E-Journal of Advanced Maintenance Vol.4No.3 (2012) 105-115
Japan Society of Maintenology

Considerations for Statistical Analysis of Nondestructive Evaluation Data: Hit/Miss Analysis

Jeremy KNOPP[1,*], Ramana GRANDHI[2,†], Li ZENG[3], and John ALDRIN[4]

[1] *United States Air Force Research Laboratory, Wright-Patterson Air Force Base, USA*
[2] *Wright State University, Dayton, OH, USA*
[3] *The University of Texas at Arlington, Arlington, TX, USA*
[4] *Computational Tools, Gurnee, IL, USA*

ABSTRACT

This paper examines recent developments in the statistical analysis of Nondestructive Evaluation (NDE) data, and suggests improvements to conventional analysis. This paper focuses on hit/miss or Bernoulli data analysis. POD evaluation is a conventional methodology used to quantify reliability of inspections in many industries. Hit/miss data is still commonly used and many developments in statistical analysis have occurred in the last few decades; thus, this paper focuses on hit/miss analysis only

KEYWORDS

Probability of detection, hit/miss analysis, Bernoulli data

Article history:
Received # 29 August 2012
Accepted # 13 November2012

1. Introduction

Hit/miss analysis or Bernoulli data analysis to be precise continues to be the most prevalent method used to determine probability of detection (POD) in practice. In general, this type of analysis requires advanced statistical methods. Attempts to quantify NDE capability began in the 1970's with studies by United States National Aeronautics and Space Administration (NASA) and the United States Air Force (USAF) [1, 2]. The latter study is probably the largest study on the reliability of NDE techniques in history. Initially, POD was analyzed using Binomial statistics. This was inadequate because POD changes as a function of flaw size. The USAF and statisticians developed methods based on Logistic regression to analyze hit/miss data [3]. The fundamental functional form of the POD curve is shown in equation 1, where 'a' is the flaw size, μ is the 50% probability of detection flaw size, also known as a_{50}, σ is a slope parameter, and Φ is the cumulative distribution function of the standard normal distribution.

$$p = \Phi\left(\frac{\ln(a) - \mu}{\sigma}\right) \tag{1}$$

The two models that were used were the log odds or logit model shown in equation 2 and the cumulative log normal or probit model shown in equation 3. The logit and probit models have different tail behavior, and since tail behavior for large flaw sizes is of most interest, determining the best model is important.

*Corresponding author, E-mail: jeremy.knopp@us.af.mil

† Present address: *Wright-Patterson AFB, OH, 45433, United States*

$$p = \frac{\exp(b_0 + b_1 \ln(a))}{1 + \exp(b_0 + b_1 \ln(a))} \qquad \text{(logit)} \qquad (2)$$

$$p = \Phi(b_0 + b_1 \ln(a)) \qquad \text{(probit)} \qquad (3)$$

Note that only the form of the function is cumulative and POD should not be interpreted as a cumulative distribution. During the development of POD analysis methods in the 1980's, the log odds model was more feasible for computational reasons. The standard reference for many years was the seminal work by Berens in the American Society of Metals (ASM) Handbook [4]. Later, this work was codified into a US Department of Defense (DOD) handbook for POD studies, which was published in 1999 [5]. The major challenge in the development of POD thus far was accurate confidence bound calculations. It needs to be emphasized that a POD curve without confidence bounds has little value.

