Modern Statistical Methods and Uncertainty Quantification for Evaluating Reliability of Nondestructive Evaluation Systems

Jeremy Scott Knopp

Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Engineering Commons

Repository Citation

This Dissertation is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
Modern Statistical Methods and Uncertainty Quantification for Evaluating Reliability of Nondestructive Evaluation Systems

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy

By

JEREMY SCOTT KNOPP
B.S.E.P, Wright State University, 2001
M.S.E, Wright State University, 2005
M.S. Wright State University, 2009

2014
Wright State University

Ramana V. Grandhi, Ph.D.
Dissertation Director

Frank W. Ciarallo, Ph.D.
Dissertation Co-Director

Committee on Final Examination:

Ramana V. Grandhi, Ph.D. (Director)

Frank W. Ciarallo, Ph.D. (Co-Director)

Michael R. Grimaila, Ph.D.

Yan Liu, Ph.D.

Pratik J. Parikh, Ph.D.

Xinhui Zhang, Ph.D.

The reliability of Nondestructive Evaluation (NDE) is an important input for risk analysis for sustainment of aging infrastructure. Reliability has typically been quantified via probability of detection (POD) studies. There are three problems with POD modeling methodologies provided in the most recent guidance on the subject:

1) Current models do not estimate the extremes of a POD curve very well because of the assumption that the POD curve approaches zero as flaw size goes to zero, and the POD curve approaches one as the flaw size goes to infinity.
2) The existing 2-parameter logit/probit models can be misused since there is not a set of diagnostics and procedures that can catch every violation of fundamental assumptions.
3) Data sets from realistic inspections often violate core assumptions in statistical models such as homoscedasticity and linearity, but statistical inference is still needed for the application.

Since one of the important inputs to risk assessment is POD, and it’s believed that the output of risk analyses can be sensitive to the tail behavior at large flaw sizes, it is worthwhile to consider better estimation procedures for the extremes of a POD curve. In this dissertation, new POD models that include lower and upper asymptotes are proposed to better model tail behavior. Transformations such as Box-Cox are proposed to mitigate violations of homoscedasticity, and bootstrapping is proposed to provide confidence bound calculations for higher order models. A case study is presented where these improvements to POD analysis are incorporated into a risk analysis. Simulation studies a presented to quantify the improvements of this work.
# TABLE OF CONTENTS

1 INTRODUCTION .................................................................................................................. 1  
1.1 Damage Tolerance Philosophy ..................................................................................... 1  
1.2 Overview of the Problem .............................................................................................. 8  
1.3 Organization of Dissertation ......................................................................................... 10

2 PROBABILITY OF DETECTION ....................................................................................... 11  
2.1 Hit/Miss Analysis ........................................................................................................ 13  
2.1.1 Binomial Based Methods ....................................................................................... 13  
2.1.2 Parametric Methods for Hit/Miss Analysis ............................................................... 14  
2.1.3 Wald Method ........................................................................................................ 17  
2.1.4 Likelihood Ratio Method ....................................................................................... 18  
2.1.5 Discussion ............................................................................................................ 19  
2.2 Case Study .................................................................................................................. 20  
2.3 Comparison Study ..................................................................................................... 21  
2.4 Classical Methods for Signal Response Analysis ......................................................... 24

3 MODERN STATISTICAL ANALYSIS OF HIT/MISS DATA ......................................... 31  
3.1 Bayesian Approach to Hit/Miss ................................................................................. 31  
3.2 Markov Chain Monte Carlo (MCMC) ........................................................................ 34  
3.3 POD Results with MCMC ......................................................................................... 36  
3.4 Challenge Data Set Example ...................................................................................... 39  
3.5 Re-Examination of Model Form via Symbolic Regression ....................................... 44  
3.6 Small Data Sets (a Case study) ................................................................................... 55

4 ADVANCED ANALYSIS OF SIGNAL RESPONSE DATA .................................. 60  
4.1 Box Cox Transformations ......................................................................................... 60  
4.2 Analysis with Physics-Inspired Model ....................................................................... 70  
4.3 Bootstrap Methods for Confidence Bound Calculation .......................................... 75  
4.4 Comparison with Hit/Miss Analysis .......................................................................... 77  
4.5 Chapter Summary ..................................................................................................... 79

5 APPLICATIONS ............................................................................................................. 81  
5.1 Experimental Cross Validation ................................................................................. 81  
5.2 Risk Analysis ............................................................................................................. 89

6 SIMULATION STUDIES ............................................................................................... 94  
6.1 Overview of Simulation Studies ............................................................................... 94  
6.2 Simulation Study 1 ................................................................................................... 94  
6.3 Simulation Study 2 .................................................................................................. 115  
6.3.1 100 observations ............................................................................................... 117
6.3.2 30 observations ................................................................. 130

7 SUMMARY AND CONCLUSIONS ............................................. 144

8 REFERENCES ........................................................................ 149
## LIST OF FIGURES

Figure 1.1: Aircraft Structure with Numerous Fastener Sites that Require Inspection. ..... 2  
Figure 1.2: Schematic for POD Evaluation Process .................................................. 7  
Figure 1.3: Schematic for the PROF Software used to make Risk Calculations .......... 8  
Figure 2.1: Hit/Miss Analysis with Wald Confidence Bounds ........................................ 22  
Figure 2.2: Hit/Miss Analysis Likelihood Ratio Confidence Bounds .............................. 23  
Figure 2.3: Scatter plot of data in Table 2.3 .................................................................. 28  
Figure 2.4: Censored Regression Performed on Data from Table 2.3 .......................... 29  
Figure 2.5: POD Curve for \( \alpha \) vs a Data from Table 2.3 ......................................... 30  
Figure 3.1: POD Curve with Both a Lower Asymptote \( \alpha \) and an Upper Asymptote \( \beta \) .. 34  
Figure 3.2: 2-parameter Probit Model with MCMC Confidence Bounds .................. 39  
Figure 3.3: Plot of observations from A6003H data set .............................................. 41  
Figure 3.4: 3-parameter Probit Model with Upper Asymptote for A6003H Data Set ..... 43  
Figure 3.5: 4-parameter Probit Model with Asymptotes for A6003H Data Set .............. 43  
Figure 3.6: Constant model used to model the A6003H data set ................................ 46  
Figure 3.7: Pareto frontier displaying the least complex solution among others .......... 47  
Figure 3.8: Linear regression model for A6003H data set ........................................... 47  
Figure 3.9: Sine wave model for A6003H data set ...................................................... 48  
Figure 3.10: Step function model for A6003H data set .............................................. 48  
Figure 3.11: Gauss function model for A6003H data set .......................................... 49  
Figure 3.12: Logistic regression model for A6003H data set ........................................ 49  
Figure 3.13: Pareto frontier showing where Logistic model ranks in complexity ............ 50  
Figure 3.14: Logistic regression model with upper asymptote .................................... 50  
Figure 3.15: Pareto frontier showing complexity of Logistic model with asymptote. ... 51  
Figure 3.16: Model with Erf and Gauss for A6003H data set ..................................... 51  
Figure 3.17: Model with Gauss and tan function for A6003H data set ....................... 52  
Figure 3.18: Model for A6003H data set containing gauss, cos, and tan functions ..... 52  
Figure 3.19: Pareto frontier showing complexity of model with additional functions ... 53  
Figure 3.20: Model for A6003H data that incorporates high level of complexity ....... 53  
Figure 3.21: Pareto frontier showing model with highest complexity for this study. ..... 54  
Figure 3.22: POD results for small data set with inspection system A .......................... 57  
Figure 3.23: POD results for small data set with inspection system B .......................... 57  
Figure 4.1: Raw Data for Eddy Current Inspection .................................................... 64  
Figure 4.2: Transformed \( \alpha \) Data with \( \lambda = 0.45 \) .............................................. 67  
Figure 4.3: Linear Regression on Transformed Data with \( \lambda = 0.45 \) ..................... 68  
Figure 4.4: POD Curve for Transformed Data with \( \lambda = 0.45 \) .............................. 69  
Figure 4.5: Model with Experimental Data ............................................................... 71  
Figure 4.6: Data with Square-Root Transform Applied ............................................. 71  
Figure 4.7: Linear Regression with Square-Root Transformed Data ........................... 72  
Figure 4.8: POD Analysis for Square Root Transformed Data ................................... 73  
Figure 4.9: Experimental Data with 2nd Order Model ............................................. 74  
Figure 4.10: Hit/Miss Analysis ................................................................................. 77  
Figure 5.1: Analysis Attempt with A9001(3)D data set ........................................... 84
Figure 6.38: Histogram of $a_{90/95}$ values for 2 parameter logit model using MCMC... 132
Figure 6.39: Histogram of POD upper bound for 3 parameter logit model... 133
Figure 6.40: Histogram of lower confidence upper bound for 3 parameter logit model. 134
Figure 6.41: Histogram of POD upper bound for 4 parameter logit model... 135
Figure 6.42: Histogram of lower confidence upper bound for 4 parameter logit model. 136
Figure 6.43: Box plot for upper bound results with 3 and 4 parameter logit models. ... 137
Figure 6.44: Box plot for lower 95% confidence bound. ........................................ 138
Figure 6.45: Bayes factor Histogram for 2 parameter and 3 parameter model... 140
Figure 6.46: Bayes factor histogram for 2 parameter and 4 parameter model. ........... 141
Figure 6.47: Comparison of 2 parameter model results from simulation studies. ....... 142
Figure 6.48: Box plot of upper bound on POD and lower confidence bounds........... 143
LIST OF TABLES

Table 2.1: Generic Hit/Miss Data Set from ................................................................. 20
Table 2.2: Comparison of Classical Hit/Miss Methods ........................................... 24
Table 2.3: Generic Signal Response Data from ......................................................... 27
Table 3.1: Bayes Factor Results of Data Analysis ....................................................... 37
Table 3.2: Performance Metrics Results .................................................................... 38
Table 3.3: Asymptote Results of Analysis ................................................................. 38
Table 3.4: Bayes Factor Results from Analysis of A6003H Data Set ....................... 42
Table 3.5: Performance Metrics Results from Analysis of A6003H Data Set .......... 42
Table 3.6: Asymptote Results from Analysis of A6003H Data Set ......................... 42
Table 3.7: Candidate models for POD model of A6003H data set ........................... 54
Table 3.8: Summary of metrics for candidate models ............................................... 55
Table 3.9: Small data set for 2 inspection systems .................................................. 56
Table 3.10: Summary results for 2 inspection systems with limited data ............... 59
Table 4.1: Bootstrap results .................................................................................... 76
Table 4.2: Competing Model Comparison Study Results ......................................... 78
Table 5.1: NTIAC data set filename index ............................................................... 88
Table 5.2: Percentage of time 2 parameter model is best fit .................................. 88
Table 5.3: Percentage of time a_{90} and a_{90/95} exists ....................................... 89
Table 5.4: Probability of failure for varying probabilities of missing large flaws and upper bound for initial crack size ...................................................................... 93
Table 6.1: Summary of simulation study 1 results ............................................... 114
Table 6.2: Results summary for simulation study 2 for 100 observations .............. 128
Table 6.3: Results summary for simulation study 2 with 30 observations .......... 139
ACKNOWLEDGEMENTS

I would like to thank my advisors Dr. Ramana Grandhi and Dr. Frank Ciarallo for helping me to develop both technically and professionally. I learned a lot by their example. In addition, I appreciate the time and energy contributed by committee members Dr. Michael Grimaila, Dr. Yan Liu, Dr. Pratik Parikh, and Dr. Xinhui Zhang.

I also appreciate the support provided by the Air Force Research Laboratory (AFRL) that enabled me to complete this degree program. In particular the Materials State Awareness Branch assisted me both on a technical and personal level.

Lastly, and certainly not least, I thank my wife Reiko and my children Tommy and Rose for the sacrifices they made while I conducted research and wrote this dissertation. I also thank my parents for providing the foundation for me to achieve this goal.

The ‘R’ software and a particular package called ‘MH1823’ Version 2.5.4.4 developed by Charles Annis were used extensively in this work. They were downloaded from www.r-project.org and www.statisticalengineering.com/mh1823 respectively at the time of this writing. The Eureqa software was used in Chapter 3, and information about it can be found here: www.nutonian.com
1 INTRODUCTION

1.1 DAMAGE TOLERANCE PHILOSOPHY

Damage tolerance based design is an approach for sustainment of aircraft which was developed in response to many catastrophic structural failures that occurred in United States Air Force (USAF) aircraft. The damage tolerance philosophy assumes that there is damage in a component initially before it becomes operational. This requires that the component be designed to tolerate such damage from a safety perspective. Since it is assumed damage will grow when the component is in service, periodic inspections are used to detect critical damage. Risk assessment requires many inputs and one of them is a quantitative measure of the performance of a nondestructive inspection (NDI). Probability of detection (POD) studies are conducted to quantify the performance of NDI methods. Chapter 2 will go into more detailed discussion of both NDI and POD.

Consider the airframe structure in Figure 1.1. Note that there are thousands of fastener sites in this component alone. Fastener sites are stress raisers that are the source of fatigue cracks, and as these cracks grow, it leads to either catastrophic failure or time consuming and expensive maintenance procedures. If the cracks are detected at small sizes, repairs can be made, and catastrophic failures can be avoided. Millions of fastener sites on many different aircraft need to be inspected with NDI methods such as eddy
current and ultrasonics, and it will be shown that the capabilities of those methods need to be quantified. The origins of structural integrity management date back to the Wright brothers. The Wright Brothers conducted tests where forces that exceeded expected loads were slowly applied to ensure the integrity of their structures [1]. These tests are known as static tests. A safety factor was added to these static tests and this basic approach was used by the USAF for nearly 50 years. In 1958, two B-47 bombers were lost on the exact same day for the exact same reason. An investigation revealed that metal fatigue caused the wings to fail catastrophically in flight. A standard static test and an abbreviated flight load survey proved the structure would support at least 150% of its design limit load. There was no assurance that the structure would survive smaller cyclic loads in actual flight. The Aircraft Structural Integrity Program (ASIP) was initiated on June 12th, 1958 in response to this pair of mishaps and others.

Figure 1.1: Aircraft Structure with Numerous Fastener Sites that Require Inspection.
The ASIP program had 3 primary objectives [1]:

1) Control structural fatigue in aircraft fleet
2) Develop methods to accurately predict service life
3) Establish design and testing methods to avoid structural problems in future aircraft systems

This led to what is known as the “safe-life” approach. This approach required a full-scale airframe fatigue test. The number of successfully tested simulated flight hours was divided by a scatter factor (usually 4) to determine the service life of the aircraft.

