JOURNAL OF AIRCRAFT
Vol. 32, No. 4, July–August 1995

Near Real-Time Approach to Statistical Flight Test

Brian L. Jones*
U.S. Air Force Institute of Technology, Wright–Patterson Air Force Base, Ohio 45433

A near real-time approach to statistical flight test is proposed that minimizes the number of test condition executions required to adequately describe system performance relative to a specified performance goal or specification. Four different statistical intervals, all based on the Student-*t* distribution and its generalization, are summarized and applied to aircraft flight test scenarios. A graphical technique is developed with applicability to near real-time mission operations. Finally, probability statements are associated with sample set statistics and applied to performance goals or specification compliance.

Introduction

PERFORMANCE flight test of aircraft systems, and some system components, is inherently statistical. Data from multiple executions of a given test condition are required to adequately characterize system performance. Examples of system performance that fall into this type of flight test are terrain following at a specified altitude and air-to-air radar target acquisition; system component evaluation would include radar altimeter and laser ranging sensor testing. Two natural questions arise from this type of flight test:

1) What is the minimum number of test condition executions required to adequately describe system performance?

2) Given a specified number of test condition executions and the sample set statistics, what is the probability (or confidence) level that the data collected represents the true system performance?

This article provides one approach to answering these questions based on statistical intervals. Furthermore, the approach given has a near real-time application for flight test mission control. The mathematical background of statistical intervals is first presented, followed by the development of a graphical technique applicable to near real-time operations. Finally, a simple example is given to demonstrate this technique.

Mathematical Development

A large amount of statistical analysis is based on the normal (or Gaussian) probability density function. Whereas the characteristics of the normal distribution are desirable for flight test applications, the assumption of an infinite sample size (or even very large) cannot be typically satisfied. However, the Student-*t* probability density function (and its generalizations) maintains the advantages of the normal distribution, but relieves the requirement for a very large sample size by adding degrees-of-freedom ν to the probability density function, where degrees of freedom (DOFs) are defined as one less than the sample set size n. The sample mean $\bar{x}$ and standard deviation s_x are defined in the usual way:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

$$s_x = \left[\frac{1}{\nu}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2} \tag{2}$$

Theoretically, with an infinite sample size, the Student-*t* distribution and the normal distribution are equal. With 30 DOFs, the maximum difference between the one-dimensional normal distribution and the Student-*t* distribution is 1.13% of the maximum value of the normal distribution; hence, the Student-*t* distribution with $\nu = 30$ will be considered equivalent to the normal distribution.

Consider a flight-test condition that has an associated upper U and/or lower L performance goal or specification. If the system performance is to meet either one or both of these goals, then the chosen criterion J must satisfy one of the following statements:

$$L \le J \le U \tag{3}$$

$$J \le U \tag{4}$$

$$L \le J \tag{5}$$

Which statement to be satisfied is a function of which of the performance goals are applicable to the flight test condition. Equation (3) applies if both upper and lower performance goals are relevant; Eqs. (4) and (5) are appropriate if only one of the upper or lower performance goals are applicable. The notation

$$L \le J \quad \text{and/or} \quad J \le U \tag{6}$$

will be used hereafter to indicate any set of equations of the form like Eqs. (3–5).

Selection of the appropriate criterion is highly dependent on the flight test considered. Hahn[1,2] summarizes six statistical intervals for a normal population that lend themselves to many applications, including aircraft flight test. Four of these intervals are taken as criterion in this article:

1) *Prediction interval to contain the mean of a future sample set with r samples* ($\bar{y}_r$). The interpretation of this interval is that in a large number of repeated experiments where this interval is computed, the claim that the interval contains the true $\bar{y}_r$ would be correct $100(1 - \alpha)\%$ of the time.

2) *Confidence interval to contain the population mean* (μ). The interpretation of this interval is that in a large number of repeated experiments where this interval is computed, the claim that the interval contains the true μ would be correct $100(1 - \alpha)\%$ of the time.

