AFRL-ML-WP-TP-2006-492

COMPARING THE EFFECTIVENESS OF $a_{90/95}$ CALCULATIONS (PREPRINT)

Charles Annis and Jeremy Knopp

SEPTEMBER 2006

MATERIALS AND MANUFACTURING DIRECTORATE
AIR FORCE RESEARCH LABORATORY
AIR FORCE MATERIEL COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433-7750

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

1. REPORT DATE *(DD-MM-YY)* September 2006	**2. REPORT TYPE** Conference Paper Preprint	**3. DATES COVERED *(From - To)***

4. TITLE AND SUBTITLE COMPARING THE EFFECTIVENESS OF $a_{90/95}$ CALCULATIONS (PREPRINT)	**5a. CONTRACT NUMBER** F33615-03-D-5204
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER 62102F
6. AUTHOR(S) Charles Annis (Statistical Engineering) Jeremy Knopp (AFRL/MLLP)	**5d. PROJECT NUMBER** 4349
	5e. TASK NUMBER 41
	5f. WORK UNIT NUMBER 05

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

Statistical Engineering Palm Beach Gardens, FL 33418-7161	Nondestructive Evaluation Branch (AFRL/MLLP) Metals, Ceramics and NDE Division Materials and Manufacturing Directorate Air Force Research Laboratory Air Force Materiel Command Wright-Patterson AFB, OH 45433-7750

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Materials and Manufacturing Directorate Air Force Research Laboratory Air Force Materiel Command Wright-Patterson AFB, OH 45433-7750	**10. SPONSORING/MONITORING AGENCY ACRONYM(S)** AFRL-ML-WP
	11. SPONSORING/MONITORING AGENCY REPORT NUMBER(S) AFRL-ML-WP-TP-2006-492

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Conference paper submitted to the Proceedings of 33rd Annual Review of Progress in Quantitative Nondestructive Evaluation (QNDE) 2006.

The U.S. Government is joint author of this work and has the right to use, modify, reproduce, release, perform, display, or disclose the work.

Paper contains color. PAO Case Number: AFRL/WS 06-2439; Date cleared: 12 Oct 2006.

14. ABSTRACT

Most practitioners see $a_{90/95}$ as a static, single-point summary of an entire inspection's capability. It purports to be the size of the target having at least 90% probability of detection in 95 of 100 POD experiments under nominally identical conditions. But in some situations the actual coverage is closer to 80%, rather than 95%, with 50% coverage being the median POD(a) curve itself. This paper discusses the two philosophies, the Wald Method, and the Loglikelihood Ratio Method, for constructing lower bounds on POD(a) curves (and therefore determining $a_{90/95}$) and compares the effectiveness of each as functions of other experimental realities such as sample size and balance.

15. SUBJECT TERMS

confidence bounds, Wald method, loglikelihood ratio, POD, probability, $a_{90/95}$

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT:**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** (Monitor)
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	14	Jeremy Knopp **19b. TELEPHONE NUMBER** *(Include Area Code)* N/A

Comparing the Effectiveness of $a_{90/95}$ Calculations

Charles Annis, P.E.[1] and Jeremy Knopp[2]

[2]Materials and Manufacturing Directorate, Air Force Research Laboratory, Dayton, OH
[1]Statistical Engineering, Palm Beach Gardens, FL 33418-7161

ABSTRACT. Most practitioners see $a_{90/95}$ as a static, single-point summary of an entire inspection's capability. It purports to be the size of the target having at least 90% probability of detection in 95 of 100 POD experiments under nominally identical conditions. But in some situations the actual coverage is closer to 80%, rather than 95%, with 50% coverage being the median POD(a) curve itself. This paper discusses the two philosophies, the Wald Method, and the Loglikelihood Ratio Method, for constructing lower bounds on POD(a) curves (and therefore determining $a_{90/95}$) and compares the effectiveness of each as functions of other experimental realities such as sample size and balance.

Keywords: confidence bounds, Wald method, loglikelihood ratio, POD, probability, $a_{90/95}$

INTRODUCTION

For POD(a) models based on *log(â) vs log(a)* data, all POD calculations are with respect to the *log(â) vs log(a)* regression, not the POD(a) space. The parameters of the POD model are used only for plotting. Thus the lower bounds constructed for the POD curve – including the calculation for $a_{90/95}$ – are not significantly different based on their method of construction, from either the Wald or loglikelihood ratio method.