The next event that caused changes in the ASIP program was the loss of an F-111 aircraft in 1969 [1]. This aircraft was expected to operate for 4,000 hours based on the safe-life approach. There was a wing separation that occurred after only 100 hours of flight. It turns out that this was because of a crack that initiated from a manufacturing defect. In response to this a two-phase program was initiated. In phase I, crack growth data in the particular material system was collected to develop a flaw growth model. Cold proof tests were used to demonstrate that critical size flaws were not present in critical forgings. The NDI techniques were improved for reinspection of components. This phase I part of the program allowed operations to resume at 80% of the designed capability. In phase II, NDI was incorporated during production. Fracture mechanics was used to determine the inspection intervals. The philosophy of damage tolerance was adopted in response to the failures of the safe-life approach. Most of the guidelines for
this approach were put in place in 1974 by ASIP. There were three options to satisfy damage tolerance requirements [1]:

1) Slow crack growth (most preferred option)
2) Fail-safe multiple load path
3) Fail-safe crack arrest

The damage tolerance approach to structural integrity used by the USAF has not changed in its essentials since 1974. Inspection requirements are in place based on initial flaw assumptions, and also inspection capability. The methods used to quantify inspection capability will be discussed in great detail in chapter 2. The situation today is that the inspection burden is increasing due to the age of the fleet. In many cases, aircraft are being operated well past twice their design lives, which is unknown territory since the fatigue tests were not conducted past this point. It’s possible that new inspections will be required in locations and components that were not anticipated. Furthermore, these inspections will need to have their capability validated in short-order to respond quickly to operational needs.

To manage the safety, readiness, and overall economic considerations for managing aircraft, risk analysis methods have been developed. According to the “Joint Service Specification Guide, Aircraft Structures” which was published in 1998 [2], the maximum acceptable frequency of catastrophic structural failure is $10^{-7}$ occurrences per flight. One of the key quantities of interest is the probability that the maximum stress in a flight will produce a stress intensity factor that exceeds the fracture toughness of the
material. This is computed based on fracture mechanics principles. The inputs to the risk analysis are probability distributions and include crack size distributions, maximum stress per flight, fracture toughness, and POD. There are also aircraft specific inputs such as number of locations, number of aircraft, hours per flight, and inspection intervals. The normalized stress intensity function and the crack growth curve are both deterministic inputs. Decisions involving safety, inspection intervals, repairs, modifications, and retirement can be analyzed quantitatively with risk analysis.

Equivalent flaw size (EFS) is a description of crack size in critical locations of interest in a structure and is modeled using a Weibull distribution [3]. Information on EFS is not easy to obtain, but risk assessment is very sensitive to this input. In particular, the results of risk assessment are sensitive to the accuracy of the extreme tails of the EFS distributions. Since data in the tail of the distribution is the most difficult to obtain, it’s extremely difficult to estimate tail behavior. It is desirable to use fracture mechanics to compute what is called the equivalent initial flaw size (EIFS) distribution. The EIFS is the distribution of defects from manufacturing before the aircraft was put in service. Repair flaw sizes are also modeled using the Weibull distribution. Fracture toughness is modeled as a normal distribution. The maximum stress per flight is modeled using extreme value statistics (the Gumbel distribution in particular).

POD is a measure of the reliability of an inspection [4]. It can be defined as follows: Given a population of cracks of size ‘a’ for a particular geometry, material,
orientation, location, and given a defined inspection system, the POD is the probability that selected cracks of size ‘a’ from the population will be detected. The models used for POD analysis are based on logistic regression and apply to two different kinds of data. Figure 1.2 shows a schematic for how POD is evaluated. The probe and flaw characteristics, instrument noise, and other aspects of data collection all contribute to POD. The signal is usually the magnitude of a voltage measurement from an NDI instrument. The first and most common form of data is known as hit/miss data. This is binary data indicating the detection or lack of detection of a flaw in an inspection. The second form of data is known in the literature as “â vs a” data. This is data where the signal response ‘â’ is recorded from an NDI instrument. The analysis methods currently available for “â vs a” data are applicable when there is a linear response between an explanatory variable such as crack length and the measurement response. Today, the lognormal or log-logistic 2-parameter POD model is commonly used [4]. The 2-parameter model assumes that the POD curve approaches one as flaw size goes to infinity, and approaches zero as flaw size goes to zero. This model can be deficient because the curve is forced to approach one for large flaw sizes, but the POD for large flaw sizes may not be one. The reasons for this range from human factors to variation in flaw morphology. The proposed research will investigate the benefits of 3 or 4-parameter models to address this deficiency. There are two outputs of risk analysis: The first output is the single flight probability of failure (SFPoF), also known as the hazard rate, h(t), as a function of time t. It is the probability of failure during a given flight at time t given that a failure has not occurred on that or any previous flight. The second output is the
cumulative distribution function $F(t)$ of the time of the first failure, $t$. A schematic of how risk analysis is conducted in a damage tolerance management framework is shown in Figure 1.3.

Figure 1.2: Schematic for POD Evaluation Process [5]
1.2 OVERVIEW OF THE PROBLEM

The purpose of this dissertation is to provide statistical inference methods that estimate the extremes of POD curves and show the impact of these methods on decision making. It is anticipated that the extreme tail of the POD curve for large flaw sizes will influence risk analysis. The risk analysis in this work has many inputs, and some of the conclusions may prove surprising. A number of concerns have been raised about POD recently, and this work attempts to address them at least in part. One concern is that the models that an are commonly used for Hit/Miss POD analysis could be used even if the assumptions of these models are violated [6]. In reality, 100% POD doesn’t exist for any flaw size, and there may be many false calls for small flaw sizes or for observations where there is no flaw. A false call is defined as an indication that there is a flaw present when there is in fact no flaw. Intuitively, if 100% POD doesn’t exist for very large flaw
sizes, then that implies there is a finite probability of a very large flaw being missed during inspection. It is anticipated that this will have a significant impact on the probability of fracture and therefore overall risk. Better analysis methods to estimate the POD for large flaw sizes are therefore desirable.

In addition to analysis methods for hit/miss, modifications are also needed to properly analyze data where the entire signal response is captured. Currently analysis of this type of data relies on two assumptions: 1) Homoscedasticity (constant variance for all flaw sizes), and 2) a linear model. A typical scenario will show that these two assumptions are often violated, and a remedy is required.

The specific contributions in this work are the following:

1) A quantitative comparison study was conducted to determine the variability in current analysis methods for Hit/Miss data.
2) New analysis methods were developed for hit/miss data to analyze data sets that should not be analyzed with the current models suggested in the literature.
3) Transformations were developed for data sets that violate the constant variance assumption necessary in “â vs a” analysis.
4) Methods were developed for the POD confidence bound calculation for situations where the relationship between the explanatory variable ‘a’ and the signal response ‘â’ is not linear.
5) Hit/miss analysis methods used to examine inspection data sets from database.
6) A sensitivity analysis showed the impact of different POD analysis methods on overall risk assessment.

7) Simulations were conducted to quantify improvements made to POD in this proposal.

1.3 ORGANIZATION OF DISSERTATION

This dissertation is organized as follows: This chapter provides an overview of where the work fits into the larger picture of aircraft sustainment. Chapter 2 discusses POD in detail and compares some of the current POD methods. Chapter 3 proposes modifications to hit/miss analysis to better model the extremes of the POD curve and provides a model selection procedure to determine the number of parameters necessary for logistic regression. Chapter 4 develops improvements in â vs a analysis to address violations of homoscedasticity and the assumption of a linear relationship between flaw and instrument response. In particular the Box-Cox transformation is used to achieve constant variance, and bootstrapping is used to compute confidence bounds for advanced models. Chapter 5 examines existing data sets to show the impact of new methodologies and also show the feasibility of incorporating this work into risk analyses. Chapter 6 contains rigorous simulation studies to quantify the improvements to POD described in chapters 2 and 3. Chapter 7 will summarize the results and provide recommendations for further research.
2 PROBABILITY OF DETECTION

For the purposes of this work, Nondestructive Evaluation (NDE) is defined according to the definition provided in [7]: “Nondestructive evaluation (NDE) is used to characterize the state or properties of components or other units of material without causing any permanent physical change to the units.” Nondestructive Inspection (NDI) is typically defined in industry as a process for determining whether a flaw is present in a component or not. So NDE refers to characterization and NDI refers to detection. Common NDI techniques include X-ray, Ultrasonics and Eddy Current. All of these techniques involve an external excitation, and a measurement of the response due to the excitation. The response will often be referred to as measurement response, voltage output, or simply measurement in this work. Statistical models are necessary because of the intrinsic variability of the inspection process. Different flaws of the same size will produce different responses in inspection equipment due to artifacts such as crack contact conditions, geometry, and orientation. The statistical analysis of such measurement responses is the focus of this entire dissertation.

The capabilities of these NDI techniques are quantified via Probability of Detection (POD) studies. POD is a statistical description of the capability of an inspection process. Inspection data can be collected in two ways: (1) A binary response (detect / no detect or
more formally, Bernoulli data) is recorded or (2) The signal response on the NDI instrument is recorded. These methods are commonly referred to as hit/miss and “â vs a” respectively. Some NDI techniques, such as fluorescent penetrant, can only provide hit/miss data. For inspections based on ultrasonic and eddy current information, a threshold is set and the inspection result is based on whether the threshold is exceeded. It is advantageous to record the voltage measurement from the instrument to evaluate POD. Both POD analysis methods assume that the detection capability is a function of flaw size, and similar functional forms can be used for the POD model with either method.

This chapter will discuss three approaches to Hit/Miss POD:

1) Binomial based methods
2) Parametric model based on Wald Statistics
3) Parametric model based on Likelihood Ratio Method

The terminology used in POD is as follows:

- \(a_{50}\) – estimate of flaw size for 50% POD
- \(a_{90}\) – estimate of flaw size for 90% POD
- \(a_{90\%5}\) – lower bound at 95% confidence for 90% POD

\(a_{90\%5}\) is a scalar quantity commonly used as a performance metric for comparison of NDE systems and risk calculations.
2.1 HIT/MISS ANALYSIS

Hit/miss analysis or Bernoulli data collection continues to be the most widespread data available used for POD evaluation 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 NASA and the United States Air Force (USAF) \[8,9\]. The later study is probably the largest study on the reliability of NDE techniques in history.

2.1.1 Binomial Based Methods

Initially, POD was analyzed using Binomial statistics, and this led to two general methods:

1) Range Interval Methods

2) “29 of 29”

In range interval methods hit/miss observations are grouped into intervals based on flaw size, and then statistical models based on the binomial distribution are used to analyze the data in each interval to establish the POD and associated lower confidence bound. For the “29 of 29” approach it was thought that if the same size flaw could be detected successfully 29 out of 29 times, then one could claim, for that particular flaw size, that the detection capability exists for 90% POD with 95% confidence \[10\]. These methods are inadequate for many reasons:

1) POD changes as a function of flaw size.
2) False calls are not taken into account.

3) The confidence bounds are greatly influenced by crack size.

4) If one crack is missed, the study will not provide an a_{90/95} result.

Since the binomial and “29 of 29” models are considered obsolete, the methods for calculating confidence bounds for those methods are not discussed here. For these obsolete methods, the POD analysis consists of using confidence intervals for the corresponding binomial distribution based on flaw size.

### 2.1.2 Parametric Methods for Hit/Miss Analysis

The USAF and statisticians developed methods based on Logistic regression to analyze hit/miss data [11]. The fundamental functional form of the POD curve is shown in Eq. (2.1), where ‘a’ is the flaw size, μ is the flaw size which can be detected with 50% probability, also known as a_{50}, σ is a slope parameter, and \( \phi \) represents a generic function. The two parameters that describe the POD curve are μ and σ. The intercept and slope of the linear model are designated b_0 and b_1 respectively. Individual observations are designated by the subscript ‘i’.

\[
p_i = \phi \left( \frac{\ln(a_i) - \mu}{\sigma} \right) = \phi(b_0 + b_1 \ln(a_i))
\] (2.1)
Many models were applied to the study from [9]. These included the following:

1) “Lockheed”
2) Weibull
3) Probit
4) Log Probit
5) Log Odds – linear scale
6) Log odds – log scale
7) Arcsine

The viability of the models was assessed using Bartlett’s test to evaluate the equality of variance and the Shapiro-Wilks W test to evaluate normality [11]. The two most appropriate models selected were the log odds or logit model shown in Eq. (2.2) and the cumulative log normal or probit model shown in Eq. (2.3). The standard cumulative normal distribution function is represented by $\Phi$. The question of model form will be reexamined in Section 3.5 with modern statistical tools.

\[
p_i = \frac{\exp(b_0 + b_1 \ln(a_i))}{1 + \exp(b_0 + b_1 \ln(a_i))} \quad \text{(logit)} \tag{2.2}
\]

\[
p_i = \Phi(b_0 + b_1 \ln(a_i)) \quad \text{(probit)} \tag{2.3}
\]
The logit and probit models have different tail behavior. Since large flaw sizes are in the tail of the flaw size distribution this behavior is of central interest in determining the best model. 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 due to the computational capability available at that time. 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 [12]. The major challenge in the development of POD at that time was accurate confidence bound calculations. It needs to be emphasized that a POD curve without confidence bounds has little value.

Confidence bound calculation methods can be divided into 2 categories:

1) simultaneous or global
2) point estimates or local

For each of those methods, there are currently two general types of statistical methods that can be used:

1) Wald [13]
2) Likelihood Ratio
2.1.3 Wald Method

The objective is to determine confidence intervals for the two parameters $\mu$ and $\sigma$. The approach for this originated from papers by Cheng and Iles [14, 15] which provided a generic approach for providing confidence bounds on cumulative distribution functions. One of the properties of maximum likelihood estimation (MLE) of $\mu$ and $\sigma$, which is designated $\hat{\theta}$, is that $\hat{\theta}$ has an asymptotically multivariate normal distribution with mean $\theta$ and variance-covariance matrix $(I(\theta))^{-1}$, where $I(\theta)$ is known as the Fisher information matrix. The Fisher information matrix is determined from the parameter estimation problem. In the theory of statistics, a quadratic form like that shown in Eq. (2.4) is asymptotically a chi-squared variable.

$$Q_1(\theta) = (\hat{\theta} - \theta)^T I(\theta) = (\hat{\theta} - \theta)$$ (2.4)

$Q_1(\theta)$ is known as a Wald statistic [13] which may not be inaccurate for large sample sizes, but it is not recommended for small sample sizes. Based on the results of simulation studies in Chapter 6, a small sample size is roughly considered to be smaller than 100 observations. Another approximation that was made is based on the fact that the information matrix is evaluated at the MLE of the parameters instead of their true unknown values.
2.1.4 Likelihood Ratio Method

As Cheng and Iles point out [14], a better method to calculate the confidence interval is to use the likelihood ratio. The definition of likelihood is the probability of the data given particular values of the parameters in the statistical model. Eq. (2.5) shows the likelihood definition where $a_i$ is the flaw size for the $i$th inspection, and $x_i = 1$ for flaw detected and $x_i = 0$ for flaw not detected.

$$L = \prod_{i=1}^{n} P(a_i, x_i)$$

Eq. (2.6) is the likelihood statistic that is used for this method. It is an important property of mathematical statistics that the ratio of two likelihoods is distributed as a chi-squared distribution. This can be used to calculate a confidence interval. It involves an optimization problem of maximizing the crack size ‘a’ given the constraint of the chi-squared statistic subtracted from the likelihood statistic.

$$Q_2(\theta) = -2\ln \frac{L(\theta)}{L(\hat{\theta})}$$

In their computations, Cheng and Iles did not directly use the likelihood ratio method. They approximated the likelihood ratio as an ellipse in parameter space for $\mu$ and $\sigma$. They did this because at that time it was too computationally intensive to directly calculate the likelihood ratio.
2.1.5 Discussion

Up until the year 2000, Berens used simultaneous confidence bounds across all crack sizes to provide bounds on the $\mu$ and $\sigma$ parameters. At some point that was deemed too conservative, and point estimates were used thereafter, such that the bounds pertained only to a specific crack size [16]. The calculations are very similar. For the likelihood ratio method, the point estimate confidence interval calculation is the same as for simultaneous confidence intervals except that the critical value of the chi-squared distribution is different.

