Comparison of the Neyer D-Optimal and 3pod Sensitivity Test Designs

Barry T. Neyer
Excelitas Technologies Corp, 1100 Vanguard Blvd, Miamisburg, OH 45342-0312

Abstract

The performances of the recently proposed 3pod sensitivity test design of Wu and Tian (2014), along with variations of the 3pod design, were compared to the original Neyer D-Optimality design (1994) used in the SenTest software through multiple simulated runs of different sizes, combinations of initial test parameters, and variations of test designs. Three performance metrics were used to evaluate the first two phases of the designs: percentage of tests without a zone of mixed responses ("wasted runs"), and the Mean Squared (truncated) Errors of both the mean and standard deviation. There was no overall design that had the best measure of performance for all metrics, sample size, or for all sets of test parameters. Designs that performed the best by one metric were not always the best when evaluated against other measures; designs that performed the best in some sets of test parameters performed worse in others. While there was little difference between various test designs for most sets of test parameters, there was a marked decrease in efficiency for the initial stage or sub phase of Wu and Tian when compared with the Neyer D-Optimal for test parameters substantially different from the underlying population. Additional simulation was conducted comparing various methods for extreme quantile estimation. The simulation showed that the c-optimal approach in the SenTest software yielded estimates with smaller MSE than the other designs studied.

Key Words: Sensitivity Test, Threshold Test, Optimal design

1. Introduction

Sensitivity tests are often used to estimate the parameters associated with latent continuous variables that cannot be measured. A typical application is determining the initiation characteristics of a hot-wire initiator used in an automotive air bag inflator. The initiator functions when current is applied to the wire embedded in energetic material. Current through the wire causes it to heat, and thus heat the energetic material close to the wire to its ignition temperature. Each initiator is assumed to have a critical initiation current level or threshold. Currents larger than this threshold level will cause the individual initiator to start the reaction to deploy the air bag (often called a success or response), while currents below this threshold will not lead to the initiator functioning (often called a failure or no response). Repeated testing of any initiator is not possible; a current that is not sufficient to cause initiation will nevertheless often cause a change in the interface between the wire and energetic material. To measure the probability of response, samples are tested at various current levels, and the reaction or lack thereof is noted.

The efficiency of a sensitivity test for providing estimates of the population parameters, or for providing estimates of extreme levels is a function of the algorithm used to pick the test levels. The population parameters estimates are often determined by computing Maximum Likelihood Estimates (MLEs) of the population parameters. Silvapulle (1981) showed that unique Maximum Likelihood Estimates (MLEs) will be obtained only if the successes and failures overlap; that is, the smallest success is smaller than the largest failure. Some of the earliest methods, such as the up-and-down or Bruceton method (Dixon and Mood, 1948) were designed, in the days before electronic calculators were available, to make estimates of the population parameters and their confidence intervals easier to calculate. Other methods, such as the Neyer D-Optimality design (often called the Neyer test and used in SenTest software) (Neyer, 1994) and the Three-phase Optimal Design of Sensitivity Experiments (dubbed 3pod) (Wu and Tian, 2014) were designed to provide efficient estimates, either of the population parameters, or of extreme levels. Both of these tests utilize an initial phase to yield unique MLEs of the parameters, and then use a D-Optimality approach to refine these estimates. As was shown by Banerjee (1980), a test method which concentrates the test levels near to the two D-Optimal points of approximately $\mu \pm 1.1381\ \sigma$ will yield statistics with variations that approach $\sigma^2/(0.392N)$ for μ and $\sigma^2/(0.507N)$ for σ, where N is the sample size.

Many other test methods, too numerous to mention here, have been adopted in the past. The reader is directed to consult the references in this paper for discussions of other methods.

This paper discusses recent simulation conducted on the Neyer test (Neyer, 1994) and the 3pod test (Wu and Tian, 2014). Wu and Tian compared these two methods, along with the Bruceton method. Their study was divided into two parts, an initial study to determine how often each of these three test methods yielded "successful" results. Quoting from Wu and Tian (2014) we have "A sequential experimental run of size n is called successful if its data satisfy the overlapping pattern in eq. (5). If not, it is said to be wasted." The simulation consisted of conducting runs until there were 1000 successful runs. Their table 2 shows that under some conditions, there were many more wasted runs than successful runs for the Up-and-Down test, many wasted runs for the Neyer (1994) test as well, but few wasted runs for the 3pod test method. The large number of wasted runs reported in Wu and Tian (2014) for the Neyer test (1994) has been briefly discussed in a paper (Ray, Roediger, and Neyer, 2014) which is summarized in the following paragraph.

Unfortunately, the algorithm that Wu and Tian used to represent the Neyer test is not the same one illustrated in the flow chart in the original paper (Neyer, 1994), nor the simulation conducted to demonstrate the results, nor does it correspond to the algorithm coded into SenTest$^{\text{TM}}$[1] and earlier Optimal$^{\text{TM}}$ that have been in use in multiple laboratories. The confusion may have been caused by the wording in Neyer (1994), which could be interpreted differently than the flow chart. "Once at least one success and failure have been obtained, a binary search is performed until the difference between the lowest success and highest failure is less than the estimate for sigma." would have been better stated as "Once at least one success and failure have been obtained, a binary search is performed whenever the difference between the lowest success and highest failure is less than or equal to the (possibly revised) guess for sigma." This paper focuses on comparing the Neyer test (the version represented by the flow chart and used in the SenTest software, hereafter called SenTest) and 3pod algorithms. The comparison was similar to that conducted in the first part of the Wu and Tian (2014) simulation study to determine the number of wasted runs, but is expanded to include more test conditions. Additional simulation was conducted comparing the various methods for extreme

[1] SenTest is sold by Neyer Software LLC.

quantile estimation. The simulation showed that the c-optimal approach yielded estimates with smaller MSE.

2. Comparison of the Neyer SenTest and 3pod test methods

As will be shown in the following discussion, the first two phases of the SenTest and 3pod algorithms are similar in many ways. Both of the algorithms can be described using the terms shown in **Table 1**. Most of the terms are identical in both Neyer (1994) and Wu and Tian (2014).

Table 1: List of common definitions

Term	Description
μ_{min}	Experimenter's guess for the Lower limit of the mean
μ_{max}	Experimenter's guess for the Upper limit of the mean
σ_g	Experimenter's guess for the standard deviation
x_i	Test level of sample *i*
y_i	Result of Test *i* (0 or 1)
k_1	Number of tests with success
k_0	Number of tests with failure
M_0	The largest x value among the y_i s with $y = 0$
m_1	The smallest x value among the y_i s with $y = 1$
x_{max}	$\max(x_1, x_2, \ldots, x_n)$
x_{min}	$\min(x_1, x_2, \ldots, x_n)$

The first phase for both tests is an initial search phase. This part of the test is designed to find the region where the responses are mixed. Wu and Tian (2014) describe three stages or sub phases of phase I in their paper; this work will divide their second stage into two different stages. The phases and stages are taken from Wu and Tian (2014). **Table 2** gives a **brief** description of some of the major differences between the two methods; the full explanation of the method requires many pages. The reader is suggested to consult the individual papers for the far more detailed explanation of each of the test methods.

