The compression test: how wrong can I be and still be right?

Richard A. Surber

The large variability in the compression test for corrugated boxes makes it critical for suppliers and customers to compare their test results and develop their own precision statements on critical test procedures.

In lab test results we trust

Sometimes it scares me when I see how much faith people put into a laboratory test result. If the answer comes from a lab, it must be right. Right?

If the answer comes from a dial indicator gauge, that's good. If the answer comes from a digital read-out gauge, that makes it even better. If you take the number off the gauge and write it down, that's good. If you have the number typed in a formal report, that makes it even better. If you have the number fed directly into a computer and typed out on a printer, that's almost as good as if the number had come down from God Himself.

People are willing to go out and take firm business positions that may involve thousands of dollars based on a laboratory number, without ever giving any thought to the truth of that number.

A test result is an estimate of the true answer. The fact is that there can be more than one correct estimate of the true answer.

Quality

People have always said they want quality; today they really mean it. This heightened interest in quality in recent years has also caused people to search for a better definition. If we want to have quality, we had better know what it is. One of the better definitions that has emerged has been Joseph Juran's description of quality as "fitness for use."

What "fitness for use" really means is:

- Does it work?
- Does it meet the customer's expectations?
- Does it do what it is supposed to do?

Surber is corporate technical director at Inland Container Corp., 4030 Vincennes Rd., Indianapolis, Ind. 46268-0937.

Another way to say it is: Does the product perform? People are moving away from descriptive statements of quality and toward performance related statements. This movement is causing important shifts in the definition of quality for the corrugated box.

One of the most important performance functions of many corrugated boxes is stackability. Our industry measures stackability via the compression test, an important measure of "fitness for use" of a corrugated box.

How good is our quality?

Everyone recognizes this is an important question. More and more of our industry's customers are setting compression limits on the boxes that we sell them. We make compression measurements to assure that we meet those limits. Our customers make compression measurements to see for themselves if we're meeting their limits.

Both measurements are estimates of the true answer. How good are these estimates? The answer can have far-reaching effects on the economics of our industry.

We determine quality by making a test measurement. We are judged as a supplier on those measurements. We may get more or less business based on those measurements. We may have boxes accepted or rejected based on those measurements.

Quality has become more important. Therefore, it is critical that the quality of our testing becomes more important.

The quality of testing

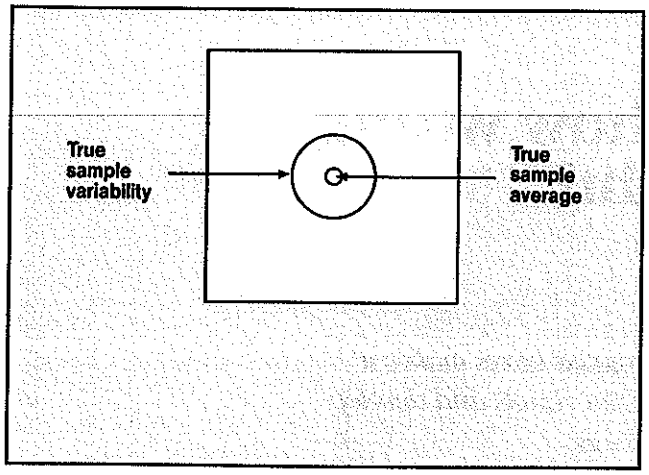
Tests can be wrong. We need to remember that testing is a process. If any of the inputs are wrong, the output will be wrong.

We judge the quality of a test answer in two different ways: accuracy and precision.