Since 2000, several modifications have been made or suggested for POD analysis. The POD/SS software developed by Berens [17] was changed to calculate the confidence bounds for each individual flaw size locally rather than to the entire POD curve. Both of these confidence bound calculations were based on the Wald Statistic [13]. Later, the likelihood ratio method was suggested for more accurate confidence bound calculations on the model parameters estimates [18, 19]. 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 [20]. 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 [21,22]. There is also parallel work being done in medical statistics similar to NDE reliability [23]. The primary requirement for hit/miss data in MIL-HDBK-1823A is that
the POD is zero as the flaw size approaches zero, and that POD is one as the flaw size approaches infinity. An example of a data set that does not meet this requirement was shown in [6]. In chapter 3, a method will be introduced to handle such a data set.

### 2.2 CASE STUDY

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 fluorescent penetrate inspection information. The methods presented here can be applied to any inspection data with binary responses.

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

Table 2.1: Generic Hit/Miss Data Set from [4]
2.3 COMPARISON STUDY

The results reported in [4] were calculated using the log odds model in Eq. (2.1), and the confidence bounds were calculated simultaneously for the entire POD curve; thus the $a_{90/95}$ number is very conservative. POD/SS software version 3 [17] was used to calculate the POD curve and the confidence bounds using the Wald method. These results can be reproduced following the calculations in [16]. The results are displayed in Figure 2.1. The key quantities $a_{50}$ (also $\mu$), $a_{90}$, and $a_{90/95}$ are indicated on this plot. The $a_{50}$ and $a_{90}$ values are the flaw sizes that would be detected 50% and 90% of the time respectively if the experiment is repeated. The lower 95% confidence bound on 90% POD is denoted as $a_{90/95}$.

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. The approach used to determine that the probit model Eq. (2.3) was a better fit will be explained in chapter 3. This calculation was done using Eq. (2.6) implemented in software known as mh1823 which is a library available for use with the ‘R’ programming language [24]. The POD curve with confidence bounds calculated using the likelihood ratio method is displayed in Figure 2.2.
Figure 2.1: Hit/Miss Analysis with Wald Confidence Bounds [17]
The numerical results are shown in Table 2.2. Both $a_{50}$ and $a_{90}$ are approximately the same as expected, but there is large discrepancy in the $a_{90/95}$ value. There is definitely a discrepancy between simultaneous and point-estimate approaches to POD as illustrated with the Wald method. The likelihood ratio method considers all flaw sizes simultaneously and has a moderately significant difference compared to the Wald simultaneous case. In chapter 3, new methods will be developed to better model tail behavior at the extremes of the POD curve. The dots represent individual observations.

Figure 2.2: Hit/Miss Analysis Likelihood Ratio Confidence Bounds
<table>
<thead>
<tr>
<th>Statistical Method</th>
<th>(a_{50})</th>
<th>(a_{90})</th>
<th>(a_{90/95})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Simulatenuous Berens ASM Result (mm)</td>
<td>2.620</td>
<td>5.340</td>
<td>21.600</td>
</tr>
<tr>
<td>Wald Point-Estimate PODSS[17]</td>
<td>2.610</td>
<td>5.252</td>
<td>8.776</td>
</tr>
<tr>
<td>Likelihood Ratio Logit[19]</td>
<td>2.613</td>
<td>5.354</td>
<td>18.550</td>
</tr>
<tr>
<td>Likelihood Ratio Probit[19]</td>
<td>2.610</td>
<td>5.252</td>
<td>17.020</td>
</tr>
</tbody>
</table>

Table 2.2: Comparison of Classical Hit/Miss Methods.

2.4 CLASSICAL METHODS FOR SIGNAL RESPONSE ANALYSIS

There is limited information in hit/miss data, so it’s desirable to record the signal response from the NDE instrument that led to the hit/miss result. If this additional information is used effectively, there will be less uncertainty in the POD result. Analysis incorporating the signal strength is commonly called “\(\hat{a} vs a\)” in the literature, where ‘\(\hat{a}\)’ is the response variable, and ‘a’ is the explanatory variable which may be crack size for example. The method developed by Berens and Hovey [25] involves two steps. First a linear measurement model shown in Eq. (2.7) needs to be determined to explain the relationship between the flaw size variable ‘a’ (usually crack size), and the signal response ‘\(\hat{a}\)’, with the assumption that \(\varepsilon \sim N(0, \sigma^2_\varepsilon)\). The slope and intercept parameters in this model are designated \(\beta_0\) and \(\beta_1\) respectively.
\[ \hat{a} = \beta_0 + \beta_1 a + \epsilon \] (2.7)

Many NDE techniques and instruments have limited sensitivity, and it is common for measurements at very small flaw sizes to be in the noise. It is difficult to differentiate signal from noise for these very small flaw sizes and these data are typically censored in the analysis. This is called left-censoring. Similarly, measurements of large flaw sizes that are in the saturation region of the instrument are also typically removed from the analysis. This is called right-censoring. Censored regression is necessary to analyze these data sets and thus ordinary least squares regression is not appropriate. The appropriate technique for this type of data is maximum likelihood (ML) estimation. Maximum Likelihood estimation with censored observations is well developed in the survival analysis literature [26]. After the measurement model is constructed, it must be transformed into a POD model via a common statistical transformation technique called the delta method. The delta method is used to approximate a random variable that is a function of an asymptotically normal statistical estimator. It is very similar to a Taylor series expansion, and explained in full detail in Section G.5.4.7 of [20].

Once again, data from [4] will illustrate how \( \hat{a} vs a \) analysis is done today. The crack size is designated ‘\( a \)’, and the signal response is ‘\( \hat{a} \)’. The source of this data is a highly automated eddy current inspection of 29 cracks in flat plates. Since this is from an automated system the variance of the signal \( \hat{a} \) is small compared to the variance of \( \hat{a} \) for
handheld inspections. In this case the saturation limit for the instrument is 4095 volts, and the signal threshold is 75 volts. In Table 2, the up arrows indicated that the response likely may have been above the recorded value, but the instrument reached saturation. Also, the down arrows indicate that no signal was recorded because their â values were below the recording signal threshold.

Figure 2.3 displays the data from Table 2.3. It shows that some observations are saturated in the large crack region. The data doesn’t appear to violate any assumptions for linear regression, but traditionally the natural log of the flaw size is taken, and sometimes the natural log of the signal response is taken. For this case, only the natural log of the flaw size is taken. The next step is to apply censored regression to determine the measurement model. The results are shown in Figure 2.4, and the values for the estimated parameters are listed on the figure as well. The measurement model is then transformed into a POD model according to the methods in [20] as shown in Figure 2.5.
<table>
<thead>
<tr>
<th>Flaw size (mm)</th>
<th>Response</th>
<th>Flaw size (mm)</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>1052</td>
<td>0.18</td>
<td>409</td>
</tr>
<tr>
<td>1.40</td>
<td>4095↑</td>
<td>0.28</td>
<td>895</td>
</tr>
<tr>
<td>0.38</td>
<td>1480</td>
<td>0.20</td>
<td>374</td>
</tr>
<tr>
<td>0.25</td>
<td>723</td>
<td>0.79</td>
<td>4095↑</td>
</tr>
<tr>
<td>0.74</td>
<td>4095↑</td>
<td>0.23</td>
<td>638</td>
</tr>
<tr>
<td>0.48</td>
<td>2621</td>
<td>0.15</td>
<td>533</td>
</tr>
<tr>
<td>0.30</td>
<td>377</td>
<td>0.08</td>
<td>150</td>
</tr>
<tr>
<td>0.23</td>
<td>223</td>
<td>0.28</td>
<td>749</td>
</tr>
<tr>
<td>0.56</td>
<td>1654</td>
<td>0.2</td>
<td>433</td>
</tr>
<tr>
<td>1.65</td>
<td>4095↑</td>
<td>0.36</td>
<td>879</td>
</tr>
<tr>
<td>0.08</td>
<td>75↓</td>
<td>0.23</td>
<td>286</td>
</tr>
<tr>
<td>0.25</td>
<td>669</td>
<td>0.23</td>
<td>298</td>
</tr>
<tr>
<td>0.18</td>
<td>374</td>
<td>0.41</td>
<td>1171</td>
</tr>
<tr>
<td>0.03</td>
<td>75↓</td>
<td>2.54</td>
<td>4095↑</td>
</tr>
</tbody>
</table>

Table 2.3: Generic Signal Response Data from [4]
Figure 2.3: Scatter plot of data in Table 2.3.
Figure 2.4: Censored Regression Performed on Data from Table 2.3
Figure 2.5: POD Curve for a vs a Data from Table 2.3.
3 MODERN STATISTICAL ANALYSIS OF HIT/MISS DATA

3.1 BAYESIAN APPROACH TO HIT/MISS

The statistical methods discussed in the previous section are state-of-the-art methods in conventional statistics. There are cases where the data suggests that the POD does not approach one as the flaw size approaches infinity. There is also the problem of false calls where the POD curve doesn’t approach zero as the flaw size goes to zero. To address these issues, it was proposed that additional parameters should be incorporated in the model [27], but inference using conventional methods is difficult. Eqs. (3.1) and (3.2) show the form of a 3-parameter model with a lower asymptote ‘α’ for the logit and probit models respectively. Eqs. (3.3) and (3.4) show the form of a 3-parameter model with an upper asymptote ‘β’. The β term is a measure of the probability of missing cracks as the crack size goes to infinity. The 4-parameter model has both terms in it and is described in Eqs. (3.5) and (3.6). Proper estimation of the upper asymptote is very important for addressing pitfalls in POD analysis.

\[
p_i = \alpha + (1 - \alpha) \cdot \frac{\exp(b_0 + b_1 \log(a_i))}{1 + \exp(b_0 + b_1 \log(a_i))} \quad \text{ (logit) } (3.1)
\]

\[
p_i = \alpha + (1 - \alpha) \cdot \Phi(b_0 + b_1 \log(a_i)) \quad \text{ (probit) } (3.2)
\]
To date, there is no literature for confidence bound calculation for models with three or four parameters. 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 [28]. The advantages of going beyond conventional statistics to Bayesian statistics are twofold: 1) Markov chain 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.

The mathematical form of Bayes’ rule is given by Eq. (3.7).
The data denoted by D follows a model M, and θ is a set of parameters in the model. P(θ|D,M) literally reads as the probability of the parameters given the data, and it is commonly called the “posterior” distribution. P(D|θ,M) literally reads as the probability of the data given the parameters, and is also known as the “likelihood”. P(θ|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. (3.7) 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 performed 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 3.1.
3.2 MARKOV CHAIN MONTE CARLO (MCMC)

The concepts of Bayesian analysis are intuitive, but there is usually a computational challenge in evaluating the integral in the bottom of Eq. (3.7). There are special cases where the prior distribution is a conjugate to the likelihood function, so the posterior can be solved analytically, but this is rare, and may explain why Bayesian analysis was not more widespread until there was easy access to computers. The problem is that the denominator of Eq. (3.7) is a high-dimensional integral, and numerical integration techniques are quickly overwhelmed, so sampling methods need to be
employed. The prior and likelihood need to be computed for random values of \( \theta \), from which the posterior can be sampled. Then statistics about the distribution can be determined.

The two most popular methods for doing this are the Metropolis-Hasting algorithm [29,30] and Gibbs sampling [31]. To illustrate these methods, Eq. (3.7) is reformulated to show that the posterior is simply proportional to the probability of the data given the model parameters multiplied by some prior thought to be known as shown in Eq. (3.8). Even though it is usually not feasible to evaluate the integral in Eq. (3.7), there is still a need to determine the posterior, and MCMC enables this by sampling the posterior without evaluating the integral.

\[
P(\theta | D, M) \equiv P(D | \theta, M)P(\theta | M)
\] (3.8)

The goal is to sample the left side of the equation without knowing the normalization constant that’s in the denominator of Eq. (3.7). The contribution made by Metropolis and his team was a significant step forward from general Monte Carlo methods where the space is sampled completely at random. Instead of sampling unrelated points, Metropolis et al. sampled a Markov chain [29]. This Markov chain is a sequence of points where each new sample depends stochastically on the previous sample. It can be proved under certain conditions that if this sequence is ergodic, then the samples are proportional to the posterior, and this is how the high dimensional normalization integral is avoided. In 1970,
this was improved upon by Hastings by adding acceptance probability criterion which means that low probability samples are mostly, but not all excluded [30]. This is conceptually very similar to the optimization technique known as simulated annealing. A recent method called “slice sampling” was proposed by Neal [32], and is used in the following work. This sampling is more efficient because the space is sampled uniformly from the region inside its density function. This method alternates uniform sampling in the vertical direction with uniform sampling from the horizontal direction. In addition to the advantages in efficiency, slice sampling is used in this work due to the ease of implementation in Matlab.

