AIAA 2002-0885

Tactical Defenses Against Systematic Variation in Wind Tunnel Testing

R. DeLoach
NASA Langley Research Center
Hampton, VA

40th AIAA Aerospace Sciences Meeting & Exhibit
14-17 January 2002
Reno, NV

For permission to copy or republish, contact the American Institute of Aeronautics and Astronautics
1801 Alexander Bell Drive, Suite 500, Reston, VA 20191

TACTICAL DEFENSES AGAINST SYSTEMATIC VARIATION IN WIND TUNNEL TESTING

Richard DeLoach*
NASA Langley Research Center, Hampton, VA 23681-2199

Abstract

This paper examines the role of unexplained systematic variation on the reproducibility of wind tunnel test results. Sample means and variances estimated in the presence of systematic variations are shown to be susceptible to bias errors that are generally non-reproducible functions of those variations. Unless certain precautions are taken to defend against the effects of systematic variation, it is shown that experimental results can be difficult to duplicate and of dubious value for predicting system response with the highest precision or accuracy that could otherwise be achieved.

Results are reported from an experiment designed to estimate how frequently systematic variations are in play in a representative wind tunnel experiment. These results suggest that significant systematic variation occurs frequently enough to cast doubts on the common assumption that sample observations can be reliably assumed to be independent. The consequences of ignoring correlation among observations induced by systematic variation are considered in some detail.

Experimental tactics are described that defend against systematic variation. The effectiveness of these tactics is illustrated through computational experiments and real wind tunnel experimental results. Some tutorial information describes how to analyze experimental results that have been obtained using such quality assurance tactics.

Nomenclature

b_i	bias error in the i^{th} observation of a sample
df	degrees of freedom
e_i	random error in the i^{th} observation of a sample
H_0	the null hypothesis
H_A	the alternative hypothesis
n	the sample size
N	the size of the population, assumed "large"
s	the sample standard deviation
s^2	the sample variance
x_i	the i^{th} independent variable, coded units
y_i	the i^{th} observation in a sample
$\bar{y}$	the sample mean
α, alpha, AoA	angle of attack, Type I inference error probability
β, beta	angle of sideslip, Type II inference error probability, net bias error due to systematic variation (equivalent to a rectification error)
μ	the population mean
ρ_m	the lag-m autocorrelation coefficient
σ	the population standard deviation
σ^2	the population variance
ξ_i	the i^{th} independent variable, physical units
Alternative Hypothesis	usually an assertion that changes in one or more variables will produce a significant change in system response.
CCD	Central Composite (Box-Wilson) Design
dwell time	time required to acquire a sample of data
MDOE	Modern Design of Experiments
Null Hypothesis	usually an assertion that changes in one or more variables will not produce a significant change in system response.
OFAT	One Factor At a Time
population	a large set of measurements from which a sample can be imagined to come
residual	difference between an observed value and some reference, such as a sample mean or a model prediction.
sample	a finite subset of available measurements from a population
sample mean	the average value of all the measurements in a sample

* Senior Research Scientist
Copyright © 2002 by the American Institute of Aeronautics and Astronautics, Inc. No copyright is asserted by the United States under Title 17, U. S. Code. The U. S. Government has a royalty-free license to exercise all rights under the copyright claimed herein for Government Purposes. All other rights are reserved by the copyright holder.

sample size	the number of measurements in a sample.
significance	the probability of an inference error due to chance variations in the data
significant	large enough to be detected with a degree of confidence satisfying specified risk tolerance levels.
Type I error	inference error committed when the null hypothesis is erroneously rejected.
Type II error	inference error committed when the alternative hypothesis is erroneously rejected.

Introduction

A key quality control strategy in conventional wind tunnel research is to hold constant all independent variables except for one that is selected for current study. Such variables as Mach number and Reynolds number might be held constant while the angle of attack is systematically varied to quantify the effect of such changes on forces and moments. Similar angle of attack sweeps are executed at other combinations of Mach number and Reynolds number, changed systematically between each sweep. This practice of holding all other variables constant while changing only one factor at a time is widely assumed to be a necessary condition for correctly associating response changes (forces, moments, pressures, etc.) with the independent variable changes that cause them. The term "one factor at a time" (OFAT) is used to describe this popular experimental method.

Experienced OFAT practitioners recognize that chance variations in the data inevitably occur in any wind tunnel test, but the usual assumption is that these fluctuations occur about a mean response that does not change significantly unless the independent variables of the test are changed. Absent this assumption, true cause and effect relationships between independent and response variables are difficult to reconcile using conventional OFAT methods. Furthermore, it is not uncommon in ground testing to assign certain convenient attributes to the unexplained variance that are often not entirely justifiable. For example, if all unexplained variance consists of random, independent variations about sample means that do not change over time, the system can be said to be in a state of "statistical control". If the system is in such a state, arbitrarily small variances in the distribution of sample means can be achieved for sufficiently large samples sizes, for example. Also, certain properties can be assigned to the distribution of experimental errors than ensure that statistics such as means and standard deviations that are based on finite data samples are reliably unbiased estimators of the corresponding parameters of the general population of interest. These are necessary conditions for "getting the right answer" (within a constant if there are bias errors) when resource constraints dictate – as they generally do – that we can only observe a subset of the entire population of (theoretically) possible observations.

A clear understanding of the necessity for independence in measurements may elude those experimentalists who are not particularly well grounded in statistical theory, and who may therefore tend to overlook such details. OFAT practitioners who understand why the prospects for reliable inference without independent experimental errors are so bleak, have a strong tendency to rely on *assumptions* (not to say *hopes*, or *prayers*) that the errors in a sequence of experimental data points are in fact statistically independent, whether they are or not. The following quote from Box, Hunter, and Hunter's seminal text on experiment design[1] provides an excellent description of this tendency: *"Statisticians frequently make the assumption of independence at the beginning of their writings and rest heavily on it thereafter, making no attempt to justify the assumption, even though it might have been thought that 'a decent respect to the opinions of mankind requires that they should declare the causes which impel them' to do so. The mere declaration of independence, of course, does not guarantee its existence."*

This paper reports the results of a recent experiment to quantify the frequency with which common assumptions of statistical independence are valid in a representative wind tunnel test. That is, we sought to determine how often the ubiquitous "declaration of independence" was violated in a typical ground-testing scenario. We begin by examining some general consequences of an unwarranted assumption of statistical independence, and consider the specific impact of such an assumption on the proper estimation of one standard ground-testing data structure, the common pitch polar.

In subsequent sections we consider plausible reasons that it may not be prudent to assume statistical independence in real world ground testing. Our results support the conclusions of standard references on experiment design[1-4], which recommend against relying upon statistical control assumptions. Instead, we follow the lead of the statistical experiment design community in counseling against OFAT techniques generally, and encourage the use of data quality assurance tactics during the execution of the experiment that can lead to reliable inferences whether the system is in a reliable state of statistical control or not. Stated slightly differently, our results support the view that while statistical control is a sufficient condition to ensure

against common causes of improper experimental inference, it is not a necessary condition.

The Role of Statistical Independence

Statistical independence is vaguely perceived in the general ground testing community (when it is perceived at all) as a "good" thing, but few in our community have a visceral understanding of what it actually means to experimental research. Its importance is rarely elevated to a level for which the researcher feels compelled to exert himself excessively to achieve it. As noted in the introduction, the general tendency is to simply assume that it exists, or more commonly, to assume that it doesn't much matter one way or the other outside the ivy towers of academe.

In this section we will attempt to describe the concept of statistical independence in terms that are meaningful to the practicing experimentalist. We will also try to make a convincing case that statistical independence is at once crucial to the success of an experiment, and also a property that cannot be reliably assumed in real data without some effort on the part of the researcher to secure it.

The Random Sampling Hypothesis

The reality of unexplained variance in experimental data forces us to describe the systems we study in terms of random variables that can assume different values, governed by a probability distribution that defines how likely it is that a given variable will assume a particular value. We can imagine a set of N measurements of such a random variable. Theoretically we can suppose that N is infinite, but for our purposes it is sufficient to assume that N is large enough to have captured the random variable at multiple instances of every level for which there is a significant probability that it will be observed. We use the term "population" to describe this collection of essentially every possible value of the random variable, with all values represented in proportion to the probability that they will actually be observed. Clearly, we could say a great deal about a random variable, and say it authoritatively, if we had such a comprehensive data set to examine. Specifically, we could infer such parameters of the population as its mean and variance with very little risk of an inference error.

Unfortunately, resource constraints generally compel us to deal with only a subset, or *sample*, of the N points that comprise the entire population. The essence of an experiment is to make reliable inferences about the parameters of the general population based on statistics from a very much smaller sample. We therefore compute such statistics as the mean and variance of a *sample*, and from those we attempt to estimate the corresponding parameters of the general *population*.

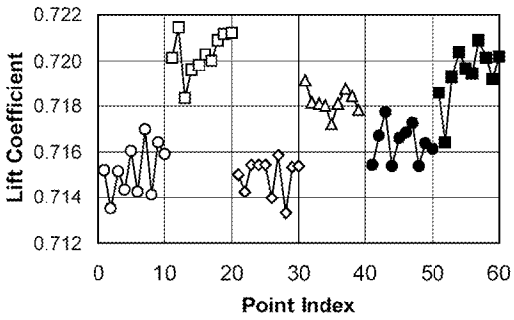

Figure 1. Points within shorter intervals can be more alike than points acquired over longer intervals. Figure from reference 5.

Any data set we acquire will be *some* sample of the population. However, to adequately represent the population from which the sample is drawn, it is crucial that the sample consist of measurements drawn *at random* from the entire population. That is, we seek a sample constructed in such a way that every condition under which a population member can be generated has an equal chance of occurring in the sample. Researchers generally assume that their data samples are representative of the general population about which they seek to make inferences. We call this ubiquitously assumed condition the *random sampling hypothesis*, and consider now some reasons to doubt that it applies in every real experimental situation.

We will show presently that the risk of making an improper inference about the population from a finite sample of data can be substantially greater if the random sampling hypothesis does not apply. Unfortunately, it often does *not* apply when real data are acquired with conventional OFAT ground-test measurement techniques that rely upon sequential independent-variable level settings to maximize data acquisition rate. The reason is that when data are acquired in a sequence, those points taken within a relatively short time interval tend to be more alike than those taken with longer intervening time intervals. Figure 1, taken from reference 5, was used in that reference to illustrate this general tendency in ground testing data. The authors distinguish between what they describe as "within-group" and "between-group" variance levels, noting the consistent tendency for between-group variance associated with longer time intervals to dominate the within-group variance associated with shorter time periods.

This can often be plausibly attributed to disturbances that tend to persist over time. For

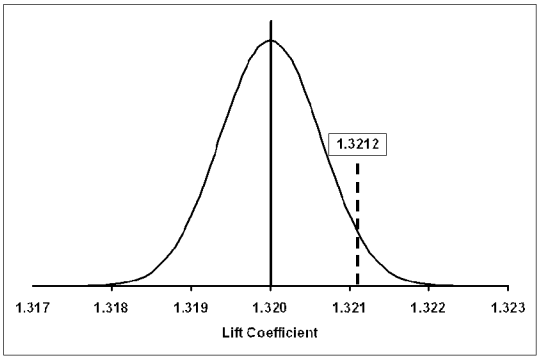

Figure 2. Distribution of sample means for 10 observations of lift coefficient, μ=1.320, σ=0.002.

example, the calibration coefficients of the force balance in a wind tunnel test can change slightly with temperature. Frictional heating as the test proceeds may therefore cause drift in the balance output or in other instruments. Subtle deviations in flow angularity can cause systematic variations in the angle of attack to be superimposed upon those generated by intentionally changing the pitch angle of the model. Thermal expansion can alter the tunnel geometry somewhat, and induce systematic variations in the wall interference effects and other changes. These are just a few of the unknown (and unknowable) sources of unexplained variation in a wind tunnel test that can persist over some extended period of time.

If such persistent disturbances are in play during the period in which a small sample of data is acquired, then deviations of the individual measurements from the sample mean or some other reference (such as the true but unknown response function relating the dependent and independent variables) will not be random. For example, if such systematic variations cause the i^{th} data point to be biased somewhat high, then the $(i+1)^{st}$ point will tend to be biased high also if it is acquired only a short time after the i^{th} point. This means that the error in the $(i+1)^{st}$ point is *not* just as likely to be negative as positive. Rather, the probability distribution of $(i+1)^{st}$ observation is affected by the level of the i^{th} observation. The errors are not independent; they exhibit some degree of autocorrelation. That is, the observations are statistically *dependent*, because the error distribution for a given observation depends on the order in which the data points were acquired. The relationship between the errors in data points "A" and "B" is not a matter of simple random chance. It depends largely on which one occurs first in the data acquisition sequence.

We will exploit this linkage between run order and the random sampling hypothesis later. For now, suffice it to say that some important statistical procedures depend upon the random sampling hypothesis for their validity, and the presence of systematic variations in the data can seriously increase the risk of inference error when conclusions are based on an erroneous assumption of independence. It is also important to note that unless our sample sizes are relatively large (no fewer than about 50 points[1]), autocorrelation computations and other direct measures of statistical dependence can be unreliable. Autocorrelation in samples no larger than a common pitch polar can go virtually undetected, and yet be sufficient to invalidate conclusions based on an assumption of independence, as we shall soon see.

The Reference Distribution

In this section we outline an objective procedure for making reliable inferences in scientific experiments *under the random sampling hypothesis*. We will later consider how this process is impacted when the random sampling hypothesis does not apply, and we will describe some practical defense tactics to ensure that the random sampling hypothesis applies even when systematic variations cannot be neglected.

Formal procedures have evolved for making scientific inferences that date from their introduction early in the 20th century by Ronald Fisher and his associates[6]. While the details obviously depend on the specific circumstances of the experiment, the general procedure is the same whether the study is in experimental aeronautics or in any other scientific or research engineering field. It begins by stipulating some state of nature and developing the distribution of sample means that would be expected if nature were in fact in that state. For example, if we wish to infer whether or not a proposed wing enhancement improves the coefficient of lift at cruise for a particular aircraft design, we might consider a sample of differential lift measurements involving the new and old wing, and ask how the mean of such a sample would be distributed if there was no change in lift. That is, we imagine many samples of differential lift data from these same two wings, with sample means that differ somewhat due to ordinary chance variations in the data. We ask how the probability distribution for the difference in those sample means would look if in fact the new wing produced no additional lift. We refer to the presumption of such a no-change state as a "null hypothesis", and use the symbol H_0 to represent it succinctly. There is always a corresponding "alternative hypothesis", H_A, which in this case would be true if the proposed wing change actually did generate additional lift. The random variable of interest in this example is the change in lift associated with the

new wing. The distribution of such a random variable that we would expect *under the null hypothesis* is known as the *reference distribution*.

Figure 2 represents a reference distribution for a special case to illustrate the concept. In this example, we assume that the lift coefficient of the old wing under cruise conditions is known to be 1.320, and the standard deviation of individual lift coefficient measurements for this model in this facility is also known, with a value of 0.002. (When we assert that the mean and standard deviation is "known" for the old wing, we simply mean that our estimates are based on a volume of replicated measurements that is so large as to be effectively "infinite", so that the uncertainty in those estimates is close enough to zero to make no practical difference.) To defend against systematic error that may be present, it is necessary to compare measurements from the new wing with measurements from the old wing, even if we know the old wing's lift. For the present, however, we assume that all unexplained variance is random, and because the lift of the old wing is known, we will only concern ourselves with measurements of the new wing's lift.

Under the null hypothesis, we would expect the distribution of sample means for the new wing to be the same as for the old wing; namely, 1.320. We know the standard deviation in the distribution of individual observations is 0.002, but what about the standard deviation in the distribution of sample means? The formula for the variance of a distribution of sample means is well known for the case in which all observations in the sample are independent, but it can be derived easily. Because this derivation is instructive for more complex cases to follow (where statistical independence cannot be assumed), we will outline it briefly here. We start with a general formula for error propagation developed in standard references[7,8] and reproduced here.

$$\sigma_y^2 = \sum_{i=1}^{k}\left[\left(\frac{\partial y}{\partial x_i}\right)^2 \sigma_{x_i}^2\right] + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k}\left(\frac{\partial y}{\partial x_i}\right)\left(\frac{\partial y}{\partial x_j}\right)\rho_{x_i x_j}\sigma_{x_i}\sigma_{x_j} \tag{1}$$

Equation 1 is useful in cases where we know the variance of the independent variables upon which some variable of interest depends, but we do not know the variance of the variable that interests us. For example, we may know the variance in measurements of the length and width of some rectangular area, but what we really want to know is the variance in an estimate of the area that is based on those uncertain length and width measurements.

Equation 1 describes the variance in a general function of k variables: $y = f(x_1, x_2, \ldots, x_k)$. It depends on σ_{x_i}, the standard deviation of the independent variables x_i, and on $\rho_{x_i x_j}$, the coefficient of correlation between the i^{th} and j^{th} independent variable.

If there is no correlation among any of the variables, the double summation term on the right of equation 1 vanishes because in that case $\rho_{x_i x_j} = 0$. We are left with

$$\sigma_y^2 = \left(\frac{\partial y}{\partial x_1}\right)^2 \sigma_{x_1}^2 + \left(\frac{\partial y}{\partial x_2}\right)^2 \sigma_{x_2}^2 + \cdots + \left(\frac{\partial y}{\partial x_k}\right)^2 \sigma_{x_k}^2 \tag{2}$$

Consider now an n-point sample mean, which can be represented as a function of n variables as follows:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \left(\frac{1}{n}\right)y_1 + \left(\frac{1}{n}\right)y_2 + \cdots + \left(\frac{1}{n}\right)y_n \tag{3}$$

From equation 3 we see that the derivative of $\bar{y}$ with respect to y_i is $1/n$ for all i. If we make the reasonable assumption that all observations in the sample have the same variance, then we can drop the distinguishing subscripts in equation 2 and call this per-point variance simply σ^2. The result of then applying (2) to (3) is that the formula for the variance in sample means is just the sum of n identical terms, each then equal to σ^2/n^2, and we have:

$$\boxed{\sigma_{\bar{y}}^2 = n\left(\frac{\sigma^2}{n^2}\right) = \frac{\sigma^2}{n} \rightarrow \sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}} \tag{4}$$

This well-known result states that the variance in the distribution of sample means decreases with the size of the sample. Among other things, it enables us to define sample sizes that drive the uncertainty in sample means to whatever level is consistent with our tolerance for inference error risk[9] (although "zero" uncertainty cannot be achieved with any finite volume of data). However, the validity of equation 4 depends on statistical independence among observations in the sample, so that $\rho_{x_i x_j} = 0$. We will later examine how the distribution of sample means changes when the observations are *not* independent, and we will also

consider the impact this has on the inference error risk we assume when we make an inference in the presence of correlation while assuming statistical independence.

We return now to our wing lift example, in which we asked how a distribution of some large number of sample means would look if each was the mean of *n* lift coefficient measurements on the new wing, acquired under the null hypothesis (no change in lift from the old wing) and under conditions for which all measurements in the sample are statistically independent. From the null hypothesis we expect the mean of this distribution to be the known lift coefficient of the old wing, which we have said is 1.320. Let us assume that our sample size is 10. We have also said that the standard deviation in individual observations of the lift coefficient is known in this example to be 0.002. Then from equation 4 the standard deviation in the distribution of sample means is expected to be $0.002/\sqrt{10} = 6.32\times10^{-4}$.

The Central Limit Theorem assures us that the distribution of a random variable that represents the sum of other random variables is approximately normal (Gaussian) if the summed variables are of comparable magnitude and satisfy other mild constraints, independent of the probability distribution of the populations from which they were drawn. We therefore adopt as our reference a normal distribution with a mean of 1.320 and a standard deviation of 6.32×10^{-4}. This is the distribution illustrated in figure 2.