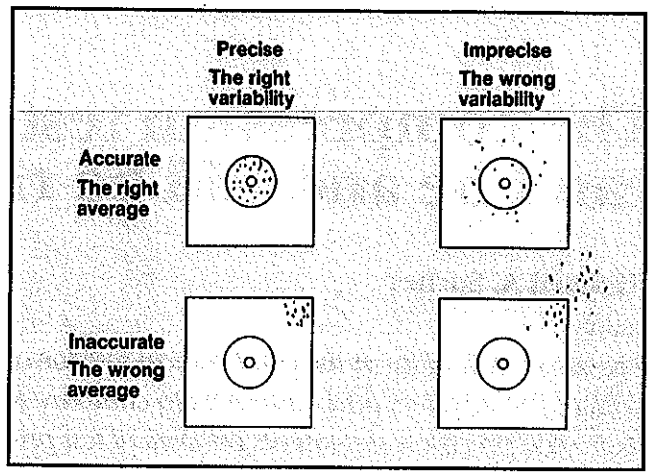
We can represent the true average and the true variability of a population on a target board (**Fig. 1**). The test can be wrong in three different ways (**Fig. 2**).

If a test procedure yields the wrong average, it is said to be inaccurate. This is easy for most people to understand. A test procedure can also yield the wrong variability.

1. The true average and variability would look like this on a target board.



2. The effects of poorly maintained and operated test equipment on the test result



When it does, it is said to be imprecise. All populations of products will exhibit variability; not all units or pieces are exactly the same. An imprecise test procedure makes us think that the variability of the product is greater than it really is.

The quality of test results mislead us in two different ways. We can get the wrong average or we can get the wrong variability.

The precision statement

Common practice includes a precision statement at the end of most TAPPI test methods. A precision statement is concerned with the variability of a test procedure, not with the accuracy.

A precision statement contains two parts: a statement of repeatability and a statement of reproducibility.

Repeatability

The repeatability portion of the statement answers this question: If I repeat the test on a homogeneous population, how well can I expect to agree with myself?

Repeatability can be used to judge:

- The ability of a test operator to repeat a measurement
- The stability of a test apparatus
- The homogeneity of a material.

Repeatability is often called instrument error. It can be affected by the variability of the material being tested (in the case of a destructive test), the measurement method, or the instrument itself.

Reproducibility

The reproducibility statement answers this question: If two labs test a homogeneous population, how well can they expect to agree?

Reproducibility can be used to judge:

- The ability of two laboratories to obtain the same test result on samples from a homogeneous population
- The closeness to which a laboratory may determine compliance with a specification.

Reproducibility is the variability caused by differences in conducting the test in different laboratories by different people on different pieces of test equipment.

TAPPI test methods

TAPPI test methods labeled "official" are required to carry a statement of repeatability. The statement of reproducibility is desirable in official methods, but not mandatory. This is a position that TAPPI may want to rethink. Official methods really do need a statement of reproducibility between laboratories. That's where the action will be in the future.

TAPPI test methods labeled "provisional" do not require statements for repeatability or reproducibility.

TAPPI has defined the precision statement in Official Test Method T-1206 titled "Precision Statement for Test Methods." It has defined a procedure for determining precision in Official Test Method T-1200 titled "Interlaboratory Evaluation of Test Methods to Determine TAPPI Repeatability and Reproducibility."

TAPPI Test Method T-804

The precision statement in TAPPI Test Method T-804, "Compression Test of Fiberboard Shipping Containers," first published as an official method in 1981, is as follows:

- Repeatability (within a lab) = 8.5% ($n = 5$)
- Reproducibility (between labs) = 11.3% ($n = 5$).

These numbers are based on an American Society of Testing Materials study published by J. G. Miles in *Materials Research and Standards* in 1966.*

How do we read these statements? Repeatability simply says that if we compare two sets of five samples from the same population, in the same laboratory, with the same operator using the same piece of test equipment, the difference between the two averages will be 8.5% or less, 95% of the time.

*Miles, J. G., Materials and Research Standards, (3): 142(1966).

Other ways to say the same thing are:

- Duplicate measurements of compression should not differ by more than 8.5%, 95% of the time.
- We have 95% confidence that duplicate measurements of compression should not differ by more than 8.5%.
- Duplicate measurements of compression should not be considered suspect unless they differ by more than 8.5%.

