INSTITUTE FOR DEFENSE ANALYSES

Sample Size Determination for Computer Simulations with Binary Outcomes

Rebecca M. Medlin, Project Leader

Kelly Duffy
Curtis G. Miller

OED Final

August 2024

IDA Product 3002814

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

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task C9082, “Cross-Divisional 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. Walter Dodson and consisted of Dr. John Haman, Dr. Breeana G Anderson, Mr. Gregory L Chesterton, Dr. Addison Adams, and Dr. Tyler Morgan-Wall, from the Operational Evaluation Division, and Dr. John Dennis from Strategy, Forces & Resources Division.

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

Dr. Heather M. Wojton, Director, Operational Evaluation Division
hwojton@ida.org • (703) 845-6811

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Product 3002814

Sample Size Determination for Computer Simulations with Binary Outcomes

Rebecca M. Medlin, Project Leader

Kelly Duffy
Curtis G. Miller

Executive Summary

Modeling and simulation (M&S) tools must undergo a process of verification, validation, and accreditation (VV&A) before their outputs can be trusted. The VV&A may involve collecting outputs from M&S tools and studying the outputs with statistical methods to understand the predictions generated by M&S tools. Space-filling designs of experiments (SFD DOE) explore the operational space of a defense-related M&S tool, then a statistical linear model or generalized linear model (GLM) may be fit to the resultant outputs to produce a metamodel describing typical outputs under factor combinations. When the response of interest is a binary outcome, such as whether a torpedo hit its target or not, we may use logistic regression GLMs to describe the relationship between factors and the probability of hit. However, when planning M&S studies, we need to determine whether an SFD can be used for estimating typical responses with an acceptable level of uncertainty.

If the statistical goal is to predict average outputs well and control the uncertainty in what the typical predictions given a set of factors will be, the statistical measure of merit we wish to control is confidence interval (CI) margin of error (MOE). Controlling the MOE, though, can be difficult in the logistic regression context since the MOE depends on the full relationship between the probability the event of interest happens (such as the probability a torpedo hits its target) and the factors under study (such as distance from the target, the target's speed, the sound velocity profile, and so on). In this presentation, we study how to control the MOE on predicted probabilities using an upper bound for the MOE. The upper bound depends on the sample size, confidence level, and design of experiments, as would any MOE. However, the MOE requires assuming only what range of probabilities we will observe rather than knowing all the probabilities at all the design points, hence loosening the assumptions a user of the upper bound needs to use the method.

Our Monte Carlo simulation studies show that the average MOE in an SFD DOE can be controlled well with our proposed upper bound. The upper bound tends to imply MOEs will be wider than actually seen in samples, thus raising the risk of collecting more data than needed to characterize an M&S tool's outputs. However, it is robust to violations of its assumptions; in particular, if the assumed range of probabilities does not capture all of the probabilities in the sample, the method still captures the final average MOEs well.

While the upper bound may be safe to use when no other method for finding sample sizes for studies involving SFDs exists, it deserves further theoretical study in order to fully understand the upper bound’s strengths and weaknesses; in particular how to characterize the probability of upper bound violation.

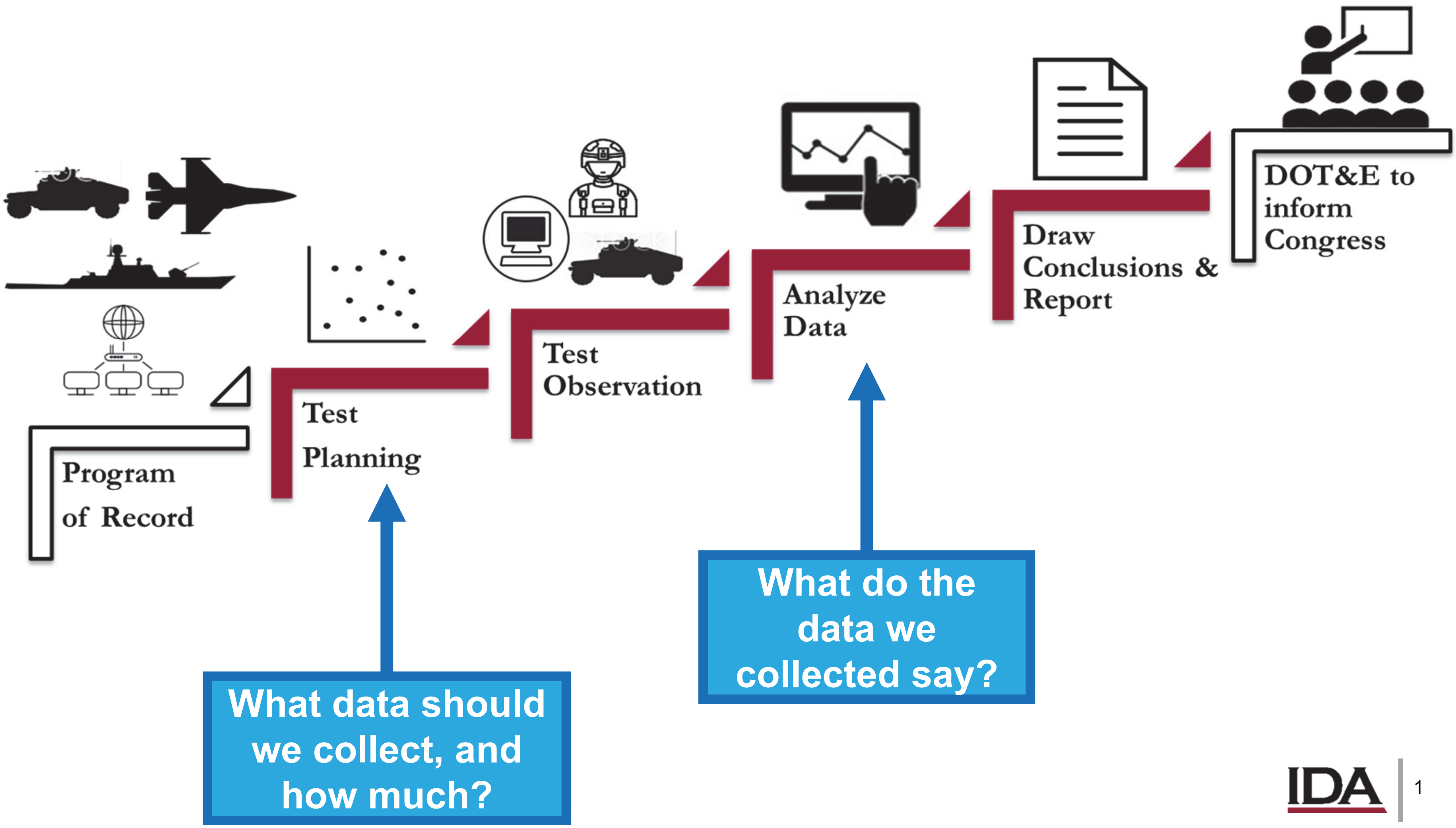
Sample Size Determination for Computer Simulations with Binary Outcomes

Curtis Miller
Kelly Duffy (University of Minnesota)

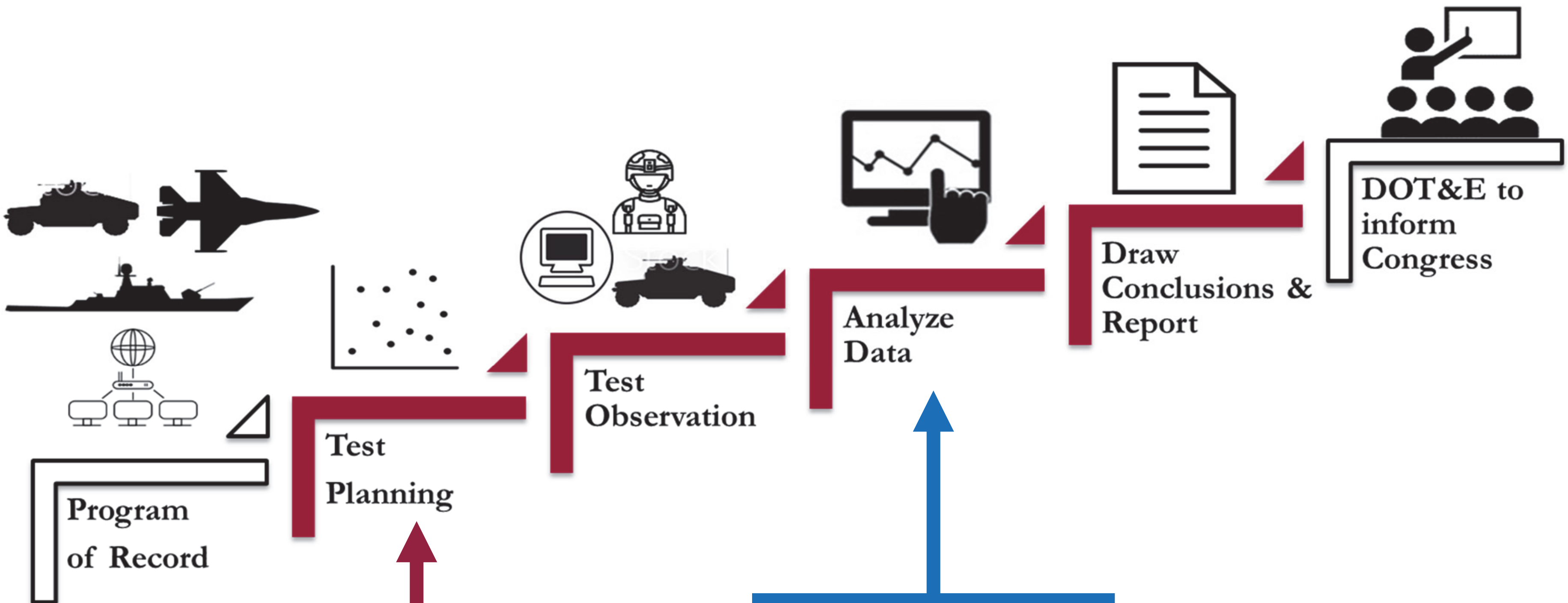
August 6, 2024

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

IDA's Operational Evaluation Division (OED) Provides Technical and Analytical Support to DOT&E



IDA's Operational Evaluation Division (OED) Provides Technical and Analytical Support to DOT&E

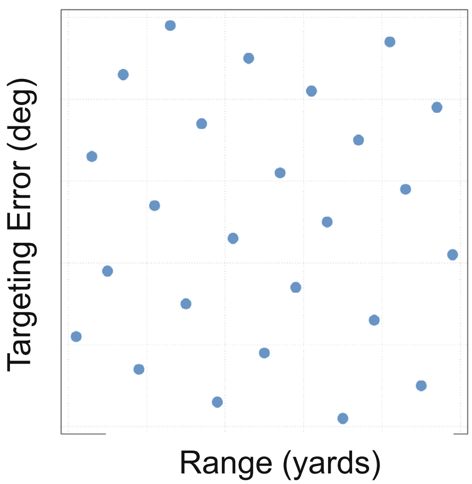


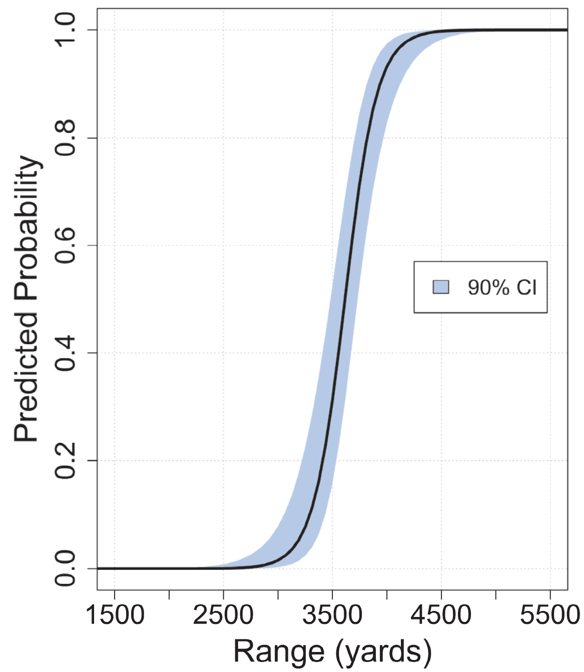
TODAY'S TALK

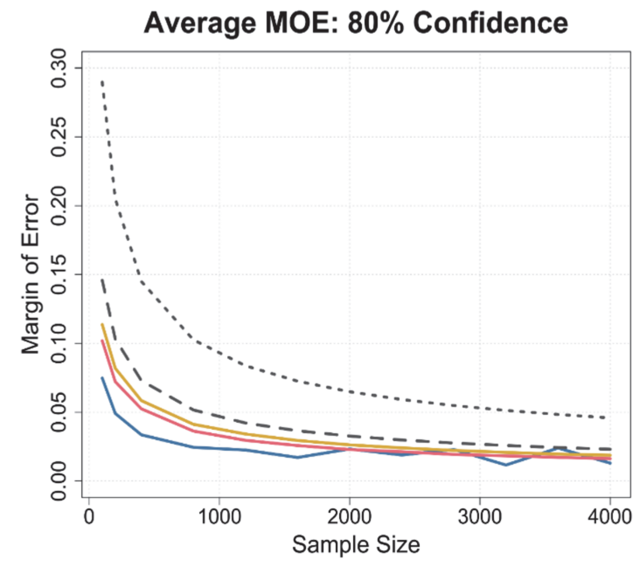
What data should we collect, and how much?

