



## STATISTICAL ANALYSIS OF MATERIAL DATA PART III: ON THE APPLICATION OF STATISTICS TO MATERIALS ANALYSIS

### Introduction

This is the last of three articles dealing with statistics and its applications to the data analysis of materials and components, as used in *Metallic Materials and Elements for Aerospace Vehicle Structures* (MIL HDBK 5) and the *Composite Materials Handbook* (MIL HDBK 17) [1,2]. The objective of this series is to discuss some ideas and philosophies underlying the use of statistical procedures that are usually not familiar to the practicing engineer. Statistics courses are often too crowded with methods and too busy explaining the how-to's, to discuss the why's and wherefore's.

In the first article of this series, we discussed some ideas dealing with random variables (R.V.), their distributions and their parameters. In the second article, we discussed some problems dealing with estimation and testing of statistical distributions and parameters, when they were unknown but estimated from a random sample. In this last article we apply some of the material discussed in the first two, in the context of statistical analyses included in MIL HDBKs 5 and 17. We hope the series will generate an on-going dialogue, through the AMPTIAC Newsletter and Web Page, where further questions and topics of interest to statistics practitioners in the area of materials data analysis may be discussed.

The whole intent of using statistical data analysis in this context stems from the need to establish types A and B tolerances for materials properties (i.e. estimates of the upper/lower first and tenth percentiles of the distribution of a population characteristic of interest). To obtain them we need to analyze samples from some material, which may come from a single source or from multiple sources or batches. Then, we need to establish the property's underlying statistical distribution, its parameters and finally to estimate the required property tolerances, according to the specific statistical model. How we implement this is the subject of the present article.

In the rest of this article, we discuss statistical procedures in [1, 2] using as guide Figure 8.3 in page 8-20 of Reference 2, (denoted as Figure 1). We discuss how, whether the data come from a single batch or whether there are two or more batches,

these are tested for potential outliers. We then see how the outliers are removed from the sample, if necessary. If there are two or more batches, we assess whether these can be pooled together (e.g. if they come from the same population). Otherwise, the desired tolerances must be obtained separately, on each individual batch. Then, whether analyzed individually or pooled, the samples need to be tested for Goodness of Fit (GoF) for three statistical distributions: Weibull, Normal and Lognormal. Finally, and once having determined the underlying distribution, we apply the corresponding method of A or B basis tolerance estimation to obtain the corresponding A or B basis allowable. Conversely, we apply nonparametric methods if neither of the above mentioned distributions fit the data.

### Establishing the Underlying Distribution and Parameters

An A or B basis allowable of a material property is an estimation of  $(\gamma_0)$ , the lower/upper first or tenth percentile of all the population values of the property. This means, with probability 0.95, ninety nine percent (A basis allowable) or ninety percent (B basis allowable) of all population values are smaller/larger than the estimate of percentile,  $\gamma_0$ . A and B allowables depend on the specific statistical distribution of the parameters of the population in question. Therefore, the estimation, with high probability and accuracy of both the underlying distribution and the corresponding parameters of the population from which the materials sample was obtained is very important. If there is a serious estimation error in this initial procedure, everything else that we do (since it is based on this) will be wrong.

In the first article we saw how  $F(x)$ , the Cumulative Distribution (CDF) Function and  $f(x)$ , the probability density function (pdf), are related to each other via:  $F(x) = \int_{-\infty}^x f(t)dt$ . Hence, two types of GoF tests exist to assess the composite hypothesis ( $H_0$ ) that a completely specified distribution  $F_0(x;\theta)$  fits a data set. One type of test compares the actual (observed) number of sample points with the corresponding expected number, obtained under the (hypothesized) pdf, for subsequent data intervals. An example of such tests is the Chi Square GoF test. The other type compares (vertical) distances between empirical,  $F_n$ , and theoretical,  $F_0$  CDF values, for the ordered sample