3) *Tolerance interval to contain a proportion of the population* (Z_p). The interpretation of this interval is that in a large number of repeated experiments where this interval is computed, the claim that the true proportion of the population contained in the interval would be at least p would be correct $100(1 - \alpha)\%$ of the time.

Presented as Paper 94-3511 at the AIAA Atmospheric Flight Mechanics Conference, Scottsdale, AZ, Aug. 1–3, 1994; received Aug. 30, 1994; revision received Dec. 18, 1994; accepted for publication Jan. 10, 1995.

*Assistant Professor of Aerospace Engineering, Department of Aeronautics and Astronautics. Senior Member AIAA.

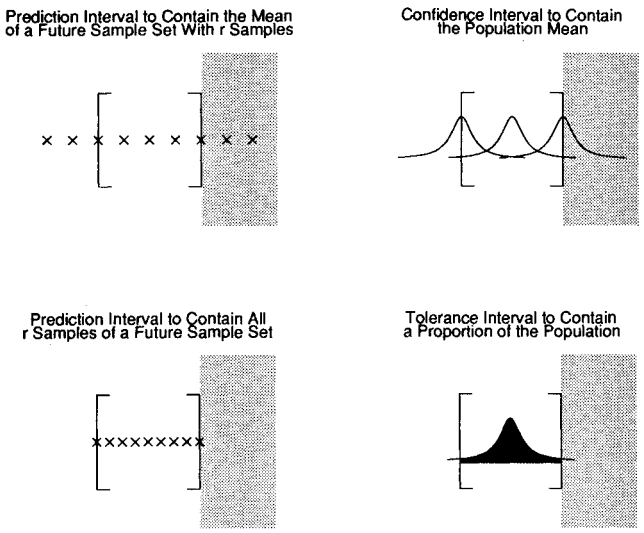

Fig. 1 Comparison of statistical intervals.

4) *Prediction interval to contain all r samples of a future sample set* (Y_r). The interpretation of this interval is that in a large number of repeated experiments where this interval is computed, the claim that the interval contains all of the r samples would be correct $100(1 - \alpha)\%$ of the time.

Of these four intervals, two of them (criterion 1 and 4) address predicting the results of a future set of samples from the same population, and two of them (criterion 2 and 3) address describing the population from which the samples have been collected. Figure 1 is a schematic of these four intervals. Also included on the figure is a one-sided performance goal (shaded area), which illustrates that samples from three of the four intervals can occur outside the interval and potentially violate the performance goal. Recall that the interval is only statistically correct $100(1 - \alpha)\%$ of the time.

In the development of each of these intervals, it is assumed that each sample (x_i; $i = 1, 2, \ldots, n$) is chosen randomly from a normally distributed population. Furthermore, for the prediction intervals, the future samples (y_i, $i = 1, 2, \ldots, r$) are assumed to be chosen randomly from the same normally distributed population and independent of the first sample set.

A prediction interval to contain the mean of a future sample set containing r samples is the appropriate interval to consider for system component flight testing. For example, component evaluation of a radar altimeter or a laser-ranging sensor would involve analysis of large data sets since these components operate at high data rates. With large data sets, it is reasonable to apply performance goals to the numerical average of the sample set; i.e., at the component level, a few outlying samples will not have a significant effect on total system performance. This interval was first considered by Baker,[3] who showed that the probability function of the deviation of $\bar{y}_r$ with respect to $\bar{x}$ was proportional to known quantities from the first sample set ($\bar{x}$ and s_x) and the Student-t distribution. Mathematically, this prediction interval can be expressed as

$$Pr\{\bar{x} - t(\nu, \gamma)[(1/r) + (1/n)]^{1/2}s_x \le \bar{y}_r \quad \text{and/or}$$
$$\bar{y}_r \le \bar{x} + t(\nu, \gamma)[(1/r) + (1/n)]^{1/2}s_x\} = 1 - \alpha \qquad (7)$$

where the notation $t(\nu, \gamma)$ indicates the 100γ percentile of the Student-t distribution with ν DOFs. Values for the Student-t distribution, $t(\nu, \gamma)$, may be found in many sources (e.g., see Fisher and Yates[4]).

