INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2023: Comparing Normal and Binary D-Optimal Design of Experiments by Statistical Power

Rebecca M. Medlin, Project Leader

Addison D. Adams
Curtis G. Miller

April 2023

Public release approved. Distribution is unlimited.

IDA Document NS D-33405

Log: H 2023-000055

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9082, "Statistics and Data Science Working Group." 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
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Kelly M. Avery, Dr. Keyla Pagan-Rivera, and Dr. Jason P. Sheldon from the Operational Evaluation Division.

For more information:
Dr. Rebecca M. Medlin, Project Leader
rmedlin@ida.org • 703-845-6731

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-33045

DATAWorks 2023: Comparing Normal and Binary D-Optimal Design of Experiments by Statistical Power

Rebecca M. Medlin, Project Leader

Addison D. Adams
Curtis G. Miller

Executive Summary

In many Department of Defense test and evaluation applications, binary response variables are unavoidable. Many have considered D-optimal design of experiments for generalized linear models. However, little consideration has been given to assessing how these new designs perform in terms of statistical power for a given hypothesis test. Monte Carlo simulations and exact power calculations suggest that D-optimal designs generally yield higher power than binary D-optimal designs, despite using logistic regression in the analysis after data have been collected. Results from using statistical power to compare designs contradict traditional design of experiments comparisons, which employ D-efficiency ratios and fractional design space plots. Power calculations suggest that practitioners that are primarily interested in the resulting statistical power of a design should use normal D-optimal designs over binary D-optimal designs when logistic regression is to be used in the data analysis after data collection

Addison Adams

Institute for Defense Analyses

Comparing Normal and Binary D-Optimal Designs by Statistical Power

Problem Statement

How should an operational test event be planned when the response of interest is a success or failure (binary data)?

When it is known prior to data collection that the response variable will be binary, how should an experiment be designed? Many common comparisons between designs rely on asymptotic results. Instead, we look at statistical power calculations to compare designs. Generally, normal D-optimal designs results in higher statistical power than binary D-optimal designs.

Comparison of Two Designs

Response surface designs are common in practice as they are an option regardless of the type of model the experimenter uses. A D-optimal design is a common response surface design that aims to minimize the generalized variance of the estimated parameters. Table 1 highlights the difference between the binary D-optimal and normal D-optimal designs.

Table 1

Test Planning

Operational Test Event

Data Analysis

	Normal Design	Binary Design
Response Model	$y_i = x_i^T\beta + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$	$y_i \sim Bernoulli(\pi_i)$ where $\pi_i = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}$
D-Criterion	Maximize the determinant of the information matrix	Maximize the determinant of the information matrix
Information Matrix	X^TX	X^TV_BX where V_B depends on β
Data Collection	Collects binary response variable	Collects binary response variable
Analysis Model	Logistic Regression	Logistic Regression

As Table 1 indicates, the two designs differ in the response model and the information matrix. The binary design correctly anticipates binary data and the analysis model – logistic regression. After the binary data is collected, the data are analyzed using logistic regression regardless of the design. However, the information matrix for the binary design depends on the unknown parameter vector β. Additionally, for the binary design, the variance matrix of the parameter estimates depends on the information matrix asymptotically, i.e. as the sample size increases to infinity; whereas for the normal design, the variance matrix is a function of the information matrix for any sample size.

Hypothetical Example for the Department of Defense (DOD)

Although it is typically preferable to use continuous response data when available, it is not always viable in DOD applications. For example, whether or not a torpedo located its target is a binary response. Suppose that it was of interest to test if the following three conditions affected the hit probability of the torpedo:

- The speed of the target boat
- If the target boat employs evasive maneuvers (yes/no)
- The depth of the target boat (deep/shallow)

For the hypothetical experiment, the binary and normal D-optimal designs are visualized in Figure 1. The normal D-optimal design produces a design with eight support points, equally weighted. The binary D-optimal design has only seven support points, with different weights.

