JRC Scientific and Technical Reports

Institute for Energy

ENIQ TGR TECHNICAL DOCUMENT

Probability of Detection Curves:

Statistical Best-Practices

ENIQ report No 41

Luca Gandossi and Charles Annis

JRC
EUROPEAN COMMISSION

Institute for Energy

24429 EN – 2010

The mission of the JRC-IE is to provide support to Community policies related to both nuclear and non-nuclear energy in order to ensure sustainable, secure and efficient energy production, distribution and use.

European Commission
Joint Research Centre
Institute for Energy

Contact information
Address: Westerduinweg 3, NL-1755 LE Petten
E-mail: luca.gandossi@jrc.nl
Tel.: +31.224.565250
Fax: +31.224.565641

http://ie.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

JRC56672

EUR 24429 EN
ISBN 978-92-79-16105-6
ISSN 1018-5593

Luxembourg: Publications Office of the European Union

Printed in The Netherlands.

European Commission
Directorate General Joint Research Centre
Institute for Energy
Petten, The Netherlands

ENIQ TGR TECHNICAL DOCUMENT

PROBABILITY OF DETECTION CURVES:

STATISTICAL BEST-PRACTICES

November 2010 ***ENIQ Report nr. 41*** ***EUR 24429 EN***

ENIQ, the European Network for Inspection and Qualification, publishes three types of documents:

Type 1 — **Consensus documents**
Consensus documents contain harmonised principles, methods, approaches and procedures and emphasize the degree of harmonisation between ENIQ members.

Type 2 — **Position/Discussion documents**
Position/discussion documents contain compilations of ideas, express opinions, review practices, draw conclusions and make recommendations for technical projects.

Type 3 — **Technical reports**
Technical reports contain results of investigations, compilations of data, reviews and procedures without expressing any specific opinion or evaluation on behalf of ENIQ.

This 'ENIQ TGR Technical Document – Probability of Detection Curves: Statistical Best Practices' (ENIQ Report No 41) is a type 3 document.

FOREWORD

The present work is the outcome of the activities of the ENIQ Task Group on Risk (TGR).

ENIQ, the European Network for Inspection and Qualification, is driven by the nuclear utilities in the European Union and Switzerland and managed by the European Commission's Joint Research Centre (JRC). It is active in the field of in-service inspection (ISI) of nuclear power plants by non-destructive testing (NDT), and works mainly in the areas of qualification of NDT systems and risk-informed in-service inspection (RI-ISI). This technical work is performed in two task groups: TG Qualification and TG Risk.

A key achievement of ENIQ has been the issuing of a European Methodology Document for Inspection Qualification, which has been widely adopted across Europe. This document defines an approach to the qualification of inspection procedures, equipment and personnel based on a combination of technical justification (TJ) and test piece trials (open or blind). The TJ is a crucial element in the ENIQ approach, containing evidence justifying that the proposed inspection will meet its objectives in terms of flaw detection and sizing capability. The assurance provided is nonetheless qualitative. Obtaining a quantitative measure of inspection reliability is becoming more and more important, as structural reliability modelling and quantitative risk-informed in-service inspection methodologies become more widely used within the nuclear industry in Europe. Such a measure is essential to quantify the reduction of failure probability, and hence risk reduction, after inspection.

The purpose of this document, aimed mostly at NDT engineers and practitioners, is threefold: (1) to provide a brief literature review of some important papers and reports; (2) to review in a simple and structured way the statistical models that have been proposed to quantify inspection reliability and to point out problems and pitfalls which may occur; and (3) to describe and recommend statistical best practices for producing POD vs size curves from either hit/miss data, or â vs. a data.

The members of the ENIQ Task Group on Risk are:

R. Alzbutas	Lithuanian Energy Institute, Lithuania
V. Chapman	OJV Consultancy Ltd, United Kingdom
D. Couplet	Tractebel, Belgium
C. Cueto-Felgueroso	Tecnatom, Spain
C. Faidy	EDF, France
R. Fuchs	Leibstadt NPP, Switzerland
L. Gandossi	JRC, European Commission, the Netherlands
J. Gunnars	Inspecta Oy, Sweden
P. Lafrenière	CANDU Owners Group, Canada
P. Luostarinen	Fortum Engineering Ltd, Finland
L. Horacek	NRI, Czech Republic
G. Hultqvist	Forsmark Kraftgrupp AB, Sweden
E. Kichev	Kozloduy NPP, Bulgaria
A. Leijon	Ringhals AB, Sweden
D. Lidbury	Serco Assurance, United Kingdom
J. Lötman	Forsmark Kraftgrupp AB, Sweden
P. O'Regan	EPRI, United States

C. Schneider	The Welding Institute, United Kingdom
K. Simola	VTT, Finland
P. Stevenson	Westinghouse Electric, United States
A. Toft	Serco Assurance, United Kingdom
A. Walker	Rolls-Royce, United Kingdom
A. Wegeland	Ringhals AB, Sweden
A. Weyn	AIB-Vinçotte International, Belgium

The authors of this report are Luca Gandossi (IE-JRC) and Charles Annis (Statistical Engineering).

The voting members of the ENIQ Steering Committee are:

R. Chapman	British Energy, United Kingdom
P. Dombret	Tractebel, Belgium
E. Martin	EDF, France
K. Hukkanen	Teollisuuden Voima OY, Finland
R. Schwammberger	Kernkraftwerk Leibstadt, Switzerland
B. Neundorf	Vattenfall Europe Nuclear Energy, Germany
J. Neupauer	Slovenské Elektrárne, Slovakia
S. Pérez	Iberdrola, Spain
U. Sandberg	Forsmark NPP, Sweden
P. Kopcil	Dukovany NPP, Czech Republic
D. Szabó	Paks NPP, Hungary

The European Commission representatives in ENIQ are L. Gandossi and O. Martin.

TABLE OF CONTENTS

1 INTRODUCTION 7

2 A LITERATURE OVERVIEW 10

2.1 SEMINAL PAPERS: UNIVERSITY OF DAYTON RESEARCH INSTITUTE 10
2.2 NONDESTRUCTIVE TESTING INFORMATION ANALYSIS CENTER (NTIAC) 10
2.3 U.S. DEPARTMENT OF DEFENSE 10
2.4 RESEARCH AND TECHNOLOGY ORGANISATION OF NATO 11
2.5 UK HEALTH AND SAFETY EXECUTIVE 11
2.6 NORDTEST 11
2.7 US NUCLEAR REGULATORY COMMISSION (NRC) 12
2.8 EPRI 12
2.9 NASA 12

3 STATISTICAL ANALYSIS OF NDE DATA 13

3.1 NECESSARY MATHEMATICAL BACKGROUND 13
3.1.1 The set-up 13
3.2 THE BINOMIAL MODEL 19
3.2.1 Binomial model: NDE capability at one crack size 19
3.2.2 Binomial model: NDE capability at multiple crack sizes 22
3.2.3 Non-Overlapping Constant Sample Size Method 24
3.2.4 Overlapping Constant Sample Size Method 25
3.2.5 "Optimised Probability" Method 25
3.2.6 Conclusions concerning the Binomial Model approach 26
3.2.7 A Final Warning about Averaging Inspector Performance 27
3.3 NDE DATA WITH INFORMATIVE SIGNAL CHARACTERISTICS: *â VS A* DATA 27
3.3.1 The "â vs a" plot 27
3.3.2 The "â vs a" plot with censoring 30
3.3.3 Probability and Likelihood 32
3.3.4 Likelihood function for censored data 32
3.3.5 Maximum Likelihood Parameter Estimates 34
3.3.6 The POD(a) relationship 34
3.3.7 Confidence bounds on POD(a) 35
3.3.8 Noise 37
3.3.9 Analyzing Noise 39
3.4 NDE DATA WITH BINARY SIGNAL RESPONSE: HIT/MISS DATA 39
3.4.1 The Parametric POD Model - Maximum likelihood analysis for hit/miss data 39
3.4.2 Joint, Marginal and Conditional probability 43
3.4.3 Estimating the GLM parameters 45
3.4.4 Confidence bounds 48
3.4.5 Separating the Influence of crack size and Inspector Capability on POD 50
3.5 SPECIAL CASES BEYOND THE SCOPE OF THIS REPORT 53
3.5.1 POD models that consider more than target size 53
3.5.2 Min(POD) > 0 or Max(POD) < 1 53
3.5.3 Field-finds 54
3.5.4 Non-normal scatter in â vs a plots 54
3.5.5 Model-Assisted POD 56
3.5.6 Bayesian Considerations 56

4 ANCILLARY TOPICS 59

4.1 ON THE INDEPENDENCE OF INSPECTIONS 59
4.2 SAMPLE SIZE REQUIREMENTS 61

5 SUMMARY AND CONCLUSIONS 62

6 ACKNOWLEDGEMENTS 63

7 REFERENCES 64

This page is intentionally left blank.

1 INTRODUCTION

In the application of a non-destructive evaluation (NDE) method there are several factors that will influence whether or not the inspection will result in the correct decision as to the presence or absence of a flaw. In general, NDE involves the application of a stimulus to a structure and the subsequent interpretation of the response to the stimulus. Repeated inspections of a specific flaw can produce different magnitudes of the stimulus response because of very small variations in setup and calibration. This variability is inherent in the process. Different flaws of the same size can produce different response magnitudes because of differences in the material properties, flaw geometry and flaw orientation. Further, the interpretation of the response can be influenced by the capability of the interpreter (manual or automatic) and by the mental acuity of the inspector (in turn, dependent on many factors such as fatigue, emotional outlook, ease of access, environment, etc.) [Berens (1989)].

Much of the modern literature on inspection reliability constantly refers to a small set of seminal papers, for instance [Berens and Hovey (1981), Berens and Hovey (1984)], which were produced in the early 1980s. A very good summary of the analytical framework that was devised to treat NDE data in order to obtain probability of detection (POD) curves is given in Berens (1989). Berens and Hovey (1981), Berens and Hovey (1984) and Berens (1989) are highly recommended to everyone wishing to gain an appreciation of flaw detection reliability. In these seminal papers, the fundamentally stochastic nature of crack detection was recognized.

A large part of the existing literature on NDE reliability has been produced in the aeronautical industry. Most of the issues involved are of course very similar. One notable exception is possibly the fact that in the nuclear industry, by the very nature of the components being inspected, the sample sizes of inspected cracks tend to be much lower. In Europe, the ENIQ methodology for inspection qualification [ENIQ (2007)] was specifically developed in the early 1990s because of the difficulty and cost of procuring or manufacturing representative flaws in test pieces in a high enough number to allow to draw quantitative (statistical) conclusions on the capability of the NDE system being investigated. Rather, the fundament of the ENIQ methodology is the Technical Justification, a document assembling evidence and reasoning providing assurance that the NDE system is capable of finding the flaws which it is designed to detect with a high enough reliability. This assurance is qualitative, and comes usually in the form of statements such as: "*Sufficient experimental verification of the procedure has been performed, on representative defects in test blocks with the correct geometry, to be confident that the procedure and equipment will find all defects which conform to the detection criteria and to specifications within the range of plausible defects"*.

The importance of obtaining a quantitative measure of inspection reliability is justified by the fact that structural reliability modelling and quantitative risk-informed in-service inspection methodologies are becoming more widely used within the nuclear industry in Europe. A measure of inspection reliability is essential to quantify the reduction of failure probability, and hence risk reduction, after inspection.

The ENIQ approach to Inspection Qualification has been very successful and it is widely applied, especially in European countries. Due to its central reliance on the Technical Justification, whose qualitative content is normally not amenable to a formal statistical analysis, such an approach has not fostered an environment where statistical models for NDE reliability could be used effectively. The purpose of this document, aimed mostly at NDE engineers and practitioners, is threefold: (1) to provide a brief literature review of some important papers and reports; and (2) to review in a simple and structured way the statistical models that have been proposed to quantify inspection reliability and to point out problems and pitfalls which may occur, (3) to describe and recommend statistical-best-practices for producing POD vs size curves from either hit/miss data, or â vs. a data (see section 3).

As discussed above, the probability of detecting a crack depends on many factors: not only factors intrinsic to the defect itself (its shape, location, roughness, material, etc.) but also factors related to the inspection system (procedure, hardware, software, operator capability, etc.). For structural integrity reasons, probability of detection is nearly always derived (and plotted) against crack size, in particular the through-wall extent of the crack. It is thus not surprising that cracks having the same through-wall extents may have different detection probabilities.

In principle, a certain probability of detection can be thought for a given crack size as the population proportion of cracks having that size that will be found. This assumption leads to a rather simple statistical model (with advantages and disadvantages), based on the binomial distribution, that will be discussed in section 3.2. It is based on considerations resulting from this very statistical model that led the ENIQ network to state that: "[…] *ENIQ argues that it is normally not possible or practicable to construct a convincing argument for inspection capability using results from test piece trials alone*". Also, it is from this model that the following statement is justified: "*For example, if 95% probability of detection of a particular defect type were required, with 95% confidence, this would require the manufacture of test specimens containing 59 examples of this defect type, and the detection of all 59 in the test piece trial. This process would have to be repeated for each defect type of concern*." [ENIQ (1998)].

It is important to note here that in the aeronautical industry this model was used only up to the middle of the 1970s, and then more or less abandoned in favour of more sophisticated ones, which we will review in more detail in sections 3.3 and 3.4. The main driver came from the realization that for a given number of specimens a much more precise POD vs size curve can be obtained with a parametric model than by connecting the dots of point estimates.

Berens and Hovey (1981) and Berens and Hovey (1984) proposed that, at each crack size, a distribution of crack probabilities exist, and the POD curve is then seen as the curve through the mean of detection probabilities. The idea is then to use a model where the whole POD curve as a function of crack size is assumed to have a functional form which is dependent on a small number of parameters (usually 2).

We will discuss the fact that this is an important assumption to make, and that it is very important that the user is well aware of the implications. Having made such an assumption, the number of data points required to obtain a full POD curve becomes much smaller. Berens and Hovey (1984) states that 60 data points (covering all crack sizes) should be enough to obtain a reliable POD curve. This is much less than the roughly 59 flaws for just

one point on the POD curve in the binomial analysis (to prove a 95% probability of detection with 95% confidence).

In Berens and Hovey (1984), the authors wrote that: "*one of the most controversial aspects of NDI[1] reliability estimation is the selection of a model for the POD function*". In Berens and Hovey (1981), they investigated six different functional forms for the POD curve, and concluded that the log-logistic model provided the best fit for the data that were analysed (data from the "Have cracks will travel" US Air force (USAF) program). They also explicitly stated that "*however, the evidence is still limited to this study*" [Berens and Hovey (1984), page 31].

We will see in section 3.3 the reason why the log-logistic model was considered good. A clear advantage it offers is its particularly straightforward analytical tractability coupled with a much more efficient use of the information contained in the binary, hit/miss data. The important point we wish to make here is the fact that in the many subsequent studies of NDE reliability we have seen, this model is nearly always used directly, and always quoting Berens and Hovey (1981), as having proved that it is the best model. The past three decades have indeed reinforced the superiority of parametric modelling to describe probability of detection as a function of target size, but we stress the fact that the user should adopt it in full awareness of the assumptions and limitations implicit in the choice.

[1] Non Destructive Inspection is the way how NDE used to be called before practitioner agreed that "inspection" was too narrow, being only the physical process of examining a part while "evaluation" also includes the inspection and the subsequent statistical assessment of the results.

2 A LITERATURE OVERVIEW

The literature on probability of detection and NDE reliability is quite extensive, especially concerning the aeronautical industry. As already mentioned, Berens and Hovey (1981), Berens and Hovey (1984) and Berens (1989) are highly recommended papers to everyone wishing to gain an appreciation of flaw detection reliability. In this section, a list of interesting or relevant documents is briefly discussed.

2.1 Seminal papers: University of Dayton Research Institute

Berens and Hovey (1981) is a report that reviews the statistical methods used to model inspection reliability up to 1980 and introduces the probabilistic model that has been mostly used since then (the assumption of a functional form for the POD curve). It is a very good report, by no means dated, that presents and explains the models in a very clear way. In addition, it contains the study that concludes that the log-logistic model is (at least for the analysed data) the one giving the best results. This conclusion is continuously referenced, in many subsequent studies, as the proof that the log-logistic model is the best to model POD curves.

Berens and Hovey (1984) has the same authors and a content which is largely similar to the report described above, with a more detailed investigation of the statistical properties of the models proposed.

Finally, Berens (1989) is a short (12 pages) article published in the ASM Metals Data Book (Volume 17), and gives a very good summary of the statistical models proposed by Berens and Hovey (1981) and (1984).

2.2 Nondestructive Testing Information Analysis Center (NTIAC)

Rummel and Matzkanin (1997) is a comprehensive data book, published by the Nondestructive Testing Information Analysis Center (NTIAC), with hundreds of POD curves for different NDE techniques and many different aeronautical components. It is intended to be a condensed reference to previously demonstrated NDE capabilities. Another interesting report published by NTIAC is Matzknin and Yolken (2001), which includes a comprehensive list of recent references.

2.3 U.S. Department of Defense

MIL-HDBK-1823A (2009) is a United States Department of Defense handbook which was recently updated (the latest release is dated 7 April 2009). This handbook provides guidance for establishing NDE procedures for inspecting flight propulsion system, airframe components, etc. The methods include Eddy Current, Fluorescent Penetrant, Ultrasonic, and Magnetic Particle testing. The models and techniques described in this document can be applied to the nuclear industry. Appendix G, in particular, provides a very detailed description of the statistical methods employed to analyse NDE data, to produce POD curves and 95% confidence bounds, of noise analysis, and of noise/detection trade-off curves. It also presents worked-out examples using real hit/miss and signal strength ($\hat{a}$) data. It can be downloaded from http://www.mh1823.com/mh1823/index.html.

Another very good document is the Damage Tolerance Design Handbook (2006), This report is available online at http://www.afgrow.net/applications/DTDHandbook/. In particular, Chapter 3 describes and compares the common NDI methods and presents a very good discussion of the statistically based demonstration programs that are required to quantify the detection capability of an NDI system.

Singh (2000) reviews three decades (from 1970 to 1999) of engineering and research efforts to quantify capability and reliability of non-destructive inspections in aerospace and other industries. This report, performed by Karta Technologies on behalf of the United States Air Force, covers nearly 150 reports and manuscripts from over 100 authors.

2.4 Research and Technology Organisation of NATO

The Research and Technology Organisation of NATO has published a comprehensive report on inspection reliability and POD curves, [RTO (2005)]. The report is targeted at the inspection of aeronautical structures. One important contribution is a detailed summary of the close relationship between NDE, fracture mechanics and airworthiness including a review of the statistical basis for many of the current approaches to inspection. This report can be downloaded from http://www.rta.nato.int/Pubs/RDP.asp?RDP=RTO-TR-AVT-051.

AGARD-LS-190 (1993) is a series of lectures sponsored by NATO, by C. Annis and S. Vukelich, aimed at providing a methodology to quantify probability of detection. The methodology includes design of experiments, specimen generation and maintenance, statistical analyses, data reduction and presentation, and the procedure required to establish a reliable probability based inspection for detecting anomalies in engine parts. The report can be downloaded from http://ftp.rta.nato.int/public//PubFullText/AGARD/LS/AGARD-LS-190///AGIA2DL5190.pdf.

2.5 UK Health and Safety Executive

The UK Health and Safety Executive has published two reports on POD curves. The first [Visser (2002)] is a review of relevant results on probability of detection and probability of sizing of defects in welded structures (offshore industry). The second [Georgiou (2006)] is a very good document whose main goal is to provide clear and understandable information on POD curves to Health and Safety Inspectors when discussing safety cases involving POD curves. The main premise of Georgiou (2006) is that a large amount of POD data is available, for instance from the National NDT Centre (UK), NORDTEST (Norway), NIL (Netherlands) and NTIAC (USA), but that POD curves produced from experimental data are not very well understood by many who use and apply them. Georgiou (2006) quotes, for example, the fact that a certain material and thickness may have been used in producing a POD curve and yet the same curve is then quoted for a range of thicknesses. In other cases, POD curves may have been developed for pipes, but they have been applied to plates or other geometries. In conclusion, Georgiou (2006) makes a point which is similar to the central aim of this report, i.e., that it is very important to question the validity of how POD curves are applied and to keep their limitations well in mind.

2.6 NORDTEST

Two NORDTEST reports dealing with NDE reliability are Førli et al. (1998) and Førli et al. (1999).

2.7 US Nuclear Regulatory Commission (NRC)

Gosselin et al. (2007) contains a chapter which describes how experts from EPRI and PNNL used industry fatigue crack detection data to develop NDE performance-based POD curves. The fatigue crack detection data were assembled from the industry Performance Demonstration Initiative (PDI) testing results at the EPRI NDE Center in Charlotte, North Carolina. The results were subsequently used to establish a database of 16,181 NDE performance observations with 18 separate fields associated with each observation.

2.8 EPRI

Selby and Harrington (2009) is a recent EPRI report that describes a methodology for extracting POD from inspection qualification records. Ammirato and Dennis (2009) address the flaw-sizing problem, and Patrick O'Regan (2010) reports that EPRI has used Generalized Linear Models, described in Section 3.1.1.2 of this report, to develop POD models.

2.9 NASA

NASA has not kept pace with the USAF and others in advances in NDE analysis, as summarized by a recent paper, Generazio (2009), and still relies on a binomial approach to analyse POD data.

3 STATISTICAL ANALYSIS OF NDE DATA

Sometimes obsolete methods and poor statistical practices are inadvertently perpetuated because the reasons for their obsolescence are not sufficiently well explained. This is especially true for methods for analyzing probability of detection. As a result, superior statistical methods, which have been available to statisticians for decades, remain largely unknown to the engineering community. In our review of historical methods we will also explain their serious shortcomings, why they are obsolete, and why they should not be used. We will also provide and explain the statistical best practices (as of 2010) to be used instead.

We recommend that the interested reader try to replicate the analyses and calculations presented in this report, as nothing aids understanding like a hands-on approach. While it is possible to implement these analyses starting from a clean sheet, just as it would be possible to program a finite-element analysis starting with first principles, in practice it is more expeditious to use existing, thoroughly tested, computer code. The emerging world standard for statistical computing and graphics is **R**.[2] Further, one of the authors of this report has written a software add-on (**mh1823 POD**) to carry out POD analyses with **R** and to help promulgate the methods presented in MIL-HDBK-1823A (2009). This software is available for free[3].

