NASA/TM–2014-218183

Interrelationships Between Receiver/Relative Operating Characteristics Display, Binomial, Logit, and Bayes’ Rule Probability of Detection Methodologies

Edward R. Generazio
Langley Research Center, Hampton, Virginia

April 2014

NASA STI Program . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NASA Aeronautics and Space Database and its public interface, the NASA Technical Report Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA Programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.
- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.
- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include organizing and publishing research results, distributing specialized research announcements and feeds, providing information desk and personal search support, and enabling data exchange services.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at *http://www.sti.nasa.gov*
- E-mail your question to help@sti.nasa.gov
- Fax your question to the NASA STI Information Desk at 443-757-5803
- Phone the NASA STI Information Desk at 443-757-5802
- Write to:
  STI Information Desk
  NASA Center for AeroSpace Information
  7115 Standard Drive
  Hanover, MD 21076-1320

NASA/TM–2014-218183

Interrelationships Between Receiver/Relative Operating Characteristics Display, Binomial, Logit, and Bayes' Rule Probability of Detection Methodologies

Edward R. Generazio
Langley Research Center, Hampton, Virginia

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

April 2014

Available from:

NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320
443-757-5802

Interrelationships Between Receiver/Relative Operating Characteristics Display, Binomial, Logit, and Bayes' Rule Probability of Detection Methodologies

Edward. R. Generazio[1]

[1]National Aeronautics and Space Administration, Hampton, VA 23681

ABSTRACT. Unknown risks are introduced into failure critical systems when probability of detection (POD) capabilities are accepted without a complete understanding of the statistical method applied and the interpretation of the statistical results. The presence of this risk in the nondestructive evaluation (NDE) community is revealed in common statements about POD. These statements are often interpreted in a variety of ways and therefore, the very existence of the statements identifies the need for a more comprehensive understanding of POD methodologies. Statistical methodologies have data requirements to be met, procedures to be followed, and requirements for validation or demonstration of adequacy of the POD estimates. Risks are further enhanced due to the wide range of statistical methodologies used for determining the POD capability. Receiver/Relative Operating Characteristics (ROC) Display, simple binomial, logistic regression, and Bayes' rule POD methodologies are widely used in determining POD capability. This work focuses on Hit-Miss data to reveal the framework of the interrelationships between Receiver/Relative Operating Characteristics Display, simple binomial, logistic regression, and Bayes' Rule methodologies as they are applied to POD. Knowledge of these interrelationships leads to an intuitive and global understanding of the statistical data, procedural and validation requirements for establishing credible POD estimates.

INTRODUCTION

The lack of understanding in discussions concerning probability of detection (POD) capabilities is revealed by the confusion in supporting statements from the nondestructive evaluation (NDE) community. The general lack of understanding by community members is best illustrated and highlighted by specific statements:

"Oh boy, confidence limits. I hate these."
"They don't use 90/50 they use 90/95."
"I should have used 90/50."
"I defer my answer to the statistician."
"I'm not a statistician."
"Our statistician does not agree with your statistician."
"A high false positive probability is only an economic concern."
"90/50 POD means that there is a 50% chance that the true POD is greater than 90% at that discontinuity size?" Responses: "No.", and "Yes.", rest of world gives blank stares.
"Confusion over common definitions continues to be an issue…"
"We never validate POD curves, we only update them."
"We have been using 29 out of 29 binomial point estimate method clandestinely for years"
"Maximum likelihood estimation has nothing to do with the confidence limits."

“A man with two watches never has the correct time.”
“Their POD analysis is wrong.”
“We use a cumulative POD method”
“We fit POD data into a three-parameter Weibull distribution instead of the standard normal distribution”
"I'm too much of a knucklehead to know this stuff."
“Should I use tolerance bounds or confidence bounds?”

Although these recent quotes are from different sources (industry, government, and academia) and venues, they do highlight that the understanding of POD methods is less than straight forward. This is the environment in which the NDE community exists today. Accepting the true nature of this environment is an important step that is needed to move forward establishing a more uniform understanding of POD methods among the NDE community and to reduce risk.

Many of the above quotes are based on a partial understanding of the environment at the time. Some are facts, some are incorrect, and some have been condensed to get a point across to those that may be laymen. Subsequently, these condensed answers are reinterpreted by others to yield even briefer statements, and so on. Many of the quotes above will make statisticians cringe with disbelief and subsequently the statisticians make concerted efforts to provide corrections or clarifications to the statements. The statisticians’ clarifications are often couched in the nomenclature of the field so that the clarifications are often misinterpreted by others. The underlining issue here is that the NDE and statistical communities are talking past each other in a *dual-ogue* rather than a dialogue.

The *dual-ogue* remains present due to the lack of an overall understanding by the NDE community of the many statistical methodologies used to estimate POD. This lack of understanding is further exasperated since each statistical methodology has its own unique niche in providing an understanding of the POD data being presented. In order to develop a comprehensive understanding of POD results being presented it is necessary to expose the interrelationships between the statistical methodologies used to estimate POD. It is expected that after comprehending this work, the reader will be able to understand the interrelationship between statistical methodologies used to estimate POD and to properly address each of the above quotes, as well as many other related statements.

This work sets the foundation for an extended discussion on POD decision making where a comprehensive, decision-support document is needed that provides guidelines on practical risk-informed decisions involving POD/confidence (CL) level utilizations. The decision guidelines needed are identified later.

STATISTICAL METHODS OVERVIEW- MODELS, PARAMETER ESTIMATION, CONFIDENCE INTERVALS, AND VALIDATION

There are several methods for estimating POD. These methods use different statistical models each having unique data requirements. Models include simple binomial method for Hit-Miss data of one discontinuity size, regression models when signal response is related to discontinuity size (also known as a vs. $\hat{a}$), and binary regression for Hit-Miss data from multiple discontinuity sizes, e.g., logistic or probit regression. The regression models may be quite complex. Linear regression is often misunderstood to mean that the explanatory variable (discontinuity size, a) only enters the mean response model of the typical form,

(1) $y = \alpha + \beta \cdot a$, where α and β are parameters, and a is discontinuity size.

However, statistical models that are linear in coefficients (α, β) that need to be estimated do not require the explanatory variables, e.g., a , themselves to be linear. That is, the mean equation,

(2) $y = \alpha + \beta \cdot a + \omega \cdot a^2$, where ω is a parameter

is a linear regression problem as long as discontinuity size a , and therefore a^2 are known beforehand .

Recently, the National Aeronautics and Space Administration (NASA) used four different statistical methodologies for establishing POD capability from Hit-Miss (binary) inspection data. These were Receiver/Relative Operating Characteristics (ROC), simple binomial, logistic regression, and Bayes' Rule methods. Different statistical models and methods are used for each of these methods.

There are arguments that some statistical methods are better than others for a variety of reasons, ranging from data availability, statistically more efficient to use less data, to historical use. These four methods will be discussed and Hit – Miss data will be used to reveal the interrelationships between these statistical methods and models. It is pointed out here that signal response models are present even when considering Hit-Miss data. Here a decision threshold on the signal response establishes the Hit or Miss decisions. This is true even for inspections such as penetrant, and radiography, etc, where the underlining signal response functions (brightness, film indication density, length, etc.) are not known.

Statistical Models

It is attractive to develop signal response models first. However, there are an unlimited number of signal response models that rely on assumptions that introduce additional uncertainty in the adequacy of the model. Therefore, this work will focus on Hit-Miss data and companion decision thresholds that are in common use today.

Estimation of Model Parameters.

There are numerous methods for estimating models parameters. Many of these methods, e.g., estimating the proportions for the binomial distribution with the sample proportion developed by

Jacob Bernoulli in 1689, mean of a normal distribution with a sample mean developed by Abraham de Moivre in 1738 and Carl Friedrich Gauss in 1809, and regression coefficients of a simple linear regression developed by Carl Friedrich Gauss in 1795 using least squares. These methods can be shown to be special cases of the more recently developed method of likelihood principle developed by R. A. Fisher between 1912 and 1922. The likelihood provides a mathematical interrelationship between statistical methods being discussed here.

Importance of Confidence Statements

Confidence statements indicate the statistical uncertainty in the estimation of parameters caused by the limited data. As the amount of data is increased the width of statistical confidence interval is decreased. Confidence intervals do not explicitly address the variability in an inspection process (Li, Spencer, & Meeker, 2012), such as, operator-to-operator variations, unless statistical models explicitly include parameters reflecting those variables. However, the confidence intervals do contain the variability of the population if adequate random sampling of the population is performed. There are multiple methods for establishing confidence statements. In general, Wald confidence intervals (Wald, 1943, Agresti & Coull, 1998 and Christner & Long & Rummel, 1988) are easy to compute and are justified on the basis of large-sample statistical theory. In some circumstances (small samples), the Wald confidence intervals may not be adequate, and likelihood based confidence intervals are more reliable.

Estimates of POD should always be accompanied by a confidence interval showing at least the lower single-sided confidence bound and confidence level. For example, a typical lower bound confidence statement may appear as: “The estimated POD exceeds 0.90 with 95% confidence at the discontinuity size of 0.080 inch (a 90/95= 0.080 inch). The method for estimating POD and the associated confidence bounds need to be explicitly stated whenever POD capability is specified. It will be shown later that the a 90/95 definition may be inadequate for failure critical systems.

Model Validation

There are a wide variety of POD models and it generally is not known *a priori* if any given model is adequate. Estimated POD models need to be validated or evaluated for adequacy. Possible ways to do this depend on the specific model where different models have different assumptions. Model validations provide guidance on the adequacy of the statistical model. For example, goodness-of-fit of a statistical model to data may imply that a model, such as the logistic form for the POD is adequate for estimating a 90% detection discontinuity size for a given set of data. However, it should not be taken as a statement that no other model should be used, nor should it be taken as a claim that it is adequate for all future uses for characterizing the same NDE type of process. Other internal and external validation methods are used to further assess the adequacy of the estimated model obtained from a given set of data.

TOP LEVEL GUIDANCE

Different methods of estimating POD may have quite different test specimen requirements. It is important to have test specimens containing discontinuity populations that are representative to the general population of real discontinuities. Often, subjective decisions are made on what discontinuity populations are required in order to claim that the test specimens are representative of the population. However, there are formal approaches to validate that the test specimen set is adequate. These formal methods include external validation of the estimated POD when the statistical POD model is known to be adequate, as well as using POD estimation methods where the proof property of the test specimen requirements has been demonstrated (Generazio, 2011).

All test procedures and processes, such as, test setup, calibration, detection threshold, testing protocol, documentation requirements, etc., are to be fixed prior to performing a POD test.

DEFINITIONS

There is some confusion among common definitions. A listing of definitions is provided here to assist in removing this confusion.

Indication – A NDE signal that exceeds a pre-specified value.

There are two levels of evaluation for inspection data. The first level is to determine the presence or lack of an indication. The second level is to evaluate each result or indication for relevance. A non-relevant designation may refer to something like an insignificant surface scratch or non-imperfections as part of the material's characteristics or the physical makeup of the component are part of the design or results from the fabrication materials or methods. These non-relevant indications do not received further classification. Relevant indications are further tested to the detection threshold and subsequently become classified as positive or negative indications. There is confusion here in regards to classifications that can be made by an inspector, who is relying only on NDE signal information versus classifications that can be made by a test monitor or experimenter who has knowledge of the flaw state independent of the NDE inspection. Here positive and negative refer to meets or exceeds and not exceeding the detection threshold, respectively.

Often it is required to record all relevant indications, including those indications that may be from non-critical discontinuity sizes. These relevant indications may not exceed the detection threshold for further classification, however, they may exceed a tracking threshold for recording.

Detection Threshold – A measured value at or above this NDE signal threshold level, a positive indication, for which a classification of the presence of a discontinuity is made. A measured value below this NDE signal threshold level is a negative or no indication. This NDE signal threshold is defined in the inspection requirements, and is often established during the engineering and development of an inspection procedure. A common NDE signal threshold level may be stated as a signal magnitude three times greater than the noise level observed when no discontinuity is present. In some organizations this NDE signal threshold is also called the acceptance threshold.

Tracking Threshold – A measured value at or above this NDE signal threshold level may be recorded as a suspect discontinuity and may be used for records.

Critical Discontinuity Size – The initial discontinuity size used in damage tolerance fracture analyses.

Non-Significant Discontinuities- All discontinuities having discontinuity sizes that are less than the critical discontinuity size.

An experimenter, with knowledge of the true flaw condition, can further classify the results of the inspection with respect to a specific discontinuity size as being either True Positive, False Positive, False Negative, or True Negative. Thus the definitions:

True Positive or Hit – The classification of a positive indication where a discontinuity of critical discontinuity size or larger exists.

False Positive – The classification of a positive indication where a discontinuity of critical discontinuity size or larger does not exist.

False Negative or Miss – The classification of a negative indication where a discontinuity of critical discontinuity size or larger exists.

True Negative – The classification of a negative indication where a discontinuity of critical discontinuity size or larger does not exist.

Note that the above classifications are made with respect to a specific detection threshold that may be set to correspond to a critical discontinuity size. A change for this quantity would result in different classifications for the results of an inspection.

Noise - The presence of constructive or destructive signal effects due to random or systematic mechanisms such as, electronic, material structure, human factors, etc. The list of mechanisms is extensive and it should be considered that all inspection data can be influenced by various noise factors. This creates an issue when a POD analysis approach ignores the influence of noise. For example the mechanism resulting in a false positive classification is often, erroneously, not considered as also affecting (either additive or subtractive) the signal responses resulting from a discontinuity.

Noise Level - signal responses due to random or systematic mechanisms such as, electronic, material structure, human factors, etc., that are present even in the absence of discontinuities. The noise level may be probabilistic but is often characterized by a maximum signal response that can be expected in the absence of a flaw.

Independent Events – Two events are independent when the occurrence of one event does not affect probability of the other event occurring. For example, signals from a discontinuity may be independent from electronic noise signals that may be observed at the time of inspection.

Mutually Exclusive and Non-Mutually Exclusive Events - Events are mutually exclusive when only one event may occur at a time. Hit or Miss events are classifications that are mutually exclusive. That is, we can not have an event classified as a Hit and also have the same event classified as a Miss. It is either one or the other. In contrast, signals due to noise and signals due to discontinuities may or may not occur at the same time, therefore signals are non-mutually exclusive signals. These non-mutually exclusive signals may be either additive or subtractive and this affects the resulting mutually exclusive classifications, such as Hit, Miss, false positive, and false negative.

Non-Mutually Exclusive Signals– Signal responses are non-mutually exclusive when responses may occur and interact simultaneously to create constructive or destructive interference. For example, a positive indication may be due to noise response or a positive indication may be due to a discontinuity signal response. Both responses may be independently too small to yield a positive indication separately, however, these two responses may be constructively additive to yield a positive indication. Non-mutually exclusiveness is assumed in the presence of high signal to noise levels or when other strong mechanisms create false positives.

.

PROBABILITY OF DETECTION AS A FUNCTION OF DISCONTINUITY SIZE IN THE PRESENCE AND ABSENCE OF NOISE

It is helpful to explore the fundamental structure of the typical relationship for probability of detection versus discontinuity size. In the discussion that follows a distinction is made between an indication that meets or exceeds the tracking threshold but does not meet or exceed the detection threshold and a positive indication that meets or exceeds the detection threshold used for identifying a discontinuity of critical discontinuity size. The distinction being made is that a positive indication is made to identify discontinuities with sizes that are at or greater than the critical discontinuity size, whereas an indication can occur for all size discontinuities that yield signal responses meeting or exceeding the tracking threshold. Although we may identify an indication as positive, this does not imply that a discontinuity of critical discontinuity size exists. Further classification will identify a positive indication as a true positive or a false positive.

Figure 1a illustrates an idealized signal response with respect to discontinuities as a red line. If there were no noise factors present one would be able to set a detection threshold, 1.5 (upper horizontal dash line), corresponding to the signal response for a discontinuity of critical discontinuity size. This would result in the step function probability of detection (POD) that is zero for non–significant discontinuities (discontinuities smaller than the critical discontinuity size) and one for all discontinuities having sizes at the critical discontinuity size or greater.

Consider the vertical lines in Figure 1b as representing the extent of noise. The noise amplitude may be additive or subtractive from the discontinuity signal response and varies from +0.5 to -0.5. If the noise were uniformly distributed along the vertical distance represented then a probability of an indication (POI) represented by the open circles and solid green line would result as shown in Figure 1b. The solid green line represents POI in reference to the detection of discontinuities having discontinuity sizes at or greater than the critical discontinuity size, whereas the open circle portion reflects the false positive probability contribution to POI with respect to the discontinuity sizes. If the noise is characterized by a Normal or Gaussian distribution then the familiar s-shaped curve (dash curve in Figure 1c) follows for the POI (with the curve still having the same interpretations below and above the critical discontinuity size). We use the label POI here where a positive indication may now be recorded due to the presence of noise and/or the presence of a discontinuity.

The noise models used here are meant to be instructive in highlighting the origin of the shape of POD curves observed in practice.

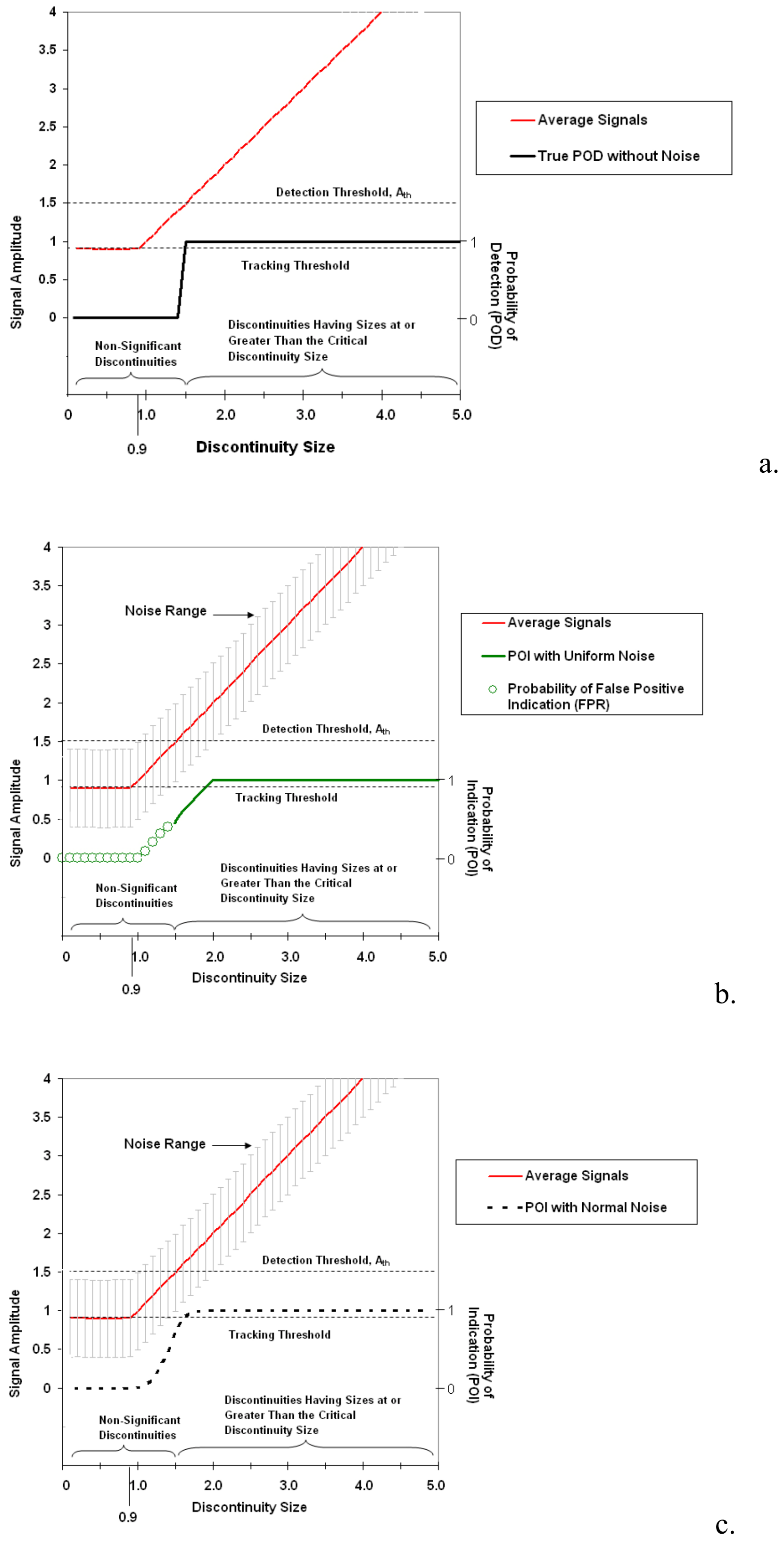


Figure 1. Probabilities as a function of discontinuity size with and without signal noise.

