selection of the statistical technique (using regression on count data). The blunders that occur before hypothesis testing are not the subject of this section. It is assumed that the analyst has learned the necessary information from theories and related empirical studies, has made sound decisions about experimental or observation field-study design, has selected the appropriate methods, checked assumptions about distributions and sampling, and is ready to interpret the results of hypothesis tests. It is at this stage of the research where this paper will aid the analyst.

Misconception 1: α Is the Most Important Error Rate

For several reasons the probability of making a type II error is often ignored. First, many introductory statistics courses—the type most engineers have taken—do not include detailed discussions about type II errors. Second, an overemphasis on reporting levels of significance (P values) has developed in the research literature (2). Finally, most statistical packages to date do not provide information on the power of statistical tests, so the information is not readily available to practicing engineers. Fortunately, current and newly proposed versions of many statistical packages are being offered with at least some limited power analysis capabilities. Engineers should take into account both α and β in application of the decision rule to the extent possible. To illustrate the importance of statistical errors associated with hypothesis testing, consider again the advance warning sign example illustrated previously. If a type I error were made after data were collected and analyzed, an engineer would conclude that the advance warning sign did divert a significant number of travelers off the freeway and that the sign was indeed effective when it actually was not. Monies may be spent on installation of variable message signs based on the premise that they are effective when they are not. In contrast, in the case of a type II error, the engineer would conclude that the advance warning sign did not significantly affect travelers when in fact it did. Then monies may not be allocated for the purchase of variable message signs even though they are effective.

Consider a more critical example. In many safety studies the effect of countermeasures on human safety is often being assessed. Suppose an engineer is conducting a statistical test of the effect of freeway concrete median barriers on fatal crashes—whether they reduce fatalities. A type I error would result in spending money on ineffective median barriers, whereas a type II error would result in failing to spend money on effective median barriers, resulting in a missed opportunity to save lives.

Type II errors often are ignored in the development of statistical models—which, of course, is embedded with statistical tests of hypotheses. Engineers might set an α level associated with a t statistic to 0.05 only to find that several variables in their models, which were thought to be important on theoretical grounds, had P values associated with t statistics greater than 0.05—and they then remove them from their model. In some cases, these variables have suffered from type II errors and should still be included in the model, especially when there is theoretical support for such variables. Systematically removing nonsignificant variables from a model, therefore, ignores the possibility of important variables and related t -tests that have suffered from type II errors.

The determination of which statistical error is less desirable depends on the research question and consequences of the errors. Because these errors are related—smaller α equals larger β , all else

being equal—careful decisions need to be made with regard to selection of α and β , and attempts need to be made to quantify β when appropriate. There are various software packages available for calculating type II error rates, and many textbooks also provide the necessary tools—see, for instance, chapter 9 of Glenberg (*1*).

Misconception 2: Hypothesis Test Results Are Unconditional

The correct interpretation of a hypothesis test is as follows: If repeated many times (many samples drawn from the population), the outcome (data) observed by the analyst/engineer and reflected in a computed test statistic (t statistic, F ratio, chi-squared) would occur x percent of the time if the null hypothesis were true. In other words, the probability of occurrence is conditional on the null hypothesis being true. If x is less than α , then the null hypothesis is rejected. When the null hypothesis is rejected, the statistical evidence suggests that the null hypothesis is not true and that some alternative hypothesis provides a better account of the data. What is important to understand (and what is commonly misinterpreted) is that the result does not provide the probability of the null hypothesis being true, nor does it provide evidence that the particular alternative hypothesis is true. In contrast, however, it provides the probability of observing the data if the null hypothesis were true.

To illustrate this important interpretive distinction, consider the following hypothetical (and vivid) scenario. Suppose a researcher is interested in testing the following question: What is the probability that a motor vehicle crash results in a fatality, given that a motorist traveling on a freeway (at high speeds) is involved in a head-on collision with an immovable object?

Upon analyzing the available data, the analyst finds that 15 percent of all fatal crashes involved motorists traveling on freeways where there was a head-on collision with an immovable object. So, is 15 percent the correct answer to the researcher's original question? From the given information the correct answer is not known, but it is likely to be as high as 80 percent. The difference in interpretation of the two related probability statements is rather fuzzy, as illustrated by the two apparently similar but very different probabilistic statements:

1. The probability that a crash is fatal, given a head-on collision with an immovable object on a freeway \approx 80 percent.
2. The probability that a crash is a head-on collision with an immovable object on a freeway, given that the crash is fatal \approx 15 percent.