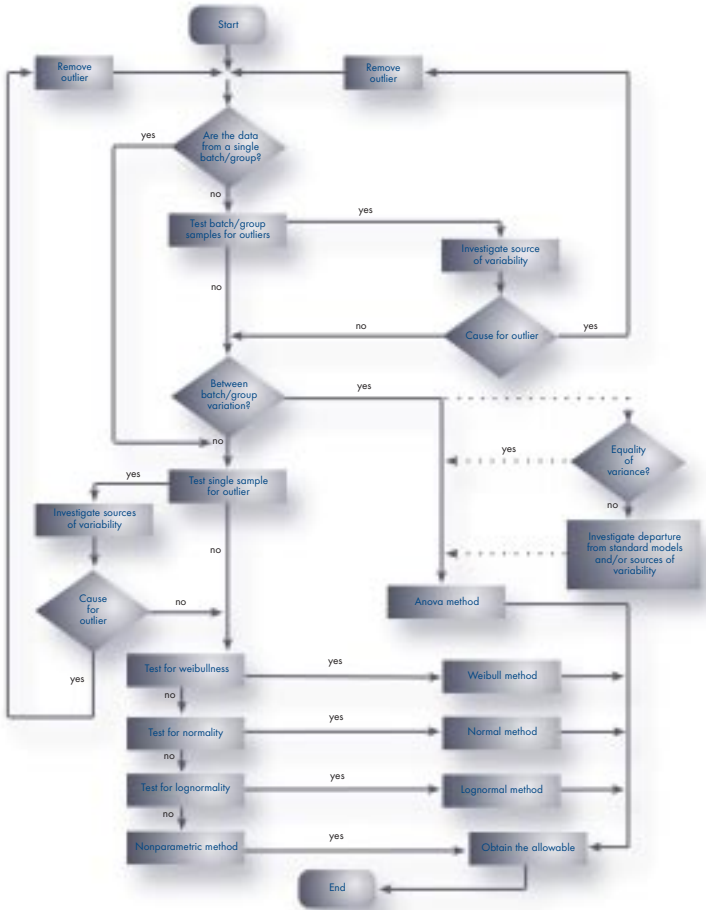


Figure 1. Computational Procedure for B-Basis Material Allowables

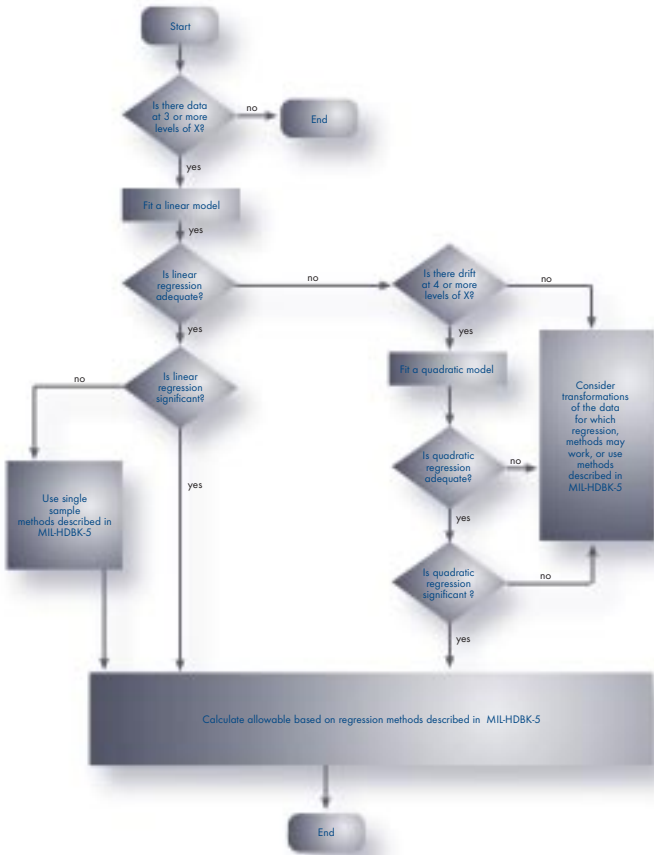


Figure 2. General Procedures for performing a regression analysis in order to calculate design allowables

# Material

## E A S E

points. An example of this type is the Anderson and Darling (A-D) test. Both of these approaches assume that the data come from a completely specified, continuous distribution  $F_0$ , with known parameter  $\theta$ . However, both these GoF approaches allow for the case when the distribution parameters are unknown, and need to be estimated from the sample, which is the most usual case in practice.