It should be noted that many POD analyses do not make the distinction of the detection of a discontinuity at or greater than the critical discontinuity size, but rather consider an indication as the same as a detection.

There can be many contributors to factors that would cause NDE signals to deviate from an idealized relationship with discontinuity size. One ever present factor is the intrinsic variations of character and morphology of same-sized discontinuities. Such variations result in true signal responses (without other sources of noise) that also vary. The signal responses from the same sized, but not identical, discontinuities may or may not reach the detection threshold. These signal varying responses are actually interrogating and exposing the character or morphology of the individual same-sized discontinuities. Therefore, when the data from same -sized, but not identical, discontinuities are used to generate a point estimates of POD, the signal responses from same-sized, but not identical, discontinuities also transform the ideal (no noise) POD step function to the familiar "S" shape POD function. Figure 1c shows the POI for discontinuities having a normal distribution of variations in discontinuity topology in the absence of noise. In this case we may use also use the label POD where a positive indication, in the absence of noise, is recorded where the signal response is only due to the presence of a discontinuity.

If the detection threshold, A_{th}, is lowered it is clear that the POI increases for discontinuities with sizes at or greater than the critical discontinuity size. However, the false positive portion of the POI will also increase. Lowering the detection threshold level to be within the noise band surrounding the non- significant discontinuities, e.g., $A_{th} = 1.25$, where the idealized average signal is no longer a factor, results in the probability of a false positive indication that can be significantly different than zero. The trade-off of increasing POI at the expense of having a large false positive probability has been addressed in various ways.

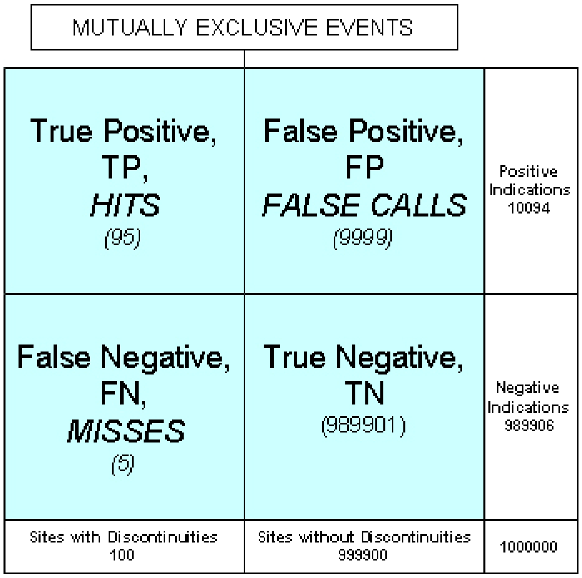
Generazio (2009, 2011) identified an acceptable maximum probability of false positive allowed (3.44%) for simple binomial applications, while Fahr, Forsyth, Bullock & Wallace (1995) provided guidance that probability of false positive should not exceed 5%. Fahr, et al recognized that the probability curves fit to the data were POI curves. He assumed the probabilistic model that an indication resulted from either of two mechanisms, a call independent of discontinuity size or a call due to the presence of a discontinuity. The resulting POI is a function of the false positive probability as well as a POD function. If the false call probability exceeded 5% Fahr backed out the POD from the POI. Spencer (1998) adds parameters to limit the maximum and minimum POD asymptotes to less than 1.0 and greater than zero, respectively to reflect not only a false call probability influence, but also the possibility that misses, like indications, could have a random component independent of discontinuity size. Spencer's extension recognizes that POD as characterized from blind inspections is actually a POI as discussed here. He proposes that "lucky hits" solely due to a mechanism creating false positives cannot be distinguished from detections due solely to a signal response to a discontinuity. Thus, the modeling of the form of the POI should recognize that the natural lower limit will be no lower than the probability of a false call.

STATISTICAL METHODS

Receiver/Relative Operating Characteristics (ROC) Display Method

The Receiver/Relative Operating Characteristics (ROC) (Tanner & Swets, 1954) method was originally developed for characterizing the capability of radar systems. The target size was not the main concern, but whether or not a hostile aircraft was in the vicinity. The ROC method is primarily a method to display estimated POI versus the false positive probability for a fixed discontinuity size and changing the decision criteria (signal threshold, in most cases). It will be shown that the POI displayed in ROC graphs can be derived from the simple binomial model.

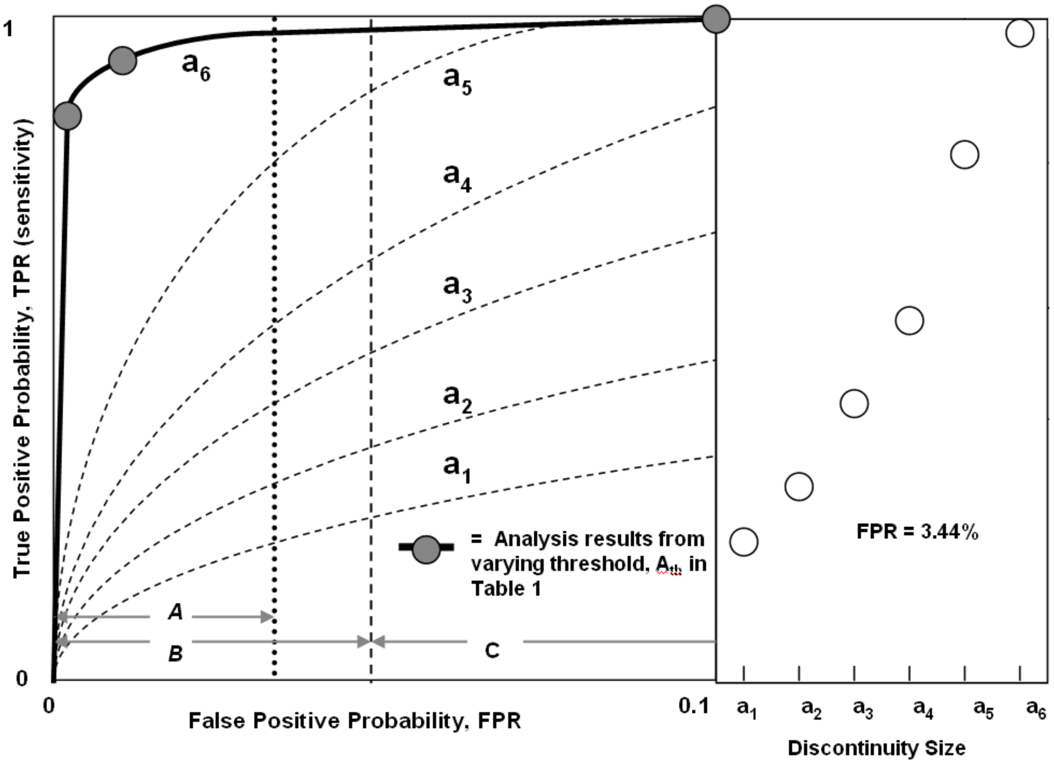
Data indicating a detection of an aircraft was collected as a function of a cathode ray tube (CRT) intensity point on the CRT screen. A fixed intensity level is used as a threshold for determining a positive indication or negative indication result. If there was an aircraft present and the CRT intensity greater than some threshold intensity, then this was recorded as a positive indication and a True Positive (a Hit). If the CRT intensity was less than the threshold, then the aircraft was undetected and this was recorded negative indication and a False Negative (a Miss). By decreasing the threshold, additional positive indications may be observed even when there was no aircraft present and these were recorded as False Positives. A negative indication result when there was no aircraft was recorded as a True Negative. For typical evaluations, increasing the threshold generally resulted in the number of False Positives to decrease, while the number of False Negatives increased. By varying the threshold, a data set was created that contains the number of True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN) as a function of the threshold level. An important requirement is that TP, FN, FP and TN are mutually exclusive events or classifications. Although the ROC display method is decades old, the data set needed for the ROC display method is typical of what is required for many POD statistical methodologies. When a single target size is used, then the data set is for that target size only. For the following discussion, targets will be referred to as discontinuities. First, consider a single discontinuity size and fixed threshold value, A_{th}. The resulting data matrix, with example data taken from Table 1, using A_2 threshold value, is shown in Figure 2a.



MUTUALLY EXCLUSIVE EVENTS

True Positive, TP, *HITS* *(95)*	False Positive, FP *FALSE CALLS* *(9999)*	Positive Indications 10094
False Negative, FN, *MISSES* *(5)*	True Negative, TN (989901)	Negative Indications 989906
Sites with Discontinuities 100	Sites without Discontinuities 999900	1000000

(a)



(b)

Figure 2. ROC Matrix: (a) Historical form of ROC matrix showing data for one discontinuity size and A_{th} = A_2. (b) ROC curve showing values (shaded circles) from Table 1 for one discontinuity size, a_6. Possible ROC curves (dashed curves) for additional discontinuity sizes a_1, a_2, a_3, a_4, and $a_{5,}$. Example true positive probability (open circles) versus discontinuity size when false positive probability is 0.0344.

For an example, suppose that there are 1,000,000 different locations that will be inspected and that there are discontinuities in 100 of these locations. For an inspection at a given location, there will be a signal response, say SR_i, where i =1 to 1,000,000. If $SR_i \geq A_{th}$, then there is a positive indication. Otherwise there is a negative indication.” If there is a positive indication on one of the locations where there is a discontinuity, we say there is a “True Positive” or “Hit.” If there is a positive indication at a location where there is no discontinuity, we say there is a False Positive” or “False Call”. If $SR_i \leq A_{th}$, then there is a negative indication. If $SR_i \leq A_{th}$ at a location with a discontinuity, then we have a “False Negative” or “Miss”. If there is a negative indication at a location where there is no discontinuity, then we have a True Negative.

Generally, if the detection threshold A_{th} is decreased, then the number of positive indications will increase, increasing the number of True Positives. Unfortunately, decreasing A_{th} generally increases the number of False Positives. Table I gives some illustrative numbers for the example, assuming that all discontinuities are of the same size (or that the probability of a indication does not depend on size).

A_{th} Threshold	TP	FP	FN	TN	Total Number of Inspected locations	TPR	FPR
$A_1 = 3$	90	999	10	998901	1000000	0.900	0.001
$A_2 = 2$	95	9999	5	989901	1000000	0.950	0.010
$A_3 = 1$	99	99999	1	899901	1000000	0.990	0.100

Table 1 Estimates of TPR and FPR for Different Detection Thresholds

The true positive probability (TPR) and the false positive probability (FPR) are given by,

(3) $$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{P}} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} = \Pr\left(\text{Positive Indication} \mid \text{Discontinuity}\right)$$

$$= 0.95 \text{ for } A_2$$

It is important to emphazise here that TPR is the probability of an indication (POI) due to any souce when a discontinuity is present. This is distinct from the probability of an indication when a discontinuity exists given no noise, that is, the true probability of detection (POD) as will be shown later.

(4) $$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{N}} = \frac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}} = \Pr\left(\text{Positive Indication} \mid \text{No Discontinuity}\right)$$

$$= 0.01 \text{ for } A_2$$

where,

TP = number of True Positives (Hits)
FP = number of False Positives (False Calls)
TN = number of True Negatives (Correct Accepts)
FN = number of False Negatives (Misses)
P = number of discontinuities
N = number of non discontinuities
P + N = total number of inspection locations

The acronyms TPR and FPR are used here in keeping with historical work. The statistical nomenclature Pr(Positive Indication | Discontinuity) has been introduced and represents the probabililty of a positive indication given a discontinuity is present, the POI. Pr(Positive Indication |No Discontinuity) is the probabililty of a positive indication given no discontinuity is present, the false positive probability. These are conditional probability statements.

Choosing a site to be inspected at random, the probability that a discontinuity is present, Pr(Discontinuity), is given by,

(5) $$\text{Pr(Discontinuity)} = \frac{\text{P}}{\text{P + N}} = \frac{\text{TP + FN}}{\text{TP + FN + FP + TN}}$$

$$= 0.0001 \text{ for } A_2$$

The probability that no discontinuity is present, Pr(No Discontinuities), is given by,

(6) $$\text{Pr(No Discontinuity)} = 1 - \text{Pr(Discontinuity)} = \frac{\text{N}}{\text{P + N}} = \frac{\text{FP + TN}}{\text{TP + FN + FP + TN}}$$

$$= 0.9999 \text{ for } A_2$$

Other conditional probability statements may be made,

The probability of no indication given a discontinuity is present, a Miss,

(7) $$\text{Pr(Negative Indication | Discontinuity)} = \frac{\text{FN}}{\text{FN + TP}}$$

$$= 0.05 \text{ for } A_2$$

The probability of no indication given no discontinuity is present, a true negative,

(8) $$\text{Pr(Negative Indication | No Discontinuity)} = \frac{\text{TN}}{\text{TN + FP}}$$

$$= 0.99 \text{ for } A_2$$

Historically, at a particular detection threshold a single matrix (Figure 2a) is developed for each discontinuity size. The TPR and FPR points for the thresholds in Table 1 are shown as a shaded circles in the ROC chart (Figure 2b). If the discontinuity size is fixed, e.g., a_6 and the detection threshold value is allowed to vary then a series of paired values of TPR and FPR are generated that describe a TPR versus FPR (solid curve in Figure 2b).

TP, FP, FN, and TN events are all mutually exclusive classifications. A tabulation (Figure 2a) of positive and negative events (sums of rows) and the number of sites with and without discontinuities (sums of columns) highlights this exclusiveness for the $A_{th} = A_2 = 2$ example where there are 1,000,000 test locations. When $A_{th} = 2$,

(9) Pr(Positive Indication | Discontinuity) = 0.95

(10) Pr(Positive Indication | No Discontinuity) = 0.01

(11) Pr(Negative Indication | Discontinuity) = 0.05

(12) Pr(Negative Indication | No Discontinuity) = 0.99

(13) Pr(Discontinuity) = 0.0001

(14) Pr(No Discontinuity) = 1 - Pr(Discontinuity) = 0.9999

Similar ROC curves may be generated for different discontinuity sizes (discontinuity sizes a_1, a_2, a_3, a_4, and a_5) to generate a family of ROC curves (dashed curves in Figure 2b), from which TPR versus discontinuity size may be obtained at fixed FPR (Figure 2b).

Interrelationship of ROC with Joint Probability Matrices

There are two joint probability matrices of interest that may be generated to highlight conditional probabilities. The first joint probability matrix addresses probability of indications in the presence and absence of a discontinuity, and the second probability matrix assumes that a discontinuity exists and addresses the probability of indications in the presence and absence of noise.

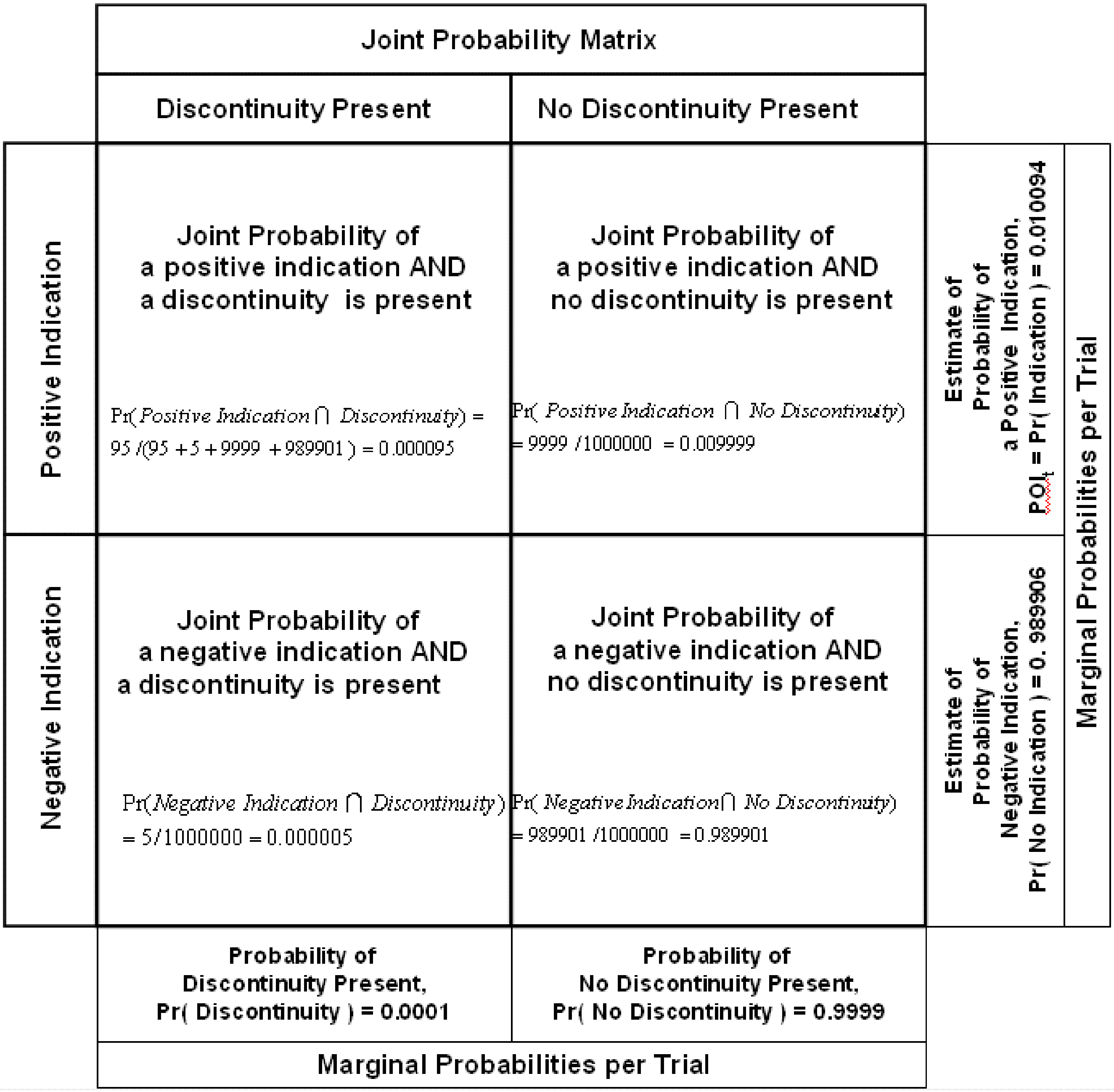
The first joint probability matrix (Figure 3) is constructed. from the ROC matrix (Figure 2a), where,

Pr (Positive Indication ∩ Discontinuity) is the joint probability of a positive indication and a discontinuity is present.