What does the data we collected say?

Bottom Line Up Front







M&S designs should ideally be space-filling, current guidance to select number of runs for these is lacking

Operationally useful measure: statistical precision of estimates

Simulation study indicates our approach to estimate statistical precision before data collection is a good way to size a test

CI: Confidence Interval; M&S: Modeling and Simulation; MOE: Margin of Error

Modeling and Simulation (M&S) Can Supplement Live Data

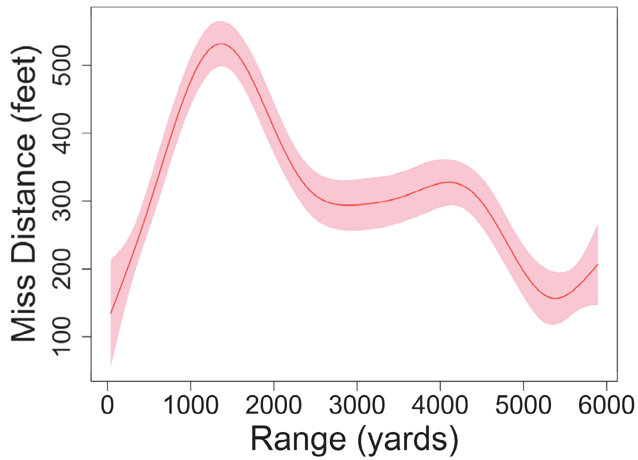
Two potential goals we might have:

- Compare live testing data to M&S
- **Explore and evaluate the behavior of the M&S itself**

A good M&S model would allow us to **make predictions across the entire space** of relevant variables:

- Test conditions that cannot be tested in real life
- **Construct a meta-model** that could potentially be used as a surrogate for the M&S itself, saving time and money

Notional Example



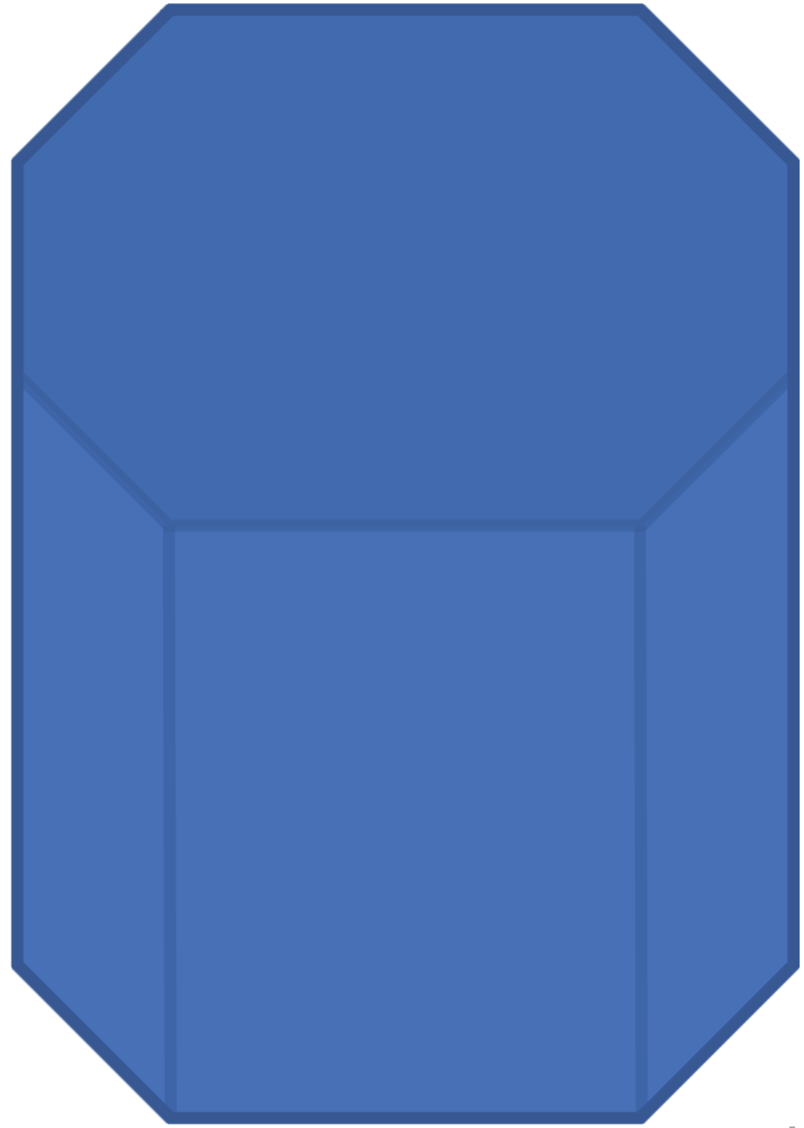
ECWAF Simulates Torpedo Performance

We are interested in running a simulation in the ECWAF where we vary several factors:

- <u>Range</u> (continuous)
- <u>Targeting error</u> (continuous)
- <u>Environment</u> (categorical: open ocean or littoral)
- <u>Target type</u> (categorical: SSN, SSK)

In each run of the simulation, the outcome is either hit or miss (*binary*)

Our primary interest is the **probability of a hit** given the conditions



ECWAF: Environment Centric Weapons Analysis Facility;
SSK: diesel-electric submarine; SSN: nuclear-powered submarine

Optimizing the Number of Runs in M&S is Essential to Both Understand Design Space and Conserve Resources

The ECWAF is a hardware-in-the-loop simulation:

- Runs in real time (~30 minutes/run)

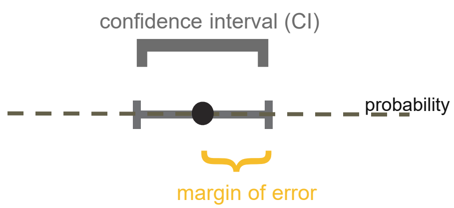
Aim: What number of runs is adequate?

ECWAF: Environment Centric Weapons Analysis Facility; M&S: Modeling and Simulation
Image from Naval Undersea Warfare Center, 2014

A Desired Uncertainty Level, Study Design, and Analysis Model are Needed to Determine the Ideal Number of Runs

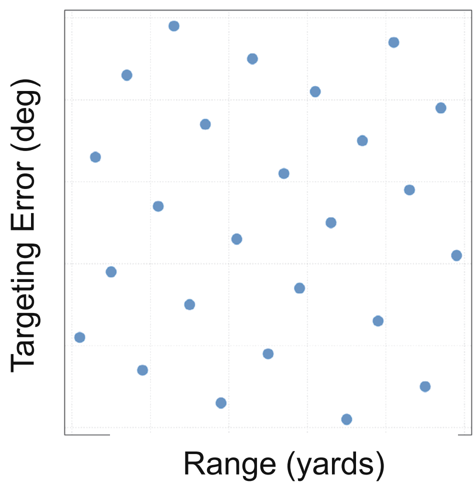
Desired Uncertainty

Margin of Error



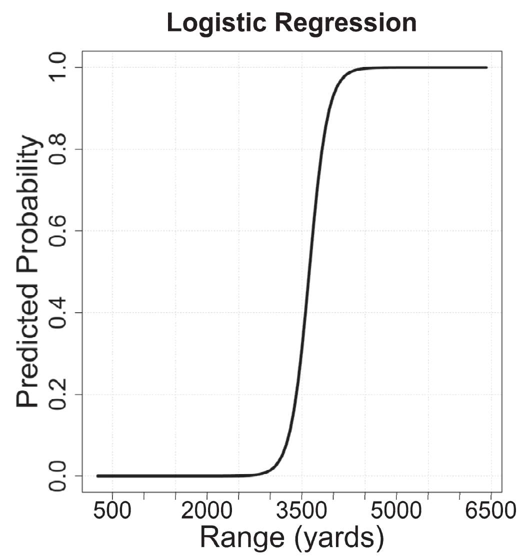
Study Design

Space-Filling Designs



Analysis Model

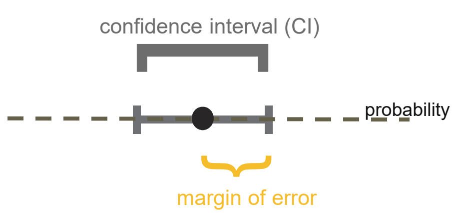
Logistic Regression



A Desired Uncertainty Level, Study Design, and Analysis Model are Needed to Determine the Ideal Number of Runs

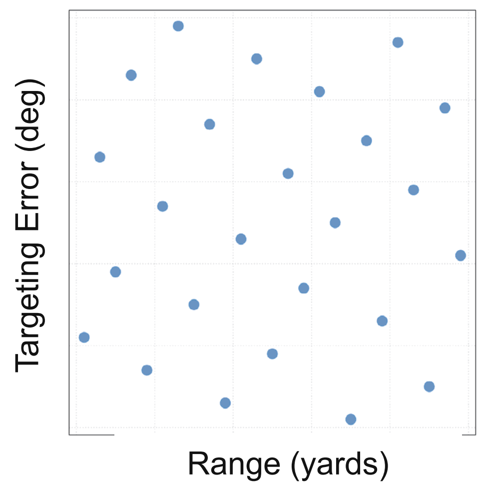
Desired Uncertainty

Margin of Error



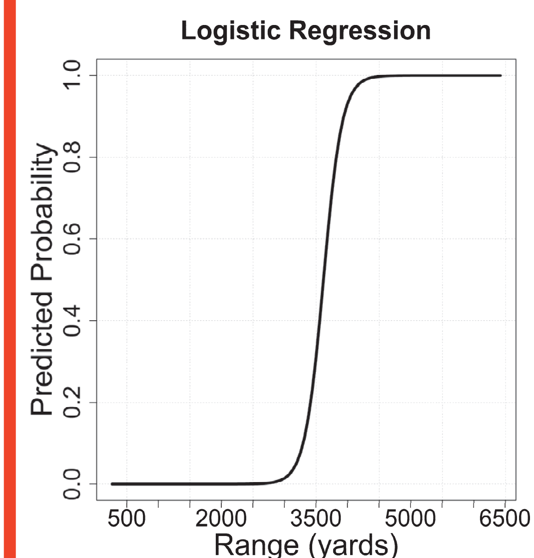
Study Design

Space-Filling Designs



Analysis Model

Logistic Regression

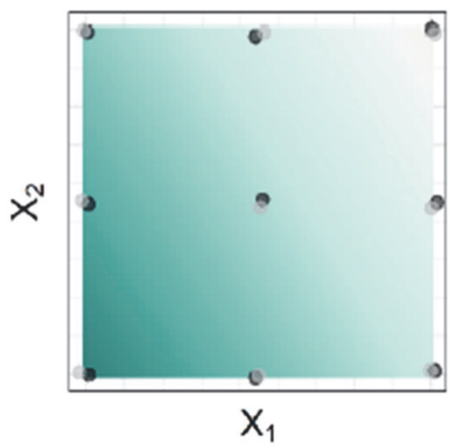


Space-Filling Designs Better Capture Deviations from Linearity

Space-filling designs (SFDs) are experimental designs that are a general principled approach to “fill” the design space

Space-Filling Designs Better Capture Deviations from Linearity

Space-filling designs (SFDs) are experimental designs that are a general principled approach to “fill” the design space

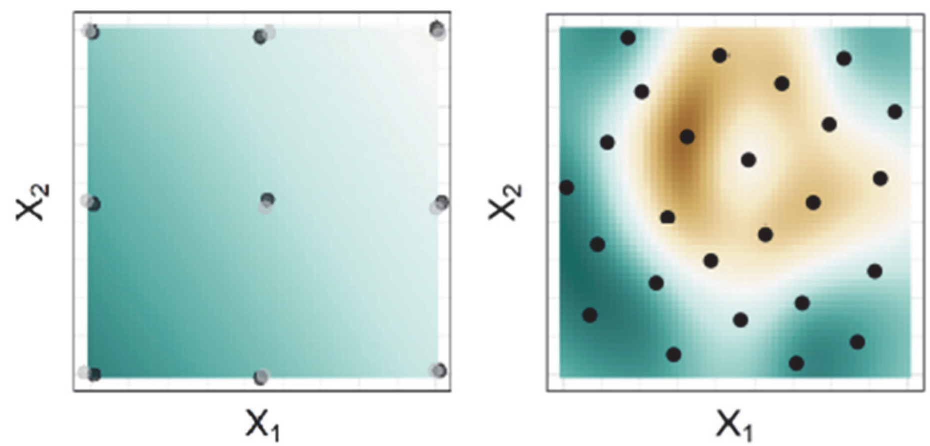


“Classical” Design
(*D*-optimal)

Figure adapted from Wojton et al. (2021)

Space-Filling Designs Better Capture Deviations from Linearity

Space-filling designs (SFDs) are experimental designs that are a general principled approach to “fill” the design space



