When Have I Done Enough, How Do I Know and How Do I Prove It?

Joel Anderson
RSI Pipeline Solutions

Pipeline Pigging and Integrity Management Conference

February 12-16, 2024

Organized by
Clarion Technical Conferences

https://doi.org/10.52202/072781-0022

https://doi.org/10.52202/072781-0022

Background

The regulations published by PHMSA in October of 2019 require non-destructive examination (NDE) to determine the material properties of pipelines that lack traceable, verifiable, and complete (TVC) records. To meet this requirement, operators must conduct verification of pipeline material properties in accordance with 49 CFR 192.607. (Assuming the pipe is to continue in service.)

In addition, section 192.607(e)(1) states that a population is a combination of material properties and attributes: Nominal Wall Thickness, grade, manufacturing process, pipe manufacturing dates and construction dates. The pipeline segments of the population do not need to be contiguous. If an operator chooses to use an alternate statistical sampling, 192.607(e)(5) allows the use of another method that has a statistically valid basis designed to achieve a 95% confidence level.

Confidence level is a statement of the probability at which an estimated parameter, such as the sample mean, is also true for the population. A common example of this would in the reporting of survey results where a statement like; "45% are in support of a measure with a margin of error of 3%". The margin of error is the confidence interval for a 95% confidence level. Meaning that if the whole population was surveyed, the true mean would be within 3% of the estimated mean 95% of the time.

The goal of a sampling program is to sample enough to be able to make a defensible decision about the population without sampling too much. Too small of a sample increases the risk of making an erroneous decision about the population, sampling too much wastes time and resources, indirectly increasing risk. elsewhere Differences will always exist between the measured value and the true value since perfect measurements are impossible. Errors of varying magnitude will exist due to random effects from any number of sources. The inevitable question arises is whether these observed differences between the sample and the hypothesized value are likely due to chance and have enough samples been completed to be able to tell the difference. If you are looking for a universal number that is applicable in all situations, you are not going to find it in this paper or anywhere else. Statements to the contrary are either misinformed or deliberately misstated. The appropriate sample size is interrelated with your acceptable risk. Changes to the acceptable risk will change the number of samples significantly. It's easy to say that the risk of an erroneous decision should be as close to zero percent as possible but if each sample costs $100,000 there is significant motivation to consider the benefit of additional samples vs. the cost. But with some knowledge of sampling statistics, though the risks haven't been eliminated, it has been lowered to an acceptable level.

This paper will discuss all the requirements to set up a sampling plan. Including how to determine if a sample is inconsistent with some value, determining sample size and comparing two samples against each other. In addition, alternate sampling plans to achieve the 95% confidence level will be covered.

Hypothesis Testing

Most elementary statistics are introduced with the concept of an urn with a known number of balls itemized by color. Everything about the urn is known and the goal is to predict what kind of data it will produce. While drawing from an urn is a good analogy to introduce statistics, it is not representative of any problem of consequence. In almost any problem outside of contrived examples and casino games, neither the size nor the exact make-up of the population is known ahead of time. Rather than predicting what data will come out of drawing from the urn, the engineer has a much different problem, that is to reason about the likely contents of the "urn" based on their existing

https://doi.org/10.52202/072781-0022

knowledge of the problem and a limited set of observations. Therefore, any statistic should not be confused with the reality of nature. A statistic is a measure of our state of knowledge about the problem at that time or more accurately a measure of our ignorance, since we are not simply given the contents of the "urn" ahead of time nor is there is any way to measure the entire population to see if we are correct. Therefore, in any hypothesis test, the only two options are to either reject the hypothesis because the data is very unlikely if the hypothesis is true or fail to reject it if the data is consistent with the hypothesis. We can never accept the hypothesis, because that would require rejecting an infinite number of alternative hypotheses. For example, if the hypothesis is that the mean value of the yield strength is 42 ksi, to accept that hypothesis would require having to prove that it's not 42.1 or 42.01 and every other infinite number of arbitrary levels of precision. A hypothesis test is analogous to a criminal trial where the defendant is pronounced either guilty or not guilty, they are never pronounced innocent. We are looking to prove beyond a reasonable doubt, not beyond all doubt. Likewise in statistics even with the best plan and execution there is some inevitable probability of making an error in the decision.

Sampling Principles

As an introduction to sampling, a few introductory sampling principles will be introduced to dispel some entrenched myths. A detailed discussion and proof of these is beyond the scope of the paper but are presented for the benefit of background information.

Principle I: The size of the population (the "lot" in sampling terms) does not appreciably change the information obtained from the sample. Assuming the lot is much greater than the sample size by at least a factor of 10. This means that the sample will contain roughly the same defective percentage regardless of the lot size. This principle is often the least understood. This is because standards like ANSI/ASQ Standard Z1.4 contain sampling charts that require differing sample sizes depending on the size of the lot and further subdivided by "inspection levels". The reason for the sample size to lot size dependency lies not in statistics but economics. As the lot size grows the economic benefit of larger samples is worth it to acquire more information about it. See Principle II.

Principle II: The information contained in a sample increases when the size of the sample increases, independent of the size of the lot. The larger the sample, the more faithful the sample will be to the population.

Principle III: The information contained in a sample does not depend on the *proportion* that the sample represents of the lot only the overall size. If we sample 10% of two different lots with the same defect rate and where one lot is larger than the other, the larger lot will have a larger sample, and will more closely resemble the overall defect rate of the population even though both samples represent the same proportion of the lot.

Measurement Uncertainty

When discussing sampling, one of the prerequisites is to discuss measurement uncertainty. A measurement is the process of assigning a quantity to a physical property. Regardless of the precision of the process, there is some level of error. The term, "error" in this context does not imply a mistake but rather the deviation between the measured quantity and the (unknown) true value. For any physical property, there is only one true mean. But if repeated samples are taken, assuming the size of the population is much larger than the size of the sample, differing sample means would be

https://doi.org/10.52202/072781-0022

calculated each time. Perfect measurements don't exist, therefore just taking an average of the sample is an inadequate description of a physical property or most anything for that matter. Since 192.607(c)(1) requires performing measurements at 5 places in 2 circumferential quadrants of the pipe it is necessary and beneficial to use the measurements to calculate not just the mean value but also the standard deviation. With only 10 measurements, a single outlier can exert significant leverage on the average. In addition to measurement error, the steel itself is not perfectly homogeneous. A variation of at least 3 - 4 ksi in the measured yield strength (YS) around the circumference or end to end is common even when using destructive testing.

