Designation: E3023 – 15

Standard Practice for Probability of Detection Analysis for $\hat{a}$ Versus a Data[1]

This standard is issued under the fixed designation E3023; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope

1.1 This practice defines the procedure for performing a statistical analysis on Nondestructive Testing (NDT) $\hat{a}$ versus a data to determine the demonstrated probability of detection (POD) for a specific set of examination parameters. Topics covered include the standard $\hat{a}$ versus a regression methodology, POD curve formulation, validation techniques, and correct interpretation of results.

1.2 The values stated in inch-pound units are to be regarded as standard. The values given in parentheses are mathematical conversions to SI units that are provided for information only and are not considered standard.

1.3 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E178 Practice for Dealing With Outlying Observations
E456 Terminology Relating to Quality and Statistics
E1316 Terminology for Nondestructive Examinations
E1325 Terminology Relating to Design of Experiments
E2586 Practice for Calculating and Using Basic Statistics
E2782 Guide for Measurement Systems Analysis (MSA)
E2862 Practice for Probability of Detection Analysis for Hit/Miss Data

2.2 *Department of Defense Document:*[3]

MIL-HDBK-1823A Nondestructive Evaluation System Reliability Assessment

3. Terminology

3.1 *Definitions of Terms Specific to This Standard:*

3.1.1 *analyst, n*—the person responsible for performing a POD analysis on $\hat{a}$ versus a data resulting from a POD examination.

3.1.2 *decision threshold, $\hat{a}_{dec}$, n*—the value of $\hat{a}$ above which the signal is interpreted as a find and below which the signal is interpreted as a miss.

3.1.2.1 *Discussion*—A decision threshold is required to create a POD curve. The decision threshold is always greater than or equal to the noise threshold and is the value of $\hat{a}$ that corresponds with the flaw size that can be detected with 50% POD.

3.1.3 *demonstrated probability of detection, n*—the calculated POD value resulting from the statistical analysis of the $\hat{a}$ versus a data.

3.1.4 *false call, n*—– the perceived detection of a discontinuity that is identified as a find during a POD examination when no discontinuity actually exists at the inspection site.

3.1.5 *noise, n*—signal response containing no useful target characterization information.

3.1.6 *noise threshold, $\hat{a}_{noise}$, n*—the value of $\hat{a}$ below which the signal is indistinguishable from noise.

3.1.6.1 *Discussion*—The noise threshold is always less than or equal to the decision threshold. The noise threshold is used to determine left censored data.

3.1.7 *probability of detection, n*—the fraction of nominal discontinuity sizes expected to be found given their existence.

3.1.8 *saturation threshold, $\hat{a}_{sat}$, n*—the value of $\hat{a}$ associated with the maximum output of the system or the largest value of $\hat{a}$ that the system can record.

3.1.8.1 *Discussion*—The saturation threshold is used to determine right censored data.

3.2 *Symbols:*

3.2.1 a—discontinuity size.

3.2.2 $\hat{a}$—the measured signal response for a given discontinuity size, a.

3.2.2.1 *Discussion*—The measured signal response is assumed to be continuous in nature. Units depend on the NDT

[1] This test method is under the jurisdiction of ASTM Committee E07 on Nondestructive Testing and is the direct responsibility of Subcommittee E07.10 on Specialized NDT Methods.

Current edition approved June 15, 2015. Published August 2015. DOI: 10.1520/E3023–15.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

[3] Available from Standardization Documents Order Desk, DODSSP, Bldg. 4, Section D, 700 Robbins Ave., Philadelphia, PA 19111-5098, http://dodssp.daps.dla.mil.

inspection system and can be, for example, scale divisions, number of contiguous illuminated pixels, or millivolts.

3.2.3 a_p—the discontinuity size that can be detected with probability p.

3.2.3.1 *Discussion*—Each discontinuity size has an independent probability of being detected and corresponding probability of being missed. For example, being able to detect a specific discontinuity size with probability p does not guarantee that a larger size discontinuity will be found.

3.2.4 $a_{p/c}$—the discontinuity size that can be detected with probability p with a statistical confidence level of c.

3.2.4.1 *Discussion*—According to the formula in MIL-HDBK-1823A, $a_{p/c}$ is a one-sided upper confidence bound on a_p. $a_{p/c}$ represents how large the true a_p could be given the statistical uncertainty associated with limited sample data. Hence $a_{p/c} > a_p$. Note that POD is equal to p for both $a_{p/c}$ and a_p. a_p is based solely on the observed relationship between the $\hat{a}$ and a data and represents a snapshot in time, whereas $a_{p/c}$ accounts for the uncertainty associated with limited sample data.

4. Summary of Practice

4.1 This practice describes, step-by-step, the process for analyzing nondestructive testing $\hat{a}$ versus a data resulting from a POD examination, including minimum requirements for validating the resulting POD curve.

4.2 This practice also includes definitions and discussions for results of interest (e.g., $a_{90/95}$) to provide for correct interpretation of results.

4.3 Definitions of statistical terminology used in the body of this practice can be found in Annex A1.

5. Significance and Use

5.1 The POD analysis method described herein is based on well-known and well-established statistical methods. It shall be used to quantify the demonstrated POD for a specific set of examination parameters and known range of discontinuity sizes under the following conditions.

5.1.1 The initial response from a nondestructive evaluation inspection system is measurable and can be classified as a continuous variable.

5.1.2 The relationship between discontinuity size (a) and measured signal response ($\hat{a}$) exists and is best described by a linear regression model with an error structure that is normally distributed with mean zero and constant variance, σ^2. (Note that "linear" refers to linear with respect to the model coefficients. For example, a quadratic model $\hat{y} = \beta_0 + \beta_1 \cdot x + \beta_2 \cdot x^2$ is a linear model.)

5.2 This practice does not limit the use of other statistical models if justified as more appropriate for the $\hat{a}$ versus a data.

5.3 This practice is not appropriate for data resulting from a POD examination on nondestructive evaluation systems that generate an initial response that is binary in nature (for example, hit/miss). Practice E2862 is appropriate for systems that generate a hit/miss-type response (for example, fluorescent penetrant).

5.4 Prior to performing the analysis it is assumed that the discontinuity of interest is clearly defined; the number and distribution of induced discontinuity sizes in the POD specimen set is known and well documented; the POD examination administration procedure (including data collection method) is well designed, well defined, under control, and unbiased; the initial inspection system response is measurable and continuous in nature; the inspection system is calibrated; and the measurement error has been evaluated and deemed acceptable. The analysis results are only valid if the $\hat{a}$ versus a data are accurate and precise and the linear model adequately represents the $\hat{a}$ versus a data.

5.5 The POD analysis method described herein is consistent with the analysis method for $\hat{a}$ versus a data described in MIL-HDBK-1823A and is included in several widely utilized POD software packages to perform a POD analysis on $\hat{a}$ versus a data. It is also found in statistical software packages that have linear regression capability. This practice requires that the analyst has access to either POD software or other software with linear regression capability.

