2009

Reliability Goodness of Fit for Oil Spills in the Gulf of Mexico

William V. Harper
Otterbein University, wharper@otterbein.edu

Thomas R. James
Otterbein University, TJames@otterbein.edu

Ted G. Eschenbach
TGE Consulting,

Reliability Goodness of Fit for Oil Spills in the Gulf of Mexico

William V. Harper[1], Ted G. Eschenbach[2], Thomas R. James[1]
[1]Mathematical Sciences, Otterbein College, One Grove Street, Westerville, OH 43081
[2]TGE Consulting, 4376 Rendezvous Circle, Anchorage, AK 99504

Abstract

Eschenbach and Harper (2006) analyzed offshore oil spills in the Gulf of Mexico with extensions to the northern seas of Alaska. This involved multiple methods including assessing what statistical distribution adequately fits the data. Empirical distribution function (EDF) statistical procedures are powerful goodness of fit tests and also provide for good visual assessments. The most powerful of the current EDF methods is the Anderson-Darling test. This paper focuses on the Anderson-Darling EDF goodness of fit procedure for both the two and three parameter Weibull distribution that is often used in reliability analysis. Excel VBA code has been developed to compute this test statistic and also the associated p-values to allow statistical significance tests. The Excel routines are available free at http://faculty.otterbein.edu/WHarper/. The functions are illustrated with Gulf of Mexico oil spill data.

Key words: Anderson-Darling, Goodness of Fit, Weibull, Oil Spills

1. Introduction

Statisticians have long proposed various statistical distributions for data; however, there has been less enthusiasm about testing if such assumptions are true. Over time a multitude of what are called goodness of fit tests have evolved. Goodness of fit procedures are used to see how well a hypothesized distribution matches the available data. The null hypothesis H_o assumes that the proposed distribution adequately agrees with the data. Due to the vagueness of the alternative hypotheses, goodness of fit tests are performed backwards from most statistical hypothesis tests. In a typical statistical null hypothesis (H_o) one generally hopes to reject it to make some assertion at a given degree of confidence that a well defined alternative hypothesis (H_a) is valid. For example H_o may be that a given medical treatment is equivalent to a placebo whereas the alternative H_a states that the treatment is better than a placebo. In goodness of fit situations where a well defined H_a is not available, the setup is such H_o states that a specific distribution fits the data. However the alternative H_a vaguely states that the specific distribution does not agree with the data. It does not specify a specific other distribution. One hopes not to reject the null hypothesis H_o and thus accept the specified distribution as adequate. This may create confusion among practitioners who may be already struggling with statistical thinking in general.

The grandfather of goodness of fit tests is the venerable chi-square test. It is the method students are most likely to see in an introductory statistics course. Karl Pearson (1900) developed the chi-square statistic with the idea of reducing the general problem of testing goodness of fit to a multinomial setting by comparing observed cell counts to the expected frequencies dependent on the assumed distribution being evaluated. It has a

long history and has been well studied. It is flexible, but it is not a powerful statistical test.

A major reason the chi-square goodness of fit procedure is less powerful than other procedures is that the observed data are not treated as individual data points in the analysis but instead values are placed in bins (also called cells or groups). Then it is only known that a certain number of observations fall somewhere in the bin but they have lost their unique value. Additionally many issues arise on how best to select the bin boundaries. Most applications use bins of equal width rather than the statistically suggested practice of making the bins equally probable. Making cells equally probable takes time, and much more daunting it also requires different binning for each hypothesized distribution.

Some other goodness of fit measures are graphical in nature and take advantage of the human mind's ability to assess if the theoretical distribution passes the TLAR method. TLAR is an occasionally used engineering acronym for *That Looks About Right*. Graphical methods have been used for years and are a great communication tool for both statisticians and others. More recently graphical measures have been combined with more formal statistical techniques to provide both visual and quantitative evaluations.

