United States Department of Agriculture

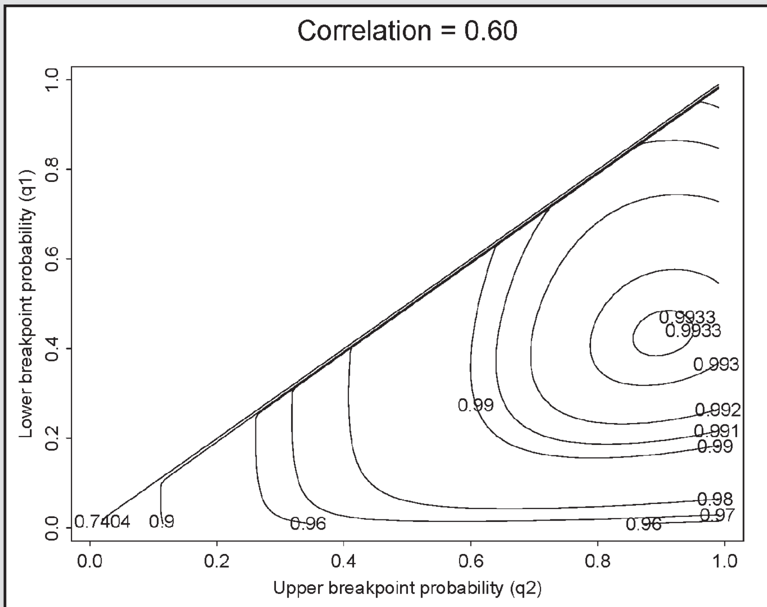
Forest Service

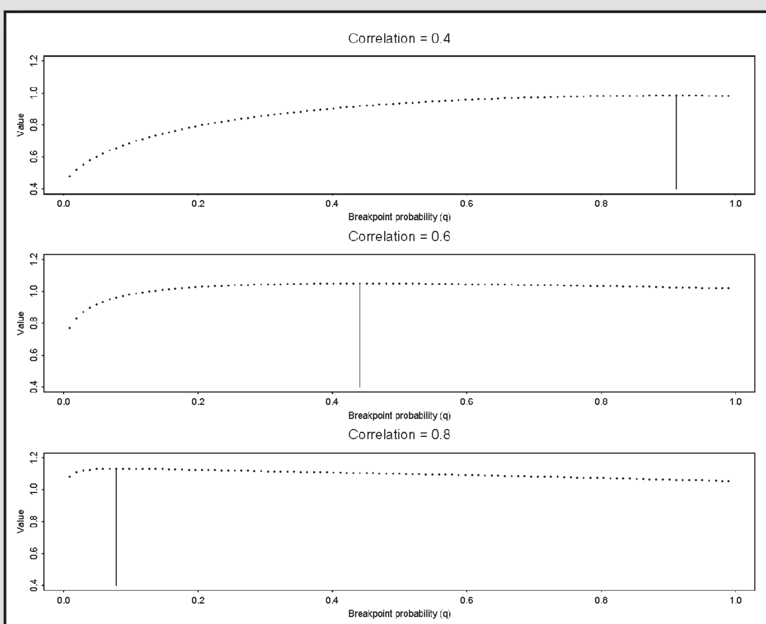
Forest Products Laboratory

Research Paper FPL–RP–651

Statistical Framework for Comparing Lumber Sorting Procedures

Steve Verrill
Richard A. Johnson
James W. Evans





Abstract

In this paper, we take a preliminary look at techniques for comparing sorts that yield grades of lumber. We propose methodology that takes into account differences in grade prices and the costs associated with misgrading. We focus on two- and three-category sorts, but our results could be readily extended. We provide web links to sample FORTRAN implementations of this methodology for the case in which the strength predictor and strength have a bivariate normal distribution and the load distribution is also normally distributed. We indicate how these approaches would have to be modified for other predictor, strength, and load distributions.

Keywords: lumber grading, correlation estimation, comparing sorting procedures, classification

Contents

Page

July 2009

Verrill, Steve; Johnson, Richard A.; Evans, James W. 2009. Statistical framework for comparing lumber sorting procedures. Research Paper FPL-RP-651. Madison, WI: U.S. Department of Agriculture, Forest Service, Forest Products Laboratory. 17 p.

A limited number of free copies of this publication are available to the public from the Forest Products Laboratory, One Gifford Pinchot Drive, Madison, WI 53726–2398. This publication is also available online at www.fpl.fs.fed.us. Laboratory publications are sent to hundreds of libraries in the United States and elsewhere.

The Forest Products Laboratory is maintained in cooperation with the University of Wisconsin.

Statistical Framework for Comparing Lumber Sorting Procedures

Steve Verrill, Mathematical Statistician
USDA Forest Products Laboratory, Madison, Wisconsin
Richard A. Johnson, Emeritus Professor of Statistics
University of Wisconsin, Madison, Wisconsin
James W. Evans, Mathematical Statistician
USDA Forest Products Laboratory, Madison, Wisconsin

1 Introduction

People working with wood and wood products are often confronted with sorting problems: To what end use category should a log be directed? What grade should be assigned to a piece of lumber? In what repair category should we place a component of an existing wood structure? Good sorting procedures obviously require performance predictors (for example, modulus of elasticity (MOE)) that are well correlated with performance (for example, modulus of rupture). However, an optimal sorting procedure will also take into account the costs associated with obtaining better predictors and the costs associated with misclassifications.

To make the most efficient use of our nation's timber supplies, we need sorting methods that are optimized. Thus we certainly need methods for comparing the quality of various sorting procedures. These methods must take into account the costs of the procedures and the economic consequences of the resulting sorts.

In this paper we take a preliminary look at techniques for comparing sorts that yield grades of structural products. We propose methodology that takes into account differences in grade prices and the costs associated with misgrading. The methodology has a decision theoretic flavor. (See, for example, Berger (1993).) We provide web links to sample FORTRAN implementations of this methodology for the case in which the strength predictor and strength have a bivariate normal distribution, and the load distribution is also normally distributed. We indicate how these approaches would have to be modified for other predictor, strength, and load distributions.

This methodology is raw, but we should be able to refine it to handle any special case. However, that refinement will need to depend on data — data that gives us better information about strength and load distributions, about prices associated with different binning schemes, about costs of the schemes, and about costs of failures.

Here we consider only two- and three-category grading schemes, but our methods are easily extended to schemes that involve more than three categories.

In the material below, in accord with statistical usage, we use the term "bin" rather than the term "category."

2 Maximizing Expected Value

In this paper, we focus on a situation in which competing sorts can involve different numbers of bins and different bin cutoffs. For a given sort, the expected value associated with the sort is calculated as (here, we are neglecting the cost of the sorting procedure)

Prob(lumber placed in bin 1) $\times$ (bin 1 price) $+ \ldots +$ Prob(lumber placed in bin k) $\times$ (bin k price)
$+$Prob(lumber placed in bin 1 and fails) $\times$ (cost of bin 1 failure) $+ \ldots$
$+$Prob(lumber placed in bin k and fails) $\times$ (cost of bin k failure)

The mathematical and numerical difficulties lie in evaluating the probabilities. The practical difficulties lie in evaluating the prices and costs. In this paper we focus on the (simpler) mathematical problem.

In Section 3 we consider the two-bin case and calculate the probability of failure given that a piece of lumber is placed in the first bin and the probability of failure given that a piece is placed in the second bin. To calculate these probabilities, we first find the fifth percentile of the strengths of the pieces of lumber placed in a bin. We divide this value by the factor 2.1 (see Section 3) and take this as the "allowable property" (see Section 3) associated with the bin. This then leads to an associated distribution of acceptable loads. Combining information about the distribution of strengths associated with a bin together with the distribution of loads permitted for that bin, we obtain a probability of failure for that bin. Given probability of failure values, and assumed prices and costs, we can then calculate the expected value associated with a sorting procedure.