Since 2000, several modifications have been made or suggested for POD analysis [5]. First, the confidence bounds for hit/miss data were deemed overly conservative because they were applied simultaneously to all the points on the POD curve. The software developed by Berens [6] was changed to calculate the confidence bounds for each individual flaw size locally rather than to the entire POD curve [7]. Both of these confidence bound calculations were based on the so called Wald Statistic [8]. Later, the likelihood ratio method was suggested for more accurate confidence bound calculations on the model parameters estimates [9, 10]. This is a modern "Gold standard" statistical method that is now feasible to implement on a personal computer thanks to advances in computational statistics. Recently, MIL-HDBK-1823 was revised to include these developments [11]. MIL-HDBK-1823A is considered the state-of-the-art guidance for conducting POD studies by the USAF and other industries that conduct POD studies [12]. There is also parallel work being done in medical statistics similar to NDE reliability [13]. The primary requirement for hit/miss data in MIL-HDBK-1823A is that POD is 0 as flaw size approaches 0, and 1 as flaw size approaches infinity. An example of a data set that does not meet this requirement was shown in [14], and methods in a Bayesian framework will be introduced to handle such a data set.

2. Review of Current Methodologies

2.1. Hit/Miss Analysis: Case Study and Terminology.

To illustrate the different methods for analyzing NDE data, the data set in [4] will be used and is shown in Table 1. This data set contains 35 observation opportunities with 13 hits and 22 misses. The flaws range from 0.200 mm to 6.990 mm. The data was derived from an evaluation of a fluorescent penetrate inspection. In practice, the statistical methods presented here can be applied to any inspection data with binary responses.

First, the commonly used terms are defined:

a_{50} – estimate of flaw size for 50% POD, also equal to μ
a_{90} – estimate of flaw size for 90% POD
$a_{90/95}$ – lower bound at 95% confidence for 90% POD

$a_{90/95}$ is a scalar quantity commonly used as a performance metric for comparison of NDE systems and risk calculations. Note that confidence bounds refer to only the specific POD experiment, and to make inference on future capability of an inspection, prediction bounds should be considered.

Table 1 Hit/Miss Data Example from [4]

Flaw size (mm)	Response	Flaw size (mm)	Response
0.2	0	2.18	1
0.23	0	2.18	0
0.25	0	2.21	0
0.38	0	2.41	1
0.46	0	2.49	0
0.51	0	2.54	1
0.58	0	2.64	0
0.64	0	2.84	1
0.99	0	2.97	1
0.99	0	3.3	0
1.02	0	4.09	0
1.42	0	4.22	1
1.63	1	4.42	1
1.85	0	4.95	1
1.98	1	5.59	1
2.03	0	6.2	1
2.06	0	6.99	1
2.13	0		

2.2. Hit/Miss Analysis according to [5, 6, 9, 10]

The results reported in [4] were calculated using the log odds model and the confidence bounds were calculated globally for the entire POD curve; thus the $a_{90/95}$ number is very conservative. POD/SS software version 3 [6] was used to calculate the POD curve and confidence bounds using the Wald method for the set of data. These results can be reproduced following the calculations in [7]. The results are displayed in Figure 1. The key quantities a_{50} (also μ), a_{90}, and $a_{90/95}$ are indicated on this plot.

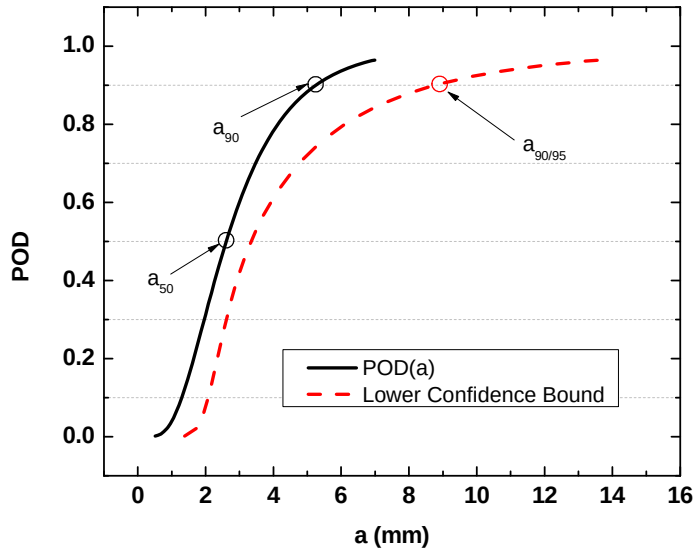

Figure 1: Analysis of data from [4] with POD/SS [6].