6. Procedure

6.1 The POD analysis objective shall be clearly defined by the responsible engineer or by the customer.

6.2 The analyst shall obtain the $\hat{a}$ versus a data resulting from the POD examination, which shall include at a minimum the documented known induced discontinuity sizes, the associated measured signal response, and any false calls.

6.3 The analyst shall also obtain specific information about the POD examination, which shall include at a minimum the specimen standard geometry (e.g., flat panels), specimen standard material (e.g., Nickel), examination date, number of inspectors, type of inspection method (e.g., Eddy Current Inspection), pertinent information about the instrument and instructions for use (e.g., settings, probe type, scan path), and pertinent comments from the inspector(s) and test administrator.

6.3.1 In general, the results of an experiment apply to the conditions under which the experiment was conducted. Hence, the POD analysis results apply to the conditions under which the POD examination was conducted.

6.4 Prior to performing the analysis, the analyst shall conduct a preliminary review of the POD examination procedure to identify any issues with the administration of the examination. The analyst shall identify and attempt to resolve any issues prior to conducting the POD analysis. Identified issues and their resolution shall be documented in the report. Examples of examination administration issues and possible resolutions are outlined in the following subsections.

6.4.1 If problems or interruptions occurred during the POD examination that may bias the results, the POD examination should be re-administered.

6.4.2 If the examination procedure was poorly designed and/or executed, the validity of the resulting data is questionable. In this case, the examination procedure design and execution should be reevaluated. For design guidelines see MIL-HDBK-1823A.

6.5 Prior to performing the analysis, the analyst shall conduct a preliminary review of the $\hat{a}$ versus a data to identify any data issues. The analyst shall identify and attempt to resolve any issues prior to conducting the POD analysis. Identified issues and their resolution shall be documented in the report. Examples of data issues and possible resolutions are outlined in the following subsections.

6.5.1 Any apparent outlying observations shall be reviewed for correctness. If a typo is identified, the typo shall be corrected prior to performing the analysis. If the value is correct, it shall be retained in the analysis and its influence on the $\hat{a}$ versus a model shall be evaluated during the model diagnostic assessment. The analyst should also reference Practice E178.

6.5.2 POD cannot be modeled as a continuous function of discontinuity size if all the measured signal responses are below the noise threshold or above the saturation threshold. If this occurs, the adequacy of the nondestructive testing system should be evaluated.

6.6 Only â versus a data for induced discontinuities shall be used in the development of the linear regression model. False call data shall not be included in the development of the linear model when using standard linear regression methods.

6.7 The analyst in conjunction with the responsible engineer shall determine the value of the noise threshold, $\hat{a}_{noise}$, and saturation threshold, $\hat{a}_{sat}$, prior to performing the analysis.

6.7.1 The value of $\hat{a}_{noise}$ is determined by performing a noise analysis. A noise analysis is typically accomplished by assessing the distribution of measured signal responses from sites with no known discontinuity (false calls) and/or measured signal responses that are not influenced by the size of the discontinuities (noise). Details on performing a noise analysis can be found in MIL-HDBK-1823A.

6.8 The analyst shall select an appropriate linear regression model to establish the relationship between $\hat{a}$ and a. Selection of a linear model may be an iterative process as the significance of the predictor variable(s) and the appropriateness of the selected model are typically assessed after the model has been developed.

6.8.1 "Linear" refers to linear with respect to the model coefficients. For example, $\hat{y}_i = b_0 + b_1 \cdot (x^2)$ and $\hat{y}_i = b_0 + b_1 \cdot x_1 + b_2 \cdot \ln(x_2)$ are linear regression models.

6.8.2 In general, only significant and uncorrelated predictor variables are included in a regression model. If more than one predictor variable is being considered for inclusion in the model, a preliminary graphical analysis of the response variable against each predictor variable may help identify which predictor variables appear to influence the response and the type of relationship (for example, direct, inverse, quadratic). In addition, a preliminary graphical analysis of all possible pairings of predictor variables shall be performed to verify independence of the predictor variables. When plotted against each other, there should be no apparent relationship between any two predictor variables.

6.8.3 The appropriateness of a selected model is determined by how well the model fits the observed data and how well the underlying regression assumptions are met.

6.9 The analyst shall use software that has the appropriate linear regression capabilities to perform a linear regression analysis on the $\hat{a}$ versus a data.

6.9.1 If censored data are present, the analyst shall do the following:

6.9.1.1 Include and identify the censored data in the analysis (according to the notation required by the software).

6.9.1.2 Use the method of maximum likelihood to estimate the model coefficients.

6.9.1.3 Verify that convergence was achieved. If convergence is not achieved, the resulting $\hat{a}$ versus a model shall not be used to develop a POD curve.

6.9.1.4 Check the number of iterations it took to converge provided that information on convergence and the number of iterations it took to converge is included in the analysis software output. If more than twenty iterations were needed to reach convergence, the model may not be reliable.

6.9.1.5 Include a statement in the report indicating that convergence was achieved and the number of iterations needed to achieve convergence.

6.9.2 If no censored data are present, the method of maximum likelihood or the method of least squares shall be used.

6.10 If included in the analysis software output, the analyst shall assess the significance of the predictor variables in the model. Only significant predictor variables should be included in the model.

6.11 Once the $\hat{a}$ versus a model is estimated, the analyst shall use, at a minimum, the model diagnostic methods listed below to assess the underlying linear regression assumptions. The methods listed below shall be performed using only non-censored data. If available, other formal diagnostic methods (noted in Appendix X1) should be used to assess the linear regression assumptions.

6.11.1 There are three main underlying assumptions in a linear regression analysis: *(1)* residuals are normally distributed with mean 0 and constant variance, σ^2, *(2)* the residuals are independent, and *(3)* the relationship is in fact linear. The residual is calculated as $e_i = y_i - \hat{y}_i$ and represents the difference between the observed result, y_i, and the predicted value, $\hat{y}_i$, for the i^{th} case. In general, the results of a linear regression analysis are not valid unless these assumptions hold. At a minimum, the following analyses of the residuals shall be performed to verify the assumptions.

6.11.1.1 A histogram of the residuals shall be constructed to assess the normality assumption and centering of the residuals. A histogram of the residuals should be roughly bell-shaped and symmetric around zero. In general, bell shape and symmetry around zero are more important than strict normality since traditional estimation procedures are typically only sensitive to large departures from normality (particularly with respect to skewness).

6.11.1.2 The constant variance and linearity assumptions shall be verified by plotting the residuals (y-axis) against the predicted values (x-axis). If the residuals fall in a horizontal band centered around zero, with no systematic preference for being positive or negative, then the assumption of constant variance and a linear relationship holds. (See Fig. X1.2 in

Appendix X1.) In general, meeting the constant variance assumption is more important than meeting the normality assumption.

