Notice

Profound Statistical Concepts: When Theory Collides with Reality

Beverly Daniels

The late George Box is famous for reminding us that "all models are wrong; some are useful."[1] The most common statistical models that are taught to most Quality practitioners are too frequently limited to theoretical distributional models and tests of statistical significance that are focused on the Normal distribution and tests of means. Certainly other distributional models and other tests of significance are given some attention. These include the Binomial and Poisson distributions and tests of significance such as the homogeneity of variance and non-parametric tests. Most statistical training readily accessible to Quality practitioners is naturally of a limited nature – there never seems to be enough time to cover all of the important aspects of statistical analysis that we will need in our careers. While some of the more practical approaches that are designed to handle non-normal and non-homogenous distributions (*e.g.* multi-vari*,* blocking and split plots, *etc*.) are taught and are available for study in the literature, their presence is too often overwhelmed by a perceived lack time to teach and a lack of time to learn. Of course, there are many statisticians, instructors, and practitioners who avoid this. ASQ was founded by many of these distinguished individuals. However, as Six Sigma has grown, this knowledge and training has become diffused by a large number of trainers who primarily teach the common ideal models.[2]

The emergence of powerful user-friendly statistical software – while a welcome addition to our arsenal of tools – hasn't been able to break this time constraint either. In fact, the ability to perform a plethora of statistical tests at the literal "push of a button" has had a further unintended consequence of a narrowing focus on some of the more popular "ideal" conditions and tests. It's not that these software packages can't perform less common analyses intended for the messy real world reality, it's that the tests based on the more commonly taught theoretical ideals are so accessible that little thought is now required to push the button. Because of the perceived time constraint, too many people get trained primarily on how to use the software and not on how to perform practical statistical analyses; they don't know how to assess the process variation to design the correct experimental test structures. Improving our knowledge of the limitations of the common models and the more useful alternatives is the responsibility of both instructors and students.

Several times a year I hear from practitioners the following statements:

- I have some data; what statistical tests can I perform?
- My distribution isn't Normal. Now what do I do?
- My p value was less than .05; why didn't my fix work?

Certainly many of you also hear finger nails on the black board every time you hear these and similar statements. But there is a contingent of instructors who limit their time to the common models and an even larger contingent of Quality practitioners who are trained in or only remember these common models.

This gap is further compounded by the almost total lack of training in the difference between an enumerative study and an analytic study as defined by Deming.[3]

The commonly taught distributional models and statistical tests of statistical significance or confidence intervals are usually taught in the ideal context of identically distributed and homogenous distributions. When confronted with what Deming referred to as an Enumerative study where we are trying to characterize a population through random sampling, these statistical approaches work very well. However, the quality practitioner is more often confronted with problems that need to be solved or prevented. These situations require what Deming referred to as an Analytical study.[4] Analytical studies are focused on understanding causal mechanisms so that we can "make predictions about the future," *i.e.* solve and prevent problems. These are not simple or easy studies to design or analyze. They require a deep understanding of both statistical variation and science. When we limit ourselves to the common statistical theories of distributional models and tests of statistical significance on Analytical studies we too often get the wrong answer.[5]

In this paper I discuss a more effective alternative for solving and preventing most real world Quality problems.

"In theory, there is no difference between theory and practice.
But, in practice, there is."
Jan L.A. van de Snepscheut[6]

This is a discussion of how our common "ideal" models are getting in the way of improvements and what we can do about it. It is meant to be thought provoking. It requires thought to achieve a deep understanding of our profound statistical concepts and to be able to overcome some of our preconceived notions of how statistics can be applied to dealing with real world problems.

Some of the common models that trip us up:

- The Normal Distribution Assumption
- Homogenous Distribution Assumption
- Independent & Identically Distributed Distributions

These models form the basis of the tests of statistical significance testing that are so commonly taught.

So what happens when our reality doesn't match the common model?
We may assume there is something wrong with our process, we manipulate the data somehow to match the common theory, or we feel paralyzed. None of this is necessary and worse it can lead to false conclusions and ineffective solutions.

The Normal Distribution

"Normality is a myth; there never was,
and never will be, a normal distribution"[7]
Robert Charles Geary

The Normal Distribution is a man-made construct; it is not a law of physics. Many processes naturally produce symmetrical "bell" shaped distributions of data, but many naturally do not. Some examples are coating thickness, fill volumes, defect rates, flatness, true position, and processes that involve tool wear or some other component that is consumed. Where is it credibly proven that processes are supposed to be Normally distributed?

The Normal Reaction

What happens when a data set fails a test for Normality? Some Quality practitioners will try to search for outliers and throw them out of the data set; some will try to transform the data; some will be stumped because they might have heard the fallacy that "there is nothing you can do if your process isn't Normal." For something that doesn't actually occur very often and really isn't needed for many tests of statistical significance, we spend an awful amount of time testing for it and agonizing over its absence.

Let's first take a brief look at some of the negative consequences of placing so much reliance on the Normality of our data.