We use such a reference distribution to make an inference in the following way. We know that *under the null hypothesis*, the population mean in this case would be 1.320. However, experimental error virtually guarantees that a 10-point sample mean will not be exactly 1.320 except by pure coincidence, *even if the null hypothesis is true*. Nonetheless, a casual inspection of figure 2 suggests that if H_0 is true, while the sample mean may not be exactly 1.320, there is an overwhelming probability that it will lie somewhere between 1.318 and 1.322. If our specific 10-point sample lies outside this range, we have good reason to reject the null hypothesis.

The dashed line in figure 2 marks a criterion by which we can objectively decide whether or not to reject the null hypothesis. There is associated with this criterion a controlled probability (controlled by the selection of the criterion level) of making an inference error that is defined by the area under the reference distribution to the right of this line. In this illustration a criterion of 1.3212 makes this area 0.05. We will acquire a 10-point sample and compute its mean, accepting or rejecting the null hypothesis depending on whether it is less than or greater than this criterion.

Because the sample mean is a random variable, there is some probability that the mean of any one 10-point sample will lie to the right of the criterion even if the null hypothesis is true. In such a case we would erroneously reject the null hypothesis, concluding that the new wing was better than the old wing even though it was not. But because the criterion was selected to ensure that the area under the reference probability distribution to the right of the criterion is only 0.05, we know that *assuming a valid reference distribution* there is only a 5% chance that we will erroneously reject the null hypothesis due to ordinary random experimental error. We can move the criterion to the right as needed to drive the inference error probability lower than this if we require greater than 95% confidence in an inference that the new wing is better than the old. (We might require greater confidence if there were large tooling costs associated with a decision to take the new wing design to production, for example.)

A complete description of this problem is more complicated than the one we have illustrated because in selecting a criterion (and also in defining the optimum sample size as it happens), we must not only account for the possibility of erroneously rejecting the *null* hypothesis, but also the possibility of erroneously rejecting the *alternative* hypothesis should our specific sample mean lie to the left of the criterion. This extension is beyond the scope of the present paper but it is described in standard references[3,4] and has also been applied to the general problem of scaling data volume requirements in ground testing[9]. For the purposes of this paper, it is simply necessary to note that the extended problem depends all the more on a reliable reference distribution. This adds to the pressure to ensure that the random sampling hypothesis is satisfied, because our estimate of the standard deviation in the reference distribution depends upon it. This, in turn, directly impacts the risk of erroneously rejecting H_0 or H_A.

Impact of Correlation

The key to an objective inference is a reference distribution that reliably describes some hypothesized state of nature. Unfortunately, when the observations in a sample of data are correlated, even mildly, the corresponding reference distribution can be different enough from the one we construct under the assumption of statistical independence to substantially inflate the inference error probability. We will demonstrate this for the case of our wing lift example in a moment, but first let us examine how correlation can affect the standard deviation in a distribution of sample means.

Distribution of Sample Means with Correlated Observations

Consider a sample of data in which the standard deviation of the distribution of individual observations is the same for each point, σ, just as in the previous wing comparison example. However, unlike the previous example in which we assumed no correlation, assume in this case that each point is correlated with the one acquired immediately before. We will further assume that correlation with all other points is zero. (The distance between points in a sequence is called the "lag", and these conditions describe a state in which only the "lag-1" or first-order autocorrelation is non-zero.)

Consider now the variance of a function $n\bar{y} = y_1 + y_2 + \cdots + y_n$, where $\bar{y}$ is the sample mean and the y_i are the n individual observations of the sample. We can use equation 1 to compute this variance, where for a lag-1 autocorrelation we have:

$$\rho_{12} = \rho_{23} = \ldots \rho_{n-1,n} = \rho_1,$$

where the subscript "1" indicates that the autocorrelation is lag-1. With all the partial derivatives equal to 1 for this case, equation 1 reduces to:

$$\begin{aligned} \sigma^2_{n\bar{y}} &= \sum_{i=1}^{n} \sigma^2 + 2\sum_{i=1}^{n-1} \rho_1 \sigma^2 \\ &= \mathrm{n}\sigma^2 + 2(\mathrm{n}-1)\rho_1\sigma^2 \\ &= \sigma^2[\mathrm{n} + 2(\mathrm{n}-1)\rho_1] \end{aligned} \qquad (5)$$

The second summation goes to *n-1* because for *n* observations in a sample there are *n-1* adjacent (and in this case, correlated) pairs.

Equation 5 gives us the variance for $n\bar{y}$, but we are interested in the variance for $\bar{y}$. We can again apply the general propagation formula of equation 1 by representing $\bar{y}$ as a function of a single variable, "$n\bar{y}$", for which we know the variance from equation 5:

$$\begin{aligned} \bar{y} = \left(\frac{1}{n}\right)(n\bar{y}) \rightarrow \sigma^2_{\bar{y}} &= \left(\frac{1}{n}\right)^2 \sigma^2_{n\bar{y}} \\ &= \left(\frac{1}{n}\right)^2 \left\{\sigma^2[n + 2(n-1)\rho_1]\right\} \end{aligned} \qquad (6)$$

After rearranging terms, this becomes:

$$\boxed{\sigma^2_{\bar{y}} = \frac{\sigma^2}{n}\left[1 + \frac{2(n-1)}{n}\rho_1\right]} \qquad (7)$$

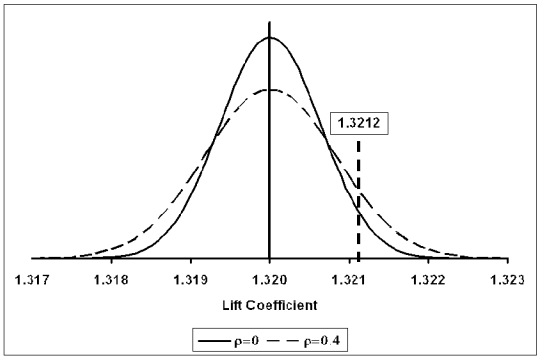

Figure 3. Distribution of 10-point sample means, μ=1.320, σ=0.002, with and without lag-1 autocorrelation of 0.4.

The term outside the square brackets on the right side of equation 7 is the familiar variance for the distribution of sample means when all observations are statistically independent, as derived in equation 4. The term inside the square brackets is a measure of how the variance is changed when the observations are not independent. It depends on the autocorrelation coefficient as expected, and it also depends on the volume of data in the sample. This should not be unexpected either, since the more points there are in the sample, the more correlated pairs there will be, and thus the greater will be the deviation from the no-correlation case.

The lag-1 autocorrelation coefficient is bounded by ±0.5, so the bracketed term in equation 7 can range from $(2n-1)/n$ on the high side to $1/n$ on the low. This means that even a relatively mild lag-1 case of autocorrelation can cause the variance to change by a factor of $(2n-1)/n \div (1/n) = 2n-1$, depending on details of the correlation. This range is substantial, even for small samples. It is 5-to-1 for as few as three points in the sample, and is about two orders of magnitude for samples of around 50 points. The larger the correlated sample, the greater the ambiguity that is introduced by correlation. The situation is exacerbated even further, of course, for common situations in which the correlation is more severe than the first-order (lag-1) case we have considered here.

Impact of Correlation on Inference Error Risk

Let us now return to our wing lift example. Equation 7 describes the variance of the actual distribution of sample means that we need to use as a reference distribution when there is first-order autocorrelation. Let us assume a lag-1 autocorrelation coefficient of 0.4. The sample size is assumed to be the

same as in the uncorrelated case: n=10, and likewise the standard deviation of individual measurements is still 0.002. Inserting these values into equation 7 yields a value for the standard deviation in the distribution of sample means for the correlated case of $8.29x10^{-4}$, which is over 30% greater than the uncorrelated case. This means that the distribution that actually corresponds to our null hypothesis of no increase in lift coefficient will be wider than the distribution we would assume if we thought all the observations were statistically independent (and thus used equation 4 to compute the variance of the distribution instead of equation 7).

The greater variance in the true distribution means that the area under the distribution to the right of the criterion we would set under the random sampling hypothesis is now larger. Figure 3 illustrates this. Recall that this area corresponds to the probability of erroneously rejecting the null hypothesis due to experimental error. In the uncorrelated case it was selected by design to be 0.05, but if correlation inflates the variance of the distribution of sample means as figure 3 illustrates, this probability increases to 0.098. In other words, the introduction of a rather mild degree of correlation has essentially doubled the probability of an inference error. It would now be twice as likely as before to take an ineffective wing design to production, for example, incurring the tooling costs and other expenses associated with such an undertaking with no prospects of producing a wing whose performance could justify these costs.

Impact of Correlation on the Quality of Sample Statistics as Population Parameter Estimators

Valid mean and variance numbers can be computed for any sample of data because they are simply determined from mechanical mathematical operations, but there is very little else we can say about those results if they apply only to an isolated sample of data. The underlying assumption in experimental research is that the sample statistics tell us something useful about the broader population from which the sample was drawn. We count on the true expectation value of the sample mean, $\bar{y}$, to be μ, the population mean, and likewise we assume that the expectation value of the sample variance, s^2, will be the population variance, σ^2. Instead, in an appendix to this paper we see that when the random sampling hypothesis does not hold, the expectation values of the sample mean and variance can be quite different. Equations for the sample mean and variance derived in the appendix are reproduced here for convenience:

$$\boxed{E\{\bar{y}\} = \mu + \beta} \tag{8}$$

$$\boxed{E\{s^2\} = \left\{\frac{[n-1-f(\rho)]}{n-1}\right\}\sigma^2 + \left(\frac{n}{n-1}\right)\left(\sigma_\beta^2 + 2\rho_{e,\beta}\sigma\sigma_\beta\right)} \tag{9}$$

The quantities μ and σ are the true population mean and standard deviation, respectively. β and σ_β^2 are, respectively, the mean (generally non-zero) deviation and the mean square deviation of the systematic errors relative to the true population mean. The function $f(\rho)$ is a generic representation of the change in population variance attributable to correlation, ρ, among the observations in the sample, as we have already considered for a special case of autocorrelation. This is the general representation for which the specific instance was derived in equation 7. Finally, $\rho_{e,\beta}$ is a coefficient describing any correlation that might exist between the random and systematic components of the unexplained variance. For example, if thermal effects caused a systematic drift in the instrumentation and a simultaneous increase in the random scatter of the data, this correlation coefficient would be non-zero.

Equation 8 shows that statistical dependence results in a bias shift in the estimation of the population mean. This is attributable to the fact that at any point in time during which the sample is being acquired, the ordinary random variations that occur in any data set are distributed not about the true population mean, but about a value that is offset from the true mean by whatever the systematic error is at that moment. In other words, the systematic error behaves as a time-varying bias error, which is precisely what it is. The quantity β in equation 8 is simply the average value of this bias error over the time interval in which the sample was acquired.

The impact of statistical dependence on variance estimates is more complicated, but two cases in the limit of large n are interesting. Assuming $f(\rho)$ is bounded for large n as it was in the special case of lag-1 autocorrelation that we developed earlier (see equation 7), for large n equation 9 reduces to:

$$E\{s^2\} = \sigma^2 + \sigma_\beta^2 + 2\rho_{e,\beta}\sigma\sigma_\beta \tag{10}$$

If the random and systematic components of the unexplained variance are uncorrelated so that $\rho_{e,b} = 0$ (which we would expect in general), this further reduces to:

$$\boxed{E\{s^2\} = \sigma^2 + \sigma_\beta^2} \tag{11}$$

That is, systematic variation can cause the expectation value of the sample variance to be a biased estimate of the population variance. These bias errors, for both the sample mean and the sample variance, are functions of transient systematic variations. This is a potential cause of irreproducibility in wind tunnel test results.

Consider now the case in which the random and systematic errors are perfectly correlated (either positively or negatively). In that case, $\rho_{e,b} = \pm 1$, and equation 10 reduces to:

$$E\{s^2\} = \sigma^2 + \sigma_\beta^2 \pm 2\sigma\sigma_\beta = (\sigma \pm \sigma_\beta)^2 \qquad (12)$$

This case has little practical interest because the concept of perfect correlation between random and systematic error is invalid, but it represents a reassuring limiting case that helps validate the derivation. If there was perfect correlation, then the systematic errors would not be systematic at all, but would simply represent an increase in the magnitude of random error for positive correlation or a decrease for negative correlation. The fact that equation 12 says precisely that provides some additional confidence in the analysis.

Equations 8 and 11 tell us that the addition of systematic error biases our estimates of both the population mean and the population variance. The expectation value of the sample mean is not the true population mean as we presume, nor is the expectation value of the sample variance the true population variance. These results are especially troubling because of our dependence upon finite samples to achieve reproducible insights into the general population.

Impact of systematic error on data structures

In the acquisition of a typical polar, angle of attack (alpha) levels are almost always varied sequentially over some range of levels from smallest to largest, despite the fact that for a typical 15-point polar, say, there are 15! - 1 = 1,307,674,367,999 (1.3+ *trillion*) other permutations of the set-point order from which to choose. The polar could be constructed from alpha levels acquired in any of these permutations, simply by plotting the data in increasing order of alpha. That is, while the data must be *plotted* as a monotonically increasing function of alpha to produce a conventional pitch polar, there is no reason in principal that it must be *acquired* in that order.

A common reason for sequential ordering is that it results in the highest possible data acquisition rate, which is widely perceived as an important productivity consideration. Also, sequential ordering minimizes hysteresis effects caused by different flow attachment mechanisms that depend on the direction of change in angle of attack. A pitch polar acquired in increasing order of alpha can differ from a similar polar acquired in decreasing order of alpha, for example, so a certain consistency is achieved when all set point levels are approached from below. Finally, researchers often claim that there are insights to be had by watching how a polar develops as it is acquired, and that this enables early detection of various potentially pathological conditions.

While sequential ordering does have the virtues of speed and consistency that are responsible for its popularity in conventional wind tunnel testing, there is a serious drawback. Sequential ordering of set-point levels ensures that the forces and moments are acquired as a function *time* as well as a function of angle of attack. If systematic variations are in play as the polar is acquired, the resulting data set will contain *both* the effects of systematic changes in alpha, *and* the effects of systematic changes in some unknown source of unexplained variation.

This is a different and much more serious condition than if the errors are all random. Random errors will cause some scatter in the data, but that scatter can be expected to occur about the true alpha dependence. That is, the true polar will be revealed, except that there will be some "fuzz band" that reflects chance variations in the data. (If there are constant bias errors, the polar may also be displaced, but the *shape* of the polar, which reveals the change in response due to a specified change in alpha, will be correct. We will have more to say about such bias errors presently.)

If the errors are systematic instead of random, the polar can actually be disfigured in various ways. This is because data acquired by setting independent variable levels sequentially in time will depend upon changing response levels attributable both to the effects of time-varying alpha settings *and* to the effects of time-varying response changes due to unknown systematic errors. That is, the polar will reflect the true *time*-dependence of the data, but this will not generally be the true *alpha* dependence if systematic variations are in play. We say in such cases that the true angle of attack effects are "confounded" by the unknown sources of systematic variation. When alpha effects are *confounded*, it means there is no way to know how much of an observed change in response is due to systematic variations in alpha, and how much is due to systematic variations in something else, such as temperature or flow angularity or instrumentation drift, or any of a large number of other sources of systematic variation that are possible when real experiments are performed.

Systematic variations are more troublesome than simple random fluctuations in the data because they complicate the association of observed effects with

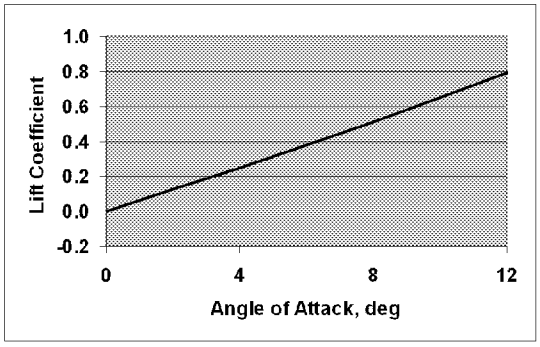

Figure 4. Hypothetical "true" lift polar.

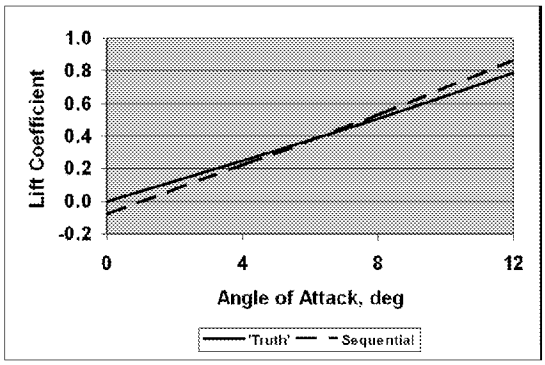

Figure 5. Effect of systematic variation on a sequential lift polar.

specific causes, wreaking havoc with our conventional perceptions of cause and effect. For example, consider a lift polar acquired over a range of pre-stall alpha settings in the presence of a persistent systematic variation that causes lift measurements made early in the polar to be biased somewhat low, while those acquired later in the polar to be biased somewhat high. If the systematic variation occurs at a constant rate, the result of its imposition on the true alpha effect is to rotate the polar counterclockwise. The slope of the measured polar is greater than the slope of the true polar, causing the lift for higher angles of attack to be overstated while the lift at lower angles of attack are understated. If the systematic variation does not occur at a constant rate, the polar can be misshaped as well as rotated. Also, certain fine structure in the polar might be attributable to variations in the systematic error that have nothing to do with angle of attack effects.

It is also possible for systematic variations to occur *between*, but not *during* two time periods. In such a case the polars will be displaced by whatever systematic change occurs between the two blocks of time. Veteran wind tunnel practitioners who have known the frustration of imperfect polar replicates have experienced these kinds of between- and within-polar systematic variations first-hand.

Random errors in individual observations impose their presence on us every time we replicate a data point because previous points serve as a reference by which to judge the latest point, but routine replication of entire polars is often limited (or omitted entirely) as a concession to demands for high-volume data collection. We have relatively infrequent opportunities to judge the shape of a polar against some comparable reference in such circumstances, and it is therefore possible to suffer from systematic errors without even knowing it. This is one reason systematic errors are so troubling – they are much harder to detect than random errors. For example, if a lift polar is rotated by systematic variation, we may not find out about this until much later – sometimes not until another tunnel entry or until subsequent flight tests have proven disappointing. Even when polars are replicated during the same experiment, if one polar differs from another there is seldom any objective basis for selecting which of the two (if either) is the "true" polar.

We will illustrate with a specific example how systematic within-polar variation can generate errors in a common pitch polar that are difficult to detect. Later, we will return to this same example to illustrate a tactical defense against such errors.

Consider a lift polar consisting of twenty observations from which lift coefficient values are estimated at each of four unique angles of attack: 0°, 4°, 8°, and 12°, so that each angle of attack is replicated 5 times. Figure 4 represents the resulting polar. For the purpose of this example, we will assume that this polar is comprised of completely error-free data. That is, we will assume that figure 4 represents the true polar, devoid of any effects of either random or systematic error.

Now assume that some systematic variation is in play while the data for this polar are acquired, and assume further that the data are acquired in the usual way, by setting angle of attack levels sequentially from smallest to largest. The systematic variation would then be superimposed upon the polar in figure 4. We postulate a large systematic error in lift coefficient of -0.10 when the first data point is acquired, that is incremented by 0.01 for each subsequent measurement. We also have assumed a random component of the unexplained variance with a standard deviation of 0.018.

