On Statistical Testing Methods for Dimension Lumber Property Monitoring

Matthew A. Arvanitis
C. Adam Senalik

Abstract

The American Society of Testing and Materials (ASTM), D1990 details standard practices for establishing and maintaining design values for dimension lumber. Included in this standard are statistical methods for detecting changes in dimension lumber properties over time. A characteristic (or design) value, also known as the allowable property, is published for each lumber property, and this value must be updated whenever changes in the resource warrant. Currently ASTM D1990 calls for the use of the Wilcoxon Rank Sum test (also known as the Mann-Whitney test) to detect changes in these properties. In essence, this test is a two-sample test designed to detect changes in the underlying populations from whence the samples were drawn. In this work, FPL researchers recommend, with justification, that this practice be revised to the use of one-sample tests that focus on detecting disparities between the current resource properties and the corresponding currently accepted design values. We detail the tests that are recommended for properties whose design values are (based on) either the mean or some quantile. We further examine the impacts, in terms of statistical errors, of these alternative tests in comparison to the current paradigm through simulations using a collection of distributions modeled from actual within-grade softwood lumber modulus of elasticity and modulus of rupture data.

Dimension lumber property monitoring is a long-standing practice intended to ensure allowable properties reflect the current state of the resources they represent. Under the current paradigm, this is accomplished with the use of a characteristic (or design) value for each lumber property. It is important to recognize that this value, which is a summary statistic describing a key parameter of the resource, does not provide complete information about the property; it merely addresses what the industry finds to be an important descriptor of the property. Currently, the testing procedure applied to identify "decreases" in the properties is not designed to detect specifically decreases in underlying parameters driving these characteristic values; it is designed to detect general distributional decreases. Strength design values (bending, tension, and compression) are dictated by the lower tail of the distribution of strength values obtained through testing. It is imperative that a designer be able to assume that any piece of dimension lumber will perform to the stated design level. For this reason, the mean and median are not as important as the lower extent of the values for the distribution.

Prior to the 1990s, design values for dimension lumber were based upon tests of small clear wood specimens. Previous testing had quantified the reduction in wood strength caused by features such as knots, slope of grain, etc. As feature size increased and/or features moved closer to critical points on the board (tension edge), so did the reduction in strength caused by the feature. Lumber with small features away from the board edge was assumed to have higher strength than lumber with larger features closer to the board edge. Grades were established by prescribing the size and location of the feature allowable for a particular grade based on both the effect on strength and/or appearance and usability. The presence of the features was determined visually, leading to the term visually graded lumber.

In the late 1970s and 1980s a large-scale sampling and testing plan was undertaken to establish design values for certain properties based on full sized dimension lumber rather than small pieces of clear wood. While the lumber grade was still dictated by the size and location of features within the board, the design values were now based upon the strength values obtained from direct testing of the lumber with those features. This was known as the in-grade process. In 1991, the design values for visually graded dimension lumber

The authors are, respectively, Research Mathematical Statistician (Corresponding Author: matthew.arvanitis@usda.gov) and Research General Engineer (christopher.a.senalik@usda.gov), USDA Forest Products Laboratory, Madison, Wisconsin, United States. This paper was received for publication in February 2025. Article no. 25-00006.

Forest Prod. J. 75(3):229–237.
doi:10.13073/FPJ-D-25-00006

based upon testing of full-sized lumber were published (Evans 2001, Green and Evans 2001).

In the early 2010s regular monitoring became required for large production commercial species groups. Monitoring typically involved gathering a sample of dimension lumber of a single size (2X4, 2X6, etc.) and grade (usually number 2) from across the production region. The monitoring sample was (usually) tested in bending, and the modulus of rupture and modulus of elasticity were compared against the in-grade data.

Between the time of the original in-grade testing and required monitoring, machines capable of performing visual grading had become available. For some commercial species groups, these advancements resulted in a narrower range of strength values. While the lower tail of the monitoring sample exceeded the lower tail of the in-grade sample, the mean of the monitoring sample was below the mean of the in-grade sample. As part of the monitoring process, the two samples were compared using the Wilcoxon rank sum test.

Hence, the narrowing of the strength distribution of the monitoring sample caused the Wilcoxon rank sum test to (correctly) indicate a reduction in the present resource relative to the in-grade population but gave little insight into whether the monitoring sample supported the existing design values, which were based upon the lower tail of both distributions (Kretschmann et al. 2014).

In this work, FPL researchers recommend a revision to ASTM D1990, Section 14 (and related sections), in which *one*-sample testing replaces the existing Wilcoxon rank sum test for detecting changes in characteristic values specifically over time (ASTM International 2019). In addition, we propose one-sample tests that are specific to the distributional parameter(s) on which the characteristic values are based, rather than the general "stochastic differences" approach offered by the Wilcoxon test. We argue that testing procedures should be, at the very least, related to the methods initially applied to establish the published characteristic values. For example, if the characteristic value is based on the 5th percentile, then the test should focus its detection capabilities on the 5th percentile, e.g., a simple non-parametric test; if the characteristic value is based on the mean, then the test should focus its detection capabilities on the mean, e.g. the one-sample *t*-test. We show by simulation that this revision will significantly reduce the likelihood of Type I and Type II errors in these statistical procedures, potentially saving countless dollars by avoiding unnecessary lumber testing and improving the overall integrity of the monitoring process.