3.1 Necessary Mathematical Background

For reasons that will become obvious later, it is necessary to consider two approaches to analyzing data collected as (x, y) pairs. The purpose of the analysis here (indeed for all analyses in this report) is to infer the behaviour of a population by studying the behaviour of a small sample taken from it. Simulated data has the advantage of arising from a known population so we can compare the two analysis approaches, not only against each other, but also against the *known true behaviour*, something not possible with laboratory data.

3.1.1 The set-up

Consider five repeat y observations, taken at each of six different values of x, 30 observations in all. The underlying true relationship between x and y is $y=\beta_0+\beta_1 x$. To simulate randomness in the experiment, to each Y is added a random error: $y=\beta_0+\beta_1 x+\varepsilon_i$. The error is normally distributed with zero mean and standard deviation 5, $\varepsilon \sim N(0,5)$[4]. The six values for X are: 10, 20, 30, 40, 50, and 60. The data are shown in Figure 1, along with the line that generated them.

The purpose of the analysis is to estimate the true line from the information contained in the samples.

[2] **R** is a *free* software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. Binary code is also available for Windows. See http://www.r-project.org/

[3] The software add-on mh1823 POD is available for free at http://mh1823.com/mh1823.

[4] In the statistics literature the symbol"~" is read "is distributed as", a shorthand notation for specifying a probability distribution.

3.1.1.1 Method 1 – Modelling Group Behaviour

The data are in groups of five observations each, so we will calculate the 5 sample means and 5 standard deviations, assuming that they have a normal distribution. Of course we *know* they are normally distributed because this is a simulation but we are developing a method to use with real data where the true line is not known. To estimate the line we connect the sample means. To estimate the lower (and upper) bound we draw normal distributions centred at the sample means and based on the sample standard deviations. The result is shown in Figure 2.

But what if the data are not conveniently taken in nice increments of X? It would be more common for the X values to be random on the X interval of interest, as in Figure 3.

In the first example the data were already in groups so taking group means and standard deviations seemed like an obvious way to proceed. Since the data in the second example are not grouped naturally, and because we need to connect the dots to estimate the underlying x, y relationship (red line) we will need an algorithm to group the data. One choice would be to take the first *n* points, then the next *n*, and so on. For our example, n=5. We group the first 5 points and compute the mean y for the group. We also compute the standard deviation. But what X value should we use? We could choose the mean of the n x values. Or we could be conservative and choose the largest x in the interval. Figure 3 and Figure 4 present the data segregated into groups by each of these two methods, with their group means represented by black dots.

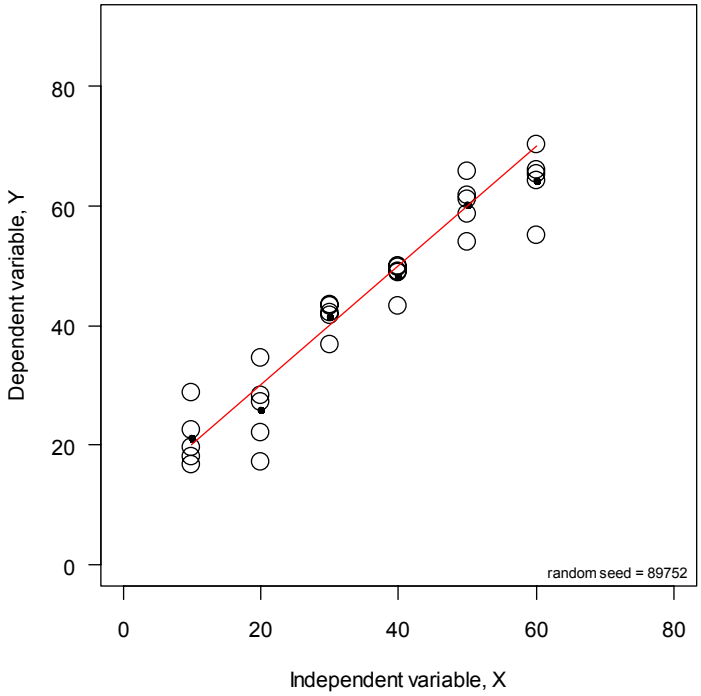

Figure 1
Simulated data to be analysed using two proposed methods, showing the sample means as small black dots, and the true x, y relationship as a straight red line.

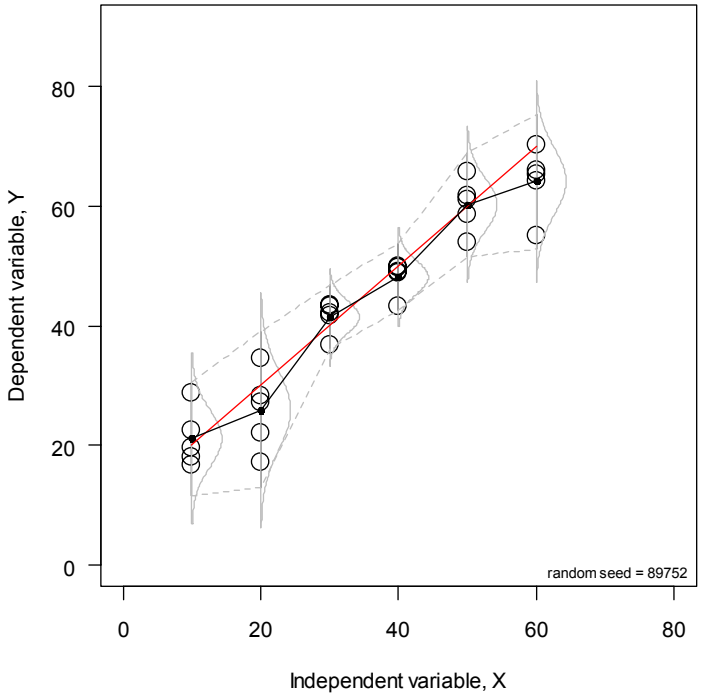

Figure 2
Method 1: estimate the mean behaviour by connecting the sample means; estimate the bounds by connecting the points at $sample.mean \pm 2 sample.sd$

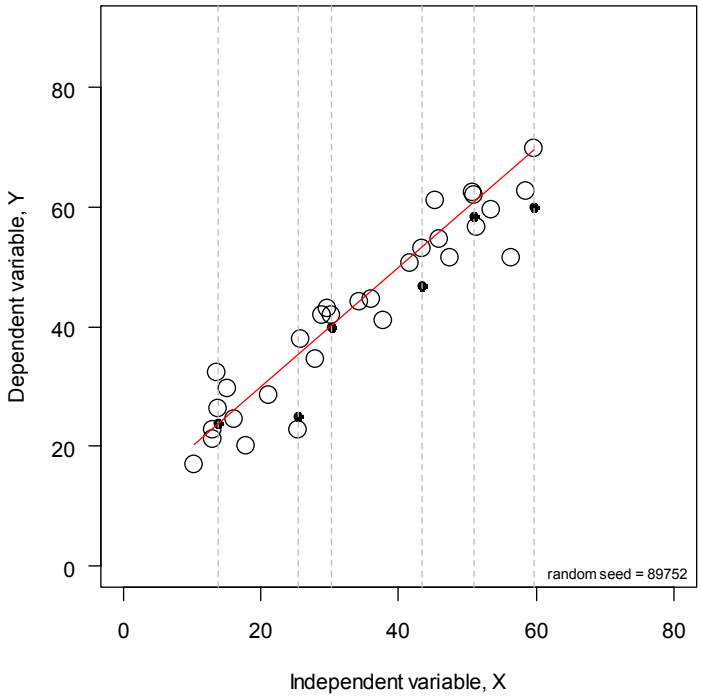

Figure 3
Simulated data showing the true x, y *relationship as a straight red line.*

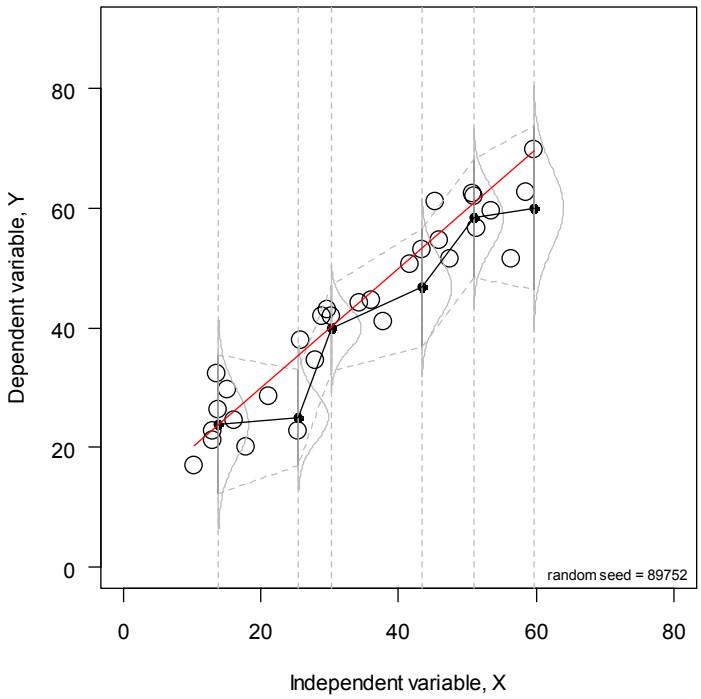

Figure 4
Simulated data showing the true x, y relationship as a straight red line, the estimated mean behaviour by connecting the group means(black dots), and the lower and upper bounds connecting the $\pm\sigma$ points.

A second algorithm for analyzing data that has no natural grouping would be to choose intervals of constant width, as compared with the previous intervals having a constant number of points in each. Anyhow, this approach will not be pursued further here.

All of the mathematical manipulations in Method 1 are valid, and Figure 3 and Figure 4 are valid analyses, but this approach begs several questions:

- Is the true underlying relationship really as crooked as it appears?
- Are the 5 standard deviations really different, or do they result entirely by chance, and might another random sample of 30 look rather different?
- Is it the best we can do?

Method 1 requires estimating 12 parameters, 6 sample means, and 6 standard deviations, and tacitly assumes that the observed behaviour is the actual behaviour. Since this is a simulation we know that the true relationship is a simple line. So the deviations from a line are only the result of random chance.

We also know that all 30 random errors came from the same density function, $\varepsilon \sim N(0,5)$. The 5 sample standard deviations in Method 1 were 4.74, 6.52, 2.71, 2.77, 4.35, and 5.60, for the naturally grouped data, and 5.81, 4.02, 3.56, 4.98, 5.00, 6.84 for the randomly spaced data[5]. The true value is 5, so even though each of those estimates of standard deviation is valid, the differences are random, and thus meaningless.

[5] Because they had the same random number seed, the random errors were the same for both groups. The standard deviations are different for the constant interval group because the Y values changed over the interval, due to the random X, while the naturally grouped data had the same Y value for all observations within that group, because they all had the same X value.

3.1.1.2 Method 2 – Parametric Modelling

The method of parametric modelling recognizes that there is some underlying reality and small deviations are due only to randomness. (Large deviations would indicate that we have chosen the wrong parametric model.) If we assume that the underlying relationship is a simple straight line, and further assume that any scatter is from the same probability distribution (and not from 6 individual distributions) we could build a parametric model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad \text{(Eq. 1)}$$

We know the x, y pairs, but we don't know the model parameters, β_0, β_1, and we do not know the distribution of the errors, ε, so we have three parameters to estimate, rather than 12. By assuming a parametric model we can use all of the data everywhere, recognizing that the apparent differences in local behaviour are not real differences.

The reader certainly recognizes that one kind of parametric model is the linear model[6], and one example of the linear model is ordinary least-squares (OLS) regression. Assuming that (Eq. 1) is true, and that the variance is constant over the range of the model, we can now analyse the data using Method 2, the parametric model with OLS regression. Figure 5 and Figure 6 compare Method 2 with Method 1.

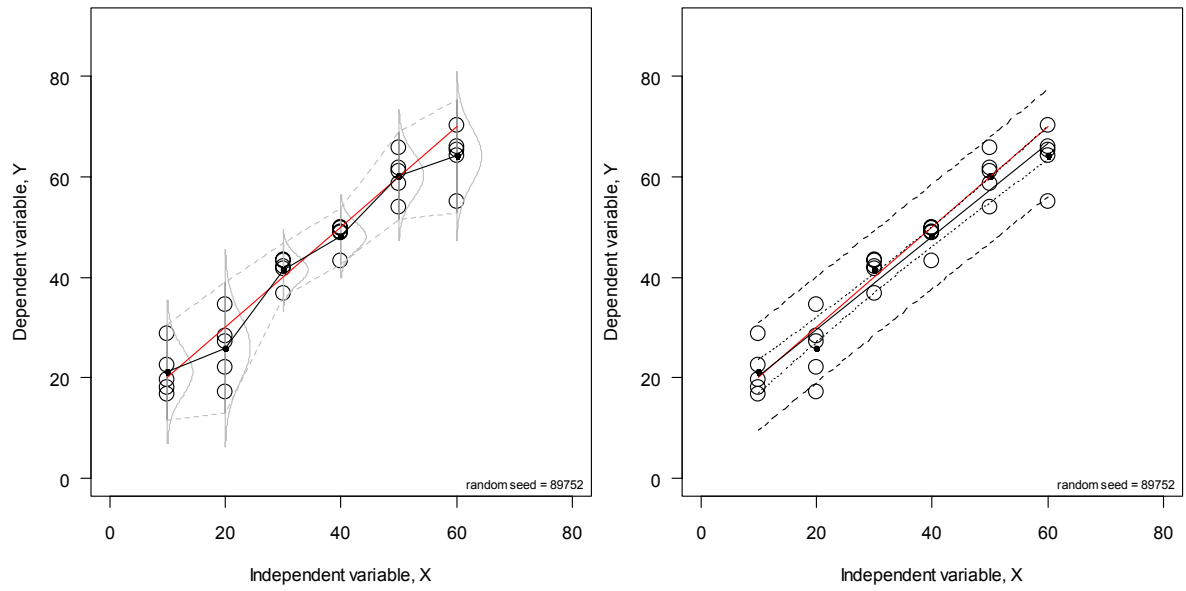

Figure 5
A parametric model can use all the data everywhere (data of Figure 2): Method 1 (left) versus Method 2 (right).

[6] Statisticians and engineers use the term "linear model" differently. In statistics a model is linear if it is linear in the model parameters. In engineering a model is linear if it is linear in x, the dependent variable. So $y=\beta_0+\beta_1 x^2+\beta_3 sin(x)$ is a linear statistical model, but $y=\beta_0+\beta_1 e^{-\beta_3 x}$ is not.

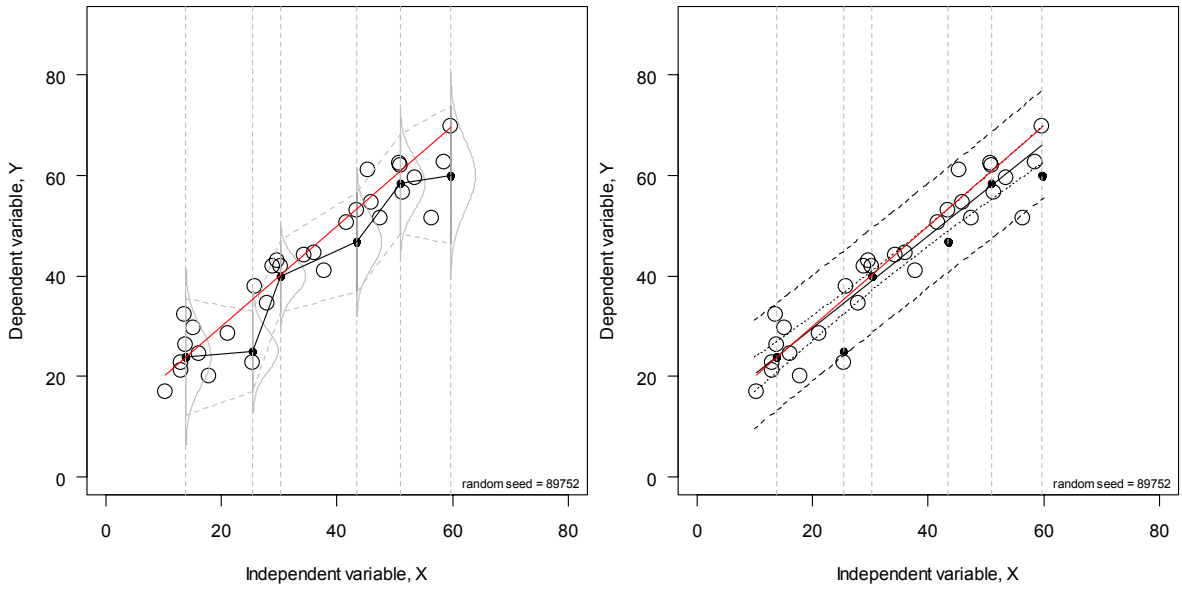

Figure 6
A parametric model can use all the data everywhere (data of Figure 3): Method 1 (left) versus Method 2 (right).

A parametric model produces a more believable description of the underlying reality, and does not tempt the unwary into trying to explain group differences that are only illusory. The parametric model is more efficient, requiring only three parameters as compared with 12, and thus provides more (30–3=27) degrees of freedom for estimating the standard deviation of the underlying variability ("error").

Notice that there are two sets of bounds on the regression plot. The inner-most bounds are the *confidence bounds on the mean line*. We would expect the confidence bounds to contain the true relationship (red line) in 95 of 100 nominally identical experiments. The outer bounds are *prediction bounds on the individuals*. We would expect the next future *single* observation to fall within the prediction bounds 95% of the time[7]. It is important to distinguish these two bounds. Confidence bounds describe how well the model captures the true x, y relationship. The prediction bounds describe the anticipated behaviour of the next *single* observation. The reason for our close attention to this distinction will become clear when we study the binomial model.

It will be clear in the following discussion that the binomial POD model relies on Method 1 and thus has all of Method 1's shortcomings, especially treating random behaviour as if it were meaningful. We will show that parametric models (Method 2) are far superior for dealing with binary data, but first, for historical reasons, we will study some older methods and discuss more in details their shortcomings.

[7] This does *not* mean that 95 of the next 100 observations will fall within the prediction bounds. It means that of 100 similar, nominally identical, experiments, the next *single* observation in 95 of the experiments would be contained within that experiment's prediction bounds. If the probability that the next single observation will be within the prediction bounds is 0.95, then the probability that the next *two* observations will both be within the bounds is 0.95x0.95 = 0.9025, so the 95% prediction bounds for a single future observation are also the approximate 90% bounds for the next two observations.

3.2 The Binomial Model

Even though the binomial model is generally inappropriate[8] for data collected from cracks of varying sizes, it is important historically, and, remarkably, is still in wide-spread use [e.g. ASME V Article 14; ASME XI Appendix VIII]. So before discussing parametric modelling, the binomial model will be investigated closely.

3.2.1 Binomial model: NDE capability at one crack size

Because of individual physical differences, individual cracks having the same size will have different detection probabilities (for a given, fixed NDE system). The recognition of this idea led to the development of the models that will be described in section 3.4.

However, a single POD for all cracks of that size can be postulated in terms of the probability of detecting a randomly selected crack from the population of all cracks of that given size. In this framework, the proportion of cracks detected in a random sample is an estimate of the POD for that size. Each experiment is seen as a Bernoulli trial, and binomial distribution theory can be used to calculate a lower confidence bound on the estimate.

This is precisely the model that was used in Gandossi and Simola (2005) and Gandossi and Simola (2007). In particular, Appendix 1 of Gandossi and Simola (2007) describes in a certain detail the estimation of the relevant statistical quantities (such as confidence bounds), both in the classical and Bayesian statistical frameworks.

It is very interesting to note that in the aeronautical industry this model was extensively considered in the 1970s but is virtually not applied anymore for the quantification of NDE reliability [Damage Tolerance Design Handbook (2006)], for reasons we are going to review in the following. It is also interesting to note that such single crack size characterisations of NDE capability were expressed in terms of a crack size for which there was at least a given POD at a defined confidence level (CL). This was called the POD/CL crack size, indicated as $a_{POD/CL}$.

Let us suppose that the target crack size is a_{NDE}. The inspection system is considered adequate if the lower confidence bound on the proportion of detected cracks exceeds the desired POD value.

Let us suppose N cracks of size a_{NDE} are inspected, and N_s are successfully detected. If p is the true (but unknown) probability of detection for the population of cracks, the number of detections is modelled by the binomial distribution. The probability of N_s detections (successes) in N independent inspections is:

$$p(successes = N_s) = \binom{N}{N_s} p^{N_s} (1-p)^{N-N_s} \qquad \text{(Eq. 2)}$$

[8] The binomial model may still be appropriate in the (albeit unusual) situation where data is only available for one particular crack size.

p can be approximated by the ratio N_s /N. This is not only a very natural choice, but it turns out that such ratio is also the unbiased, maximum likelihood estimate[9] (MLE) of the true value of p.

$$\hat{p} = \frac{N_s}{N} \qquad \text{(Eq. 3)}$$

$\hat{p}$ may be the best estimate of p we have after carrying out the set of N trials, but alone it does not tell us much about the true value of p, unless N is large. $\hat{p}$ is only an approximation of p, it is a *point estimate of p*. To see why $\hat{p}$ alone is not necessarily very informative it is enough to consider the limit case in which N=1. Then we will have either $\hat{p}$=0 (the single trial was a failure) or $\hat{p}$=1 (the single trial was a success), but this does not tell us much about p. Increasing the sample size, i.e. the value of N, of course helps, but the information offered by a point estimate such as $\hat{p}$ is necessarily limited.

A confidence interval can be built around $\hat{p}$. The purpose of using an interval estimator, rather than a point estimator such as $\hat{p}$, is to have some confidence of capturing the parameter of interest. Moving from a point estimator to a confidence interval estimator entails sacrificing some precision in our estimate of POD (it is now an interval rather than a single point) but results in increased confidence that our assertion is correct.

In our problem of determining the probability of detection of the given NDT system, it is of particular interest to find a lower bound on the values of p, which would provide us with a conservative estimate of the capabilities of the system.