Probability plots are a well grounded graphical method. The probability plot (as with any visual method) is best evaluated with a trained eye especially in the tails of the distribution. However, poor fits as well as excellent fits are obvious even to the newly initiated data detective.

Superimposition of a probability density function over the empirical histogram has long been a well established tool. It is a useful method but one that is visually harder to assess than seeing if the data fits on a straight line as done in the aforementioned probability plot. Perhaps its main utility is showing the shape of the distribution suggested by the observed data and assessing if a proposed theoretical model passes a TLAR test.

More recent developments in goodness of fit are called EDF procedures. EDF stands for the empirical distribution function. It is labelled $F_n(x)$ with its mathematical definition below where n is the number of observations.

$F_n(x) = \frac{number\ of\ observations \le x}{n}; -\infty < x < \infty$. Some label this ECDF for empirical cumulative distribution function. Whereas the probability density function (pdf) is compared to the observed histogram in the prior paragraph (with issues of binning for the histogram), the EDF does not bin and uses each data value separately. It compares the observed (empirical) cumulative distribution function to the theoretical cumulative distribution function. Generally it is easier to assess fit with the cumulative probabilities used in EDF procedures than the pdf vs. histogram approach of the prior paragraph.

2. Minerals Management Service Gulf of Mexico Oil Spill Data

Eschenbach and Harper (2006) studied both pipeline and platform oil spills in off-shore waters for the U.S. Department of Interior (DOI). Most of the data came from the Gulf of Mexico. One of the research objectives was to analyze the existing data and suggest way to migrate the findings to the northern slope of Alaska with the likely future growth of

off-shore drilling in areas such as Prudhoe Bay. The funding for the work was provided by DOI's Minerals Management Service (MMS) Alaska Outer Continental Shelf Region. Prior published work by Anderson & LaBelle (2000) examined spills of at least 1,000 barrels of oil. Eschenbach and Harper argued that while the consequences of smaller spill sizes may not have as much environmental impact as the larger spills, it was critical to analyze all available data so that the causes may be identified and measures to improve processes put into place to mitigate any future spills. Spills sizes down to 50 barrels are recorded in MMS spreadsheets. A database is not kept for oil spills less than 50 barrels. A barrel of oil is 42 U.S. gallons (approximately 159 litres or 35 U.K. Imperial gallons).

The resulting 36 pipeline spills over 1964 – 2005 were felt to provide a more comprehensive evaluation than the 16 spills of 1,000 barrels or more previously used as the basis for off-shore pipeline spill assessments. Nonetheless, there are unresolved issues still to be studied. For example, some of the Gulf of Mexico spills were caused not by human error but by hurricanes not found in the icy waters north of Alaska. Additionally potential new threats found in Alaska must be quantified including ice keel gouging, strudel scour, upheaval buckling, and thaw settlement. Thus as proposed in the research, only portions of the data analyzed may be used to assess the likelihood of spills on the northern Alaska slope.

The scope of the MMS work was broad and many topics were investigated, but as with many projects there were additional future areas for subsequent study and follow-on. As part of this project, Eschenbach & Harper fit statistical distributions to oil spill volume data. Both platform and pipeline spill data sets plus additional data sets were used in this new analysis to develop our Anderson-Darling goodness of fit Excel VBA software that is available free on the web.

Reliability is the likelihood than an item or a system will perform its intended function under specific conditions without failure for a given period of time. The Weibull distribution is one of the key distributions in reliability analysis. This is for multiple reasons including the flexibility of the Weibull distribution shape. It can represent what is often call the bathtub reliability curve that starts with decreasing failure rates during what is often termed infant mortality, then constant failure rate (bottom of the bathtub), and finally the right hand side of the bathtub with increasing failure rates due to wear out. Thus reliability is not a static probability, but instead is a dynamic assessment over time. While the Weibull plays a large role in reliability, its flexibility in shape from an exponential (a special case of the Weibull) to a near normal or lognormal has found the Weibull to be an adaptive distribution for statistical modeling of many real world variables.