6.12 The analyst shall use at a minimum the methods listed below to assess the goodness-of-fit, influential points, and multicollinarity among predictor variables. If available, more formal methods (noted in Appendix X1) should be used.

6.12.1 A plot of predicted values versus actual values shall be used to assess goodness-of-fit. The plotted points should fall roughly on the $y = x$ line. Plotted points deviating from the $y = x$ line in a systematic way may be an indication of poor fit.

6.12.2 The analyst shall assess the influence of data that appears to be outlying on the established $\hat{a}$ versus a model. The histogram of the residuals and plot of the residuals versus predicted values can help identify outlying values. The influence of a suspected outlying value shall at a minimum be evaluated by removing the outlying value from the data and re-running the analysis to assess its influence on the $\hat{a}$ versus a model. A data point is said to be influential (or have high leverage) if its exclusion from the analysis has a relatively large effect on the $\hat{a}$ versus a model. Both analysis results (with and without the outlying data) shall be included in the report along with a discussion of the impact to the resulting POD curve and confidence bound (if applicable).

6.12.3 If the model includes more than one predictor variable, a graphical analysis shall be performed to verify independence of the predictor variables. (This step may be done during model selection as described in Appendix X1.)

6.13 The responsible engineer shall determine the value of $\hat{a}_{dec}$ that is most appropriate with respect to end use of the POD analysis results. A value for the decision threshold is required to create a POD curve. The value must be greater than or equal to the value of the noise threshold. That is, $\hat{a}_{dec} \geq \hat{a}_{noise}$.

6.14 The analyst shall use the decision threshold to determine a POD value for each discontinuity size given the established relationship between $\hat{a}$ and a, the formula for which can be found in Appendix X1. The resulting POD values shall be plotted against discontinuity size to produce a POD Curve.

6.14.1 POD curves tend to be *s*-shaped when a simple linear regression model is selected.

6.14.2 If more than one predictor variable is included in the model, POD is a response surface rather than a single curve.

6.14.3 The analyst shall determine the most appropriate way to plot the results.

6.15 If a c% level of confidence is specified by the responsible engineer or the customer, the analyst shall put a c% lower confidence bound on the POD curve by calculating a c% lower confidence bound on the $\hat{a}$ versus a model fit. Methods for constructing a confidence bound around a regression fit can be found in MIL-HDBK-1823A as well as statistics text books on linear regression.[4]

6.15.1 If, for example, the objective of the analysis is to determine the discontinuity size that can be detected with 90% probability and 95% confidence, denoted $a_{90/95}$, then the analyst shall put a 95% lower confidence bound on the POD curve by calculating a 95% lower confidence bound on the $\hat{a}$ versus a model fit. The formula for the 95% lower confidence bound on the POD curve, which is based on the 95% lower confidence bound around the regression fit, can be found in Appendix X1.

6.16 The analyst shall analyze any false call data and shall report the false call rate.

6.16.1 The responsible engineer or the customer shall clearly define what constitutes a false call.

6.16.2 A distributional analysis of false call or noise data, or both, is typically performed to assess the false call rate, a discussion of which can be found in MIL-HDBK-1823A.

6.17 Acceptable false call rates shall be determined by the responsible engineer or by the customer.

7. Report

7.1 At a minimum the following information about the POD analysis shall be included in the report.

7.1.1 The specimen standard geometry (e.g., flat panels).

7.1.2 The specimen standard material (e.g., nickel).

7.1.3 Examination date.

7.1.4 Number of inspectors.

7.1.5 Type of inspection (e.g., Eddy Current).

7.1.6 Pertinent information about the instrument and instructions for use (e.g., settings, probe type, scan path).

7.1.7 Any comments from the inspector(s) or test administrator.

7.1.8 The documented known induced discontinuity sizes.

7.1.9 The associated measured signal responses, including information about censored data.

7.1.10 Any false calls.

7.1.11 The linear regression model describing the relationship between the observed $\hat{a}$ versus a data and confidence bound (if applicable).

7.1.12 A statement indicating that convergence was achieved and the number of iterations to convergence, if maximum likelihood estimation was used.

7.1.13 A statement about the model diagnostic methods used and conclusions.

7.1.14 The estimate of the error around the regression fit (calculated as the square root of the mean square error, which is typically included in the software output).

7.1.15 Summary of the noise analysis and rationale for selection of the decision threshold.

7.1.16 A plot of the resulting POD curve and confidence bound (if applicable).

7.1.17 Specific results of interest as required by the analysis objective (e.g., $a_{90/95}$).

7.1.18 Any deviations from the POD examination procedure or standard POD analysis.

7.1.18.1 If the POD examination was re-administered, the original results and rationale for re-administration shall be documented in the report.

7.1.18.2 If a discontinuity is removed from the analysis, the specific discontinuity and rationale for removal shall be documented in the final report.

[4] Neter, J, Kutner, M, Nachtsheim, C, Wasserman, W. *Applied Linear Statistical Models*, The McGraw-Hill Companies.

7.1.18.3 If the impact of outlying data was assessed, the results shall be included in the report along with an explanation.

7.1.19 Summary of false call analysis, including a definition of what constitutes a false call, the false call rate, and the method used to estimate the false call rate.

7.1.20 Name of analyst and company responsible for the POD calculation.

8. Keywords

8.1 a-hat vs. a; ahat vs. a; eddy current inspection; eddy current POD; Linear Regression; POD; POD analysis; probability of detection; regression

ANNEX

(Mandatory Information)

A1. TERMINOLOGY

A1.1 Definitions:

A1.1.1 a_{90}—the discontinuity size that can be detected with 90% probability.

A1.1.1.1 *Discussion*—The value for a_{90} resulting from a POD analysis is a single point estimate of the true value based on the outcome of the POD examination. It represents the typical value and does not account for variability due to sampling or inherent variability in the inspection system, which is always present.

A1.1.2 $a_{90/95}$—the discontinuity size that can be detected with 90% probability with a statistical confidence level of 95%.

A1.1.2.1 *Discussion*—The value for a_{90} resulting from a POD analysis is an estimate of the true a_{90} based on the outcome of the POD examination. If the examination were repeated, the outcome is not expected to be exactly the same. Hence the estimate of a_{90} will not be the same. To account for variability due to sampling, a statistical confidence bound with a 95% level of confidence is often applied to the estimated value for a_{90}, resulting in an $a_{90/95}$ value. POD is still 90%. The 95% refers to the ability of the statistical method to capture (or bound) the true a_{90}. That is, if the examination were repeated over and over under the same conditions, the value for $a_{90/95}$ will be larger than the true a_{90} 95% of the time. In practice the POD examination will be conducted once. Using a 95% confidence level implies a 95% chance that the $a_{90/95}$ value bounds the true a_{90} and a 5% risk that the true a_{90} is actually larger than the $a_{90/95}$ value.