3.3 POD RESULTS WITH MCMC

The data in chapter 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 [33]. Mathematically, the Bayes factor is the ratio of the marginal likelihoods of the competing models. For example, the Bayes factor for Model i vs. Model j is \( \frac{ML(\text{Model } i)}{ML(\text{Model } j)} \). Since the marginal likelihood indicates plausibility of a model, if the Bayes factor is larger than 1, it means that Model 1 is more plausible, or, in other words, Model 1 is more strongly supported by the data. Using this tool, the
candidate models can be compared in pairs and the best model will be determined. The key to implementing this method is to calculate the marginal likelihoods which can be realized by using MCMC algorithms. The marginal likelihood and corresponding Bayes factors are displayed in Table 3.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal Likelihood</th>
<th>Bayes factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 parameter Logit</td>
<td>2.78E-08</td>
<td>2.109</td>
</tr>
<tr>
<td>2 parameter Probit</td>
<td>5.86E-08</td>
<td>1.000</td>
</tr>
<tr>
<td>3 parameter lower bound Logit</td>
<td>0.39E-08</td>
<td>14.976</td>
</tr>
<tr>
<td>3 parameter lower bound Probit</td>
<td>0.14E-08</td>
<td>42.339</td>
</tr>
<tr>
<td>3 parameter upper bound Logit</td>
<td>0.71E-08</td>
<td>8.269</td>
</tr>
<tr>
<td>3 parameter upper bound Probit</td>
<td>0.33E-08</td>
<td>17.914</td>
</tr>
<tr>
<td>4 parameter Logit</td>
<td>0.04E-08</td>
<td>130.735</td>
</tr>
<tr>
<td>4 parameter Probit</td>
<td>0.31E-08</td>
<td>18.692</td>
</tr>
</tbody>
</table>

Table 3.1: Bayes Factor Results of Data Analysis

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 3.2 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 hit/miss data for small sample sizes such as this one. Table 3.2 shows the $a_{50}$, $a_{90}$, and $a_{90/95}$ values for each of the models. As is often the case, the $a_{50}$ and $a_{90}$ values do not differ significantly. There are differences in the $a_{90/95}$ values. These differences are not
surprising because the methods of their determination are based on different assumptions and approximations as discussed in Chapter 2. Table 3.3 shows the estimates for the lower and upper asymptotes for the 3 and 4-parameter models. Note that the upper asymptote is nowhere close to 90%, so these results address the concerns in [6] about POD evaluation of this particular data set, and should also serve as a warning of what may happen if 2-parameter models are used indiscriminately.

<table>
<thead>
<tr>
<th>Model</th>
<th>$a_{90}$</th>
<th>$a_{40}$</th>
<th>$a_{90/95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 parameter Logit</td>
<td>2.611</td>
<td>5.333</td>
<td>10.136</td>
</tr>
<tr>
<td>2 parameter Probit</td>
<td>2.611</td>
<td>5.353</td>
<td>10.075</td>
</tr>
<tr>
<td>3 parameter lower bound Logit</td>
<td>3.072</td>
<td>5.253</td>
<td>10.355</td>
</tr>
<tr>
<td>3 parameter lower bound Probit</td>
<td>3.352</td>
<td>5.153</td>
<td>10.415</td>
</tr>
<tr>
<td>3 parameter upper bound Logit</td>
<td>2.191</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 parameter upper bound Probit</td>
<td>1.951</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4 parameter Logit</td>
<td>2.473</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4 parameter Probit</td>
<td>2.429</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wald point-wise estimates - PODSS[17]</td>
<td>2.610</td>
<td>5.252</td>
<td>8.776</td>
</tr>
<tr>
<td>Likelihood Ratio Probit [24]</td>
<td>2.610</td>
<td>5.252</td>
<td>17.020</td>
</tr>
</tbody>
</table>

Table 3.2: Performance Metrics Results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lower asymptote</th>
<th>Upper asymptote</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 parameter lower bound Logit</td>
<td>0.163</td>
<td>-</td>
</tr>
<tr>
<td>3 parameter lower bound Probit</td>
<td>0.189</td>
<td>-</td>
</tr>
<tr>
<td>3 parameter upper bound Logit</td>
<td>-</td>
<td>0.612</td>
</tr>
<tr>
<td>3 parameter upper bound Probit</td>
<td>-</td>
<td>0.577</td>
</tr>
<tr>
<td>4 parameter Logit</td>
<td>0.116</td>
<td>0.662</td>
</tr>
<tr>
<td>4 parameter Probit</td>
<td>0.113</td>
<td>0.623</td>
</tr>
</tbody>
</table>

Table 3.3: Asymptote Results of Analysis
3.4 CHALLENGE DATA SET EXAMPLE

One of the data sets referred to in [6] is identified as A6003H, as shown in Figure 3.3, which has 184 observations [34]. The data sets that begin with ‘A’ from this database are eddy current inspections. 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 [20]. 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 [3], the likelihood ratio method for both logit and probit models [19], 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 3.4. 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 an \( a_{90/95} \) estimate does not exist 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 3.5 since it is also an acceptable model.
Figure 3.3: Plot of observations from A6003H data set [34].
### Table 3.4: Bayes Factor Results from Analysis of A6003H Data Set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Marginal Likelihood</th>
<th>Bayes Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 parameter Logit</td>
<td>0.0001039E-27</td>
<td>68,210</td>
</tr>
<tr>
<td>2 parameter Probit</td>
<td>0.0000003418E-27</td>
<td>20,730,000</td>
</tr>
<tr>
<td>3 parameter lower bound Logit</td>
<td>0.0000007304E-27</td>
<td>970,300</td>
</tr>
<tr>
<td>3 parameter lower bound Probit</td>
<td>0.000000368E-27</td>
<td>19.26</td>
</tr>
<tr>
<td>3 parameter upper bound Logit</td>
<td>0.373E-27</td>
<td>18.98</td>
</tr>
<tr>
<td>3 parameter upper bound Probit</td>
<td>7.087E-27</td>
<td>1.000</td>
</tr>
<tr>
<td>4 parameter Logit</td>
<td>0.794E-27</td>
<td>8.921</td>
</tr>
<tr>
<td>4 parameter Probit</td>
<td>0.357E-27</td>
<td>19.85</td>
</tr>
</tbody>
</table>

### Table 3.5: Performance Metrics Results from Analysis of A6003H Data Set.

<table>
<thead>
<tr>
<th>Model</th>
<th>(a_{50}) (mm)</th>
<th>(a_{90}) (mm)</th>
<th>(a_{90/95}) (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 parameter Logit (MCMC)</td>
<td>2.042</td>
<td>3.403</td>
<td>3.953</td>
</tr>
<tr>
<td>2 parameter Probit (MCMC)</td>
<td>1.980</td>
<td>3.815</td>
<td>4.451</td>
</tr>
<tr>
<td>3 parameter lower bound Logit (MCMC)</td>
<td>2.051</td>
<td>3.452</td>
<td>3.992</td>
</tr>
<tr>
<td>3 parameter lower bound Probit (MCMC)</td>
<td>2.031</td>
<td>3.832</td>
<td>4.432</td>
</tr>
<tr>
<td>3 parameter upper bound Logit (MCMC)</td>
<td>2.051</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3 parameter upper bound Probit (MCMC)</td>
<td>2.051</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4 parameter Logit (MCMC)</td>
<td>2.083</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4 parameter Probit (MCMC)</td>
<td>2.091</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 3.6: Asymptote Results from Analysis of A6003H Data Set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Lower Asymptote</th>
<th>Upper Asymptote</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 parameter lower bound Logit (MCMC)</td>
<td>0.030</td>
<td>-</td>
</tr>
<tr>
<td>3 parameter lower bound Probit (MCMC)</td>
<td>0.034</td>
<td>-</td>
</tr>
<tr>
<td>3 parameter upper bound Logit (MCMC)</td>
<td>-</td>
<td>0.922</td>
</tr>
<tr>
<td>3 parameter upper bound Probit (MCMC)</td>
<td>-</td>
<td>0.921</td>
</tr>
<tr>
<td>4 parameter Logit (MCMC)</td>
<td>0.045</td>
<td>0.921</td>
</tr>
<tr>
<td>4 parameter Probit (MCMC)</td>
<td>0.044</td>
<td>0.921</td>
</tr>
</tbody>
</table>
Figure 3.4: 3-parameter Probit Model with Upper Asymptote for A6003H Data Set.

Figure 3.5: 4-parameter Probit Model with Asymptotes for A6003H Data Set.
3.5 RE-EXAMINATION OF MODEL FORM VIA SYMBOLIC REGRESSION

As mentioned in Chapter 2, the early work in POD modeling compared many candidate models using the Bartlett’s test to evaluate the equality of variance and the Shapiro-Wilks test to evaluate normality. There are now modern tools based on recent advances in evolutionary computation combined with surrogate models that can investigate the viability of many more models than could have been investigated a few decades ago when the original research was conducted. In this chapter a software package called Eureqa [35] is used to revisit the model-form for POD models. The Eureqa software has been used to quickly determine governing equations from experimental data [36]. It’s based on major advances in evolutionary computation. In particular, a coevolution algorithm of fitness predictors is used [37]. This addresses a long standing challenge in evolutionary computation and the algorithm is able to reduce fitness evaluation cost while maintaining evolutionary progress. One of the keys to this is the use of approximate fitness calculations [38] in the context of symbolic regression [39]. In this arrangement, functional expressions are represented as a binary tree of primitive operations. For example, common operations such as abs, exp, log, add, sub, mult, and div are a subset of the possible operations that can be combined or mutated. The common concepts in genetic programming such as crossover are also incorporated. A goodness of fit measure of a subset of experimental observations and inputs is quickly calculated using a fitness predictor. In summary, this software enables millions of
approxiations to candidate models to be evaluated according goodness of fit criteria supplied by the user.

The dataset A6003H data set which was investigated in the previous section is used again here. Recall, that this data set originated from a typical eddy current inspection where there were a significant number of misses for large flaw sizes, which made it difficult to model the inspection capability with a 2 parameter logit or probit model. Figure 3.6 shows the raw data and the first candidate solution provided by Eureqa which simply a constant function at 0.654. The Pareto frontier is shown in Figure 3.7, and note that the solution with the least complexity, but the greatest error is the constant function and identifiable by the large dot in the plot. The next major change in complexity leads to the linear model shown in Figure 3.8. The sine wave, step, and Gauss function models appear in Figure 3.9, Figure 3.10, and Figure 3.11 respectively. In Figure 3.12, the familiar 2 parameter logistic functional form is shown and has a complexity value of nine shown in the Pareto frontier graph in Figure 3.13. This is where the Pareto frontier gets very interesting for the purposes of model selection, because the complexity of nine is close to the point of inflection, which is a common method to determine the best tradeoff between model error and complexity. Next, the logistic regression model with the upper asymptote is shown in Figure 3.14. The Pareto frontier graph in Figure 3.15 indicates that the complexity of this model is 11 and there are no significant reductions in error as the complexity increases beyond this. This is one more
piece of evidence that the logistic model-form with an asymptote is a reasonable model for this type of inspection data. Figure 3.16 and Figure 3.17 show models with additional complex functions such as Erf and Gauss that don’t really reduce the error substantially. Figure 3.18 more clearly illustrates the risks over fitting the data. Figure 3.19 shows that this over fitting does little to reduce the error. Figure 3.20 and Figure 3.21 further illustrate the dangers of over fitting with no benefit to error reduction. The results are tabulated in Table 3.7 and Table 3.8.

Figure 3.6: Constant model used to model the A6003H data set.
Figure 3.7: Pareto frontier displaying the least complex solution among others.

Figure 3.8: Linear regression model for A6003H data set.
Figure 3.9: Sine wave model for A6003H data set.

Figure 3.10: Step function model for A6003H data set.
Figure 3.11: Gauss function model for A6003H data set.

Figure 3.12: Logistic regression model for A6003H data set.
Figure 3.13: Pareto frontier showing where Logistic model ranks in complexity.

Figure 3.14: Logistic regression model with upper asymptote.
Figure 3.15: Pareto frontier showing complexity of Logistic model with asymptote.

Figure 3.16: Model with Erf and Gauss for A6003H data set.
Figure 3.17: Model with Gauss and tan function for A6003H data set.

Figure 3.18: Model for A6003H data set containing gauss, cos, and tan functions.
Figure 3.19: Pareto frontier showing complexity of model with additional functions.

Figure 3.20: Model for A6003H data that incorporates high level of complexity.
Figure 3.21: Pareto frontier showing model with highest complexity for this study.

<table>
<thead>
<tr>
<th>model</th>
<th>solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$hm = 0.654$</td>
</tr>
<tr>
<td>2</td>
<td>$hm=0.395+0.078*a$</td>
</tr>
<tr>
<td>3</td>
<td>$hm=\sin(0.329*a)$</td>
</tr>
<tr>
<td>4</td>
<td>$hm=\text{step}(a-2.09)$</td>
</tr>
<tr>
<td>5</td>
<td>$hm=\text{gauss}(1.567/a)$</td>
</tr>
<tr>
<td>6</td>
<td>$hm=\text{logistic}(4.264*a-8.811)$</td>
</tr>
<tr>
<td>7</td>
<td>$hm=0.929*\text{logistic}(4.903*a-9.947)$</td>
</tr>
<tr>
<td>8</td>
<td>$hm=0.92*\text{step}(a-1.645)<em>\text{erf}(0.129</em>a^2)$</td>
</tr>
<tr>
<td>9</td>
<td>$hm=0.926*\text{erf}(0.123<em>a^2)</em>\text{gauss}(728.8*\text{gauss}(a^2))$</td>
</tr>
<tr>
<td>10</td>
<td>$hm=\text{gauss}((1.786-\cos(\tan(a))))/a)$</td>
</tr>
<tr>
<td>11</td>
<td>$hm=0.9601*\text{gauss}((0.0952-\cos(\tan(a)))/(\sqrt{a})<em>\text{gauss}(13.49</em>\text{gauss}(a)))$</td>
</tr>
</tbody>
</table>

Table 3.7: Candidate models for POD model of A6003H data set.
Table 3.8: Summary of metrics for candidate models

<table>
<thead>
<tr>
<th>model</th>
<th>$R^2$</th>
<th>r</th>
<th>ME</th>
<th>MSE</th>
<th>MAE</th>
<th>com</th>
<th>obj</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.41E-05</td>
<td>2.95E-10</td>
<td>0.654</td>
<td>0.227</td>
<td>0.454</td>
<td>1</td>
<td>-119.05</td>
</tr>
<tr>
<td>2</td>
<td>0.155</td>
<td>0.394</td>
<td>1.041</td>
<td>0.192</td>
<td>0.406</td>
<td>5</td>
<td>-144.38</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>0.669</td>
<td>0.993</td>
<td>0.136</td>
<td>0.289</td>
<td>6</td>
<td>-200.3</td>
</tr>
<tr>
<td>4</td>
<td>0.487</td>
<td>0.739</td>
<td>1</td>
<td>0.117</td>
<td>0.117</td>
<td>7</td>
<td>-223.86</td>
</tr>
<tr>
<td>5</td>
<td>0.489</td>
<td>0.712</td>
<td>0.964</td>
<td>0.116</td>
<td>0.271</td>
<td>8</td>
<td>-226.51</td>
</tr>
<tr>
<td>6</td>
<td>0.565</td>
<td>0.757</td>
<td>1</td>
<td>0.099</td>
<td>0.164</td>
<td>9</td>
<td>-250.53</td>
</tr>
<tr>
<td>7</td>
<td>0.576</td>
<td>0.759</td>
<td>0.929</td>
<td>0.096</td>
<td>0.192</td>
<td>11</td>
<td>-254.71</td>
</tr>
<tr>
<td>8</td>
<td>0.587</td>
<td>0.766</td>
<td>0.92</td>
<td>0.094</td>
<td>0.188</td>
<td>23</td>
<td>-256.82</td>
</tr>
<tr>
<td>9</td>
<td>0.583</td>
<td>0.764</td>
<td>0.926</td>
<td>0.095</td>
<td>0.189</td>
<td>29</td>
<td>-257.5</td>
</tr>
<tr>
<td>10</td>
<td>0.613</td>
<td>0.785</td>
<td>0.976</td>
<td>0.088</td>
<td>0.196</td>
<td>36</td>
<td>-269.61</td>
</tr>
<tr>
<td>11</td>
<td>0.661</td>
<td>0.813</td>
<td>0.959</td>
<td>0.077</td>
<td>0.157</td>
<td>72</td>
<td>-289.31</td>
</tr>
</tbody>
</table>

3.6 SMALL DATA SETS (A CASE STUDY)

Following the work presented thus far, a case study is proposed that will show that the Bayes factor approach to model selection can be misused. This is particularly true for small data sets. Table 3.9 displays a data set with 23 observations from two different inspection systems. The inspection was an eddy current inspection for the detection of cracks around fastener sites. It may be obvious to the analyst that claiming an $a_{90}$ or $a_{90,95}$ based on only seven hits in the case of system B is absurd. Note that all known statistical routines performing logistic regression will not display any warning for these data sets. The analysis is conducted as described in section 3 of this chapter.
Table 3.9: Small data set for 2 inspection systems.

