INSTITUTE FOR DEFENSE ANALYSES

Statistical Methods for Defense Testing

Dean Thomas, *Project Leader*

Kelly M. Avery
Matthew R. Avery
Laura J. Freeman

December 2017

Approved for public release.
Distribution is unlimited.

IDA Non-Standard Document
NS D-8893

Log: H 2017-000710

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task 229990, Test Science Applications, for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
Technical review was performed by Yevgeniya K. Pinelis of the Operational Evaluation Division.

For more information:
Dean Thomas, Project Leader
dthomas@ida.org • (703) 845-6986

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-8893

Statistical Methods for Defense Testing

Dean Thomas, *Project Leader*

Kelly M. Avery
Matthew R. Avery
Laura J. Freeman

Statistical Methods for Defense Testing

Matthew R. Avery, Kelly M. Avery, Laura J. Freeman

Institute for Defense Analyses, VA, USA

Keywords: Defense, Operational Testing, Design of Experiments, Modeling & Simulation

Abstract: In the increasingly complex and data-limited world of military defense testing, statisticians play a valuable role in many applications. Before the DoD acquires any major new capability, that system must undergo realistic testing in its intended environment with military users. Although the typical test environment is highly variable and factors are often uncontrolled, design of experiments techniques can add objectivity, efficiency, and rigor to the process of test planning. Statistical analyses help system evaluators get the most information out of limited data sets. Oftentimes new or complex analysis techniques are needed to support the goal of characterizing or predicting system performance across the operational space. Finally, the growing need for computer models or simulations to supplement live testing also means that these models must be appropriately validated before their output can be deemed sufficient for use. Statistical design and analysis techniques are essential for rigorous evaluation of these models.

1. Introduction

From fighter aircraft and satellites to submarines and ground vehicles, the United States Department of Defense (DoD) acquires some of the world's most complex systems. These systems push the limits of existing technology and span a wide range of domains. Before it is fielded, each one of these systems must be evaluated to determine whether it will be effective in enabling military users to accomplish missions.

Operational testing (OT) is the final stage in a series of such evaluations. OT involves military operators using the production-representative system to complete realistic combat scenarios. One unique characteristic of OT is that there are many uncontrollable variables. Operational users, the operational environment (e.g., changes in the weather), and the evolving context of the mission all introduce variability.

Historically, OT was often conducted as “freeplay” exercises, or as a series of hand-selected scenarios based on the most common use conditions, or simply “what we did last time.” While these approaches produced tests that were operationally realistic, they lacked the scientific process needed to ensure that testing is both efficient and able to characterize system performance across a diverse range of conditions. For complex defense systems, performance often depends on interactions of independent variables. Historical test strategies are typically inadequate to support the estimation of such interactions.

By involving a statistician in all phases of the testing process, the DoD can ensure they are conducting tests objectively and efficiently, maximizing the information gained from data analysis, and reporting results with an appropriate level of certainty. The following sections describe in more detail the role statisticians play and the techniques they use to support test design (Section 2), data analysis (Section 3), and modeling and simulation validation (Section 4).

2. Design of Experiments

Design of experiments (DOE) is a scientific, structured, objective test methodology for answering key questions of test, including which and how many points are needed to test a hypothesis of interest[1]. DOE is useful for covering a large input space efficiently; testers don’t necessarily need to conduct test runs in every possible combination of conditions in order to meaningfully describe relationships between each factor and the response.

DOE encompasses many different design types that support various types of data, analytical goals, and levels of fidelity. Operational tests are by nature highly stochastic, and testers seek to characterize or explain sources of variability; thus the most commonly used DOE techniques fall in the “classical” design family and include full factorials, fractional factorials, optimal designs, and response surface designs[2].

Since operational tests should be as realistic as possible, testers often cannot, or do not wish to, control certain variables that they think may have an effect on the response of interest. This makes at least a portion of the test design seem more like an observational study approach, where testers simply record what happened with particular variables during test. However, this doesn’t preclude testers from using DOE to plan an adequate test, and actively checking off runs as they occur. If the end of test is approaching and there are still significant gaps in the design space, testers can consider executing more control to force collection of the missing data, or even extending the test so that the required data is collected naturally. Despite the highly variable and often observational nature of OT, statisticians can help ensure that the data collected supports analysis that includes hypothesized variable interactions, and limits the implications of correlation.