One of the challenges of reading Weibull literature search is keeping track of both the Weibull parameter notation and the terminology. See Harper, James, Eschenbach, and Slauson (2008) for more details. Below are both the pdf and cdf formulation for our 3-parameter Weibull. In terms of the subsequent Minitab 15 plot labels α is the scale, β is the shape, and γ is the threshold. For a 2-parameter Weibull, γ is 0. Weibull distributions often underlie reliability evaluations.

$$pdf: f(x)=\beta\alpha^{-\beta}(x-\gamma)^{\beta-1}e^{-((x-\gamma)/\alpha)^{\beta}} \text{ for } x>\gamma, 0 \text{ otherwise. } cdf: F(x)=1-e^{-((x-\gamma)/\alpha)^{\beta}}$$

Figures 1 and 2 based on 36 pipeline spills illustrate the utility of the plotted EDF in assessing goodness of fit visually instead of the more common histogram versus the probability density function approach. Figure 1 is a fairly standard approach of comparing a histogram of the pipeline spills to a potential statistical probability density function for a 3-parameter Weibull. Visually it is difficult even for the trained statistical eye to accurate compare the thin line of the hypothesized distribution to the observed histogram. While one can see that both the observed data and the theoretical model are highly skewed with a long tail to the high values, it is hard to say much more.

Figure 2 is the EDF (or Empirical CDF) for the same 36 observed values versus the same theoretical 3-parameter Weibull cumulative distribution function. It is easy to see in Figure 2 where the smooth theoretical CDF curve deviates from the step function EDF. Figure 2 shows a reasonable fit between the observed data and the proposed 3-parameter Weibull though this will not be as obvious to those that have not used these methods especially for relatively small to moderate sized data sets.

Examining the cumulative probabilities (the vertical or y axis in Figure 2), it can be seen that the observed data (the EDF) represented by the step function deviates from the smooth theoretically proposed Weibull in the 60-80% cumulative probability range. Since the step function is to the right of the smooth theoretical curve in this range, this shows the actual data does not attain these cumulative probabilities as soon as the theoretical model. Similarly examine the horizontal or x axis for pipeline oil spills of 15,000 or larger. For the oil spill range depicted, the observed oil spills for 15,000 to 45,000 have cumulative probabilities somewhat higher than predicted by the theoretical smooth curve. In a related note since cumulative probabilities sum to 100%, the theoretical model has a higher probability of spills larger than 45,000 barrels than observed by the 36 spills.

