

Analysis of Variance in the Modern Design of Experiments

Richard DeLoach*

NASA Langley Research Center, Hampton, Virginia, 23681

This paper is a tutorial introduction to the analysis of variance (ANOVA), intended as a reference for aerospace researchers who are being introduced to the analytical methods of the Modern Design of Experiments (MDOE), or who may have other opportunities to apply this method. One-way and two-way fixed-effects ANOVA, as well as random effects ANOVA, are illustrated in practical terms that will be familiar to most practicing aerospace researchers.

I. Introduction

The Modern Design of Experiments (MDOE) is an integrated system of experiment design, execution, and analysis procedures based on industrial experiment design methods introduced at the beginning of the 20th century for various product and process improvement applications. MDOE is focused on the complexities and special requirements of aerospace ground testing. It was introduced to the experimental aeronautics community at NASA Langley Research Center in the 1990s as part of a renewed focus on quality and productivity improvement in wind tunnel testing, and has been used since then in numerous other applications as well. MDOE has been offered as a more productive, less costly, and higher quality alternative to conventional testing methods used widely in the aerospace industry, and known in the literature of experiment design as “One Factor At a Time” (OFAT) testing. The basic principles of MDOE are documented in the references¹⁻⁴, and some representative examples of its application in aerospace research are also provided⁵⁻¹⁶.

The MDOE method provides productivity advantages over conventional testing methods by generating more information per data point. This is accomplished by changing the levels (values) of multiple factors (independent variables) for each new data point, rather than changing only one factor at a time. OFAT testing methods, by definition, change the level of only one variable (or factor) at a time in going from one data point to the next. The corresponding changes in the levels of measured responses such as forces and moments in a wind tunnel test can then be unambiguously related to the change that was made in the only independent variable that was altered. By contrast, MDOE methods rely upon a kind of “multitasking” to minimize data volume, in which several independent variables are changed simultaneously for each new data point. It can be said that MDOE methods make each data point “work harder” in this way, and this greater average “work done per data point” comprises the added efficiency that translates into MDOE cost savings. Fewer data points are required when each point works harder, which reduces direct operating costs and also cycle time—the real cost driver in many practical research and technology development scenarios. This also facilitates certain quality improvements by allowing sufficient time to invoke such quality assurance tactics as replication, randomization, and blocking, as discussed and illustrated in the references cited earlier.

Those who are unfamiliar with experimental methods that change more than one factor at a time can be understandably anxious when multiple independent variables are changed simultaneously. After all, if angle of attack and angle of sideslip are both changed before the next data point in a wind tunnel test is acquired, how is it possible to know what portion of the resulting change in forces and moments is attributable to the change in angle of attack and what portion was due to the change in sideslip angle?

Such anxiety is justified when one thinks in terms of the impact on a single data point. However, if multiple data points are in hand (a *sample* of data rather than an individual data point), the effects of multiple factor changes can in fact be reliably partitioned. This highlights a key distinction between MDOE and OFAT testing methods. Much of the insight derived from a conventional test is achieved by simply *displaying* the data in various ways, including some very clever representations that bring to light a variety of interesting effects. While MDOE practitioners also utilize graphical displays, most of the insights achieved in an MDOE experiment are the result of some *analysis* that

* Senior Research Scientist, Aeronautical Systems Engineering Branch, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA, 23681, Associate Fellow.

has been applied to the data, rather than a simple display of it. The analysis extends beyond visual representations of responses as a function of various factor levels, and generally consists of a variety of mathematical operations that are performed on the data.

One common analytical method invoked in MDOE testing is the partitioning of response effects according to the factor levels that are simultaneously changed to produce them when an MDOE test matrix is executed. This paper is intended for the reader unfamiliar with such partitioning methods, both to reassure such a reader that it is possible to do so, and as an elementary tutorial on how to proceed. Experimentalists equipped with the tools to partition responses measured when multiple factors are changed per data point are able to profit by the substantial increases in productivity that derive from the speed with which one can traverse the design space when one is no longer limited to exploring it one factor at a time. This paper focuses specifically on one such partitioning tool, the analysis of variance (ANOVA), and is intended as a tutorial introduction to the analytical methods of the Modern Design of Experiments (MDOE).

ANOVA techniques, while somewhat tedious if performed by hand, are fundamentally straightforward, easy to understand, and directly applicable in many practical aerospace research applications. It is unfortunate that they are so often treated in standard references in an intimidating and largely unapproachable way. The author has for some time advocated the use of formal experiment design methods in aerospace testing, and one of the greatest obstacles to embracing this method is a lack of understanding of the analytical methods that it requires, which largely entails ANOVA methods in one form or another. ANOVA methods will be illustrated in this paper using ground test data acquired in circumstances familiar to most practicing aerospace researchers.

Section II is intended to motivate the remainder of the paper by describing certain productivity and quality weaknesses of conventional aerospace ground testing. In Section III, a simple one-way analysis of variance introduces the concept of variance components in a sample of experimental data, and the fact that some components can be explained while others cannot. Section IV illustrates two-way ANOVA methods with a typical ground-testing example of practical interest. Differences between one-way and two-way ANOVA methods are highlighted. In Section V, a random effects ANOVA is illustrated in the context of developing a multivariate response model, and the difference between fixed-effects and random-effects ANOVA is outlined. Section VI discusses some practical implications of partitioning the variance in a sample of data, with concluding remarks offered in Section VII.

II. Issues in Conventional Experimental Design

In the introduction it was noted that MDOE methods contrast with conventional experimental procedures rooted in what is known in the literature of experiment design as the One Factor At a Time (OFAT) method. The OFAT method is commonly applied in wind tunnel testing, in which all factors (independent variables) except one (often angle of attack, initially) are held at constant levels while the one factor chosen to be varied is changed systematically through a prescribed range of levels, usually at uniform intervals that may change from one sub-range of the variable to another. Various responses of interest such as forces, moments, and pressures, are measured at each new level of the factor that is being changed systematically. The level of one of the factors previously held constant through this process (angle of sideslip, for example) is then incremented, and the process is repeated until all scheduled combinations of both variables have been set. A third factor (Mach number, say) is then set to a new level and in this example all combinations of angles of attack and sideslip are set as before, and so on, until the wind tunnel has been physically set to every scheduled combination of every independent variable of interest.

The OFAT method suffers from multiple weaknesses that adversely impact its productivity. Ironically, OFAT countermeasures intended to overcome productivity issues also adversely impact the quality of the experimental results produced by this method.

A. OFAT Productivity Problems

Productivity problems with conventional testing methods are best illustrated with a representative example. Consider a relatively simple wind tunnel test featuring a moderate number of independent variables. Let us say that there are six such variables to be examined—perhaps angle of attack, Mach number, angle of sideslip, flap deflection, aileron deflection, and speed brake deflection. More ambitious wind tunnel tests than this are of course common, featuring more than six independent variables, but this will serve as an illustration.

Let us postulate some typical ranges and increments for these six factors. The angle of attack (alpha) might be set in one-degree increments from a low level of -10° to a high level of $+20^\circ$, for a total of 31 levels. It might be desirable to examine this angle of attack range for angles of sideslip (beta) from -10° to $+10^\circ$ in 2° increments, so 11 levels of sideslip angle and therefore $11 \times 31 = 341$ combinations of alpha and beta. A half dozen Mach number

levels would not be atypical, requiring a total of $6 \times 341 = 2046$ individual combinations of Mach number, sideslip angle, and angle of attack in this typical illustration.

Control surface deflections are often changed in finer increments at lower angles and coarser increments for larger angles. A typical pattern might be $0^\circ, \pm 5^\circ, \pm 10^\circ, \pm 15^\circ, \pm 30^\circ,$ and $\pm 45^\circ$. Hardware for each of the specified fixed angles of the test is typically fabricated so for the 11 deflections settings for flap and aileron in this example, there would be hardware available to support investigations of 121 unique settings of these control surfaces. Let us assume five discrete speed brake deflection angles, so $5 \times 121 = 605$ configurations of the aircraft would be possible, and 2046 unique combinations of Mach, alpha, and beta for each. That's a grand total of $605 \times 2046 = 1,237,830$ individual combinations of tunnel state, model attitude, and configuration variables that would have to be physically set to examine every factor combination one factor at a time. Not every combination would necessarily be of interest, although in the early stages of an investigation especially, it is often unclear where the boundaries are that separate regions of greater and lesser interest.

If we assume a representative average pitch-pause data acquisition rate of seven polars per hour, this test would require just under 10 hours to execute the 66 polars (six Mach numbers by 11 sideslip angles) planned for each configuration. If we add two hours for a configuration change, this test matrix would require 1.5 eight-hour shifts for each of the 605 configurations, or more than 900 shifts. The actual required time would have to be much longer to account for tunnel down time, maintenance, periodic wind-off zeros and instrument re-calibration, and any quality assurance runs added to the test matrix to provide replicates for directly assessing random error. Unfortunately, only 20 shifts are available in a typical tunnel entry lasting four weeks—40 if there are two shifts per day. There is obviously a massive disconnect between the wind tunnel resources one might reasonably expect to be available, and the resources that would be required to examine every factor combination in a test even as unremarkable in its scope and complexity as the relatively simple example cited here. In realistic wind tunnel tests involving more combinations of factor levels, and especially additional configuration variables, available resources permit the acquisition of an even smaller fraction of the total number of factor levels that would be required for a comprehensive OFAT examination of the variables. The practical reality is that resource constraints inevitably force the OFAT practitioner to leave a considerable amount of information on the table.

This illustrates a fundamental contradiction inherent in OFAT wind tunnel testing, which is that OFAT practitioners presumably know sufficiently *little* about some system under study to justify the high cost of testing (costs that can range over \$1,000,000 per test in some instances, and even higher when the cost of capital is included in the cycle-time expense accounting), and yet they know the test article sufficiently *well* to infer its important operating characteristics from an examination of the tiny fraction of unique factor combinations that time permits when such an examination is conducted one factor at a time. Furthermore, they know precisely *which* fraction comprises the key factor combinations to examine, and can with acceptable confidence ignore the rest of the information that could be acquired with less stringent resource constraints. OFAT practitioners are not generally inclined to dwell upon the fact that unanticipated and possibly vital insights might be gleaned from an examination of the vast number of factor combinations that must go unexamined because of the productivity limitations of OFAT testing.

B. OFAT Quality Problems

Data acquisition rate is a primary focus in OFAT testing because of the urgency to acquire as much data as possible in a given tunnel entry. This urgency is quite justifiable given the fundamental inefficiency of the OFAT method in examining complex, multivariate responses; there can never be enough time to set every factor combination of interest, as illustrated in the previous section. This need for speed adversely impacts the quality of an OFAT test in two important ways (besides the obvious fact that “haste makes waste”). Both reflect a competition between the desire to set as many factor combinations as possible, and the need to devote some of the available data acquisition time to quality assessment and quality assurance activities.

A well-designed experiment features some number of replicated measurements to quantify the irreducible random error (also called “pure error”) that is responsible for elements of the unexplained variance in any experiment. Ideally, replicates should be acquired at combinations of higher and lower levels of all of the factors under study. This provides a representative estimate of the pure error, as well as evidence to support or refute assumptions about the homogeneity of the pure error that may be necessary to justify certain types of analysis.

Unfortunately, replicates are often given a relatively low priority in OFAT testing, with decidedly finite data acquisition resources allocated first to filling the cells of what for practical purposes is an infinite multi-dimensional test matrix. Faced with the choice of stopping to cover familiar ground again by replicating data previously acquired, or plowing ahead with factor combinations as yet unset, the OFAT practitioner often finds it harder to resist the

temptation to expand the envelope of the test matrix than to break off the quest for new data long enough to quantify random error in what has already been acquired.

Often a generally unsatisfactory compromise is reached in which a relatively few replicates are acquired for conditions that are easy to set but not necessarily representative of the entire test. For example, the decision may be made to replicate certain runs acquired with control surfaces in a neutral position (no deflection), but not to acquire replicates at other combinations of control surface settings that would require additional time-consuming configuration changes. In some especially unfortunate cases, simple repeat points are acquired in lieu of genuine replicates. In such cases the operator acquires multiple data points without changing the levels of any of the independent variables so that set-point errors go undetected, and random error is generally understated.

Another way in which high-speed data acquisition in OFAT testing adversely impacts quality is that it forecloses options to ensure independence in a sequence of measurements. We say that a measurement error in one data point is independent of the error in an earlier point only if knowledge of the earlier error would provide no information about the current error. Such independence is, unfortunately, elusive in the real world of wind tunnel testing. The reason is that there are inevitable systematic (not random) changes wafting through the measurement environment that persist for extended periods of time. If such changes cause the current lift measurement to be too high, say, then it is more likely that the next measurement will likewise be too high than too low. The two measurements are not independent in such circumstances.

Persisting, non-random changes may be due to instrument drift, or flow angularity changes caused by the tendency for dirt to accumulate more on the lower half of flow alignment screens than the upper half. Frictional heating can flow expansion, which can in turn cause test section walls to expand slightly, resulting in a slowly varying error in the wall interference correction that persists in one direction for an extended period of time. Because the Young's modulus of the sting is temperature dependent, it can become slightly more compliant over time, resulting in slightly more sting bending and a systematic change in true angle of attack. Leading-edge trip dots wear over time and require periodic redressing. Facility personnel experience both fatigue effects and learning effects that can cause their performance to change systematically over time. These are but a few of an uncountable number of possible systematic effects, most of which are likely to be unanticipated or accounted for in a given test. Furthermore, unlike random errors that reveal their presence every time a data point is replicated, these systematic errors can go undetected in replicates made over relatively short intervals of time, or they can be erroneously attributed to random error when replicates are acquired over longer time intervals.

The impact of any one systematic effect may not be large in an absolute sense, but such effects do not have to be very large to completely overcome the fractional drag count error budgets that characterize modern wind tunnel testing. (A half-drag-count error budget requires that unexplained variance not exceed a few parts per million of the total variance in a typical wind tunnel test). In modern wind tunnel facilities precision is often so good that true random error is quite small compared to the systematic component of the unexplained variance in a given sample of data.

The OFAT speed imperative demands that successive levels of certain factors (angle of attack, say) be set sequentially, since this is fastest way to traverse a given range of factor levels. Sequential factor level changes are also often prescribed simply for operator convenience, or because it is the most orderly way specify a test matrix. Why not acquire all the Mach 0.70 polars first, then the 0.75 polars, then 0.80, etc?

There would be no problem with the strategy of changing factor levels sequentially with time if there was no systematic unexplained variance in the data. Unfortunately, whenever such effects are in play they become confounded with any factor effects that are estimated with sequential factor settings. For example, if the Mach number is increased systematically with time while the magnitude of the experimental error in lift is also changing systematically, it is not possible to know how much of the resulting change in lift is due to the change in Mach number and how much is due to whatever is causing the systematic error. Furthermore, while it is not generally obvious when such systematic effects are in play, they can have a significant adverse impact on the reproducibility of wind tunnel data. This is because the dominant sources of systematic error are likely to be different from test to test, either in their fundamental nature, or in the magnitude and direction that they change with time, or in the rate at which they change.

Consider the systematic change in total temperature that is due to frictional heating when a large mass of air traverses the tunnel circuit for hours on end, often at very high speeds. Assume for a moment that the angle of attack package is neither insulated nor temperature compensated and the changing temperature causes the bias and sensitivity calibration constants (which are in fact temperature dependent) to change, resulting in a systematic shift over time in the indicated angle of attack. This translates into a systematic change over time in the estimated lift coefficient.

Assume for the sake of this illustration that the effect of the change is to cause the indicated lift to be greater than the true lift, by an amount that is progressively larger as time goes by. Then, relative to the sample mean, lift coefficients estimated from data acquired later will be biased higher, while lift coefficients estimated from data acquired earlier will be biased lower. If the angle of attack is likewise changed sequentially, indicated lift coefficients for the higher angles of attack (acquired later) will have a *positive* bias relative to the sample mean and indicated lift coefficients for the lower angles of attack (acquired earlier) will have a *negative* bias relative to the sample mean. Under such circumstances, a pres-stall lift polar would experience an undetectable counter-clockwise rotation. It would still have the appearance of a lift polar acquired without systematic error, but its slope would be greater. If the systematic error occurred at a non-uniform rate, it could introduce artificial structure in the sequential lift polar that might be mistaken for evidence of some complex (non-existent) aerodynamic phenomenon. Such a change is likely to go undetected until some future test in which an ostensibly identical lift polar will be discovered to have a different slope or different structure, discounting the highly improbable case in which identical systematic effects afflict both sets of data.