3. Discussion of the differences between the SenTest and 3pod test methods

3.1 Stage I1, obtain at least 1 response and 1 no response

The first stage is designed to find the appropriate experimental range by obtaining at least one response and one non-response.

There are two main differences between the I1 stages of the two test methods. The first difference is that 3pod picks the first two test levels at the start of the test, whereas SenTest picks a single test level based upon the results of all previous test results. Thus, 3pod has the possibility of achieving a zone of mixed results after testing the first two items, with two additional tests to perform if this happens. However, using the recommended minimum range for μ_{max} - $\mu_{\text{min}} \geq 6\ \sigma_g$, these first two test levels would be at least 3 σ_g apart. The highest probability of getting sample 1 to respond and sample 2 to not respond would be if the test levels were equidistant from the population mean. In such a case when the σ_g exactly matches population standard deviation, the probability of achieving overlap on the first 2 tests is approximately 0.45%. When the range for μ_{max} - μ_{min} is larger than the minimum, or if the population standard deviation were smaller than

σ_g, the probability would be even lower. However, if the population standard deviation were much larger than σ_g, the probability approaches 25%.

Table 2: Table 2: Brief description of the differences between the 2 methods

Phase / Stage	Goal / End condition	SenTest (Neyer)	3pod (Wu and Tian)
I1	Obtain both 0 & 1 (k_0 & $k_1 > 0$)	$x_1 = \frac{1}{2}\mu_{\min} + \frac{1}{2}\mu_{\max}$ $x_{i+1} = \max\{(\mu_{\max} + x_{\max})/2, x_{\max} + 2\sigma_g, 2\,x_{\max} - x_{\min}\}$ until $y_{i+1} = 1$ $x_{i+1} = \min\{(\mu_{\min} + x_{\min})/2, x_{\min} - 2\sigma_g, 2\,x_{\min} - x_{\max}\}$ until $y_{i+1} = 0$ Expand test range exponentially	$x_1 = \frac{3}{4}\mu_{\min} + \frac{1}{4}\mu_{\max}$ $x_2 = \frac{1}{4}\mu_{\min} + \frac{3}{4}\mu_{\max}$ $x_{i+1} = \mu_{\max} + 1.5(i\text{-}1)\,\sigma_g$ until $y_{i+1} = 1$ $\mu_{\min}$ -1.5(i-1) σ_g until $y_{i+1} = 0$ Expand test range linearly
I2(i) (b)	Reduce m_1- M_0 return to this stage when σ_g changes	Whenever $m_1 - M_0 > 1\,\sigma_g$: x_{n+1} = MLE μ (σ free) Binary search	Whenever $m_1 - M_0 \geq 1.5\,\sigma_g$: x_{n+1} = MLE μ (σ fixed @ σ_g)
I2(i) (c,d)	Achieve overlap $m_1 < M_0$	x_{n+1} = D Optimal point using σ_g for σ. Set $\sigma_g = 0.8\,\sigma_g$ and re-enter I2(i)(b) if no overlap.	$x_{n+1} = m_1 + 0.3\,\sigma_g$ ($k_0 > k_1$) or $x_{n+1} = M_0 + 0.3\,\sigma_g$ ($k_0 \leq k_1$) Use other one if no overlap. Set $\sigma_g = \frac{2}{3}\,\sigma_g$ and re-enter I2(i)(b) if no overlap
I3	Enhance overlap	Not applicable	Test one item at $(M_0 + m_1)/2$ or two items at $(M_0 + m_1)/2 \pm \sigma_g$
II	D-optimality	Choose level which maximizes determinant of Fisher information matrix	Choose level which maximizes determinant of Fisher information matrix

The second main difference is in the algorithm used to expand the search region when the first few test levels result in all responses or all no responses. Consider the case that the first 2 test results are no responses. The 3pod test would choose $x_3 = \mu_{\max} + 1.5\,\sigma_g$, $x_4 = \mu_{\max} + 3\,\sigma_g$, $x_5 = \mu_{\max} + 4.5\,\sigma_g$, and so on, increasing the test range linearly (after the first 2 tests) with the test number until a response is achieved. The SenTest method would choose $x_2 = \frac{1}{4}\mu_{\min} + \frac{3}{4}\mu_{\max}$, $x_3 = \mu_{\max}$, $x_4 = 2\,\mu_{\max} - \mu_{\min}$, increasing the test range exponentially with the number of tests until a response is achieved. The SenTest initial search phase was designed to efficiently obtain both responses and non-responses if the parameter guesses are close to the population parameters, while also ensuring that the algorithm finds both responses and non-responses when the parameter guesses differ from the population by orders of magnitude.

The SenTest stage I1 algorithm would be expected to be completed with fewer samples than 3pod when the population mean was well outside the interval $[\mu_{\min}, \mu_{\max}]$. The simulation reported later validates this expectation.

3.2 Stage I2(i)(b), reduce the separation interval

This stage is designed to reduce the length of the separation interval [M_0, m_1] ($M_0 < m_1$) to less than (or equal to) a specified multiple of σ_g. This stage is always used in the SenTest method, and is used in the 3pod method in all cases except where overlap is obtained on the first 2 tests as described above. This stage is unique among the stages in that the exit criteria depend on a relative measure instead of absolute criteria; furthermore, it can be re-entered after passing the exit criteria when σ_g is reduced.

There are 2 main differences between the two methods. The 3pod method finds the MLE for μ using $\sigma = \sigma_g$. SenTest uses a binary search. This is equivalent to finding the MLE for μ, allowing σ to vary; because the MLE for μ is the middle of the separation interval when there is no overlap. When the separation interval is large compared with σ_g there is little difference between the 2 methods. The second difference is that the 3pod method may switch to the next stage slightly earlier because it switches at a larger multiple of σ_g.

3.3 Stage I2(i)(c,d), achieve overlap

This stage is designed to achieve an overlap interval [m_1, M_0], where $m_1 < M_0$; it ends when overlap is achieved. SenTest uses the D-Optimal approach of finding the test point that maximizes the determinant of the Fisher information matrix, with μ chosen as the midpoint of the separation interval and σ set equal to σ_g. To insure that the algorithm achieves overlap, if overlap is not achieved SenTest multiples σ_g by 0.8, and returns to the previous stage, I2(i)(b), to test for the ending condition. 3pod uses an alternate approach of testing 0.3 σ_g outside both limits of the separation interval. If overlap is not achieved with either of the two test levels, then 3pod multiples σ_g by 2/3, and returns to the previous stage, I2(i)(b), to test for the ending condition.

Designs using this 3pod stage would be expected to perform slightly better (worse) than SenTest when the guess for σ was significantly larger (smaller) than the population value. The simulation reported later validates this expectation.

3.4 Stage I3, enhance the overlapping regions

This phase is intended to enhance the overlap region by obtaining more points in this region. If $M_0 - m_1 \geq \sigma_g$, one test is run at $(M_0 + m_1)/2$; otherwise tests are run at the 2 levels $(M_0 + m_1)/2 \pm 0.5\ \sigma_g$. SenTest has no corresponding stage.