The effect of the postulated systematic error on the first data point is that it will be biased low by 0.10 in addition to the effect of random variation. The next will be biased low by 0.09, and so on, until the systematic error term reaches zero for the 11th reading.

The next reading will then be biased 0.01 too high, the one after that will be 0.02 too high, and so on, until the 20th reading is acquired, with a positive bias of 0.09. Table I displays these errors. Figure 5 compares the lift polar of figure 4 with an identical polar consisting of a first-order polynomial function of alpha fitted to the data of figure 4, upon which the systematic errors of Table I are superimposed along with random errors drawn from a normal distribution with mean of zero and standard deviation of 0.018. The systematic error has caused a rotation in the polar, as described previously.

It is important to note that without the "true" polar to serve as a reference in figure 5, there would be no way to tell that a systematic variation had caused us to generate an incorrect lift polar. We would overstate the lift at high angle of attack, understate it at low angle of attack, and have no way of knowing that our lift measurements were systematically biased. It is this "stealth" aspect of systematic variation that makes it so hard to detect and therefore so easy to ignore.

Table I. Systematic error in a conventional sequential lift polar

Run Order	Angle of Attack	Systematic C_L Error
1	0	-0.10
2	0	-0.09
3	0	-0.08
4	0	-0.07
5	0	-0.06
6	4	-0.05
7	4	-0.04
8	4	-0.03
9	4	-0.02
10	4	-0.01
11	8	0.00
12	8	0.01
13	8	0.02
14	8	0.03
15	8	0.04
16	12	0.05
17	12	0.06
18	12	0.07
19	12	0.08
20	12	0.09

Impact of systematic error on response models

We often wish to model responses such as forces and moments by developing mathematical response functions to describe how they depend on various independent variables. We might use least-squares regression techniques to fit the data to a specific model, or absent any candidate model, a general Taylor series can be used to represent the unknown functional dependence. The Taylor series is typically truncated to include only terms of high enough order to assure an adequate fit without fitting noise.

It is convenient to code the independent variables as a prelude to developing a response model, by applying a linear transformation that both scales and centers the variables. If ξ_i represents an independent variable in physical units and ξ_{imin} and ξ_{imax} are the upper and lower limits of the range of this variable, then the following transformation will map ξ_i into x_i, a coded variable that ranges from -1 to +1, and is 0 at the midpoint of the range.

$$x_i = \frac{\xi_i - \frac{1}{2}\left(\xi_{i\max} + \xi_{i\min}\right)}{\frac{1}{2}\left(\xi_{i\max} - \xi_{i\min}\right)} \tag{13}$$

For example, the alpha values for a pitch polar spanning the range of -4° to +10° can be coded by substituting these values for ξ_{imin} and ξ_{imax} in this formula, yielding $x_i = (\xi_i - 3)/7$. So the center of the alpha range, $\xi_I = 3°$, codes into $x_i = 0$, the upper and lower limits of +10° and -4°, respectively, code into ±1, and all other alpha values for this polar code into x_i values in the range of ±1. After such a variable transformation, a second order Taylor series in two variables would be of this form:

$$\begin{aligned} y &= b_0 + b_1 x_1 + b_2 x_2 \\ &\quad + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2 \end{aligned} \tag{14}$$

where y is some response of interest (e.g., lift), the x_i are the independent variables (e.g., alpha and Mach number), and the b_i are regression coefficients proportional to the partial derivatives of the Taylor series that we numerically determine by a least-squares fit to the data.

The coefficient for each term in the series is subjected to the same formal inference procedure as described above. Based on the uncertainty associated with each coefficient, a reference distribution is determined under a null hypothesis that the expectation value of the coefficient is zero. If the magnitude of the coefficient determined through least-squares regression is small enough that the reference distribution under the null hypothesis suggests its departure from zero can be attributed to simple chance variations in the data, that term is dropped from the Taylor series model.

A detailed tutorial on regression is beyond the scope of this paper but in brief, the variance of a

reference distribution for each regression coefficient under the null hypothesis is determined from a corresponding element of the diagonal of a covariance matrix and is proportional to σ^2, the population variance for the response we measure. Unless the random sampling hypothesis holds, if we estimate this parameter from the sample variance we will be in error by equation 9. Furthermore, each coefficient is determined from a weighted, linear combination of the y_i that comprise all of the observations in the sample. These estimates will be biased if the random sampling hypothesis fails, by equation 8. In such a case, systematic variation will induce errors in both the estimate of the regression coefficient and the variance of the distribution by which it will is evaluated to test the null hypothesis, increasing the risk of an inference error. The result of such an error would be to retain extraneous terms in the model (if the null hypothesis is erroneously rejected for one or more coefficients), or to fail to include significant terms (if the alternative hypothesis is erroneously rejected).

For example, if the null hypothesis for the b_{11} term in equation 14 is erroneously rejected, we would assume that b_{11} was zero when in truth it was not. We would therefore drop it from the model, failing to correctly predict curvature in the x_1 variable. Likewise, if there actually was no significant curvature and we failed to reject the null hypothesis for b_{11}, we would incorrectly forecast curvature for x_1. In either case, not only would our response model fail to make accurate predictions, but we might also lose valuable insights into the underlying physics of the process.

The effects of biased estimates of the mean (equation 8) are felt in a special way in the b_0 coefficient of equation 14. This coefficient is a constant that represents the y-intercept of the response function. After the coding transformation of equation 13, it is computed by simply averaging all of the response measurements in the data set. Hypothesis testing is not normally applied to assess whether this term is real or not, although this can be done in circumstances for which an objective test is desired to determine if the response model passes through the origin (i.e., to determine if b_0 can be reliably distinguished from 0). For example, if the response function represents a calibration curve relating the output of a transducer to its input, the intercept term is expected to be zero in cases where zero input should produce zero output. In such cases, a rejection of the null hypothesis for b_0 may indicate that the calibration function needs to be improved.

The effect of within-polar systematic variation is to bias the y-intercept per equation 8, so that not only is the shape of the response function misrepresented because of terms that are erroneously dropped or retained due inference errors in assessing the reality of the individual coefficients, but the level to which changes in the response function is referenced can be either too high or too low, biasing all predicted response values accordingly.

In summary, the random sampling hypothesis is necessary for developing reliable response models from experimental data. If we acquire data for which the random sampling hypothesis does not hold, we can generate response models that are both misshaped and biased.

Evidence of Autocorrelation in Real Experimental Data

We have introduced the random sampling hypothesis and described some of the consequences of acquiring data when it does not apply. These are conditions for which data points are more alike when they are acquired over shorter intervals than longer intervals. Such conditions can be attributed to systematic sources of unexplained variance that persist over time, such as temperature effects, instrumentation drift, etc.

We have seen that when the random sampling hypothesis does not apply, sample statistics such the mean and variance are not reliable estimators of the corresponding population parameters. We have also seen that the risk of inference errors is inflated under such conditions, and that common data structures such as a pitch polar can be shifted, rotated, or bent due to within- and between-polar systematic variation, disguising true underlying stimulus/response relationships.

Given the substantial negative impact that systematic variation can have of sequentially acquired data, it is important to ask just how frequently such conditions exit in typical ground test experiments. If they are sufficiently rare, we may be justified in ignoring them on a cost/benefit basis, notwithstanding the fact that they can be troublesome when they are present. The argument in such a case would be that it is not cost-effective to "chase ghosts" that are not likely to harm us. We will note the results of a long-term investigation at NASA Langley Research Center that provides convincing evidence that systematic variations are not rare, and we will also summarize the results of a recent wind tunnel experiment in which one of the specific objectives was to quantify how often systematic variations can be detected in a representative wind tunnel experiment if an effort is made to do so. However, it is worth noting first that a substantial volume of anecdotal evidence already exists to support the notion that systematic variations are routinely recognized, if only implicitly, in conventional wind tunnel testing.

Much of the standard operating procedure in a conventional wind tunnel test is devoted to countermeasures against systematic variations that are implicitly understood to be in play. For example, the prudent wind tunnel researcher seldom lets more than an hour elapse between wind-off zeros. This is tacit recognition of the fact that various subtle instabilities in the measurement systems can have a cumulative effect over prolonged periods of time. The intent of frequent wind-off zeros is to minimize this effect by essentially resetting the system to a constant reference state periodically. Unfortunately, this procedure does nothing to defend against adverse affects caused by meandering systems *between* wind-off zeros. There is an inherent assumption that if the zeros are acquired frequently enough, the system will not have had time to shift far enough between zeros to be of serious concern, but perceptions of what constitutes "frequently enough", "far enough", and "serious" are generally left to the subjective judgment of the researcher. There is no guarantee that some effect did not come into play between zeros to invalidate the random sampling hypothesis for much of the data acquired in that period.

Wind-off zeros are just one of a number of standard wind tunnel operating procedures that reveal a general cognizance of persisting systematic variations and the need to establish formal procedures to defend against them. Data systems are routinely calibrated over short time intervals, for example; daily calibrations are common, and calibrations as often as every few hours are not unusual. Clearly this would be unnecessary in a stable environment in which nothing ever changed over time. Frequent model inversions to quantify flow angularity are also a staple of conventional wind tunnel testing. Again, the reason is clear. It is only necessary to make regular corrections for flow angularity under conditions for which the flow angularity changes over time. The same can be said for the reason that electronic pressure instrumentation is calibrated so frequently during a typical wind tunnel test, and why occasional adjustments are made to automated control systems to minimize set-point errors. "Things change" is one of the most reliable maxims in all of ground testing.

The effect of changes that persist over prolonged periods of time is to invalidate the random sampling hypothesis, with the attendant adverse effects documented earlier. These effects result from conditions in which the differences between replicates acquired over longer periods are not the same as the differences between replicates acquired over shorter periods. A wind tunnel testing technology development of major significance has been the careful documentation over a period of years by Hemsch and colleagues at NASA Langley Research Center that routine differences do in fact exist between what they describe as "within-group" and "between-group" variance estimates[5,10]. "Within-group" observations are those acquired over relatively short periods of time – minutes, typically – in which the variance can be attributed primarily to ordinary chance variations in the data that result in common random error. The "between-group" variance is associated with ordinary random error *plus* the effects of changes in within-group sample means that can be attributed to systematic variation persisting over relatively long time periods. The magnitude of the between-group variance has been shown by Hemsch and his associates to consistently and substantially exceed the magnitude of within-group variance. Typically, between-group variance estimates are factors of 2-3 times as large as within-group variance estimates, or more. The fact that such large differences are so consistently reported between short-term and long-term variance estimates is an indication that systematic variation is both common and significant in typical ground testing scenarios. Equation 11 is simply an analytical representation of what Hemsch discovered empirically – that over prolonged periods of time the variance due to ordinary random errors, σ^2, is augmented by a component due to systematic sources of variance, σ_{β}^2, that persist over time. The expectation value of sample variance reflects both of these effects, rendering the sample variance an unreliable estimator of population variance when only random variations are assumed.

An experiment to detect autocorrelation

An experiment was recently conducted at Langley Research Center in which a number of conventional lift polars were replicated, but with a slight variation to facilitate the detection of within-polar systematic variation that can otherwise be so difficult to detect, as noted above. (It turns out that this alteration in data acquisition procedure is also key to eliminating the adverse effects of systematic variation, as will be developed below.)

In this experiment, ten commercial jet transport wing configuration settings were examined at each of seven Reynolds numbers, all at a constant Mach number. For the seventy combinations of Reynolds number and configuration, two lift polars were acquired. One was a conventional polar in which angle of attack levels were set in increasing order in the usual way. In the other polar, the same angle of attack levels were set as in the conventional polar, but the order was set at random. In addition, one angle of attack (6°) was replicated several times. There were four 6° settings in half of the randomized polars, and eight 6° settings in the other half. The replicates were interleaved among the single-level alpha settings and were therefore uniformly distributed among all the points in the

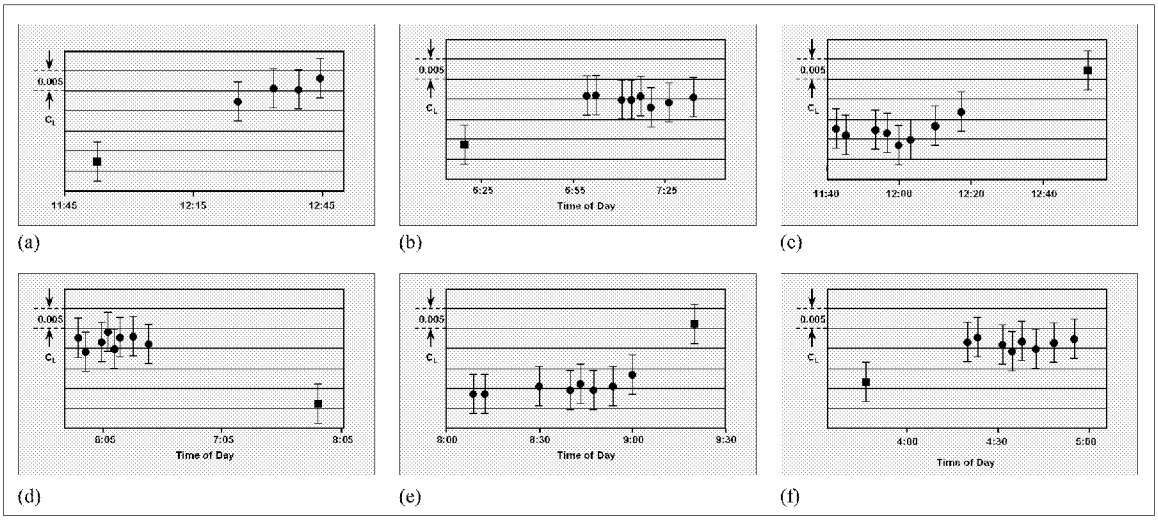

Figure 6. Time history of C_L replicates. Circles: randomized polar; square: sequential polar. (a) Between-polar variation with a positive slope continues into second polar. (b) Between-polar systematic variation of roughly 200% of the entire 0.005 error budget. (c) Onset of systematic variation in mid-polar, continuing between polars. (d) Between-polar systematic variation. (e) Systematic variation begins at the end of one polar and continues between polars. (f) Systematic between-polar variation.

randomized polar. Care was taken to ensure that all set-points were approached from below to eliminate hysteresis effects, by first going to a "home state" alpha setting below the smallest alpha level in the polar whenever the next point in the randomized sequence was at a lower alpha value than last point. The randomized and sequential polars were run back to back, with the order selected at random.

Systematic variation could be detected in this experiment in two ways. First, the replicated lift measurements could be plotted as a function of time. Absent systematic variation, these points should all be the same within experimental error, and their time histories should be generally featureless, displaying no particular trend. On the other hand, pronounced within-polar systematic variation should result in some structure in the time history of replicates acquired in the same polar. Between-polar variation should result in a significant displacement between the replicates acquired in the randomized polar, and the single lift point acquired at 6° in the conventional polar.

Plotting time histories of replicates has the disadvantage that it is a subjective way of establishing the presence of systematic variation, which can be in the eye of the beholder. Those who are inclined to fear systematic variation might see pronounced trends in time histories that seem featureless to those inclined not to be bothered by such effects. (Incidentally, this weakness is not confined to the search for correlation in replicate time histories. It applies whenever subjective judgment is the basis for conclusions drawn from the examination of graphs and other representations of data.) Nonetheless, a few examples of time histories are presented in figure 6 in which reasonably pronounced trends would have to be acknowledged by even the most reluctant observer. These trends reveal both within-polar and between-polar systematic variations that are not simply substantial portions of the entire 0.005 error budget declared for this experiment, but which are in fact significant *multiples* of the entire budget. In these figures, the circular symbols represent points acquired in the randomized polar at α=6°, and the square symbol represents the single α=6° point acquired in the corresponding conventional polar.

A technique for a less subjective approach to detecting systematic variations was outlined above in the general discussion of scientific inference. In short, we define a null hypothesis which in this case is, "No correlation among the observations in the sample", we construct a reference distribution representing how a relevant statistic should be distributed under that hypothesis, and we either reject the null hypothesis or not, depending on whether the observed value of that statistic is or is not generally within the range of values that would be expected if the null hypothesis were true.

This objective inference procedure leads to conclusions that are based upon a set of procedures and criteria agreed upon before the data are acquired. It has the virtue that it reduces our dependence on pure subjective judgment, which is vulnerable to subconscious prejudices and also to the conflicting judgments of others who may simply be inclined to see things differently[†].

In this specific study, the lift data from the randomized polars were fitted to polynomial functions of alpha serving as Taylor series representations of the unknown functional dependence of lift on alpha, as described above. The analysis was restricted to pre-stall alpha ranges for which the alpha dependence is dominated by the first-order term in a least-squares regression. However, smaller second-order terms were often found to be significant by this procedure, and so were third order terms on occasion. No significant terms of order four or higher were observed. The resulting first-, second-, or third-order polynomial functions of alpha were subjected to a battery of standard goodness-of-fit tests to assess their adequacy. The central criteria were that the magnitude of the unexplained variance be acceptably low (standard error no greater than 0.0025 in lift coefficient for an average "two-sigma" value of 0.005 over the alpha range), and that the residuals be randomly distributed about the fitted curve.

This latter criterion ensures that when residuals are plotted as a function of alpha, they are randomly distributed about zero and therefore contain no information to suggest that an alternative model would better represent the alpha dependence. We would conclude under such circumstances that additional alpha terms in our truncated Taylor series would not improve the fit.

Likewise, if the residuals plotted as a function of time are randomly distributed about zero there will be no information to suggest that an alternative model would better represent the *time* dependence we assumed in the lift model. Of course, we assumed no time dependence when we fit the lift data only to alpha, but this is equivalent to fitting the data to both alpha and time and discovering that the regression coefficients of all time-related terms in our model are zero. Featureless plots of residuals as a function of time would tend to validate this model and, analogous to the featureless plots of residuals against alpha, would support the conclusion that no additional (non-zero) time terms would improve it. We could infer in that case that as we acquired our data, the lift coefficient was changing only as a function of alpha and not also as a function of time.[‡] The random sampling hypothesis could be assumed for such data as there would be no net difference in the residuals of points acquired over relatively short time intervals and those acquired over longer intervals.

A total of seven standard hypothesis tests for correlation were applied to the residual time histories of each of the seventy randomized polars. These tests are listed in Table II. Detailed descriptions of each test can be found in standard references and are beyond the scope of this paper, but the basic approach was the same as for any formal hypothesis test: We formulated a null hypothesis that some measure of autocorrelation was zero and established a reference distribution for how that quantity would be distributed due to chance variations in the data if the null hypothesis were true. Observed values that were different from zero by more than could be reasonably attributed to ordinary chance variations in the data were interpreted as evidence of non-zero autocorrelation. We concluded that those polars were acquired under conditions for which the random sampling hypothesis did not apply.

Table II: Tests for correlated residuals

1	Durbin-Watson Test
2	Swed-Eisenhart Runs Test
3	Pearson's Product Moment Test
4	Test of Significant Trend in Residuals
5	Serial Correlation Test
6	Wilcoxon-Mann-Whitney Test
7	Spearman Rank Correlation Test

A significance level of 0.05 was specified. The number of polars for which the null hypothesis of no correlated residuals was rejected at this level varied

[†] As a practical matter this reduced reliance upon subjective judgment is not universally embraced, especially among those who regard "judgment" as a major element of their contribution to the research process. The intent here is not to suggest that judgment is irrelevant to those possessed of statistical expertise, but to say simply that objective techniques for making inferences can bring additional clarity to the judgment process. It also helps defend us against the prejudices of others, and our own.

