Assessing risk due to small sample size in probability of detection analysis using tolerance intervals

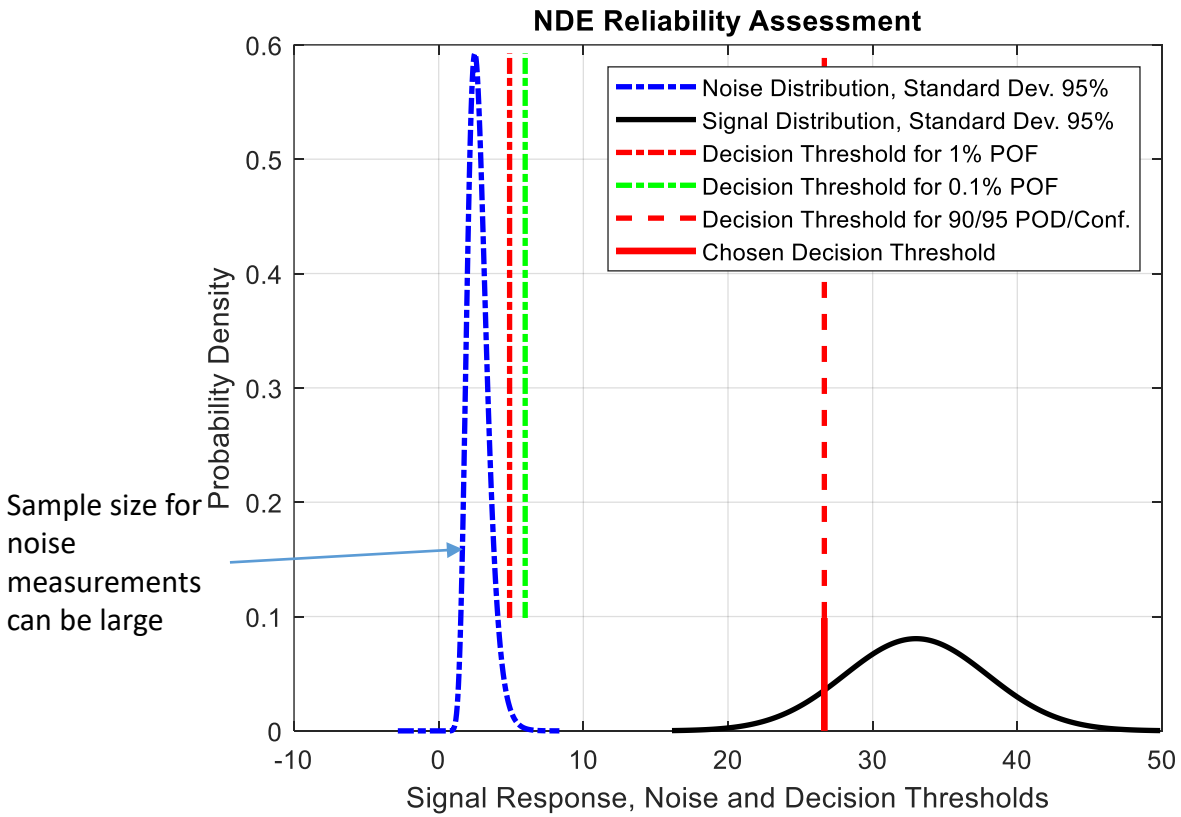
SPIE Smart Structures + NDE 2022

March 8, 2022

Ajay Koshti

NASA Johnson Space Center

Background



An illustration of NDE reliability assessment in **Limited Sample (LS) POD** analysis

POD - Probability of Detection
POF – Probability of False Positive

- LS POD[1] analysis **originally** used signal responses from **nominally fixed target size flaws**
- Noise is measured as a signal response outside flaw area
- Signal response distribution can be described by mean and standard deviation
- **Assumes that the signal response sample is RANDOM to the population.**
- Objective is to determine **decision threshold** that meets POD and POF reliability conditions
- Probability and Confidence in describing signal response distribution or predicting whether a given signal response value belongs to the population distribution depends upon **sample size of signal response measurements**.
- Creating a large sample size for independent signal repose measurements for flaws (e.g. 1 measurement per flaw) may be expensive, time consuming and/or impractical.
- Hence, use of LS POD is attractive to make the NDE qualification or reliability assessment **more practical, less expensive and requiring less time.**
- **However, small sample size poses a risk.**

[1] - Koshti, A. M., "Using requirements on merit ratios for assessing reliability of NDE flaw detection," SPIE Smart Structures and NDE, Proc. SPIE 11593, (2021).

Issue

- Sampling risk in LS POD comes from **small sample size** (e.g. 6) and **sample bias**
  - Small samples are not likely to be **random** to the population or **representative** of the population and are likely to be biased
  - Biased samples have lower standard deviation compared to the population
  - LS POD Analysis based on **small biased sample of target size flaws** can lead to overestimation of POD

Objective of LS POD Analysis

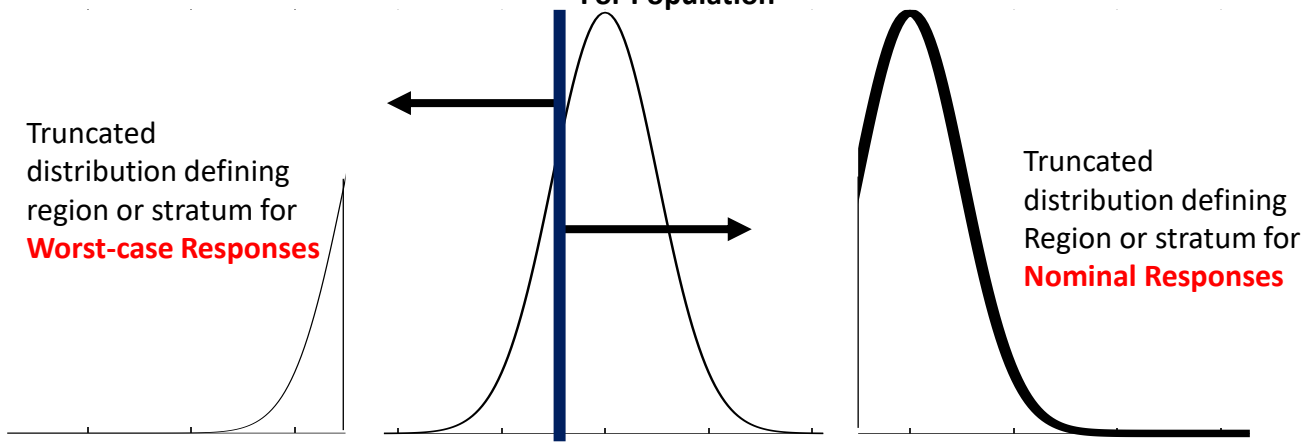
- Primary objective of LS POD analysis is to determine signal response **decision threshold** such that there is at least 90% (i.e. POD) population signal response data (with 95% confidence) above this threshold.
- Decision threshold $\hat{a}^{sample}_{dec.thr.90/95}$ at POD/Conf. 90/95 of data is computed from the sample.
- Decision threshold ($\hat{a}^{population}_{dec.thr.90/95}$) at POD/Conf. 90/95 of data for population is a **theoretical quantity**.
- Following inequality shall be true for LS POD (or any POD analysis) in order to accept LS POD analysis and to mitigate sampling error.

$$\hat{a}^{sample}_{dec.thr.90/95} \leq \hat{a}^{population}_{dec.thr.90/95}$$

- **In other words, $\hat{a}^{sample}_{dec.thr.90/95}$ shall be a conservative estimation of $\leq \hat{a}^{population}_{dec.thr.90/95}$**