As an example, assume 10 replicate measurements of the YS are taken with a mean of 40 ksi and a standard deviation of 3.5 ksi, on a pipe having an assumed specified minimum YS of 42 ksi. Without accounting for the uncertainty, one might conclude that that the YS is less than what was on record, which may not be correct decision. The way to account for the uncertainty in the mean is with the confidence interval.

$$CI = \mu \pm t\frac{s}{\sqrt{n}} \qquad (1)$$

where: μ = the average of the sample, s = the sample standard deviation, n = the number of measurements, t = t-value based on a probability of 0.975, and $n - 1$ degrees of freedom (t = 2.26)[1]

The 95% confidence interval of the mean for this example, working in units of ksi, would be:

$$40 \pm 2.26\left(\frac{3.5}{\sqrt{10}}\right) = 40 \pm 2.50 \qquad (2)$$

The confidence interval of the mean would extend from 37.5 to 42.5 ksi therefore a true yield strength of 42 ksi cannot be excluded based on the measurements and the YS of record is not rejected. This indicates that the test could then be counted as being successful for the purposes of verifying the material properties. This is why a measurement that is different should not be treated as inconsistent without considering the uncertainty in the measurement.

Confidence Intervals

A common misconception in the interpretation of the confidence interval is that a 95% confidence interval means that there is a 95% probability that the true value is within those bounds. The actual meaning of the confidence interval is that if somebody took repeated samples of a given size and calculated the confidence interval each time, 95% of the confidence intervals would include the true value. This is demonstrated in Figure 1 where 100 samples each with a size of 10 were taken from a population with a known mean of 42. For each sample, the mean and confidence interval is then calculated. Because this plot is based on simulated data, we know the mean is 42, yet exactly half of the sample means are less than the true mean, half are larger and 95 of the 100 confidence intervals include the true value. In practice someone is not going to take 100 different samples, they are

[1] The 95% confidence interval refers to the middle 95% of the distribution from 2.5% to 97.5% and the degrees of freedom is the number of samples minus one, (n - 1). The t-distribution is used in statistics to estimate the population parameters for small sample sizes. The formula for the t-distribution is quite complicated, it is shown in the appendix and the t-value is almost never calculated manually, it is either looked up in statistical tables or more commonly calculated using built-in formulas in a spreadsheet or statistical software.

https://doi.org/10.52202/072781-0022

probably going to take one. So, the confidence interval is not a probability statement about the mean but a statement about the precision of our estimate of the mean, not the accuracy.

The Central Limit Theorem guarantees that near identical results to these (proportion above and below mean and percentage of confidence intervals) will occur regardless of the distribution of the original data, even if it is highly skewed. This demonstrates why simply using the sample mean is a poor predictor of the population mean. If the sample mean is used to decide if the true mean is above or below some value, you are guaranteed 50% of the time the decision will be wrong and 100% of the time it will not equal the true mean. Keep in mind that the true value is an unknown since the engineer is only taking a sample from a population of unknown parameters.

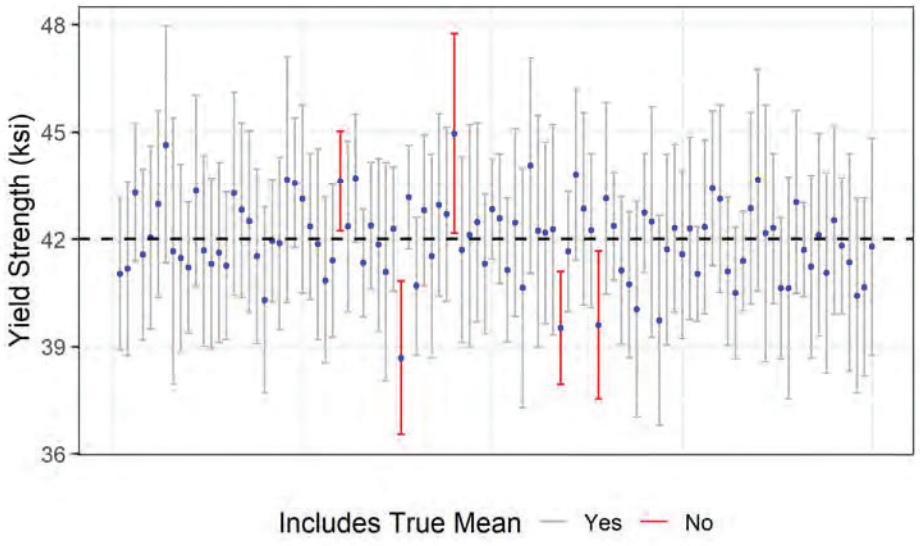

Figure 1. Confidence Intervals

Comparing Two Samples

The previous example is a one-sample test. In the alternate scenario where two samples are being compared against each other, a similar process is employed. But instead of comparing a sample against a hypothesized value as in the one-sample test, the purpose is to compare one sample against the other to determine if they could have reasonably come from the same population. The goal is not to determine consistency with any specific value but rather if it is a reasonable hypothesis that the two samples could have come from a single population with a common mean. There are several different variations of two-sample t-tests but the relevant one to material verification would be an independent two-sample t-test. In this circumstance, it is assumed that each sample has no influence on the other (independence), the variances are equal, and both samples are the same size.