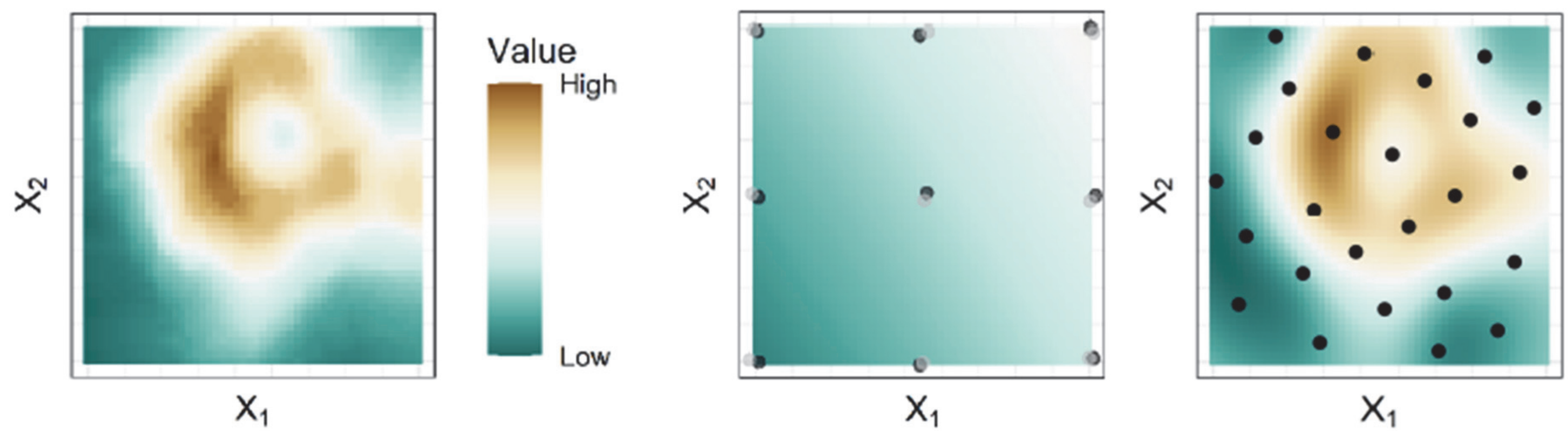
“Classical” Design (*D*-optimal)

Space-Filling Design

Figure adapted from Wojton et al. (2021)

Space-Filling Designs Better Capture Deviations from Linearity

Space-filling designs (SFDs) are experimental designs that are a general principled approach to “fill” the design space



Underlying **Truth**

“Classical” Design (*D*-optimal)

Space-Filling Design

Figure adapted from Wojton et al. (2021)

Current Available Approaches are Not Sufficient

Some rules of thumb are available for choosing the number of runs in space-filling designs:

- e.g., 10*d* where *d* is the number of design factors[1,2]
  - Not rigorously evaluated

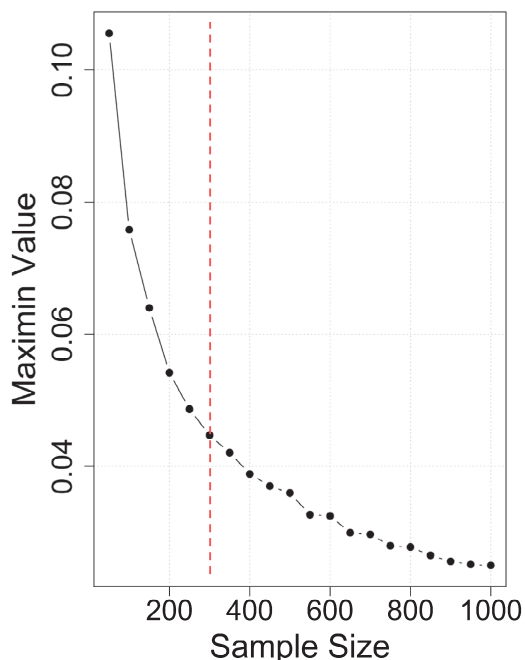
[1] Loeppky et al., 2009; [2] Wojton et al., 2021

Current Available Approaches are Not Sufficient

Some rules of thumb are available for choosing the number of runs in space-filling designs:

- e.g., 10*d* where *d* is the number of design factors[1,2]
  - Not rigorously evaluated
- e.g., Looking for the "elbow" in a plot of a space-filling metric like "maximin" used in creating the design[2]
  - This is a design-based metric, does **not** tell us anything about *precision of results*

Maximin Values Across Sample Sizes



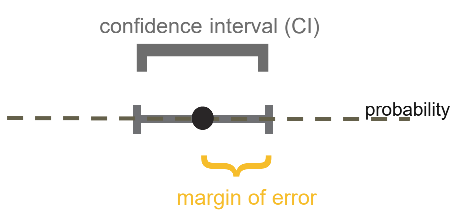
Maximin: A measure of the *minimum* pairwise distance between design points (after *maximization*)

[1] Loeppky et al., 2009; [2] Wojton et al., 2021

A Desired Uncertainty Level, Study Design, and Analysis Model are Needed to Determine the Ideal Number of Runs