It is important to understand that such (composite) hypothesis  $H_0$ , when rejected, may imply more than one alternative possibility. For example, when we reject  $(H_0)$  the hypothesis that a data set comes from the Normal distribution, with specific mean  $\mu$  and variance  $\sigma^2$ , it may occur that: (i) the distribution is not Normal, even when the mean and variance may be the ones stated; (ii) the distribution is indeed Normal, but the mean, or the variance, or both, are not the ones stated in  $H_0$ ; (iii) none of the stated assumptions, i.e. neither the distribution nor parameters, are as assumed in  $H_0$ . It is also important to remember that, when  $H_0$  is not rejected, we have not found enough grounds to question the validity of  $H_0$  (the assumptions made) and hence we assume  $H_0$  is correct. The A-D GoF test, for one [3] or several samples [4], is highly regarded among univariate GoF tests. Its asymptotic distribution (i.e. for large sample sizes) has been thoroughly studied. The sample sizes required in the statistical procedures discussed in [1, 2] are large enough for the use of these asymptotic distributions.

Following Figure 1, we first screen each sample individually for potential outliers. This is done using the MNR test, which singles out such unusually high/low observations in the data set. The outliers must then be checked for accuracy (clerical errors) or experimental implementation (materials test procedures) problems. If some such problems are detected, the error must be corrected or the data point discarded. If no error is ascertained, the data point must be retained in the sample. If there is more than one sample, the MNR procedure is first implemented in each and then in the combined sample.

Three statistical distributions are tested for GoF. The Weibull is tested first. It is justified for theoretical reasons in the derivation of materials properties and by a long practice. Its shape and scale parameters are estimated from the data. If the A-D test rejects that the data come from the Weibull distribution, the Normal distribution is then tested. If the A-D GoF test also rejects that Normal is the underlying distribution, then the data is tested for Lognormal. If all three mentioned distributions fail to fit the data set, a nonparametric method must be implemented to obtain the allowables. However, if the sample size is less than 29, the Hanson-Koopmans method must be used.

If working with a single sample, the above procedure will be

implemented with the standard A-D GoF test. If working with more than one, the k-sample A-D GoF test is implemented to assess the hypothesis  $(H_0)$  that all samples (batches) come from the same population. In the affirmative case, we pool all the batches into a single, combined sample from which we obtain the desired allowables. If A-D rejects  $H_0$  then different allowables must be obtained for each batch, using ANOVA [5] methods. We need more than two batches when implementing the ANOVA procedures. If only two batches are available, we must assess whether they can be pooled together or we must wait for additional data and form three or more batches.

Finally, if the property of interest is associated with other (predictor) measurements, then ( $\delta$ ) regression methods can be employed. One must first verify that the regression model assumptions (i.e. independence and identically distributed observations, linearity, normality) are met. If so, we can obtain the model parameter estimates. We must also check the model appropriateness. If the general linear model (GLM) is applicable then either the ANOVA or the regression procedure will provide the desired allowables. We use the sample batch (in the ANOVA case) or the covariant (predictor) setting (in the regression case).

### The Case of Bivariate Data

We have seen that we can work with a single or with several samples (batches). If data come from the same population (i.e. the A-D GoF test does not reject  $H_0$ ) we can pool samples and obtain the allowables for the combined data set. However, if A-D rejects  $H_0$  (that all samples come from the same population) these become bivariate data. Now, each data point provides two pieces of information: one is its materials property measurement and the second, its batch or grouping number.