This paper is organized as follows. First, we explain the purposes of the monitoring program and the current testing paradigm before detailing our recommended revision to the program guidance in ASTM D1990. We then exhibit a systematic simulation that mimics the testing procedure normally followed under the monitoring program and present results of said simulations for both the Wilcoxon and the alternative tests for both modulus of elasticity (MOE) and modulus of rupture (MOR). In the last section, we provide some concluding remarks and restate our recommendations for updates to the monitoring program. Appendices discuss some mathematical modeling details and further expand on some findings in the paper.

Background

A characteristic value can assume one of three possible forms, each computed from a sample of the resource property:

1. The sample mean, used for the modulus of elasticity (MOE),
2. The sample median[1], or
3. The 75% Lower Tolerance Limit (LTL) for the 5th percentile, used for the modulus of rupture (MOR).

It should be noted that while these are the values assumed by the characteristic values, if perfect information was available about the properties, they would be given their corresponding population values; that is, the population mean would be used for the mean, the population median would be used for the median, and the population 5th percentile would be used for the LTL of the 5th percentile. This is important to consider when performing statistical test, and it is the reason why we use the unknown population parameters in the statement of our hypotheses discussed later.

Lumber properties can change over time. These changes can be due to a variety of reasons, e.g., advances in forest management practices, growth rates, mill processes and technologies, or grading procedures. The monitoring program was instituted as a result of the recognition of these factors. Kretschmann et al. (1999) stresses this point:

> Independent grading agencies have a strong interest in determining whether significant change (particularly a decrease in material properties) in the lumber resource has occurred.

ASTM D1990 defines *monitoring* as

> a periodic review of a subset of structural properties of a lumber cell to determine if a potential downward shift from the assigned values indicates a need for an evaluation or reassessment, or both, of allowable properties developed with this practice.

ASTM D1990 further states that a fundamental purpose of monitoring is to

> determine if there is sound evidence to believe that there has been a change in the product performance sufficient to justify an evaluation . . . or a reassessment.

Evaluations and reassessments are primarily performed to establish new characteristic values.

Therefore, through the monitoring program detailed in the standard, grading agencies are tasked with assessing whether a characteristic value comports with the current state of the resource. The test that has long been recommended by the standard is the Wilcoxon rank sum test, a nonparametric two-sample test that is designed to detect general stochastic differences between the corresponding populations. There are two problems with this:

1. The Wilcoxon test does not specifically address the parameter of the property distribution that drives the characteristic value. Rather, it measures general probabilistic differences between two populations: the original from which a sample was collected and the characteristic value computed, and

[1] The median is not currently used as the basis for any softwood lumber characteristic value, but we include it here for consistency in accordance with ASTM D1990, 3.2.2.

the current from which a sample is presented. It does not address specifically the mean, nor the median, nor the 5th percentile, nor any other specific parameter of the distribution. Therefore, used as a test whose sole purpose is to assess whether a particular one of these parameters is reasonably represented by the currently published value, the Wilcoxon is not ideal.

2. The Wilcoxon is a 2-sample test. Once again, we are interested in assessing whether the currently published value, however and whenever it may have been first computed, comports with the current resource property. The only information relevant to this question is that currently published value and any information which pertains to the *current* population, e.g., a representative sample of said property. Under no circumstance would a second sample, particularly any that does not pertain to the current population, be useful to answering this question. Such unnecessary use of data can only serve to add variation, making it more difficult for the test to detect a difference.

To address this issue, FPL researchers recommend a revision to the current testing procedures in this standard. In the following sections, we present the proposed testing procedures, justify them, and provide simulation evidence to support them.

Proposed Testing Revisions

Before explaining the specifics, we first must fully understand the types of statistical error and their consequences in the context of lumber property monitoring.

- Type I error. Also known as "false positives," Type I error events occur when the null hypothesis, H_0, is rejected despite the fact that it holds. Generally, Type I error is easily controlled, since, in performing the hypothesis test, assuming H_0 presents knowable conditions. For example, suppose that we wish to know if the average height of males in France is greater than the same in Nigeria. To do this, we would construct a *t*-test with H_0 stating the average male height in both countries is the same. We choose equality since it is the circumstance most difficult to distinguish from the state we are trying to detect; that is, the "closest" circumstance to a violation of H_0: Nigerian males are at least as tall as French males. This assumption has reasonably knowable consequences, and we therefore are able to make tangible predictions based on those consequences which can then be compared to the data. In lumber monitoring testing, H_0 would state that the published characteristic value is no larger than the current corresponding population value. Thus, a Type I error event occurs when the test suggests that the characteristic value needs to be revised (is too high), even though it doesn't. This results in the industry needing to engage in (usually expensive) reevaluation efforts (also described in D1990) to obtain a revised characteristic value.
- Type II error. Also known as "false negatives," Type II error events occur when H_0 is not rejected even though it does not hold. Unlike Type I error, Type II error is difficult to control, for the conditions under which it can occur are largely unknown or can be drawn from a large collection of possibilities. Using the previous heights example, Type II error can occur when the average heights of Nigerian males is 2 cm less than that of the French, but it can also occur if that number was 2 mm. These two scenarios present very different Type II error rates with the latter being much higher under the same testing conditions. For this reason, assessing actual Type II error rates is difficult and generally requires undesirable assumptions. For lumber property monitoring, a Type II error event occurs when the test suggests the characteristic value is representative of the resource property when, in reality, it is excessive. While this does not cost money in unnecessary testing procedures, it does pose a safety risk, since the engineering community and regulatory agencies rely on these values to design structures with lumber and maintain building codes. In this sense, it can be said that, while Type I error can lead to unnecessary expense, Type II error can lead to even more serious consequences.