The α percent lower confidence bound, p_{CL}, on the estimate of POD can be obtained as the solution to the following equation [Casella and Berger (1990)]:

$$p_{CL} = \sup\left\{ p : \sum_{i=0}^{N_s-1} \binom{N}{i} p^i (1-p)^{N-i} \geq \alpha \right\} \qquad \text{(Eq. 4)}$$

The interpretation of p_{CL} as a lower confidence bound is as follows. If the experiment (comprising the inspection of N cracks of size a_{NDE}) was *completely and independently repeated* a large number of times, α percent of the calculated lower bounds would be less than the true value of p. In other words, there is α percent confidence that p_{CL} from a single experiment will be less than the true value. Solutions to (Eq. 4) can be found in many textbooks (*cf.* Box, Hunter and Hunter, 1978).

Several objections to the use of the binomial approach to quantifying inspection capability have been raised [Berens and Hovey (1981, 1984), MIL-HDBK-1530 (1996)]:

1. The choice of a particular POD and confidence limits are normally made on a rather arbitrary basis. For example, the Damage Tolerance Design Handbook (2006) quotes JSSG (2006) (a US Department of Defense Joint Service Specification Guide for

[9] The words "estimate", "estimation" and "estimating" have a special meaning in the statistics literature that is different from the common usage meaning "to guess". Because model parameters are unknown, we need a mechanism for determining numerical values for them based on the observed data. Statisticians call this process "estimation" and the resulting value is a parameter "estimate". To distinguish a parameter, whose value is unknown, from a parameter *estimate*, a known value, statisticians use a caret over the parameter to indicate that

Aircraft Structures) where 90/95 (i.e. 90% POD with 95% confidence) values were used. The Damage Tolerance Design Handbook (2006) concludes that 90/95 limits were selected because higher POD or confidence limit values would have required much larger sample sizes in the demonstration programs for the analysis methods being used. A 95 percent confidence limit is assumed to provide the required degree of conservatism, but there is no sound justification for such a choice.

2. A POD/CL limit is not a single, uniquely defined number but rather a statistical quantity. Any particular POD/CL estimate is only one realisation from a conceptually large number of repeats of the demonstration program. Berens and Hovey (1981) showed there can be a large degree of scatter in these POD/CL estimates and the scatter depends on the POD function, analysis method, POD value, confidence level and number of cracks in the demonstration program.
3. The POD/CL characterization is not related to the size of cracks that may be present in the structure after an inspection. Calculating the probability of missing a large crack requires knowledge of both the POD curve, POD(a), for all crack sizes and the crack size distribution of the cracks in the population of structural components being inspected.

For the reasons outlined above quantifying inspection capability in terms of the entire POD(a) function has evolved as the preferred method in the aeronautical and aerospace industry. This approach is described below in sections 3.3 and 3.4.

3.2.1.1 "29 of 29"

It is a property of the binomial distribution, determined using (Eq. 4), that if the true probability of success is 90%, then the probability of observing 29 successes in 29 opportunities is only 0.047, since at least one failure would be anticipated. It is entirely correct to conclude that if 29 out of 29 successes are obtained in an experiment, the lower bound on the true value of p is above 0.9 with 95% confidence. Unfortunately this fact has been misused to justify the practice of "29 of 29". If the inspector correctly finds all 29 nominally identical cracks in 29 attempts, then he has demonstrated 90% POD with 95% confidence for that crack size, which is called $a_{90/95}$.

Unfortunately the "29 of 29" single point POD determination is still being used in spite of its often-discussed deficiencies, which include:

1. There is no procedure for what to do when fewer than 29 cracks are found. What happens in practice is that the inspector is given a second (or third) opportunity to find all 29. Of course the binomial statistical justification for 29 of 29 is violated since what is really demonstrated is 29/(29n) where n > 1 is an integer.
2. The practice makes no accommodation for real differences in crack size so any influence of size on POD is ignored, and specimens having larger, more detectable cracks are grouped with others in some range of sizes. Choosing the single size with which to label the range of sizes is also left to the practitioner, although the largest size in the interval is the customary choice.

---

it is an estimate. For example, an unknown probability might be referred to as "p" while its estimate is "$\hat{p}$," which is read "p-hat."

3. In practice it is difficult to make 29 specimens that are nominally identical because it is hard to stop growing a small crack before it becomes a large crack, and machining away the specimen surface to make the crack smaller results in a more open, more easily detected crack, even if the crack size is within specification.
4. Even if everything worked perfectly, the test would only provide an estimate for a single crack size. The detection probability might also be at least 90% for a crack half that size. There is no way to tell.

To summarize, the "29 of 29" procedure is misleading and should not be used. We comment on it here as a caution to the reader.

The fixed crack size approach (“29 of 29”) is still sanctioned by NASA for quantifying inspection capability of vendors [*cf.* Salkowski (1993)] in spite of its often-discussed deficiencies. Even as recently as 2009 NASA continues to promote inappropriate binomial methods [Generazio (2009)].

3.2.2 Binomial model: NDE capability at multiple crack sizes

In this category of experiments, many components covering a range of crack sizes are inspected once and the results are used to estimate the POD as a function of crack size with confidence limits.

This approach can be seen as an extension of the binomial model described in the previous section (3.2.1) if we have a set of neatly defined sets of cracks, each set being characterised by a different crack size and with all cracks of the set having *exactly* the same size. For instance, if our sample size consists of five sets of 30 cracks each, 5x30=150 cracks total, with the first 30 cracks having 2mm size, the second 30 cracks having 3 mm size, etc, the model described in the previous section could be applied to the results obtained in each crack set. That would lead to the determination of five p_{CL} values that, as a whole, could be seen as a rudimentary POD curve.

In general, it will not be possible to have inspected cracks that so neatly belong to such well defined sets. More likely, the cracks will span a continuous interval of crack sizes. The main idea of this approach is still to group the cracks in intervals of crack size, and to assume that all cracks within a specified interval have approximately the same POD. This immediately causes a problem: increasing the range of the size interval to include a greater number of cracks will improve the resolution in $\hat{p}$ but, because it also results in fewer intervals, will result in a poorer resolution in crack size. Choosing a more narrow size range improves the size resolution, but at a cost of resolution in $\hat{p}$.

With this limitation in mind it would nonetheless be possible to model the fraction of cracks detected in an interval with the binomial distribution, as described in the previous section, and assign the lower confidence bound to the crack size at the upper end of the interval.

Various methods have been developed to form these crack size intervals, ranging from very simple (partitioning the range of data into equal intervals) to more sophisticated ones. These are reviewed in Berens and Hovey (1981) and briefly in the following sections. It is interesting to present at the same time an example to show the limitations of this approach when applied in practice. The example is the same as shown in Berens and Hovey (1981), i.e. the

inspection results for eddy current (EC) inspections of etched fatigue cracks in aluminium flat plates (from Yee *et al*, (1976), page D-63).

3.2.2.1 Range Interval Method

In the Range Interval Method, crack size intervals are defined with equal lengths across the range of data. Due to the somewhat random nature of the crack sizes in the components being inspected, the interval constructed according to this method will very likely contain different numbers of cracks. For this reason, the estimate of the POD curve and its lower confidence bound can exhibit an erratic behaviour.

Let us move to the example. The data are tabulated in Table 1 (note: the interval lengths are reported in milli-inches in the original reference but are here converted in millimetres), already subdivided in intervals. The table also reports the values of $\hat{p}$, i.e. N_s/N, as defined in (Eq. 3), and of p_{CL}, obtained using (Eq. 4) with a confidence level α=95%. The results are plotted in Figure 7.

Table 1 Probability of detection of etched fatigue cracks in aluminium flat plates [Yee, et al, (1976), page D-63], grouped according to the Range Interval Method

Interval (mm)		Number of cracks, N	Detections, N_s	$\hat{p}$	p_{CL} (α=95%)
min	max				
0.18	0.56	13	1	0.08	0.00
0.64	0.91	18	2	0.11	0.02
0.97	1.30	23	3	0.13	0.04
1.37	1.70	46	30	0.65	0.52
1.73	2.08	53	36	0.68	0.56
2.11	2.46	38	36	0.95	0.84
2.49	2.82	18	16	0.89	0.69
2.92	3.20	17	16	0.94	0.75
3.28	3.58	19	17	0.89	0.70
3.63	3.99	15	15	1	0.82
4.01	4.34	3	3	1	0.37
4.62	4.70	3	3	1	0.37
4.83	5.00	2	2	1	0.22
5.89	5.89	1	1	1	0.05
6.12	6.27	3	3	1	0.37
6.30	6.65	17	17	1	0.84
6.81	6.99	3	3	1	0.37
7.09	7.37	7	7	1	0.65
7.49	7.72	6	6	1	0.61
7.87	8.18	10	10	1	0.74
8.20	8.53	12	12	1	0.78
8.59	8.94	11	11	1	0.76
9.04	9.19	4	4	1	0.47
9.40	9.68	5	5	1	0.55
9.75	9.98	2	2	1	0.22
10.36	10.36	1	1	1	0.05
10.82	10.82	1	1	1	0.05
11.23	11.23	1	1	1	0.05
11.28	11.28	1	1	1	0.05
11.63	11.99	8	8	1	0.69

The solid dots represent, for each crack size interval, the value of $\hat{p}$. The empty dots represent the value of the 95% lower confidence bound, p_{CL}. The latter show an extremely erratic behaviour, due to small sample sizes in certain intervals, even though all cracks greater than 3.6mm were detected ($\hat{p}$=1). For example, the very low confidence bound at 5.89mm resulted from the fact that particular interval contained only one crack. Even though detected (and therefore with $\hat{p}$=1), the lower 95% confidence bound on *p* for a sample size of 1 is 5%. The same happens at crack sizes of 10-11mm.

This curve is not useful from an engineering perspective. Even though the binomial analysis was correct mathematically, statistically it represents a suboptimal use of the data so that the artificial grouping of cracks results in an unusable lower bound. The parametric model method described in Section 3.4 avoids this difficulty.

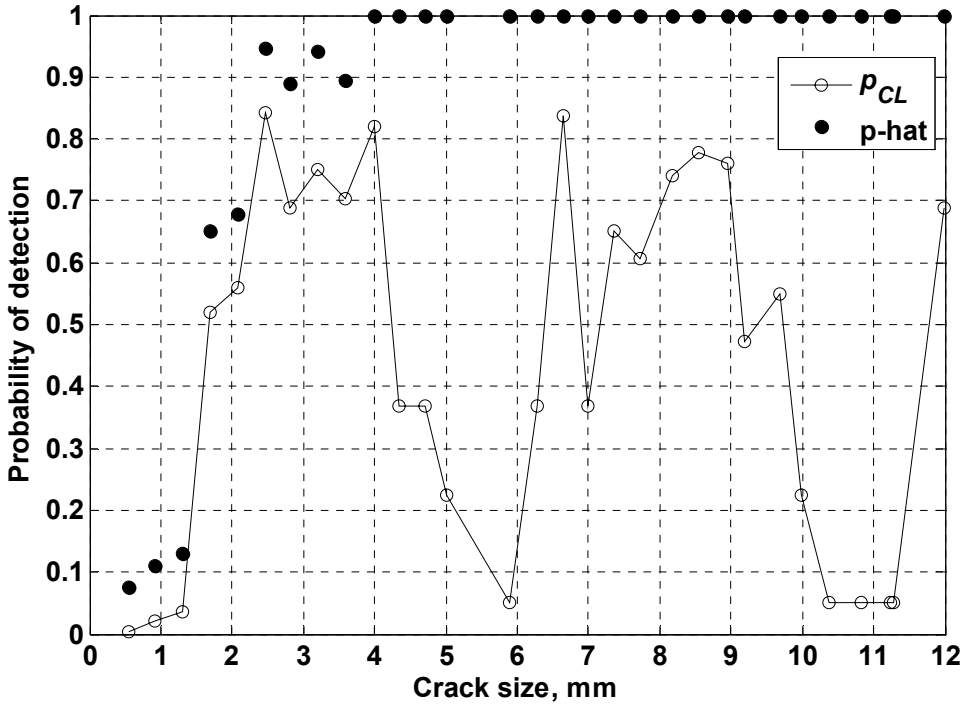

Figure 7
Probability of detection of etched fatigue cracks in aluminium flat plates (eddy current inspections), data from Yee, et al, (1976), page D-63, Range Interval Method.

3.2.3 Non-Overlapping Constant Sample Size Method

To avoid the problems resulting from a sample size which varies from interval to interval, the lengths of the intervals can be changed so that each contains the same number of cracks. This method is called the Non-Overlapping Constant Sample Size method.

If N is the number of cracks assigned to each interval, the intervals are constructed in the following way. The longest N cracks form the first group. These are removed from the total set of data, and the following longest N cracks are assigned to the second group. This process is repeated until all the cracks have been assigned to a group. The intervals thus formed are non-overlapping, they all contain the same number of flaws (except the last one, with the shorter cracks) and they have different widths (depending on the data).

Figure 3 of Berens and Hovey (1981) shows the results of analysing the data of Table 1 using this method, using a value N=60 to define each interval. (This plot is not reproduced here for brevity). The situation was markedly improved. For instance, the lower confidence bound obtained by grouping the data in this way results in a monotonically increasing function. Anyhow, Berens and Hovey (1981) pointed out several problems. First of all, lower confidence bound points are plotted at the longest crack size of the interval, thus introducing an unknown degree of conservatism in the NDE capability. Further, Berens and Hovey (1981) concluded that despite the example at hand resulting in a monotonically increasing POD function, other data sets could still result in erratic behaviour, i.e. the monotonicity is a function of the data and not of the analysis method. Finally, this method reduces the number of intervals for analysis and therefore leaves longer gaps to be filled by interpolation. Increasing the intervals can be achieved by choosing a smaller value for N, but this in turn reduces the sample size in each interval and results in lower confidence bounds.

3.2.4 Overlapping Constant Sample Size Method

To obtain narrower intervals with a relatively large number of cracks in each interval, the intervals can be defined to overlap. Each interval contains the same number of cracks but the same crack can belong to more than one interval [Yee *et al*, (1976)], even though this obviously violates the implicit statistical assumption that the observations (groups) are independent.

The intervals are created in the following way. The longest N cracks are assigned to the first group. Of these, the smallest T percent is grouped with the next N-NxT cracks. The process is then repeated until all cracks have been assigned to at least one interval. T is the percentage of overlapping. If for instance T =50%, half of the cracks of each groups are also assigned to the next one.

Page D-65 of Yee, *et al* (1976) reports the data of the example now grouped according to this method, with N=60 and T =50%. The method is thus called the overlapping 60-points method. The situation shows a clear improvement when compared to the POD curve plotted in Figure 7. The curve is (nearly) monotonic and shows the expected plateau at longer crack sizes, but again erratic behaviour and unknown conservatism (because the points are plotted at the longest crack size of the interval) can result. Finally, Berens and Hovey (1981) state that using the results from a particular inspection in more than one interval results in deriving confidence limits from non-independent data sets with unknown effects on the total measure of NDE capability.

3.2.5 "Optimised Probability" Method

One last grouping method is presented in Yee *et al* (1976) and reviewed in Berens and Hovey (1981) is an algorithm devised to obtain the highest possible lower confidence bounds. This was named the Optimised Probability Method. This method, which is based on dubious statistical manipulation, is not described in detail here. Berens and Hovey (1981) conclude that the confidence bounds obtained with this method are better behaved than those obtained with the previous methods, but that such behaviour is obtained at the expense of unknown statistical validity of the POD curve across all ranges of crack sizes. The overlapping of intervals entails inspection results that are analysed more than once. Any crack belonging to two intervals is used in calculating the confidence bound for both

intervals. Confidence bounds that share the data are correlated and the influence of this correlation on the POD curve as a function of crack size is unknown.

3.2.6 Conclusions concerning the Binomial Model approach.

The methods described in this section suffer serious deficiencies. The POD curves obtained can show very erratic behaviour. Since they are based on the binomial distribution, the confidence bounds are greatly influenced by the method used for assigning cracks to the intervals. The confidence bounds are as much influenced by the analysis method as they are by the data. Crucially, these methods provide only limited inference on the entire POD curve if this curve is required for further studies such as probability of failure analyses. Finally, and most important of all, it will be shown in Section 3.4 that the parametric model method described therein is superior to the binomial model by any measure, and it does not suffer from any of the several deficiencies affecting the binomial model.

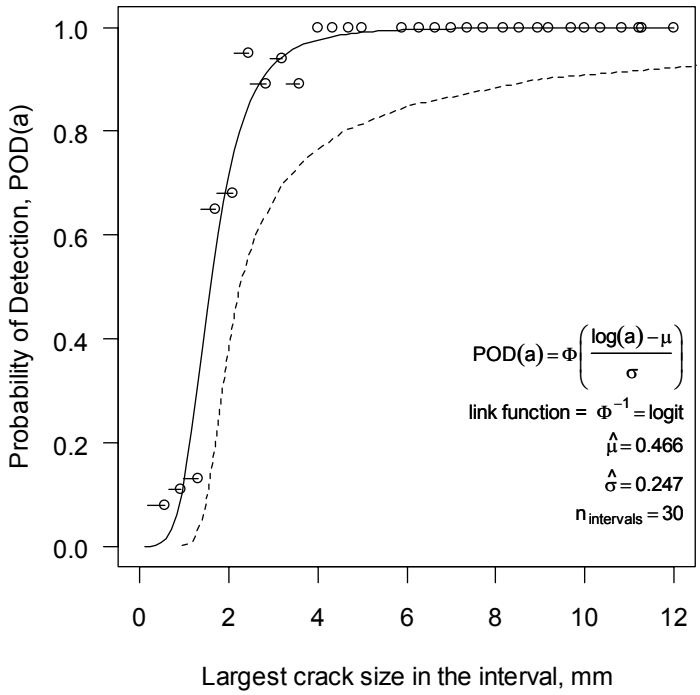

Figure 8
Parametric model for the proportion data in Table 1. The horizontal lines indicate the width of the size interval associated with the average POD (open dots). The parametric model is discussed in detail in Section 3.4.

Because it will be necessary to discuss the parametric model for continuous response ($\hat{a}$ *vs* a) data to build the foundation for extending the model to binary, hit/miss, data, we provide a preview here. Figure 8 shows the parametric model description of the data in Table 1, with the maximum likelihood estimate of its lower 95% confidence bound, for comparison with the plot in Figure 7. The parametric model is as clearly superior to the foregoing methods as Method 2 (OLS regression) was to Method 1 in the introduction to this topic (section 3.1), and for the same reasons: a parametric model can use all the data everywhere, whereas the binomial method can only describe the average POD in a very local neighbourhood, where it is more vulnerable to the vagaries of randomness.

3.2.7 A Final Warning about Averaging Inspector Performance

Much of the historical literature reports only the average POD from a number of repeated inspections of the same test object by multiple inspectors having varying abilities. This makes the data nearly unusable because it obscures the influence of target size with the sometimes erratic performance of the inspectors.

Of course measuring the performance of the inspectors is crucially important. This is why they should be evaluated individually and not averaged with many others. Individual evaluation makes it possible to identify superior performers, so that their techniques can be studied and emulated. It also makes it possible to identify poor performers so that they can be afforded remedial training. Averaging destroys this possibility.

Plotting the group average PODs on a POD vs size plot (as in Figure 7) gives the mistaken impression that such POD "data" can be analysed using the same methods that are appropriate for real observations. They cannot. Remember: you cannot observe a probability. You can only observe an average. To illustrate the difference we will re-examine the data presented in Lewis, *et al* (1978) which is extremely informative because it reports *all* of the data, not just summary averages. To do so we will need to use the parametric model which is described in section 3.4, and so will defer the discussion to that section.

3.3 NDE data with informative signal characteristics: $\hat{a}$ *vs* a data

Although historically hit/miss methods preceded $\hat{a}$ *vs* a methods, and at first glance the methods appear to have little in common, the underlying statistical concepts are very closely related[10]. We will show that the parametric model for binomial data is a generalization of the parametric model described here. Therefore we will address the $\hat{a}$ *vs* a situation first.

3.3.1 The "$\hat{a}$ *vs* a" plot

Figure 9 illustrates the relationship between POD and signal strength, for a given detection criterion. If the signal, $\hat{a}$, is greater than the detection criterion (often called the detection threshold), $\hat{a}_0$, the crack of size a is detected. The probability that $\hat{a} > \hat{a}_0$ depends on signal strength, which itself depends on the size of the crack, and on the random scatter associated with the size versus strength relationship.

Test specimens with known characteristics (crack size, location, orientation, etc.) are inspected and the signal amplitude, $\hat{a}$, (or other meaningful characteristic) is recorded for each crack of size a, and plotted as in Figure 9. Notice that that this "$\hat{a}$ *vs* a" plot may also be log($\hat{a}$) vs a, $\hat{a}$ vs log(a) , or log($\hat{a}$) vs log(a), depending on the behaviour of the data, Figure 10.

Contrary to common perception, there is no single choice (like *log-log*) that works for all cases, so each case must be analysed individually. That means plotting all four cases and

[10] The parametric model describing binary data was introduced in a seminal paper by Nelder and Wedderburn (1972). The binomial methods for POD predate the introduction of the GLM (Generalized Linear Model) but were being adopted by the aircraft industry by the early 1980s. A working group from the USAF, University of Dayton Research Institute, Pratt&Whitney, General Electric Aircraft Engines, and Allied-Signal (now Honeywell), produced MIL-HDBK-1823, "Nondestructive Evaluation System Reliability Assessment". While it would be some years before an official publication was available, the draft became the *de facto* standard for establishing quantitatively the effectiveness of inspections by measuring POD. The document was completely updated in 2009, as MIL-HDBK-1823A.

choosing the best fit. There may be some situations where none of the transformations provides a useful fit, and professional statistical help is then recommended, but in nearly all situations one of these cases will be adequate. In the following discussions we will use "$\hat{a}$ *vs* a" to mean any of these four modelling choices.