Pr(Positive Indication ∩ No Discontinuity) is the joint probability of a positive indication and a discontinuity is not present.

Pr(Negative Indication ∩ Discontinuity) is the joint probability of negative indication and a discontinuity is present

Pr(Negative Indication ∩ No Discontinuity) is the joint probability of negative indication and a discontinuity is not present.



Joint Probability Matrix				
	Discontinuity Present	No Discontinuity Present		
Positive Indication	Joint Probability of a positive indication AND a discontinuity is present Pr(*Positive Indication* ∩ *Discontinuity*) = 95 /(95 +5 +9999 +989901) = 0.000095	Joint Probability of a positive indication AND no discontinuity is present Pr(*Positive Indication* ∩ *No Discontinuity*) = 9999 /1000000 = 0.009999	Estimate of Probability of a Positive Indication, POI = Pr(Indication) = 0.010094	Marginal Probabilities per Trial
Negative Indication	Joint Probability of a negative indication AND a discontinuity is present Pr(*Negative Indication* ∩ *Discontinuity*) = 5/1000000 = 0.000005	Joint Probability of a negative indication AND no discontinuity is present Pr(*Negative Indication* ∩ *No Discontinuity*) = 989901 /1000000 = 0.989901	Estimate of Probability of Negative Indication, Pr(No Indication) = 0. 989906	
	Probability of Discontinuity Present, Pr(Discontinuity) = 0.0001	Probability of No Discontinuity Present, Pr(No Discontinuity) = 0.9999		
	Marginal Probabilities per Trial			

Figure 3. Joint Probability Matrix for A_2 case of Table1

The joint probabilities are directly obtained by dividing the number of events in the ROC matrix quadrants by the total number events of the matrix. For example, the joint probability Pr(Positive Indication ∩ Discontinuity) in the upper left quadrant of the joint probability matrix is obtained by dividing the number of events, 95, in the upper left quadrant of the ROC matrix by the total number of events in the ROC matrix (Figure 2a) , 1,000,000 to yield ,

(15) Pr(Positive Indication ∩ Discontinuity) = 0.000095,

the joint probability of an indication and a discontinuity is present. A similar procedure is done for the other quadrants of the joint probability matrix. Numerical estimates for these probabilities are shown in Figure 3. The marginal probabilities shown in the margins of Figure 3 for discontinuities and indications are obtained by summing the appropriate columns and rows. Marginal probabilities are per test location Marginal probabilities are not conditional probabilities. The conditional probability of a positive indication given a discontinuity of size a is present, POI(a), is obtained from joint probability Pr(Positive Indication ∩ Discontinuity) by dividing Pr(Indication ∩ Discontinuity) by the marginal probability that a discontinuity exists, Pr(Discontinuity)) to yield,

(16) $$\mathrm{POI} = \Pr(\text{ Positive Indication} \,|\, \text{Discontinuity }) = \frac{\Pr(\text{ Positive Indication} \cap \text{Discontinuity })}{\Pr(\text{ Discontinuity })}$$

$= 0.95$ for A_2

Similarly for the other quadrants in the joint probability matrix,

(17) $$\Pr(\text{ Positive Indication} \,|\, \text{No Discontinuity }) = \frac{\Pr(\text{ Positive Indication} \cap \text{No Discontinuity })}{\Pr(\text{ No Discontinuity })}$$

$= 0.01$ for A_2

(18) $$\Pr(\text{ Negative Indication} \,|\, \text{Discontinuity }) = \frac{\Pr(\text{ Negative Indication} \cap \text{Discontinuity })}{\Pr(\text{ Discontinuity })}$$

$= 0.05$ for A_2

and

(19) $$\Pr(\text{ Negative Indication} \,|\, \text{No Discontinuity }) = \frac{\Pr(\text{ Negative Indication} \cap \text{No Discontinuity })}{\Pr(\text{ No Discontinuity })}$$

$= 0.99$ for A_2

These conditional probabilities are identical to those obtained by using the TP, FP, FN, and TN relationships (Equations (3), (4), (7), and (8)) for the ROC method. Therefore, the relationship between ROC conditional probabilities and the joint probability matrix, is shown to be direct by combining the marginal probabilities with the joint probabilities. Given marginal probabilities, after getting one joint probability, the other joint probabilities can be obtained by subtraction. Marginal probabilities may also be derived from the conditional probabilities. For example, the probability of a positive indication per location is,

(20) $$\begin{aligned}\mathrm{POI} = \Pr(\text{ Positive Indication }) = {}& \Pr(\text{ Positive Indication} \,|\, \text{Discontinuity }) \cdot \Pr(\text{ Discontinuity}) + \\ & \Pr(\text{ Positive Indication} \,|\, \text{No Discontinuity }) \cdot \Pr(\text{ No Discontinuity})\end{aligned}$$

= 0.010094,

The marginal probability of negative indication per location is,

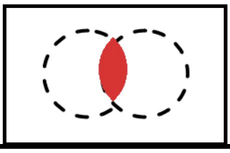
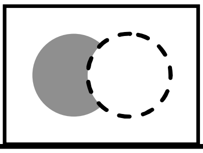
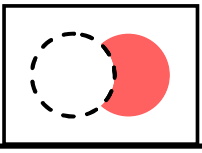
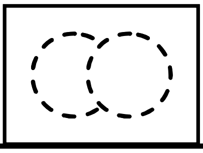
(21) $$\Pr(\text{ Negative Indication }) = 1 - \Pr(\text{ Positive Indication })\ .$$

The interrelationship between the ROC matrix shown in Figure 2a and the joint probability matrix shown in Figure 3 is an important one, and many interrelationships between probability related metrics may be established via the joint probability matrix once it is generated.

The second joint probability matrix (Figure 4) is constructed assuming that a discontinuity exists and addresses the probability of indications in the presence and absence of noise.

Joint Probability Matrix

Numerical Example:
Positive indication can be due to discontinuity (red) or noise (grey) or both.
For a given discontinuity size, under the assumption that discontinuity indications are independent of noise indications,

<table>
<tr><td></td><td>Positive Indication
from Discontinuity, <b>ID</b></td><td>Negative Indication
from Discontinuity, <b>ID^{C}</b></td><td></td></tr>
<tr><td>Positive Indication from Noise, <b>IN</b></td><td><b>Pr(ID and IN | A) =
Pr(ID ∩ IN | A)</b>
<b>Indication from both Noise and Discontinuity</b>
$\Pr(ID \cap IN|A) = \Pr(ID|A)\cdot\Pr(IN|A)$
$= 0.00949495$
</td><td><b>Pr(not ID and IN | A) =
Pr (ID^{C} ∩ IN | A)</b>
<b>Indication from Noise Only</b>
$\Pr(ID^C \cap IN|A) =$
$\Pr(IN|A) - \Pr(ID|A)\cdot\Pr(IN|A) = 0.00050505$
</td><td><b>Probability of Indication Due To Noise and Given a Discontinuity,
Pr(IN | A) = 0.01 = FPR</b></td></tr>
<tr><td>Negative Indication from Noise, <b>IN^{C}</b></td><td><b>Pr(ID and not IN | A) =
Pr(ID ∩ IN^{C} |A)</b>
<b>Indication from Discontinuity Only</b>
$\Pr(ID \cap IN^C|A) =$
$\Pr(ID|A) - \Pr(ID|A)\cdot\Pr(IN|A) = 0.94$
</td><td><b>Pr(not ID and not IN |A) =
Pr(ID^{C} ∩ IN^{C} | A)</b>
<b>Negative Indication</b>
$\Pr(ID^C \cap IN^C|A) = 1 - \Pr(ID|A) - \Pr(IN|A)$
$+\Pr(ID|A)\cdot\Pr(IN|A) = 0.05$
</td><td><b>Probability of Negative Indication Due to Noise and Given a Discontinuity,
Pr(IN^{C} | A) = 1- Pr (IN | A) = 0.99</b></td></tr>
<tr><td></td><td><b>Probability of a Positive Indication Due to Discontinuity and Given Noise, Estimate of Pr(ID | A) = 0.94949495</b></td><td><b>Probability of Negative Indication Due to Discontinuity and Given Noise, Estimate of Pr(ID^{C} | A) = 1- Pr(ID | A)
= 0.05050505</b></td><td></td></tr>
</table>

Note: Given marginal probabilities, after getting one joint probability, the other joint probabilities by subtraction

Figure 4. Joint Probability Matrix for A$_2$ case of Table 1 highlighting probability of indication due to discontinuity and probability of an indication due to noise.

We introduce some acronyms here for various events so that the statistical relationships are more readily interpreted:

ID = positive indication from discontinuity

ID^c = negative indication from discontinuity, c denotes complement

IN = positive indication from noise

IN^c = negative or no indication from noise

A = discontinuity is present

In order to construct the joint probability matrix shown in Figure 4, we note that the probability of a positive indication from noise is independent of whether a discontinuity is present and therefore the probability of indication from noise given a discontinuity is equal to the false positive probability, Equation (4),

(22) $\Pr(\, IN \big|\ A\,) = FPR$

$= 0.01$ for A$_2$

and the probability of a negative indication from either discontinuity or noise when a discontinuity is present, Equation (7),

(23) $\Pr(ID^c \cap IN^c \big| A) = \Pr(\text{ Negative Indication } | \text{ Discontinuity})$

$= 0.05$ for A$_2$,

are used.

Following the rules for joint probability matrices:

The marginal probabilities must equal the sum of the respective row or column quadrant probabilities. The sum of the marginal probabilities of the matrix rows must equal one, and the sum of the marginal probabilities of the matrix columns must equal one.

The sum of the respective rows are,

(24) $\Pr(ID \cap IN \big| A) + \Pr(ID^c \cap IN \big| A) = \Pr(\, IN \big|\ A\,)$

(25) $\Pr(ID \cap IN^c \big| A) + \Pr(ID^c \cap IN^c \big| A) = \Pr(IN^c \big| A) = 1 - \Pr(IN \big| A)$

and the sum of the respective columns are,

(26) $\Pr(ID \cap IN \left| A \right.) + \Pr(ID \cap IN^c \left| A \right.) = \Pr(ID \left| A \right.)$

(27) $\Pr(ID^c \cap IN \left| A \right.) + \Pr(ID^c \cap IN^c \left| A \right.) = \Pr(ID^c \mid A)$

The sum of the marginal probabilities of the matrix rows is,

(28) $\Pr(IN^c \mid A) + \Pr(IN \left| A \right.) = 1$

and the sum of the marginal probabilities of the matrix columns must equal one.

(29) $\Pr(ID \left| A \right.) + \Pr(ID^c \mid A) = 1$

If $\Pr(IN \left| A \right.)$ and $\Pr(ID \left| A \right.)$ events are independent then,

(30) $\Pr(ID \cap IN \left| A \right.) = \Pr(IN \left| A \right.) \cdot \Pr(ID \left| A \right.)$ and similarly

(31) $\Pr(ID^c \cap IN^c \left| A \right.) = \Pr(IN^c \mid A) \cdot \Pr(ID^c \mid A)$

Using equations (22), (23), (30), (31) and the above rules, we have the remaining probabilities to complete the matrix,

(32) $\Pr(ID \cap IN \left| A \right.) = 0.00949495$

(33) $\Pr(ID^c \cap IN \left| A \right.) = 0.00050505$

(34) $\Pr(ID \cap IN^c \left| A \right.) = 0.94$

(35) $\Pr(ID \mid A) = 0.94949495$

(36) $\Pr(ID^c \mid A) = 0.05050505$

The probability of an indication due to any souce when a discontinuity is present is obtained by summing three quandrants of the joint probability matrix having indications to yield,

(37) $\Pr(ID \cap IN \left| A \right.) + \Pr(ID^c \cap IN \left| A \right.) + \Pr(ID \cap IN^c \left| A \right.) = TPR$

$= 0.95$ for A_2

and this is the same value obtained by using Equation (3).

Also note that assuming independence of indications from the two sources, discontinuities and noise, then equation (37) may be written as,

(38) $TPR = \Pr(ID \left| A \right.) \cdot \Pr(IN \left| A \right.) + \Pr(IN \left| A \right.) - \Pr(ID \left| A \right.) \cdot \Pr(IN \left| A \right.) + \Pr(ID \left| A \right.) - \Pr(ID \left| A \right.) \cdot \Pr(IN \left| A \right.)$

(39) $TPR = \Pr(ID \left| A \right.) + \Pr(IN \left| A \right.) - \Pr(ID \left| A \right.) \cdot \Pr(IN \left| A \right.)$

or

(40) $TPR = POI = POD + FPR - POD \cdot FPR$ also shown in Petrin, Annis, & Vukelich, (1993)

where,

(41) $POD = \Pr(ID|A)$, the probability of an indication solely due to the presence of a discontinuity and not due to noise, as defined by Petrin, et al., (1993), and

(42) $FPR = \Pr(IN|A)$, the probability of an indication due to noise when a discontinuity exist.

Importance of False Positives

Before continuing developing the interrelationship between additional POD-related metrics, it is important to further discuss the designation of a false positive.

The mechanism creating false positives may be random or systematic. The mechanism may be due to electronic or other noise in the inspection system due to equipment, process control, environment, and human factors, such as fatigue, etc. At first evaluation one may believe that having a non-zero false positive probability only affects costs and this is anecdotally highlighted by having a manager perform a test to estimate POD. Components with and without discontinuities are given to the manager for testing. Suppose that 50% of the components have a critical discontinuity, while the rest of the components are without discontinuities. The manager's test results indicate that all components have discontinuities. Since the manager found all of the components with discontinuities, then the estimated POD capability of the manager is 1.0 (Equation (3)). Further, the manager found discontinuities on all the components without discontinuities to yield a probability of false positive of 1.0 (Equation (4)). The manger rejects all components all the time and this is very costly.

There is another side to this issue. The mechanism creating false positives may also inhibit or enhance the true test response (response from a discontinuity only). A dramatic, but not common, physical example is when the sensing system generates a large signal every twenty-ninth measurement. The presence of this mechanism increases both FP and TP values as a positive indication is being made independent of the presence of a discontinuity. The ramification here is that with high noise levels TP is inflated due to mechanisms creating false positives, and the estimate of POD is no longer accurate. Fracture critical inspections require the level of true POD to support qualification of inspection systems and personnel. Any artificial inflation of POD due to mechanisms creating FP must be addressed where an artificial inflation POD now affects both costs and life. There are multiple ways to address an artificially inflated POD. One approach is to increase the threshold detection level (upper horizontal line in Figure 1) in an effort minimize false positives, another method is to utilize the relationship developed by Petrin, C.,et al., (1993) to correct the artificial inflated POD.

An issue arises in practical applications where FPR is rarely zero, so that the degree of inflation for binomial point estimates of POD needs to be evaluated. Twenty-nine Hits out of twenty-nine trials are required to demonstrate a 90/95 POD capability at a critical discontinuity size when the POD is also verified to be monotonic for all discontinuities sizes greater than the critical discontinuity size (NASA, 2008). It has been shown that if there is one “lucky Hit” per 29 trials, then the presence of this “lucky Hit” yields an inflated POD. The number of “lucky hits” in an inspection would be random, but the average number is one when FPR is 3.44%. Generazio (2009, 2011) has proposed that when FPR is less than or equal to 3.44% at 95% confidence, then the binomial estimate of POD is an adequate methodology. This adequate region is shown as region A in Figure 2b. Note that in order to have a 95% confidence interval on FPR to be as low as 0.0344 when no false calls are made requires at least 86 test sites without discontinuities.

Binomial Estimate of POD and FPR Calculations for Single Discontinuity Size

The binomial method (Rummel, 1982) for estimating POD has been used for decades by NASA for the inspection of Space Shuttle, International Space Station (ISS), and satellite systems, and this method is currently used at NASA the next generation of space systems. The binomial method is fairly straight forward, and is usually applied when determining point estimates of POD for a given discontinuity size. The data requirements for using the binomial methodology for NASA applications are described elsewhere [Generazio, 2009 & 2011]. If the false call probability is significant then a binomial test for estimating POD yields an inflated POD (Generazio, 2009 & 2011) or the POI.

In the following discussion two different mathematical approaches are used to determine binomial point estimates of POD and the lower confidence bound on POD. The two methods are included here to cover the use of lookup tables that are included in currently accepted Standards and to cover the more modern approach, identified in italics, that utilizes the ready access to computational power for evaluating complex integrals.

The binomial point estimate of POD is given by the ratio of the number Hits to the number of sites with discontinuities tested,

$$(43)\ \Pr(\text{ Positive Indication}\,|\,\text{Discontinuity }) \ = \frac{\text{Number of Hits}}{\text{Number of Test Sites with Discontinuities}}$$

$$= \frac{95}{100} = 0.95 \text{ for } A_2$$

The ROC display methodology uses the same relationship (Equation (3)) for binomial estimation of POD.

The one-sided lower confidence bound on POD may be determined by the ”exact” Clopper-Pearson method (Clopper & Person, 1934) based on the Binomial distribution. This procedure is referred to as ”exact” because it uses the Binomial distribution and finds the probability points that satisfies the coverage (i.e., confidence level) statement. However, for other values of the POD the actual coverage probability can be much larger than nominal confidence level and for this reason the Clopper-Pearson interval may be considered ”wastefully conservative” (Brown, Cai & DasGuupta, 2001).The lower confidence bound is given by the Clopper-Pearson relationship (Rummel, 1982) uses a lookup table and is given by,

(44) $\mathrm{LCB} = \dfrac{\mathrm{TP}}{\mathrm{TP} + (\mathrm{FN} + 1) \times \mathrm{F}_\gamma(\mathrm{f}_1, \mathrm{f}_2)},$

$\text{where } \mathrm{F}_\gamma(\mathrm{f}_1, \mathrm{f}_2) = 1.83 \text{ when } \left\{\begin{matrix} \mathrm{f}_1 = 2 \times (\mathrm{FN} + 1) \\ \mathrm{f}_2 = 2 \times \mathrm{TP} \end{matrix}\right\}$ and γ = 0.95

Here the F-distribution quantile $\mathrm{F}_\gamma(\mathrm{f}_1, \mathrm{f}_2)$ is obtained from the F-distribution statistical table (Weast, 1970), and γ is the nominal confidence level that is being required.

An equivalent equation utilizing the Beta distribution is given by

(45) $LCB = \boldsymbol{I}^{-1}(1-\gamma; TP, FN+1)$*, the inverse of the incomplete Beta function given by*

(46) $I(x;a,b) = \dfrac{B(x;a,b)}{B(1;a,b)}$*, where* $B(x,a,b) = \int_0^x t^{a-1}(1-t)^{(b-1)}\,dt$ *.*

There is no theoretical difference for the results of equations (44) and (45) when the exact values for the given F distribution are known. However, the use of tables will often require interpolation or bounding values to be used.