In Section 4 we consider the three-bin case.

In both Sections 3 and 4 we assume that the joint distribution of the strength predictor and the strength is bivariate normal. However, as we note in Section 5, our results are easily extended to the case in which the strength distribution is lognormal.

3 Bivariate Normal, Two-Bin Case

Let Y denote the strength, and let X denote the strength predictor used to sort wood specimens into bins. We assume that X and Y have a joint bivariate normal distribution. X could be explicit (for example, an MOE measurement), or it could be implicit in the judgment of a human grader.

Let x_{b} denote the breakpoint. Thus lumber pieces for which $X \leq x_{\mathrm{b}}$ will be placed in the lower bin. Lumber pieces for which $X > x_{\mathrm{b}}$ will be placed into the upper bin. x_{b} is the qth quantile of the predictor distribution for some q. That is,

$$\mathrm{Prob}(X \leq x_{\mathrm{b}}) = q$$

where X is a normally distributed random variable with mean μ_x and standard deviation σ_x. Thus

$$\mathrm{Prob}\left(\frac{X-\mu_x}{\sigma_x} \leq \frac{x_{\mathrm{b}}-\mu_x}{\sigma_x}\right) = q$$

or

$$\frac{x_{\mathrm{b}}-\mu_x}{\sigma_x} = \Phi^{-1}(q)$$

where Φ denotes the standard normal cumulative distribution function (cdf).

Probability Density Function (pdf) of the Lower Bin Strengths

The cdf of the lower bin items is given by

$$\mathrm{Prob}(Y \leq y | X \leq x_{\mathrm{b}}) \quad = \quad \int_{-\infty}^{y} \int_{-\infty}^{x_{\mathrm{b}}} \text{bivariate normal pdf } ds\, dt/q$$

$$
\begin{aligned}
&= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp\left(-0.5 \times (t-\mu_y)^2/\sigma_y^2\right) \\
&\quad \times \Phi\left(\left[\Phi^{-1}(q) - \rho\left(\frac{t-\mu_y}{\sigma_y}\right)\right] / \sqrt{1-\rho^2}\right) dt/q
\end{aligned}
$$

so the pdf of the lower bin strengths is

$$
\begin{aligned}
f_1(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp\left(-0.5 \times (y-\mu_y)^2/\sigma_y^2\right) \\
&\quad \times \Phi\left(\left[\Phi^{-1}(q) - \rho\left(\frac{y-\mu_y}{\sigma_y}\right)\right] / \sqrt{1-\rho^2}\right) /q
\end{aligned}
$$

where q is the probability that a piece of lumber is placed in the lower bin, μ_y and σ_y are the mean and standard deviation of the strength distribution, and ρ is the correlation between the strength predictor and strength.

The fifth percentile, $y_{.05,1}$, of the lower bin of strengths can be found by solving the equation

$$
.05 = \int_{-\infty}^{y_{.05,1}} f_1(y)\, dy
$$

This only requires a normal cdf inverse routine,[1] a normal cdf routine,[2] a 1-d numerical integration routine (we use the SLATEC[3] routine dqags), and a zero-finding routine (we use the MINPACK[4] routine hybrd1).

Note that there can be numerical difficulties here. The numerical integration routine yields a result that is accurate only up to a tolerance set by the user. Thus, the result will not be a perfectly smooth function of the fifth percentile. Theoretically, the zero-finding routine assumes smoothness of the function for which a zero is being determined. However, the disconnect between the properties of the result produced by the numerical integration and the properties needed by the zero-finder can be bridged by requiring a tolerance for the numerical integration that is smaller than the tolerance required of the zero finder. In our (double precision) programs, we used a 0.0000001 tolerance for the numerical integration program and a 0.0001 tolerance for the zero-finder.

The mean of the lower bin strengths is given by

$$
\text{mean}_1 = \int_{-\infty}^{\infty} y f_1(y)\, dy
$$

This requires a normal cdf inverse routine, a normal cdf routine, and a 1-d numerical integration routine.

Probability of a Lower Bin Failure

Here we must make a series of assumptions. Our computer program permits these to be altered. First we need to input a factor (for example, 2.1[5]) by which the fifth percentile of the strength distribution is divided to obtain an allowable property. Then we assume that this allowable property is equal to a specified quantile (for example, the 0.99 quantile) of the load distribution. (In practice

[1]See, for example, norminv.f at `http://www1.fpl.fs.fed.us/sortsim.html`

[2]See, for example, norcdf2.f at `http://www1.fpl.fs.fed.us/sortsim.html`

[3]See, for example, `http://www.netlib.org/slatec/`

[4]See, for example, `http://www.netlib.org/minpack/`

[5]This factor appears, for example, in the ASTM standards D 1990 and D 2915 (ASTM International 2007a,b).

this means that if the allowable property calculated for a bin is less than the chosen quantile of a known load distribution, then specimens from that bin are not certified for use in situations that see that load distribution.) This gives us one constraint on the load distribution. For two-parameter load distributions, we need one other constraint to specify the distribution completely. For example, if the load distribution is normal, and we are given the associated coefficient of variation (cv), then we can calculate the mean, $\mu_{\mathrm{L},1}$, and variance, $\sigma_{\mathrm{L},1}$, of the load distribution:

Setting

$$y_{.05,1}/2.1 = \mu_{\mathrm{L},1} + \Phi^{-1}(0.99) \times \mu_{\mathrm{L},1} \times \mathrm{cv}$$

we have

$$\mu_{\mathrm{L},1} = y_{.05,1}/\left(2.1\left(1 + \Phi^{-1}(0.99) \times \mathrm{cv}\right)\right)$$

Also, by the definition of coefficient of variation,

$$\sigma_{\mathrm{L},1} = \mu_{\mathrm{L},1} \times \mathrm{cv}$$

Now, assuming a normally distributed load distribution and statistical independence between the load and strength distributions, given that a piece of lumber is sorted into the lower bin, the probability of failure, $p_{\mathrm{F},1}$, is

$$\begin{aligned} p_{\mathrm{F},1} &= \int_{-\infty}^{\infty} f_1(y) \int_y^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_{\mathrm{L},1}} \exp\left(-0.5 \times (z - \mu_{\mathrm{L},1})^2/\sigma_{\mathrm{L},1}^2\right) dz\, dy \qquad (1) \\ &= \int_{-\infty}^{\infty} f_1(y) \left(1 - \Phi\left((y - \mu_{\mathrm{L},1})/\sigma_{\mathrm{L},1}\right)\right) dy \end{aligned}$$

To perform this calculation we need a normal cdf inverse routine, a normal cdf routine, and a 1-d numerical integration routine.

Upper Bin Strengths and Probability of Failure

We argue as we did in the lower bin case to obtain the pdf of the upper bin strengths:

$$\begin{aligned} f_2(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp(-0.5 \times (y - \mu_y)^2/\sigma_y^2) \\ &\quad \times \left[1 - \Phi\left(\left[\Phi^{-1}(q) - \rho\left(\frac{y - \mu_y}{\sigma_y}\right)\right] / \sqrt{1 - \rho^2}\right)\right] / (1 - q) \end{aligned}$$

The fifth percentile, $y_{.05,2}$, of the upper bin of strengths can be found by solving the equation

$$.05 = \int_{-\infty}^{y_{.05,2}} f_2(y)\, dy$$

This requires a normal cdf inverse routine, a normal cdf routine, a 1-d numerical integration routine, and a zero-finding routine.