Censoring "outliers": Outlier detection was intended for static data sets (enumerative studies) not for data streams (analytical studies). Most outlier tests rely on the Normal distribution as the distributional model and also use a 95% coverage limit. If the data sets are large enough, some data that fits the distribution but lies beyond the 95% limits will be flagged as an outlier. These data points are not outliers in the sense that they don't belong; they are simply extreme values. Beyond that issue, we must think about what an outlier test actually tells us. A data point that is flagged as an outlier to any theoretical distribution doesn't mean that the data point is invalid.[8] It doesn't mean that the customer won't get that value. When engaged in problem solving the outlier data might be our best opportunity for solving the problem.[9] When looking to censor data from any data set the only values we should remove are those that are validated as being:

- Impossible results
- Mis-recorded or typos
- Results from an invalid measurement event
- Misread measurements

Transforming the data: Transformations can hide the data and divert us from understanding the data to trying to make our data set fit some theoretical model so we can apply tests of statistical significance. While transformations can work to allow us to use certain tests of statistical significance, they require special skills to perform correctly. Moreover, if the process is not homogenous the answer will be incorrect; no transformational formula will save you from non-homogeneity. In today's world many processes are not homogenous, especially those that are creating defects. Since many problem solving efforts do not require tests of statistical significance[10], attempting to transform our data may be a waste of time.

Which comes first, the data or the model? An incorrect choice of statistical model prior to understanding the data can lead to errors in data collection, analysis, and interpretation. It can lock you into a theoretical model that doesn't approximately describe your data and it can hide the truth hidden in the data. If your data doesn't fit a model, it's the selection of a model that is incorrect, it's not the data (or the process) that are 'wrong'. For some problems, distributional model fitting is at best wasted effort for the Quality professional. At worst it could result in incorrect conclusions and generate distrust in the Quality sciences.

So what can we do? The late Ellis Ott was continually exhorting his students to understand the technical background of the problem, collect some data, draw some plots of the data, and think about what you have learned and what action you should take. Dr. Ott wanted to teach his students statistical thinking: plot your data and think about your data.[11]

Let's look at an example of a problem that involved seriously non-Normal data but was able to be solved with statistical thinking rather than forced theory .

Example 1, Noisy Gears (Non-Normal Data with unequal variances)

Two gears engage at high rpms. Frequently each of these gears are rotating at different speeds (revolutions per minute: rpm). During the prototype testing a loud noise occurs when the gears engage. Observation shows that the gears are not meshing smoothly and the gears are grinding against each other. There is no obvious cause for this delayed engagement.

The team's first step was to establish a measurement system for the delayed engagement or noise. The first decision involved the conditions of the test: at which delta rpm would the test be run? There was no quick way to assess what range of conditions could be experienced in the field and so the engineers determined the likely range of deltas and a Measurement System Analysis (MSA) would test gears across the range to determine the delta that would provide the best information. There were two choices for the measurement of the noise itself: duration and amplitude. Amplitude is measured in decibels and is a logarithmic function. Amplitude only measures one aspect of the noise – its loudest point and is an indirect measure of the force of the collision. Duration is a measure of how long it takes the gears to mesh, which is a function of the physical features that are prolonging meshing as well as the delta rpm at the

moment of engagement of the gears. Duration is often a non-normal distribution and is bounded at 0.

The measurement system analysis had a different structure than a typical one since this is a measurement of a function. Five gear sets were selected for the study and were randomly engaged twice at five different delta rpms. Both duration and amplitude were measured for each event (Figure 1).

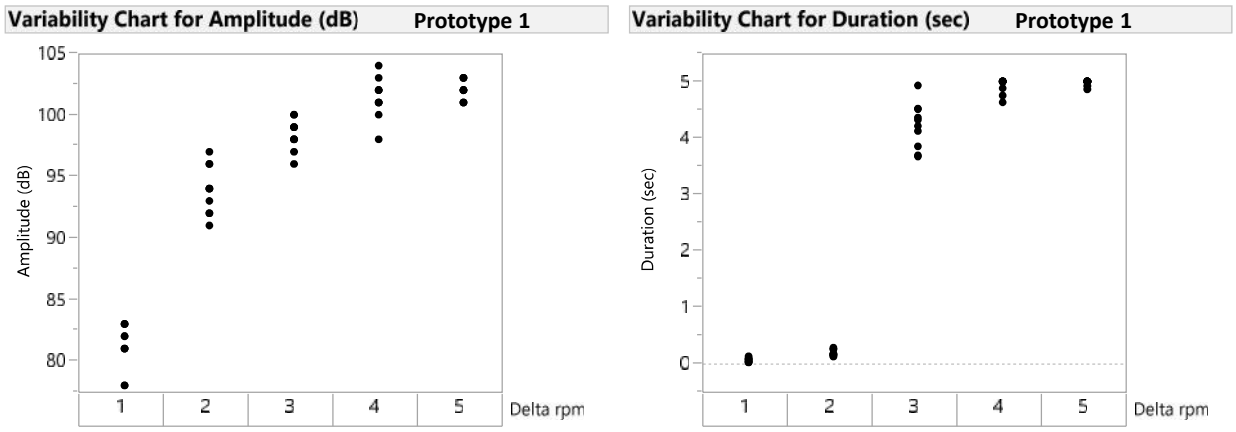

Figure 1: Response of noise with increasing difference in delta rpm between gears

There is more to the selection of the appropriate measure of the problem than ease of collection, analysis or passing an MSA; the measurement must be a true representation of the failure. In this case duration provides a better measure of what is actually occurring with the gears – it matches the physics of what is happening. While the customer's initial perception of the quality of the vehicle will be effected by the noise, it is the potential for physical damage to the gears by prolonged grinding that poses the largest threat. Duration of the noise is also a measure of how long it takes the gears to mesh which is a function of the delta rpm and the features that are inhibiting meshing. Additionally, the data indicate that the larger the delta between the two gear rpms, the longer (and louder) the noise is. When the MSA data are plotted on a Youden plot we see a clearer picture of the noise and gain even further insight into the causal mechanism (Figure 2). This clarity comes from graphing the actual data in context of the study design rather than parameter estimates from a common theoretical statistical analysis.