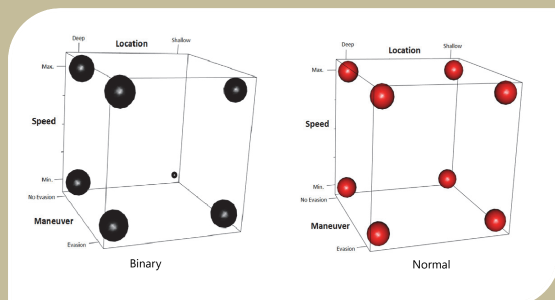


Figure 1

Method of Comparison

- The two designs are compared via statistical power.
  - Traditional comparisons include D-efficiency or Fraction of Design Space plots, both of which rely on asymptotic theory.
- The analysis model is logistic regression:

$$y_i \sim Bernoulli(\pi_i) \text{ where } \pi_i = \frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)} \text{ and } x_i^T\beta = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$$

- The model utility test is used and is described as:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0 \text{ vs } H_a: \text{at least one } \beta_i \text{ is non-zero.}$$

- The test statistic used is the likelihood ratio test.
- The binary D-optimal design was calculated using the true unknown parameters. The resulting design is locally optimal.
- Power is calculated via Monte Carlo simulation with logistic regression in the analysis.

Simulated Power Results

Figure 2 demonstrates the simulated power results for the two designs from Figure 1 as sample size increases. The normal D-optimal design outperforms the binary D-optimal design in terms of statistical power.

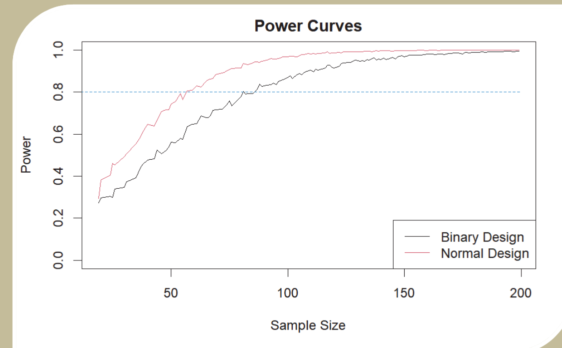


Figure 2

Exact Power Results

Power can be calculated exactly. The intercept term β_0 is held fixed at 2 and the sample size is held fixed at 12. For simplicity, β_3 is dropped for the model. Power is calculated over the parameter space, $(\beta_1, \beta_2) \in \mathbb{R}^2$. Power is illustrated for the two designs by heat plots in Figure 3.

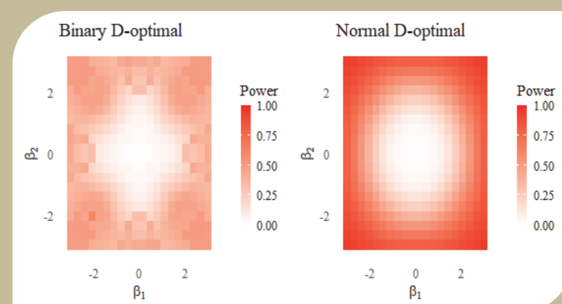


Figure 3

Generally, the normal D-optimal design outperforms the binary D-optimal design in terms of statistical power.

Conclusions

- The normal D-optimal design generally yields higher power than the binary D-optimal design despite logistic regression being used in the data analysis.
- The results presented here are based on the model utility test hypothesis; however, different hypotheses such as standard pairwise comparisons lead to similar results.
- The results presented here are based on the likelihood ratio test; however, different test statistics such as the Wald test lead to similar results.
- Standard comparison methods such as the D-efficiency ratio and fraction of design space plots favor the binary D-optimal design. These comparisons contrast with those based on statistical power.
- The results presented here are based on the logit link function. Different link functions may lead to different results, e.g. there is a link function such that information matrices are equivalent, and thus the designs are equivalent.

Acknowledgements

Thank you to Dr. Curtis Miller and Dr. Rebecca Medlin for their mentorship.