Designs using this 3pod stage would be expected to perform slightly better (worse) than SenTest for estimating the mean (standard deviation) since the test points chosen by this design would typically be closer to the mean than the 2 D-Optimal points. There should be no difference between designs that use this 3pod stage and those that don't in the determination of wasted runs. The simulation reported later validates these expectations.

4. Simulation Plan

It was relatively easy to incorporate the code to perform a sensitivity test according to the 3pod method into a previous version of the SenTest software called Optimal. In order to determine the effects on efficiency, the 4 stages discussed previously were coded separately, with the option to configure the software at run time to incorporate either the SenTest or 3pod version of each of the 4 stages. To get the original SenTest, the simulation ran the test with the switch "/q0". To turn on just 3pod stage I1, while keeping the rest as SenTest, required the switch "/q1". 3pod Stage I2(i)(b) alone required the switch "/q2". To use 3pod as defined with all 4 stages based upon 3pod requires the switch "/q15". Simulation was conducted on all 16 possible combinations of the 4 stages.

The simulation was conducted similarly to the original Neyer (1994) scheme with a given sample size and a fixed number of 4000 runs, as opposed to the Wu and Tian (2014) approach of conducting tests until there were 1000 runs with an overlap region. The larger number of runs (4000 to 1000) was chosen to reduce the statistical variation in the results. The population parameters were "fixed" at a mean of 0 and a standard deviation of 1. To account for any differences in test efficiency due to the exact position of the mean with respect to the test limits, the population mean was changed for each test by a mean offset with a uniform distribution between ± 1 σ. Tests were run with different sets of the test parameters $\mu_{\min}$, $\mu_{\max}$, and σ_g. All of the simulations were run with test parameters $\mu_{\min} = \mu_g - 4\sigma_g$ and $\mu_{\max} = \mu_g + 4\sigma_g$, where μ_g is the experimenter's guess for the mean, and σ_g is the experimenter's guess for the standard deviation. Simulation runs were conducted with $\sigma_g = (\frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8)\ \sigma$ and $\mu_g = \mu$, as well as $\sigma_g = \frac{1}{8}\ \sigma$ and $\mu_g = \mu +$ (2, 4, 8) σ. The simulation was repeated for each of the designs studied using the same threshold values. Threshold files were generated that contained 100 threshold values distributed as $N(0,1)$. These threshold files were used by the test software to determine the response. Thus, tests with one sample size have the same initial test levels and responses as all tests with a smaller sample size. Moreover, the simulation with different sets of experimenter guesses and different test designs all used the same set of data.

Three different metrics were used to evaluate the performance of the tests: the percentage of time the test failed to yield an overlap (similar to the "wasted runs" in Wu and Tian, 2014), the Mean Squared (truncated) Error (MStE) of the estimate of the mean, and the MStE of the estimate of the standard deviation. Because there are often "wild" estimates of the parameters when analyzing tests, especially with small sample sizes, the errors for any test were truncated to be no larger than 2 σ (2 since $\sigma = 1$) for all of the simulations reported in the next section.

Simulation Results

The simulations with 4000 runs showed similar results to the runs of 1000; the only difference is that the curves were smoother with the runs of 4000 versus the runs of 1000. Thus, the results reported here are mainly of the simulation with 4000 runs. The graphs in Figure 1 and Figure 2 show the results of some of the simulations for determining the percentage of tests that resulted in no overlap. The other graphs show similar results. In most cases, the 3pod design was **slightly** more efficient at yielding overlap results compared with Sentest. The right graph in Figure 1 shows an example of the largest advantage of 3pod over SenTest. However, the SenTest design is much more efficient at yielding overlap when the experimenter's guess for the range of the mean is far from the population mean, as shown in the right graph of Figure 2.

To test which stage(s) of 3pod were responsible for this behavior, simulations were run with the SenTest stage I1 and the rest of the 3pod stages (/Q14) as well as 3pod stage I1 and the rest of the SenTest stages (/Q01). The simulation clearly showed that the weak performance of 3pod for the stage I1 when the mean was far from the experimenter's guess was due to the linear search algorithm of 3pod stage I1 as shown in the right graph of Figure 3. The left graph shows one example where the 3pod method (/Q15) was less efficient than /Q14. This ranking of wasted runs occurred for almost all of the simulations conducted and for most of the mixtures of SenTest and 3pod stages. The weak performance of the 3pod stage I1 was also found when evaluated by the ability to estimate the population parameters.

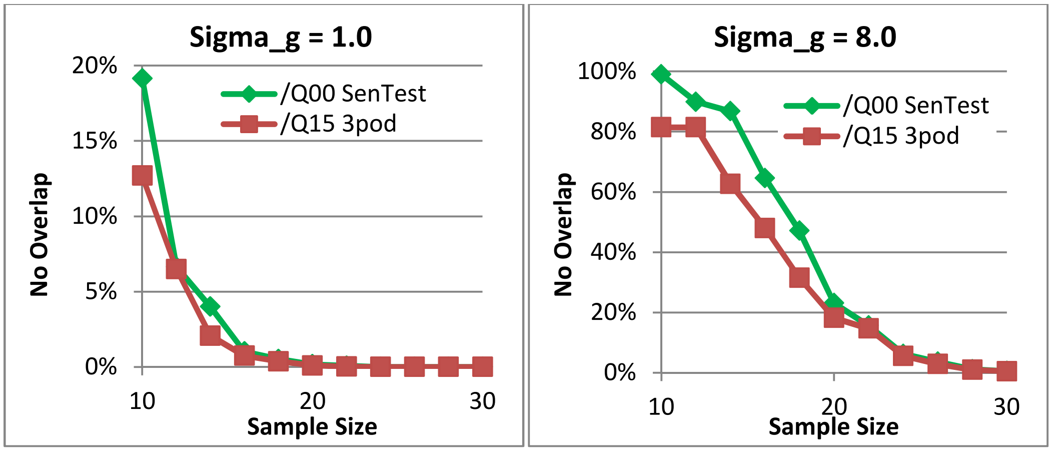

Figure 1: Fraction of "wasted runs" when $\mu_g = \mu$ and $\sigma_g = 1$ (left), and $\sigma_g = 8$ (right)