Next, the likelihood ratio method was used to determine POD. Here, the probit model was used because it was determined that the probit model provided a better fit. This calculation was done using software known as mh1823 software which is a library available for use with the 'R' programming language [15]. The POD curve with confidence bounds calculated using the likelihood ratio method is displayed in Figure 2. The numerical values are reported later in the paper in Table 3. Visual inspection of the results reveals that both a_{50} and a_{90} are approximately the same as expected, but there is big discrepancy in the $a_{90/95}$ value. This discrepancy is a result of the Wald method using point estimates, and the likelihood ratio considering all flaw sizes simultaneously.

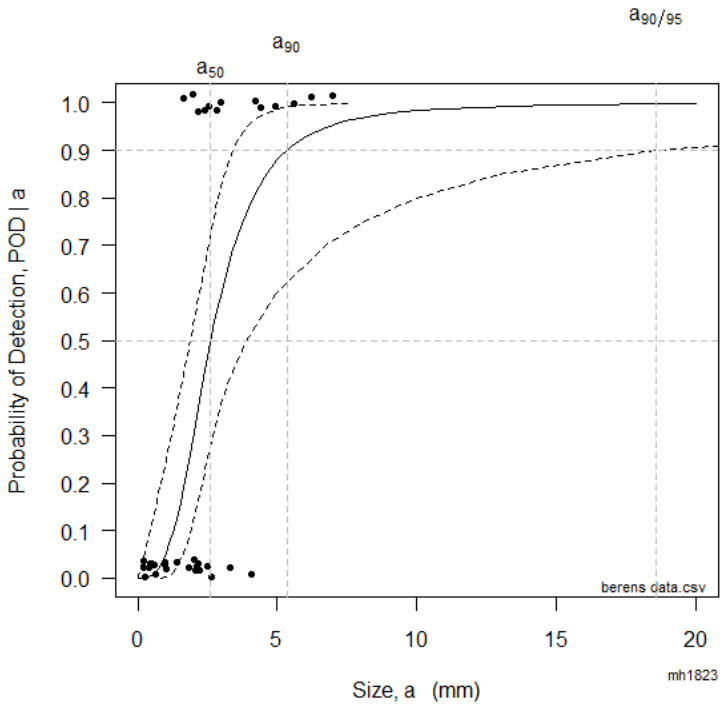

Figure 2: Analysis of data in [4] with likelihood ratio method [10]

3. New Methodology

3.1 Hit/Miss Analysis using Markovchain Monte Carlo Simulation

The two statistical methods used in the previous section are modern state-of-the-art methods in conventional statistics. There are cases where the data does not suggest that the POD goes to 1 as the flaw size goes to infinity, and there is also the problem of false calls where the POD curve doesn't go to 0 as the flaw size goes to 0. To address these issues, additional parameters must be incorporated in the model, but inference using conventional methods is difficult. In addition, recent efforts in model-assisted POD have led to the consideration of Bayesian statistical methods to incorporate information from physics-based models and expert opinion. The advantages of going beyond conventional statistics to Bayesian statistics are twofold: 1) Markovchain Monte Carlo (MCMC) simulation allows more complicated POD models to be used because it facilitates parameter estimation and confidence bound computation, and 2) prior information from expert opinion and physics-based models can be incorporated in the POD study. Recently, a new method for analyzing hit/miss data was proposed using MCMC simulation [16]. Details on the application of MCMC simulation for parameter estimation can be found in [17]. The mathematical form of Bayes' rule is given by

$$P(\theta \mid D,M) = \frac{P(D \mid \theta,M)P(\theta \mid M)}{P(D \mid M)} = \frac{P(D \mid \theta,M)P(\theta \mid M)}{\int_{\theta} P(D \mid \theta,M)P(\theta \mid M)}, \tag{4}$$