The mean of the upper bin strengths is given by

$$\mathrm{mean}_2 = \int_{-\infty}^{\infty} y f_2(y)\, dy$$

This requires a normal cdf inverse routine, a normal cdf routine, and a 1-d numerical integration routine.

Making the same assumptions as in the lower bin case, we have

$$\mu_{\mathrm{L},2} = y_{.05,2} / \left(2.1\left(1+\Phi^{-1}(0.99)\mathrm{cv}\right)\right)$$

and

$$\sigma_{\mathrm{L},2} = \mu_{\mathrm{L},2} \times \mathrm{cv}$$

and, given that a piece of lumber is sorted into the upper bin, the probability of failure, $p_{\mathrm{F},2}$, is

$$\begin{aligned} p_{\mathrm{F},2} &= \int_{-\infty}^{\infty} f_2(y) \int_y^{\infty} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_{\mathrm{L},2}} \exp\left(-0.5 \times (z-\mu_{\mathrm{L},2})^2/\sigma_{\mathrm{L},2}^2\right) dz\, dy \\ &= \int_{-\infty}^{\infty} f_2(y)(1-\Phi\left((y-\mu_{\mathrm{L},2})/\sigma_{\mathrm{L},2}\right)) dy \end{aligned}$$

To perform this calculation we need a normal cdf inverse routine, a normal cdf routine, and a 1-d numerical integration routine.

Value

Let P_1 denote the price of a lower bin lumber piece and P_2 denote the price of an upper bin lumber piece. Let C_1 denote the cost of a lower bin failure and C_2 the cost of an upper bin failure. Then the expected value, V, per piece of lumber from this sort is given by (here, we are neglecting the cost of the sorting procedure)

$$V = (q \times P_1) + ((1-q) \times P_2) - (q \times p_{\mathrm{F},1} \times C_1) - ((1-q) \times p_{\mathrm{F},2} \times C_2)$$

The computer program must permit a user to specify P_1, P_2, C_1, and C_2. Ideally the user should be able to specify them as functions of q. Currently we have implemented a necessarily (in the absence of data and/or the assistance of an economist) unrealistic approach. We take as the unit of value the price that one would receive per piece of lumber for a lot of lumber that had mean strength equal to the population mean. Then P_1 is taken to be the ratio of lower bin mean strength to population mean strength. P_2 is taken to be the ratio of upper bin mean strength to population mean strength. C_1 and C_2 were taken to be, for example, 1000 and 10000, reflecting the fact that upper bin failures would have more serious consequences than lower bin failures, and that failures could have repercussions that were much more serious than the cost of a simple replacement of a piece of lumber.

Note that when we take price per piece of lumber to be proportional to bin mean strength, then the $qP_1 + (1-q)P_2$ portion of V does not depend on q. That is,

$$q\int_{-\infty}^{\infty} y f_1(y)\, dy/\mu_y + (1-q)\int_{-\infty}^{\infty} y f_2(y)\, dy/\mu_y = \int_{-\infty}^{\infty} (\mu_y + \sigma_y y)\frac{1}{\sqrt{2\pi}}\exp(-y^2/2) dy/\mu_y = 1$$

regardless of the value of q. Thus, in this case, value is dependent on q only through $-qp_{\mathrm{F},1}C_1 - (1-q)p_{\mathrm{F},2}C_2$.

4 Bivariate Normal, Three-Bin Case

This is a simple extension of the two bin case. Let $x_{\mathrm{b},1} < x_{\mathrm{b},2}$ be the strength predictor breakpoints and $q_1 < q_2$ be the corresponding quantiles, so that

$$\mathrm{Prob}(X \le x_{\mathrm{b},i}) = q_i$$

and

$$(x_{\mathrm{b},i} - \mu_x)/\sigma_x = \Phi^{-1}(q_i)$$

where Φ denotes the standard normal cumulative distribution function. The probability density function (pdf) of the bin 1 ($X \leq x_{\mathrm{b},1}$) strengths is

$$\begin{aligned} f_1(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp(-0.5 \times (y - \mu_y)^2/\sigma_y^2) \\ &\quad \times \Phi \left(\left[\Phi^{-1}(q_1) - \rho \left(\frac{y - \mu_y}{\sigma_y} \right) \right] / \sqrt{1 - \rho^2} \right) / q_1 \end{aligned}$$

The pdf of the bin 2 ($x_{\mathrm{b},1} < X \leq x_{\mathrm{b},2}$) strengths is

$$\begin{aligned} f_2(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp(-0.5 \times (y - \mu_y)^2/\sigma_y^2) \\ &\quad \times \left[\Phi \left(\left[\Phi^{-1}(q_2) - \rho \left(\frac{y - \mu_y}{\sigma_y} \right) \right] / \sqrt{1 - \rho^2} \right) - \Phi \left(\left[\Phi^{-1}(q_1) - \rho \left(\frac{y - \mu_y}{\sigma_y} \right) \right] / \sqrt{1 - \rho^2} \right) \right] / (q_2 - q_1) \end{aligned}$$

The pdf of the bin 3 ($x_{\mathrm{b},2} < X$) strengths is

$$\begin{aligned} f_3(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp(-0.5 \times (y - \mu_y)^2/\sigma_y^2) \\ &\quad \times \left[1 - \Phi \left(\left[\Phi^{-1}(q_2) - \rho \left(\frac{y - \mu_y}{\sigma_y} \right) \right] / \sqrt{1 - \rho^2} \right) \right] / (1 - q_2) \end{aligned}$$

The fifth percentiles, $y_{.05,1}, y_{.05,2}$, and $y_{.05,3}$, of the three strength populations can be found by solving the equation

$$.05 = \int_{-\infty}^{y_{.05,i}} f_i(y)\, dy$$

for $i = 1, 2, 3$.

This requires a normal cdf inverse routine, a normal cdf routine, a 1-d numerical integration routine, and a zero-finding routine.

The means of the three bin strength populations are given by

$$\mathrm{mean}_i = \int_{-\infty}^{\infty} y f_i(y)\, dy$$

for $i = 1, 2, 3$. This requires a normal cdf inverse routine, a normal cdf routine, and a 1-d numerical integration routine.

The probabilities of failure (here we are assuming normally distributed loads) associated with the three populations are given by

$$p_{\mathrm{F},i} = \int_{-\infty}^{\infty} f_i(y)(1 - \Phi\left((y - \mu_{\mathrm{L},i})/\sigma_{\mathrm{L},i}\right))dy$$

for $i = 1, 2, 3$, where the $\mu_{\mathrm{L},i}$'s and $\sigma_{\mathrm{L},i}$'s are calculated as in the two-bin case (by setting the (for example) 0.99 quantile of the load distribution equal to the fifth percentile of the strength distribution divided by (for example) 2.1 and by assuming a particular value for the load distribution coefficient of variation). To perform these calculations we need a normal cdf inverse routine, a normal cdf routine, and a 1-d numerical integration routine.

The value in the three bin case has expectation (here, we are neglecting the cost of the sorting procedure)

$$V = q_1 P_1 + (q_2 - q_1) P_2 + (1 - q_2) P_3 - q_1 p_{\mathrm{F},1} C_1 - (q_2 - q_1) p_{\mathrm{F},2} C_2 - (1 - q_2) p_{\mathrm{F},3} C_3$$

where the P_i's are the prices associated with the three bins, the $p_{\mathrm{F},i}$'s are the probabilities of failure, and the C_i's are the costs of failure.

