The goal of the STAT T&E COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.
# Table of Contents

Executive Summary ....................................................................................................................................... 2  
Introduction .................................................................................................................................................. 2  
   The Benefits of a Designed Experiment .................................................................................................... 2  
   Multiple Data Analysis Methods May Be Needed .................................................................................... 2  
Summary Statistics ........................................................................................................................................ 3  
Single Aggregate Distribution and Percentage above Threshold ................................................................. 4  
Factor Effects; Best and Worst Performance ................................................................................................ 5  
Variation at a Single Design Point .................................................................................................................. 6  
Significant Effects, Regression Modeling, and Prediction ............................................................................. 6  
Analysis in Greater Detail ................................................................................................................................ 8  
Conclusion ..................................................................................................................................................... 8  
References .................................................................................................................................................... 8
Executive Summary
Test and evaluation (T&E) activities ultimately support programmatic decisions. Decision-quality information requires an assessment of the system against the requirement as well as perspective on the data to identify risk. The use of scientific test and analysis techniques (STAT) will improve the breadth and depth of information collected during testing utilizing technique like design of experiments (DOE) is employed. Simple statistics like the mean and standard deviation will not adequately inform the decision maker. The ability to separate and estimate effects and interactions is critical to evaluating performance. Reporting the best and worst performing conditions and the local variability quantifies the performance margin (risk) in the system and provides perspective for the decision-maker. This paper presents a methodical analysis process to leverage the DOE data set and provide comprehensive data analysis.

Keywords: data analysis, reporting, Department of Defense, scientific test and analysis techniques

Introduction
“A rigorous and efficient T&E program provides early knowledge of developmental and operational issues” (Defense Acquisition Guidebook, 2014). Decision-quality information requires an assessment of the system against the requirement as well as perspective on the data to identify risk. The use of scientific test and analysis techniques (STAT) will improve the breadth and depth of information collected during testing especially where design of experiments (DOE) is employed. DOE is a powerful test methodology conceived with an equally powerful analysis capability. But, unlike typical industrial examples where the goal is often to determine a single point of optimum manufacturing conditions, DOD testing strives to assess performance against the requirement across the entire operational space. A methodical analysis process is presented to leverage the DOE data set and provide comprehensive information for the decision maker. This paper requires the reader have completed at least a basic DOE course and understand the fundamentals of test design and analysis (for maximum utility).

The Benefits of a Designed Experiment
Well designed experiments offer many advantages to the tester especially in terms of data analysis. DOE is not just a method to create test points but a method to create test points purposely arranged to enable detailed analysis of the response behavior under varying factors and levels. The ability to separate and estimate effects and interactions is critical to evaluating performance and permits reporting beyond mere averages and aggregates. Factor levels that negatively impact performance can be investigated in order to focus future improvements. Predictive modeling can also assist in the development of tactics and techniques. For a DOD-specific DOE test planning reference see OSD STAT COE (2014).

Multiple Data Analysis Methods May Be Needed
Consider a requirement to “launch a projectile at least 32 yards, 80% of the time.” We know that repeated tests under one set of conditions will display variability in the response and changing factor levels will certainly impact the response. Simple statistics like the mean and standard deviation will not adequately inform the decision maker about these changes. Reporting the best and worst performing conditions and local variability goes a long way toward explaining the amount of performance margin
(risk) the system possesses. Figure 1 outlines the analysis concepts that support comprehensive reporting on DOE data sets.

![Figure 1: Multiple Components of Analysis](image)

The process of combing through the data should be methodical and thorough. The following sections detail the concepts in Figure 1, provide further references, and discuss strengths and limitations. These steps are presented in an ordered approach.

**Summary Statistics**

Reporting the average, data range, and standard deviation are simple statistics to calculate for any data set (designed or not) but they can be misleading. When one considers the range of conditions under which weapons systems operate, it becomes obvious that simple statistics are not sufficient for drawing conclusions. Without considering factor levels and their impact on the response the variance may be so large as to make any inferences about averages, proportions, and other simple statistics from the data highly suspicious. The reason for a DOE is to go beyond mere summary statistics. Figure 2 shows a table of basic statistics including the confidence bounds. This information should be reported with the data set since it is applicable. However, decisions should not hinge on this information without deeper analysis described later.
Figure 2: Basic Statistical Information

For instance, the mean of 39.4 passes the minimum requirement of 32.0 and the lower bound for the mean (36.3) passes as well. However, the distribution of the data may indicate that some very low values are being offset in the mean by some very large values. Remember, the bounds on the mean do not indicate the range of values in the data but rather under repeated use of the sampling method, 95% of such intervals would contain the true mean. Also, this information cannot address the “80%” portion of the requirement. Therefore, deeper analysis is required to make an informed decision and the benefits of having conducting a DOE have not yet been realized.

Single Aggregate Distribution and Percentage above Threshold

Viewing the data set in histogram form and/or fitting a distribution provides additional insight. Figure 3 shows a histogram for 40 distances (green boxes) along with a fitted distribution (blue line). Remember that the number of data points, bin width, and distribution of the data will impact the histogram image produced and the quality of the fitted curve.

Beware that presenting the data in this manner implies a single population. A population is a set of objects (outcomes in this case) of a similar nature. Our distance data may not possess this “similar

\[\text{http://www.statistic.com/glossary&term_id=812}\]
nature” because they were collected under varying conditions (on purpose) however the requirement treats all results as if from a single population. The analytical software easily fits the blue distribution line and estimates the associated parameters (Lognormal, mean=3.6, sigma=0.24). This information can be used to estimate that 75% of results are greater than 32 yards. This result indicates the system DOES NOT meet the 80% portion of the requirement despite the mean value clearly meeting the minimum of 32 yards portion. But again, the distribution alone does not fully describe the factor effects in the DOE.