Using the TP and FN values shown in Figure 2a, and γ = 0.95, we have for equation (44), the 95% lower confidence bound on Pr(Positive Indication | Discontinuity) is

(47) $\mathrm{LCB} = 0.896$

or more rigorously from Equation (45),

(48) $LCB = \boldsymbol{I}^{-1}(0.05; 95, 6) = 0.8977$

For this conservative Clopper-Pearson 95% confidence interval procedure and discontinuity size, if LCB POD is 0.90, then this confidence level procedure has a probability of at least 0.95 to give a one-sided lower confidence bound for the POD point that exceeds the true (unknown) 90% POD point. In general, the smallest discontinuity size where the one-sided lower 95% confidence bound on POD exceeds 0.90 is referred to as the 90/95 POD point or the $a_{90/95}$ discontinuity size. The symbolic ratio 90/95 refers to the one-sided lower confidence bound of 0.90 on the estimated POD and the 95 refers to the confidence level, 95%. When including the confidence interval statement, the discontinuity size using this criterion may now be designated as $a_{90/95}$. Fracture critical applications generally require that inspection systems demonstrate capability by showing that the 95% one-sided lower bound on POD at the critical discontinuity size, $a_{90/95}$, be 0.9 or greater.

The binomial point estimate of the probability of a false call is given by the ratio of the number false positives to the number of sites without discontinuities tested, e.g., from Figure 2a,

(49)

$$\Pr(\text{ Positive Indication} \,|\, \text{No Discontinuity }) \;=\; \frac{\text{Number of False Calls}}{\text{Number of Test Sites without Discontinuities}}$$

$$= \frac{9999}{999900} = 0.0100$$

Bounds for the FPR follow the same rationale as that for the POD but now an one-sided upper bound is required rather than a lower bound and the rates are with respect to the FP and TN, that is the non-flawed specimens. The 95% one-sided upper confidence bound on FPR, is given by (Rummel, 1982),

$$(50)\,UCL = \frac{(FP+1)\cdot F_{\gamma}(f_1, f_2)}{(TN)+(FP+1)\cdot F_{\gamma}(f_1, f_2)},$$

where $F_{\gamma}(f_1, f_2) = \left\{ \begin{matrix} f_1 = 2\cdot(FP+1) \\ f_2 = 2\cdot(TN) \end{matrix} \right\}$ and using data from Figure 2a, and $\gamma = 0.95$ we have

$$\mathbf{(51)}\,UCL = \frac{(10000)\cdot F_{0.95}(f_1, f_2)}{(989901)+(10000)\cdot F_{0.95}(f_1, f_2)},$$

where $F_{0.95}(f_1, f_2) = 1,$ with $\left\{ \begin{matrix} f_1 = 2\cdot(10000) \\ f_2 = 2\cdot(989901) \end{matrix} \right\}$ and using $FP+1$=10000, $TN = 989901$,

We have the 95% one-sided upper confidence bound on FPR,

(52)UCL = 0.0100

or more rigorously,

$$\textit{(53)}\,UCL = I^{-1}(\gamma; FP+1, TN)$$

$$\textit{(54)}\,UCL = I^{-1}(0.95;10000,989901) = 0.0102$$

These UCL values, equations (52) and (54), are very nearly the same as the estimated FPR of 0.0100 due to the large amount of data. Binomial estimation data and methods can be used to generate ROC matrices from the elementary counts of TP, FP, TN, and FN taken at several different discontinuity sizes, and subsequently estimate TPR, FPR, and their companion 95% confidence bonds.

An important issue arises with large data sets when using look up tables. For large data sets, the F-distribution quantile, $\mathrm{F}_{\gamma}(\infty, \infty)$ might be selected from a look-up table. In this case $F_{\gamma}(\infty, \infty)$ has a value of one and is independent of the confidence level γ, and this highlights the importance of using more exact methods.

The reader should be aware that software programs are available that automatically calculate the exact confidence intervals (Pezzullo 2010).

Binomial Method for Multiple Flaw Sizes

For fracture critical applications, the POD must exceed 0.90 with 95% confidence for all discontinuity sizes larger than a binomial point estimate of POD at $a_{90/95.}$ In order to verify that this is true, the binomial method may be applied in an iterative fashion (Generazio 2011) for different discontinuity sizes and grouping of discontinuity sizes to estimate the POD as a function of discontinuity size. It has been shown that under strict testing protocol, that if a 95% one-sided lower bound of single binomial point estimate of POD meets or exceeds 0.90 at a discontinuity size, $a_{90/95}$, then the estimate of POD may be extended to larger discontinuity sizes when an adequate number (25) of large discontinuities are included in the test so that the POD is verified to be monotonic above the $a_{90/95}$ point (Generazio 2011). The general procedural for validating monotonicity of POD above the $a_{90/95}$ point is mathematically simple but mathematically intensive (Generazio 2009, 2011, 2012) for binomial estimation of POD. The reader is referred to existing cited publications detailing the procedure.

One and Two-Sided Confidence Bounds, and Confidence Intervals

Before moving into parametric curve fitting methods it is important to highlight the differences in confidence statements. Figure 5 shows a likelihood function for a quantity, X . The confidence that X is between the two-sided lower and upper bounds is 90%. Alternatively, a one-sided lower bound is also identified for which X meets or exceeds the one-sided lower bound is 95%. Where now the upper tail of the likelihood function is also included in determining the confidence. A similar statement may be made of 95% one-sided upper bound. The two-sided bounds contain an interval as determined from the data, while the one-sided bounds have one end of the interval open to all possible values of X in the given direction. Figure 5 is notionally showing a normal symmetric function, however, likelihood functions generally are asymmetric.

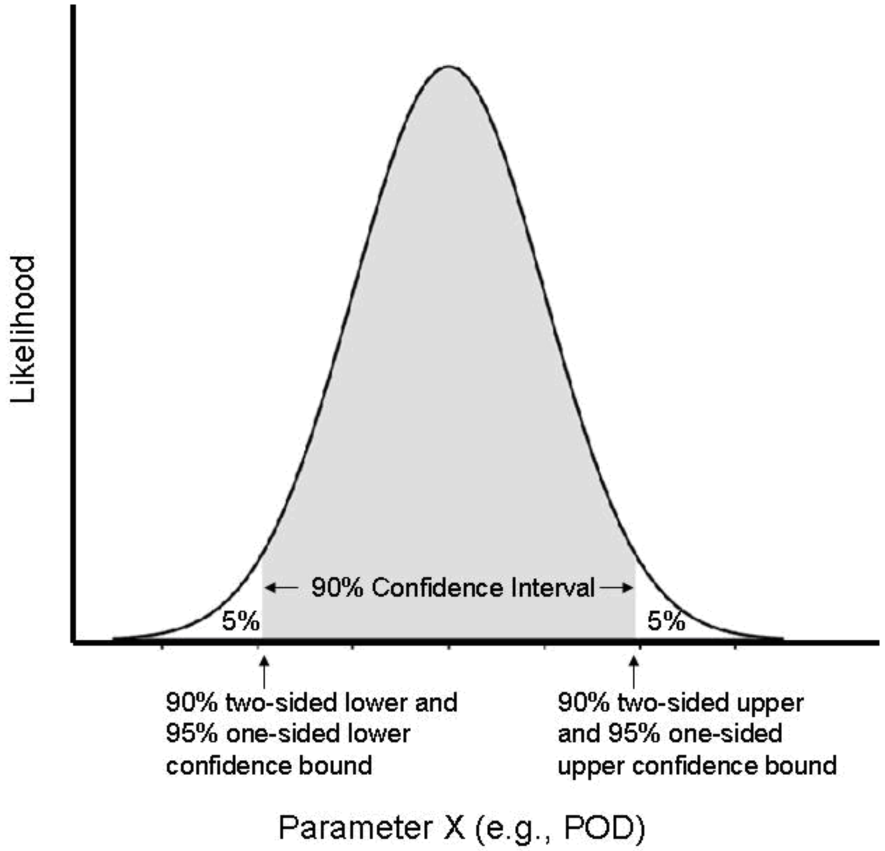


Figure 5. Likelihood as a Function of Parameter X

Multi-parameter Logistic Maximum Likelihood Method (Logit-ML)

Before engaging in the detailed calculations the need for logistic regression, the selection of points, and criteria for the acceptable POD versus size need to be identified. The original need for the use of logistics regression in estimating POD came into being when it was observed that test specimens were expensive and having hundreds of test specimens for each structural and material configuration was not a viable option. Best efforts were made to produce specimens having discontinuity sizes in the range of interest, however, since POD testing is often done to establish the capability of the inspection system for applications to a specific material and structural configuration, the number of samples and discontinuity sizes produced falls short of being the optimum selections. Generazio (2011), and Wald (1943) have shown that by using sequential statistics, sample sets selection may be optimized to be the most efficient for binomial analysis. However, a proven methodology for establishing optimum POD sample sets for logistic regression has not been shown. The NDE community uses 0.9 POD at critical discontinuity size with 95% confidence (90/95 POD) as the target inspection capability. The origin of this 90/95 POD requirement is from damage tolerance fracture analysis for fracture critical flight hardware. It is noted here that for logistics regression, it is assumed that the POD is always increasing with discontinuity size. In contrast, for binomial analysis this assumption is not made so that verification that the POD is increasing with discontinuity size is required (NASA 2008 and Generazio 2011) when accepting binomial estimates of POD capability.

The two parameter Logistic statistical binary regression model assumes that the POD is always increasing with discontinuity size and is commonly written as,

$$P(a) = \frac{e^{\alpha + \beta \ln(a)}}{1 + e^{\alpha + \beta \ln(a)}} \tag{55}$$

where a is discontinuity size, and α and β are the two parameters that are to be estimated using maximum likelihood procedures. Although this function describes a cumulative distribution, this function should not be confused with cumulative probability functions (Hald, 1952) that are often used and derived from binomial, normal, or Weibull probability distributions of a measured values, such as time to failure or bearing diameters. Equation (55) is not a cumulative POD function, as the discontinuity size is not a random variable.

The term likelihood refers to the likelihood of data given parameters, and in this case parameters α and β. The likelihood is proportional to density for continuous random variables. For a discrete random variable the likelihood is the probability of the outcome. Therefore, for a single inspection the likelihood is either the probability of a hit or the probability of a miss. Assuming each inspection outcome is independent of the other inspection outcomes the total likelihood is the product of the individual inspection likelihoods.

The likelihood function for or binary Hit-Miss data is,

$$(56)\quad L(\alpha\,,\beta\,) = \prod_{i=1}^{N} (\,P(\,a_i\,)\,)^{d_i}(\,1 - P(\,a_i\,)\,)^{(\,n_i\,-d_i\,)}$$

where N is the number of different discontinuity sizes, d_i is the number of detections at the discontinuity size a_i, and n_i is the number of discontinuities of size a_i and the probability of detection, P, is given by equation (55).

The maximum of L, when it exists, occurs at the same point as maximum of the natural logarithm of L, $\mathrm{Ln}(L)$, and using this relationship makes the math somewhat easier. The maximum value of

$$(57)\quad LL = \mathrm{Ln}\left[L(\,\alpha,\,\beta\,)\right]$$

is what is needed. We have,

$$(58)\quad LL = \mathrm{Ln}\left[\prod_{i=1}^{N} (\,P(\,a_i\,)\,)^{d_i}(\,1 - P(\,a_i\,)\,)^{(\,n_i\,-d_i\,)}\right]$$

$$(59)\quad LL = \sum_{i=1}^{N} d_i \cdot \ln\left[P(\,a_i\,)\right] + (\,n_i\, - d_i\,)\ln\,\left[1 - P(\,a_i\,)\,\right]$$

and for the logistic model,

$$(60)\quad LL = \sum_{i=1}^{N} d_i \cdot \ln\left[\frac{e^{\alpha + \beta\,\ln(\,a_i\,)}}{1 + e^{\alpha + \beta\,\ln(\,a_i\,)}}\right] + (\,n_i\, - d_i\,)\ln\left[1 - \frac{e^{\alpha + \beta\,\ln(a_i)}}{1 + e^{\alpha + \beta\,\ln(a_i)}}\right]$$

The parameters are α and β , and these parameters are to be varied to obtain the maximum LL . There are a variety of ways to find this maximum. It is important that the reader does not get lost in the mathematical procedures. Some maximization methods are more complex than others exhibiting more efficiency, including procedures for solving simultaneous equations that may or may not yield a solution, or methods that start with trial values of α and β and adjusts these parameter values by non-uniform amounts to minimize the variance between successive estimates of the derivatives $\frac{\partial LL}{\partial \alpha}$ and $\frac{\partial LL}{\partial \beta}$. At the true maximum likelihood point, $\frac{\partial LL}{\partial \alpha}$ and $\frac{\partial LL}{\partial \beta}$ are simultaneously exactly zero, if the maximum point is in the interior of the parameter space. Other methods are more straight forward, easy to implement, and always yield a solution. One simple method is to perform a grid search that simply varies the values of the two parameters over a wide range of possibilities until the maximum LL is observed. The grid mesh size may be adjusted to achieve better precision of the estimated parameters. A grid search will be used in this work as it may be readily implemented by the reader without relying on proprietary software generated by others. Once the values of the parameters at the maximized LL are known, $\hat{\alpha}$ and $\hat{\beta}$, then the estimated POD model (Logit-ML) is given by,

(61) $$\mathrm{P}(\,\mathrm{a}\,) = \frac{\mathrm{e}^{\hat{\alpha} + \hat{\beta}\ln(\,\mathrm{a}\,)}}{1 + \mathrm{e}^{\hat{\alpha} + \beta\ln(\,\mathrm{a}\,)}}$$

for all discontinuity sizes, a .

There are several ways to determine confidence intervals and we have discussed the Clopper-Pearson method for a single binomial parameter. Other popular methods include the Wald and relative likelihood methods. These methods have different properties and yield credible representative intervals depending on the character of the original data. There is general agreement that the likelihood ratio confidence interval procedure is better that the Wald approach but the likelihood ratio method is computationally more difficult. The Wald approach, essentially uses a quadratic approximation to the log likelihood to simplify computations (Meeker and Escobar, 1995).

Relative Likelihood Confidence Interval Method

Before establishing confidence bounds using the relative likelihood confidence interval method it is warned here that the sample size is critical in establishing the credibility of the intervals obtained. For small sample sizes, e.g., less than 100, obtained confidence bounds may be in error. Harding and Hugo (2003) examine the relative likelihood confidence interval method and provide guidance for when the method is applicable for a given sample size.

The likelihood confidence intervals may be obtained from the relative likelihood ratio (Meeker and Escobar, 1998, chapter 8).

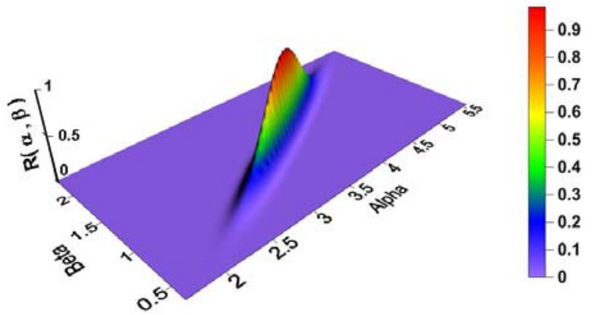
(62) $$R(\alpha , \beta) = \frac{L(\alpha , \beta)}{L(\hat{\alpha} , \hat{\beta})}$$

or

(63) $$R(\alpha , \beta) = \frac{\exp\left(\sum_{i=1}^{N} d_i \cdot \ln\left[\frac{e^{\alpha + \beta\ln(a_i)}}{1 + e^{\alpha + \beta\ln(a_i)}}\right] + (n_i - d_i)\ln\left[1 - \frac{e^{\alpha + \beta\ln(a_i)}}{1 + e^{\alpha + \beta\ln(a_i)}}\right]\right)}{\exp\left(\sum_{i=1}^{N} d_i \cdot \ln\left[\frac{e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}{1 + e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}\right] + (n_i - d_i)\ln\left[1 - \frac{e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}{1 + e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}\right]\right)}$$

Note that R is a ratio of likelihoods - not log-likelihoods, and that the denominator is a constant since $\hat{\alpha}$ and $\hat{\beta}$ are known from the estimated model.

A three dimensional surface plot of R(α, β) and contours at constant values of R(α, β) are shown in figure 6a and b, respectively, for typical data set. The maximum of R(α, β) is 1.0 and occurs at R($\hat{\alpha}$, $\hat{\beta}$). From figure 6b it is observed that there is a large selection of α, β pairs that may be used to generate a constant value of R(α, β).



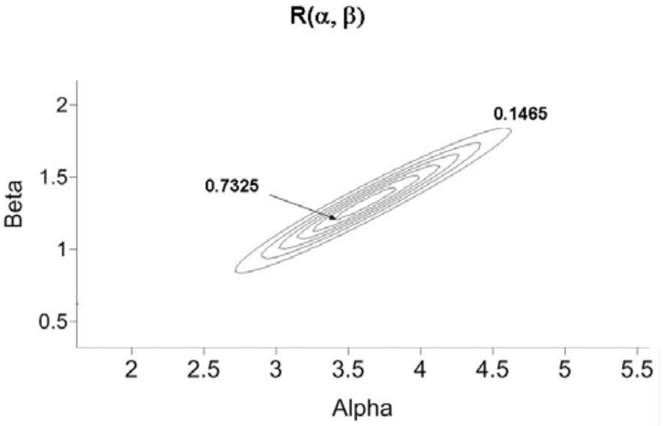


Figure 6a Relative Likelihood Ratio **Figure 6b Contours of Relative Likelihood Ratio**

In order to define the likelihood ratio confidence procedure for determining the confidence interval for P_0, it is helpful to perform a change of variables, replacing α, with

$$\alpha = \ln\left(\frac{P_0}{1 - P_0}\right) - \beta\ln(a_0),$$

where P_0 is the value of POD at a fixed discontinuity size a_0. This change replaces the parameter α with the new parameter P_0 and allows the determination of confidences bounds on P_0 (Meeker and Escobar, 1998, chapter 8, pages 182-183). Therefore, the re-parameterized likelihood ratio

(64)

$$R(P_0, \beta) = \frac{\exp\left(\sum_{i=1}^{N} d_i \bullet \ln\left[\frac{e^{\ln(\frac{P_0}{1-P_0}) - \beta\ln(a_0) + \beta\ln(a_i)}}{1 + e^{\ln(\frac{P_0}{1-P_0}) - \beta\ln(a_0) + \beta\ln(a_i)}}\right] + (n_i - d_i)\ln\left[1 - \frac{e^{\ln(\frac{P_0}{1-P_0}) - \beta\ln(a_0) + \beta\ln(a_i)}}{1 + e^{\ln(\frac{P_0}{1-P_0}) - \beta\ln(a_0) + \beta\ln(a_i)}}\right]\right)}{\exp\left(\sum_{i=1}^{N} d_i \cdot \ln\left[\frac{e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}{1 + e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}\right] + (n_i - d_i)\ln\left[1 - \frac{e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}{1 + e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}\right]\right)}$$

is a function of P_0 and β and is shown in Figure 6c.