The formula for a independent two-sample t-test is as follows:

$$t = \frac{\bar{X_1} - \bar{X_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (3)$$

https://doi.org/10.52202/072781-0022

$$Where:$$

$$s_p = Pooled\ standard\ deviation = \sqrt{\frac{s_{X_1}^2 + s_{X_2}^2}{2}}$$

$$\bar{X}_{1,2} = Mean\ each\ sample$$

$$n_{1,2} = Size\ of\ each\ sample$$

$$s_{X_{1,2}} = Standard\ deviation\ of\ each\ sample$$

In this case, the null hypothesis (the Null) is that the means of the two samples are equal. The Null is always based on the assumption that the hypothesis is true and it remains so until the data shows it to be very improbable. To accomplish this, the value arrived at in Equation 3 is the t-statistic which is compared against the critical t-value at a probability of $1 - \alpha/2$ with $\nu = n_1 + n_2 - 2$ degrees of freedom. For an $\alpha = 0.05$ with 10 samples in each group, $\nu = 18$, the critical t-value at a probability of $1 - \alpha/2 = 0.975$ is 2.10. Then if $|t| > t_{1-\alpha/2,\nu}$ then the Null (H_0) is rejected in favor of the alternative hypothesis H_A. The critical t-value can be calculated from the t-distribution shown in Equation 10 (see Appendix) or more commonly looked up in a table or calculated with built-in formulas in any spreadsheet[2]. This test determines if it is plausible based on the variation in the samples that the mean is greater *or* less than the the other in what is known as a two-sided test. This process is an example of a Null Hypothesis Statistical Test (NHST).

Statistical Tests and Errors

In the previous discussion, the sample size was given. But when embarking on a sampling program the first question is: "how large does the sample need to be?" To answer that question, a discussion of statistical errors is necessary. Any NHST starts with the assumption that there is no difference between groups or between a sample and a reference value. In these tests, there are two types of errors possible. They go by the creative names of Type 1 (symbolized with the Greek letter α) and Type 2 (symbolized with β) errors. But the meanings of the two are often misunderstood when setting up a sampling program and interpreting the results. To help in the comprehension of the implications of these, the two types of errors will be discussed, what they mean and how they are used as part of planning a testing program.

An error of the first type, also known as a Type 1 error, is rejecting the Null when it is true (false negative). With very few exceptions, the true mean is an unknown value, but it is possible to calculate the likelihood of falsely rejecting the Null. An error of the second type, a Type 2 error, is failing to reject the Null when it is false (i.e. you didn't reject the Null when you should have, a false positive). The plot in Figure 2 shows graphically Type 1 and 2 errors (note that sometimes Type 1 and 2 are designated by their Roman numerals I & II). Where H_0 is the distribution based on the null hypothesis with H_A being the alternate hypothesis distribution. The darker shaded region is the error bounds for α and β respectively. The boundary between the darker and lighter shaded areas would be the threshold to reject or not reject the Null. If it was rejected at that point, the potential Type 1 error would be the dark blue area and the Type 2 is the dark orange area under the curve. These errors are because even though at the decision threshold the Null is very unlikely it is still possibly true since the distribution continues. So, there is still a small probability that the Null or the alternative Hypothesis is true.

[2] In Excel, the formula would be *T.INV(0.975,18)*.

https://doi.org/10.52202/072781-0022

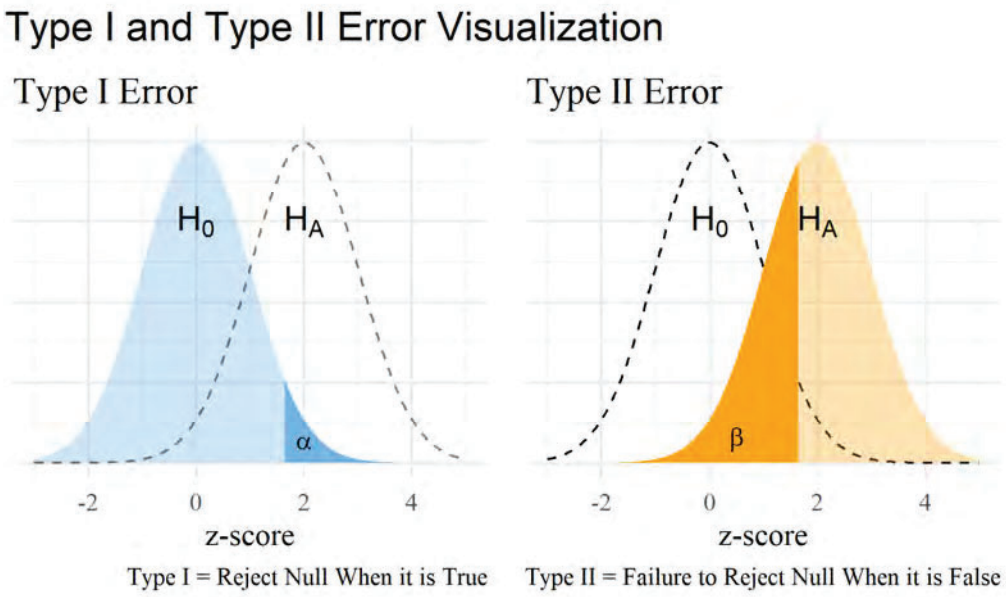

Figure 2. Type 1 and 2 errors

It is evident from Figure 2 that Type 1 and 2 errors are interrelated. Minimizing the Type 1 by setting the threshold for rejection higher (i.e. a lower α), increases the likelihood of a Type 2 for a given sample size. So, this begs the question of which error is worse. The answer is, as in most of statistics, it depends. Typically, the Type 1 error is set at 0.05 in most tests because it is usually more consequential. For example, if a new drug was being tested, the Null would be that there is no difference in outcome versus the placebo then a Type 1 error would be determining that the new drug has an effect on a disease when it really doesn't, and the Type 2 error would be rejecting that the drug had an effect when it did. The consequences of the Type 1 far outweigh the Type 2 in that scenario. But the 0.05 is strictly a convention or typical value, not an absolute law chiseled in stone. If a criminal trial were a statistical test, the Null would be that the defendant is innocent (innocent until proven guilty) and the desirable outcome would be that the likelihood of judging an innocent defendant guilty (Type 1 error) to be as small as possible even at the expense of increasing the likelihood of a guilty defendant going free. Setting a low α value means that you are requiring a high level of evidence before the Null is rejected.