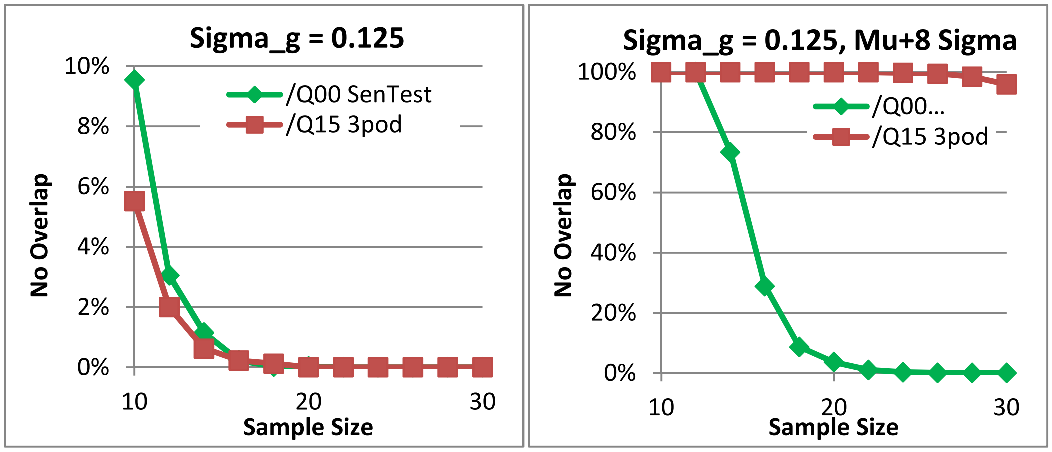

Figure 2: Fraction of "wasted runs" when $\sigma_g = 1/8$ and $\mu_g = \mu$ (left), and $\mu_g = \mu + 8\ \sigma_g$ (right)

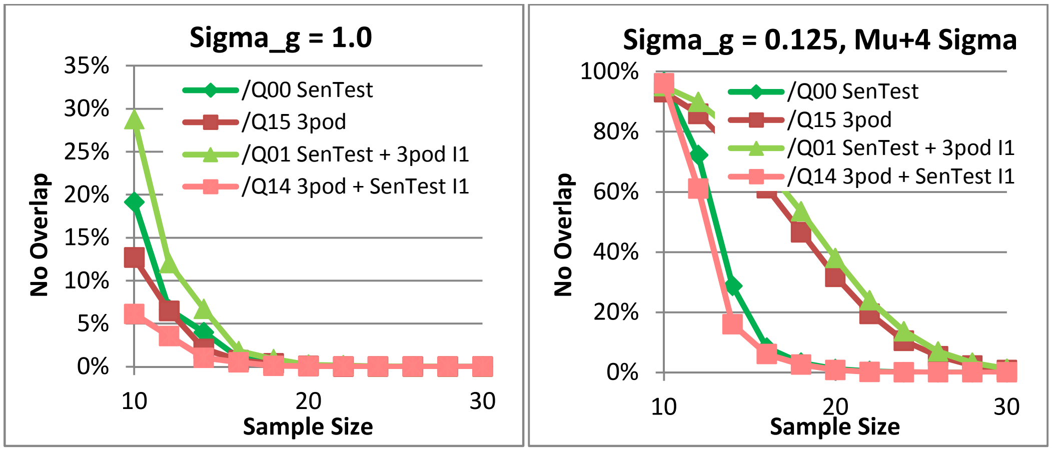

Figure 3: Fraction of "wasted runs" when $\mu_g = \mu$ and $\sigma_g = 1$ (left), and $\mu_g = \mu + 4\ \sigma_g$ and $\sigma_g = 1/8$ (right)

As will be shown in the following graphs, efficiency in reaching overlap results is not strongly correlated with efficiency in determining parameter estimates. The major exception is that the designs that perform poorly in finishing stage I1 also perform poorly in determining parameter estimates.

The graphs in Figure 4 through Figure 13 show the results of the simulations for determining the estimate of the mean and standard deviation for 6 of the 16 test designs studied. The 6 designs chosen for the graphs were the 5 best performing test designs plus 3pod. The graphs show the results expressed in units of σ^2 / MStE (μ), and σ^2 / MStE (σ) where MStE is the Mean Squared truncated Error, with a truncation of ± 2 σ. This truncation was rarely needed in the data analysis, but would protect against a single wild estimate of the parameter. The figures show that when $\sigma_g \leq \sigma$, most of the test designs yield estimates for μ with greater precision than expected for a D-Optimal test. This can be explained by the fact that the estimate for σ is biased towards the low side, resulting in testing closer to the mean than the 2 D-Optimal points, which results in greater efficiency for estimating the mean at the expense of less efficiency for estimating the standard deviation. Except for the cases where the estimated range for the mean is far from the actual population mean (see Figure 12 and Figure 13) the efficiency curves of the 16 variations of test design are all similar, with each design performing the best for some combination of population parameter guesses and/or sample size.