The confidence interval to contain the population mean is just a special case ($r = \infty$) of the prediction interval to contain the mean of a future sample set. Hence, it has the same applicability to component flight test as did the previous interval and mathematically, Eq. (7) reduces to

$$Pr\{\bar{x} - t(\nu, \gamma)(1/\sqrt{n})s_x \le \mu \quad \text{and/or}$$
$$\mu \le \bar{x} + t(\nu, \gamma)(1/\sqrt{n})s_x\} = 1 - \alpha \qquad (8)$$

A tolerance interval to contain a proportion of the population is the appropriate interval to consider for total system flight test, rather than system component flight test. For example, air-to-air radar target detection or weapons system accuracy are typical flight test applications. These types of evaluations seek to quantify system performance, but can allow a few outlying samples without "safety-of-flight" implications. Hahn[2] concluded this interval could be expressed in terms of the noncentral t distribution. Its probability statement is

$$Pr\{\bar{x} - g(z_p\sqrt{n}, \nu, \gamma)(1/\sqrt{n})s_x \le Z_p \quad \text{and/or}$$
$$Z_p \le \bar{x} + g(z_p\sqrt{n}, \nu, \gamma)(1/\sqrt{n})s_x\} = 1 - \alpha \qquad (9)$$

where $z_p\sqrt{n}$ is the noncentrality parameter, and z_p is the percentile value for a normal distribution. Values for the noncentral t distribution, $g(z_p\sqrt{n}, \nu, \gamma)$, may be found in many sources (e.g., see Owen[5]). In Eq. (9), Z_p represents the expression

$$Pr\{\mu - K_p\sigma \le Z_p \quad \text{and/or} \quad Z_p \le \mu + K_p\sigma\} \ge p \qquad (10)$$

where K_p determines the magnitude of the tails of the normal distribution. Although not considered in this article, Owen[6,7] presents a method for controlling the percentages of each tail independently.

Hahn[1] calls a prediction interval to contain all r samples of a future sample set an "astronaut's interval." A typical astronaut, assigned to fly on a future spaceflight, is generally not concerned with what has happened on past flights or even what might happen on the average of future flights; he is more interested in the worst that might happen on his particular flight. This prediction interval is applicable to total system evaluations that necessitate "safety-of-flight" predictions. Additional flight test examples would be a flight control system demonstration or an evaluation of automatic terrain following system performance. Hahn[8,9] considered this interval in detail and showed that it can be expressed in terms of the multivariate Student-t probability density function with correlation matrix Σ. Mathematically, this yields the probability statement

$$Pr\{\bar{x} - u(r, \nu, \Sigma, \gamma)[1 + (1/n)]^{1/2}s_x \le Y_r \quad \text{and/or}$$
$$Y_r \le \bar{x} + u(r, \nu, \Sigma, \gamma)[1 + (1/n)]^{1/2}s_x\} = 1 - \alpha \qquad (11)$$

In Eq. (11), Y_r represents the expression (Y_1 and Y_2 and . . . and Y_r). Values for the multivariate Student-t distribution, $u(r, \nu, \Sigma, \gamma)$, may be found in many sources (e.g., see Krishnaiah and Armitage[10]).

Notice that all of the mathematical representations for these intervals, Eqs. (7–9) and (11), can be written in the form

$$Pr\{\bar{x} - ks_x \le J \quad \text{and/or} \quad J \le \bar{x} + ks_x\} = 1 - \alpha \qquad (12)$$

where J and k are summarized in Table 1. Also included in Table 1 are the references used in this article where these k-factors have already been tabulated.