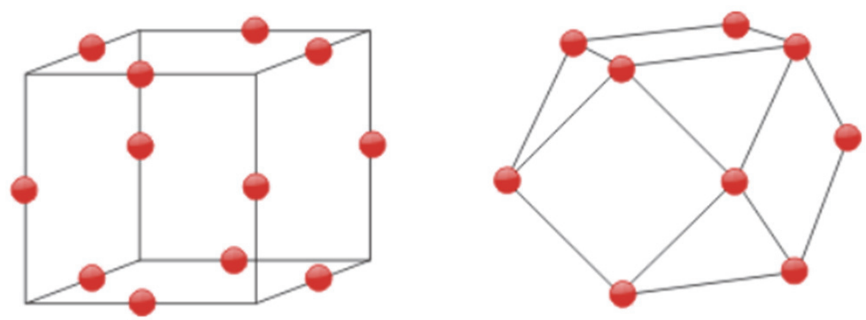
References

- Russell, K. G. (2018). *Design of experiments for generalized linear models*. Chapman and Hall/CRC.
- Box, G. E.; Hunter, W. H.; & Hunter, S. (1978). *Statistics for experimenters* (Vol. 664). New York: John Wiley and sons.
- Johnson, R. T., & Montgomery, D. C. (2009). Choice of second-order response surface designs for logistic and Poisson regression models. *International Journal of Experimental Design and Process Optimisation*, *1*(1), 2-23.
- Morgan-Wall, T., & Khoury, G. (2021). Optimal Design Generation and Power Evaluation in R: The skpr Package. *Journal of Statistical Software*, *99*, 1-36.
- Hothorn, T.; Zeileis, A.; Farebrother, R. W.; Cummins, C.; Millo, G.; Mitchell, D.; & Zeileis, M. A. (2015). Package 'lmtest'. *Testing linear regression models. https://cran. r-project. org/web/packages/lmtest/lmtest. pdf. Accessed*, *6*.
- Ozol-Godfrey, A.; Anderson-Cook, C.; & Robinson, T. J. (2008). Fraction of design space plots for generalized linear models. *Journal of Statistical Planning and Inference*, *138*(1), 203-219.
- Zahran, A.; Anderson-Cook, C. M.; & Myers, R. H. (2003). Fraction of design space to assess prediction capability of response surface designs. *Journal of Quality Technology*, *35*(4), 377-386.
- Burke, S. E.; Montgomery, D. C.; Anderson-Cook, C. M.; Silvestrini, R. T.; & Borror, C. M. (2021). Optimal designs for dual response systems for the normal and binomial case. *Quality and Reliability Engineering International*, *37*(7), 3034-3054.
- McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27-38.

DATAWorks 2023: Comparing Normal and Binary D-Optimal Designs by Statistical Power