Note that, as in the two bin case, if we take price per piece of lumber to be proportional to bin mean strength then the $q_1 P_1 + (q_2 - q_1) P_2 + (1 - q_2) P_3$ portion of V does not depend on q_1 and q_2.

5 Lognormal/Normal Case

Here we consider the case in which $\ln(Y)$ (the log of the strength) and X (the untransformed predictor) have a bivariate normal distribution. The point to be made is that, numerically, this is no more difficult a case than the pure bivariate normal case. For example, in this case the bin 1 pdf is simply

$$g_1(y) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_z} \exp(-0.5 \times (\ln(y) - \mu_z)^2 / \sigma_z^2) \times \frac{1}{y} \times \Phi\left(\left[\Phi^{-1}(q) - \rho\left(\frac{\ln(y) - \mu_z}{\sigma_z}\right)\right] / \sqrt{1 - \rho^2}\right) / q$$

where $y > 0$, and μ_z, σ_z^2 are the expectation and variance of $\ln(Y)$. Given this pdf, the estimates of fifth percentile, mean, and probability of failure can be obtained by using the same numerical analysis routines used in the pure bivariate normal case.

6 Non-Normal Load Distributions

Because we are assuming that the load and strength distributions are independent, the load distribution simply enters as a multiplier (as in Equation 1) so there is no new numerical difficulty (we still need only 1-d numerical integration routines) regardless of the nature of the load distribution. Of course the method of specifying the load distribution might need to change. We would still set the 0.99 (for example) quantile of the load distribution equal to a fifth percentile of the strength distribution divided by 2.1 (for example) and then impose a second constraint (for a two-parameter load distribution) but the constraint might not be the coefficient of variation constraint used in the normal case.

7 Pre-Specified Load Distributions

In the approach described above we make use of a "maximum permissible load distribution" that is completely specified by the fifth percentile of the strength distribution and one other assumption (for a two-parameter load distribution) about a parameter of the load distribution. However, one could take an alternative (and presumably more realistic) approach. One could specify a load distribution that is based on real world wind/snow/... loadings. This load distribution would not in general have its 0.99 (for example) quantile matched with the allowable property value obtained from a bin strength distribution.

If we did take such an approach, however, it would not complicate the numerics at all. The load distribution would still just yield a multiplier in the integrand of the 1-d numerical integration that yields a probability of failure (see, for example, Equation 1).

8 Estimating the Correlation, ρ, between a Predictor and Strength

To perform the calculations specified in the preceding sections, we need an estimate of the correlation, ρ, between the strength predictor and the strength. In the case in which the predictor X is explicit, the calculation of this estimate is trivial.[6] When X is implicit[7] one can use maximum likelihood techniques (Cox, 1974).

In the two-bin, implicit Normal/Normal case we can take X to be N(0,1). There are four parameters to be estimated — μ_y, σ_y^2, q, and ρ where the implicit breakpoint is at $x_{\mathrm{b}} = \Phi^{-1}(q)$, i.e., specimens are placed in bin 1 if $x \leq x_{\mathrm{b}}$ and in bin 2 if $x > x_{\mathrm{b}}$. For the expected value work, we need estimates of μ_y, σ_y^2, q, and ρ. The data is of the form

$$x \leq x_{\mathrm{b}}, y = y_1, \ldots, x \leq x_{\mathrm{b}}, y = y_{n_1}, x > x_{\mathrm{b}}, y = y_{n_1+1}, \ldots, x > x_{\mathrm{b}}, y = y_{n_1+n_2}$$

The corresponding likelihood is

$$\prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp(-(y_i - \mu_y)^2/(2\sigma_y^2)) \Phi \left(\frac{\Phi^{-1}(q) - \rho(y_i - \mu_y)/\sigma_y}{\sqrt{1-\rho^2}} \right)$$

$$\times \prod_{i=n_1+1}^{n_1+n_2} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma_y} \exp(-(y_i - \mu_y)^2/(2\sigma_y^2)) \left(1 - \Phi \left(\frac{\Phi^{-1}(q) - \rho(y_i - \mu_y)/\sigma_y}{\sqrt{1-\rho^2}} \right) \right)$$

This can be maximized to yield estimates of the four parameters. To do so we need a normal cdf inverse routine, a normal cdf routine, and an unconstrained nonlinear optimization routine (we use UNCMIN, see `http://www1.fpl.fs.fs.us/optimization.html`). We have checked this in both the two- and three-bin cases (in the three-bin case, the likelihood has a third factor that corresponds to the middle bin) with simulation programs. The approach yields reasonable estimates of the parameters in the cases that we considered. (We used a sample size of 100. For small sample sizes, the approach might not work as well.) The FORTRAN programs, two.f and three.f, used to perform the simulations can be viewed at `http://www1.fpl.fs.fed.us/sortsim.html`.

9 Estimating the Correlation between Two Predictors — Aside

This section is really in the nature of an aside. If we choose to compare sorting schemes on the basis of expected value, then we do not need to calculate the correlations between different strength predictors. However, to estimate value it is necessary to determine load distributions, the costs of running different sorting schemes, prices as functions of sorting schemes, and costs due to failures. Given the difficulty of these tasks, one might choose to measure the similarity of two sorting schemes by the correlation between the strength predictors used in those schemes. If both strength predictors were explicit (for example, two different measures of stiffness), then this correlation calculation would be trivial. If one or both of the strength predictors were implicit, then we would have to be more careful. However, in principle, one should be able to generalize the maximum likelihood calculation described in the preceding section. In this case we assume that strength and the two predictors have a trivariate normal distribution. As an example, for two implict predictors and sorting schemes that have two bins, the data would be of the form

$$x_1 \leq x_{\mathrm{b},1}, x_2 \leq x_{\mathrm{b},2}, y = y_1, \ldots, x_1 \leq x_{\mathrm{b},1}, x_2 \leq x_{\mathrm{b},2}, y = y_{n_1}$$

[6] That is, $\hat{\rho} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$.

[7] For example, X might represent the eyeball judgment of a human grader, and no predictor X value would be explicitly measured or calculated.

$$x_1 \le x_{\mathrm{b},1}, x_2 > x_{\mathrm{b},2}, y = y_{n_1+1}, \ldots, x_1 \le x_{\mathrm{b},1}, x_2 > x_{\mathrm{b},2}, y = y_{n_1+n_2}$$

$$x_1 > x_{\mathrm{b},1}, x_2 \le x_{\mathrm{b},2}, y = y_{n_1+n_2+1}, \ldots, x_1 > x_{\mathrm{b},1}, x_2 \le x_{\mathrm{b},2}, y = y_{n_1+n_2+n_3}$$

$$x_1 > x_{\mathrm{b},1}, x_2 > x_{\mathrm{b},2}, y = y_{n_1+n_2+n_3+1}, \ldots, x_1 > x_{\mathrm{b},1}, x_2 > x_{\mathrm{b},2}, y = y_{n_1+n_2+n_3+n_4}$$

