Bayesian Revision of Residual Detection Power

Richard DeLoach[1]
NASA Langley Research Center, Hampton, Virginia, 23681

This paper addresses some issues with quality assessment and quality assurance in response surface modeling experiments executed in wind tunnels. The role of data volume on quality assurance for response surface models is reviewed. Specific wind tunnel response surface modeling experiments are considered for which apparent discrepancies exist between fit quality expectations based on implemented quality assurance tactics, and the actual fit quality achieved in those experiments. These discrepancies are resolved by using Bayesian inference to account for certain imperfections in the assessment methodology. Estimates of the fraction of out-of-tolerance model predictions based on traditional frequentist methods are revised to account for uncertainty in the residual assessment process. The number of sites in the design space for which residuals are out of tolerance is seen to exceed the number of sites where the model actually fails to fit the data. A method is presented to estimate how much of the design space in inadequately modeled by low-order polynomial approximations to the true but unknown underlying response function.

I. Introduction

Throughout most of the 20th Century, a key tactical objective of wind tunnel testing has been to acquire high quality data in as great a volume as resource constraints permit. Wind tunnels have been regarded, explicitly or implicitly, as "data factories" designed to achieve this end.

This industrial model of tunnel-as-factory has influenced how quality and productivity are perceived in wind tunnel testing, with concepts common in industrial settings borrowed extensively by the experimental aeronautics community. For example, industrial production quality has been associated with manufacturing process uniformity since Shewhart's early work on this subject[1], and Taguchi later stressed that product uniformity should be independent of production conditions[2]. The concept of "continuous improvement" promoted by W. Edwards Deming and others has likewise had considerable influence on industrial quality and productivity in the US and abroad, notably in Japan[3].

These industrial quality and productivity concepts have been adapted to experimental aeronautics under the tunnel-as-factory model. The result has been that quality and productivity in wind tunnel testing is generally associated with the quality and productivity measures for the perceived product of the test; namely, the data. That is, the quality of a wind tunnel test is generally assessed in terms of the experimental error in the data, and productivity is typically assessed in terms of data volume acquired. Such rate metrics as "polars per hour" and "points per test" have been commonly used to assess productivity with a view to maximizing the production of data, the perceived end-product of the test.

The author has been a critic of this tunnel-as-data-factory model, with its attendant industrial concepts of quality through minimized experimental error, and productivity through high data volume. This criticism originates with a conviction that, contrary to assumptions broadly held in experimental aeronautics, the acquisition of quality data in high volume is not a principal objective of wind tunnel testing. Data acquisition is, of course, a crucial element of the wind tunnel testing process, but it represents an element of the *approach* we take to achieving the true objective. It is not the objective itself.

The objective of a wind tunnel test is to acquire enough new knowledge about the test article that we can adequately predict its future behavior within the design space of the test. The traditional approach to achieving this objective has been to require all test conditions of interest to be physically established in the test, making direct measurements at all of these conditions. In reality, there is seldom nearly enough resources available to do this.

Consider, for example, a relatively modest wind tunnel test featuring only six independent variables; say angle of attack, angle of sideslip, Mach number, Reynolds number, and the deflection of a couple of control surfaces. For

[1] Senior Research Scientist, NASA Langley Research Center, MS 238, 4 Langley Blvd, Hampton, VA 23681, Associate Fellow, AIAA.

computational convenience, let us assume that each variable is to be set at 10 levels. There would then be $10^6 = 1{,}000{,}000$ unique factor combinations, of which there is time in a typical four-week tunnel entry to set, perhaps, 5,000. That means that 99.5% of the design space will go unexplored in this example. (The author has been performing similar calculations in actual wind tunnel tests for decades, and this percentage is representative of how much information is "left on the table" in a conventional wind tunnel test focused only on data collection).

It is not uncommon for experimental aerodynamicists to cite their expertise in prioritizing factor combinations of interest to explain how a successful wind tunnel test can still be executed despite the small percentage of factor combinations there is time to physically set in a typical tunnel entry. Nonetheless, it strains credulity to suggest that absolutely nothing of interest lies in over 99% of the factor combinations that go unexplored, and that there are no surprises lurking in such a huge portion of the total design space. On the contrary, it is not uncommon in the latter stages of an aircraft development program for hindsight to reveal that regions of the design space that were sparsely explored during prior wind tunnel testing are in fact at least as interesting as those where more data were acquired.

A wind tunnel is more properly regarded as a laboratory than as a data factory. We can conduct experiments in this laboratory that yield new knowledge about the test article. Specifically, we can analyze the data acquired with a relatively compact test matrix designed for this purpose, to learn enough about the test article that its behavior can be adequately predicted anywhere within the test design space. This includes all points in the design space where measurements were physical made, and all other points that could not be directly assessed because of constraints on time and money.

The transition from high-volume data collection to aeronautical experimentation is facilitated by exploiting response surface modeling (RSM) methods to fit a sample of data, typically using regression techniques. This process results in mathematical models for the responses measured in a test (e.g., forces, moments, pressures) as a function of the independent variables that define the test matrix, such as angle of attack, angle of sideslip, Mach number, and Reynolds number.

The emergence in recent years of powerful, low-cost data acquisition programs for personal computers makes it quite easy to generate relatively complex regression models. Such a model can be created with a few mouse-clicks after the data are pasted from a spreadsheet, say. As a practical matter, most of the effort in producing a reliable response surface model is not expended in developing the regression model. The effort lies in validating the model.

All of all the information that can be extracted from the data about the quality of a regression model is contained in the residuals—the differences between response model predictions and physical measurements made at the same site in the design space. Other information may be available from sources external to the test, but everything that can be learned about the quality of the result from the test itself is to be found in the residuals. For this reason, rather a lot of care goes into analyzing them in a typical wind tunnel response surface modeling experiment.

A procedure has been proposed for assessing the quality of a response surface model for the case in which residual tolerance levels have been established[4]. By this procedure, the residuals are regarded as a series of Bernoulli trials with binary outcomes of "pass" or "fail." A residual is said to "pass" if it is within tolerance; that is, if its magnitude is smaller than the prescribed level of acceptable tolerance, and it is said to "fail" if its magnitude equals or exceeds this prescribed tolerance level.

The tolerance level is therefore defined as the magnitude of the smallest residual that is too large to be acceptable. For example, if a wind tunnel customer declares his quality specification to be "one half drag count, with 95% confidence," he is declaring that an empirical estimate of the drag coefficient for some prescribed combination of independent variable levels will be satisfactory if it deviates from the "true" drag coefficient by less than half a count.

"Truth" in wind tunnel testing is of course an elusive commodity, but if the drag has been estimated with a response surface model, then a physical measurement of drag at the same site in the design space is commonly regarded as an acceptable surrogate for "truth." In this example, the quality criterion reduces to a simple assertion that residuals of a response surface model must be less than half a drag count for the response model predictions to be regarded as acceptable.

This paper was motivated by some curious results observed in the assessment of certain wind tunnel response surface models. Specifically, response models that generally reproduced pitch polars to the customer's satisfaction and gave other evidence of fitting the data well would nonetheless appear to generate slightly more out-of-tolerance residuals than anticipated. This lead to a reexamination of the RSM quality assessment process, in which Bayesian statistics were invoked to explain the apparent discrepancies. This Bayesian perspective provides some surprising insights into the RSM quality assessment process.

Section II of this paper begins with a review of quality assurance in response surface modeling. Quality assurance is distinguished from quality assessment in that it simply encompasses those procedures designed to achieve a particular end, while quality assessment entails some evaluation of the success in accomplishing that end.

In Section III a particular quality assessment strategy for response surface modeling is outlined. Section IV introduces Bayes' Theorem and the impact of inference error on conclusions based on necessarily imperfect observations. Section V describes the application of Bayesian inference specifically to response surface model quality assessment, citing two case studies. Section VI presents concluding remarks.