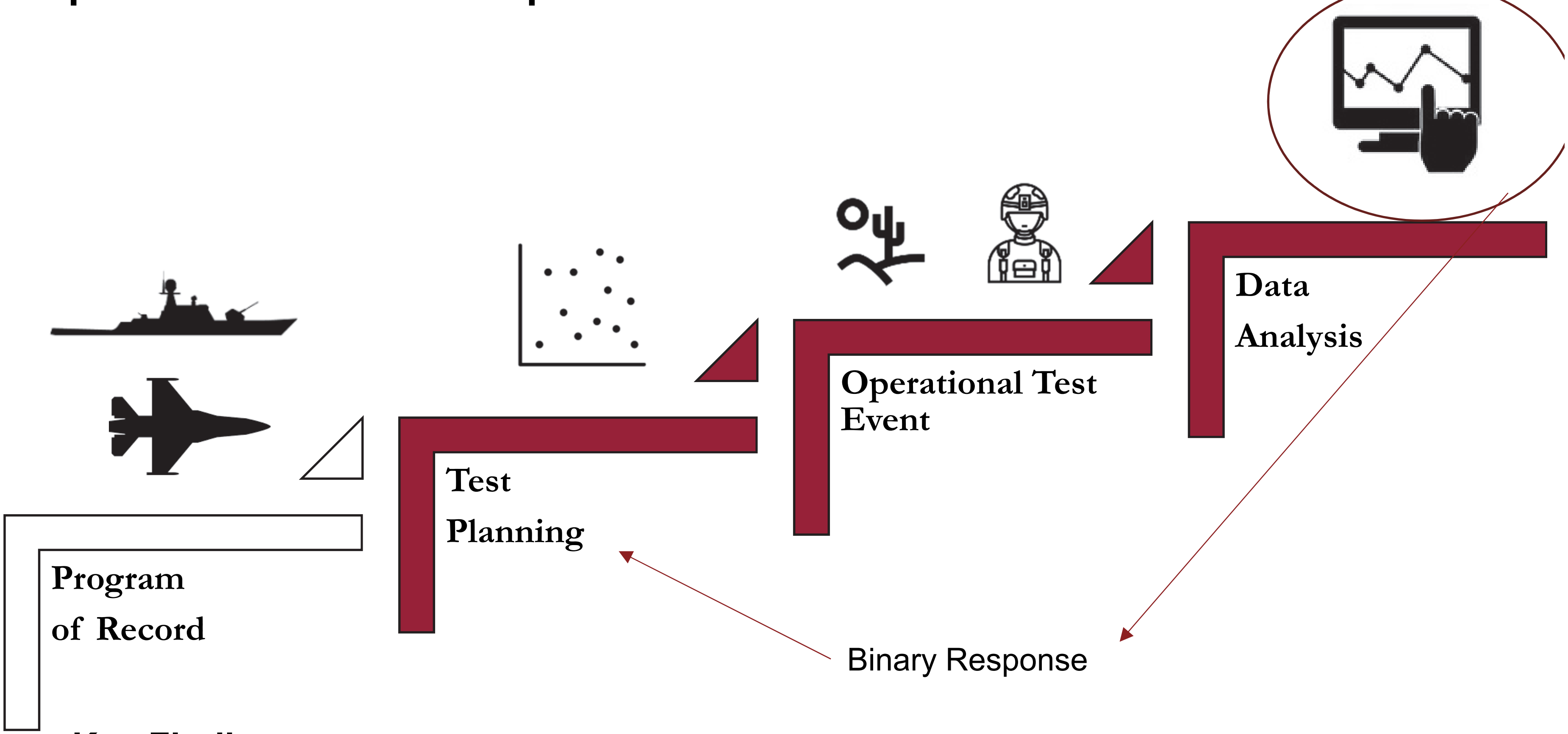
Addison D. Adams

April 2023



Institute for Defense Analyses
730 East Glebe Road • Alexandria, Virginia 22305

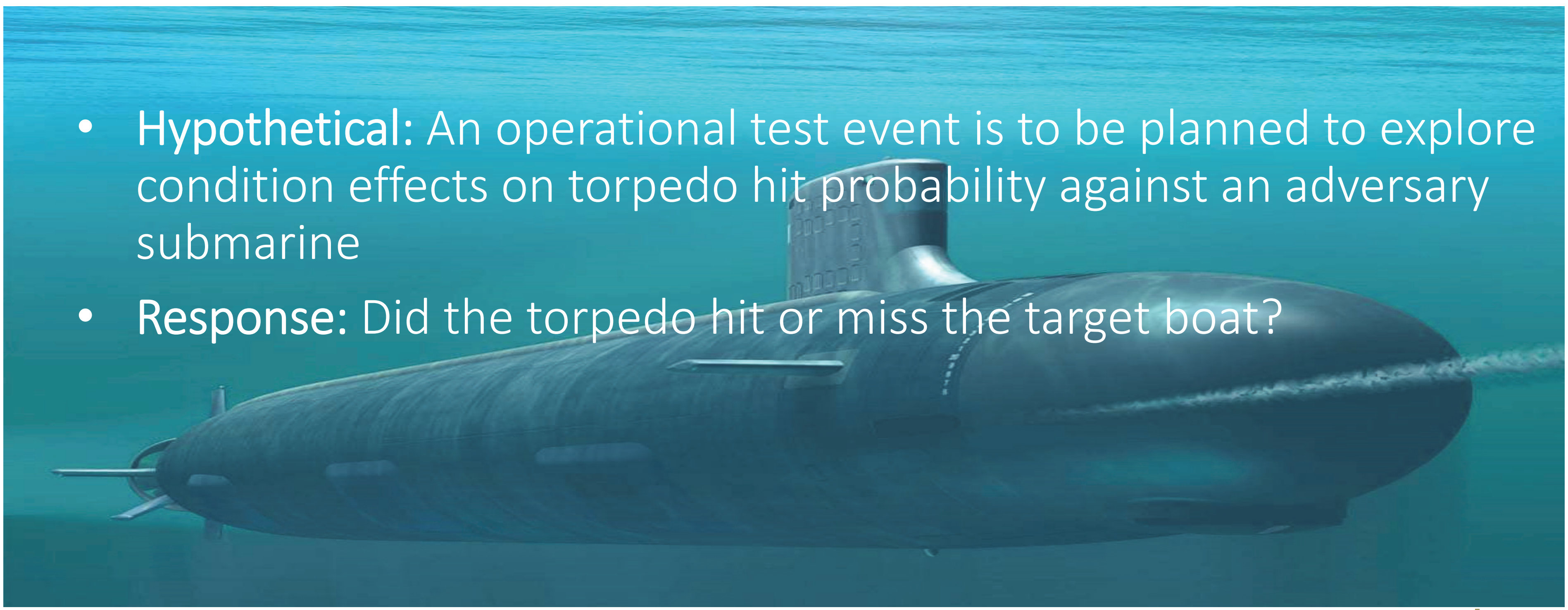
Central Question: How should an operational test event be planned when the response of interest is a success or failure?



- **Key Findings**:
  - We compare the binary D-optimal design with the normal D-optimal design by statistical power.
  - Standard design comparisons favor a binary design.
  - Generally, a normal D-optimal design results in higher statistical power than a binary D-optimal design.