[‡] Had alpha levels not been set in random order but instead had been monotonically changed with a constant data acquisition rate in the usual way, time and alpha effects would be completely confounded so that the time dependence of the residuals could not be distinguished from the alpha dependence. In such a case, plotting the residuals against time would provide no additional information beyond plotting them against alpha, because the two plots would be identical except for the scale labels for the abscissas.

somewhat from test to test. Out of 70 randomized polars a minimum of 11 and a maximum of 21 displayed evidence of correlated residuals, depending on the specific test, but the average was 17 times out of the 70 polars examined, or about 24%. That is, on average in 24% of the polars, we were able to say with at least 95% confidence that the random sampling hypothesis did not apply.

It is possible that the particular conditions of this test were unusually conducive to systematic variation and that 17 correlated polars out of 70 is unusually high. On the other hand, it is equally plausible that this specific test was conducted under conditions for which systematic variation would be expected to occur less often than usual, and that 17 correlated polars out of 70 represents a lower limit on what might typically be expected. We computed that the range of Bernoulli trial success probabilities must lie between 15% and 35% to observe 17 successes out of 70 at least 95% of the time. (Anything less that 15% would generate 17 or more cases out of 70 less often than 95% of the time and anything more than 35% would generate 17 or more cases out of 70 more often than 95% of the time.)

That is, we asked what the probability would have to be of an individual polar having correlated residuals if more than 17 out of 70 would be observed no more than 2.5% of the time. The answer is 15%. We likewise asked what the probability would have to be of an individual polar having correlated residuals if *less* than 17 out of 70 would be observed no more than 2.5% of the time. The answer to that was 35%. We therefore concluded that, given an observation of 17 correlated polars out of 70 tested, we could say with 95% confidence that the random sampling hypothesis would be expected to fail between 15% to 35% of the time.

Figure 7 illustrates for each of the seven specific tests noted in Table II what specific upper and lower limits were computed for the probability of an individual polar being acquired under conditions for which was random sampling hypothesis does not hold. There is some variability from test to test, but taken as a whole these tests support the general conclusion that to the extent that this test could be regarded as representative, correlated residuals could be expected between 15% and 35% of the time, or in roughly every 7^{th} polar at best, and every 3^{rd} polar at worst.

There are two reasons to suspect that these percentages are lower limits on the true frequency with which the systematic variations are in play in wind tunnel testing. First, the time series analysis methods used could only test for within-polar systematic variation. The time histories in figure 6 suggest that between-polar systematic variation occurs as at least as often as within-polar systematic variation.

Secondly, none of the correlation tests were very sensitive for samples as small as the number of pre-stall

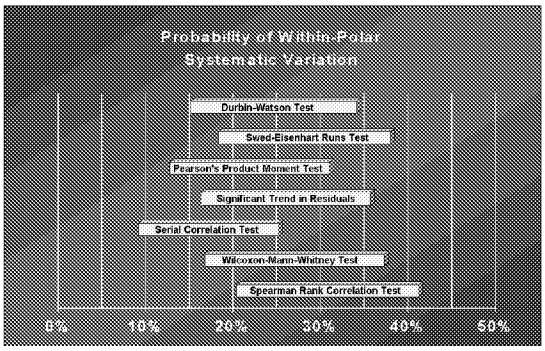

Figure 7. 95% confidence intervals for probability of systematic within-polar variation, various statistical tests.

points in a lift polar. The degree of correlation therefore had to be quite severe to register in these tests. It is quite likely that correlations large enough to be troublesome, but too small to be reliably detected with such small sample sizes, occur more often than the 15% to 35% range quoted here.

Tactical Defenses against Systematic Unexplained Variance

We have found that there can be serious consequences if we assume that the random sampling hypothesis applies when it does not, and we have also seen that the random sampling hypothesis probably fails too often in wind tunnel testing to safely take it for granted. It is not unlikely that much of the difficulty in achieving reliably reproducible wind tunnel results is due to variably biased estimates of population means and variances, caused by improper assumptions of random sampling.

Fortunately for the 21^{st}-century experimental aeronautics community, random sampling has been sufficiently elusive in other experimental circumstances besides wind tunnel testing that over the years certain effective tactics have been developed to defend against the adverse effects of its unwarranted assumption. Savvy experimentalists in other fields have long recognized that the random sampling hypothesis is simply too unreliable to count on consistently. They assume as a matter of course that it will not apply naturally when they acquire data, and take proactive measures to impose random sampling on their samples, using techniques to be described in this section.

Randomization: A Defense Against Within-Sample Systematic Variation

The problems induced by unstable sample means were first recognized by Ronald Fisher and his

associates early in the last century[6]. He proposed a conceptually simple yet effective tactical defense against the adverse impact of systematic within-sample variation. Fisher's idea was to set the levels of all independent variables in random order. We have already seen in the above discussion of residual time histories that randomization decouples the components of system response that can be attributed to intentional changes in the independent variables from changes that are due to unexplained sources of systematic variation that change as a function of time. This permits us to see the true dependence of system response on the changes we make in the independent variables, clear of the effects of any systematic variations. We can exploit this decoupling by adopting randomization as part of the standard operating procedure of a prudent researcher, as recommended by Fisher.

Table III. Systematic error in a randomized lift polar

Run Order	Angle of Attack	Systematic C_L Error
1	4	-0.10
2	4	-0.09
3	**0**	**-0.08**
4	8	-0.07
5	**0**	**-0.06**
6	12	-0.05
7	8	-0.04
8	12	-0.03
9	4	-0.02
10	**0**	**-0.01**
11	8	0.00
12	8	0.01
13	12	0.02
14	4	0.03
15	**0**	**0.04**
16	8	0.05
17	12	0.06
18	**0**	**0.07**
19	4	0.08
20	12	0.09

To illustrate the effect of randomization in a common wind tunnel scenario, we revisit the lift polar example considered earlier, in which systematic within-polar variation rotated the polar as in figure 5. Recall in that case that the rotation of the polar was caused by superimposing on the true alpha dependence a

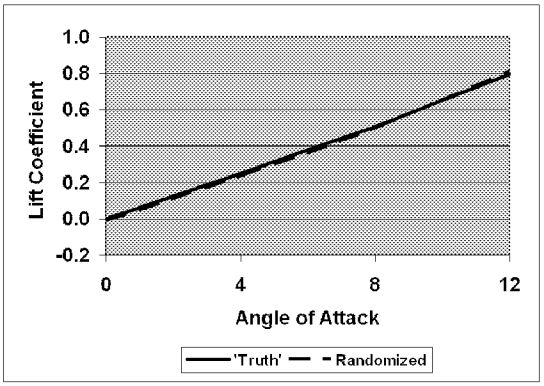

Figure 8. Effect of systematic variation on a randomized lift polar.

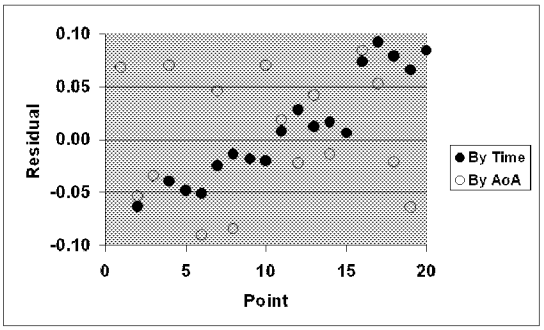

Figure 9. Residuals from randomized polar in order of alpha and in run order (by time).

systematic variation changing at a constant rate during the time the polar was acquired, starting with a lift coefficient bias error of -0.1 for the first data point and incrementing by 0.01 for each of the 19 remaining points in the polar, as in Table I.

Imagine now that we repeat the experiment under exactly the same conditions, except that we will set the angle of attack levels in random order, as indicated in Table III.

In the standard order case of Table I, zero degrees angle of attack is set in the first five measurements, where the bias error in lift coefficient was -0.1 through -0.06, for an average of -0.08. When we randomize the angle of attack set-point order however, zero degrees is set in the third, fifth, tenth, fifteenth, and eighteenth measurement as highlighted in Table III, where the systematic errors are -0.08, -0.06, -0.01, +0.04, and +0.07, respectively. The average systematic error in lift coefficient for these $\alpha=0^{o}$ points is -0.008, an order of magnitude less than in the conventional sequential-order case.

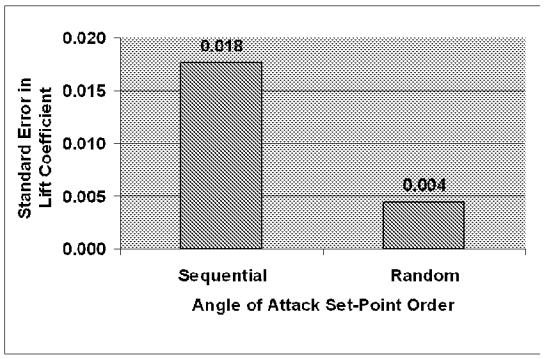

Figure 10. Impact of angle of attack set-point order on standard error in lift coefficient in presence of systematic variation.

The systematic component of error is so much less when the alpha levels are set in random order than when they are set sequentially because randomizing causes some of the lower-alpha values to be acquired earlier – when the systematic errors are negative – and some to be acquired later – when the systematic errors are positive. The same is true for the higher alpha values – some are acquired earlier and were therefore negative while some were acquired later and were positive. As a result, the systematic variation is converted to ordinary random fluctuations occurring above and below an unbiased estimate of the true polar shape. These random errors are minimized by replication, which tends to cause positive and negative random errors to cancel.

Figure 8 plots the polar constructed from a first-order polynomial fit to the data associated with randomized angle of attack set-point levels and compares it to the "true" polar of figure 4. A comparison of figures 8 and 5 shows that when systematic within-polar variations are in play, a randomized polar compares much better with the true polar than one acquired with sequential alpha settings.

Figure 9 displays the residuals from the randomized polar plotted two ways. The open circles show the residuals plotted as a function of increasing angle of attack while the filled circles show the residuals plotted in run order; that is, as a function of time. Absent any systematic variation during the polar, the run order would be irrelevant and both the plot as a function of alpha and the plot as a function of time would be featureless for a good fit to angle of attack. However, figure 9 quite clearly shows structure in the residual time history that reveals how earlier points were biased lower while later points were biased higher. Notwithstanding this systematic bias, the featureless plot of residuals against alpha suggests that there is no substantive deficiency in the modeled alpha dependence.

One option available to the researcher to further improve precision is to de-trend the residual time history, essentially correcting for the time-varying component of the unexplained variance. However, this option is only available in cases for which the systematic variation is sufficiently extreme to be seen unambiguously in a residual time history of a sample as small as that acquired in a polar. The more common case is for systematic variation large enough to invalidate the random sampling hypothesis to be nonetheless too subtle to detect by visual inspection.

We can quantify the improvement wrought by randomization in this illustration because we have stipulated what the "true" polar is, and can therefore quantify differences between this and the polars acquired under systematic variation for both the sequential and randomized set-point ordering schemes. We compute a mean square error, $\hat{\sigma}^2$, by summing the squared deviations of the observed polar from the true polar and dividing by the number of observations – 20 in this case.[§] The prediction error will depend on alpha, being smaller near the center of the polar than near the ends, but the average mean square error over all points in the sample is $p\hat{\sigma}^2/n$, where p is the number of parameters in the model (2), and n is the number of points in the sample (20). The square root of this is the standard prediction error for lift coefficient (the "one-sigma" error in model prediction); 0.018 for the sequential polar and 0.004 for the randomized polar, as figure 10 shows.

Randomization has reduced the model prediction error by more than a factor of 4. To have achieved a similar reduction through replication alone would have required 16 times as much data, even assuming the original errors were random. For the relatively modest cost associated with acquiring the data in a random alpha sequence, the same benefit could be achieved without acquiring any additional data. In any case, given that the errors were largely systematic, no amount of additional data would have reduced the error significantly when the angle of attack levels were set sequentially, so conventional replication alone would have been ineffective no matter how much extra data had been acquired. This is another reason that gratuitously maximizing data volume is no guarantee of

[§] If we were computing this mean square error from a *fitted* estimate of the true polar, we would divide by 18 rather than 20, since in that case two degrees of freedom would be lost to estimates of the slope and y-intercept for a first-order function of alpha. In this case, the slope and y-intercept are assumed to be known, so no degrees of freedom are lost in estimating them.

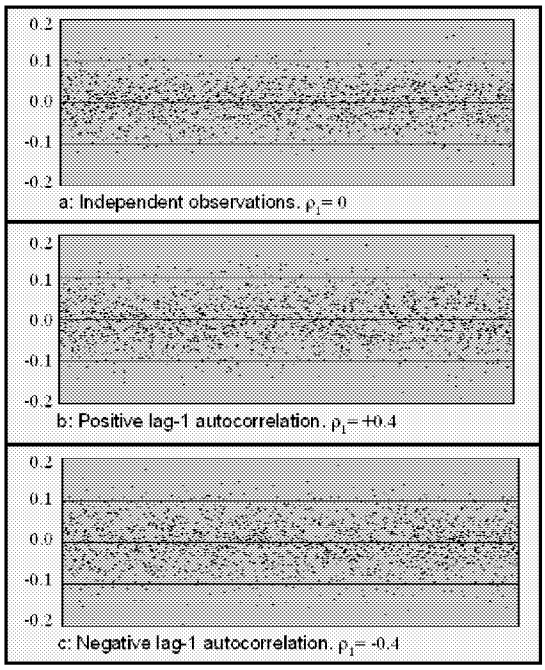

Figure 11. Time history of 1000 normally distributed C_L errors with three states of lag-1 autocorrelation.

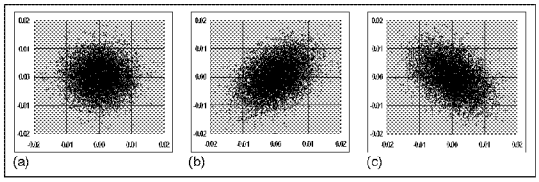

Figure 12. Plots of e_i versus e_{i-1} for three states of lag-1 autocorrelation, data drawn from a normal distribution with mean of 0 and standard deviation of 0.005. (a) $\rho_1 = 0$. (b) $\rho_1 = +0.4$. (c) $\rho_1 = -0.4$.

reduced inference error risk unless precautions have been taken to ensure the random sampling hypothesis.

Note also that each data point was modeled in this example as the sum of three quantities: the true lift coefficient as represented by figure 4, a large systematic error from tables I and III, and a random error drawn from a normal distribution with mean of zero and standard deviation of 0.018. The corresponding standard deviation in the distribution of sample means would be in this case $0.018/\sqrt{20} = 0.004$. To three decimal places, this is the same numerical value as the mean standard error in the prediction for the randomized polar, as figure 10 shows. This standard error in the prediction includes both this random component of error, plus any residual systematic error, so to three decimal places randomization in concert with replication has essentially eliminated the systematic component of error, notwithstanding the fact that it was the dominant source of error at the start. Clearly the dominant error is now random, and the random sampling hypothesis is therefore restored.

Note that randomization has defended us against a serious systematic error that we are not likely to have detected had we acquired the data in sequential order. Our first inking of a problem might not have occurred until much later, perhaps in another wind tunnel test or in a flight test, when it would have become apparent that our predictions of lift coefficient overstated the lift substantially at higher angles of attack and understated it at lower angles of attack. Such easily avoidable errors are often attributed in hindsight to poor wind tunnel facility performance when they are ultimately detected, but a rather more honest appraisal might include the failure of the researcher to use standard available precautions in the design of the experiment to insure against systematic errors that invalidate the random sampling hypothesis.

In this example we considered an exaggerated systematic error simply to be able to resolve its effects graphically in a comparison with the true polar, as in figure 5. Obviously the true polar is never available in real experiments to make such a comparison, and we are usually faced with systematic variations that cause errors much smaller than 0.100 in lift coefficient, which is as much as two orders of magnitude larger than the entire error budget in some high-precision performance wind tunnel tests, for example. We will therefore now consider more realistic levels of error in which the correlation may be too subtle to detect readily. We will examine the impact of such subtle correlation on our ability to extract reliable inferences from finite samples of data.

Figure 11 presents three hypothetical lift coefficient error time histories, each with 1,000 observations. In figure 11a, the errors are drawn from a normal distribution with a mean of 0 and a standard deviation of 0.005, satisfying requirements that would not be atypical for a common wind tunnel configuration test. All of the errors are independent of each other so there is no correlation among the individual points. In other words, the random sampling hypothesis would be valid for any data set characterized by this error time history.

Figures 11b and 11c display time histories that differ from the one in 11a in that a mild positive correlation is introduced in 11b, and a mild negative correlation is introduced in 11c. Specifically, the points in 11b and 11c were generated to ensure that each was

correlated with the one immediately preceding it. To produce the data set in 11b and 11c we set $e_i = d_i + ke_{i-1}$, where d_i is the i^{th} error without correlation, drawn from the same normal distribution with mean of zero and standard deviation of 0.005 that produced 11a. Thus e_i is the i^{th} error with correlation, and k is a constant chosen in this case to generate a lag-1 autocorrelation coefficient, ρ_1, of +0.4 for 11b and -0.04 for 11c, computed as follows:

$$\rho_1 = \frac{\sum_{i=1}^{n-1}[(y_i - \bar{y})(y_{i+1} - \bar{y})]}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (15)$$

All three time histories in figure 11 appear qualitatively indistinguishable, despite differences in the degree of autocorrelation. This illustrates how difficult it is to detect correlation without a special effort to do so. A graphical method for revealing lag-m autocorrelation is to plot the i^{th} residual against the $i\text{-}1^{st}$ residual. In figure 12, such plots were constructed using the corresponding data from figure 11. The uncorrelated data points generate a symmetrical pattern but the positively and negatively correlated points display a pronounced positive and negative slope, respectively.

We now construct a null hypothesis that we know to be true, which is that the difference in two ten-point sample means drawn from the data sets of figure 11 is zero. (Each individual point was drawn from a normal distribution with a mean of zero, so any 10-point sample mean has an expectation value of zero.) We construct a reference distribution corresponding to this null hypothesis in the usual way, and use it to determine if the number of times the null hypothesis is rejected is more or less than would be expected. In this case, the test statistic is constructed as follows:

$$t = \frac{\bar{y}_2 - \bar{y}_1}{s_p^2\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (16)$$

where the numerator contains the difference in the two sample means, n_1 and n_2 are the sample sizes (both 10 in this case), and s_p^2 is the pooled sample variance, computed as follows:

$$s_p^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2} \qquad (17)$$

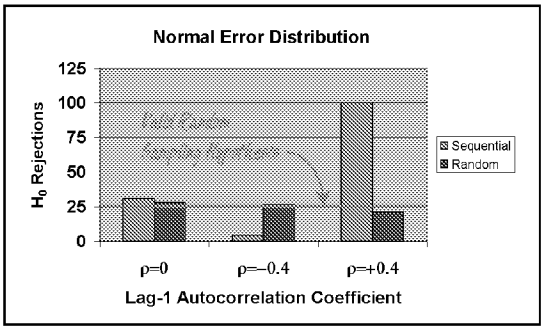

Figure 13. Impact of set-point order on inference error probability for different levels of correlation. Normal error distribution.

where s_1 and s_2 are the standard deviations estimated from the observations in the two samples.

Under the assumption of random sampling from normal populations, the statistic in (16) follows a t-distribution with $n_1 + n_2 - 2 = 18$ degrees of freedom, which serves as a reference distribution. We can compare this with a critical t-statistic corresponding to a significance level of 0.05, say, and accept or reject the null hypothesis depending on whether the computed t-statistic is less than or greater than the critical value. See the above discussion of reference distributions for more details.