In some circumstances a Type 2 error has a higher consequence, and it is appropriate to set a lower evidence threshold to reject the Null by setting a lower β value. The ability of a test to correctly reject the Null is known as the power and is equal to $1 - \beta$. In this alternate scenario (ignoring any code requirements for the sake of demonstration) suppose there are two lots of pipe in a pipe yard, one with documentation and another without and you suspect that the yield strength is the same for the two lots, so a test program is set up to test them and compare the results. If it is judged that the Null (i.e. that they are the same yield strength) can't be rejected, then the undocumented lot will be used along with the documented one for a project. In this case, the Type 1 error would be rejecting that they are the same when they really are and consequently not using the undocumented pipe, a cost consequence but not a safety one. But a Type 2 error would imply that the undocumented lot gets used in the project under the assumption that the material properties are the same when they are not, a potential safety consequence. In this case, the consequences of a Type 2 are much greater than

https://doi.org/10.52202/072781-0022

a Type 1 therefore a higher power test would be desirable, creating a lower threshold of evidence before you reject the Null. There are two ways to increase the power of a test, the first is to lower the acceptable Type 2 error, which in turn raises the Type I error, and the other is to increase the sample size. See Section 4.1 for a more detail explanation. In this scenario, it was decided a higher Type 1 error was acceptable for the benefit of higher likelihood of properly rejecting the Null. If a higher Type 1 is not an acceptable risk, then the only other alternative is to increase the sample size to get the required power.

One-Tail vs. Two-Tail Test

Upon conclusion of the discussion of error types it is necessary to cover the types of tests and how they are used. The t-test which was introduced earlier is a statistical test that compares the means of two groups or one group relative to a hypothesized value. A one-tailed t-test is used when the researcher is interested in only one direction of difference between the two groups. For example, if we want to know whether the mean of Group A is significantly greater (or less than) than Group B, we will use a one-tailed t-test. A two-tailed t-test, on the other hand, is used when the researcher is interested in any difference between the two groups in either direction. For example, if we want to know whether the mean of Group A is significantly different from Group B (either greater or less than), we will use a two-tailed t-test.

In a one-tailed test, the entire alpha level is in one tail, while in a two-tailed test, the alpha level is split in half between the two tails. One-tailed tests detect an effect in one specific direction, while two-tailed tests consider effects in both directions. One-tailed tests require prior knowledge about the direction of the effect, whereas two-tailed tests do not. If the researcher is using a one-tailed test, then the critical t-value as was discussed in the section on t-tests for rejecting the Null is based on $1-\alpha$ vs. $1-\alpha/2$ for the two-sided. The difference in the two tests is shown in Figure 3, plots A & B are examples of one-tail tests for left and right tails respectively and C shows a two-tail test.

https://doi.org/10.52202/072781-0022

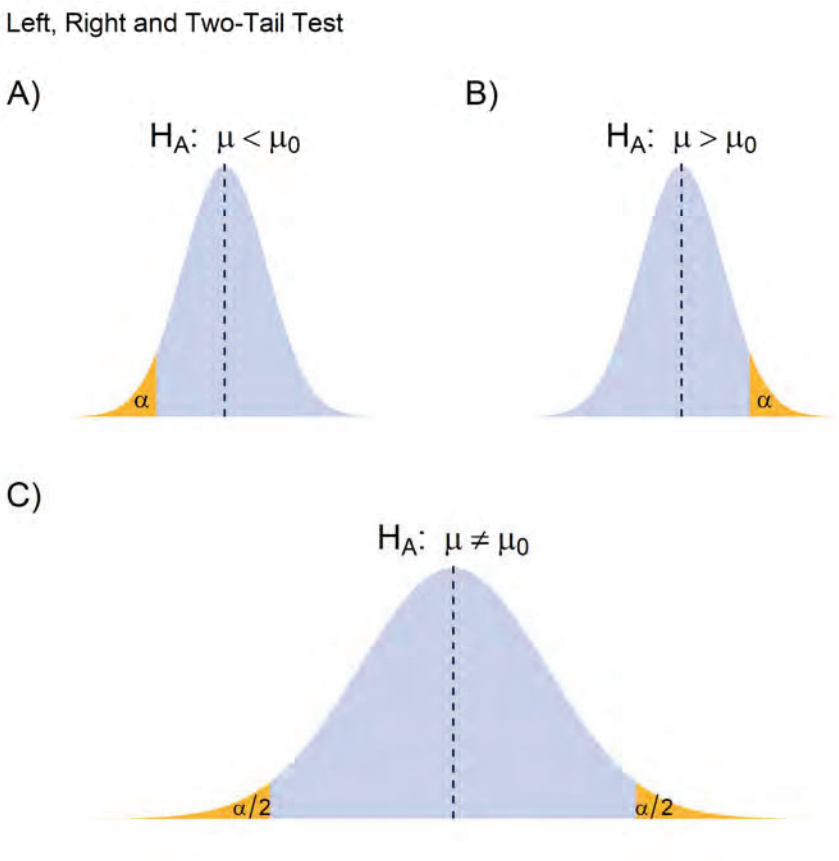

Figure 3. One-Tail vs. Two-Tail Test

Sample Size

If the first question in a sampling program is, "how do I know?", the second question that instinctively comes up when a testing program is being considered is, "how much is enough?". The answer to this is dependent on the answer to several questions. The first one is how big is the effect that the researcher wants to detect. The effect size can be thought of as how much of a difference between the sample and the hypothesized value is a practical concern. There is always some difference, how big a difference does it need to be before it is considered significant? The other question that must be answered is what is the acceptable error rate? This ultimately involves defining your acceptable risk, it is impossible to reduce the potential error to zero outside of sampling the entire population, so the person designing the sampling plan must decide what is the acceptable error rate they can live with and what are the risks. Since there is always going to be some difference, the question the researcher needs to decide ahead of time is how large the difference can it be before it becomes a concern. If the testing program is trying to detect if the mean value is 10 vs 10.5 in two samples, that is much smaller effect size and requires a much larger sample than trying to differentiate 10 from 15. The size of your sample is also dependent on the required power of the test, which is also set by the researcher ahead of time, recall that the power of a test is simply $1 - \beta$ and that β is the Type 2 error rate as discussed in the section on error types. The larger the sample size, the larger the power of the test (i.e. the smaller the Type 2 error). The power can be calculated after the data is collected but normally the person planning the sampling program will specify the required test power and significance which will in turn determine the size of the sample required. Generally, (again, not an absolute) it is