where the data follows a model M, and θ is a set of parameters in the model. $P(\theta|D,M)$ literally reads as the probability of the parameters given the data, and it is commonly called the "posterior" distribution. $P(D|\theta,M)$ literally reads as the probability of the data given the parameters, and is also known as the "likelihood". $P(\theta|M)$ is the "prior" distribution of the parameters, which represents prior information/expert knowledge of the model. $P(D|M)$ is commonly called the "evidence" or "marginal likelihood" under the assumed model, which can be calculated by an integration. In this work, no special prior information is used, but this framework has the flexibility to include prior information in future work. The parameters are estimated by sampling from the posterior distribution in Eq. (4) through MCMC simulation. The benefit of using this method is that the parameter estimates and confidence bounds for more complicated models can be computed more easily. Moreover, model selection can also be conducted to determine the best model for the data by checking the marginal likelihood which is a popular indicator of the "strength" of the assumed model. The additional models that will be considered include 3 and 4-parameter models. A 3-parameter model will have either a lower asymptote (α) or an upper asymptote (β), and a 4-parameter model will have both α and β. The 4-parameter case is depicted in Figure 3.

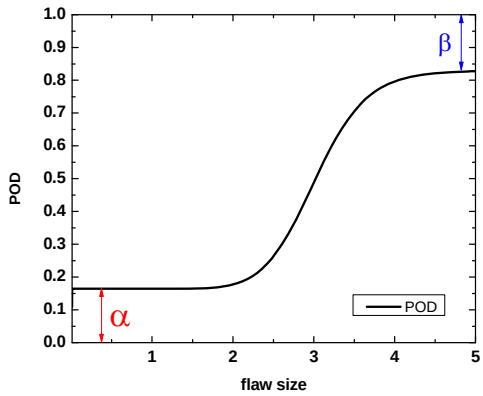

Figure 3: POD curve with both a lower asymptote α and an upper asymptote β.

3.2 Results of POD Analysis Using MCMC Computation

The data in section 2 will now be analyzed using MCMC computation for the confidence bounds. The 2-parameter logit and probit models are calculated first. Next, 3-parameter models with a lower and upper asymptote are calculated for both logit and probit cases. Finally 4-parameter models for both logit and probit cases are calculated. A model selection technique commonly used in Bayesian Statistics is known as the Bayes' factor. The Bayes' factor is the ratio of marginal likelihoods evaluated for different models. The marginal likelihood and corresponding Bayes factors are displayed in Table 2. The probit models have a slightly higher marginal likelihood compared to the logit models. All Bayes' factors for the analysis of Berens data are computed with the marginal likelihood for the 2-parameter probit model in the numerator and the alternative model in the denominator. The 2-parameter probit model shown in Figure 4 is the best fit according to the Bayes' factor. It should be noted, that this isn't overwhelming evidence as the Bayes factor of the 2-parameter probit model vs. the 2-parameter logit model is not large, and caution should be taken when drawing conclusions about Bernoulli data for small sample sizes such as this one. Table 3 shows the a_{50}, a_{90}, and $a_{90/95}$ values for each of the models.

Table 2: Bayes Factor Results of Analysis of Berens Data from [4]

	Marginal Likelihood	Bayes factor
2 parameter Logit	2.78E-08	2.109
2 parameter Probit	5.86E-08	1.000
3 parameter lower bound Logit	3.92E-09	14.976
3 parameter lower bound Probit	1.39E-09	42.339
3 parameter upper bound Logit	7.09E-09	8.269
3 parameter upper bound Probit	3.27E-09	17.914
4 parameter Logit	4.49E-10	130.735
4 parameter Probit	3.14E-09	18.692