Signal Response Population Regions, Nominal versus Worst-Case

Example 1: 30/70 Probability Split Partition For Population

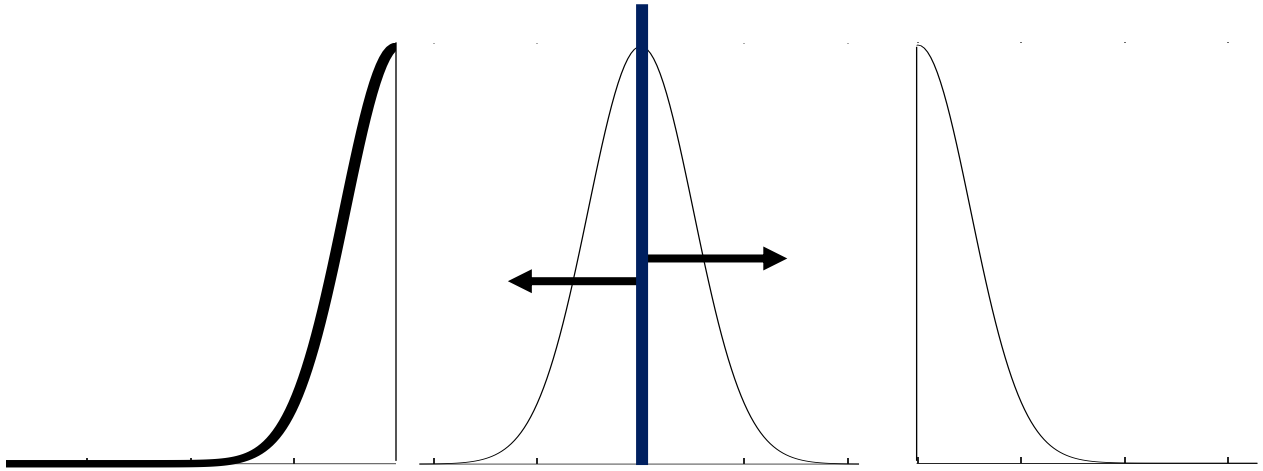


Split between Nominal and Worst-Case Signal Response Values in Population Distribution

- Higher responses are attributed to nominal flaw and part conditions with chance of occurrence e.g. ~50-70%
  - NDE signal response sample is likely to be nominal response sample, if nominal parameters are used to make flaw specimens
- Worst-Case or lower responses are attributed to off-nominal flaw and part conditions with chance of occurrence e.g. ~30-50%

Sampling fraction for Worst-Case = 0.30
Sampling fraction for Nominal case = 0.70

Example 2: 50/50 Probability Split Partition for Population



Split between Nominal and Worst-Case Signal Response Values in Population Distribution

Differences Between Population and Sample Signal Responses, and Sampling Definitions

Population Signal Responses from Target Size Flaws	Sample Signal Responses from Target Size Flaws
Real flaws	**Artificially manufactured flaws**
In real parts	Specimens are made using controlled fabrication process
All part surface geometries (cylindrical, spherical and flat, fillet) are assumed	Simple specimen surface geometry compared to part (e.g. flat)
All applicable material types are assumed	One material type/alloy is used (nominal)
All applicable surface finishes are assumed	Fixed value smooth surface finish is assumed (nominal)
Applicable variation in flaw morphology is assumed	Flaw morphology is controlled by controlling flaw manufacturing process (nominal)
All applicable orientations of flaws are assumed	Nominal orientation of flaws is assumed

- **Due to differences between population and sample, sample may not be fully representative of the population, i.e. sample may have a bias.**

Definitions

- A **representative sample (e.g. stratified sample)** is a group **or** set chosen from a larger statistical population according to specified characteristics.
- A **random sample** is a group **or** set chosen in a **random** manner from a larger population.
    - Both representative and random normal samples are acceptable for k_1 factor statistics, although a representative sample has less variance in results and reduces magnitude of error.
- In statistics, **sampling bias** is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher relative sampling probability than others.
    - It results in a biased sample, a non-random sample of a population.
    - If sampling bias is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling.
    - If a sample is neither random nor representative, it may be a **biased** sample.

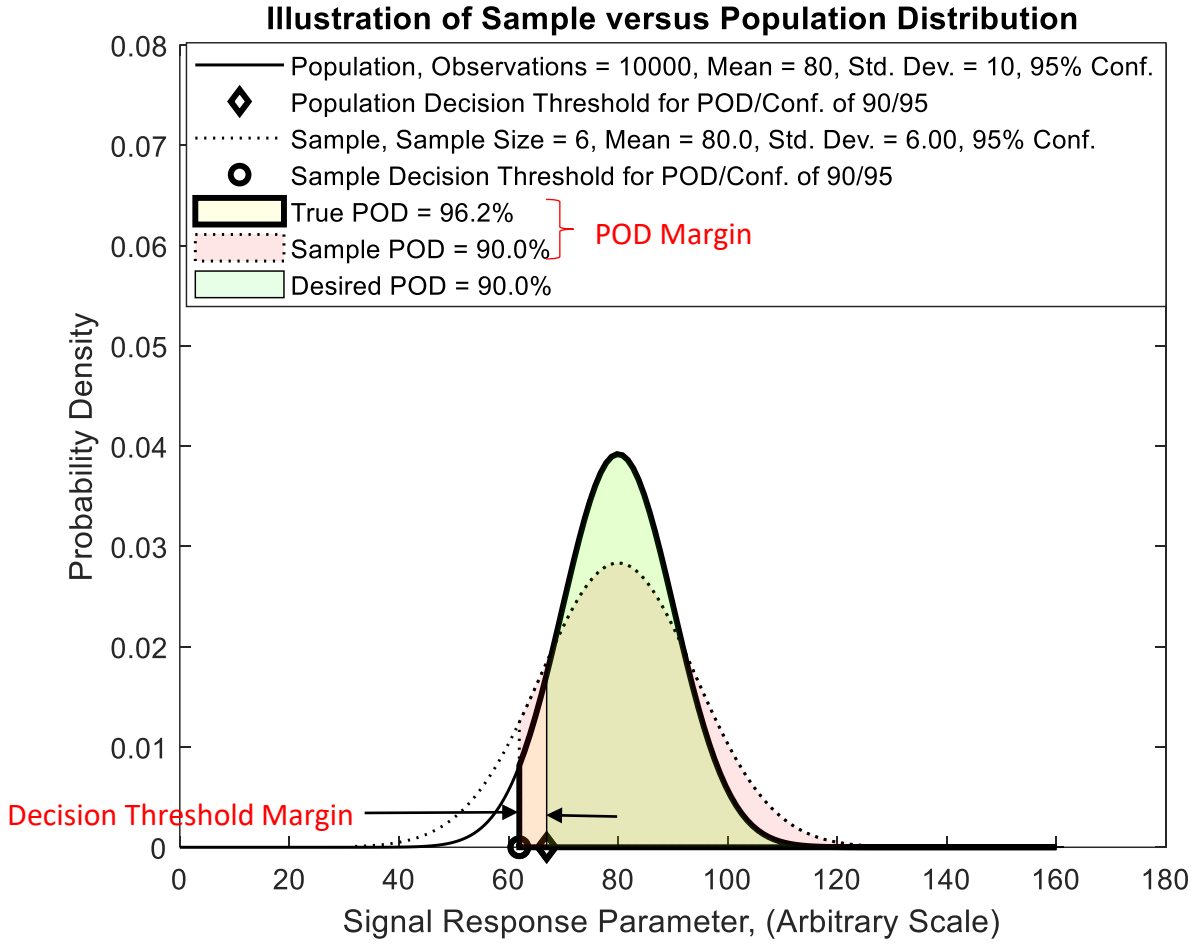
LS POD Concerns

- Small Sample Concern
  - Higher variability of decision threshold for small sample size (e.g. 6)
  - Higher sampling error in POD values for small sample size (e.g. 6)
    - Note: 90/95 POD/confidence is assured for both random and representative samples
- Non-random Sample Concern
  - A small sample generated using fabricated flaw specimens is not likely to be random
    - due to well controlled process of fabricating flaw specimens to nominal parameters and inadequacy of sample size to accommodate all factors that affect signal response
  - If a sample is biased to higher signal response values, it causes overestimation error in POD which is further compounded by small sample concern.