The corresponding likelihood is

$$\prod_{i=1}^{n_1} \int_{-\infty}^{x_{\mathrm{b},1}} \int_{-\infty}^{x_{\mathrm{b},2}} f(y_i, z_1, z_2) dz_2\, dz_1$$

$$\times \prod_{i=n_1+1}^{n_1+n_2} \int_{-\infty}^{x_{\mathrm{b},1}} \int_{x_{\mathrm{b},2}}^{\infty} f(y_i, z_1, z_2) dz_2\, dz_1$$

$$\times \prod_{i=n_1+n_2+1}^{n_1+n_2+n_3} \int_{x_{\mathrm{b},1}}^{\infty} \int_{-\infty}^{x_{\mathrm{b},2}} f(y_i, z_1, z_2) dz_2\, dz_1$$

$$\times \prod_{i=n_1+n_2+n_3+1}^{n_1+n_2+n_3+n_4} \int_{x_{\mathrm{b},1}}^{\infty} \int_{x_{\mathrm{b},2}}^{\infty} f(y_i, z_1, z_2) dz_2\, dz_1$$

where $f(r, s, t)$ is the pdf of a trivariate normal distribution with $\mu_2 = \mu_3 = 0$ and $\sigma_2 = \sigma_3 = 1$. Thus

$$f(r, s, t) = \frac{1}{(2\pi)^{3/2}} \frac{1}{\sigma_y \sqrt{1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23}}} \exp(-\mathrm{arg})$$

where

$$\mathrm{arg} = \begin{pmatrix} (r - \mu_y)/\sigma_y \\ s \\ t \end{pmatrix}^T A \begin{pmatrix} (r - \mu_y)/\sigma_y \\ s \\ t \end{pmatrix} /2$$

$$A = \begin{pmatrix} 1 - \rho_{23}^2 & \rho_{13}\rho_{23} - \rho_{12} & \rho_{12}\rho_{23} - \rho_{13} \\ \rho_{13}\rho_{23} - \rho_{12} & 1 - \rho_{13}^2 & \rho_{12}\rho_{13} - \rho_{23} \\ \rho_{12}\rho_{23} - \rho_{13} & \rho_{12}\rho_{13} - \rho_{23} & 1 - \rho_{12}^2 \end{pmatrix} /(1 - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 + 2\rho_{12}\rho_{13}\rho_{23})$$

ρ_{1j} is the correlation between the strength and predictor j, $j = 1, 2$, and ρ_{23} is the correlation between the two predictors. We would need to estimate μ_y, σ_y, $x_{\mathrm{b},1}$, $x_{\mathrm{b},2}$, ρ_{12}, ρ_{13}, and ρ_{23}. This would require an unconstrained nonlinear optimization routine and a 2-d numerical integration routine (see, for example, GAMS class H2b1a1 at `http://gams.nist.gov/serve.cgi/Class/H2b1a1/`).

10 Uncertainty in the Value of a Sort

The calculated value of a sort depends on μ_y, σ_y, the q's, ρ, the parameters of the load distribution, lumber prices, and the costs of failures. (Note that the cost of running the sorting scheme also needs to be substracted from the value of the sort. We have not done so in our sample programs.) There will be estimation uncertainties in μ_y, σ_y, the q's, and ρ. In our current approach the load distribution that a piece of lumber could see is specified exactly given the fifth percentile of the strength distribution and an assumption about a parameter of the load distribution. However, as noted in the remarks on a "pre-specified load distribtion," one could use a load distribution estimated from data. There would then be estimation uncertainties in the parameters of the load distribution.

The estimated uncertainties in the parameters propagate through to uncertainties in the value of the sort. Since our maximum likelihood estimation procedures would yield a covariance matrix for our estimates of μ_y, σ_y, q, and ρ, and we would also have a covariance matrix for the load distribution parameter estimates, we could use the "delta method" to obtain an approximate variance for the value estimate. We would obtain our estimates of the partial derivatives of value with respect to the parameters via simple numerical difference estimates.

11 Example Computer Programs

We provide example programs to estimate value at `http://www1.fpl.fs.fed.us/sortsim.html`. The programs are called loss2.gr.f and loss3.gr.f. The loss2.gr.f program handles the two bin case. The loss3.gr.f program handles the three-bin case. Both programs assume that prices are proportional to bin means. (This is done solely for the purpose of explication. In a more realistic case prices as functions of breakpoints would be determined by market data and an economist's analysis.) Thus in both cases the optimal breakpoints are dependent solely on the the failure costs and load distributions.

In Figure 1 we plot value versus q for predictor/strength correlations 0.4, 0.6, and 0.8 in the two-bin case. In these plots, the vertical lines mark the q that leads to a maximum value. It is clear that when there is a high correlation between predictor and strength (so we are unlikely to put a poor piece into the top bin and thus to incur the serious costs associated with a bin 2 failure) then it is worthwhile to lower q and put a higher percentage of the population into the upper bin. When ρ is low so we have a higher probability of mistakenly placing a poor piece into the upper bin, it is better to increase the q and to assign a smaller percentage of the pieces into the "high quality" bin. Here we are assuming that $\mu_y = 4000$, the strength coefficient of variation equals 0.2 so $\sigma_y = 800$, the divisor of the strength fifth percentile is 2.1, the load distribution is normal, the coefficient of variation of the load distribution is 0.15, the cost of a bin 1 failure is 1000 times the standard price of a piece of lumber, and the cost of a bin 2 failure is 10000 times the standard price. Also note that, as we would expect, as the correlation between the strength predictor and the actual strength increases, the maximum value of the sorted material increases. (As the correlation between predicted and actual strength increases, there are fewer misclassifications of poorer pieces into the upper bin so the costs associated with failures decrease.) Figure 2 corresponds to the same situation except that the costs associated with failure are now 500 and 50000 times the standard price of a piece of lumber (lower bin failures have less severe consequences than in the first case, upper bin failures have more serious consequences than in the first case). It is clear from Figure 2 that the increase in the ratio of the severities of the two kinds of failure now leads one into being more conservative. (For $\rho = 0.6$ or 0.8, the optimal q is higher in Figure 2 than in Figure 1. Note that the $\rho = 0.6$ plot is the middle plot in Figure 1 and the upper plot in Figure 2.)

In Figures 3 - 5 we plot contours of value versus q_1, q_2 for predictor, strength correlations 0.6, 0.7, and 0.8 in the three bin case. It is again clear that when ρ is higher the optimal breakpoints put a higher percentage of the pieces in the top bin, and when ρ is lower the optimal breakpoints put a higher percentage of the pieces in the lowest bin. Here we are again assuming that $\mu_y = 4000$, the strength coefficient of variation equals 0.2 so $\sigma_y = 800$, the divisor of the strength fifth percentile is 2.1, the load distribution is normal, and the coefficient of variation of the load distribution is 0.15. We also assume that the cost of a bin 1 failure is 100 times the standard board price of a piece of lumber, the cost of a bin 2 failure is 1000 times the price, and the cost of a bin 3 failure is 10000 times the price.

12 An Alternative Model

Because most lumber failures in light-frame structures will not lead to the catastrophic failure of the structure, it can be argued that there will generally be no cost to a manufacturer for the failure of a piece of lumber. Also, it can be argued that there is little flexibility in bin boundaries. Instead, for example, machine stress rated 1800f 1.6E lumber corresponds to a relatively fixed quantile of the strength/MOE distribution, and a producer cannot easily alter this boundary. Instead a producer's product must regularly pass screening trials. For example, periodically, the producer must sample n pieces of lumber and proof-load them at a stress level L. If more than k of the pieces fail, then the producer's product must undergo a second round of testing. If the product fails this second round of tests, the costs to the producer can be considerable ("production stops, a bigger sample is required to reestablish boundaries, and more quality control samples might be taken for a while after production restarts"[8]). Under this scenario, the costs to a producer associated with a given set of predictor boundaries are the costs associated with qualification testing and the costs associated with setting a predictor qualification value high so that fewer lumber pieces qualify as high price pieces. If a producer raises the bar, the producer will receive a lower price for some of its lumber, but if the producer sets the bar too low, the producer will be forced to bear certification costs too frequently and, potentially, to deal with irritated customers (not, perhaps, because of failures but because of discards).