However unrealistic (and morbid) this scenario, it illustrates the common interpretive error made with hypothesis tests. Often the hypothesis test provides not the probability of greatest interest but an interesting and related hypothesis. In fact, the apparently similar probabilistic statements can reflect markedly different probabilities.

Consider the hypothesis test comparing the means of off-ramp traffic flows before and after installation of a variable message sign (as used in the previous example). Assume that the engineer is able to reject the null hypothesis (that means are equal) at the 5 percent level of significance. A correct interpretation would be as follows: "Given the assumption that the traffic flow means are equivalent (the null hypothesis), if an engineer repeatedly collected off-ramp traffic flows during before and after periods, their means would differ by the amount observed in about 5 percent or fewer of the cases by random chance alone. This is sufficient objective evidence to reject the null

hypothesis as a good account of the data" (given the engineer's subjective application of α). An incorrect interpretation of the results is "we are 95 percent sure that the means of the two traffic volume distributions differ," because that is not the probability that has been calculated.

Misconception 3: Hypothesis Test Conclusions Are Correct

Conducting hypothesis tests comes with errors predetermined by the engineer. Type I errors (rejecting the null hypothesis when it is true) will be committed for about α percent of tests where the null hypothesis is rejected. Similarly, type II errors (failing to reject the null hypothesis when it is false) will be committed for about β percent of the tests when the null hypothesis is not rejected. Only the long-term likelihood of the error is knowable and not whether the analyst made an error on any particular test.

Consider the following. If, for example, over the course of a study, a researcher were to conduct 300 independent hypothesis tests using $\alpha = 0.05$ and $\beta = 0.10$, and there were 100 rejections of the null hypothesis, then there should be $100 \times 0.05 = 5$ type I errors and $200 \times 0.10 = 20$ type II errors—a total of 25 errors in the course of testing hypotheses. Thus, over this given course of study the researcher would have made 5 false claims of success and would have incorrectly concluded failure on 20 occasions. Unfortunately, there is insufficient information for the engineer to know which of the conclusions were correct and which were incorrect. Interestingly, this is one of the arguments against data snooping and data mining. If researchers search long and hard enough through statistical testing (modeling, testing means), then they are bound to make type I errors and claim success eventually.

What becomes evident is that by conducting increasing numbers of hypothesis tests the chance of obtaining an incorrect rejection of the null hypothesis becomes increasingly likely. If the careless engineer discards (or neglects to report) the failures—the inability to reject the null hypotheses—then the results will be misleading. To combat this problem, the engineer can reduce the error rates by dividing by the number of tests to be conducted in a particular study. For instance, for a study in which 200 tests are to be conducted the engineer could use $\alpha = 0.05/200 = 0.00025$ on each individual test, which would result in about $200 \times 0.00025 = 0.05$ type I errors. Of course, one must balance out the effect this would have on statistical power and would be able to implement such corrections only when sample sizes were significantly large, effect sizes were large, or variances were small.

Continuing with the example, assume off-ramp traffic volumes were collected during 100 different time periods. Instead of pooling the data, the engineer decides to conduct 100 different hypothesis tests for difference in means, each with a 5 percent α and a 35 percent β . Of the 100 tests, 80 showed rejection of the null hypothesis (the "no difference in mean volumes" hypothesis is rejected). Because there were 80 rejections of H_0 , there were about $80 \times 0.05 = 4$ incorrectly rejected hypotheses. But there were also 20 null hypotheses that could not be rejected, and so there should be about $20 \times 0.35 = 7$ type II errors. Thus, in all likelihood, there were $80 - 4 + 7 = 83$ occasions when the variable message sign successfully diverted travelers. It is impossible to determine specifically from the data which observation periods were successful and which were not.

Misconception 4: Statistical Significance (Not Effect Size) Determines Practical Significance

Unfortunately, it is all too easy to focus too heavily on P values, the probability of committing a type I error provided in most statistical package outputs, instead of effect size. The difference between α and P values is subtle. α is the probability of committing a type I error and is a constant selected by the engineer before a hypothesis test is conducted. It is a subjective value that is intended to delineate chance from nonchance outcomes. P values are provided in most statistical package output. They are specific α levels associated with particular test results, where the α associated with an individual test result is calculated exactly. Often, a large number of highly significant statistical tests can mislead an engineer into claiming early success. Actually, statistical significance is a necessary but not sufficient condition for ensuring practical significance.