Proposed alternative tests

In this section we explain the testing procedures that FPL researchers propose replace the existing Wilcoxon test in ASTM D1990. Set α to be the desired level of significance (that is, the maximum acceptable Type I Error probability); in the standard, this value is currently set to $\alpha = 0.05$. Also, suppose the (adjusted) sample from the current resource property is $\boldsymbol{X} = (X_1, X_2, \ldots, X_n), n \in \mathbb{Z}^+$. For each of the three cases for the characteristic value, we present a specific test:

1. **Mean.** Suppose the published characteristic value is μ_0, and the actual unknown mean of the current population is μ. If the underlying distribution of the property can reasonably be assumed to be well-behaved, that is, continuous, and possessing a finite second moment (finite variance), then a reasonable (and statistically powerful) test for the mean is the one-sample *t*-test:

$$H_0 : \mu \geq \mu_0.$$

Set

$$t = \frac{\overline{X} - \mu_0}{\sqrt{\frac{s^2}{n}}},$$

where $\overline{X} = n^{-1}\sum_{j=1}^{n} X_j$, $s^2 = (n-1)^{-1}\sum_{j=1}^{n}(X_j - \overline{X})^2$, and, under H_0, $t \sim t(n-1)$; that is, t follows a *t*-distribution with $n-1$ degrees of freedom. Therefore, if $t < t_{\alpha.n-1}$, where $t_{\alpha.n-1}$ is the αth quantile of a *t*-distribution with $n-1$ degrees of freedom, then H_0 may be rejected at the α level of significance. This test can be particularly robust to deviations from its assumptions when n is large. In the case of lumber monitoring, $n \approx 360$, a value sufficient to overcome even very unusual distributional forms, which, as it happens, have rarely, if ever, been historically exhibited by MOE.

2. **Median.** Suppose the published characteristic value is m_0, and the actual unknown median is m. If the underlying distribution of the resource property can be assumed to be continuous, then a simple nonparametric test may be applied to assess the median:

$$H_0 : m \geq m_0$$

Set $U_{1-\alpha}$ to be the distribution-free $100(1-\alpha)\%$ upper tolerance limit (UTL) for the median; that is,

$$U_{1-\alpha} = \inf \{X_{(j)} : P(X_{(j)} > m) > 1-\alpha\},$$

where

$$P(X_{(j)} > m) = \left(\frac{1}{2}\right)^n \sum_{i=0}^{j-1} \binom{n}{i}.$$

Then, H_0 may be rejected at the α level of significance whenever $U_{1-\alpha} < m_0$. Because it is nonparametric, this test is highly robust to deviations from typical underlying distributions. Founded on a century of statistical theory, this test guarantees containment of Type I error, and its specific focus on the median suggests its success in terms of Type II error as well.

3. **5th Percentile.** Suppose the published characteristic value is q_0 (that is, the 75% LTL for the 5th percentile computed from some previous sample), and the actual unknown 5th percentile of the current population is q. If the underlying distribution of the resource property can be assumed to be continuous, then a test similar to that for the median can be applied:

$$H_0 : q \geq q_0.$$

Set $V_{1-\alpha}$ to be the distribution-free $100(1-\alpha)\%$ UTL for q; that is,

$$V_{1-\alpha} = \inf \{X_{(j)} : P(X_{(j)} > q) > 1-\alpha\},$$

where

$$P(X_{(j)} > q) = \sum_{i=0}^{j-1} \binom{n}{i} \left(\frac{1}{20}\right)^i \left(\frac{19}{20}\right)^{n-i}.$$

Then, H_0 may be rejected at the α level of significance whenever $V_{1-\alpha} < q_0$.

Remark. There has been some discussion in the community about the statistical consequences of using a LTL as a design value for lumber properties, to include some dissent expressed by the authors. In the authors' opinions, this test serves as a compromise on the matter. That is, we are no longer forwarding the concern of using an LTL in this way; rather, we believe this test uses the LTL in a reasonable and sustainable way, though it still results in the following minor consequence. This test is attempting to detect the event that the current unknown actual 5th percentile is less than the currently accepted design value. That design value is a *biased* estimate of the unknown actual 5th percentile obtained from the previous (usually In-Grade) sample. By its design, that 75% LTL has a 25% probability of being greater than the actual 5th percentile of the previous population and, importantly, the current actual 5th percentile in the case that it has not changed. Hence, not only will this test tend to detect cases in which the unknown actual 5th percentile has decreased over time, but it also has the secondary effect of tending to "correct" that 25% of cases in which the original design value was excessive due only to random chance. The long-term consequence of the latter is that the actual probability associated with the design value will tend to be greater than the reported 75%. Though favorable, this is an unavoidable consequence of using an LTL as a design value under this testing regime. Further discussion is found in Appendix B.

