Citizens Exhibits 38

RAS 14355

CORRO-CONSULTA
Rudolf H. Hausler
rudyhau@msn.com

8081 Diane Drive
Kaufman, TX 75142
Tel: 972 962 8287 (office)
Fax: 972 932 3947
Tel: 972 824 5871 (mobile)

Memorandum
August 16, 2007

Response To The Questions About Statistics

I. Some Background on the Origin of Statistics

- Collecting data of all sorts is a basic human and/or societal occupation. The data one collects are very simply observations put into a quantitative form. For instance, the child may separate out the red from the green pebble at the beach and then count them

- Statistics is a tool to organize and describe some innate properties of data, i.e. properties of the observations one has made.

- The most prevalent observations are measurements, and the most important property of measurements, any measurement, is the fact that they are not absolutes, as we tend to assume, but in fact estimates. Hence, the data we collect are estimates and statistics is a tool to describe certain properties of these estimates.

- What are these properties? Take an example: We have a corroded surface, which is characterized by pits. (Pitting is a most prevalent form of corrosion). We would like to know how deep the pits are. We take a micrometer, for instance, or a microscope, and start measuring the depth of pits. We measure 87 pits and acquire 87 data points each describing the depth of a pit. We now need to describe what we have done in some way that is more concise and more understandable than a collection of 87 data points. We take recourse to some statistical tools.

- The first thing we do is to calculate the **average pit depth**, also called the mean. The next day the boss comes with another piece of corroded surface, from another structure, and wants to know whether the pits are the same, or maybe were caused by a different phenomenon. So, we start all over again measuring pit depth but only take 47 pit depth measurements. We quickly calculate the average of the 47 data points and because the new average is slightly different but we then have a statistical problem – **how much confidence can we have that the two averages (means) are indeed different or come from the same universe of pit depths.**

Template=SECY-028

SECY-02

II. What Statistics can do and can't do

- One needs to recognize that statistics is a set of rigorous mathematical equations to answer certain very specific questions one may have about a collection of data. However, one also needs to recognize that as with all mathematical theories, they are based on assumptions, and one always needs to test whether the assumptions in a given mathematical procedure actually reflect the true nature of the phenomenon under scrutiny. Example: if the variation of the pit depths is truly normal, then Gaussian statistics will apply because they are based on the observation that measurements close to the mean are more frequent than the ones further removed. Since AmerGen and NRC have decided that Gaussian statistics are what applies to the sandbed corrosion problem, that is what we will discuss first. (An alternative will be discussed later on).

- The basic assumption underlying Gaussian statistics is the notion that the variations within the universe of data, which characterize a particular parameter, are distributed normally as defined above. The universe (also sometimes called the population) therefore can be represented by the well-known bell-curve. This simply means that the frequency of data close to the mean is higher than the frequency of data removed from the mean.

- The Gaussian or Normal Distribution curve shown in Figure 1 is rigorously described by a mathematical formula which simply says that the logarithm of the probability density p(y) (or how often a value occurs within its population) of a particular value is a quadratic function of the difference between the value and the mean of the population.

$$p(y) = const. \cdot 1/\sigma \cdot e^{-[(y-\eta)^2/2\cdot\sigma^2]}$$

where y = attribute (value to be measured)
σ = standard error
η = mean of the population

III. The Error Measurement

- Since one can never measure the entire population, but only samples of the population, the average of the sample (the sample mean) and the standard deviation of the sample (standard error)[1] become **estimates** of the true population mean and the true probability density distribution, i.e. the true standard deviation. This, however, is only true if the samples have been selected randomly. The difference between a value y and the mean of the population is often called the error. However, the error is a complex function of a lot of things, some can be controlled others cannot. Therefore we would prefer to call this difference ((y-η)

[1] The variance is the square of the standard deviation

the variability in the data. The variability is composed (in the simplest case) by the "innate" reproducibility of the instrument (the accuracy of the instrument [2)]) and by the natural variation of the measured parameter (pit depth as a function of location on the surface).