All of this implies that the difference will be greater than 8.5% for 5% of the duplications made.

Reproducibility is read in the same way. If we compare two sets of five samples from the same homogeneous population in two different laboratories, the differences between the two averages will be less than 11.3%, 95% of the time. We could also say that compression measurements made in two different labs should not be considered suspect unless they differ by more than 11.3%. We could also say that the differences will be greater than 11.3% on 5% of the testing occasions and still should not be considered suspect.

What all this says to me is that compression testing is not a very exact science.

What does it all mean?

Where does all this statistical gobbledygook come from? It comes from two decisions that have been made in advance in regard to evaluating the data. The first decision is to make the assumption that the test errors follow a normal distribution around the true average (**Fig. 3**). The second decision is to choose a confidence limit of 95%. Other confidence limits could be chosen; if they were, the answers we get would be different for the same data. What we have done when we make these two decisions is to say that our estimates of the true answer through multiple tests will look like **Fig. 4**.

What this means is that 95 out of 100 duplicate tests will fall within this range.

The current repeatability and reproducibility statements for Test Method T-804 are represented in this format in **Figs. 5 and 6**.

These repeatability and reproducibility statements bother me. It is not uncommon for a buyer and seller to be less than this far apart (11.3%) and have a serious disagreement. As an example, the box maker could test and get 105% of the target and say, "I pass." The box user could test and get 95% of target and say, "You fail." Actually, the two tests are within 11.3% of each other, and there is no statistical reason to consider them to be different.

FSCTC 1988 review

When Test Method T-804 came up for its normal five-year review, the Fiberboard Shipping Container Testing Committee (FSCTC) decided that it was time to review the precision statement. Surely the numbers were not really that large.

I wrote a letter to all members of FSCTC and asked if they would volunteer their laboratory to participate in the study. If they volunteered, they were asked to fill out two forms. On the first, they were asked to describe their test equipment. On the second form, they were asked to describe their test procedure.

Twelve laboratories volunteered. The labs represented the following groups:

- Seven corrugated companies
- Three labs within one of these companies
- A corrugated box user
- An independent lab
- The Institute of Paper Science and Technology.

These laboratories represented a wide assortment of test equipment, as shown in **Table I**.

The load was applied to the boxes being tested via a mechanical drive on eight of the machines and via a hydraulic drive on the other four machines. The force applied during the test was measured by a balance beam mechanism in six of the test machines and by an electronic load cell in the remaining six.

Surprisingly, these labs also utilized a wide assortment of flap-sealing techniques (**Table II**).

Test Method T-804 outlines a very specific flap-sealing procedure, yet only 2 of the 12 labs followed the procedure on a regular basis.

Because of the moisture hysteresis effects on paper, Test Method T-804 also specifies that boxes be preconditioned at a low humidity, followed by conditioning at standard conditions. Table II shows that only 9 of the 12 labs precondition on a regular basis. All 12 laboratories do condition in a standard condition atmosphere on a regular basis.

I asked that all labs both precondition and condition those boxes tested for precision statement data.