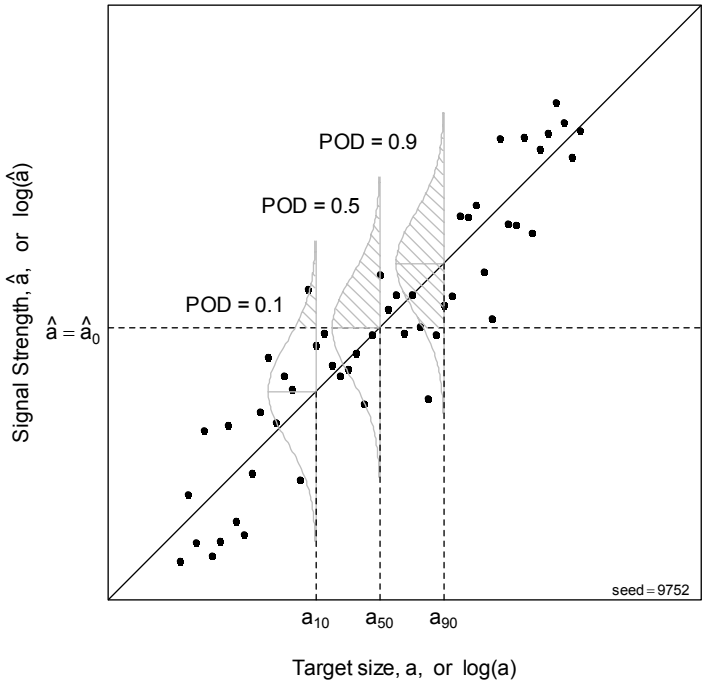

Figure 9
The POD is the fraction of the scatter density that is above the decision criterion.

The validity of $\hat{a}$ *vs* a modelling depends on four conditions:

1. *The $\hat{a}$ vs a model must look like the measured data, i.e. it must follow a similar trend*. Figure 10 shows 4 possible modelling combinations using a straight-line, $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$. If the straight line is not a reasonable representation of one of these, then the subsequent POD curve will be wrong. (More involved models are possible but beyond our scope here, see section 3.5.1)

2. *The variance must be uniform about the $\hat{a}$ vs a line.* That means that the scatter cannot be narrow at one end and wide at the other, or that the scatter is above the line at the ends and below in the middle.

3. *The observations must be uncorrelated*. That means there should not be any influence on $\hat{a}$ except size, caused by, for example, changing equipment settings (e.g. EC probes) for small cracks, or changing operators half-way through the test.

4. *The errors must be (approximately) normal*. Because of the Central Limit Theorem, this condition is nearly always met in practice, but there are situations where it is not. For example if a logarithmic transformation is necessary for the resulting â vs a relationship to be effectively represented by $y_i=\beta_0+\beta_1 x_i+\varepsilon_i$, the resulting errors might be unavoidably transformed from being normal to being skewed.

In practice all four restrictions are almost always easily met. If they are not met – and a plot of $\hat{a}$ *vs* a such as Figure 10 shows a violation of any one of these - the resulting POD curve will be wrong.

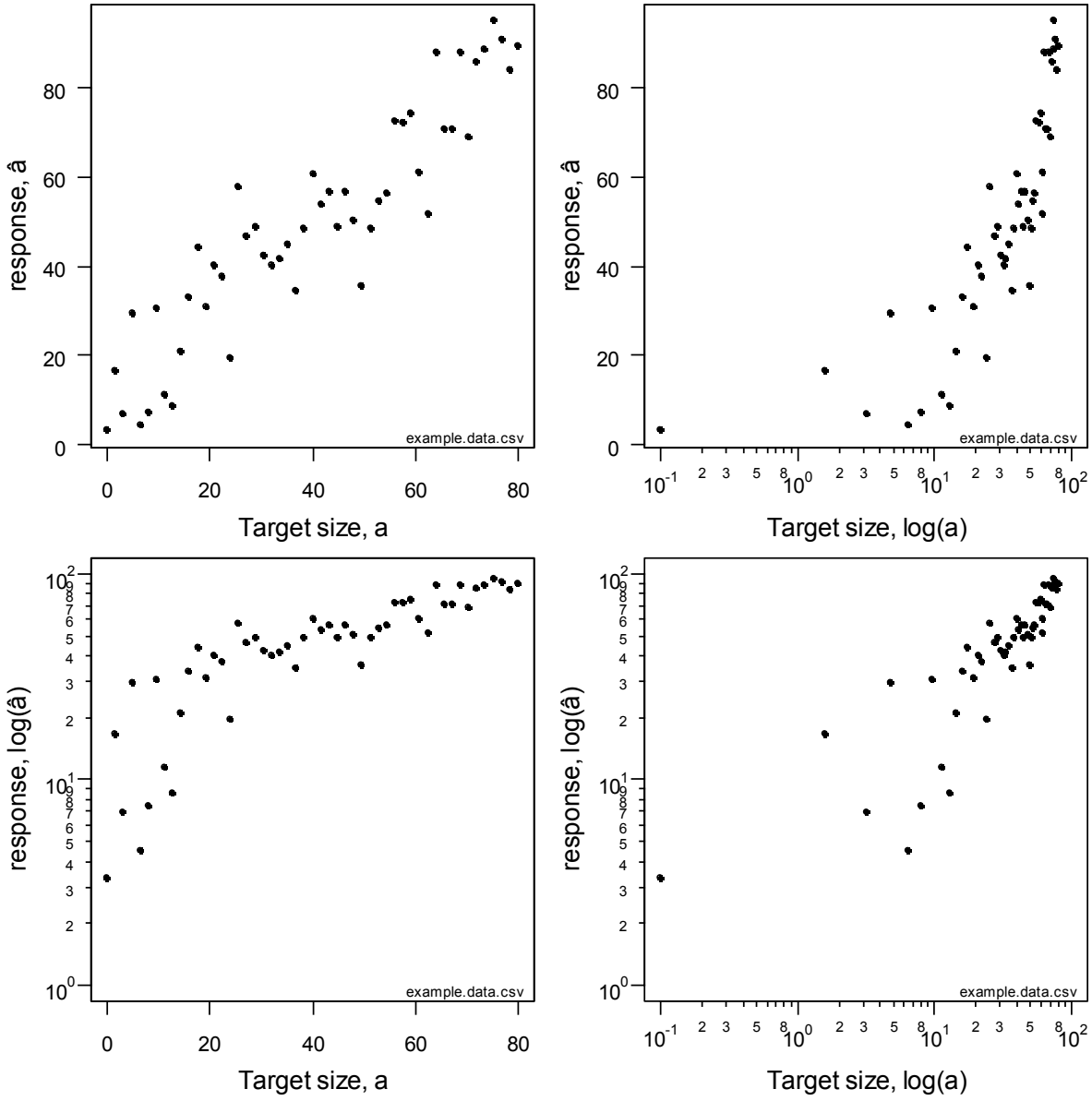

Figure 10
Choosing the right transformation depends on the particular data.
Log-log is <u>not</u> always the best transformation.

The relationship of POD and size, for a given $\hat{a}_0$, can be shown on the more familiar POD vs size curve as in Figure 11.

We describe $\hat{a}$ *vs* a with a linear model (regression) in (Eq. 5):

$$\hat{a} = \beta_0 + \beta_1 a \qquad \text{(Eq. 5)}$$

where β_0, β_1 are the intercept and slope. The individual observations, $\hat{a}_i$, are encumbered with an uncertainty, ε_i, which is the random component of the NDE signal and $\varepsilon \sim N(0,\tau)$. This is read "$\varepsilon$ has a normal distribution with zero mean and standard deviation, τ". We use τ rather than the more familiar σ to avoid confusion with the POD(a) model parameters.

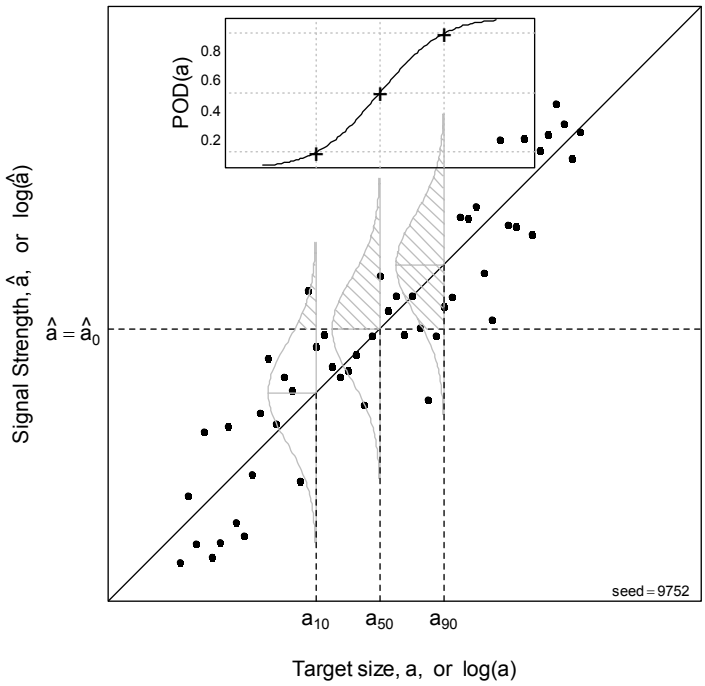

Figure 11
The POD vs a relationship can be plotted as the fraction of the scatter density, POD, that is above the decision criterion (the shaded areas), for a given size, a.

For many datasets, it is not possible to use ordinary least-squares regression (OLS) to estimate the parameters of the parametric model for the *â vs a* data pairs because the *â* data are censored.

3.3.2 The "*â vs a*" plot with censoring

An observation is *censored* if it is only known to be in some range but its exact value is unknown. Left-censored observations are in the range [$-\infty \le y \le y_{censor}$), right-censored ($y_{censor} \le y \le \infty$] and interval-censored ($y_{left\text{-}censor} \le y \le y_{right\text{-}censor}$).

With many kinds of inspections the signal response is censored. For example the observed amplitude cannot be greater than 100% of screen height. The lower value is often chosen to avoid background noise. So for some very small cracks the signal will be obscured by the noise and the response is left-censored. For very large cracks the true signal response may be unknown but greater than 100% of the maximum setting, and thus right-censored. Sometimes the signal is no longer linear with size at very long cracks and the response may be treated as being censored at a value where deviation from linearity is no longer acceptable. In all of these situations using ordinary least squares regression would produce very skewed results as is illustrated in Figure 12. This is a result of incorrectly substituting the censoring value for the unobserved true response, rather than treating the observations correctly as being censored.

The censored regression, which is correct, is presented in Figure 13. The censored data were not thrown away or ignored. Rather their likelihood function was re-defined to account for their unknown value being anywhere below the censoring line.

It is no longer possible to delay a discussion of probability and likelihood and how we use likelihood to estimate the parameters of the $\hat{a}$ *vs* a model here, and later the parameters of the hit/miss POD model.

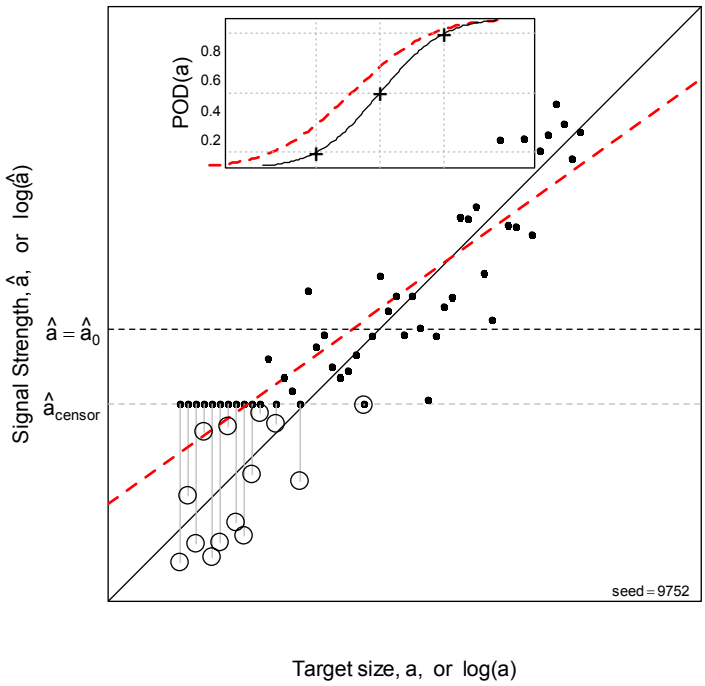

Figure 12
OLS Regression produces bogus results (red dashed line) when the data are censored.

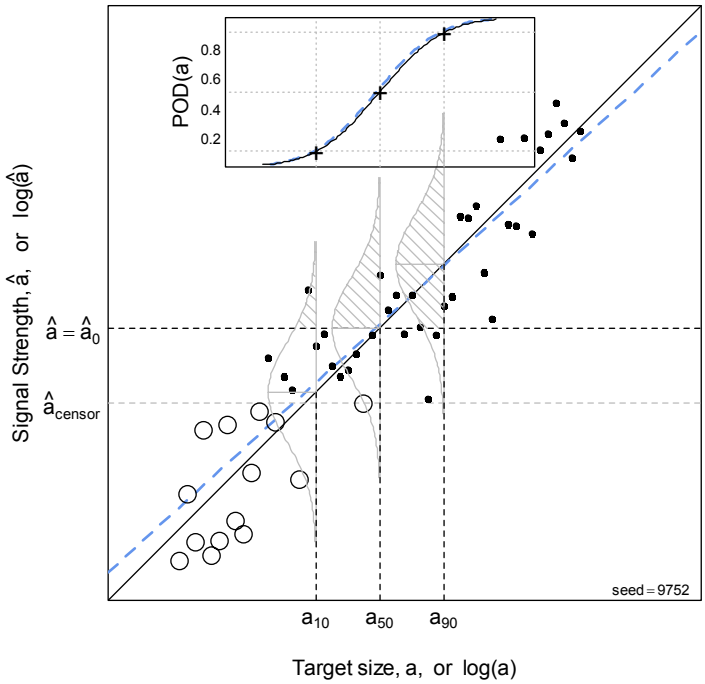

Figure 13
Censored regression (blue dashed line) correctly accounts for observations with actual responses obscured by background noise.

3.3.3 Probability and Likelihood

Probability and likelihood are mirror images of one another. Consider the familiar normal distribution:

$$p(x \mid \mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{(Eq. 6)}$$

This can be interpreted as the probability density of an unknown, x, given the known values for the mean and standard deviation, parameters μ and σ. By comparison, if μ and σ are not known, but we have a collection of known observations, $\mathbf{X}=x_1, x_2, \ldots x_n$, (Eq. 6) becomes:

$$L(\mu,\sigma \mid \mathbf{X}) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\mathbf{X}-\mu}{\sigma}\right)^2} \qquad \text{(Eq. 7)}$$

This is the *likelihood* of μ and σ, given the observations, $\mathbf{X}$. The mathematical formulations are identical. The only difference is in what is known and what is not known.

Formally we can say that given a vector of continuous random variables, $\mathbf{X}=x_1, x_2, \ldots x_n$, that depend on model parameters, $\boldsymbol{\theta}=\theta_1,\theta_2,\ldots,\theta_k$, then $f(\mathbf{X}|\boldsymbol{\theta})$ is the *probability density function* of $\mathbf{X}$ given $\boldsymbol{\theta}$, and $L(\boldsymbol{\theta}|\mathbf{X})$ is the *likelihood* of $\boldsymbol{\theta}$ given $\mathbf{X}$. The functional form of f and L is the same.

Less formally *likelihood is the probability of the observed data*. It is the probability that the experiment turned out the way that it did for a given set of parameters $\boldsymbol{\theta}$. Numerically, it is the ordinate of the probability density evaluated at a point x, for uncensored observations.

3.3.4 Likelihood function for censored data

As stated earlier, an observation is *censored* if it is only known to be in some range but its exact value is unknown. Left-censored observations are in the range $[-\infty \le y \le y_{censor})$, right-censored $(y_{censor} \le y \le \infty]$ and interval-censored $(y_{left\text{-}censor} \le y \le y_{right\text{-}censor})$.

Censored observations could be anywhere within the censored region, so rather than using the ordinate at a point to define the likelihood, we define the likelihood for a censored observation as being *all* ordinates in that region, *i.e.* the integral of the function over the censored region, (Eq. 8) and (Eq. 9).

For left-censored points:
$$L(\mu,\sigma \mid x) = \int_{-\infty}^{x=\text{left censor}} \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \text{(Eq. 8)}$$

For right-censored points:
$$L(\mu,\sigma \mid x) = \int_{x=\text{right censor}}^{\infty} \frac{1}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \qquad \text{(Eq. 9)}$$

For the parametric model describing how $\hat{a}$ changes with a we have (Eq. 1), which locates the mean of the probability distribution in Figure 14, so in (Eq. 8) and (Eq. 9), $\mu = \beta_0+\beta_1 \cdot a$.

If likelihood were looked at as a probability, the likelihood of a censored observation would be the probability of the observation being in the censored (shaded) region.

Comparing (Eq. 7) with (Eq. 8) and (Eq. 9) we see that the constant $1/\sqrt{2\pi}$ has disappeared. That constant is to make the integral of (Eq. 6), which is a probability, equal to one (or 100%). This brings up an important difference between probabilities and likelihoods: probabilities must sum to one, whereas likelihoods do not have that restriction. For that reason the value of the likelihood by itself is meaningless. It only gains meaning when it is compared to another likelihood.

Figure 14 compares the various definitions of likelihood and the regions where they apply.

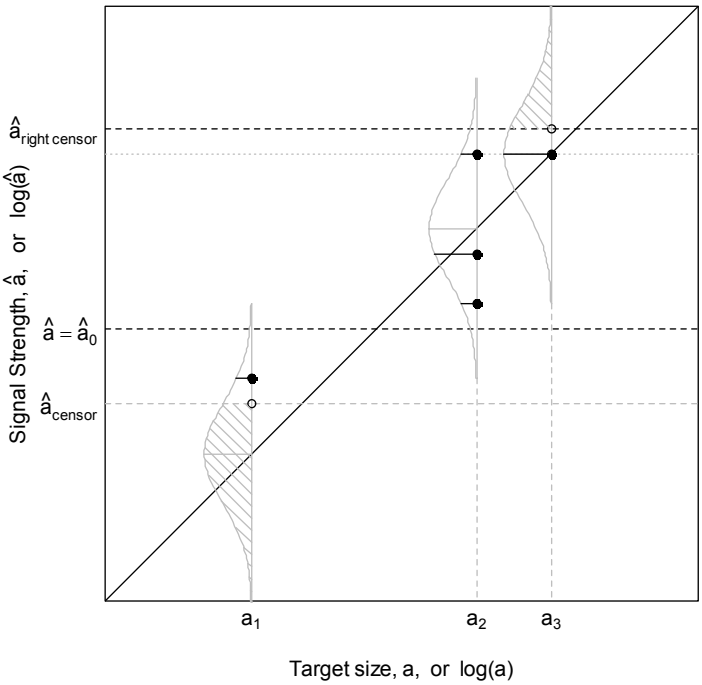

Figure 14
The likelihood of an uncensored observation is the ordinate of the likelihood function; for censored observations it is the integral of the likelihood function over the censored interval.

Figure 14 requires some discussion: at a_1 there are two observations, from two different test specimens. One is uncensored (black dot), and the other is left-censored (open dot). The open dot is plotted at the censoring value only for convenience since its true value is unknown, other than being below the censoring value somewhere. The likelihood for the censored observation (the shaded region) is greater than the likelihood for the known observation at the same size, a_1, because the expected response at a_1 (which is given by the line relating a with $\hat{a}$) is much lower than the censoring value. In other words, for such a small a_1 we would expect the response to be below the censoring value, and indeed it has a greater likelihood than the uncensored observation at the same size.

There are three observations, from three different test specimens, at a_2. None is censored, so the likelihood for each is shown as the ordinate at each $\hat{a}$ value.

At a_3, we have two observations, one censored, the other not. Their likelihoods are the shaded region and the ordinate, respectively. Notice that there is another observation with the same response, $\hat{a}$, at size = a_2. They do not have the same likelihood, even having the

same response, because their expected responses are a function of size (the line) and they are responses from different size targets.

3.3.5 Maximum Likelihood Parameter Estimates

Consider Figure 14 again. The likelihood that the black line is in the right place is the product of all the individual likelihoods that contributed to the overall likelihood that $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\tau}$ are the best estimates for β_0, β_1 and τ. Clearly, the red line is a poor description of the data. The parameter values that maximize the total likelihood are called, not surprisingly, *maximum likelihood estimates.*

When there is no censoring, the maximum likelihood parameter estimates are *exactly* the values determined by ordinary least-squares regression[11], which is comforting since the theory does not conflict with that well-respected method used by engineers for 200 years.

In practice dealing with the product of likelihoods is extremely tedious. It is much simpler to use the logarithm of the likelihood, since the maximum log(likelihood) will occur at the same parameter values as for the likelihood itself. Computing with a sum of logarithms is much easier. There is another reason to use log(likelihood), which is also written for brevity as log-likelihood. Recall that the likelihood only has meaning when compared with another likelihood computed using competing parameter estimates. The most important comparison is the likelihood ratio, the ratio of a likelihood to the maximum likelihood. The log of the likelihood ratio has very useful statistical properties: the log likelihood ratio has an asymptotic *chi-square* (χ^2_{df}) distribution, which can be used to construct confidence bounds on MLE parameter estimates, and by extension, bounds on the POD curve itself. We will defer that discussion until a later section (3.4.4), where it is used with hit/miss data.

3.3.6 The POD(a) relationship

We determine POD(a) from Figure 9, (Eq. 5) and (Eq. 6), which leads to (Eq. 10):

$$POD(a_i) = P(\hat{a}_i > \hat{a}_0) = 1 - \Phi_{norm}\left(\frac{\hat{a}_0 - (\beta_0 + \beta_1 a_i)}{\tau}\right) \qquad \text{(Eq. 10)}$$

where $\Phi_{norm}(z)$ is the standard normal cumulative density function.