To express these statistical intervals in terms of the performance goals, consider Fig. 2 where both upper and lower performance goals are shown. Combining Eqs. (6) and (12) yields a relationship between the statistical interval and the performance goals

$$L + ks_x \le \bar{x} \quad \text{and/or} \quad \bar{x} \le U - ks_x \qquad (13)$$

Table 1 Summary of *k*-factors

J	k	Tabulation
$\bar{y}_r$	$t(\nu, \gamma)[(1/r) + (1/n)]^{1/2}$	Hahn[1]
μ	$t(\nu, \gamma)(1/\sqrt{n})$	Hahn[1]
Z_p	$g(z_p\sqrt{n}, \nu, \gamma)(1/\sqrt{n})$	Natrella[11]
Y_r	$u(r, \nu, \Sigma, \gamma)[1 + (1/n)]^{1/2}s_x$	Hahn[8,9]

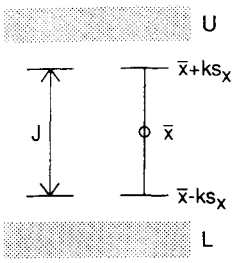

Fig. 2 Comparison of statistical intervals with respect to the performance goals.

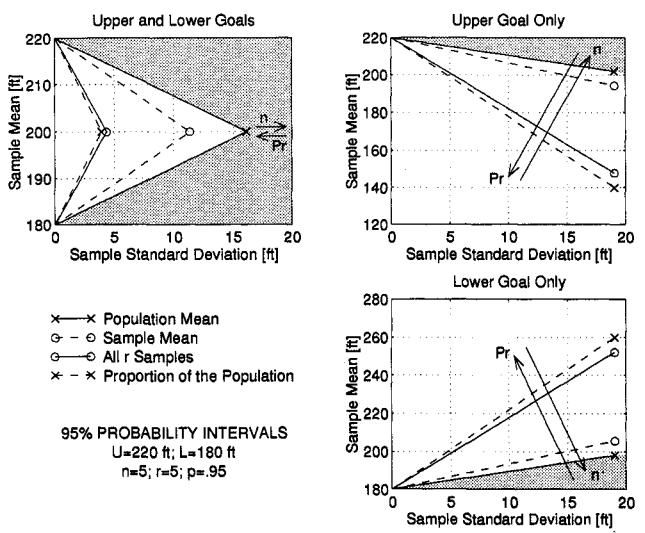

Fig. 3 Performance wedge construction.

Notice that the criterion J enters this statement through the appropriate value of k. Further, Eq. (13) carries the associated probability statement given by Eq. (12). Figure 3 demonstrates Eq. (13) for the four types of statistical intervals and the three combinations of performance goals; in this example, upper and lower performance goals of 220 and 180 ft, respectively, were assumed with a 95% probability level selected. The units in this example were feet, but any units can be associated with $\bar{x}$ and s_x as long as they are consistent. The sample set contained five samples and the future sample set being considered had five samples.

For flight test conditions where both upper and lower performance goals apply, the result is a wedge where if the intersection of the sample mean and standard deviation falls inside (nonshaded area), both goals will be met. For flight test conditions where only an upper (lower) performance goal apply, the sample mean and standard deviation must fall below (above) the line for the goal to be met. Also shown on Fig. 3 is trend information as the number of samples and the probability level increase. For conditions where upper and lower goals apply, additional samples have the effect of "stretching" the wedge by moving the tip point to the right; increased probability "shrinks" the wedge by moving the tip to the left. For single performance goals, additional samples increase the goal-satisfied region while increased probability decreases the goal-satisfied region.

Figure 3 also provides a comparison between the four statistical intervals. The least-conservative interval will always be the confidence interval to contain the population mean followed by the prediction interval to contain the mean of a future sample set with r samples. The prediction interval to contain all r samples of a future sample set and the tolerance interval to contain a proportion of the population will be more conservative than the other two intervals, although not necessarily in that order. The order of these last two intervals is dependent on the number of future samples r and the proportion of the population p selected. In Fig. 3, for the selected values of r and p, the prediction interval is more conservative than the tolerance interval, but this is not necessarily true in general. Nevertheless, the more conservative criteria does correlate to the more restrictive regions of the performance goals being satisfied. Hahn[1] gives additional comparison information between the different statistical intervals.