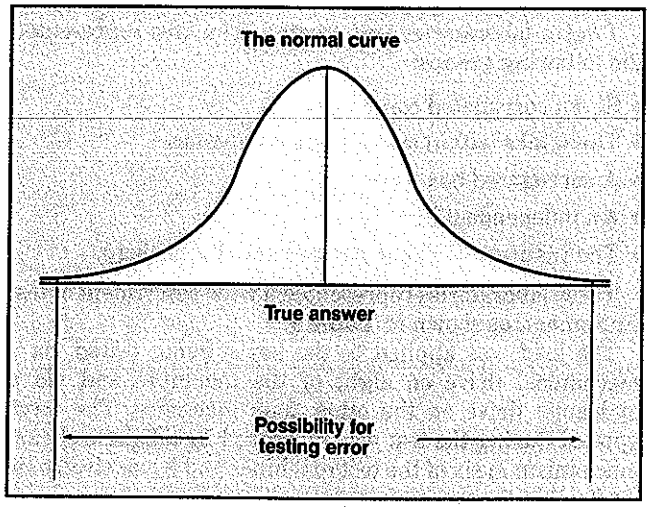
Closing the flaps presented another problem. I found that only 3 of the 12 labs even had the capability to seal flaps according to the TAPPI method, which I have decided to call the classic method. The classic method is unwieldy. It requires the use of sealing board and two preconditioning steps. It appears that laboratory managers decided over the years that the important issue was to seal the flaps. I wanted to see if this change in sealing technique had caused any change in the variability of the test.

I asked the three labs that could seal according to the classic method to also test a set of samples sealed with hot melt adhesive, which is currently the most common technique used.

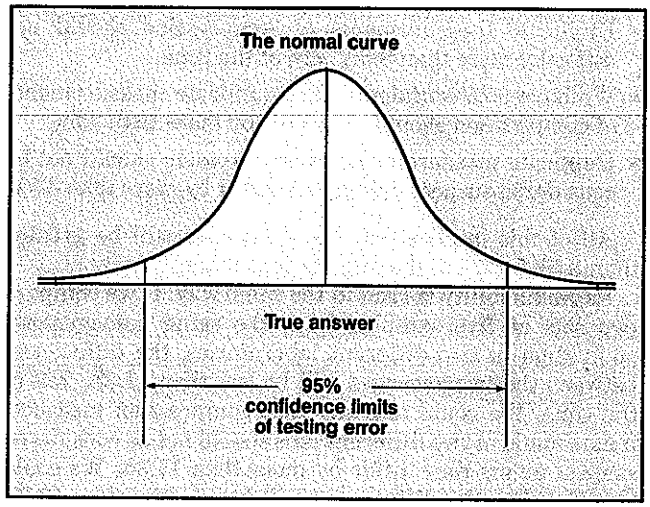
I asked the other labs to seal according to their standard technique. The laboratory that normally tested with unglued flaps was asked to use hot melt adhesive.

Repeatability and reproducibility studies are straightforward when evaluating a technique for measuring diameter or for measuring brightness. The same item can be measured more than one time by the same operator using the same gauge to get repeatability. That same item can be measured by another operator in another lab to get reproducibility. In nondestructive test situations, it is possible to get a good estimate of the true measuring error.

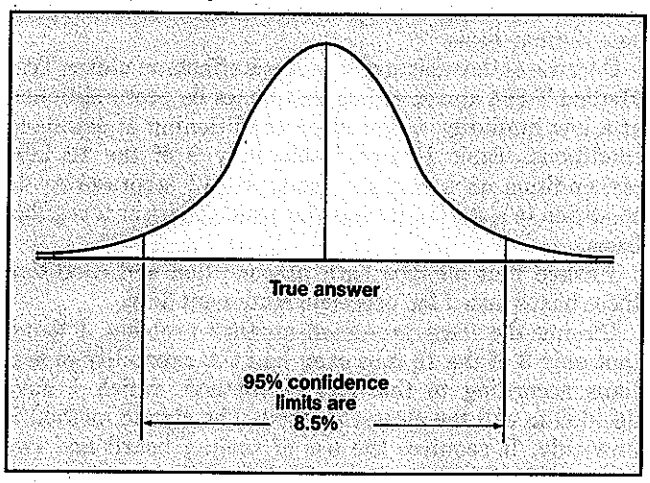
3. Test errors follow a normal distribution around the true average.