From (Eq. 10) we can construct the familiar POD(a) curve, Figure 15. The POD(a) model parameters, μ, σ, are related to the $\hat{a}$ *vs* a model parameters by

$$\mu = \left(\frac{a_0 - \beta_0}{\beta_1}\right) = a_{50} \qquad \text{(location parameter)}$$
$$\sigma = \frac{\tau}{\beta_1} \qquad \text{(shape parameter)} \qquad \text{(Eq. 11)}$$

[11] OLS estimates are *exactly* ML estimates. Not approximately equal; exactly equal. Demonstrating that the maximum likelihood estimate of the sample mean is the familiar least squares estimate, $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, is beyond the scope of this report.

so that

$$POD(a) = \Phi_{norm}\left((a-\mu)/\sigma\right) = \Phi_{norm}(z) \qquad \text{(Eq. 12)}$$

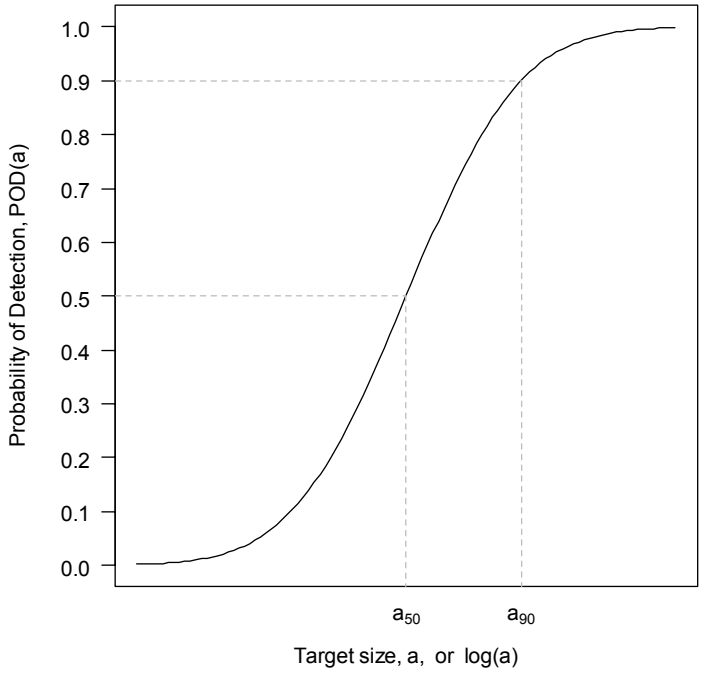

Figure 15
POD(a) resulting from the $\hat{a}$ vs a relationship in Figure 1.

3.3.7 Confidence bounds on POD(a)

The confidence bounds on the POD(a) relationship are not computed from the confidence bounds on the $\hat{a}$ *vs* a plot, which is why they are not shown. They are superfluous and are not used as they lead to confusion.

The confidence bounds on the POD(a) relationship are determined from the joint distribution of the POD vs a model parameters, μ, σ, which are unknown and must be computed from the joint distribution of the $\hat{a}$ *vs* a linear model parameters β_0, β_1 and the observed scatter, τ. This is accomplished using the Delta method, one of the most powerful and most commonly used transformations in applied statistics. A detailed discussion of the Delta method is beyond the scope of this report, and we refer the interested reader to MIL-HDBK-1823A (2009) for a complete explanation.

After considerable statistical manipulation, the lower bound on the POD(a) curve is constructed using the Wald Method[12], which although perhaps not recognized by name, should be familiar from any statistics class:

$$y_\alpha = \mu_y - 1.645\sigma_y \qquad \text{for } \alpha = 0.95 \qquad \text{(Eq. 13)}$$

[12] Abraham Wald (1902-1950) was a Hungarian statistician who reasoned that if x is some number of standard deviations from the mean, based on knowing μ, then, knowing x, the unknown μ must be within a similar neighbourhood of x.

Where μ_y is the mean of variable y, σ_y its standard deviation, and -1.645 a numerical constant corresponding to the α=0.95 quantile of the normal distribution. For two dimensions (μ, σ) the calculation is analogous. For a given value of POD, a_{POD}:

$$a_{POD} = \mu_{POD} + z_{POD}\,\sigma_{a_{POD}} \qquad \text{(Eq. 14)}$$

With z_{POD} being a constant dependent on the required level of confidence (e.g. -1.645 for 95% or -2.326 for 99% confidence) and $\sigma_{a_{POD}}$ the standard deviation of a_{POD}. The latter quantity is derived from the covariance matrix for $\hat{\mu},\hat{\sigma}$, and is equal to

$$\sigma_{a_{POD}} = \left(\sigma_{\mu}^2 + 2z_{POD}\times\sigma_{\mu\sigma} + z_{POD}^2\times\sigma_{\sigma}^2\right)^{1/2} \qquad \text{(Eq. 15)}$$

It is worth it to digress briefly here to explain that in N dimensions the variance becomes a $N\times N$ covariance matrix which includes the individual variances on the diagonal, and their covariances (i.e. how they vary with each other) on the off-diagonal. The covariance matrix is symmetrical because how X changes with Y is the same as how Y changes with X. The variance of X is defined as the average of the squared differences between X_i and μ_X:

$$\sigma_X^2 \equiv \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2 \qquad \text{(Eq. 16)}$$

The covariance of X and Y has an analogous definition:

$$\sigma_{XY} \equiv \frac{1}{n}\sum(X - \mu_X)(Y - \mu_Y) \qquad \text{(Eq. 17)}$$

Figure 16 shows a POD(a) curve and the 95% lower bound obtained with the method described above.

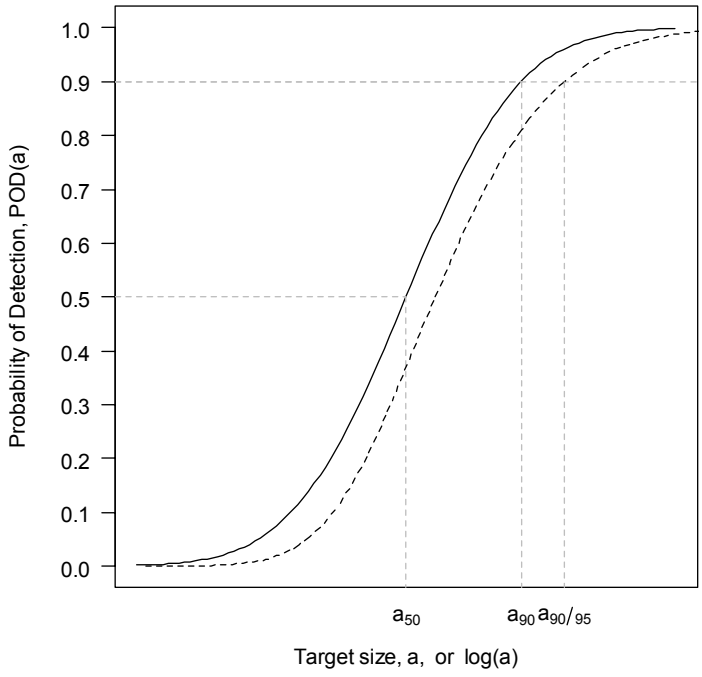

Figure 16
POD(a) curve with 95% lower bound

3.3.8 Noise

Noise is defined as signal responses that contain no useful target characterization information. Thus noise appears on a $\hat{a}$ *vs* a plot as responses, $\hat{a}$, that are random with respect to size, a, Figure 17. Noise is ubiquitous and it must therefore be considered in any POD analysis. To do otherwise would be irresponsible.

Looking again at Figure 9 it would be tempting to lower the decision criterion, $\hat{a}_0$, and thereby increase the POD at every flaw size. But the random component of the signal response, $\hat{a}$, is not the only random component that influences the effectiveness of an inspection. The other is random background noise, and every POD study should quantify the behaviour of noise as an integral part of the overall data-gathering experiment.

Figure 17 reconsiders figure 2 in the presence of noise in a hypothetical example. It is immediately obvious that the existing probability of False Positive would increase from its current value, 3%, if $\hat{a}_0$ were made smaller. This value (3%) represents the percentage of the area under the distribution for noise that exceeds the decision threshold. The noise distribution is obtained from the noise measurements using censored regression since many of the noise measurements will be left censored. 3% in the hypothetical example is likely to be too high already. The example shows that a modest lowering of $\hat{a}_0$ from would increase the probability of a false positive by 10 times.

It is an unfortunate fact that the engineer responsible for setting the value for $\hat{a}_0$ and the engineer responsible for conducting the field inspections may not communicate effectively.

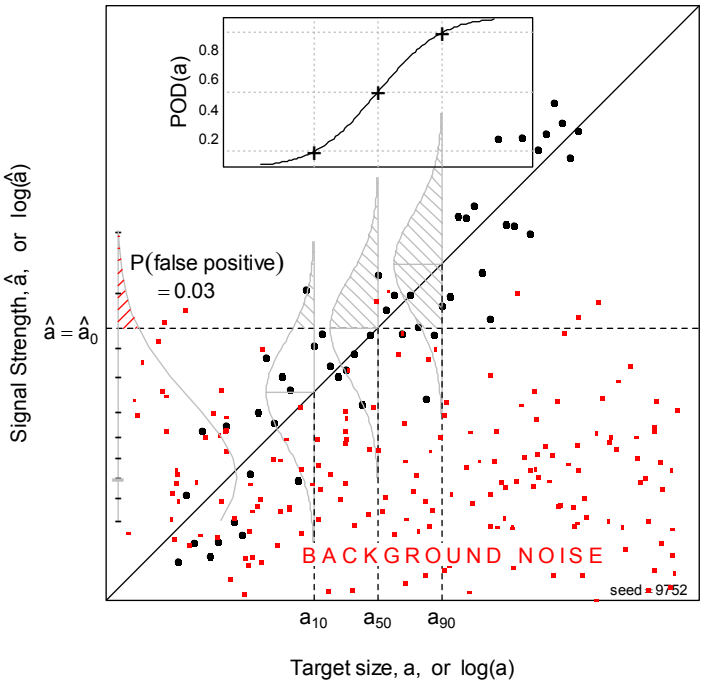

Figure 17
Choosing a smaller decision criterion increases the Probability of False Positive (PFP).

This is what can happen. To meet his fracture mechanics life requirement, the structural engineer needs the inspection to find a small crack of size, a_0, and he sets the inspection

threshold criterion, $\hat{a}_0$, sufficiently low that he achieves the required POD at that size. (In the aircraft and flight propulsion industries a_0 is often chosen to be $a_{90/95}$). What the structural engineer does not know is that threshold value is in the middle of the inspection background noise. The service engineer then sets up the inspection and finds an untenable false-positive rate, perhaps 50%. To meet his throughput requirements, the field engineer raises the inspection threshold to be above the noise so that he will not reject every other part. The actual POD is much lower than the structural engineer thinks it is, and this is often discovered only after an unexpected service failure.

All this trouble can be avoided by insisting that a noise analysis be presented with any POD analysis. With Figure 17 as a guide, it is a straight-forward task to plot PFP vs $\hat{a}_0$ to see the trade-off with improved POD. The mh1823POD[13] software can support this task. A more complete discussion of noise with respect to $\hat{a}$ *vs* a analysis can be found in MIL-HDBK-1823A (2009).

3.3.8.1 How to collect Noise Data

Collecting noise data as part of the NDE experiment adds very little to the costs or time required, and provides immeasurable benefits (avoiding a premature failure, for instance).
Since noise is a signal with no information about the target, it is easy to collect noise data: perform the inspection protocol in a structurally similar, *unflawed* region of the test piece[14] at the same time, and with the same test setup, as the designed test.

A rule of thumb is to collect *at least* three times noise information as target information. In some inspections the background noise is low and thus more difficult to quantify. Having a great deal of noise data mitigates this problem. Since it is difficult (or impossible) to duplicate the initial test environment later, noise must be collected concurrently with the target response data.

Collecting noise data for hit/miss inspections (like Fluorescent Penetrant) is only slightly more complicated. In this situation, an inspector must specify the location (within some tolerance) as well as registering a “hit.” To collect noise data, an area of the test piece is designated for inspection but contains no target.

The inspector should be unaware of both the number and location of the targets in any inspection trial.

3.3.8.2 Exogenous Noise

Exogenous noise associated with inspecting new parts is often rather different from noise encountered during inspection of parts returning from field service. New parts are pristine. Used parts are often dirty, corroded, scratched, or otherwise disfigured, and the resulting background noise has a deleterious effect on the probability of false positive that must be accounted for. Thus, whenever there is a radical change in the inspection milieu, it is prudent to collect new noise data to represent the new situation. This is often inconvenient, but it is seldom costly, whereas ignoring the problem is often both.

---

[13] See discussion on page 1 (including footnote 3).
[14] Test pieces must contain targets with *known* characteristics. That is why using “field-finds” will result in bogus POD curves. See section 3.5.3. Since the location of the target is known, it is easy to inspect where the target is not located to gather noise data.

3.3.9 Analyzing Noise

â vs a noise

To estimate the probability that noise will be greater than some a_0 value it is necessary to determine the characteristics of the noise distribution. Noise data is not always normal. Furthermore, with $\hat{a}$ data, much of the noise signal will be below a_0 and thus censored. The estimation of the parameters of the noise distribution requires a censored regression, with the censored responses regressed against a column vector of ones. The easiest way to accomplish this is to use existing software, such as mh1823POD[15], which is free. With the noise model parameter estimates, the $P(\hat{a}_{noise} > a_0)$ can be directly calculated.

Hit/Miss noise

"Hits" associated with no target are noise and the probability of false positive can be estimated using the binomial distribution. An approximation, based on median rank, is

$$PFP = (n + 0.5)/(N + 1) \qquad \text{(Eq. 18)}$$

where n=0,1,2, ... N, is the number of incorrect "hits" and N is the total number of inspection opportunities[16]. For example if 100 non-flawed inspection areas were inspected and 5 spurious "hits" recorded, the probability of false positive would be estimated to be approximately 0.045. If zero false indications were recorded in 100 opportunities the PFP would be estimated to be about 0.005.

3.4 NDE data with binary signal response: hit/miss data

Binary responses carry no information except found or not-found. In this section we review the method developed to analysis NDE data in the form of hit/miss data.

3.4.1 The Parametric POD Model - Maximum likelihood analysis for hit/miss data

It might appear that the analysis of *â vs a* data, where the response is continuous, is quite different from the analysis of hit/miss data, where the response is binary, but the two are very similar. Continuous response data can be modelled using the familiar ordinary least-squares (OLS) regression, although a generalization of the definition of likelihood is necessary if any of the observations are censored. The analysis of *â vs a* data was discussed in Section 3.3.

Binary response data can also described with a regression model, a generalization of the linear model. For ordinary regression we say that $y=f(X)$. To generalize, we need some function of y that can *link* (through the probability of the outcome) the binary response to the function of x, $g(y)=f(X)$. This generalization is called, not surprisingly, a Generalized Linear Model (GLM). Obviously, for ordinary regression, $g(y)=y$.

3.4.1.1 The logit link

The most useful link function is the logistic function:

$$f(X) = g(y) = \log\left(p/(1-p)\right) \qquad \text{(Eq. 19)}$$

[15] See discussion on page 1 (including footnote 3).

[16] (Eq. 18) estimates the median rank of the response. A slightly different equation describing the mean ranks is used for plotting binomial probabilities (e.g. Figure 8) [Meeker and Escobar (1998)]. Note that the POD plotting positions are not used in any numerical computation of POD and only serve to help illuminate the data.

The response $g(y)$ is now continuous [$-\infty \le g(y) \le \infty$]. Ordinary regression methods are still not appropriate because ordinary least-squares (OLS) regression requires the variance to be constant, and that is not true for binary data since the variance of the response is equal to $p(1-p)$.

Because the logistic link (also called *logit* or *log-odds*) is the most common and most useful, it deserves some further discussion. The "odds" are defined as the probability of occurrence of a binary outcome divided by the probability of non-occurrence:

$$odds \equiv \frac{p}{1-p} \qquad \text{(Eq. 20)}$$

The log of the odds (hence *log-odds*) is the logit:

$$\log(odds) \equiv \log\left(\frac{p}{1-p}\right) \qquad \text{(Eq. 21)}$$

The log-odds POD model is then

$$\log\left(\frac{POD(a)}{1 - POD(a)}\right) = \beta_0 + \beta_1\, a \qquad \text{(Eq. 22)}$$

or also, commonly:

$$\log\left(\frac{POD(a)}{1 - POD(a)}\right) = \beta_0 + \beta_1 \log(a) \qquad \text{(Eq. 23)}$$

Whether or not to transform *size* logarithmically depends entirely on the data being modelled, so no universal transformation is recommended, even though much of the older NDE literature *assumes* log(*size*), with no justification. To avoid the appearance of perpetuating this practice we will use $h(a)$ to mean either a, or $\log(a)$, depending on the data. Solving (Eq. 23) for POD(a) produces:

$$POD(a) = f(a, \boldsymbol{\theta}) = \frac{e^{\beta_0 + \beta_1 \cdot h(a)}}{1 + e^{\beta_0 + \beta_1 \cdot h(a)}} \qquad \text{(Eq. 24)}$$

Where $\boldsymbol{\theta} = (\beta_0, \beta_1)^T$. It should be obvious from (Eq. 24) that this formulation of $POD(a)$ is easily evaluated in closed-form, which is historically interesting since computing power was not always so readily available and inexpensive as it is today.

Unfortunately the parameters, $(\beta_0, \beta_1)^T$, have no obvious physical interpretation and so it is convenient to re-parameterize as (Eq. 25):

$$POD(a) = f(a, \boldsymbol{\theta}) = \Phi_{link}^{-1}\left(\frac{x-\mu}{\sigma}\right) \qquad \text{(Eq. 25)}$$

where for the logit link:

$$\Phi_{link}\left(\frac{x-\mu}{\sigma}\right) = \log\left(\frac{POD(a)}{1-POD(a)}\right) \qquad \text{(Eq. 26)}$$

Now:

$$f_1(x \mid \beta_0, \beta_1) = \log\left(\frac{POD(a)}{1 - POD(a)}\right) = f_2(a \mid z), \text{ where } z = \left(\frac{x-\mu}{\sigma}\right) \quad \text{(Eq. 27)}$$

Since $f_1(x \mid \beta_0, \beta_1) = f_2(a \mid z)$ then:

$$\beta_0 + \beta_1 x = \frac{x-\mu}{\sigma} \quad \text{(Eq. 28)}$$

Solving for (μ, σ) in terms of $(\beta_0, \beta_1)^T$, shows that

$$\frac{x-\mu}{\sigma} = \left(\frac{1}{\sigma}\right)x + \left(\frac{-\mu}{\sigma}\right) \quad \text{(Eq. 29)}$$

which means that

$$\beta_1 = \frac{1}{\sigma} \text{ and } \beta_0 = \frac{-\mu}{\sigma} \quad \text{(Eq. 30)}$$

so that

$$\sigma = \frac{1}{\beta_1} \text{ and } \mu = -\beta_0\sigma \quad \text{(Eq. 31)}$$

μ and σ have useful physical interpretations. μ is the size, or log(size), at which POD = 0.5. σ is the inverse of the GLM regression slope. This is illustrated in Figure 18.

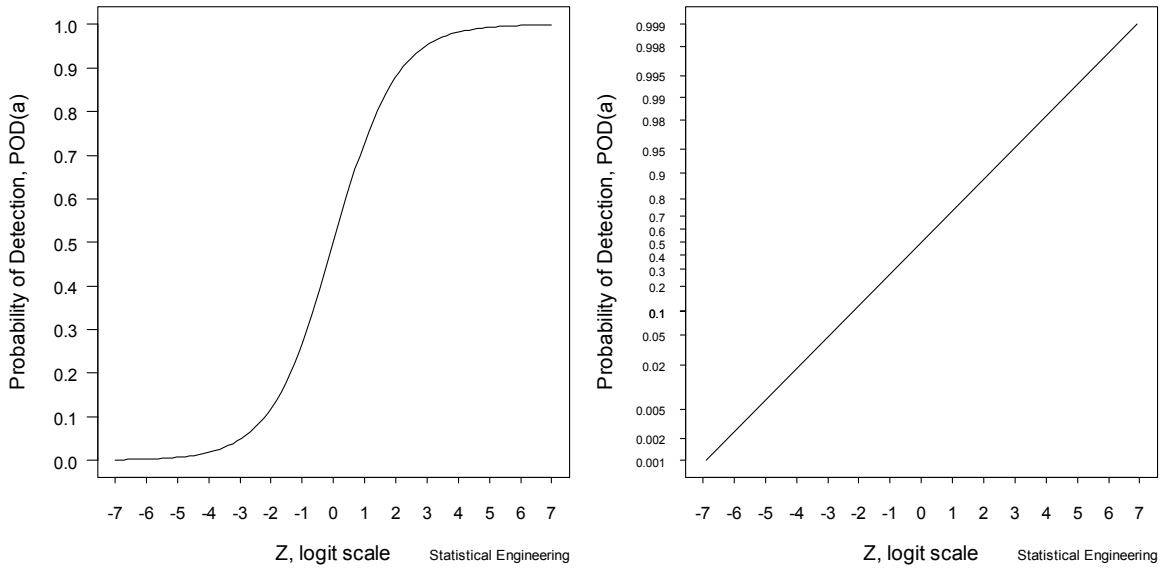

Figure 18
The "S" shaped POD(a) curve plots as a straight line on the logit grid.
The location parameter μ corresponds to the x value at POD=0.5.
The scale parameter σ on the left is 1/slope on the right.

3.4.1.2 The probit link:

A similar, but distinctly different, link function is the *probit* or Gaussian.

$$\Phi_{probit}\left(\frac{x-\mu}{\sigma}\right)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad \text{(Eq. 32)}$$

It must be remembered that while $\Phi(z)$ has the functional form of the standard normal equation, it is *not* a probability density. It does *not* indicate the probability that a crack will reside in some size range, $(x-\delta x)\le x\le(x+\delta x)$, but it is used to describe the POD(a) curve because the function has a useful "S" shape. The probit model does not have a closed-form for $POD(a|z)$, but this is seldom problematic since almost all numerical analysis programs have the function built-in.

3.4.1.3 Comparison of logit and probit link functions

These two link functions, logit and probit, while similar, have significant differences, especially in their extremes. Consider Figure 19, that demonstrates that the probit and logit links are very similar for $(0.1\le p\le 0.9)$ but they differ considerably in the extremes. The probit link is more computationally sensitive to "errors" in the tails, making it more vulnerable to lack-of-fit. The two plots of Figure 19 represent the same data, but the plot on the left uses a Cartesian y-axis, whereas the plot on the right uses a probability y-axis[17]. In Figure 19, a $\sqrt{3}$ probit is also plotted. This makes the logit look like the probit over the central range of probabilities for the purposes of comparison.

It might be argued that if there is not much point in discussing differences, if a special grid is needed to see them. In reality, it is important because the probit function is more sensitive to "improbable" outcomes. For example, if the true probability of detection at size a = a_0 is 90%, there is a 10% probability of a miss, due to chance alone. The contribution to the likelihood of a hit is 0.9, and the contribution of a miss at that crack size is 0.1, so an algorithm to maximize the likelihood would move away from p=90% to a lower value, to increase from 0.1 the likelihood contribution caused by that (improbable) miss.