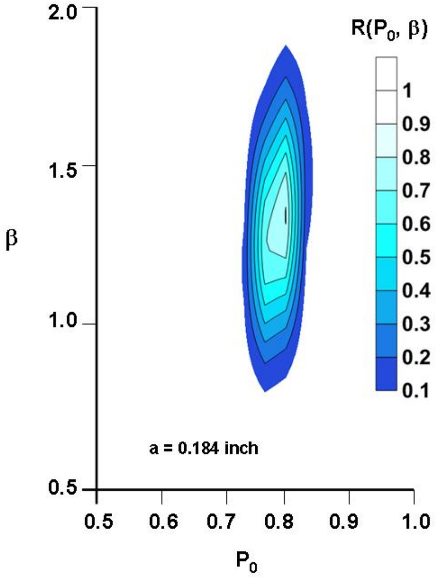


Figure 6c. The parameterized likelihood ratio, R (P_0, β) when a_0 = 0.184.

Using this function, for given values of a_0 one can find a confidence interval for P_0 by maximizing out the nuisance parameter beta.

Operationally, for given a_0, β may be varied to obtain a maximum value of R, $R(P_0)$ the profile likelihood, where

(65) $$R(P_0) = \max_{\beta}\left[R(P_0, \beta)\right]$$

Once the profile likelihood is established the two-sided interval with 95% confidence may be determined as those values of P_0 where

(66) $$R(P_0) > e^{-\left[\frac{\chi^2_{(0.95:1)}}{2}\right]} = 0.1465001,$$

is true. Here, for a 95% two-sided confidence interval, $\chi^2_{(0.95:1)}$ is the 0.95 quantile of the chi-square value with one degree of freedom, [Hines & Montgomery, 1972; and Weast, 1970]. The 95% confidence interval is obtained by determining the P_o values where $R(P_o)$ > 0.1465001.

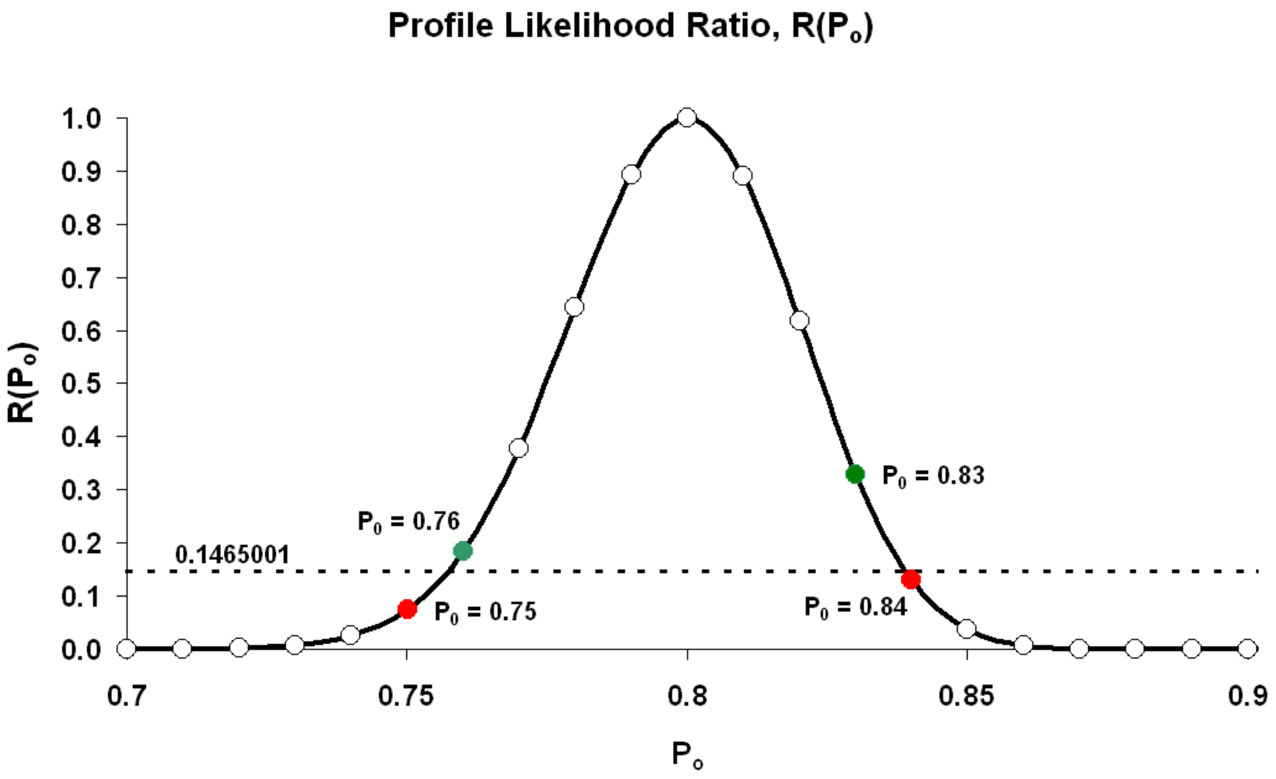


Figure 6d Profile likelihood ratio $R(P_0)$ **as a function of P_o.**

The two-sided upper (0.83) and lower (0.76) bounds are shown in Figure 6d where

(67) $R(P_0) > e^{-\left[\frac{\chi^2_{(0.95:1)}}{2}\right]} = 0.1465001$ for P_0 between these bounds.

Again, it is important to note that we are working with a two-sided confidence interval, that has upper and lower bounds (or limits), while a one-sided confidence “interval” only has a single upper or lower bound (or limit). For example the endpoints of the interval for P_0 that satisfy

(68) $R(P_0) > e^{-\left[\frac{\chi^2_{(0.90:1)}}{2}\right]} = 0.2585227$

establish a 90% two-sided confidence interval. However, assuming that half of the remaining 10%, or 5% is assigned to each of the tails, the lower limit of the interval can be taken as a 95% one-sided lower bound and the upper limit of the interval as a 95% upper bound.

The confidence statements need to clearly specify if they are one-sided (bound) or two-sided interval (with bounds or limits), and the level of confidence.

In an effort to provide some intuition on what is happening when determining the profile likelihood, an interim step will be discussed. Here $R(P_0, \beta)$ will be constrained by using a fixed $a_0 = 0.184$, holding P_0 at a few selected values (0.75, 0.76, 0.80, 0.82, 0.83) while allowing β to vary. In this manner separate profiles of the relative likelihood may be generated for each P_0.

Typical $R(P_0, \beta)$ curves with $a_0 = 0.184$ are shown in Figure 6e. The individual maximums of each profile likelihood $R(P_0, \beta)$ are the $R(P_0)$ for the P_0 listed in Figure 6e. Since $P_0 = 0.80$ is the Logit ML estimate of POD at $a_0 = 0.184$, then $R(0.80) = 1$. Given a particular $P_{0,}$ if the maximum of the profile likelihood exceeds 0.14650001, then that P_0 is within the 95% confidence interval. By examining the $R(P_0, \beta)$ at various $P_{0\,,}$ the bounds of the confidence interval are estimated. For example, when $P_0 = 0.75$ or $P_0 = 0.84$, $R(P_0, \beta)$ (red curves in Figure 6e) does not exceed 0.14650001. When $P_0 = 0.76$ or $P_0 = 0.83$, $R(P_0, \beta)$ (green curves in Figure 6e) does exceed 0.14650001, therefore there is 95% confidence that the true probability is within these bounds, so that $P_0 = 0.76$ or $P_0 = 0.83$, define the 95% confidence interval at $a_0 = 0.184$. The solid curve is $R(0.8, \beta)$. This process is continued for all P_0 to yield the full profile likelihood (solid curve in Figure 6d) at $a_0 = 0.184$ from which the 95% confidence interval may be estimated. The red and green markers in Figure 6d refer to the peak values in Figure 6e. Continuing in the same fashion for all values of a_0, the Logit-ML and respective relative likelihood confidence intervals are estimated as shown in Figure 7. A finer grid mesh size may be used to yield confidence bounds closer to 0.1465001. We see that constrained maximum likelihood estimation is integral to establishing relative likelihood confidence intervals.

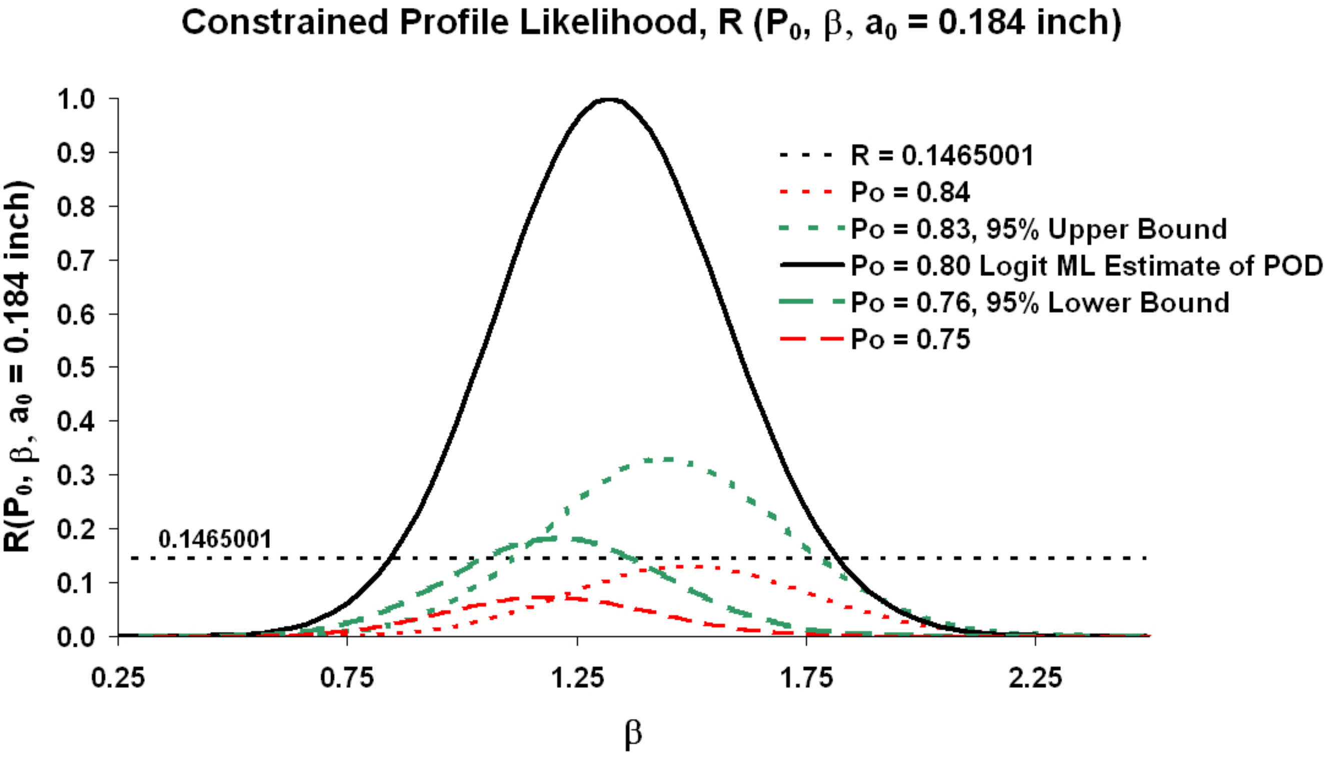


Figure 6e. Constrained profile likelihood ratio as a function of β and five different P_o values . The constraint is at one flaw size. The maximum of the lower green curves (P_o = 0.76 and 0.83) exceed $R\ (\ P_0\ ) > e^{-\left[\frac{\chi^2_{(0.95:1)}}{2}\right]} = 0.1465001$ and estimate the upper and lower bounds with 95% confidence. In contrast, the maximum of the lower red curves (P_o = 0.75 and 0.84) do not exceed $R\ (\ P_0\ ) > e^{-\left[\frac{\chi^2_{(0.95:1)}}{2}\right]} = 0.1465001$.

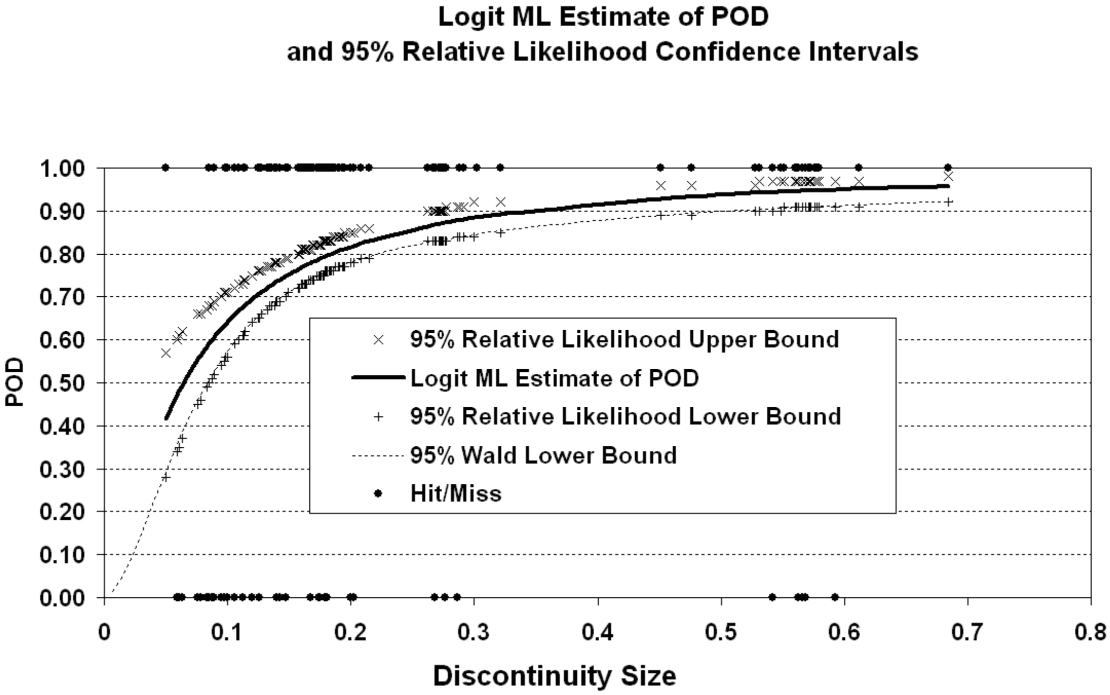


Figure 7. Logit ML estimate of POD (solid curve), relative likelihood two-sided upper and lower bounds (X and +), and the Wald one-sided lower bound with 95% confidence (dashed curve). Hits and Misses are 1 and 0, respectively.

The maximum likelihood method is invariant and the previous change of variables could just as well have been to replace the parameter α by fixing P_0 at 0.9 and considering the discontinuity size, a_0, as the new parameter. This change allows the determination of confidences bounds on a_0 (Meeker and Escobar, 1998, chapter 8, pages 182-183). The implication is that the bounding curves can interchangeably be considered as bounds on probability of detection at fixed flaw sizes (vertical bounds) or bounds on flaw sizes at fixed probabilities (horizontal bounds) This relationship is not generally true for Wald confidence bounds.

Wald Confidence Interval Method

The Wald method (Wald, 1943) for determining confidence intervals may also be used. The Wald method is based on estimating the variance-covariance matrix of the parameters using the inverse of the observed information matrix. The Wald confidence interval can be viewed as an approximation of the likelihood interval based on a quadratic approximation to the log likelihood profile function (Meeker and Escobar 1995). The Wald method is also known as the normal approximation method for establishing confidence intervals. It is instructive to show that confidence intervals quantify the uncertainty of the estimated POD in a manner similar to that used for propagation of errors for a function of multiple variables. In both cases the variances and covariances of the input quantities are needed. For the two parameter logit function, the variance relation and an estimate of the variance of $\hat{\alpha}$, variance of $\hat{\beta}$ and the covariance $\hat{\alpha}$ and $\hat{\beta}$, labeled $(\alpha)_{\text{var}}$, $(\beta)_{\text{var}}$, and $(\alpha\beta)_{\text{covar}}$, respectively, are needed.

The Wald lower 95% confidence bound (Figure 7, dashed curve) on $\hat{\text{P}}(\text{a})$ is obtained (see Appendix A) as a function of variances,

(69) $$\text{Wald}_{\text{LCB}}(\,\hat{\text{P}}(\text{a})\,) = \frac{e^{\hat{\alpha} + \hat{\beta}\ln(\text{a}) - 1.64\sqrt{(\alpha)_{\text{var}} + 2\ln(\text{a})\ (\alpha\beta)_{\text{covar}} + \ln(\text{a})^2\ (\beta)_{\text{var}}}}}{1 + e^{\hat{\alpha} + \hat{\beta}\ln(\text{a}) - 1.64\sqrt{(\alpha)_{\text{var}} + 2\ln(\text{a})\ (\alpha\beta)_{\text{covar}} + \ln(\text{a})^2\ (\beta)_{\text{var}}}}}\ .$$

The prior discussion primarily addressed the procedural steps highlighted by heavy borders in Figure 8. It is important to realize that even though an estimated POD model is obtained, the estimated model may be inadequate (Agresti, 2002). This is a critical area and it is often assumed that confidence interval statements are tests for adequacy of the model. Confidence interval statements indicate the statistical uncertainty due to sampling variability of the data, but assume the POD model form is given. For large samples, the width of a confidence interval will shrink to zero. However, confidence statements do not address whether the estimated POD model is adequate.

There is confusion about the acronym 90/50 that is often quoted. This acronym is a misnomer, where the original intent was to indicate the 0.9 POD point estimate, and therefore the 0.9 POD point estimate should be used without the 90/50 acronym. For Figure 7, the 0.9 POD point estimate is 0.342.

The 0.9 POD point estimate is independent of the confidence bound procedure. It has been suggested that a 0.9 POD point estimate be cited as an inspection capability requirement, however, unless you have a sufficient amount of data, a point estimate can be misleading and sometimes seriously so. A 0.90 POD requirement level may be demonstrated by using test samples with only ten discontinuities and allowing one Miss. The 0.9 POD point estimate and the 90/95 POD reflect an estimate and the uncertainty of that estimate, however, with more data these values will be closer together. The confidence value reflects the degree of conservativeness and there is no difference in acceptance of identical POD estimates based on 10 discontinuities of data versus the one based on 100 discontinuities if the confidence limits (bounds) are the same.

However, we know that the calculated bounds will be closer to the estimates with increased data, for a given confidence level. In practice, if very large bounds exist about a POD estimate, then accepting the POD estimate to validate the capability of the inspection system for fracture critical application is done with high risk.