https://doi.org/10.52202/072781-0022

desirable for a test to have enough samples to achieve at least a power of 80% which corresponds to a 20% Type 2 error. But as was discussed previously, the size of the acceptable β is dependent on the relative consequences of Type 2 vs Type 1 error. The reason the increased sample size creates increased power of the test is that as the sample size grows, the variability in the mean shrinks, making it easier to differentiate between the two hypotheses. Note that the uncertainty in the mean is not linear with sample size, it's inversely proportional to, $\sqrt{n}$ this means that if you want to shrink your uncertainty in half, it will require a sample four time as large. The power versus sample size for various effect sizes and the number of samples required to reach 80% power is shown in Figure 4. The effect size (d) is calculated by taking the difference of the two means and dividing by the pooled standard deviation (pooled standard deviation calculation is shown in Equation 3). The three effect sizes shown in Figure 4 correspond to what is roughly categorized as the thresholds for a small, medium, and large effect sizes respectively. These are the general conventions for categorization of the effect size. The necessary effect size is dependent on the application and the preferences of the decision maker. These example plots are for a two-sample, two-sided t test. This demonstrates the influence effect size has on required sample size. A "large" effect can be detected with as little as 26 samples where a "small" effect requires over 15 times that amount at 394. In summary, the required sample size is dependent on the size of effect the researcher wants to detect and the acceptable Type 1 and 2 error rate.

https://doi.org/10.52202/072781-0022

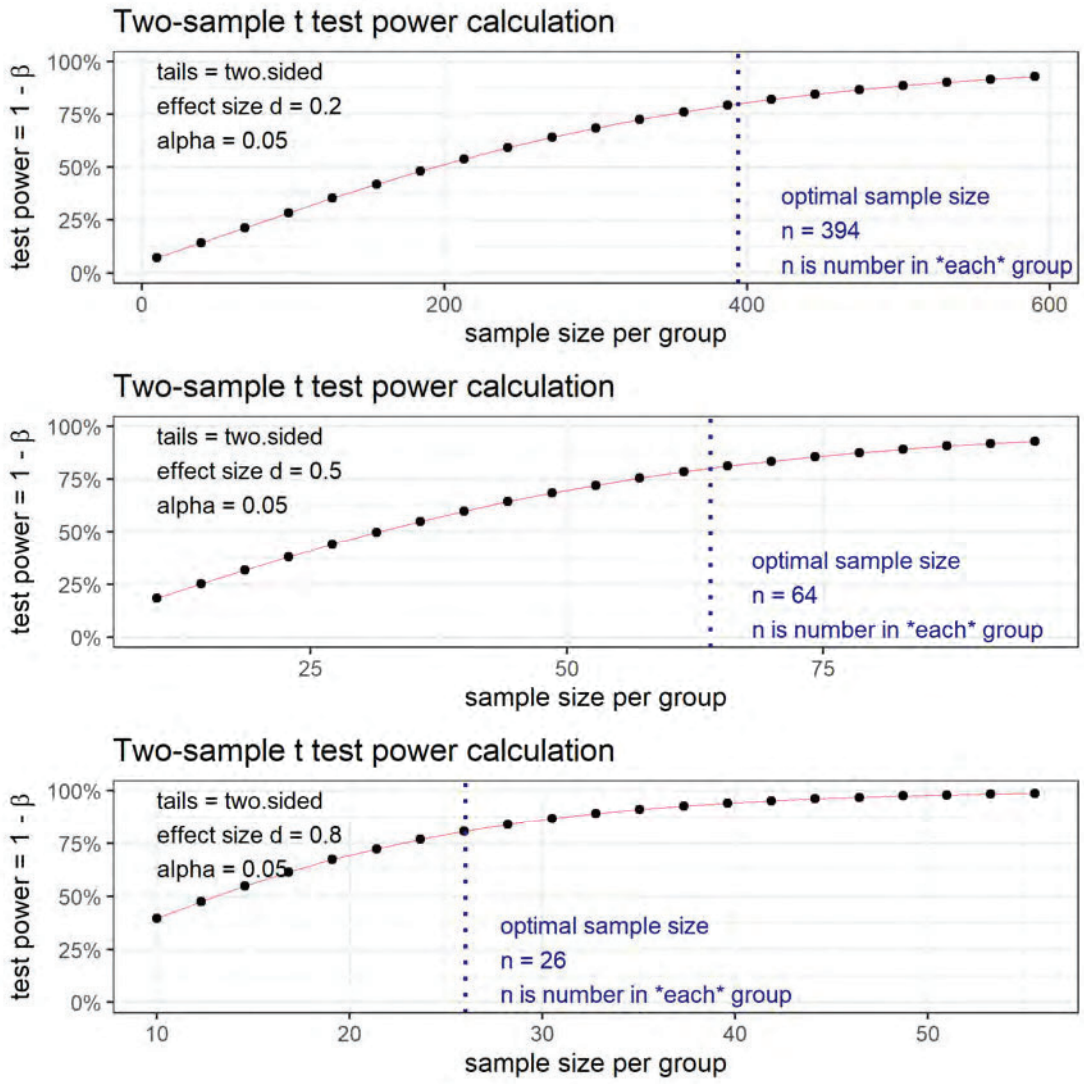

Figure 4. Sample Size vs. Power for Two Sample t-test

It is evident from Figure 4 that the effect size makes a significant difference on the size of the sample required to reach a power of 80%. While more data points is always better, the cost of collecting additional data is a real world concern and should be weighed against the benefit of further reducing the likelihood of a Type 2 error relative to the current sample size.