All three of these tests are based on well-established statistical theory and are not at all novel.

Simulations

As stated earlier, to assess Type I and II errors for each of these testing regimes and compare these results to those produced by the Wilcoxon Rank Sum test, we must make a collection of assumptions that may not be terribly desirable. To minimize the likelihood that these results do not represent reality, we will apply assumptions that the authors believe most likely represent the lumber property monitoring environment.

Model selection and construction

In this study, we address the two properties that are currently used as the basis for characteristic values of lumber properties: the mean and the 5th percentile. MOE and MOR, respectively, are properties whose corresponding characteristic values are based on these. One notable feature of these two properties is that they are weakly dependent in the context of softwood lumber. Therefore, we will simulate these two properties using a bivariate statistical model that is fit to actual observed data from the field. The data set is from a recent study, but, because it constitutes proprietary information, we divulge here only that it includes observed MOE and MOR values (adjusted for moisture and size) from just under 500 boards, the authors have concluded that this data represents typical results from the industry, and the values have been normalized so that the mean MOE is 1M psi and the 5th percentile of MOR is 1,000 psi so as to obscure the source of the dataset. Given the inherent weak dependence relationship between MOE and MOR, we choose to model the distributions as the marginals of a bivariate distribution. Using the process outlined in Appendix A, we constructed 7,500 models which varied slightly from one another. Contour plots for two of the models from the database are shown in Figure 1.

Simulation method

Each simulation consisted of formulating datasets of size 360 from two randomly selected models of MOE and MOR. To choose which of the two cases will be assigned to the in-Grade population, we designed a simple algorithm that stochastically chooses the assigned roles of the samples based on the known difference between the corresponding parameters. The larger the difference between the parameters, the more likely the In-Grade sample would be assigned the stronger sample. This tends to shed more light on Type II error in extreme cases of differences and Type I error in cases with small differences.

Datasets of size 360 are randomly generated from the models (one In-Grade MOE, one monitoring MOE, with corresponding In-Grade MOR, and monitoring MOR samples), and, for both properties, a characteristic value is constructed in accordance with ASTM D1990 from the sample