- This is demonstrated in Table 1. During the 2006 refueling outage the external UT measurements from 1992 were repeated. We had already shown [3)] that even though there was a slight bias between the 1992 and 2006 data (of about 20 mil) which might have been interpreted as ongoing corrosion, this bias was not statistically significant in view of the inherent variability of the data (individual UT measurements). In 2006, additionally, duplicate and triplicate measurements were made externally in some bays. From these repeat measurements it was possible to estimate the standard deviation associated with the error of the measurement only. The results are summarized in Table 1 below.[4] In the columns headed by 'std. dev. variability' we calculated the spread of the remaining wall thicknesses between pits as a standard deviation. It is easy to understand that the wall thickness measurements will vary from pit to pit because one could not expect corrosion to be uniform over the entire surface [5)]. Hence one finds the wall thicknesses vary within certain limits as expressed by the standard deviation [6)].

Now, it must be also understood that the variability of the data arises from two sources: a) the actual variation of the pit depths (residual wall thickness) and the measuring error. If only wall thicknesses are measured, the two effects are hopelessly confounded. In the present case, in 2006 duplicate measurements were made in some cases as shown in Table 1. It was then possible to estimate the measuring error form these repeated measurements. As Table 1 shows the measuring error (Std. Dev. from 2006 Repeat Measurements) is smaller than the variability of the wall thickness measurements themselves. This is of course anticipated since the latter contain, as mentioned above, both the error as well as the variability of the wall thicknesses. It is interesting to note that the measurement error depends on the roughness of the surface. Indeed, if the error is plotted against the variability a straight line results which extrapolates roughly to 0.01 inches at 0.01 inch (see Figure 5), i.e. the instrument error one would expect, from the manufacturers specifications, if the surface had not been corroded and were still in a pristine state.

[2)] UT instruments are generally said to be accurate within +/- 1 to 2% of wall thickness, where newer instruments are more accurate than the older ones.

[3)] R. H. Hausler Memorandum to Richard Webster, Esq. April 25, 2007, Figure 7

[4] The standard deviations derived from repeat measurements shown in Table 1 differ slightly from those previously presented, because I have used a more rigorous calculation method than previously.

[5)] In acid for instance one could anticipate and observe that the thinning of the probe (wall) is uniform over the entire surface with no evidence of pitting. In neutral solution where the build-up of corrosion product layers can be anticipated corrosion will not be uniform over the surface and localized attack can be anticipated.

[6)] If the pit depths, or by implication the remaining wall thickness measurements, are normally distributed, characterizing the variability with a standard deviation makes sense. If the distribution is a different one (which we suspect) then the data spread should be characterized differently.

IV. Definition of Confidence Limits and Statistical Testing

Referring back to Figure 1, it turns out that about 66% of all data belonging to the universe of data estimated by the standard deviation are within the mean m +/- 1s and 95% of the data are roughly within m +/- 2s. The exact multiple of s to be used is derived from the Student's t-distribution, which takes account of our imprecise knowledge of the true population variance. The multiples of s that encompass a certain portion of the data decrease as the number of samples increase. At the limit, when the number of samples is large, the students-t distribution becomes equivalent to the classic Gaussian distribution

Now it may arise that one asks of a particular measurement whether it belongs to this universe. If the point is more than two s removed from the mean there is a certain probability that it does not belong and vice versa. One can specify that probability with the Student's t-test, which calculates the chance that a sample would deviate that much from the population mean. The student t-test may also compare the means of two sets of measurements to calculate the chance that the difference between the means is caused by random variation.

Similarly one may ask the question whether the variabilities of two sets of data are the same. The comparisons are made on the basis of the variances (s^2) rather than the means and the test is called the F-test. Imagine two machines turning out bolts. The mean length of the bolts is the same for both machines, but the variance, i.e. the spread of the length measurement is different. The F-test indicates the probability that the difference in the variance is real. It might tell us that one of the machines needs to be better adjusted. Similarly, one might ask the question whether the variance of the measured residual wall thicknesses for Bay 5 in 2006 is different from that of Bay 19. If the F-test returns a probability of between 95 and 99%, this might tell us something about the corrosion mechanism or the cause of corrosion in the two Bays.

Getting back to the simple comparison of two measurements, one may for instance posit the hypothesis that the mean m_1 does not belong to the same universe of data as m_2. Based on the Student t-test one may find that there is a 75% probability that the hypothesis is true. Customarily one would not accept this as sufficiently significant. If how ever there was a 95% probability for m_2 to belong to a different universe, one would very likely accept the hypothesis. Happily, these statistical probabilities have been calculated and are available in tables. Even better, nowadays, computer codes have made life very easy for the experimental statistician (see for instance Fig 2 to be discussed later).