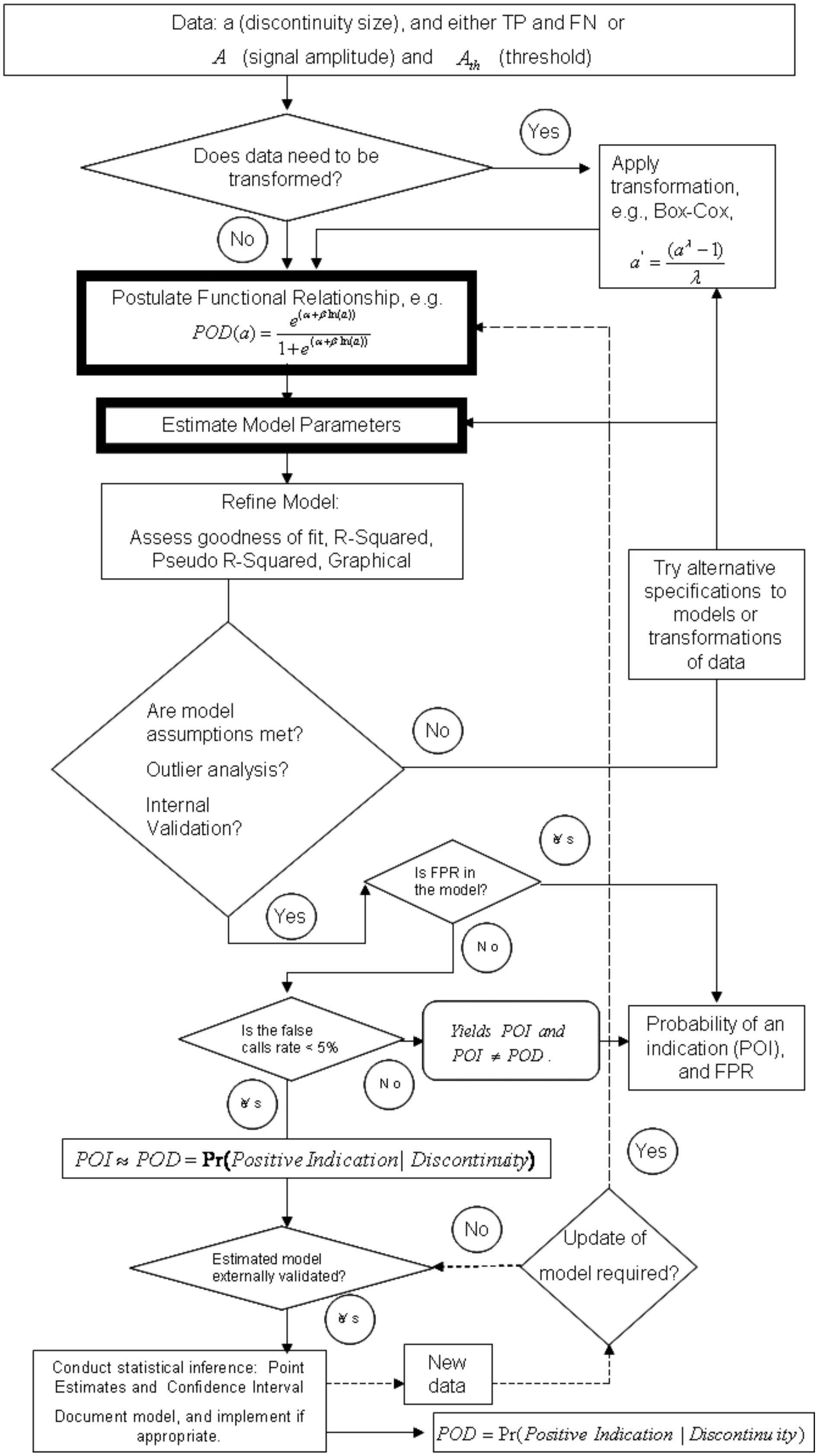

Data: a (discontinuity size), and either TP and FN or
A (signal amplitude) and A_{th} (threshold)
Does data need to be transformed?
Yes
Apply transformation, e.g., Box-Cox,
$a' = \frac{(a^{\lambda}-1)}{\lambda}$
No
Postulate Functional Relationship, e.g.
$POD(a) = \frac{e^{(\alpha+\beta\ln(a))}}{1+e^{(\alpha+\beta\ln(a))}}$
Estimate Model Parameters
Refine Model:
Assess goodness of fit, R-Squared, Pseudo R-Squared, Graphical
Try alternative specifications to models or transformations of data
Are model assumptions met?
Outlier analysis?
Internal Validation?
No
Yes
Is FPR in the model?
Yes
No
Is the false calls rate < 5%
Yields POI and $POI \neq POD$.
Probability of an indication (POI), and FPR
No
Yes
$POI \approx POD = \mathbf{Pr}(Positive\ Indication \mid Discontinuity)$
Yes
No
Update of model required?
Estimated model externally validated?
Yes
Conduct statistical inference: Point Estimates and Confidence Interval
Document model, and implement if appropriate.
New data
$POD = \Pr(Positive\ Indication \mid Discontinuity)$

Figure 8. Flow diagram of binary Logit regression ML estimation procedure

Although there are guidelines and computer software packages (e.g., Mil-HDBK-1823A, 2009) for automating and performing a maximum likelihood analysis for determining POD, there are several additional critical steps (Figure 8) that need to be taken to assure that the generated estimated POD model is adequate before implementing the estimated model. Many of these critical steps are not fully covered when relying on POD software generated by others. It is also important to recognize that different methods of analysis may be used for different versions of the same software. This becomes critical when tracking, comparing, or correcting previously obtained and accepted POD capability results. This highlights the importance of fully understanding the basics of maximum likelihood analysis, as detailed in this document, so that reliance on POD software generated by others does not create risk when meeting specific inspection requirements. It is recommended that any personnel, analyzing inspection data by maximum likelihood methods and reporting POD capabilities, be able to demonstrate prior personal maximum likelihood analysis capability to estimate POD from relevant data sets without the use of maximum likelihood POD analysis software (shareware or otherwise) generated by others. This is an obvious recommendation. For example, statisticians working in the area of POD already generate their own individual software routines for maximum likelihood POD analysis, and POD software generated by one statistician is not often relied upon by another statistician.

Four critical steps are brought out in figure 8.

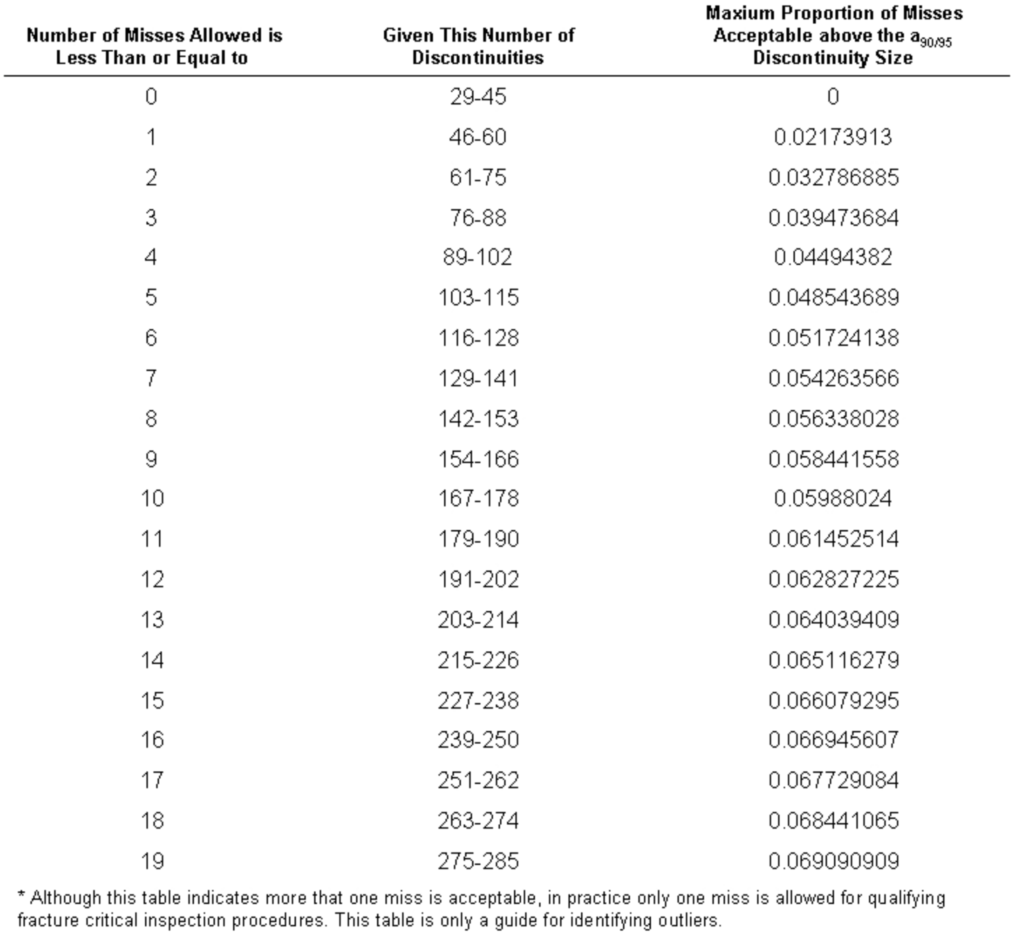
The first critical step is to verify model assumptions. For the two parameter Logit-ML of binary data, these assumptions are:

- There is no classification error in the binary responses. That is, the data are recorded properly.
- The observations are independent.
- The explanatory (independent) variables are measured without error.
- The assumed relationship between POD and the explanatory variable is adequate.
- The independent variables are not linear combinations of each other.

A determination of the properties of the model, such as POD is increasing with discontinuity size and $0 \leq \text{POD} \leq 1.0$, are to be considered when selecting models.

Significant outliers are explained as recording or data collection errors or model variances. There is no exact definition of what constitutes an outlier; so that determining whether or not an observation is an outlier is a subjective exercise. Deletion of outliers is not recommended when estimating POD, where the underlining model and measurement errors are not known *a priori*. Outliers resulting from instrument reading errors may be deleted unless human factors are also being evaluated. Methods for performing an outlier analysis include evaluating the number of observations that exceed those expected at three standard deviations from the mean assuming normally distributed observations, or graphical methods. For fracture critical applications, the proportion of the number of Misses to number of discontinuities above the 90/95 POD discontinuity size should not be excessive. It is proposed here that when the one-sided upper 95% confidence bound for the binomial proportion of the number of Misses for the number of discontinuities above the 90/95 POD discontinuity size is greater than 0.1, then these Misses may be outliers and further evaluation is necessary.

Figure 9 and table 2 are generated from equation (44) where the equivalent 95% one-sided upper bound on probability of a miss is just one minus the 95% one-sided lower bound probability of a hit (TPR). By acceptable it is meant that the Misses observed may not be outliers. However, it is stated here that no more than one Miss is allowed in practice for failure critical inspections, and all Misses are to be evaluated for cause.



Number of Misses Allowed is Less Than or Equal to	Given This Number of Discontinuities	Maxium Proportion of Misses Acceptable above the $a_{90/95}$ Discontinuity Size
0	29-45	0
1	46-60	0.02173913
2	61-75	0.032786885
3	76-88	0.039473684
4	89-102	0.04494382
5	103-115	0.048543689
6	116-128	0.051724138
7	129-141	0.054263566
8	142-153	0.056338028
9	154-166	0.058441558
10	167-178	0.05988024
11	179-190	0.061452514
12	191-202	0.062827225
13	203-214	0.064039409
14	215-226	0.065116279
15	227-238	0.066079295
16	239-250	0.066945607
17	251-262	0.067729084
18	263-274	0.068441065
19	275-285	0.069090909

* Although this table indicates more that one miss is acceptable, in practice only one miss is allowed for qualifying fracture critical inspection procedures. This table is only a guide for identifying outliers.

Table 2. Maximum Proportion of Misses Acceptable above the a90/95 Discontinuity Size

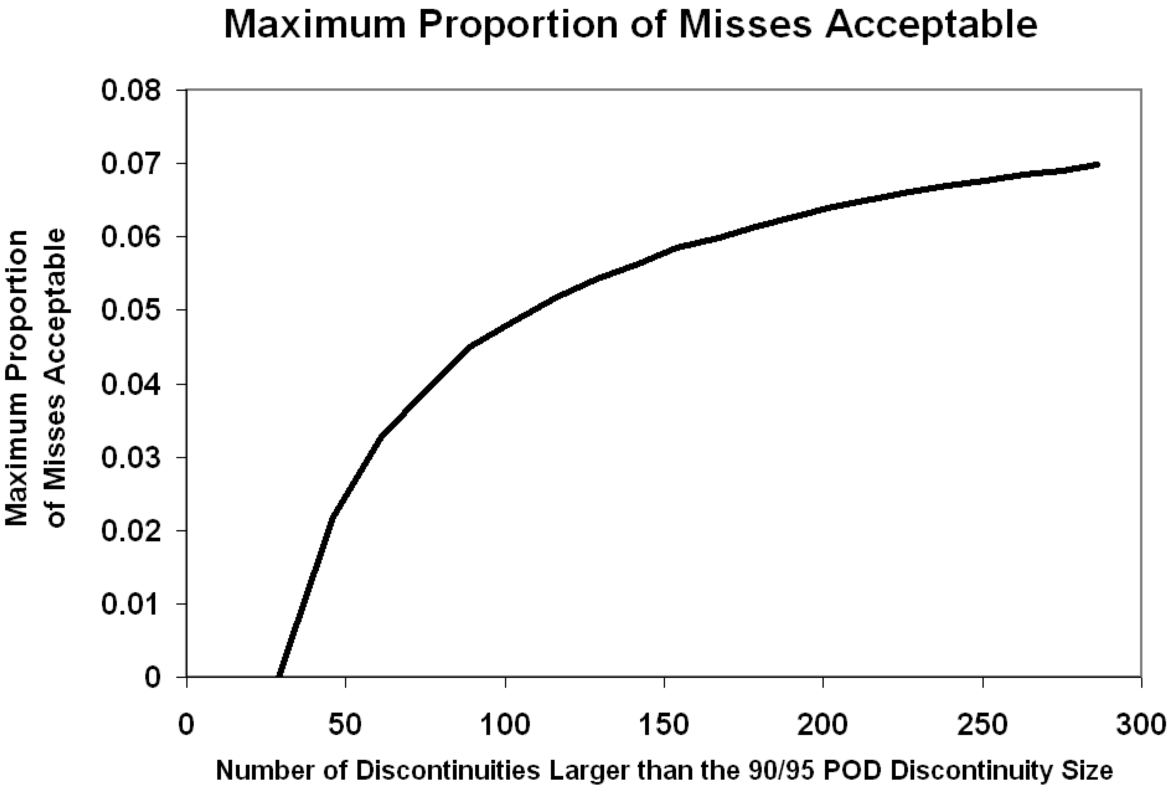


Figure 9. Acceptable Proportion of Number of Misses to Number of Discontinuities Larger than the 90/95 POD Discontinuity Size.

A second critical step is to determine the need for a refined model that may provide a better representation of the data. A refined model may include additional parameters such as that proposed by (Spencer, 1998). The Akaike (1974) information criterion (AIC) may be used to evaluate the relative goodness of fit when comparing proposed statistical models. The AIC is given by,

(70) $\mathrm{AIC} = 2k - 2\ln(L)$

where k is the number of parameters in the statistical model, L is the maximized value of the likelihood function for the estimated model. Given a set of proposed models for the data, the preferred model is the one with the minimum AIC value. Transformations of the discontinuity size variable (see Figure 8) may also be used to assist in refining the model. The likelihood ratio test (Meeker and Escobar, 1998, chapter 8) may also be used to provide a formal comparison between nested models.

A third critical step is to perform an internal validation of the model. Methods include cross validation method and jackknife approaches (Steyerberg, 2009). If the internal validation does not yield adequate results, then an alternative model is to be postulated. These internal validation methods explore the sensitivity of the statistical model by utilizing a selection of random sample subsets. For POD analysis, a substantial proportion (typically 95%, and with high confidence) of the estimated models generated from sample subsets are to be entirely above the lower confidence bound established for the full sample set.

A fourth critical step is to perform an external validation to assure that the estimated model is adequate for the population for which it is intended. There are two different methods that may be used for external validation. One method verifies that the new data is consistent with the estimated model. The other method, verifies that the new data is consistent with the old data. The first method requires generating a new estimated parameters for the given model from the new data, then the old and new estimated parameters are compared to evaluate if there is a statistical significance between them. The second method requires that the new data be added to the old data, and evaluating if there is a significant change in the model parameters estimated from the combined data sets. When additional data are available a validated estimated model may be updated and the procedures are repeated.

Additional guidance on diagnostics, model selection, goodness of fit, and assessment may be found at BIOST 515 (2004) and Pregibon (1981). There are many software packages available that carry out logistics regression analysis and are useful for various logistic regression alternatives (see Pezzullo 2006, and Winbugs). Winbugs is used by NASA in risk analysis. The reader should be careful in downloading and using software (commercial or shareware) and verify that software is cleared by your information technology security organizations.

Bayes' Rule

Bayes' rule is a method for computing certain conditional probabilities and is often cited as a procedure for computing certain POD metrics. Bayes' rule is given by,

(71) $\Pr(\mathrm{A_k} | \mathrm{B}) = \dfrac{\Pr(\mathrm{A_k})\mathrm{P}(\mathrm{B}|\mathrm{A_k})}{\Pr(\mathrm{A_1})\Pr(\mathrm{B}|\mathrm{A_1}) + \Pr(\mathrm{A_2})\Pr(\mathrm{B}|\mathrm{A_2}) + ... + \Pr(\mathrm{A_m})\Pr(\mathrm{B}|\mathrm{A_m})}$

where

$\mathrm{A_1}, \mathrm{A_2}, ..., \mathrm{A}_m$ = set of mutually exclusive events such that $\Pr\left(\bigcup_1^m A_i\right) = 1$

m = number of mutually exclusive events, and

$\Pr(\mathrm{A_k}|\mathrm{B})$ = Probability of $\mathrm{A_k}$ given an event B

$\Pr(\mathrm{A})$ is referred to as the prior distribution and $\Pr(\mathrm{A}|\mathrm{B})$ is the posterior distribution of A, given B and can be thought of at $\Pr(\mathrm{A})$ updated with the knowledge that B has occurred. Bayes' rule links a conditional probability $\Pr(\mathrm{A}|\mathrm{B})$ to its inverse $\Pr(\mathrm{B}|\mathrm{A})$ to provide the relationship between $\Pr(\mathrm{A}|\mathrm{B})$ and $\Pr(\mathrm{B}|\mathrm{A})$.

Bayes' Rule Applied to the Conditional Probability of a Positive Indication

We have two mutually exclusive events of a discontinuity being present, and a discontinuity not being present.

Bayes' rule, equation (71) may be used to determine the conditional probability of interest. Using the data from Figures 2a to obtain parameters values (Equations (3), (4),(7), and (8)) we have,

(72) $\text{Probability of a discontinuity existing, } \Pr(\text{ Discontinuity }) = 0.0001$

(73)
$$\begin{aligned}\text{Probability of a discontinuity not existing } &= \Pr(\text{ No Discontinuity })\\ &= 1 - \Pr(\text{ Discontinuity }) = 0.9999\end{aligned}$$

Probability of a positive indication given a discontinuity exists,

(74) $\text{POD} = \Pr(\text{ Postive Indication} \mid \text{Discontinuity }) = 0.95$

Probability of a positive indication given no discontinuity exists, probability of false positive,

(75) $\text{FPR} = \Pr(\text{ Positive Indication} \mid \text{No Discontinuity }) = 0.01$

Using Bayes' rule, the posterior probability of a discontinuity being present given a positive indication,

(76)
$$\Pr(\text{ Discontinuity} \mid \text{Positive Indication }) = \frac{\Pr(\text{ Discontinuity })\Pr(\text{ Positive Indication} \mid \text{Discontinuity })}{\Pr(\text{ Discontinuity })\Pr(\text{ Positive Indication} \mid \text{Discontinuity }) + \Pr(\text{ No Discontinuity })\Pr(\text{ Positive Indication} \mid \text{NoDiscontinuity })}$$

$$= \frac{\Pr(\text{ Positive Indication} \mid \text{Discontinuity })\Pr(\text{ Discontinuity })}{\Pr(\text{ Positive Indication })}$$

Therefore,

(77)
$$\Pr(\text{ Discontinuity} \mid \text{Positive Indication }) = \frac{\Pr(\text{ Discontinuity })(\text{ POD })}{\text{POI}}$$

and we obtain the probability of a discontinuity present given a positive indication,

(78)
$$\Pr(\text{ Discontinuity} \mid \text{Positive Indication }) = \frac{(0.95)(0.0001)}{(0.95)(0.0001)+(1 - 0.0001)(0.01)} = \frac{0.000095}{0.010094} = 0.009411532$$

that is identical to that obtained from the joint probability matrix (Figure 3),

(79)
$$\frac{\Pr(\text{ Positive Indication } \cap \text{Discontinuity })}{\Pr(\text{ Positive Indication })} = \frac{0.0000095}{\text{POI}} = \frac{0.000095}{0.010094} = 0.009411532$$

Bayes' rule methodology has a direct relationship with the joint probability matrix (Figure 3).