With hit/miss data things are very different, and the effectiveness of lower bound calculations depends on the sample size, the balance of the target sizes (how many are on either side of the POD inflection), the sensitivity of the inspection (as indicated by the steepness of the POD(a) relationship) and how all these are influenced by the method for constructing lower bounds.

This paper discusses the two philosophies for constructing lower bounds on POD(a) curves (and therefore determining $a_{90/95}$) and compares the effectiveness of each as functions of other experimental realities such as sample size and balance.

WHAT IS MEANT BY "CONFIDENCE?"

While we are not concerned with ordinary regression in this paper, we must digress briefly to discuss it since the Wald method of placing confidence bounds on POD(a) curves is based on methods that are only valid for ordinary linear regression.

Requirements for a valid Ordinary Least-Squares Regression Model

There are four *mandatory* requirements for ordinary regression and all four must be satisfied.

1. Linearity of the parameters: Nonlinear functions of X variables are permitted, such as X^2 or log(X), but the model parameters, β, must appear alone and untransformed. In

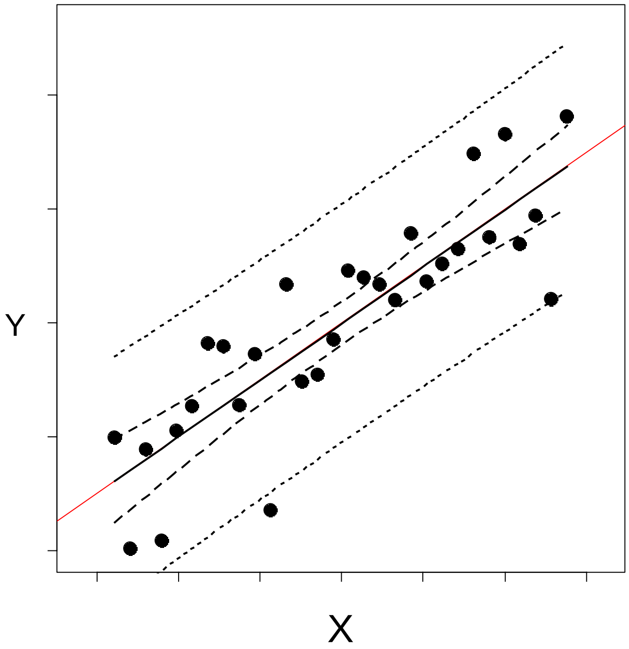
other words the relationship between the response **y** and the controlling variables **X** can be nonlinear, so long as the relationship of **y** with respect to the model parameters, $\boldsymbol{\beta}$ is linear. For example $y = \beta_0 + \beta_1 \sin(x^2)$ is a linear model; $y = \beta_0 + \exp(\beta_2 x)$ is not.

2. Uniform variance (homoscedasticity): $\text{var}(y_i \mid \mathbf{X}) = \sigma^2, \quad i = 1, 2, 3, \cdots, n$
3. Conditionally uncorrelated observations: $\text{cov}(y_i, y_j \mid \mathbf{X}) = 0, \quad (i \neq j)$
4. Normal errors: $(y_1, y_2, \cdots, y_n) \mid X$ have a multivariate normal distribution.

If *any* of these requirements is not met, the resulting regression model will be invalid and conclusions based on it will necessarily be in error. The most often violated requirement is for uniform variance.

Figure 1 depicts an ordinary regression. There are two sets of bounds. The outermost bounds are on the individual observations, and are called "prediction bounds." We are not concerned with prediction bounds here. The innermost bounds are on the regression model (the solid line) itself, and are called "confidence bounds." We would expect that the true line to fall within these 95% confidence bounds in 95 of 100 future experiments like the one that produced these data. We are concerned with the performance of confidence bounds, not on an ordinary regression, as in this figure, but on POD(a) model produced by hit/miss data.

Figure 1 Ordinary Least-Squares Regression Showing 95% Confidence Bounds (innermost lines)



HOW TO BUILD CONFIDENCE BOUNDS ON A LEAST SQUARES REGRESSION.

We calculate the estimated response, $\hat{y}$, from the regression equation, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. We don't know the true values for the parameters and thus must rely on their estimates, $\hat{\beta}_1, \hat{\beta}_2$. We are interested in the variability of $\hat{y}$ as a consequence of the variability in $\hat{\beta}_0, \hat{\beta}_1$. Since $\hat{y}$ involves a sum and a product we need some statistical background.

