CWPA Proceedings, 2009, 105-119

Prepared for CWPA, October 2009

Research Needs in Wood Preservation – Thoughts on In-plant Quality Control

Scott W. Conklin
Universal Forest Products, Inc.

Executive Summary

The variability of wood is very large. Some basic concepts from statistics are used to quantify this variability. Data show relative standard deviation of wood cores to be in excess of 40%. The current quality control requirements – standardized over 40 years ago – do not adequately deal with this variability and result in very inconsistent results. Improved quality control based on greater numbers of samples are needed to improve the quality of treated wood products produced under today's standards.

Our primary tool for determining if a charge of wood meets a particular standard is to take a prescribed number of samples using a boring bit and test those samples. This is true of American Wood Protection Association (AWPA) standards, ICC-ES Evaluation Reports (ESR) and many Canadian Standards Association (CSA) standards. Because wood is not a uniform, homogenous material it is possible, just possible, that each sample may be a bit different. This variability is of critical importance to (1) the way treating plant operators pass and fail a charge, (2) the way treating companies and/or preservative suppliers develop warranties and (3) the way the entire industry provides quality products to the industry. However, this variability and its implications are not well understood.

The goal of this paper is to develop some rudimentary statistical concepts, apply those concepts to samples obtained from treated wood, and suggest ways to perhaps improve the systems currently used by most of our industry to "pass" or "fail" charges of treated wood.

This paper will focus on testing for retention, but the concepts are equally true for penetration testing.

Basic Concepts from Statistics

Let me first say my education was in Chemical Engineering and I have spent my entire adult career in the wood treating business. I am not, and will never claim to be, a statistician. Hopefully, in the following I will keep to concepts basic enough to not get myself in too much trouble.

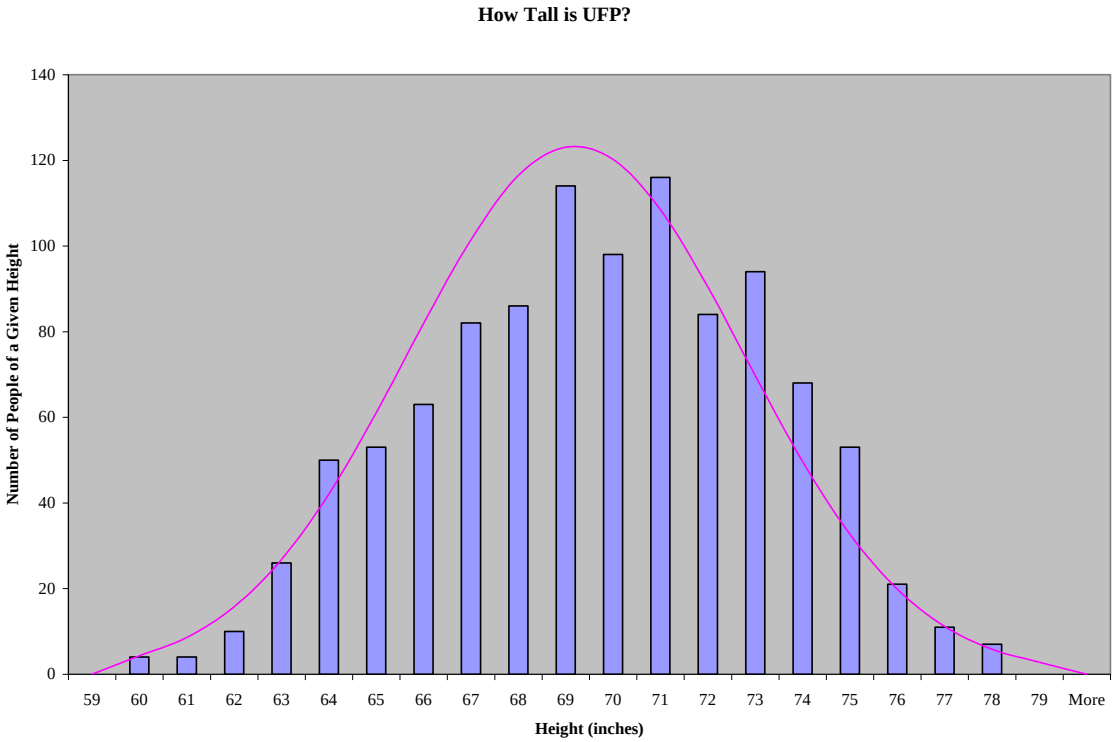
Rather than jumping right into a discussion of wood samples and assay results, I'd like to develop the basic concepts using a more intuitive set of data. Consider a data set based on how tall people are, specifically the height of people that work for Universal Forest Products. I asked each person in our corporate office and field offices to tell me how tall they are. In all, I have the height of 1,044 people. This is the *population* we are going to talk about. Obviously, everyone in our population, everyone in our offices, is not the same height; some are tall, some are short. Statistics provides tools to both describe the population and draw inferences about the population.

There is a very typical pattern associated with the kind of variation we see when talking about how tall people are. It is called a *normal distribution* and when we graph it we get a *bell curve*. Generally, it shows that heights cluster around the average height for the population and the further you get away from the average, the less people there are who have that height. In our population, the average height was just over 5 foot 9 inches (69.2 inches to be exact). There are lots of people who are right around 5'9", quite a few who are 5'11", not nearly as many who are 6'2" and darn few who are 6'6".