Table 3: Performance Metrics Results of Analysis of Berens Data from [4]

	a_{50}	a_{90}	$a_{90/95}$
2 parameter Logit	2.611	5.333	10.136
2 parameter Probit	2.611	5.353	10.075
3 parameter lower bound Logit	3.072	5.253	10.355
3 parameter lower bound Probit	3.352	5.153	10.415
3 parameter upper bound Logit	2.191	-	-
3 parameter upper bound Probit	1.951	-	-
4 parameter Logit	2.473	-	-
4 parameter Probit	2.429	-	-
Berens ASM result	2.620	5.340	21.600
PODSS[6]	2.610	5.252	8.776
MH 1823 software Logit [15]	2.613	5.354	18.550
MH 1823 software Probit [15]	2.610	5.252	17.020

Table 4: Asymptote Results of Analysis of Berens Data from [4]

	Lower asymptote	Upper asymptote
3 parameter lower bound Logit	0.163	-
3 parameter lower bound Probit	0.189	-
3 parameter upper bound Logit	-	0.612
3 parameter upper bound Probit	-	0.577
4 parameter Logit	0.116	0.662
4 parameter Probit	0.113	0.623

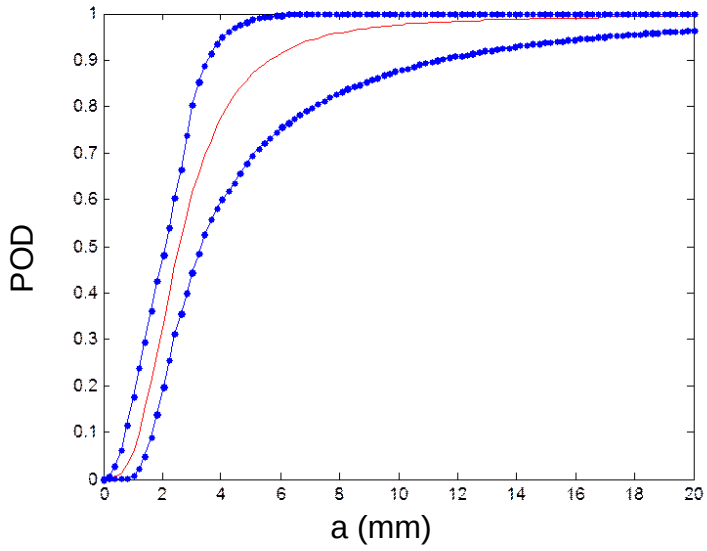

Figure 4: 2-parameter probit model with MCMC confidence bounds for Berens data [4].

4. A Difficult Data Set

4.1 Second example data set

One of the data sets referred to in [14] is called A6003H, as shown in Figure 5, which has 184 observations [18]. Visual inspection of the data reveals that there are many misses for larger flaw sizes. Since the 2-parameter models force the POD curve to go to 1 for large flaw sizes, and 0 for small flaw sizes, it is not recommended to use conventional methods suggested in [11]. A new parameter that represents an upper asymptote will need to be added to the POD model. It may also be necessary to add a lower asymptote that will provide some measures of false calls.

The data was analyzed with 11 different models. These include the Wald bounds [6], the likelihood ratio method for both logit and probit models [15], and logit and probit models for 2-parameter, 3-parameter with lower bound, 3-parameter with upper bound, and 4-parameter models that have both lower and upper bounds. The results of the analysis are listed in Table 5, 6 and 7. The marginal likelihood was largest for the 3-parameter probit model with an upper bound which is shown in Figure 6. In fact, the evidence for an upper bound is overwhelming. The Bayes' factor for comparing the 3-parameter probit model with the 2-parameter probit model is 2.073 E+07 which indicates that an upper asymptote for the POD curve is absolutely necessary. The upper asymptote is 0.921, but the lower confidence bound never reaches 0.9 so there is no $a_{90/95}$ estimate for this data set. The author recommends evaluating the quality of fit for multiple models to avoid drawing the wrong conclusions for a POD study. The 4-parameter model is also shown in Figure 7 since it is also an acceptable model.