ANOVA is the procedure used to establish whether k batches of n elements each have the same mean, or whether the group means differ. The assessment is made via comparing two estimates of the variance. One estimate is obtained using the variance estimator within groups. The other is obtained using the variances between groups. If all k group means are equal, then these two variance estimators are close (for both estimate the same parameter) and their ratio is unit. If means differ, the ratio of these two variance estimators is statistically different from unit. From here, we get the ANOVA or Analysis of Variance model:

$$y_i = \mu + \alpha_i + \epsilon_i; 1 \leq i \leq n; 1 \leq j \leq k$$

where  $\alpha_i$  is the contribution of the ith sample (group) to the general mean  $\mu$ , and  $\epsilon_i$  is the error term which is distributed normally, with mean 0 and variance  $\sigma^2$ . Under  $H_0$ , all group means are equal, hence all  $\alpha_i = 0; 1 \leq j \leq k$ .

One crucial ANOVA assumption is that all group variances are equal. This assumption must be tested before implementing the ANOVA procedure. If the test fails (i.e. there is reason to believe that not all groups have the same variance  $\sigma^2$ ) then data transformation or other procedures must be implemented before or instead of ANOVA [7].

Another important ANOVA consideration is the number of data points ( $n_j$ ;  $1 \leq j \leq k$ ) per group. ANOVA works better under "balanced" designs (i.e.  $n_j = n$ ). This means that all  $k$  groups should be of equal size  $n$ . For example, think of the sample size  $n$  as the amount of information, of the  $k$  groups as informants and of the statistical test as an assessment procedure based on information provided by  $k$  different informants. Optimally, we would like to give equal weight to all informants' contribution and not receive more information from some (possibly biased) informants, over the others.

In practice, samples are often of different sizes. To correct this problem we use "effective" sample sizes ( $n^*$ ) obtained via the formula:  $n^* = (N \cdot n^*) / (k \cdot 1)$ ; where  $n^* = \sum n_j^2 / N$ ;  $N = \sum n_j$  and  $1 \leq j \leq k$ . When  $n_j = n$  (i.e. all groups have the same size) then  $n^* = n$ . However, when  $n_j \neq n$  (group sizes differ) then  $n^* < n$  (i.e. the test procedure is less efficient, for the samples are suboptimal). In statistical analysis we strive to obtain the most efficient and unbiased assessment (test) from the data (information).

In ANOVA, the bivariate data is categorical (qualitative). Each data point  $P_i = \{Y_i, j\}$  yields both the materials property measurement  $Y_i$  and its corresponding group  $j$ . The group is not quantitative. When both measurements are quantitative, i.e. when  $P_i = \{X_i, Y_i\}$ , the association between the two variables can be established in a better way: via regression.

Assume now that two quantitative measurements,  $X_i$  and  $Y_i$ , are obtained from each data point  $P_i$ ;  $1 \leq i \leq n$ . And that  $X_i$  is easier, faster, cheaper or more accurately obtained than  $Y_i$ , if  $X$  and  $Y$  are associated, we can use such relation to obtain an alternative or improved estimation of  $Y_i$  through  $X$ . This is the idea behind regression analysis.

We start by plotting  $Y_i$  vs  $X_i$  for each  $1 \leq i \leq n$ . If there is no association between variables  $X$  and  $Y$  (null hypothesis  $H_0$ ) then the resulting set of points  $P_i = \{X_i, Y_i\}$ , will be uniformly and randomly scattered all over the plane. If we draw two lines (one vertical through the average of the projections over the X-axis; one horizontal, through the average of the projections over the Y-axis) then we divide the plane into four quadrants. Under  $H_0$ , the set of points  $P_i$  will be equally and randomly distributed among these four quadrants.