Confidence Intervals for this Bayes' Rule Example

The $a_{90/95}$ discontinuity size is obviously dependent on the procedure used to estimated POD and to the procedure used to determine the confidence intervals. For this example, the probability estimate is from a simple ratio of number of successes divided by the number of trials so the Clopper-Pearson method may be used to determine confidence intervals.

The Interrelationship with the ROC display, Joint Probability Matrix, Binomial, Logit-ML, and Bayes' Rule POD Methods

From the joint probability matrix (Figure 4), the probability of a positive indication given a discontinuity of size a and given noise is given by equation (40). It has been reported as a rule of thumb (Petrin, et al., 1993), when FPR exceeds 5% (high noise levels) for mulit-parameter curve fits, the probability of a positive indication will be inflated and not adequately reflect the probability of detection.

The probability of an indication from any source when a discontinuity exists is given by

(80) $POI = \Pr(ID|A) + \Pr(IN|A) - \Pr(ID|A) \cdot \Pr(IN|A) = POD + FPR - POD \cdot FPR$

where POD is the probability of detection from a discontinuity indication.

Therefore, the difference is given by,

(81)POI - POD = FPR·(1 – POD).

When FPR is low and POD is high there is little difference in equating POD to POI. When FPR is less than 5% (low noise levels) then

(82) $POI \approx POD$

For example, when FPR is 0.01 (1%) and POD(a) is 0.95, then,

(83) $$\begin{aligned} POI &= POD + FPR - POD \cdot FPR \\ &= 0.95 + 0.01 - (0.95) \cdot (0.01) = 0.9505 \end{aligned}$$

However, POD is less than POI. This highlights the need to make corrections when estimates of inspection capabilities when FPR is non-zero. If FPR is increased to 5%, then there can be up to a 5% difference between and POI and POD. The difference becomes progressively larger as FPR increase or POD decreases. Figure 10 highlights the inflationary effect that FPR in equation (80) has on the corrected POD for several values of POI.

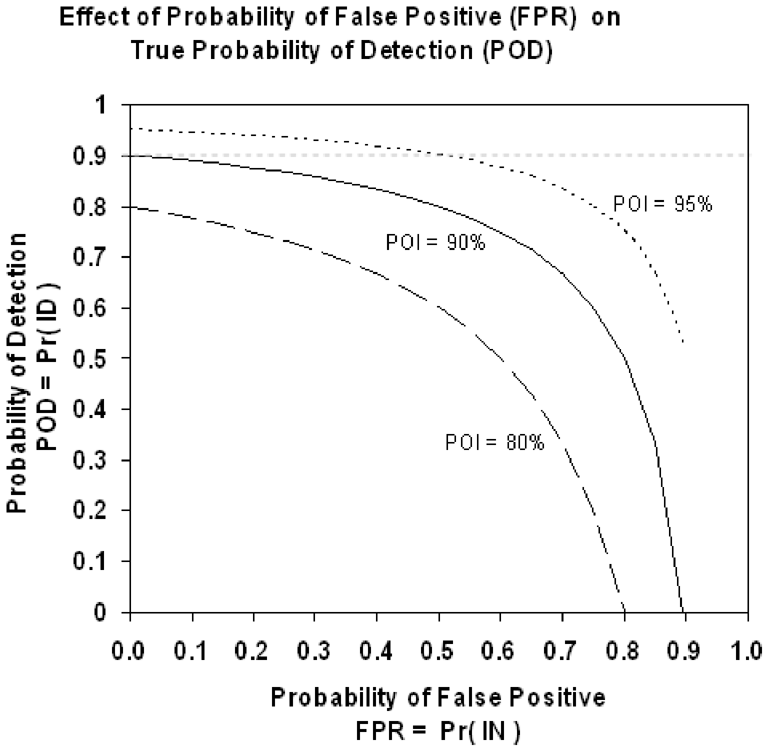


Figure 10. Effect of Probability of False Positive (FPR) on Probability of Detection (POD) when POI = 0.9.

For example, if the observed POI is 0.9 and Pr(IN) is 0.1 (10%), from Equation (80), then the corrected POD is Pr(ID) is 0.89. This inspection system would not be acceptable for fracture critical inspection requirements requiring 0.9 POD at 95% confidence.

It has been shown for the region A in Figure 2b, that the binomial point estimate method yielding POI can also be assumed to yield POD when FPR is small. Further, the multi-parameter logistic maximum likelihood estimates also yield POD when low noise levels exist. This is very important as the ordinate (vertical) variable in the ROC graph changes from POI to POD when low noise levels are maintained. That is, direct estimates of POD may be obtained from binomial and multi-parameter curve fits when low noise levels exist, in Figure 2 regions B and A, respectively. Outside of these regions, corrections need to be applied to the estimated POI in order to determine the estimated POD. Establishing a large signal to noise ratio is one method for assuring measurements of estimated POD.

This issue is also important when interpreting Standards for inspections. The NASA requirement for 0.90 POD of a specific discontinuity size at 95% confidence ($a_{90/95}$) for critical inspections is based on POD and not on POI. This is evidenced by the internal NASA practice of qualifying inspectors and inspection systems by demonstrating $a_{90/95}$ inspection capability, while not allowing excessive unexplained false positives. For a typical qualification POD test, a set of at least 29 sample sites with discontinuities and a set of 86 sample sites without discontinuities are included in the test. This practice allows for sufficient test data to verify both the $a_{90/95}$ POD inspection capability and that the false call probability is less than 3.44% with high confidence.

It is also important not to focus on the occurrence of false positives during testing as it is preferred that the inspection or inspector erroneously identify a suspect discontinuity, rather than miss discontinuities with discontinuity sizes at or greater than the critical discontinuity size. A judgment has to be made as to how many false positives are allowed to pass a qualification POD test. Although acceptable upper bounds on the false positive probabilities have been specified in the previous text, there are judgments that may be made explaining the occurrence of individual false positive classifications. Therefore the adherence to meeting the false positive probability requirements exactly may be waived with documented justification, while realizing that excessive false positives implies that the measure of POI that is used as a measure of POD may be overly inflated. The acceptance of POI adds risk if POI is used to qualify inspection personnel and inspection systems.

Another significant factor that is not readily seen when comparing POD methodologies, is that estimated POD may in fact be quite different depending on the method used. For example, a binomial point estimate of POD using 29 sites with same-sized discontinuities often results in an estimated POD of one (1.0) and a one-sided lower 95% confidence bound of 0.9. This meets the $a_{90/95}$ inspection capability as long as it is also verified that POD exceeds 0.9 with 95% confidence for discontinuities larger than $a_{90/95}$ (Generazio 2011). As a practical matter, it is unusual that more than 46 sites having discontinuities with sizes at the critical discontinuity size will be used in a binomial based qualification process to establish POD capability. Forty six is the number of discontinuities having the critical discontinuity size that would allow up to one miss for establishing a one-sided lower 95% confidence bound of 0.9. In this case, the binomial estimate of POD will be approximately 0.978 with a 95% confidence one-sided lower bound of 0.90. In sharp contrast, a maximum likelihood curve fit method (Christner, et. al., 1988) that includes discontinuity sizes ranging from very small to very large, may, at the discontinuity size $a_{90/95}$ yield an estimated POD close to the one-sided lower confidence bound of 0.90, as shown in Figure 11.

Both of the above methods demonstrate $a_{90/95}$ capability, however, for fracture and failure critical systems the binomial point estimated POD is in the range 0.978 to 1.0 providing added assurance when compared to 0.90 estimated POD from curve fitting methods. Binomial-based qualification testing for fracture critical applications yields an estimated POD at or above 0.978. Figure 11 shows the Logit-ML estimated POD obtained at $a_{90/95}$ discontinuity sizes for 437 data sets (NTIAC 1997). Logit-ML estimated POD at a $_{90/95}$ discontinuity sizes are as low as 0.93. Typical data sets with identical number of samples (open diamonds in Figure 11) exhibit Logit-ML estimated POD as high as 0.99 and as low as 0.93. In practice, the maximum number of misses allowed in a binomial point estimate is one, where one miss out of 46 sites having discontinuities with sizes at the critical discontinuity size will yield an acceptable 90/95 POD capability when the POD is verified to be monotonic above the $a_{90/95}$ critical discontinuity size. The horizontal line in Figure 11 is this worst case binomial estimated probability of 0.978 at $a_{90/95}$ discontinuity size, so that for all fracture critical inspections the binomial estimated POD is at or above 0.978. The implication here is that 90/95 POD criterion may not be sufficiently adequate for fracture critical system, rather the criteria should be that the estimated POD should meet or exceed 0.978 with a 95% one-sided lower bound on that probability meeting or exceeding 0.90.

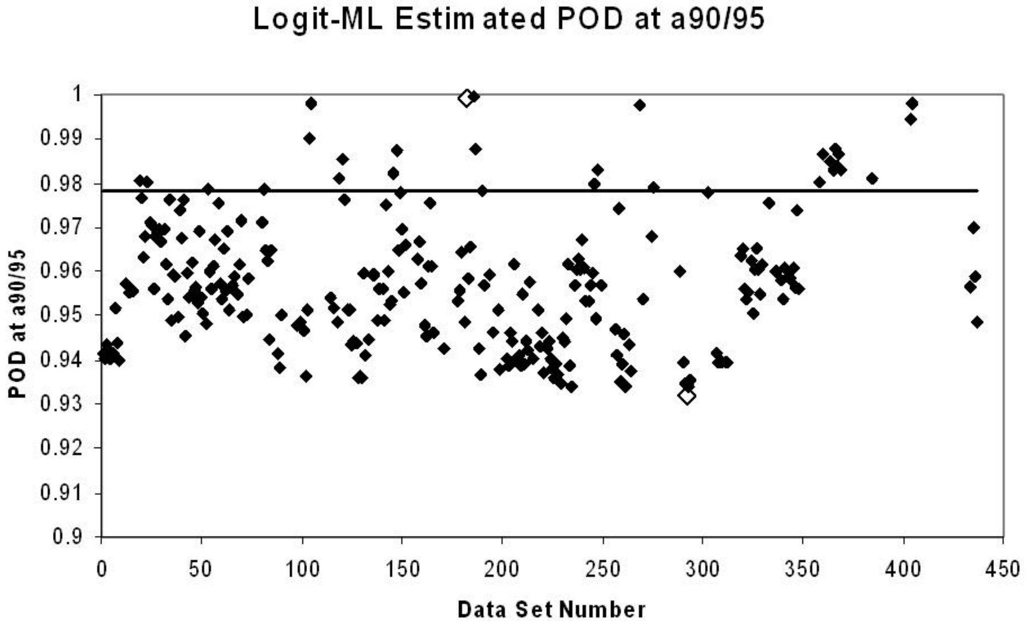


Figure 11. Logit-ML Estimated POD at $a_{90/95}$ from NTIAC (1997). Open diamonds refer to data sets each having 325 samples. The horizontal line is the NASA minimum binomial estimated POD (0.978) accepted in practice at $a_{909/95}$ discontinuity size for failure critical applications.

This is actually intuitive to all. If a single fracture critical component fails due to the presence of discontinuities with sizes at or greater than the critical discontinuity size and this failure results in catastrophic losses, is it appropriate to have used an inspection qualification system that has an estimated POD to find only 90% of the discontinuities have sizes at or greater than the critical discontinuity size?

Tolerance

Statistical confidence intervals are placed on single valued parameters that are fit to a model and reflect the uncertainty of estimation due to the statistical nature of the data used for the estimation. A tolerance interval differs in that it bounds the range of data which contains a specific proportion of a naturally varying population. The size of a confidence limit is due to sampling error and will shrink to zero as the data sample size increases. A tolerance interval depends both on the actual variation in a population, as well as the sampling error and will approach the specific proportion for which the tolerance interval was performed as the sampling size increases.

For example, a specified NDE inspection process may be implemented by a population of different inspectors. Each inspection can be considered to have its own effective POD, not all the same. Therefore, the probability of detection for a given discontinuity size would have a naturally occurring distribution. A lower 0.90 tolerance bound level for the probability of detection at a 95% confidence would be that probability, with 95% confidence that at least 90% of the inspectors exceed. It has been recently proposed that naturally occurring inspector-to-inspector variation should be explicitly considered and that tolerance bounds for POD is more appropriate than the confidence bounds that have traditionally been used (Li, et al., 2012).

Comparing Binomial, Logit, ROC, Bayes' Rule POD Numerical Results

Different properly executed procedures may produce POD estimates that do not appear to agree when compared to each other or when comparing specific published POD capabilities. The reader should avoid shopping around to select the most attractive POD values. Rather, the decision to rely on one POD methodology over another is to be supported by understanding the conservativeness of the procedure and adequacy of models used to generate the POD estimates. There is always a trade off between the confidence in credible results and the conservativeness required.

Decision Guidelines

This work has set the foundation for an extended discussion on POD decision making where a comprehensive, decision-support document is needed that provides guidelines on practical risk-informed decisions involving POD/confidence (CL) level utilizations. These decision guidelines need to include:

1. Criteria for actually selecting a POD value and the associated confidence level (CL), i.e. a POD/CL pair (e.g., situations where a 99%/95% pair may not be stringent enough and where a 80%/50% may be too stringent or visa versa) ,
2. Consideration of the actual probabilistic, consequential risk of a given pair POD/CL which includes the probability of actually having the discontinuity, the probability of not detecting it, and the consequences of not detecting it,
3. The tradeoff between the POD value and the associated value for CL (e.g., when is it better to have a high POD and lower CL and vice versa),
4. The concepts of a *conditional* false positive associated with the complement of the POD versus the *unconditional* probability of false positive, where the unconditional includes the probability of having the discontinuity. NASA HDBK 8739.19.19-4 “Estimation and Evaluation of Measurement and Decision Risk” recommends evaluation and use of the *unconditional* false positive in selecting and deciding upon acceptance limits.
5. When the different methods of estimation matter and when they don't and guidelines for their practical applications.
6. When the Logistic (or Probit) method is useful for estimation of the POD for particular discontinuity sizes that are not directly obtainable but that can be consequential.
7. What sample sizes are needed to achieve a given POD/CL

8. How the Bayes approach can be used for the Binomial and Logistic to incorporate prior information to improve the POD/CL. Also how to actually form and estimate the prior from past knowledge and history.

Baseline decision guidelines are provided here, and it is recognized that there are many different types of risk assessment procedures that have additional requirements. These include probabilistic risk assessment (NASA 2011) and development of advanced physics based risk assessments methods (Generazio 1994) that utilize probabilities in complex environments to evaluate trade offs and establish benefits.

SUMMARIZING

The ROC 2x2 matrix of data, joint probability matrix, and companion probability estimates are the statistical concepts that provides the direct relationship between ROC display, binomial, logistic regression, and Bayes' rule POD methodologies. The concept of probability, conditional probabilities, joint probabilities, and marginal probabilities are discussed as they relate to POD methodologies. When inspection response signals have high noise levels, then important conditions on the applications of these POD methodologies are to be met. Specifically, probability methods that yield probability of an indication POI can also be assumed to yield POD when false positive probability (FPR) is small

For fracture critical applications, the binomial, two parameter logistic regression, and Bayes' rule are directly useful for determining probability of detection when the probability of false positive is low. Inspection Standards for fracture critical applications need to specify the acceptance of probability of detection (measurements with low noise levels) in the criteria so that inflated POD capabilities are not erroneously relied upon. Probability of an indication capability assessments are to be rejected for meeting inspection Standards when measurements accompanied by high noise levels. ROC is a method of display for probability data. It is shown that the ROC ordinate axis changes from probability of detection to probability of an indication depending on the probability methodology being displayed and the magnitude of the probability of false positives.

The availability of a large number of methods for computing POD-related metrics leads to estimates of POD and lower confidence bounds that may appear to be quite different, so it is important to clearly identify methods used to estimate POD and confidence bounds. Variations in results may be attributable to estimating methods, confidence level and procedures to determine confidence intervals, statistical models, and other assumptions used.

Estimates of POD should always be accompanied by a confidence interval showing at least the one-sided lower confidence bound and confidence level that will express the statistical uncertainty (i.e., the uncertainty due to limited data). There are multiple confidence bound procedures available and all describe uncertainty in estimates resulting because we have limited data. It is important to recognize that confidence intervals do not reflect variability in the process. The a_{90} point on a POD plot (sometimes incorrectly called $a_{90/50}$) is an estimate of the discontinuity size that will be detected with probability 0.90. The point where the one-sided lower 95% confidence bound on POD exceeds 0.90 is referred to as the 90/95 POD point or the $a_{90/95}$ discontinuity size. When likelihood methods are used to determine confidence bounds, the $a_{90/95}$ point is also an upper confidence bound on the discontinuity size that will be detected with probability 0.9. Confidence statements only describe statistical uncertainty arising from limited data, therefore estimated POD models need to be validated or evaluated for adequacy A non-zero probability of a false positive is not solely an economic concern. A non-zero false positive probability inflates the estimated POD and adds risk when not addressed properly. This risk may become critical when signal response mechanisms have high noise levels. To avoid confusion, definitions are to be specified when discussing POD data, analysis, and methods.

The accepted criteria requiring that the estimated POD shall meet or exceed 0.9 with a 95% one-sided lower confidence bound on that probability is not adequate for fracture critical system, rather the accepted criteria should be that the estimated POD shall meet or exceed 0.978 with a 95% one-sided lower confidence bound on that probability that meets or exceeds 0.90 for all discontinuities at or larger than the critical discontinuity size..

It is recommended that any personnel analyzing inspection data and reporting POD results for decision-making that have been obtained by maximum likelihood methods be able to demonstrate prior personal knowledge of maximum likelihood principles and not simply use POD analysis software (shareware or otherwise) as black boxes.

This work has set the foundation for initiating discussions on POD decision making where a comprehensive, decision-support document is needed to provide detailed guidelines on practical risk-informed decisions involving POD/confidence (CL) level utilizations.

The reader should now be able to properly address the italicized statements in the introduction as well and many other related statements.