From the definition of variance it can be shown that the variance of a sum is $\text{var}(U + V) = \text{var}(U) + \text{var}(V) + 2\,\text{cov}(U, V)$ and the variance of a product of a constant and a variable is $\mathrm{var}(aU) = a^2\,\mathrm{var}(U)$. Thus the variance of the expected value of regression response $\hat{y}$ is

$$\mathrm{var}(\hat{y}) = \mathrm{var}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) = \mathrm{var}\left(\hat{\beta}_0\right) + 2x\,\mathrm{cov}\left(\hat{\beta}_0, \hat{\beta}_1\right) + x^2\,\mathrm{var}\left(\hat{\beta}_1\right) \qquad \text{equation 1}$$

From which the 95% Wald confidence bounds on $\hat{y}$ can be constructed:

$$\hat{y}_{\alpha=0.95} = \hat{y} + 1.645\, sd_{\hat{y}} = \hat{\beta}_0 + \hat{\beta}_1 x + 1.645\sqrt{\mathrm{var}(\hat{y})} \quad \text{where 1.645 is z(0.95)} \qquad \text{equation 2}$$

Before investigating the performance of the Wald bounds, questionably applied to POD(a), which is an OLS regression, we need to consider a better alternative – confidence bounds based on the loglikelihood ratio criterion.

HOW THE LOGLIKELIHOOD RATIO CRITERION WORKS

Likelihood is “the probability of the data.” It is proportional to the probability that the experiment turned out the way it did. So some POD model parameters are more likely than others because they explain the inspection outcome better than other values. We choose the “best” parameters, i.e. those that maximize the likelihood. These are called the maximum likelihood parameters estimates.

If we choose slightly different values, the resulting likelihood diminishes. As a consequence of the Central Limit Theorem, the ratio of the logs of the new values to their maximum values, the loglikelihood ratio, Λ, has an asymptotic chi-square density. That provides a means for constructing likelihood ratio confidence bounds: Move the POD(a) model parameters away from their maximum values but not too far – only until the criterion is reached. In other words, values of the parameters that are “close” to the best estimates are plausible, but values that are “far” are unlikely to describe the data. The asymptotic behavior of Λ provides a way of determining what is meant by “close.”

Figure 2 POD(a) curve based on Hit/Miss Data, showing Wald confidence bounds (narrow) and Loglikelihood Ratio Bounds

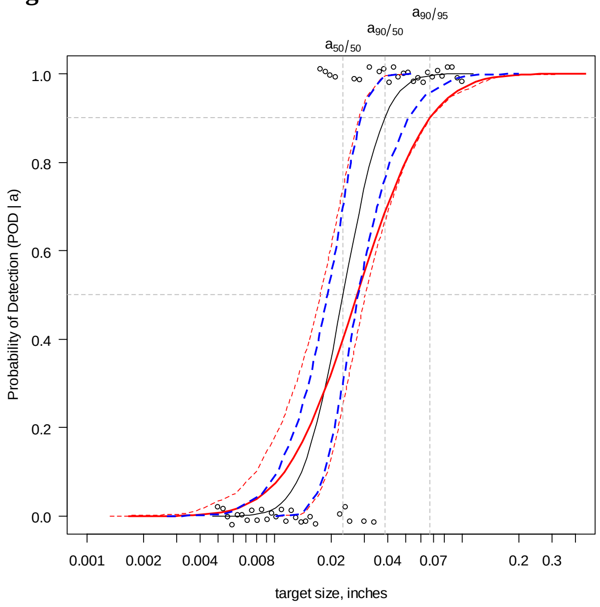
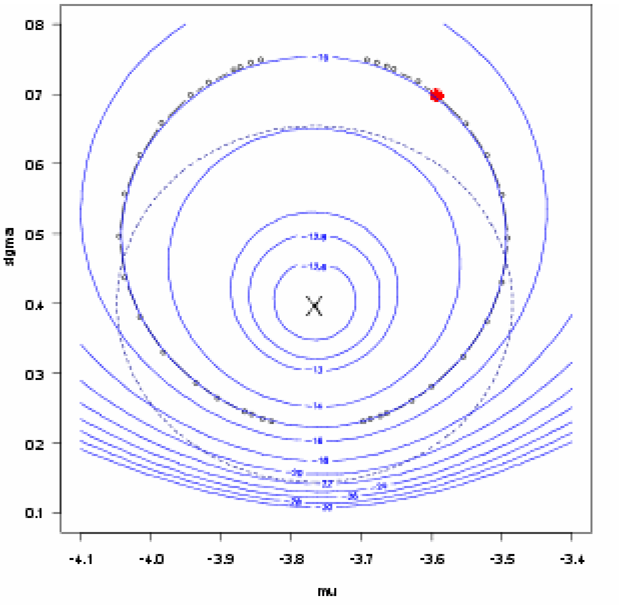