II. Quality Assurance in Response Surface Modeling

The fidelity with which a regression model can be made to represent the data is a function of the volume of data employed in the regression. Assuming the response being modeled exhibits certain mild constraints that are common in practical circumstances—that it can be represented by a continuous function with derivatives that exist everywhere, for example—a Taylor series could be fitted, the order of which is limited only by the volume of unique data points used in the regression. An infinite number of unique points would permit all of the coefficients of an infinite Taylor series to be quantified, representing the underlying response exactly. A polynomial approximation to the underlying function of order *n-1* can be constructed from *n* data points which would pass through each data point in the regression, thus fitting all of the data exactly.

Fortunately, it is neither necessary nor even desirable to fit all of the data exactly. This is because the data will inevitably exhibit experimental error, with the higher order terms of a complex regression model serving only to represent the noise in the data. A lower order polynomial generally suffices to represent the ways in which forces, moments, or pressures change systematically with changes in the independent variables (angle of attack, etc.), especially over suitably constrained domains of the independent variables.

It is necessary to acquire a minimum number of data points for each term in such a polynomial to ensure an adequate quality of fit. The fit quality of a response model is defined by a tolerance level quantifying the maximum acceptable difference between a model prediction and a physical measurement made at the same site in the design space. It is also defined by certain minimum levels of confidence, as will now be described.

A. Response Model Quality Specifications

Quality in wind tunnel testing has been associated historically with the data that are acquired in the test. A high quality conventional wind tunnel test requires that there be relatively little variance in the data beyond that which is intentionally induced by planned changes in the independent variables.

By contrast, quality in a formally designed wind tunnel test is associated with low inference error risk. The objective of a designed wind tunnel test is to acquire enough knowledge of the test article that its future behavior can be adequately predicted. That is, we wish to be able to infer responses such as forces, moments, and pressures within some specified tolerance, and to be able to do so for all independent variable combinations of interest throughout the design space of the test, with a specified level of confidence.

The quality of a properly designed experiment depends on the probability that response predictions are within the specified tolerance. Note that this is independent of the quality of the *data*, per se. An experiment that generates high quality data but fails to reliably predict the future behavior of the test article is a failure, no matter how pristine the data. Likewise, an experiment that enables us to consistently forecast to future behavior of the test article is a success, no matter how much unexplained variance there may be in the raw data. As we will demonstrate, any level of random variation in the data itself can be overcome by acquiring a sufficient volume of data.

We follow a common convention by using the term "residual" to designate the difference between a response estimated by a model prediction and one estimated at the same design space site by a physical measurement. There is always some uncertainty in either type of estimate. If we have established a tolerance of δ, say, and the residual at a given site equals or exceeds δ, we will infer that the model is inadequate to predict the response *within tolerance* at that site. Likewise, if the residual is less than δ, we will infer that the model is adequate to predict the response within tolerance. The prediction tolerance therefore represents the magnitude of the smallest unacceptable residual. That is, we require residuals to be smaller than (and not equal to) δ.

Depending on the size of the residuals relative to δ, we will infer that the model either is or is not capable of making predictions at various sites in the design space that are within tolerance, but because of the uncertainty in the predicted response as well as the measured response, either inference may be right or wrong. If we are right in either case, we will have made a proper inference. But there are two possible ways to be wrong. We might erroneously indict the model on the basis of a single-point prediction when the mean of a large number of residuals would have revealed a within-tolerance response prediction. Likewise, we might erroneously validate the model based on a single-point prediction when the mean of a large number of residuals would have exceeded the specified tolerance level. In either case, we will have made an inference error.

We can drive the probability of making either inference error arbitrarily low by acquiring enough data. However, we cannot drive either inference error risk to zero with any finite volume of data. An important element of the modern design of wind tunnel experiments is to specify enough data to ensure that inference error probabilities do not exceed levels prescribed by the experimenter. This inference error risk management is accomplished by *scaling* the experiment; that is, by specifying the minimum volume of data necessary to achieve tolerance and inference error risk specifications.

B. Minimum Data Volume Requirements

Let α represent some acceptable probability of erroneously indicting a residual as too large, and let β represent some acceptable probability of erroneously claiming that a residual is within tolerance. Note that these two probabilities do not have to be equal, and in fact they should reflect the experimenter's assessment of the consequences of committing each error. For example, it might be regarded as a more serious error to validate the model at a site where its prediction is actually out of tolerance than to indict the model at a site where its response prediction is in fact within tolerance. The latter error may result in some added expense and duplication of effort in an unnecessary search for an improved response model when an adequate model was already in hand. While not a desirable outcome, it is probably not as serious as validating a model that makes erroneous response predictions. In such a case one might prefer to specify a value of β that is smaller than α.

To summarize, the quality of a wind tunnel response surface modeling experiment is less dependent on the quality of the measurement data than the quality of response predictions we are able to make after learning enough about the test article to model its future behavior. The quality of response predictions is assessed in terms of the experimenter's tolerance for *prediction error*, δ, and his tolerance for *inference error risk*, α and β.

The researcher specifies acceptable levels of α, β, and δ and then estimates the volume of data required to ensure that response predictions reflect these specifications. The smaller these values are, the more data that will be required. As noted earlier, the minimum volume of data necessary to ensure specified levels of α and β for a given δ will also depend on the quality of the measurement environment as quantified by its unexplained variance, σ^2. Here, σ is the usual standard deviation associated with a sample of genuine replicates. The unexplained variance can be estimated from previous experience in the proposed tunnel, or from replicates acquired for this purpose. We will show that the minimum required data volume is directly proportional to the unexplained variance, which is in harmony with intuitive expectations.

Figure 1 illustrates how tolerances for prediction error, δ, and for inference error risk, α and β, are related to each other through two formal hypotheses used to assess prediction residuals. We first postulate a null hypothesis that there is no difference at a given site in the design space between the response estimated from a model prediction and the response estimated from a direct measurement. That is, the null hypothesis, H_0, states that the true residual is zero. There is a reference distribution associated with this hypothesis that is represented by the normal probability distribution on the left in Fig.1. This distribution has a mean zero and a variance equal to the model prediction variance.

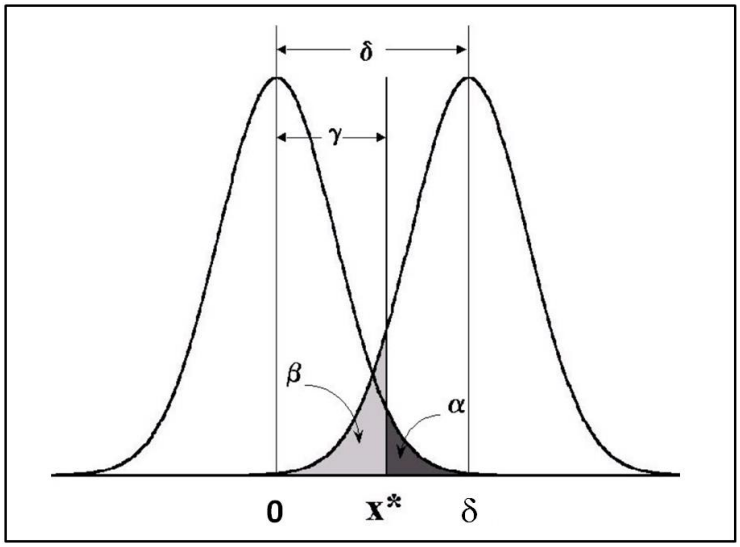

Figure 1. Reference distributions for null hypothesis (left) and alternative hypotheses (right) describing model prediction residuals

Under the null hypothesis that the model predicts perfectly at the design-space site where it is being evaluated, the H_0 reference distribution reflects the dispersion in estimates of the residual that will be induced by ordinary chance variations in the data. Even if the model predicts perfectly and the true residual is zero, experimental error will result in estimates of the residual that can be greater than or less than zero, however, the larger the magnitude of the estimated residual under H_0, the less likely it will be to occur. This is what the H_0 reference distribution on the left of Fig. 1 displays.

