MEASURING DIFFERENCES AMONG PROBABILITY OF DETECTION CURVES

Charles Annis and Kimberly Erland

United Technologies
Pratt & Whitney
P.O. Box 109600
West Palm Beach, Florida 33410-9600

ABSTRACT

Probability of Detection (POD) curves are compared by two statistical methods to quantify system-to-system differences. The first method assesses performance among a group of inspection systems through an adaptation of statistical analysis of variance (ANOVA). The second method uses a chi-squared statistic to test for a difference between two systems. Examples using eddy current data are given for each technique.

INTRODUCTION AND BACKGROUND

For more than a decade the Air Force and gas turbine engine manufacturers have joined others in the NDE community in efforts to develop quantitative measures of nondestructive evaluation. Much of the broad scope of recent work in QNDE has been reported in annual symposia sponsored jointly by the Air Force and the Center for NonDestructive Evaluation at Iowa State University and can be found in Thompson and Chimenti, [1]. Annis [2] modeled probability of detection by comparing the apparent size of a crack, $\hat{a}$, with its actual size, a , and observing that the distribution of their quotient $\hat{a}/a$, was lognormal, and centered at $\hat{a}/a = 1$. Berens and Hovey [3] refined this crude approach by plotting $\hat{a}$ vs. a on log-log coordinates and noting a simple linear relationship between apparent and actual cracksize: $\ln \hat{a} = \alpha + \beta \ln a$. Furthermore, they noted that the residuals are normally distributed which permits describing the POD vs. a relationship as a cumulative normal function with mean θ_1 and standard deviation θ_2, where

$$POD(a) = 1 - Q\left[\frac{x - \theta_1}{\theta_2}\right] \tag{1}$$

and Q is the standard normal survivor function. The parameters θ_1 and θ_2 are related to the $\hat{a}$ vs. a relationship by

$$\text{mean } \theta_1 = \frac{\ln a_{th} - \hat{\alpha}}{\hat{\beta}}, \text{ and}$$

$$\text{standard deviation } \theta_2 = \frac{\hat{\sigma}}{\hat{\beta}}$$

where $\hat{\alpha}, \hat{\beta}, \hat{\sigma}$ are parameter estimates for the $\hat{a}$ vs a relationship, and a_{th} is the system threshold.

Now, as long as there existed some apparent cracksize, $\hat{a}$, for each actual size, a, Equation (1) allowed estimation of the POD function.

In any real inspection, however, some fatigue cracks may be too small to be detected by the inspection apparatus. The system output signal, $\hat{a}$, is not zero, it is just indiscernible from the noise. These misses have no associated $\hat{a}$ value and so are said to be left-censored. Similarly, cracks which are sufficiently large can overwhelm the system, resulting in a saturated signal. Again, the apparent size, $\hat{a}$, is unknown, other than that it exceeds some saturation level, $\hat{a}_{sat}$. These saturated observations are said to be right-censored.

Annis and Erland [4] and Berens and Hovey working together expanded the $\hat{a}$ vs. a analysis procedure to accommodate these censored observations. An Air Force gas turbine industry NDE reliability team has been assembled to formalize the entire methodology and produce a MIL-STD for POD analysis.

In addition to providing a statistically rigorous method for assessing POD, the normal formulation of equation (1) also suggests ways for measuring differences among POD curves.

Comparing POD Curves: ANOVA

Mathematical models of NDE capability are used to assess jet engine component reliability. In analyzing POD vs. a relationships, it is often desirable to compare results of differing systems such as changes in basic system configurations or operator-to-operator differences. In the past these assessments relied heavily on the well-known "inter-occular trauma test" for engineering significance*. We suggest two, perhaps less colorful but more pragmatic, statistical assessments.

An analysis of variance is possible by using the $POD(a)$ model location and scale parameters, θ_1 and θ_2. The within-column variance, and column-to-column variance are computed in the usual way, with the pooled within-column variance given as

$$s_W^2 = \frac{1}{(n_t - j)} \sum_j \theta_{2j}^2 n_j$$

where n_t is the total number of observations, and the column-to-column variance is

$$s_C^2 = \frac{1}{(j-1)} \sum_j n_j (\bar{\theta}_{1j} - \bar{\bar{\theta}}_1)^2$$

where $\bar{\bar{\theta}}_1$ is the grand mean of the j column means.

The ratio $F = s_C^2 / s_W^2$ is then formed to test for a difference among POD curve means. ANOVA comparisons can also be done with a censored data regression analysis, such as CENSOR (cf. Nelson, [5]). Since it is sometimes necessary, however, to compare POD curves for which the original raw $\hat{a}$ vs a data no longer exist, this method, and the one described in the following section, have proved useful.