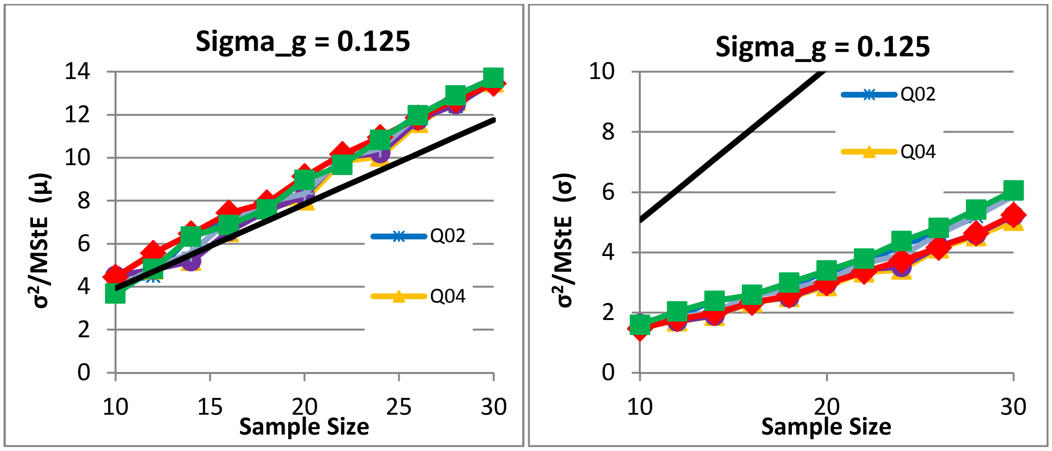

Figure 4: Efficiency of determining μ (left) & σ (right) when σ_g = 1/8

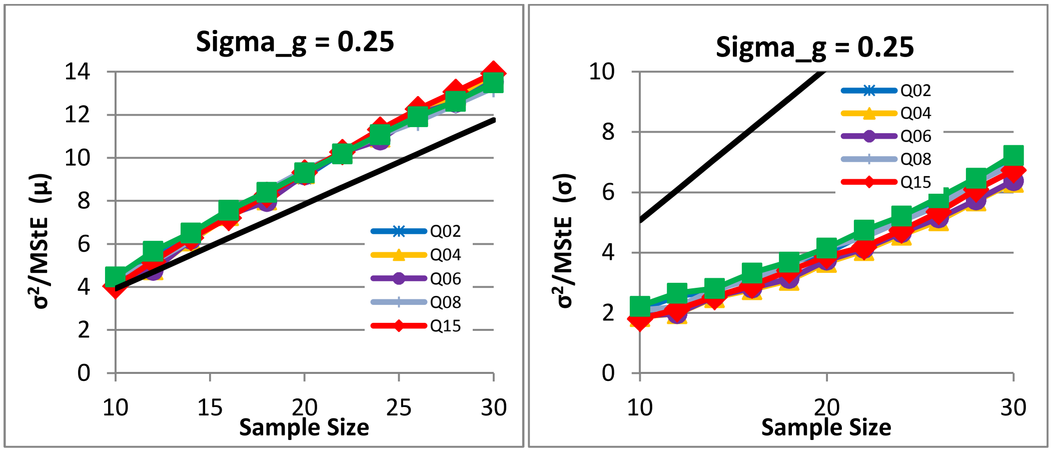

Figure 5: Efficiency of determining μ (left) & σ (right) when σ_g = ¼

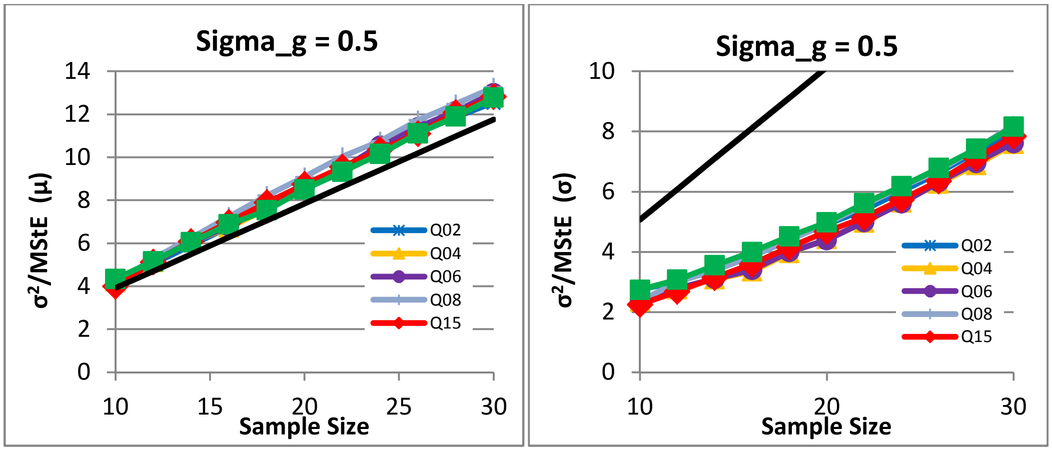

Figure 6: Efficiency of determining μ (left) & σ (right) when σ_g = ½

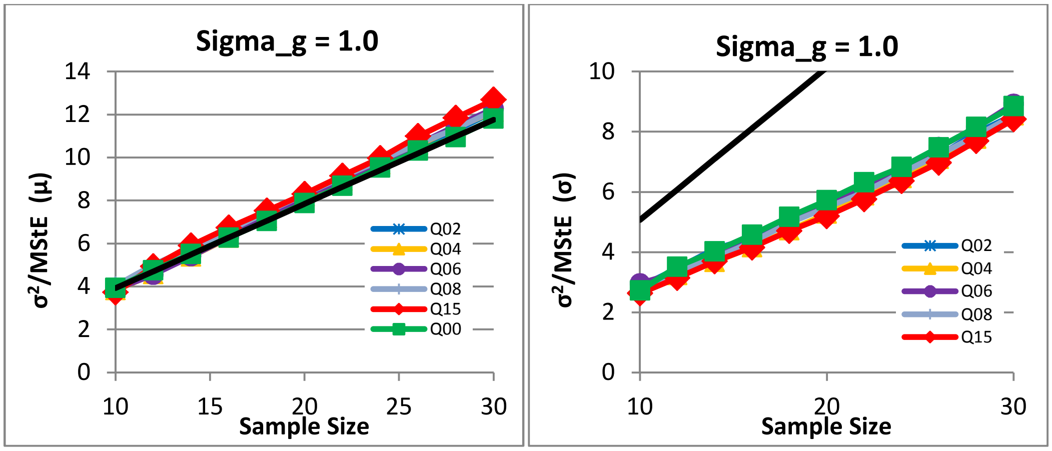

Figure 7: Efficiency of determining μ (left) & σ (right) when σ_g = 1

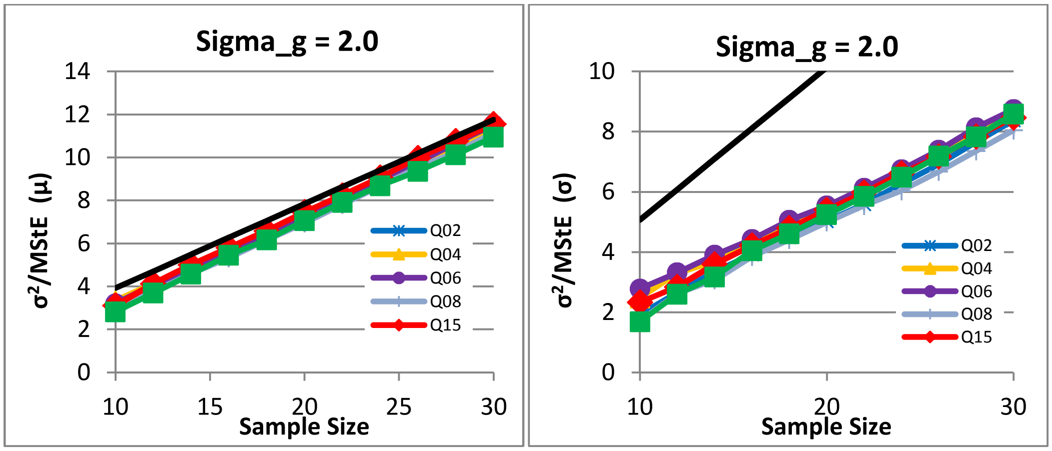

Figure 8: Efficiency of determining μ (left) & σ (right) when σ_g = 2

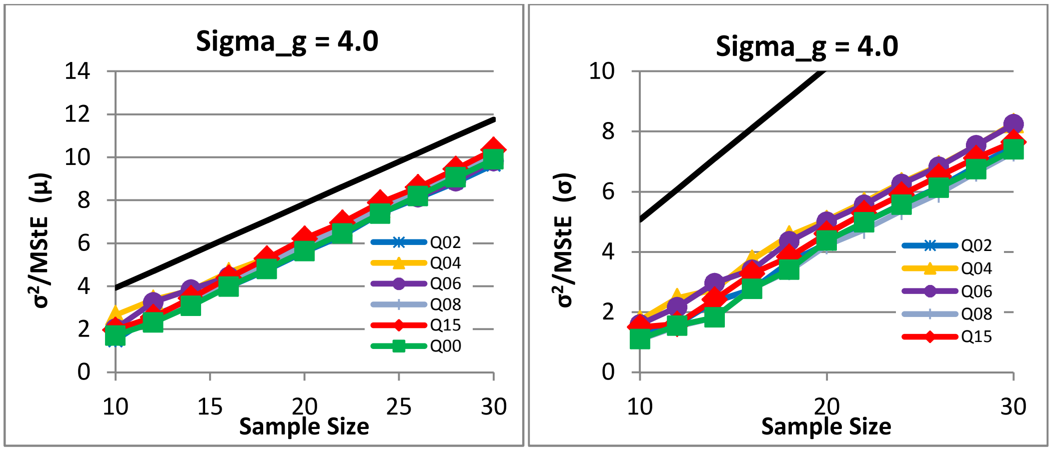