The 95% confidence limits based on the 95% probability of a specific hypothesis being accepted as true is often used in science as an indicator that it is unlikely that the observed effect is caused by random variation. However, there is no set standard in the literature or in engineering that imposes such limits. There may be certain recommended practices, which embrace this limit as practical but not as imperative. The reason for this is quite understandable. From a practical point of view 95% confidence may be too little if the consequences of drawing an erroneous conclusion are large. In such cases, the required confidence limit could as well be 99%, or in the case of GE, which advertised 5s as the a company wide standard, it is more like 99.9%. What is, however vitally important is risk assessment. To perhaps clarify the difference between probability (confidence) and risk contemplate the following situation: A blind man crossing a major thoroughfare has a one in a hundred chance of being hit by a car. The confidence level therefore is 99% that he will make it across. The risk he takes, however, is unacceptable, because he would cross thousands of roads in his lifetime and so a he would have a very high chance of being hit. Hence, assessing the confidence one may have in acquired data is only the first step, albeit an important one, in assessing risk. Generally, it is the process of risk assessment that imposes the confidence level to be used.

In the pipeline industry both internal and external corrosion damage is assessed by ILI (internal line inspection, also sometimes referred to as intelligent line inspection). Pigs equipped with sensors (either UT or magnetic flux leakage – and in earlier years mechanical calipers) are pushed through a pipeline, and the responses of the sensors recorded, to be downloaded and interpreted after the run. The API (American Petroleum Institute) has prepared a Standard for In-Line Inspection Systems Qualification [7]. The 65 page documents standardizes the entire process. We are here only interested in the statistical handling of the data. As the title indicates it is the "system" that is being standardized. The basis question is: "how accurate is the instrumentation in the pigs"? This is being demonstrated by verification of the indicated corrosion (or other) anomalies where the buried line can be accessed. The anomalies are located by means of the distance measurements (made by the pig) and verified with an alternate method or instrument. The paradigm for verification is as follows: The pig is specified (by the manufacturer) to detect anomalies with an accuracy of +/-10% of wall thickness. Verification occurs with a certain additional error, say 5%. If the anomaly is verified within the pooled accuracies it is accepted, otherwise it is rejected. The aim of this verification is to be 95% confident that the measurements of the pig have identified at least 80% of the anomalies correctly and within the set specifications. The thinking is not applicable to the Oyster Creek drywell shell problem, because no effort has been made to independently verify the corrosion anomalies. At Oyster Creek it is assumed that the UT measurements are correct. The Pipeline example however shows that a 95% confidence limit is considered adequate, albeit not imperative.

There is power in Numbers

There is another side to using statistics. As indicated above, by repeating measurements one can determine the standard deviation (accuracy) of the

[7] API – 1163 Qualification of In-Line Inspection Systems, 2004

instrument and calculate a mean for the particular point that was measured. The more measurements that go into the mean the more accurate the mean becomes. Expressed in a formula:

$$s_m = \frac{s_x}{\sqrt{n}}$$

where s_m = standard deviation of the mean
s_x = standard deviation of a single measurement
n = number of measurements used for the mean [8]

Applied to the problem at hand this would mean that the more wall thicknesses are determined, the greater the confidence we can have in the mean. However, this will not change the distribution (spread) of the wall thicknesses, it will only better define it and better define the average wall thickness. But structures do not fail by averages. Just likes storms do not destroy villages, but extreme storms do. Structures, like pipelines, for instance, fail where the deepest pit is located. So, how do we apply this thinking to the drywell?