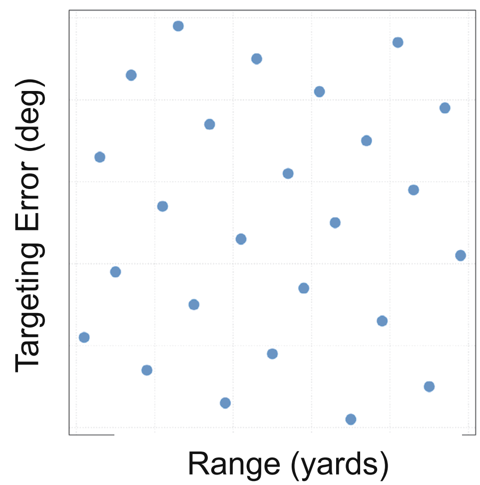
Desired Uncertainty

Margin of Error



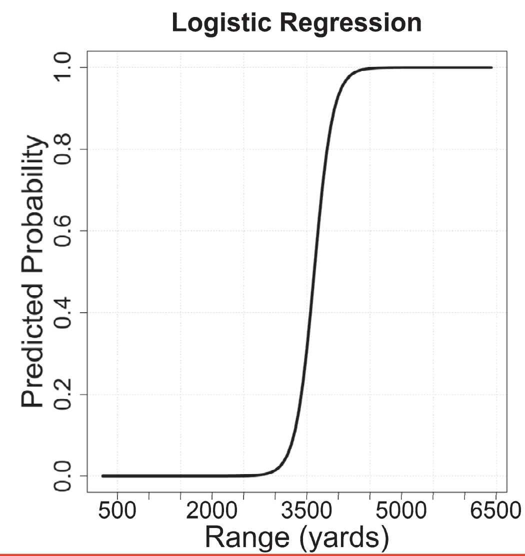
Study Design

Space-Filling Designs

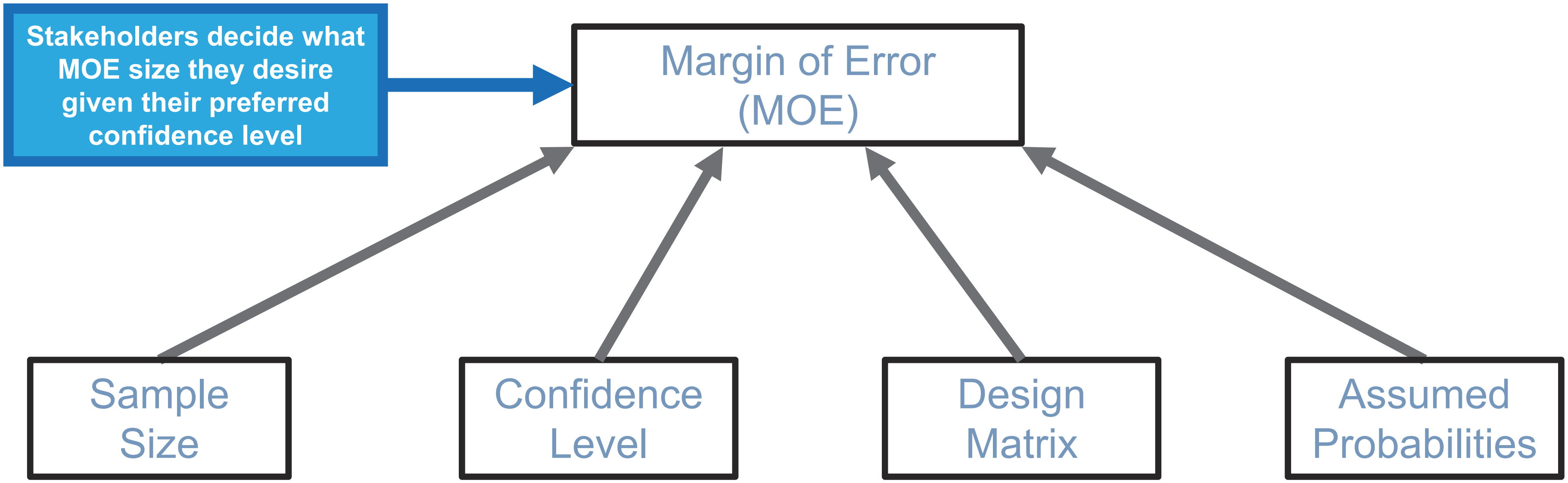


Analysis Model

Logistic Regression



Factors That Affect Confidence Interval Margin of Error



Sample Size

sample size ↑, MOE ↓

What we want to determine based on the other factors

Confidence Level

confidence ↑, MOE ↑

Known, specified in formula

Design Matrix

complicated

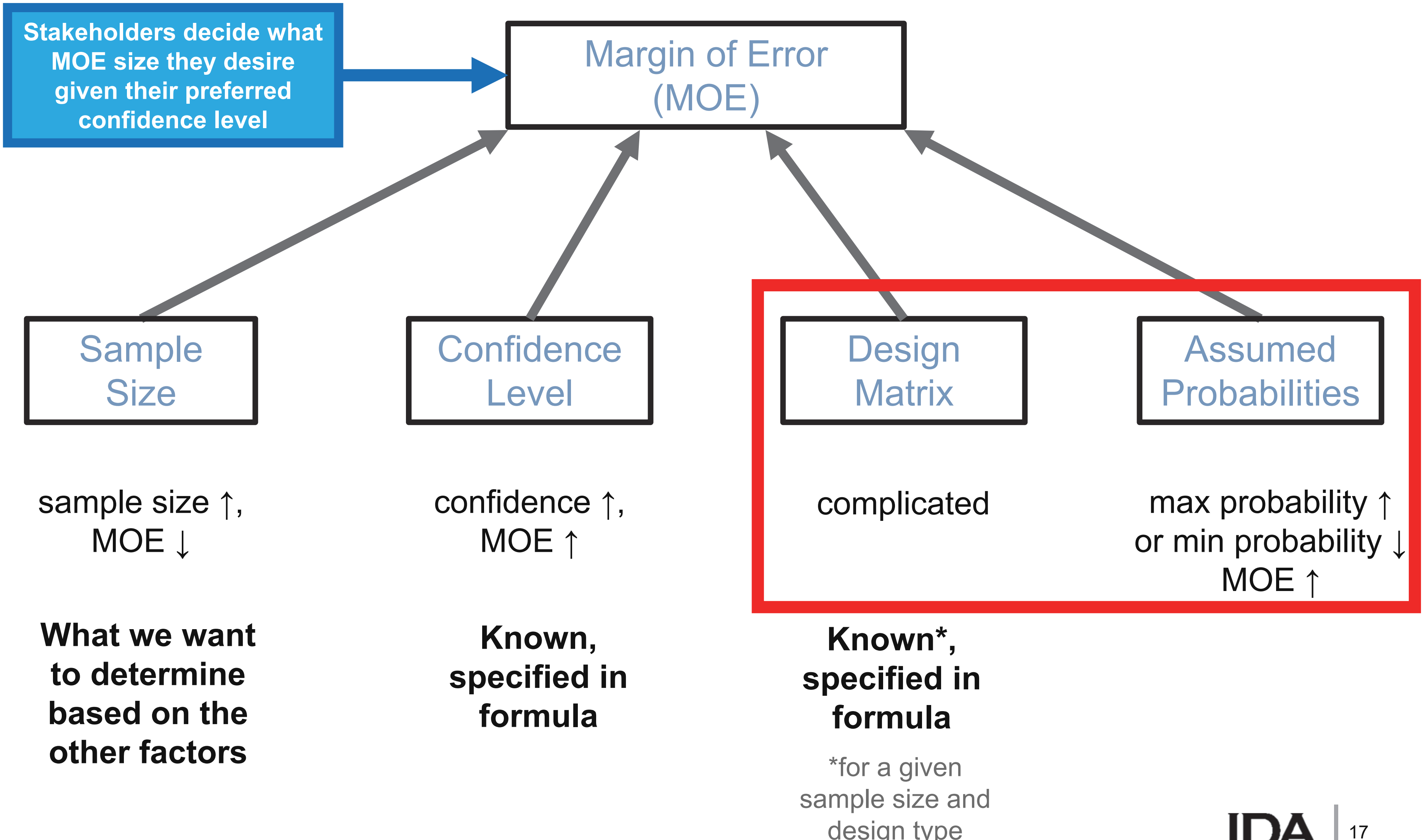
Known*, specified in formula

*for a given sample size and design type

Assumed Probabilities

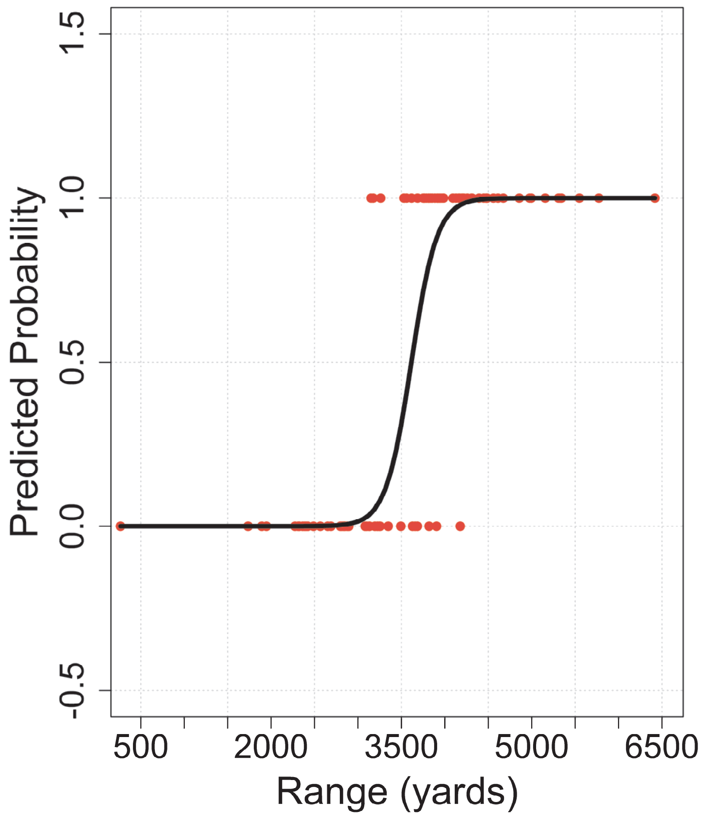
max probability ↑ or min probability ↓ MOE ↑

Factors That Affect Confidence Interval Margin of Error



Logistic Regression Models are Used to Study Torpedo Hits and Misses

Logistic regression models are used to estimate probabilities (constrained between 0 and 1) when we have a binary response

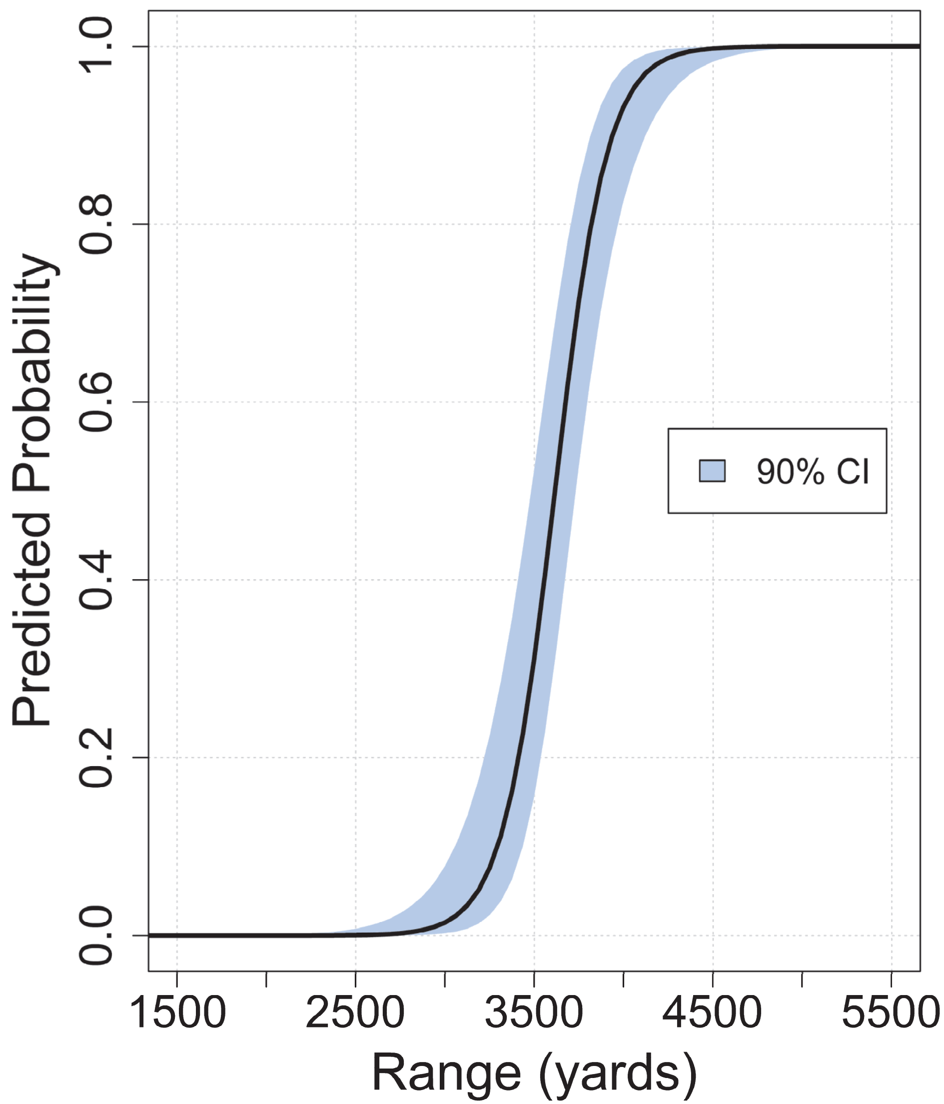


$$\text{hit\_prob}_i = \eta(\beta_0 + \beta_1 \text{range}_i)$$

$$\boxed{\eta = \frac{1}{1 + e^{-x}}}$$

Most common and well-known among weapons testing analysts

Confidence Intervals Assess Uncertainty in Logistic Regression Estimates



We Derived an Oracle Upper Bound on the Margin of Error

Variance of Predicted ($\hat{\pi}$) Versus True (π) Probabilities

$$[\pi_i(1-\pi_i)]^2 x_i^T (X_n^T W_n X_n)^{-1} x_i$$

Requires knowing the model parameters

W_n is the weight matrix $\mathrm{diag}(\pi_i(1-\pi_i): 1 \leq i \leq n)$, where π_i is the probability at the design points (here, diag means diagonal matrix with listed entries)

x_i is the vector of points in the design space at which predictions $\hat{\pi}$ are made

$X_n = (x_1, \dots, x_n)^T$ is the design matrix

We Derived an Oracle Upper Bound on the Margin of Error

Variance of Predicted ($\hat{\pi}$) Versus True (π) Probabilities

$$[\cancel{\pi}_i(1-\cancel{\pi}_i)]^2 x_i^T (X_n^T W_n X_n)^{-1} x_i$$

$$\boxed{\leq \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} x_i (X_n^T X_n)^{-1} x_i^T}$$

Oracle Upper Bound

W_n is the weight matrix $\mathrm{diag}(\pi_i(1-\pi_i): 1 \leq i \leq n)$, where π_i is the probability at the design points (here, diag means diagonal matrix with listed entries)

x_i is the vector of points in the design space at which predictions $\hat{\pi}$ are made

$X_n = (x_1, \dots, x_n)^T$ is the design matrix

We Derived an Oracle Upper Bound on the Margin of Error

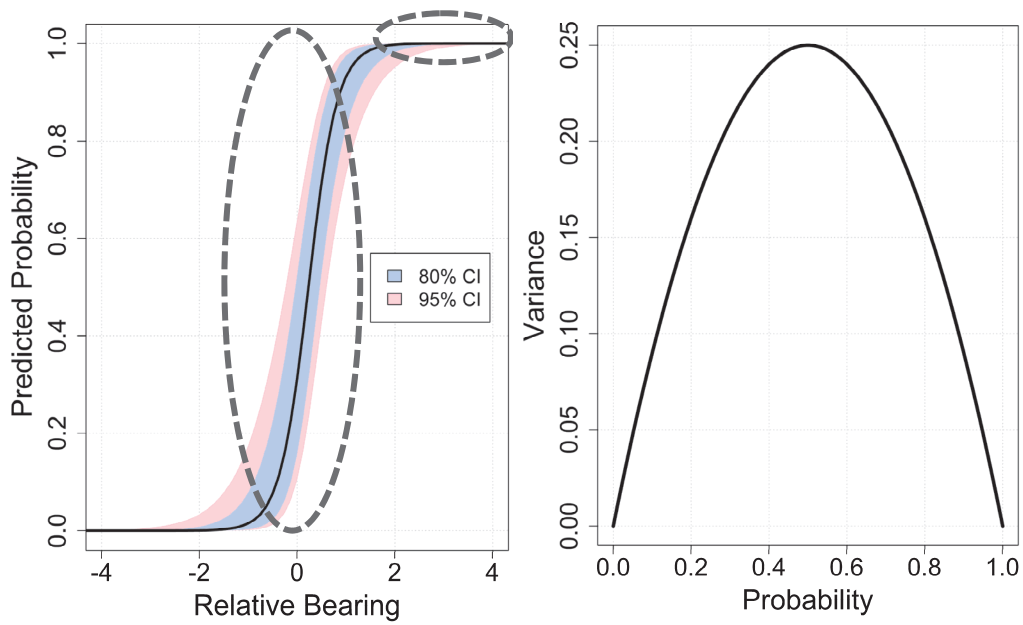
$$\frac{1/4^2}{\underline{\pi}(1-\underline{\pi})}\mathrm{diag}(X_n(X_n^T X_n)^{-1}X_n^T)$$

$\boldsymbol{\underline{\pi}}(\mathbf{1}-\boldsymbol{\underline{\pi}})$ is the smallest variance in the weight matrix (determined by π_i or $1-\pi_i$)

X_n is the matrix of n points in the design space at which predictions $\hat{\pi}$ are made

Here diag means the diagonal entries of a matrix

MOEs Tend to Be **Wider** for Probabilities Near the **Middle**



CI: Confidence Interval; MOE: Margin of Error

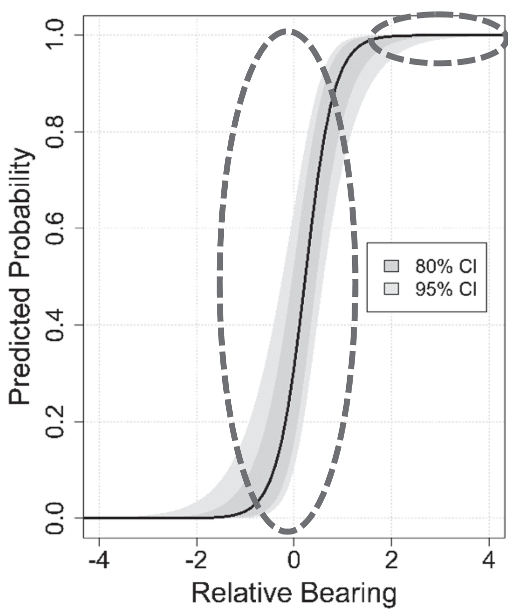
We Derived an Oracle Upper Bound on the Margin of Error

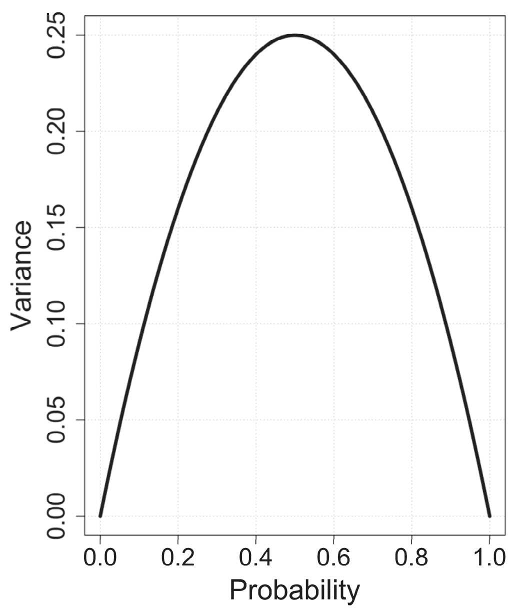
$$\frac{1/4^2}{\underline{\pi}(1-\underline{\pi})}\mathrm{diag}(X_n(X_n^T X_n)^{-1}X_n^T)$$

$\underline{\boldsymbol{\pi}}(\mathbf{1}-\underline{\boldsymbol{\pi}})$ is the smallest variance in the weight matrix (determined by π_i or $1-\pi_i$)