**Factor Effects; Best and Worst Performance**

Viewing the raw data is a very useful way to look for trends and examine the output. Figure 4 is a scatterplot of all 40 responses (distance) against each design factor. The red points are values below threshold. Plotting the raw data is always a good idea and may be useful in identifying gross trends and independently confirming statistical analysis. These plots are particularly useful when many factors are involved because only two factors can be depicted in three dimensional space (as depicted in Figure 1). All the responses (black dots) are shown but the other factor levels that contributed to each response point are not immediately obvious in this plot so drawing conclusions via graphical interpretation must be done carefully. The means (green diamonds) superimposed on this plot appear to indicate that factors ANGLE and VELOCITY are significant (to this response) due to the high slope between levels. This will be examined statistically later.

![Figure 4: Scatterplot of a Designed Experiment](image)

Additionally, the design/data table can be used to explicitly identify the factor combinations that result in the best and worst performance. One can compare and contrast these factor levels and comment on whether the results were predictable or unexpected. Furthermore, the frequency with which the system is operated at these conditions should be noted. Performance (good or bad) at rarely seen conditions should not influence the decision-maker as much as results under more commonly encountered conditions.
Scatter plots are only a single graphical tool among many available to the analyst. A large list of tools is available at the StatSoft website (StatSoft, 2014). Largely graphical in nature, exploratory data analysis (EDA) was pioneered by John Tukey (Tukey, 1977) and an introduction to the field can be explored at the National Institute of Standards and Technology EDA website (NIST Section 1, 2014).

**Variation at a Single Design Point**

Response variation measured at a single design point provides valuable information about expected accuracy (in this case). Since regression modeling generally assumes constant variation throughout the space we can easily derive this value from analysis of variance (ANOVA) for the factorial design that was executed. The ANOVA table shown in Figure 5 shows a Mean Square Error of 7.408 (red box) which is equivalent to the local variance (local variance=standard deviation squared; square root of 7.408 yds$^2$ = 2.72 yds) at any design point (Montgomery et al, 2012).

![Figure 5: ANOVA Table](image)

If there are requirements for accuracy or repeatability then this value can assist in addressing them. Also, this information may be of significant interest for the user when tactics and procedures are developed.

**Significant Effects, Regression Modeling, and Prediction**

Significant effects are also determined through the ANOVA process. Figure 6 shows two main effects and four two-factor interactions explain the responses (highlighted in red). Note that factors ALTITUDE, WIND, and TEMPERATURE are not significant but remain included because they are included in the interactions.
Regression is the process of creating an empirical model to mathematically describe the response as a function of the factors. The parameter estimates for the significant factors directly populate the regression model in Figure 7. This is easily calculated and manipulated using DOE software.

Distance = 39.7 + 8.0*ANGLE + 0.56*ALTITUDE + 2.4*VELOCITY - 0.25*WIND + 0.66*TEMPERATURE + 2.7*ANGLE*VELOCITY – 0.95*ANGLE*WIND + 0.98*ANGLE*TEMPERATURE – 2.6*ALTITUDE*WIND

Entering any factor values between the upper and lower design limits will predict performance for that point, not just at the design points (NIST Section 4, 2014). This is particularly useful for comparing
results with other testing conducted in the same factor space but at different factor levels. In operational testing, where factor control is often more difficult, being able to compare the regression predictions to the actual OT data is extremely useful for reporting results and conclusions. In addition, Monte Carlo simulation can be employed with the model to calculate a finer estimate of the percentage of points that exhibit performance above or below certain values, providing another method to assess the 80% portion of the requirement. This model estimates that 96% of all values in the factor space fall above 32. Because only ANGLE and VELOCITY are significant main effects the plot in Figure 1 turns out to be an accurate representation of the surface in just three dimensions. One can clearly see that a large portion of the green surface is above the horizontal (value of 32) grid so the 96% estimate is expected. The additional insight provided from the regression model indicates the system DOES PASS the requirement with some margin, while the fitted distribution did not. Note that the fit of the regression model (how well it explains the data) must be taken into account but that is not described here.

Analysis in Greater Detail

The process outlined above is truly a minimum and the particulars of a complex weapon system test program may require significant analysis beyond what is presented here. Curvature in the response, optimization, hard to change factors, and data outliers all require an experienced analyst. DOE software tools provide many automated methods to present and analyze data but one must ensure they are employed properly. Sometimes custom analysis tools are required to handle unique designs and data sets. In all cases it is incumbent on the practitioner to consult a STAT specialist when the complexity of the test design goes beyond the basics. The efficiency and effectiveness that DOE provides is best leveraged when the whole problem is well understood including requirements, objectives, responses, factors, constraints, and limitations. These must all be considered in the design process to ensure the data generated will sufficiently address the requirement.

Conclusion

The use of DOE provides rigor and depth in data collection and analysis sufficient to address DOD requirements and support programmatic decisions. A comprehensive and methodical analysis approach permits reporting beyond mere averages and aggregates and effectively informs requirements, risk, and variation in the operating region. A combination of summary statistics, aggregate data calculations, ANOVA, use of the regression model, and graphical data techniques will support production of the most useful test reports.

References


Tukey, John, *Exploratory Data Analysis*, Addison-Wesley, 1977