Multi-level designs are also common in OT. Testers often plan their test and build their primary DOE around a handful of mission-level factors that cover the space where the system will operate. But the analysis these testers are actually interested in conducting is based on much lower-level, more granular data. A good design strategy will cover the operational envelope while also ensuring that sufficient data is collected in each sub-mission area to support the desired analysis.

The application of DOE to OT is nontraditional in a number of ways, but is valuable for adding objectivity to the test planning process, building efficient tests, and ensuring that a rigorous analysis is possible after the test is complete. The statistical community tends to think too narrowly about the application of DOE; the core principles still apply in settings outside of a pristine lab environment.

3. Analysis

Given the complexity of the environment in which many DoD systems are expected to operate and the often limited data available from OT, it is important to use this data efficiently. This typically means fitting a statistical model, such as linear regression. Commonly employed in other fields, statistical models allow analysts to make inferences across a wide range of conditions and leverage information from the full set of data simultaneously. The most common tools used in OT analysis will be familiar to trained statisticians, though the ways in which they are applied may not be. When the results from operational test are being reported, the focus is on prediction and inference about the response variable rather than on parameter estimates, as is common in academic fields.

Common analytical tools used for analyzing operational test data include linear models[2] of many varieties, as well as categorical data analysis[3]. Simple linear regression, generalized linear models, and mixed models are all useful in operational test. Random effects and mixed models can be used when the operational environment creates substantial uncontrolled variability. For example, aircraft will exhibit performance variation on different sorties. The most logical approach for modeling this variation is to treat sortie as a random effect, analogous to subject-level variability in biomedical research. The biostatistician may be interested in the effect of a new drug on blood pressure levels, and between-patient variation is a source of noise rather than a subject of direct investigation. An operational tester may similarly be interested in the effect of a countermeasures system on the miss distance of enemy surface-to-air missiles fired at a fighter aircraft, and differences in miss distance from one sortie to the next (whether due to the performance of the pilot, aircraft, or threat simulator responsible for "firing" the surface-to-air missile) isn't of direct interest.

One important difference is how the results from these analyses are reported. To continue with the drug trial analogy introduced above, the biostatistician's report will likely include a regression table showing the parameter estimates and p-values.[1][4] These are useful for addressing the types of questions biostatistical research is often concerned with ("Did the intervention improve my chosen measure of patient health?" and "Is the difference statistically significant so that I can hope for FDA approval?"), but these types of tables fail to present results in a way that is meaningful for DOD decision-makers. Instead,

[1] See, for example, Tables 3 and 5 in Casiglia et al., 2002

analysis of DOD systems focuses on response-level inference. Plots showing estimated performance across the range of values covered in the DOE illustrate the missions and conditions in which a system meets or fails to meet its performance thresholds. Prediction intervals are often preferred over confidence intervals for illustrating to the warfighter the range of outcomes they can expect in combat. System requirements are often given in the form of quantiles[2] or probabilities,[3] making estimation of these values important.

For evidence that these quantities are not as popular as parameter estimates, one need look no further than the built-in tools in most statistical packages. With R as an example, most linear modeling packages (e.g., lm, glm, lme4) have built-in reporting of parameter estimates and associated p-values. Predicted point estimates and associated confidence intervals are available for some packages but not all. But prediction intervals and quantile and probability estimates typically require the investigator to do additional work. This need has led to the development of a new R package, ciTools,[5] which addresses these shortfalls for many popular linear model functions. Having these quantities readily available makes it easier for analysts to answer the most common types of questions in DOD testing.

While statistical models are applicable to data collected from observational studies, they are most potent when used with efficiently designed experiments. OT data collected from a DOE and analyzed using an appropriate statistical model can maximize the information gained from test and provide valuable information to decision-makers and warfighters. Properly applying the tools in a statistician's toolbox allows analysts to identify

[2] e.g., "The 90th percentile of miss distances for the new Howitzer must be no greater than X meters."

[3] e.g., "The system must be sufficiently reliable to complete a 720-hour mission without failure 80 percent of the time."

differences in system performance across the operational space, identify areas where a system should be improved, and evaluate system requirements.

4. Modeling & Simulation

Operational evaluations increasingly rely on M&S to supplement live testing. The complexity of modern military systems and the environment in which they operate means that live testing is often expensive or even impossible, since certain threats or combat scenarios simply cannot be replicated on test ranges. Thus, M&S capabilities (including but not limited to digital computer models, hardware-in-the-loop simulations, and threat emulators) are critical to fully characterizing a system's capabilities.