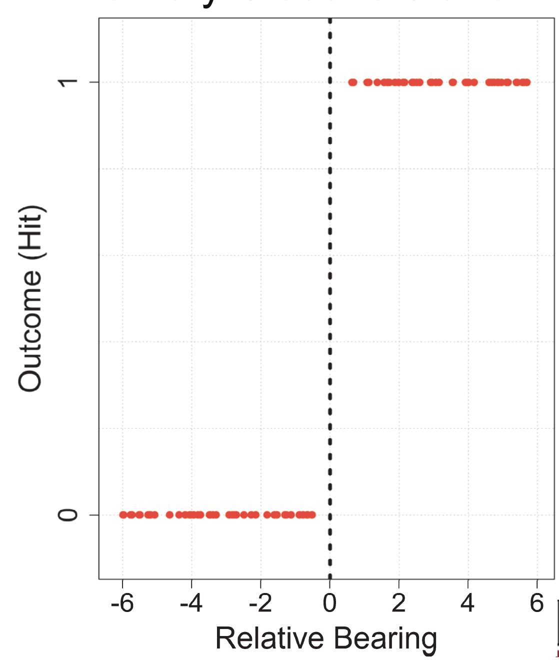
X_n is the matrix of n points in the design space at which predictions $\hat{\pi}$ are made

MOEs Tend to Be **Wider** for Probabilities Near the **Middle**



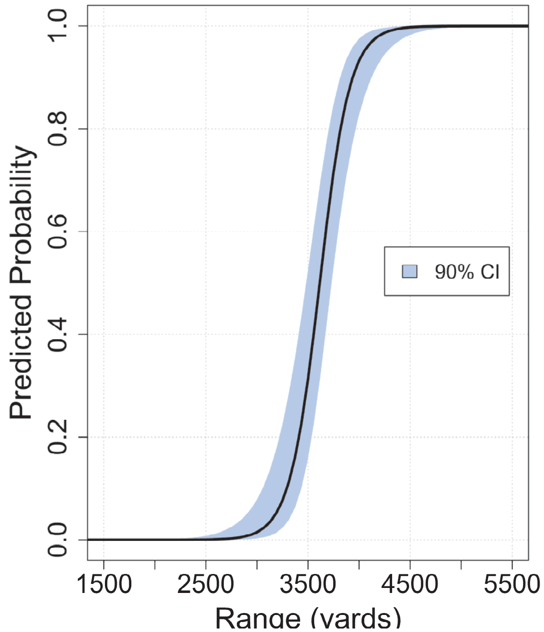


Logistic Regression MLEs May Not Exist when **All Probabilities** Are Very Close to **0 and 1**



CI: Confidence Interval; MLE: Maximum Likelihood Estimate; MOE: Margin of Error

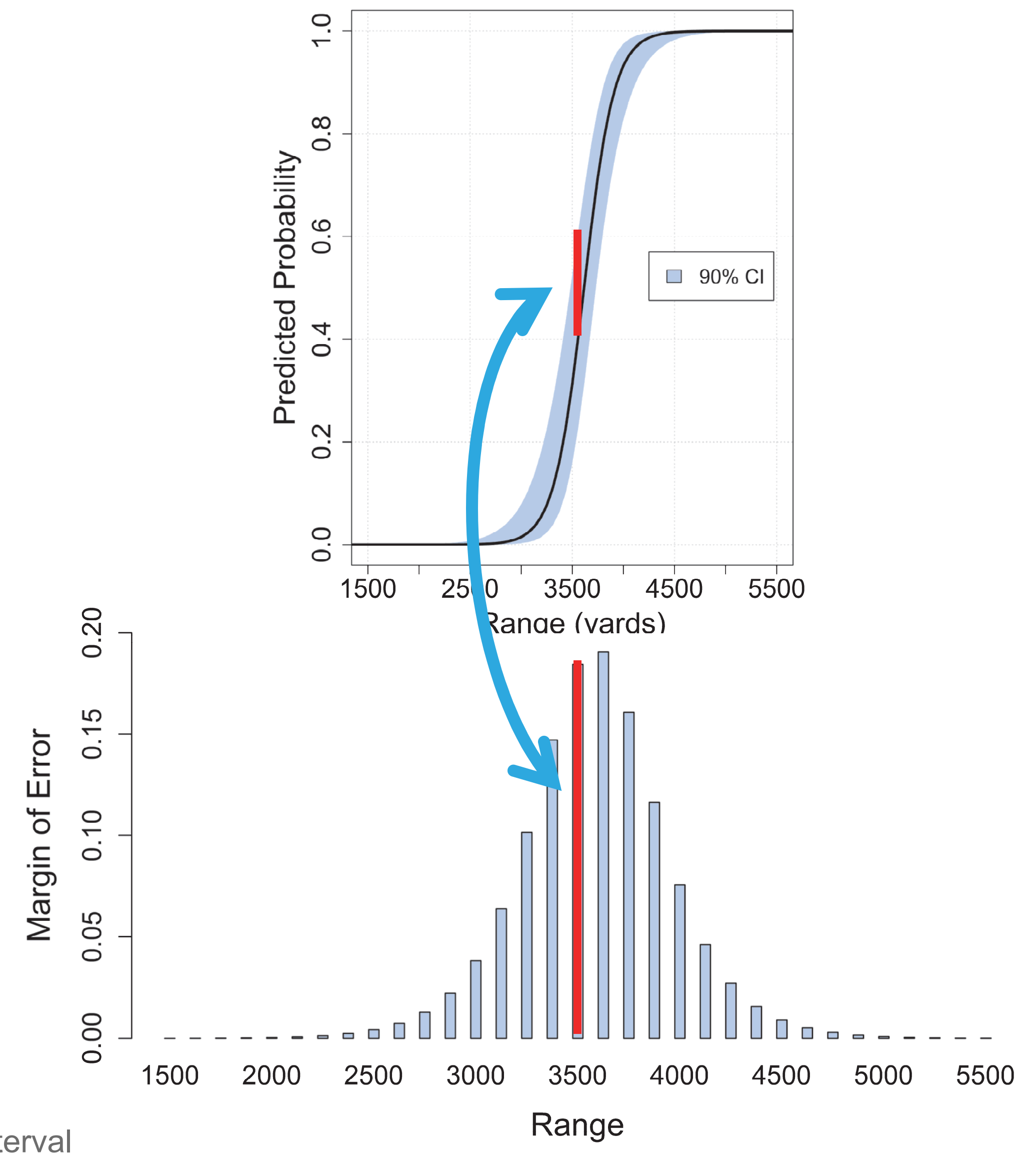
An Oracle Upper Bound on the Margin of Error Can Be Determined for Each Point Before Data Collection



Estimating the margin of error **requires knowing the model**

CI: Confidence Interval

An Oracle Upper Bound on the Margin of Error Can Be Determined for Each Point Before Data Collection

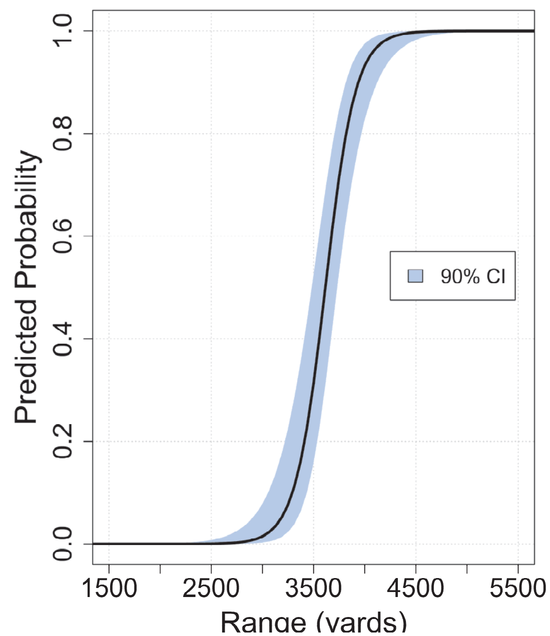


Estimating the margin of error **requires knowing the model**

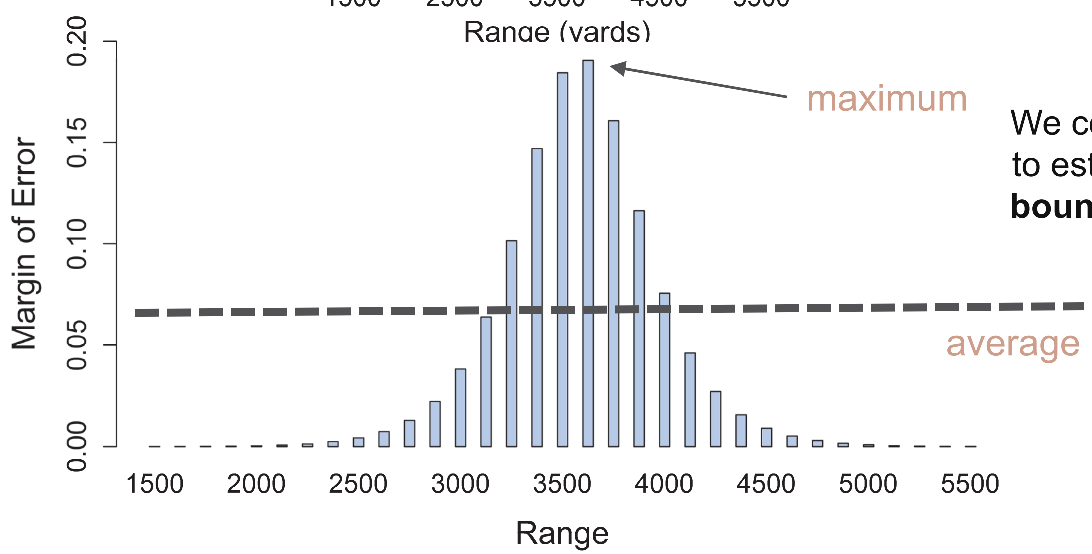
We computed an expression to estimate an **oracle upper bound** to the margin of error at each point

CI: Confidence Interval

An Oracle Upper Bound on the Margin of Error Can Be Determined for Each Point Before Data Collection

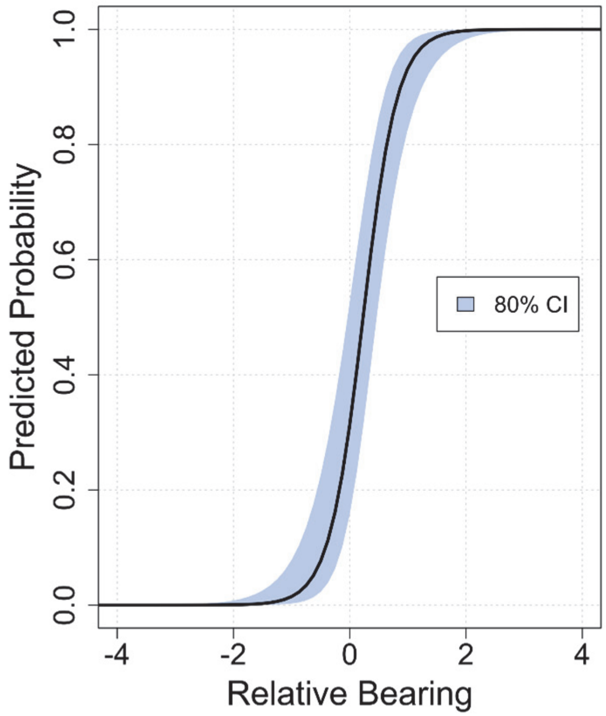


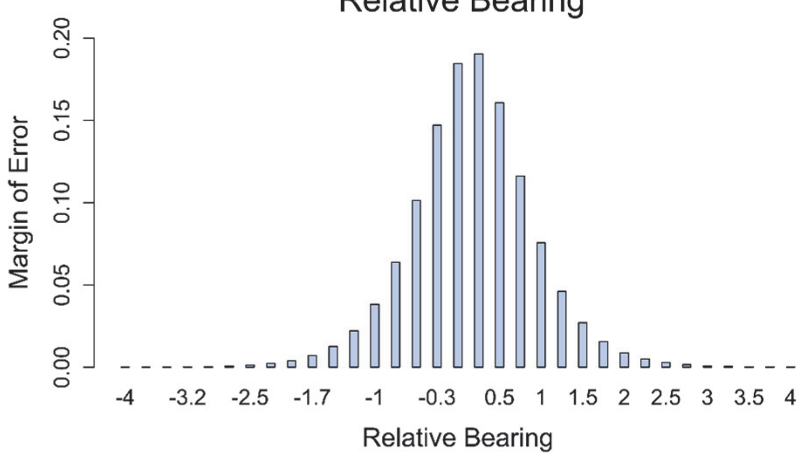
Estimating the margin of error **requires knowing the model**



We computed an expression to estimate an **oracle upper bound** to the margin of error at each point

Oracle Upper Bound on the Margin of Error Evaluates All Points in the Design Space





CI: Confidence Interval

$$W_n X_n^T (X_n^T W_n X_n)^{-1} X_n W_n \preccurlyeq \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} X_n (X_n^T X_n)^{-1} X_n^T$$