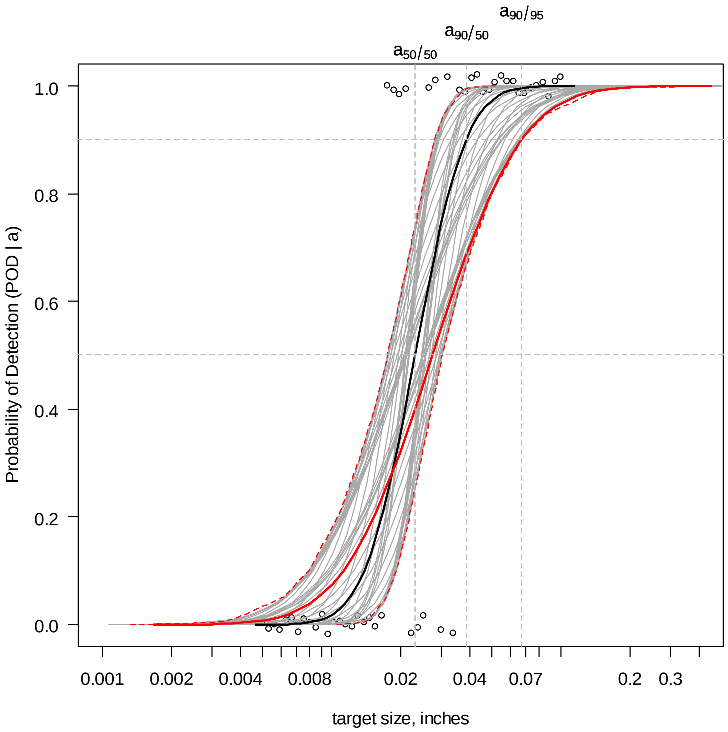
Figure 3 Loglikelihood Contour Plot Showing the MLE (X) and the 95% Confidence Contour.



Consider the POD(a) curve in Figure 2, represented by the solid line. Two model parameters determine the line: μ which locates the curve horizontally and is, for a log(x) model, the log of the size having 50% probability of detection, and σ, which is the inverse of the POD curve's "slope." Please remember that even though the equation for the POD(a) curve is the cdf for a normal density, and the parameters are those of a normal density, *there is no statistical significance* to this. If there were, then the curve would describe the cumulative probability of *existence* of a target of size *a*, and not the probability of *finding* a target of that size, given that it exists. Thus two numbers, μ and σ, describe the curve. In this example they are (-3.766, 0.3992). The resulting loglikelihood is -12.468. (The units for loglikelihood are immaterial.).

Next consider a plot of the loglikelihood for different values of μ and σ, shown in Figure 3. Moving the pair from their mle position (the large **X**) changes the loglikelihood, as illustrated by the contour lines. One of the contours, shown by the alternating lines and dots, is the 95% confidence bound for the parameter estimates based on these data. In other words, the true μ, σ pair is expected to be contained within such a confidence ellipse in 95% of future experiments like this one. We now construct POD curves for all the points along the 95% confidence ellipse in Figure 3. These are shown in Figure 4

Figure 4 POD(a) Curves for Parameter Values on the 95% Confidence Ellipse



The envelope of all these POD(a) curves represents the confidence bounds on the POD(a) curve. It is interesting to note the dark solid line intersecting POD = 0.90, and a = $a_{90/95}$ which corresponds to the large dot on that confidence contour in Figure 3.