Approach to Mitigate Sampling Risk

- Use a **validated sampling scheme** to mitigate risk from small biased sample or to reduce sampling error
  - Some approaches include using a **representative sample** which reduces variance in POD estimates, which also reduces magnitude of error. A representative sample can be used in LS POD
  - Other approach is to use a **sample of smaller size flaws called nominal sub-target size flaws** and qualify the sampling scheme
  - It is proposed that Monte Carlo Sampling sensitivity analysis may be performed to design/validate a sampling scheme and mitigate or reduce sampling risk

Illustration of Representative Sample versus Population - Concept only



$$\hat{a}_{90/95,dec.thr} = \hat{a}_{mean} - k_{1,90/95}\sigma$$

$\hat{a}_{mean}$ =Mean of signal responses
σ = Standard deviation of signal responses
$k_{1,90/95}$= k1 tolerance factor

- Representative sample has both low and high value readings in correct proportion
- Representative sample is like a random sample in its effect on POD estimation except the tolerance ranges on 90/95 decision threshold are smaller.
- **LS POD assumes that the sample is random or representative of the population**
- This is an example of unbiased sample
- LS POD will provide 95% confidence for minimum 90% POD for analysis based on representative sample

Illustration of Nominal Sample and Its Influence on POD Estimation – Concept only

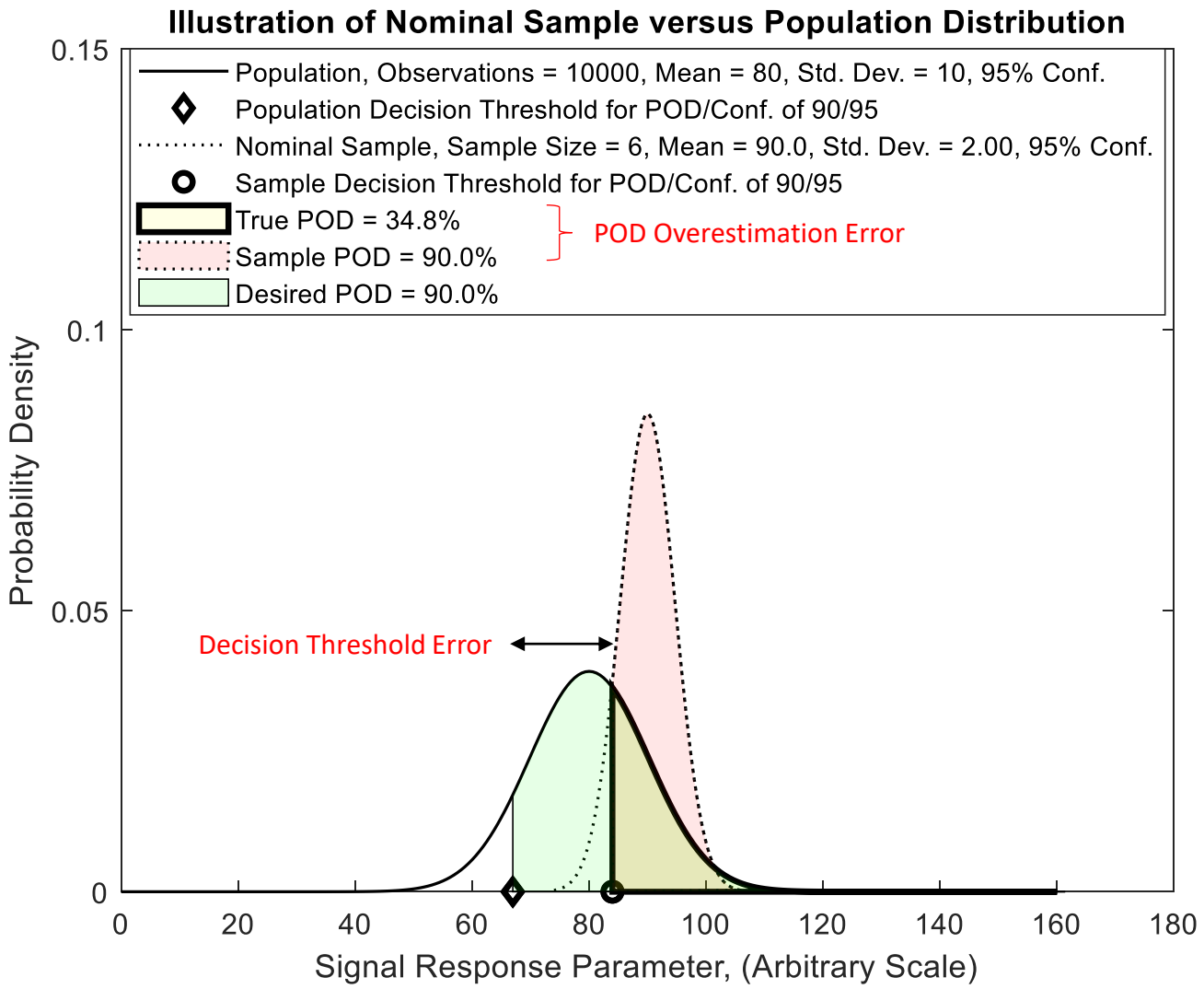


Illustration of sample overestimating POD

- Nominal sample has only high value readings above a probability partition due to nominal conditions (smooth flat surfaces from one alloy and low variability in flaw morphology) used in making specimens.
- This is an example of sample bias.
- LS POD assumes that the sample is representative of or random from the population
- LS POD results using nominal sample may not be acceptable based on sampling sensitivity analysis.

Illustration of Worst-Case Sample and Its Influence on POD Estimation - Concept only

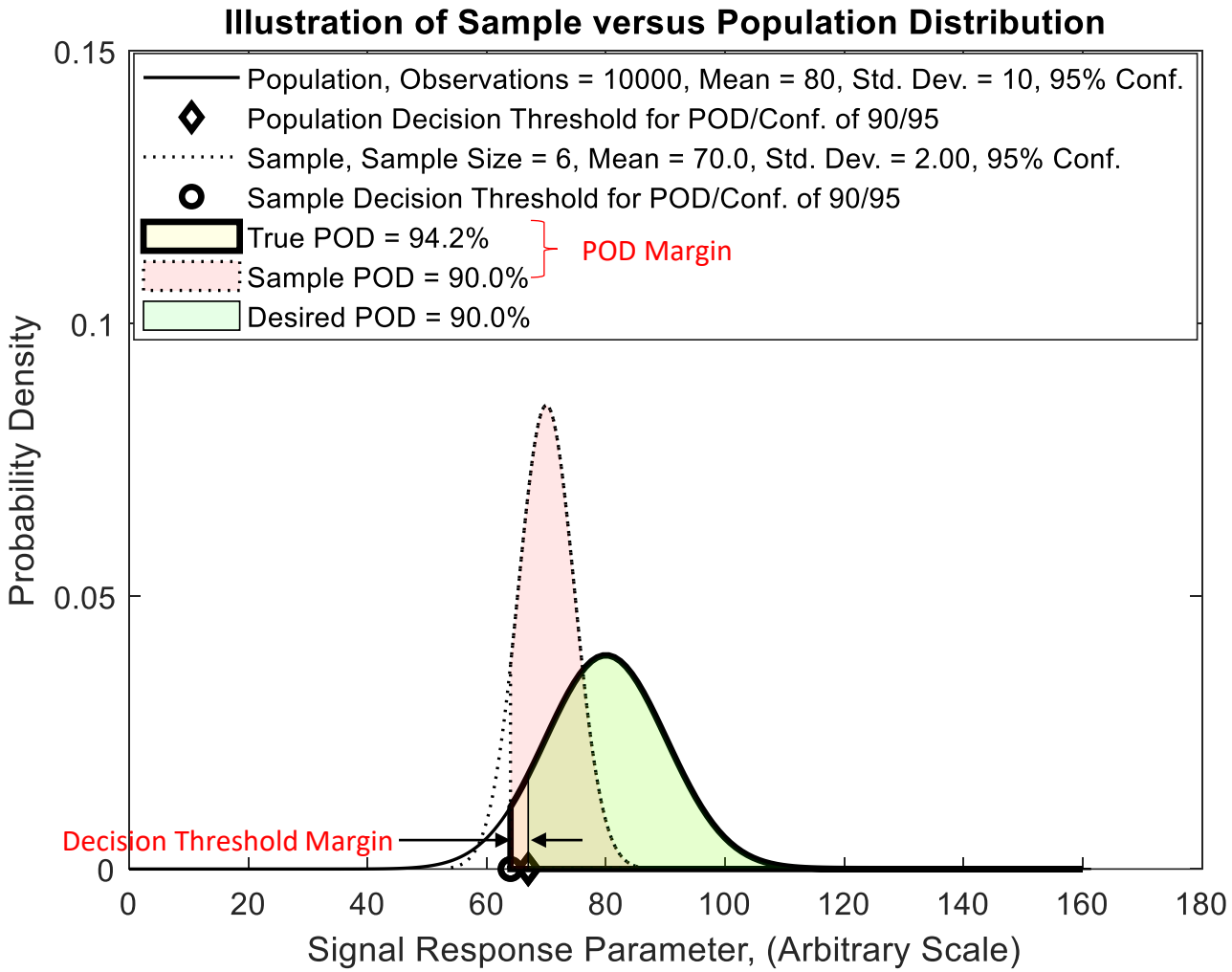
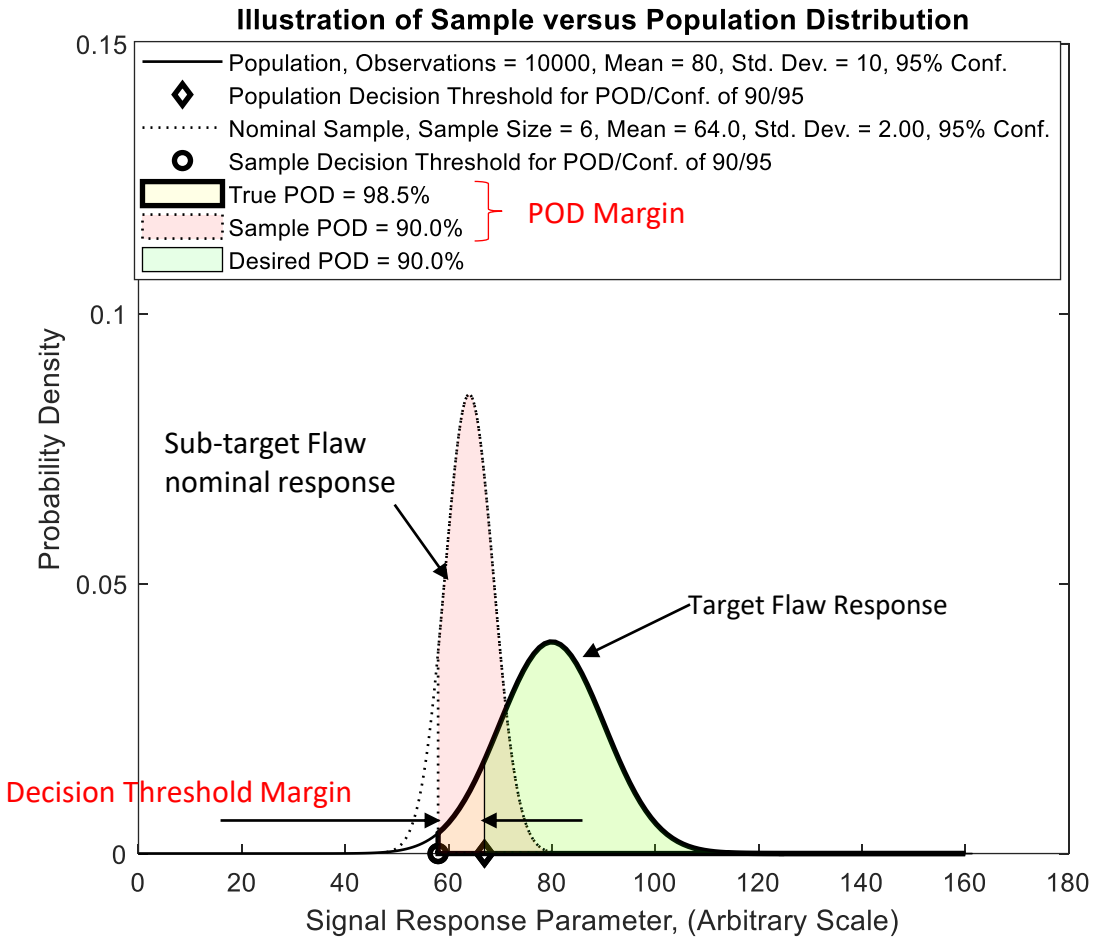