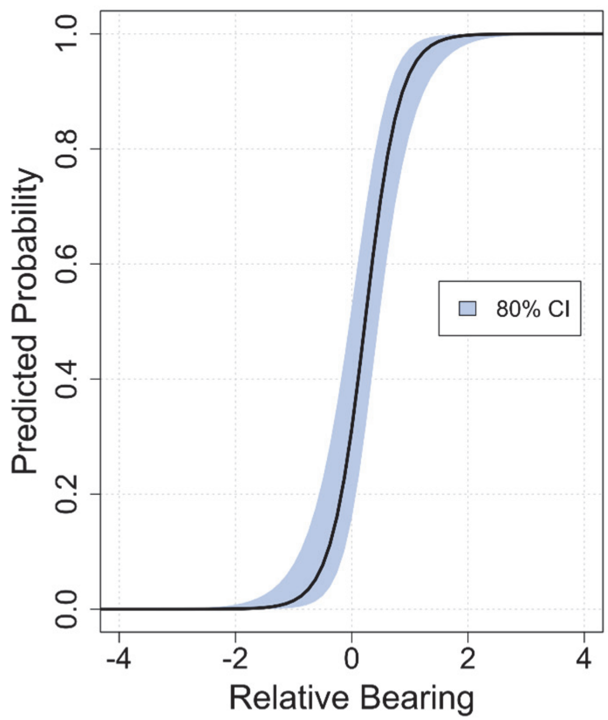
W_n is the weight matrix $\mathrm{diag}(\pi_i(1-\pi_i): 1 \leq i \leq n)$, where π_i is the probability at the design points (here, diag means diagonal matrix with listed entries)

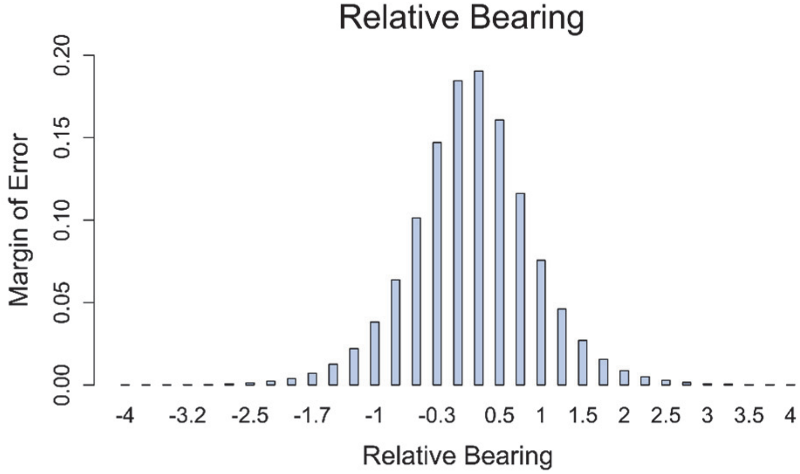
X_n is the matrix of n points in the design space at which predictions $\hat{\pi}$ are made

$\underline{\pi}(1-\underline{\pi})$ is the smallest variance in the weight matrix W_n

$\preccurlyeq$ refers to Löwner partial ordering for matrices

Oracle Upper Bound on the Margin of Error Evaluates All Points in the Design Space





CI: Confidence Interval

$$W_n X_n^T (X_n^T W_n X_n)^{-1} X_n W_n \preccurlyeq \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} X_n (X_n^T X_n)^{-1} X_n^T$$

This upper bound may be exceeded when we estimate W_n in a sample

W_n is the … $\le$ $i \le n$), wh… design po… matrix with listed entries)

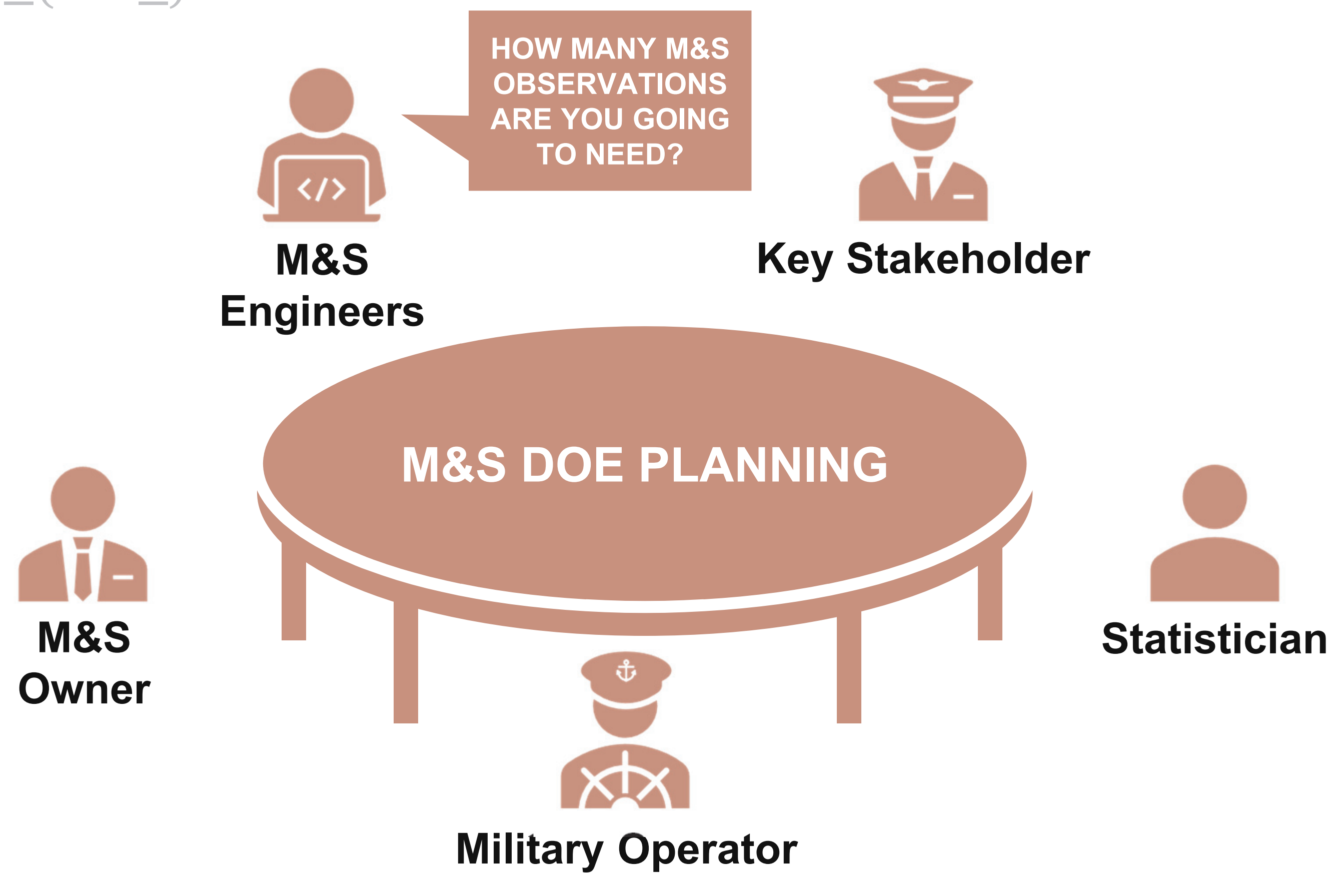
X_n is the matrix of n points in the design space at which predictions $\hat{\pi}$ are made

$\underline{\pi}(1-\underline{\pi})$ is the smallest variance in the weight matrix W_n

$\preccurlyeq$ refers to Löwner partial ordering for matrices

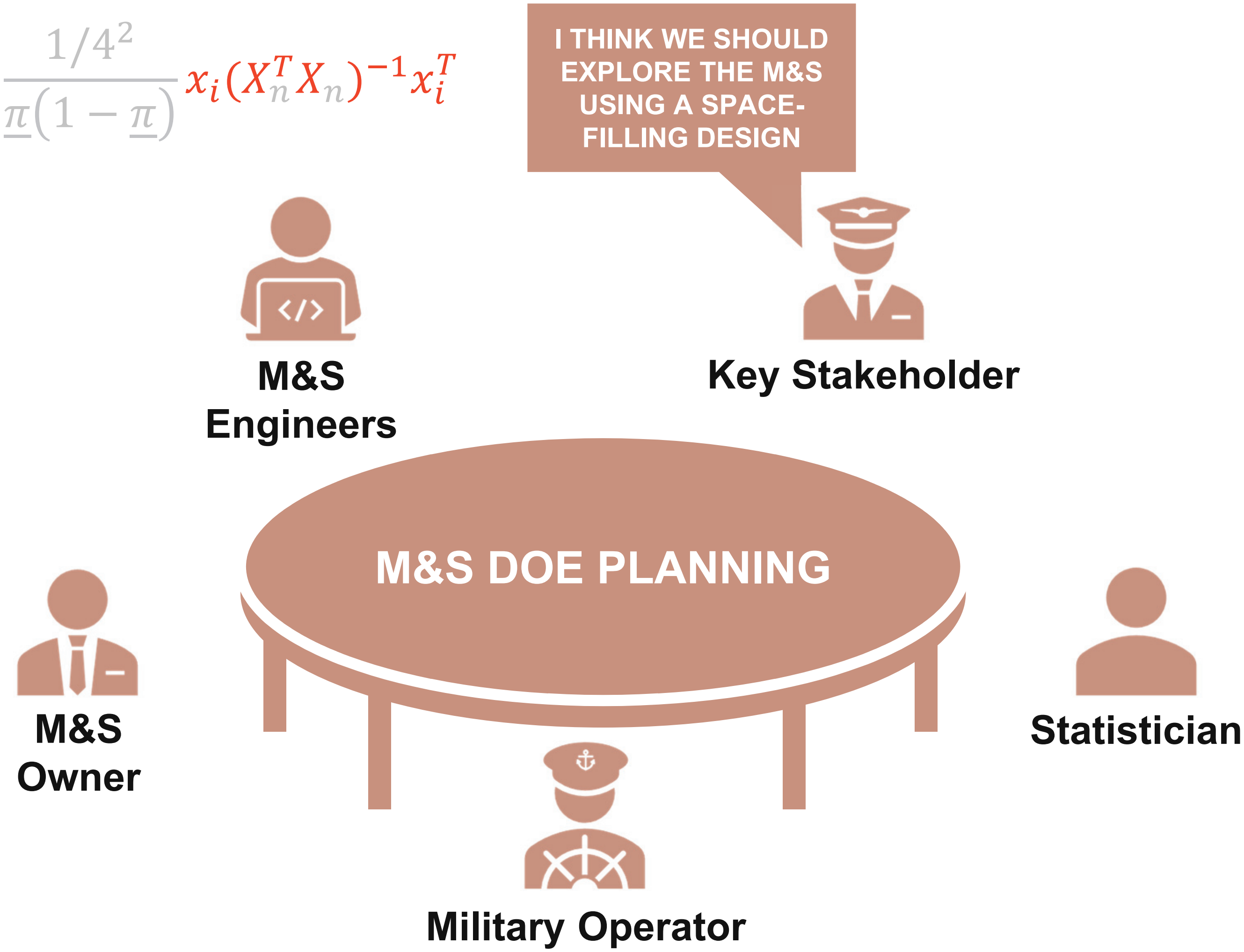
The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

$$\overline{m} \le \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} x_i (X_n^T X_n)^{-1} x_i^T$$



The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

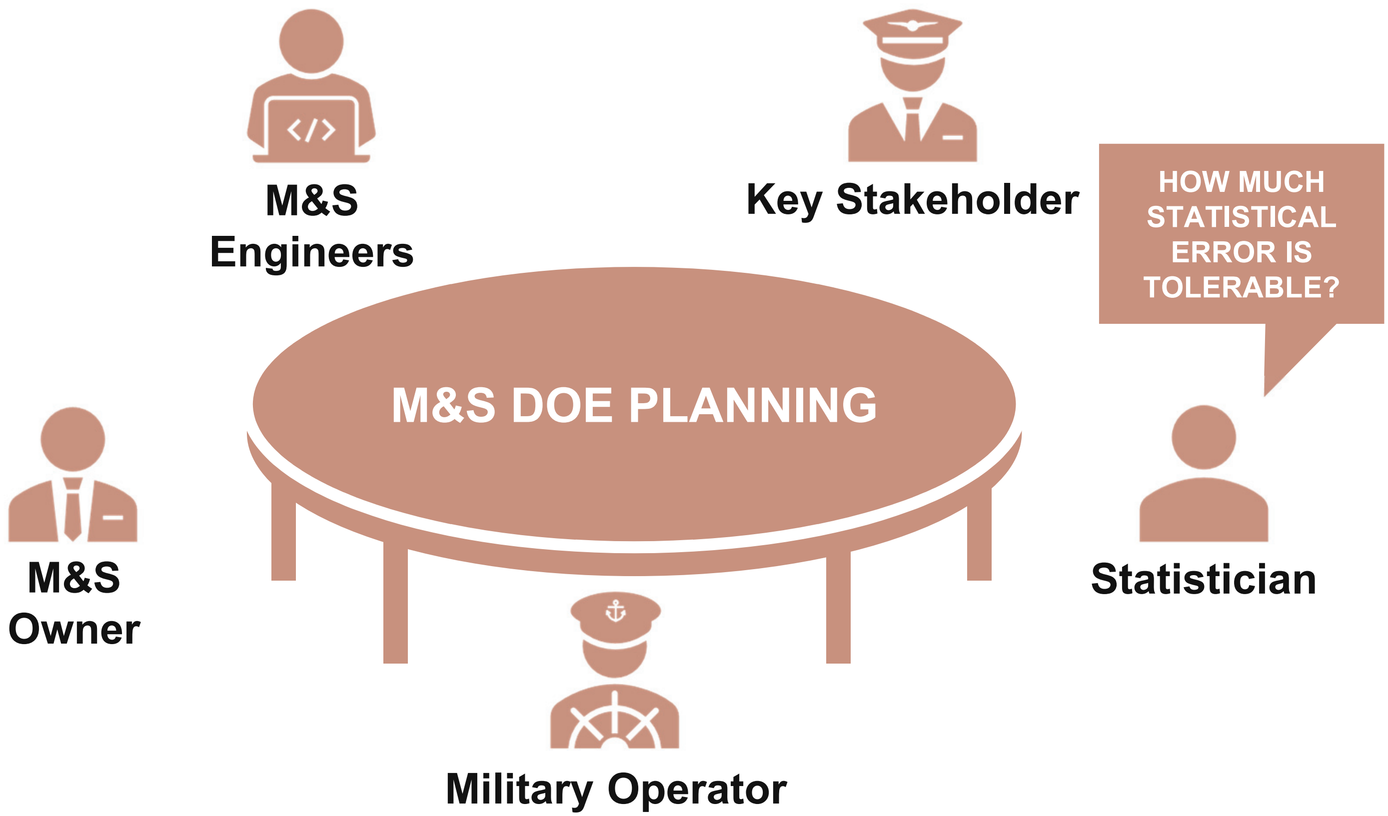
$$\overline{m} \le \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} x_i (X_n^T X_n)^{-1} x_i^T$$