A1.1.3 $a_{90/50}$—the discontinuity size that can be detected with 90% probability with a statistical confidence level of 50%.

A1.1.3.1 *Discussion*—Using a one-sided 50% confidence bound implies a 50% chance that the $a_{90/50}$ value bounds the true a_{90} and a 50% risk that the true a_{90} is actually larger than the $a_{90/50}$ value. Given this, $a_{90/50}$ is really the same as a_{90}.

A1.1.4 *censored data, n* —A censored data point is one in which the value is not known exactly.

A1.1.4.1 *Discussion*—The two most common types of censoring encountered in an $\hat{a}$ versus a POD analysis are right-censored and left-censored. A right-censored data point is one in which there is a lower bound y_i for the ith response. That is, the exact response value is somewhere in the interval (y_i, ∞). A left-censored data point is one in which there is an upper bound y_i for the ith response. That is, the exact response value is somewhere in the interval $(-\infty, y_i]$. In practice, right-censoring occurs when the signal generated by a large flaw saturates the system. For example, suppose that the maximum amplitude that can be reported by an inspection system is 25. The underlying assumption is that the measured signal increases as flaw size increases. If the measured signal from a large flaw exceeds 25, the response for that flaw is $(25, \infty)$. In other words, the exact measured signal is some amplitude to the "right" of 25. Note that the censoring in this case is predetermined by the limitations of the instrument electronics. Right-censored data is identified in an $\hat{a}$ versus a POD analysis by the saturation threshold. Left-censoring occurs in practice when the inspection system cannot distinguish the signal generated by a small flaw from inherent system noise or material noise, or both. For example, suppose that the noise threshold is 1 division. That is, any signal below 1 division is indistinguishable from noise. If the measured signal from a small flaw falls below 1, the response for that flaw is recorded as (0,1). In other words, the exact measured signal is some amplitude to the "left" of 1, or within the noise. Note that the censoring in this case is predetermined by inherent noise in the inspection system. Left-censored data is identified in an $\hat{a}$ versus a POD analysis by the noise threshold.

A1.1.5 *histogram, n*—graphical representation of the frequency distribution of a characteristic consisting of a set of rectangles with area proportional to the frequency. **Terminology E456, Practice E2586**

A1.1.5.1 *Discussion*—While not required, equal bar or class widths are recommended for histograms according to Practice E2586. A histogram provides information on the central tendency of the distribution, reveals the amount of variation in the data, provides information on the shape of the distribution, and reveals potential outlying values.

A1.1.6 *linear regression model, n*—any theoretical model built of the form $Y_i = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 \ldots \beta_{p-1} \cdot x_{p-1} + \varepsilon_1$, where Y_i is the response for case i; x_1, x_2, …, x_{p-1}, are the predictor variables; p is the number of regression coefficients; β_0, β_1, …, β_{p-1} are the regression coefficients; and ε_i are the random errors

that are assumed to be independently and identically distributed and follow a normal distribution with mean zero and constant variance, σ^2.

A1.1.6.1 *Discussion*—"Linear" regression means linear with respect to the coefficients, β_0, β_1, …, β_{p-1}. For example, $\hat{y}_i = b_0 + b_1 \cdot (x)^2$ and $\hat{y}_i = b_0 + b_1 \cdot x_1 + b_2 \cdot \ln(x_2)$ are linear regression models. However, $\hat{y}_i = b_0 \cdot \exp(\beta_1 \cdot x_1)$ is not a linear regression model. Simple linear regression refers to any linear model that includes a single predictor variable. For example, $\hat{y}_i = b_0 + b_1 \cdot x$ and $\ln(\hat{y}_i) = b_0 + b_1 \cdot \ln(x)$ are simple linear regression models. Estimates of linear regression coefficients can be obtained using the method of least squares or the method of maximum likelihood. If censored data are present, the method of maximum likelihood must be used.

A1.1.7 *linear regression with censored data, n*—a specialized linear regression modeling technique used when the response variable is not known exactly.

A1.1.7.1 *Discussion*—The method of maximum likelihood is used to estimate the model coefficients. Failure to treat censored data correctly can have a significant impact on the regression model, resulting in a poor model and misleading predictions.

A1.1.8 *measurement systems analysis (MSA), n*—any of a number of specialized methods useful for studying a measurement system and its properties. **Terminology E456, Guide E2782**

A1.1.9 *method of least squares, n*—a technique of estimation of a parameter which minimizes $\sum e^2$, where e is the difference between the observed value and the predicted value derived from the assumed model. **Terminologies E456, E1325**

A1.1.10 *method of maximum likelihood, n*—estimation method that finds the values of the parameters of interest, denoted by θ, that maximize the likelihood function, $L(\theta \mid x)$.

A1.1.10.1 *Discussion*—The method of maximum likelihood chooses values for the parameters of interest (for example, regression model coefficients) that are most consistent with the sample data. The likelihood function is directly derived from the joint probability function of the observed data, written as a function of the model parameters: $f(x \mid \theta) = L(\theta \mid x)$. The method of maximum likelihood finds the parameter values θ for which the value of the likelihood function is the largest, indicating a high probability given the observed data. It has been shown in theory that maximum likelihood estimates are optimal in large samples under standard regularity conditions. Typically the log of the likelihood function, $\ln(L(\theta \mid x))$ is used instead for computational convenience. While more computationally intensive than the method of least squares, the method of maximum likelihood is a more versatile estimation method since it can handle not only a wide variety of models but also a wide variety of data types, including censored data. If the distribution of the error term is specified and follows a location-scale distribution other than the normal distribution (for example, Weibull or lognormal), then the method of maximum likelihood can be used to obtain estimates of the regression model coefficients. There are two procedures for finding maximum likelihood estimates: analytical and iterative numerical search. When censored data are present, the only option is to use an iterative numerical search procedure since a closed form solution does not exist with the analytical procedure. The iterative numerical procedure searches for a solution to the system of equations from which the estimates of the model coefficients are derived. The procedure iterates until a convergence criterion is met, at which point estimates of the model coefficients are obtained from the last iteration. If no solution exists to the system of equations from which the model coefficients are derived, then the procedure will not reach convergence. Some statistical analysis software may produce estimates of the model coefficients even though the convergence criterion has not been met. These estimates are based on the last iteration. However, they are likely to be erroneous and should not be used. When no censored data are present, the values obtained for the parameter estimates using the method of maximum likelihood are the same as those obtained using the method of least squares and have the properties of all least squares estimators.

A1.1.11 *outlying observations, n*—an observation that appears to deviate markedly in value from other members of the sample in which it appears. **Practice E178, Terminology E456**

A1.1.12 *probability plot, n*—used to assess whether or not a particular continuous distribution fits continuous data by plotting what is expected under the assumed distribution against what is actually observed.