Figure 3 has some additional interesting features. Notice that the maximum likelihood estimates (the big **X**) are *not* in the center of the loglikelihood contours. As the sample size is increased the resulting contours contract toward the MLEs and the contours become symmetrically centered asymptotically, but for this smaller sample (n=51) the contour is decidedly *not* symmetric. There is another ellipse (dotted line) that is centered at the MLE values. That is the Cheng and Iles (Cheng and Iles, 1983) approximation to the confidence contour. For small sample sizes it is a poor approximation, as is evident here.

APPLYING THE WALD METHOD TO THE POD(a) CURVE (which is questionable)

We have discussed Wald bounds for an ordinary regression, but fitting a probit model to binary data is not an ordinary regression. The Wald confidence equations can used in a situation for which they do not apply – but the resulting bounds do not have the advertised 95% coverage, and in all cases (that we have investigated) are *anti*-conservative. Rather than having 95% coverage, their coverage is usually much less, approaching 80% in many instances. Here is why Wald methods violate the assumptions on which they are based:

1. The range of the response is [0,1]. OLS assumes $-\infty < y < \infty$ as a consequence of requirement 4 (above) for normal errors. (The domain of a normally distributed variable is infinite).
2. OLS requires normally distributed errors; Hit/Miss errors are binomial.
3. OLS requires constant variance; The variance of a binomial is $p(1-p)$. The variance is greatest at *p*=0.5 (variance = $0.5 \times 0.5 = 0.25$) and approaches zero as *p* approaches either zero or one. ($0 \times (1-0) = 1 \times (1-1) \Rightarrow 0$)
4. Finally, *size* is the independent variable, yet the conventional formulation has size as a function of POD for purposes of constructing a Wald lower bound on POD. Of course it would not be possible to fit a model with size as the *dependent* variable since the responses, forced to be the independent variable, are either zero or one, so some statistical manipulation is required to translate the covariance matrix for the parameter values from a generalized linear model with a probit link to those having the MH1823 formulation.

Now, ignoring the uncomfortable fact that the binary response does *not* result in a constant variance, equation 2 appears to be applicable for computing bounds on cracksize, *a*, for a given POD. Let $\Phi(\cdot)$ be the normal cdf function, then $POD(a) = \Phi\left([a-\mu]/\sigma\right)$ so that $a = \mu + \sigma\, \Phi^{-1}\left(POD(a)\right)$, and thus $a_{0.90} = \mu + 1.282\,\sigma$. Finally we calculate the 95% confidence on $a_{0.90}$: $a_{0.90/0.95} = a_{0.90} + 1.645\, sd_{a\,0.90}$ where $sd_{a\,0.90}$ is coaxed from equation 1 [1] and the covariance of the parameters μ, σ, which can be obtained either as the inverse of the negative matrix of second partials of the loglikelihood surface, Figure 3, or, more simply, from the covariance of the glm parameters via the delta method.

Therefore constructing confidence bounds on the probit POD(a) model using OLS methods is questionable at best. Perhaps not surprisingly, how well Wald bounds perform depends on the situation:

1. The number of hit/miss targets,
2. The distribution of sizes of the targets
3. The performance of the inspection being used, as quantified by
   a) The location of the POD(a) curve with respect to the sizes being inspected for, whether it is "balanced," (half to the right of the POD(a) midpoint; half to the left).
   b) The shape (steepness) of the POD(a) curve.

COMPARING THE EFFECTIVENESS OF POD(a) 95% CONFIDENCE BOUNDS

[1] Perhaps surprisingly, this construct *IS* valid when the POD(a) curve is based on a valid *â vs. a* regression, with its constant error variance, since it is only a mathematical transformation from the valid Wald confidence bounds on the *â vs. a* regression in *â vs. a* space to the corresponding points in POD(a) space. Simulation studies also bear this out.

To compare the performance of the Wald and Loglikelihood ratio confidence bounds we simulated hit/miss experiments having sample sizes of 30, 45, 60 and 200. For each sample size, and for different combinations of POD parameter values, 1,000 experiments were simulated, using target sizes from 5 to 125 mils, uniformly distributed logarithmically. The same sizes were used for all simulations. (Another set of experiments used randomly spaced target sizes, but these were computationally problematic and are not reported here.) The sizes are immaterial, only that they cover a representative size range, since the POD parameters locate the POD curve relative to the size range.