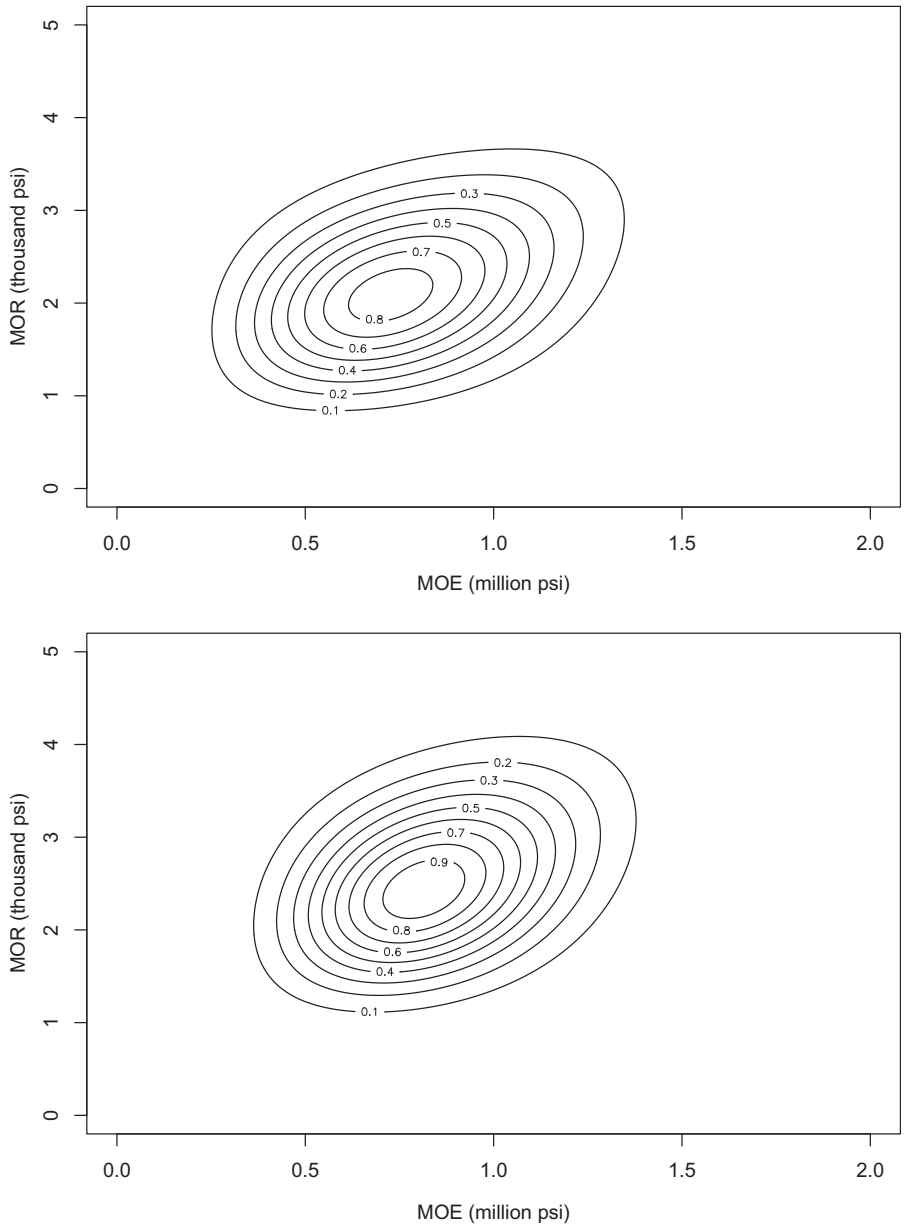

Figure 1.—Contour plots of two scaled Olkin & Liu Bivariate Beta densities from the database: $BB_s(6.63, 5.44, 10.63, 2.42, 5.54)$ (top), and $BB_s(8.69, 8.33, 16.36, 2.99, 6.31)$ (bottom).

representing the In-Grade population. Each pair fell into one of two basic cases:

1. Cases in which H_0 holds; that is, cases where the characteristic value calculated from the In-Grade sample is less than or equal to the corresponding population parameter of the distribution used to generate the monitoring sample. In these cases, only Type I error is possible.
2. Cases in which H_0 does not hold. Specifically, these cases ranged from exhibiting a negligible decrease in the characteristic value to exhibiting a 25% decrease (between In-Grade and monitoring). In these cases, only Type II error is possible.

For each of the two properties, we performed the Wilcoxon test and the alternative one-sample test assigned to the corresponding property described above. This procedure was repeated four times for each of 25 million random choices of property distributions: once at each of the $\alpha = 0.05, 0.10, 0.15, 0.2$ significance levels. This may sound excessive. However, as will be seen in the next section, we needed to observe statistically significant error rates across a continuum of differences in μ and μ_0 for MOE and between q and q_0 for MOR. The results are shared in the next section.

Remark. The authors should mention that we have previously endorsed the use of the melded random quantile difference (MRQD) test (Arvanitis 2022) for lumber property monitoring when the characteristic value is based on a population quantile. In light of further consideration and investigation, we now retract this position for two reasons:

1. Under some, possibly many, circumstances, the test can be highly conservative in terms of Type I error. Because MOR

follows a rather well-behaved distribution, it is one such case of this conservatism. This, we now believe, is problematic because it leads to unnecessarily excessive Type II error which, as previously stated, presents elevated safety risks.

2. It is a two-sample test. As discussed earlier in this work, using the original (typically In-Grade) sample serves only to introduce unnecessary noise with no accompanying benefit.

For these reasons, we have chosen not to include the MRQD test in this study. While the test can certainly be useful for many other applications, possibly even some involving softwood lumber properties, e.g., fact-finding and other research endeavors, we are no longer of the opinion that it is appropriate for softwood lumber property monitoring applications.

Results

For each value of α, the process outlined above was repeated in its entirety 25 million times, for each of which it was determined whether a Type I, Type II, or no error occurred for both the MOE and MOR properties. In addition, the actual percentage disparity between the In-Grade-determined characteristic value and the actual corresponding parameter of the monitoring population was also recorded, where it was negative if it decreased (H_0 did not hold) or positive if it increased (H_0 held). MOE and MOR results are shown in Figures 2 and 3, respectively.

In all cases, a small but significant improvement in Type II error is observed for MOE. Though the Wilcoxon test exhibits less Type I error, the *t*-test contains Type I error to α, as prescribed. For MOR, the results are unsurprisingly more significant. Type I error for the Wilcoxon test far exceeds α and is thus substantially smaller for the alternative test. This is not because the Wilcoxon test is a poor or ineffective test; rather, it is because the Wilcoxon test is not designed to address the null hypothesis as stated. For Type II error, the Wilcoxon test performs better for small differences only (by sacrificing Type I error control).