A1.1.12.1 *Discussion*—If the data closely follow the reference line and are within the 95% confidence bounds (if included on the plot), then the assumed distribution is considered a reasonable fit to the data. This visual assessment holds for any probability plot. There are also more formal statistical hypothesis tests, such as the Anderson-Darling (AD) test, that can be performed to assess the fit of the selected distribution. More detail on probability plots can be found in Practice E2586.

A1.1.13 *regression, n*—the process of estimating parameter(s) of an equation using a set of data. **Terminology E456, Practice E2586**

A1.1.13.1 *Discussion*—See Practice E2586 for a general overview of simple linear regression analysis.

A1.1.14 *residual, n*—observed value minus fitted value, when a model is used. **Terminology E456, Practice E2586**

A1.1.15 *statistical confidence, n*—the long run frequency associated with the ability of the statistical method to capture the true value of the parameter of interest.

A1.1.15.1 *Discussion*—Statistical confidence is a probability statement about the statistical method used to estimate a parameter of interest—e.g., the probability that the statistical method has captured the true capability of the inspection system. The opposite of statistical confidence can be equated to risk. For example, a statistical confidence level of 95% implies a willingness to accept a 5% risk of the statistical method yielding incorrect results—e.g., there is a 5% risk that the wrong conclusion has been drawn about the capability of the inspection system. (See Practice E2586, section 6.19.1 for more detail.)

A1.1.16 *statistical confidence bound, n*—a one-sided or two-sided bound around a single point estimate representing the variability due to sampling.

A1.1.16.1 *Discussion*—According to the formula in MIL-HDBK-1823A, $a_{p/c}$ is a one-sided upper confidence bound on a_p. $a_{p/c}$ represents how large the true a_p could be given the statistical uncertainty associated with limited sample data. In general, a confidence bound is a function of the amount of data, the scatter in the data, and the specified level of statistical confidence. When the sample size increases, statistical uncertainty decreases (all else held constant). That is, given an infinite amount of data (e.g., an infinite number of flaw sizes adequately distributed across a POD specimen set), $a_{p/c}$ will approach a_p because the statistical uncertainty goes away. It is important to note that a statistical confidence bound on a_p only accounts for variability due to sampling. It does not account for inherent process variability. In order to capture inherent process variability, a tolerance bound should be used. As opposed to a confidence bound, a tolerance bound will always differ from the point estimate because process variability cannot be eliminated by increasing the sample size.

A1.1.16.2 *Discussion*—The term "statistical confidence bound" in this standard is equivalent to the term "confidence interval" in Terminology E456 and Practice E2586.

APPENDIX

(Nonmandatory Information)

X1. POD ANALYSIS PROCESS

X1.1 POD Analysis

X1.1.1 Fig. X1.1 shows a flowchart of POD Analysis for $\hat{a}$ versus a data.

X1.2 Additional Commentary on the $\hat{a}$ versus a POD Analysis Process as illustrated in Figure X1.1 and its Significance

X1.2.1 *Define POD Analysis Objective:* In general, the objective of a POD analysis is to determine the relationship between discontinuity size and POD. Based on the established relationship, the objective may be to determine the discontinuity size that can be detected with a given probability p and specified statistical confidence level c, denoted $a_{p/c}$. It is important for the analyst to have a clear understanding of the specific analysis objective prior to performing the analysis.

X1.2.2 *Obtain POD Demonstration Test Data and Examination Specifics:* In general, the results of an experiment apply to the conditions under which the experiment was conducted.

X1.2.3 *Conduct Preliminary Review of Examination Procedure and Data:*

X1.2.3.1 If an experiment is not properly designed and executed, the data collected are subject to question and likely invalid. Invalid data cannot be corrected through a statistical analysis. Hence, any results from a statistical analysis of invalid data will be invalid as well.

X1.2.3.2 In general, a graphical assessment of the data should be performed prior to conducting a statistical analysis to become familiar with the data (for example resolution, distribution, correlations). A graphical assessment can also help identify potential outlying observations. See Practice E178.

X1.2.3.3 Prior to conducting a POD examination on a nondestructive inspection system that generates a measurable response, a Measurement Systems Analysis (MSA) is recommend to assess the adequacy of the measurement system in terms of its repeatability and reproducibility. See Guide E2782 for information about and guidelines for performing an MSA.

X1.2.3.4 Examples of examination procedures or data issues, or both, and possible resolutions can be found in sections 6.4 and 6.5.

X1.2.4 *Determine Noise Threshold and Saturation Threshold:*

X1.2.4.1 The analyst and responsible Engineer shall determine the appropriate value for the noise threshold, $\hat{a}_{noise}$, based on the noise analysis results. Example values include the 97.73 percentile or 99.87 percentile of the noise distribution, which corresponds to a +2σ or +3σ value respectively.

X1.2.4.2 The saturation threshold, $\hat{a}_{sat}$, is the largest signal value that the system can record.

X1.2.5 *Select Model:* Selection of a linear model may be an iterative process as the significance of the predictor variable(s) and the appropriateness of the selected model is typically assessed after the model has been developed.

X1.2.5.1 A linear model with a single predictor variable, commonly referred to as a simple linear regression model, is typically expressed as $\hat{y}=b_0+b_1\cdot x$, where x is the continuous predictor variable, b_0 is the intercept, b_1 is the slope, and $\hat{y}$ is the expected response given x. "Linear" refers to linear with respect to the model coefficients. Hence, $\hat{y}$ and x may represent transformations of the raw data. With respect to POD, for example, $\hat{y} = \hat{a}$ or $\hat{y}= \ln(\hat{a})$ and $x = a$ or $x = \ln(a)$, resulting in four possible simple linear models: $\hat{a}=b_0+b_1\cdot a$, $\hat{a}=b_0+b_1\cdot\ln(a)$, $\ln(\hat{a})=b_0+b_1\cdot a$, and $\ln(\hat{a})=b_0+b_1\cdot\ln(a)$. If predictor variables other than discontinuity size are quantifiable factors (either continuous or categorical), a linear regression model with more than one predictor may be used.