Because the null hypothesis is known to be true in this case, we can expect it to be rejected in a two-sample t-test only because of chance variations in the data. Since we selected our reference t-statistic to correspond to a significance level of 0.05, this should occur in about 5% of the cases we try. Each data set in figure 11 has 1000 data points so we can compare a total of 500 unique pairs of 10-point samples. We would expect chance variations in the data to cause erroneous rejections of the null hypothesis $0.05 \times 500 = 25$ times under the random sampling hypothesis. Of course, we would not expect every individual 500-pair set of data to produce precisely 25 rejections every time, any more than we would expect 50 flips of a coin to produce precisely 25 heads every time. (If a fair coin is flipped 50 times, there is a 95% probability that heads will appear between 18 and 32 times, or in a range of ±7 times about the expected value of 25.)

We did in fact conduct the above-described t-test for a difference in means using all 500 unique pairs of 10-point samples that could be extracted from each of the three 1000-point data sets in figure 11. We conducted the test in two ways. First, we acquired samples in sequential order, comparing the mean of the first 10 points to the mean of the second 10 points for

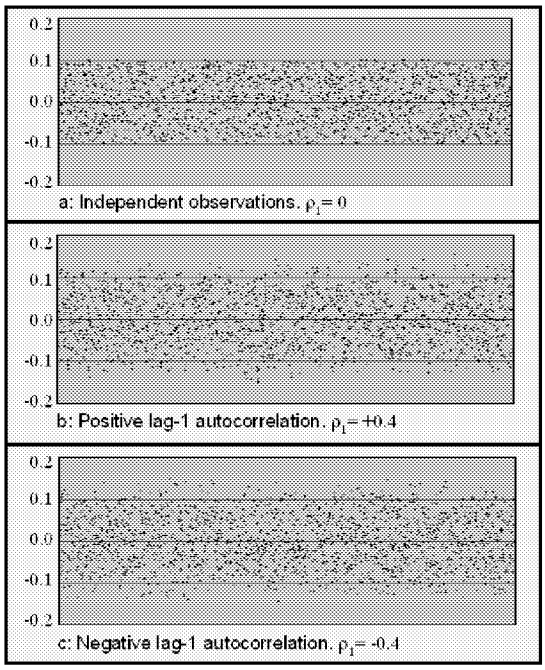

Figure 14. Time history of 1000 uniformly distributed C_L errors with three states of lag-1 autocorrelation.

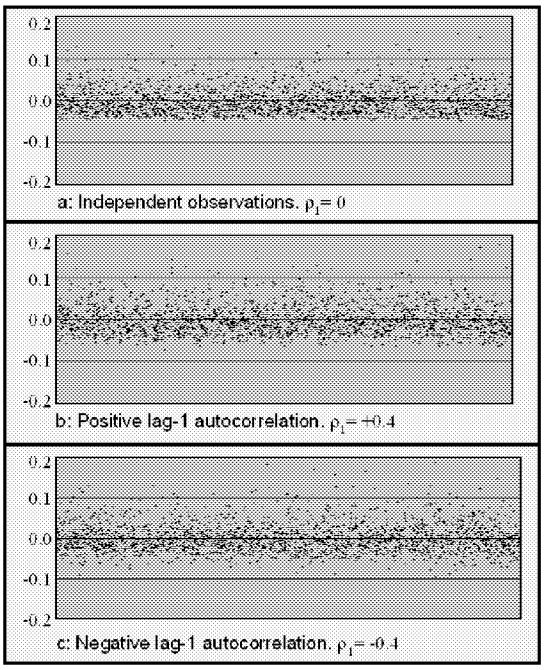

Figure 15. Time history of 1000 C_L errors drawn from a skewed (4-df chi-squared) distribution with three states of lag-1 autocorrelation.

the first comparison, the mean of points 21-30 and the mean of points 31-40 for the second, and so on until the entire 1000-point data set was exhausted and 500 unique 10-point sample pairs were compared. We then repeated the process after completely randomizing the order of the 1000 data points in each of the three data sets in figure 11. The number of times the null hypothesis was erroneously rejected is displayed in figure 13.

Figure 13 shows that when the data points were independent, sequential and random ordering resulted in 31 and 28 rejections of the null hypothesis, respectively, reasonably close to the expected value of 25. Likewise, for the points sampled in *random* order both positively and negatively correlated data yielded in the neighborhood of 25 rejections (26 and 21, respectively). However, when the correlated data were sampled in *sequential* order, the t-test for a difference in means gave results that were substantially different from what would be expected under the random sampling hypothesis. The negatively correlated data generated only *four* rejections of the null hypothesis, and there were *one hundred* rejections when the data were positively correlated.

We have an expectation of 25 rejections of the null hypothesis when a criterion corresponding to a significance of 0.05 is used with 500 trials. However, this expectation is predicated on two assumptions. One is that the random sampling hypothesis holds, but the other is that the sample means are drawn from a normal distribution. We are relying on the Central Limit Theorem for assurance that the difference in two 10-point sample *means* is drawn from a normally distributed population, regardless of the distributional properties of the populations from which the two individual samples were drawn. We therefore expect the quantity in (16) to follow a t-distribution if the random sampling hypothesis holds. However, it is possible that the difficulty we are having with the sequential ordering of correlated data is not that the points are not independent, but that for relatively small (10-point) samples the sample means fail to approximate a normal distribution adequately. Therefore the solution might not necessarily be to ensure independence in the individual observations, but instead to rely on larger sample sizes (more data) for a better approximation of a normal distribution via the Central Limit Theorem. To test this, we repeated the above experiment with two additional 1000-point data sets. In one of them, the individual data points were all drawn from a uniform distribution. That is, the probability distribution was not the familiar bell-shaped Gaussian, but a rectangular "boxcar" distribution. In the other 1000-point data set, the points were drawn from a chi-squared distribution with four degrees of freedom, which is highly skewed. Since both of these

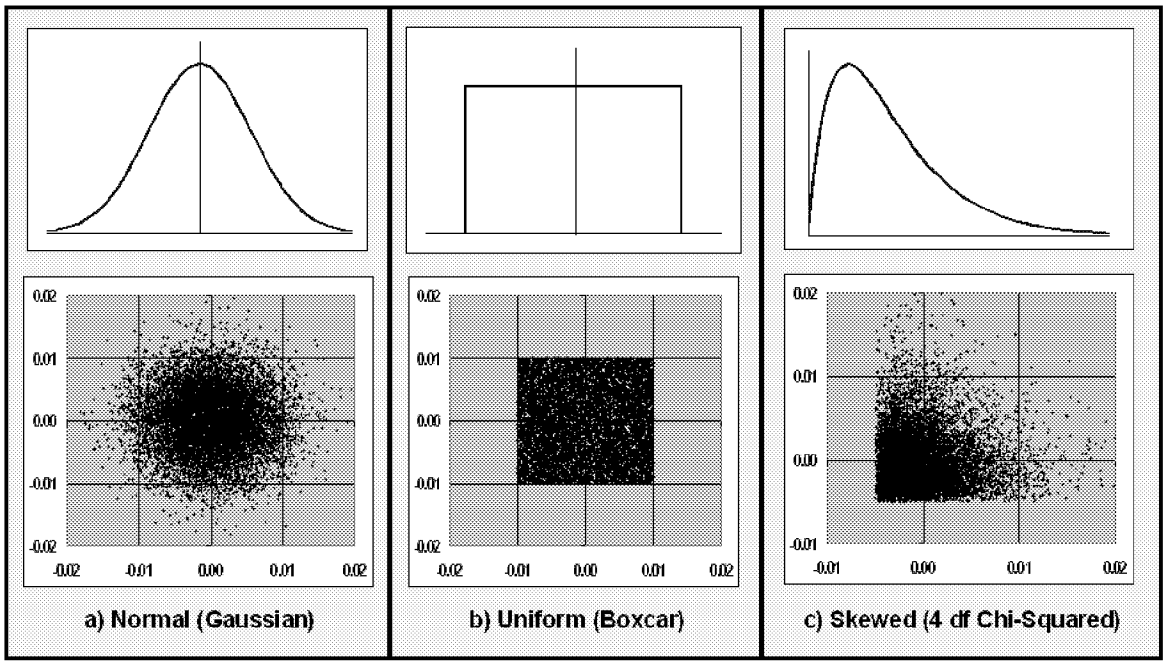

Figure 16. *Independent* lift coefficient errors. Plots of e_i versus e_{i-1} for samples with lag-1 autocorrelation coefficient $\rho_1 = 0$, drawn from three population distributions.

distributions are substantially different from normal, we would expect to see some change for the worse in the number inference errors generated in the two-sample t-test.

Figures 14 and 15 show the error time histories with no correlation, positive correlation, and negative correlation, for the uniform and skewed distributions, respectively. Note that just as in the case of errors drawn from a normal distribution, it is difficult to tell by inspection which time histories are independent, which exhibit positive correlation, and which are negatively correlated. Figures 16-18 compare the corresponding lag-1 autocorrelation plots with the case of normally distributed errors considered already. The uniform (boxcar) distribution produces relatively sharply defined boundaries in the autocorrelation plot as expected from its discontinuous boundaries, and the autocorrelation plots for the chi-squared distribution are much more densely populated in the lower left than in the upper right. Points in the lower left of the chi-squared autocorrelation plots correspond to two relatively small values occurring in succession, while points in the upper right correspond to two relatively large values occurring in succession. Because a low degree-of-freedom chi-squared distribution is strongly skewed to the right (long tail to the right, mode or peak shifted to the left), smaller values are relatively likely to be drawn and larger values are less likely. Two small values in a row are therefore not unlikely to occur, while it is it is much less likely that two high values will occur in a row, because single occurrences of high values are relatively rare. This explains the high density of points in the lower left of the chi-squared autocorrelation plots and the low point density in the upper right of these plots.

Figures 19 and 20 show how many times the null hypothesis was erroneously rejected for the case of individual errors being uniformly distributed and highly skewed, respectively, for lag-1 autocorrelation coefficients, ρ_1, of 0, +0.4, and -0.4. We see the same general pattern as in figure 13, where the errors were drawn from a normal distribution. That is, for all three distributions randomization ensures the random sampling hypothesis regardless of correlation. When correlated data are acquired as two successive samples, the assumption of random sampling leads to substantially more or substantially fewer rejections of the null hypothesis than one would actually encounter with negative or positive correlation, respectively. Since these conclusions held for all three distributions, the data in figures 13, 19, and 20 are averaged across distributions and presented in figure 21, which clearly shows how randomization stabilizes the inference error

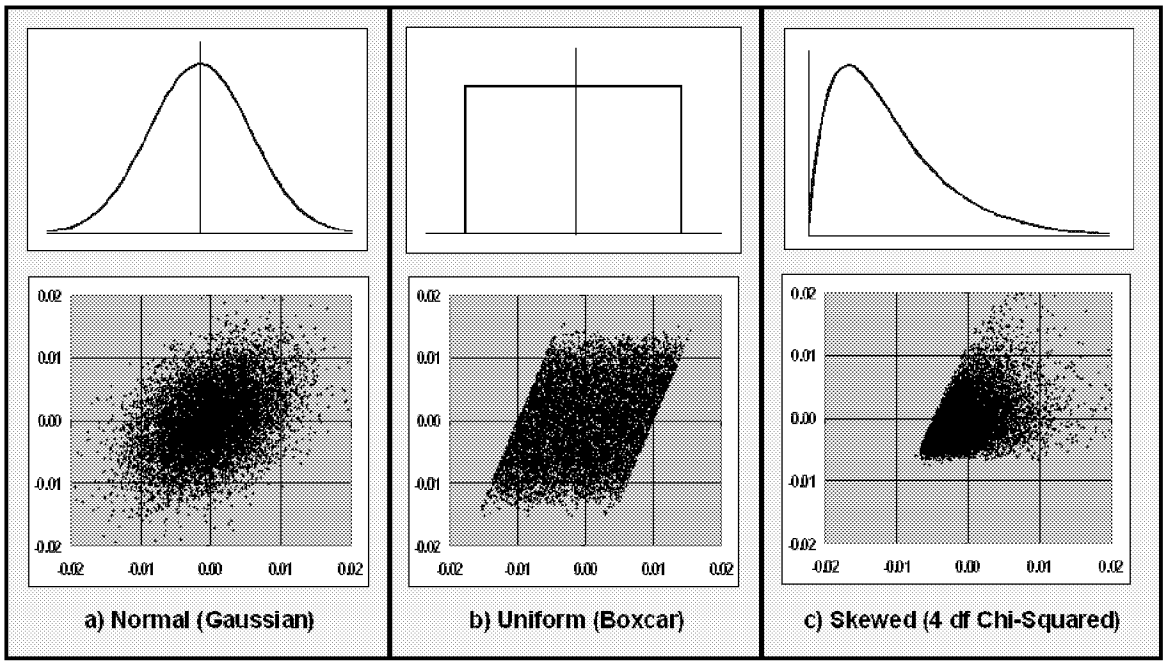

Figure 17. *Correlated* lift coefficient errors. Plots of e_i versus e_{i-1} for samples with lag-1 autocorrelation coefficient $\rho_1 = +0.4$, drawn from three population distributions.

risk (probability of erroneously rejecting the null hypothesis) when the random sampling hypothesis fails.

To understand why negative correlation results in fewer H_0 rejections than expected while positive correlation results in so many more, consider the variance of the difference between two successive averages of n observations serially correlated at lag 1. This is given by Box, et al.[1] as:

$$V(\bar{y}_B - \bar{y}_A) = \frac{2\sigma^2}{n}\left[1 + \left(\frac{2n-3}{n}\right)\rho_1\right] \qquad (18)$$

The term outside the square brackets is the familiar formula for the variance of the difference in two n-point sample means in the absence of correlation, which follows directly from applying the general error propagation formula, (1), to $\Delta\bar{y} = \bar{y}_B - \bar{y}_A$. The term inside the square brackets indicates how this variance changes when ρ_1 is not 0. This is analogous to equation 7, which describes the variance of a single n-point sample under lag-1 autocorrelation.

For the cases examined here, where n is 10 and ρ_1 is +0.4 and -0.4, the corresponding numerical values of the bracketed term are 1.68 and 0.32, respectively. The square roots of these terms, which indicate roughly[**] how much the widths of the distributions of measure t-statistics change due to non-zero ρ_1, are 1.30 and 0.57, respectively. Under the random sampling hypothesis, the critical t-statistic for an 18 df two sample t test at a significance of 0.05 is 2.101. If we assume this reference t-statistic while the width of the measured t distribution increases by 30% due to positive correlation, the area under this distribution to the right of the reference t-statistic will increase. This represents half of the percentage of erroneous rejections there will be of the null hypothesis (the other half coming from negative t-values less than -2.101). This is why positive correlation increases the number of erroneous rejections of H_0. Likewise, when negative correlation reduces the width of the measured t distribution to 57%

[**] Equation 18 assumes a known variance ("infinite" df) while the reference distribution used in this computational experiment was based on variance estimates with 18 degrees of freedom. In the former case the reference distribution is normal, while in the latter it is an 18-df t distribution. The t-distribution is somewhat broader than the normal so the influence of correlation is somewhat greater than equation 18 suggests, but for 18 df this difference is sufficiently small that we neglect it.

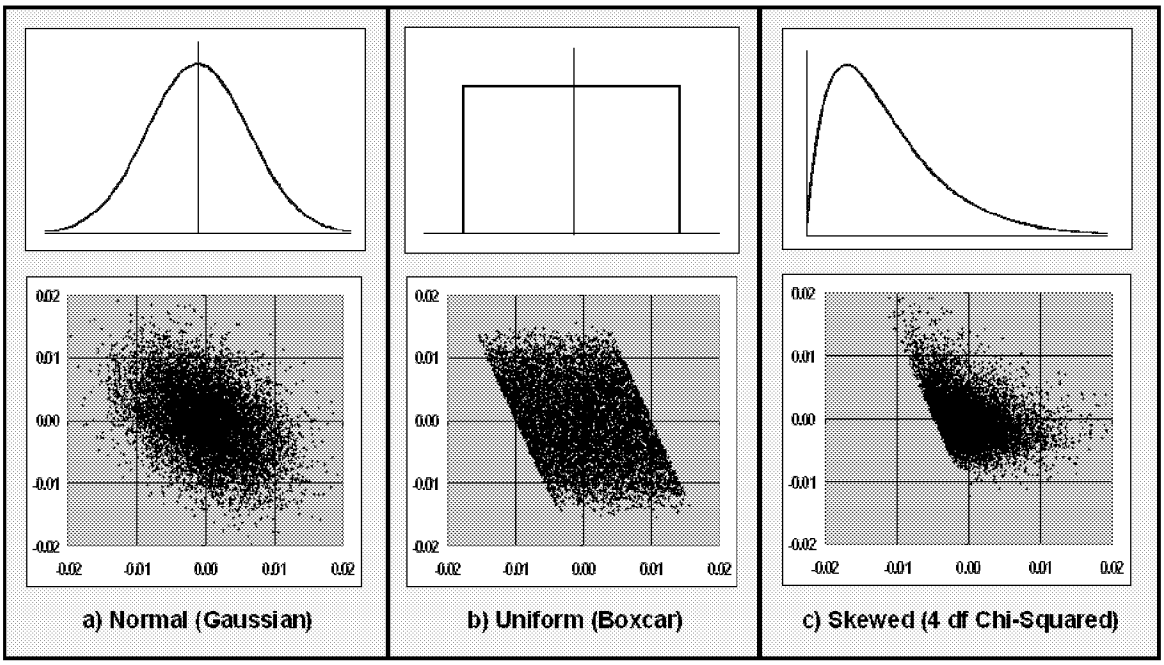

Figure 18. *Correlated* lift coefficient errors. Plots of e_i versus e_{i-1} for samples with lag-1 autocorrelation coefficient $\rho_1 = -0.4$, drawn from three population distributions.

of the non-correlated case, this reduces the area under that distribution outside a fixed reference interval centered on the mean, causing a reduction in the number of H_0 rejections due to random error. Figure 22 illustrates schematically how correlation-induced contractions and expansions in the width of measured t distributions can change the risk of inference error when the reference t-statistic assumes uncorrelated observations.

Negative correlation results in a kind of self-correction in the data, in which negative errors tend to be followed by positive (or less negative) errors, and

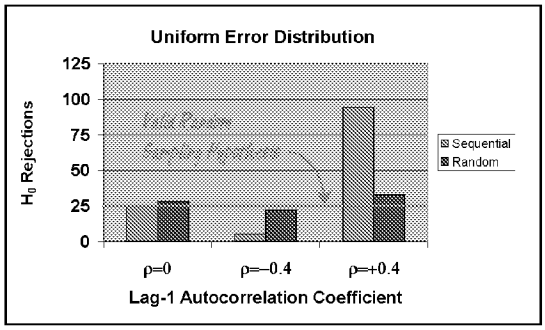

Figure 19. Impact of set-point order on inference error probability for different levels of correlation. Uniform error distribution.

positive errors tend to be followed by negative (or less positive) errors. The tendency for random errors to cancel is thus enhanced somewhat, which reduces the width of the measured t-distribution. This is why negative correlation tends to reduce the number of erroneous rejections of the null hypothesis. Unfortunately, this same process tends to produce "false alarms" at a higher rate, in which differences in sample means too small to be of concern are flagged as significant. For example, an experiment to compare the lift of two wing designs might result in an erroneous conclusion that there is some significant difference in performance when in fact neither wing is genuinely superior to the other. This could lead to unjustified production decisions or other undesirable consequences.