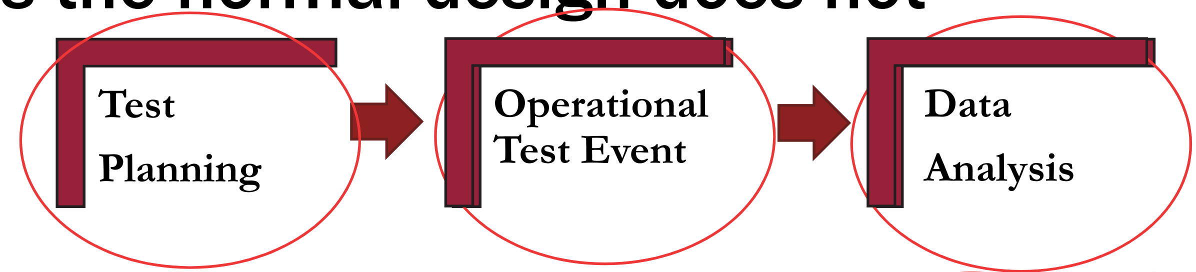
Design of experiment for torpedo hit probabilities

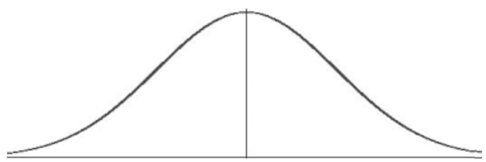
- A **design of experiment** (DOE) is the planning of an experiment with the statistical analysis in mind
- **D-optimal:** A design which minimizes the generalized variance of the parameter estimates

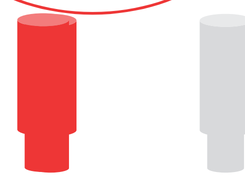


- **Hypothetical:** An operational test event is to be planned to explore condition effects on torpedo hit probability against an adversary submarine
- **Response:** Did the torpedo hit or miss the target boat?