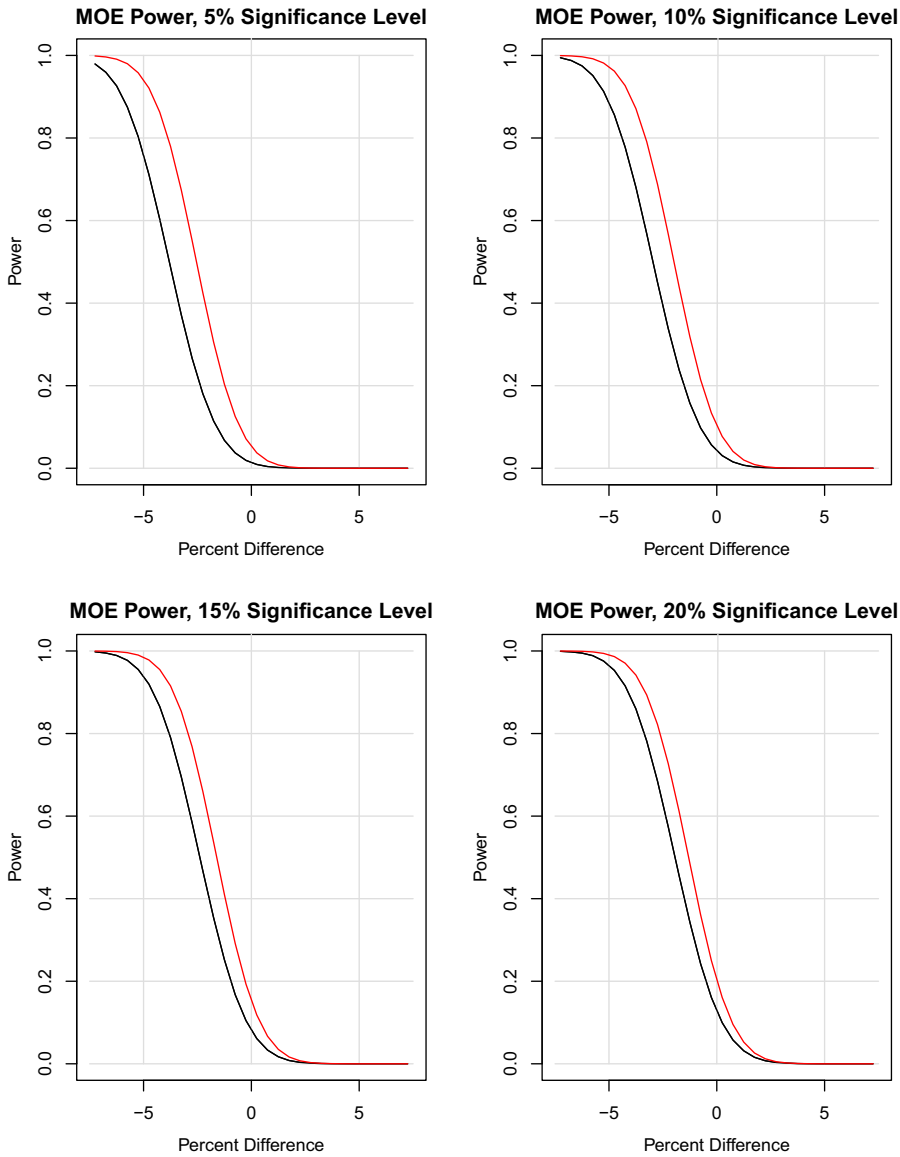

Figure 2.—Power plots for MOE. The Wilcoxon test is shown in black while the alternative test is shown in red.