It turns out to be relatively easy to convert the insidious effects of systematic error into rather benign components of random error, simply by altering the order that the data are acquired. Details are beyond the scope of this paper, but are presented in the references^{3,4,6,15}. Such quality assurance tactics, which are pro forma in MDOE testing, require a departure from the sequential ordering that maximizes data acquisition rate.

C. A Solution to the OFAT Productivity/Quality Problem

The OFAT method is relatively unproductive because time constraints permit so few factor combinations to be examined one at a time. The results it produces are of relatively low quality because the OFAT practitioner is under so much time pressure to make a larger dent in what is for practical purposes an infinite number of factor combinations that he cannot take the time to properly assess experimental uncertainty through genuine replicates, or to engage in other quality assure tactics that ensure statistical independence at the expense of reduced data acquisition rate. To state the problem in informal but descriptive terms, the OFAT practitioner's quality and productivity problems both stem from the fact that the data points are not "working hard enough." By that it is meant that each data point carries with it relatively little information about the system under study; namely, the information necessary to describe the effects of a single factor change. With so little information packaged into each data point, a considerable volume of data is required to fully characterize the complex dependencies of forces, moments, and pressures on tunnel state, model attitude, and model configuration variables; significantly more data than can be acquired within the resource constraints of a typical wind tunnel test.

As noted in the introduction, the MDOE solution to this problem is to force each data point to "work harder" by carrying information about the effects of changing more than one factor at a time in each data point. An analysis of variance (ANOVA) can be applied later to properly attribute various response changes to the factors that were changed simultaneously. The next section establishes the concept of variance and its components, and introduces analysis of variance with the simplest example, a "fixed effects," "one way" ANOVA.

III. One-Way Analysis of Variance

We begin with an elementary review of variance, which will also serve to introduce terminology used in the remainder of the paper. Variance is a quantifiable measure of the degree to which individual data points are dispersed about some reference value in a sample of data. While the reference can be arbitrary, a common convention is to use the sample mean. Table 1 displays a sample of 15 genuine replicates of the lift coefficient measured under ostensibly identical conditions for each measurement. That is, for each of the 15 measurements the angle of attack, Mach number, control surface deflections, Reynolds number, angle of sideslip, vehicle configuration, and all other variables under the experimenter's control were set to as nearly identical values from point to point as could be established. Nonetheless, *for reasons we cannot entirely explain*, no two measurements of the lift coefficient are identical. We may be tempted to attribute the differences to set-point errors that are due to practical difficulties in replicating the angle of attack and other independent variables perfectly from point to point, and in fact set point error is expected to contribute to the response variance. However, even in the absence of set point error, there would still be a finite level of unexplained variance.

Every sample of experimental data is afflicted with unexplained variance to some degree, and it is because there is such unexplained variance in experimental results that there is also uncertainty in response estimates. Of the 15 different values of lift coefficient in Table 1, which can we say is, or is closest to, the "true value"? It is customary to assume that the unexplained variance is attributable entirely to random fluctuations about the true response, in which case averaging these 15 estimates of lift coefficient is expected to cause positive and negative experimental

errors to cancel to some degree, resulting in an estimate that is more reliable than any individual measurement. However, the unexplained variance may in fact feature a systematic component as well as a random component, as we will discuss presently.

Table 1. A Sample of Lift Coefficient Replicates.

Replicate	CL
1	0.1451
2	0.1636
3	0.1507
4	0.1534
5	0.1585
6	0.1390
7	0.1515
8	0.1557
9	0.1519
10	0.1638
11	0.1511
12	0.1602
13	0.1512
14	0.1633
15	0.1544

We can quantify the variation among individual data points in such a sample as Table 1 in terms of the sum of squared deviations (SS) from some convenient reference such as the mean. Table 2 provides the computational details for this example.

Table 2. Calculation of the Sum of Squares from Data in Table 1.

Point	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	0.1451	-0.0091	8.358E-05
2	0.1636	0.0094	8.771E-05
3	0.1507	-0.0036	1.272E-05
4	0.1534	-0.0008	7.206E-07
5	0.1585	0.0043	1.823E-05
6	0.1390	-0.0152	2.318E-04
7	0.1515	-0.0027	7.559E-06
8	0.1557	0.0015	2.283E-06
9	0.1519	-0.0023	5.303E-06
10	0.1638	0.0095	9.091E-05
11	0.1511	-0.0031	9.629E-06
12	0.1602	0.0060	3.562E-05
13	0.1512	-0.0030	9.008E-06
14	0.1633	0.0091	8.232E-05
15	0.1544	0.0002	4.616E-08
Mean:	0.1542	SS:	6.774E-04

The number of “degrees of freedom,” df , is the smallest number of points required to estimate the sum of squares, SS. When the reference used to compute the SS is the mean of an n -point sample, there are $n-1$ degrees of freedom because the value of the n^{th} point can be found by subtracting the sum of the first $n-1$ points from the

known sum of all points, which is just n times the mean. Therefore, if the mean is known, only $n-1$ points are needed to compute the sum of squares.

There are $n-1=14$ degrees of freedom associated with the 15-point sum of squared residuals about the mean in this example. The variance, or mean square (MS), is defined as the sum of squares per degree of freedom, so for this example:

$$\sigma^2 = \frac{SS}{df} = \frac{6.774 \times 10^{-4}}{14} = 4.839 \times 10^{-5} \quad (1)$$

The standard deviation (or standard error) in the distribution of replicates is just the square root of the mean square.

$$\sigma = \sqrt{\frac{SS}{df}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{6.774 \times 10^{-4}}{14}} = \sqrt{4.839 \times 10^{-5}} = 0.0070 \quad (2)$$

We noted above that the unexplained variance might feature a systematic component as well as the random component that is generally associated with unexplained variance in a wind tunnel test. We now consider an example of such systematic variation, which will be used to illustrate how an analysis of variance can be used to partition the random and systematic components.

Now assume that the 15 replicates in Table 1 were acquired in three tunnels with five measurements from each tunnel, as in Table 3. Here we assume that the same test article has been installed in all three tunnels for the purpose of assessing between-tunnel differences, and to the extent possible all conditions within each tunnel are identical.

Table 3. Ostensibly Identical Lift Coefficient Replicates Acquired in Three Tunnels.

Point	Tunnel		
	1	2	3
1	0.1390	0.1602	0.1585
2	0.1451	0.1515	0.1633
3	0.1519	0.1512	0.1534
4	0.1557	0.1544	0.1636
5	0.1511	0.1507	0.1638

The fundamental concept behind an analysis of variance in this case is that the total variance in the sample of data displayed in Table 3 can be partitioned into one component associated with within-column variation and another component associated with between-column variation. The within-column variation is attributable to ordinary chance variations in the data (random error), as before. We expect that random error will also be responsible for some degree of between-column variation, but if the tunnels are essentially identical, the amount of variation observed *between* columns (from tunnel to tunnel) will not be large compared to the within-tunnel variation. On the other hand, if the between-column component of the total variance is large compared to the within-column component, this can be taken as evidence of a systematic between-tunnel difference.

An analysis of variance in this example consists of computing the within- and between-column components of variance and comparing them. As in the case of the total variance, each component is computed as the ratio of a sum of squares (SS) and a corresponding number of degrees of freedom (df).

We compute the within-column SS by adding together the SS computed for each column. Each column SS is computed precisely as in the case of the total variance, by adding the squared differences between each value in a given column and the mean of that column. The computational details for the first column in Table 3 (Tunnel 1) are displayed in Table 4. Similar calculations are performed for the data from tunnels two and three of Table 4 to obtain their sums of squares.

Table 4. Sum of Squares Calculations for Tunnel 1.

Point	y_i	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	0.1390	-0.0096	9.162E-05
2	0.1451	-0.0035	1.218E-05
3	0.1519	0.0033	1.122E-05
4	0.1557	0.0072	5.131E-05
5	0.1511	0.0025	6.499E-06
Mean:	0.1486	SS:	1.728E-04

Because there are $n=5$ measurements in the sample from each tunnel, there are $n-1=4$ df for each of the three tunnels. The total df associated with the estimate of within-column variance is therefore 12. As indicated above, the sums of squares are likewise added across columns, as in Table 5.

Table 5. Calculation of Within-Column Variance.

Variance Component	Tunnel			Total
	1	2	3	
SS	1.728E-04	6.299E-05	8.278E-05	3.186E-04
df	4	4	4	12

With the within-column sum of squares and degrees of freedom in hand, the within-column variance follows immediately as the ratio of the former to the latter:

$$\sigma_{within}^2 = \frac{SS_{within}}{df_{within}} = \frac{3.186 \times 10^{-4}}{12} = 2.655 \times 10^{-5} \quad (3)$$

The within-column standard error is just the square root of this within-column variance, which has a numerical value of 0.0052. This represents a pooled standard deviation in lift coefficient estimates across the three tunnels.

From the data in Table 3, it is easy to compute the average lift coefficient based on five replicated in each tunnel. The three means are 0.1486 for Tunnel 1, 0.1536 for Tunnel 2, and 0.1605 for Tunnel 3. The fact that there is clearly variance in these means is not particularly interesting, since ordinary random error in a finite and relatively small sample of replicates from each tunnel virtually guarantees this result. A much more interesting question is whether the between-tunnel variance is large compared to the within-tunnel variance, indicating some systematic difference in lift coefficient estimates from one tunnel to another that is too large to attribute to simple random error. This is the question we are trying to answer with an analysis of variance.

Variance across columns is called “treatment variance” (from ANOVA’s early roots in agricultural research, in which ostensibly identical fields of crops were treated with different types of chemical fertilizers to assess their impact on crop yield). To compute the treatment variance for this example, we begin in the usual way by computing the sum of squared deviations of each column mean from that average of all three means. There is one slight wrinkle, however, necessitated by the fact that the number of rows and the number of columns is different. Each SS is based on the *average* of a number of individual measurements, so we weight the SS for each column by the number of measurements in that column and then sum:

$$SS_{treatment} = \sum_{columns} n_i (\bar{y}_i - \bar{y})^2 \quad (4)$$

Table 6 summarizes the treatment SS calculations for this example.

Table 6. Calculation of the Treatment SS.

Tunnel	Column			Residuals		
	Sum	Rows	Average	Value	Squared	Weighted
1	0.7429	5	0.1486	-0.0057	3.195E-05	1.597E-04
2	0.7680	5	0.1536	-0.0006	3.927E-07	1.964E-06
3	0.8026	5	0.1605	0.0063	3.943E-05	1.971E-04
Grand Mean:			0.1542	Treatment SS:		3.588E-04

There are $n-1=2$ df associated with the three-column treatment SS, so the between-columns (treatment) variance is:

$$\sigma_{\text{treatment}}^2 = \frac{SS_{\text{treatment}}}{df_{\text{treatment}}} = \frac{3.588 \times 10^{-4}}{2} = 1.794 \times 10^{-4} \quad (5)$$

At this point we have computed the SS and df describing the variance of the entire ensemble of 15 data points, as well as individually for the within- and between-column variance components. These calculations are summarized in Table 7:

Table 7. Summary of Variance Components.

	Within Columns	Between Columns	Total
df	12	2	14
SS	3.186E-04	3.588E-04	6.774E-04

Note that the within- and between-column df sum to the total df, and the within- and between-column SS sum to the total SS. This is a general property of the variance components so that once any two components are known the third can be determined by simple arithmetic. Note also that while the SS and df add, the variance components (ratios of SS to df) do *not* add.

At this point we have a 12-df estimate of a within-column sum of squares equal to 3.186×10^{-4} and representing random error variance, and a 2-df estimate of a between-column sum of squares equal to 6.778×10^{-4} and representing systematic tunnel-to-tunnel variance. We wish to know if the difference is significant. Here, as in all ANOVA applications, the term “significant” has no value judgment associated with it, in that a “significant” result does not denote a result of special importance. The term “significant” is simply used to describe an inference that can be made with some prescribed level of confidence. That is, a significant effect in the ANOVA sense is one that is unambiguously detectable, because it is relatively large compared to the uncertainty in estimating it. So in this example we would like to know if there clearly is a difference from one tunnel to another that can be unambiguously detected in the presence of such random error variance as characterizes the data from the individual tunnels.

In formal terms, we establish a null hypothesis that can be expressed for this example as follows, where μ_i represents the mean of the five lift coefficient replicates measured in the i^{th} tunnel:

$$H_0: \mu_1 = \mu_2 = \mu_3 \text{ (No between-tunnel differences)}$$

There is a corresponding alternative hypothesis, H_A , stating that even given the existing experimental error, the lift coefficient means for at least two of the tunnels can be distinguished with some prescribed level of confidence, taken in this example to be 95%.

The ratio of the between-column to within-column variance is drawn from an F-distribution under commonly occurring experimental conditions¹⁷ (that the experimental errors are normally and independently distributed with mean of zero and a constant variance, for which we rely upon the Central Limit Theorem and an assumption that time-independent bias errors have been eliminated by calibration, say). The F-distribution provides a reference probability density function against which we can compare our measured ratio of variance components to test the null hypothesis that there is no significant difference between the components.

An F-statistic is constructed as a ratio of two variance estimates. Under the null hypothesis of no significant difference between the numerator and denominator variances, the F-statistic is drawn from a skewed distribution that is part of a family of curves. The shape of each curve depends on the number of degrees of freedom associated with the variance estimates in the numerator and denominator. Figure 1 displays the F-distribution for 12 denominator degrees of freedom and 2, 5, and 10 numerator degrees of freedom.

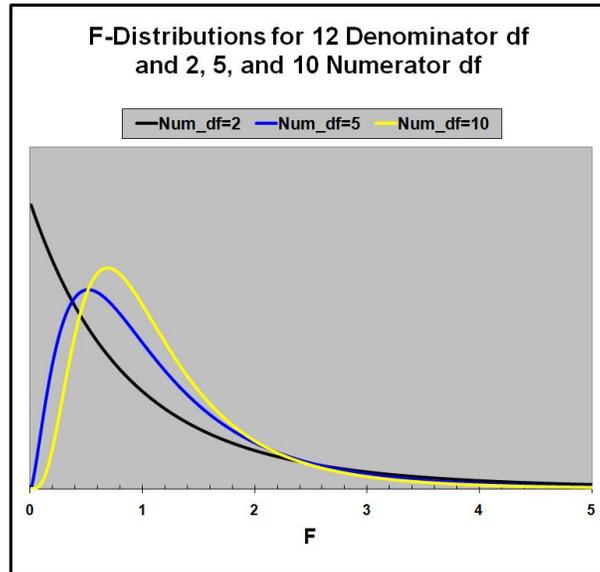


Figure 1. Selected members of a family of F-distributions with 12 denominator degrees of freedom.

The curves in Fig. 1 are probability density functions for which the area under each curve is equal to 1. For the F-distribution corresponding to the numerator and denominator degrees of freedom of the F-statistic we are evaluating (2 and 12), we establish a critical F-value at a location on the abscissa for which the area under the F-distribution to the right of this critical value is a probability that corresponds to our inference error risk tolerance. The inference error risk tolerance corresponding to the 95% confidence level specified in this example as acceptable is 0.05, and for the area under the F-distribution for 2 and 12 degrees of freedom to be 0.05 to the right of the critical F-value, it must be located at $F=3.885$. Such critical values are tabulated in standard statistical references¹⁷⁻¹⁹ and can be estimated using readily accessible commercial software packages. For example, the `FINV(prob, num_df, denom_df)` workbook function in Excel can be used to establish the critical value for this example. `FINV(0.05, 2, 12)` returns 3.885.

If the measured F-value exceeds 3.885, we can reject the null hypothesis that there is no significant difference distinguishing the between-column and within-column variance components. In such a case, we are entitled to conclude with no more than a 5% chance of an inference error that there is in fact a systematic difference in lift coefficient estimates between at least two of the tunnels that is causing the between-column variance to be significantly larger than that which can be attributed to random error.

In the current example, we have found that the within-column variance is only 2.655×10^{-5} (see discussion following Table 5), but that the between-column variance component describing tunnel-to-tunnel changes is 1.794×10^{-4} (discussion following Table 6). Our measured F-value is the ratio of these two variance estimates, which is $(1.794 \times 10^{-4}) / (2.655 \times 10^{-5}) = 6.758$, comfortably to the right of the 3.885 critical value. We therefore reject the null hypothesis and infer with at least 95% confidence that there is in fact a between-tunnel difference that is large enough to detect in the presence of experimental error.

We call the area to the right of the test F-statistic (6.758) a “p-value.” In this example, the p-value is 0.0108, and it represents the probability that one would observe an F-value of 6.758 or larger due simply to random fluctuations in the data if the null hypothesis were true. The probability that this would occur just over 1%, suggesting that it is unlikely that an F-value this large would occur by chance. We infer therefore that the between-tunnel variance is large enough compared to experimental error to conclude with little chance of error that between-tunnel effects are real. The p-value can also be computed using Excel or other readily available software packages. The FDIST(x, num_df, denom_df) Excel workbook function will return the p-value for a measured F-statistic, x, and specified values of the numerator and denominator degrees of freedom. FDIST(6.758, 2, 12) returns a value of 0.0108.