Illustration of Sample Underestimating POD

- Worst-Case sample has only low value readings below a probability partition due to Worst-Case conditions (rough curved surfaces, alloy with high noise, and tight gap flaw) used in making specimens
- LS POD assumes that the sample is representative of the population.
- This is an example of sample bias.
- LS POD results using Worst-Case sample may be acceptable based on sampling sensitivity analysis.

Illustration of Nominal Sub-target Flaw Response Sample - Concept only

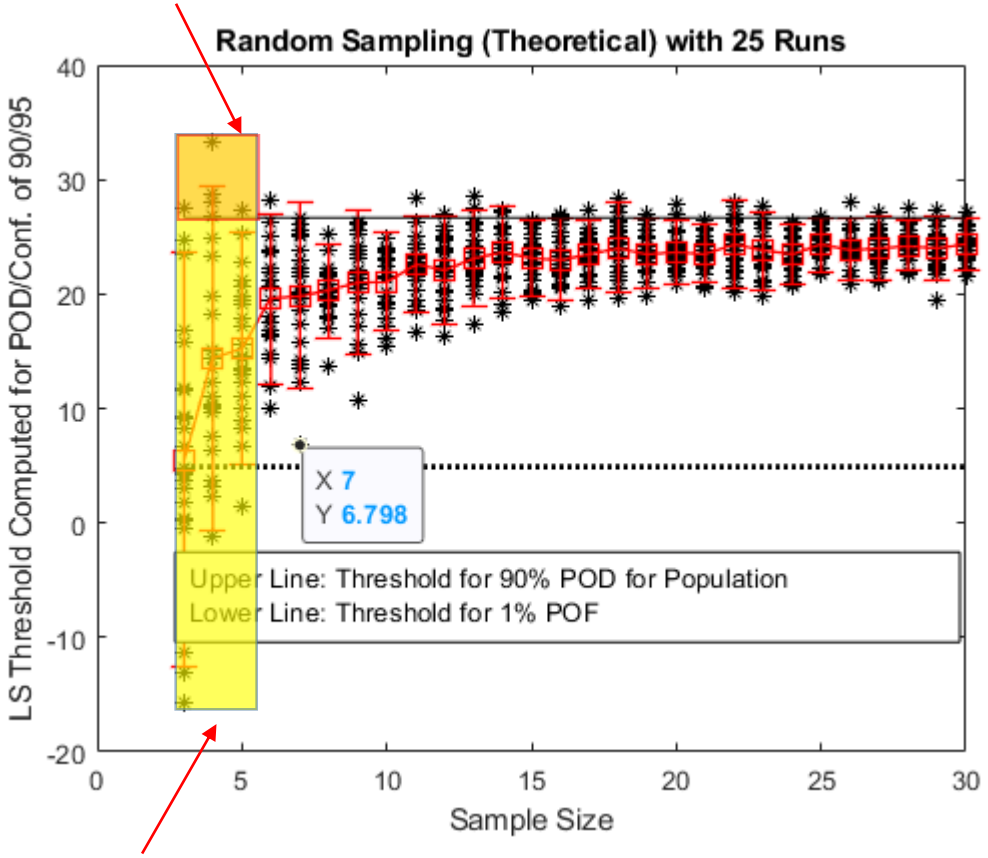


- Nominal sub-target sample has only low value readings compared to target signal responses for population
- LS POD results using Nominal sub-target sample may be acceptable based on sampling sensitivity analysis.

Illustration of Nominal Sub-target Sample Underestimating POD

D. Random Sampling – Sampling Sensitivity Analysis

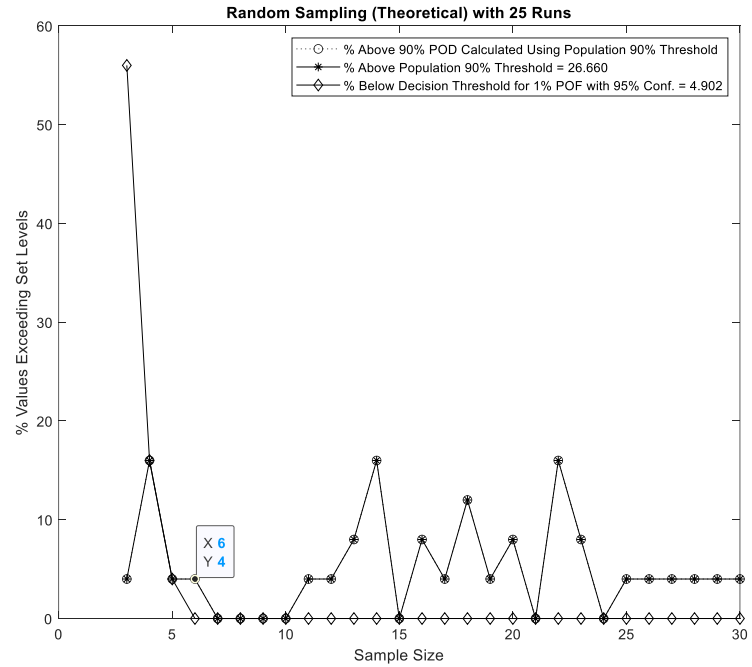
Higher sampling error in decision threshold or POD values for small sample size



Higher variability of decision threshold for small sample size

- About 5 % decision thresholds are not acceptable as they do not provide POD of 90%.

- Repetitive type D sampling (25 runs) indicates that
  - Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1%.
  - Sample size of 6 does not provide decision threshold > decision threshold at 90% POD for Population. Meets criteria for providing minimum POD/Conf. 90/95. 5 % decision thresholds are not acceptable, as they do not provide POD of 90.
  - Recommend using lower than calculated decision threshold to improve POD/Conf. as magnitude of error is relatively large.
- Both POD and POF criteria are met indicating that the validation is acceptable.