Figure 9: Efficiency of determining μ (left) & σ (right) when $\sigma_g = 4$

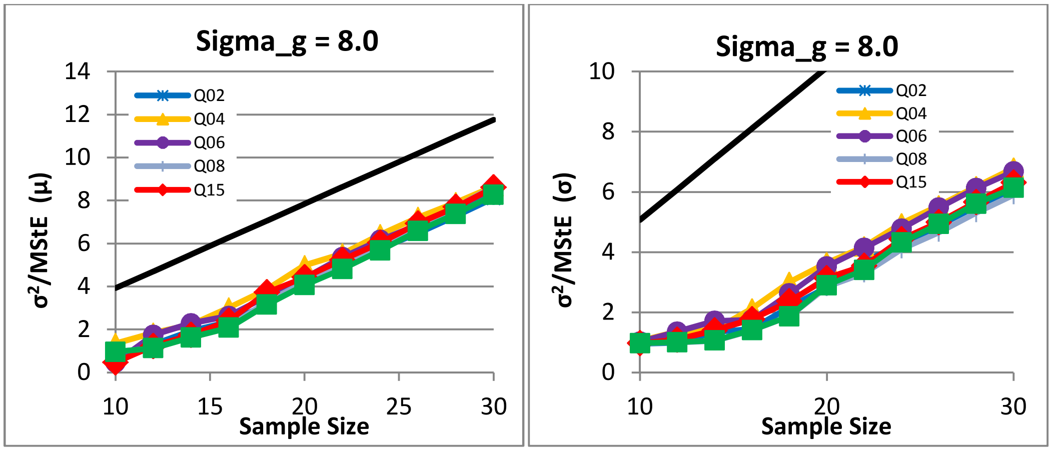

Figure 10: Efficiency of determining μ (left) & σ (right) when $\sigma_g = 8$

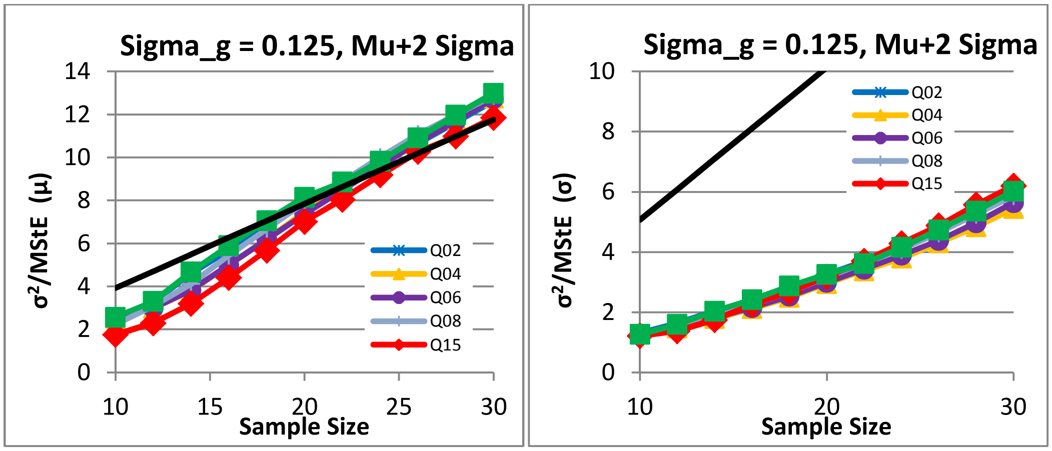

Figure 11: Efficiency of determining μ (left) & σ (right) when $\sigma_g = 1/8$ & $\mu_g = 2\ \sigma$

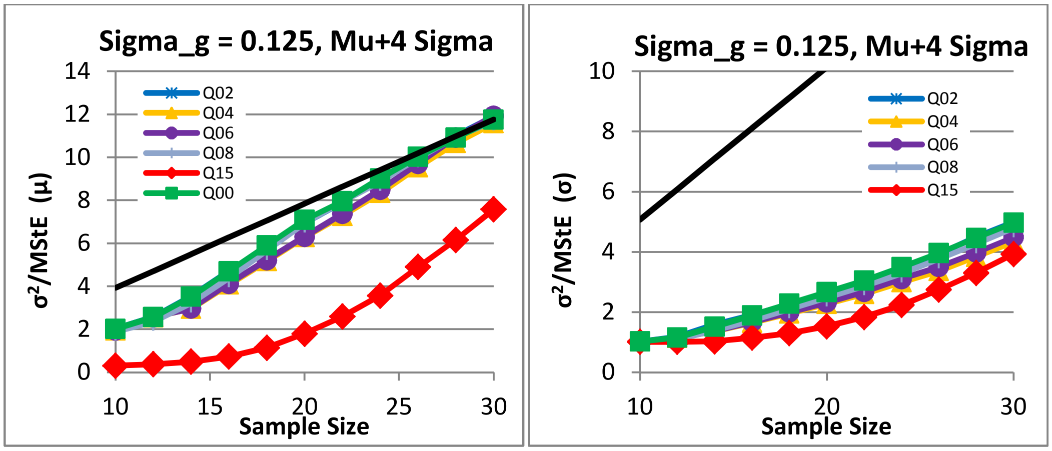

Figure 12: Efficiency of determining μ (left) & σ (right) when $\sigma_g = 1/8$ & $\mu_g = 4\ \sigma$

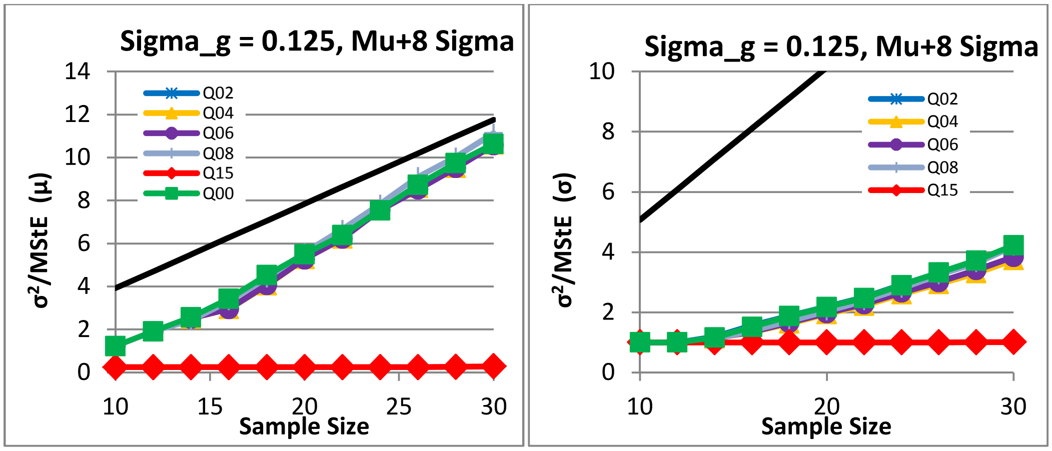

Figure 13: Efficiency of determining μ (left) & σ (right) when $\sigma_g = 1/8$ & $\mu_g = 8\ \sigma$