The reference distribution on the right in Fig. 2 corresponds to an alternative hypothesis, H_A, that the residual is just out of tolerance; that is, that it has a magnitude of δ. The mean of this distribution is displaced from the mean of the H_0 distribution by δ, and has the same variance. That is, there will be the same model prediction variance due to experimental error whether the predicted response estimate coincides with the measured response estimate or is displaced from it by δ.

Even though Fig. 2 displays two reference distributions in the same figure, only one will describe a given experimental situation. That is, either the model will predict the measurement precisely, or there will be some difference, but both conditions cannot hold simultaneously. We make a formal inference in evaluating the model prediction at a given site in the design space by rejecting either the null hypothesis of its alternative.

The quantity x^* in Fig. 1 is a critical value used as a criterion for deciding which hypothesis to reject. If the residual is large enough to equal or exceed x^*, we will reject the null hypothesis and infer that the model prediction is not within tolerance. Such an inference could be erroneous if ordinary chance variations in the data conspired to generate an artificially large residual. The shaded area under the H_0 reference distribution to the right of x^* has a value of α, and represents the probability that this will happen. That is, α is the probability that we will erroneously reject the null hypothesis and indict a residual as being out of tolerance when it is not.

If the residual is less that x^*, we will reject the alternative hypothesis and infer that the model prediction is within tolerance. The shaded area under the H_A reference distribution to the left of x^* has a value of β, and represents the probability that such an inference will be incorrect because of some unfortunate combination of experimental errors that combine algebraically to understate the residual.

We can estimate data volume requirements with the aid of Fig. 1 by recognizing that the sizes of α and β depend on the width of the reference distributions, which are determined by the volume of data used in the regression. The greater the volume of data, the narrower the reference distributions will be, and thus the smaller α and β will be. The key to a well-designed experiment is to specify enough data to drive α and β to acceptably low levels, but not to acquire substantially more data than this so that direct operating costs and cycle time can be kept to a minimum.

The minimum volume of data required can be estimated easily with the aid of Fig. 1. Note that x^* is located a distance γ to the right of zero. If σ_M is the standard error in model predictions that defines the widths of the H_0 and H_A reference distributions, then we can express this distance as a multiple of σ_M. We will use z_α to designate this multiple. That is, x^* is a distance of $z_\alpha\sigma_M$ to the right of zero. Note that z_α increases as α decreases. That is, the farther x* is to the right, the less likely it is that a residual large enough to fall to the right of it will occur by chance if the null hypothesis is actually true.

We can also express in multiples of σ_M the distance x^* is to the left of the mean of the H_A distribution. If we use z_β to designate this multiple, then x^* is a distance of $z_\beta\sigma_M$ to the left of δ in Fig. 1. Consulting the figure, we therefore have:

$$\delta = z_\alpha\sigma_M + z_\beta\sigma_M \tag{1}$$

It has been shown[5] that the average prediction variance across all points used to fit a polynomial regression model has the same functional form for any order of polynomial and for any number of independent variables fitted by the polynomial:

$$\sigma_M^2 = \left(\frac{p}{n}\right)\sigma^2 \tag{2}$$

where p is the number of terms in the model including the intercept, n is the number of fitted data points, and σ is the ordinary measurement standard error.

Square Eq. (1) and insert Eq. (2):

$$\delta^2 = \left(z_\alpha + z_\beta\right)^2 \sigma_M^2 = \left(\frac{p}{n}\right)\left(z_\alpha + z_\beta\right)^2 \sigma^2 \tag{3}$$

Solving Eq. (3) for n:

$$n = p\left[\left(z_\alpha + z_\beta\right)^2 \frac{\sigma^2}{\delta^2}\right] \tag{4}$$

At least one degree of freedom is required for each term in a regression model, so we must have at least p points to fit a p-term regression model. The bracketed term in Eq. (4) features metrics for data quality, σ, the experimenter's tolerance for prediction error, δ, and inference error risk, α and β. It represents a multiplier of the minimum volume of data, p, that is necessary to achieve specified quality standards in a given measurement environment.

Equation (4) can be simplified by expressing the tolerance, δ, as a multiple of the standard error, σ. One such multiple has an especially attractive physical interpretation. The 95% Least Significant Difference represents the smallest discrepancy between two response estimates that can be resolved with at least 95% confidence. It is related to σ as follows:

$$95\%\ \mathrm{LSD} = \left(2\sqrt{2}\right)\sigma \tag{5}$$

If, by definition, two physical measurements displaced by the 95% LSD cannot be resolved with more that 95% confidence, then a physical measurement and a response model prediction differing by the same amount are indistinguishable in this same sense. Inserting Eq. (5) for δ into Eq. (4) eliminates both σ and δ from the data volume specification, and results in the following formula for minimum data volume as a function of the two inference error risk specifications, α and β:

$$n = p\left[\frac{1}{2}\left(\frac{z_\alpha + z_\beta}{2}\right)^2\right] \tag{6}$$

Assume a well-fitted regression model; that is, a regression model that passes through the "the center of the data" in such a way that fitted data points are distributed above and below model predictions only because of random error, plus possibly some systematic error that is insignificant compared to the random error. If such a model is based on fitting a volume of data specified by Eq. (4), it will be expected to predict responses with a tolerance of δ such that individual residuals exceed x^* with a probability not in excess of α. Likewise, chance variations are not expected to result in a truly out-of-tolerance residual appearing to be within tolerance (smaller than $x^* = z_\alpha \sigma_M$) more than $(1 - \beta)\%$ of the time. We test these assertions during the assessment of the model, by counting the residuals that are within tolerance and those that are out of tolerance.

III. Quality Assessment in Response Surface Modeling

We distinguish between quality *assurance*, and quality *assessment*. Quality *assurance* entails proactive measures undertaken to meet specified quality requirements, such as scaling the experiment so that enough data are acquired for a regression model constructed by fitting those data will have the requisite precision to make adequate predictions. This topic was addressed in the previous section. Quality *assessment* in response surface modeling entails performing various tests to establish whether a candidate model does in fact meet specified quality requirements. We discuss certain aspects of quality assessment in this section.

A candidate response surface model is typically subjected to a battery of tests to detect systematic (not random) departures of model predictions from the data. Such departures, called lack-of-fit errors, can occur when the

underlying response is too complex to be adequately represented by the fitted model, for example. The fact that lack-of-fit errors are systematic rather than random can be exploited to detect them.

A comprehensive discussion of lack-of-fit detection methods is beyond the scope of this paper, but standard texts on response surface modeling cover this subject[5-7]. Lack-of-fit errors that are large compared to random errors suggest that a different model might fit the data better, perhaps one with higher-order terms or other variations that might be indicated by patterns in the residuals. Only when no significant lack-of-fit error is evident in the residuals is the model typically regarded as a candidate for further consideration.

To assess the quality of a candidate response surface model which has been shown to be devoid of significant lack of fit, it would be useful to have an objective process by which to determine whether residuals are within acceptable limits. In the previous section we took some pains to compute the volume of data needed to ensure that residuals would lie within some prescribed error tolerance limit with a specified probability. When the model fits the data well, the probability that a given residual will lie within the error tolerance is $1 - \alpha$.

If we regard physical measurements as surrogates for the true response at various sites within the design space, then assessing the quality of the model seems to reduce to a simple counting exercise. Can we not simply count the number of residuals that are within tolerance, divide by the total number of residuals, and compare with $1 - \alpha$? For example, if we have designed the experiment with $\alpha = 0.05$ and we examine 100 residuals, should we not demand that at least 95 of them be within tolerance as a condition for validating the model?