DOE: Design of Experiments; M&S: Modeling and Simulation

The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

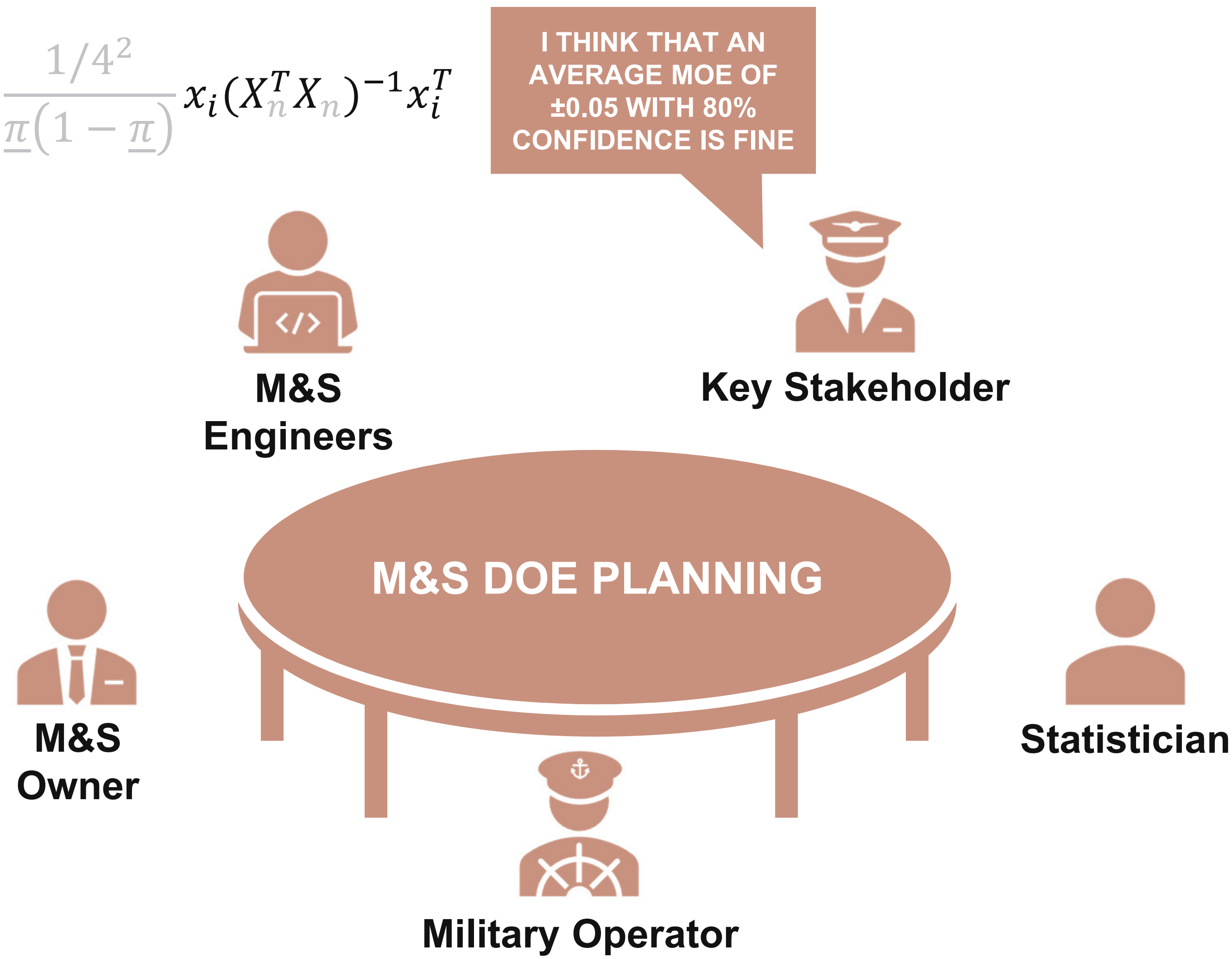
$$\overline{m} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}\left(1-\underline{\pi}\right)} x_i (X_n^T X_n)^{-1} x_i^T$$



DOE: Design of Experiments; M&S: Modeling and Simulation

The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

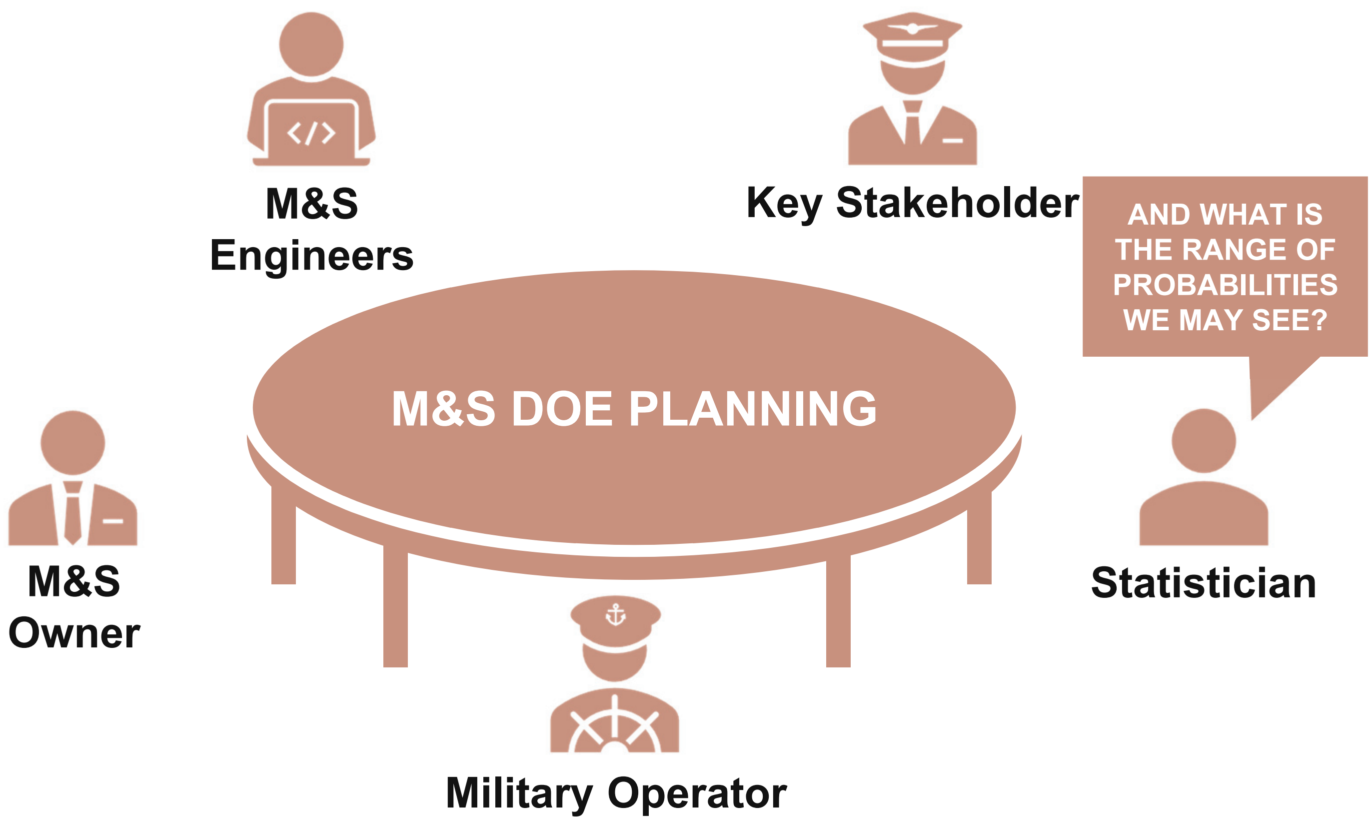
$$\overline{m} \le \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} x_i (X_n^T X_n)^{-1} x_i^T$$



DOE: Design of Experiments; M&S: Modeling and Simulation; MOE: Margin of Error

The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

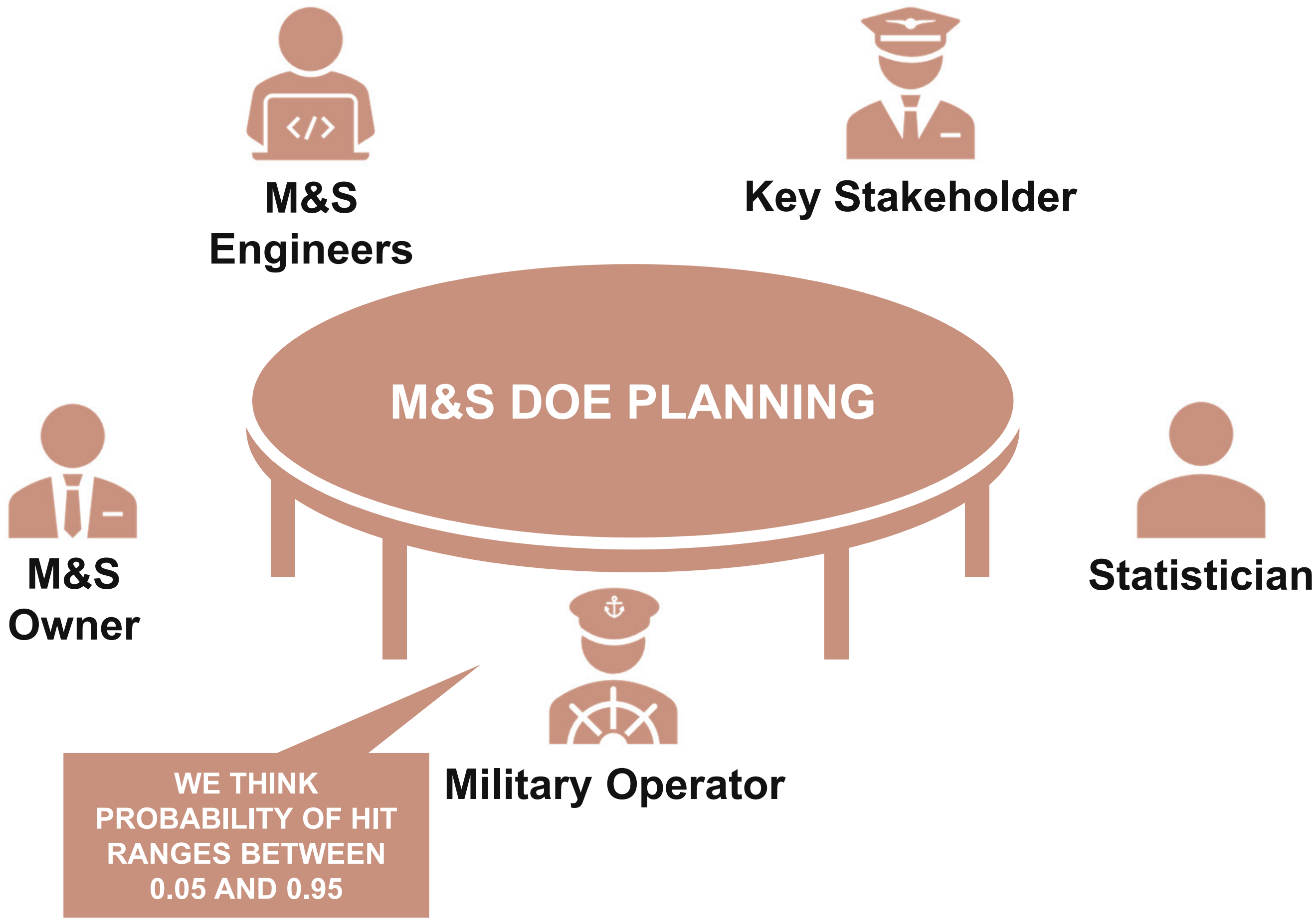
$$\overline{m} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} x_i (X_n^T X_n)^{-1} x_i^T$$