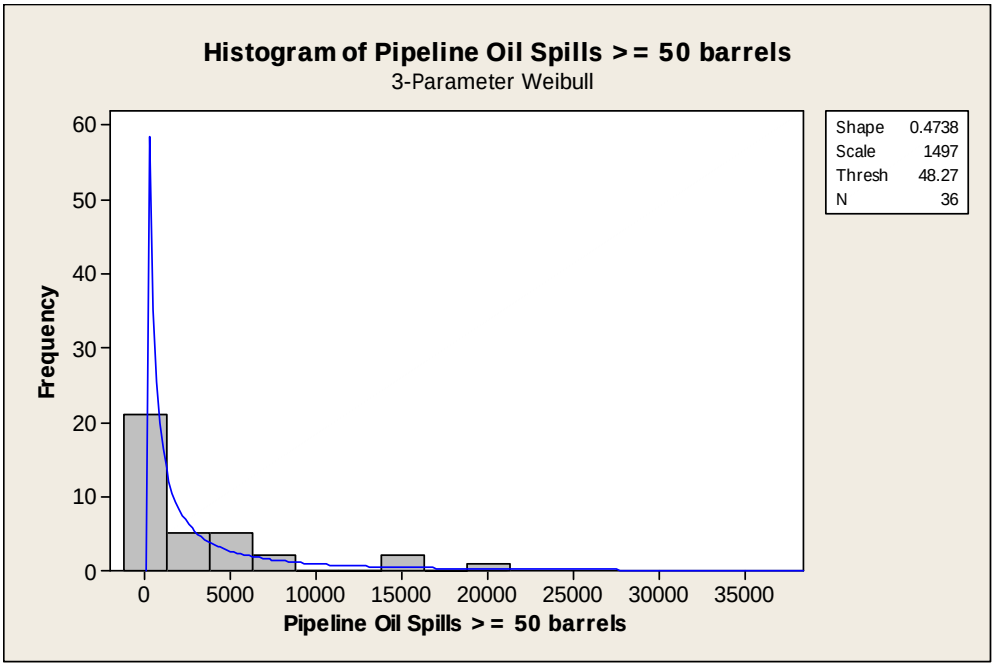

Figure 1: Gulf of Mexico Pipeline Oil Spill Histogram with Superimposed Theoretical 3-Parameter Weibull Probability Density Function

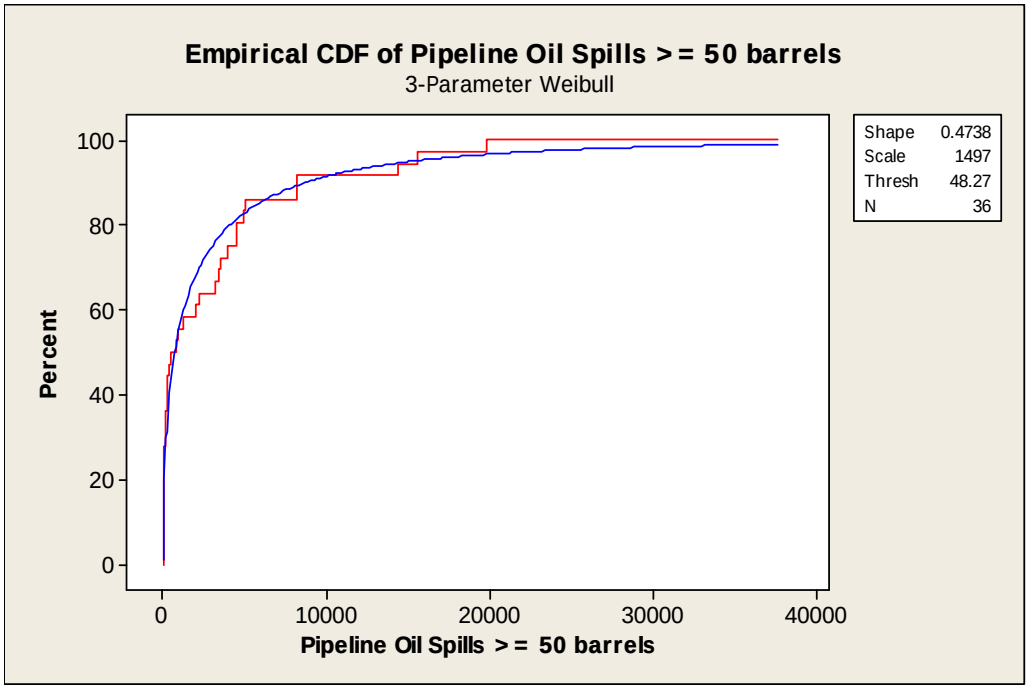

Figure 2: Gulf of Mexico Pipeline Oil Spill Empirical Distribution Function with Superimposed Theoretical 3-Parameter Weibull Cumulative Distribution Function.

EDF statistical procedures are statistically powerful and complement the visual images such as Figure 2 above. The most powerful of the current EDF methods is the Anderson-Darling test. This paper focuses on the Anderson-Darling EDF goodness of fit procedure for both the two and three parameter Weibull distribution. The estimation of the Weibull parameters used in our VBA software is based on the maximum likelihood estimation; however, the Anderson-Darling goodness of fit procedures are applicable regardless of the estimation methodology.