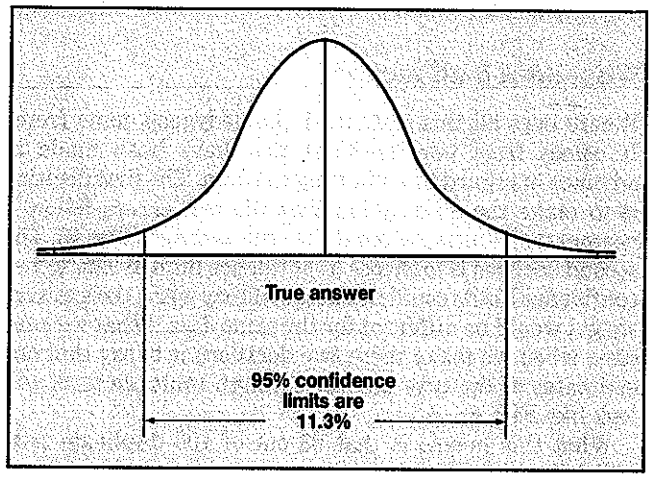
4. Test error distribution with a 95% confidence limit



5. Current repeatability statements for Test Method T-804



6. Current reproducibility statements for Test Method T-804



I. Types of compression testers studied

Number	*Manufacturer*	*Max. capacity, lb*	*Platen sizes, in.*
1	Baldwin/Satec	60,000	30x42
4	Tinius Olsen	30,000	60x72
2	MTS	30,000	60x60
1	Gaynes	20,000	48x72
1	Gaynes	20,000	48x48
1	TMI (modified)	10,000	36x48
1	Homemade	6,250	26x30
1	National Forge	5,000	24x36

II. Flap sealing and conditioning techniques in normal use in 1988

Technique	*No. of labs*
Flap sealing	
T 804	2
Hot Melt	5
Clips or Clips/Tape	2
Stitched or Stitched/Clips	2
Unsealed	1
Conditioning	
Precondition	9
Condition (50%/73°F)	12

The compression strength of a box can only be measured once. In such a case, different samples must be used to determine repeatability and reproducibility. The sample differences introduce variation due to materials, which we want to keep as small as possible. When studying a destructive test procedure, samples must be as homogeneous as the process allows; they must be processed under the same conditions.

To get a meaningful idea of precision, I had to have boxes manufactured as uniformly as possible. I chose to work

with the following box:

- 200 lb test (42-26-42)
- C flute
- RSC
- 20-1/2 in. × 15-7/8 in. × 12 in.

Another engineer and I went to the plant to supervise the manufacture and sampling of the boxes used in the test.

In an attempt to keep product variability to a minimum, we made many checks of the order being run prior to actually taking samples for testing purposes.

On the corrugator, we checked single-face and double-face glue patterns, board caliper, flat crush, and both single-face and double-face pin adhesion. Adjustments were made as necessary. All of the corrugated sheets for this study were taken from one position off the corrugator running at a constant speed of 700 ft/min over a two-minute period.

On the flexo folder gluer, we checked slot depth correctness and uniformity, caliper crush in printed areas, as well as caliper loss from the feed roll, pull rolls, and folding belts. We also checked for proper joint gluing and folding uniformity. The stack of sheets from the corrugator was run consecutively through the flexo folder gluer over about a three-minute period.

The boxes off the flexo were numbered in order of production. A random numbers table was used to select the boxes for testing at each laboratory.

Three of the labs tested 10 boxes each using the classic sealing method and 10 boxes each using a very specific hot melt glue pattern. The other 9 labs tested 10 boxes each using their normal sealing technique.

Results

The repeatability and reproducibility results are shown in **Table III**. There are several interesting points.

First, there were no statistically significant differences in average values among the 12 laboratories. Comparison of the data from the three labs that conducted the test by both the classic and the hot melt sealing methods shows that the hot melt method is not any more variable than the classic method.

A look at all of the repeatability data shows that, as long as a single method is used within a laboratory, the closing procedure has no negative effect on repeatability.