Sampling Types for used in Sampling Scheme Sensitivity Analysis

- Sampling Types used in Monte Carlo Simulation
- A. Nominal and Worst-Case Sampling for Target flaw
- B. Worst-Case Sampling for Target Size Flaw
- C. Nominal Case Sampling for Target Size Flaw (Special case of A or F)
- D. Random Sampling for Target Size Flaw (Theoretical) (Special case of E)
  - Used as a baseline for comparison
- E. Random Sampling for both Target and sub-target Size Flaws (Theoretical)
- F. Nominal Sampling for both Target and sub-target Size Flaws

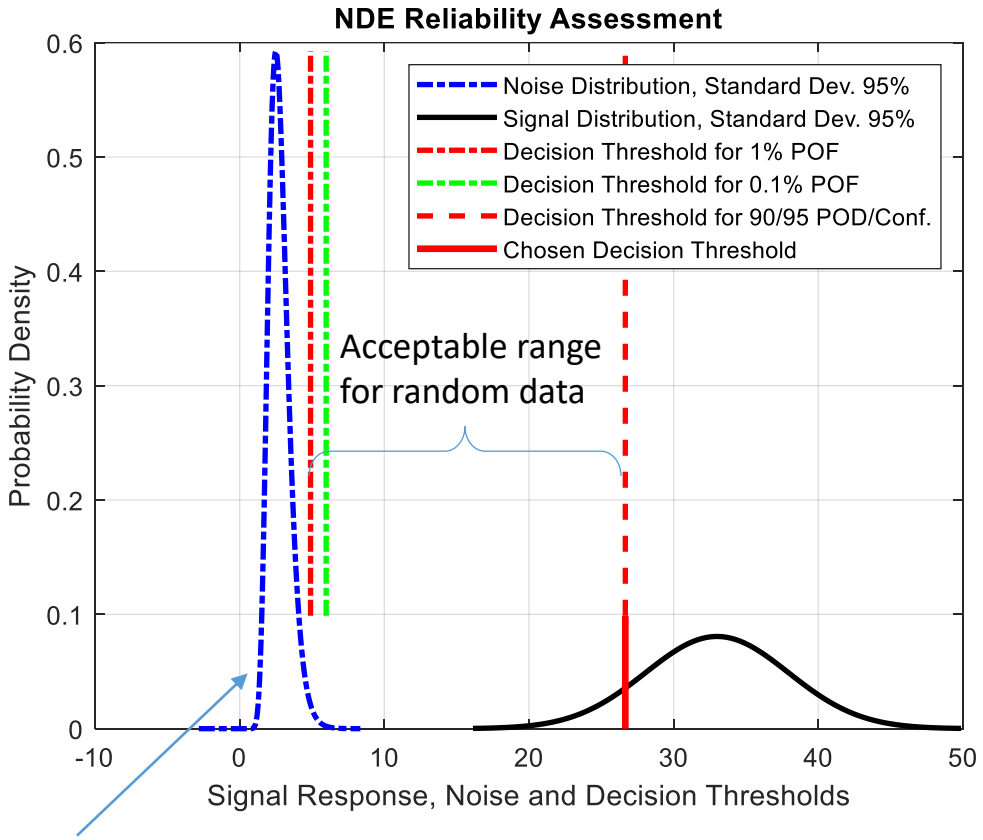
Assessing Sampling Schemes for LS POD for Mitigating Sampling Risk

- Type A - Nominal and Worst-Case Sampling
  - Probability split of nominal/Worst-Case between 50/50 to 70/30 is assumed to be reasonable.
  - This is most straight forward from analysis point of view as the goal here is to create representative sample.
  - When both nominal and worst-case values are represented in their probability of occurrence then a **representative sample** is created. LS POD k1-statistics works for a representative sample.
  - Measuring Worst-Case signal response values from real specimens may be challenging due to challenges in making the corresponding specimens.
- Type B - Worst-Case Sampling
  - Standard deviation of a Worst-Case sample is lower than that of population. The mean responses are lower. Monte Carlo analysis indicates that reliable LS POD may be validated.
  - Measuring Worst-Case signal response values from real specimens may be challenging due to challenges in making the corresponding specimens.

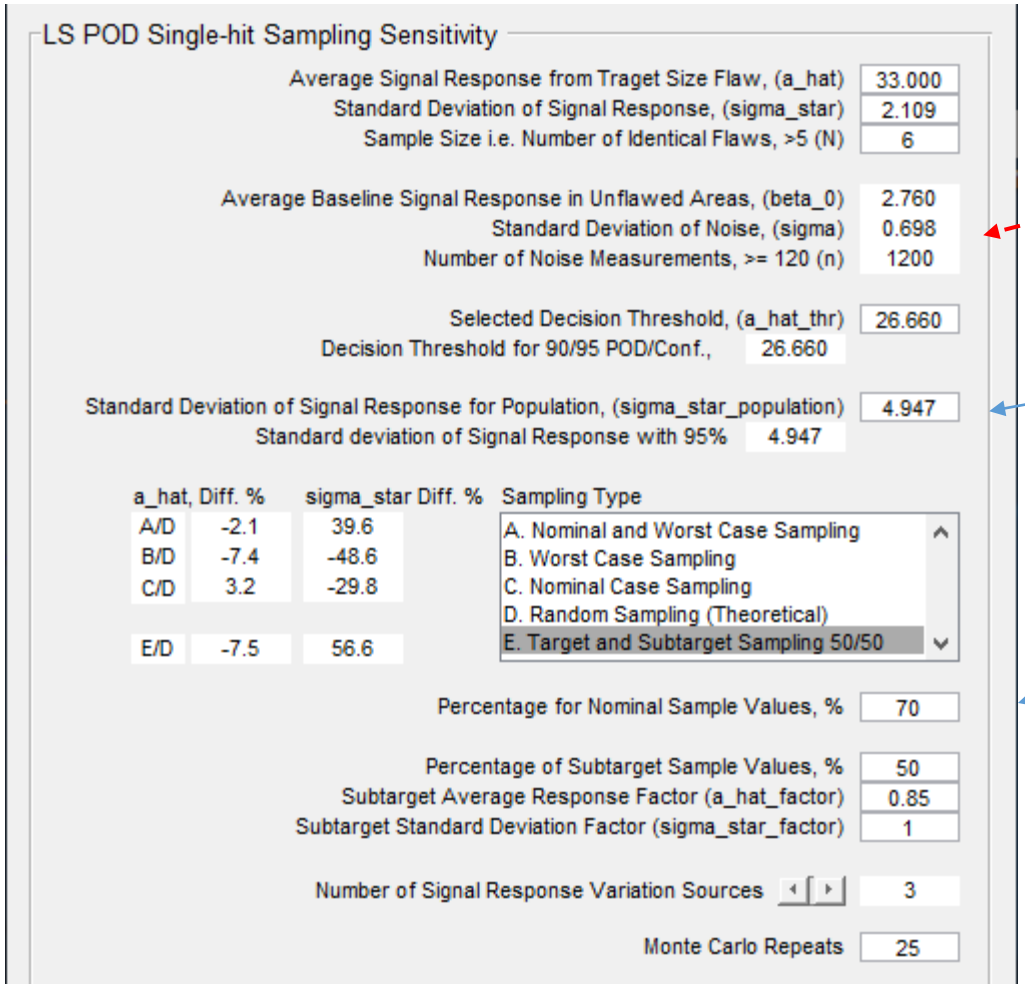
Assessing Sampling Schemes for LS POD for Mitigating Sampling Risk

- Type C - Target Flaw Nominal Sampling
  - Does not work and should be avoided
- Type D - Random Sample
  - It is not possible get a random sample for small sample size
- Type E - Random Target and Sub-target Sampling
  - This option is theoretical for a small sample size
- Type F - Nominal Sampling for both Target and sub-target Sampling
  - **Nominal Sampling for sub-target only is a special case of Type F sampling**
    - Sub-target flaw is a smaller size flaw and nominal sub-target samples are easy to make
    - Sampling size and analysis conservatism can be assessed through Monte Carlo Simulation
      - Standard deviation of a nominal sample is lower than that of representative sample. But it can be adjusted using nominal-random factor to be equivalent to the representative sample. The mean responses are lower. Monte Carlo analysis indicates that reliable LS POD may be validated.