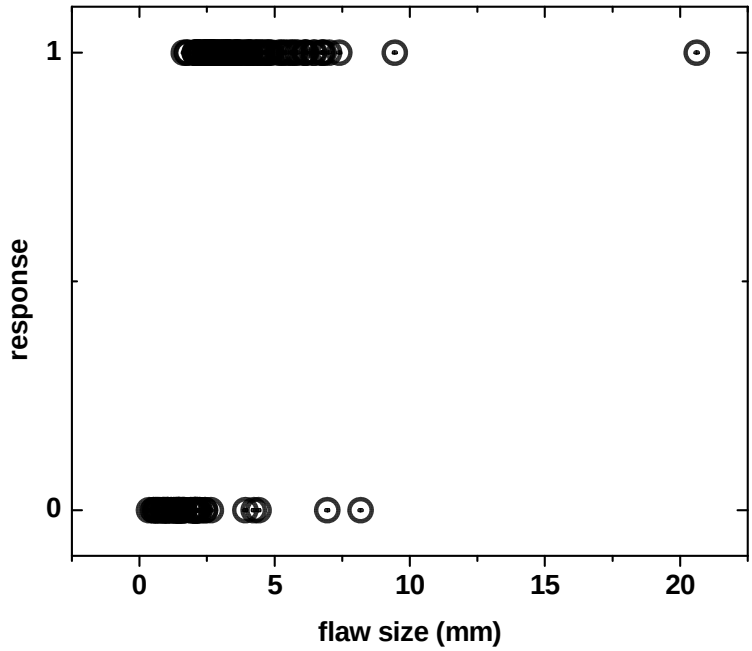

Figure 5: A6003H data set [16].

Table 5: Bayes Factor Results from Analysis of A6003H data set.

Model	Marginal Likelihood	Bayes Factor
2 parameter Logit	1.039E-31	6.821E+04
2 parameter Probit	3.418E-34	2.073E+07
3 parameter lower bound Logit	7.304E-33	9.703E+05
3 parameter lower bound Probit	3.680E-34	1.926E+07
3 parameter upper bound Logit	3.734E-28	1.898E+01
3 parameter upper bound Probit	7.087E-27	1.000E+00
4 parameter Logit	7.944E-28	8.921E+00
4 parameter Probit	3.571E-28	1.985E+01

Table 6: Performance Metrics Results from Analysis of A6003H data set.

	a_{50} (mm)	a_{90} (mm)	$a_{90/95}$ (mm)
2 parameter Logit (MCMC)	2.042	3.403	3.953
2 parameter Probit (MCMC)	1.980	3.815	4.451
3 parameter lower bound Logit (MCMC)	2.051	3.452	3.992
3 parameter lower bound Probit (MCMC)	2.031	3.832	4.432
3 parameter upper bound Logit (MCMC)	2.051	-	-
3 parameter upper bound Probit (MCMC)	2.051	-	-
4 parameter Logit (MCMC)	2.083	-	-
4 parameter Probit (MCMC)	2.091	-	-
PODSS[6]	2.032	3.797	4.340
MH 1823 software Logit [15]	2.038	3.439	4.157
MH 1823 software Probit [15]	2.032	3.798	4.586

Table 7: Asymptote Results from Analysis of A6003H data set.