It is convenient to gather all of the ANOVA calculations in a table as in Table 8.

Table 8. ANOVA Table for Investigation of Between-Tunnel Differences.

<i>Source of Variation</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>	<i>P-value</i>	<i>F crit</i>
Between Columns	3.588E-04	2	1.794E-04	6.758	0.0108	3.885
Within Columns	3.186E-04	12	2.655E-05			
Total	6.774E-04	14				

For each variance component the table displays the sum of squares, the degrees of freedom, and the variance (or “Mean Square,” MS). The fact that the component sums of squares and degrees of freedom add up to the total sum of squares and total degrees of freedom for the entire ensemble of data is clearly illustrated. In addition, the measured F-value and p-value are displayed. One typically examines an ANOVA table such as this to seek variance components with p-values below the specified significance level (0.05, in this example). Since such components display a low probability of generating their corresponding F-values by chance, we conclude that they are due to some systematic effect in addition to ordinary chance variations in the data (random error).

The standard format for an ANOVA table includes the columns described here. We have also added the critical F-value corresponding to the 0.05 significance level for this example.

While the computational details for this analysis of variance have been presented in some detail to illustrate how multiple effects within a sample of data can be partitioned, such analyses are seldom performed by hand. The analysis of variance illustrated here can be performed automatically with the Excel ANOVA data analysis tool, and also with ANOVA tools found in many other readily available analytical software packages.

The analysis of variance performed here has partitioned the total unexplained variance into components attributable to different sources of error (random error and systematic between-tunnel effects), and has established a high probability that significant differences exist across the tunnels for which data were acquired. However, the ANOVA method is silent on the question of which tunnels if any are statistically indistinguishable and which differ from each other significantly. A large number of multiple comparison tests have been developed to identify dissimilar treatment means. One such test compares differences in treatment means with Fisher’s Least Significant Difference (LSD), which represents the smallest difference between two means that can be resolved with $(1-\alpha) \times 100\%$ level of confidence (alpha = 0.05 in this example). Assuming c columns, each with r rows, and with a within-column variance (also known as an error mean square) of MS_E , Fisher’s least Significant Difference can be computed as follows:

$$LSD = t_{\alpha/2, c(r-1)} \sqrt{\frac{2MS_E}{r}} \quad (6)$$

There are $c=3$ columns and $r=5$ rows in the current example and the mean square error, which is just the within-columns mean square from the ANOVA table, is 2.655×10^{-5} . The t-value is a tabulated coverage factor that depends on a specified number of degrees of freedom ($c \times (r-1)$, in this case) and a specified significance level. For $\alpha=0.05$:

$$LSD = t_{0.025, 12} \sqrt{\frac{2(2.655 \times 10^{-5})}{5}} = 2.179 \times 0.0033 = 0.0071 \quad (7)$$

Any two tunnels with 5-point sample means for lift coefficient that differ by 0.0071 or more are different by this test. Table 9 displays differences in sample means among the three tunnels in this example.

Table 9. Comparison of All Pairs of Lift Coefficient Sample Means for Three Tunnels

		1	2	3
		0.1486	0.1536	0.1605
1	0.1486	0	0.0050	0.0119
2	0.1536	-0.0050	0	0.0069
3	0.1605	-0.0119	-0.0069	0

The magnitude of the difference in lift coefficient sample means between tunnels 1 and 2 is only 0.0050, within the LSD of 0.0071. We conclude that there is no significant difference between these two tunnels. Tunnels 1 and 3 differ by an absolute value of 0.0119, enough to be resolved by the Fisher Least Significant Difference test. We therefore conclude that tunnels 1 and 3 did not produce the same values of lift coefficient. While there is not quite a significant difference between tunnel 2 and tunnel 3 by the LSD test, their difference of 0.0069 is very close to the LSD value of 0.0071. We conclude that with the volume data now in hand we are unable to detect a difference between tunnels 2 and 3 with 95% confidence, but that the acquisition of more data might permit a more definitive assertion about the similarity or difference between tunnels 2 and 3. Figure 2 displays how the means of the lift coefficient estimates differ from one tunnel to the next, relative to the pooled error variance across all tunnels as reflected in the widths of the probability density functions in this figure.

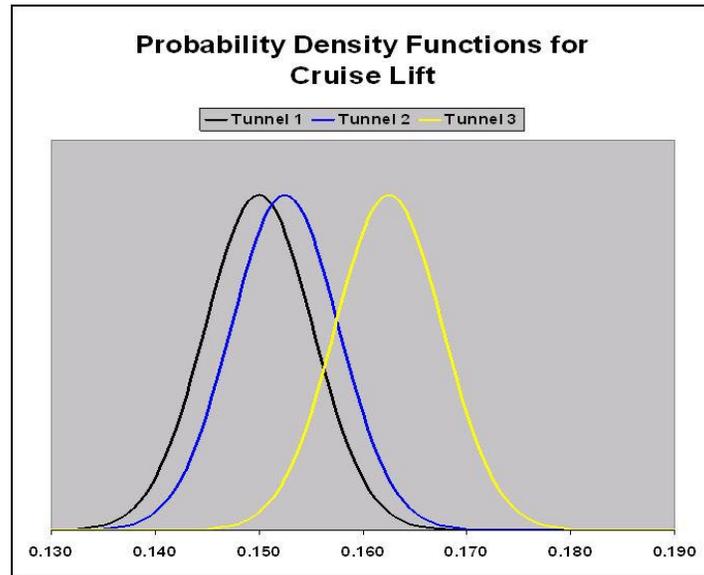


Figure 2. Probability density functions for estimates of lift coefficient from three tunnels.

IV. Two-Way Analysis of Variance

The ANOVA in the previous section is described as a *one-way* analysis of variance because only a single factor is involved besides random variation. That is, in a one-way ANOVA we seek to discover if variations in the levels of a single factor are large compared to random error. The factor in this example was tunnel selection.

Now consider a series of replicated lift coefficient polars acquired in a single tunnel at different times throughout the test. We can say for purposes of analysis that the data in each polar were acquired in a different “block” of time. We can then gauge how stable the tunnel is by examining the data to see if there are any significant *block effects*,

where a block effect is defined as systematic difference or bias shift between the data acquired in one block and the data acquired in another. This is analogous to the ANOVA we performed in the previous section, testing for significant differences from one tunnel to another.

Table 10 displays 10 pre-stall lift polar replicates acquired at various intervals of time in the same facility over a period of just over 17 days. All 10 polars are expected to be the same within experimental error, having been acquired with identical set-point values. That is, the same set-point levels for Mach number, control surface deflections, sideslip angle, and all other independent variables were specified for all 10 of the polars in Table 10, and each polar featured the same sequence of angle of attack settings.

Table 10. Ten Ostensibly Identical Pre-Stall Lift Coefficient Polars

Alpha	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
-2	-0.1881	-0.1809	-0.1850	-0.1801	-0.1863	-0.1774	-0.1654	-0.1711	-0.1759	-0.1702
-1	-0.1289	-0.1180	-0.1245	-0.1190	-0.1260	-0.1166	-0.1078	-0.1115	-0.1170	-0.1088
0	-0.0721	-0.0576	-0.0638	-0.0632	-0.0684	-0.0609	-0.0499	-0.0573	-0.0608	-0.0501
1	-0.0171	-0.0044	-0.0097	-0.0095	-0.0131	-0.0074	0.0015	-0.0054	-0.0055	0.0030
2	0.0365	0.0431	0.0414	0.0460	0.0407	0.0477	0.0530	0.0484	0.0461	0.0574
3	0.0896	0.0985	0.0971	0.1033	0.0944	0.1041	0.1096	0.1042	0.1029	0.1136
4	0.1442	0.1545	0.1527	0.1593	0.1495	0.1599	0.1644	0.1598	0.1578	0.1686
5	0.2010	0.2099	0.2051	0.2128	0.2049	0.2139	0.2186	0.2140	0.2107	0.2248
6	0.2557	0.2611	0.2574	0.2639	0.2577	0.2658	0.2720	0.2666	0.2647	0.2791
7	0.3055	0.3123	0.3101	0.3144	0.3087	0.3173	0.3211	0.3185	0.3163	0.3294
8	0.3560	0.3667	0.3634	0.3688	0.3621	0.3725	0.3760	0.3724	0.3733	0.3867
9	0.4144	0.4271	0.4249	0.4289	0.4202	0.4328	0.4399	0.4294	0.4325	0.4492
10	0.4785	0.4855	0.4821	0.4921	0.4812	0.4954	0.4992	0.4899	0.4902	0.5088
11	0.5426	0.5478	0.5453	0.5552	0.5436	0.5574	0.5616	0.5526	0.5491	0.5699
12	0.6036	0.6112	0.6094	0.6163	0.6057	0.6173	0.6225	0.6142	0.6118	0.6347
13	0.6608	0.6678	0.6678	0.6741	0.6662	0.6742	0.6797	0.6715	0.6683	0.6917
14	0.7146	0.7228	0.7211	0.7286	0.7229	0.7284	0.7343	0.7257	0.7222	0.7455
15	0.7672	0.7726	0.7730	0.7804	0.7743	0.7807	0.7866	0.7785	0.7751	0.7992
16	0.8192	0.8242	0.8256	0.8302	0.8223	0.8314	0.8367	0.8302	0.8257	0.8515

Note that Table 10 has the same general structure as Table 3 except that the number of rows and columns is different. There are rows and columns of lift coefficient measurements just as before, and we would like to invoke an analysis of variance to determine if there is any systematic variation between columns, just as before. If we applied the analysis of the previous section to the data in Table 10, it would result in the following ANOVA table:

Table 11. One-Way ANOVA Applied to Data of Table 10.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Columns (Runs)	9.869E-03	9	1.097E-03	0.011	1.0000	1.932
Within Columns (Runs)	17.9880	180	9.993E-02			
Total	17.9979	189				

The small F and large P for between-column variation suggest that there is no significant difference among the column means, so negligible difference from run to run. There are some red flags, however. The square root of the within-column mean-square (variance) is 0.3161. This should represent the standard deviation in random lift coefficient error (“one sigma value”), and yet it is two orders of magnitude larger than typical standard error values. Also, the between-column sum of squares is only about 1/20th of 1% of the within-column sum of squares. If the within-column sum of squares is a measure of irreducible random error as in the earlier one-way ANOVA, it seems quite unlikely that variations from one polar to the next would display so much less variation than that.

The solution to this apparent puzzle is that the one-way analysis of variance applied in the previous section must be extended slightly to account for an important distinction between the data in Tables 10 and 3. In Table 3, the row-to-row variation within any one column was attributable only to the chance variations that are caused by random

experimental error. In Table 10, the within-column changes in lift coefficient are due not only to random error, but also to a systematic change in angle of attack from row to row. Because the systematic change in lift coefficient from row to row was induced by changes in the angle of attack about which we are aware and which we in fact intentionally caused, we describe the row-to-row component of the total variance of Table 10 as “explained.”

In a perfect world, all the variance in a sample of data would be explained by changes intentionally induced by the experimenter. In practice, when we subtract from the total variance all that we can explain, there is always some residual, unexplained variance remaining, which is the reason that all experimental results are accompanied by some degree of uncertainty. In this example, at least some part of the residual variance is expected to be attributable to random error, but it is also possible that there is some unexplained variation from column to column, notwithstanding the fact that the columns represent ostensibly identical polars. It is this possibility that we wish to investigate. Under these circumstances we must use what is called a *two-way analysis of variance*, as will be illustrated in this section.

A two-way ANOVA further partitions the total variance in the sample of data beyond that which is performed in a one-way ANOVA. In a one-way ANOVA we are interested in only one potential source of variation besides the ubiquitous random experimental error but in a two-way ANOVA, there is more than one source of variation beyond random error.

To perform the two-way ANOVA, we will proceed much as we did with the one-way ANOVA, by computing first the total sum of squares and degrees of freedom for the entire sample of data and then for the row-wise and column-wise components of the total variance. However, we will provisionally attribute *both* the column-wise and row-wise variations to explained sources of variance. We will say that the row-wise variation is explained by changing angle of attack and the column-wise variation—to the extent that it differs from ordinary chance variations in the data—is explained by some systematic, persisting change occurring over time during the test.

Let $c=10$ be the number of columns in Table 10 and let $r=19$ be the number of rows. Then the row and column sums of squares are computed as follows:

$$SS_{rows} = \sum_{i=1}^r c(\bar{y}_i - \bar{\bar{y}})^2$$

$$SS_{columns} = \sum_{j=1}^c r(\bar{y}_j - \bar{\bar{y}})^2$$
(8)

where \bar{y}_i and \bar{y}_j are the means of the i^{th} row and j^{th} column, respectively, and $\bar{\bar{y}}$ is the “grand mean” of all data points in all rows and columns. Table 12 displays the row and column SS computational details, with $\bar{\bar{y}}$ highlighted.

Table 12. Sums of Squares for Two-Way ANOVA of Data from Table 10.

Alpha	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10	Row Means	Residuals	Squared Residuals	Columns	Weighted SS
-2	-0.1881	-0.1809	-0.1850	-0.1801	-0.1863	-0.1774	-0.1654	-0.1711	-0.1759	-0.1702	-0.1781	-0.5042	0.254	10	2.543
-1	-0.1289	-0.1180	-0.1245	-0.1190	-0.1260	-0.1166	-0.1078	-0.1115	-0.1170	-0.1088	-0.1178	-0.4440	0.197	10	1.971
0	-0.0721	-0.0576	-0.0638	-0.0632	-0.0684	-0.0609	-0.0499	-0.0573	-0.0608	-0.0501	-0.0604	-0.3866	0.149	10	1.495
1	-0.0171	-0.0044	-0.0097	-0.0095	-0.0131	-0.0074	0.0015	-0.0054	-0.0055	0.0030	-0.0068	-0.3329	0.111	10	1.109
2	0.0365	0.0431	0.0414	0.0460	0.0407	0.0477	0.0530	0.0484	0.0461	0.0574	0.0460	-0.2802	0.078	10	0.785
3	0.0896	0.0985	0.0971	0.1033	0.0944	0.1041	0.1096	0.1042	0.1029	0.1136	0.1017	-0.2245	0.050	10	0.504
4	0.1442	0.1545	0.1527	0.1593	0.1495	0.1599	0.1644	0.1598	0.1578	0.1686	0.1571	-0.1691	0.029	10	0.286
5	0.2010	0.2099	0.2051	0.2128	0.2049	0.2139	0.2186	0.2140	0.2107	0.2248	0.2116	-0.1146	0.013	10	0.131
6	0.2567	0.2611	0.2574	0.2639	0.2577	0.2658	0.2720	0.2666	0.2647	0.2791	0.2644	-0.0618	0.004	10	0.038
7	0.3055	0.3123	0.3101	0.3144	0.3087	0.3173	0.3211	0.3185	0.3163	0.3294	0.3154	-0.0108	0.000	10	0.001
8	0.3560	0.3667	0.3634	0.3688	0.3621	0.3725	0.3760	0.3724	0.3733	0.3867	0.3698	0.0436	0.002	10	0.019
9	0.4144	0.4271	0.4249	0.4289	0.4202	0.4328	0.4399	0.4294	0.4325	0.4492	0.4299	0.1038	0.011	10	0.108
10	0.4785	0.4855	0.4821	0.4921	0.4812	0.4954	0.4992	0.4899	0.4902	0.5088	0.4903	0.1641	0.027	10	0.269
11	0.5426	0.5478	0.5453	0.5552	0.5436	0.5574	0.5616	0.5526	0.5491	0.5699	0.5525	0.2263	0.051	10	0.512
12	0.6036	0.6112	0.6094	0.6163	0.6057	0.6173	0.6225	0.6142	0.6118	0.6347	0.6147	0.2885	0.083	10	0.832
13	0.6608	0.6678	0.6678	0.6741	0.6662	0.6742	0.6797	0.6715	0.6683	0.6917	0.6722	0.3460	0.120	10	1.197
14	0.7146	0.7228	0.7211	0.7286	0.7229	0.7284	0.7343	0.7257	0.7222	0.7455	0.7266	0.4004	0.160	10	1.603
15	0.7672	0.7726	0.7730	0.7804	0.7743	0.7807	0.7866	0.7785	0.7751	0.7992	0.7788	0.4526	0.205	10	2.048
16	0.8192	0.8242	0.8256	0.8302	0.8223	0.8314	0.8367	0.8302	0.8257	0.8515	0.8297	0.5035	0.254	10	2.535
Column Means	0.3149	0.3234	0.3207	0.3264	0.3190	0.3282	0.3344	0.3279	0.3257	0.3413	0.3262			Row SS:	17.987
Residuals	-0.0113	-0.0028	-0.0055	0.0003	-0.0072	0.0020	0.0082	0.0017	-0.0005	0.0151					
Squared Residuals	1.272E-04	7.904E-06	3.001E-05	6.288E-08	5.202E-05	4.186E-06	6.739E-05	3.020E-06	2.904E-07	2.273E-04					
Rows	19	19	19	19	19	19	19	19	19	19	Column SS				
Weighted SS	2.417E-03	1.502E-04	5.703E-04	1.195E-06	9.883E-04	7.954E-05	1.280E-03	5.739E-05	5.517E-06	4.319E-03	9.869E-03				

The total sum of squares is computed just as before:

$$SS_{total} = \sum_{i=1}^r \sum_{j=1}^c (y_{ij} - \bar{y})^2 \quad (9)$$

We estimate the error sum of squares by subtraction, exploiting the fact that the error SS and the SS for rows and columns all sum to the total SS:

$$SS_{error} = SS_{total} - (SS_{rows} + SS_{columns}) \quad (10)$$

The row degrees of freedom are $r-1=19-1=18$ and the column degrees of freedom are $c-1=10-1=9$. The total degrees of freedom are $rc-1=189$. We can also compute the error degrees of freedom by subtraction, exploiting the fact that degrees of freedom for rows, columns, and error all sum to the total degrees of freedom. The error degrees of freedom are then $df_{total} - (df_{rows} + df_{columns}) = 189 - (18 + 9) = 162$. We can also compute the error df from this formula:

$$df_{error} = (rows \times columns) - rows - columns + 1 = (19 \times 10) - 19 - 10 + 1 = 162 \quad (11)$$

As before, we gather the SS and df data into a table and compute the variance or mean square by dividing the SS by the df. We compute F for both the row and column variance by dividing the mean square of each by the error variance, comparing each such F with the corresponding critical F value obtained from statistical tables or computed with Excel as described for the one-way ANOVA or using other readily available software. The p-values are likewise computed as described in the one-way ANOVA. We gather all of the calculations into an ANOVA table as before.