That is the situation if the model is log-logistic (log-odds) and p(z)=0.9 at $a=a_0$. If the model is the probit (Gaussian) then p(z) = 0.986, not 0.9. The contribution to the likelihood of that miss at $a=a_0$ is much smaller, now 0.014, and the model will need to change more to increase it, making the probit model more vulnerable to randomness. So as a matter of experience, the logit link is used unless there are compelling reasons (physical or statistical) for using the probit link.

Figure 19 also illustrates why using a Cartesian y-axis for probability can obscure important aspects of the POD model. For that reason conventional Cartesian y-axis POD curves (left plot) should be used for presentation only; computational exposition should use the probability y-axis based on the appropriate link function (right plot).

---

[17] Just like a log-axis is an axis where the plotting position is proportional to the logarithm, a probability axis plots probabilities. In Figure 19 the y-axis of the plot on the right is based on the logit distribution, but one can make a probability axis for any density. For instance, a normal probability axis would transform (and plot) 0.1%, 1%, 10% and 50% at values y (= z) = -3.090232 , -2.326348, -1.281552, and 0, respectively, since that is how many standard deviations away from the mean those percentages are.

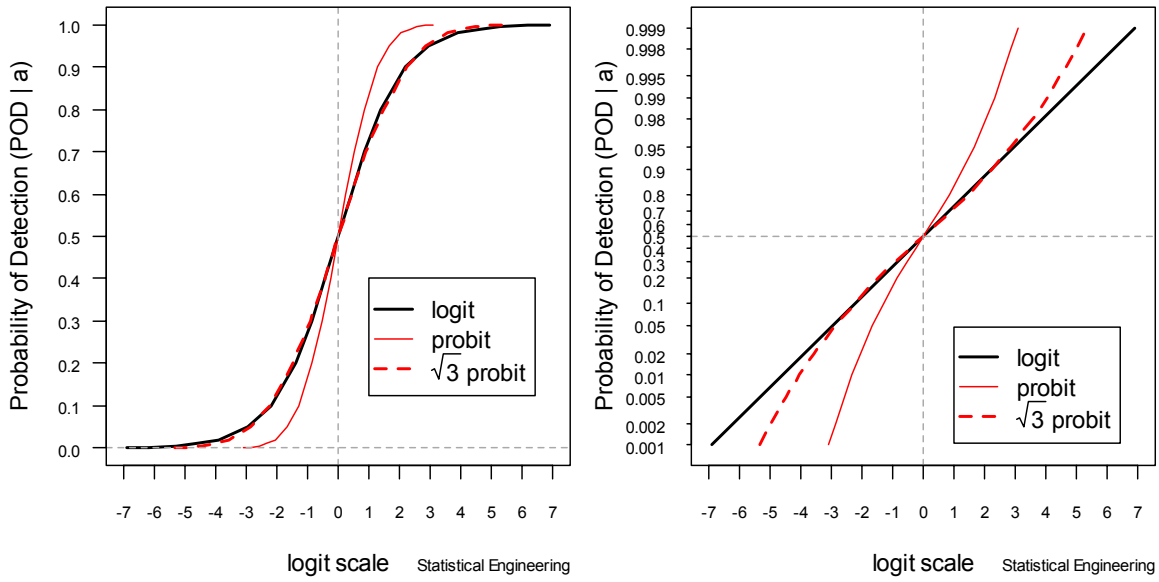

Figure 19
The logit and probit models are similar for $(0.1 \le p \le 0.9)$ but they differ considerably in the extremes.

3.4.1.4 Other link functions

Both the logit and probit links are symmetrical, i.e. $f(Z)=f(-Z)$. There are two other link functions commonly found in the statistics literature, the log-log function $g(y)= -log(-log(p))$, which is skewed right, and the complementary log-log function, $g(y)= log(-log(1-p))$, which is skewed left. Data which is skewed right can usually be modelled using a symmetrical link after taking the log, which is why log(size) is often appropriate.

3.4.1.5 Choosing the appropriate link function

Except in special cases (discussed later), where min(POD)>0 and/or max(POD)<1, either the probit or logit, in combination with either x or log(x) will describe nearly all POD vs size relationships arising from binary data. In practice all four combinations are used, with the decision then based on which does the best job based on the behaviours of the *deviance*, which is the binary analogue of the residual summed square error for continuous data. Smaller deviance indicates a better model fit. The deviance is computed from the log likelihood ratio, as was discussed above.

3.4.2 Joint, Marginal and Conditional probability

Before proceeding further, we need to briefly review the concepts of joint, marginal, and conditional probabilities.

*A **joint probability*** is the probability of two (or more) events happening in conjunction. Consider first the idea of a probability density (or distribution): $f(x|\theta)$ where f is the probability density of x, given the distribution parameters, θ. For a normal distribution, $\theta=(\mu, \sigma)^T$ where μ is the mean and σ is the standard deviation. This is also called a probability density function (pdf). The area (integral) under the pdf curve between specified values of x is equal to the

probability of the random variable occurring in that interval, and corresponds to the cumulative distribution function (cdf), $F(x|\theta)$.

A joint probability density with two or more variables is called a multivariate distribution. It is often summarized[18] by a vector of parameters, which may or may not be sufficient to characterize the distribution completely. For example, the normal is summarized (sufficiently) by a mean vector and covariance matrix.

*A **marginal probability*** is the unconditional probability of one of the events (say x) happening, regardless of the other (y). It is expressed by $f(x|\theta)$ where f is the probability density of x, for all possible values of y, given the distribution parameters, θ. The marginal probability is determined from the joint distribution of x and y by integrating over all values of y (this is called "integrating out" the variable y). Figure 20 illustrates joint, marginal, and conditional probability relationships.

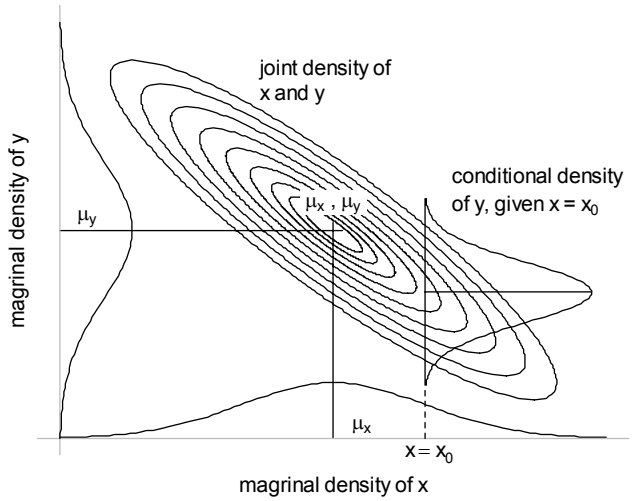

Figure 20
How the Joint, Marginal, and Conditional distributions are related.

Conditional probability: $f(x \mid y,\theta)$ where f is the probability of x by itself, given a specific value of the variable y, and the distribution parameters, θ. If x and y represent events A and B, then $P(A|B) = n_{AB}/n_B$, where n_{AB} is the number of times both A and B occur, and n_B is the number of times B occurs. $P(A|B)=P(AB)/P(B)$, since $P(AB)=n_{AB}/N$ and $P(B)=n_B/N$ so that $P(A|B)=(n_{AB}/N)/(n_B/N)=n_{AB}/n_B$ (N being the size of the overall population).

Note that, in general, the conditional probability of A given B is *not* the same as B given A. For example, the probability of having 4 legs, given that the animal is a dog, is very close to 1 (allowing for disfiguring accidents), $P(4\text{ legs} \mid \text{dog}) = 0.999$ (hypothetically), but the probability that an animal is a dog, given that it has 4 legs, is much less, say $P(\text{dog} \mid 4\text{ legs}) = 0.01$.

[18] "Summarize" has a special meaning in statistics. Statisticians have a concept of "sufficiency" that implies that all the necessary information is known about a distribution. For example, sufficient statistics for the normal distribution are the mean and standard deviation, as they fully define the distribution. A distribution is thus "summarized" by a set of parameters if these represent sufficient statistics.

3.4.3 Estimating the GLM parameters

We recommend using existing statistical code for executing routine statistical procedures. We also strongly recommend understanding what the code is doing, and for this reason we describe here the workings of the maximum likelihood criterion for estimating model parameters.

In the following, we present an example in which we apply the statistical concepts studied to hit/miss results from a real inspection. We assume an underlying model that relates POD to target size, the logit model.

$$POD(a) = f(a, \boldsymbol{\theta}) = \Phi_{link}^{-1}\left(\frac{x-\mu}{\sigma}\right) \qquad \text{(Eq. 33)}$$

The likelihood associated with a target of size x is $POD(x)$, and the likelihood of a miss is 1-$POD(x)$. We want values for μ and σ that will maximize the sum of the logs of the individual likelihoods. Knowing that $\hat{\mu}$ and $\hat{\sigma}$ have a joint density, as in Figure 20, we could search the density surface to locate the maximum but for hit/miss data the computational method used in practice, iteratively re-weighted least-squares, is more robust than a likelihood surface.

Just as ordinary least-squares parameter estimates *are* maximum likelihood estimates (when the errors are Gaussian, with no censoring), parameter estimates based on iteratively reweighted least-squares (IRLS) for generalized linear models are also equal to ML estimates (within numerical methods tolerance). The method is to perform an ordinary least-squares (OLS) regression on the transformed model, with the observations weighted by the model variances at the x values, $p(1\text{-}p)$. The model is then re-evaluated using the new parameter estimates, and new values for p, and thus new weights. The process is continued until the change in weights is zero (within tolerance). This method is faster and more computationally stable than searching for a maximum on the log-likelihood surface, and of course, produces identical results, when the search method can find a solution.

Table 2 presents hit/miss results from a real inspection. The same data are found as EXAMPLE 3 hm.csv in MIL-HDBK-1823A (2009). There are 134 observations but the largest ID number is 161.

The data, and the MLE fit, are plotted in Figure 21. Figure 22 plots the log-likelihood ratio surface showing how the likelihood gets smaller with distance from the most likely values for μ and σ (0.1156, 0.0251), given the hit/miss data (the negative of the log-likelihood ratio is plotted, to avoid negative signs). The MLEs are identified by the "**+**" in the centre of the figure. The likelihood ratio at that location is one. These values for μ and σ, when plotted on the familiar POD vs size grid, produce Figure 21.

Table 2 – Example hit/miss dataset

ID	size (inches)	hit/miss (1/0)	ID	size (inches)	hit/miss (1/0)	ID	size (inches)	hit/miss (1/0)
1	0.21	1	57	0.09	0	116	0.188	1
2	0.095	0	58	0.057	0	117	0.085	0
3	0.195	1	59	0.082	0	118	0.215	1
4	0.095	1	61	0.079	0	119	0.175	1
5	0.207	1	62	0.034	0	120	0.242	1
6	0.102	0	65	0.145	1	121	0.255	1
7	0.141	1	66	0.165	1	122	0.275	1
8	0.115	0	67	0.15	1	123	0.262	1
9	0.08	0	70	0.15	0	124	0.22	1
10	0.019	0	74	0.144	1	125	0.2	1
11	0.324	1	75	0.26	1	126	0.125	1
13	0.08	1	76	0.315	1	127	0.207	1
14	0.325	1	77	0.3	1	128	0.24	1
15	0.082	0	78	0.065	0	129	0.24	1
16	0.084	0	79	0.07	0	130	0.24	1
17	0.18	1	80	0.07	0	131	0.247	1
18	0.176	1	81	0.25	1	132	0.211	1
19	0.152	1	82	0.2	1	133	0.155	1
20	0.085	1	83	0.32	1	135	0.265	1
21	0.116	1	84	0.352	1	136	0.13	1
22	0.09	1	85	0.085	0	137	0.14	1
23	0.138	1	86	0.045	1	138	0.05	0
24	0.125	0	87	0.025	0	140	0.08	1
25	0.067	0	88	0.213	1	141	0.07	0
26	0.075	0	89	0.167	1	142	0.075	0
31	0.155	0	91	0.023	0	143	0.08	0
32	0.117	0	93	0.067	0	144	0.087	0
33	0.317	1	94	0.065	0	145	0.05	0
34	0.345	1	95	0.025	0	146	0.225	1
35	0.25	1	96	0.215	1	147	0.179	1
36	0.407	1	97	0.19	1	148	0.21	1
37	0.355	1	99	0.177	0	149	0.212	1
38	0.3	1	100	0.034	0	150	0.125	1
39	0.32	1	101	0.182	1	151	0.137	1
40	0.32	1	102	0.187	1	152	0.13	1
41	0.111	1	103	0.235	1	153	0.21	1
42	0.152	1	104	0.235	1	154	0.194	1
43	0.142	0	105	0.216	1	155	0.202	1
45	0.157	1	106	0.195	1	156	0.212	1
46	0.185	1	107	0.187	1	157	0.257	1
48	0.097	0	111	0.235	1	158	0.182	0
49	0.125	0	112	0.262	1	159	0.216	1
52	0.091	0	113	0.205	1	160	0.21	1
53	0.11	1	114	0.227	1	161	0.125	0
55	0.18	1	115	0.207	1			

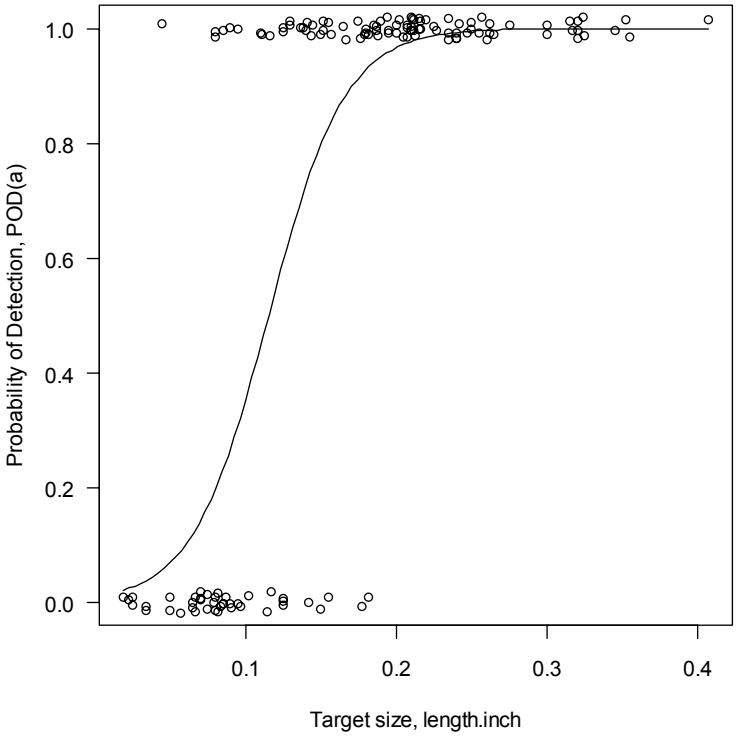

Figure 21
The logit model for POD vs size for the data in Table 2. Parameter estimates are MLE, determined using iterated re-weighted least-squares.

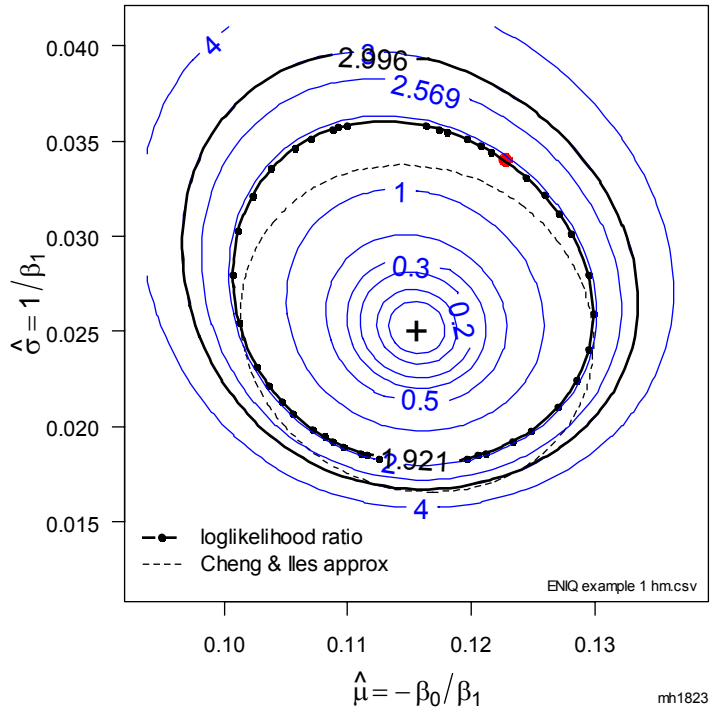

Figure 22
The log-likelihood ratio surface for the logit model parameters (μ and σ) for the data in Table 2. The plot is of negative log(likelihood ratio) where the maximum is log(1)=0.

3.4.4 Confidence bounds

Since we will need to use the probability distribution of the log-likelihood ratio to construct POD confidence bounds for hit/miss data we first review some of the properties of maximum likelihood estimators.

3.4.4.1 Asymptotic Properties of Maximum Likelihood Estimators[19]

The Central Limit Theorem

A very interesting and useful property of simple averages is that the distribution of an average tends to be normal, even when the distribution from which the average is computed is decidedly non-normal. Thus, the Central Limit theorem (CLT) is the foundation for many statistical procedures, including Quality Control Charts: the distribution of the phenomenon under study does not have to be normal because its average will be[20]. Furthermore, this normal distribution will have the same mean as the parent distribution, AND, variance equal to the variance of the parent divided by the sample size.

MLEs are asymptotically multivariate normal (MVN)

As the sample size increases, the joint distribution of MLEs becomes multivariate normal, just as the distribution of a sample average becomes normal as a consequence of the central limit theorem. That means that the multivariate normal distribution can be used to construct confidence bounds on maximum likelihood estimators. The MVN covariance matrix is asymptotically the negative inverse of the Fisher Information matrix, which itself can be estimated as the matrix of second partial derivatives of the log-likelihood with respect to the model parameters. In practice, however, constructing these so-called Wald-type bounds is not as efficient as using the asymptotic behaviour of the log-likelihood ratio.

Log-likelihood ratio has an Asymptotic χ^2_{df} Distribution

As the sample size increases, the log-likelihood ratio has an asymptotic chi-square distribution, with degrees-of-freedom equal to the number of parameters in the model, specifically:

$$-2\log\left(\frac{L(\theta_0)}{L(\theta)}\right) \sim \chi^2_{1-\alpha;df} \qquad \text{(Eq. 34)}$$

where 1-α is the confidence probability, df are the degrees-of-freedom of the model (*i.e.* the number of model parameters), and θ is the maximum likelihood model against which other models, θ_0, are being compared. Since by definition $\theta > \theta_0$, the ratio in (Eq. 34) is always less than one, hence the logarithm is always negative. The negative sign thus makes the log-likelihood ratios positive and comparable to the always-positive chi-square.

Furthermore, the log-likelihood ratio approaches χ^2_{df} more rapidly than the joint distribution of MLEs approaches MVN. Thus, for a given sample size the log-likelihood ratio produces

[19] See for example Casella and Berger (1990). For a concise summary of useful results from the theory of mathematical statistics see the Appendix of Meeker and Escobar (1998).

[20] The distribution of an average will tend to be normal as the sample size increases, regardless of the distribution from which the average is taken *except* when the moments of the parent distribution do not exist. All practical distributions in engineering have defined moments, and thus the CLT applies.

confidence bounds on the parameter estimates that are closer to nominal than those constructed using the MVN.

Back to the example, and the calculation of a confidence bound. Figure 22 also shows that while the values at "**+**" are the best (in the sense of maximizing the probability of the experiment generating the data that was observed), there are other values near "**+**" that are still plausible. How near is "near"? The isocline at 1.921 encloses 95% of the plausible[21] values for the two model parameters. Each point on that isocline is a (μ, σ) pair. Plotting all the corresponding POD(a) curves results in Figure 23.

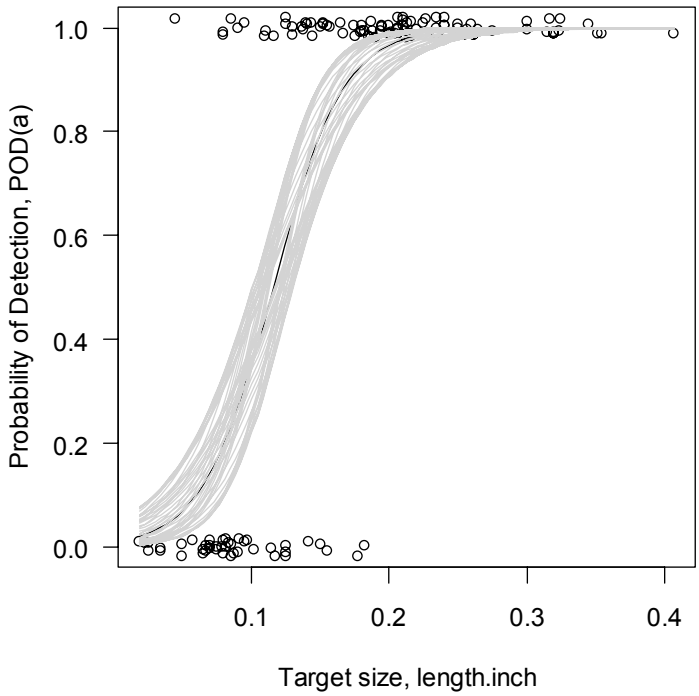

Figure 23
Plotting the POD vs size for each of the μ, σ pairs in Figure 12 shows all the plausible values that are supported by the data.

Rather than plot all the curves we plot only their envelope, Figure 24. The choice of a single degree of freedom as the likelihood ratio criterion is related to the historical fixation on a single value to represent an inspection, $a_{90/95}$.

In the early 1980s, an approximation to the log-likelihood ratio, suggested by Cheng and Iles (1982), was used, and is also shown in Figure 22 as the dashed line ellipse. Since it is always possible to construct the actual log-likelihood surface, and since the approximation often lacks fidelity (as in this example), the use of an approximation is now discouraged in favour of estimating the actual log-likelihood surface.

[21] Based on a single degree-of-freedom chi-square criterion, which corresponds to a single point on the POD curve, usually taken to be at $a_{90/95}$ for historical reasons.