The three labs that used hot melt in comparison to the classic method were instructed to use a very specific hot melt glue pattern. The other four labs used whatever pattern is normal for them. A comparison of line 2 to line 3 shows that reproducibility increased slightly. This could be due to differences in sealing or other differences in test technique that we don't know about.

One of the most significant findings appears as we compare line 3 to line 4. This shows that as we include a wide array of sealing techniques, the uncertainty of the comparisons between labs increases as well. This makes a strong statement that if two labs are going to be making comparative tests, it is extremely important that they use exactly the same test procedure.

Line 2 is interesting to me for another reason. It shows that careful attention to test technique can result in a reproducibility figure that approaches repeatability. This is true in spite of the fact that these three machines represent two machinery manufacturers as well as a homemade model. They also represent the widest platen size range, the widest maximum loads, both mechanical and hydraulic drive mechanisms as well as both balance beam and electronic load cell weighing mechanisms.

This suggests that as companies develop strong partnership relationships, they owe it to themselves to measure the precision statement that applies to their two specific laboratories.

For the purpose of the new version of Test Method T-804 we will have to accept the following:

- Repeatability (within a lab) = 7.0% (n = 5)
- Reproducibility (between labs) = 10.6% (n = 5).

My original thought was that through very exact sampling, I would find much lower repeatability and reproducibility factors than were reported as a result of the 1966 work. These 1988 numbers are somewhat lower, but not much.

Even though these between-lab confidence limits are wide, there is no guarantee that they will be met. The two technicians must be well trained in the test technique, and the two labs must be following identical techniques to assure that even these wide confidence limits are met.

A problem

Let's say, just for the purposes of discussion, that we have a specification of a target value which we will call 100% with limits for the average of ±10%. We represent this schematically in **Fig. 7**.

If we overlay the confidence limits for reproducibility onto this schematic, we begin to see the problem.

If the true compression value of the order equals 100% of the target, our two labs will have no trouble if they test and get values at the extremes of the test confidence limits (**Fig. 8**).

If the true average is toward the lower end of the acceptable range, the two labs could easily disagree on whether or not the order passes (**Fig. 9**).

The true average value is 93%, which is an acceptable value for this specification. The two labs could test, and one could get 98% as an average, and the other could get 88% as an average. Both have made valid estimates of the true average, but one laboratory would say that the order fails and the other would say it passes.

This is bad enough, but it is actually worse than it seems. Metrologists calculate something called test capability. The 95% confidence limits for capability are the square root of the sum of the variance for both repeatability and reproducibility. I estimate the capability figure to be 12.7% for the compression test.

Metrologists look at test capability as a percentage of the tolerance of the value being measured. In this example, the tolerance is 20% and the procedure capability is 12.7%. The capability width takes up 12.7/20 of the tolerance width, about 64%. As a rule of thumb, metrologists state that if the test capability is less than 30% of the tolerance, then the effect of measurement error on decisions can be ignored. These data say that we can't ignore the current

III. Results of 1988 study

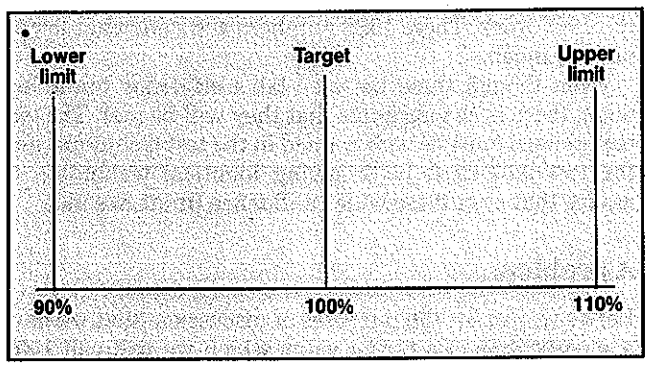
Sealing method	No. of labs	Compression average, lb	Repeatability	Reproducibility
Classic	3	810	7.1%	9.3%
Hot melt	3	805	7.5%	8.5%
Hot melt	7	818	7.0%	10.6%
Most common for each lab	12	840	6.5%	16.2%
Current statement in T-804	11	...	8.5%	11.3%

Repeatability and reproducibility are for averages of 5 samples.