How the power is calculated is dependent on the type of data that is being analyzed and all circumstances won't be covered in this paper, there are several online resources and texts that discuss it in detail. But there are also numerous statistical power calculators online that can be used. The example of a more common case involving two normal distributions like H_0 and H_A shown in Figure 2 will be worked here.

https://doi.org/10.52202/072781-0022

Normal Distribution Sample Size Calculation

In this example, the Null is that the mean value of the YS is 45 ksi and the goal is to be able to detect a difference of more than ± 3 ksi at a 95% confidence level and a power of 80%. Therefore, the researcher wants to be able to reject the Null when the estimated mean of the population is less than 42 or greater than 48 ksi. Based on industry data, it is known that the typical standard deviation for the YS is 6 ksi.

Since the comparison is one population against a reference value this is a one-sample test. As a simplification to the problem, the normal distribution will be used rather than the t-distribution. The t-distribution is required when the standard deviation is unknown or can't be estimated. But recall that the t-value is dependent on the sample size which we are trying to estimate so it requires an iterative process of guessing a sample size and calculating the power and adjusting the sample size based on the calculated power. For even moderately large samples (generally n > 30) the difference between the t and normal distribution is small and diminishes as the sample size gets larger. Therefore, since the standard deviation can be estimated, we will use the normal distribution.

The first step is to calculate the critical *z*-values at $1-\alpha/2$ and $1-\beta$ which in this case would be 1.96 and 0.84 respectively. Then for a normal distribution the minimum sample size for a two-sided test is as follows. For simplicity, the formula is presented here without derivation. If the reader is interested in how this formula is arrived at, see https://www.itl.nist.gov/div898/handbook/prc/section2/prc222.htm for more background.

$$N = \left(Z_{1-\alpha/2} + Z_{1-\beta}\right)^2 \left(\frac{\sigma}{\delta}\right)^2 \qquad (4)$$

$Where$:
$\sigma = Standard\ Deviation\ (6\ ksi)$
$\delta = Minimum\ difference\ between\ H_0\ and\ H_A\ that\ you\ want\ to\ detect\ (3\ ksi)$
$Z = Critical\ Z\ value\ for\ normal\ distribution$
$\alpha = Type\ 1\ error$
$\beta = Type\ 2\ error$

Putting this all together:

$$N = (1.96 + 0.84)^2 \times (6/3)^2 \approx 31.4$$

Since a fractional sample has no physical meaning, the number is rounded up to the next whole number of 32. If the goal is only a one-sided test the same formula is used except for $Z_{1-\alpha/2}$ is replaced with $Z_{1-\alpha}$.

Expanded Sampling Requirements

As part of this testing requirement under 192.607(e)(4) if a sample is found to be inconsistent with available information or existing expectations or assumed properties used for operations and maintenance in the past, *"the operator must establish an expanded sampling program. The expanded sampling program must use valid statistical bases designed to achieve at least a 95% confidence level that material properties used in the operation and maintenance of the pipeline are valid."* The regulation does not define what is a "statistical valid bases" or how to get to a 95% confidence level. This section will discuss the meaning of confidence level and how it applies to this regulation and

https://doi.org/10.52202/072781-0022

Confidence level is not a probability of the hypothesis being true but rather a statement of how likely we are to make the correct decision based on the testing. In essence, PHMSA is asking the operator to prove to a 95% certainty that the inconsistent material does not exist in the rest of the population. It is a common saying that you can't prove a negative, which is true. However, if an estimate is made (or assumed) about the population it can be shown what the probability of observing certain results are if the estimate was true. Therefore, it can be shown that something is highly improbable based on results, but the negative still can't be proven absolutely.

Statistically Valid Bases

If an inconsistent result is found (however "inconsistent" is defined by the operator) and the operator must implement an expanded sampling plan per 192.607(e)(4), then each verification dig can be thought of as a pass/fail test for whether inconsistent results are found or not. In statistics, this type of sample where the criterion is a binary choice, is known as a Bernoulli trial and is modeled with the binomial distribution. This distribution will determine a probability of k successes in n independent trials if the probability of success per trial is p. A "success" is whatever is being counted. A "success" is not necessarily a positive outcome. In the expanded sampling context, a "success" would be finding an inconsistent material property in a verification dig. The binomial equation is shown as:

$$\binom{n}{k} p^k (1-p)^{n-k} \qquad (5)$$

Where the $\binom{n}{k}$ is what is called the binomial coefficient which calculates the number of ways k successes in n trials can be distributed. In an expanded sampling program, the desired number of successes (inconsistent results) is zero, but the number of samples required to achieve a 95% certainty is not. It would be straight forward to show the probability of observing zero successes in a number of trials, but this doesn't show how likely it would be to find something if additional samples were taken. For instance, the probability of zero successes in one trial is 95% if the inconsistent material is 5% of the population. But even the casual observer would realize that doesn't prove anything, and the result of that one trial could simply be the result of random chance. That's equivalent to saying 95% of all cars are white because the first one you saw was white. To show the probability of **not** finding something in future trials it is necessary to calculate the opposite, the probability of finding **one or more** successes in n trials as shown in Figure 5. Since the total probability of all possible outcomes must sum to one for a given sample size, the probability of finding one or more is one minus the probability of finding zero.