With positive correlation, positive errors tend to be followed by positive (or less negative) errors, and negative errors tend to be followed by negative (or less positive) errors. There is thus a reduced tendency for random errors to cancel that broadens the distribution of measured t-statistics, resulting in more erroneous rejections as we have seen. This tends to reduce the sensitivity of experiments, making it more difficult to detect subtle effects that may be important. For example, with positive correlation an experiment to compare the lift of two wing designs might result in an erroneous conclusion that there is no significant

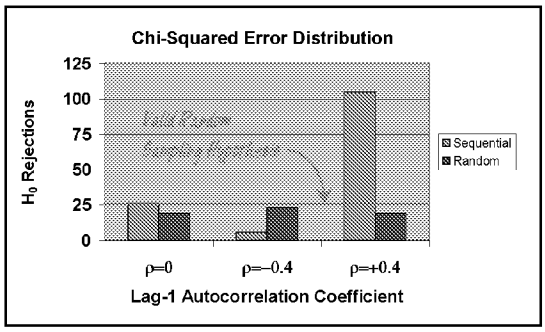

Figure 20. Impact of set-point order on inference error probability for different levels of correlation. Chi-Squared error distribution

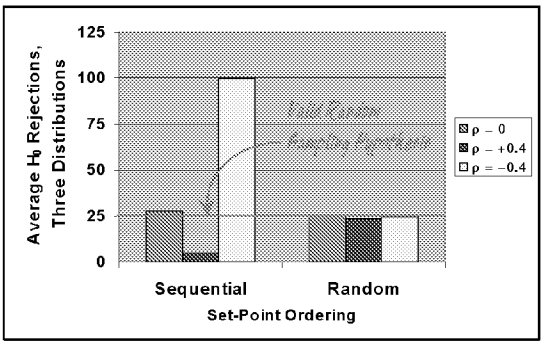

Figure 21. Impact of set-point order on inference error probability for different levels of correlation: Average number of erroneous H_0 rejections for errors drawn from normal, uniform, and skewed distributions.

difference in performance when in fact one wing actually is genuinely superior to the other. This could result in a lost opportunity to exploit the benefits of the superior wing.

In all the cases we examined, whether there was positive correlation, negative correlation, or no correlation at all, and whether the errors were drawn from a normal distribution, a uniform distribution, or a skewed distribution, randomizing the set-point order produced the general levels of inference error risk that one would anticipate if the random sampling hypothesis were valid. Only in the case of completely independent observations does a failure to randomize result in expected inference error risk levels. Positive correlation is more common in wind tunnel testing than negative correlation, meaning that Type I inference errors (erroneous rejection of the null hypothesis) are more common than Type II errors (erroneous rejection of the alternative hypothesis). In the context of the comparative wing performance example, correlation would make it more likely in a real experiment to miss a significant improvement in lift than to claim some improvement when none existed, simply because correlations are more likely to be positive than negative.

The results of the above computational experiments demonstrate that even mild correlation among observations in a sample of data can adversely impact the results of standard statistical tests that assume the random sampling hypothesis. They also suggest that reliable inference error risk predictions are not influenced so much by the distributional details of the population from which the errors are drawn as by the independence of the individual observations in the sample. In particular, if the observations are not independent, then larger samples (more data) that might produce a better approximation to a normal distribution of sample means because of the Central Limit Theorem, will do nothing to ensure reliably predictable inference error risk levels. Randomization has been shown to stabilize the inference error risk about predictable levels regardless of the population from which the errors are drawn, and regardless of correlation among observations. This is one more way that randomization defends against the adverse effects of systematic variation in a data sample, and another reason that randomization is a recommended standard operating procedure in experimental disciplines that focus upon inference (knowledge) rather than simple high-volume data collection.

Blocking: A Defense Against Between-Sample Systematic Variation

We saw in the last section that significant enhancements in quality can be achieved by permuting the order in which observations are recorded in an experiment. Specifically, we saw that randomizing the order that independent variable levels are set can reduce the unexplained variance in an experiment and also

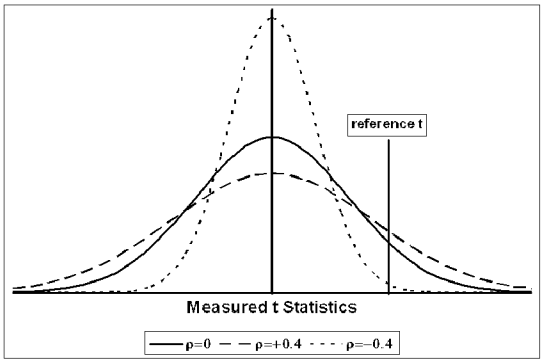

Figure 22. Impact of correlation on inference error risk.

Table IV: Test matrix to support a second-order C_L response surface experiment, independent variables in physical and coded units

SET POINT			CODED			CL	ELAPSED
BLOCK	ALPHA	BETA	BLOCK	ALPHA	BETA		TIME, Min
1	12	0	-1	0	0	0.5379	0.00
1	10	4	-1	-1	1	0.4503	1.10
1	12	0	-1	0	0	0.5374	1.36
1	14	4	-1	1	1	0.6337	2.00
1	12	0	-1	0	0	0.5399	3.16
1	14	-4	-1	1	-1	0.6293	3.40
1	12	0	-1	0	0	0.5390	4.96
1	10	-4	-1	-1	-1	0.4462	7.14
2	9.17	0	1	-1.414	0	0.4102	8.21
2	12	0	1	0	0	0.5377	8.89
2	12	-5.66	1	0	-1.414	0.5439	10.02
2	12	0	1	0	0	0.5396	11.16
2	12	5.66	1	0	1.414	0.5491	12.29
2	12	0	1	0	0	0.5393	13.82
2	14.83	0	1	1.414	0	0.6658	14.11
2	12	0	1	0	0	0.5407	15.28

minimize the probability of false alarms and missed effects. We will now examine another way to select the order of observations to further improve the quality of experiment results. We call this technique *blocking*.

A recent wind tunnel experiment at NASA Langley Research Center was designed to characterize the forces and moments on a generic winged body over a relatively narrow range of angles of attack and angles of sideslip. Table IV presents the test matrix in run order, with independent variables listed in both physical and coded units, per equation 13. Lift coefficients computed from measurements at each set-point are included in the table, as well as the elapsed time for each point relative to the start of the sample.

The set-point levels are not uniformly randomized in this design. Rather, they are clustered into two "blocks" of points, with points randomized within each block. It is this blocking scheme that we will examine in some detail in this section.

The design in Table IV is a very efficient design for fitting second-order response models called a Central Composite Design (CCD) or Box-Wilson design, after its developers[11]. In this experiment, the ranges of the independent variables were sufficiently restricted that response function terms of order three and higher were believed to be negligible, which is a good scenario in which to apply the CCD. Figure 23 is a general schematic representation of a two-variable CCD, in which the set points are plotted as coded units in what is called the inference space or design space of the experiment. This space is simply a Cartesian coordinate system in which each axis represents one of the independent variables, so that every point in the space corresponds to a unique combination of the independent variables. The eight points near the center are in fact collocated replicates at (0,0), drawn in the figure to show how many center points there are. The filled circles are points acquired in Block 1 and the stars are points acquired in Block 2. Note that half the center points are acquired in one block and half in the other. All points in Block 1 were run before any points in Block 2.

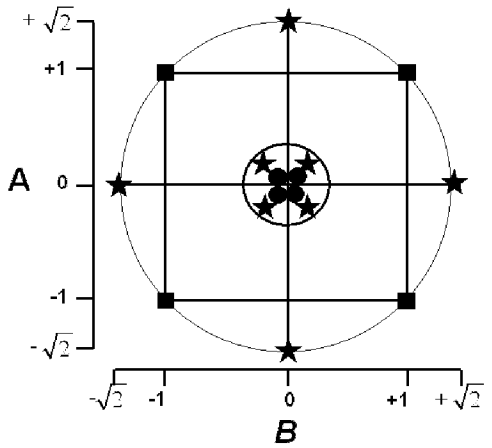

Figure 23. Orthogonally blocked Central Composite (Box-Wilson) Design in two variables with four center points per block.

Table V. Regression coefficients for full, unblocked, second-order C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H0 Coeff=0	Prob > |t|
Intercept	5.389E-01	1	5.48E-04		
A-AoA	9.099E-02	1	5.48E-04	165.97	< 0.0001
B-Sideslip	1.967E-03	1	5.48E-04	3.59	0.005
A^2	-1.051E-03	1	5.48E-04	-1.92	0.0842
B^2	3.188E-03	1	5.48E-04	5.81	0.0002
AB	6.350E-05	1	7.75E-04	0.082	0.9363

While this paper focuses on the quality aspects of formal experimental execution tactics, we note in passing that designs such as the CCD also enhance productivity. Only 16 points are required in this design to cover the whole range of both alpha and beta. An OFAT design would typically involve multiple alpha polars, each at a different beta set-point, with each individual polar featuring approximately as many points as the entire 16-point CCD design requires. This much additional data adds to the both the expense and the cycle time of a wind tunnel experiment, reducing productivity. To facilitate the acquisition of so many data points in as little time as possible, OFAT practitioners are forced to set the angles of attack sequentially to maximize data acquisition rate, guaranteeing by this ordering the greatest possible adverse impact of within-polar systematic variation on the alpha dependence. The sideslip angles are typically set in monotonically increasing order as well, likewise guaranteeing the greatest possible adverse impact of between-polar systematic variation on the beta dependence. Thus, OFAT methods often manage to minimize both productivity and quality simultaneously, an accomplishment all the more noteworthy for the substantial expense required to achieve it.

We will begin with an analysis of the data in Table IV that does not take blocking into account. We fit the response data to a full second order polynomial in the two coded variables as in equation 14, generating estimates for the coefficients of this model and also the uncertainties in estimating them. We use standard regression methods outlined in an earlier discussion of the impact of systematic variation on response surface estimates. Table V is a part of a computer-generated output from such a regression analysis. For each of the six terms or "factors" in the model, a numerical estimate is made of both the coefficient and the "one-sigma" uncertainty in estimating it (its "standard error"). The intercept factor in this table is the b_0 term in equation 14, and factors A and B in the table correspond to variables x_1 and x_2 in equation 14, which are the angles of attack and sideslip, respectively, in this model.

The column in Table V labeled "t for H_0 Coeff = 0", contains measured t-statistics referenced to a null hypothesis that the true value of the coefficient is zero. These are computed by dividing the coefficient estimate by the standard error. The t-statistics thus represent how far the coefficient is from zero in standard deviations. Large t-statistics imply that the estimated coefficients are large enough relative to the uncertainty in estimating them that they are unlikely to appear non-zero only due to experimental error, and are therefore probably real. The right-most column in

Table VI. Regression coefficients for reduced, unblocked C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H0 Coeff=0	Prob > |t|
Intercept	5.389E-01	1	5.48E-04		
A-AoA	9.099E-02	1	5.48E-04	165.97	< 0.0001
B-Sideslip	1.967E-03	1	5.48E-04	3.59	0.005
A^2	-1.051E-03	1	5.48E-04	-1.92	0.0842
B^2	3.188E-03	1	5.48E-04	5.81	0.0002
AB	6.350E-05	1	7.75E-04	0.082	0.9363

Table V contains "p-statistics" that represent the probability that a coefficient as large as the one estimated could occur entirely due to chance variations in the data if the true value of the coefficient is zero.

Coefficients with large t-statistics have small p-statistics. For example, the first-order AoA term in this model features a t-statistic of more than 165, indicating that this coefficient estimate is more than 165 standard deviations to the right of zero. Assuming random sampling, the probability that such a result could be due to ordinary chance variations *under the null hypothesis* is infinitesimal, or as the computer output coyly describes it, "< 0.0001". The miniscule probability that the linear AoA term is zero (or conversely, the substantial size of the t-statistic for this term in the model) confirms what subject matter specialists already know: that lift has a strong first-order dependence on angle of attack.

By contrast, note that the AB interaction term has a very small t-statistic. The size of the coefficient is much smaller than the standard error in estimating it (only about 8.2% of the standard error), and there is more than a 93% chance that such a small value could result from experimental error if the coefficient was actually zero. We are therefore unable to conclude from the data that alpha and beta interact over the ranges tested. That is, we cannot say over this range of variables that a given change in alpha will produce a different change in lift at one beta than another.

The quadratic AoA term also looks quite small. It is less than 2 standard deviations away from 0 so we are unable to distinguish it from zero with at least 95% confidence. We therefore drop this term from the model also, concluding that at least over the range of alpha examined, we are unable to detect curvature with sufficiently high confidence to retain a term for it. Table VI displays the regression coefficients for a *reduced* C_L response model. A reduced model features only the terms that we can infer are non-zero with sufficient confidence to satisfy our inference error risk tolerance. We declare this risk level in advance, and use it as a criterion for accepting or rejecting candidate model terms

The reduced model now features only four terms, but each one is highly likely to be non-zero. We therefore have some reason to believe that this model may adequately represent the data. The reduced model is:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{22} x_2^2 \qquad (19)$$

Before we accept equation 19 as an adequate representation of the data, numerous additional tests would typically be applied. A full discussion of all model adequacy tests that are normally applied in a response surface experiment such as this is well beyond the scope of this paper. We will examine one, however, called a *lack of fit test*, to highlight the role that blocking can play in improving the fit.

The lack of fit test begins by computing the total variance of the data sample in the usual way. The sum of squared deviations of each observation from the sample mean is divided by the minimum degrees of freedom required to compute the sum of squares – *n-1* for an *n*-point sample. The total variance is then partitioned into explained and unexplained components using analysis of variance (ANOVA) methods. We would like all of the variance to be explained by the model, but in reality there is always a component of unexplained variance that is responsible for the uncertainty that inevitably attaches to response predictions we make with the model.

To assess the quality of the model, we are interested in further examining the unexplained variance. The unexplained variance is non-zero because even a reasonably good model will not go precisely through each point in the data sample. There will generally be some residual for each point. However, a non-zero residual can be explained in two ways: It is possible that the model is correct and the residual is due simply to random variations in the data. It is also possible that the data point is correct and the model is simply wrong at that point. That is, the point

Table VII. ANOVA table for reduced, unblocked C_L response model.

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F
Model	6.63E-02	3	2.21E-02	8066.23	< 0.0001
Residual	3.29E-05	12	2.74E-06		
Lack of Fit	2.33E-05	5	4.65E-06	3.38	0.072
Pure Error	9.63E-06	7	1.38E-06		
Cor Total	6.64E-02	15			

may be in the wrong place or the response surface may be in the wrong place; it is difficult to say which is true by inspection alone. In practice, both explanations usually apply, but in different degrees. It is important to decide which of these factors is driving the unexplained variance because the choice of remedial action is different for one case than the other. If the unexplained variance is due primarily to random variations in the data, additional replicates can be acquired to average out the random variation. We say in such a case that the unexplained variance is due to *pure error*. If the unexplained variance is due primarily to an inadequate model, however, we would have to re-examine the model to see if additional terms or other changes might improve it. We say in such a case that the model suffers from *lack of fit*.

We determine whether we have a lack of fit problem or a pure error problem by further partitioning the unexplained component of the total variance into pure error and lack of fit components, again using ANOVA techniques that are beyond the scope of this paper but which are readily available in standard references on the subject[12-15]. The analysis of variance culminates in an ANOVA table describing the various components of the total variance. Table VII is a computer-generated ANOVA table for the reduced lift model described by equation 19 and Table VI.

The first column in the ANOVA table identifies various sources of variance. A sum of squares is computed for each component, as is the corresponding number of degrees of freedom. The "Mean Square" column is the ratio of the sum of squares and degrees of freedom ("DF") for each source of variance. These are the actual variance components, which we examine in the following way: The first row in the table, labeled "Model", describes the component of the total variance that can be explained by the candidate model. The second row corresponds to the total unexplained or residual variance – that portion due to changes that the researcher cannot attribute to any known source. We first examine the ratio of explained to unexplained variance, which is listed in the fifth column, labeled "F Value". The explained volume is over 8000 times larger than the unexplained variance, which gives us confidence that we are not simply fitting noise. That is, changes in the independent variables are forecasted by our model to produce changes that are substantially larger than experimental error. The model F-statistic is therefore a measure of signal to noise ratio. The p-statistic in the last column is the same as we encountered earlier. It represents the probability that an F-statistic his large could be due simply to chance. The fact that this is so low for the variance explained by the model suggests that we are very likely to have an adequate signal to noise ratio.

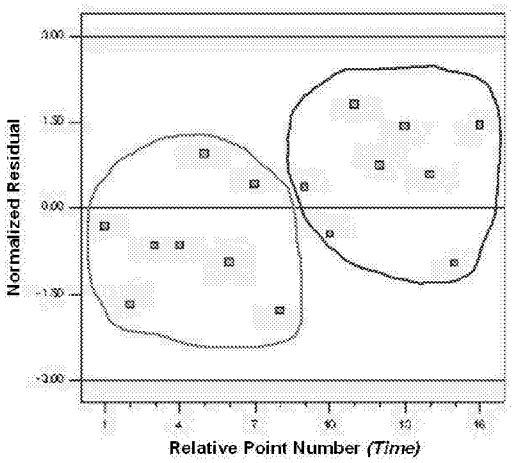

Figure 24. Residual time history of unblocked, reduced C_L response model, showing a slight between-block shift in sample means.

The residual or unexplained variance is further partitioned into pure error and lack of fit components by the ANOVA process, as indicated in Table VII. Again we construct an F-statistic by taking the ratio of two variance components — the lack of fit and pure error components of the residual variance in this case. We see from Table VII that this ratio is 3.38 for our model. This says that the variance attributable to lack of fit is over 3 times as large as the variance due to pure error, a troubling sign that the model is not an equivalent representation of the data, which is the objective of our modeling efforts.

Figure 24 provides a clue as to why the model may be suffering from lack of fit. This is a plot of normalized residuals in run order, which is a surrogate for time. The first eight points were acquired as a block and so were the second eight points. Recall that the set-points were randomized within blocks, which accounts for the fact that there is no particular pattern in the residuals within each block. However, while the mean of all of the residuals is zero (by definition of the mean), it seems as if the mean of the first block is slightly less than the mean of the second, suggesting that some kind of time-varying systematic error is afoot that is causing the block means to trend. (Note that had we not randomized the set-point order, the best-fit regression procedures would have incorporated this unexplained systematic error into the regression coefficients. We would have generated a "good fit" to an erroneous model.) In short, figure 24 suggests that our lack of fit may be due to "block effects" –

Table VIII. ANOVA table for full blocked C_L response model.

Factor	Coefficient Estimate	DF	Standard Error	t for H0 Coeff=0	Prob > |t|
Intercept	5.39E-01	1	4.47E-04		
Block 1	-0.0008				
Block 2	0.0008				
A-AoA	9.10E-02	1	4.47E-04	203.76	< 0.0001
B-Sideslip	1.97E-03	1	4.47E-04	4.40	0.002
A^2	-1.05E-03	1	4.47E-04	-2.35	0.043
B^2	3.19E-03	1	4.47E-04	7.14	< 0.0001
AB	6.35E-05	1	6.31E-04	0.10	0.922

systematic between-block shifts in sample means – which we will now remove.

To remove block effects, we augment the model in equation 14 by adding a *blocking variable*, making the response model a function now of three variables rather than two. The model we now fit to the data is an extension of equation 14, as follows:

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2 + cz \tag{20}$$

The blocking variable, z, is assigned a value of -1 for one of the blocks and +1 for the other. The assignment can be arbitrary, as long as it is consistent throughout the analysis. Note that the coefficient of the blocking variable represents an increment to the intercept term, b_0, quantifying how the mean level changes from block to block. That is, since z takes on only two discrete values, ±1, equation 20 reduces to:

$$y = (b_0 \pm c) + b_1 x_1 + b_2 x_2 + b_{12} x_1 x_2 + b_{11} x_1^2 + b_{22} x_2^2 \tag{21}$$

We can therefore generate in effect *two* response functions, one applying to each of the blocks. The functions are identical except for the intercept term, which is adjusted to reflect the different mean levels in each block. Table VIII is the computer-generated output of a regression analysis in which the model in equation 20 that includes the additional blocking variable was fit to the data of Table IV. (Note in Table IV that the possible need of a blocking variable was anticipated in the design of the experiment.)