V. Margins and Extremes

Sandia recently made a new study of the integrity of the drywell. The result was that the safety factor for the undegraded shell was 3.85 while for the degraded shell is was 2.15, or 7.5 % above the minimum safety factor specified by the ASME code. The input to the calculation for the degraded shell was all the external thickness data from 1992 some of which are shown in Table 1 below. The first question we may ask is whether the 100 odd data points truly represent the state of the corroded drywell. The answer is: not likely, but that is speculation, perhaps. However, we do know that the standard deviation of the error for the individual measurement is of the order of 0.029 inches in the more heavily corroded areas, the ones of interest. This means that the 95% confidence interval for individual wall thickness measurements is of the order of 0.058 or also about 7.5% of the remaining wall thickness. The model took no account of this uncertainty. Furthermore, the Sandia model has other non-conservative features and was designed to provide accurate absolute predctions. Specifically, the model did not attempt to model the actual shapes and placement of the observed corrosion features, and it did not use the 2006 data which all agree is more accurate and shows the drywell shell is on average approximately 0.02 inches thinner than measured in 1992 data. Therefore, the Sandia model does not establish any margin with a high degree of confidence.

[8] This formula essentially says that the means of samples are more narrowly distributed than the samples themselves. This is a direct consequence of the "Central Limit Theorem".

In addition, we question whether the buckling models should take explicit account of other uncertainties, including variation in nominal wall thickness [9)], variations in tensile strength, variation in temper properties, inclusions in the steel sheet (of which the UT tests seem to have found a relatively large number), and perhaps many more. It may be that the safety factor of 2 is designed to take account of these, but it appears that designers err on the safe side to ensure that this requirement is met with a very high degree of certainty at the outset, when the variation in wall thickness is much less than it is after 40 years of corrosion.

There are two criteria of primary interest with respect to the integrity of the Oyster Creek Dry well. First is the buckling criterion discussed above and addressed by the Sandia study. Then there is a pressure criterion, which says that a corroded area thinner than 0.536 inches shall be greater than 0.49 inches and not larger than 2.5 inches in diameter. In this case only one relatively small area corroded fairly deep could in fact lead to non-compliance with the safety requirements. The trouble is one does not know where this spot could be. One therefore tries to extract from the available data whether such a spot could exist and with what probability.

In the first attempt to answer the question one would probably examine the available data for "normalcy" (i.e. normal, or random, distribution). Figure 2 shows a histogram generated in the SAS software called JMP. While the computer program churns out a result it is up to the operator to decide whether the result justified the underlying assumption. In this case one clearly recognizes, even without any further statistical tests, that the assumption of "normalcy" is not fulfilled. (In fact statistical tests, which are not shown in the printout do confirm this). The aspect of a histogram often depends on the "bin size" (i.e. the width of the intervals chosen for the density counts). Figure 3 shows that reducing the bin size to 25 from 50 in Fig. 2, leads to the same conclusion, namely that the data are not normally distributed. It would therefore not be prudent to use the statistical data from Figure 2 and calculate the 2.5% probability for the lowest wall thickness, 790 – 2*112 = 564 mils. Of course, one could have asked for the 1% probable thinnest thickness which would have been around 790 – 3*112 = 454 mils or thin enough to violate the safety requirement. In view of the high stakes in these considerations one may then want to explore other approaches.

VI. Extreme Value Statistics

An alternate approach is in the application of extreme value statistics, which does require the data to be normally distributed. Figure 4 shows the data from Bay 13 (external measurements made in 1992). The theory requires that when the ordered data are plotted against a double logarithmic function of the reduced or relative order of the data points (the reduced variate) in the series, a straight line is obtained. Figure 4 shows the result. The correlation would appear to be

[9)] while the nominal thickness may well have been 1.154 inches the manufacturer's tolerances vary from 10 mils to as much as 2 to 3%.

considerably better than under the assumption of normalcy (Fig. 2). The regression function of the straight line can be used to extrapolate values, which would have been obtained if more points had been measured. Thus, if 37 points had been measured, one might have observed a point of the order of 490 mils residual wall thickness. Larger number of data points might well have included even lower wall thicknesses. This is a disturbing result because it indicates that there is a significant chance that the drywell shell would not contain the gases in an accident condition.

Extreme value statistics, as I understand it, was developed in order to predict damage from extreme weather conditions. For example the 100-year flood plain is based on such predictions. These predictions only say how high water may rise if the extreme amount of rain falls. They do not predict when this may happen. Similarly, extreme value statistics applied to corrosion only says that there may be a pit deeper than all the others with a certain probability based on the number of observations, but it does not say where this pit (or damage) may be unless it has actually been measured.