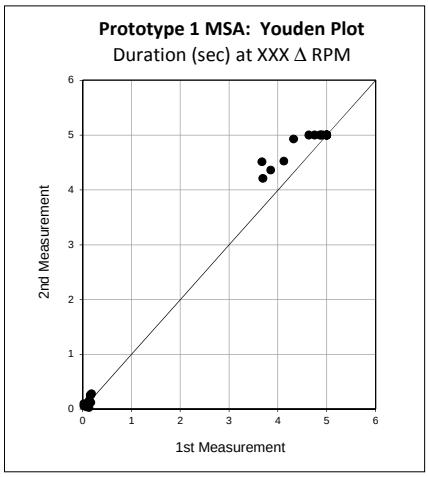

Figure 2: Youden Plot of MSA results

A second view of the data that combines both the delta rpm and the resulting duration shows the effect of test condition and the measurement repeatability in the same chart (Figure 3).

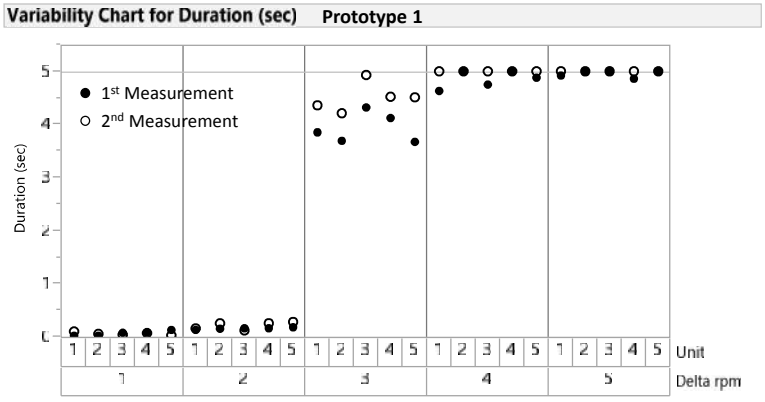

Figure 3: Multi-Vari Chart of Noise MSA showing both the delta rpm and the repeated measures of noise duration

This plot allows us to dissect the variation seen in the previous chart. A substantial hint as to what is happening is seen when we look at the repeated measurements of each unit at each delta rpm across the range of delta rpm settings. It is important to note that the 5 units were completely randomized for the 1st and 2nd run across delta rpm settings. (*i.e.* the lower delta rpm settings were *no*t run sequentially before the higher delta rpm settings.) The second engagement noise duration is longer than the first at high delta rpms. Why? This is the important question. This insight comes from graphing the actual data in context of the study design. In this case the results lead us to test under the worst case condition of the highest delta rpm. This will provide much better detection of the causal mechanism.

Occasionally, the MSA will also provide insight into the causal mechanism as with this study. Physical observation of the parts reveals that the parts are being damaged. The data indicates that at high delta rpm settings it takes longer to fully engage due to the physical damage. By looking at the damaged areas the features that could create the noise were apparent. A Design of Experiments (DoE) was planned with these 3 features of interest.

In the interim, a second prototype had been built independently of this investigation. It was tested and the noise duration was substantially less at the higher delta rpm settings (Figure 4). The 3 factors that were suspected based on the damage were all changed in the second prototype design, confirming that the original selection of factors was appropriate.

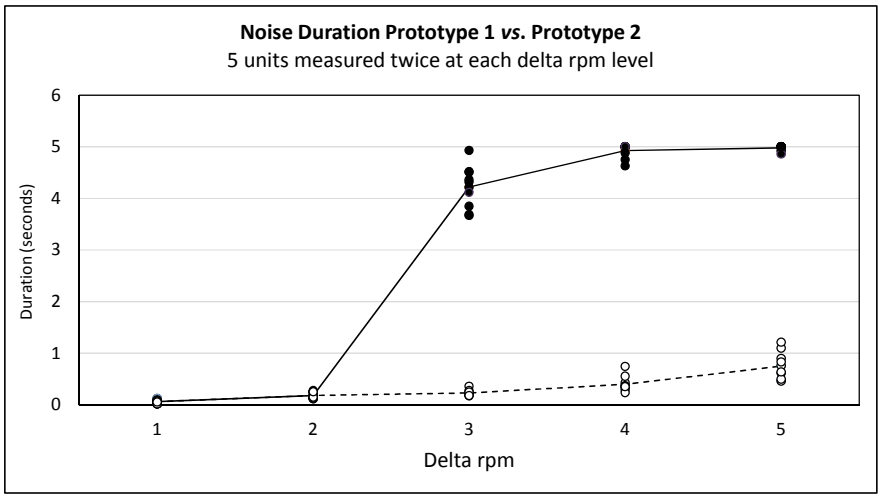

Figure 4: Performance of Prototype 1 (solid line) vs. Prototype 2 (dashed line)

The study design was a 2^3 full factorial (3 factors at 2 levels). The sample size was 2 gears per condition. Why 2? There was no sample size equation used; not that these aren't useful at times it's just that in this case the difference in performance between prototype 1 and 2 was so patently obvious at a sample size of 2 gears. Testing would occur at only the highest delta rpm level and each gear would be tested twice in a random test order. The test would be truncated at 5 seconds to prevent catastrophic damage to the gears (Table 1).