Table VIII is similar to Table V, which presents the results of the regression analysis for the unblocked case, but there are both obvious and subtle differences. The obvious difference is that Table VIII has coefficients for the two blocks, which are equal in magnitude and opposite in sign. These are the values that the cz term in equation 20 assumes in each of the blocks. They represent how much the response function must be shifted in each block from a value of b_0 that would split the difference between the two blocks. In this case, the first block is about 8 counts below the grand mean of all the data, and the second block is about 8 counts above it. The block effect is defined as the difference between these two levels, which is about 16 counts. This is not very large in absolute terms, but it is large enough to completely consume a 10-count error budget, which is commonly

Table IX. Regression coefficients for reduced, blocked C_L response model

Factor	Coefficient Estimate	DF	Standard Error	t for H0 Coeff=0	Prob > |t|
Intercept	5.39E-01	1	4.24E-04		
Block 1	-7.78E-04	1			
Block 2	7.78E-04				
A-AoA	9.10E-02	1	4.24E-04	214.66	< 0.0001
B-Sideslip	1.97E-03	1	4.24E-04	4.64	0.001
A^2	-1.05E-03	1	4.24E-04	-2.48	0.033
B^2	3.19E-03	1	4.24E-04	7.52	< 0.0001

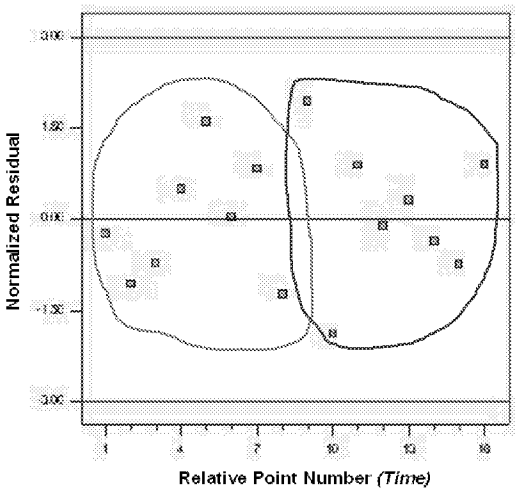

Figure 25. Residual time history of *blocked*, reduced C_L response model, showing no block effects.

specified in precision lift performance testing. Note also that the pure error variance in Table VII is 1.38×10^{-6}. The square root of this, 0.0012, is the pure error standard deviation. The fact that we have a 16-count block effect imposed upon chance variations with a 12-count standard deviation helps explain why the unblocked model failed to represent the data within pure experimental error.

A more subtle but crucial difference between the regression coefficients listed in Tables V and VIII is that while the numerical values of the coefficients are identical, the standard errors are much smaller for the blocked case than the unblocked case. The t-statistics are larger and the corresponding p-statistics are smaller. That is, blocking reveals the same regression coefficients, but permits them to be seen with greater precision. One result in this specific case is that the quadratic AoA term that could not be resolved before blocking is now comfortably more than two of the new, reduced standard deviations away from zero, permitting us to assert with at least 95% confidence that curvature in the AoA variable is real. Therefore, after the keener insight afforded by blocking, we can confidently retain the quadratic alpha term in the model, and infer that there actually is curvature in alpha over the range of angles of attack that we examined. We continue to omit the interaction term from the model, however.

Table IX is a computer-generated regression analysis table for the blocked, reduced model (interaction term dropped). Comparing with the unblocked reduced model of Table VI reveals that blocking has produced a reduction of more than 27% in the standard errors of the coefficients.

Figure 25 displays the residual time history of the reduced, blocked model. Note that as before, randomization has ensured that there are no within-block trends in the residuals, suggesting that the within-block errors are independent. Blocking has now removed the systematic difference between the two blocks that was apparent in figure 24.

Table X is the ANOVA table for the blocked, reduced model. Compare this with the ANOVA table for the corresponding unblocked case in Table VII. Recall that is was the rather significant lack of fit in Table VII that prompted a further analysis, which led to the discovery of a block effect and motivated the blocking analysis that removed it. The comparison of Table VII with Table X reveals that the lack of fit F-statistic that was 3.38 before blocking is now only 1.05. That is, before blocking the lack of fit component of the unexplained variance was over three times as large as the pure error component, but after blocking they are comparable. Note that we can never expect to produce a model with a lack of fit component that is significantly *smaller* than the pure error component, simply because the fit is limited by the quality of the data. (The model cannot be made better than the data that produced it without fitting the noise.) We can only aspire to generate models that do not significantly increase the unexplained variance beyond the pure error component, which appears to be the situation in this case only after blocking the data.

A comparison of Tables VII and X reveals that the model F-statistic is much larger for the reduced, blocked model than for the reduced, unblocked model

Table X. ANOVA table for reduced, blocked C_L response model.

Source	Sum of Squares	DF	Mean Square	F Value	Prob > F
Block	9.68E-06	1	9.68E-06		
Model	6.63E-02	4	1.66E-02	11541.20	< 0.0001
Residual	1.44E-05	10	1.44E-06		
Lack of Fit	5.91E-06	4	1.48E-06	1.05	0.456
Pure Error	8.47E-06	6	1.41E-06		
Cor Total	6.64E-02	15			

(11541 versus 8066). This suggests that there has been a significant increase in the explained variance attributable to blocking. Blocking does increase the explained variance, in that components of formerly unexplained variance can now be explained as block effects. However, there is a complicating factor in this instance. Blocking permitted us to include the quadratic alpha term in the model because by converting so much unexplained variance to variance that could be explained by the block effect, the residual unexplained variance was sufficiently reduced that the quadratic alpha term could be clearly resolved where it could not be resolved before blocking. So part of increase in explained variance is due to the blocking directly, but part of it is also because we converted some of the unexplained variance to explained variance when we added the extra term in the model that "explains" curvature effects in alpha.

To make a fair assessment of how blocking impacts the explained variance, a blocked model was analyzed in which the quadratic alpha term was dropped. This made it identical to the unblocked case with the single exception of blocking. Dropping the A^2 term reduced the model F-statistic from 11541 to 10480, reflecting the loss of the quadratic alpha's contribution to the explained variance. The relevant comparison, however, is between 10480 for the case of blocking and 8066 for the model that is identical in every respect *except* blocking. Note that the F statistic is simply the ratio of the variance explained by the model to the residual variance. Since now, except for blocking, the two models are identical (they contain the same independent variable terms), the ratio of F-statistics is the ratio of equivalent residual variances. In this case, that ratio is 8066/10480 = 0.77. The residual variance goes as *1/n*, so to achieve an equivalent increase in precision by conventional replication alone would require an increase in data volume (and associated cycle time) of a factor of 1/0.77 = 1.30. Blocking the data, which requires essentially no additional resources beyond the workload required to plan for it, has achieved in this case an increase in precision that would have required 30% more resources by conventional means.

We can also compare the blocked and unblocked cases on the basis of uncertainty in the model predictions. The uncertainty associated with predictions made by any linear regression model depends on the combination of independent variables for which the prediction is made, but the average variance across all points in the design space used to generate the model has been shown (e.g., by Box and Draper[12]) to be independent of the details of the model and equal simply to $p\sigma^2/n$, where p is the number of parameters in the model, n is the number of points used to fit the data, and σ^2 is the variance in the response. The square root of this is the average standard error ("one-sigma" uncertainty) associated with the model predictions. If we use the residual mean square from the ANOVA table as an unbiased estimator of the response variance (justified under an assumption of the random sampling hypothesis that randomization assures), then the average standard prediction error for the unblocked model is $\sqrt{4\times\left(2.74\times10^{-6}\right)/16} = 0.00083$. The blocked model has 5 terms instead of 4, which tends to increase the mean standard error of the prediction (because each term in the model carries with it some uncertainty). However, the residual mean square is less, because a portion of the otherwise unexplained variance attributable to alpha curvature effects is now converted to a component of explained variance, and also blocking has explained additional components of variance that were formerly unexplained. The average standard prediction error for the blocked model is $\sqrt{5\times\left(1.44\times10^{-6}\right)/16} = 0.00067$. Blocking has reduced the uncertainty in predictions from 8.3 counts to 6.7 counts, a 19% increase in precision that is obtained essentially for free.

The standard error in prediction even before blocking was relatively small in this case and it might be argued that efforts to further improve precision by blocking are unnecessary. First, it should be noted that there is no way to forecast in advance how large the block effects will be, so that blocking is a prudent precaution against systematic errors in any case. In this case the block effect was merely due to the influence of some unknown source or sources of systematic variation persisting for no more than about 15 minutes. (See elapsed time values in Table IV.) The block effects could easily be much greater if the blocks were separated further in time, especially if there was some identifiable change from block to block. For example, the block effects might have been greater if a facility shut-down and start-up had occurred between blocks such as occurs overnight, or if the two blocks were acquired by different shifts in a multi-shift tunnel operation.

Secondly, whether block effects can be considered negligible or not depends on the precision requirements of the experiment. In this case, an unblocked prediction standard error of 0.00083 has a "two-sigma" value of 0.0016 to two significant figures. This is ample precision to satisfy the requirements of many stability and control studies, for example, where precision requirements no more stringent than 0.005 in lift coefficient are common. (Even so, it is the sum of block effects *plus all other error sources* that must be maintained below 0.005, so while block effects alone may not be important in such a case, they could

contribute to "the straw that breaks the camel's back".) In performance testing where precision requirements are often 0.001, minimizing block effects in this example would be much more important.

Again we emphasize that systematic variations persisting over periods as short as fifteen minutes were sufficient in this case, drawn from an actual wind tunnel test that was not atypical, to introduce errors large enough to consume much of the error budget in precision wind tunnel performance testing. This is why precautions such as randomization and blocking are so important in such applications, and why experimental results obtained without these quality assurance tactics are so often difficult to reproduce within the desired precision from test to test within a given tunnel, and certainly across tunnels.

The reader may protest that we understated the standard error in predictions for the blocked model in the numerical example we considered above, because we failed to count the blocking variable as one of the parameters in the model. We dropped this term and its associated variance component from the model altogether because we are not interested in predicting the lift coefficient for one specific block of time or the other. Rather, we are interested in an overall estimate of the lift coefficient. We therefore use b_0 as the intercept term in equation 21 rather than either b_0+c or b_0-c. The rationale for this is that we have no reason to assume that one block is more representative of the long-term mean state of the tunnel than the other, and the average of the two is more likely to be a better approximation than either extreme.

We noted earlier that the regression coefficients for the blocked and unblocked models were identical, and the only difference caused by blocking was to improve the precision in estimating the coefficients. The importance of this result is easy to overlook on first read, and deserves to be highlighted. It means that even in the presence of block effects (and independent of how large those effects are, as it turns out), it is possible to recover the precise model we would have obtained if there had been absolutely no block effects in the data whatsoever! Not only are the model predictions the same but the actual coefficients are as well, meaning that no matter how large the block effects are, they will have no influence at all on our ability to predict responses, nor on the insights we can achieve into the underlying physics of the process. This is quite remarkable, and of enormous practical significance given the ubiquitous nature of block effects in real experimental situations with stringent precision requirements, as is common in performance wind tunnel testing. There is a great potential for exploiting blocking to minimize test-to-test and tunnel-to-tunnel variation that is yet to be tapped by the experimental aeronautics community.

To achieve these results requires that the blocking be performed in a special way that makes the blocking variable *orthogonal* to all other terms in the model. This is because changes to orthogonal terms in a model have no impact on the coefficients of other terms to which they are orthogonal. In particular, setting the coefficient of an orthogonal blocking variable to zero (dropping it from the model) has no affect on the rest of the terms in the model.

Orthogonal blocking in a second-order design such as this one (an experiment designed to produce a response model with no more than second order terms) requires that two rather mild conditions be met. The first is that the points within each block be themselves orthogonal. This is achieved when the products of all independent variables for each data point sum to zero. Consulting the two columns of coded independent variables in Table IV, it is clear that this condition is met in both blocks for the Central Composite Design. The second condition is that within each block, the sum of squared distances of each point from the center of the design space must be such that the ratio of these quantities from block to block is the same as the ratio of the number of points in each block. This condition is met in a Central Composite Design by adjusting the number of points in the center of the design and the distance that each "star" point is from the center of the design space in the second block. For a two-variable CCD, assigning the same number of center points to each block and setting the star points a distance from the design center equal to the square root of two is one way to ensure that the blocks identified in Table IV are orthogonal. Geometrically, this places all points either at the center of the design space or on a circle with its origin at the center of the design space. See figure 23.

The reader may ask why blocking is necessary when randomization has already been represented as an effective defense against systematic variation. Why did we not simply randomize the 16 points in Table IV, rather than dividing them into two blocks and randomizing within blocks?

One reason is that organizing the experiment as a series of small, orthogonal blocks makes it convenient to halt testing on block boundaries whenever it is necessary to do so, secure in the knowledge than any bias in response measurements that may materialize across blocks can be eliminated. We therefore break for lunch on a block boundary, schedule any tunnel entries to occur on block boundaries, end daily operations on a block boundary, and change shifts on block boundaries. All within-test subsystem calibrations are scheduled on block boundaries, including periodic calibrations of the data system, all wind-off zeros, all model inversions, and so on. Also, if some unforeseen event causes an unscheduled suspension of tunnel operations, we resume operations

not at the last data point in the test matrix, but at the last block boundary.

Another reason for blocking in addition to randomizing is that while randomization ensures the independence of observations necessary to make reliable inferences and converts hidden systematic errors to visible random errors that can be readily minimized by replication, blocking actually *eliminates* elements of the unexplained variance. That is, randomization ensures the proper shape of representations of cause-effect relationships between system response changes and changes in the independent variables (e.g., polars), while blocking enhances the precision with which such relationships are represented.

Finally, orthogonal blocking gives us a tool to reveal the degree of systematic variation ongoing in our experiment. This also enables us to quantify an important component of bias errors that is generally overlooked in wind tunnel testing. Because block effects can be completely eliminated from our characterizations of system response when the blocks are orthogonal, there is no reason to fear them. This liberates us to design our experiments to exploit block effects by using them as "tracers" to quantify systematic variation. Each block effect comprises an additional degree of freedom that can be used to assess the between-group variance that causes a (usually ignored) component of bias error in a typical wind tunnel test. This is the bias error due to relatively long period variations in sample means that are caused by the kinds of persisting systematic effects we have discussed (temperature effects, instrumentation drift, flow angularity changes, etc).

The strategy for quantifying systematic variation with block effects is to randomize the order in which blocks are executed, in addition to randomizing the order that points are set within a block. For example, the two blocks in Table IV were both acquired at one Mach number, and typically there would be a similar pair of blocks for each of a number of other Mach numbers. It is only moderately more trouble to execute a single block before changing to another Mach number than to execute both blocks at once, so the experiment could be organized as a series of, say, 10 blocks (corresponding to five Mach numbers), with the order that each block is executed determined at random.

Not only would this reduce the impact of systematic variation on the quantification of Mach effects, it would also ensure a relatively broad spectrum of time intervals between blocks, enabling us to quantify block effects over shorter periods and longer periods as well. This could provide valuable insights into the nature of these systematic effects and the performance of the facility, perhaps indicating ways to reduce the systematic variation. The information could also provide a more quantitative basis for deciding how often to perform such tasks as data system calibrations and wind-off zeros. In any case, this strategy would enable us to sample block effects over a relatively wide range of conditions, providing a reasonable estimate of the contribution that systematic variation makes to within-test bias errors. These are the effects that result in uncertainty in the absolute level of the intercept term (e.g., b_0 in equation 14) that serves as the reference level about which our response models predict changes due to changes in the independent variables.

Results and Discussion

Most of the results of this paper have been discussed as they were developed, but a few important points are reemphasized here.

There is an emerging consensus within the experimental aeronautics community that the objective of wind tunnel testing is not simply to acquire data in high volume, but to make specific, reliable inferences about the system under study. That is, there is a growing realization that wind tunnel tests are conducted to learn new things, not simply to "get data".

The importance of increasing knowledge through wind tunnel testing has never been disputed, of course. Rather, it has been taken for granted that efforts to maximize data collection rates are necessary to facilitate the greatest number of reliable subsequent inferences that can be drawn from the data. This has resulted in a focus on speed, and the emergence of various rate-related productivity metrics in ground testing such as "polars per hour", "data points per test", etc. The attitude during the test execution phase is often that the most effective way to facilitate future analyses is to acquire as much data as possible while the means to do so are available.

Unfortunately, there is a tradeoff between speed and quality that imparts hidden costs to this high-speed data collection strategy. We do not refer to the conventional "haste makes waste" argument that continuous efforts to hurry a process can generate careless errors, although that, too, is a consideration. Rather, we mean that the standard practice of setting monotonically increasing, sequential levels of the independent variables to maximize data acquisition rate incurs a quality penalty when subtle, persistent systematic variations are in play. Systematic (non-random) variations invalidate the random sampling hypothesis for data samples acquired this way. The random sampling hypothesis is a prerequisite for making reliable scientific inferences, which is a different activity altogether from simple high-volume data collection.

Blocking and randomization are quality assurance tactics that can augment the traditional quality

enhancement procedure of replication when systematic variations are in play, by ensuring the validity of the random sampling hypothesis. Unfortunately, the costs of randomization and blocking are often more apparent during execution than are their benefits. We seldom see the subtle systematic variations from which good experimental technique defends us. The temptation is therefore always there to abandon good technique in the name of expediency or convenience. Those who resist such temptations are more likely to be rewarded with reproducible results than those who succumb. Researchers who recognize and overtly defend against systematic variation also enjoy the intangible peace of mind that comes from knowing that they can control the quality of the experiment through its design, without depending exclusively on the state of the facility to ensure a quality result. They know that while Nature is probably visiting one unknown systematic variation or another on the experiment at any given time, the design of the experiment – like good anti-virus software – is working in the background to protect them.

Concluding Remarks

This paper has considered the role of independent observations in experimental data. It has shown that results obtained when observations are not independent can be unreliable. It has also cited experimental evidence demonstrating that observations are often not independent in wind tunnel testing. Certain effective tactics have been described that were developed in other research fields and are available in experimental aeronautics to defend against the adverse effects of conducting experimental research in environments for which observations may not be independent. The specific conclusions of this report are as follows:

1) When an effect that persists over time biases an observation in one direction, the errors in subsequent observations are more likely to be in the same direction than the opposite direction. Under these conditions, observations replicated over a shorter time interval are more alike that observations replicated over longer time periods.

2) Temperature effects, subtle changes in flow angularity and wall effects, and drift in instrumentation and data systems are all effects that persistent over time in wind tunnel testing.

3) Wind tunnel practitioners implicitly acknowledge the existence of persisting error sources through standard operating procedures that include frequent wind-off zeros, model inversions, and data system calibrations.

4) A large body of experimental evidence is in hand to demonstrate that replicates acquired in wind tunnels over shorter time periods are more alike than replicates acquired over longer time periods, and that the uncertainty introduced into experimental results by ordinary random variations in the data are small compared to the uncertainty caused by systematic variations that persist over time.

5) The results of at least one experiment designed to quantify such effects suggests that systematic variation can occur in 15% to 35% of the polars acquired in a representative wind tunnel test.

6) When systematic variations are in play, sample means and sample variances are not unbiased estimators of population means and variances.

7) Systematic variations are generally more difficult to detect than random variations unless a special effort is made to do so. For example, regression and other "best fit" methods tend to absorb systematic errors into the estimates of model coefficients, generating results that display only the random error component of the total unexplained variance, leaving the systematic component undiscovered.

8) The bias in sample statistics caused by persisting systematic variations in wind tunnel test environments is a function of factors that do not persist indefinitely. This can result in experimental data that might not be subsequently reproduced with the precision demanded of modern wind tunnel testing.