Figure 3.22 and Figure 3.23 displays the results of a POD analysis according to the Likelihood Ratio method. Visual inspection reveals that the POD doesn’t come close to approaching 90% for any model with an upper asymptote. This is intuitive because there is so little data and it is binary, so inference is difficult.
Figure 3.22: POD results for small data set with inspection system A.

Figure 3.23: POD results for small data set with inspection system B
Table 3.10 displays the results for all the models that were considered. In the left column the model is identified with the model form, followed by the number of parameters, and if the lower asymptote \( \alpha \) or upper asymptote \( \beta \) is included in the model, it is listed. ML is the maximum likelihood and BF the Bayes factor. Surprisingly, performing the analysis according to the Bayes factor approach selects a 2 parameter model for both data sets. Clearly, this is one possible scenario where this method can also lead to misleading results. This result inspired more in depth simulation studies that will be conducted in Chapter 6 to provide recommendations concerning sample size.
<table>
<thead>
<tr>
<th></th>
<th>α</th>
<th>β</th>
<th>ML</th>
<th>a₉₀</th>
<th>a₉₀/₉₅</th>
<th>BF</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Inspection System A</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logit 2</td>
<td>-</td>
<td>-</td>
<td>0.0028738</td>
<td>3.3799</td>
<td>4.1086</td>
<td>1.203598</td>
</tr>
<tr>
<td>Probit 2</td>
<td>-</td>
<td>-</td>
<td>0.0034589</td>
<td>3.285</td>
<td>3.8801</td>
<td>1</td>
</tr>
<tr>
<td>Logit 3 α</td>
<td>0.062975</td>
<td>-</td>
<td>0.00015701</td>
<td>-</td>
<td>-</td>
<td>22.02981</td>
</tr>
<tr>
<td>Probit 3 α</td>
<td>0.065564</td>
<td>-</td>
<td>0.000009891</td>
<td>-</td>
<td>-</td>
<td>349.7017</td>
</tr>
<tr>
<td>Logit 3 β</td>
<td>-</td>
<td>0.77195</td>
<td>0.00045975</td>
<td>-</td>
<td>-</td>
<td>7.523437</td>
</tr>
<tr>
<td>Probit 3 β</td>
<td>-</td>
<td>0.76569</td>
<td>0.000010246</td>
<td>-</td>
<td>-</td>
<td>337.5854</td>
</tr>
<tr>
<td>Logit 4</td>
<td>0.069811</td>
<td>0.75305</td>
<td>0.000014377</td>
<td>-</td>
<td>-</td>
<td>240.5857</td>
</tr>
<tr>
<td>Probit 4</td>
<td>0.064687</td>
<td>0.75928</td>
<td>0.00003591</td>
<td>-</td>
<td>-</td>
<td>96.32136</td>
</tr>
<tr>
<td><strong>Inspection System B</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logit 2</td>
<td>-</td>
<td>-</td>
<td>0.0031594</td>
<td>2.1738</td>
<td>2.885</td>
<td>1</td>
</tr>
<tr>
<td>Probit 2</td>
<td>-</td>
<td>-</td>
<td>0.0030043</td>
<td>2.098</td>
<td>2.7902</td>
<td>1.052993</td>
</tr>
<tr>
<td>Logit 3 α</td>
<td>0.1459</td>
<td>-</td>
<td>0.0010256</td>
<td>-</td>
<td>-</td>
<td>3.080538</td>
</tr>
<tr>
<td>Probit 3 α</td>
<td>0.15467</td>
<td>-</td>
<td>0.0014058</td>
<td>-</td>
<td>-</td>
<td>2.247404</td>
</tr>
<tr>
<td>Logit 3 β</td>
<td>-</td>
<td>0.8979</td>
<td>0.0001799</td>
<td>-</td>
<td>-</td>
<td>17.56198</td>
</tr>
<tr>
<td>Probit 3 β</td>
<td>-</td>
<td>0.8779</td>
<td>0.00021658</td>
<td>-</td>
<td>-</td>
<td>14.58768</td>
</tr>
<tr>
<td>Logit 4</td>
<td>0.1242</td>
<td>0.90188</td>
<td>0.000018893</td>
<td>-</td>
<td>-</td>
<td>167.226</td>
</tr>
<tr>
<td>Probit 4</td>
<td>0.12785</td>
<td>0.90577</td>
<td>0.0001075</td>
<td>-</td>
<td>-</td>
<td>29.38977</td>
</tr>
</tbody>
</table>

Table 3.10: Summary results for 2 inspection systems with limited data.
4 ADVANCED ANALYSIS OF SIGNAL RESPONSE DATA

4.1 BOX COX TRANSFORMATIONS

The two major requirements for $\hat{a}$ vs $a$ analysis are a linear relationship between flaw size and signal response, and homoscedasticity. A logarithmic transformation is commonly applied to data to remedy any violations of fundamental assumptions in linear regression. If the logarithmic transformation fails to provide a set of data suitable for regression analysis, there are currently limited options other than hit/miss analysis. It is also possible that a logarithmic transformation can address a violation of constant variance, but this isn’t always the case.

Given that these conditions of linearity and homoscedasticity are often not met with real NDI data, it is useful to explore remedial measures such as transformations so that the full signal response of NDI data can be used more frequently in practice. In addition, if the linear assumption is not met after the data is transformed, there is an additional question of how to properly put confidence bounds on a POD result that is derived from a more complicated measurement model. A case study problem is presented in this chapter to explore these issues in POD evaluation.
Prior work on detecting subsurface cracks in multi-layer airframe structures used novel methods to extract features that were used as \( \alpha \) values for POD analysis [40-42]. A preliminary model-assisted POD study was conducted based on those efforts [43-45]. In the previous work, hit/miss analysis was used to analyze this data because visual inspection indicated that there was a violation of the constant variance assumption and possibly the linear assumption.

In this work, a Box-Cox transformation will be applied to the data which mitigates, at least in part, concerns about heteroscedasticity. If constant variance can be achieved with this transformation and the linear assumption is met, then \( \alpha \) vs \( \alpha \) analysis can be performed according to the methods set forth in Berens’ classic work on the subject [4]. If constant variance is achieved, but the linear assumption is not met, then methods need to be developed for more complicated models.

It was difficult to determine via visual inspection of the data whether a linear model was most appropriate for the data set, so additional modeling and simulation studies have been conducted to determine the model form of the response that can be expected with this type of inspection. While the model itself is not used in this study, it inspired the use of a second order linear model; thus it is referred to as a “physics-inspired” model rather than a physics-based model. Lastly, it has been found that bootstrapping is a very easy
and useful method for providing confidence bounds on POD curves, and its use will be illustrated with some examples.

The experimental problem of interest is the detection of cracks under installed countersunk fasteners in airframe structures. The description of the data and how it was processed is provided in detail in prior papers [41, 42]. In brief, a conventional eddy current probe was raster scanned over the entire specimen and the vertical component of the voltage was recorded. The scan resolution was 0.25 mm in the x and y direction. The sample set contained over 300 fastener sites with cracks in the 1st layer and 2nd layer at the faying surface. In this work, only the 1st layer cracks are considered, and there are a total of 171 observations. The dimensions for the thickness of the top and bottom layers measured 3.96 mm and 2.54 mm respectively. Conductivities of 1.87 E7 S/m for the aluminum layers and 1.79 E6 S/m for the titanium fasteners were considered. The radius of the fastener hole was 4.04 mm. The probe was operated at 0.6 KHz, well within the specifications of the probe of 0.200 KHz and 20 KHz. The coil dimensions are 6.0 mm for height, 3.0 mm for inner radius and 6.0 mm for outer radius. The number of windings was estimated to be 1000. A corner crack model for the first layer was considered with the assumed aspect ratio length, a, to width, b, of 1:1. Crack lengths in the experimental samples were between 0.7 to 4.3 mm. Since there were 132 unflawed fastener sites, the intrinsic variance due to the inspection system could be estimated.
Model-based image processing methods were used to extract features in the scans that correlate to flaw size [41, 42]. This model-based approach essentially fits models based on first-principles to image data in order to enhance crack indications in the presence of coherent noise from the fastener site, adjacent fasteners and panel edges. The final step is to extract a quantitative metric associated with the crack condition off-axis from each fastener site center. Since the model-based approach to remove the fastener response and extract measures of the crack response is imperfect, the response for no-crack conditions will have some variation around a value of zero. More details on the crack feature extraction procedure used to acquire the data reported here can be found in reference [42].

This same analysis process was applied to all experimental and simulated data to facilitate proper comparison. The raw data is displayed in Figure 4.1. A previous analysis of data from these samples used binary logistic regression because visual inspection of the data revealed that the homoscedasticity assumption was violated. It is clearly observed that the variance increases as a function of flaw size. There is also another current study investigating a similar set of inspection data, with an alternative approach to the statistical analysis [46].
In this analysis, ‘â’ is the magnitude of the eddy current signal response, and ‘a’ refers to crack length. For cases where there is a relationship between the mean response and variance, the Box-Cox transformation is used to stabilize the variance. This method assumes that the relationship between the error variance $\sigma_i^2$ and the mean response $\mu_i$ can be described with a power transformation on â in the form of Eq. (4.1). The subscript ‘i’ refers to particular crack sizes.

Figure 4.1: Raw Data for Eddy Current Inspection.
The new regression model in Eq. (4.2) will include the additional $\lambda$ parameter which will also need to be estimated. The subscript ‘$i$’ refers to the $i^{th}$ observation.

\[ \hat{a}' = \hat{a}^\lambda \]  
\[ \hat{a}_i^\lambda = \beta_0 + \beta_1 a_i + \epsilon_i \]  

Following a method outlined in Kutner et al [47], a numerical search procedure is set up to estimate $\lambda$. The $\hat{a}$ observations are first standardized so that the order of magnitude error sum of squares isn’t dependent on the value of $\lambda$. The observations are standardized according to Eq. (4.3) and Eq. (4.4), where $g_i$ are the standardized observations.

\[ g_i = \frac{1}{\lambda c^{\lambda-1}} (\hat{a}_i^\lambda - 1), \quad \lambda \neq 0 \]  
\[ g_i = c(\ln(\hat{a}_i^\lambda)), \quad \lambda = 0, \]  

where $c = (\prod \hat{a}_i)^{1/n}$, and $n$ is the total number of observations, which is the geometric mean of the observations. Once these standardized observations are obtained, they are then regressed on ‘$a$’, which in this case is crack length, and then the sum of squares error (SSE) is obtained. The optimization problem is formulated such that the objective is to
minimize SSE with λ as a single parameter to be adjusted. Microsoft Excel’s Solver add-in was used to determine \( \lambda_m \), which is the value of \( \lambda \) minimizes SSE.

For this data, \( \lambda_m = 0.45 \) is the transformation that minimizes the SSE. Note that if \( \lambda = 0.5 \), it is simply a square root transformation. This procedure only provides a general estimate of a preferred transformation, so for the sake of using a familiar transformation, further analysis will use the square root transform. Both values of \( \lambda \) will be used to provide an idea of the sensitivity of POD results to the choice of transformation. The transformed data, \( \hat{a} \) vs a analysis, and the POD curve are shown in Figure 4.2, Figure 4.3, and Figure 4.4 respectively for \( \lambda = 0.45 \). The left censor value is selected to be 0.13, the right censor is not used, and the detection threshold is set to 0.23. The following parameter estimates are obtained for the linear regression model: \( \beta_0 = 0.166 \), \( \beta_1 = 0.045 \) and \( \tau = 0.026 \), where \( \tau \) is the regression standard deviation. The \( a_{90} \) value is 2.176 mm and the \( a_{90/95} \) value is 2.327 mm. Recall that since \( \beta_0 \) and \( \beta_1 \) are unknown, the estimates of those parameters are used for confidence bound calculation according to the Wald method. The confidence bounds for the regression model are calculated using Eq. (4.5).

\[
\hat{a} = \hat{\beta}_0 + \hat{\beta}_1 a \pm 1.645\sqrt{\text{var}(\hat{a})} \quad (4.5)
\]
Figure 4.2: Transformed \( \hat{a} \) Data with \( \lambda = 0.45 \)
Figure 4.3: Linear Regression on Transformed Data with $\lambda = 0.45$
The next step is to transform the measurement model into a POD model. The method is commonly called the “delta method” in the statistics literature. Section G.5.4.7 in [20] provides complete details on how this calculation is performed. Also note that

Figure 4.4: POD Curve for Transformed Data with $\lambda = 0.45$
there is an additional set of outer bounds in Figure 4.3 that represent prediction bounds. This can be interpreted as the bounds for future observations for individual cracks of a given size.