X1.2.5.2 In general, appropriateness is determined by how well the model fits the observed $\hat{a}$ versus a data and how well the analysis assumptions are met. If a simple linear model is selected, for example, start with a plot of $\hat{a}$ (y-axis) against a (x-axis) using a Cartesian scale. A natural log transformation is common in POD analysis. Re-create the plot looking at all possible combination of Cartesian and base e logarithmic scale for the y-axis and x-axis respectively: Cartesian versus

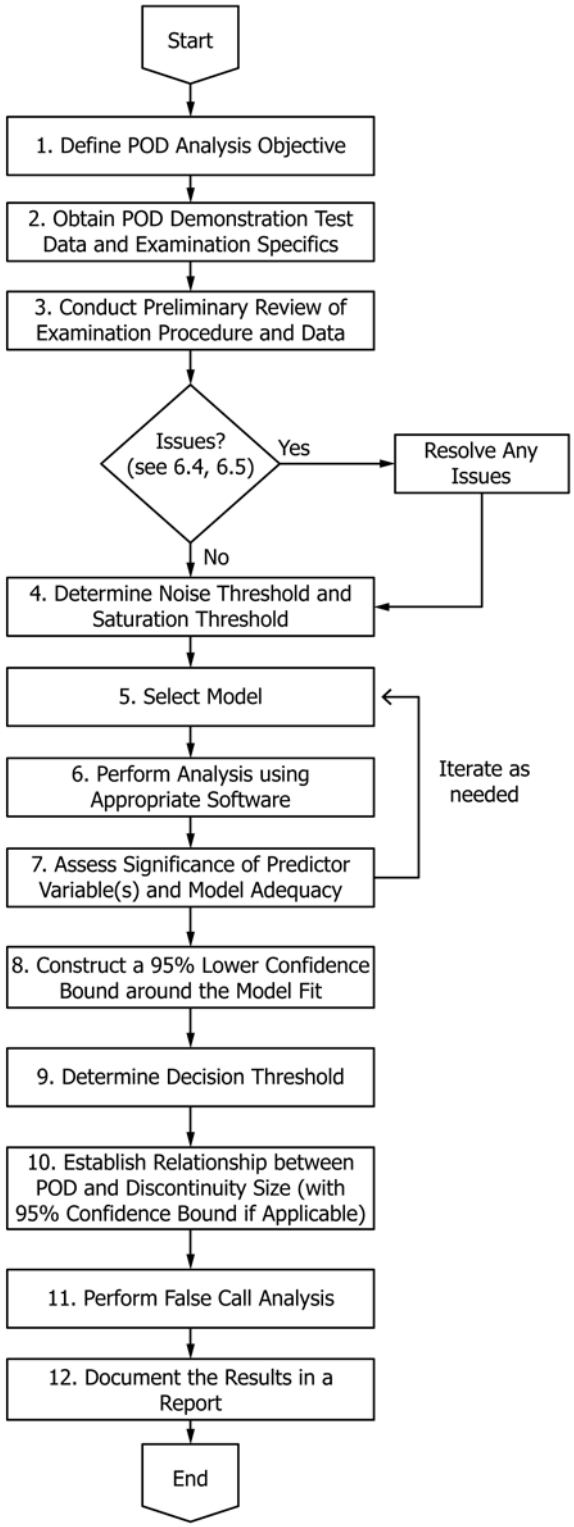

FIG. X1.1 Flowchart of POD Analysis *â* versus *a* Data

Cartesian, logarithmic versus Cartesian, Cartesian versus logarithmic, and logarithmic versus logarithmic. Select the simple linear model based on the plot in which the data appears most linear. This method is consistent with that described in MIL-HDBK-1823A.

X1.2.5.3 Model selection for linear regression models with more than one predictor variable can be facilitated during the analysis using formal statistical model building techniques such as forward selection, backward selection, or stepwise regression.[4]

X1.2.5.4 Other statistical regression models exist and may be more appropriate in some cases than the standard linear regression model for determining the demonstrated POD for $\hat{a}$ versus a data. It is the analyst's responsibility to select the appropriate statistical model and verify that all underlying assumptions associated with the selected model hold.

X1.2.6 *Perform Analysis Using Appropriate Software:* POD specific software or statistical software is commonly used to perform an analysis on $\hat{a}$ versus a data in order to establish a functional relationship between $\hat{a}$ and a, on which the POD curve is based.

X1.2.6.1 Prior to performing the POD analysis, the analyst shall format the data as required by the software used to conduct the analysis. For some software this may require the analyst to perform a transformation of the predictor variable prior to running the analysis. For example, if the natural log of discontinuity size is used as the predictor variable, then the analyst may need to create a new variable column for the natural log of discontinuity size prior to running the analysis. If censored data are present, the analyst shall verify if censored data are handled by the software and the format required by the software to identify censored data.

X1.2.6.2 Though the software performs the complex calculations, it does not check the validity of analysis inputs or outputs. The analyst is responsible for ensuring that the analysis inputs (e.g., data, model formulation) are correctly specified and that the underlying model assumptions hold. Treating the software as a "black box" can lead to seriously misleading conclusions about the inspection capability of the system. Hence, it is critical that the practitioner have a basic understanding of the complete analysis process, including the underlying statistical methods and techniques for validating the results.

X1.2.6.3 The standard states that if more than twenty iterations were needed to reach convergence when censored data are present, the model may not be reliable. This criterion was selected to be consistent with several well known software packages. The criterion of twenty is used in Minitab® statistical software and PODv3. MIL-HDBK-1823A uses a criterion of twenty-five.

X1.2.7 *Assess Significance of Predictor Variable(s) and Model Adequacy:*

X1.2.7.1 Only significant and uncorrelated variables are included in a linear regression model. The significance of a predictor variable is assessed after a model is selected and analysis performed. An Analysis of Variance (ANOVA) is typically used to evaluate the significance of the model as a whole. ANOVA uses an F-test to test the significance of the linear regression relation between the response variable and predictor variable(s). The p-value resulting from the F-test is used to assess the significance of the model as a whole. A p-value less than or equal to 0.05 implies evidence of statistical significance of the model as a whole. A t-test for significance or confidence interval for a regression coefficient can be a useful tool in assessing the significance of an individual predictor variable in the model. A t-test produces a p-value, which is used to judge how close the regression coefficient is to zero. A p-value less than or equal to 0.05 implies evidence that the predictor variable is a statistically significant contributor to the model. A two-sided 95% confidence interval can also be used to assess the significance of an individual predictor variable in the model. If the 95% confidence interval does not include zero, then the predictor variable is a statistically significant contributor to the model. The same conclusion about the significance of a predictor variable will be drawn from a t-test at the 0.05 level of significance and a 95% confidence interval using a confidence coefficient based on the t distribution.

X1.2.7.2 Fig. X1.2 shows three residual plot examples. If the residuals fall in a horizontal band centered around zero, with no systematic preference for being positive or negative as shown in Fig. X1.2(a), then the assumption of constant variance and a linear relationship holds. A "rainbow" (or inverted "rainbow) pattern as shown in Fig. X1.2(b) indicates that a linear relationship is not appropriate. A megaphone shape, either increasing (as shown in Fig. X1.2(c)) or decreasing, indicates non-constant variance.

X1.2.7.3 A plot of predicted values versus actual values to assess goodness-of-fit is most useful when more than one predictor variable is included in the model. When the model includes only one predictor variable, it is relatively easy to graphically assess how well a simple linear model fits the observed data. This is not typically the case when more than one predictor variable is included in the model.

X1.2.7.4 The analyst should use other standard statistical model diagnostic methods recommended for assessing the adequacy of a linear regression model beyond those described in this standard. These methods are performed after the model has been developed. Examples include but are not limited to those described in the following subsections.