VII. Extreme Value Statistics in Industry and Risk Assessment

Extreme value statistics is beginning to be used in the pipeline industry. This writer has used the approach to calculate the corrosion rate (pitting rate) in pipelines based on successive scans and evaluation of the resulting data according to extreme value statistics. Pipelines are scanned by intelligent pigs, using either mechanical calipers (rarely used any more), UT, or Magnetic Flux Leakage (MFL) technology. This technology has been growing rapidly with advances in the respective areas. Successive MLF scans for instance rarely record the same corrosion feature twice. There are a number of reasons for this too numerous to go into here. However, because of this, extreme value statistics is the only means to compare successive scans and estimate an overall corrosion rate, It must be remembered, however, that it is individual pits that corrode, not the ensemble of pits, and one initially assumes that all the pits are subject to the postulated corrosion mechanism. If different conditions prevail along the pipeline, then the data may have to be partitioned appropriately and analyzed separately.

VIII. Some General Remarks

I hope that these very brief remarks make it clear that the application of statistics (any statistics) in the oilfield is a difficult problem. However, in view of the fact that some recent incidences have led to the criminalization of negligence with respect to corrosion prevention of structures in the public sector, and with DOT and EPA setting rules, companies have begun to realize that failures are becoming less and less acceptable. However, while diligence has been legislated, there are currently no standards with respect to the certainty required. This means it is left to the individual companies to assess risk based on the probability predictions extracted from the data collected.

This gets us right back to risk assessment, and the probability of the "extreme corrosion damage", and the corrosion rate based upon such estimates, are only a small part of the overall input into the Monte Carlo simulations that calculate the value of the risk the company is bearing and in the end dictate corporate behavior.

In conclusion, it appears to us, that similarly to corporate behavior, where the responsibility lies squarely on the shoulders of the engineer and responsible personnel, the nuclear industry, and in particular the NRC, should take a much more sophisticated look at uncertainty and risk. Here the NRC Staff initially required the drywell shell to meet the deterministic design criteria for both pressure and buckling with a very high degree of certainty. However, the Staff seem to have drifted from this stance to approving the safety of the proposed relicensing, when compliance with those same criteria can no longer be established with any certainty. Although NRC Staff at one point required AmerGen to show margin with a nominal 97.5% confidence and the Staff appear to espouse that standard in the SER and in their testimony, they failed to apply it in practice.

Finally, I would like to point out that this writer at least recognizes the large investment in the nuclear industry, recognizes the complexity of the installations, and appreciates the accident free (or near accident free) operations up to this point. I also am of the opinion that safe nuclear power generation, safe spent fuel handling, and safe spent fuel storage, and reprocessing are in the future of the country. However, just like it has evolved in the pipeline industry, in the refining industry, and in offshore oil and gas production, the personnel involved in the nuclear industry must be held to the highest standards of technical and ethical judgment, which must trump corporate imperatives (see for instance the BP Alaska debacle). Relying on past performance when predicting future behavior of 40-year-old installations is not enough. Aging management has to be a lot more sophisticated and must require the industry to demonstrate it can meet safety-related requirements on an on-going basis with a high degree of certainty.

Table 1

Comparison of the Data Spread in the External UT Measurements with the Standard Deviation of Duplicate or Triplicate Measurements at the same Spot

Bay	1992 average			2006 Average			
	No of Data Points	Measure ments	Std. Dev. as Variability	No of Data Points	Measure ments	Std. Dev. as Variability	Std Dev. from 2006 Repeat Measurements
5	8	0.994	0.053	8	0.96	0.0386	0.017
7	7	1.004	0.043	7	1.007	0.027	0.017
15	11	0.816	0.054	11	0.81	0.053	0.023
19	10	0.889	0.08	10	0.848	0.083	0.029