The optimal approach to take under this model will depend on the prices that can be charged for the different lumber quality categories and the costs, in customer satisfaction and/or additional testing, that are incurred because of poor lumber properties or qualification test failures.

The methods in the current paper can be extended to this new situation. However, the probabilities of failure will no longer depend on in-service load distributions. Instead they will depend on probabilities associated with qualification tests. For example, the probability that an individual bin 2 piece of lumber will fail in a qualification run will be

$$p_{\mathrm{F}} = \int_{L}^{\infty} f_2(y)dy$$

and the associated probability that a manufacturer will fail a qualification test is

$$1 - \mathrm{Prob}(k \text{ or fewer failures})$$

$$= 1 - \sum_{i=0}^{k} \begin{pmatrix} n \\ i \end{pmatrix} (1 - p_{\mathrm{F}})^{n-i} \times p_{\mathrm{F}}^{i}$$

We will pursue this approach in detail in a subsequent paper.

13 Summary

To compare lumber grading schemes, we have developed raw methodology that we should be able to refine to handle any special case. However, that refinement will need to depend on data — data that gives us better information about predictor, strength, and load distributions, about lumber prices associated with different binning schemes, about costs of running different schemes, and about costs of failures.

[8]Conversation with Dr. David W. Green, Emeritus Supervisory Research General Engineer, USDA Forest Products Laboratory

REFERENCES

ASTM. 2007a. Standard practice for establishing allowable properties for visually graded dimension lumber from in-grade tests of full-size specimens. D 1990-00. West Conshohocken, PA: ASTM International.

ASTM. 2007b. Standard practice for evaluating allowable properties for grades of structural lumber. D 2915-03. West Conshohocken, PA: ASTM International.

Berger, J.O. 1993. *Statistical Decision Theory and Bayesian Analysis*. New York: Springer. 617 p.

Cox, N.R. 1974. Estimation of the correlation between a continuous and a discrete variable. *Biometrics* **30**: 171–178.

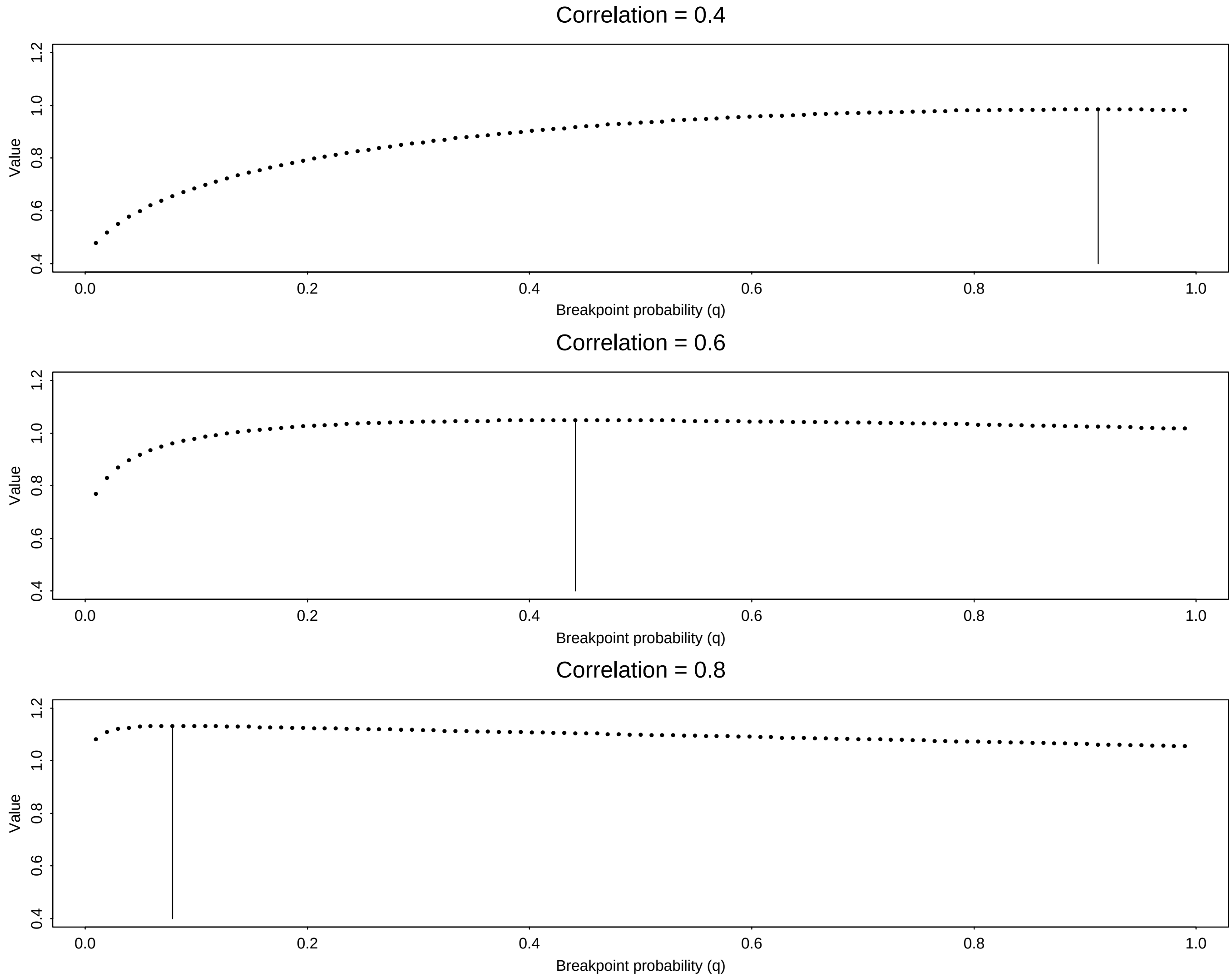


Figure 1: Plots of the sort value versus the probability (q) associated with the breakpoint for correlations 0.4, 0.6, and 0.8, and misclassification multipliers 1000 and 10000