https://doi.org/10.52202/072781-0022

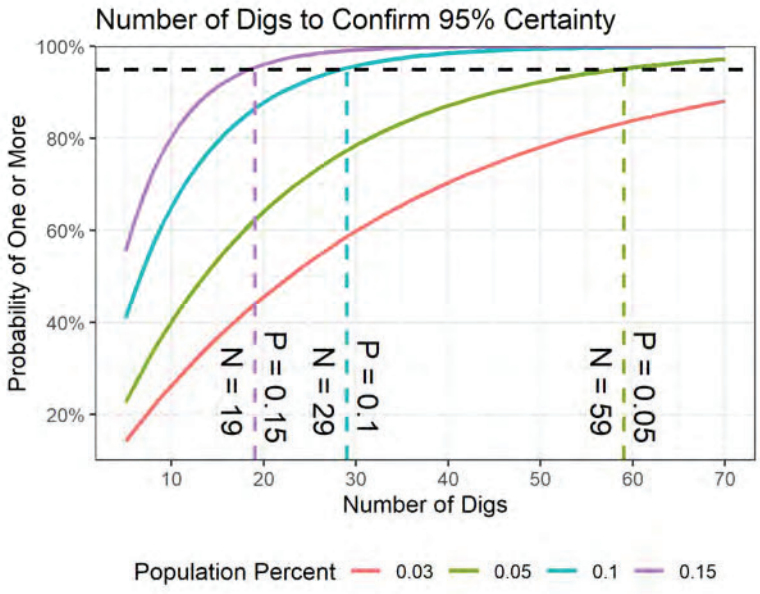

Figure 5. Cumulative Probability

To facilitate this calculation, to estimate or assume the percent of the population that might be represented by the inconsistent sample, and this become the probability of success in Equation 5 and the number of successes is set to zero. Then the sample size is iterated until one minus the probability is greater or equal to 95%[3]. The reason for this is that the proportion of the population affects the probability of finding it in future samples. The larger the proportion it represents, the easier it would be to find in fewer trials and likewise, the smaller the proportion, the more trials it would take to find it even once. What Figure 5 illustrates is that if no inconsistent samples are found by the time the probability reaches 95%, then the likelihood of finding additional ones in future digs is 5% or less if the proportion of population estimate is correct. For example, if it is hypothesized that the proportion of the population that is inconsistent is 5% then 59 samples would be required to get to a 95% probability of observing one or more of them. If after 59 examinations, none are observed, then it can be concluded that there is a 95% probability that the inconsistent material is 5% or less of the overall population. If an operator wanted to demonstrate that the population of inconsistent material was a smaller percentage, continued sampling would be required. In Figure 5 the line for 3% of the population (bottom curve) doesn't reach 95% even after 75 trials. To demonstrate the proportion is 3% or less would take almost 100 samples before reaching 95% probability.

Alternate Sampling Plan

Section 192.607(e)(2) requires one excavation per mile for each population rounded up to the nearest whole number up to a maximum of 150 excavations if the population is more than 150 miles. If an operator chooses to use an alternate statistical sampling, 192.607(e)(5) allows the use of another method that has a statistically valid basis designed to achieve a 95% confidence level.

[3] The binomial distribution is built into Excel, using that to do the iteration will make this process easier and repeatable.

https://doi.org/10.52202/072781-0022

The method presented here will provide a robust, defensible method to update the beliefs in the value of material property such as yield strength of an unknown material based on new information provided by sequential NDE testing. However, with the application of statistics, specifically Bayes' Theorem, it can be shown that a very high level of confidence can be achieved with a relatively small number of tests. This paper will discuss how to determine yield strength with far fewer tests than is currently required under 192.607(e)(2).

Bayes' Theorem

Intuitively, people know that as they acquire more samples that agree with their hypothesis, the probability of it being true should increase as well and it shouldn't take up to 150 samples. That intuition is correct and the way to show this increase in belief is through Bayes' Theorem.

The goal of an alternate sampling program is to demonstrate the the probability of a hypothesis being true after observing some data. The optimum tool to accomplish this goal is Bayes' Theorem. This was originally referred to as the inverse probability problem. In traditional frequentist statistics, the hypothesis is chosen and the goodness of fit of the data is judged. In Bayesian statistics, the data is given, and the likelihood of the hypothesis is judged by how well it fits the data. Frequentist methods require much larger samples because the user must effectively pretend that they don't have any prior knowledge of the likelihood of the hypothesis being true. Whereas Bayes' Theorem works well when there is sparse data, but prior information is available. The key aspect of Bayes' Theorem is the ease with which beliefs can be updated based on new data. The reason for the choice of this methodology is that it allows the incorporation of prior knowledge of the problem. Bayesian statistics starts with a prior distribution, referred to as simply "the prior", and is shifted as new information is acquired. The prior can be thought of as the starting point that is conditioned (updated) in proportion to the weight of new evidence.

Nomenclature

In the following example when the notation of P followed by parentheses such as P(x) this represents the probability of the independent event happening, the probability of observing x in this case. When there are two variables separated by a vertical bar such as P(H $|$ x) this is a conditional probability. In this example it would be the probability of the hypothesis, H given that x (the data) has been observed. It is read as the probability of H given x.

The most basic form of Bayes Theorem is represented by the following equation.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \qquad (6)$$

There are two main components to Bayes' Theorem, the hypothesis (the criterion of interest) and the evidence (data). When there is only a binary condition being considered (e.g., true/false, sick/not sick) Bayes' Theorem can be written conceptually as:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} = \frac{P(E|H)P(H)}{P(E|H)P(H) + P(E|H')P(H')} \qquad (7)$$

https://doi.org/10.52202/072781-0022

Where:

P(H|E) = Probability of hypothesis is true given the evidence (e.g. probability of disease given a positive test)

P(H) = Probability that hypothesis is true overall (e.g. percentage of population that have the disease)

P(E) = Probability of observing the evidence if the hypothesis is true. In the case of disease diagnosis, it would be the P(positive test if disease is true)+P(positive test if no disease), i.e. True positive rate + False positive rate

P(E|H) = Conditional probability of observing the evidence if the hypothesis is true. (e.g. probability of a positive test if the disease is present)

P(E|H') = Conditional probability of observing the evidence if the hypothesis is false (e.g. probability of positive test if the disease is not present)

P(H') = Probability the hypothesis is false (e.g. probability the disease is not present)

In the above construct, H' represents the hypothesis being false.