X1.2.7.4.1 In addition to a histogram, a normal probability plot of the residuals should also be assessed to verify the normality assumption. A normal probability plot is a more

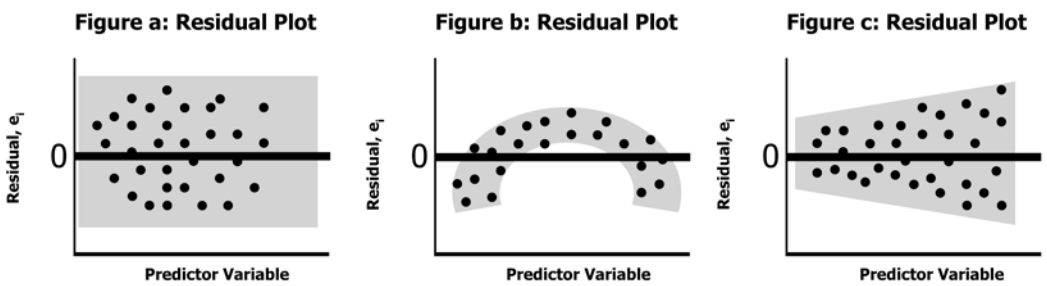

FIG. X1.2 Example Residual Plots

rigorous diagnostic tool for assessing a distributional assumption. A detailed description of and method for constructing a normal probability plot can be found in Practice E2586. More formal statistical tests, such as the Anderson Darling Test, to assess distributional assumptions are also available and recommended.

X1.2.7.4.2 In addition to plotting the residuals against the predicted values as illustrated in Fig. X1.2, there are also more formal statistical tests that are recommended to assess linearity and non-constant variance such as Levene's Test for Constant Variance and F-Test for Lack-of-Fit respectively. Note that the F-Test assumes that the normality and constant variance assumptions hold.

X1.2.7.4.3 If available in the analysis output, the R^2 and Adjusted R^2 statistics should be used to verify goodness-of-fit. When the model includes a single predictor variable, R^2 shall be used. By definition, $0 \le R^2 \le 1$. $R^2 = 1$ would imply that the change observed in the response variable is a direct result of manipulation of the predictor variable. A low R^2 value may imply, for example, that there may be other explanatory variables that should be considered for inclusion in the model. When the model includes more than one predictor variable, Adjusted R^2 shall be used. Adjusted-R^2 is preferred to R^2 because R^2 will always increase when more explanatory variables are added. Adjusted-R^2 is interpreted in the same way as R^2 when assessing goodness-of-fit.

X1.2.7.4.4 Influential observations are not necessarily wrong or even outlying values. However, misleading or wrong conclusions may be drawn due to a single influential observation. If available, more formal diagnostic measures for influential observations should be used to identify influential observations. Several measures of influence which are commonly used include Leverage, Standardized Leverage, Delta β, Standardized Delta β, and Cook's Distance.

X1.2.7.4.5 A correlation analysis should accompany the graphical analysis to test for statistically significant correlations among the predictor variables (also known as multicollinearity). More formal diagnostic measures, such as Variance Inflation Factor (VIF), should be used if available to determine whether multicollinearity may be present.

X1.2.7.5 More detailed descriptions of the methods described in X1.2.7 can be found in most statistics text books that cover linear regression analysis.[4] Most statistics software packages with linear regression modeling capabilities have the standard model diagnostic methods built-in.

X1.2.8 *Construct a 95% Lower Confidence Bound Around the Model Fit:* A confidence bound on linear regression model reflects the long run frequency associated with the ability of the statistical method to capture the true relationship between $\hat{a}$ and a. In general, statistical confidence is a probability statement about the statistical method used to estimate a parameter of interest (the regression model coefficients in this case). The opposite of statistical confidence can be equated to risk. For example, a statistical confidence level of 95% implies a willingness to accept a 5% risk of the statistical method yielding incorrect results—for example, there is a 5% risk that the wrong conclusion has been drawn about the relationship between $\hat{a}$ and a.

X1.2.9 *Determine Decision Threshold:* $\hat{a}_{dec} = \hat{a}_{noise}$ implies that any signal above the noise is interpreted as a find.

X1.2.10 *Establish Relationship between POD and Discontinuity Size (with 95% Confidence Bound if Applicable):*

X1.2.10.1 Formula for Calculating POD for the simple linear regression model $\hat{a} = b_0 + b_1 \cdot a$, where Φ is the cumulative standard normal distribution function, b_0 and b_1 are the estimated model coefficients, and $\hat{\sigma}$ is the estimated standard deviation around the predicted response:

$$POD = P(\hat{a} > \hat{a}_{dec}) \quad \text{(X1.1)}$$

$$= 1 - P(\hat{a} \le \hat{a}_{dec})$$

$$= 1 - \Phi\left(\frac{\hat{a}_{dec} - \hat{a}}{\hat{\sigma}}\right) = 1 - \Phi\left(\frac{\hat{a}_{dec} - (b_0 + b_1 \cdot a)}{\hat{\sigma}}\right)$$

X1.2.10.2 Formula for Calculating POD for the simple linear regression model $\ln(\hat{a}) = b_0 + b_1 \cdot \ln(a)$

$$POD = P(\hat{a} > \hat{a}_{dec}) \quad \text{(X1.2)}$$

$$= 1 - P(\hat{a} \le \hat{a}_{dec})$$

$$= 1 - \Phi\left(\frac{\ln(\hat{a}_{dec}) - \ln(\hat{a})}{\hat{\sigma}}\right) = 1 - \Phi\left(\frac{\ln(\hat{a}_{dec}) - (b_0 + b_1 \cdot \ln(a))}{\hat{\sigma}}\right)$$

X1.2.10.3 See Fig. X1.3 for an illustration of calculating POD.

X1.2.10.4 A similar formula is used to calculate POD with 95% confidence. The formula used to construct the 95% lower confidence bound on the model fit is used in place of the model fit. Using the simple linear regression model $\hat{a} = b_0 + b_1 \cdot a$ as an example:

$$POD = 1 - \Phi\left(\frac{\hat{a}_{dec} - (95\%\ \text{lower bound formula}}{\hat{\sigma}}\right) \quad \text{(X1.3)}$$

X1.2.10.5 See Fig. X1.4 for an illustration of calculating POD with 95% confidence.

X1.2.10.6 The resulting POD values with 50% confidence and 95% confidence are plotted against their respective discontinuity size to produce a POD Curve and 95% lower bound on the POD curve as illustrated in Fig. X1.5.

X1.2.11 *Perform False Call Analysis:* This standard does not limit the use of other methods aside from those described in MIL-HDBK-1823A for assessing false call rate, provided that they are appropriate.

X1.2.12 *Document the Results in a Report:* The report should contain enough information such that the analysis results may be reproduced.