Table 13: Two-Way ANOVA Applied to Data of Table 10.

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	17.987105	18	9.993E-01	176626.1	0	1.668
Columns	0.0098692	9	1.097E-03	193.8	6.247E-82	1.938
Error	0.0009165	162	5.658E-06			
Total	17.997891	189				

We can now draw objective inferences from the F and p-values in the ANOVA table. For the row-wise variance we note that the F of 176,626 substantially exceeds the critical F ($\alpha=0.05$) of 1.668, and from the p-value we see that the probability that an F-value this large could occur due simply to random variations of the magnitude observed in the data is “zero,” to within the computer’s prodigious ability to resolve small numbers. This provides reassuring if not exactly unexpected confirmation of the rather unremarkable fact that lift coefficient does indeed vary significantly with angle of attack, an inference we are entitled to make in this case with what is effectively 100% confidence (1-p).

The row-wise variance is not interesting in and of itself in this case, but it is necessary to compute this explained-variance component in order to isolate the more interesting *unexplained* variance, consisting of a random error component and a possible component associated with column-to-column variation. We performed this ANOVA for the express purpose of objectively determining whether column-to-column variation is significant.

The column F of 193.8 is two orders of magnitude greater than the critical F ($\alpha=0.05$) of 1.938, and from the p-value we see that the probability that an F-value this large could occur due to chance variations in the data is vanishingly small (on the order of 10^{-82}). We therefore reject the corresponding null hypothesis of no column-to-column variation, and conclude with at least 95% confidence that two or more otherwise identical polars acquired in different blocks of time differ from each other to a greater degree than can be attributed to ordinary chance variations in the data.

As in the case of the one-way ANOVA, we are only able to establish that two or more columns differ significantly from each other. We cannot say just from the ANOVA results which columns (polars) are similar and which are different. We note, however, that under the null hypothesis of no differences in polars acquired in one block of time or another, each column mean should be the same within experimental error. The average value of a sample of lift coefficients measured at different angles of attack has a physical significance that will be explained shortly, and it also serves as a convenient reference for quantifying variation from column to column.

We again invoke Fisher’s Least Significant Difference formula to compute a threshold for the magnitude of the difference in column means that we can resolve, given the degree of random error in the data. This formula, given earlier, is reproduced here for convenience and to reflect the relevant degrees of freedom in this two-way ANOVA:

$$LSD = t_{\alpha/2, rc-r-c+1} \sqrt{\frac{2MS_E}{r}} \quad (12)$$

For $r=19$ rows, $c=10$ columns, and a significance level, α , of 0.05, the t-statistic that serves as a coverage factor in this formula corresponds to 162 degrees of freedom and has a value of 1.975, very close to the limiting value of 1.960 that corresponds to an infinite number of degree of freedom. The Mean Square Error, MS_E , has a value of 5.658×10^{-6} , from Table 13. Inserting these values, we compute a LSD of 0.0015 for column means in this example. By this test, any two columns with means that differ by 0.0015 or more can be distinguished with at least 95% confidence.

Applying this criterion to the columns of data for which means are displayed in Table 12, we discover the remarkable fact that of the 45 possible pairs of column means, all but three pairs differ by more than Fisher’s LSD, and can therefore be declared different with at least 95% confidence. Figure 3 is a dot-diagram of column means for this example, with run-number labels. Red bars are displayed over pairs of column means that cannot be distinguished from each other with at least 95% confidence by Fisher’s Least Significant Difference test. Note that only runs 4 & 9, 4 & 8, and 8 & 6 are indistinguishable. The rest of these ostensibly identical lift polars all have unique mean values. This figure suggests why the ANOVA results indicate with such low probability of an inference error (low p-value) that we are entitled to reject the null hypothesis of no significant difference among polar means.

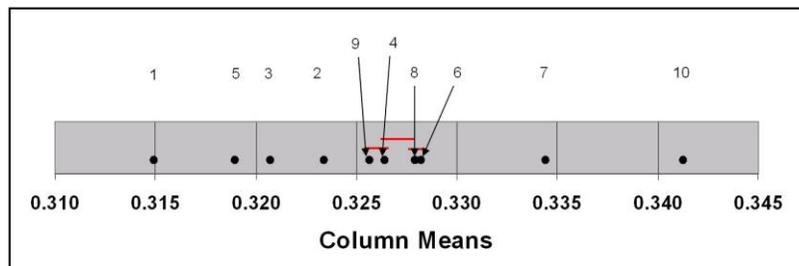


Figure 3. Dot diagram of column means. Red bars span indistinguishable pairs.

It was noted earlier that the polar means have a certain physical significance. This physical interpretation of the means provides an important insight into what is happening when polar means vary with the block of time in which the polar is acquired.

It is customary to apply a linear transformation to prescribed factor levels that serves to “center” and also to “scale” the independent variables. Centering simply transforms the variables into a symmetrical range about a convenient reference such as zero. This decouples the y-intercept of fitted models from other regressors in the model, which permits an independent consideration of the shape of the response model, and the overall level as dictated by the y-intercept. Changes that may impact the shape of the model have no effect on the y-intercept when the independent variables are centered, and vice-versa.

Scaling maps the independent variable ranges into a specified range. A common and convenient range runs from -1 to +1. Such scaling prevents computational round-off errors when dealing with factors that span vastly different ranges in physical engineering units (for example, Reynolds numbers in the millions and flap-wing gaps that are in fractions of a millimeter). Scaling also facilitates a direct comparison of factor effects in terms of the fraction of full range over which each factor is varied in the test.

A transformation that both centers the factor levels and scales them to a range of ± 1 is as follows:

$$x = \frac{\xi - \frac{1}{2}(H + L)}{\frac{1}{2}(H - L)} \quad (13)$$

where x is the factor level in coded (centered and scaled) units, ξ is the factor level in physical units, H is the physical value for $x=+1$, and L is the physical value for $x=-1$. When linear regression models are fitted as a function of factors that have been coded with this transformation, the y -intercept is computed by averaging the fitted response levels used in the regression. Therefore in the current example, if the lift coefficient data from each polar is fitted as a function of angle of attack, the y -intercept of each such model will be equal to the polar mean. The effect of time-varying y -intercept terms is illustrated in Fig.4.

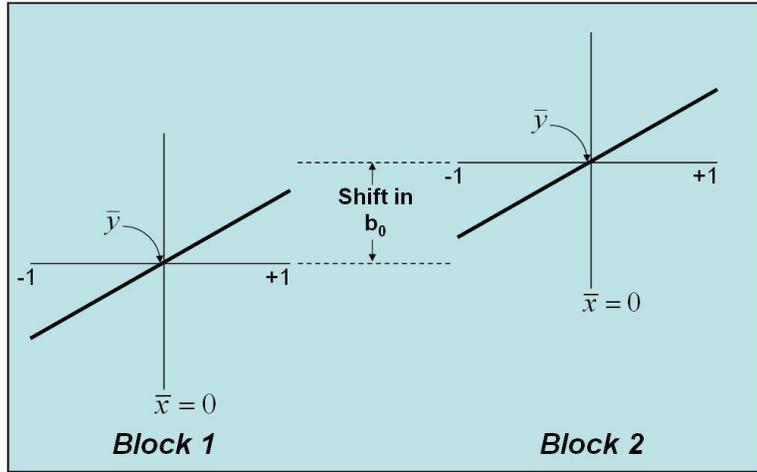


Figure 4. Shift in y -intercept, b_0 , of a regression model fitted from replicated data acquired in two blocks when a significant block effect is in play.

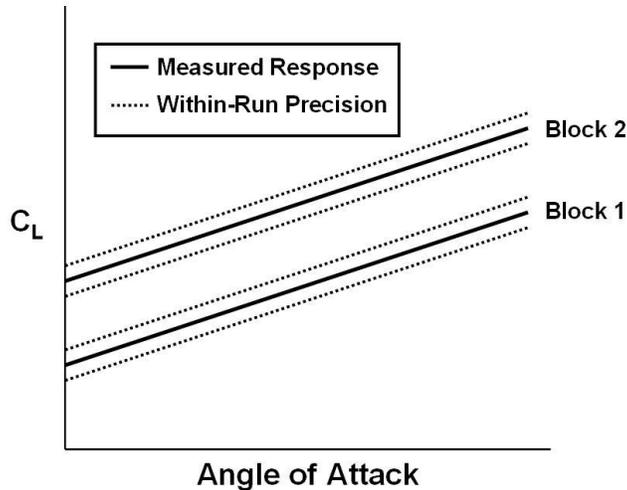


Figure 5. Block effects frustrate efforts to generate results that are reproducible to within irreducible chance variations in the data.

Block effects as illustrated in Fig. 4 are not uncommon in wind tunnel testing, notwithstanding how often they go undetected during a particular test. They tend to impose themselves during efforts to reproduce data acquired in an earlier block of time. Figure 5 illustrates how block effects such as those detected in the data examined here can influence ostensibly identical polars acquired in different blocks of time.

To further examine the block effects revealed in this sample of data, we establish a convenient reference y-intercept equal to the average of all ten column means and determine how much each individual y-intercept is shifted from this mean. Table 14 lists these block effects.

Table 14. Shifts in y-intercepts of 10 Replicated Lift Polars in Different Blocks of Time.

<i>Run</i>	b_0	\bar{b}_0	b_0 Shift
1	0.3149	0.3262	-0.0113
2	0.3234	0.3262	-0.0028
3	0.3207	0.3262	-0.0055
4	0.3264	0.3262	0.0003
5	0.3190	0.3262	-0.0072
6	0.3282	0.3262	0.0020
7	0.3344	0.3262	0.0082
8	0.3279	0.3262	0.0017
9	0.3257	0.3262	-0.0005
10	0.3413	0.3262	0.0151

The polars analyzed in this section were acquired over the period of time indicated in Table 15. It is instructive to examine how the block effects that have been detected change with time.

Table 15. Days and Times That Replicated Polars Were Acquired

<i>Run</i>	<i>Day/Time</i>
1	Day 0, 3:00 PM
2	Day 2, 7:30 AM
3	Day 2, 2:00 PM
4	Day 3, 7:00 PM
5	Day 5, 11:00 AM
6	Day 9, 11:00 AM
7	Day 13, 9:00 AM
8	Day 14, 11:00 AM
9	Day 17, 4:30 PM
10	Day 18, 8:07 AM

Recall that Fisher's Least Significant Difference for the polar means under examination was computed to be 0.0015. The LSD differs from a 95% confidence interval half-width by the square root of two, so we can make a rough estimate of the uncertainty in individual polar means by dividing the LSD of 0.0015 by $\sqrt{2}$, yielding 0.0011. In Fig. 6 we plot the block shifts of Table 14 against the times listed in Table 15.

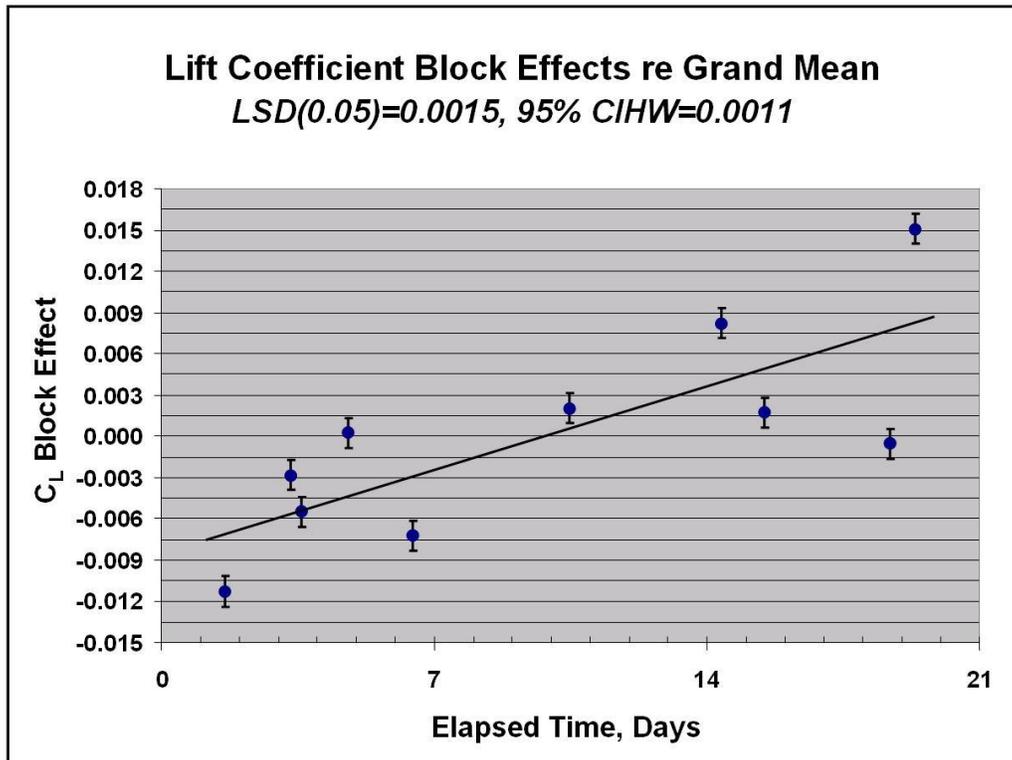


Figure 6. Block effects reveal unexplained systematic variation over time. Grid lines spaced at LSD of 0.0015 so points farther apart than this can be resolved with 95% confidence.

The slope of the trend line in Fig. 6 was estimated as nine counts per day (0.0009/day), with a 95% confidence interval half width of six counts. Since the 95% confidence interval that therefore ranges from three to fifteen counts per day (0.0003/day to 0.0015/day) does not contain zero, we infer that that a real trend is in fact in play, and that some systematic change persisted over the 17+ days during which data were acquired that resulted in progressively higher lift coefficient measurements.

Note that the error bars on each point in Fig. 6 represent the irreducible random error of the measurements, and that the systematic error revealed by the trend line is a substantial *multiple* of the random error. It is the much smaller random error that is normally detected and reported in conventional (OFAT) wind tunnel tests, while systematic error represented by the trend line in Fig. 6 generally goes undetected and unreported. The between-columns F value of 193.8 in the ANOVA table in Table 13 is the ratio of a variance component due to between-polar variations plus ordinary random error, to a component attributable to random error alone. The square root of that 193.8 F value is 13.9, which is the ratio of the corresponding standard errors. That is, the standard error from run to run is 13.9 times greater than the within-run standard error as indicated in Fig. 7.