LS POD Analysis of Selected Simulated Data, Sample Size = 6



Lognormal Noise
Minimum 95% confidence in fit



Measured noise

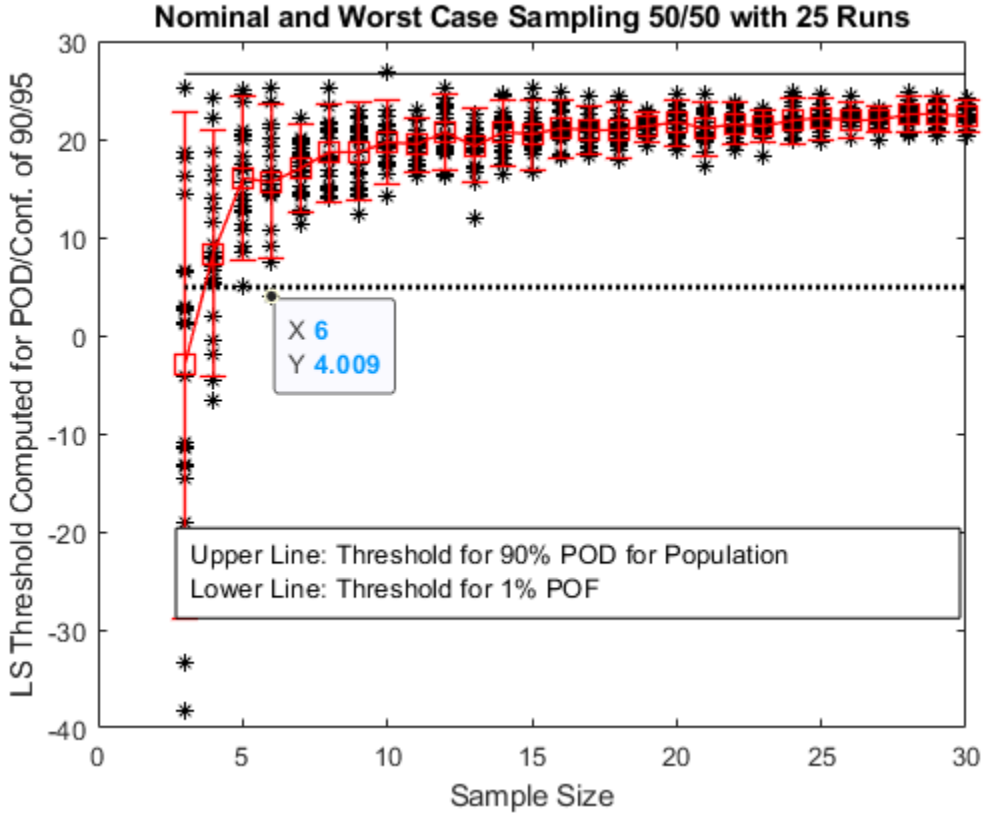
Assumes as truth for population with 95% conf. Conservative assumption

For Type A, B and C

For Type E

Note: Repeats are limited to 25 to save computing time but 500 or more are recommended.

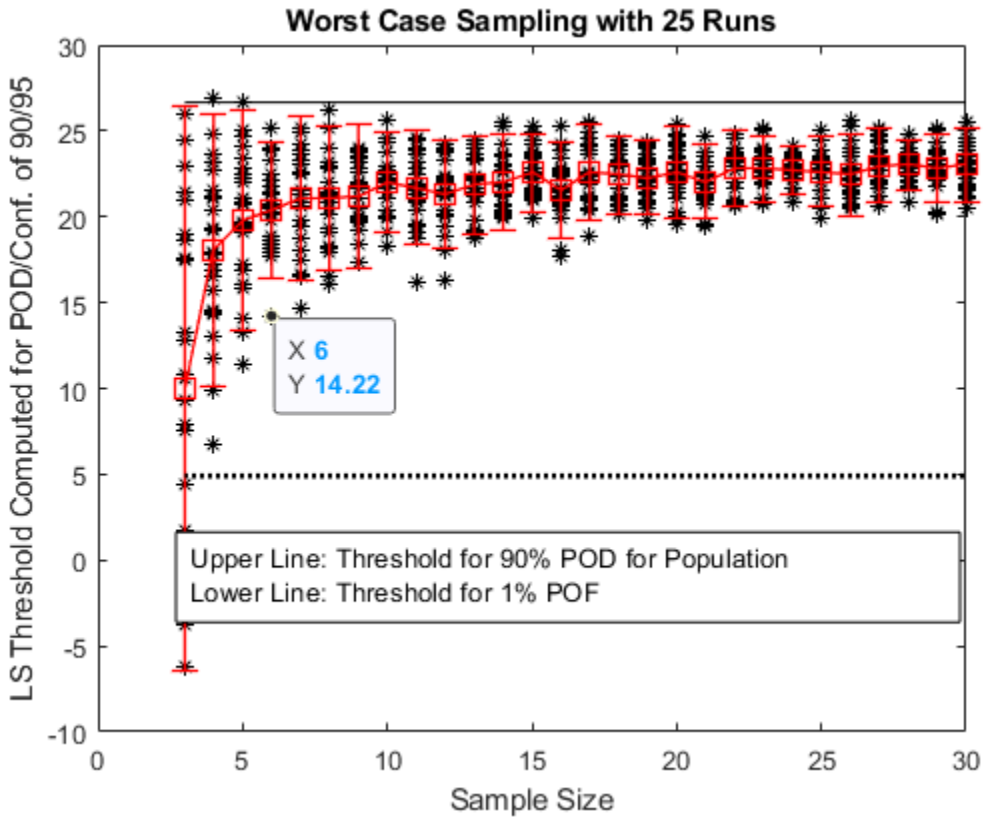
Type A - Nominal and Worst-Case Representative/Conservative Sampling



Red error bars are for 95% data (cumulative one-sided)

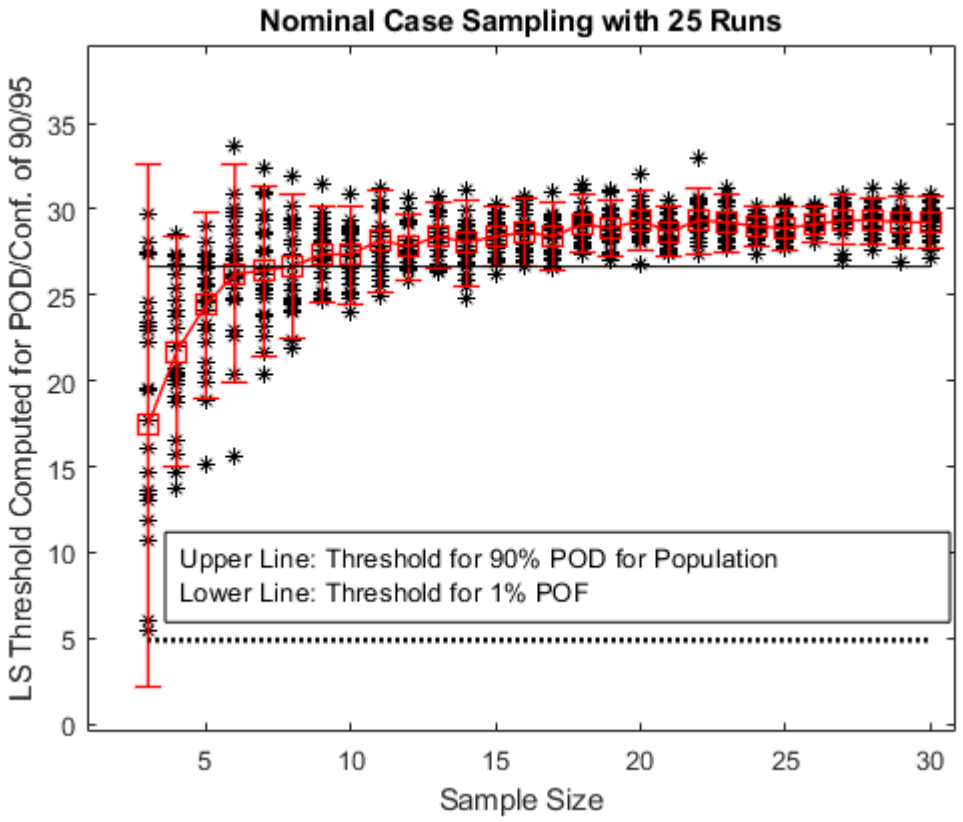
- Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1%.
- Sample size of 6 does not provide decision threshold > decision threshold at 90% POD for Population. Meets criteria for providing minimum POD/Conf. 90/95.
- Both POD and POF criteria are met indicating that the validation is robust.