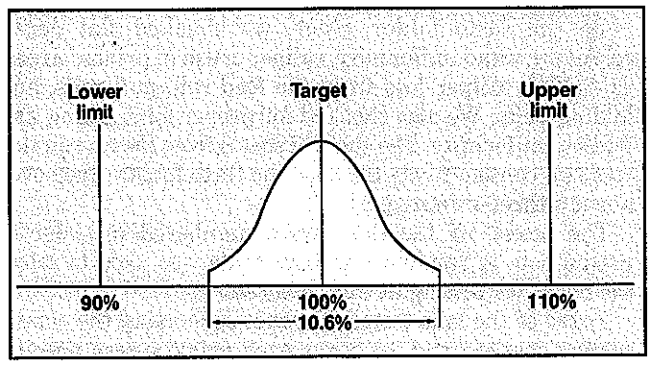
IV. Reproducibility for different numbers of test samples

No. of samples	Reproducibility, %
1	23.7
2	16.8
5	10.6
10	7.5
20	5.3
40	3.8

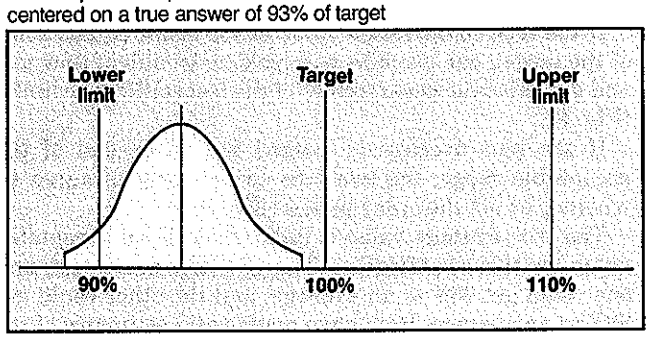
7. Compression specification



8. Compression specification with 95% confidence limits of reproducibility centered on a true answer of 100% of target



9. Compression specification with 95% confidence limits of reproducibility centered on a true answer of 93% of target



V. Reproducibility for different numbers of test samples

No. of samples	Reproducibility 95% limit	Reproducibility 50% limit
1	23.7	8.1
2	16.8	5.7
5	10.6	3.6
10	7.5	2.6
20	5.3	1.8
40	3.8	1.3

limits of test uncertainty as we make decisions to accept or reject product.

Much of the test uncertainty of the compression is likely due to product variability, but this experiment was not set up to separate test variability from product variability.

Two helpful hints

There are two things we can do to help offset the high degree of uncertainty that we have in the answers we get from the compression test.

One is that we can test more than five samples. There is a relationship between the 95% confidence limits of the test answer and the number of samples tested. This relationship is:

$$95\%\ \text{confidence limits} = \pm 1.96\ s_{repo}/n^{1/2}$$

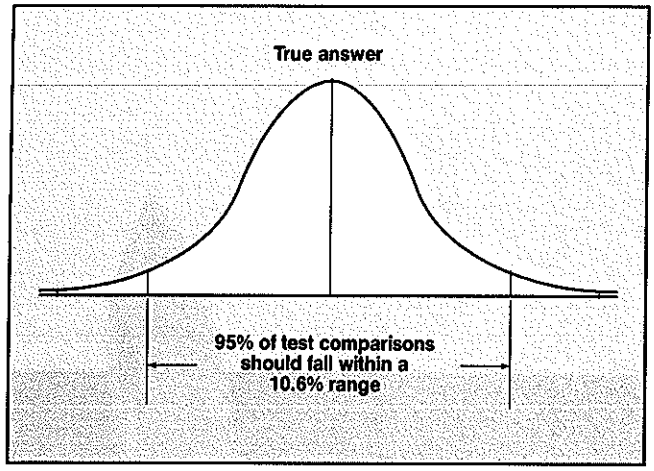
The value, s_{repo}, is one standard deviation of the reproducibility error. This is the central limit theorem, which says that if we test more samples from a population we will get an estimate that is closer to the true average.