Notice that there are no ”prediction” bounds in Figure 24. The confidence bounds enclose the true POD(a) curve in 95 of 100 similar experiments[22]. Prediction bounds, by contrast, are for a *single* future observation. *But you cannot observe a probability* – the next hit/miss result will be either a hit or a miss, so it makes no sense to talk about a prediction interval for POD (unless the interval is (0≤POD≤1), which, while true, is not particularly useful). The concept of a prediction interval for POD has no meaning. This obvious fact was less obvious when binomial POD(a) curves plotted the fraction found in a size interval as a point, as if it were a single observation rather than a summary of many observations (see Figure 7).

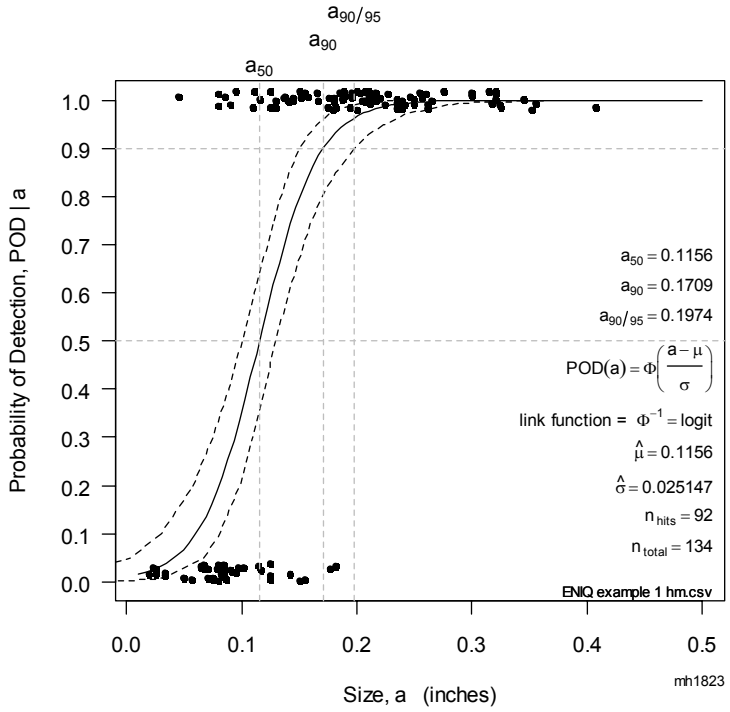

Figure 24
The POD vs size curve for the data in Appendix, example 1,showing the 95% confidence bounds and the values for a_{50}, a_{90}, and $a_{90/95}$.

3.4.5 Separating the Influence of crack size and Inspector Capability on POD

Lewis, *et al* (1978) report the results of 58 inspectors using eddy current inspection (ECI) to interrogate 42 test objects having cracks of various sizes. Fortunately, they report the raw data, which we re-examine here, Figure 25, left. Using the parametric model, Figure 25, right, compares a simulation of a single inspector’s repeated inspections of 42 cracks with the performance of the 58 ECI inspectors in Lewis, *et al* (1978). The crack sizes for the simulations were the same as those in Lewis, *et al* (1978). Inspections with negative slopes (see Figure 26) have been removed from further consideration.

[22] The confidence is with the procedure itself, and not with an individual application of that procedure. So for this realization the true POD curve is within the confidence bounds or it is not. There is no probability involved. Tossing a fair coin will show “heads” 50% of the time, but after any individual toss, the coin will show either “heads” or “tails.” There is no probability involved after the coin has been tossed, or after the NDE experimental data have been collected. The probability is associated with the *procedure* of tossing the coin, or the procedure for statistically estimating the confidence bounds for a POD model.

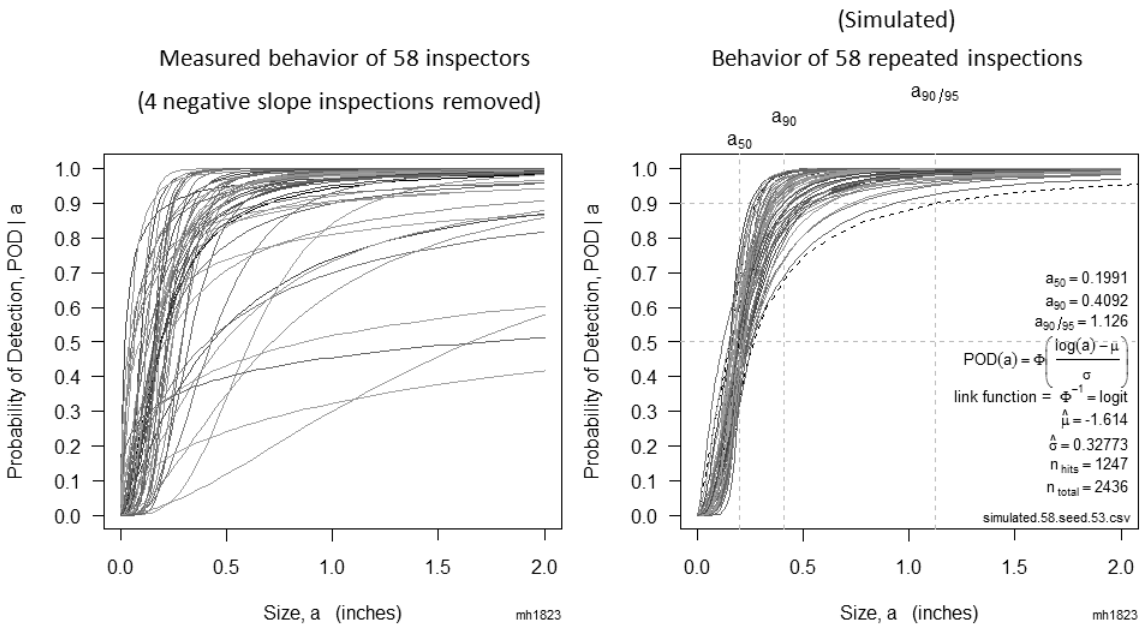

Figure 25
The inspectors do not have equal skill. POD variability results from inspection randomness and inspector-to-inspector differences (left). Repeated inspections, with only inspection randomness, display considerably less variability (right).

3.4.5.1 About the Eddy Current Inspection Data

One sample, with a crack of 1.92 inches, was either missed or not reported by 57 of the 62 inspectors and was not considered in the original document. We have omitted that specimen from our analyses.

Four of the 62 inspections resulted in negative values for the slope, indicating that the inspector's ability to detect a crack got worse as the crack size got bigger. Two of these were interesting for another reason: Inspector 1113 "found" nearly all (38) of the 42 cracks, while Inspector 1807 found only 1. We have removed from consideration the inspections having negative slopes (0605, 1113, 16E7, and 1807) although they do appear in Figure 26.

To interpret the model parameters we can borrow from pharmacology where logistic models are used to describe drug effectiveness. a_{50} is an indication of the inspector's *capability*, so smaller values are better. The model slope, β_1, describes the inspector's *sensitivity*, where a unit change in size produces a β_1 increase in the log of the odds of detection, so larger is better. It is easier to see the combined effects by plotting the model as POD vs size.

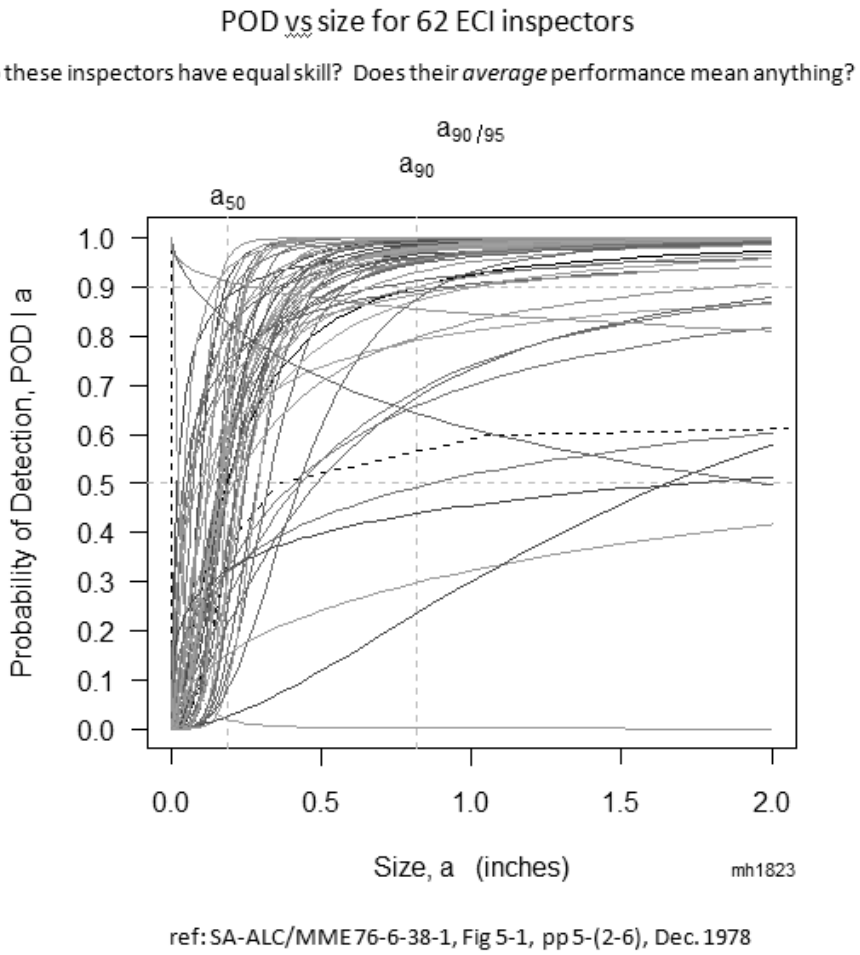

Figure 26
The inspectors do not have equal skill. Their performances should not be averaged.

3.4.5.2 Conclusion

The inspectors do not have equal skill. POD variability results from inspection randomness and inspector-to-inspector differences. Averaging these two influences obscures the relative influence of each, so POD "data" that are averages, rather than raw hit/miss results, are suspect (see Figure 7). It is useful to remember that the average weight of a grape and a grapefruit tells you very little about either, so averaging the performance of a very good inspector with an incompetent one may produce a mediocre average POD curve, but a future inspection will be performed by only one or the other inspector, so knowing their average performance is not helpful. Testing many inspectors can provide exceedingly useful information for finding effective inspection techniques, or determining who needs remedial training, but averaging inspector results tells us nothing, may hide serious deficiencies, and results in a distorted assessment of the inspection's effectiveness.

3.5 Special cases beyond the scope of this report

There are some special cases of which the reader should be aware. These are beyond the scope of this report, but are briefly discussed in the following.

3.5.1 POD models that consider more than target size

(Eq. 5), for $\hat{a}$ *vs* a data, and (Eq. 24), for hit/miss data provide a simple linear relationship between signal and size, $y=\beta_0+\beta_1 x$. This idea can be extended for other POD-controlling parameters, such as target morphology, location, orientation, chemical composition, density, reflectivity, and the like. An expanded model might be $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2+\ldots$ where, say, x_1 is target size, and x_2 is subsurface distance in a UT inspection, for example. Because these influences often interact, the model includes an interaction term for x_1 and x_2. Careful planning is required in these situations so that an understanding of the physics of the problem can instruct the mathematical form of the model. For example, it might be known that UT signal strength is inversely proportional to depth, so that in this example, $x_2=1/depth$. As a rule, the amount of data required to provide meaningful model parameter estimates increases rapidly as the number of variables is increased: four times as much data may be required for a two-parameter model as for a one-parameter model.

3.5.2 Min(POD) > 0 or Max(POD) < 1

If the POD model goes to zero, but the data do not (for example because of background noise) then using the methods in this report will result in erroneous, often non-conservative POD curves. If the POD model goes to one, but the data do not (for example because of inspection impediments, such as difficult access to the inspection site) then using these methods is not recommended. Figure 27 illustrates these situations.

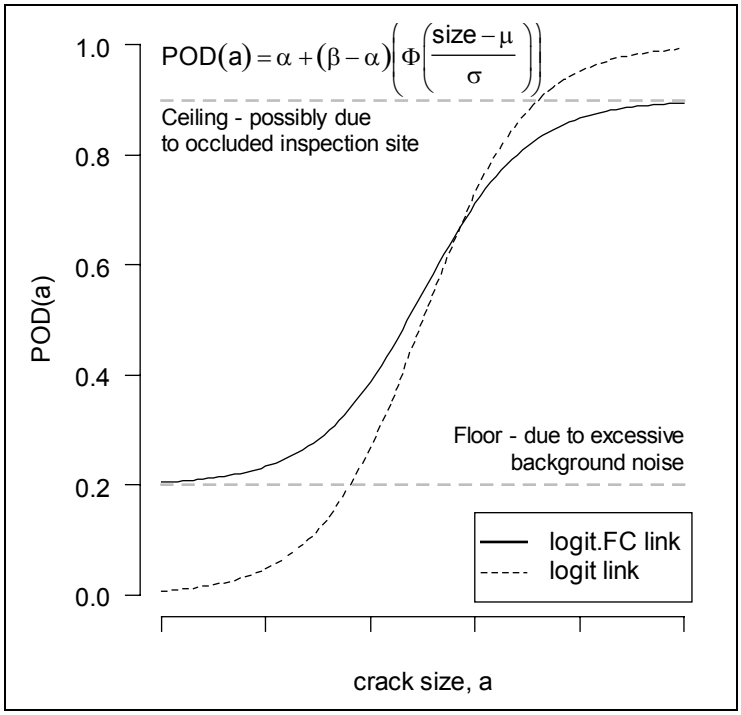

Figure 27
POD(a) model with POD "floor" (α) and POD "ceiling" (β) (FC = "floor, ceiling").

As with any engineering analysis, if the model does not fit the data, then it is the wrong model and should not be used. The methods in this report are applicable to a very large fraction of problems encountered with analyzing NDE data, but not *all* data. Special situations require specialized statistical methods that are beyond the scope of this report.

3.5.3 Field-finds

"Field-finds" are cracks discovered in components that have experienced service in the field, and are subsequently used as NDE specimens. Superficially this does not seem unreasonable, but field-finds are the most detectable cracks in a given size range, since those are those found by the inspection. It would be irresponsible to use these most-easily detected to represent all of the cracks because the inspection's capability most likely will be grossly overestimated, claiming for example 90% POD when the true detectability is closer to 50%.

The statistical properties of field-finds are radically different from POD demonstration specimens. The methods discussed in this report require that you know the entire population of cracks to be inspected. So even if a crack is missed in a POD demonstration, its size and other characteristics are known because it is a laboratory specimen. Special methods (censored regression) can then be used with known misses to determine the $\hat{a}$ *vs.* a (signal vs. size) relationship, and from that the POD(a) function.

On the other hand, with field-finds, it is clear what was found, *but it is not known what was missed*. Thus, the methods of ordinary regression and censored regression cannot be used because their underlying requirements are not satisfied. Using the methods presented herein to analyse field-finds from an unknown population of cracks will grossly over-estimate the Probability of Detection for field inspections, as shown in Figure 29.

Figure 28 plots all the data in this simulated example, and the POD(a) curve which is derived is based on the methods presented in this report. The same data are used in Figure 29, but only using those that were “detected” (a simulated threshold, $\hat{a}$=80, was used). The data below the threshold (plotted in grey) are therefore unknown to the analysis. Since only the observations above $\hat{a}$=80 are known, these pull the left side of the line sharply upward (the data that would otherwise pull the line back into place are not known). Furthermore, since a considerable fraction of the data is missing, the observed $\hat{a}$ *vs* a scatter appears much smaller (probability densities in Figure 29) than it really is (Figure 28). The result is a woefully optimistic, non-conservative POD(a) curve.

In conclusion: field-finds require special statistical methods. Using the methods in this report with specimens gleaned from field-finds *will* produce a POD curve which will be *wrong*.

3.5.4 Non-normal scatter in â vs a plots

In section 3.3.1 we listed four conditions necessary for valid $\hat{a}$ *vs* a modelling:

1. The â vs a model must look like the data.
2. The variance must be uniform about the â vs a line.
3. The observations must be uncorrelated.
4. The errors must be (approximately) normal.

In most cases these restrictions are easily met, but if any one of them is not, the resulting POD curve will be wrong. There is an entire area of applied statistics that worries about what to do when these conditions for using the Linear Model are violated. Statistical methods for dealing with exceptions to all of these requirements exist, e.g. Bates and Watts (2007), but they are beyond the scope of this report.

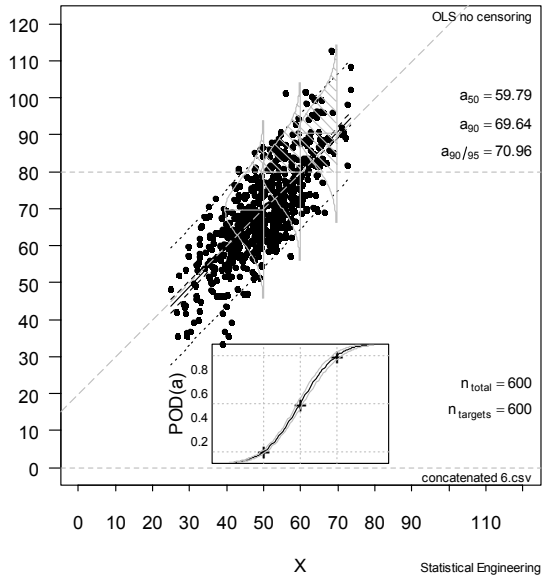

Figure 28
Correct POD vs a is based on all the data.

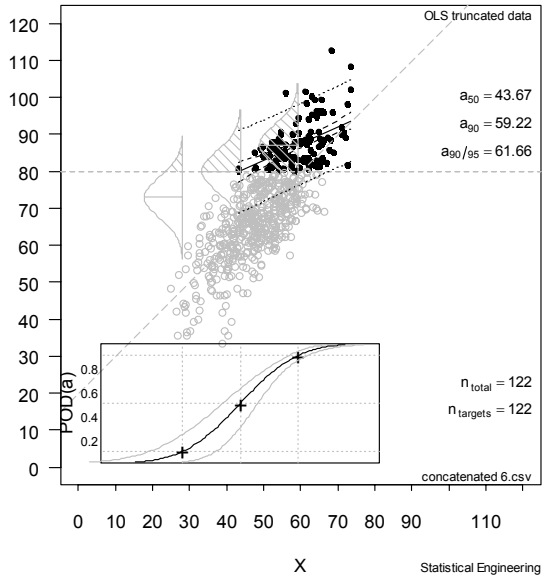

Figure 29
Non-conservative POD vs a resulting from improperly using OLS with field-finds.

3.5.5 Model-Assisted POD

Model-Assisted POD (MAPOD) is the Holy Grail of NDE practitioners and has been pursued, at great expense, for more than 30 years, including efforts that are on-going at the time of writing. Applicable results are always promised to be *almost* ready but have yet to materialize in a sufficiently well-developed form to allow use outside of the laboratory. Using mathematics and physics in place of (or to augment) laboratory specimens could greatly reduce the time and money necessary to provide POD(a) curves and is a worthy goal. It would be prudent, however, not to delay any NDE program to wait for something useable from MAPOD.

Thompson, et al (2009) summarises recent advances in MAPOD applications.

3.5.6 Bayesian Considerations

Bayesian statistics, named for the work of Rev Thomas Bayes (1702–1761) that was published posthumously, actually predates the more familiar frequentist[23] statistics. All the early work on chance, probability, and reliability began with the work of Pierre-Simon Laplace (1749–1827) who used a form of Bayes Theorem.

Bayes Theorem begins with a statement of knowledge prior to performing the experiment. Usually this prior is in the form of a probability density. It can be based on physics, on the results of other experiments, on expert opinion, or any other source of relevant information. It is desirable to improve this state of knowledge, and an experiment is designed and executed to do this. Bayes Theorem is the mechanism used to update the state of knowledge to provide a posterior probability distribution. The mechanics of Bayes Theorem can sometimes be overwhelming, but the underlying idea is very straightforward: Both the prior (often a prediction) and the experimental results have a joint distribution, since they are both different views of reality.

Let the experiment be A and the prediction be B. Both have occurred, AB. The probability of both A and B together is $P(AB)$. The law of conditional probability says that this probability can be found as the product of the conditional probability of one, given the other, times the probability of the other. That is

$$P(A|B) \times P(B) = P(AB) = P(B|A) \times P(A) \qquad \text{(Eq. 35)}$$

if both $P(A)$ and $P(B)$ are non-zero. Simple algebra shows that:

$$P(B|A) = P(A|B) \times P(B) / P(A) \qquad \text{(Eq. 36)}$$

The mathematics in equation (Eq. 36) assumes that events A and B each have a single probability. While true in many cases, in most situations the events are better described with probability densities. The underlying idea is still the same, but the arithmetic is much more tedious, as we will show. Until recently (the past two decades) computational difficulties were

[23] *Frequentist statistics* is the familiar "statistics" taught in University survey courses to engineers, physicists and other scientists, often as STAT 101. It get its name from the frequentist definition of probability as the long-run frequency of occurrence, $P(A)=n/N$, where n is the number of times event A occurs and N is the total number of opportunities. Bayesians, in contrast, see probability as a measure of the plausibility of an event, given incomplete knowledge. Although the two schools disagree on the fundamental definition of probability, the rules of the probability calculus are the same for both.

a severe detraction from the utility of Bayesian methods. Current ubiquitous inexpensive computing power has greatly mitigated this difficulty.

Let θ be an estimate of the *single-valued* $POD(a,...)$ where the size, a, and all other characteristics of the inspection are fixed, and $P(\theta)$ is the prior probability density. $P(\theta)$ is what is known about θ before the data, x, are collected. $P(\theta|x)$ is the posterior distribution of θ, and is what is known later, incorporating information in the data. Bayes's Theorem for a single continuous random variable is then:

$$P(\theta \mid x) = \frac{P(x \mid \theta)P(\theta)}{P(x)} \qquad \text{(Eq. 37)}$$

where

$$P(x) = \int P(x \mid \theta)\, P(\theta)\, d\theta \qquad \text{(Eq. 38)}$$

This becomes a multiple integral when the number of dimensions of θ increases, and as a matter of practice it can be prohibitively difficult to evaluate, even using numerical methods. Perhaps surprisingly, however, onerous computational difficulties are not the primary impediment.

Often the prior belief about θ is represented as a Beta density, and this was used by Gandossi and Simola (2005), "Framework for the quantitative modelling of the European methodology for qualification of non-destructive testing". There the authors presented a model in which Bayes theorem could be used to incorporate prior information based on expert opinion, with some ensemble of hit/miss data taken at one single target size, and with all other factors also fixed.