Histograms are a type of graph that allows us to visualize the variation that exists in a given population. Graph 1. Each bar in the graph represents the number of people (shown on the left or "y axis") who are a given height (shown on the bottom or "x axis"). Now, it is not a perfect bell shape as shown by the red line because (1) heights are "roughly" normally distributed[1], and (2) even at 1,000 plus, this is a small population.

[1] Entry for "Normal Distribution", Wikipedia, 8-20-2009

Graph 1 - UFP Height Histogram



Now graphs are great, but we need to be able to describe the population with numbers to really be useful. So…here comes the statistics!

Population: We have already used this term. It refers to the collection of values being discussed. It this example, the "population" is the height of each individual at UFP.
Normal Distribution: This is also called a Gaussian distribution. It can be expressed by a formula which depends only on the mean and the standard deviation. When graphed in a histogram, it produces a bell curve.
Average: This is the same as our every-day understanding of the term. For our population, the average height is 69.2 inches. In statistics, this is more properly called the "mean."
Standard Deviation, S: A measure of the variability or dispersion. A small standard deviation means most values are very close to the average value; a large standard deviation means the values are spread out. When graphed, small standard variation gives you a tall skinny peak; large standard variation gives you a low fat peak. For our population, the standard deviation is 3.6 inches.
Standard Deviation of the Sample Average, S_{ave}: Often the data we have are the average values from different samples of the population. This is exactly like the Standard Deviation, *S*, except the data are sample averages rather than direct samples of the population itself.
Sample Size, n: The number of samples from the population which are combined and averaged to produce S_{ave}.

Table 1 presents the average and standard deviation for our 1,044 employees' heights broken into different sample sizes. So a sample size of two shows what happens when we pair everyone up and only report their average height. Notice that while the average calculated for the full population is exactly the same, the standard deviation gets smaller as the sample size gets bigger.

Table 1 – Effect of Sample Size on Distribution

n, sample:	**1**	**2**	**3**	**5**	**10**
count:	1,044	522	348	208	104
average:	69.2	69.2	69.2	69.2	69.2
max:	78.0	75.6	73.9	73.4	71.5
min:	59.6	61.9	63.3	65.0	66.3
S_{ave}, stdev:	3.56	2.54	1.95	1.44	1.03
S, stdev	3.56	3.59	3.38	3.23	3.26

For a normal distribution, the following holds true[2]:

$$S_{ave} = S/\text{square root}(n)$$

Notice in Table 1 the values of S for each group are slightly different than the true value which we know to be 3.65. For example, S based on 10 sample lots (n=10) is 3.26. This is because it is not a "perfect" normal distribution, but it is still an excellent estimate of S for each sample size presented.

It is important to note that while our data gets "tighter" (smaller and smaller S_{ave}) as our sample size n increases, it does not change the true standard deviation S of our population; the population is still just as variable as ever. You can also see this by looking at the maximum and minimum values. Using 10 sample averages, the tallest value is only 71.5" and the shortest is only 66.3" but you know that the tallest people in the population were really 78.0" and the shortest were 59.6". The population doesn't change just because we increase the sample size.

So, how many cores do you take out of that charge?

At Last, Sampling A Charge of Treated Wood

Wood treating is a batch process. In the best case scenario, we place a lot of pieces of wood which are all the same size, same species and from the same mill into a pressure cylinder, apply our vacuum-pressure-vacuum process, and then pull it back out of the cylinder. This batch is typically called a "charge." Now the big question: does our charge meet the standard? If we are treating SYP lumber with CCA to AWPA's UC3B, we need a retention of 0.25 pcf. So we take our 20 samples, process in accordance with the AWPA standard and our result is 0.27 pcf. Excellent! The charge passes.

[2] "Applied Statistics", Neter et al., Allyn and Bacon, 1993, p. 267

Now let's put on our statistics hat.

- The *population* is the total number of coring locations we could have used. A charge of lumber in the US will typically have over 2,000 pieces of wood and each piece has perhaps 10 (perhaps a lot more) places you could take a sample; so our *population* is 20,000.
- Our *sample* size, *n*, is 20.
- If we take another sample, will we get the same result? If we knew the *standard deviation*, *S*, of the population, we could use statistics to predict how likely it would be to get a similar result. If *S* is small, we know our graph is tall and skinny and there is a darn good chance our next sample will look a lot like the first sample. On the other hand, if *S* is rather large…

Well, I don't want to leave you in suspense: *S* is enormous! If we sample that charge 100 times, we will get 100 very different results. Sometimes they would be above 0.25 pcf (we pass!) and sometimes they would be below 0.25 pcf (rats)…so I ask you, does the charge pass or fail?