Consider further the case when both upper and lower performance goals apply. The wedge that results is a function of confidence level and DOFs, but will always have its vertices at

$$(0, U), \quad (0, L), \quad \{[(U - L)/2k], [(U + L)/2]\} \tag{14}$$

Additionally, it will always be symmetric about

$$\bar{x} = (U + L)/2 \tag{15}$$

If, as in many specifications, the performance goals can be written as a percentage of some nominal value

$$U = [1 + (u/100)]x_{\text{nom}} \tag{16}$$

$$L = [1 + (l/100)]x_{\text{nom}} \tag{17}$$

where both u and l can be above (>0) or below (<0) the nominal value, then, Eq. (13) can be written as

$$l + k\varepsilon_{s_x} \le \varepsilon_{\bar{x}} \quad \text{and/or} \quad \varepsilon_{\bar{x}} \le u - k\varepsilon_{s_x} \tag{18}$$

where $\varepsilon_{\bar{x}}$ is the percent error in the sample mean referenced to x_{nom}, and ε_{s_x} is the percent standard deviation normalized with respect to x_{nom}. Equation (18) is preferred because it is a relative equation; hence, the absolute numerical value of the performance goals and sample set statistics do not have to be considered. The end result of a wedge containing the region where the performance goals are satisfied does not change. Finally, the vertices of the wedge, Eq. (14), become

$$(0, u), \quad (0, l), \quad \{[(u - l)/2k], [(u + l)/2]\} \tag{19}$$

Flight Test Technique

The near real-time approach presented in this article is a simplified version of sequential life testing, which is defined as hypothesis testing in which the course of action is reassessed as observations become available. Sequential life testing has been predominantly applied in the reliability and manufacturing fields, but has application in many disciplines such as aircraft flight test. Many current texts (e.g., see Kapur and Lamberson[12]) cover the theory of sequential life testing. Sequential life testing is "simplified" in this scenario because of the small number of samples typically collected during performance flight tests.

During near real-time operations, the question at hand is "When can repeated test condition executions be terminated because, in a statistical sense, no additional information relative to satisfying a given performance goal is being obtained?" A graphical technique is proposed to address this question. To illustrate this technique, consider the case where both upper and lower performance goals apply to the specified flight test condition. For a given probability level and criterion, performance wedges can be constructed, as previously discussed, ranging from the minimum number of test condi-

tion executions planned to the maximum number of test condition executions allowed by the test program; typically, this range will be between 3 and 10 samples. As test conditions are executed, the sample set statistics can be plotted against these performance wedges. The sample set statistics will lie in one of three regions:

1) The first region is inside the performance wedge corresponding to the actual number of samples collected. If inside this region, testing should be terminated because to the probability level and criterion specified, the sample set has met the performance goals. Additional samples would be beneficial, but not necessary in addressing the question of did the system performance meet specification.

2) The second region is outside of the performance wedge corresponding to the actual number of samples collected, but inside the performance wedge associated with the maximum number of test condition executions allowed by the test program. In this situation, test condition execution should continue because the performance goals have not been met to the specified probability level for the given criterion. Additional samples are required to be able to address system performance relative to a given specification.

3) The last region is outside of the performance wedge corresponding to the maximum number of test condition executions allowed by the test program. In this case, test condition execution should be terminated because to the specified probability level, the performance goals will not be met, even if additional samples are collected. Hence, one should then answer the second question posed in this article and quantify to what probability level the performance goals were satisfied, albeit lower than the desired probability level.

One inherent assumption is made with this technique, namely, that the sample set statistics remain constant between test condition executions. This assumption is not necessarily good for extremely small sample sizes. For the two cases presented in the next section ($n = 4$), this assumption will be quantified.