Figure 1

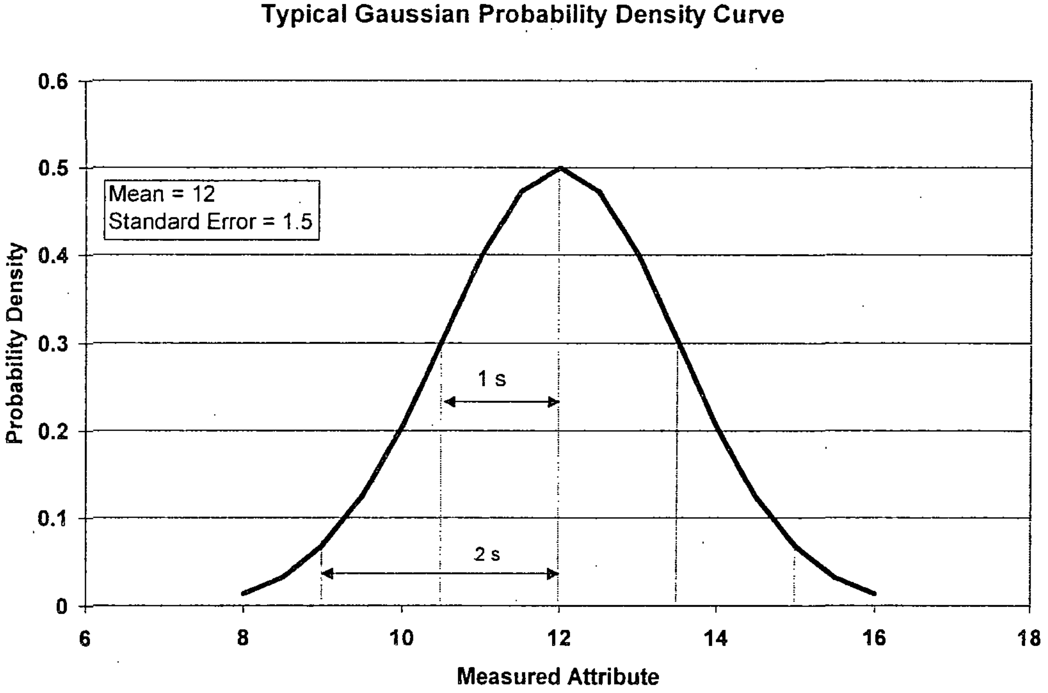
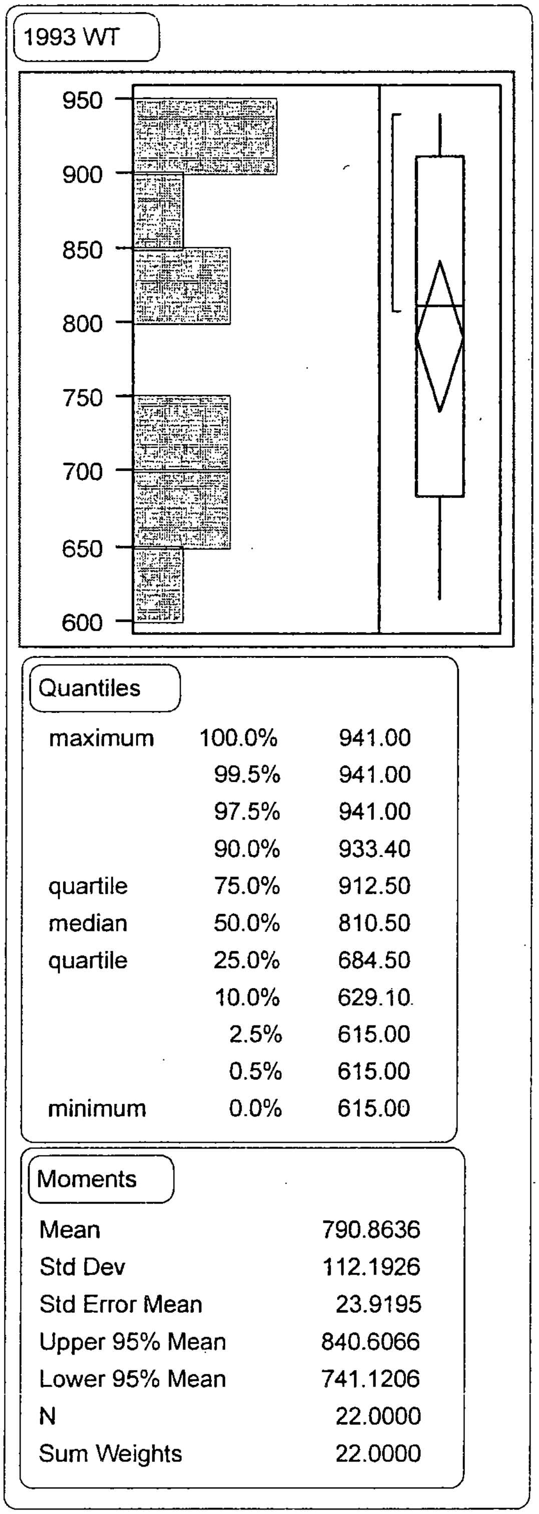