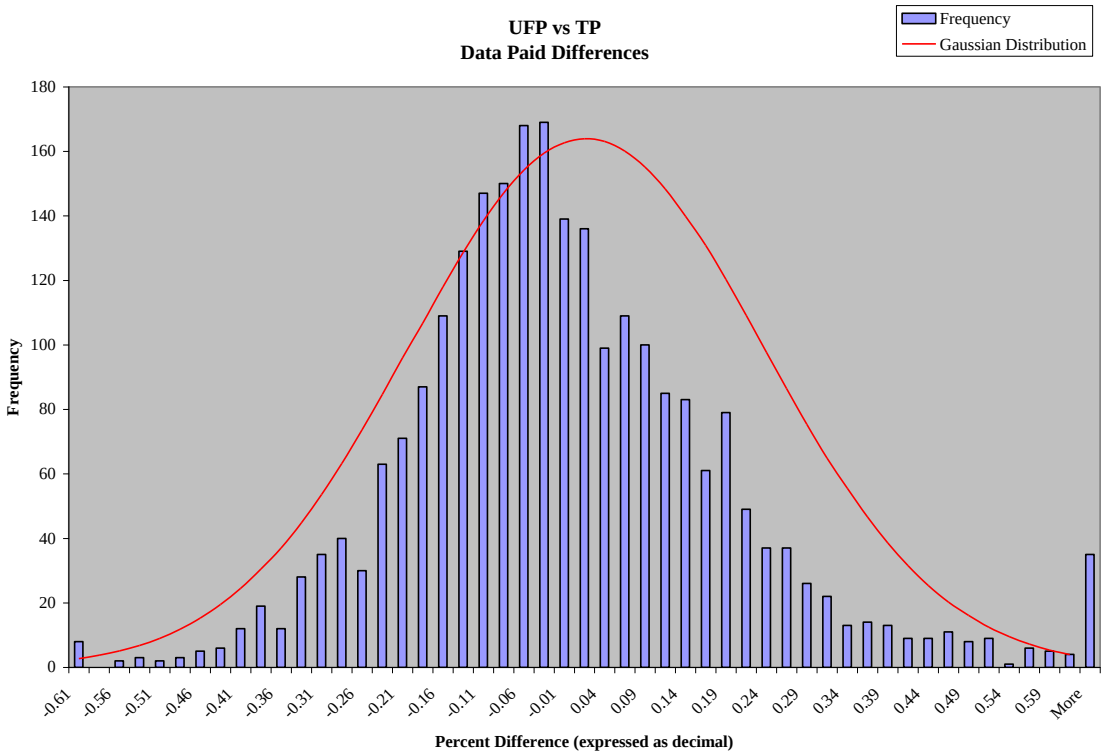
Estimating the Standard Deviation of Wood Cores

Using in-house data and literature values, we can calculate the standard deviation of the sample average, S_{ave}, from which, we can estimate the standard deviation, *S*, of the individual cores. The first set of in-house data comes from our in-house QC coupled with our third party inspection. In its simplest form this is two different samples of the same charge. If there were no variability, these two results would agree. If they disagree, it is an expression of variance.[3] In total, there were 2,500 data-pairs. Before trying to quantify the variance from this data set, we need to determine if there is any systematic change: do the assays generally go up or go down with time? This data says, "no". Assays go up, assays go down, assays move all around! The average difference is -1.67%, but the spread in the data quickly reveals that this is not a significant number.

[3] Throughout this discussion, it is impossible to separate analytical variance from variance caused by sampling/wood except to point out that the magnitude is larger than well accepted values for analytical variance.

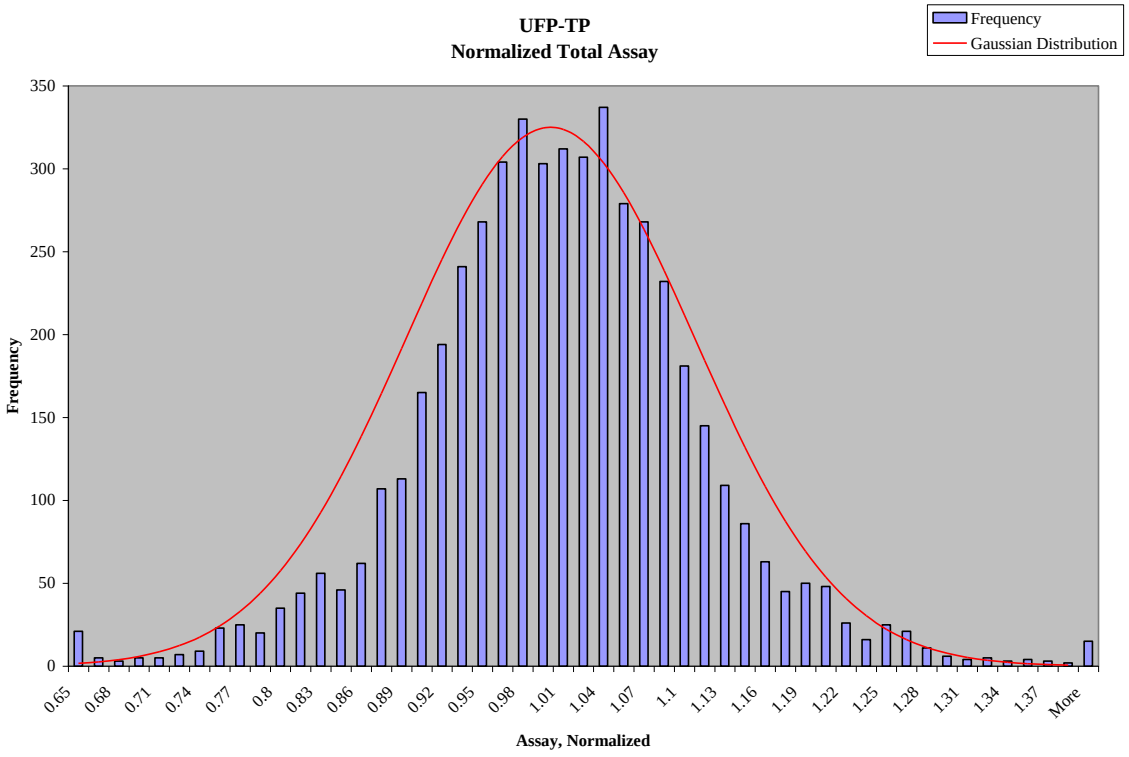
Graph 2 is a histogram showing the percent difference between each pair to determine if the difference between the two readings is purely an expression of variance or is the result of a systematic mechanism. The bars are the actual data; the red line is a normal distribution calculated from the average and standard deviation of the data set.