Model	Lower Asymptote	Upper Asymptote
3 parameter lower bound Logit (MCMC)	0.030	-
3 parameter lower bound Probit (MCMC)	0.034	-
3 parameter upper bound Logit (MCMC)	-	0.922
3 parameter upper bound Probit (MCMC)	-	0.921
4 parameter Logit (MCMC)	0.045	0.921
4 parameter Probit (MCMC)	0.044	0.921

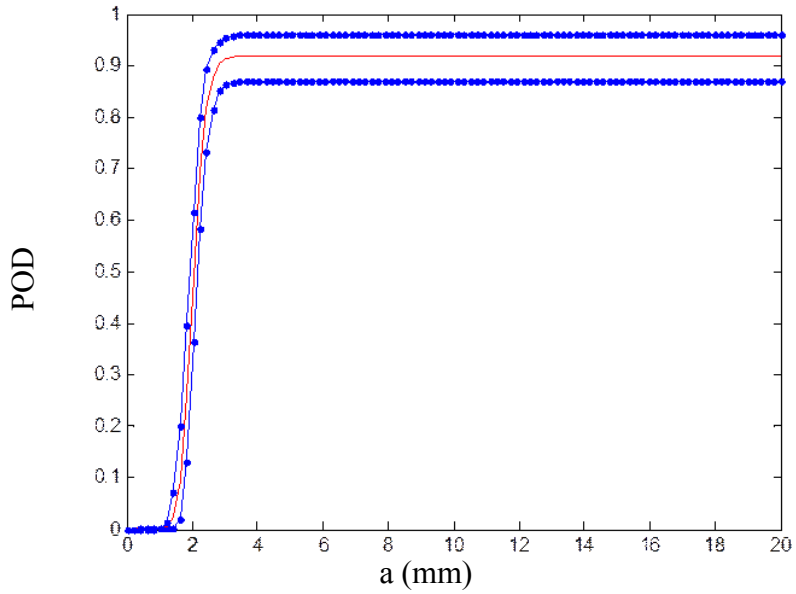

Figure 6: 3-parameter probit model with upper asymptote for A6003H data set.

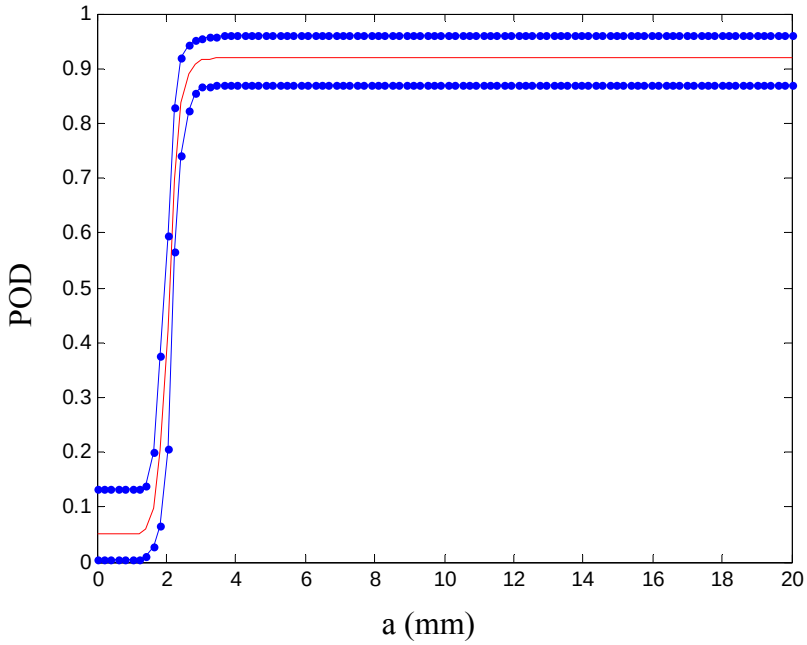

Figure 7: 4-parameter probit model with lower and upper asymptotes for A6003H data set.

5. Conclusion

The historical development of POD evaluation was summarized in this paper. Improvements to standard guidance were proposed. In particular, MCMC computation to facilitate parameter estimation and confidence bound calculations is suggested. POD can definitely be used to compare the performance of NDE systems. Caution should be taken when using POD for risk management of assets. Some statisticians have recommended sample sizes over 300 for hit/miss analysis. The sample size requirement is an open issue, but considering more models may ultimately reduce the size necessary. For the first data set considered, the sample size was only 35, but the 3 and 4 parameter model analysis did not suggest any problems. The second data set was large, but the models with 3 and 4 parameters showed that the $a_{90/95}$ estimate does not exist. Analyzing the A6003H data set with only a 2-parameter model will lead to a fictitious $a_{90/95}$ value. The author recommends consulting with a professional statistician for POD studies.