There is no one design that is the most efficient for all or even most combinations of population estimates and sample sizes studied, but it is possible to look at all of the simulation to see whether the effect of one or more of the 3pod stages provided an improvement to the SenTest design. All 16 possible combinations of the 4 stages were analyzed to look for the design that was most efficient. The Q00 (SenTest) design was the most efficient design for 42 combinations of population estimates and sample sizes studied. The next biggest number was 25 for Q02 & Q12. The Q00 (SenTest) design also had the highest average Sigma^2/MSE, with an average of 5.5725, Q02 was 2nd with 5.5477, and Q08 was 3rd with 5.5334. (See Table 3.) The main driver for the improved ranking of the Q00 design is the better performance on the estimate for the standard deviation. Extreme quantiles depend more on estimates of the standard deviation than the mean; the simulation suggests that starting with Q00 (SenTest) instead of Q15 (3pod) should provide estimates of extreme quantiles that have smaller MSE. Looking at the individual stages, it appears that there was a consistent decrease in efficiency for all designs when using 3pod phase I1, and a slight decrease in efficiency when using any of the other 3pod phases. In general, the decrease in efficiency was larger the more of the 3pod stages were used.

Table 3: Comparison of the efficiency of the Various Test Designs

Design	Best Design			Average Sigma^2/MSE		
	Best Mu	Best Sigma	Best Total	Avg Mu	Avg Sigma	Avg Total
Q00	15	27	42	7.1628	3.9821	5.5725
Q01	1	16	17	6.0896	3.7415	4.9155
Q02	10	15	25	7.1522	3.9432	5.5477
Q03	0	16	16	6.1091	3.7435	4.9263
Q04	6	10	16	7.1901	3.8621	5.5261
Q05	9	1	10	6.3105	3.6563	4.9834
Q06	3	15	18	7.1495	3.9140	5.5318
Q07	0	3	3	6.2066	3.7483	4.9775
Q08	22	2	24	7.2374	3.8293	5.5334
Q09	2	1	3	6.1512	3.6188	4.8850
Q10	7	1	8	7.1849	3.7858	5.4853
Q11	1	3	4	6.1802	3.5838	4.8820
Q12	14	11	25	7.1462	3.7551	5.4507
Q13	17	1	18	6.3200	3.6173	4.9687
Q14	6	2	8	7.0942	3.7755	5.4348
Q15	7	1	8	6.2922	3.6645	4.9784

Estimating Extreme Quantiles

A significant portion of the Wu and Tian (2014) paper is devoted to the third and final phase of their design which is optimized to provide an efficient estimate of an extreme quantile. Such an approach was very briefly discussed in the original D-Optimality paper by Neyer (1994), but was not the main focus of that work. SenTest and previously Optimal have had the ability to be run optimized for finding a single quantile.

It is relatively easy to modify the D-Optimal algorithm to one designed to find the point that maximizes information about a single quantile, say x_p, where p represents the probability. Maximizing information about x_p instead of the determinant of the information matrix results in a *c*-optimal design. For probabilities less extreme than the two Sigma-Optimal points (approximately $\mu \pm 1.575\ \sigma$, or within the approximate range 5.8% – 94.2%) *c*-optimality is achieved by testing at the point estimate. In this case, the test design is similar to the MLE recursive method of Wu (1985). Note however, that the estimate for x_p is computed from the MLE of the population parameters, and not by the next stimulus . For more extreme quantiles, *c*-optimality is achieved by testing at the two Sigma-Optimal points, with the ratio of the upper and lower points determined by how extreme the level is; this ratio approaches 50% as x_p approaches 0 or 1.

Simulation of various designs was conducted to determine their efficiency in estimating extreme quantiles. There are several slightly different approaches for this effort. The first approach was to use a method similar to that described in Wu and Tian (2014): perform a D-optimal design for a fixed quantity, and then to switch to a *c*-optimal design for the remaining samples. Tests conducted to this approach are labeled Neyer D#$_1$-C#$_2$ in Table 4. The second approach was to perform the phase 1 of the Neyer D-optimal design and then to switch to the *c*-optimal design instead of the D-optimal design at the start of phase II (once overlap occurred). This approach is identified as Neyer P# in Table 4. A third approach is to start with the *c*-optimal approach in stage I2(i)(c,d), testing at the c-optimal point computed by using the estimate σ_g for σ. This approach is identified as Neyer C# in Table 4. The final approach is to use the D-Optimal method throughout. Table 4 shows the results of the analysis; the data for the 3pod, Wu, and RMJ tests were copied from Wu and Tian (2014) with the results for μ_g = 9, 10, 11 averaged since there was little variation between the simulations for different μ_g. As before, the simulation conducted for this paper had a uniform distribution between ± 1 for μ_g.

Table 4: Comparison of various methods for estimating extreme quantiles.

					RMSE				
Method	D	C	Total	Prob Level	Sigma_g = 0.5	Sigma_g = 1.0	Sigma_g = 2.0	Sigma_g = 3.0	Sigma_g = 4.0
Neyer D40	40	0	40	0.1	0.4379	0.4335	0.4417	0.4550	0.4634
Neyer D25-C15	25	15	40	0.1	0.3815	0.3691	0.3815	0.3837	0.3881
Neyer C40			40	0.1	0.3355	0.3346	0.3415	0.3468	0.3567
Neyer P40			40	0.1	0.3333	0.3307	0.3394	0.3458	0.3569
3pod (25,15)	25	15	40	0.1	0.4408	0.4511	0.4788	0.4493	0.4514
Wu			40	0.1	0.4284	0.3643	0.3960	0.4318	0.4800
RMJ	0	40	40	0.1	0.3043	0.2656	0.3082	0.3576	0.4371
Neyer D60	60	0	60	0.01	0.5358	0.5402	0.5411	0.5517	0.5562
Neyer D25-C35	25	35	60	0.01	0.4796	0.4715	0.4660	0.4783	0.4817
Neyer D30-C30	30	30	60	0.01	0.4881	0.4717	0.4731	0.4843	0.4893
Neyer C60			60	0.01	0.4721	0.4581	0.4776	0.4795	0.4823
Neyer P60			60	0.01	0.4725	0.4513	0.4628	0.4666	0.4639
3pod (25, 35)	25	35	60	0.01	0.5549	0.5682	0.5633	0.5028	0.5109
3pod (30,30)	30	30	60	0.01	0.5791	0.5727	0.5902	0.5502	0.5575
Wu			60	0.01	1.2803	1.1792	1.7099	1.1916	1.2667
RMJ	0	60	60	0.01	0.4600	0.4102	0.5875	2.3509	4.9470
Neyer D80	80	0	80	0.001	0.5832	0.5883	0.5721	0.5917	0.5973
Neyer D25-C55	25	55	80	0.001	0.5127	0.5100	0.5174	0.5077	0.5228
Neyer D35-C45	35	45	80	0.001	0.5324	0.5164	0.5158	0.5240	0.5290
Neyer C80			80	0.001	0.5175	0.5007	0.5128	0.5109	0.5182
Neyer P80			80	0.001	0.5063	0.5083	0.5141	0.5297	0.5294
3pod (25,55)	25	55	80	0.001	0.7863	0.7909	0.7862	0.6899	0.7425
3pod (35,45)	35	45	80	0.001	0.8283	0.8077	0.7807	0.7307	0.7527
Wu			80	0.001	1.3367	1.8104	2.5391	2.0050	2.2416
RMJ	0	80	80	0.001	0.6964	0.6027	1.4926	5.0104	8.4440

Inspection of Table 4 shows that the Neyer C and P designs had similar efficiency. Similarly, the 2 Neyer D-C designs were comparable, with the slightly better performance for the design that devoted a larger fraction to the *c*-optimal design. A relative ranking of efficiency would be:

Neyer P ≈ Neyer C > Neyer D-C > Neyer D > 3pod > Wu

This ranking is similar for all three extreme levels, except that Wu > Neyer D > 3pod for the sample size of 40 and 10% probability level. The RMJ method is not included in this ranking since the results depend critically on the starting level.