Graph 2 – UFP-TP Data Pairs Percent Difference Histogram



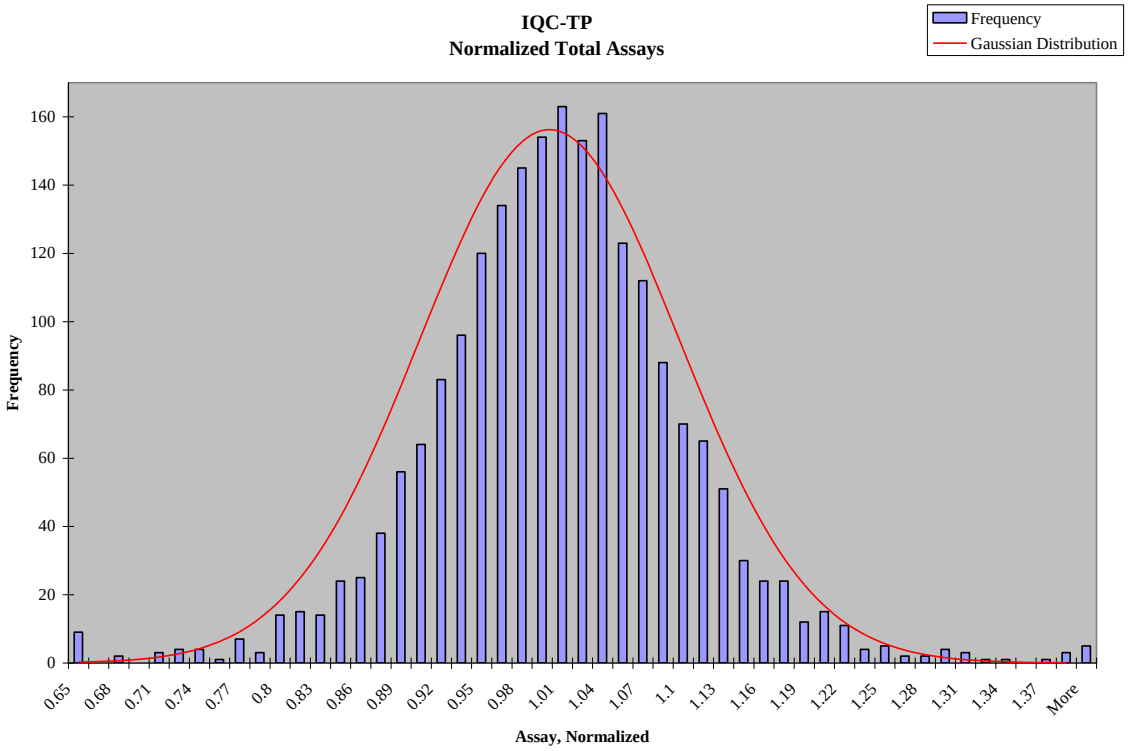
Having concluded there is no systematic change, we can use the data to evaluate variance. The variance was estimated by normalizing each data pair around its average value to a value of one. Doing so, the relative standard deviation of the average, S_{ave}, is 10.8%. Given that n is 20 for all of these samples, we calculate S to be 48%. The normalized values along with the corresponding normal distribution are presented in Graph 3.

Graph 3 - UFP-TP Data Pairs Histogram



In order to eliminate lab-to-lab error, a subset of this data set was created replacing the UFP result with the TP's analysis of the same sawdust. This is the IQC-TP data set and consists of approximately 1,100 samples. After the same treatment, this data produced a relative standard deviation of the average, S_{ave}, of 9.7%. Given that n is 20 for all of these samples, we calculate S to be 43%. The normalized values along with the corresponding normal distribution are presented in Graph 4.

Graph 4 - IQC-TP Data Pairs Histogram



The second set of data, the Janesville data set, comes from a test designed to determine if assays changed with time, by sampling boards over-time[4], and quantify gradient, by sampling five assay zones[5], by cycle type[6] and finally board-to-board[7] variance. Ultimately, the data showed no change with time (up, down, all-around), no change with zone, no change with cycle and board-to-board variance was really just wood variance. Each assay came from taking samples from center, left and right of a given board and combining these three samples to get an average assay. All in all, this project produced 500 assays from 1,500 samples.