Unfortunately, residuals are random variables. If α truly is 0.05, it simply means that in a large number of similar experiments, we would expect 95 successes out of 100 trials more often than any other number as long as the model fitted the data well; that is, as long as there was no significant systematic departure of the response model from the data. But we would find in such a series of evaluations that 94 successes occurred almost as often as 95, even when the model is perfectly adequate. It would not be reasonable to reject a response model that predicted response levels within a prescribed tolerance level 94 times out of 100 but validate it if 95 residuals out of 100 were within tolerance. As a practical matter, the tolerance specification could probably be relaxed an insignificant amount to pick up one more residual, permitting the model to be rigorously validated.

The prediction interval test is an example of a binomial process that is characterized by a specified number of trials, an assumed probability of success that is the same for each trial, and an observed number of successes. There is a probability distribution that describes this process, which has in common with other probability distributions a family of Critical Values corresponding to specified probability levels.

For example, let us use Eq. (6) to estimate the volume of data we would need to acquire if our intent was to fit a 4^{th}-order polynomial in four factors well enough that prediction errors would not exceed the 95% Least Significant Difference more than 5% of the time, and that there would be a 99% probability that if the model truly was inadequate at a given site, the residual at that site would be out of tolerance.

There are 70 terms in a full fourth-order polynomial in four independent variables, so $p = 70$, $\alpha = 0.05$ $\beta = 0.01$ in Eq. (6), which specifies a data sample with a minimum of $n = 161$ points. Of these 161 residuals, how many must be within tolerance to validate the model with say, 99% confidence? The critical value of the binomial distribution that is associated with 99% confidence is 146 when there are 161 trials and the probability of success on any one trial is 0.95.

This means that if at each of 161 sites in the design space there actually is a 95% probability that the model will predict responses within the specified tolerance, there is a 99% probability that not more than 161 – 146 = 15 of the residuals will be out of tolerance due to simple chance variations in the data. If there are more residuals out of tolerance than this, we can infer with no more that 1% chance of an inference error that the model discrepancies are attributable to some systematic (not random) effect, the most likely cause of which is some imperfection in the response model that introduces systematic departures from the measured data.

We call "146" in this example the "Critical Binomial Number (CBN)." Critical Binomial Numbers are tabulated in standard statistical references, or they can be computed with readily available software. The CRITBINOM workbook function in Excel returns the Critical Binomial Number, for example. For the model assessment illustrated here, "=CRITBINOM(161,0.95,0.01)" returns the value 146. For any volume of data, the CBN can be used to estimate the minimum number of residuals that must be within tolerance to validate the model at some prescribed level of confidence.

IV. Inference Error in a Response Model Adequacy Assessment and an Introduction to Bayes' Theorem

We assess response model adequacy by examining residuals at various sites throughout the design space. We infer that the model makes an adequate prediction at that site if the residual is within some prescribed tolerance and we infer that the model prediction is inadequate if it is not.

We define a model prediction at a specified cite in the design space to be *adequate* if any difference between it and a physical measurement at the same site is small enough to be attributed to ordinary random experimental error. If the difference is large enough that it is sufficiently unlikely for it to be due to random error only, we assume that there is some systematic lack-of-fit error in play attributable to inadequacy of the response model at that site.

It is possible, and in fact rather to be expected, that absent a level of fitting perfection unlikely to be achieved with a simple low-order graduating function, even a "good" regression model that fits the data adequately almost everywhere in the design space will display systematic error in some subset of it. We seek a model that makes predictions within tolerance in an acceptably high percentage of the design space (say, 95%). However, we recognize that there is some probability of an inadequate prediction when a low-order polynomial serves as a surrogate for the unknown true underlying response function.

From examining the residual at a given site, we must infer whether the model prediction is adequate or not. It is possible for either inference to be in error, simply due to random experimental error. The model may be perfectly adequate at a given site, but the residual observed there could be artificially large because of some unlikely measurement error. Likewise, a within-tolerance residual might provide a false sense of security because of an equally unlikely experimental error that deflates the residual. Absent an infallible test of model adequacy, it is possible to make an inference error about the adequacy of a response model whether we validate it or impeach it. More to the point, there is always some non-zero probability that such an inference error will occur. It is therefore necessary to consider not only the fact that the response model may or may not be adequate, but that our assessment of it can be right or wrong in either case. Fortunately, there is a mechanism available to incorporate this uncertainty in the assessment methodology into the overall assessment of the model's adequacy. We can use Bayes' Theorem for this purpose.

We will provide an overview of Bayes' Theorem before delving into its application to response model adequacy assessment, but we first offer a few additional remarks to motivate what follows. It is incorrect to regard Bayesian inference, the process we will demonstrate here, as an optional refinement of an otherwise adequate assessment that can be achieved with conventional frequentist thinking about how often events occur in nature. Unless we explicitly take into account the imperfections in our assessment methodology, as well as the imperfections they are designed to assess, our inference is virtually guaranteed to be wrong. It is simply a matter of degree, but in an alarming number of cases the conclusions that are reached absent such considerations can be very wrong indeed.

Bayesian inference is appealing for scientific research because it explicitly accounts for the distinction between what is true in nature, and what is merely inferred to be true on the basis of necessarily imperfect assessments. When we use Bayesian inference to evaluate the probability that "A" is true given that we have observed "B," we not only consider the probability of observing "B" when "A" is true, but also the probability of observing "B" when "A" is *not* true. In our application of Bayesian inference, we will not only take into account the probability that a residual will be out of tolerance when the response model actually is inadequate at a given site in the design space, but also the probability that a residual will be out of tolerance even when the response model actually is adequate.

The following section provides a concise review of Bayes' Theorem, and how it is a natural consequence of ordinary conditional probability concepts familiar to all. This will be followed by a demonstration of the basic idea of Bayesian inference, and then a discussion of its relevance in response surface model adequacy assessments.

A. Review of Bayes' Theorem

Bayes' Theorem is easily derived from the definition of conditional probability, as expressed in terms of a joint probability. The joint probability of events A and B is represented as $P(A \cap B)$ and defined as the probability that events A and B both occur. We define the prior probability of event "B," written as P(B), as the probability that "B" will occur independent of whether event "A" occurs. We can now express both the conditional probability of "A" given "B," and the conditional probability of "B" given "A" in terms of their joint and prior probabilities, as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \tag{7}$$

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \tag{8}$$

From Eqs. (7) and (8) we have

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{9}$$

which is the well-known product rule for probabilities. It leads directly to Bayes' Theorem:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{10}$$

The quantity P(B|A)/P(B) is often described as the normalized likelihood function. We say, then, that the conditional probability of "A" given "B" is just P(A) times this normalized likelihood function, where P(A), called the "prior probability of A", is simply our best estimate of the probability that "A" is true, absent any empirical evidence, "B." The importance of the likelihood function in a Bayesian analysis derives from the fact that it depends on "B", and thus represents the mechanism by which the prior probability of "A" is modified by empirical evidence represented by "B." An illustration will clarify these points.

B. Illustration of Bayes' Theorem: Major League Baseball Steroid Scandal

The brief derivation of Bayes' Theorem presented above is taken from a survey paper on potential applications of Bayesian inference in aerospace research which the author published in 2008[10]. The following motivational illustration is also drawn from that source. It serves as a near perfect analogy of the application of Bayesian inference to response surface modeling assessment, which will be treated after this example is presented as a foundation.

Consider how Bayesian inference could have altered key conclusions in an investigation of steroid use in major league baseball in 2007. This investigation was accompanied by significant publicity, and amounted to a severe indictment of the sport and many of its key players.

At the time of the scandal, 12 players had experienced the league's 10-day suspension rule for first-time substance abuse offenses, a number nominally consistent with the Commissioner's assertion that 1.2% of the players had used steroids during the previous season, based on a combined American League and National League roster of 854 players.

We will augment this meager information with what are believed to be reasonable suppositions to illustrate how a Bayesian analysis might be performed on the steroids-in-baseball scandal. (The central points of this illustration are not dependent on the suppositions.)