Type B - Worst-Case Sampling

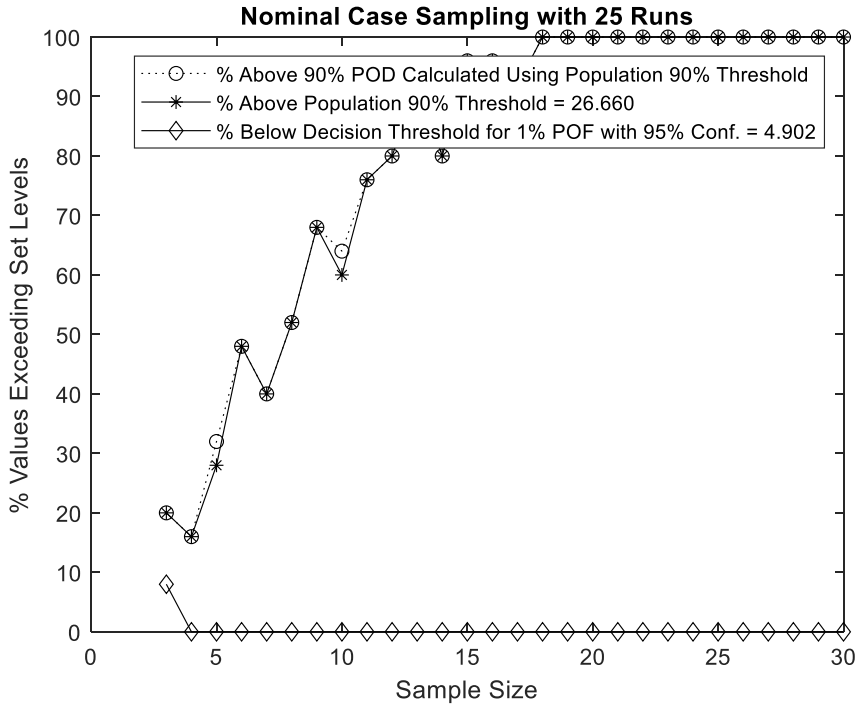


- Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1%.
- Sample size of 6 does not provide decision threshold > decision threshold at 90% POD for Population. Meets criteria for providing minimum POD/Conf. 90/95.
- Both POD and POF criteria are met indicating that the validation is robust.

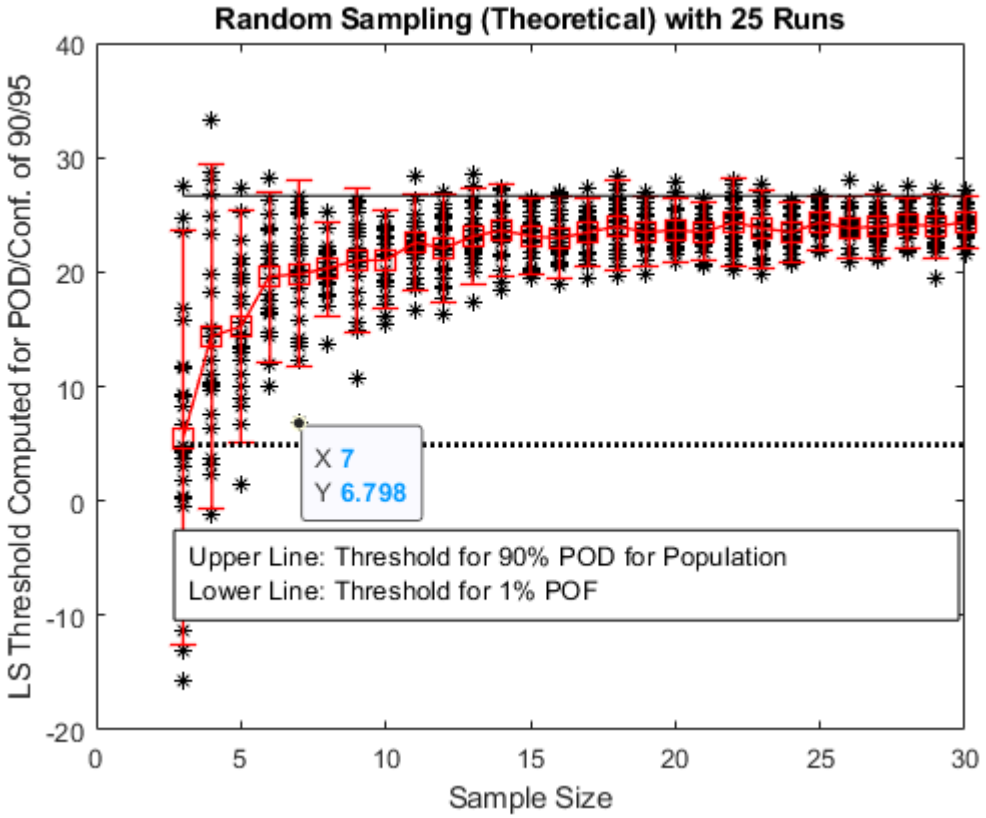
Type C - Nominal Value Sampling



- Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1%.
- Sample size of 6 does not provide decision threshold lower decision threshold at 90% POD for Population with 95% confidence. Does not meet criteria for providing minimum POD/Conf. 90/95.
- Indicates unacceptable validation.

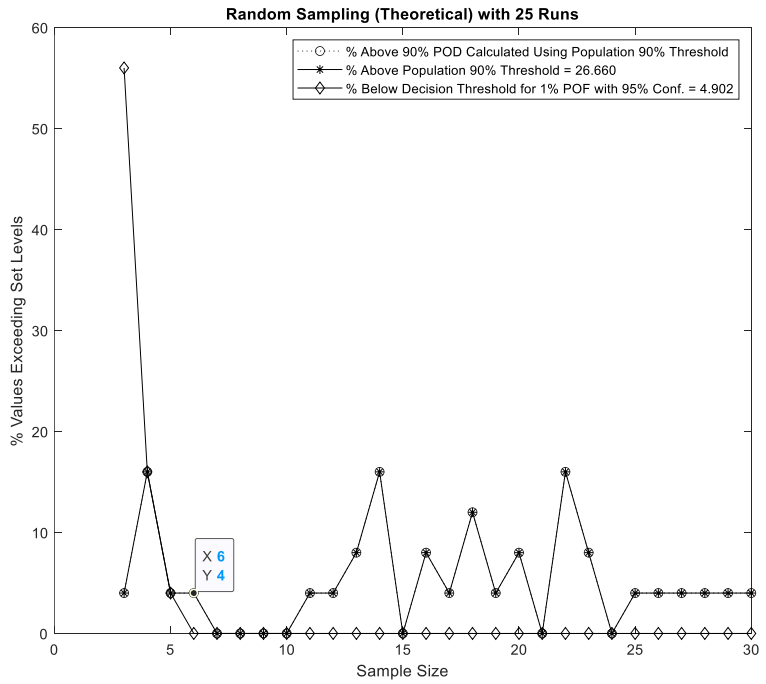


Type D - Random Sampling

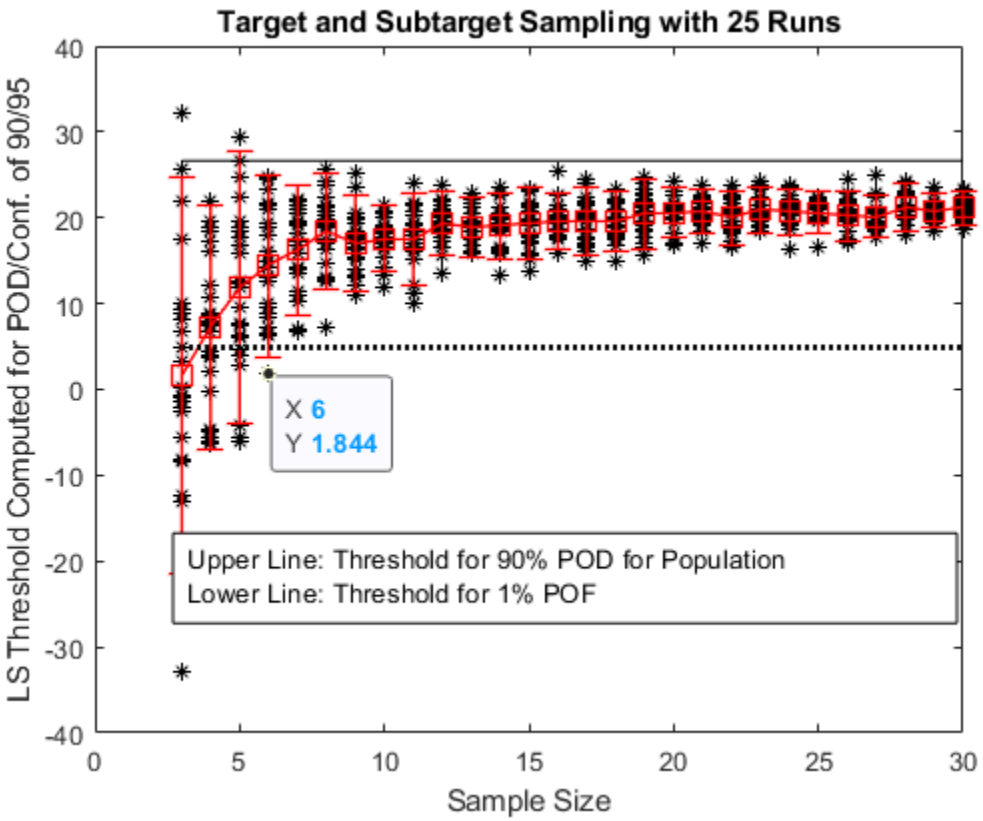


- About 5 % decision thresholds are not acceptable as they do not provide POD of 90%.