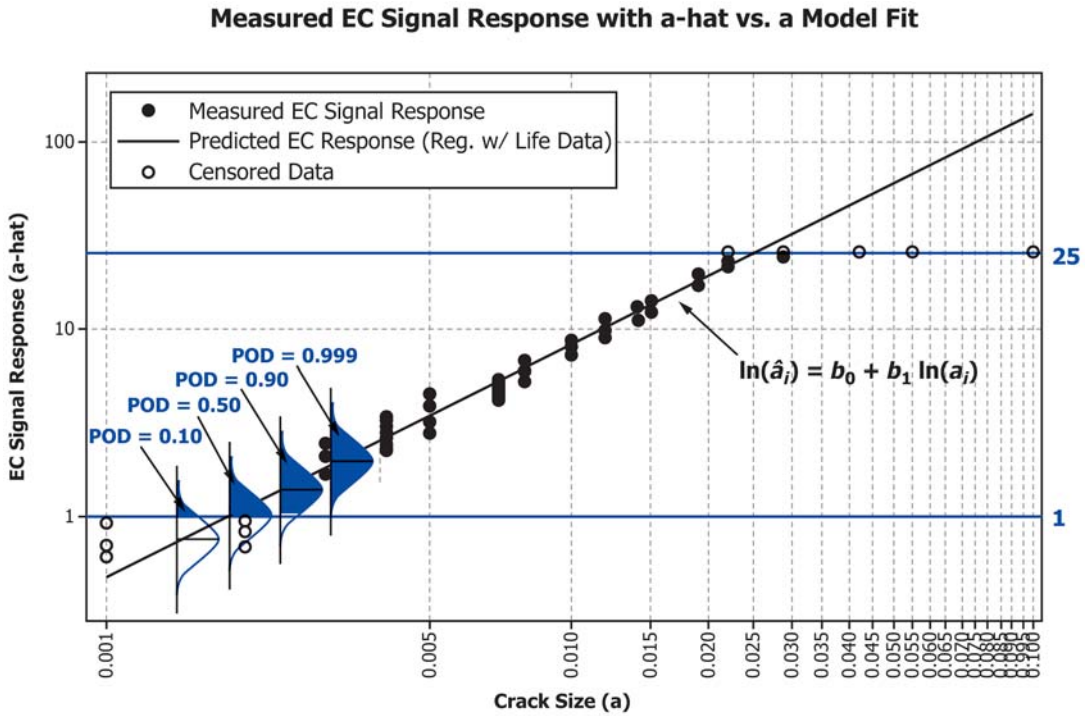

FIG. X1.3 POD with 50% Confidence is the Area Under the Normal Curve Above $\hat{a}_{dec}$, Where the Normal Curve is Centered Around the Model Fit

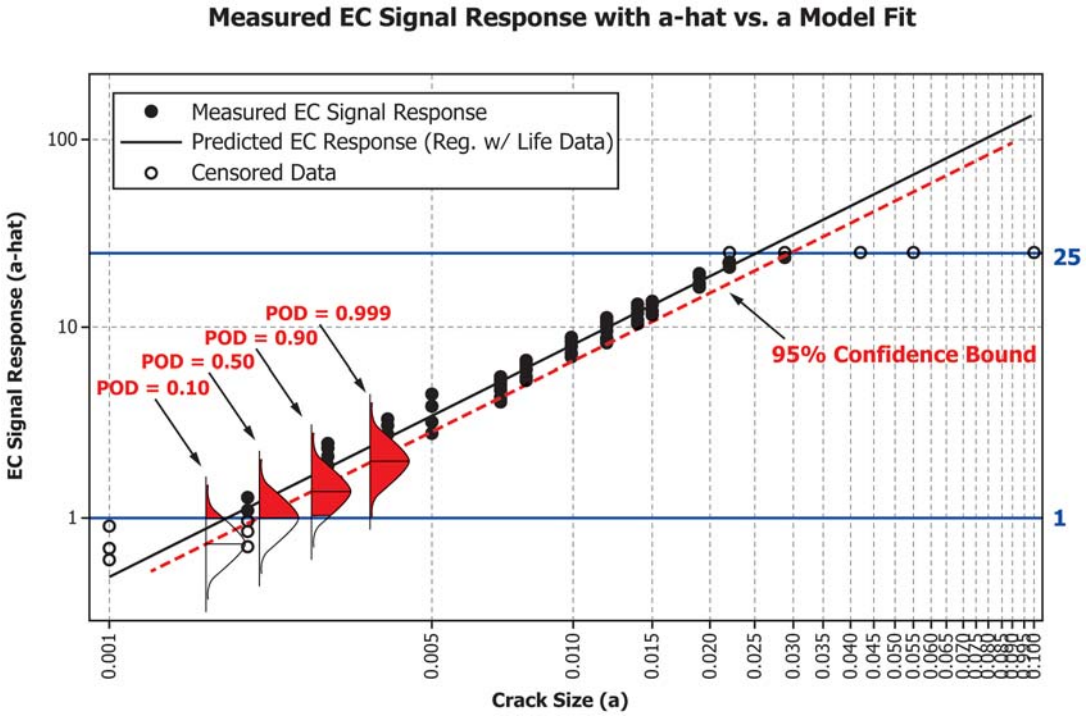

FIG. X1.4 POD with 95% Confidence is the Area Under the Normal Curve Above $\hat{a}_{dec}$, Where the Normal Curve is Centered Around the 95% Lower Bound on the Model Fit

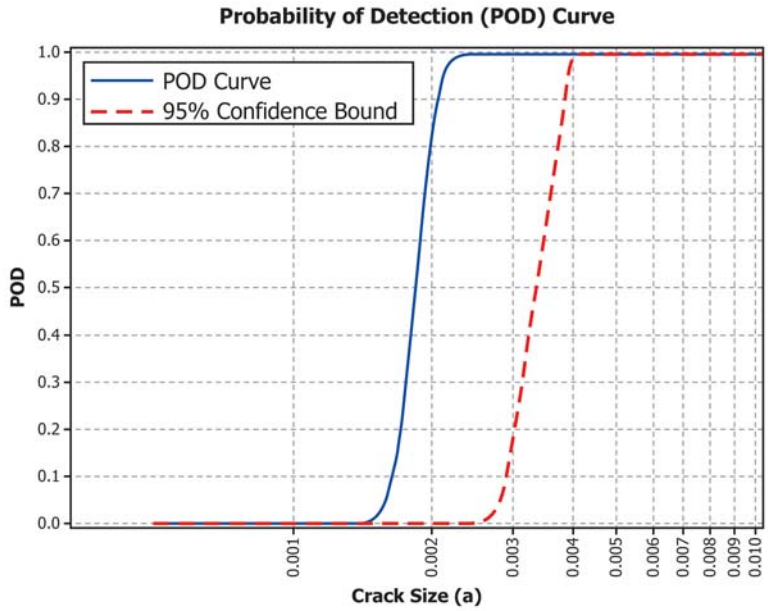

FIG. X1.5 POD Curve and 95% Lower Confidence Bound