Let $X_1, \ldots, X_n$ random sample of size n with $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ as the order statistics. The cumulative distribution function (CDF) of X is F(x). The empirical distribution function (EDF) as explained earlier is

$$F_n(x) = \frac{\text{number of observations } \le \text{ x}}{n}; -\infty < x < \infty. \text{ In particular,}$$

$$F_n(x) = 0, \; x < X_{(1)}; \quad F_n(x) = \frac{i}{n}, \; X_{(i)} \le x < X_{(i+1)}, i = 1, \ldots, n-1; \quad F_n(x) = 1, \; X_{(n)} \le x.$$

$F_n(x)$ is a step function with height changes of 1/n based on the observed order statistics of the data. The EDF $F_n(x)$ is the proportion of observations less than or equal to x. The CDF F(x) is the probability of an observation less than or equal to x based on the assumed theoretical distribution. If the assumed theoretical distribution is correct, then $F_n(x)$ provides a consistent estimator of F(x).

3. Well known EDF Statistics

An EDF statistic measures the gap between $F_n(x)$ and $F(x)$ and is based on the vertical differences between $F_n(x)$ and $F(x)$. These statistics fall into either the supremum class or the quadratic class. The supremum statistic most well known is the Kolmogorov-Smirnov goodness of fit D statistic. The two EDF statistics used to compute D are D^+ and D^-. D^+ is the largest vertical difference when $F_n(x)$ is greater than $F(x)$. D^- is the largest vertical difference when $F_n(x)$ is smaller than $F(x)$. Mathematically, $D^+ = \sup_x\{F_n(x) - F(x)\}$ while $D^- = \sup_x \{F(x) - F_n(x)\}$. The Kolmogorov-Smirnov test statistic D is defined as $D = \sup_x|F_n(x) - F(x)| = \max(D^+, D^-)$. Kolmogorov (1933) and Smirnov (1939) are the original sources for Kolmorgov-Smirnov test though it is well covered in many text books such as Banks, Carson, Nelson, Nicole (2004).

The focus of this work is on the second class known as the quadratic EDF statistics. Much of what follows is based on D'Agostino and Stephens (1986). The quadratic class is based on the Cramer-von Mises family $Q = n\int_{-\infty}^{\infty}[F_n(x) - F(x)\}^2\psi(x)dF(x)$ where the function $\psi(x)$ weights the squared difference $\{F_n(x) - F(x)\}^2$. When $\psi(x) = 1$ the statistic is the Cramer-von Mises statistic. If $\psi(x) = [\{F(x)\}\{(1 - F(x)\}]^{-1}$ the statistic is the Anderson-Darling (1954) statistic, sometimes called A^2, or more commonly AD.

Computing formulas are based on using the Probability Integral Transformation. When F(x) is the true distribution of X, the random variable Z = F(X) is uniformly distributed between 0 and 1. This is an often used concept when generating pseudo-random numbers for many statistical distributions. Z has a Uniform (0,1) distribution function $F^*(z) = z$, $0 \le z \le 1$. Suppose that a sample $X_1, \ldots, X_n$ gives values $Z_i = F(X_i)$, i = 1, …, n, and let $F_n^*(z)$ be the EDF of the values z_i.