The same analysis is conducted for the $\lambda = 0.5$ or square root transformation. The detection threshold for this transformation is 0.195, and the left censor value is 0.14. The transformed data, $\hat{a}$ vs $a$ analysis, and the POD curve is shown in, Figure 4.6, Figure 4.7, and Figure 4.8 respectively. The following parameter estimates are obtained for the linear regression model: $\beta_0 = 0.135$, $\beta_1 = 0.043$ and $\tau = 0.024$. The $a_{90}$ value is 2.102 mm and the $a_{90/95}$ value is 2.257 mm.

4.2 ANALYSIS WITH PHYSICS-INSPIRED MODEL

There is some precedent for using a physical model of an inspection to improve the evaluation of POD in ultrasonic inspections [48]. In the work of Thompson and Meeker, a “kinked” regression model was developed to describe the impact of hard-alpha inclusions on POD. In particular, the physics model provided a better understanding of the small flaw regime. If the flaw is significantly smaller than the ultrasonic wavelength, it is in the Rayleigh scattering regime which has a cube relationship with the flaw dimensions. Thus, two different linear models were used depending on the flaw size. Figure 4.5 shows the signal response based on simulations in VIC-3D® [40]. The same image processing methods were applied to the simulated data. This type of response is
not quite linear, so the next analysis will be with a 2nd order regression model shown in Eq. (4.6). The \( \varepsilon \) symbol is a normally distributed error term.

Figure 4.5: Model with Experimental Data.

Figure 4.6: Data with Square-Root Transform Applied
Figure 4.7: Linear Regression with Square-Root Transformed Data
Figure 4.8: POD Analysis for Square Root Transformed Data
One additional test was used to investigate whether the $a^2$ term provided any advantage. Uncensored regression was performed on the data so that the adjusted R-square values could be calculated and used to determine if the model fit improved with the addition of an extra term. It is not appropriate to calculate an R-squared value using censored data. The statistical significance or p-value of $a^2$ in the standard regression model is 0.001, and the adjusted R-square value for the model including $a^2$ is 0.7754 which is slightly above 0.7619 which is for the model that includes only ‘a’, so this is

$$\hat{a} = \beta_0 + \beta_1 a + \beta_2 a^2 + \epsilon$$  \hfill (4.6)
more evidence that there is good reason to include it in the model. Given the square root transform or $\lambda = 0.5$, the estimates for $\beta_0$, $\beta_1$, and $\beta_2$ are 0.137, 0.027, and 0.005 respectively. $\tau$ is 0.0229. The censored regression has the same left censor and threshold as the first order model. The $a_{90}$ value for this second order model is 2.277 mm. Figure 4.9 shows the fit of the quadratic model. There are no published procedures to find the $a_{90,95}$ value for this type of model. The next section introduces a very useful bootstrapping method to address this issue.

### 4.3 Bootstrap Methods for Confidence Bound Calculation

The algorithm to generate confidence bounds on models with additional complexity is quite simple. The main idea is to use “sampling with replacement”, which interestingly wasn’t used much in the statistics community until relatively recently [49], and has been used with good success in engineering applications [50, 51]. To illustrate how bootstrap confidence bounds are calculated, and to verify against standard methods, it is necessary to go back to a previous $\hat{a}$ vs $a$ analysis where the confidence bound calculation method is well established. To verify, the case of the transformation parameter $\lambda = 0.5$ with the threshold set to 0.195 and the left censor equal to 0.14 is used. This time a new data set is generated by the sampling with replacement of the original data. This new set is used to calculate $a_{90}$, and this process is repeated 1,000 times. The $a_{90}$ results are then sorted in ascending order. For the case of 1,000 samples, the $950^{th}$ $a_{90}$ value is considered the value for $a_{90,95}$. Table 4.1 summarizes the results of this process.
No significant difference in $a_{90}$ exists, and although there is a slight difference in $a_{90/95}$, the bootstrap results are on the conservative side. Based on these results, it doesn’t seem necessary to sample more than 1,000 times. This bootstrap approach was applied to the 2nd order model. Figure 4.9 shows the fitted 2nd order model with the transformed $\lambda = 0.5$ data. The $a_{90/95}$ value using the bootstrap method with 1,000 samples is 2.472 mm.

One of the advantages of adding $a^2$ to the model is that there is less dependence on subjective decisions regarding censoring values and threshold values. The small flaw region is better represented with this model. Future work will involve a sensitivity study of the left censor value and threshold and the impact they have on the $a_{90}$ and $a_{90/95}$ results.

<table>
<thead>
<tr>
<th>Method</th>
<th>$a_{90}$</th>
<th>$a_{90/95}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald Method</td>
<td>2.102 mm</td>
<td>2.257 mm</td>
</tr>
<tr>
<td>Bootstrap 1,000</td>
<td>2.096 mm</td>
<td>2.281 mm</td>
</tr>
<tr>
<td>Bootstrap 10,000</td>
<td>2.099 mm</td>
<td>2.299 mm</td>
</tr>
<tr>
<td>Bootstrap 100,000</td>
<td>2.099 mm</td>
<td>2.297 mm</td>
</tr>
</tbody>
</table>

Table 4.1: Bootstrap results
4.4 COMPARISON WITH HIT/MISS ANALYSIS

Since, the data have been examined with â vs a analysis and also with a 2nd order linear model, it is interesting to compare it with hit/miss Bernoulli analysis since that is
still overwhelmingly used to this day. The analysis will be conducted in two different ways. At most one false call was recorded in the previous analysis, so in the analysis shown in Figure 4.10, the number of false calls is forced to one by setting the detection threshold to 0.187. At this threshold, $a_{90} = 1.72$ mm and $a_{90/95} = 2.04$ mm which are considerably smaller than the corresponding POD parameters for the other types of analysis. Secondly, the threshold is lowered substantially to 0.167 so that two additional flaws are detected, and this results in eleven false calls. Even smaller POD parameters are determined with $a_{90} = 1.498$ mm and $a_{90/95} = 1.907$ mm. Note that this was performed with the transformed data with $\lambda = 0.5$, but it was also performed with the original data, and the exact same POD parameters were obtained corresponding with the number of false calls at one and eleven.

<table>
<thead>
<tr>
<th>analysis method</th>
<th>$\lambda$</th>
<th>left censor</th>
<th>detection threshold</th>
<th>False calls</th>
<th>$a_{90}$ (mm)</th>
<th>$a_{90/95}$ (mm)</th>
<th>$a_{90} - a_{90/95}$ (% difference)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st order linear</td>
<td>0.45</td>
<td>0.13</td>
<td>0.23</td>
<td>0</td>
<td>2.176</td>
<td>2.327</td>
<td>6.9%</td>
</tr>
<tr>
<td>1st order linear</td>
<td>0.5</td>
<td>0.14</td>
<td>0.195</td>
<td>1</td>
<td>2.102</td>
<td>2.257</td>
<td>7.3%</td>
</tr>
<tr>
<td>1st order linear</td>
<td>0.5</td>
<td>0.195</td>
<td>0.195</td>
<td>1</td>
<td>2.269</td>
<td>2.53</td>
<td>11.5%</td>
</tr>
<tr>
<td>2nd order linear</td>
<td>0.5</td>
<td>0.14</td>
<td>0.195</td>
<td>1</td>
<td>2.277</td>
<td>2.472</td>
<td>8.5%</td>
</tr>
<tr>
<td>2nd order linear</td>
<td>0.5</td>
<td>0.195</td>
<td>0.195</td>
<td>1</td>
<td>2.197</td>
<td>2.428</td>
<td>10.5%</td>
</tr>
<tr>
<td>hit/miss</td>
<td>1</td>
<td>0.187</td>
<td>1</td>
<td>1.72</td>
<td>2.04</td>
<td></td>
<td>18.6%</td>
</tr>
<tr>
<td>hit/miss</td>
<td>1</td>
<td>0.162</td>
<td>11</td>
<td>1.498</td>
<td>1.907</td>
<td></td>
<td>27.3%</td>
</tr>
</tbody>
</table>

Table 4.2: Competing Model Comparison Study Results
4.5 CHAPTER SUMMARY

Multiple statistical analysis methods were used to examine data from an eddy current inspection of fastener sites in multi-layer structures. There were notable differences in $a_{90}$ and $a_{90/95}$ estimates for the different models. The hit/miss model contains the least information, but produces the most attractive POD. The lower the $a_{90/95}$ value, the longer the inspection intervals, so a lower $a_{90/95}$ value is desirable. There must be confidence that the $a_{90/95}$ is determined accurately with as little uncertainty as possible. No hard conclusions can be made about this single case study, but it does show that in at least one real case, the hit/miss results may be optimistic when compared to analysis that contains more information.

It is also interesting to note that the physics-inspired model produced similar results for the POD parameters of interest regardless of the chosen value of the left censor. Further investigations will systematically study the effect of censoring on linear and higher order models. Preliminary evidence suggests that the $a_{90/95}$ value may be invariant to the choice of the left censor value if the small flaw region is modeled adequately.

As more sophisticated models begin to be used in analysis of inspection data, bootstrapping is an easy and accurate way to produce confidence bounds on POD results.
This was demonstrated for the usual $\hat{a}$ vs $a$ analysis which provided confidence in the bootstrap approach. It was practical to use this method for putting confidence bounds on the second order model.
5 APPLICATIONS

5.1 EXPERIMENTAL CROSS VALIDATION

The vast majority of inspection data sets are still in the form of hit/miss. This next investigation will investigate how the new statistical methods proposed in chapter 3 deviate from prior methods for real data sets from [34]. A total of 45 data sets were analyzed including 15 eddy current, 15 ultrasonic, and 15 x-ray inspections. Table 5.1 displays the index of data sets from the NTIAC collection and the corresponding filename used in this work. The NTIAC file names that beginning with a letter ‘A’ are associated with eddy current, ‘D’ are associated with ultrasonics, and ‘F’ are associated with X-ray.

The purpose of analyzing these 45 data sets is to understand how often the statistical approach in Chapter 3 determines that a 3 or 4 parameter model provides a better model fit than the conventional 2-parameter model. The files were randomly selected from the data base. Only the logit link for 2, 3, and 4 parameter models were evaluated. The criterion for the deciding whether or not enough evidence exist for model selection purposes is a Bayes factor of 3. If the Bayes factor is above 3, then there is enough evidence to suggest a better model fit for the purpose of this section.
The details of the study will be illustrated in detail for three data sets. The first is identified as A9001(3)D from the NTIAC database, and it is considered to be a data set that violates fundamental assumptions necessary for a proper POD evaluation. The second is identified as A1002BL from the NTIAC database, and it is considered to be a well behaved data set. The third is identified as A1002CL and it is also considered to be a well behaved data set.

As shown in Figure 5.1, there is a major overlap between the find and no find data. There is no clear separation of what can and cannot be detected. Forcing a 2-parameter model on this type of data set will provide an estimate of \(a_{90}\) and a value for \(a_{90/95}\), but they’re fictitious. The Bayes factor for the 3-parameter logit model with an upper asymptote is 1.76E+06 which strongly supports the presence of an upper asymptote. The 3-parameter logit model for the A9001(3)D data set is displayed in Figure 5.2. The value of the upper asymptote \(\beta\) is 0.681 and the lower 95% confidence limit on \(\beta\) is 0.638.

The A1002BL data set shown in Figure 5.3 is a well behaved data set where there is a clear separation between what crack size can be detected and what cannot. Even in such a case as this one, the Bayes factor in support of the 3 parameter model is 5.13. This could be because of the 2 misses around 7 and 8 mm. The \(a_{90}\) value estimated by the 2 parameter model and the 3 parameter model is 1.888 mm and 1.645 mm respectively.
The $a_{90/95}$ value for the 2 parameter model and the 3 parameter model is 2.239 mm and 1.985 mm respectively. The upper asymptote $\beta$ is estimated to be 0.974 and the lower 95% confidence bound on $\beta$ is 0.922. This is a case where using the 2-parameter model is not considered by the author as a fatal error; nonetheless, it is wise to use the model with the most evidence supporting it.

The A1002CL data set is again clearly well behaved as shown in Figure 5.5. There are in fact no large misses, and the separation between detect and no detect is easily observed. In this case the Bayes factor in support of the 2-parameter model is 4.061. It is also interesting to note that the $a_{90}$ and $a_{90/95}$ values evaluated by the 2 and 3-parameter models differ only by 100ths of a millimeter. This is about as ideal of POD data set as one will find. The value of the upper asymptote $\beta$ is 0.975 and the 95% lower confidence bound on $\beta$ is 0.974. This is about as high of a value $\beta$ value that can be found in any of the 45 data sets.
Figure 5.1: Analysis Attempt with A9001(3)D data set

Figure 5.2: 3-Parameter Model Fit for A9001(3)D Data Set
Figure 5.3: 2-parameter Model Fit for A1002BL Data Set
Figure 5.4: 3-parameter Model Fit for A1002BL Data Set
Figure 5.5: 2-parameter Model Fit for A1002CL Data Set

Based on analysis of all the data sets, the 2 parameter model fit was considered superior to the 3 parameter model fit in only 42.222% of these data sets. The 2 parameter model fit was considered superior to the 4 parameter model fit in 33.333% of these data sets. The 2 parameter model fit was considered superior to the 3 or 4 parameter models in 48.888% of the data sets. This implies that the conventional 2 parameter model was not the best model fit for roughly half of the data sets analyzed.
<table>
<thead>
<tr>
<th>dissertation file</th>
<th>NTIAC</th>
<th>dissertation file</th>
<th>NTIAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A9001(3)D</td>
<td>24</td>
<td>D1003CD</td>
</tr>
<tr>
<td>2</td>
<td>A9001(3)L</td>
<td>25</td>
<td>D1001BL</td>
</tr>
<tr>
<td>3</td>
<td>A9002(3)D</td>
<td>26</td>
<td>D1001CL</td>
</tr>
<tr>
<td>4</td>
<td>A9002(3)L</td>
<td>27</td>
<td>D1002AL</td>
</tr>
<tr>
<td>5</td>
<td>A9003(3)D</td>
<td>28</td>
<td>D1002BL</td>
</tr>
<tr>
<td>6</td>
<td>A9003(3)L</td>
<td>29</td>
<td>D1002CL</td>
</tr>
<tr>
<td>7</td>
<td>A1001AL</td>
<td>30</td>
<td>D1003AL</td>
</tr>
<tr>
<td>8</td>
<td>A1001BL</td>
<td>31</td>
<td>F5001(3)D</td>
</tr>
<tr>
<td>9</td>
<td>A1001CL</td>
<td>32</td>
<td>F5001(3)L</td>
</tr>
<tr>
<td>10</td>
<td>A1002AL</td>
<td>33</td>
<td>F30651AD</td>
</tr>
<tr>
<td>11</td>
<td>A1002BL</td>
<td>34</td>
<td>F30651BD</td>
</tr>
<tr>
<td>12</td>
<td>A1002CL</td>
<td>35</td>
<td>F30653AD</td>
</tr>
<tr>
<td>13</td>
<td>A1003AL</td>
<td>36</td>
<td>F8001(3)D</td>
</tr>
<tr>
<td>14</td>
<td>A1003BL</td>
<td>37</td>
<td>F8001(3)L</td>
</tr>
<tr>
<td>15</td>
<td>A1003CL</td>
<td>38</td>
<td>F8001(3)D</td>
</tr>
<tr>
<td>16</td>
<td>D1001AD</td>
<td>39</td>
<td>F8002(3)L</td>
</tr>
<tr>
<td>17</td>
<td>D1001BD</td>
<td>40</td>
<td>F8003(3)D</td>
</tr>
<tr>
<td>18</td>
<td>D1001CD</td>
<td>41</td>
<td>F8003(3)L</td>
</tr>
<tr>
<td>19</td>
<td>D1002AD</td>
<td>42</td>
<td>F20852AD</td>
</tr>
<tr>
<td>20</td>
<td>D1002BD</td>
<td>43</td>
<td>F20852BD</td>
</tr>
<tr>
<td>21</td>
<td>D1002CD</td>
<td>44</td>
<td>F20852CD</td>
</tr>
<tr>
<td>22</td>
<td>D1003AD</td>
<td>45</td>
<td>F22202AD</td>
</tr>
<tr>
<td>23</td>
<td>D1003BD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.1: NTIAC data set filename index