Experimental Design for Noisy Gears		
Design Element	Purpose	Consequence
3 factors, 2 levels Prototype 1 (+) and Prototype 2 (-)	The +/- level assignment was based on the noise duration for each level. The choice of the two 'known' levels was to ensure that the results could be replicated and if replicated, the critical factor would then be identified.	Standard 2^3 full factorial design.
Sample Size: 2 gears per condition	This gives us 6 independent units for each individual factor and level condition	Provides statistically significant sample size
Testing occurred at the highest delta rpm setting	The worst case condition for failure provides the largest contrast between levels and provides our best opportunity to see true differences	The variances for the two levels are not equal at the worst case delta rpm
Each unit is run twice for each condition	This provides a measure of the damage that is a result of the energy expended.	The two repeated measures will create results that are not independent
The testing is stopped at 5 seconds	This minimizes damage to the test parts	Creates a truncated data set

Table 1: Experimental Design Structure for Noisy Gears

This design will produce data that are non-normal, non-symmetrical and bounded by zero, the two repeated measurements of each part are not independent (although there is some independence between the two parts in each condition and replication across the conditions), the testing is truncated at 5 seconds and the variances of the two levels are not equal.

Results: While we could try to determine the correct statistical analysis, we could also try to plot the data and look at it first (Figure 5).

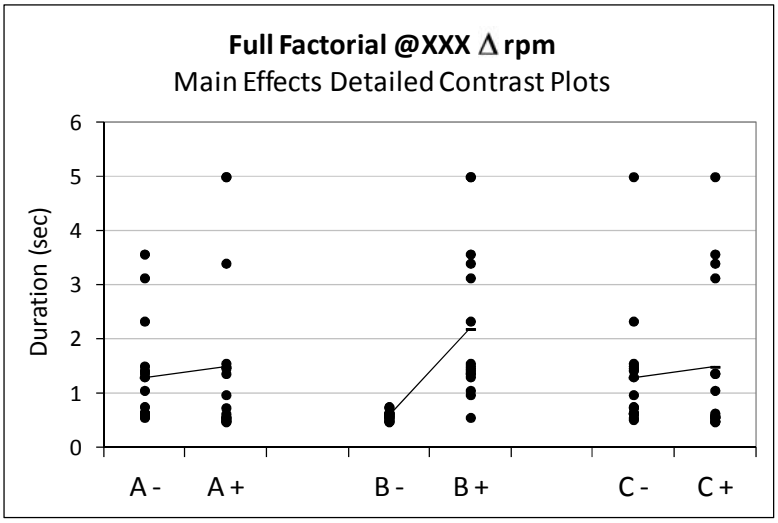

Figure 5: The main effects plot for noise duration

This plot says it all. Factor B is the controlling factor for noise. As Deming pointed out many times, there is no statistical test of significance that will increase our belief in this result or change the action that we take next. With this knowledge the team was able to alter Factor B such that the noise and the time to mesh the gears was minimal and within the specified range.

In Analytical studies, experimental structure and replication provide confidence. Sophisticated statistical tests can help us tease out small effects but solid experimental structures that account for process variation may tease out small differences with much smaller sample sizes. Further, in my experience, industrial quality practitioners are usually looking for large differences.

Homogenous Process Streams

Most tests of statistical significance rely less on a distributional assumption and more on the homogeneity of the process. If the process under study is non-homogenous, statistical tests of significance will correctly detect the non-homogeneity but will not necessarily gives us a practically useful answer. A homogenous process will produce parts or results that will appear to vary randomly without trends, shifts or cycles. A homogenous process is one in which the factors that primarily determine the location of the data are the same factors that primarily determine the piece to piece variation. The largest component of variation is part to part or event to event and the variation of the sample means will be a function of the sample size and the population standard deviation ($\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$).

Not all homogenous processes are Normal and not all Normally distributed processes are homogenous.

For enumerative studies we create homogeneity by random sampling. Enumerative studies are concerned with characterizing a static data set and the action taken will be on that data set. A prime example of this is acceptance sampling. An enumerative study is not intended to make predictions about the future performance of a process. The majority of the work that a quality professional engages in is to determine the causes of poor performance and take action to improve the future performance. This is the purpose of an analytical study. Since many process streams are not homogenous, the first step in an analytical study is to understand the non-homogeneity so we can design studies that won't result in a misleading conclusion from common or misapplied statistical analysis.[12]

Example 2 Stacking Faults (Non-Homogenous categorical data)

Stacking faults are a particular type of defect in a silicon wafer used to manufacture semiconductor devices. An additional silicon layer is grown on a silicon wafer's surface to produce a very clean substrate for subsequent fabrication steps. The additional layer is an epitaxial growth. If there are contaminants or structural imperfections at the surface of the wafer prior to the growth of the epitaxial layer the contaminants can result in a larger structural defect that propagates through the epitaxial layer. These defects are collectively referred to as stacking faults. If the fault is in the wrong location, the device may fail at subsequent testing steps or may pose a reliability risk. The stacking fault inspection involves microscopically checking the surface of the wafer at 5 locations on each of 3 wafers. The maximum average allowable count is 10 faults. Seven of the last 25 lots were rejected and scrapped for exceeding this limit.