There are several reasons for the difference in rankings for this work compared with the work of Wu and Tian (2014). The first is that Wu and Tian used a different version of the D-optimal test as mentioned earlier in this paper. Second, the work reported in this section uses analysis that depends **explicitly on the probability distribution function**, whereas the 3pod and RMJ methods have no **explicit** distribution dependence. As will be discussed in the next section, the absence of explicit distribution dependence does not mean that the results are not extremely dependent on the exact form of the distribution.

Estimating Extreme Quantiles with Relatively Small Sample Sizes

It is difficult to obtain accurate estimates of extreme quantiles that are truly independent of both the assumed distribution and the test parameters with only a small sample. There are two general sensitivity test approaches that are used to estimate a quantile. Approach 1 is to *assume* that the probability of response when tested at test parameter x is $M(x)$, and that $M(x)$ is a *known* cumulative distribution function, one that can be characterized by a small number of unknown population parameters. The approach is to conduct tests at various levels throughout the distribution to estimate these parameters and thus to be able to estimate the probability of response $M(x)$ as a function of x. This is the approach used by SenTest for the results in Table 4. The accuracy of the probability response curve or the probability at any given point depends critically on the *assumed* form of $M(x)$.

Approach 2 is to assume that M(x) is an unknown cumulative distribution function whose first derivative is known or can be guessed at one particular point x_p such that $M(x_p) = p$ for probability p. Test levels are chosen in the vicinity of the x_p, sometimes according to the design $x_{n+1} = x_n - a_n(y_n - b_n)$, where b_n -> p for large n. In many cases the test points will cluster around x_p, with the result that a good estimate for x_p is x_{n+1}. This approach is not *explicitly* distribution dependent; however, the constants a_n and b_n are calculated assuming a given distribution. Moreover, the efficiency of these methods depends critically on a choice for x_1 that is close to the true value, as well as having a sample size large enough so that there are a reasonable number of both responses and non-responses in the vicinity of x_p. The 3pod design uses a mixed approach of starting with a distribution specific test designed to provide estimates of the population parameters. The starting point for the point estimate search is calculated using this *assumed* distribution.

Unfortunately, there is no guarantee that the function $M(x)$ is the simple distribution function of a few parameters that are the basis of the distribution independence methods. Moreover, there is no guarantee that it is distribution function, or even that it is monotonic. The response of many engineered products is often governed by several parameters with distributions that may be approximated as normal, and thus $M(x)$ may appear normal in the center of the distribution. However, most engineered products have specific limits for each of these parameters, and it is not unusual to receive material that is distributed close to one of these limits. If the cut off parameter is the largest contributor to the distribution, the distribution will not resemble a Gaussian. In such a case, there may be quite a few parameters required to characterize the distribution. In addition, the population could have some units that are defective and will not respond no matter what the stimulus is. The percentage of defects is unlikely to be governed by the same set of parameters as those that govern the non-defective center of the population. Finally, the response physics may change over the region tested, so that a higher stimulus has a reduced probability of response. For these reasons, unless the sample size is sufficiently large so that there are a number of responses and non-responses in the vicinity of x_p, any of the test methods will have at least an *implicit distribution* or *experimenter guess dependence*.

While the problem of determining the sample size necessary to provide reasonable estimates of an extreme quantile, x_p, independent of distribution or experimental parameters is difficult to analyze analytically, it is straight forward to analyze the opposite case of estimating the probability p of a given extreme level x_p. The confidence intervals can be calculated exactly. To obtain a 90% confidence interval of size [0.5p to 1.5p] for probabilities at least as extreme as 0.1, the sample size must be large enough that there are at least 10 of each of responses and non-responses. For example, a sample size of 10,000 would be required to obtain a ± 50% estimate of the probability p @ 90% confidence when $p = 0.1\%$. It is not unreasonable to assume that the sample size needed to determine the level x_p that has a probability of response of p is at least as large as the sample size needed to estimate the probability p if independence of distribution and experimental parameters is required.

It is also possible to analyze a slightly different situation. If it were possible to measure the threshold for each individual item, then it would be possible to estimate an extreme quantile by measuring a large number of parts, and picking the largest value as the estimate of the quantile. For example, to determine the 99.9% quantile, we would need a sample size of 1000 to obtain an unbiased estimate, and to obtain a 95% confidence level, a sample of 2995 would be required (almost 3 times the inverse of the smaller of p and q). In the case of sensitivity testing where there is less information to be gathered by testing each item, a multiplier larger than 3 would be expected.

As the two preceding paragraphs suggest, a very large sample would be required to reliably estimate extreme quantiles that are independent of distribution and experimental parameters. Such large sample sizes are rarely used in practice, with the result that the estimate of the quantile is strongly dependent on the initial experimenter's guess as well as the distribution.

References

Banerjee, K. S. (1980), "On the Efficiency of Sensitivity Experiments Analyzed by the Maximum Likelihood Estimation Procedure Under the Cumulative Normal Response," Technical Report ARBRL-TR-02269, U.S. Army Armament Research and Development Command, Aberdeen Proving Ground, MD.

Dixon, W. J. and Mood, A. M. (1948), "A Method for Obtaining and Analyzing Sensitivity Data," Journal of the American Statistical Association, 43, pp. 109-126.

Joseph, V. R (2004), “Efficient Robbins-Monro procedure for binary data,” Biometrika 91, 461-470.

Neyer, B. T (1994), “A D-optimality-based sensitivity test,” Technometrics 36, 61-70.

Ray, D. M., Roediger, P. A., and Neyer, B. T (2014) “Commentary: Three-phase optimal design of sensitivity experiments,” Journal of Statistical Planning and Inference.

Silvapulle, M. J. (1981), "On the Existence of Maximum Likelihood Estimators for the Binomial Response Models," Journal of the Royal Statistical Society B, 43, pp. 310-313.

Wu, C. F. J. (1985), “Efficient sequential designs with binary data,” J. Am. Stat. Assoc. 80, 974-984.

Wu, C. F. J. and Tian, Y. (2014), “Three-phase optimal design of sensitivity experiments,” Journal of Statistical Planning and Inference.