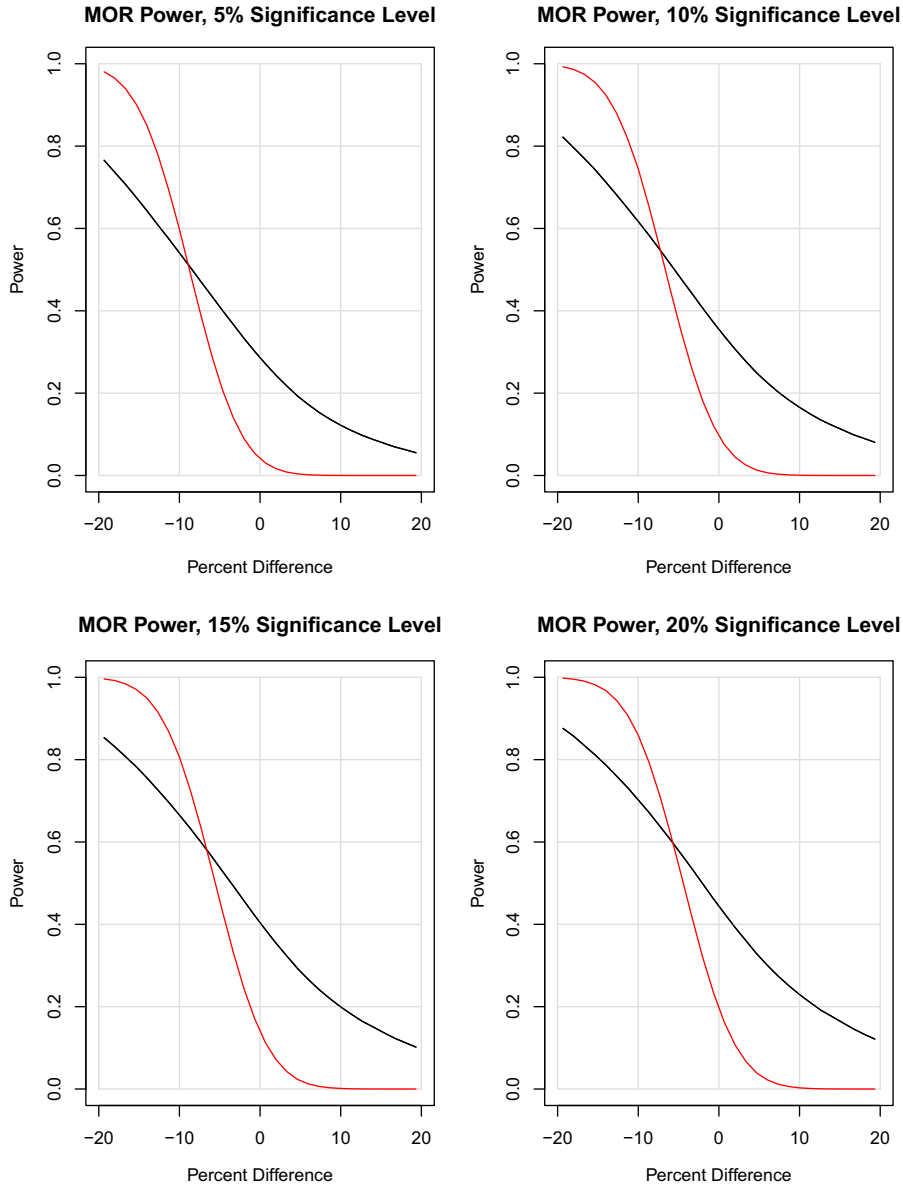

Figure 3.—Power plots for MOR. The Wilcoxon test is shown in black while the alternative test is shown in red.

Discussion

Among those involved with maintaining ASTM D1990, there are predominately two identified goals of monitoring. One goal is to detect a downward shift in the wood resource; the other is to determine if the sample supports current published design values. The producer is concerned with both goals; the consumer is likely concerned only with the latter. Over time, producers must have confidence that the resource properties are stable. If not, plans must be made to harvest the resource from different areas or go through the arduous process of modifying design values. This is important to the producer, but of less concern for the consumer. Day to day, both producers and consumers must have confidence that the product has the claimed strength to avoid potential life-safety hazards.

The Wilcoxon test provides insight into shifts in the resource but is not useful when determining if the current resource supports the published design values. Using the Wilcoxon test it is possible that no downward shift in the resource property is detected, but the lower tail of the strength distribution may have decreased below the level that can support current design values. Conversely, the Wilcoxon test may detect a downward shift in the resource property even though the current resource supports the design values. To provide a full picture of the current state of the resource, a method to evaluate whether the resource supports the current published design values must be used.

Conclusion

In this work, it has been shown that a simpler statistical test addresses the question of whether a reevaluation of the characteristic values of dimension lumber properties is necessary in a manner that is significantly better than the current testing regime. For MOE, the one-sample *t*-test is recommended to replace the Wilcoxon test, and for MOR, the simple non-parametric test outlined herein is recommended to replace the Wilcoxon test.

It should be noted that more work must be done to assess the consequences of treating the samples as random samples without regard to the possibility of clustering and other forms of dependence within the samples that the sampling procedures suggest may, at times, manifest. Independent study of these issues and their impacts on monitoring results is a necessary step to ensure the integrity of the lumber monitoring program.

Literature Cited

Arvanitis, M. 2022. Small-sample solution to the two-sample problem for quantiles using melded random confidence intervals. FS Research Paper, 713, 1–15.

ASTM International 2019. Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens (ASTM D1990-19(2019)). ASTM International.

Evans, J. W. 2001. Procedures for developing allowable properties for a single species under ASTM D1990 and computer programs useful for the calculations, Vol. 126. US Department of Agriculture, Forest Service, Forest Products Laboratory.

Green, D. W. and J. W. Evans 2001. Evolution of standardized procedures for adjusting lumber properties for change in moisture content, Vol. 127. US Department of Agriculture, Forest Service, Forest Products Laboratory.

Kretschmann, D., D. DeVisser, K. Cheung, B. Browder, and A. Rozek. 2014. Maintenance procedures for north American visually-graded dimension lumber design values. In: *World Conference on Timber Engineering, Quebec City*, Canada. p. 8.

Kretschmann, D. E., J. W. Evans, and L. Brown. 1999. Monitoring of visually graded structural lumber, Vol. 576. US Department of Agriculture, Forest Service, Forest Products Laboratory.

Olkin, I. and Liu, R. 2003. A bivariate beta distribution. *Statistics & Probability Letters*, 62(4):407–412.

Appendix

Appendix A: Methodology for constructing the model database

To simulate realistic samples for MOE and MOR, we have used a recent dataset of (paired) MOE and MOR observations that includes just under 500 samples and has been deemed typical across the softwood lumber industry by expert opinion. The only alteration made to the dataset for purposes of this study has been to scale the MOE observations to have a mean of 1 million psi and scale the MOR observations to have a 5th percentile of 1,000 psi. Because these two properties tend to exhibit a weak positive correlation, we model them together with a bivariate distribution and can subsequently generate simulated samples of any size from that model.