ACKNOLEDGEMENTS

The author wishes to thank statisticians, Dr. William Q. Meeker of Iowa State University and Dr. Floyd W. Spencer of Sfhire for clarifying statistical concepts, procedures, and detailed critical reviews of this document. I also want to thank Dr. William Vesely of NASA for his extended review of this document, and Ward D. Rummel of D & W Enterprises for his unwavering support, encouragement, and recommendations.

APPENDIX A

A infinitely differentiable function may be represented by a Taylor series expansion about the maximum likelihood estimates of the parameters, so that for the log likelihood profile, we have,

(84)

$$\mathrm{LL} = \left.\mathrm{LL}\right|_{\hat{\alpha},\hat{\beta}} + \left.\frac{\partial \mathrm{LL}}{\partial \alpha}\right|_{\hat{\alpha},\hat{\beta}} (\alpha - \hat{\alpha}) + \left.\frac{\partial \mathrm{LL}}{\partial \beta}\right|_{\hat{\alpha},\hat{\beta}} (\beta - \hat{\beta}) + \frac{1}{2}\left[\left.\frac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \alpha}\right|_{\hat{\alpha},\hat{\beta}} (\alpha - \hat{\alpha})^2 + \left.\frac{\partial^2 \mathrm{LL}}{\partial \beta \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} (\beta - \hat{\beta})^2 + 2\left.\frac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} (\alpha - \hat{\alpha})(\beta - \hat{\beta}) \right] + \ldots$$

Where LL and it's partial derivatives are evaluated at the estimated maximum likelihoods $\hat{\alpha}$ and $\hat{\beta}$.

The first term is the estimated maximum likelihood value and a constant. The first partial derivatives are maximized out at the maximum likelihood and are zero. The terms with second partial derivatives are the second order estimates of variance. Higher order terms are presumed to be small and are neglected to yield a quadratic approximation to LL described by,

$$(85)\ \left.\mathrm{LL}\right|_{\hat{\alpha},\hat{\beta}} + \frac{1}{2}\left[\left.\frac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \alpha}\right|_{\hat{\alpha},\hat{\beta}} (\alpha - \hat{\alpha})^2 + \left.\frac{\partial^2 \mathrm{LL}}{\partial \beta \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} (\beta - \hat{\beta})^2 + 2\left.\frac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} (\alpha - \hat{\alpha})(\beta - \hat{\beta}) \right]$$

Equation (85) may also be represented in matrix form,

$$\left.\mathrm{LL}\right|_{\hat{\alpha},\hat{\beta}} + \frac{1}{2}\begin{pmatrix} (\alpha - \hat{\alpha}) \\ (\beta - \hat{\beta}) \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha^2}\right|_{\hat{\alpha},\hat{\beta}} & \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} \\ \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} & \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \beta^2}\right|_{\hat{\alpha},\hat{\beta}} \end{pmatrix} \begin{pmatrix} (\alpha - \hat{\alpha}) \\ (\beta - \hat{\beta}) \end{pmatrix} =$$

$$\left.\mathrm{LL}\right|_{\hat{\alpha},\hat{\beta}} + \frac{1}{2}\begin{pmatrix} (\alpha - \hat{\alpha}) & (\beta - \hat{\beta}) \end{pmatrix} \begin{pmatrix} \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha^2}\right|_{\hat{\alpha},\hat{\beta}} & \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} \\ \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} & \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \beta^2}\right|_{\hat{\alpha},\hat{\beta}} \end{pmatrix} \begin{pmatrix} (\alpha - \hat{\alpha}) \\ (\beta - \hat{\beta}) \end{pmatrix}$$

where

$$(86) - \begin{pmatrix} \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha^2}\right|_{\hat{\alpha},\hat{\beta}} & \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} \\ \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \alpha \partial \beta}\right|_{\hat{\alpha},\hat{\beta}} & \left.\dfrac{\partial^2 \mathrm{LL}}{\partial \beta^2}\right|_{\hat{\alpha},\hat{\beta}} \end{pmatrix} = \begin{pmatrix} \mathrm{II}_{11} & \mathrm{II}_{12} \\ \mathrm{II}_{21} & \mathrm{II}_{22} \end{pmatrix}$$

is the information matrix. The variance of the $\hat{\alpha}$ and $\hat{\beta}$ estimators are calculated by the inverse of the information matrix.

The information matrix is the negative of the expected value of the Hessian matrix, where the Hessian is the matrix of second derivatives of the log likelihood with respect to the parameters, α and β. We have,

$$\begin{pmatrix} \mathrm{II}_{11} & \mathrm{II}_{12} \\ \mathrm{II}_{21} & \mathrm{II}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} (\alpha)_{\mathrm{var}} & (\alpha\beta)_{\mathrm{covar}} \\ (\alpha\beta)_{\mathrm{covar}} & (\beta)_{\mathrm{var}} \end{pmatrix}, \quad \text{where } \mathrm{II}_{12} = \mathrm{II}_{21},$$

to obtain the variances and covariance,

$$(87) \quad (\alpha)_{\mathrm{var}} = \frac{\mathrm{II}_{22}}{(\mathrm{II}_{11}\mathrm{II}_{22} - \mathrm{II}_{12}{}^{2})}$$

$$(88) \quad (\beta)_{\mathrm{var}} = \frac{\mathrm{II}_{11}}{(\mathrm{II}_{11}\mathrm{II}_{22} - \mathrm{II}_{12}{}^{2})}$$

$$(89) \quad (\alpha\beta)_{\mathrm{covar}} = \frac{-\mathrm{II}_{12}}{(\mathrm{II}_{11}\mathrm{II}_{22} - \mathrm{II}_{12}{}^{2})}.$$

Applying the above procedure to the log-likelihood function,

$$(90)\ \mathrm{LL} = \sum_{i=1}^{N} d_i \cdot \ln\left[\frac{e^{\alpha + \beta \ln(a_i)}}{1 + e^{\alpha + \beta \ln(a_i)}}\right] + (n_i - d_i)\ln\left[1 - \frac{e^{\alpha + \beta \ln(a_i)}}{1 + e^{\alpha + \beta \ln(a_i)}}\right]$$

we have,

$$(91) - \left.\frac{\partial^2 \mathrm{LL}}{\partial \alpha^2}\right|_{\hat{\alpha},\hat{\beta}} = \sum_{i=1}^{N}\left[\frac{e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}{\left(1 + e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}\right)^2}\right] n_i = \mathrm{II}_{11}$$

$$(92) - \left.\frac{\partial^2 \mathrm{LL}}{\partial \beta^2}\right|_{\hat{\alpha},\hat{\beta}} = \sum_{i=1}^{N}\ln(a_i)^2\left[\frac{e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}}{\left(1 + e^{\hat{\alpha} + \hat{\beta}\ln(a_i)}\right)^2}\right] n_i = \mathrm{II}_{22}$$

(93) $$-\left.\frac{\partial^2 LL}{\partial\alpha\partial\beta}\right|_{\hat{\alpha},\hat{\beta}} = \sum_{i=1}^{N} \ln(a_i)\left[\frac{e^{\hat{\alpha}+\hat{\beta}\ln(a_i)}}{\left(1+e^{\hat{\alpha}+\hat{\beta}\ln(a_i)}\right)^2}\right] n_i = II_{12}$$

Using equations (91), (92), and (93) in equations (87), (88), and (89) the variance and covariance are obtained. These variances will be used to estimate the standard errors for determining the one-sided lower confidence bounds using the Wald method.

The large sample normal approximation for the distribution of estimators is also referred to as the Wald method. We seek the Wald 95% lower confidence bound on

(94) $$\hat{P}(a) = \frac{e^{\hat{\alpha}+\hat{\beta}\ln(a)}}{1+e^{\hat{\alpha}+\beta\ln(a)}}$$

The one-sided 100(1-γ)% lower bound on a scalar function Y is given by,

(95) $$Y(a)_{LB} = \hat{Y}(a) - z_{(1-\gamma)} \cdot \text{stde}$$

where stde is the local estimate for the standard error, σ_Y of $\hat{Y}$, $z_{(1-\gamma)}$ is the 1-γ quantile of the standard normal distribution (Meeker and Escobar, 1998, pg. 628). With γ = 0.05 for the 95% one-sided confidence bound, $z_{0.95}$ = 1.64.

Let

(96) $$\hat{Y}(a) = \hat{\alpha} + \hat{\beta}\ln(a)$$

then the standard deviation σ_Y of $\hat{Y}(a)$ may be estimated from (Bevington, 1969),

(97) $$\sigma_Y{}^2 \simeq \sigma_\alpha{}^2\left(\frac{\partial Y}{\partial\alpha}\right)^2 + \sigma_\beta{}^2\left(\frac{\partial Y}{\partial\beta}\right)^2 + 2\sigma_{\alpha\beta}{}^2\left(\frac{\partial Y}{\partial\alpha}\right)\left(\frac{\partial Y}{\partial\beta}\right)$$

where $\sigma_\alpha{}^2$, $\sigma_\beta{}^2$, and $\sigma_{\alpha\beta}{}^2$ are the variances $(\alpha)_{\text{var}}$, $(\beta)_{\text{var}}$, and $(\alpha\beta)_{\text{covar}}$, respectively.

Taking the derivatives of Equation (96), Equation (97) becomes,

(98) $$\sigma_Y \simeq \sqrt{\sigma_\alpha{}^2 + \ln(a)^2\ \sigma_\beta{}^2 + 2\ln(a)\,\sigma_{\alpha\beta}{}^2}$$

and by using the variances obtained from log likelihood profile, Equations (87), (88), and (89), we have,

(99) $$\text{stde} = \sigma_Y \simeq \sqrt{(\alpha)_{\text{var}} + \ln(a)^2 \; (\beta)_{\text{var}} + 2\ln(a)\;(\alpha\beta)_{\text{covar}}}$$

combining Equations (95), (96), and (99)

(100) $$Y(a)_{LB} = \quad \hat{\alpha} + \hat{\beta}\ln(a) - 1.64\sqrt{(\alpha)_{\text{var}} + 2\ln(a)\;\;(\alpha\beta)_{\text{covar}} + \ln(a)^2\;\;(\beta)_{\text{var}}}$$

The upper one-sided bound on the same scalar function follows similarly and is given by

(101) $$Y(a)_{UB} = \quad \hat{\alpha} + \hat{\beta}\ln(a) + 1.64\sqrt{(\alpha)_{\text{var}} + 2\ln(a)\;\;(\alpha\beta)_{\text{covar}} + \ln(a)^2\;\;(\beta)_{\text{var}}}$$

It is reemphasized here that even though the upper and lower bounds are determined, these are one-sided 95% confidence bounds with open ended intervals, and not two-sided 95% confidence bounds enclosing an interval. Taken together the interval (Y_{LB}, Y_{UB}) would be considered a 90% two-sided confidence bound.

The Wald lower 95% confidence one-sided bound (Figure 7) on $\hat{P}(a)$ is now obtained as a function of variances,

(102) $$\text{Wald}_{LCB}(\,a\,) = \frac{e^{\hat{\alpha} + \hat{\beta}\ln(a) - 1.64\sqrt{(\alpha)_{\text{var}} + 2\ln(a)\;\;(\alpha\beta)_{\text{covar}} + \ln(a)^2\;\;(\beta)_{\text{var}}}}}{1 + e^{\hat{\alpha} + \hat{\beta}\ln(a) - 1.64\sqrt{(\alpha)_{\text{var}} + 2\ln(a)\;\;(\alpha\beta)_{\text{covar}} + \ln(a)^2\;\;(\beta)_{\text{var}}}}}\;.$$

REFERENCES

Agresti, A., Coull, B. A., Approximate is Better than “Exact” for Interval Estimation of Binomial Proportions, *The American Statistician*, Vol. 52, No. 2, May 1998

Agresti, Alan, Categorical Data Analysis, John Wiley & Sons, Hoboken, New Jersey, 2nd ed., pg, 174, 2002

Akaike, Hirotugu "A new look at the statistical model identification". IEEE Transactions on Automatic Control 19 (6): 716–723, (1974).

Bevington, Philip R., *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill, New York, New York, 1969

Brown, Lawrence D., Cai, T. Tony, and DasGupta, Anirban, Interval Estimation for a Binomial Proportion, *Statistical Sciences*, Vol. 16, No. 2, pp. 101-133

Christner, B. K., D. L. Long and W. D. Rummel, “NDE Detectability of Fatigue-Type Cracks in High-Strength Alloys,” MCR-88-1044, Martin Marietta Astronautics Group, September 1988.

Clopper, C.; Pearson, E. S. (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial". *Biometrika* 26: 404–413. doi:10.1093/biomet/26.4.404

BIOST 515, *Diagnostics and model checking for logistic regression,* Lecture 14, February 19, 2004, http://courses.washington.edu/b515/l14.pdf

Generazio, E. R., *Technology Benefit Estimator (T/BEST): User's manual*, NASA-TM-106785, December 1994

Generazio, Edward R., Binomial Test Method for Determining Probability of Detection Capability for Fracture Critical Applications, NASA/TP–2011-217176, September 2011.

Generazio, E. R., “Design of Experiments for Validating Probability of Detection Capability of NDT Systems and for Qualification of Inspectors,” *Materials Evaluation*, Vol. 67, No. 6., 2009, pp. 730–738, and “Errata”, *Materials Evaluation*, Vol. 67, No. 6, 2009, pp. 730 - 738.

Generazio, E. R., Validating Design of Experiments for Determining Probability of Detection Capability (DOEPOD) for Fracture Critical Applications, *Materials Evaluation*, Vol. 69, No. 12., December 2011, pp 1399 – 1407

Generazio, Edward R., Directed Design of Experiments for Validating Probability of Detection Capability of a Testing Systems, patent US 8108178 B2, January 31, 2012

MIL-HDBK-1823A, Nondestructive Evaluation (NDE) System, Reliability Assessment, Nondestructive Evaluation System Reliability, Department of Defense, April 7, 2009, superseding MIL-HDBK-1823, Nondestructive Evaluation System Reliability Assessment, Department of Defense, April 30, 1999

Fahr, A., Forsyth, D., Bullock, M., and Wallace, W., POD Assessment of NDI Procedures Using a Round Robin Test, AGARD-R-809, January 1995

Hald, A., *Statistical Theory with Engineering Applications*, John Wiley & Sons, New York, New York, 1952

Hines, William W. and Montgomery, Doulas C., *Probability and Statistics in Engineering and Management Science*, 2nd Ed., pg. 954, John Wiley & Sons, New York, New York, 1972

C. A. Harding, and Hugo, G. R. *Statistical Analysis of Probability of Detection Hit/Miss Data for Small Data Sets*, Review of Progress in Quantitative Nondestructive Evaluation, V. 22B , AIP Conference Proceedings, Melville, New York, July 2003

Li, M., Spencer, F. W. , and Meeker, W. Q., “Distinguishing Between Uncertainty and Variability in Nondestructive Evaluation,” , *Review of Progress in Quantitative Nondestructive Evaluation*, AIP Conf. Proc. 1430, 1725-1732, 2012.

Meeker, W. Q. and Escobar, L. A., *Statistical Methods for Reliability Data*, John Wiley and Sons, New York, New York, 1998

Meeker, William Q., Escobar, Luis A., Teaching About Approximate Confidence Regions Based on Maximum Likelihood Estimate”, *The American Statistician*, February 1995, Vol. 49, No. 1

NASA, NASA-STD-5009, Nondestructive Evaluation Requirements for Fracture Critical Metallic Components, National Aeronautics and Space Administration, 7 April 2008

NASA SP-2011-3421, *Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners*, Second Edition, December 2011

NTIAC, Nondestructive Evaluation (NDE) Capabilities Data Book, 3rd ed., NTIAC DB-97-02, Nondestructive Testing Information Analysis Center, November 1997

Petrin, C., Annis, C. A., and Vukelich, S. I., A Recommended Methodology for Quantifying NDE/NDI Based on Aircraft Engine Experience, AGARD-LS-190, April 1993.

Pezzullo, John C., *Logistic Regression Calculating Page*, http://statpages.org/logistic.html, 2006

Pezzullo, John C., *Exact Binomial and Poisson Confidence Intervals*, http://statpages.org/confint.html, 2010

Pregibon, D., *Logistic Regression Diagnostics*, Annals of Statistics, Vol. 8, No.4, 1981, pp. 705-724.

Rummel, W. D., “Recommended Practice for Demonstration of Nondestructive (NDE) Reliability on Aircraft Production Parts,” *Materials Evaluation*, Vol. 40, No. 8, 1982.

Spencer, F., Identifying Sources of Variation for Reliability Analysis of Field Inspections, SAND-98-0980C, 1998

Steyerberg, Ewout W., *Clinical Prediction Models, A practic al Approach to Deve lopment, validation, and Updating*, Springer Science and Business Media, 2009, Chapter 17.

Wald, A., Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large, *Transactions of the American Mathematical Society,* Vol. 54, No. 3 (Nov., 1943), pp. 426-482

Weast, Robert C. , editor*, Handbook of Mathematical Tables, 4th Ed.,* The Chemical Rubber Company, Cleveland, Ohio, 1970, pg. 967 and 970

Wilson P. Tanner Jr., John A. Swets, A Decision-Making Theory of Visual Detection, *Psychological Review*, Vol 61, Issue 6, November 1954, Pages 401-409

Winbugs, http://www.mrc-bsu.cam.ac.uk/software/bugs/ and
and Simple Logistic Regression Program Using Winbugs,
http://www.medicine.mcgill.ca/epidemiology/joseph/courses/EPIB-621/bayeslogit.pdf

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 01- 04 - 2014	**2. REPORT TYPE** Technical Memorandum	**3. DATES COVERED** *(From - To)*

4. TITLE AND SUBTITLE

Interrelationships Between Receiver/Relative Operating Characteristics Display, Binomial, Logit, and Bayes' Rule Probability of Detection Methodologies

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Generazio, Edward R.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

724297.40.44.07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER

L-20385

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-0001

10. SPONSOR/MONITOR'S ACRONYM(S)

NASA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

NASA/TM-2014-218183

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Category 38
Availability: NASA CASI (443) 757-5802

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Unknown risks are introduced into failure critical systems when probability of detection (POD) capabilities are accepted without a complete understanding of the statistical method applied and the interpretation of the statistical results. The presence of this risk in the nondestructive evaluation (NDE) community is revealed in common statements about POD. These statements are often interpreted in a variety of ways and therefore, the very existence of the statements identifies the need for a more comprehensive understanding of POD methodologies. Statistical methodologies have data requirements to be met, procedures to be followed, and requirements for validation or demonstration of adequacy of the POD estimates. Risks are further enhanced due to the wide range of statistical methodologies used for determining the POD capability. Receiver/Relative Operating Characteristics (ROC) Display, simple binomial, logistic regression, and Bayes' rule POD methodologies are widely used in determining POD capability. This work focuses on Hit-Miss data to reveal the framework of the interrelationships between Receiver/Relative Operating Characteristics Display, simple binomial, logistic regression, and Bayes' Rule methodologies as they are applied to POD. Knowledge of these interrelationships leads to an intuitive and global understanding of the statistical data, procedural and validation requirements for establishing credible POD estimates.

15. SUBJECT TERMS

Failure; NDE; Nondestructive evaluation; POD; Probability of Detection; Statistical analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			STI Help Desk (email: help@sti.nasa.gov)
					19b. TELEPHONE NUMBER *(Include area code)*
U	U	U	UU	65	(443) 757-5802