The underlying POD model is binomial which was discussed in Section 3.2.2 and shown to be inappropriate for realistic situations where POD is recognized to depend on target size and other factors.

To expand on Gandossi and Simola (2005), consider POD now as a function of size, and also a function of the mathematical model used to relate probability of detection to size. Let

$$POD(a) = f(a, \boldsymbol{\theta}) = \Phi_{link}^{-1}\left(\frac{x-\mu}{\sigma}\right) \qquad \text{(Eq. 39)}$$

$\boldsymbol{\theta}$ now is a two parameter column vector $\theta=(\mu, \sigma)^T$ and the prior for $\boldsymbol{\theta}$ is the *joint* probability density for μ and σ. Notice that the joint probability is more difficult to define than the single-value used previously because, in addition to defining the marginal probabilities of μ and σ, we also need to define their covariance matrix, and in nearly all real situations model parameters are *not* independent, so their covariance function is non-zero. A further complication is how to define a prior probability density for the choice of link function.

Here the Bayesian approach becomes problematic in practice. It is not difficult to elicit an expert's opinion on the prior POD for some fixed target size and other fixed inspection conditions because NDE experts can relate to the idea of POD. It is much more difficult (and

perhaps infeasible) to elicit an NDE expert's meaningful opinion on the probability density for the statistical model parameter μ, or σ, and more difficult still to ask for their opinion on their covariance and joint density. Maximum likelihood estimators are asymptotically multivariate normal, so a MVN joint density is often used to model a joint prior for model parameters. That means the NDE expert would be asked to estimate five parameters: two for the joint mean, $\boldsymbol{\mu}$, and three for the covariance matrix, **VAR**.

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_\mu \\ \mu_\sigma \end{bmatrix} \qquad \text{(Eq. 40)}$$

$$\mathbf{VAR} = \begin{bmatrix} \sigma_\mu^2 & \sigma_\mu \sigma_\sigma \\ \sigma_\mu \sigma_\sigma & \sigma_\sigma^2 \end{bmatrix} \qquad \text{(Eq. 41)}$$

Even if it was possible to find NDE experts confident enough to provide these numbers, a question mark would certainly remain on the validity and meaningfulness of such estimates[24]

These are genuine impediments to using Bayesian methods with parametric models in NDE analysis, and are therefore beyond the scope of further discussion in this report.

[24] There is an active area of statistical research concerned with eliciting expert opinion, *c.f.* Meyer and Booker (2001).

4 Ancillary topics

4.1 On the independence of inspections

As discussed in Chapter 3, the statistical model proposed by Berens and Hovey (1981) is based on the idea of grouping the sources of uncertainties affecting the signal response *â* into two distinct groups:

1. the variability in the mean *â* from flaw to flaw;
2. the variability in *â* from inspection to inspection of the same flaw.

The material properties, the flaw location, geometry, orientation, etc. are strictly associated with individual flaws, and do not change from inspection to inspection. All these causes of uncertainties affect the first type of source of variation. Human factors (the operator and its skills, attentiveness, mental attitude, health) and equipment factors (calibration, transducer variability, etc.) usually vary from inspection to inspection even of the same flaw, and therefore affect the second type of source of variation.

Under these assumptions, the analysis of NDE reliability data must follow different paths according to whether the data are for single inspections per crack or multiple inspections per crack.

The US Department of Defense Handbook on NDE Reliability Assessment [MIL-HDBK-1823A (2009)] presents an interesting discussion on this subject, under a meaningful title: "*A common misconception about statistics and POD – "Repeated inspections improve POD"*", and comes to the following conclusion: "*Repeated inspections in an NDE demonstration provide information about the inspection, not the specimens. Therefore repeated inspections of a field component provide no further information about its fitness for service*". We quote in full this discussion from MIL-HDBK-1823A (2009).

> *"The erroneous conventional thinking that POD can be improved by looking at the same item repeatedly using the same inspection system is based on a misunderstanding of simple statistics.*
>
> *Since specimens are expensive to fabricate and maintain, and since more is better than fewer, it is sometimes suggested that repeated inspections of the same specimens might be a way of increasing the effective sample size. Unfortunately this idea doesn't stand up to scrutiny. To illustrate this, consider the thought experiment of "inspecting" a barrel of apples to determine the proportion of red and green apples.*
>
> *Of course we could empty the barrel and count all the apples but this is often either too costly or otherwise infeasible. Thus, as with other NDE problems, we replace exhaustive enumeration by sampling. If we draw a random sample of n apples we can estimate the proportion of red apples as number of red apples divided by the total number of red and green apples, i.e. #red/n. If the number of apples in the barrel is much larger than the size of the sample, n, then the total number of red apples in a sample has a binomial density. (This is because there are only two possible outcomes, red or green, the probability of red for any apple is constant, the size of the sample, n, is fixed, and the "inspections" are assumed independent.) For large n, the distribution of the sample proportion of reds is asymptotically normal, centered at the true proportion of reds, p, and having a variance of $\sigma^2=pq/n$, where p is the proportion of reds, and q is the proportion of greens, and q=1−p. Knowing the estimated proportion, and its variance, permits construction of a confidence interval where the number of*

samples, n, has a clear and quantifiable influence – the confidence interval for the estimate of the proportion of red apples can be made as narrow as needed, by choosing the appropriately large number of specimens (apples), n.

Now consider using fewer than n samples. An apple is selected at random from the barrel and Inspector 1 reports that it is red. A second inspector also declares it to be red. Inspectors 3, 4, and 5 also examine the apple and all concur that it is indeed red. How much more is known now about the proportion of reds in the barrel after these five "inspections" than was known after the first "inspection?" Answer: Nothing. While multiple inspections will provide insight into the consistency of our observations, they provide zero further illumination concerning the proportion of red apples in the barrel. You can't decrease the number of samples by multiple inspections because the "inspections" are not independent, as was implicitly assumed in this example. (In NDE analysis independence is almost always assumed, rightly or wrongly, but unfortunately this is left unsaid, and thus often overlooked or ignored, sometimes with unfortunate consequences.)" (from MIL-HDBK-1823A (2009), pages 36-37)

The conclusion that it is not possible to "*decrease the number of samples by multiple inspections because the "inspections" are not independent*" is fundamentally correct, but this example is slightly misleading. It is clear that the determination of the colour of an apple drawn from the barrel is, in the example, free from mistakes. Inspectors 1 to 5 examine the apple and all concur that it is red (or green). In an NDE experiment, the situation is not so straightforward. A crack which is "drawn from the barrel" will be inspected and – typically – sometimes will be found and sometimes will not be. This is the variability which is inherent and so typical of an NDE examination. So, multiple inspections of the same crack (by different inspectors, or by the same inspector on a different day, with a newly calibrated equipment, etc.) can add interesting information concerning the detection probability of that individual crack.

For several NDE systems, the inspector plays a significant role in influencing the detection probability. For instance, if the signal response is not in the form of a single quantitative value $\hat{a}$ (which can be easily compared with a decision threshold value $\hat{a}_{dec}$) but is for instance in the form of a complex 2-D pattern which must be qualitatively assessed by the inspector, the latter will have a definite influence on whether a crack is identified or not from an ambiguous indication. In this case, it could be argued that a truly independent inspection will be one where all the factors affecting the variation are randomly and independently chosen. In other words, the data should be formed by a set (of adequately large sample size, *N*) of cracks of various sizes covering the whole range of interest, each inspected once by a single inspector (randomly chosen among the set of representative inspectors).

On the other hand, in practice an experimenter would elect to have that set of *N* specimens inspected by as many inspectors as possible, in order to also assess the influence of inspector. A method called Linear Mixed Model (beyond the scope of this report, cf. e.g. McCulloch *et al.* (2008)) could then be used for statistically modelling the influence of inspector as a probability distribution of inspector capabilities (so that the resulting model would add to the parameters describing crack-to-crack variability others describing the inspectors). In practice, grouping the inspectors together and treating their variability as part of the overall variability is much easier, and will not be too far off in most cases (assuming that all inspectors are equally competent). It would then be important not to forget that the number of degrees of freedom is really limited by the number of cracks, not the number of cracks times the number of inspectors.

4.2 Sample size requirements

With the obsolete single-point POD method based on the binomial distribution, it was easy to compute the number of specimens necessary to claim some α percentage confidence on a given single value for POD, using (Eq. 4). This is not possible with parametric models because the answer depends on the inherent scatter in the data, with respect to the parametric model used to describe it. Nonetheless we offer some rules-of-thumb.

For hit/miss data experience has shown that 60 specimens is often adequate, and using fewer often results in confidence bounds, while valid, that are too broad to be useful. (As an extreme example, knowing that $a_{90/95}$ is less than, say 5 cm, may be true – it's less than 10 cm also - but not helpful). For $\hat{a}\ vs\ a$ data, 30 specimens is a practical minimum requirement.

In any case it is useful to conduct a numerical simulation beforehand of any proposed statistical experiment, like a POD demonstration experiment. This can uncover, at very little cost, possible problems with the experiment's design, and also show if the experiment's goals can be accomplished at all with the experiment's resources. The reader should be forewarned however: such simulations are often faulty themselves because they are vulnerable to GIGO - Garbage in, Garbage out – assuming as given what you are trying to prove.

5 Summary and Conclusions

This paper has summarized the current state-of-the-art in statistical analysis of NDE data, beginning with a review and critique of early efforts (1970s) to produce Probability of Detection curves based on single-point binomial methods. We present worked-through examples that expose their severe limitations. While statisticians have been aware of these deficiencies for decades, much of the engineering community has remained aloof, and as recently as 2009 continues to promulgate old, inferior methods in new engineering specifications. We hope to change this situation with the thorough discussion presented herein, on the reasons why binomial methods should be retired in favour of more efficient and effective methods.

We described two methods, based on statistical best-practices (as of 2010), for extracting the maximum amount of useful information in NDE data to produce POD vs size curves: (1) one for data for which the signal contains useful information about the target (e.g. its size) and (2) one for data that conveys only a binary result (e.g. found or not found). The first type of data is called â vs a data, the second type is called hit/miss data.

While seemingly having little to do with one another, the methods for analyzing each are related: â vs a data use a linear relationship between size and signal, to compute a probability of detection as the probability that the signal at some target size will exceed a decision threshold. Analysis of binary data is based on a generalization of the linear model (GLM) and posits an underlying relationship between size and POD. Both methods have been part of the mainstream statistical literature for more than two decades. We also note that powerful software implementing these methods is available for free.

We also presented a re-analysis of some historical data that illustrates the GLM method and demonstrates how much more information can be extracted from a given collection of observations. This led us to a discussion of the influence of inspector, the effectiveness of re-inspection (it is not as effective as it might seem), sample size requirements, and Bayesian considerations.

Finally, we conclude with a look to the future that includes MAPOD - Model-Assisted Probability of Detection: an evolving methodology for augmenting physical specimens with mathematical models of the underlying physics of inspection to provide POD vs size curves in situations where relying on specimens alone is either too costly, too time consuming, or both.

6 Acknowledgements

The authors would like to express their thanks to the following individuals for reviewing earlier drafts of this paper and for providing very valuable comments. In alphabetical order: Robertas Alzbutas (LEI, Lithuania), Christiane Bruynooghe (JRC-IE, The Netherlands), Bob Chapman (EDF Energy, United Kingdom), Hans Martinsen (Ringhals NPP, Sweden), Anders Richnau (Ringhals NPP, Sweden), Charles Schneider (TWI, United Kingdom), Kaisa Simola (VTT, Finland), Håkan Söderstrand (SQC Swedish Qualification Centre AB, Sweden), Iikka Virkkunen (Trueflaw Ov, Finland).

7 References

AGARD-LS-190 (1993), "A Recommended Methodology for Quantifying NDE/NDI Based on Aircraft Engine Experience," North Atlantic Treaty Organization, Advisory Group for Aerospace Research and Development, series of lectures by Charles Annis and Sharon Vukelich, available for download at http://ftp.rta.nato.int/public//PubFullText/AGARD/LS/AGARD-LS-190///AGIA2DL5190.pdf.

Ammirato, F., and Dennis, M. (2010) "Development of POD and Flaw Sizing Uncertainty Distributions from NDE Qualification Data to Support Probabilistic Structural Integrity Assessment for Dissimilar Metal Welds", 8th International Conference on NDE in Relation to Structural Integrity for Nuclear and Pressurised Components, Berlin, 29 September - 1 October, 2010.

Bates, D.M. and Watts, D.G. (2007), **Nonlinear Regression Analysis and Its Applications**, Wiley.

Berens, A.P., and Hovey, P.W. (1981), "Evaluation of NDE reliability characterisation", AFWAL-TR-81-4160, Vol. 1, Air Force Wright-Aeronautical Laboratories, Wright-Patterson Air Force Base.

Berens, A.P. and Hovey, P.W. (1984), "Flaw Detection Reliability Criteria. Volume 1. Methods and Results", AD-A142 001, DAYTON UNIV OH RESEARCH INST, Final technical report.

Berens, A.P. (1989), "NDE Reliability Data Analysis", in Non-destructive Evaluation and Quality Control: Qualitative Non-destructive Evaluation, ASM Metals Data Book, Volume 17, ASM International.

Box, G.E.P., Hunter, W.G. and Hunter, J. S. (1978), **Statistics for Experimenters**, Wiley.

Casella, G. and Berger, R. (1990), **Statistical Inference**, Duxbury Press.

Cheng, R.C.H. and Iles, T.C. (1983), "Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables," *Technometrics*, vol. 25, no. 1, pp 77 – 86.

Damage Tolerance Design Handbook (2006), contributors: Peggy C. Miedlar, Paul A. Wawrzynek, Mary Schleider, Jerzy P. Komorowski, Matthew Creager, Don Locke. Available for download at http://www.afgrow.net/applications/DTDHandbook/.

ENIQ (1998), "ENIQ recommended practice 3: strategy document for technical justification", ENIQ Report No. 5, JRC-Petten, EUR 18100/EN; 1998.

ENIQ (2007), "European methodology for qualification of non-destructive testing: third issue", EUR 22906 EN, 2007.

Førli, O. et al. (1998), "Guidelines for NDE reliability determination and description", Nordtest NT Technical Report 394. Available for download at http://www.nordicinnovation.net/nordtestfiler/nt394.pdf.

Førli, O., Ronold, K.O. et al (1999), "Guidelines for development of NDE acceptance criteria", Nordtest NT Technical Report 427. Available for download at http://www.nordicinnovation.net/nordtestfiler/nt427.pdf.

Gandossi, L. and Simola, K. (2005), "Framework for the quantitative modelling of the European methodology for qualification of non-destructive testing", International Journal of Pressure Vessels and Piping, Volume 82, Issue 11, November 2005, Pages 814-824.

Gandossi, L. and Simola, K. (2007), "A Bayesian Framework for the Quantitative Modelling of the ENIQ Methodology for Qualification of Non-Destructive Testing", JRC Technical report, EUR 22675 EN, 2007. Available for download at http://safelife.jrc.ec.europa.eu/eniq/docs/eur_reports/EUR-22675.pdf.

Generazio, E.R. (2009), "Design of Experiments for Validating Probability of Detection Capability of NDT Systems and for Qualification of Inspectors," *Materials Evaluation*, June, 2009.

Georgiou G. (2006), "Probability of Detection (PoD) curves - Derivation, applications and limitations", HSE Research Report 454. Available for download at http://www.hse.gov.uk/RESEARCH/rrhtm/rr454.htm.

Gosselin, S.R., Simonen, F.A., Heasler, P.G., Doctor, S.R. (2007), "Fatigue Crack Flaw Tolerance in Nuclear Power Plant Piping – A Basis for Improvements to ASME Code Section XI Appendix L", NUREG/CR-6934, PNNL-16192. Available for download at http://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6934/cr6934.pdf.

JSSG (2006), Joint Service Specification Guide, Aircraft Structures, Department of Defense, 30 October 1998.

Lewis, W.H., Dodd, B.D., Sproat, W.H., and Hamilton J.M. (1978), "Reliability of Nondestructive Inspections – Final Report". Report No. SA-ALC/MEE 76-6-38-1. United States Air Force, San Antonio Air Logistics Center, Kelly Air Force Base, Texas. Available for download at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA072097&Location=U2&doc=GetTRDoc.pdf.

Matzknin, G A and Yolken, HT (2001), "Probability of detection (PoD) for non-destructive evaluation (NDE)", NTIAC-TA-00-01.

McCulloch, C.E., Searle, S.R, and Neuhaus, J.M. (2008), **Generalized, Linear, and Mixed Models**, 2nd Ed., Wiley.

Meeker, W.Q. and Escobar, L.A. (1998), **Statistical Methods for Reliability Data**, Wiley.

Meyer, M.A. and Booker, J.M. (2001), **Eliciting and Analyzing Expert Judgment: A Practical Guide**, SIAM (Society for Industrial and Applied Mathematics).

MIL-HDBK-1530 (1996), Military Handbook - General Guidelines for Aircraft Structural Integrity Program, 4 November 1996, available for download at http://terpconnect.umd.edu/~sanford/milhandbook.pdf.

MIL-HDBK-1823A (2009), "Nondestructive Evaluation System Reliability Assessment," Standardization Order Desk, Building 4D, 700 Roberts Avenue, Philadelphia, PA 19111-5094. Available for download at http://mh1823.com/mh1823.

Nelder, J.A. and Wedderburn, R.W.M., (1972), "Generalized Linear Models", *Journal of the Royal Society* A 135, 370-84.

O'Regan, Patrick (2010), personal communication with Luca Gandossi (15 October, 2010).

RTO (2005), "The Use of In-Service Inspection Data in the Performance Measurement of Non-Destructive Inspections", TECHNICAL REPORT TR-AVT-051, The Research and Technology Organisation (RTO) of NATO, March 2005. Available for download at http://www.rta.nato.int/Pubs/RDP.asp?RDP=RTO-TR-AVT-051.

Rummel, W.O. and Matzkanin, G.A. (1997), "Nondestructive Evaluation (NDE) Capabilities Data Book", NTIAC: DB-97-02, Nondestructive Testing Information Analysis Center (NTIAC), Austin, Texas.

Salkowski, C. (1993), “NDT and the Aging Orbiter.” FAA Inspection Reliability Workshop, Atlantic City, NJ.

Selby, G., Harrington, C. (2009), "Materials Reliability Program: Development of Probability of Detection Curves for Ultrasonic Examination of Dissimilar Metal Welds (MRP-262) – Typical PWR Leak-Before-Break Line Locations". EPRI, Palo Alto, CA. 1019088.

Singh, R. (2000), "Three Decades of NDI Reliability Assessment", Report No. Karta-3510-99-01. Available for download at http://www.cnde.iastate.edu/MAPOD/Reference%20Documents/Karta%20pod%201970-1999.pdf.

Thompson, R. Bruce, Brasche, L. J., Forsyth, D. S., Lindgren, E., Swindell, P., Winfree, W., (2009), “Recent Advances in Model-Assisted Probability of Detection”, in the Proceedings of the 4th European-American Workshop on the Reliability of NDE, Berlin, Germany, 23 – 26 June 2009. Available for download at http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20090025450_2009023843.pdf.

Visser, W. (2002), POD/POS curves for non-destructive examination, HSE Offshore Technology Report 2000/018. Available for download at http://www.hse.gov.uk/research/otohtm/2000/oto00018.htm.

Yee, B.G.W., Chang, F.H., Covchman, J.C., Lemon, G.H., Packman, P.F., (1976), ass "Assessment of NDE reliability data", NASA report NASA-CR-134991, Oct 1, 1976. Available for download at http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19760026437_1976026437.pdf.

This page is intentionally left blank.

European Commission

EUR 24429 EN – Joint Research Centre – Institute for Energy
Title: ENIQ TGR TECHNICAL DOCUMENT – PROBABILITY OF DETECTION CURVES: STATISTICAL BEST-PRACTICES

Author: Luca GANDOSSI (DG-JRC-IE)
Charles ANNIS (Statistical Engineering)

Luxembourg: Publications Office of the European Union
2010 – 68 pp. – 21 x 29.7 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-16105-6

Abstract
In the application of a non-destructive evaluation (NDE) method there are several factors that will influence whether or not the inspection will result in the correct decision as to the presence or absence of a flaw. In general, NDE involves the application of a stimulus to a structure and the subsequent interpretation of the response to the stimulus. Repeated inspections of a specific flaw can produce different magnitudes of the stimulus response because of very small variations in setup and calibration. This variability is inherent in the process. Different flaws of the same size can produce different response magnitudes because of differences in the material properties, flaw geometry and flaw orientation. Further, the interpretation of the response can be influenced by the capability of the interpreter (manual or automatic) and by the mental acuity of the inspector (in turn, dependent on many factors such as fatigue, emotional outlook, ease of access, environment, etc.).
Much of the modern literature on inspection reliability constantly refers to a small set of seminal papers, derived from studies mainly carried out in the aeronautical industry. Most of the issues involved are of course very similar. One notable exception is possibly the fact that in the nuclear industry, for the very nature of the components being inspected, the sample sizes of inspected cracks tend to be much lower. In Europe, the ENIQ methodology for inspection qualification was specifically developed in the early 1990s because of the difficulty and cost of procuring or manufacturing representative flaws in test pieces in a high enough number to draw quantitative (statistical) conclusions on the capability of the NDE system being investigated. Rather, the fundament of the ENIQ methodology is the Technical Justification, a document assembling evidence and reasoning providing assurance that the NDE system is capable of finding the flaws which it is designed to detect with a high enough reliability. This assurance is qualitative, and comes usually in the form of statements such as: "Sufficient experimental verification of the procedure has been performed, on representative defects in test blocks with the correct geometry, to be confident that the procedure and equipment will find all defects which conform to the detection criteria and to specifications within the range of plausible defects".
The purpose of this document, aimed mostly at NDE engineers and practitioners, is threefold: (1) to provide a brief literature review of some important papers and reports; (2) to review in a simple and structured way the statistical models that have been proposed to quantify inspection reliability and to point out problems and pitfalls which may occur, and (3) to describe and recommend statistical best practices for producing POD vs size curves from either hit/miss data, or â vs. a data.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

JRC
EUROPEAN COMMISSION

LD-NA-24429-EN-C

Publications Office

ISBN 978-92-79-16105-6