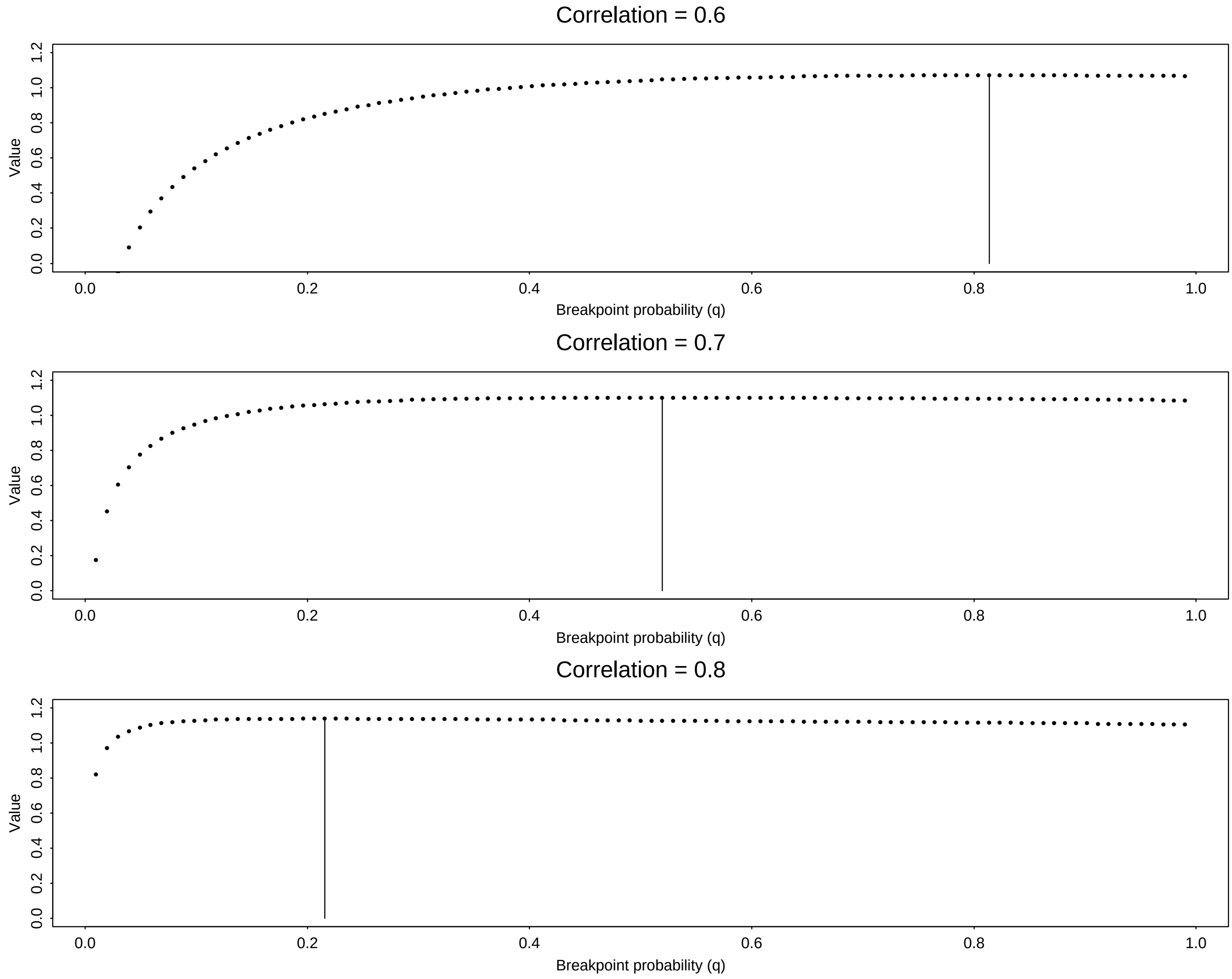


Figure 2: Plots of the sort value versus the probability (q) associated with the breakpoint for correlations 0.6, 0.7, and 0.8, and misclassification multipliers 500 and 50000

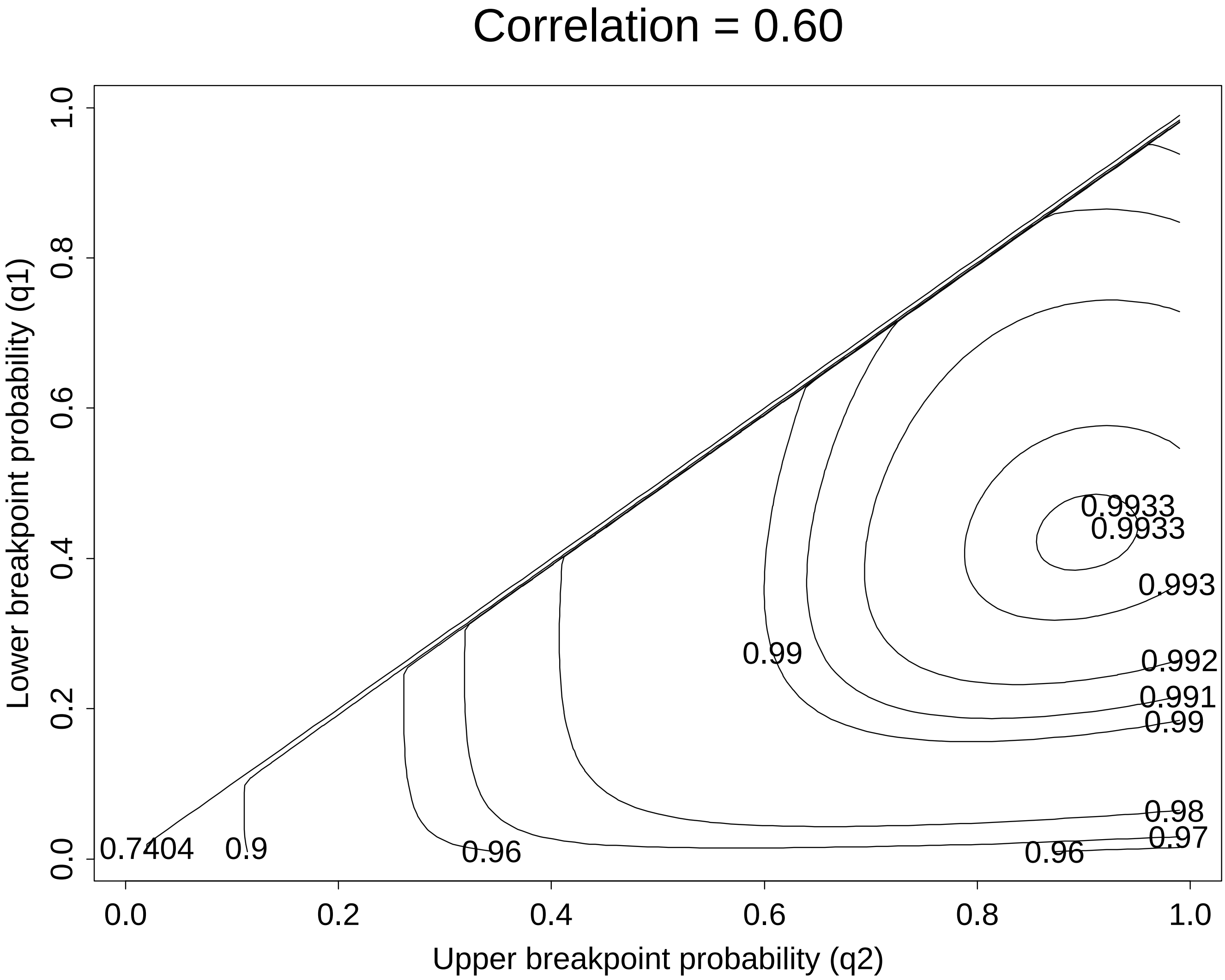


Figure 3: Contour plot of value versus lower breakpoint probability (q_1) and upper breakpoint probability (q_2) for predictor/strength correlation 0.6, and misclassification multipliers 100, 1000, and 10000. The maximum value in this case is attained when $q_1 = 0.43$ and $q_2 = 0.90$; that is, when the lower breakpoint is set at the 43rd percentile of the predictor distribution and the upper breakpoint is set at the 90th percentile of the predictor distribution.