<table>
<thead>
<tr>
<th></th>
<th>2 par / 3 par</th>
<th>2 par / 4 par</th>
</tr>
</thead>
<tbody>
<tr>
<td>% Bayes factor &gt; 3</td>
<td>42.222%</td>
<td>33.333%</td>
</tr>
</tbody>
</table>

Table 5.2: Percentage of time 2 parameter model is best fit


<table>
<thead>
<tr>
<th></th>
<th>3 par</th>
<th>4 par</th>
</tr>
</thead>
<tbody>
<tr>
<td>% a_{90} exists</td>
<td>55.56%</td>
<td>44.44%</td>
</tr>
<tr>
<td>% a_{90/95} exists</td>
<td>37.78%</td>
<td>31.11%</td>
</tr>
</tbody>
</table>

Table 5.3: Percentage of time $a_{90}$ and $a_{90/95}$ exists

5.2 **RISK ANALYSIS**

Chapters 2 through 4 include significant contributions to estimating POD, which is one of the critical inputs for risk analysis. In particular, the methods developed in this work provide a much better estimate of the behavior of the POD curve for large flaw sizes and also for small flaw sizes. It is believed that higher probability of missing large flaws implies an increased probability of failure towards the end of life for an asset. False calls may increase the cost of maintaining an asset.

This chapter will provide an outline of how structural risk analyses are conducted for aircraft components managed with the damage tolerance approach. The motivation for such studies is to provide the manager responsible for safe operation of an aircraft fleet with quantitative information to make decisions such as when to inspect, replace, repair, and retire a component or system. The particular problem in mind is cracks around fastener sites in aging aircraft structures. In chapters 2 through 4, significant improvements were made for quantifying inspection capability. Even with good POD
studies, it is quite another challenge to estimate crack size distributions in structures from field inspections alone. Other sources of information beyond inspection data are needed to infer anything about the crack size distribution. A couple of decades ago, a tool called Probability of Fracture (PROF) was developed by the University of Dayton Research Institute to assist decision makers. The risk calculations in this chapter follow the methods developed for PROF [52].

Safety is managed by the probability of fracture (POF) defined in Eq. (5.1), where \( \sigma \) is stress intensity factor, \( K_c \) is fracture toughness of the material, \( \beta \) is a geometric factor and usually a function of crack depth ‘a’. The POF occurs when a stress produces a stress intensity factor that exceeds the fracture toughness for a part.

\[
P[\sigma_{max} \geq \sigma_{cr}] = P \left[ \frac{K_c}{\sqrt{\pi a \cdot \beta(a)}} \right] \tag{5.1}
\]

Fundamentally Eq. (5.1) is the calculation needed to determine POF, but there are many inputs in realistic scenarios that need to be incorporated. PROF requires 9 inputs:

1) Peak Stress/Flight - It’s assumed that the largest stress encountered in flight will cause the fracture if there is fracture. This is modeled in terms of a Gumbel distribution.

2) POD function – This was the topic for chapters 2 through 4 of this dissertation, but in PROF, only the parameters \( \mu \) and \( \sigma \) are used for risk calculations.
3) Fracture Toughness – This information can be found in damage tolerance handbooks for many geometries.

4) Aircraft Parameters – past and expected usage of the fleet of aircraft.

5) a vs K/σ – Relationship between stress intensity factor, stress, and crack size.

6) a vs time – crack growth is a stochastic phenomena, but PROF uses a deterministic correlation between flight hours and crack size. This usually fits an exponential function.

7) Initial crack sizes – Known as the equivalent initial crack size (EIFS). Any valid cumulative distribution can be an input. Weibull is a common option.

8) Repair crack sizes – Assumes repairs leave a crack and this is considered a uniform distribution to be conservative.

9) Usage intervals – This the period of time between inspections.

The inputs to the risk analysis mentioned in chapter 1 and in the paragraph above are probability distributions and include crack size distributions, maximum stress per flight, fracture toughness, and POD. One of the inputs is the equivalent initial flaw size (EIFS), which is obtained by looking at fractography data and using crack growth curves to determine what the state of damage was when the component was manufactured. There is significant uncertainty associated with determining EIFS, and it may be useful to determine how sensitive POF is to EIFS compared to how sensitive POF is to the upper asymptote of a POD curve.
The following example shows the probability of failure calculation for 3 different probabilities of missing a large flaw (1-β), and 3 different values for the largest possible crack size in the EIFS. Figure 5.6 shows an example of a probability of failure calculation with the probability of missing a large flaw set at α = 0.05, and the largest allowed crack size in the EIFS distribution set to 0.01 mm.

Table 5.4 summarizes the probability of failure for three different sizes of largest flaws in the EIFS, and also three different values of 1-β. This is a preliminary examination of the impact of probability of missing large flaws on risk analysis compared to uncertainty in the EIFS. The results indicate that the sensitivity of risk to the upper asymptote β of the POD curve certainly depends on the nature of the crack size distributions. It also depends on factors not investigated in this dissertation such as the spectrum of the loading.
Figure 5.6: Probability of failure for example risk analysis scenario

<table>
<thead>
<tr>
<th>1-β \size upper bound (mm)</th>
<th>0.005</th>
<th>0.01</th>
<th>0.02</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2.5339E-6</td>
<td>2.4494E-6</td>
<td>2.5504E-6</td>
</tr>
<tr>
<td>0.05</td>
<td>1.5175E-6</td>
<td>3.0257E-4</td>
<td>6.1244E-4</td>
</tr>
<tr>
<td>0.10</td>
<td>2.5504E-6</td>
<td>1.1550E-3</td>
<td>2.6763E-3</td>
</tr>
</tbody>
</table>

Table 5.4: Probability of failure for varying probabilities of missing large flaws and upper bound for initial crack size.
6 SIMULATION STUDIES

6.1 OVERVIEW OF SIMULATION STUDIES

Chapter 3 and 4 introduced new methods for calculating POD curves and the confidence bounds associated with them. Chapter 5 summarized the process of evaluating risk and how improvements to POD calculations may impact risk assessment. This part of the effort will investigate the coverage of existing and improved POD methods. The coverage probability of a confidence interval is defined as the proportion of the time that the interval contains the true value of interest [53]. This chapter will include two simulation studies. The first study will minimize complexity and will be based on a basic linear regression model with additive noise and à vs a analysis and hit/miss analysis will be performed according to traditional and new methods to demonstrate how such a simulation study should be conducted. The second study will include both additive and multiplicative noise. This type of noise is similar to noise in real experiments. Lastly, tolerance intervals will be discussed in light of the results from the simulation studies.

6.2 SIMULATION STUDY 1

The first simulation study is designed to have minimum complexity in order to illustrate how such a study should be conducted. The linear model parameters of the
transformed data from chapter 4 section 2 will used as the basis for this study. In particular the model fit shown in Figure 4.7 is used. The y-intercept $\beta_0$ is 0.135, the slope $\beta_1$ is 0.043, and the additive noise $\tau = 0.024$. Figure 6.1 shows a data set with 5000 observations generated according to the parameter values. A linear regression model with the detection threshold set at 0.195 and the left censor set at 0.14 is fit to provide a “true” value of $a_{90}$. In this sense, there is an approximate population value for $a_{90}$ to compare $a_{90}$ and $a_{90/95}$ values evaluated for smaller samples from this “population”. The results for the model parameters are slightly different than the input parameters and are shown in Figure 6.2. Figure 6.3 displays the value of “true” $a_{90}$ which is 2.086.
Figure 6.1: Synthetic data set created based on linear model fit for chapter 4 section 2.
Figure 6.2: Linear regression model for “population” generated for simulation study.

Figure 6.3: POD evaluation for “population” generated for simulation study.
In this study, a comparison of “â vs a” and hit/miss analysis is of interest. The â data is transformed to hit/miss simply by considering anything over the detection threshold 0.195 a hit and anything under the detection threshold a miss. The transformed data of the “population” along with the hit/miss analysis is shown in Figure 6.4.

The next step is to sample from this set of 5,000 observations 29 at a time for a total of 30 samples. The goal of this is to compare the coverage of the different analysis methods discussed in chapters 2 and 3. To illustrate what the data sets consisted of, only the first two samples are shown. Figure 6.5 - Figure 6.10 show the linear model fit, POD evaluation, and hit/miss analysis for the first and second data sets out of the 30.

![Figure 6.4: Hit/miss data of “population” with POD analysis.](image)
Figure 6.5: Linear model fit for first sample out of total of 30.
Figure 6.6: POD analysis for first sample for total of 30.

Figure 6.7: Hit/miss analysis for first data set out of 30.
Figure 6.8: Linear model fit for second data set out of 30.

Figure 6.9: POD evaluation for second data set out of 30.
The 30 data sets were analyzed using “â vs a” analysis, likelihood ratio method for hit/miss and the MCMC method for hit/miss. For each case, histograms for the a_{90} and a_{90/95} values are constructed. Recall that the “true” a_{90} values for this simulation study is 2.086, therefore two items are of interest: 1) The percentage of values of a_{90/95} that are less than 2.086 and 2) the scatter of the a_{90} values.

Figure 6.10: Hit/miss analysis for second data set out of 30.

Figure 6.11 and Figure 6.12 display the histograms for the a_{90} and a_{90/95} values associated with the â vs a analysis respectively. Roughly half (43.333%) of the a_{90} values are less the true mean which is expected. Only 10% of the a_{90/95} values are less than the true mean.
Figure 6.11: Histogram of $a_{90}$ values for $\hat{a}$ vs $a$ analysis.
Figure 6.12: Histogram of $a_{90/95}$ values for \( \hat{a} \) vs analysis.

Figure 6.13 and Figure 6.14 shows the histograms for the $a_{90}$ and $a_{90/95}$ values computed using the Likelihood Ratio method for hit/miss analysis. The percentage of $a_{90}$ values less than the true mean is 60%, and none of the $a_{90/95}$ values are less than the true mean.
Figure 6.13: Histogram of $a_{90}$ values for hit/miss analysis with likelihood ratio method.
Figure 6.14: Histogram of $a_{90/95}$ values for hit/miss analysis with likelihood ratio method.

Figure 6.15 and Figure 6.16 shows the histograms for the $a_{90}$ and $a_{90/95}$ values computed using the MCMC method for hit/miss analysis. The percentage of $a_{90}$ values less than the true mean is $66.667\%$, and $16.667\%$ of the $a_{90/95}$ values are less than the true mean.
Figure 6.15: Histogram of $a_{90}$ values for hit/miss analysis with MCMC method.
Before summarizing the results, one additional study was conducted with only “â vs a” analysis. This study was conducted with only 10 observations. Hit/miss analysis is not possible with only 10 observations, so there is an intrinsic advantage of using “â vs a” analysis. Here, 46.666% of the a_{90} values are below the true mean and 16.666% of the a_{90/95} values are below the true mean. This is comparable to having 30 observations of hit/miss data. Note that the samples were uniformly distributed and the â data had constant variance as a function of flaw size, so this may represent a best case scenario.
Figure 6.17 shows the linear regression fit for the $\hat{a}$ data with 10 observations, and Figure 6.18 displays the POD analysis associated with it. Figure 6.19 and Figure 6.20 shows the histograms of the $a_{90}$ and $a_{90/95}$ values associated with this study with 10 observations.

Figure 6.17: Linear model fit for data set with 10 observations.
Figure 6.18: POD analysis for data set with 10 observations.
Figure 6.19: Histogram of $a_{90}$ values for data sets with 10 observations.
To summarize the results of this simulation study, the coverage is presented in box plots shown in Figure 6.21 and Figure 6.22 and the numerical results are shown in Table 6.1. The left column contains the quantity that is computed along with the type of analysis. Signal response analysis is simply referred to as ‘â’ in the first two rows and also the last two rows. The likelihood ratio method for hit/miss is identified by hmLR. Markovchain Monte Carlo uses the common MCMC acronym. As expected, the box plot associated with the “â vs a” analysis with 30 observations has the least variance. The
variance for hit/miss analysis using MCMC is about the same as that of the “â vs a” analysis with only 10 observations. It is encouraging to see that the mean a₉₀ of all the methods is very close to the true a₉₀ value. The coverage for the a₉₀/₉₅ values is as follows: The “â vs a” analysis with 30 observations provided 90% coverage. The likelihood ratio method for hit miss analysis provided 100% coverage. The MCMC method with the 2 parameter logit model and “â vs a” analysis with 10 observations both provided 83.33% coverage.

Figure 6.21: Box plot chart for a₉₀ values from simulation study.
Table 6.1: Summary of simulation study 1 results.