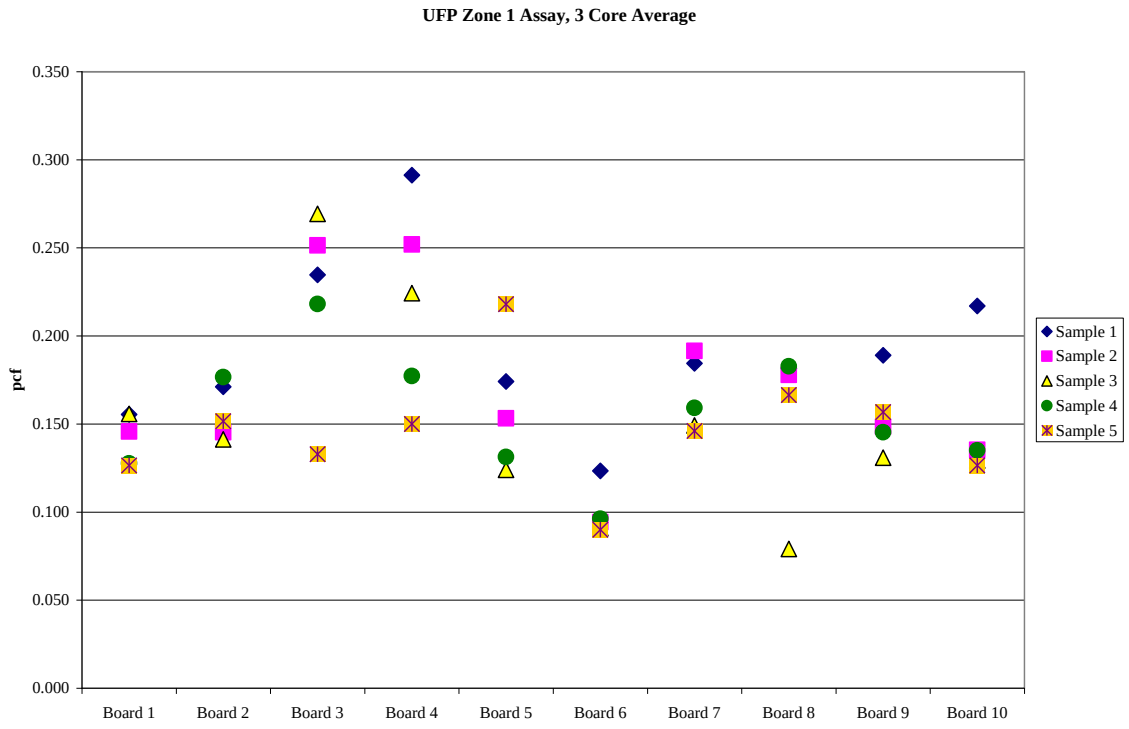
Although statistical analysis suggested that the outer assay zone was significantly different from the other assays zones in the "MFC" treatment, and that sample time 1 was significantly different

[4] Samples were spread over 34 days starting with a sample immediately after treatment.
[5] Assay zones were half inch increments going from 0.0" to 2.5".
[6] Two different modified full cell cycles were used designated "UFP" & "MFC."
[7] 2 charges each with 10 boards, all from the same unit, same mill treated.

from the other sample times in the "UFP" treatment, the estimates of sample variance are very consistent (consistently scary that is) no matter how you slice and dice the data.

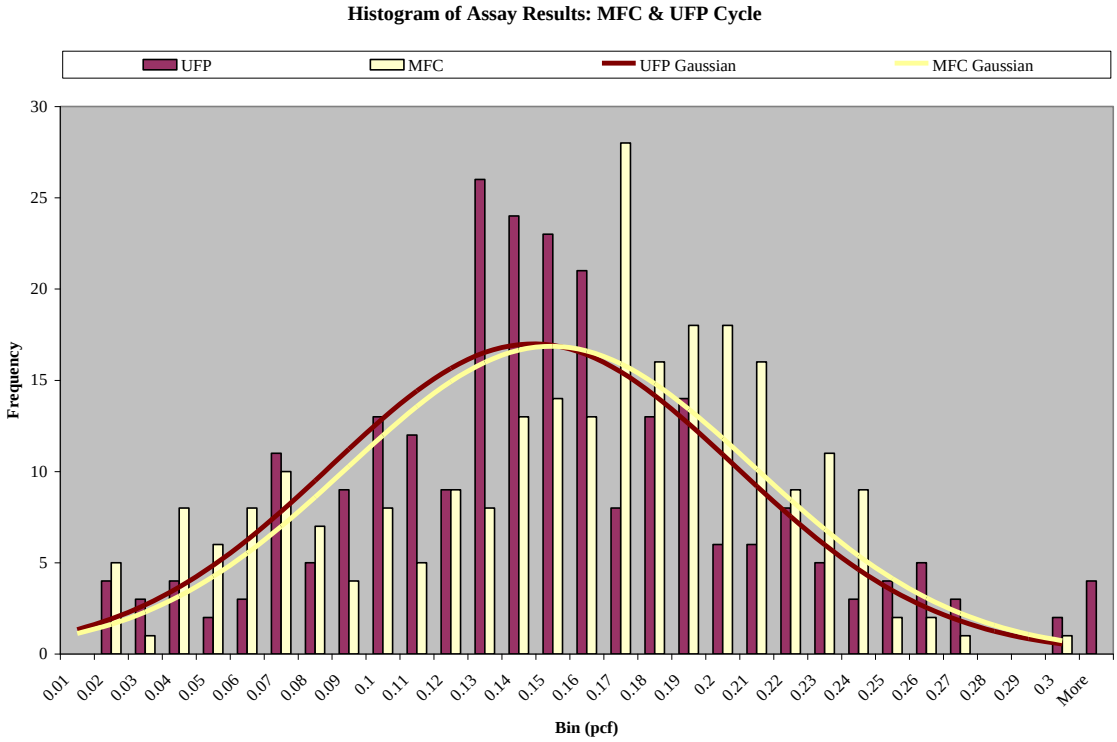
Graph 5 shows the results from one of the two charges for the outer (0.0" to 0.5") assay zone. All 50 assay values are shown: five different samples taken and different times for each of the ten boards. If this charge needed to be 0.15 pcf, does it pass or fail?