Sequential Updating

The method for sequential updating of a Bayesian inference starts with the prior probability (prior) of the event happening and a probability of the test correctly identifying the value of interest. Using this prior information and known test accuracy an updated probability is calculated after each round of testing. Then after each test the updated posterior becomes the prior for the next test. This is illustrated with the following example. The prior is simply a statement of the strength user's belief that the hypothesis is true based on what is known a priori about the pipeline. This can be based on design standards and operating pressure. It's not intended to be a long-running frequency but rather how confident the user is in the hypothesis based on any prior information before seeing the data.

The average person uses prior probability all the time without thinking about it. If someone makes an extraordinary claim about something regardless of the subject, you are skeptical. Why? Because you placed such a small prior probability on it being true based on experience and knowledge. Even after you learn that the person making the claim is truthful 90% of the time you would remain unconvinced. This is because even the 90% rate of truthfulness isn't enough to overcome the very small prior you placed on the claim. You didn't consult a database of such claims to develop the prior, you knew from general knowledge and prior experience that the probability of truthfulness of that claim was very low. Your prior is just the starting point, as data is accumulated it will move the updated probability either toward or away from the prior depending on the outcome.

In this example, nondestructive testing is used to determine an unknown yield strength (YS). It is known from qualification testing that the tool will produce results that are either correct or conservative 60% of the time. Based on prior knowledge of the operating pressure and construction practices in the time frame of its installation it is believed that the true YS is 42 ksi or greater (in other words, the specified minimum yield strength is that corresponding to a grade X42). But given that the YS is not TVC, the initial belief of that is limited to a skeptical 10%. The initial belief level is somewhat arbitrary, but it should be dependent on the level of prior information that is available before testing.

https://doi.org/10.52202/072781-0022

Assuming the first test is successful (YS ≥ 42), the probability of being YS ≥ 42 ksi (YS_{42}) in untested samples is updated according to Equation 6 as follows.

$$p(YS_{42}) = \frac{(0.10)(0.60)}{(0.10)(0.60) + (1 - 0.10)(1 - 0.60)} = 14.3\% \qquad (8)$$

The 14.3% then becomes the prior for the next test. If the second test is successful then the updated probability becomes

$$p(YS_{42}) = \frac{(0.143)(0.60)}{(0.143)(0.60) + (1 - 0.143)(1 - 0.60)} = 20.0\% \qquad (9)$$

Then each test is updated in the same way until the probability of success meets or exceeds 95% as shown in Figure 6. In this example, the probability that $YS \geq 42$ in the untested population of similar pipe after 13 samples. This example assumes that the population is homogeneous, and any variability is due to the uncertainty in the measurement. The actual number of confirming samples required will depend on the prior probability used and the accuracy of the NDE method used.

One caveat of this calculation is that it assumes that the population is homogeneous, and you are only attempting to infer if it meets some threshold value. We are not concerned about what the true value is, only the binary hypothesis that it meets or exceeds some threshold.

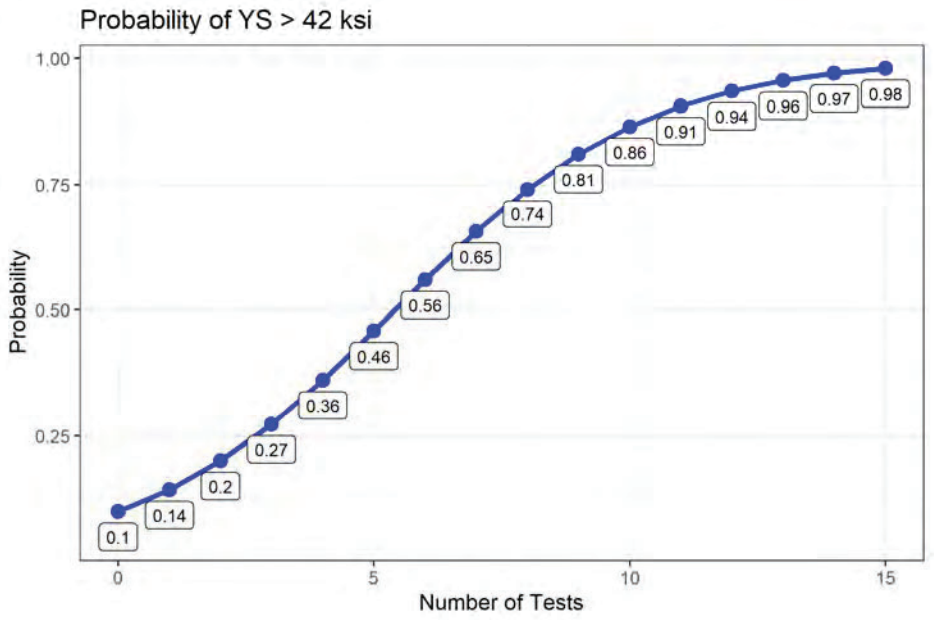

Figure 6. Updated Beliefs

Conclusion

All the methodologies presented here are defensible and can be explained from the first principles of statistics. Your selection of the error threshold is dependent on the consequences tied to a Type 1 or Type 2 error. The two errors are intertwined with each other, and it is impossible to minimize both at the same time for a given sample size. It is necessary to pick your poison of which has the lesser consequence, a false positive or false negative and set your threshold accordingly. In addition,

https://doi.org/10.52202/072781-0022

a methodology of updating an initial belief based on sequential testing using Bayesian statistics is presented along with the process of determining the number of samples needed for an expanded sampling program.

These principles explain how to set up a robust sampling program to meet the objectives and minimize potential errors. Following this guidance will allow the end-user to not only define their sampling program but more importantly be able to explain the process and what was done and why.

Appendix

t-distribution formula

The probability density function (pdf) of the t-distribution with ν degrees of freedom is given by:

$$f(t|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad (10)$$

Where Γ is the gamma function:

$$\Gamma(z) = \int_0^\infty x^{z-1}\, e^{-x} dx \qquad (11)$$

https://doi.org/10.52202/072781-0022