An engineer was assigned to test a new cleaning process to reduce the contaminants on the wafer surface. The first test was run on one production lot using the proposed process. A randomly selected lot that was cleaned with the current process (produced during the same time as the experimental lot) was also tested as a control. All of the wafers in both lots were

tested for stacking faults. The original analysis involved the commonly applied t-test on the number of observed stacking faults on each wafer from both lots.

Although the p value was less than 0.05, no one really believed the conclusion that the new process was better than the current process as the results for the two lots were so similar (Figure 6). Before abandoning the new cleaning method, the Quality Engineer (QE) took a stab at the analysis.

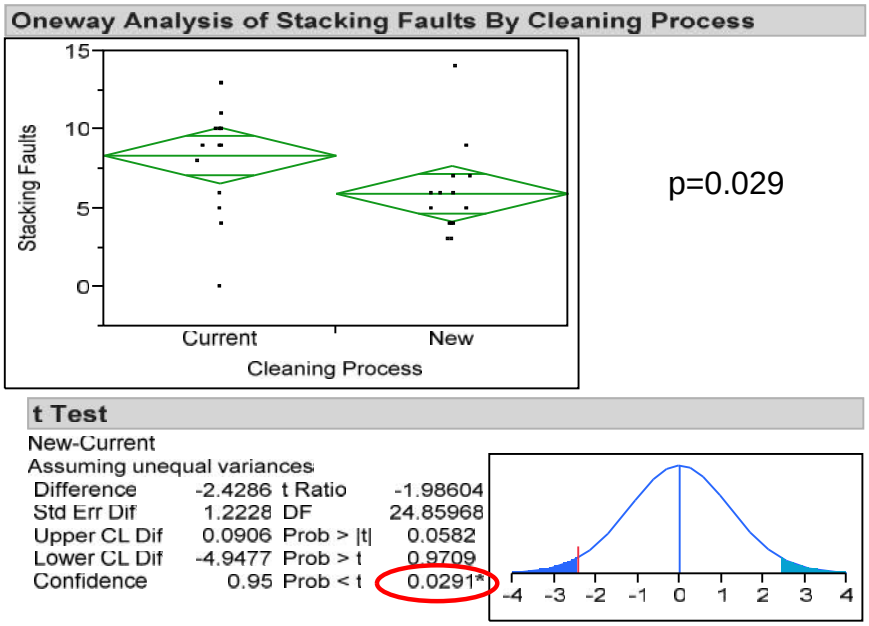

Figure 6: t-test with confidence intervals for the new cleaning process vs. the current process

The QE recognized that the data were not continuous, but categorical counts. The better theoretical choice for a data model was the Poisson distribution and the QE redid the analysis using Poisson confidence intervals (Figure 7).

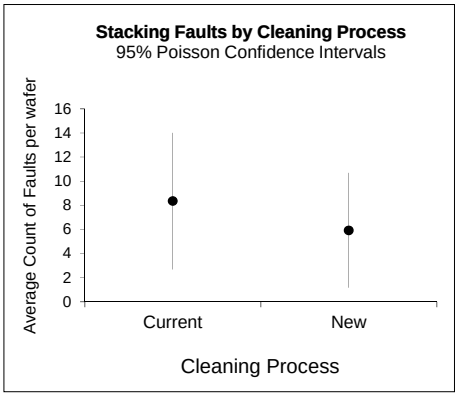

Figure 7: Poisson confidence intervals for the new cleaning process vs. the current process

This analysis resulted in a conclusion that the two processes were not statistically significantly different. Now we have two conflicting conclusions – which one is correct? In this case, it turned out that wasn't the relative question. It's not about the model or the statistical analysis; it's about the structure of the experiment in relation to the variation of the process.

A quick control chart of the stacking fault counts for the 25 previous lots reveals that the process is not in statistical control; it is non-homogenous (Figure 8).

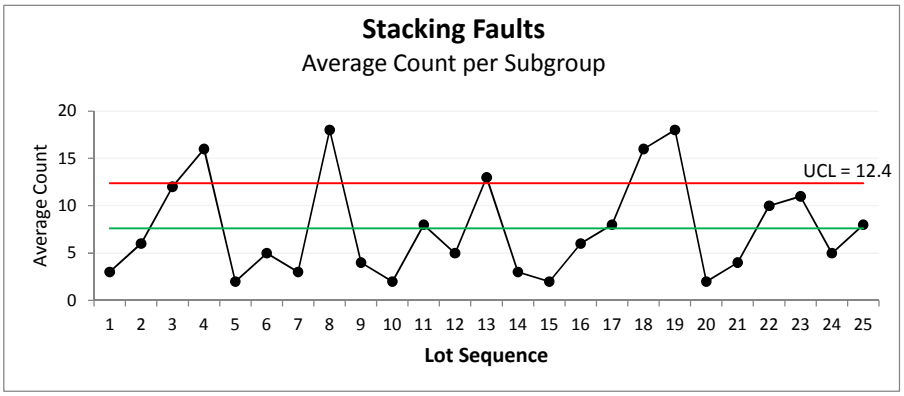

Figure 8: Control chart for stacking fault inspection. Note: each data point is not a single count. It is the average count of 3 wafers. The chart shown is a u chart with a subgroup size of 3.