EDF statistics can now be calculated from a comparison of $F_n^*(z)$ with the uniform distribution for Z. For values z and x related by z = F(x), the corresponding vertical differences in the EDF diagrams for X and for Z are equal; that is,

$$F_n(x) - F(x) = F_n^*(x) - F^*(x) = F_n^*(z) - z$$

EDF statistics calculated from the EDF of the Z_i that is compared with the uniform distribution take the same values as if calculated from the EDF of the X_i, compared with F(x). The Anderson-Darling computational formulas involve the Z-values arranged in ascending order, $Z_{(1)} < Z_2 < \ldots < Z_{(n)}$, i.e., these are the Z order statistics. Both formulas give identical results though the second formula is easier to program.

$$AD = -n - (1/n)\sum_i (2i-1)[\ln Z_{(i)} + \ln\{1 - Z_{(n+1-i)}\}]$$
$$AD = -n - (1/n)\sum_i [(2i-1)\ln Z_{(i)} + (2n+1-2i)\ln\{1 - Z_{(i)}\}].$$

Since $Z_i = F(x)$, $Z_i = 1 - e^{-((x_i-\gamma)/\alpha)^\beta}$ for the three parameter Weibull where typically parameter estimates for α, β, and γ are also computed from the available data.

Once the AD statistic has been computed for a given set of data and a hypothesized Weibull distribution, a table can be used to see if the one-tailed test statistic is in the critical region. If the observed test statistic is in the critical region for a specified α level, then the Weibull distribution is rejected and found not to adequately fit the data. Stephens (1977) presents a table of AD critical values for the extreme value distribution. Since the natural log transformation of the Weibull is the minimum extreme value distribution, this table may also be used for the 2-parameter Weibull goodness of fit. This is also covered in D'Agostino and Stephens (1986). Lockhart and Stephens (1994) provide similar tabular results for the 3-parameter Weibull. Anderson-Darling critical values in both publications are based on Monte Carlo simulations. Lockhart and Stephens (1994) say the results are adequately accurate for samples sizes $n \geq 10$. For the 2-parameter case, Stephens (1977) provides a modified Anderson-Darling test statistic often shown in today's software as AD* in which AD* = $AD(1 + 0.2/\sqrt{n})$.

Minitab personal communication indicated that they compute the associated p-values by interpolation of the above tables. An unknown source developed the p-value formula below based on AD* which may be found in Romeu and Grethlein (2000) as well as other sources listed along with this reference. Interestingly, Dr. Romeu who also wrote the relevant section of MIL-STD-17 which is the first source we could find for this p-value formula states the following in a personal e-mail in 2009: "I just used the formula (which I imagine is an asymptotic result), and it works well under the assumptions it requires. This is as much as I can tell you, regarding this subject." If any reader can track the original source, please let us know.

$$p-value = 1/\{1+\exp[-0.1+1.24\ln(AD^*)+4.48(AD^*)]\}$$

We programmed the above Weibull based steps in Excel VBA and compared our AD, AD* and p-value results where possible to both Minitab 15 and Palisade's @Risk 5.0 for Excel. @Risk for Excel does not calculate p-values with its Anderson-Darling output. For the data sets tested and shown in the table below it is surprisingly found that while the p-value formula was developed for AD*, it is generally more accurate for AD. Hence in our VBA code, we output only AD and the associated p-value based on the formula modification shown below. We find this works well for either the two or three parameter Weibull. The source for each set of data is documented along with the data in the free web file.

$p-value$ that we recommend for $\mathrm{AD} = 1/\{1+\exp[-0.1+1.24\ln(AD)+4.48(AD)]\}$

Data set	*Sample Size, n*	*Weibull 2 or 3 parameter?*	*Minitab p-value*	*AD p-value using formula*	*AD* p-value using formula*
MMS Pipeline Spills	36	3	0.260	0.260	0.239
MMS Platform Spills	78	3	< 0.005	0.000114	0.0000915
Table 9	6	2	> 0.250	0.650	0.605
Table 7	6	2	> 0.250	0.435	0.380
MatLab	50	2	< 0.010	0.005	0.004
LS 1	10	3	> 0.500	0.647	0.612
LS 2	15	3	0.124	0.124	0.103
MTB by hand Weibull	10	2	0.071	0.086	0.067

4. Summary

After a brief background on statistical goodness of fit tests, the empirical distribution function (EDF) was introduced. Advantages of comparing the EDF to a proposed distribution's CDF versus the more traditional histogram to pdf were given. The adaptation of the most powerful EDF member, the Anderson-Darling AD statistic, to both the two and three Weibull distribution was documented. An unknown original source of a p-value approximation for the modified Anderson-Darling AD* statistic was tested on multiple data sets. It is recommended that this p-value approximation not be used on AD* but instead on the original AD statistic. A free web zip file at http://faculty.otterbein.edu/WHarper/ provides many Weibull analysis tools including Excel VBA functions for our recommended Anderson-Darling procedure.