Binary design anticipates binary data in data analysis, whereas the normal design does not



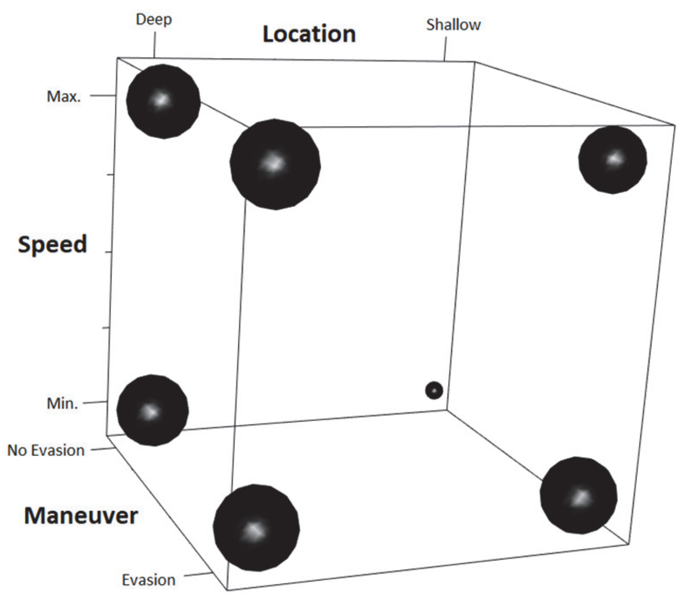




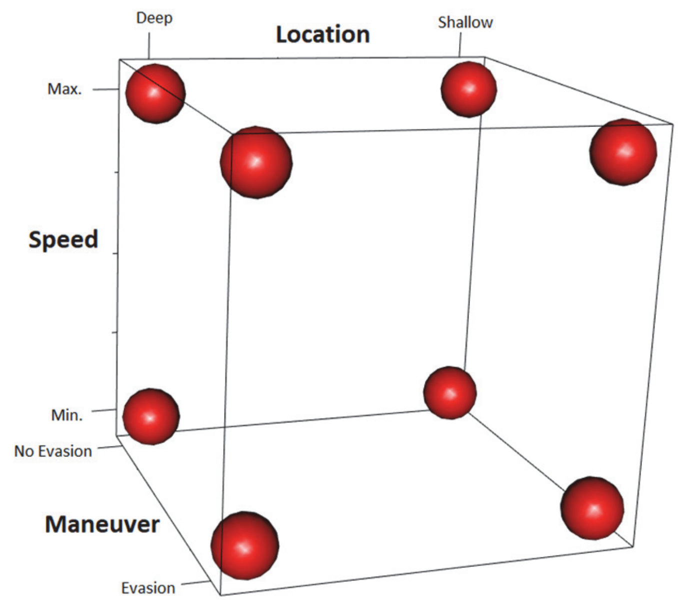
Normal design	Binary design
• Anticipates normal data in data analysis • Optimizes precision of parameter estimates assuming normal data	• Anticipates binary data in data analysis • Optimizes precision of parameter estimates assuming binary data
• Collects binary response in operational test event • Uses logistic regression in data analysis	• Collects binary response in operational test event • Uses logistic regression in data analysis