The current process did not even have a Poisson distribution because the probability of a fault is not constant from wafer lot to lot. A t-test or One way ANOVA aren't appropriate because the data are counts not continuous data and the Poisson confidence intervals and any transformation won't be valid because the process isn't homogenous.

Why is the process non-homogenous? Lets' consider the science for a moment. Silicon wafers are cut from ingots of silicon that are grown from a silicon seed. Each silicon seed is different and the growth conditions will vary slightly from ingot to ingot. Wafers are then sliced from the ingot to form a wafer lot. The cutting tool used to slice each ingot is subject to wear and will not cut as cleanly as it wears. Theoretically, the science tells us that imperfections in the silicon structure will be relatively homogenous within an ingot and different from ingot to ingot. The cutting process will also introduce contaminants and surface imperfections that are relatively homogenous within a wafer lot but different from wafer lot to lot. The number of stacking faults will be partially dependent on the initial contamination of the wafers and the structural imperfection of the ingot. Although it would have been better to fix the causes of the variation and substantially reduce the stacking fault rates, this performance was considered state of the art at the time and the ingot supplier was not inclined to take this effort on. So the semiconductor manufacturer was stuck with improving their cleaning process.

A requirement of a t-test or any other test of statistical significance is that the parts within each level are independent of each other. Although there were 8 wafers used for the new process

and the current process, the results were not independent due to the homogeneity within the wafer lots. So regardless of the statistical test, the basic requirement of independence was violated. Additionally, it is possible that the first experiment chose a wafer lot that was highly contaminated compared to the control lot making the initial comparison invalid.

A better experiment would be one that accounts for the actual physics of the process. The new design is a "matched pair" block design[13] described in Table 2.

Experimental Structure for Stacking Fault Cleaning Method Comparison	
Design Element	Purpose
5 different wafer lots each from a unique ingot	This provides replication across different ingots that will have different contamination and imperfection loads
Select 3 wafer pairs from each lot The 3 pairs are taken from different locations within the wafer lot (ingot location)	This provides repeated measures within a wafer lot and allows the assessment of the theory of homogeneity within a wafer lot
Each pair contains two wafers that were adjacent in the ingot	These wafers will be as alike as possible in their starting condition in terms of structural imperfections from the ingot growth, imperfections induced from the cutting operation and contamination left by the cutting operation
A wafer from each pair is randomly selected for the new cleaning method or the current method; all other processing of the pairs within a wafer lot is the same	The random selection within a pair randomizes any temporal differences in the ingot and cutting operation
The 3 pairs from each wafer lot are cleaned together. Each wafer lot is cleaned in separate setups and the order of cleaning is randomized	This provides replication across different cleaning conditions for each lot while providing repeatability within a lot. (Standard process is to clean all wafers from a single lot together)
Stacking fault counts are taken at the standard 5 locations on each wafer	This keeps the data in the same scale as the inspection so results can be easily translated to the expected improvement in performance

Table 2: Experimental Design Structure for Stacking Fault Cleaning Method Comparison

The results of the experiment are plotted in Figure 9.

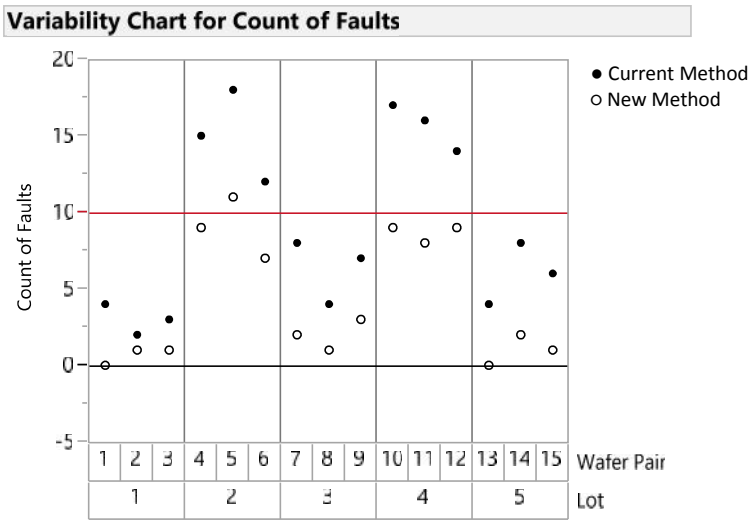

Figure 9: Results of the "matched pair" study showing the effect of the new cleaning method in direct comparison to the current method

The best statistical test of significance may not be obvious, but here the graphical display of the appropriately structured experiment makes the statistical mathematics redundant. Within each pair the new cleaning method results in fewer stacking faults than with the current method. This improvement is replicated with all 3 pairs within an ingot and across multiple ingots with different silicon seeds, growth conditions, sawing conditions and cleaning events. The probability of this happening simply by chance is vanishingly small.

An alternative graphical approach is to plot the difference between the two methods (Figure 10). This plot clearly shows that the amount of improvement using the new method is dependent on the underlying (uncleaned) fault rate. There is more improvement in raw counts when a wafer has more contamination and damage. The improvement in counts is naturally less when a wafer has few faults.