Family of distributions for modeling MOE & MOR

Given the slight dependency between MOE and MOR, the fact that both properties' observed values must be non-negative, and that both typically exhibit uni-modality with varying degrees (and sometimes directions) of skewness, we have constructed a scaled version of the Olkin & Liu Bivariate Beta family (Olkin and Liu 2003). With scaling, this family has a total of five parameters, three of which can describe a positive dependence relationship. Because the Beta distribution is supported on the unit interval, the scale of each marginal represents a theoretical maximum value of the property, which, once again, is realistic since there certainly exists such an unknown quantity for both of these two properties. The density of this family is

$$f(x_1,x_2) = \frac{\left(\frac{x_1}{\eta_1}\right)^{\alpha_1-1}\left(\frac{x_2}{\eta_2}\right)^{\alpha_2-1}\left(1-\frac{x_1}{\eta_1}\right)^{\alpha_1+\beta-1}\left(1-\frac{x_2}{\eta_2}\right)^{\alpha_2+\beta-1}}{\eta_1\eta_2 B(\alpha_1,\alpha_2,\beta)\left(1-\frac{x_1x_2}{\eta_1\eta_2}\right)^{\alpha_2+\alpha_2+\beta}}.$$

$$I\{0<x_j<\eta_j, j=1,2\},$$

where $\alpha_1, \alpha_2, \eta_1, \eta_2 > 0$, $\beta > 1$, and $B(\cdot)$ is the generalized beta function,

$$B(a_1,a_2,\ldots,a_k) = \frac{\prod_{i=1}^{k}\Gamma(a_i)}{\Gamma\left(\sum_{i=1}^{k}a_i\right)},$$

with $\Gamma(\cdot)$ being the gamma function. We will denote this distribution by $BB_s(\alpha_1,\alpha_2,\beta,\eta_1,\eta_2)$. It can be shown that the marginals follow scaled beta distributions with the following densities:

$$f_j(x_j) = \frac{\left(\frac{x_j}{\eta_j}\right)^{\alpha_j-1}\left(1-\frac{x_j}{\eta_j}\right)^{\beta-1}}{\eta_j B(\alpha_j,\beta)}\ I\{0<x_j<\eta_j\}$$

for $j = 1, 2$. For any dataset, maximum likelihood estimates may be obtained via numerical optimization for the parameters of the $BB_s(\alpha_1,\alpha_2,\beta,\eta_1,\eta_2)$ family. Clearly β is the only one of the five parameters to be common between the marginals, and it, together with mild contributions from α_1 and α_2, governs the dependence relationship between the two marginals. The last two parameters, η_1 and η_2, represent the scales of the marginals; that is, the maximum possible values of the corresponding random variables. These two parameters bear no impact on the dependence relationship between the marginals. For the following database, we have assigned MOE to X_1, the first marginal, and MOR to X_2, the second.

Model database

The goal is to form a database of varying distributions (parameter sets) which can plausibly represent the same resource. This was achieved by randomly selecting 100 observations at a time from the dataset described in the beginning of this appendix and obtaining parameter estimates for the Scaled Olkin & Liu Bivariate Beta distribution, based on those observations. We repeated this process 7,500 times to complete the database. The resultant MOE means varied by about ±10% and MOR 5th percentiles varied by about ±25%.

Appendix B: Impact of the 5th percentile test on the LTL

As stated previously, following a monitoring examination, the actual probability associated with the design value will tend to be greater than the reported 75%. This appendix further details why.

Suppose that the actual unknown 5th percentile of the population, q, has not changed over time. Further, suppose p_2 is the probability of rejecting H_0 when H_0 does not hold,

that is, the statistical power of the test, and set p_1 to the probability of rejecting H_0 when H_0 holds, that is the probability of Type I error. It can therefore be concluded that

$$P(q_{new} < q) = \frac{3(4 + p_2 - p_1)}{16}$$

where q_{new} is the original design value, q_0, if H_0 was not rejected (and therefore no reevaluation was done) or the newly updated design value if H_0 was rejected (and a reevaluation done). Now, since q is a fixed value regardless of whether H_0 is rejected, by the design of the one-sample 5th percentile test, it must be that $p_1 < p_2$, so that $P(q_{new} < q) > 0.75$, with an upper bound of 0.9375. However, this is after only one monitoring evaluation; it further increases after multiple evaluations. For example, if $p_1 = 0.2$, and $p_2 = 0.6$, then, after one monitoring test, the probability for the LTL representing the design value rises to 0.825; after two, 0.8625. In addition, while this probability would increase over multiple monitoring tests, it would near an asymptotic value within only a few such tests, e.g., for the present example, the limiting probability is 0.9. Now, under no circumstance would we know the values of p_1 and p_2 because we do not know the value of q, nor do we know the exact distribution of MOR, and we therefore cannot determine (or even bound) the probability that the UTL is less than q_0, so we can only guess what the actual probabilities might be. What we can be certain about is that the probability associated with the LTL (the characteristic value) will remain at least 75% regardless of p_1 and p_2 or how many monitoring tests are performed on the resource over the years. This effect is, of course, in addition to the disparity between the assigned probability and the actual probability associated with the corresponding order statistic representing the LTL, which may be significantly higher due to the inherent granularity of nonparametric tolerance limits.