The square root of the error mean square from Table 13, 5.658×10^{-6} , is only 0.0024. This represents the actual irreducible within-polar random error in lift coefficient. The mean square for run-to-run variations is 1.097×10^{-3} from Table 13, the square root of which is 0.0031. This is the between-run standard error, which is an order of magnitude larger. These calculations reveal that for this data sample, average variations from run to run dominated the unexplained variance, with the random error component being essentially negligible by comparison. This is not uncommon in wind tunnel testing.

We note in passing that for these same ten polar replicates, significant block effects similarly exhibiting a pronounced trend with time were also observed for drag coefficient and the coefficients of pitching moment and rolling moment. No significant trends were detected in the coefficients of yawing moment or side force.

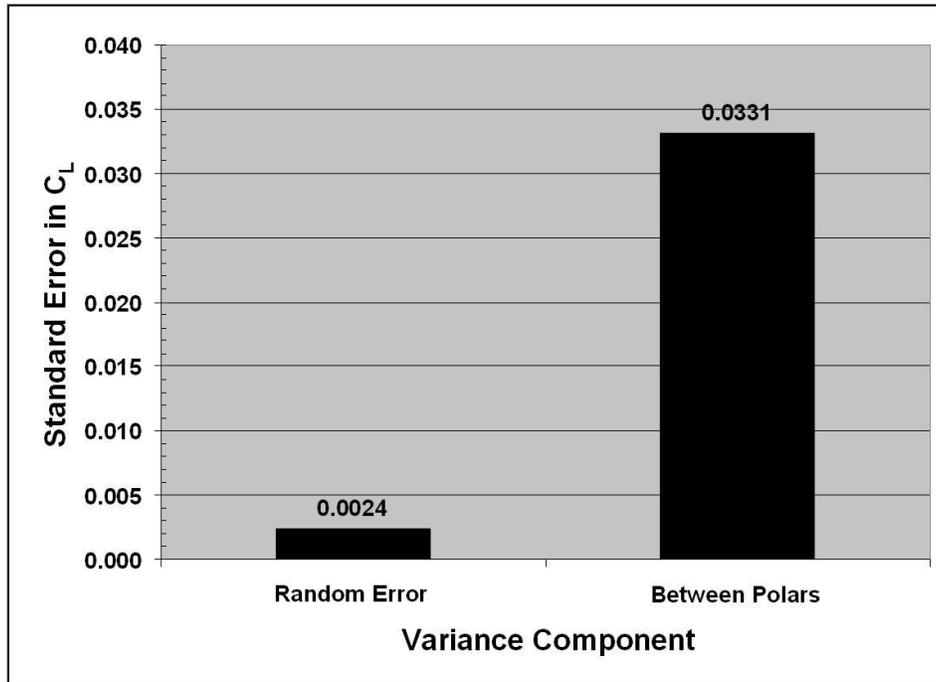


Figure 7. Relative contributions of within- and between-polar components of unexplained variance.

V. Regression and the Random Effects Analysis of Variance

The previous two sections described how an analysis of variance is applied in the common situation in which specific treatments are chosen by the experimenter. The one-way ANOVA dealt with three specific tunnels. The two-way ANOVA dealt with replicate polars acquired in 10 specific blocks of time during a single test. These are examples of what is known as a *fixed-effects* ANOVA, in which the results apply to the specific treatment means considered in the analysis. So, for example, the results of the one-way ANOVA apply only to the three specific tunnels from which data were acquired for that analysis. They cannot be extrapolated to other tunnels. That is to say, we cannot assume that the between-columns (cross-tunnel) variance quantified in that analysis characterizes differences among arbitrary groups of tunnels.

Likewise, the two-way ANOVA performed earlier applies to the replicate polars analyzed in specific blocks of time for one wind tunnel test. The precise results of this analysis cannot be generalized to all wind tunnel tests, beyond a reasonable inference that similar effects may be observed under similar circumstances in other tests. For example, we cannot infer from this analysis that lift coefficients increase systematically with time in all wind tunnel tests just because they did in this one. We may know from experience, as indeed the author does know, that unexplained variance in a wind tunnel test consists generally of both systematic and random components, with the systematic component almost always dominating the random component as in this test. However, this is not an inference we can make from a single fixed-effects analysis of variance.

We consider now an alternative scenario in which the treatments analyzed in the ANOVA are regarded as a sample that is drawn from a larger population. In this situation we are less interested in the specific treatment means than in making inferences about the population as a whole. This type of analysis, known as a *random-effects* ANOVA, is applicable in situations of practical interest in experimentation in which a subset of all possible levels of an independent variable are included in the experiment. Such situations arise in response surface modeling applications, for example, where we wish to understand how some response of interest (a force or moment in a wind tunnel test, say) depends on various factors or more generally, on various regressors that each consist of some function of one or more factors. We are not as interested in making inferences about a specific level of a given regressor as we are about such questions as whether changes in the level of that regressor have a significant influence on system response. We therefore focus on the response variance that is induced when the regressor changes over some prescribed range. It is variance in this sense that interests us in a random effects ANOVA, rather

than the variance associated with differences in the specific treatment means that are central to a fixed-effects ANOVA. The calculations used in a random effects analysis of variance are identical to those used in a fixed effects analysis of variance, but the results apply to a general population of treatment levels rather than to specific treatment levels.

As with the fixed-effects ANOVA discussed earlier, in a random-effects ANOVA we are interested in mean square values that consist of the ratios of sums of squares and degrees of freedom. We test the ratio of various treatment mean squares to some reference error mean square using a reference F distribution to infer the significance of each treatment mean square.

We will consider a response surface modeling application in this section, in which familiar concepts developed in the earlier discussion of fixed-effects ANOVA are extended. The central concept is that an ensemble of experimental data is characterized by its variance, which can be partitioned into explained and unexplained components. The unexplained component, which is responsible for experimental uncertainty, can be further partitioned into random and systematic components. This was illustrated earlier, when the one-way ANOVA revealed a systematic component of unexplained variance attributable to tunnel differences, and the two-way ANOVA revealed a systematic component of unexplained variance attributable to some unknown, persisting change that was in play during a wind tunnel test. The explained variance can also be partitioned into components, each associated with changes associated with a given regressor in a proposed response model. The essence of ANOVA for response surface modeling is to determine for each candidate regressor if its mean square exceeds the error mean square sufficiently to reject a null hypothesis that the term makes no significant contribution to model response predictions. Regressors of progressively higher order can be added to the model and tested against this criterion until the addition of further terms to the model makes no significant difference in its predictive capability.

D. Explained and Unexplained Variance in a Regression Context

Consider the sample of data displayed in Fig. 8. There are $n=19$ measurements in this sample, and no two are identical. The sample mean is indicated, as is the formula for computing the total variance of the sample.

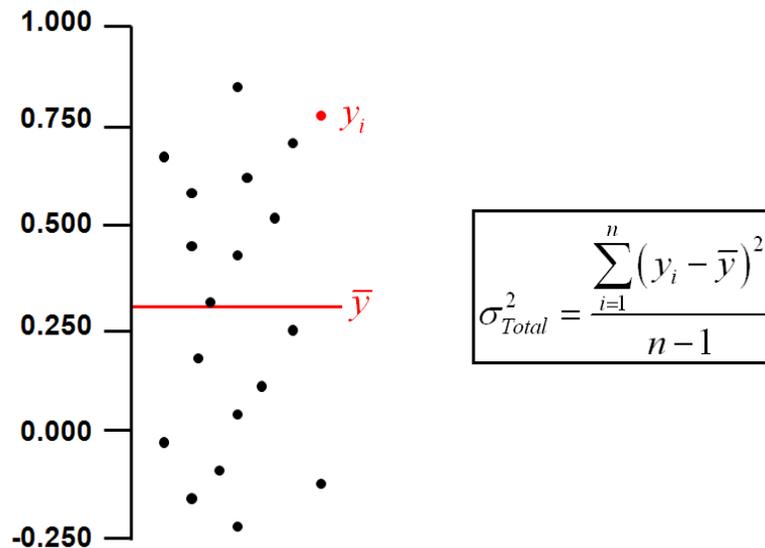


Figure 8. A sample of data with variance.

Now consider Fig. 9, which displays the identical data set of Fig. 8 in the dashed rectangle on the left. On the right of Fig. 9, identically the same 19 data points are displayed, except that each point is shifted to the right by varying degrees. The ordinal values of the data points remain unchanged, so the mean of the sample on the right is the same as the mean of the sample on the left, and so is the total variance. That is, there is identically the same “scatter” in the data on the right as on the left. And yet the data on the right appear in some sense to be more orderly. We say that we have “explained” much of the variation in the data in terms of some explanatory variable that is displayed on the abscissa in coded units in the range of -1 to +1.

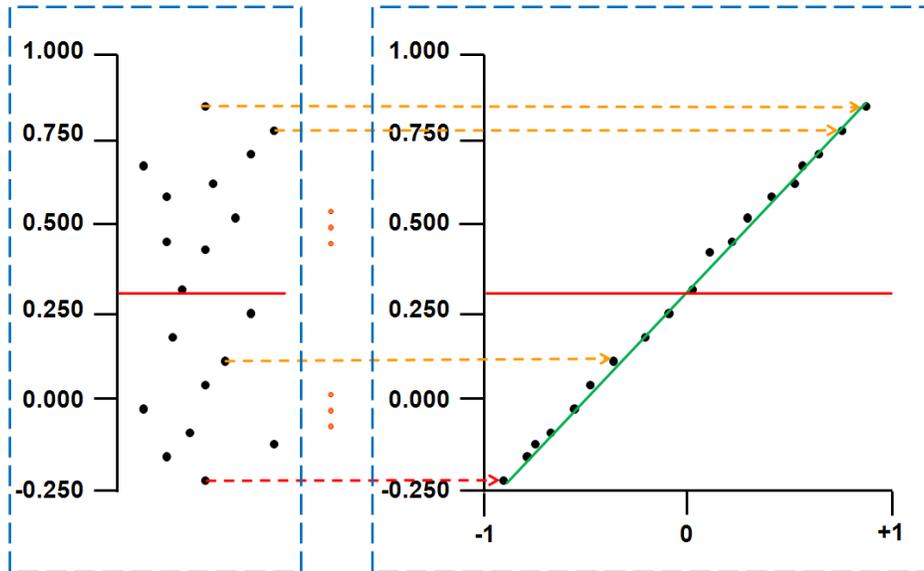


Figure 9. “Explained variance” in a sample of data.

We wish to quantify that part of the total variance that has been explained by the first-order response model represented by the straight line in Fig 9. We proceed as before, by calculating the sum of squares and degrees of freedom associated with the explained component of the total variance. The ratio is the explained mean square, or variance. Figure 10 illustrates how it is calculated.

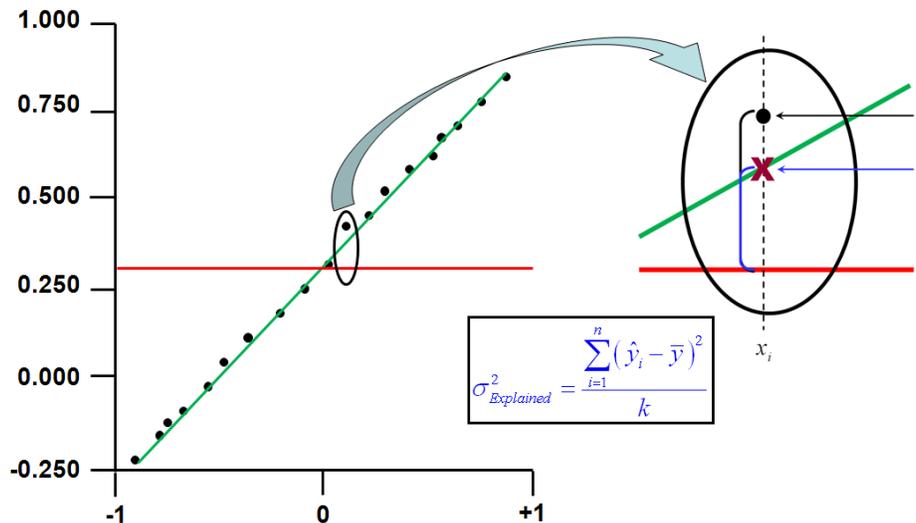


Figure 10. Explained component of the total sample variance.

Note in Fig. 10 that the explained sum of squares in the numerator of the variance formula is of the same general form as the formula for the total sum of squares for the entire data sample. Instead of squaring the difference between each *measured* value and the sample mean and then summing all the squared values, we square the

difference between each *predicted* value and the sample mean before summing. In a perfect world in which every measured data point fell exactly on the regression line, the explained and total sum of squares would be identical.

The quantity k in the formula for explained variance in Fig. 10 is the number of regressors in the equation of the response model. This is the number of degrees of freedom associated with the explained variance, which is just the number of terms in the response model not counting the intercept term. For the straight-line model in this example, the only regressor is the slope of this line, so $k=1$.

Given the data displayed in Fig. 8 and again in Figs. 9 and 10, and given the formulas for total and explained sums of squares and degrees of freedom, it is possible to compute those quantities, and by subtraction calculate the sum of squares and degrees of freedom for the unexplained component of the total variance. Again we exploit the fact that the explained and unexplained degrees of freedom sum to the total degrees of freedom, and likewise for the sums of squares. Note that there are $n-1$ total degrees of freedom given the mean, and k degrees of freedom for the explained variance, so there are $n-k-1$ degrees of freedom for the unexplained variance. We assemble the sums of squares and degrees of freedom in an ANOVA table, and compute mean squares, F-statistics, and p-values just as before. Table 16 is the ANOVA table for this example:

Table 16. ANOVA Table for First Order Response Model in One Variable.

Source of Variation	SS	df	MS	F	p-value
Model	1.792753716	1	1.792753716	46867	< 0.0001
Residual	0.000650279	17	3.82517E-05		
Total	1.793403995	18			

We see from the large F and small p-value that the straight-line model displayed in Figs. 9 and 10 explains a great deal of the total variation in the data sample. If we establish a threshold significance level of 0.05 corresponding to a 95% level of confidence, we see that the p-value is well below this threshold and so we are entitled to conclude with at least 95% confidence that the addition of a linear explanatory term to the model reduces the unexplained variance significantly. The situation represented in Fig. 8 is one in which we have not attempted to explain the variance in the data with any regressor. We have a very primitive model of the response in such a case, as follows:

$$y = b_0 \tag{14}$$

where b_0 is the sample mean, a constant with a value of 0.3261 in this case. Absent any other regressor information, this would be the best available representation of the data. The associated unexplained variance would simply be the total sum of squares, 1.7934, divided by the total df, 18, or 9.063×10^{-2} , and the standard error would be the square root of that, or 0.3156. One would therefore report the response with a “one-sigma” uncertainty as 0.3261 ± 0.3156 .

Note that adding a single term to the response model, consisting of a first order function of a variable we will call x_1 , dramatically reduces the unexplained variance. If the original “sample mean” model is augmented with another term as follows:

$$y = b_0 + b_1 x_1 \tag{15}$$

The unexplained variance drops three orders of magnitude, from 9.063×10^{-2} to 3.825×10^{-5} , the residual mean square in Table 16. Here, the quantity b_1 is of course the slope of the best straight line fit to the data. The corresponding standard error is the square root of the unexplained variance, or 0.0062. This is a substantial improvement over the standard error that existed before the second term was added to the model (about a factor of 50), and represents the degree to which the addition of this term helped further “explain” the variance displayed in the data.

It is reasonable to enquire if the model can be further improved by adding additional regressors. For example, one might consider the addition of a second regressor consisting of a quadratic term, resulting in the following second-order function of the variable x_j :

$$y = b_0 + b_1x_1 + b_{11}x_1^2 \quad (16)$$

The b_{11} term is a coefficient determined by linear regression. Progressively higher-order terms could be added to the model, each “explaining” a further increment of the total variance in the data until a limit is reached in which an 18th-order polynomial is used to fit these 19 data points. Such a model would pass through every data point, driving to zero that portion of the total variance in this sample of data that is unexplained by the response model. Such an extreme model would not be attractive, however, since we understand that some portion of the total variance is due to experimental error and is therefore necessarily unexplained. Fitting progressively higher-order models simply results in fitting the noise in the data sample, and provides no useful insight into the underlying physical process described by the relationship of the system’s response to the regressors in the response model.

Note that we are not constrained to examine only higher-order terms involving the same independent variable. It is possible to apply the same response modeling process to functions of more than one variable. For example, if the data sample in this example had been acquired while a second variable, x_2 , was also being changed, we would want to examine how adding first- and high-order functions of both variables would increase the explained variance and reduce the unexplained variance. We now consider this more general case.