9) Bias in sample variance caused by unrecognized systematic variation has the effect of increasing the risk of inference error by generating a different set of circumstances than is assumed when null hypotheses and corresponding reference distributions are developed for formal hypothesis testing under the assumption that all observations are independent.

10) Systematic variations can rotate or disfigure wind tunnel polars. They can also be responsible for fine structure within a polar that is unrelated to independent variable effects.

11) Systematic variation over extended periods can result in block effects that cause polars acquired in one block of time to be significantly displaced from polars acquired in a later block of time.

12) Systematic variations sufficient to consume significant fractions (and often significant multiples) of the entire error budget can occur over relatively short periods of time in wind tunnel tests. It is not uncommon for such variations to occur over periods that are not long compared to the time to acquire a typical polar, for example.

13) It is possible to impose independence on experimental data even in the presence of systematic variations that persist over time, by setting the independent variable levels in random order. This technique, widely used in other fields besides experimental aeronautics for most of the 20th century, decouples independent variable effects from the effects of changes occurring systematically over time.

14) Randomization ensures that sample statistics are unbiased estimators of their corresponding population parameters.

15) Randomization ensures that inference error risk can be reliably assessed during the design of an experiment, by ensuring that reference distributions of selected test statistics faithfully represent conditions that exist when proposed null hypotheses are true. These considerations are expected to assume greater relevance as the focus of wind tunnel testing shifts more toward scientific inference, with less reliance upon simple high volume data collection.

16) Randomization has been shown to stabilize the inference error risk about predictable levels regardless of the population from which the errors are drawn, and regardless of correlation among observations.

17) There is an inherent conflict between speed and quality in wind tunnel testing, in that test matrices designed to maximize data acquisition rate seldom coincide with those that are designed to maximize the quality of experimental results.

18) The ultimate intent of high-volume data collection is to minimize the risk of finishing an experiment with insufficient information to adequately characterize the system under study. Fortunately, it is possible to drive inference error probabilities well below typically accepted levels by acquiring significantly fewer data points than common high-volume data collection strategies generate. This implies that it is generally possible to design test matrices that maximize research quality while still acquiring ample data to drive the probability of inference errors acceptably low.

19) Blocking an overall wind tunnel test matrix into particular clusters of independent variable settings can facilitate a subsequent analysis of the data that enables substantial portions of the unexplained variance attributable to unknown systematic variation ("block effects") to be eliminated from the experimental results.

20) Blocking is commonly used in other experimental research fields to enhance precision to a degree that would otherwise require significantly more data to achieve through conventional replication. Blocking therefore has the potential to ameliorate the adverse effects on data volume of designing test matrices to maximize research quality rather than data collection rate.

21) It is possible to use block effects estimates as "tracers", to characterize the overall degree of systematic variation in a wind tunnel test. This information can help facility personnel identify possible sources of systematic variation and can also quantitatively inform decisions about how often it is necessary to impose such common defenses against systematic variation as wind off zeros, model inversions, and subsystem calibrations.

22) Randomization and blocking are tactical defenses against systematic variation that have the same potential for guaranteeing quality enhancements in experimental aeronautics as they have been providing in other experimental research disciplines since their introduction by Ronald Fisher over 80 years ago.

23) Tactical defenses against systematic variation would be useful in any case, but will become increasingly important as an evolving consensus emerges about the objective of experimental aeronautics. That consensus is that experimental aeronautics is conducted to acquire knowledge and insight, and not simply to "get data" in as great a volume as resources limitations permit.

24) The approach to wind tunnel testing changes dramatically when one recognizes that the objective is to maximize the volume and quality of scientific inferences, rather than the volume and quality of individual data points. Tactics such as sequential settings of independent variables, which can achieve high data volume at the expense of reliable insight, are expected to fall

increasingly into disfavor by experimentalists sophisticated enough to fully appreciate both the importance of independent observations in scientific research, and the limited data volume often necessary to ensure acceptably small inference error risk.

Appendix

Impact of Statistical Dependence on the Utility of Sample Statistics as Reliable Estimators of Population Parameters

We depend upon the estimates we make of such statistics as the mean and standard deviation of relatively small data samples for information about the larger populations that interest us, but which are simply too large to quantify directly, given realistic resource constraints. In this appendix, we examine the expectation values of the sample mean and sample variance under conditions for which the random sampling hypothesis does not hold. In such circumstances, some degree of correlation exists among individual observations in the sample and they cannot be said to be statistically independent. This can occur when the unexplained variance in a set of data contains a systematic component superimposed upon the ubiquitous random errors that are well known to characterize any real data set. Systematic effects are in play when conditions are such that observations made over a short interval are more like each other than they are like observations made at some later time. This can be due to thermal effects that persist over time, or drift in the instrumentation and data system, or any of a large number of other unknown and unknowable sources. Because this condition in which short-term variance is smaller than long-term variance is not rare in wind tunnel testing, it behooves us to examine more closely how it affects our use of sample statistics to estimate population parameters.

Sample Mean.

Consider first the sample mean, $\bar{y}$. If systematic variation is in play while the data sample is acquired, the i^{th} observation will consist of the sum of the population mean, μ, the usual random component of unexplained variance, e_i, plus a systematic component of unexplained variance, b_i. That is, $y_i = \mu + e_i + b_i$. Let E{x} represents the expectation value of x for any x.

Then

$$E\{\bar{y}\} = E\left\{\frac{\sum_{i=1}^{n}(y_i)}{n}\right\} = E\left\{\frac{\sum_{i=1}^{n}(\mu + e_i + b_i)}{n}\right\} \quad \text{(A-1)}$$

$$= \frac{1}{n}E\left\{\sum_{i=1}^{n}\mu + \sum_{i=1}^{n}e_i + \sum_{i=1}^{n}b_i\right\}$$

$$= \frac{1}{n}E\{n\mu\} + \frac{1}{n}E\left\{\sum_{i=1}^{n}e_i\right\} + \frac{1}{n}E\left\{\sum_{i=1}^{n}b_i\right\} \quad \text{(A-2)}$$

$$= \mu + \frac{1}{n}\left\{\sum_{i=1}^{n}E\{e_i\}\right\} + \frac{1}{n}\left\{\sum_{i=1}^{n}E\{b_i\}\right\} \quad \text{(A-3)}$$

The expectation value for the component of random error associated with the i^{th} observation in a sample, e_i, is 0 (first summation term in A-3). The expectation value for the component of systematic error associated with the i^{th} observation in a sample, b_i, we will call β (second summation term in A-3). The value of β will depend on the details of the systematic variation but it will not be zero in general. Electronic engineers will recognize β as a kind of "rectification error". It represents a component of unexplained variance that is not completely cancelled out by replication in the same way as random errors because it is systematic – more akin to a bias error than a random error. Therefore we have:

$$E\{\bar{y}\} = \mu + 0 + \frac{1}{n}\left\{\sum_{i=1}^{n}\beta\right\} = \mu + \frac{1}{n}(n\beta) \quad \text{(A-4)}$$

or

$$\boxed{E\{\bar{y}\} = \mu + \beta} \quad \text{(A-5)}$$

If by coincidence the systematic variation exhibits a time history during the sample interval that causes early contributions be exactly canceled by later ones, say, then it is possible for β to be zero even in the presence of systematic error. However, from equation A-5 we see that the expectation value of the sample mean can only be relied upon to be an unbiased estimator of the population mean, μ, when there is no systematic component of the unexplained variance.

Sample Variance.

Consider now the impact of systematic error on the expectation value of the sample variance, s^2, where we follow the common convention by using Arabic characters to refer to sample statistics and Greek characters to refer to population parameters. We begin with the mechanical formula for sample variance:

$$E\{s^2\} = E\left\{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}\right\} = \left(\frac{1}{n-1}\right)E\left\{\sum_{i=1}^{n}(y_i-\bar{y})^2\right\} = \left(\frac{1}{n-1}\right)E\{SS\} \qquad \text{(A-6)}$$

Here we use SS to denote the sum of squared deviations from the sample mean:

$$E\{SS\} = E\left\{\sum_{i=1}^{n}(y_i-\bar{y})^2\right\} = E\left\{\sum_{i=1}^{n}y_i^2 + \sum_{i=1}^{n}\bar{y}^2 - 2\bar{y}\sum_{i=1}^{n}y_i\right\} = E\left\{\sum_{i=1}^{n}y_i^2 + n\bar{y}^2 - 2n\bar{y}^2\right\} \qquad \text{(A-7)}$$

or

$$E\{SS\} = E\left\{\sum_{i=1}^{n}y_i^2\right\} - n\bar{y}^2 \qquad \text{(A-8)}$$

We will consider each term on the right of equation A-8 in turn. We begin with the summation term, noting as before that the i^{th} observation, y_i, consists of the population mean, μ, plus the i^{th} components of random and systematic error, e_i and b_i respectively:

$$E\left\{\sum_{i=1}^{n}y_i^2\right\} = E\left\{\sum_{i=1}^{n}(\mu+e_i+b_i)^2\right\} \qquad \text{(A-9)}$$

or

$$E\left\{\sum_{i=1}^{n}y_i^2\right\} = E\left\{\begin{array}{l}\sum_{i=1}^{n}\mu^2 + \sum_{i=1}^{n}e_i^2 \\ +\sum_{i=1}^{n}b_i^2 + 2\mu\sum_{i=1}^{n}e_i \\ +2\mu\sum_{i=1}^{n}b_i + 2\sum_{i=1}^{n}(e_ib_i)\end{array}\right\} \qquad \text{(A-10)}$$

The first term on the right of equation A-10 is just $n\mu^2$, since μ – the population mean – is a constant. The second term is a sum of squared random deviations. In the limit of large n, this is just $n\sigma^2$, from the definition of population variance as a sum of squared random deviations divided by n. So the second term is just n times the random component of the unexplained variance. Similarly, the third term is $n\sigma_\beta^2$, or n times the systematic component of the unexplained variance. The fourth term is zero, since all random variations must sum to zero by definition of the mean. In this case, the random variations that occur at a particular point in time are not generally distributed about the true population mean. Instead, they are distributed about a value that is displaced from the true mean by the value of the systematic error at that time. In other words, the systematic variation acts like a time-varying bias error.

The fifth term is also zero, again because the mean has associated with it a constraint that all residuals sum to zero. In this case, the sample mean is biased from the true mean by an amount that causes the sum of the systematic residuals to sum exactly to zero.

The last term on the right of equation A-10 features the sum of the products of two deviations, one random and one systematic. In the limit of large n, the covariance between two variables, z_1 and z_2 is defined as the mean of the product of $z_1 - \bar{z}_1$ and $z_2 - \bar{z}_2$, or:

$$Cov(z_1, z_2) = \frac{\sum_{i=1}^{n}[(z_{1i}-\bar{z}_1)(z_{2i}-\bar{z}_2)]}{n}$$

Therefore the last term on the right of A-10 is just n times the covariance between the random and systematic error components of the unexplained

variance. It is a measure of the degree, if any, to which the random and systematic errors influence each other.

It is customary to normalize the covariance by dividing it by the product of the associated standard deviations, generating a dimensionless correlation coefficient, ρ, that ranges between ±1. For our case:

$$\rho_{e,\beta} = \frac{Cov(e,\beta)}{\sigma\sigma_\beta}$$

where the unsubscripted σ represents the standard deviation of the random component of the unexplained variance and σ_β represents the standard deviation of the systematic component of the unexplained variance with respect to the population mean, μ. A positive correlation coefficient would indicate that an increase in systematic error tends to be accompanied by a corresponding increase in random error. Likewise, if an increase in systematic error tends to be accompanied by a decrease in random error, the correlation coefficient is negative. It is zero if the systematic and random components of the unexplained variance are independent of each other.

Therefore, for the last term on the right of equation A-10 we have:

$$E\left\{2\sum_{i=1}^{n}(e_i b_i)\right\} = 2n \times Cov(e,\beta) = 2n\rho_{e,\beta}\sigma\sigma_\beta \qquad \text{(A-11)}$$

We can combine all of this into a rewriting of equation A-10 as follows:

$$E\left\{\sum_{i=1}^{n} y_i^2\right\} = n\mu^2 + n\sigma^2 + n\sigma_\beta^2 + 0 + 0 + 2n\rho_{e,\beta}\sigma\sigma_\beta \qquad \text{(A-12)}$$

or

$$\boxed{E\left\{\sum_{i=1}^{n} y_i^2\right\} = n\left(\mu^2 + \sigma^2\right) + n\left(\sigma_\beta^2 + 2\rho_{e,\beta}\sigma\sigma_\beta\right)} \qquad \text{(A-13)}$$

This is the first term on the right of equation A-8. We will now consider the second term, $n\bar{y}^2$. From the definition of a mean, we have:

$$n\bar{y}^2 = \sum_{i=1}^{n} \bar{y}_i^2 \qquad \text{(A-14)}$$

Here, $\bar{y}_i$ is the i^{th} sample mean in a distribution of sample means, rather than the i^{th} individual point in a sample. We will let e_i' represent the deviation of the i^{th} point from the population mean, *assuming systematic errors are present*. That is, e_i' includes whatever effect systematic errors have on the distribution of sample means. Then:

$$n\bar{y}^2 = \sum_{i=1}^{n} \bar{y}_i^2 = \sum_{i=1}^{n} (\mu + e_i')^2 = \sum_{i=1}^{n} \mu^2 + 2\mu\sum_{i=1}^{n} e_i' + \sum_{i=1}^{n} e_i'^2 \qquad \text{(A-15)}$$

As in earlier derivations, the first term is $n\mu^2$ (μ^2 is a constant) and the second is 0 (from definition of the mean). The third term is a sum of squared deviations, which we recognize as the product of n and the corresponding variance, by definition of the variance as the sum of squared deviations divided by n. We will call this variance σ'^2, where the prime indicates that this is the variance in a distribution of sample means that has been affected in some way by the presence of systematic error in the samples, and not just random error. That is, this variance will be something like equation 7 in the main text, which described the special case of a lag-1 autocorrelation among the observations in a sample, except that in this case no special restrictions are placed on the nature of the correlation (i.e., it is not constrained to be simply first-order or lag-1). We will represent this variance in the presence of systematic error as follows:

$$\sigma'^2 = \frac{\sigma^2}{n}\left[1 + f(\rho)\right], \qquad \text{(A-16)}$$

where $f(\rho)$ is defined for this representation as a function of the correlation that exists among observations in a sample when systematic variation is present, and is such that $f(\rho)=0$ when $\rho=0$. In that case the variance in the distribution of sample means reverts back to the familiar form we derived in the main text in equation 4.

We have, then:

$$\begin{aligned} n\bar{y}^2 &= n\mu^2 + 0 + n\sigma'^2 \\ &= n\mu^2 + \sigma^2[1+f(\rho)] \end{aligned} \quad \text{(A-17)}$$

We now insert equations A-17 and A-13 into equation A-8:

$$\begin{aligned} E\{SS\} = n\left(\mu^2 + \sigma^2 + \sigma_\beta^2 + 2\rho_{e,\beta}\sigma\sigma_\beta\right) \\ -\left\{n\mu^2 + \sigma^2[1+f(\rho)]\right\} \end{aligned} \quad \text{(A-18)}$$

or, after gathering terms:

$$\begin{aligned} E\{SS\} = \sigma^2[n-1-f(\rho)] \\ + n\left(\sigma_\beta^2 + 2\rho_{e,\beta}\sigma\sigma_\beta\right) \end{aligned} \quad \text{(A-19)}$$

We insert equation A19 into equation A-6:

$$\boxed{\begin{aligned} E\{s^2\} &= \left(\frac{1}{n-1}\right)E\{SS\} \\ &= \left\{\frac{[n-1-f(\rho)]}{n-1}\right\}\sigma^2 \\ &\quad + \left(\frac{n}{n-1}\right)\left(\sigma_\beta^2 + 2\rho_{e,\beta}\sigma\sigma_\beta\right) \end{aligned}} \quad \text{(A-20)}$$

Equation A-20 represents the expectation value of the sample variance when observations within the sample are correlated due to the kinds of systematic variation that can occur when the random sampling hypothesis does not hold. This is a very ugly function of the systematic error, and certainly the condition upon which we depend when we use sample statistics to estimate population parameters; namely, that $E\{s^2\}=\sigma^2$, does not hold in this case. However, if the random sampling hypothesis is valid, then the second term on the right of A-20 vanishes because all the terms related to systematic error are then zero, and the portion of the first term on the right within braces goes to one because $f(\rho)$ also goes to zero. The result is:

$$E\{s^2\} = \sigma^2 \quad \text{(A-21)}$$

That is, the expectation value of the sample variance is in fact the population variance as we require, *but only in the absence of systematic variation within the sample.*

Acknowledgements

This work was supported by the NASA Langley Wind Tunnel Enterprise. Dr. Michael J. Hemsch and the Langley data quality assurance team are acknowledged for documenting consistent distinctions between within-group and between-group variance in wind tunnel testing over years of studying a wide range of tunnels. The present paper was motivated in part by this comprehensive documentation of the frequency of occurrence and magnitude of systematic variations in wind tunnel data. Dr. Mark E. Kammeyer of Boeing St. Louis is gratefully acknowledged for discussions highlighting the importance of bias error in wind tunnel data. The staff of the National Transonic Facility and the ViGYAN low speed tunnel provided invaluable assistance in the acquisition of data used in this report.

References

1) Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building.* New York: Wiley.

2) Cochran, W. G. and Cox, G. M. (1992). *Experimental Designs. 2nd ed.* Wiley Classics Library Edition. New York: Wiley.

3) Montgomery, D. C. (2001). *Design and Analysis of Experiments, 5th ed.* New York: Wiley.

4) Diamond, W. J. (1989). *Practical Experiment Designs for Engineers and Scientists, 2nd ed.* New York: Wiley.

5) Hemsch, M., et al. "Langley Wind Tunnel Data Quality Assurance – Check Standard Results (Invited)". AIAA 2000-2201. 21st AIAA Advanced Measurement and Ground Testing Technology Conference, Denver, CO. 19-22 June 2000.

6) Fisher, R. A. (1966). *The Design of Experiments, 8th ed.* Edinburgh: Oliver and Boyd.

7) Coleman, H. W. and Steele, W. G. (1989). *Experimentation and Uncertainty Analysis for Engineers.* New York: Wiley.

8) Bevington, P.R. and Robinson, D.K. (1992, 2nd Ed). *Data Reduction and Error Analysis for the Physical Sciences.* New York: McGraw-Hill.

9) DeLoach, R. "Tailoring Wind Tunnel Data Volume Requirements through the Formal Design of Experiments". AIAA 98-2884. 20th AIAA Advanced Measurement and Ground Testing Technology Conference, Albuquerque, NM. June 1998.

10) Hemsch, M.J. "Development and Status of Data Quality Assurance Program at NASA Langley Research Center – Toward National Standards". AIAA 96-2214. 19th AIAA Advanced Measurement and Ground Testing Technology Conference, June 1996.

11) Box, G. E. P., and K. B. Wilson (1951). On the experimental attainment of optimum conditions, *J. Roy. Stat. Soc., Ser. B,* 13, 1.

12) Box, G.E.P and Draper, N.R. *Empirical Model Building and Response Surfaces.* (1987). New York: John Wiley and Sons

13) Myers, R.H. and Montgomery, D.C. (1995). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments.* New York: John Wiley & Sons.

14) Draper, N.R. and Smith, H. (1998, 3rd ed): *Applied Regression Analysis.* New York: John Wiley and Sons.

15) Montgomery, D.C. and Peck, E.A. (1992, 2nd Ed). *Introduction to Linear Regression Analysis*. New York: John Wiley and Sons.