Figure 2: histogram for external wall thickness measurements in Bay 13
(Generated in JMP)



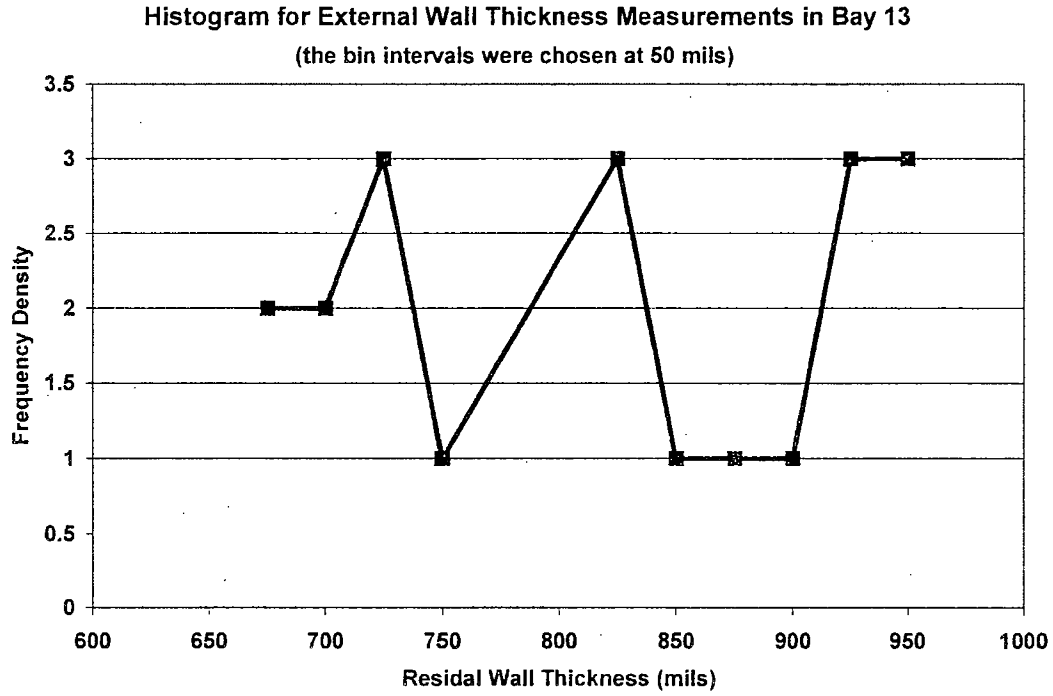
Quantiles

maximum	100.0%	941.00
	99.5%	941.00
	97.5%	941.00
	90.0%	933.40
quartile	75.0%	912.50
median	50.0%	810.50
quartile	25.0%	684.50
	10.0%	629.10
	2.5%	615.00
	0.5%	615.00
minimum	0.0%	615.00

Moments

Mean	790.8636
Std Dev	112.1926
Std Error Mean	23.9195
Upper 95% Mean	840.6066
Lower 95% Mean	741.1206
N	22.0000
Sum Weights	22.0000