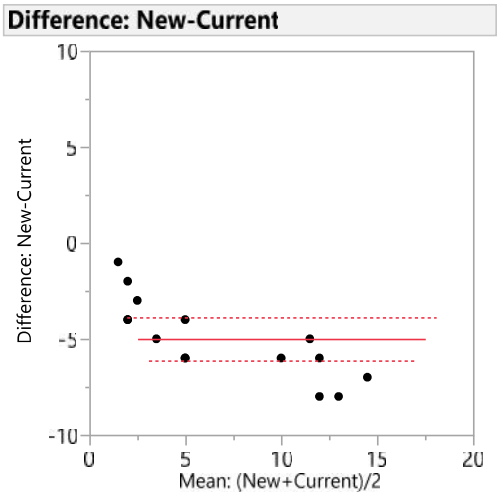

Figure 10: Method comparison plot of difference between the new and current cleaning method

Are there statistical tests of significance that would fit this situation? Of course there are, but look at the charts again – do we need them? What additional value would one of these tests provide? Would a precise p value change the decision? The confidence in our conclusion that the new method is better than the old method comes from replication and the structure of the study, not from a p value or confidence interval.

"…levels of significance furnish no measure of belief in a prediction. Probability has use; tests of statistical significance do not." W. Edwards Deming[14]

The important decision is whether or not the new cleaning method is worth implementing. Taking a closer look at the improvement (Table 3), we can see that the two lots that would have been rejected with the current cleaning method were passing with the new method.

Data Table for Matched Pair Comparison of Current Cleaning Method *vs.* New Method					
Lot	Pair	Current	New	Lot Average	
				Current	New
1	1	4	0	3	0.67
	2	2	1		
	3	3	1		
2	4	15	9	15.33	9
	5	18	11		
	6	13	7		
3	7	8	2	6.33	2
	8	4	1		
	9	7	3		
4	10	17	9	15.67	8.67
	11	16	8		
	12	14	9		
5	13	4	0	6	1
	14	8	2		
	15	6	1		

Table 3: Experimental results of new vs. current cleaning method

How can we use this data to predict the improvement we might expect to see in the future?

There are several methods available to us. We can use the method comparison chart and a simple calculation to make a rough estimate of the improvement we expect to see when lots would otherwise have counts above 10. Looking at the method comparison chart (figure 10) shows roughly two groups of data (below 10 and above 10) with two different average levels of improvement (based on the current method results). Our historical data shows that recent failing lots have average counts between 11 and 20. The average improvement for the two lots that had counts above 10 using the current method is 6.7. The average improvement for the 3 lots with counts less than 11 is 4.3. This provides a rough estimate of the improvement. Using this information, we can predict the level of improvement we might have had with our historical data (Figure 11).

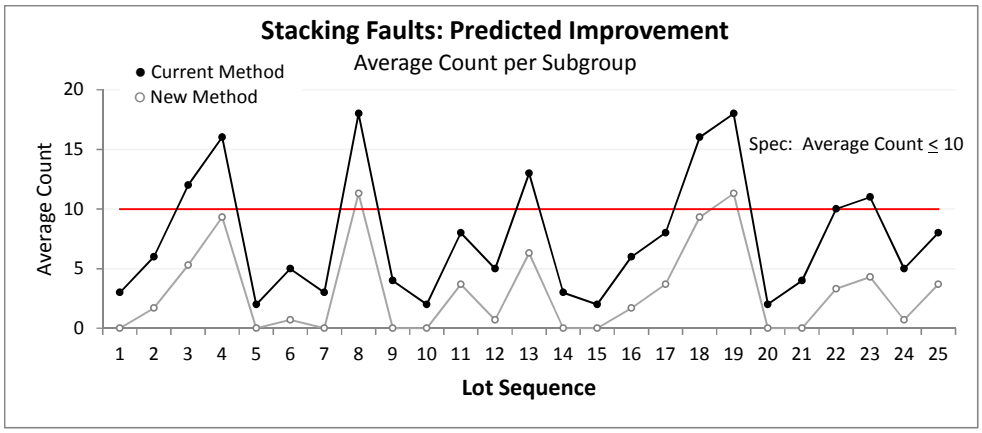

Figure 11: Average amount of improvement for historical data

Five of the seven failing lots would have had acceptable stacking fault counts with the new method. Additionally, the overall stacking fault rate improves for every lot, so that the die yield loss and field reliability failures would also be expected to improve.

Can we provide a more precise estimate? Yes, but it is still an estimate no matter how precise we make it. Is there any added value to a more precise estimate? We have a fairly small data set; so more sophisticated approaches may not provide a better estimate. The nature of the data set again provides some serious challenges to a choice of statistical modelling: There are only 5 lots and the 3 values within a lot are not independent. While the current method results correlate to the new method results they do not have the traditional 'dependent' relationship of the common linear regression model. Certainly there are techniques to deal with this, but will they provide more confidence in the answer? All estimates at this point will be wrong; additional statistical manipulation will only increase the precision of the estimate; it cannot increase the accuracy. An example of trying to be more precise with the data we have is to calculate the best fit line formula with the Current method results as the independent variable (X) and the New method results as the dependent variable (Y). In this example we use all of the individual wafer counts with the paired wafers serving as the X:Y values (Figure 12). The resulting equation of the line is used to predict the improvement using the historical baseline (figure 13). As can be seen the change in the predicted improvement is trivial.