Graph 5 – Janesville Test Assays, UFP Cycle, Assay Zone 1



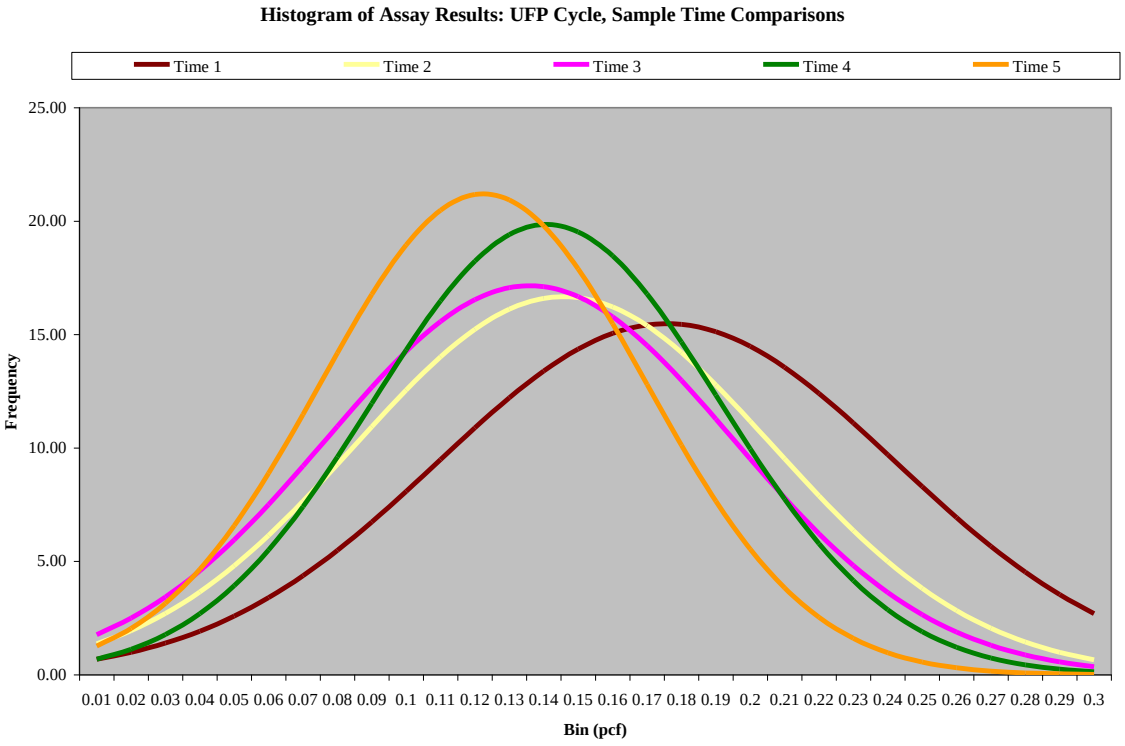
Examining the data using histograms provides a good picture of the data. Graph 6 compares the two different treatments showing both the actual values and normal distributions based on the average and standard deviation of each charge. There doesn't appear to be much difference and the assays are all over the map. These were both nominally 0.15 pcf charges. Notice that some of the assays came back as low as 0.02 pcf…just as the normal distribution predicts. The "tails" of the distribution are there, both high and low, like it or not.

Graph 6 – Janesville Data Set Histogram, all assays



Statistical tests[8] of the data from the UFP cycle showed there was no difference between the assay zones, nor was there any difference between the later four sample times, but that the first set of samples were significantly higher than later samples. While this can be seem in the following plot (Graph 7), you can also see that there is a very wide range of results in all data sets.

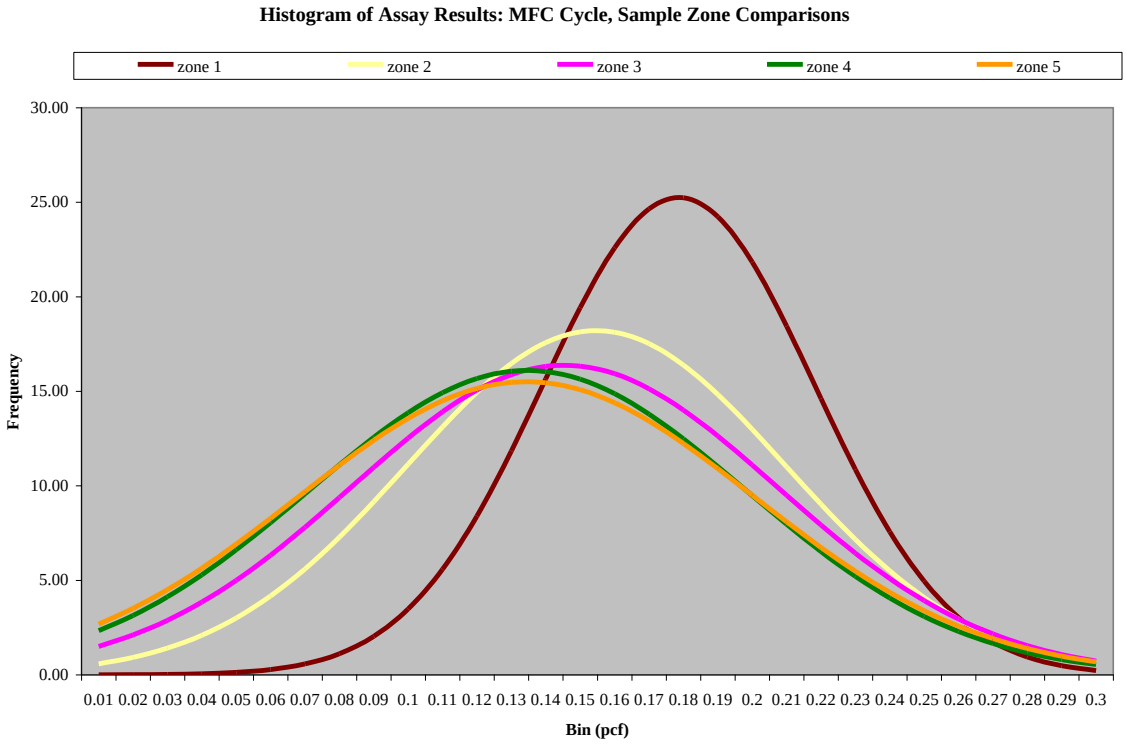
Graph 7 – Janesville Data Set, UFP Cycle, Sample Time Distributions



[8] Analysis of variance using Two-Way ANOVA with replication.