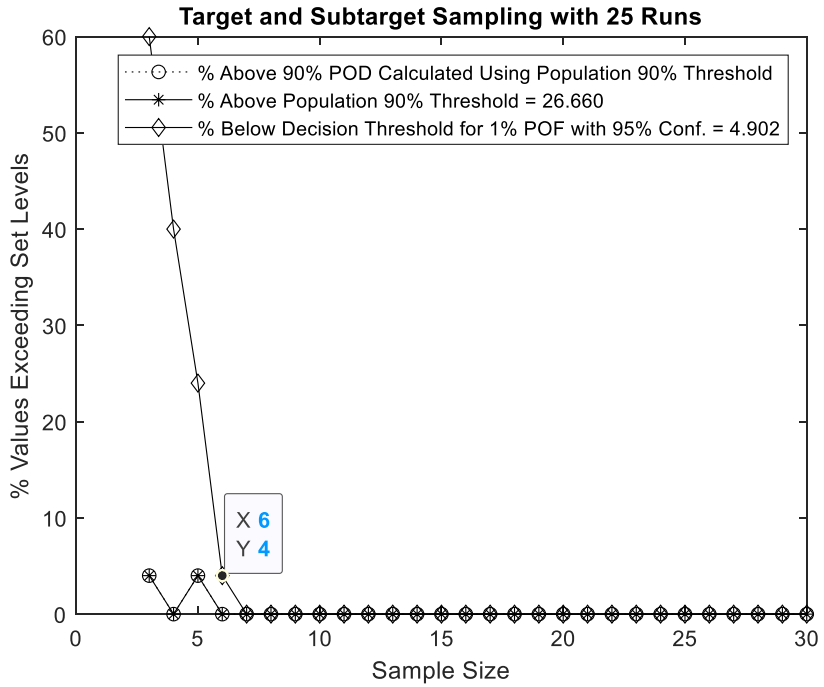
- Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1%.
- Sample size of 6 does not provide decision threshold > decision threshold at 90% POD for Population. Meets criteria for providing minimum POD/Conf. 90/95. 5 % decision thresholds are not acceptable as they do not provide POD of 90.
- Recommend using lower than calculated decision threshold to improve POD/Conf.
- Both POD and POF criteria are met indicating that the validation is acceptable.



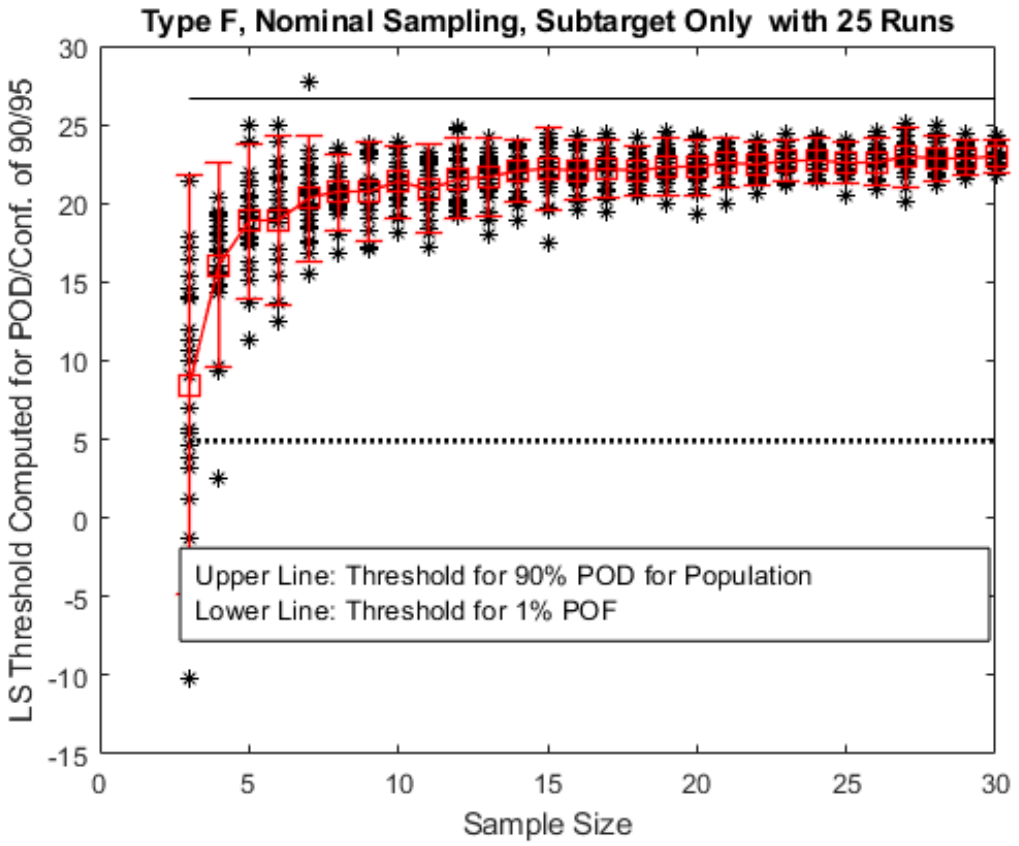
Type E - Target and Sub-target Random Sampling



- Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1% for .
- Sample size of 6 does not provide decision threshold > decision threshold at 90% POD for Population. Meets criteria for providing minimum POD/Conf. 90/95.
- Both POD and POF criteria are met indicating that the validation is robust.



Type F – Nominal Sampling of Sub-target Flaw only



- Sample size of 6 provides decision threshold ≥ 1% POF decision threshold. Meets the criterion for POF < 1% for .
- Sample size of 6 does not provide decision threshold > decision threshold at 90% POD for Population. Meets criteria for providing minimum POD/Conf. 90/95.
- Both POD and POF criteria are met indicating that the validation is robust.
- Although type F sampling does not provide a representative sample, it can provide a conservative sample that can be used for LS POD analysis
- Can be used successfully to create equivalent random sample properties for sub-target flaws
- Type F sampling can be designed to be conservative.
- Recommended as a lower risk option if it is not practical to produce Worst-Case signal responses

Observations from Sampling Runs

- Type A i.e. nominal-worst 50/50 split sampling
  - **Standard deviation of signal response measurement is most robust and conservative**
  - **Recommended as the lowest risk option if Worst-Case values can be measured**
- Type B i.e. Worst-Case sampling also may provide adequate decision threshold
  - May pose difficulty in measurement on a small sample
- Type D random sample
  - It is not possible get a random sample for small sample size
- Type C Target flaw Nominal sampling
  - Does not work and should be avoided
- Type E Random Target and sub-target Sampling
  - Has benefits of type F but is not recommended because it is not possible get a random sample for small sample size
- Type F i.e. Target and sub-target nominal sampling is more practical to all above types
  - **Although type F sampling does not provide a representative sample, it can provide a conservative sample that can be used for LS POD analysis**
  - **Recommended as a lower risk option if it is not practical to produce Worst-Case signal responses i.e. Type A**

Conclusions

- Sampling sensitivity analysis can be used to assess sampling risk in LS POD results and to validate a sampling scheme
- Type A, i.e. Nominal and Worst-Case target flaw sampling can create representative sample, which can be directly used in LS POD analysis
- Type F, i.e. Target and sub-target nominal sampling is more practical than Type A sampling
  - Although, type F sampling does not provide a representative sample, it can provide a conservative sample that can be used for LS POD analysis
  - Type F is recommended as a lower risk option, if it is not practical to produce worst case signal responses needed in type A option