E. Partitioning Factor Effects—The Extra Sum of Squares Principle

The introduction described how practitioners of the Modern Design of Experiments improve productivity by ensuring more information per data point than in a conventional one-factor-at-a-time experiment. This is achieved by changing *more* than one factor at a time, which permits the design space to be traversed much more quickly. That, combined with the use of response surfaces rather than reliance upon an exhaustive enumeration strategy requiring that the test facility be individually configured for every individual factor combination of interest, results in significant productivity advantages. The question that originated this tutorial introduction to the analysis of variance still remains: If we change multiple factor levels per data point, how can we tell how much of the resulting system response change is due to changes in one factor, and much is due to another? We now address that question in the context of analysis of variance, with a description of the extra sum of squares principle.

Consider Fig 11, which compares the MDOE design space for an experiment in two factors with the equivalent OFAT design space.

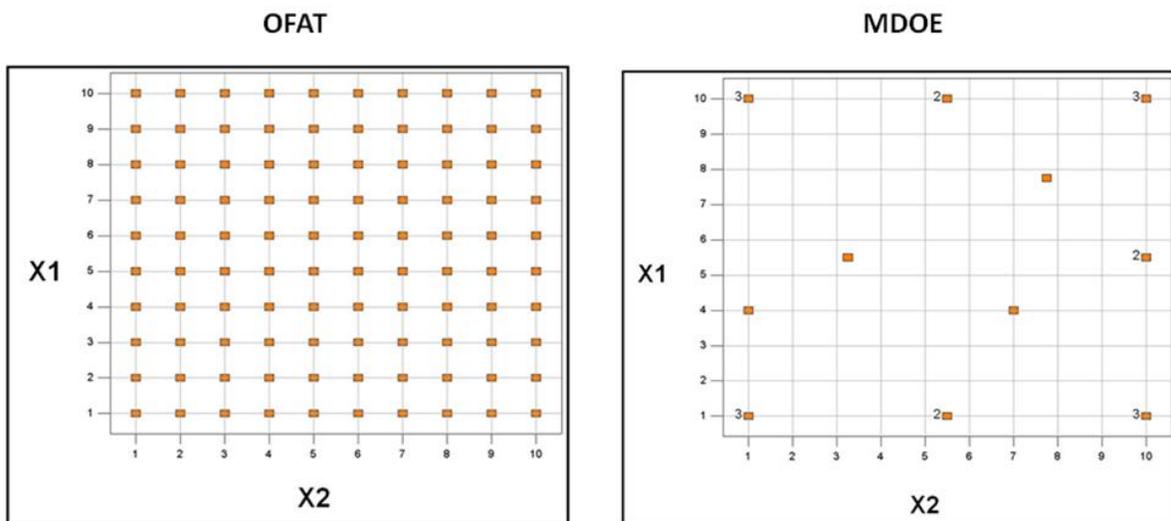


Figure 11. Comparison of MDOE and OFAT design spaces for a two-factor experiment. Numbers next to selected MDOE sites indicate replication. Factor levels are coded.

The site selection process in an MDOE experiment (process of deciding how many and which factor combinations to set) is beyond the scope of the current paper but methods and practical aerospace examples are described in the references^{2, 11, 12, 14, 26-30}. Suffice it to say that for the not-atypical comparison illustrated in Fig. 11, the MDOE design offers three advantages. The most obvious is that it requires fewer points and so consumes less cycle time and direct operating costs. The second is that it features replication, which facilitates a direct estimation of irreducible random error. Furthermore, the replicates are acquired at a variety of combinations of low and high factor levels, providing a robust estimate of random error levels throughout the design space. Finally, the site selection decisions are made to improve quality, not data acquisition rate or operator convenience. The specific factor levels in the MDOE design space were chosen to minimize uncertainty in the coefficients of a regression model used to fit the data.

Coded system responses from the OFAT test design on the left of Fig. 11 are plotted in Fig. 12 to display how the system responds to changes in both factors x_1 and x_2 . The nature of the specific response, which might be a force or moment from a wind tunnel test, for example, is unimportant for the purposes of this illustration, which applies for any system response, no matter how non-linear, as long as the response is continuous. Discontinuous system responses are treated by truncating the inference space so as to place the discontinuity on a subspace boundary, with response surfaces developed on either side of the boundary. It is problematic to estimate responses at the discontinuity, but no more so in an MDOE response surface experiment than in an OFAT experiment.

The response surface representation in Fig. 12 is a common data structure in MDOE testing, but in OFAT testing the data are typically displayed as a family of functions of one variable, with the response plotted as a function of x_1 , say, for various fixed levels of x_2 . It is largely a matter of individual taste as to which display type is most effective and this is not the issue here. The central question is whether it is possible to reproduce the OFAT results acquired with the relatively large data volume (100 points) on the left of Fig. 11 and displayed in Fig. 12, with the much more parsimonious test matrix represented on the right of Fig. 11. We answer this question by illustrating how the analysis of variance is used to extract individual factor effects when both factors are changed on each data point.

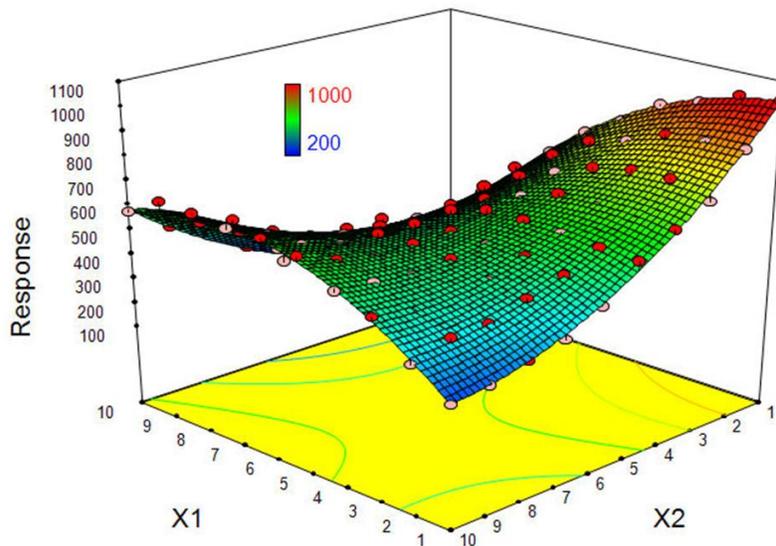


Figure 12. Response measurements acquired with a 100-point OFAT test matrix.

The MDOE test matrix displayed graphically in Fig. 11 is presented in Table 17 in run order. Note the absence of systematic structure that characterizes OFAT test matrices; these points are selected to minimize uncertainty in a response model representation, which seldom leads to the kind of rectilinear symmetry displayed in the OFAT design space of Fig. 11. This is an important distinction between MDOE and OFAT experiment designs. Not all test matrices are created equal, and OFAT matrices, in addition to be larger and therefore more costly and more time-

consuming to execute, also tend to generate greater experimental uncertainty, as will be further touched upon presently.

Returning to the question of how to exploit ANOVA methods to objectively infer relationships between system response and independent variable levels, we can begin by postulating a simple response model, testing it for adequacy, augmenting it with additional regressors if necessary, repeating the adequacy assessment, and continuing until a response model is achieved that adequately represents system response over the range of independent variables considered. We rely upon hypothesis testing that is facilitated by the analysis of variance at each step in this process. A detailed example of this process will now be provided to illustrate how the calculations are performed, but the reader should keep in mind that in practice all of these calculations are performed essentially simultaneously by computer, using commercially available software³¹⁻³⁵.

Table 17: MDOE Test Matrix with Points in Run Order.

X1	X2	Measured
7.75	7.75	496.4
1	10	264.5
10	10	546.7
1	10	280.5
1	1	1059.6
1	1	1014.9
1	10	246.6
1	5.5	483.4
10	1	119.5
1	1	1020.6
10	10	589.2
10	5.5	184.1
5.5	3.25	622.6
10	1	97.9
5.5	10	629.8
5.5	10	636.6
4	7	545.8
10	1	129.3
10	10	607.6
4	1	883.8
10	5.5	174.7
1	5.5	493.2

We begin with the total sum of squares and total degrees of freedom (given the mean), as computed earlier for other data samples. There are 22 data points so $n-1=21$ degrees of freedom, and the corresponding sum of squares is 1,887,913 for this data set. Some of this total sum of squares is inevitably attributable to experimental error but we wish to find a linear combination of regressors, each a function of x_1 and x_2 , which comprises a response model that can be used to explain most of the rest of the total sum of squares.

Note that it is unreasonable to expect to account for 100% of the sum of squares not associated with experimental error, since this would imply a perfect response model. We typically use low-order polynomial “graduating functions” that serve to approximate the true but unknown underlying functional relationship between the response of interest and the independent variables, but these are in fact only approximations. The intent is to approximate the responses “adequately” (that is, within some acceptable level of precision), but we cannot do so “perfectly” without a priori knowledge of the true underlying response function, which of course is unknown.

For pedagogical reasons, we will start the analysis with $y=b_0$ as the model (where b_0 is just the mean of the data in Table 17) and add terms of increasing complexity to this model. However, in actual applications we generally begin by postulating a more complex model. We may not know the details of the actually response dependence on the factors we have examined in the experiment, but we can rely upon subject matter expertise to know that there is at least a high probability that both factors will be in the model (else they would not have been included in the test matrix). We can also reasonable assume that the dependence is likely to be higher than first-order, a forecast that can be made with some confidence about essentially any response of interest in an aerospace experiment for which the

factors have been changed over ranges of practical interest, and certainly about all forces and moments in a wind tunnel test. Nonetheless, we will begin at the beginning with a simple $y = b_0$ model, which we augment initially with a simple first order term in x_1 . That is, we will consider initially this model: $y = b_0 + b_1x_1$. We use regression methods to estimate the b_i model coefficients, and then apply the model to estimate responses for all 22 data points in the sample. The difference between each predicted response and the sample mean is squared and the 22 squared values are summed to compute the explained sum of squares for this candidate response model, just as described in the prior sections on fixed-effects ANOVA. For the data sample displayed in Table 17, the result is an explained SS of 397,401 corresponding to the first-order model in x_1 . Since all of the variance in the original $y = b_0$ model was unexplained, we attribute the explained sum of squares of the $y = b_0 + b_1x_1$ model to x_1 . There is one degree of freedom associated with this term, so the mean square and the sum of squares is the same.

The SS for unexplained variance is computed as before, by subtracting the explained SS from the total SS. In this case, the unexplained SS is therefore $1,887,913 - 397,401 = 1,490,511$. As noted in the previous section there are $n-k-1$ degrees of freedom associated with the component of variance explained by a model with k regressors beyond the intercept term, so for $n=22$ data points and $k=1$ regressors in the model, the unexplained variance has 20 degrees of freedom. The unexplained or error mean square is thus $1,490,511/20 = 74,526$ for this model.

We now know that the mean square associated with the addition of a linear x_1 term to the model is 397,401 and the error mean square is 74,526. We construct an F statistic by taking the ratio of these two mean square values: $F = 397,401/74,526 = 5.332$. We compare with a critical F-value associated with a significance level of 0.05, one numerator df, and 20 denominator df, which is 4.351. Since the measured F value is greater than the critical F, we reject the null hypothesis of no significant difference in the two mean squares, and conclude with at least 95% confidence that some of the total variance in the data can be explained by variations that were made in x_1 over the range that it was changed during the test. We therefore retain the linear x_1 term in the model.

We now provisionally add a linear x_2 term to the model: $y = b_0 + b_1x_1 + b_2x_2$ and repeat the process. We find that the unexplained SS decreases from 1,490,511 to 1,425,456. The difference of $1,490,511 - 1,425,456 = 65,055$ is the sum of squares we associate with the first-order x_2 term. This is an illustration of what is known as the Extra Sum of Squares Principle. Each new term explains more of the total variance, and the unexplained SS is therefore reduced. Equivalently, the explained SS is increased by the same amount. We attribute the change in the SS to the newly added term, and use it to perform an F-test for the significance of that term just as before. That is, we generate a mean square value by dividing the extra sum of squares by the degrees of freedom associated with this addition to the model (one regressor so 1 df), and form an F value by dividing this by a newly computed error mean square. The error mean square will be the ratio of an error SS equaling the prior error SS less the extra sum of squares, to an error df equal to the prior error df reduced by 1. In this case, the MS associated with the x_2 term is 65,055 and the error MS is $1,425,456/19 = 75,024$, so the corresponding F value is $65,055/75,024$, or 0.867. This compares with the critical 0.05 F-value for 1 numerator df and 19 denominator df of 4.381. The fact that the measured F is less than the critical F suggests that we are unable to reject the null hypothesis of no first-order x_2 effect, and should therefore drop the x_2 term from the model.

It is at this point that we should remind ourselves of an obligation to bring our subject matter expertise to bear as well as our statistical analysis capability. It is unlikely that the x_2 variable had absolutely no effect in this test, so why the apparent insignificance of adding it to the response model? An examination of the error mean square sheds some light on this question.

The square root of the 75,024 error mean square is 273.9. This represents a standard error (“one sigma uncertainty”) that is 54% of the sample mean of 505.8. It is highly unlikely that an adequate model would fit the data with so little precision. We also have the benefit of 11 replicates in this experiment (see numbers next to selected sites in the design space displayed in Fig. 11). The standard deviation computed from those 11 replicates is only 19.8, an order of magnitude smaller than the standard error of the regression when only the first order x_1 and x_2 terms are included in the model. We conclude that the apparent insignificance of the x_2 term may be due to the large unexplained variance associated with this simple first-order model, rather than the x_2 term contributing negligibly to response predictions. We know that the unexplained variance will continue to erode as additional terms are included in the model, so that the first-order x_2 term may in fact emerge as significant in a more complex model. We therefore provisionally retain this term, and press on to add additional regressors.

Having at this stage incorporated the x_1 and x_2 first-order terms, we examine the interaction between them, computing the explained sum of squares for a model that contains an x_1x_2 interaction term in addition to the first order terms considered thus far:

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 \quad (17)$$

After estimating the b_i model coefficients by linear regression and computing the sums of squares as before, we see that the addition of the interaction term to the model has a considerable impact on the explained and unexplained sum of squares, increasing the former and decreasing the latter by 1,192,775. The prior unexplained sum of squares was 1,425,456, so the inclusion of an interaction term explains a substantial portion of the formerly unexplained variance, reducing the unexplained sum of squares to 232,681 with $n-k-1 = 22-3-1 = 18$ df. The error mean square is thus $232,681/18 = 12,927$. Since there is only one df for the interaction term for which a sum of squares of 1,192,775 is associated, the SS and mean square are the same, and the F statistic for this term is then $1,192,775/12,927 = 92.27$, which is considerably larger than 4.414, the critical F for a significance of 0.05, 1 numerator df, and 18 denominator df. We therefore reject the null hypothesis of no significant difference between the mean square for the interaction term and the error mean square, and conclude with at least 95% confidence that the interaction term makes a significant contribution to the response prediction.

Note also that the substantial reduction in unexplained variance achieved by introducing the interaction term to the response model has essentially improved the “signal-to-noise ratio” affecting the way in which all candidate model terms are evaluated. For example, the first-order x_2 term had an F-statistic of 0.867 when it was examined earlier, but the reduction in error mean square resulting from the addition of the interaction term increases this to 4.204. Likewise, the F-statistic for the x_1 term increases from 5.332 to 30.44. Since more of the previously *unexplained* variance is now *explained*, we are able to see with greater clarity the contributions of the individual candidate model terms.

We continue this process with the addition of candidate terms of progressively higher order, evaluating the impact that each new term has on the explained and unexplained components of the total variance in the data sample. Eventually a point of diminishing returns is reached, when the unexplained variance has been driven to levels sufficiently low that there is no significant difference with irreducible levels of ordinary random error. If further precision were required, it would be necessary to acquire additional replicates. There are also circumstances in which an excessively complex model would be necessary to drive residual variance levels to acceptably low levels by the process that has been described here. For example, it is possible that response measurements were made at too few unique levels to support a model of the required order. The MDOE test matrix of Table 17, displayed graphically in Fig. 11, features five levels of x_1 . It is therefore not possible to fit higher than pure fourth-order terms in x_1 . The x_2 variable is likewise limited since it was set at only six levels. There is thus an upper limit on the number of terms one can continue to add as the response model is being constructed. If that limit is confining, it may be necessary to augment the test matrix with measurements made at alternative sites in the design space. Another common remedy is to divide the design space into subspaces within which lower-order models can be constructed. The response is then represented as a piecewise continuous response model spanning multiple design subspaces.