Statistical tests[9] of the data from the MFC cycle showed there was no difference between the sample times, nor was there any difference between the deeper four sample zones, but that the outer zone samples were significantly higher than the inner zone samples. Again, while this can be seen in the following plot (Graph 8), you can also see that there is a very wide range of results in all data sets.

Graph 8 – Janesville Data Set, MFC Cycle, Sample Zone Distributions



Ultimately, when these data were used to determine relative standard deviation, there was little impact on the numbers. The following values of S_{ave} were determined:

UFP Cycle	41%
UFP Cycle, excluding sample time 1	40%
MFC Cycle	40%
MFC Cycle, excluding sample zone 1	44%

Overall, this data indicates an S_{ave} of 40% and an S (0.40 x sqrt(3)) of 69%.

Finally, just to prove these numbers are not completely out of left field, allow me to pull data from an excellent paper by Schultz et al., from 2004[10]. They reported several different sets of

[9] Analysis of variance using Two-Way ANOVA with replication.
[10] "Biocide retention variation of southern pine", Tor Schultz et al., Forest Products Journal, Vol. 54, No. 3

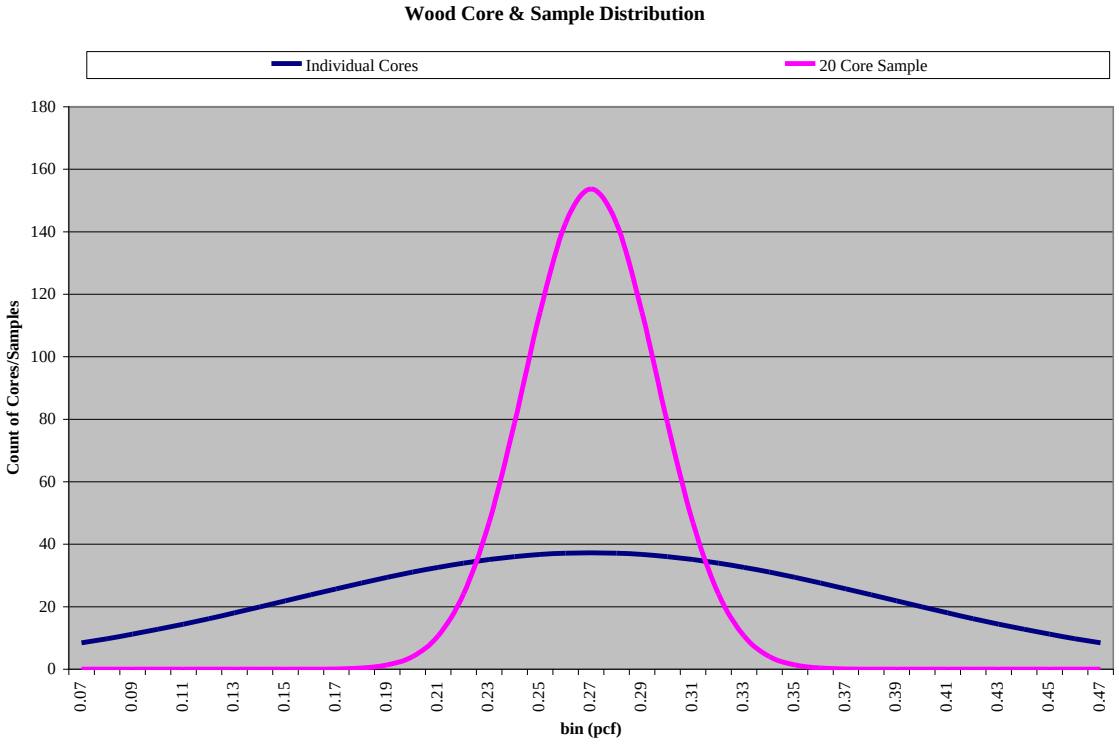
data which are germane to this topic. They provided data for individual boards in commercial charges as well as analysis of test stakes which yield the following:

Source	*n*	S_{ave}	*S*
Commercial, <2" thick	20	9.7%	43%
Commercial, >2" thick	20	19%	85%
Test Stakes	30	11%	60%

The best case for us poor treaters is a standard deviation of the "population" of cores of 43% with much of the data indicating it is considerably higher than that. So what do the "core" population and a 20-core sample look like from a charge of SYP lumber treated with CCA to a retention of 0.25 pcf? Even better, lets say the charge was somewhat over-treated to a retention of 0.27 pcf. Graph 9 is the now familiar histogram showing, for the first time, the distribution of the population of cores rather than averages of some number of cores.

Graph 9 also points out a technical flaw, which is beyond the scope of this paper to sort out. Specifically, the S we calculate is so large that the distribution of individual cores cannot be a truly normal distribution. Statistics provides tools for dealing with this situation in a technically valid way and this would be an excellent topic for future research but it does not change the conclusion that the individual cores are highly variable.