Standard DOE comparisons favor the binary design

Binary Design



Normal Design

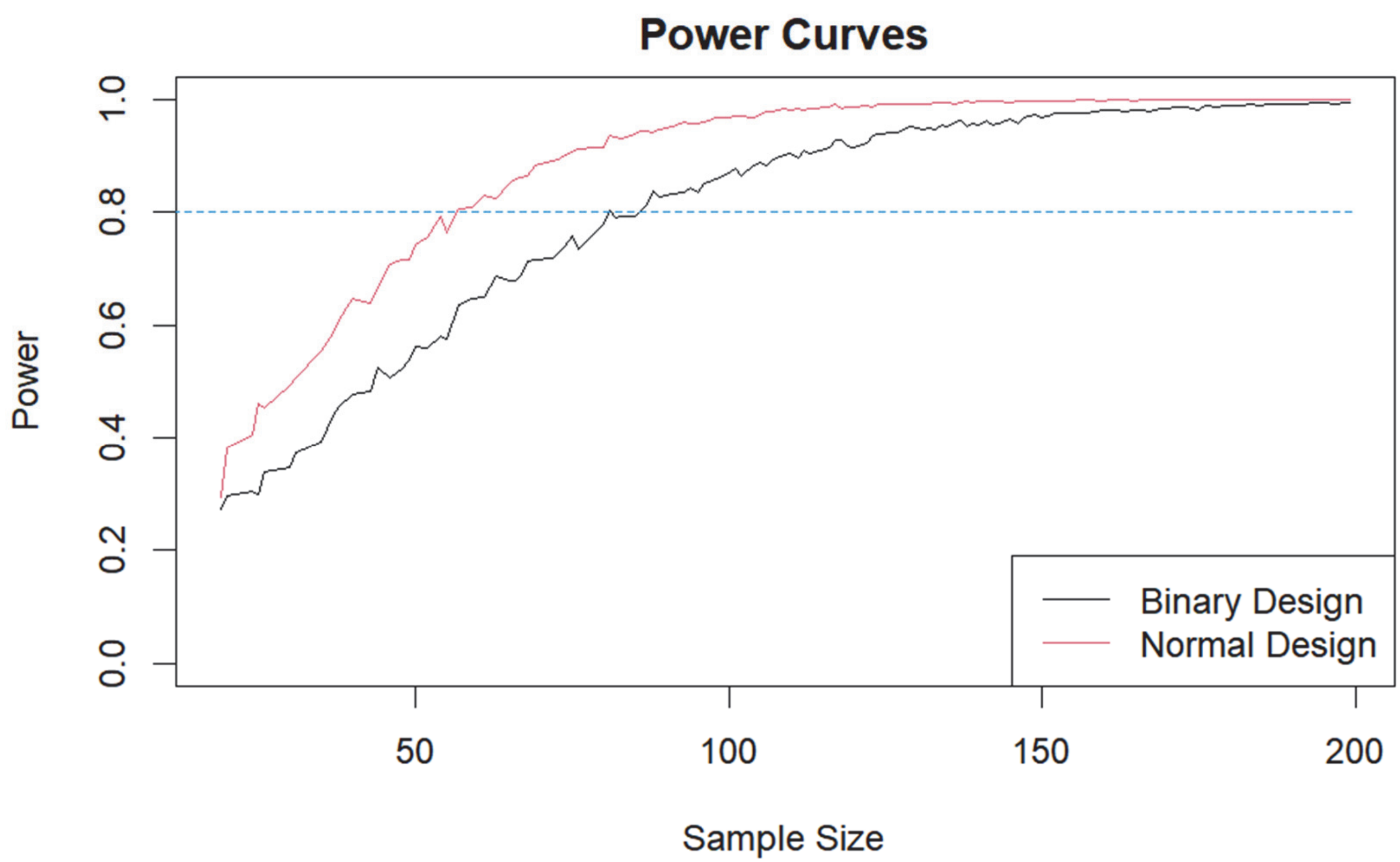


- Standard DOE comparisons favor the binary design including:
  - D-efficiency
  - Fraction of Design Space (FDS) plots
- Standard DOE comparisons rely on an large sample size (asymptotic theory)

DOE: Design of Experiment

Binary design underperforms in power analysis

- Power: The probability of detecting that at least one factor affects hit probability given that there is indeed an effect*
- Binary design underperforms normal design in power analysis



* Model utility test hypothesis with likelihood ratio statistic ($\alpha = 0.05$)

Primary Findings and Next Steps

Primary Findings

- Normal D-optimal designs outperform* the binary D-optimal designs despite using logistic regression in the data analysis
- Standard DOE comparisons favor the binary D-optimal design

Next Steps

- Consider a new optimality criterion
- Better understand why the normal design outperforms the binary design

* Performance is measured by power calculations.