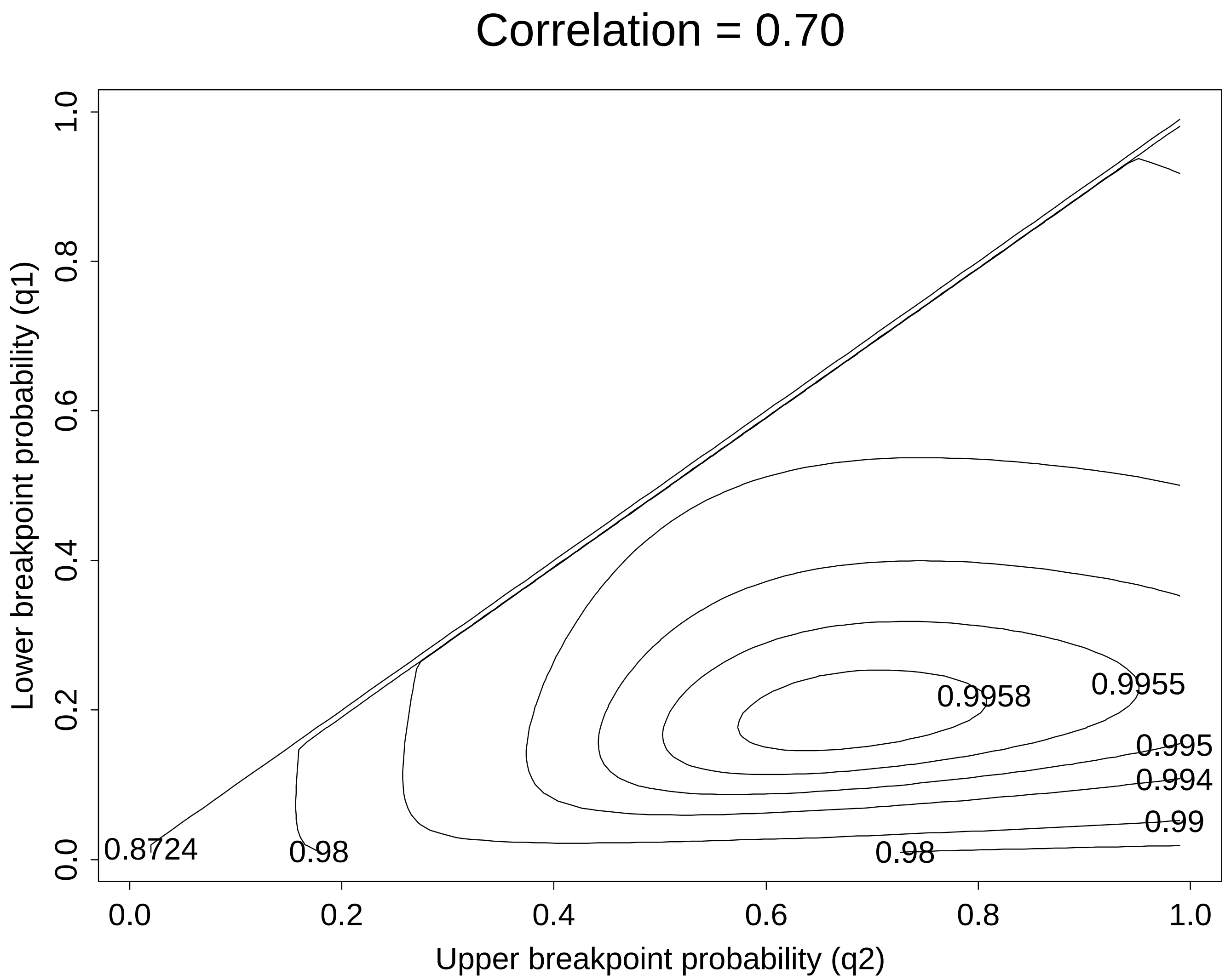


Figure 4: Contour plot of value versus lower breakpoint probability (q_1) and upper breakpoint probability (q_2) for predictor/strength correlation 0.7, and misclassification multipliers 100, 1000, and 10000. The maximum value in this case is attained when $q_1 = 0.20$ and $q_2 = 0.68$; that is, when the lower breakpoint is set at the 20th percentile of the predictor distribution and the upper breakpoint is set at the 68th percentile of the predictor distribution.

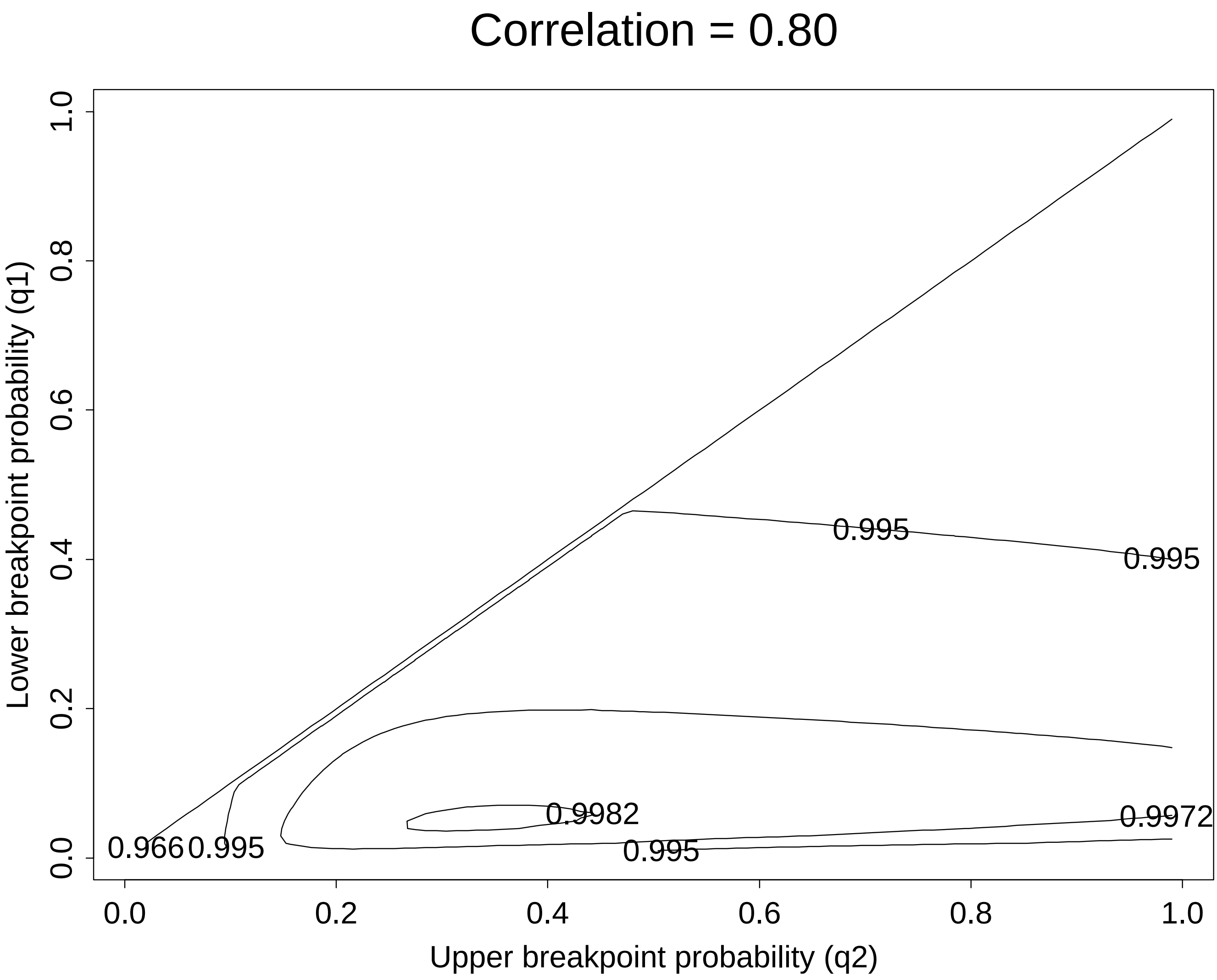


Figure 5: Contour plot of value versus lower breakpoint probability (q_1) and upper breakpoint probability (q_2) for predictor/strength correlation 0.8, and misclassification multipliers 100, 1000, and 10000. The maximum value in this case is attained when $q_1 = 0.05$ and $q_2 = 0.33$; that is, when the lower breakpoint is set at the 5th percentile of the predictor distribution and the upper breakpoint is set at the 33rd percentile of the predictor distribution.