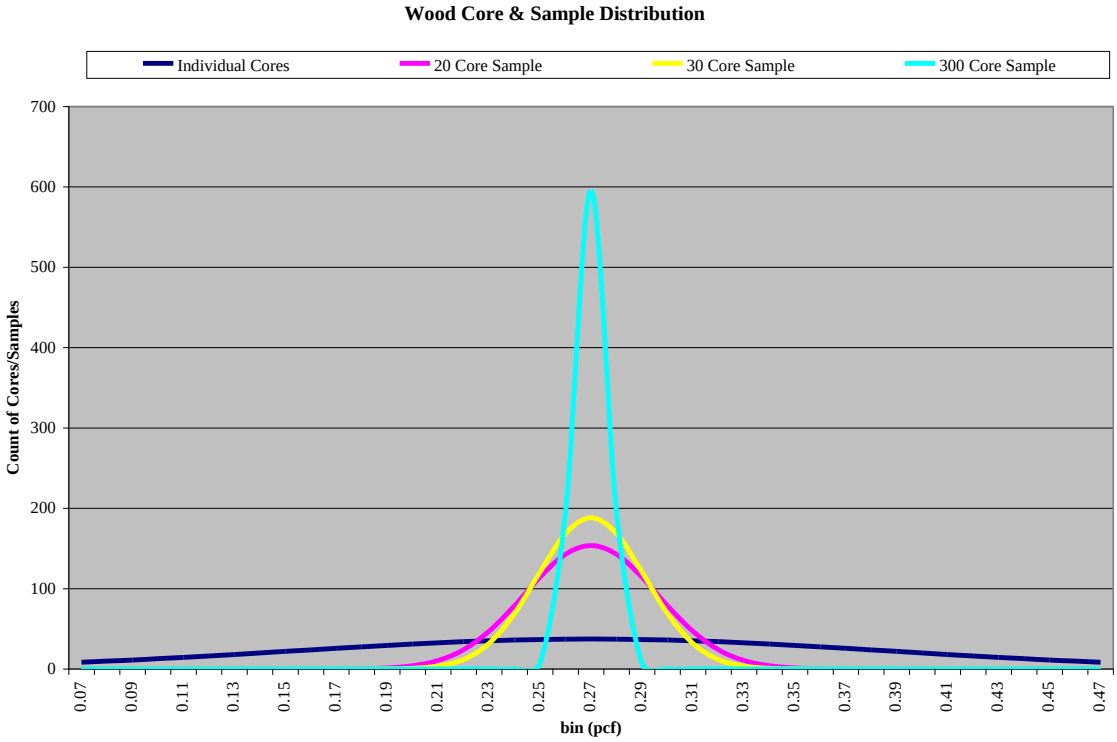
Graph 9 – Normal Distributions based on *S*, S_{ave}, and *n*



The true average retention of this charge is 0.27 pcf. By definition, it "passes" assay. Yet if this charge was sampled repeatedly, roughly 1 out every 6 times sampled, it would fail. So does the charge pass or fail? What are the implications for warranty programs?

The only way to improve the reliability of the result, or perhaps the repeatability of the result, is to increase the sample size.

Graph 10 – Normal Distributions based on *S*, S_{ave}, and *n*



Only by taking 300 cores for each sample would you achieve a situation where the charge essentially always passed. Even then, by definition, a charge that was truly treated to a retention of 0.25 would pass half the time and fail half the time. Furthermore, taking 300 cores for each charge is probably not realistic. The focus of our QC programs needs to shift from assessing individual charges to assessing (and controlling) the process. If a plant treated 10 charges during the course of the day and pulled 30 cores from each charge, that plant has 300 cores that could be used to make process changes.

Implications for Quality Control

The variance of individual cores is an inherent aspect of our treated wood products. The variability between boards and within boards in a charge is very, very big. The only weapon we have is sample size. The more samples we take, the more samples we base process changes on – because only process changes can affect the actual quality of product in the market place – the better the product.

Today, programs require 20 cores to be taken. The charge is passed or failed based on those results. Looking at the overall production from the plant, if the plant treats five charges in a day, 100 cores are taken. Generally, every charge is to be sampled by the plant but allowances have been created that potentially reduce the number of samples. Third party programs sample only 5% of the charges.

Perhaps the most counter-productive element of today's QC program is the fact that the treater is allowed to "pull the tags" on a charge that fails. This essentially disconnects the QC sampling results from process control. This, coupled with the very low number of samples in play in the first place leaves us with a 50 year old program in need of a complete re-evaluation, if not re-invention.

Many plants are now run by treating computers which collect key process information in databases. Better tools need to developed for treaters which utilize all of the samples they collect (and they need to collect more samples, not less) to make process changes. The option of ignoring samples that you don't like needs to be removed. The option of trying to control your process based on your third party results (which many treaters do) needs to be removed.

We need the technical voices of our industry to study and define the variability we see in wood. We need to develop systems for assessing the quality of wood coming out of a plant that are much more likely to give consistent results. Most importantly, those systems need to be tied to process control.

In 1999, we proposed a system called "Continuous Sampling" to AWPA. A watered-down version of it was accepted (AWPA Standard M3-05, Part B, 6.5). I believe most people considered it an attempt to create a loophole and pass a charge that "fails" retention. They missed the point. But I'll admit Continuous Sampling did not go far enough. Changes are needed to the entire system from in-plant testing to third party programs. Penetration needs to be addressed as well as bringing in other process information. There is no single area of research and development that could have a greater impact on the quality of wood products produced by our industry.