Effect size is the magnitude of a difference being tested (the difference in traffic volume means), the slope coefficient in a linear model (the rate of change of trips with respect to household income), and the difference between observed and expected cell frequencies in a contingency table analysis, to name a few.

To illustrate, suppose that in testing the effectiveness of the variable message sign the engineer claims success by showing a 2 percent increase in off-ramp use during peak travel times to a 99.99 percent level of significance—a P value of 0.01. Assuming that the three-lane facility carries 2,000 vehicles per lane per hour (vplph) during peak times, and the one-lane off-ramp carries 1000 vplph, this amounts to an increase of $1000 \times 0.02 = 200$ vehicles per hour on the off-ramp, or a diversion of about 66 vplph on the mainline—a diversion of about 3 percent of all mainline vehicles. Whether a 3 percent diversion of mainline vehicles has practical significance depends on the specific location. However, at a location where a known high proportion of motorists could use the alternative route information (which presumably is a good candidate location for sign placement), a 3 percent diversion could well be considered a failure and not a success. Thus, successfully demonstrating statistical significance by no means ensures practical significance.

The need to assess effect size is perhaps the biggest argument in favor of confidence intervals. Confidence intervals are used to quantify the range of effect size so that engineers can assess the practical significance of the statistical findings. Confidence intervals are a natural extension of hypothesis testing and are extremely useful to aid in the interpretation and presentation of research findings. In fact, some argue that confidence intervals should be used in place of hypothesis testing in all applications (3).

Misconception 5: Nonsignificant Hypothesis Test Results Are Unimportant

Many engineering researchers (and nonengineering researchers) have the impression that the inability to reject the null hypothesis is an undesirable result. Although understandable, this perception is ill conceived. The hypothesis test gives us an objective account of the likelihood of observing data given certain assumptions or preconceptions about the process under study. No matter the result of particular hypothesis tests, the engineers depart with more knowledge than they had before the test. Failure to reject the null hypothesis merely suggests that one of several explanations provides an account of the results: (a) the expected effect, relationship, or difference did not manifest itself in these data; (b) a type II error

occurred; (c) the sample size was too small to discern an effect; (d) the effect size was too small to detect; (e) the inherent variability in the data is too large to discern an effect; or (f) there is no effect.

A failure to reject a null hypothesis leaves the engineer with a number of possible responses. First is to believe the result as truth and report the findings [accepting account (a) as explanation]. Considering the off-ramp flow example, under consistent failings to reject the null hypothesis the engineer might conclude that the variable message sign, which provided route diversion information to motorists, was ineffective in its attempts at diverting motorists. Suppose, however, that similar variable message signs installed elsewhere were shown to be effective, and so the engineer was hesitant to accept the statistical evidence.

The second possible response is to deny the results as truth and accept one of the latter four explanations for the results. A type II error may have occurred. However, disaggregation of the data to single time periods might provide evidence against this explanation, as one would expect to be able to reject at least some hypotheses if a type II error occurred on the aggregate data. If hypothesis tests were conducted for each of the 100 days of off-ramp data, an engineer might rightly conclude that a large number of type II errors are likely. In this case explanations (c), (d), and (e) are likely candidates to explain the consistent failings to reject the null hypothesis. Figure 1 shows the relationship between effect size—the difference between μ_{null} and $\mu_{\text{alternative}}$ —and α and β . As the true difference between distribution means becomes smaller (with variance fixed), our ability to identify it as significant decreases. In other words, as $\mu_{\text{alternative}}$ moves toward μ_{null} , it becomes decreasingly likely that an outcome under the alternative hypothesis will be observed in α and more likely that it will be observed in β . With regard to the off-ramp flows under route diversion, as the true effect of the variable message sign gets smaller, the ability to detect it (assuming constant variance in traffic flows) becomes increasingly difficult. It should be apparent that sample size (which directly affects sampling variance), variance, and effect size are all manifestations of the same difficulty—an inability to reject the null hypothesis.