Acknowledgement

The authors wish to thank the Air Force Office of Scientific Research (AFOSR) and Dr. David Stargel in particular for supporting this research under task number 11RX15COR. The R software environment for statistical computing and graphics was used for all statistical computations involving Figure 2. R is open-source (free) software and is available to download here: http://www.r-project.org.

References

[1] W. D. Rummel, P. H. J. Todd, S. A. Frecska, and R.A. Rathke, "The Detection of Fatigue Cracks by Nondestructive Testing Methods," Technical Report NASA CR 2369, National Aeronautics and Space Administration Martin Marietta Aerospace, (1974)

[2] W. H. Lewis, W. H. Sproat, B. D. Dodd, and J. M. Hamilton, "Reliability of Nondestructive Inspections – Final Report. Technical Report SA-ALC/MME 76-6-38-1, San Antonio Air Logistics Center, (1978)

[3] Berens, A.P. and P.W. Hovey, "Evaluation of NDE Reliability Characterization," Final Report AFWAL-TR-81-4160, University of Dayton Research Institute, (1981)

[4] A. P. Berens, NDE Reliability Data Analysis, American Society for Metals Handbook Nondestructive Evaluation and Quality Control, Vol 17, pp. 689-701, ASM International, (1989)

[5] U.S. Department of Defense, MIL-HDBK-1823 A Nondestructive Evaluation System Reliability Assessment, (1999)

[6] POD/SS version 3, University of Dayton Research Institute, (1999)

[7] A. P. Berens, "Probability of Detection (POD) Analysis for the Advanced Retirement for Cause (RFC)/ Engine Structural Integrity Program (ENSIP) Nondestructive Evaluation (NDE) System Development," Technical Report AFRL-ML-WP-TR-2001-4010, Air Force Research Laboratory, (2000)

[8] A. Wald, Selected Papers in Statistics and Probability, (1955)

[9] C. A. Harding, and G. R. Hugo, "Experimental Determination of the Probability of Detection Hit/Miss Data for Small Data Sets", Rev. Prog. Quant. Nondestr. Eval, Vol 22, pp. 1823-1844, (2003)

[10] C. Annis, and J. S. Knopp, "Comparing the Effectiveness of a90/95 Calculations," Rev. Prog. Quant. Nondestr. Eval, Vol 26, pp. 1767-1774, (2007)

[11] U.S. Department of Defense, MIL-HDBK-1823A Nondestructive Evaluation System Reliability Assessment, (August 2010)

[12] L. Gandossi, and K. Simola, “Derivation and Use of Probability of Detection Curves in the Nuclear Industry,” Insight Vol 52, pp. 657-663, (2010)

[13] D. Collett, Modelling Binary Data, Chapman and Hall, (2002)

[14] E. R. Generazio, “Validating Design of Experiments for Determining Probability of Detection Capability for Fracture Critical Applications”, Materials Evaluation, Vol 69, pp. 1399-1407, (2011)

[15] C. Annis, “Statistical Best-Practices for Building Probability of Detection (POD) models” R package mh1823, version 2.5.4.1, http://statisticalengineering.com/mh1823

[16] J. S. Knopp and L. Zeng, Statistical Analysis of Hit/Miss Data, Materials evaluation (accepted for publication 2012)

[17] F. Kojima, J. S. Knopp, "Inverse Problem for Electromagnetic Propagation in a Dielectric Medium using Markov Chain Monte Carlo Method", International Journal of Innovative Computing Information and Control, vol 8 (3), pp. 2339-2346, (2012)

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