Let us assume that a steroid test is "95% accurate," by which it is meant that if a player who uses steroids is subjected to this test, he will test positive 95% of the time. This means that 5% of steroid users will pass this test. Let us also assume that the test will exonerate non-users 95% of the time, which nonetheless means that 5% of those who do not use steroids will be falsely accused.

Under these circumstances, one might be tempted to conclude that if a player fails such a test, there is a 95% probability that he does use steroids. However, this is not the case. We will use Bayes' Theorem to show that if a player does fail this test, the probability that he is a steroid user is actually much less that 95%.

Let "A" represent a baseball player who uses steroids and let "B" represent the case of a failed drug test. We are interested in the conditional probability that a player actually is guilty, given that he has failed the test. That is, we are interested in P(A|B). We can compute this using Bayes' Theorem.

We will use for this illustration the figure cited by the Commissioner of 1.2% to represent the fraction of players who actually used steroids prior to any testing. Let us say also that prior to any testing, this was also our estimate of the probability that a randomly selected player would be a steroid user. We can use this information to compute how many users *and non-users* will test positive.

If we assume that 0.012 of the league uses steroids and 0.95 of those will test positive, then $0.012 \times 0.95 = 0.0114$ of the league will be users who test positive. On the other hand, $1 - 0.012 = 0.988$ is the fraction of the league that is clean in this example, of which 0.05 will wrongly test positive. So $0.988 \times 0.05 = 0.0494$ is the fraction of the league that will be falsely accused. We will have $0.0114 + 0.0494 = 0.0608$ as the total fraction of the league that will test positive (roughly 52 players), of which only 0.0114×854, or about 10, will actually be steroid users. Therefore, P(A|B), the probability that a player actually does use steroids given that he has tested positive, is only about 10 in 54, or roughly 18.5%!

The probabilities of false negatives and false positives were estimated in this example for the sake of illustration, although they are believed to be not unrealistic. The total number of positive tests computed in this example—52—is consistent with numbers reported in the media at the time of the scandal for active players said to have tested positive for steroid use.

The calculations presented in this example were rounded to ensure an integer number in each category of accused players (correctly accused and falsely accused). We can more rigorously estimate P(A|B) by invoking Bayes' Theorem, after first making a slight modification to the way it is expressed in Eq. (10). For computational convenience we re-cast P(B) in Eq. (10) by first noting that

$$P(B) = P(A \cap B) + P(\bar{A} \cap B) \tag{11}$$

where the bar over "A" implies "not A". Eq. (11) simply states that the probability of "B" independent of "A" is the probability that "B" occurs when "A" occurs plus the probability that "B" occurs and "A" does not occur. From Eq. (8) with obvious extensions to the "not A" case, Eq. (11) becomes

$$P(B) = P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A}) \tag{12}$$

and Bayes' Theorem as expressed in Eq. (10) becomes

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})} \tag{13}$$

Absent any testing and subsequent application of Bayesian inference, our best estimate of the probability that a randomly selected player will be a steroid user is the prior probability, P(A) = 0.012 introduced earlier. We wish to revise this probability by testing each player and incorporating those test results into our assessment. We will do this by inserting known values into Eq. (13) to compute a revised probability that a player uses steroids if he fails the steroid test.

The probability that a player would test positive for steroids given that he is an actual user is $P(B \mid A) = 0.95$. The probability that a player would test positive for steroids given that he is *not* a user, $P(B \mid \bar{A})$, is 0.05. The probability that a randomly selected player is clean is

$$P(\bar{A}) = 1 - P(A) = 1 - 0.012 = 0.988 \tag{14}$$

Inserting these numbers into Eq. (13) yields the following result:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \bar{A})P(\bar{A})} = \frac{0.95 \times 0.012}{(0.95 \times 0.012) + (0.05 \times 0.988)} = 0.1875 \tag{15}$$

This compares favorably with the approximate solution of 18.5% obtained earlier by rounding numbers of players to integers.

These results illustrate how much difference there can be between what is true and what is inferred in a test. The empirical evidence that a player failed the steroid test would cause the probability of his being a user to increase from the prior probability of 1.2%, estimated in the absence of any empirical evidence, to 18.75% when test results were properly taken into account, an increase of a factor of more than 15. Nonetheless, it is surprising that there is only an 18.75% probability that a player who tests positive with this test will actually be a steroid user, given the "95% reliability" of the test.

This apparent discrepancy is resolved by recognizing that we wish to know the probability that a player uses steroids given he has failed the drug test, P(A|B), but the "95%" figure actually represents the reverse of this, P(B|A). It is the probability that a player will fail the drug test given that he uses steroids. The erroneous assumption that P(A|B) = P(B|A) occurs often enough to have a name. It is known in forensic science as the Prosecutor's

Fallacy, and occurs when inference error probabilities are not explicitly taken into account as they are when Bayesian inference is applied.

The same kind of erroneous inferences can be made about the quality of a response model if we assess it only by counting out-of-tolerance residuals. To make a proper assessment, we must also explicitly account for inference error probabilities. Unfortunately, we are often in error when we make the common and seemingly reasonable assumption that a response model predicts adequately at (and only at) at design space sites where residuals are within tolerance. As we shall see, to make such an assumption is to be guilty of the Prosecutor's Fallacy.

V. Bayesian Inference in a Response Model Quality Assessment

Let us now apply Bayes' Theorem to the problem of response surface model quality assessment. The baseball steroid example presented in the previous section provides a roadmap for doing so.

A. Two Examples of Response Model Quality Assessment

The possibility that Bayesian inference could be applied in response model quality assessment was motivated in part by recent experiences in performing a Critical Binomial Number assessment on the results of a particular wind tunnel response surface modeling experiment. Figure 2 displays a set of residuals associated with a pitching moment response model developed from data acquired in this test. These residuals were estimated at 41 randomly selected sites within the design space. Measurements made at these sites were used exclusively to assess the quality of the response model, and were not used to generate the response model, which was based on an independent sample of data acquired in the same test. The red lines represent tolerance levels specified by the principal investigator, which corresponded to the upper and lower limits of a 95% prediction interval estimated from error degrees of freedom in an analysis of variance performed on the test data.

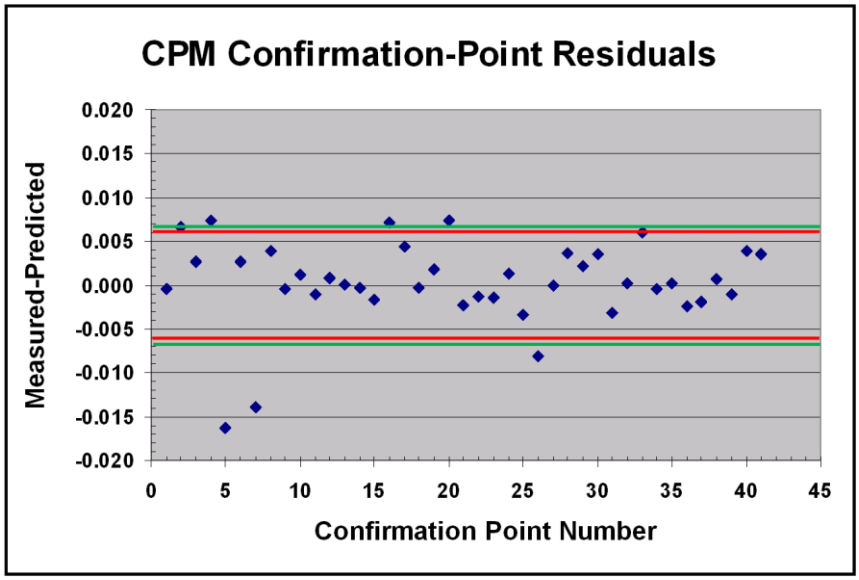

Figure 2. Confirmation-point residuals for wind-tunnel pitching moment response model. Red line: "Experimental Error" tolerance level. Green line: Critical Binomial Number limits.