<table>
<thead>
<tr>
<th>Method</th>
<th>μ</th>
<th>σ²</th>
<th>median</th>
<th>% error</th>
<th>% coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_{90}$ (â)</td>
<td>2.097</td>
<td>0.040</td>
<td>2.104</td>
<td>0.529%</td>
<td>56.667%</td>
</tr>
<tr>
<td>$a_{90}/95$ (â)</td>
<td>2.351</td>
<td>0.046</td>
<td>2.346</td>
<td>11.287%</td>
<td>90.000%</td>
</tr>
<tr>
<td>$a_{90}$ (hmLR)</td>
<td>1.998</td>
<td>0.109</td>
<td>2.009</td>
<td>4.380%</td>
<td>40.000%</td>
</tr>
<tr>
<td>$a_{90}/95$ (hmLR)</td>
<td>3.011</td>
<td>0.276</td>
<td>2.916</td>
<td>30.731%</td>
<td>100.00%</td>
</tr>
<tr>
<td>$a_{90}$ (MCMC 29 points)</td>
<td>2.017</td>
<td>0.168</td>
<td>2.031</td>
<td>3.397%</td>
<td>33.334%</td>
</tr>
<tr>
<td>$a_{90}/95$ (MCMC 29 points)</td>
<td>2.886</td>
<td>1.029</td>
<td>2.586</td>
<td>27.730%</td>
<td>83.334%</td>
</tr>
<tr>
<td>$a_{90}$ (MCMC 30 points)</td>
<td>2.017</td>
<td>0.168</td>
<td>2.034</td>
<td>3.428%</td>
<td>33.334%</td>
</tr>
<tr>
<td>$a_{90}/95$ (MCMC 30 points)</td>
<td>2.879</td>
<td>0.993</td>
<td>2.584</td>
<td>27.533%</td>
<td>83.334%</td>
</tr>
<tr>
<td>$a_{90}$ (â) (10 obs)</td>
<td>2.060</td>
<td>0.092</td>
<td>2.100</td>
<td>1.277%</td>
<td>53.334%</td>
</tr>
<tr>
<td>$a_{90}/95$ (â) (10 obs)</td>
<td>2.444</td>
<td>0.152</td>
<td>2.466</td>
<td>14.643%</td>
<td>83.334%</td>
</tr>
</tbody>
</table>
For this simulation, everything was kept as simple as possible, so the results obtained for this study should be viewed as a best case scenario and caution should be taken when extending the lessons learned from this study to more complicated measurement and noise models. The study did encourage the use of more data sets than 30.

### 6.3 SIMULATION STUDY 2

The second simulation considers a more complicated noise model that includes both additive and multiplicative noise as represented in Eq. (6.1). This additive and multiplicative noise model resembles realistic inspection data, and is used for the purpose of generating a synthetic data set useful for simulations purposes only. The additive noise component is designated by $\varepsilon_{\text{add}}$, and the multiplicative is designated by $\varepsilon_{\text{mult}}$.

\[
\hat{a} = \beta_0 + \beta_1 a (1 + \varepsilon_{\text{mult}}) + \varepsilon_{\text{add}} \tag{6.1}
\]

This model along with the other parameters in simulation study 1 was used to create another data set to resemble a population as shown in Figure 6.23. In this case, $\beta_0 = 0.13546$, $\beta_1 = 0.043$, $\sigma^2 = 0.000576$, $\varepsilon_{\text{mult}} = 0.316$, and $\varepsilon_{\text{add}} = 0.0316$. 
In this study 100,000 observations were simulated, then the proportion of observations above the detection of threshold of 0.195 were determined in intervals of 1,000 observations. If 100,000 simulations approaches what can be considered a population for this inspection, then the interval of observations with 90% proportion above 0.195 can be considered the true \( a_{90} \) value. It was determined based on this method.
that $a_{90}$ is 2.907 mm. The coverage of this value with different models will be investigated.

![Graph showing proportion of observations above detection threshold.](image)

Figure 6.24: Proportion of observations above detection threshold.

### 6.3.1 100 Observations

Based on experience, it is predicted that 100 observations of hit/miss data randomly sampled should provide appropriate coverage for POD. The first study will generate samples with 100 observations, and determine coverage. As done in simulation study 1,
the histograms for the \( a_{90} \) and \( a_{90/0.95} \) values are displayed in Figure 6.25, Figure 6.26, Figure 6.29, Figure 6.30, Figure 6.33, and Figure 6.34 for the models where these values exist. Also, the histograms for the lower and upper asymptotes for the three and four parameter models are shown in Figure 6.27, Figure 6.28, Figure 6.31, and Figure 6.32.

Figure 6.25: Histogram of \( a_{90} \) values for 2 parameter logit model.
Figure 6.26: Histogram of $a_{90/95}$ values for 2 parameter logit model.
Figure 6.27: Histogram of upper bound on POD for 3 parameter logit model.
Figure 6.28: Histogram of lower 95% confidence bounds.
Figure 6.29: Histogram of $a_{90}$ values for 3 parameter logit model.
Figure 6.30: Display of the 11 $a_{90/95}$ values determined with 3 parameter model.
Figure 6.31: Histogram of upper bound of $a_{90}$ values.
Figure 6.32: Histogram of lower 95% confidence bound for 4 parameter model.
Figure 6.33: Histogram of $a_{90}$ values for 4 parameter model.
Figure 6.34: Histogram of $a_{90/95}$ values for 4 parameter model.
Another analysis that can be done to investigate the amount of evidence favoring one model to another is to look at the distribution of Bayes factors for competing models. Figure 6.35 shows a plot of the histogram of Bayes factors for comparing the 2 parameter model with the 3 parameter model. It is interesting to note that 75% of all the data sets favored the 2 parameter model, and most of the favorable Bayes factors were between 1 and 30, so while there is positive evidence favoring the 2 parameter model, the evidence is not overwhelming strong. The situation is similar with when comparing the 2 parameter model to the 4 parameter model as is done in Figure 6.36. Once again, 66% of the data sets favor the 2 parameter model, but the evidence is not overwhelmingly strong.

Table 6.2: Results summary for simulation study 2 for 100 observations.

<table>
<thead>
<tr>
<th></th>
<th>2 parameter logit model</th>
<th>3 parameter logit model</th>
<th>4 parameter logit model</th>
</tr>
</thead>
<tbody>
<tr>
<td>% coverage true (\text{a}_{90})</td>
<td>94%</td>
<td>100%</td>
<td>96%</td>
</tr>
<tr>
<td>average (\text{a}_{90})</td>
<td>3.084</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>average (\text{a}_{90,95})</td>
<td>4.330</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>average upper bound on POD</td>
<td>-</td>
<td>93.264%</td>
<td>90.669%</td>
</tr>
<tr>
<td>average upper bound on lower 95% confidence</td>
<td>-</td>
<td>84.588%</td>
<td>82.599%</td>
</tr>
<tr>
<td>% Bayes factor &gt; 3</td>
<td>-</td>
<td>76%</td>
<td>66%</td>
</tr>
<tr>
<td>Average Bayes Factor</td>
<td>-</td>
<td>85.75</td>
<td>41.06</td>
</tr>
</tbody>
</table>
Figure 6.35: Bayes factor for (model 2 / model 3)
6.3.2 30 observations

The study conducted in 6.3.1 is repeated in this section with only 30 observations.

Once again the histograms for the $a_{90}$, $a_{90/95}$, and upper bound values are provided in Figure 6.37 - Figure 6.42. As expected and detailed in the summary table, greater variability in the results of these evaluations is evident for the smaller sample size of 30 observations.
Figure 6.37: Histogram of $a_{90}$ values for 2 parameter logit model.
Figure 6.38: Histogram of $a_{90/95}$ values for 2 parameter logit model using MCMC.
Figure 6.39: Histogram of POD upper bound for 3 parameter logit model.
Figure 6.40: Histogram of lower confidence upper bound for 3 parameter logit model.
Figure 6.41: Histogram of POD upper bound for 4 parameter logit model.
Figure 6.42: Histogram of lower confidence upper bound for 4 parameter logit model.

Figure 6.43 displays a box plot showing the range of asymptotes for the mean POD curve. Note that the vast majority of the asymptotes are below 90% POD. Figure 6.44 shows the box plot for the range of lower 95% confidence bounds on the mean POD curve. The vast majority of these lower confidence bounds range from 60% to 75% POD.
Figure 6.43: Box plot for upper bound results with 3 and 4 parameter logit models.
Table 6.3 displays the summary of results for the simulation study with 30 observations. The coverage of all the models is quite good, and it seems that the confidence bounds are quite conservative for small data sets. The downside of this is that inspection intervals may be set more frequently than necessary if small data sets are used for POD evaluation. It is also noted that it is difficult to do a meaningful comparison of model-form using the Bayes factor method because the results are mixed. For example, the ratio of the marginal likelihood of model 2 and model 3 is only above one 59% of
time, and the histogram in Figure 6.45 shows that the majority of cases don’t overwhelmingly support the 2 parameter model. The same is true when looking at the ratio of the marginal likelihoods of model 2 and model 4 as depicted in Figure 6.46.

<table>
<thead>
<tr>
<th></th>
<th>2 parameter logit model</th>
<th>3 parameter logit model</th>
<th>4 parameter logit model</th>
</tr>
</thead>
<tbody>
<tr>
<td>% coverage true a₀₀</td>
<td>85.417%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Average a₀₀</td>
<td>3.039</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average a₀₀/₀₅</td>
<td>7.680</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Average upper bound on POD</td>
<td>-</td>
<td>80.316%</td>
<td>85.646%</td>
</tr>
<tr>
<td>Average upper bound on lower 95% confidence</td>
<td>-</td>
<td>64.315%</td>
<td>70.058%</td>
</tr>
<tr>
<td>% Bayes factor &gt; 3</td>
<td>-</td>
<td>59%</td>
<td>54%</td>
</tr>
<tr>
<td>Average Bayes Factor</td>
<td>-</td>
<td>19.96</td>
<td>18.48</td>
</tr>
</tbody>
</table>

Table 6.3: Results summary for simulation study 2 with 30 observations.
Figure 6.45: Bayes factor Histogram for 2 parameter and 3 parameter model.
Figure 6.46: Bayes factor histogram for 2 parameter and 4 parameter model.

To conclude, the simulation studies with 30 and 100 observations are compared with each other in Figure 6.47 and Figure 6.48. There doesn’t appear to be a statistically significant difference in the $a_{90}$ values for the 2 parameter model, but the variance for the 30 observation case is about double the 100 observation case. There is definitely more conservatism with the $a_{90/95}$ value for the 30 observation case and the variance is approximately 4 times that of the 100 observation case.
Similarly, the results for the mean POD asymptotes and the 95% lower confidence bounds indicate that the results for the 30 observation case are very conservative even though there is good coverage. The 100 observation case is still conservative, but much more reasonable.

Figure 6.47: Comparison of 2 parameter model results from simulation studies.
Based on the results of this simulation study, it is recommended that at least 100 samples be used for hit/miss analysis. Fortunately, in this simulation study, the 30 observation case led to conservative results, which is desirable from a risk assessment point of view. Recall that in the small sample study in Chapter 3, this was not the case. In that case, a 2 parameter model was selected using the Bayes factor approach that led to a very questionable $a_{90\%, 95\%}$ value.
7 SUMMARY AND CONCLUSIONS

The following contributions were made in this dissertation:

1) Improved methods for modeling of the tails of a POD curve were presented. This was done by using the Bayes factor to select among many competing models with different tail characteristics.

2) A new method for confidence bound calculation was implemented for hit/miss analysis using MCMC simulation.

3) A transformation was presented to mitigate violations of key assumptions in linear regression, so that signal response data could be analyzed in more situations.

4) A new method for confidence bound calculation was implemented for signal response data where the model may not necessarily be linear using bootstrapping.

5) The impact of the upper asymptote on a POD curve on probability of failure was demonstrated.

The focus of this dissertation was modeling the extreme behavior of a POD curve. To do this, it was necessary to consider both hit/miss and signal response data. It was also necessary to consider different models. The methods developed will enable better analysis of all types of inspection data in any industry. A noise mixture model was also
proposed as the basis for simulation studies to provide a measure of coverage of the methods developed.

The terminology of “physics-inspired” model was chosen carefully. Using the physics-based model directly would have required modeling the model discrepancy. This is typically done with a Gaussian process model, but much work in this area has produced confidence bounds that are too wide to be useful.

The areas suggested for future work are the following:

1) Physics-based modeling to assist in POD evaluation
2) Sensitivity analysis of coverage as a function of sample size
3) Sensitivity analysis of coverage as a function of departures from normality
4) Sensitivity analysis of coverage as a function of sample distribution of flaw size
5) Nonparametric statistical methods for POD evaluation
6) Tolerance intervals for POD curve
7) Classical design of experiments for POD studies
8) Bayesian design of experiments for POD studies
9) Incorporating human factors data from previous POD studies into new POD studies
10) Leverage crack growth models to estimate crack size distributions based on field inspection data and prior POD studies
One of the ultimate goals of this research is to use models to supplement experimental data to reduce cost and time of POD studies. It is hoped that the introduction of MCMC for computing confidence bounds will be a first step in implementing Bayesian model calibration to enable model-assisted POD. Future work may include Bayesian analysis using model calibration methods proposed by Kennedy and O’Hagan [54], and consideration of rectification [55]. This will allow the physics-based model to be used directly as opposed to using physics-inspired models.

One of the most common questions with regard to POD studies is “What sample size is required?” The coverage probabilities introduced in this work provide very good insight into this question. The simulation studies performed in this work can be extended beyond the two sample sizes (30 and 100) used in this work to develop guidance for sample size decisions. Of course, the nature of the statistical distribution of flaw sizes is influential in coverage and must be addressed simultaneously with the question of sample size.

Nonparametric statistical methods were also not discussed in this work. This may be another area of research that may provide advantages. It would certainly interesting to perform a comparison study of coverage of parametric and nonparametric methods together.
Another important topic that should be investigated is the use of tolerance intervals for risk analysis. The author is not an expert in risk analysis, but it seems that the engineer responsible for structural integrity may be interested in the answering a question like this: What is the POD for all future inspections with 95% confidence. The confidence intervals discussed in this work refer to the mean. In standard linear regression, there is the concept of a prediction interval that answers questions pertaining to future observations. For logistic regression, there is no concept of prediction interval, because it doesn’t make sense to predict a probability; however, a tolerance interval should exist. Tolerance intervals are very difficult to calculate even for standard linear regression [56, 57].

Future work should also include design of experiments (DOE) and sample size requirements. Simulations studies for 30 observations and 100 observations were performed in this work. The difficulty in providing sample size requirements is mainly the uncertainty in variance and its relationship to flaw size. Two directions for DOE development include classical DOE methods and Bayesian DOE meetings that guides the experiment as data is being collected to maximize inference at the area of interest on the POD curve.
It is well known that human factors have significant influence on POD capability. One open area of research is to merge prior human factors elements of POD studies into new studies as a way of minimizing the man hours required for a new POD study. Finally, research should be conducted to update POD capability based on field inspection data and crack growth models.
8 REFERENCES

1. Negaard, G.R. “The History of the Aircraft Structural Integrity Program.” Aerospace Structures Information and Analysis Center (ASIAC), Wright-Patterson AFB, Ohio. 1980


16. Berens, A.P. “Probability of Detection (POD) Analysis for the Advanced Retirement for Cause (RFC)/ Engine Structural Integrity Program (ENSIP)

17. POD/SS version 3, University of Dayton Research Institute, (1999)


35. Eureqa Formulize: Downloaded from www.nutonian.com