How The Simulations Were Conducted

1. Hit/miss responses were simulated for each experiment based on the "true" POD curve and the target size.
2. The best-fit POD(a) model was determined for these "data" based on the maximum likelihood criterion.[2]
3. The Wald and loglikelihood ratio $a_{90/95}$ sizes were determined.
4. The computed $a_{90/95}$ values were compared with the "true" a_{90}. We want at least 95% of the $a_{90/95}$ values to be larger than the "true" a_{90} since that is what is meant by "95% confidence."

Figure 5 shows an example of a "true" POD(a) curve. The POD axis is Cartesian, but it is more informative to plot POD using a probability axis because it illuminates the high and low probability occurrences. The results of 1,000 simulations are shown in Figure 6, using a probability y-axis.

Figure 5 "True" POD(a) Curve Showing a range of target sizes from 5 to 125 mils, with POD Centered

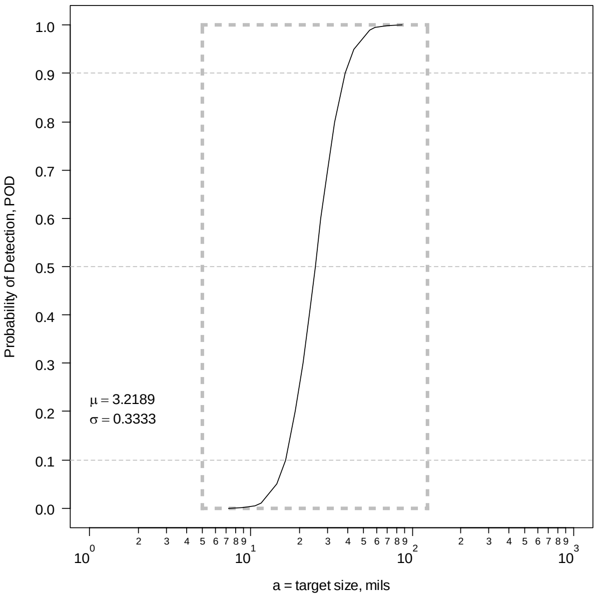


Figure 8 The Wald Method is Anti-conservative by about 3X while the LR Method is slightly conservative.

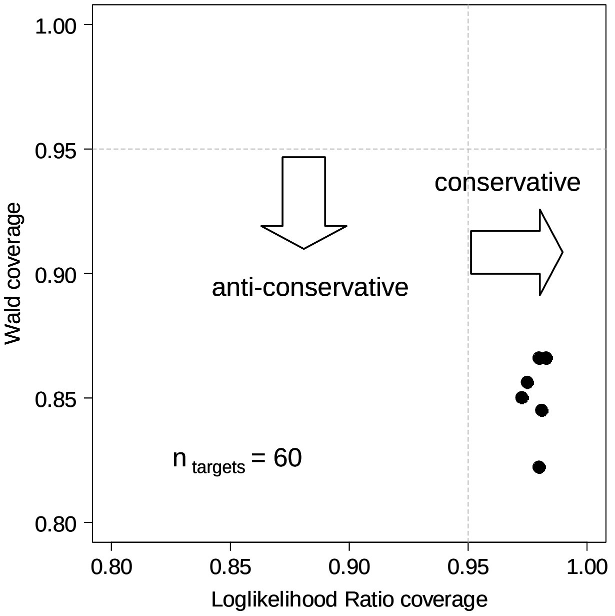
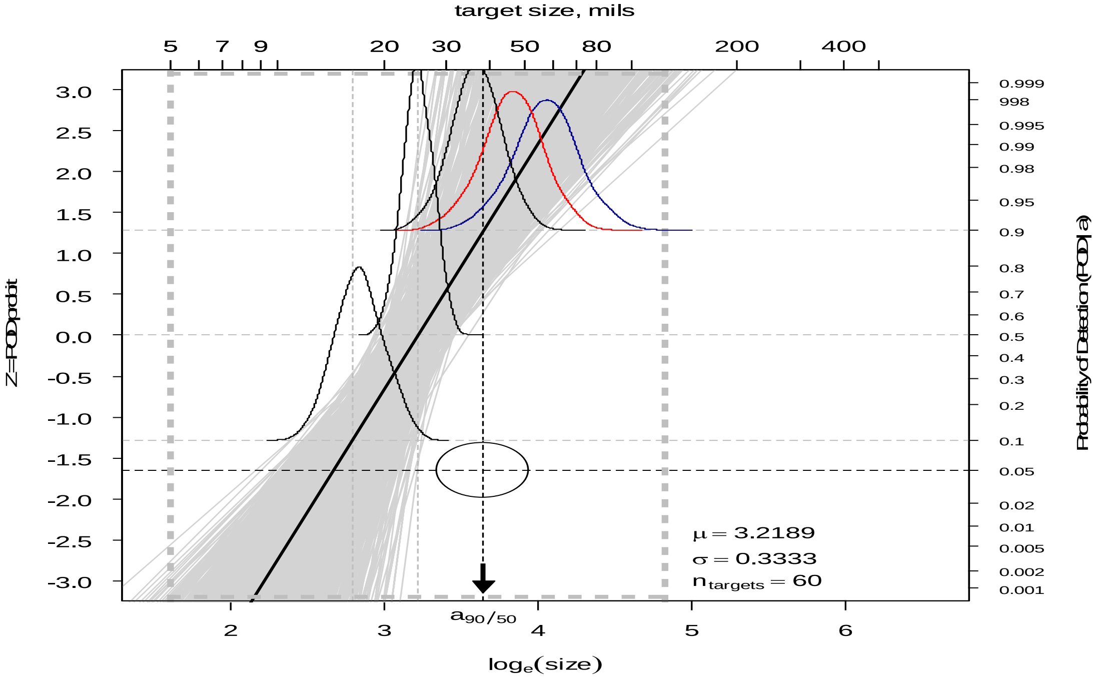