ANOVA Examples

This procedure was used to assess the influence of the different operators, eddy current probes, and specimen orientations for a semi-automated inspection system. Annis and Erland presented results of 10 semi-automated eddy current inspections, reproduced here as Table 1, and used subsequently as examples.

---

* Data are plotted. If the resulting relationship hits you between the eyes, it's significant.

Table 1

Model Parameters for Semi-Automated Inspections

Test	a_{50}	θ_2	α	β	σ	n_i
A1	0.00498	0.2693	7.5271	1.4195	0.3822	30, 3, 2
B1	0.00526	0.2343	7.7306	1.4733	0.3452	30, 3, 2
B2	0.00489	0.2642	7.9070	1.4863	0.3926	30, 3, 2
B3	0.00473	0.3070	7.3941	1.3812	0.4240	30, 3, 2
C	0.00474	0.1968	8.4873	1.5859	0.3120	30, 3, 4
G	0.00484	0.2549	7.6671	1.4384	0.3666	30, 3, 3
H	0.00503	0.3070	7.7186	1.4585	0.4477	30, 4, 2
I1	0.00557	0.2379	7.7638	1.4956	0.3558	30, 4, 3
I2	0.00520	0.2012	8.2517	1.5691	0.3157	30, 3, 4
I3	0.00596	0.4662	7.2437	1.4142	0.6594	30, 6, 1

Notes:

1. $a_{50} = e^{\theta_1}$, cracksize at 50% POD.
2. Inspections A1, B1, B2, B3 are operator 1, repeat tests. Probe and system calibration, unchanged.
3. Inspection C changed probe.
4. Inspections G and H changed specimen orientations.
5. Inspections I1, I2, and I3 are operator 2, repeat tests.
6. n_1 = total observations, n_2 = data in noise, n_3 = saturations.

The inspections designated A1, B1, B2, B3 are repeated evaluations of the (unchanged) NDE system. The same operator performed all four inspections using the same eddy current probe. Next, the inspection probe, and therefore system calibration parameters were changed, and designated as inspection C. Inspections G and H changed the physical orientation of the fatigue-cracked specimens being inspected. All system parameters were identical to inspection C. Finally, to assess the human contribution to NDE variability, inspections I1, I2, I3 were performed by a new operator. Results are summarized in Table 1. A representative plot of the POD vs a relationship (Test A1) is provided as Figure 1.

Inspections A1, B1, B2, B3 were performed by the same operator using an unchanged NDE system. Engineers find it reassuring when their statistics corroborate their physics. One would not expect much difference in POD capability for an unchanged system. It is interesting, then, to note that $F < 1.0$ for inspections A1, B1, B2, B3 in Table 2, a reassuring result. In this case the repeated inspections are not so much independent samples from identical populations as recapitulations of the same sample.

The second inspection operator seems to have had difficulty with his final inspection, I3. This is evidenced by the behavior of the F-statistic ($F = 1.99$), computed from all groups collectively, which exceeds the critical value ($F_{0.95} = 1.88$). When inspection I3 is removed from consideration, the behavior of the remaining nine inspections is within expected variability. These results are summarized in Table 2.

Table 2

ANOVA Comparing Various Inspections

	A1, B1, B2, B3	All groups	All but I3
s_C^2	0.0592	0.1655	0.0851
s_W^2	0.0754	0.0832	0.0674
F	0.7855	1.9907	1.2880
$F_{0.95}$	2.68	1.88	1.94

Comparing POD Curves: χ^2

A related problem, but one requiring more stringent comparison criteria, is to determine if one system differs from another to any significant extent. As schematically illustrated in Figure 2, systems A and B have identical location parameters (means). Their scale parameters indicate a profound difference in system performance, however, with B having the greater capacity to discriminate between flaws larger, or smaller, than a_{50}. System B will detect most cracks larger than a_{50}, while ignoring smaller, harmless microstructural artifacts.