On the other hand, if there is an association between  $X$  and  $Y$  (i.e. if we reject  $H_0$ ) the number of points in each quadrant will dif-

fer. If there is a positive association (i.e. when  $X$  increases/decreases, so does  $Y$ ) then the points will tend to cluster in the upper right and lower left quadrants. If there is a negative association between  $X$  and  $Y$ , the points will cluster in the upper left and lower right quadrants. The indicator "covariance between  $X$  and  $Y$ ," characterizing such relationship, is defined as:  $\text{Cov}(X, Y) = S_{xy} = \sum \{x_i - \bar{x}\} \{y_i - \bar{y}\} / (n-1)$ ; where  $\bar{x}$  and  $\bar{y}$  denote the corresponding sample averages. The covariance indicator is positive when a positive association between  $X$  and  $Y$  exists; negative when a negative association exists and zero if no association exists.

As a measure of association between two variables, the covariance is difficult to interpret because it depends heavily on the units in which variables  $X$  and  $Y$  are being measured. The correlation coefficient, defined as  $r_{xy} = S_{xy} / S_x S_y$  (where  $S_x$  is the sample standard deviation of variable  $x$ ) is a "normalized" covariance. It measures the association between  $X$  and  $Y$  just as the covariance does. However,  $-1 \leq r_{xy} \leq 1$ , is "dimensionless" and easy to interpret. In addition,  $r_{xy}$  is a measurement of "linear" association between  $X$  and  $Y$ . Hence, if  $r_{xy} > 0$  (and close to unity) there is a "linear" trend that models the association between  $X$  and  $Y$ , with positive slope. If  $r_{xy} < 0$  (and close to -1) this linear trend has a negative slope.

It is therefore very useful to obtain an estimate of such a linear trend (called the linear regression) when it exists. The linear regression is then used to obtain a better estimate of  $Y$  (the dependent variable) given a value of  $X$  (the predictor). In mathematical terms:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i; \quad 1 \leq i \leq n$$

Such is the scheme for simple linear regression. The multiple regression model is just an extension of the above, when there are two or more "predictor" variables,  $X_1, X_2, \dots$ .

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \epsilon_i; \quad 1 \leq i \leq n$$

As before with ANOVA,  $\epsilon_i$  is the error term which is distributed normally, with mean 0 and variance  $\sigma^2$ . The  $\beta_j$ ;  $0 \leq j \leq k$ , are called the regression coefficients. Figure 9.6.3, "General Procedures for Performing a Regression Analysis in Order to Calculate Design Allowables" (on page 9-209 of Reference 1) will be used here (denoted as Figure 2) to explain this regression analysis process, in the context of obtaining A and B basis allowables.

It is important to notice how, given an adequate sample of any R.V., it is always possible to obtain an estimate of its parameters (including A and B basis tolerances). However, if one R.V. (say  $Y$ ) has a strong association with another (say  $X$ ) we can use this extra information to obtain an improved estimation of the parameter for  $Y$  (i.e. one with a smaller variance).

In both regression and ANOVA models, the allowables for the dependent variable  $Y$  are obtained by "improving" the estimation via the information added by the predictor variable  $X$ , in Regression

or by the grouping into "treatments," in ANOVA. Such methods allow us to obtain more refined A and B basis allowables than we would have, had we obtained them directly from the dependent variable Y.

Regression analysis uses three or more ( $k \geq 3$ ) levels of measurements for the predictor variable X. If there are less than three levels in the measurements of X, we cannot proceed until more data (levels) are gathered. If there are enough levels, we can fit a Linear Regression. The dependent (or response) variable Y for which we are deriving the allowables, is now a function of the k predictor or independent variables,  $X_1, \dots, X_k$ .