At each step of the model-building process described here, an error mean square was constructed from the unexplained sum of squares and its corresponding degrees of freedom. The square root of the error mean square is the standard error of the regression, comprising a “one-sigma” representation of the model’s prediction uncertainty. Fig. 13 shows how this quantity decreased in the current example as each new term was added to the model. Each new term explained an additional fraction of the total sample variance, reducing the unexplained variance by the same amount. The standard error was therefore reduced with the addition of each new term, until all six terms of a full quadratic model in two independent variables were included in the model. At that point, the standard error had been reduced to the level of the standard deviation in genuine replicates acquired as part of the test matrix. The tentative addition of a mixed third-order term made no substantive further reduction in the standard error, and it was decided to retain only the terms of the full quadratic model.

$$y = b_0 + b_1x_1 + b_2x_2 + b_{12}x_1x_2 + b_{11}x_1^2 + b_{22}x_2^2 \quad (18)$$

Figure 13 also reveals a number of interesting insights into the response under investigation. The non-linear nature of the response is revealed by the relatively sharp drop in residual standard error after the first-order terms in the model are augmented by second-order terms. The interaction term causes an especially sharp drop, indicating that the effect that either variable has on the response is strongly dependent upon the level of the other factor. There is also clearly curvature revealed by the fact that the addition of quadratic terms further reduces the residual standard error.

Before leaving the discussion of how an analysis of variance is used to partition the effects of multiple factor changes, we note that the Extra Sum of Squares Principle can be used to compute the sum of squares associated with a specific model term in more than one way. The discussion in this example has centered on what is called a

“sequential” or Type I calculation of the sums of squares. This calculates sums of squares in sequence. The SS for a term is computed by subtracting from the total sum of squares the sum of squares for all terms of equal or lower order already in the model. The sums of squares for individual terms will sum to the total explained sum of squares, but the value of individual terms can be dependent on the order in which the terms were added to the model.

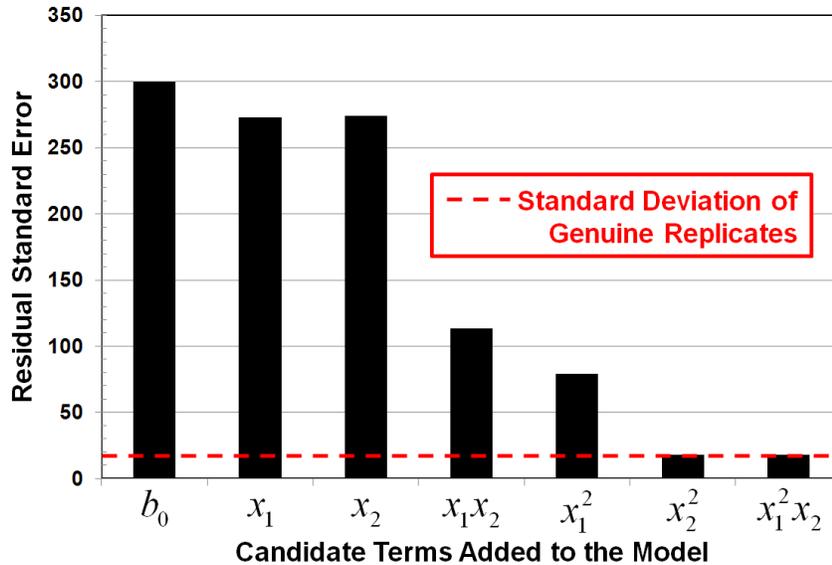


Figure 13. Residual standard error in response model after the addition of progressively higher-order terms. No significant improvement after all quadratic terms included.

A slightly more generalized form of calculation for the sum of squares is called the “Partial” or “Type III” method, in which the SS for a given term is computed by subtracting from the total SS the SS for all other terms in the model, regardless of their hierarchical order relative to the current term. This is the method often used by commercial statistical software, however it has the weakness that for non-orthogonal models, the SS for the individual terms may not add up the total explained SS. Table 18 is the ANOVA table for this example, computed using a Type III calculation of the sum of squares. The reader can compare sums of squares values in this table to estimates made earlier with a sequential approach. There are some differences in the sums of squares, but the F and p-values that result from the mean square calculations lead to the same inferences in this case; namely, that all terms in the full quadratic model are significant, and terms of higher order are not.

Table 18. ANOVA Table for Full Quadratic Response Surface, Type III SS.

Source of Variation	SS	df	MS	F	p-value
Model	1,882,598	5	376,520	1133	< 0.0001
x_1	373,686	1	373,686	1125	< 0.0001
x_2	81,178	1	81,178	244.4	< 0.0001
x_1x_2	1,149,097	1	1,149,097	3459	< 0.0001
x_1^2	154,943	1	154,943	466.4	< 0.0001
x_2^2	101,292	1	101,292	304.9	< 0.0001
Residual	5,315	16	332		
Total	1,887,913	21			

For completeness we note that a “Type II” or “Classical” sum of squares for a candidate model term is equal to the reduction in the unexplained SS caused by adding that term after all other terms have been added to the model except those that contain the term being tested. So, for example, to perform a classical calculation of the SS

associated with the x_2 term in the full quadratic model of this example, we would first compute the explained SS for a model consisting only of the intercept and terms that did not contain x_2 (the intercept plus the linear and quadratic x_1 terms), then we would subtract that from the SS explained by the same model but including the x_2 term.

In commonly occurring experimental conditions, the inferences that are made with respect to the significance of candidate terms in a model will not depend critically on the method used to compute the sum of squares. The decision to retain or reject certain marginally significant terms may be influenced to some degree by the computational method, but as a practical matter the presence or absence of marginally significant terms in the model will tend to have relatively little influence on response predictions.

F. Model Confirmation

This section has dealt with the application of concepts from the analysis of variance to the problem of partitioning explainable variance in a sample of experimental data into components that can be attributed to individual independent variables. Building such a model requires some level of judgment to decide, for example, how high an order of model to use to represent the data. However, much of the activity associated with building a response model relies upon relatively well-established procedures that are largely automated. The majority of the workload associated with developing a response model is expended not in constructing the model, but in validating it. This topic extends well beyond the scope of the present paper, except to note that dozens of tests of a candidate model are typically applied before it is offered as a representation of some system response. We illustrate one such test to close this discussion of model-building.

The acquisition of what are called “confirmation points” is highly recommended in a response surface modeling experiment. These are points that are acquired at the same time as the data used to fit the response model, but that are not used in the regression computations. They are instead held in reserve to test the model by providing the means to compare response predictions with measured response values at various sites in the design space. In this example, the OFAT data acquired at sites illustrated on the left of Fig. 11 provide an ideal collection of confirmation points by which to test the quadratic model developed in this section using regression and ANOVA concepts.

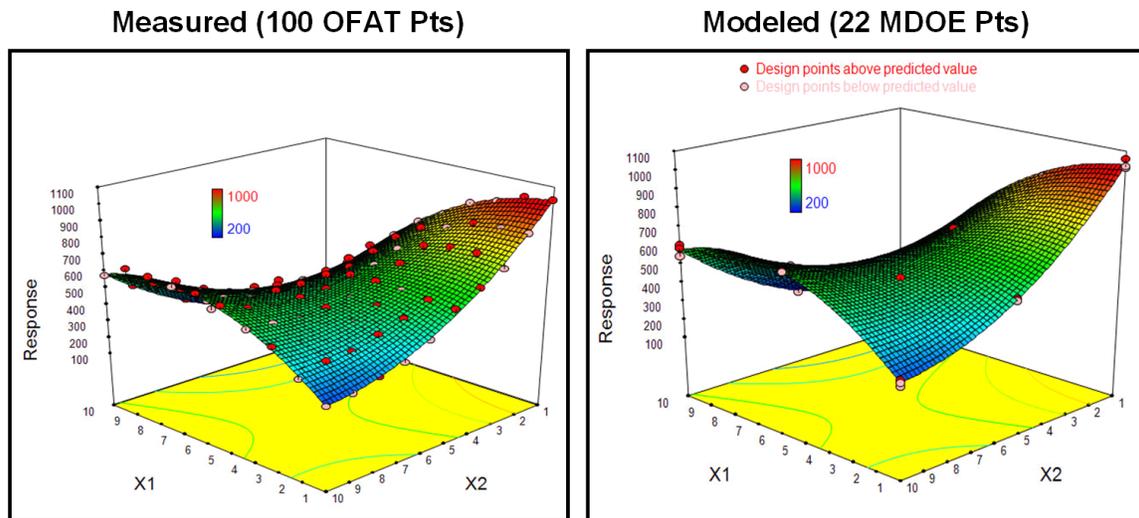


Figure 14. Comparison of response model with 100 OFAT confirmation points.

Figure 14 compares the response surface obtained from the 100 OFAT data points on the left of Fig. 11 with the quadratic response model fitted to the 22 data points on the right of Fig. 11. Qualitatively there is a great deal of similarity. To examine the comparison more closely, one can simply set one of the independent variables in the response model to a level corresponding to one held constant in the OFAT test, and predict the responses for the tested range of the other variable.

While this too is a topic that is beyond the scope of the current paper, a significant advantage of response surface modeling is that it is possible to predict not only the response for a given combination of factor levels, but also the

lower and upper limits of a specified precision interval. Figure 15a presents the lower and upper limit of a 95% prediction interval for $x_1=1$ and x_2 ranging from 1 to 10. Our claim is that for these combinations of factor levels, there is a 95% probability that measured response levels will fall above the lower limit and below the upper limit.

Figure 15b shows the 10 OFAT responses measured at $x_1=1$, with 95% confidence interval error bars for each point. This is a conventional format for data display.

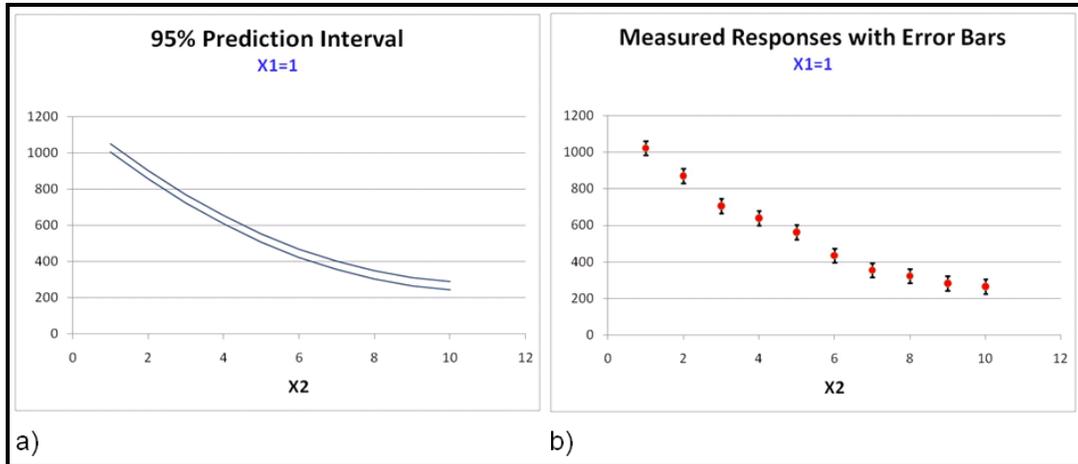


Figure 15. Comparison of MDOE response model precision interval with OFAT confirmation points.

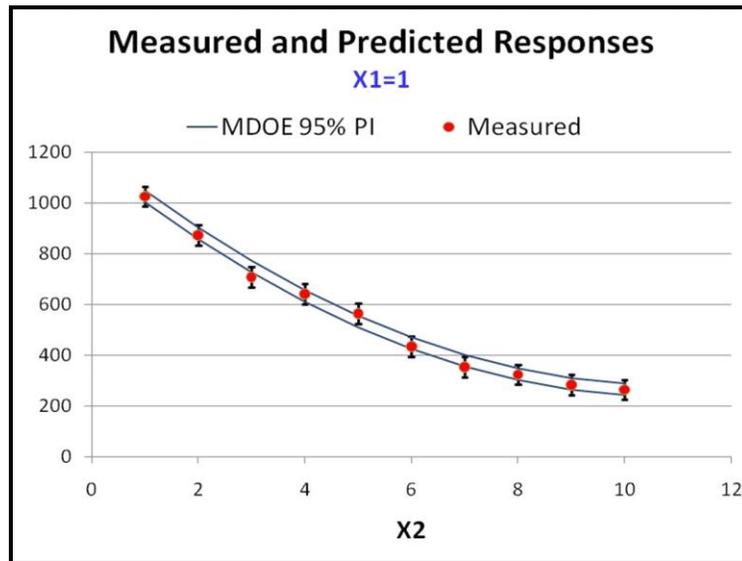


Figure 16. Comparison of measured and predicted response values.

Figures 15a and b are superimposed in Fig. 16. Clearly the quadratic response model represents system responses within experimental error for these factor combinations.

This comparison was made for $x_1=1$, but of course the same type of comparison can be made at other values of x_1 , corresponding to other slices through the response surface parallel to the x_2 axis. Likewise, the rectilinear array of OFAT measurements displayed in Fig. 11 permits us to make confirmation-point comparisons just as easily with slices through the response surface that are parallel to the x_1 axis at fixed levels of x_2 . Figure 17 displays a half a dozen “slices” parallel to each of the two axes, all indicating good agreement between the response model and the measured response data it purports to represent.

Figures 11 and 14–17 illustrate how in this case an MDOE test matrix featuring only 22 measurements at 11 unique combinations of two independent variables could be used to forecast response measurements made at the 100 design-space sites of a corresponding OFAT test. The productivity advantage is not quite 100:22 since the MDOE test matrix is not optimized for speed, and it typically takes about 1.5 to 2.5 times as long to execute an MDOE test matrix as it would take to execute the equivalent OFAT test matrix. It would have taken as long to acquire the 22 MDOE data points in this example as it would take to acquire 33 to 55 OFAT points, depending on the specific rate differential, which represents a reduction in cycle time and attendant direct operating costs of a factor of roughly two to three.

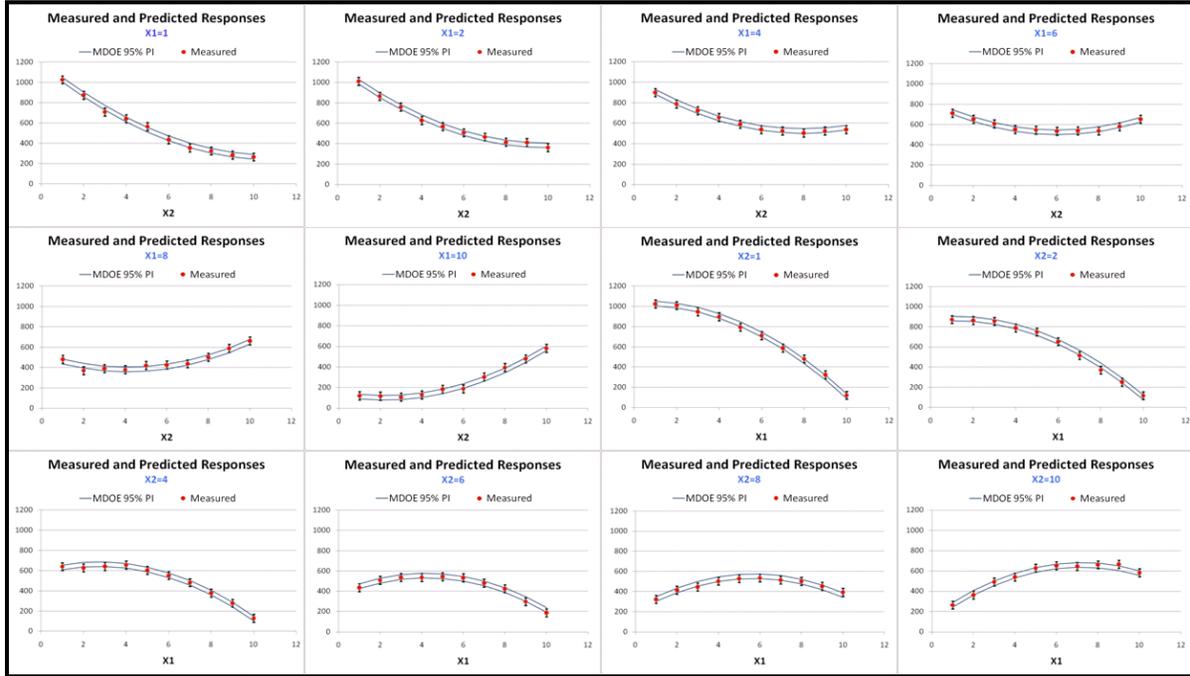


Figure 17. Comparison of response model predictions and measured confirmation points.