The Critical Binomial Number for these 41 confirmation-point residuals was 35, assuming as a success criterion a 95% probability that a given residual was within tolerance, and that a 99% confidence level for the model assessment. Eight of the residuals in Fig. 2 are outside the tolerance limits, so the 33 successes achieved in 41 trials was insufficient to declare the model adequate in this test.

The green lines in Fig. 2 represent adjusted tolerance limits for which the critical binomial number test would have been successful and the model declared adequate, as these lines do encompass the Critical Binomial Number of 35 residuals. Moving the tolerance limits from the red lines to the green lines represents a very small adjustment. This, plus the fact that the pattern of residuals is generally featureless, with residuals distributed more or less randomly about a mean of zero, suggests that the response model is good, but apparently not quite good enough to pass the Critical Binomial Number test.

There were no other indications that the model had poor fit characteristics, and the general character of measured pitching moment polar plots was well represented by the response model. The difficulty this model had in passing the Critical Binomial Number test was therefore unanticipated.

A similar result was recently brought to the author's attention by a student at the Instituto Superior Técnico in Portugal[11], who had tried to assess a response model—coincidentally also for pitching moment—by comparing the number of sites in his design space where residuals were out of tolerance with a Critical Binomial Number representing the maximum number of sites expected to display such out-of-tolerance residuals given the volume of data he had specified for the regression. He had properly applied the quality assurance methods described above to scale the experiment for specified inference error risks associated with a prescribed tolerance for prediction error. His procedures were based on methods documented in Ref. 8, in support of a Master's thesis titled "Aircraft Wind Tunnel Characterization using Modern Design of Experiments." Notwithstanding his careful quality assurance tactics, when he used a fitted model to predict pitching moment at 65 verification points acquired independently of the data used to fit the model, thirteen of the 65 points were out of tolerance. The CBN was 59, which would have only allowed for 6 residuals to be out of tolerance.

These two examples raise the same question. In the first example, residuals acquired at eight randomly selected sites in the design space were clearly out of tolerance. That is, residuals at eight sites failed the test for fitting adequacy by being larger than the prescribed fitting tolerance, δ. Likewise, in the second example residuals at thirteen sites also failed the test for fitting adequacy because residuals estimated there were out of tolerance.

The question remains: Is the model in fact inadequate at all eight sites in the first example and at all thirteen sites in the second? Is it not possible that in one or both cases, assessment errors have resulted in an inflated number of sites with an apparently inadequate fit just as assessment errors in the baseball example overstated the number of indicted players?

B. Extension of Bayesian Inference to Response Model Quality Assessment

The CBN analyses noted here have elements in common with the major league baseball example described above. For example, for a response model giving every other indication of a good fit, most of the residuals should actually be within tolerance with a relative few "guilty" (out of tolerance) residuals, just as most of the players were clean in the baseball example. However, even if an assessment mechanism has a relatively low probability of falsely identifying within-tolerance residuals as out of tolerance, the absolute number so impeached can be relatively large if enough sites in the design space are tested.

This was the reason for the unexpected results in the baseball example. Because of the large roster of players and relatively small number of steroid users, there was a sizable pool of innocent players subjected to inevitable inference errors in the testing. Even with a relatively low inference error probability, the absolute number of falsely accused players could be relatively large because so many innocent players were tested.

Compounding matters, the absolute number of guilty players identified in the steroid testing would have been relatively small if the total number of guilty players in the league were small. This would be true even if the test was effective enough to properly identify a relatively large fraction of guilty players. The result, as the calculations above illustrate, would have been a surprising percentage of accused players who in fact were innocent.

Just as the steroid test was imperfect, so is the use of residuals to identify design space sites where the model is inadequate. Residuals estimated at each site represent a single instance of the difference between two random variables, each of which is drawn from a universe of possible values distributed about some unknown true value. The two random values are each empirical estimates of some response such as pitching moment, one comprised of a single-point measurement and one comprised of a model prediction based on some finite number of individual measurements. Both response estimates are necessarily imperfect because of experimental error, and the difference between them is therefore an imperfect indicator of how well the response model represents the data.

For these reasons there is always some non-zero probability, however low, that an out of tolerance residual does not indicate a site for which the model is inadequate. The reverse is also true; the response model does not necessarily make adequate predictions at every site for which a residual is within tolerance. However, if the number of sites where the model is adequate is large compared to the number of sites where it is not, the absolute number of falsely indicted sites will be large compared to the number of sites that are unjustifiably validated. The result will be a net surplus of falsely indicted design space sites.

To see how this works in the case of a response model adequacy assessment, let ε represent the (unknown) fraction of the design space where the model truly is inadequate; that is, where true residuals would in fact be out of tolerance. Let α and β represent the inference error probabilities illustrated in Fig. 1. That is, let α represent the fraction of the design space where the model is truly adequate but experimental error results in an overstated residual and an improper inference that the model is inadequate. Let β represent the fraction of the design space

where the model is actually inadequate, but where experimental error results in residuals that are nonetheless within tolerance, resulting in improper inferences that the model has made adequate response predictions at those sites.

It is a common assumption that models are inadequate where residuals are out of tolerance and adequate where they are not. However, this will only be reliably true when inference errors (α and β) are zero. We can use Bayes' Theorem to estimate the probability that a model actually is inadequate at a site for which the residual is out of tolerance, given non-zero values for α and β.

Let "A" represent some design site for which the model is truly inadequate, and let "B" represent a residual that is out of tolerance. We wish to know P(A|B), the probability that the model predicts inadequately at a site, given that the residual there is out of tolerance.

We have defined ε to be the unknown fraction of the design space where the model truly is inadequate. That is,

$$P(A) = \varepsilon \tag{16}$$

Since β is the probability that a residual will be within tolerance at a site for which the model is actually inadequate, the quantity $1 - \beta$ is the probability that a residual will be out of tolerance given that the model is inadequate. That is,

$$P(B \mid A) = 1 - \beta \tag{17}$$

We have defined α to be the probability that a residual will be out of tolerance given that the model is adequate, so

$$P(B \mid \bar{A}) = \alpha \tag{18}$$

Since ε represents the (unknown) fraction of the design space where the model is inadequate, the probability the model is adequate at a given site is as follows, from Eq. (16):

$$P(\bar{A}) = 1 - P(A) = 1 - \varepsilon \tag{19}$$

Inserting Eqs. (16-19) into Eq.(13) yields the following result:

$$P(A \mid B) = \frac{(1-\beta)\varepsilon}{(1-\beta)\varepsilon + \alpha(1-\varepsilon)} \tag{20}$$

Note the following:

- $(1 - \beta)\varepsilon$ = the probability that a residual is out of tolerance when the model is inadequate
- $\alpha(1 - \varepsilon)$ = the probability that a residual is out of tolerance when the model is adequate

The denominator of Eq. (20), comprised of the sum of these two probabilities, is just the probability that the residual at a given site is out of tolerance whether the model is inadequate there or not. That is, the residual can be due to lack of fit in the model, or to chance variations in the data.

$$\textit{Probability that residual is out of tolerance} = (1-\beta)\varepsilon + \alpha(1-\varepsilon) \tag{21}$$

Eq. (21) represents the probability that a residual will be out of tolerance *whether the model predicts adequately at that site or not*. So Eq. (20) is

$$P(A \mid B) = \frac{\textit{sites where model does not fit}}{\textit{sites where residual out of tolerance}} \tag{22}$$

Eq. (20) demonstrates that it is only when $\alpha = 0$ that P(A|B) = 1. That is, an out of tolerance residual at some site in the design space does not imply with certainty that the model does not predict adequately at that site, as long as there is some probability that chance variations in the data are responsible for the residual being out of tolerance.