We found that reproducibility for $n = 5$ for the compression test is 10.6%. The reproducibilities for other values of n are shown in **Table IV**.

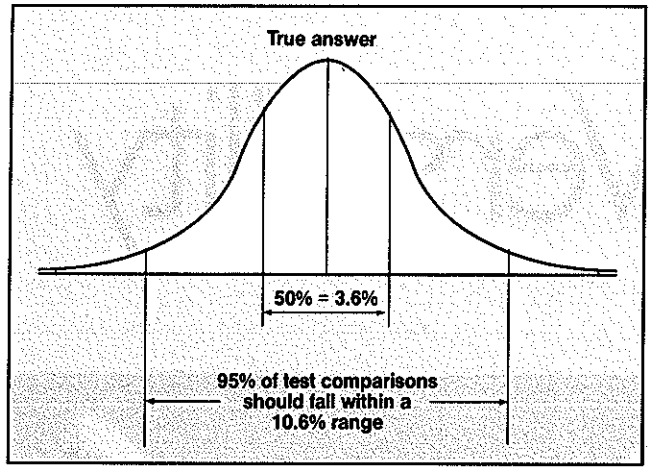
It is somewhat traditional in our business to test five samples to try to represent a population of boxes. When we do this, we have a 10.6% uncertainty between labs. If the two labs test more than five samples, they will have a reduced level of uncertainty about their comparison.

Testing more samples is a lot of work. In today's

10. The traditional way of thinking about reproducibility



11. The 50% limits for the reproducibility value of the compression test are 3.6%.



atmosphere of partnerships between companies, there may be a better way to think about reproducibility limits. **Figure 10** shows the traditional way of thinking about reproducibility.

This says that test determinations between labs can vary by as much as 10.6% and still be valid estimates of the same answer. This is frustrating because you can't even argue until you differ by more than 11%.

While true, you have to remember that this is the extreme case. Two labs don't have to differ by that much, and, in fact, they shouldn't differ by that much very often.

The 10.6% is a value to be used if you don't have any other knowledge of the other lab to which you are being compared.

Secondly, as we develop true business partnerships, we will be comparing test results with suppliers and customers more than once.

One useful technique is to understand the 50% limits. The 50% limits are approximately one-third of the 95% limits. The 50% limits for the reproducibility value of the compression test are 3.6% (**Fig. 11**).

This says that any two laboratories making a series of duplicate determinations of five tests and comparing average values should not differ by more than 3.6% in 50% of the testing comparisons made. When we add the 50% limits to the reproducibility values shown in Table IV, we get **Table V**.

In a single lab, the 50% limits for repeatability are 2.4%.

I suggest that the addition of the 50% limits to the TAPPI precision statement would be useful to many struggling with the uncertainty of test procedures.

As quality becomes more and more important to our business partnerships, companies are going to have to develop a better understanding of how to make the important measures of quality. We must understand that test answers will always have some degree of uncertainty about them. If two companies are going to compare test results, they must assure themselves that they are using identical test procedures—just to keep the uncertainty to a minimum. Sometimes even the minimum uncertainty is large. As we have seen in the compression test, the degree of uncertainty grows even larger when two labs use slightly different methods of testing.

As we develop partnerships, companies need to go to the trouble to compare their test results and develop their own precision statements on critical test procedures. Not doing so could add a great deal of unnecessary cost to the overall system of product distribution.□

Received for review July 26, 1989.

Accepted Aug. 11, 1989.

Presented at the 1989 Corrugating Conference.