If the acquisition community, and ultimately the warfighter, are going to depend on M&S results to inform decisions, testers should thoroughly investigate how well the M&S represents the system of interest across its operational space. This process of determining the degree to which a model provides an accurate representation of the real world is known as validation.[6] Statisticians can provide useful input to the validation process in both the design and analysis phases.

While many techniques can be used to support the validation of a model,[7] a portion of any validation should include a quantitative comparison of the simulation output to available live data. Such an analysis provides an objective statistical measure of the differences between the M&S and the live test, quantifies the uncertainty in the model, and can identify areas of high risk that might require additional testing. Understanding these differences and the uncertainty associated with them will inform whether the simulation is adequate for the intended use. It is important to note that a simulation need not match live data exactly in order to be useful.

Design of experiments (DOE) techniques (introduced in section 2) can be used to justify which, and how many, data points are required to support a meaningful comparison of live data to M&S output. In this case, matched classical designs (e.g., factorial or optimal) in both the live and simulation environments are ideal for facilitating regression-based statistical analyses and quantifying differences between live and M&S data across various operational conditions.

DOE techniques can also be used to support characterization of the simulation itself. Even before comparing M&S output to live data, it is important to understand how changes to input variables drive changes in output, and to determine whether those changes are reasonable. This practice is known as sensitivity analysis. Design for computer experiments methodologies, such as space filling designs,[8] can be used to build an empirical model (or statistical emulator) of the M&S and facilitate prediction analysis, heighten understanding of the most important factors, and inform future live testing.

Strong statistical designs in both the M&S and live environments enable rigorous validation analyses. If a DOE was performed, regression modeling techniques are generally the most powerful since they are designed to optimally leverage all information in the presence of factors. Even without the benefit of matched designs, robust non-parametric techniques, such as the Kolmogorov-Smirnov test[9] and Fisher's Combined Probability Test,[10] are widely applicable and can provide useful insights for validation.

Since much of the current statistical literature[11][12] tends to ignore the limited live data problem that is prevalent in defense testing, much more work can and should be done to explore the applicability and performance of other techniques. In particular, finding ways to combine limited empirical results with imperfect simulation predictions is worthy of further study.

5. Conclusion

Defense testing is complex and resource intensive, so data should be collected and evaluated in the most efficient and meaningful way possible. Statistical methods for test design and analysis provide a means for doing so. As our systems become even more complex, and leverage levels of autonomy, the defense community needs to continue to evolve our application of state-of-the-art statistical methodologies to confront these new challenges.

6. References

[1] Montgomery, D.C. (2008). Design and analysis of experiments, John Wiley & Sons, New York, NY.

[2] Johnson, R. T., Hutto, G. T., Simpson, J. R., & Montgomery, D.C. (2012). Designed experiments for the defense community, *Quality Engineering*, **24(1)**, 60-79.

[3] Agresti A. (2002). Categorical Data Analysis 2nd edition, John Wiley & Sons, New York, NY.

[4] Casiglia E., Tikhonoff V., Mazza A., Piccoli. A. & Pessina A.C. (2002). Pulse pressure and coronary mortality in elderly men and women from general population, *Journal of Human Hypertension* **16**, 611–620.

[5] Haman, J, & Avery, M. (2017). ciTools: Confidence or Prediction Intervals, Quantiles, and Probabilities for Statistical Models. R package version 0.2.1.

[6] Department of Defense (2009). DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A), DoD Instruction 5000.61.

[7] Sargent, R.G. (2003). Verification and Validation of Simulation Models, *Proceedings of the 2003 Winter Simulation Conference*, 37-48.

[8] Park J. (1994). Optimal Latin-hypercube designs for computer experiments, *Journal of statistical planning and inference* **39.1**, 95-111.

[9] Stephens, M.A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, *Journal of the American Statistical Association* **69.347**, 730-737.

[10] Fisher, R.A. (1925). Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh Scotland.

[11] Kleijnen, J.PC. (1995). Verification and validation of simulation models, *European journal of operational research* **82.1**, 145-162.

[12] Harmon S. Y. and Youngblood S.M. (2005). A proposed model for simulation validation process maturity, *The Journal of Defense Modeling and Simulation* **2.4**, 179-190.