What is required is a test to see if a difference between the two curves can be detected with some degree of confidence. This can be accomplished using a χ^2 test which considers differences in both location and scale parameters. It is true that ML estimators, $\hat{\theta}$, have an asymptotically multivariate normal distribution with mean θ and variance-covariance matrix $[I(\theta)]^{-1}$ (cf. Kendall and Stuart, [6] or Cramer, [7]) and consequentially that

$$\Omega(\theta) = (\hat{\theta} - \theta)^T I(\theta)(\hat{\theta} - \theta) \tag{2}$$

is asymptotically a chi-squared variable with k degrees of freedom for a k-parameter model. The expected Fisher information for a two parameter normal model, is

$$I(\theta_1,\theta_2) = \frac{n}{\theta_2^2}\begin{pmatrix} k_0 & -k_1 \\ -k_1 & k_2 \end{pmatrix} \text{ where } k_o = 1,\ k_1 = 0,\ k_2 = 2 \text{ (Bury, [8])},$$

which upon substitution into equation [2] gives

$$\Omega(\theta_1,\theta_2) = \frac{n}{\theta_2^2}(\hat{\theta}_1 - \theta_1)^2 + \frac{2n}{\theta_2^2}(\hat{\theta}_2 - \theta_2)^2, \tag{3}$$

where n is the number of observations in the smaller sample, and $\hat{\theta}$ is the system with fewer observations.

χ^2 Examples

Assuming inspection A1 to represent baseline capability, θ , the chi-squared test was performed comparing it to the remaining nine inspection models. The contributions of deviations from the location and shape parameters, θ_1 and θ_2 , are given in Table 3 along with their sum, $\Omega(\theta)$, which was compared to a critical value of $\chi^2_{0.95} = 5.99$. (Cheng and Isles [9]) demonstrate that $\Omega(\theta)$ approaches its asymptotic behavior rapidly, with cdf error less than 0.005 for $(1 - \alpha) \geq 0.9$ and $n \geq 20$.)

Table 3

Chi-Squared Comparisons with Inspection A1

Test	$\Omega_{\theta 1}$	$\Omega_{\theta 2}$	$\Omega(\theta)$	
B1	1.238	1.013	2.251	
B2	0.138	0.022	0.159	*Note*: $\Omega_{0.95} = 5.99$
B3	1.097	1.176	2.273	
C	1.009	4.349	5.358	*
G	0.336	0.172	0.508	
H	0.041	1.176	1.217	
I1	5.186	0.816	6.001	*
I2	0.773	3.837	4.610	
I3	13.349	32.075	45.425	*

As with the ANOVA, the χ^2 test also indicates no significant differences among the first four system evaluations: baseline A1, and three repeat evaluations B1, B2, and B3.

Replacing eddy current probe, inspection C, did result in an improved POD capability, since the cracksize at 50% POD, a_{50} , is smaller than baseline, as is the scale parameter, θ_2 (see Table 1). Smaller θ_2 means that POD increases more rapidly through a_{50} than does the baseline. However, these differences are not significant at the 95% confidence level, as seen from Table 3.

Changing the orientation of the fatigue-cracked inspection specimens, evaluations G and H, had little effect (see Table 3).

The second inspection operator had less success than the first. Inspections I1, and especially I3, are worse (larger a_{50}) than baseline. According to Table 3 these differences are significant.

SUMMARY

Differences between and among POD curves can be assessed by capitalizing on the normal formulation of the *POD(a)* function, Equation (1). An ANOVA for comparing several inspection systems, and a more stringent chi-squared procedure have been described. Both tests can be used with censored data, and examples have been provided illustrating this.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Bruce Rasmussen, Thomas Cooper, and Joseph Moyzis of the Air Force Materials Laboratory, WPAFB, OH, and other members of the Air Force/Industry working group.

REFERENCES

1. Thompson and Chimenti, editors, (1982-1986) - Review of Progress in Quantitative Nondestructive Evaluation, Vol. I (1982), Vols. 2A,B (1983), Vols. 3A,B (1984), Vols. 4A,B (1985), Vols. 5A,B (1986), Plenum Press, New York.
2. Annis, C. et.al. (1979) 1st Executive Review "Concept Definition-Retirement-for-Cause of F100 Rotor Components", 1 August 1979.
3. Berens, A.P. and P.W. Hovey (1983), "Flaw Detection Reliability Criteria", Final Report, UDR-TR-83-137, - University of Dayton Research Institute.
4. Annis, C. and K. Erland (1987), "Estimating the Probability of Crack Detection from Data with Right- and Left-Censored Observations", submitted for publication, and offered as a working document for the Air Force NDE Reliability Team; to be presented at the American Statistical Association Annual Meeting, New Orleans, August 1988.
5. Nelson, W. (1982) - Applied Life Data Analysis, New York: John Wiley.
6. Kendall and Stuart (1961) - The Advanced Theory of Statistics, Vol. 2: Inference and Relationship, London: Charles Griffin.
7. Cramer, H. (1946) - Mathematical Methods of Statistics, Princeton, New Jersey: Princeton University Press.
8. Bury, K.V. (1975) - Statistical Methods in Applied Science, New York: John Wiley.
9. Cheng and Iles (1983) - "Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables", Technometrics, Vol. 25, No. 1, February, 1983.