To address the second question posed in this article and alluded to above, consider constructing performance wedges, but rather than specifying probability level and criterion, select number of samples and criterion to be held constant. In this way, probability level can be varied on the graph and overlaying the sample set statistics will yield the probability level to which the performance goals have been satisfied.

Application

A simple example will demonstrate the usefulness of this technique. Consider a flight test condition with both upper and lower performance goals specified as ±10% error. Figure 4 shows performance wedges for each of the criterion con-

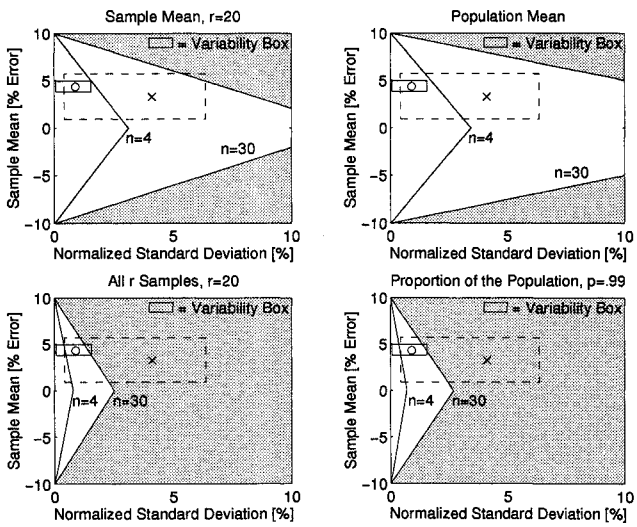

Fig. 4 Application performance wedges ($Pr = 0.99$).

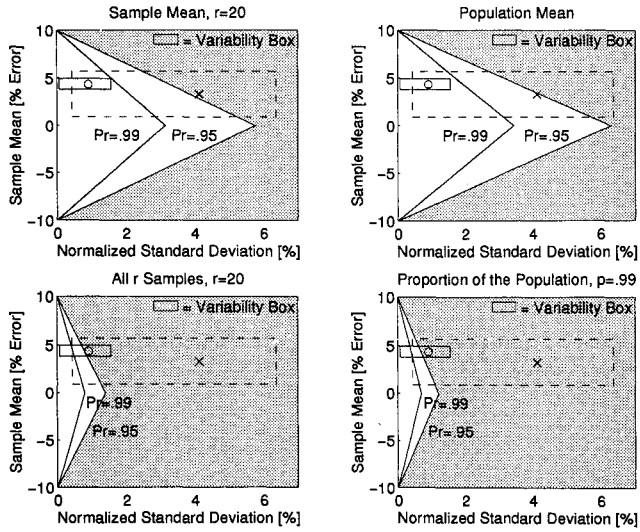

Fig. 5 Application performance wedges ($n = 4$).

sidered in this article with a probability level of 0.99 assumed. Two wedges are shown on each figure corresponding to $n = 4$ and $n = 30$; a sample set size of 30 was considered equivalent to a normal distribution. Typically, 30 samples is a much larger sample set than one can expect to collect in performance flight test. The statistics from two different test conditions of actual flight test data are indicated on Fig. 4 by the symbols ○ and ×. Each symbol represents the sample set statistics after four executions of the respective test condition. In addition, to address the assumption that the sample set statistics remain constant between test condition executions, the variability of the sample set statistics is indicated by the boxed area surrounding the two symbols (solid for ○; dashed for ×). That is, the statistics of the sample set would remain inside of the respective variability box, regardless of the order in which the samples were executed.

For the prediction interval to contain the mean of a future sample set with r samples ($r = 20$) and the confidence interval to contain the population mean, applying the proposed technique dictates terminating test condition execution for the sample set indicated by a ○, because the performance goals have been satisfied; continuing data collection for the sample set indicated by a × is also indicated. The variability of the sample set statistics, although somewhat large, does not affect these conclusions. Applying the technique to the other two intervals shows continued testing for the ○ sample set and test termination for the × sample set. Test termination resulted because with even up to 26 additional samples, the performance goals would not be met. Again, the variability of the sample set statistics does not affect the decision to continue or terminate testing.