References

Anderson, T.W., & Darling, D.A. (1952). "Asymptotic Theory of Certain 'Goodness of Fit' Criteria Based on Stochastic Processes", *Annals of Mathematical Statistics*, Vol 23, pp 193-212.

Anderson, Cheryl McMahon, and LaBelle, Robert P. (2000), "Update of Comparative Occurrence Rates for Offshore Oil Spills," *Spill Science & Technology Bulletin*, (Vol. 6, No. 5/6), pp. 303-321.

Banks, Jerry; John Carson; Barry L. Nelson; and David Nicol, (2004), *Discrete-Event System Simulation 4th* Prentice-Hall, ISBN 978-0131446793.

D'Agostino, Ralph B., and Michael A. Stephens (1986), *Goodness of Fit Techniques*, Marcel Dekker, Inc., New York. ISBN 0-8247-7487-6.

Eschenbach, Ted G., and Harper, William V. (2006), Alternative Oil Spill Occurrence Estimators for the Beaufort/Chukchi Sea OCS (Statistical Approach), OCS Study MMS 2006-059, http://www.mms.gov/alaska /reports/2006rpts/2006_059.pdf.

Harper, William V., Thomas R. James, Ted G. Eschenbach, Leigh Slauson (2008), "Maximum Likelihood Estimation Methodology Comparison for the Three-

Parameter Weibull Distribution with Applications to Offshore Oil Spills in the Gulf of Mexico," JSM Proceedings, Statistical Computing Section, American Statistical Association, Denver, CO, pp. 1231-1238.

Kolmogorov, A. N. (1933) "On the Empirical Determination of a Distribution Function," (Italian) *Giornale dell'Instituto Italiano degli Attuari*, 4, 83-91.

Lockhart, Richard A., and Stephens, Michael A. (1994), "Estimation and Tests of Fit for the Three-Parameter Weibull Distribution," *Journal of the Royal Statistical Society, Series B (Methodological)*, (Vol. 56, No. 3), pp. 491-500.

Pearson, Karl. (1900), On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling, *Philosophical Magazine*, **50**, 157-175. Available at http://www.economics.soton.ac.uk/staff/aldrich/New%20Folder/kpreader1.htm

Romeu, Jorge L. and Christian E. Grethlein (2000), A Practical Guide to Statistical Analysis of Material Property Data: A State of the Art Review, AMPT-14, Advanced Materials and Process Information Center (AMPTIAC). The p-value formula is also found in MIL-HDBK-17/1F (Composite Materials Handbook Volume 1 – Polymer Matrix Composite Guidelines for Characterization of Structural Material), 17-Jun-2002, Section 8.3.4.2.2. Additionally this formula may be found in web publications http://www.theriac.org/pdfs/startsheets/a_dtest.pdf, http://theriac.org/DeskReference/viewDocument.php?id=60, and http://www.mathworks.com/matlabcentral/fx_files/15745/1/AnDarWtest.m.

Smirnov, N. V. (1939), "On the Estimation of the Discrepancy Between Empirical Curves of Distribution for Two Independent Samples." (Russian) *Bulletin of Moscow University*, 2, pp. 3-16.

Stephens, M.A. (1977), "Goodness of Fit for the Extreme Value Distribution", *Biometrika*, Vol 64, No. 3), Dec, 1977, pp. 583-588.