One is entitled to ask if such an improvement in productivity comes at a cost in quality. If the 100 OFAT data points were analyzed as a response surface, with six df consumed by the model leaving 94 residual df available to assess uncertainty, then the OFAT uncertainty estimates would in fact be smaller than in the MDOE analysis, by just over a factor of two. However, the uncertainty in OFAT tests is generally assessed on a per-point basis, with the standard deviation of a relatively small number of replicates used to estimate some standard deviation (in those instances in which any replicates are specified at all). In this example, none of the OFAT points was replicated, but it might be assumed that some estimate of standard deviation was available from prior experience in this facility.

In an MDOE analysis such as this one, the total degrees of freedom not consumed in estimating the model are available to assess the uncertainty. In this example there were $n=22$ total data points and only $p=6$ were required to fit a full quadratic model in two independent variables, leaving 16 df to assess uncertainty. The prediction uncertainty for a response surface model is a function of site location in the design space as well as the unexplained variance in the data, but the average prediction variance over all fitted points is described by the same formula no matter the order of the polynomial or the number of independent variables.

$$\bar{\sigma}_y^2 = \left(\frac{p}{n} \right) \sigma_{data}^2 \quad (19)$$

where p is the number of parameters in the model (six in this example), n is the number of data points fitted (22), and σ_{data} is the ordinary standard deviation in unexplained variance that characterizes the data. The square root of this is the standard error in the prediction model. Since it is not possible to fit a p -parameter model with fewer than $n=p$ points, the p/n term in parentheses will never be greater than one and will generally be less than one. In this

case, the standard error in model prediction is $\text{SQRT}(6/22) = 0.52$ times the standard deviation in the raw data that is used for an OFAT standard error. Standard precision intervals that are proportional to the standard error will be reduced by this amount.

Figure. 18 compares MDOE and OFAT cycle time and uncertainty assuming an OFAT data acquisition rate that is 2.5 time faster than MDOE. The reduction in data acquisition time under this assumption is accompanied by a similar reduction in uncertainty. (It should be noted that total cycle time includes other activities besides data acquisition that are not influenced by efficiencies in the design of the experiment.)

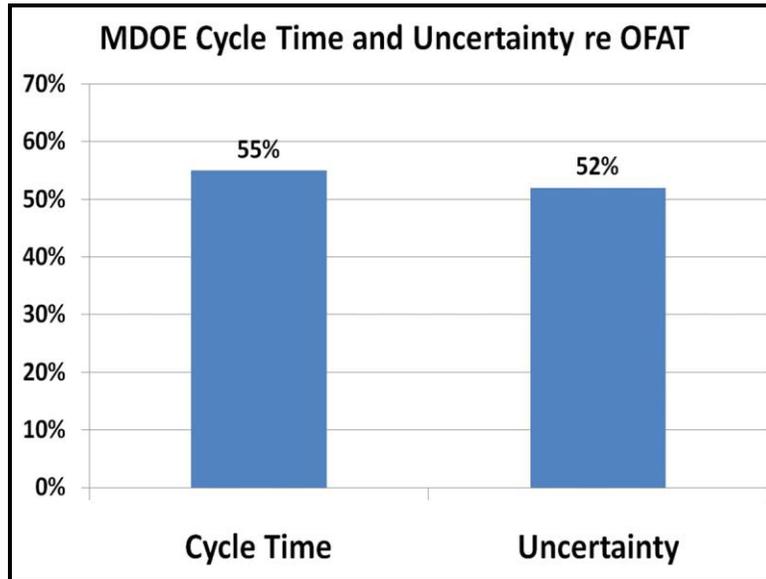


Figure 18. MDOE/OFAT quality and productivity comparison for experiment designs of Fig. 11.

VI. Discussion

It was possible in the examples cited in this paper to achieve certain insights with an analysis of variance that might otherwise go undetected. We discuss a few of those in this section.

The example used to describe two-way analysis of variance in Section IV illustrated how the variance in a sample of experimental error that is NOT induced by changes we intentionally make (the so-called “unexplained” variance in the data) is comprised of both a random component (ordinary chance variations in the data) and a systematic component that is caused by slowly varying effects that can persist over time intervals that are relatively long compared to the time it takes to acquire a polar in a wind tunnel test, say. Partitioning the unexplained variance into random and systematic components is an important exercise for a number of reasons. First, systematic sources of unexplained variance are difficult to detect with replication, which is the most common means of assessing experimental error. Systematic variance might go undetected altogether if the interval of time over which replicates are acquired is too short for the systematic effect to have resulted in a significant change.

Secondly, if replicates are acquired over an interval long enough for systematic variation to have played a role in affecting measurements of system response, such variation can easily be mistaken for a component of ordinary random error. This is problematic because of the distributional characteristics of random and systematic variation. Calculations of such dispersion metrics as the standard deviation are still valid when there are systematic error effects in play, but assumptions such as “two-sigma encompasses 95% of the observations” are only valid if the observations are normally distributed. A significant systematic component of unexplained variance changes the distributional properties of the unexplained variance, adversely affecting such assumptions.

Perhaps the most important reason to be aware of a systematic component of the unexplained variance is that is that systematic variation reduces the independence of experimental errors in a time-series of data. If some effect is in play that tends to gradually and systematically elevate successive measured values, then the error in a given data point is not independent of the error in the prior point. In such a case, if the last measurement is higher than the true value, the next measurement is more likely to be too high than to be too low. Correlated (non-independent)

experimental errors bias the location and dispersion statistics (means and standard deviations) that characterize a sample of data. The result is that these sample statistics are NOT unbiased estimators of the true population parameters we rely upon them to estimate.

For example, assume that a wind tunnel is in a state in which there is initially no systematic unexplained variance, but that at some time during the acquisition of a series of replicates of cruise lift coefficient, systematic error effects come into play that cause successive measurements to be progressively higher. This might be due to instrument drift, temperature changes, flow angularity shifts, or an uncountable number of other possible systematic (not random) changes that occur without the knowledge of the experimenter. In the presence of such systematic variation, the sample mean will be biased high and the sample standard deviation will also be inflated relative to the case in which there is no systematic variation. Experimental errors do not simply cancel in this case because the unexplained variance does not consist of random fluctuations about a mean that represents an unbiased estimate of the true lift. The longer one acquires data when such systematic variation is in play, the more the sample mean will be displaced from the true value, and the greater the standard deviation of the sample will be inflated. That is, under circumstances in which we are dealing with data for which the experimental errors are not independent, the more data we acquire the less reliable will be the result.

Systematic unexplained variance has no doubt contributed significantly to historical difficulties in consistently generating reproducible, high precision results in experimental aeronautics. MDOE experiment designs that minimize data volume requirements and therefore allow time to acquire the replicates that can reveal stealthy systematic errors, combined with the ANOVA tools routinely used in an MDOE analysis to partition the unexplained variance into random and systematic components, have the potential for generating much more realistic uncertainty estimates for wind tunnel testing that are typically reported today.

The implication is not that our tunnels are somehow less capable than we imagined, but that there are natural effects in play that conventional experimental methods do not take into account. These effects can be treated with professional experimental methods and do not represent an especially worrisome threat when such methods are applied. However, they can substantially impact quality and reproducibility when less robust experimental methods are applied, such as the OFAT method commonly used in wind tunnel testing today.

Much of the focus in the design and execution of an MDOE test matrix is on quality assurance tactics designed to assure statistical independence so that reliable results can be achieved even in the presence of systematic error. Such tactics are beyond the scope of the current paper but are treated exhaustively in the literature of experiment design²¹⁻²⁵. Practical applications of such tactics are describe in the references, see for example^{3-6, 20}.

In the one-way analysis of variance used to partition unexplained variance in the lift data acquired in three different tunnels, a component of variance attributable to tunnel differences was discovered that was over six times as large as the variance component attributable to ordinary random error. It is important to recognize that the experimental uncertainty associated with the cross-tunnel variance in this example would exist whether data were acquired in multiple tunnels or not. Acquiring data in multiple tunnels can explicitly reveal additional components of unexplained variance that are responsible for uncertainty in the results obtained in this example. However, if reasons for the cross-tunnel differences cannot be determined and we are left to conclude that the results from any of the tunnels is equally plausible, the variation across these tunnels represents a legitimate source of uncertainty about the true lift coefficient whether we know it is there or not. An analysis of variance involving multiple tunnels has the potential to provide objective insights into a source of uncertainty in experimental aeronautics that often goes unrecognized, and that could explain much of the historic difficulty in reliably reproducing experimental results from one facility to another.

VII. Concluding Remarks

The intent of this paper has been to relieve the reader of any anxiety about coping with experimental data samples in which multiple factors have been changed simultaneously. This is a necessary prerequisite to reaping the productivity benefits that accrue from increasing testing efficiency by working through a test matrix quickly, using multiple factor changes per data point.

The analysis of variance methods used to partition explained variance into components attributable to different independent variables can also be used to partition unexplained variance into random and systematic components. The systematic component can be objectively tested to determine if it is significantly greater than the random component, in which case sequential measurements are not independent and sample statistics such as means and standard deviations are guaranteed to be biased with respect to the corresponding population parameters they purport to represent. Furthermore, these bias errors will depend on local conditions that are not likely to be reproduced precisely in subsequent experiments with the same test article.

It is the author's considered opinion that this mechanism is responsible for much of historic difficulty that has been experienced in generating reproducible, high precision results in experimental aeronautics. The ANOVA methods outlined here in an introductory and tutorial way provide the tools to objectively assess the degree to which sources of systematic unexplained variance are in play in a given test. A general awareness of how common and how large such systematic effects are in even the most carefully executed test, combined with a familiarity with well-established quality assurance tactics designed explicitly to ameliorate the effect of such errors, is expected to result in more reliably reproducible results in all facets of empirical aerospace research where a state of statistical control may prove transient, or elusive altogether.

Acknowledgments

This work was supported by the Aeronautics Test Program Office of the National Aeronautics and Space Administration. The author wishes to acknowledge stimulating discussions on ANOVA-related topics with Dr. Mark Kammeyer and Messrs. Robert Dowgwilllo and Bradley Osbourne of the Boeing Company.

References

- ¹DeLoach, R., "Applications of Modern Experiment Design to Wind Tunnel Testing at NASA Langley Research Center," AIAA 98-0713, 36th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 1998.
- ²DeLoach, R., "Tailoring Wind Tunnel Data Volume Requirements Through the Formal Design Of Experiments," AIAA 98-2884, 20th Advanced Measurement and Ground Testing Conference, Albuquerque, New Mexico, June 1998.
- ³DeLoach, R., "Improved Quality in Aerospace Testing Through the Modern Design of Experiments (Invited)," AIAA 2000-0825, 38th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2000.
- ⁴DeLoach, R., "Tactical Defenses Against Systematic Variation in Wind Tunnel Testing," AIAA 2002-0885, 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 14–17, 2002.
- ⁵DeLoach, R., "A Factorial Data-Rate and Dwell-Time Experiment in the National Transonic Facility," AIAA 2000-0828, 38th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2000.
- ⁶DeLoach, R., Hill, J. S., and Tomek, W. G., "Practical Applications of Blocking and Randomization in a Test in the National Transonic Facility" (invited), AIAA 2001-0167, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.
- ⁷Morelli, E.A., and DeLoach, R., "Response Surface Modeling Using Multivariate Orthogonal Functions" (invited), AIAA 2001-0168, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.
- ⁸Underwood, P., Everhart, J., DeLoach, R., "National Transonic Facility Wall Pressure Calibration Using Modern Design of Experiments" (invited), AIAA 2001-0171, 39th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 2001.
- ⁹Parker, P., and DeLoach, R., "Response Surface Methods for Force Balance Calibration Modeling," 19th International Congress on Instrumentation in Aerospace Simulation Facilities, Cleveland, Ohio, August 2001.
- ¹⁰Danehy, P. M., DeLoach, R., and Cutler, A.D., "Application of Modern Design of Experiments to CARS Thermometry in a Supersonic Combustor," AIAA 2002-2914, 22nd AIAA Aerodynamic Measurement Technology and Ground Testing Conference, St. Louis, Missouri, June 24–26, 2002.
- ¹¹Morelli, E. A., and DeLoach, R., "Ground Testing Results Using Modern Experiment Design and Multivariate Orthogonal Functions (Invited)," AIAA 2003-0653, 41st AIAA Aerospace Sciences Meeting & Exhibit, Reno, Nevada, January 6–9, 2003.
- ¹²Dowgwilllo, R. M., and DeLoach, R., "Using Modern Design of Experiments to Create a Surface Pressure Database From a Low Speed Wind Tunnel Test," AIAA 2004-2200, 24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, Portland, Oregon, June 28–30, 2004.
- ¹³Albertani, R., Stanford, B., DeLoach, R., Hubner, J. P., and Ifju, P. S., "Wind Tunnel Testing and Nonlinear Modeling Applied to Powered Micro Air Vehicles with Flexible Wings," *AIAA Journal of Aircraft*, Summer 2007.
- ¹⁴Erickson, G. E., and DeLoach, R., "Estimation of Supersonic Stage Separation Aerodynamics of Winged-Body Launch Vehicles Using Response Surface Methods," 26th International Council of Aeronautical Sciences, Anchorage, Alaska, Sep 14–19, 2008.
- ¹⁵DeLoach, R., Marlowe, J. M., and Yager, T. J., "Uncertainty Analysis for the Evaluation of a Passive Runway Arresting System," AIAA-2009-1156, 47th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 5–8, 2009.
- ¹⁶DeLoach, R., and Lyle, K. H., "An Airbag Landing Dynamics Experiment Using the Modern Design of Experiments," AIAA-2009-0622, 47th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 5–8, 2009.
- ¹⁷Laplin, R. L., *Probability and Statistics for Modern Engineering*, Prindle, Weber, and Schmidt, Boston, 1983.
- ¹⁸Montgomery, D. C., Runger, G. C., and Hubele, N. F., *Engineering Statistics*, John Wiley and Sons, New York, 1998.
- ¹⁹More, D. D., and McCabe, G.P., *Introduction to the Practice of Statistics*, 3rd Ed., W. H. Freeman and Company, New York, 1999.
- ²⁰DeLoach, R., "Impact of Systematic Unexplained Variance on a Balance Calibration," 25th AIAA Aerodynamic Measurement Technology and Ground Testing Conference, San Francisco, California, June 2006.
- ²¹Fisher, R. A., *The Design of Experiments*, 8th Ed., Oliver and Boyd, Edinburgh, 1966.
- ²²Box, G. E. P., Hunter, W. G., and Hunter, J. S., *Statistics for Experimenters*, 2nd Ed., John Wiley & Sons, New York, 2005.
- ²³Cochran, W. G., and Cox, G. M., *Experimental Designs*, 2nd Ed., Wiley Classics Library Edition, Wiley, New York, 1992.

- ²⁴Montgomery, D. C., *Design and Analysis of Experiments*, 7th Ed., John Wiley & Sons, New York, 2009.
- ²⁵Diamond, W. J., *Practical Experiment Designs for Engineers and Scientists*, 2nd Ed., Wiley, New York, 1989.
- ²⁶DeLoach, R. "Formal Experiment Design as a Tool to Automate Aerospace Ground Testing (Invited)," Tenth Annual Spring Research Conference on Statistics in Industry and Technology, University of Dayton, June 4–6, 2003.
- ²⁷DeLoach, R., "Putting Ten Pounds in a Five-Pound Sack: Configuration Testing with MDOE," 21st AIAA Applied Aerodynamics Conference, Orlando, Florida, June 23–26, 2003.
- ²⁸DeLoach, R., Cler, D., Graham, B., "Fractional Factorial Experiment Designs to Minimize Configuration Changes in Wind Tunnel Testing," AIAA 2002-0746. 40th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 14–17, 2002.
- ²⁹Rhode, M.N., and DeLoach, R., "Hypersonic Wind Tunnel Calibration Using the Modern Design of Experiments," 41st AIAA/ASME/SAE/ASEE Joint Propulsion Conference and Exhibit, Tucson, Arizona, July 10–13, 2005.
- ³⁰DeLoach, R., "The Modern Design of Experiments for Configuration Aerodynamics: A Case Study," AIAA-2006-0923, 44th AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada, January 9–12, 2006.
- ³¹Design Expert, Software Package, Ver. 7.03, StatEase, Inc., Minneapolis, Minnesota, 2006.
- ³²Minitab, Software Package, Ver. 14.2, Minitab, Inc., State College, Pennsylvania, 2003.
- ³³JMP, Software Package, Ver. 6.0.3, SAS Institute, Cary, North Carolina, 2006.
- ³⁴Statistica, Software Package, Ver. 7.1, StatSoft, Inc., Tulsa, Oklahoma, 2006.
- ³⁵MATLAB, Software Package, Ver. 7.0 (R14), The Mathworks, Inc., Natick, Massachusetts, 2004.