DOE: Design of Experiment

References

- Russell, K. G. (2018). *Design of experiments for generalized linear models*. Chapman and Hall/CRC.
- Box, G. E.; Hunter, W. H.; & Hunter, S. (1978). *Statistics for experimenters* (Vol. 664). New York: John Wiley and sons.
- Johnson, R. T., & Montgomery, D. C. (2009). Choice of second-order response surface designs for logistic and Poisson regression models. *International Journal of Experimental Design and Process Optimisation*, *1*(1), 2-23.
- Morgan-Wall, T., & Khoury, G. (2021). Optimal Design Generation and Power Evaluation in R: The skpr Package. *Journal of Statistical Software*, *99*, 1-36.
- Hothorn, T., Zeileis, A.; Farebrother, R. W.; Cummins, C.; Millo, G.; Mitchell, D.; & Zeileis, M. A. (2015). Package 'lmtest'. *Testing linear regression models. https://cran. r-project. org/web/packages/lmtest/lmtest. pdf. Accessed*, *6*.
- Ozol-Godfrey, A.; Anderson-Cook, C.; & Robinson, T. J. (2008). Fraction of design space plots for generalized linear models. *Journal of Statistical Planning and Inference*, *138*(1), 203-219.
- Zahran, A.; Anderson-Cook, C. M.; & Myers, R. H. (2003). Fraction of design space to assess prediction capability of response surface designs. *Journal of Quality Technology*, *35*(4), 377-386.
- Burke, S. E.; Montgomery, D. C.; Anderson-Cook, C. M.; Silvestrini, R. T.; & Borror, C. M. (2021). Optimal designs for dual response systems for the normal and binomial case. *Quality and Reliability Engineering International*, *37*(7), 3034-3054.
- McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27-38.

How Power Analysis was conducted

- Power analysis is conducted via Monte Carlo simulation using the anticipated parameters
- Null Hypothesis: All non-intercept parameters are zero
  - Alternative Hypothesis: At least one non-intercept parameter is non-zero
- Test Statistic is Likelihood Ratio (R package lmtest)
- Parameter estimation is done using the firth correction
  - R package mbest to modify the glm object
- Additionally, Wald and Likelihood Ratios were assessed with ANOVA type III and produced similar results

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**	
XX-04-2023	Final	**START DATE**	**END DATE** April 2023

4. TITLE AND SUBTITLE
DATAWorks 2023: Comparing Normal and Binary D-Optimal Design of Experiments by Statistical Power

5a. CONTRACT NUMBER	**5b. GRANT NUMBER**	**5c. PROGRAM ELEMENT NUMBER**
Separate Contract		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
	C9082	

6. AUTHOR(S)
Adams, Addison, D.; Miller, Curtis, G.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
Institute for Defense Analyses 730 East Glebe Road Alexandria, Virginia 22305	NS D-33405 H 2023-000055

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**	**11. SPONSOR/MONITOR'S REPORT NUMBER**

12. DISTRIBUTION/AVAILABILITY STATEMENT
This publication has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES
Project Leader: Medlin, Rebecca M.

14. ABSTRACT
In many Department of Defense (DoD) Test and Evaluation (T&E) applications, binary response variables are unavoidable. Many have considered D-optimal design of experiments (DOEs) for generalized linear models (GLMs). However, little consideration has been given to assessing how these new designs perform in terms of statistical power for a given hypothesis test. Monte Carlo simulations and exact power calculations suggest that D-optimal designs generally yield higher power than binary D-optimal designs, despite using logistic regression in the analysis after the data have been collected. Results from using statistical power to compare designs contradict traditional DOE comparisons which employ D-efficiency ratios and fractional design space (FDS) plots. Power calculations suggest that practitioners that are primarily interested in the resulting statistical power of a design should use normal D-optimal designs over binary D-optimal designs when logistic regression is to be used in the data analysis after data collection.

15. SUBJECT TERMS
Design of Experiment; Operational Testing; Statistics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	SAR	

19a. NAME OF RESPONSIBLE PERSON	**19b. PHONE NUMBER**
Rebecca Medlin	703-845-6731