Such a regression model is adequate if it is statistically significant, i.e. if the F-test rejects the null hypothesis ( $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ ). Not all model independent variables, however, need be statistically significant (i.e.  $\beta_j \neq 0$  for some j). Some predictor variables ( $X_i$ ) may be highly significant (i.e. have a coefficient  $\beta_i \neq 0$ ) while others may be redundant (i.e. not significant or  $\beta_i = 0$ ). Nevertheless, the regression (equation) model as a whole may remain statistically significant. To handle these situations, we use variable selection methods. Through them, redundant variables are weeded out and the resulting regression model improves.

If there exist four or more levels of measurements for the independent variable (X) then it is possible to fit a quadratic regression model to the data:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i ; 1 \leq i \leq n$$

If both the linear and the quadratic regression models are statistically significant, we need to compare them to assess which one is more adequate. The better of the two regressions is then retained and the A and B basis allowables are calculated with it.

Regression models are sensitive to violations of their statistical assumptions (e.g. normality, equality of variance). These assumptions must be checked before the model can be correctly used. Analysis of regression residuals ( $\epsilon$ ) allow us to verify the normality of the data. One can use the A-D GoF test or graphical methods (e.g. Q-Q plots). If the normality assumption is rejected, data should be transformed [7] and the regression model recalculated.

The assumption of equality of variance  $\sigma^2$  can also be checked via the analysis of residuals ( $\epsilon$ ) either using statistical tests such as Bartlett's or through graphical analysis. If any of these procedures indicate that variances are not equal, the data should be transformed and the regression model should be recalculated because regression models are based on two procedures. First, an optimization process selects a function such that the sums of the squares of the distances to each data point ( $\sum \epsilon_i^2$ ) is minimum. Second, a statistical model (i.e. distributional assump-

tions) is imposed on such distances ( $\epsilon_i$ ). If an invalid regression model is used (one where the assumption of independence, normality and equality of variance of the  $\epsilon_i$  is not met) then the confidence levels and confidence intervals derived (which are the basic statistical contribution) are no longer those provided by the regression model.

For example, the point estimator for  $Y_i$  (given  $X_i$ ) is always valid (due to the optimization part of the regression procedure). However, if there are violations of the distributional assumptions of the residuals, the confidence (interval) estimation for  $Y_i$  and the probabilistic statements (tests of significance for the regression coefficients) are no longer exact or valid.

## Summary

We cannot, nor do we intend to reduce the extensive statistical theory associated with these analyses to three short review articles. We have only tried to discuss some of the ideas and concepts behind the derivation and use of several statistical procedures in MIL HDBKs 5 and 17. We have done this to facilitate and encourage their use by practicing engineers who need to deal with them on a regular basis.

We strongly believe that, by enhancing the practitioner's understanding of statistics and its underpinnings, we will encourage a more frequent and better use of statistical methods. We hope to have contributed toward such an objective among our newsletter readers.

## Bibliography

1. Metallic Materials and Elements for Aerospace Vehicle Structures MIL HANDBOOK 5G. November 1994.
2. Composite Materials Handbook MIL HANDBOOK 17. 1D
3. Anderson, T.W. and D. A. Darling. "A Test of Goodness of Fit." JASA. Vol. 49 (1954). Pages 765-769.
4. Scholz, F. W. and M. A. Stephens. "K-Sample Anderson-Darling Tests." JASA. Vol. 82 (1987). Pages 918-924.
5. Box, G.E.P., W. G. Hunter and J. S. Hunter. Statistics for Experimenters. John Wiley. NY 1978.
6. Draper, N. and H. Smith. Applied Regression Analysis. John Wiley, NY. 1980.
7. Dixon, W. J. and F. J. Massey. Introduction to Statistical Analysis. McGraw Hill. NY. 1983.

*Editors note:* The entire three part series entitled "Statistical Analysis of Materials Data" is available as a PDF file which can be downloaded from the AMPTIAC website.

ADVANCED MATERIALS AND PROCESSES TECHNOLOGY

EMAIL: [amptiac@iltri.org](mailto:amptiac@iltri.org)  
<http://amptiac.iltri.org>

PHONE: 315.339.7117  
FAX: 315.339.7107