Note also from Eq. (10) that the Prosecutor's Fallacy cited earlier, that P(A|B) = P(B|A), only holds for the theoretical case in which P(A) = P(B); that is, when the probability that a residual is out of tolerance at a given site is the probability that the response model predicts inadequately at that site. As long as there is some probability, α, that a residual can be out of tolerance at a given site even when there is no lack of fit error there, the Prosecutor's Fallacy will remain a fallacy.

Further insights can be gleaned from Eq. (21) by expanding the denominator:

$$P(A|B) = \frac{(1-\beta)\varepsilon}{(1-\beta)\varepsilon + \alpha(1-\varepsilon)} = \frac{(1-\beta)\varepsilon}{\varepsilon - \beta\varepsilon + \alpha - \alpha\varepsilon} = \frac{(1-\beta)\varepsilon}{(\varepsilon+\alpha) - \varepsilon(\alpha+\beta)} \tag{23}$$

Assume a reasonably well-fitted model for which a sufficient volume of data has been fitted to ensure that residual assessment inference errors are small. That is, assume ε, α, and β are all small. Then the second term in the denominator is negligible compared to the first. For example, assume that the model actually predicts responses within tolerance at 95% of the design-space sites, so that $\varepsilon = 0.05$. Assume also that a sufficient volume of data has been acquired per Eq. (4) so that $\alpha = 0.05$and $\beta = 0.01$. Then $\varepsilon + \alpha = 0.05 + 0.05 = 0.10$ in Eq. (23), and $\varepsilon(\alpha + \beta) = 0.05(0.05 + 0.01) = 0.003 = 0.03(\varepsilon + \alpha)$. Since we also have that $\beta << 1$, we can rewrite Eq. (23) to an excellent approximation as

$$P(A|B) \approx \frac{\varepsilon}{\varepsilon + \alpha} \tag{24}$$

Eq. (24) makes it clear that a residual can be out of tolerance two ways, while the model can only be inadequate one way. The model is inadequate to predict the response at a given site only if there is a statistically significant lack of fit error at that site. However, the residual can be out of tolerance due to lack of fit (probability ε), or because of ordinary chance variations in the data (probability α). We do not want to indict the model simply because chance variations generated an out-of-tolerance residual. Eq. (21) and in its simpler form, Eq. (24) indicate what fraction of the out-of-tolerance residuals actually do imply a poorly fitting response model. For the numerical example cited above in which $\alpha = \varepsilon = 0.05$, only about half of the residuals are expected to be out of tolerance due to lack of fit. The other half were acquired at sites where the model predicts adequately, but chance variations in the data caused the residual to be out of tolerance.

C. Revisiting the Critical Binomial Number Analyses

It is now clear why the Critical Binomial Number test is so difficult to pass. If we define a successful Bernoulli trial as one in which the residual lies within upper and lower bounds of a 95% prediction interval, for example, and then assign a per-trial success probability of 0.95 when we estimate the Critical Binomial Number, we are implicitly assuming that in 100% of the design space the model fits the data so well that all out-of-tolerance residuals are due to nothing more than chance variations in the data. If this is in fact the only mechanism in play, then indeed the CBN would represent the minimum number of successes (residuals within the 95% prediction interval) that we would expect for a given total number of trials. However, if systematic (not random) lack of fit error is in play, this would tend to increase the number of Bernoulli failures. Equivalently, it would reduce the per-trial success probability to something less than 95%.

An ideal regression model would feature enough higher-order terms to ensure that systematic lack of fit errors are negligible. Considerable effort is in fact expended to ensure this during the model construction. However, the response models are typically just low-order polynomial approximations over a limited range of the independent variables for some complex underlying function. Some lack of fit, however small, is inevitable because of all the higher-order terms in an infinite polynomial series representation of the underlying function that must be neglected due to resource constraints, if nothing else.

Rather than using a CBN analysis to determine whether the model under evaluation experiences any is lack of fit, it may be more sensible to simply stipulate that it inevitably does at some level, and then estimate in what fraction of the design space such lack of fit errors prevent the model from adequately predicting responses. Realistically, that fraction can never be zero for any imperfect response model approximation, but it may be small

enough to be acceptable. As a practical matter, few aerodynamicists are likely to reject as inadequate a model that can predict responses within tolerance in, say, 95% of the entire design space.

Consider the example cited in discussions surrounding Fig. 2. In this example, Eight residuals were out of tolerance out of a total of 41 randomly selected design-space sites were residuals were estimated. We assume a Critical Binomial Number of 41 – 8 = 33 as ask what the minimum per-trial success probability would have to be to achieve 33 successes in 41 trials, assuming we wish to be 99% confident of our estimate.

There are various ways to perform this calculation but perhaps the easiest is to invoke the CRITBINOM workbook function in Excel, where we find that CRITBINOM(41, 0.918, 0.01) = 33. This indicates that the associated success probability is 91.8%. Call this probability of success, p_s. This means that $1 - p_s$ is the probability that a residual is out of tolerance, so by Eq. (21) we have

$$1 - p_s = (1-\beta)\varepsilon + \alpha(1-\varepsilon) \tag{25}$$

Solving Eq. (25) for ε,

$$\varepsilon = \frac{(1-\alpha) - p_s}{(1-\alpha) - \beta} \tag{26}$$

Continuing with the current example cited in discussions surrounding Fig. 2, we have $p_s = 0.918$, $\alpha = 0.05$, and $\beta = 0.01$. Inserting these into Eq. (26):

$$\varepsilon = \frac{(1-\alpha) - p_s}{(1-\alpha) - \beta} = \frac{(1-0.05) - 0.918}{(1-0.05) - 0.01} = 0.034 \tag{27}$$

Eq. (27) indicates that in this example, the model will predict poorly with a probability of only 0.034, which is within the prescribed α specification of 0.050. Putting it another way, the model is expected to predict responses without significant systematic lack of fit errors in 96.6% of the design space.

We can use the results of Eq. (27) to estimate P(A|B) with Eq. (20):

$$P(A|B) = \frac{(1-\beta)\varepsilon}{(1-\beta)\varepsilon + \alpha(1-\varepsilon)} = \frac{(1-0.01)0.034}{(1-0.01)0.034 + 0.05(1-0.034)} = 0.411 \tag{28}$$

Recall that "A" signifies a design space site where the residual is out of tolerance because of model imperfections that result in genuine lack of fit at that site. Recall also that "B" signifies a residual that is out of tolerance. If we were to indulge in the Prosecutor's Fallacy, denying the possibility of inference errors in assessing whether or not an out-of-tolerance residual implies an inadequate model prediction, we would assume that an out-of-tolerance residual automatically implies that the model fits the data poorly. In the current example, this would suggest that the model is inadequate at 8 sites out of 41 tested. However, Eq. (28) indicates that given the eight residuals observed to be out of tolerance, only $0.411 \times 8 = 3.3 = 3\pm1$ of these cases are due to systematic error in the model. The remaining 5±1 out-of-tolerance residuals can be attributed to ordinary chance variations in the data.

Let us exclude the three residuals in this example that are attributed to systematic lack of fit error and ask if the remaining 5 residuals are consistent with what we would expect for the 95% prediction interval tolerance we are claiming. That is, we ask what the CBN would be for 41-3 = 38 trials for which the success probability is 0.95. Again using Excel, we find that CRITBINOM(38,0.95,0.01) = 32, implying that up to 38-32 = 6 Bernoulli trials out of 38 could fail even if the per-trial success probability is 95%. This compares with the 5±1 forecasted to be out of tolerance due to random error after we use Eq. (28) to estimate that 3±1 of the 8 total residuals observed to be out of tolerance are due to systematic error. That is, we conclude that the model satisfies precision requirements associated with a 95% prediction interval in regions of the design space where systematic error is not an issue, which we estimate to be 96.6% of the design space per Eq. (27).