Figure 6 shows the "true" POD vs size relationship as the sold dark line. For reference horizontal dotted lines are drawn at POD = 0.05, 0.01, 0.5, and 0.9, and vertical dotted lines are drawn at a_{10} a_{50}, and at a_{90}, the target sizes corresponding to POD = 0.1, 0.5 and 0.9, respectively. The line at a_{90} is drawn darker since we are interested in how many of the predicted $a_{90/95}$ values

[2] The R software environment for statistical computing and graphics was used for all computations and plots herein. R is open-source (free) software and is available to download here: **`http://www.r-project.org/`**

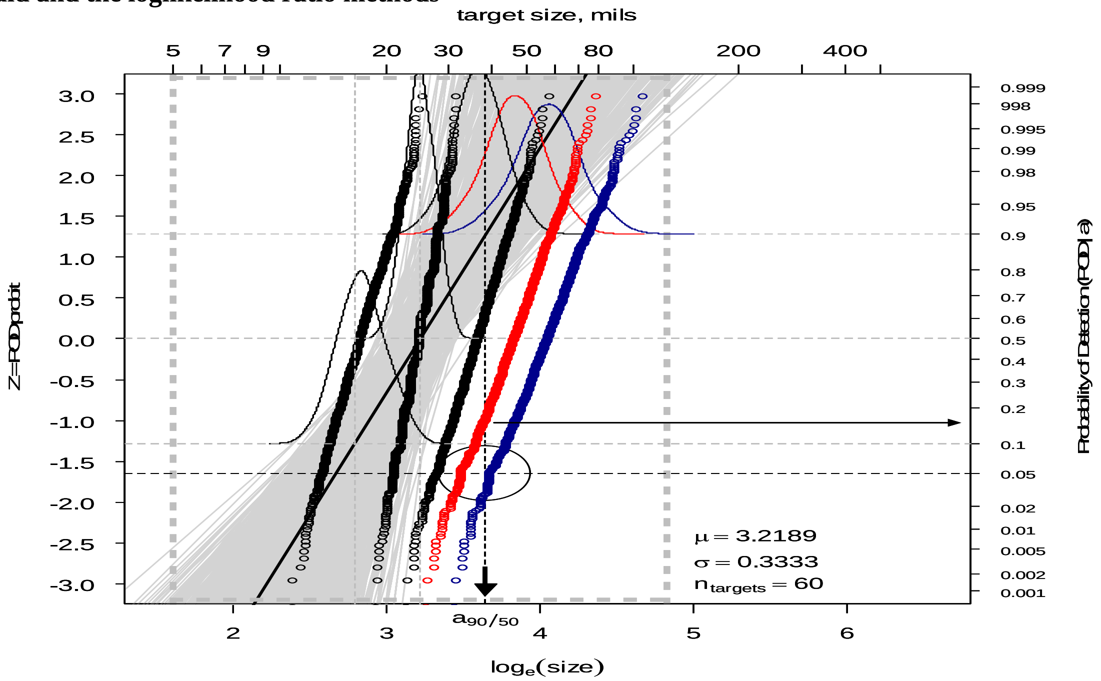
are smaller than a_{90}. For an effective 95% confidence bound, only 5% of those calculations should be less than the “true” value for a_{90}.