In this last case, the second question posed by this article comes into consideration. Figure 5 shows performance wedges for a constant sample set size ($n = 4$) and varying probability level. Also shown are the sample set statistics and their respective variability boxes from the two test conditions. For the prediction interval to contain the mean of a future sample set with r samples ($r = 20$) and the confidence interval to contain the population mean, the × sample set essentially satisfies a probability level of 0.95. For the other two intervals, a lower probability level would have to be determined for the given performance goals. Nevertheless, in this manner and for this number of samples, a probability level statement can be associated with each sample set statistics.

Conclusions

A near real-time approach to statistical flight test has been proposed that minimizes the number of test condition exe-

cutions required to adequately describe system performance relative to a specified performance goal or specification. This approach took advantage of statistical intervals based on the Student-t probability distribution and its generalizations. In this manner, probability statements could be associated with the sample set statistics when considering performance goals or specification compliance. Four different statistical criterion were considered, each having an unique applicability to aircraft flight test.

Acknowledgments

The author would like to thank D. E. Reynolds and B. W. Woodruff of the Department of Mathematics and Statistics at the Air Force Institute of Technology for their many helpful comments and suggestions concerning this article.

References

[1]Hahn, G. J., "Statistical Intervals for a Normal Population, Part I. Tables, Examples and Applications," *Journal of Quality Technology*, Vol. 2, No. 3, 1970, pp. 115–125.

[2]Hahn, G. J., "Statistical Intervals for a Normal Population, Part II. Formulas, Assumptions, Some Derivations," *Journal of Quality Technology*, Vol. 2, No. 4, 1970, pp. 195–206.

[3]Baker, G. A., "The Probability that the Mean of a Second Sample Will Differ from the Mean of a First Sample by Less than a Certain Multiple of the Standard Deviation of the First Sample," *Annals of Mathematical Statistics*, Vol. 6, 1935, pp. 197–201.

[4]Fisher, R. A., and Yates, F., *Statistical Tables for Biological, Agricultural, and Medical Research*, 6th ed., Hafner Publishing, New York, 1963, p. 46.

[5]Owen, D. B., "A Survey of Properties and Applications of the Noncentral t-Distribution," *Technometrics*, Vol. 10, No. 3, 1968, pp. 445–478.

[6]Owen, D. B., "Control of Percentages in Both Tails of the Normal Distribution," *Technometrics*, Vol. 6, No. 4, 1964, pp. 377–387; also *Technometrics*, Vol. 8, No. 3, 1966, p. 570 (Errata).

[7]Owen, D. B., "Variables Sampling Plans Based on the Normal Distribution," *Technometrics*, Vol. 9, No. 3, 1967, pp. 417–423.

[8]Hahn, G. J., "Factors for Calculating Two-Sided Prediction Intervals for Samples from a Normal Distribution," *Journal of the American Statistical Association*, Vol. 64, No. 327, 1969, pp. 878–888.

[9]Hahn, G. J., "Additional Factors for Calculating Prediction Intervals for Samples from a Normal Distribution," *Journal of the American Statistical Association*, Vol. 65, No. 332, 1970, pp. 1668–1676.

[10]Krishnaiah, P. R., and Armitage, J. V., "Percentage Points of the Multivariate t-Distribution," Applied Mathematics Research Lab., ARL Rept. 65-199, Wright–Patterson AFB, OH, Oct. 1965.

[11]Natrella, M. G., *Experimental Statistics*, National Bureau of Standards Handbook 91, U.S. Government Printing Office, Washington, DC, 1963, pp. T10–T15.

[12]Kapur, K. C., and Lamberson, L. R., *Reliability in Engineering Design*, Wiley, New York, 1977, pp. 342–365.