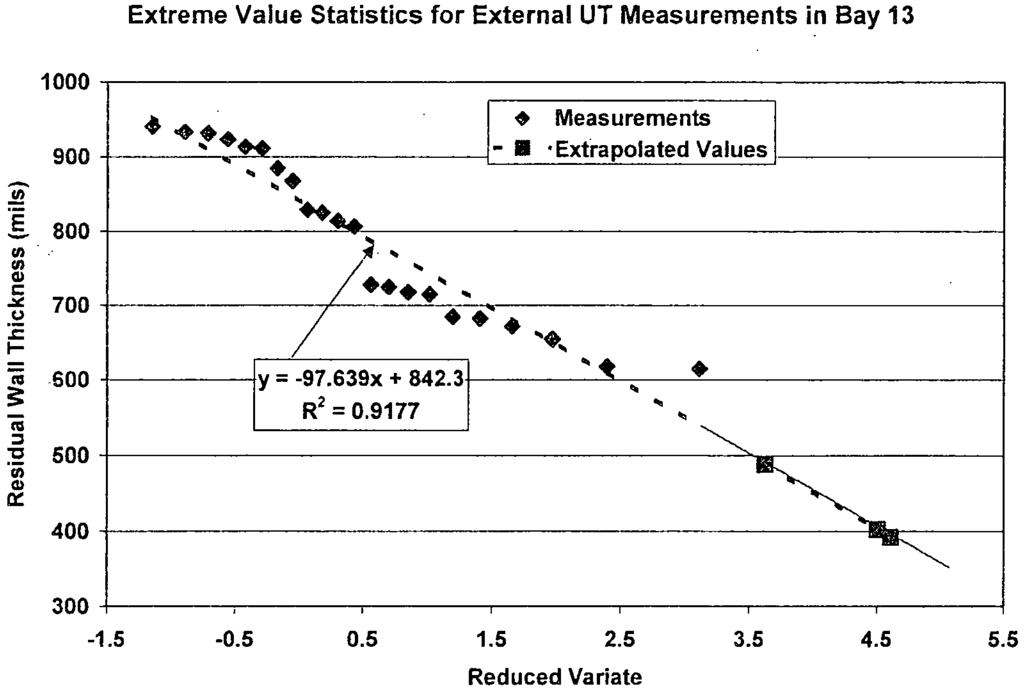
Figure 3



Histogram for External Wall Thickness Measurements in Bay 13

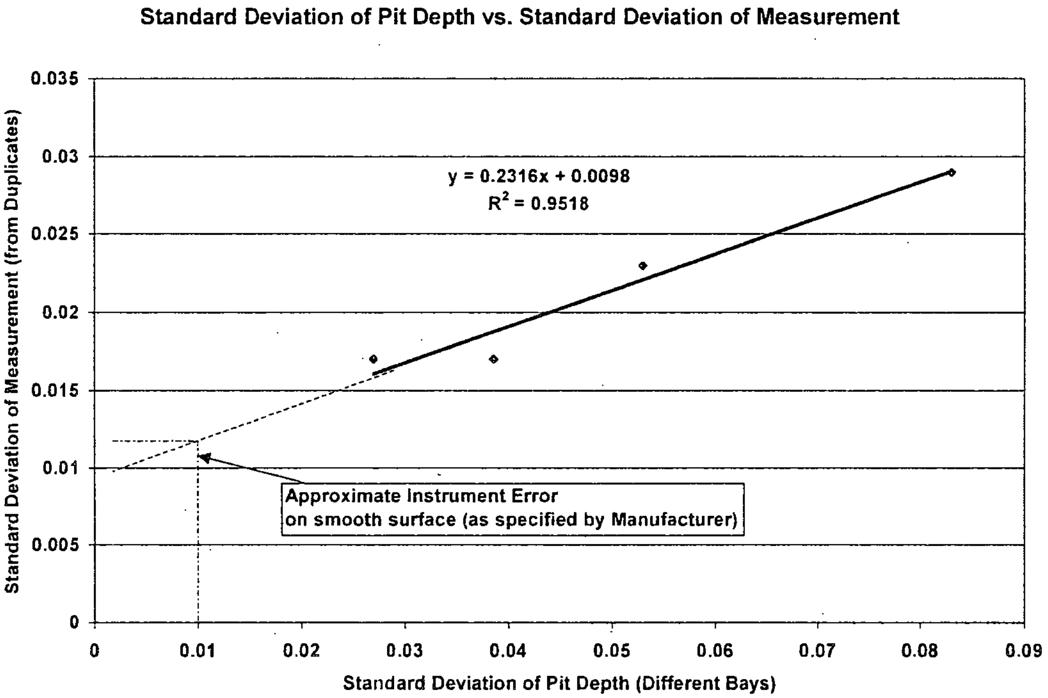
(the bin intervals were chosen at 50 mils)

Figure 4



Extreme Value Statistics for External UT Measurements in Bay 13

Figure 5


Standard Deviation of Pit Depth vs. Standard Deviation of Measurement
y = 0.2316x + 0.0098
R^2 = 0.9518
Approximate Instrument Error
on smooth surface (as specified by Manufacturer)
Standard Deviation of Measurement (from Duplicates)
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
Standard Deviation of Pit Depth (Different Bays)