We can perform a similar analysis on the second test case described above. In that example, a response model was constructed from a sample of data scaled for inference error probabilities of $\alpha = 0.05$ and $\beta = 0.01$. It was tested

with a 95% prediction interval tolerance and subjected to a Critical Binomial Number analysis featuring 65 trials, of which 13 were out of tolerance.

Using the CRITBINOM function in Excel, we compute the Critical Binomial Number for this case as CRITBINOM(65,0.95,0.01) = 59, meaning that no more than 65 - 59 = 6 residuals should have been out of tolerance, about half as many as were actually observed. There is only a 1% chance of error if we conclude from this test that the probability of success in individual trials is not 95% as assumed.

This suggests that some systematic error associated with the model is inflating the prediction error we would otherwise expect if all residuals were due exclusively to ordinary random error. The question is whether all 13 out-of-tolerance residuals were the result of an inadequate model, or whether some of them could be out of tolerance because of inference errors in assessing the residuals.

We have already established that to regard every out-of-tolerance residual as evidence of an imperfection in the response model is to be guilty of the Prosecutor's Fallacy. Some of these residuals can be out of tolerance due to chance variations in the data that do not reflect negatively on the model.

We begin as before, by estimating the per-trial success probability in a series of Bernoulli trials that is expected to result in an observed number of failed tests out of a known number of trials. In this case there were 13 failures in 65 trials, and by iterative calculations using Excel's CRITBINOM function we determine that CRITBINOM(65,0.900, 0.01) = 13. We infer therefore that the per trial probability of success, p_s, is 0.900, not 0.950 as we would have assumed if only random errors were in play. We insert this value of p_s, plus known values for α and β, into Eq. (26) to compute ε, the fraction of the design space for which the observation of an out-of-tolerance residual is expected to signify true lack of fit.

$$\varepsilon = \frac{(1-\alpha) - p_s}{(1-\alpha) - \beta} = \frac{(1-0.05) - 0.900}{(1-0.05) - 0.010} = 0.053 \tag{29}$$

Eq. (29) indicates that in this example, the model will predict poorly with a probability of 0.053, which is on the order of the prescribed α specification of 0.050. That is, the model is expected to predict responses without significant systematic lack of fit errors in 94.7% of the design space, or nominally for 95% of all model predictions.

We can use the results of Eq. (29) to estimate P(A|B) with Eq. (20) as before:

$$P(A \mid B) = \frac{(1-\beta)\varepsilon}{(1-\beta)\varepsilon + \alpha(1-\varepsilon)} = \frac{(1-0.01)0.053}{(1-0.01)0.053 + 0.05(1-0.053)} = 0.527 \tag{30}$$

From Eq. (22) above we interpret this result to mean that a little over half of the out-of-tolerance residuals reflect imperfections in the response model. The rest can be attributed to inference errors associated with the uncertainty in declaring a residual to be within tolerance or out of tolerance. For the example we are currently considering, Eq. (30) suggests that only 52.7% of the 13 out-of-tolerance residuals are due to model imperfections, or roughly 7±1. The remaining 6±1 out-of-tolerance residuals can be attributed to random error.

We exclude the 7 residuals due to genuine model imperfections from further CBN analysis to examine what the CBN would be for trials that are expected to feature only random error. That is, we ask what the CBN would be for 65-7 = 58 trials for which the success probability is expected to be 0.95. Again, using Excel we find that CRITBINOM(58,0.95,0.01) = 51, implying that up to 58-51 = 7 Bernoulli trials out of 58 could fail even if the per-trial success probability was 95%. This compares with the 6±1 forecasted by Eq. (30) to be out of tolerance due to random error.

The net assessment in this example is that in 65 trials we would expect 6 or 7 residuals to be out of tolerance due to chance variations in the data (ordinary random error), and about the same to be out of tolerance due to systematic lack of fit error, for a total of 12 – 14 residuals out of tolerance. This compares with the 13 residuals actually observed to be out of tolerance in the 65-trial test that was conducted. We anticipate that the model will predict responses with a precision consistent with a 95% prediction interval throughout roughly 95% of the design space.

VI. Concluding Remarks

This paper has focused on Bayesian inferences that affect the probability that an out-of-tolerance residual reliably implies a poorly fitting model at the site where the residual is observed. Such inferences are influenced by both by the probability of erroneously declaring a residual to be out of tolerance, and by the probability of erroneously declaring a residual to be within tolerance. Both probabilities can be driven below specified limits by

scaling the regression experiment appropriately; that is, by acquiring a sufficient volume of data. The relationship between inference error probability and data volume was reviewed.

Contrary to commonly held suppositions, the number of sites where a response surface model fits poorly is different from the number of sites where the residual is out of tolerance. In general, the number of design-space sites where the model actually does fit poorly is less than the number of sites where residuals are out of tolerance. That is, it is possible for a residual to be out of tolerance even at a site where the model fits the data adequately. This non-intuitive condition arises from the fact that there is always some uncertainty in assessing whether a given residual is within tolerance or not. In commonly occurring circumstances, a response surface model might actually fit the data adequately in over half of the design space sites for which residuals are out of tolerance.

We have also outlined a procedure for estimating the fraction of the design space for which residuals are expected to be too large to attribute to simple random error. At a given site in the design space there is bound to be some non-zero probability that a low-order polynomial approximation to the true underlying function will fail to predict the true response within some specified tolerance. The comparison of model predictions with direct measurements is subject to an extra dollop of uncertainty because such residuals reflect imperfections in the response model as well as ordinary measurement imperfections.

The question is not whether a prediction by a necessarily imperfect response model will mirror a necessarily imperfect measurement at the same site in the design space, but how often the difference between the two will be too great to attribute to simple random error. That is, we ask how often (or for what percentage of the design space) the model truly fails to fit the data. The adequacy of the model then depends on whether this fraction is low enough to be acceptable.

Even if we establish that the response model fits the data well enough in an acceptably high percentage of the cases, the question remains as to whether such modeling imperfections result in errors that are too great to be acceptable. This question is best addressed by conventional residual analysis, in which the residual variance is quantified and compared to prescribed specifications.

Acknowledgements

This work was supported by the Engineering Directorate at NASA Langley Research Center.

References

[1]Shewhart, Walter A[ndrew]. (1931). *Economic Control of Quality of Manufactured Product*. New York: D. Van Nostrand Company. ISBN 0-87389-076-0

[2]Taguchi, G., Chowdhury, S., Wu,Y.: "Taguchi's Quality Engineering Handbook." Wiley-Interscience, 2004. ISBN-13: 978-0471413349

[3]Dening, W. E.: *Quality, Productivity, and Competitive Position*. Massachusetts Inst Technology, June 1982 ISBN-13: 978-0911379006

[4]DeLoach, R., "Assessment of Response Surface Models Using Independent Confirmation Point Analysis" AIAA 2010-741, 48th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, Florida, January 4-7, 2010

[5]Box, G. E. P., and Draper, N. R., *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd Ed., John Wiley and Sons, New York, 2007.

[6]Montgomery, D. C., Peck, E. A., and Vining, C. G., *Introduction to Linear Regression Analysis*, 4th Ed., John Wiley and Sons, New York, 2006.

[7]Myers, R. H., and Montgomery, D. C., Response Surface Methodology: Process and Product Optimization Using Designed Experiments, John Wiley and Sons, New York, 1995.

[8]DeLoach, R., "Response Surface Modeling Tolerance and Inference Error Risk Specifications: Proposed Industry Standards (Invited)" AIAA 2012-2859. 28th AIAA Aerodynamic Measurement Technology and Ground Testing Conference. New Orleans, LA. June 25-28, 2012.

[9]Dr. Mark Kammeyer, The Boeing Company. Personal communication.

[10]DeLoach, R.: "Bayesian Inference in the Modern Design of Experiments", AIAA-2008-0847. 46th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV Jan 7-10, 2008.

[11]Mr. João Dias, Instituto Superior Técnico, Lisbon, Portugal. Personal communication.