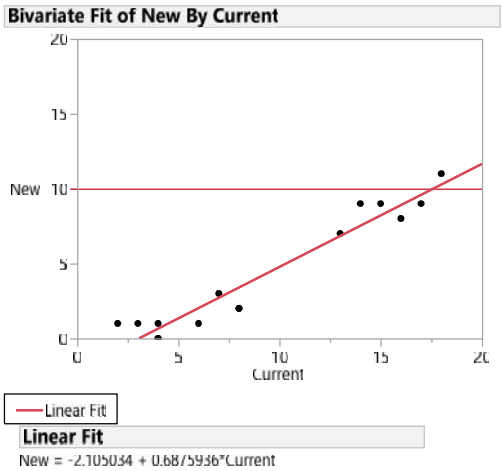

Figure 12: Bivariate fit of the new cleaning method vs. the current cleaning method from the matched pairs experiment

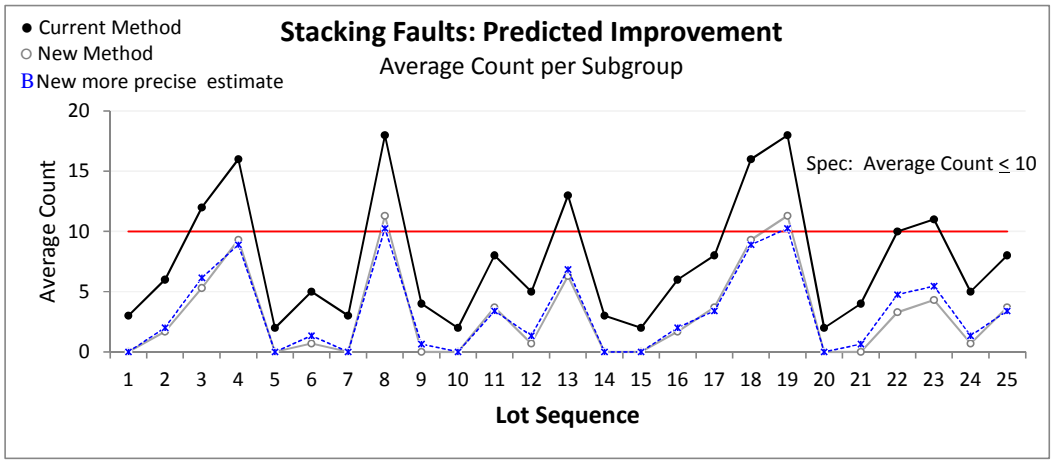

Figure 13: Expected improvement of the historical baseline using different estimates of the expected improvement

Certainly there will be improvement but it is not complete; there will still be stacking faults. Is it enough? We could run more experiments to gather more data to increase our confidence in the estimate. But will it be worth it? The answer is in the science not the statistics. If the new cleaning method is relatively inexpensive, easy to implement and poses no risk of adverse consequences, any improvement will be welcome. If the method is expensive and time consuming to implement and/or poses a high risk of adverse consequences, we will certainly want to increase our confidence in the amount of improvement we can expect to see. This may require us to better understand the cost of lot rejections, the effect on individual die yield and device reliability before we can make a better informed decision. This would require a second more complex experiment and/or analysis of existing data. This particular cleaning method was

inexpensive, easy to implement and had no adverse consequences. The method was implemented without additional analyses or experiments and the expected improvement was achieved.

As far as the laws of mathematics refer to reality, they are not certain,
and as far as they are certain, they do not refer to reality.

Albert Einstein[15]

As quality practitioners it is well worth our time to learn about analytical studies and the study designs and tools that help us perform more effective studies. We need to learn how to ask appropriate questions regarding the variation of our processes and how to better analyze the results of our experiments. This requires us to broaden our knowledge beyond the common statistical models and to apply more statistical thinking. Remember, statistics without physics is gambling; science without statistical structure is psychics.

---

[1] Box, George E. P.; Draper, Norman R., Empirical Model-Building and Response Surfaces, 1987 p. 424, Wiley

[2] Schwinn, David, "*Teaching Statistics that Help not Hinder Management*", Quality Digest, September 2012

[3] Deming, W. Edwards, *"On Probability as a Basis for Action"*, The American Statistician, 1975, Vol. 29, No. 4, pp146-152

[4] *Ibid*

[5] Stauffer, Rip, "*Render unto Enumerative Studies*…", Quality Digest, July 2013

[6] Rosenberg, Doug, Stephens, Matthew "*Use Case Driven Object Modeling with UML Theory and Practice*" 2007 Apress, p. xxvii

[7] Geary, R. C., "*Testing for Normality*", Biometrica, Vol. 34, pp. 209-242

[8] Wheeler, Donald, "*Probability Models do not Generate Your Data*", Quality Digest, March, 2009

[9] Wheeler, Donald, "*All Outliers are Evidence*", Quality Digest, May, 2009

[10] Deming, W. Edwards, *"On Probability as a Basis for Action"*, The American Statistician, 1975, Vol. 29, No. 4, pp146-152

[11] Neaubauer, Dean V., "*Pl-Ott the Data! A retrospective look at the contributions of master statistician Ellis R. Ott*", Quality Digest, May 2007

[12] Daniels, Beverly, "*Overcoming Doubt and Disbelief*", Six Sigma Forum Magazine, November, 2012

[13] Moen, Ronald D., Nolan, Thomas, W., Provost, Lloyd P., "Quality Improvement through Planned Experimentation" 2nd Edition, McGraw-Hill, 1999, chapter 4

[14] Deming, E. Edwards, Forward to Statistical Method from the Viewpoint of Quality Control by Walter A. Shewhart, 1986, Dover reprint

[15] Albert Einstein's Address on "Geometry and Experience" at the Prussian Academy of Sciences in Berlin on January 27, 1921