Figure 6 Representative Simulation results, Showing 1000 Best-Fit POD(a) Models based on the "True" Relationship (solid line)



We have 1,000 simulated values for a_{10}, a_{50} and a_{90} and their empirical densities are plotted in Figure 6 at their respective POD values. Also plotted at POD = 0.9 are the empirical densities for the Wald $a_{90/95}$ and the loglikelihood ratio value for $a_{90/95}$. For this example, $n_{targets}$ = 60, POD centered in the target sizes, and a “slope” of 3, we observe that a little less than 5% of the loglikelihood ratio calculations are less than a_{90}, so the loglikelihood ratio calculation is slightly conservative. About 18% of the Wald bounds are too small, however, meaning that the Wald calculation is anti-conservative by more than three fold. This is more easily seen in Figure 7.

Figure 7 Representative Simulation results, Showing 1000 Best-Fit POD(a) Models with a90/95 for both the Wald and the loglikelihood ratio methods



Since the right y-axis can also be used as a probability of occurrence axis, we plot on Figure 6 the 1,000 calculated values for $a_{90/95}$ for both the Wald and the loglikelihood ratio methods, producing Figure 7. Also plotted are the 1,000 values for a_{10}, a_{50}, and a_{90}. The horizontal lines and the heavy solid line use the right y-axis for POD; the points use the axis for cumulative probability of occurrence. The region of interested is highlighted as a circle centered at p=0.05 and a_{90}=true value, since we want 95% of the $a_{90/95}$ estimates to be at least this large.

The comparative performance of the two methods is summarized in Figure 8 for $n_{targets}$ = 60, and values for μ which place the "data" with 1/3, 1/2, and 2/3 of the sizes to the left.of μ, and values for σ of 1, 0.5, 0.333, representing shallow to moderate POD rise.

RESULTS and CONCLUSIONS

These results are not universal and depend on the sample size, distribution of target sizes, and the characteristics of the inspection system. For sample sizes of 200 (which are quite infrequent in practice) the coverage of the Wald bound is much closer to the nominal 95%. For sample sizes of 45 and 30 it is closer to 80%, which is very anti conservative since 50% is the coverage of the median POD line itself. But sample smaller than n = 60 are also plagued by difficulties in estimating the parameter values themselves and for that reason MH1823 has recommended at least 60 targets for hit/miss studies.

Comparing the Wald method with the LR method is not quite fair since the Wald calculation is for only a single point, $a_{90/95}$ (even thought it is routinely used to construct the entire 95% confidence bound) while the LR method is valid for the entire POD(a) curve, as was demonstrated in Figure 4. As a rule of thumb, for moderate sample sizes, bounds on the entire line are about 25% to 30% wider than single point bounds. (c.f. DeGroot, 1989)

FUTURE WORK

The Wald method is a good news/bad news story. The good news is that it is faster and computationally more stable than the loglikelihood ratio method that relies on finding an iterative solution for $a_{90/95}$. The bad news is that it's wrong. At least it is wrong much more often than its advertised 5%, often as much as 20% or more. (Remember, using the median value of a_{90}, will be too small 50% of the time, so an error of 20% is four times too large and unacceptable.) The errors are anti-conservative which compounds the problem.

REFERENCES:

1. Cheng and Iles, "Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables", *Technometrics*, Vol. 25, No. 1, February, 1983
2. Cheng and Iles, "One-Sided Confidence Bands for Cumulative Distribution Functions", *Technometrics*, Vol. 30, No. 2, May, 1983
3. De Groot, Morris, *Probability and Statistics*, 2nd Ed, Addison Wesley, 1989
4. Kutner, Michael, and Christopher J. Nachtsheim, John Neter, William Li, *Applied Linear Statistical Models*, 5th ed., McGraw-Hill/Irwin, 2005
5. Mood, Graybill and Boes, *Introduction to the Theory of Statistics*, 3rd, ed, McGraw-Hill, 1974
6. R Core Development Team (2006), R is a free software environment for statistical computing and graphics, `http://www.r-project.org/`