Fortunately the engineer can minimize the effect of problems related to sample size, variance, and effect size by careful experimental design and data collection. Pretests can be used to sample from the population and obtain initial estimates of both effect size and sample variance. These results can then be used to design a larger engineering study to minimize the likelihood of rejecting the null hypothesis when it is indeed false. Under these well-planned engineering studies, the inability to reject null hypotheses is more likely to reflect truth instead of deficiencies in experimental or observational design.

Once an outcome of a well-planned engineering study is determined, either favoring or rejecting null hypotheses, valuable information is obtained from the study. If variable message signs were shown to be effective, perhaps they should be installed at other locations where route diversion information would be of use to motorists and system managers. If variable message signs are shown to be ineffective, then perhaps there is a design flaw or operations problem at the site. Perhaps motorists were not provided with the right kind of information, perhaps motorists were given false information (alternative routes were not faster), or perhaps insufficient time was allowed to diverge to the off-ramp. In either case, the statistical results of an engineering investigation provide valuable insight and information for engineers to forge ahead and make engineering improvements. It is critically important for engineers to report the results of research, despite the statistical outcome. Failure to

report failures may result in other researchers spending money to discover the same result—an inefficient way to conduct research. Only through the collective accumulation of research findings can objective assessments of phenomena under investigation be made.

CONCLUSIONS

Five common misconceptions surrounding the conduct of statistical hypothesis tests have been examined. Many of the misconceptions, or pitfalls, can be remedied with careful application and interpretation of the findings. The recommendations in this regard are as follows:

1. The type and frequency of statistical results need to be carefully selected. α and β error rates for type I and type II errors, respectively, are inversely related, and thus a trade-off between the two must be made. Both error rates are related to sample size, effect size, and variability of the data. In some cases type I is the more serious error, and in other cases type II errors are more serious. Software packages that compute statistical power ($1 - \beta$) are becoming more readily available.
2. Researchers should recognize the conditional nature of hypothesis tests and should not interpret the P value of a statistical test as the probability of the null hypothesis being false. Often the probability of the null hypothesis being false is very different from the P value associated with a test.
3. The conduct of multiple hypothesis tests increases the likelihood of obtaining an error. Unfortunately, it is not known whether the result of any particular statistical test is an error. Researchers should be aware of the cumulative effect of conducting hypothesis tests and adjust error rates if necessary; they should understand the consequences of data snooping and mining activities.
4. The P values provided in many statistical tests do not indicate practical significance; they are merely a first step. The magnitude of the effect is the most important indicator of practical significance, and it should be explicitly considered in model development and statistical analyses.
5. Researchers should report both successes and failures in research. It is a misconception that failure to reject the null hypothesis

(nonsignificant t statistics, difference in means or variances, small differences between observed and expected) in research equates to failure of the research itself. In fact only through complete reporting of all tests can objective accumulated knowledge be generated about a phenomenon under study.

There are additional analytical techniques that can be used to remedy some of the problems identified here. One that has already been discussed is the adjustment of α levels to account for multiple hypothesis tests being conducted within and across studies.

Another powerful analytical technique, useful for overcoming the conditional nature of hypothesis tests, is application of Bayes' techniques. Although the technical details are beyond the scope of this paper, Bayesian statistics offer a way to compute the probability of the null hypothesis being true, given the data, which is usually the primary interest of researchers. Recall that the classic statistical hypothesis testing machinery provides the probability of the data, given that the null hypothesis is true, which often is a very different probability. Bayesian statistical methods can be found in many textbooks, some of which are devoted entirely to Bayesian methods.

It is intended that this primer on hypothesis testing be used to improve the quality and reporting of empirical research conducted in the field of transportation. Many aspects of hypothesis testing covered here are not covered in detail in elementary texts and so may be out of reach for many practicing engineers and engineering researchers.

REFERENCES

1. Glenberg, A. M. *Learning from Data, An Introduction to Statistical Reasoning*, 2nd ed. Lawrence Erlbaum Associates, Mahwah, N.J., 1996, chap. 9.
2. Abelson, R. P. *Statistics as Principled Argument*. Lawrence Erlbaum Associates, Mahwah, N.J., 1995, pp. 39–41.
3. Harlow, L. L., S. Mulaik, and J. Steiger, eds. *What If There Were No Significance Tests?* Lawrence Erlbaum Associates, Mahwah, N.J., 1997.

Publication of this paper sponsored by Committee on Statistical Methodology and Statistical Computer Software in Transportation Research.