DOE: Design of Experiments; M&S: Modeling and Simulation

The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

$$\overline{m} \leq \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}(1-\underline{\pi})} x_i (X_n^T X_n)^{-1} x_i^T$$

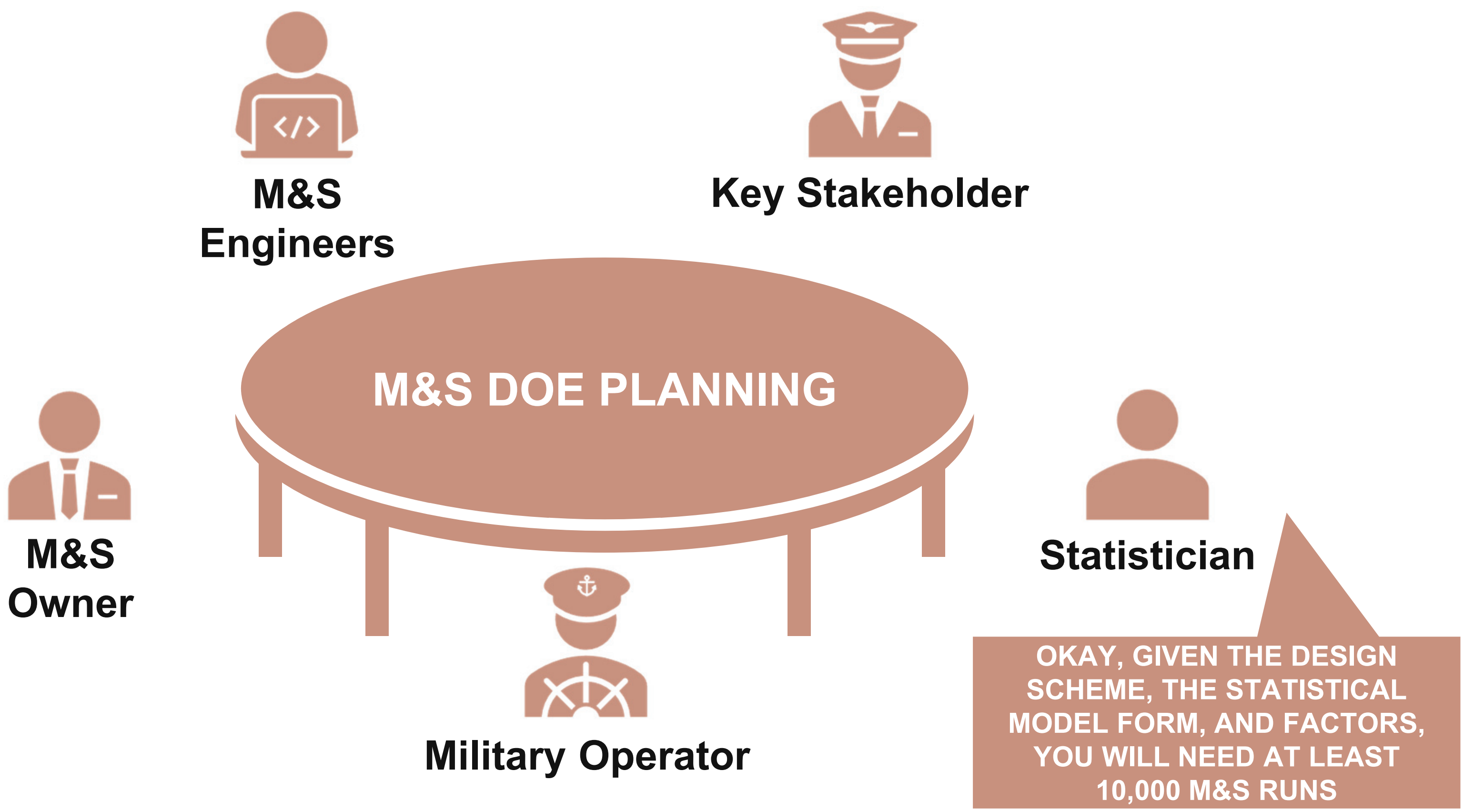


DOE: Design of Experiments; M&S: Modeling and Simulation

The Proposed Method is Easy to Use with Key Stakeholders in the Test Design Process

$$\overline{m} \le \frac{1}{n}\sum_{i=1}^{n} \frac{1/4^2}{\underline{\pi}\left(1-\underline{\pi}\right)} x_i \left(X_n^T X_n\right)^{-1} x_i^T$$

Now use R to generate candidate SFDs for multiple run sizes

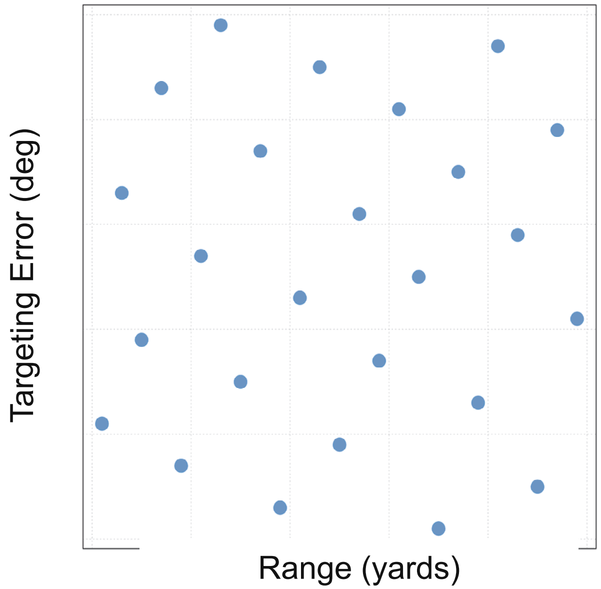


DOE: Design of Experiments; M&S: Modeling and Simulation; SFD: Space-Filling Design

Simulation Design*

Generated fixed design matrixes (X_n) from a MaxPro-optimized[1] space-filling design (2 categorical variables, 2 continuous)†

- Sample sizes of n = 100, 200, 400, 800, … , 3600, 4000



*Additional simulation studies looking at the effect of varying the beta values, number of categorical factors, etc. were conducted but not presented

†The design is a sliced Latin hypersquare design coupled with a factorial design for the categorical variables that then is optimized with the MaxPro metric

[1] See Joseph et al., 2015

Simulation Design*

Generated fixed design matrixes (X_n) from a MaxPro-optimized[1] space-filling design (2 categorical variables, 2 continuous)†

- Sample sizes of *n* = 100, 200, 400, 800, … , 3600, 4000

Evaluated three conditions of true probabilities:

- True probabilities ranged from **0.25-0.75**
- True probabilities ranged from **0.05-0.95**
- True probabilities ranged from **0< to >1**

This is done by generating models with random coefficients (from beta distributions) to represent the (unknown) population model, so the population model has these characteristics in the design space

*Additional simulation studies looking at the effect of varying the beta values, number of categorical factors, etc. were conducted but not presented

†The design is a sliced Latin hypersquare design coupled with a factorial design for the categorical variables that then is optimized with the MaxPro metric

[1] See Joseph et al., 2015

Simulation Design*

Generated fixed design matrixes (X_n) from a MaxPro-optimized[1] space-filling design (2 categorical variables, 2 continuous)†

- Sample sizes of *n* = 100, 200, 400, 800, … , 3600, 4000

Evaluated three conditions of true probabilities:

- True probabilities ranged from **0.25-0.75**
- True probabilities ranged from **0.05-0.95**
- True probabilities ranged from **0 to 1**

For each condition, performed 1,000 Monte Carlo replications generating data from the population models, fit a logistic regression model, and calculated the MOE at each point. Then evaluated:

- Whether the *empirical* MOE **exceeded** the *oracle* MOE
  - Both maximum and average values
  - **Oracle MOEs** based on probability ranges **0.25-0.75** and **0.05-0.95**
- **Difference** between the *observed* MOE and *oracle* MOE

*Additional simulation studies looking at the effect of varying the beta values, number of categorical factors, etc. were conducted but not presented

†The design is a sliced Latin hypersquare design coupled with a factorial design for the categorical variables that then is optimized with the MaxPro metric

MOE: Margin of Error

[1] See Joseph et al., 2015

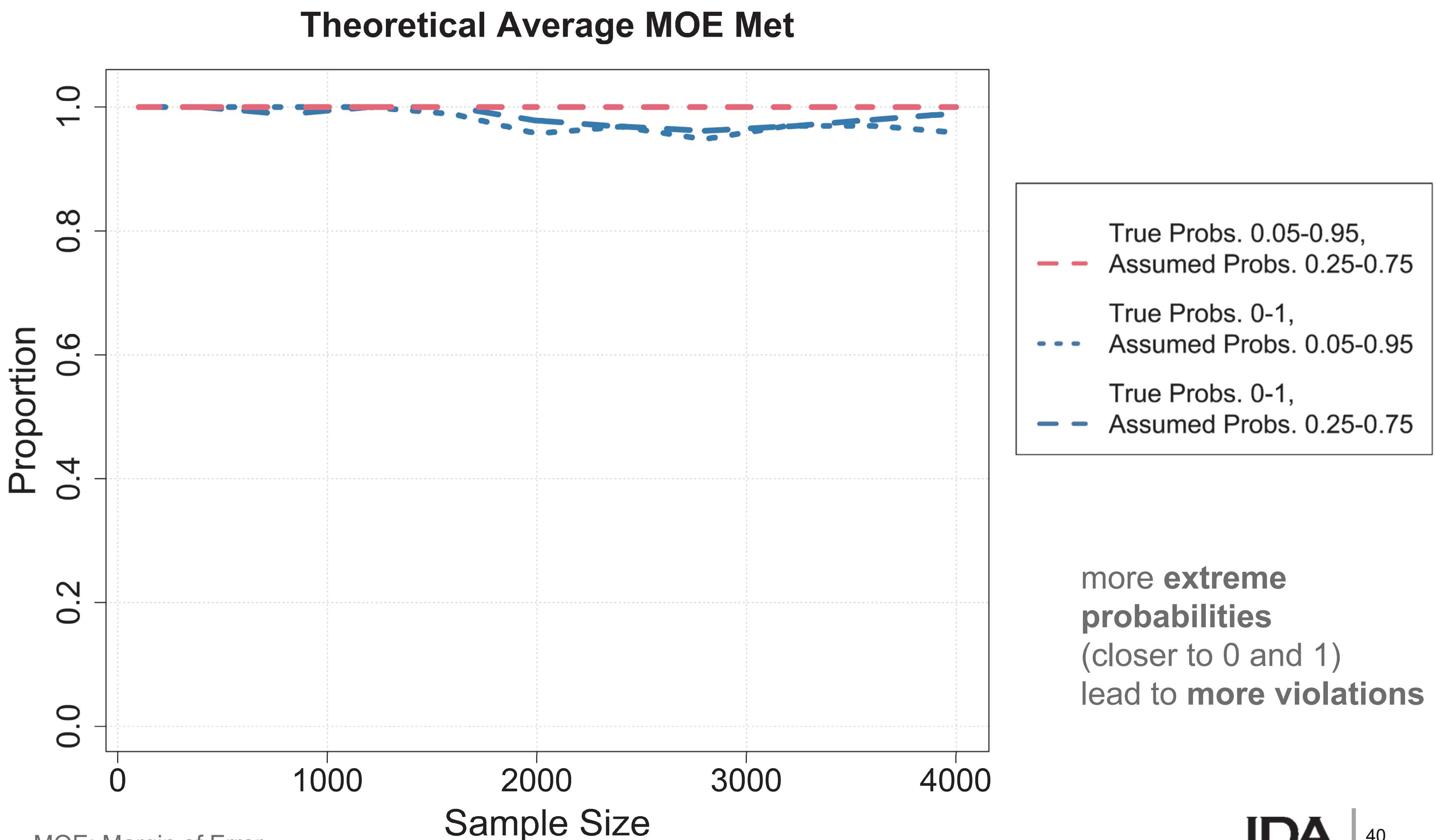
Average Oracle MOE was Unlikely to be Exceeded in Simulation Study, Regardless of Assumed Probabilities

Assumed Probability Range	True Probability Range	Percentage of Simulation Replications in Which MOE Was *Not* Exceeded
Assumed Probabilities Correct		
0.25-0.75	0.25-0.75	100%
0.05-0.95	0.25-0.75	100%
0.05-0.95	0.05-0.95	100%
Assumed Probabilities Incorrect		
0.25-0.75	0.05-0.95	100%
0.05-0.95	0.00-1.00	97.99%
0.25-0.75	0.00-1.00	98.16%

Results were similar for all tested confidence levels and across sample sizes – **suggests MOEs and number of runs** using this approach **are not too small**

MOE: Margin of Error

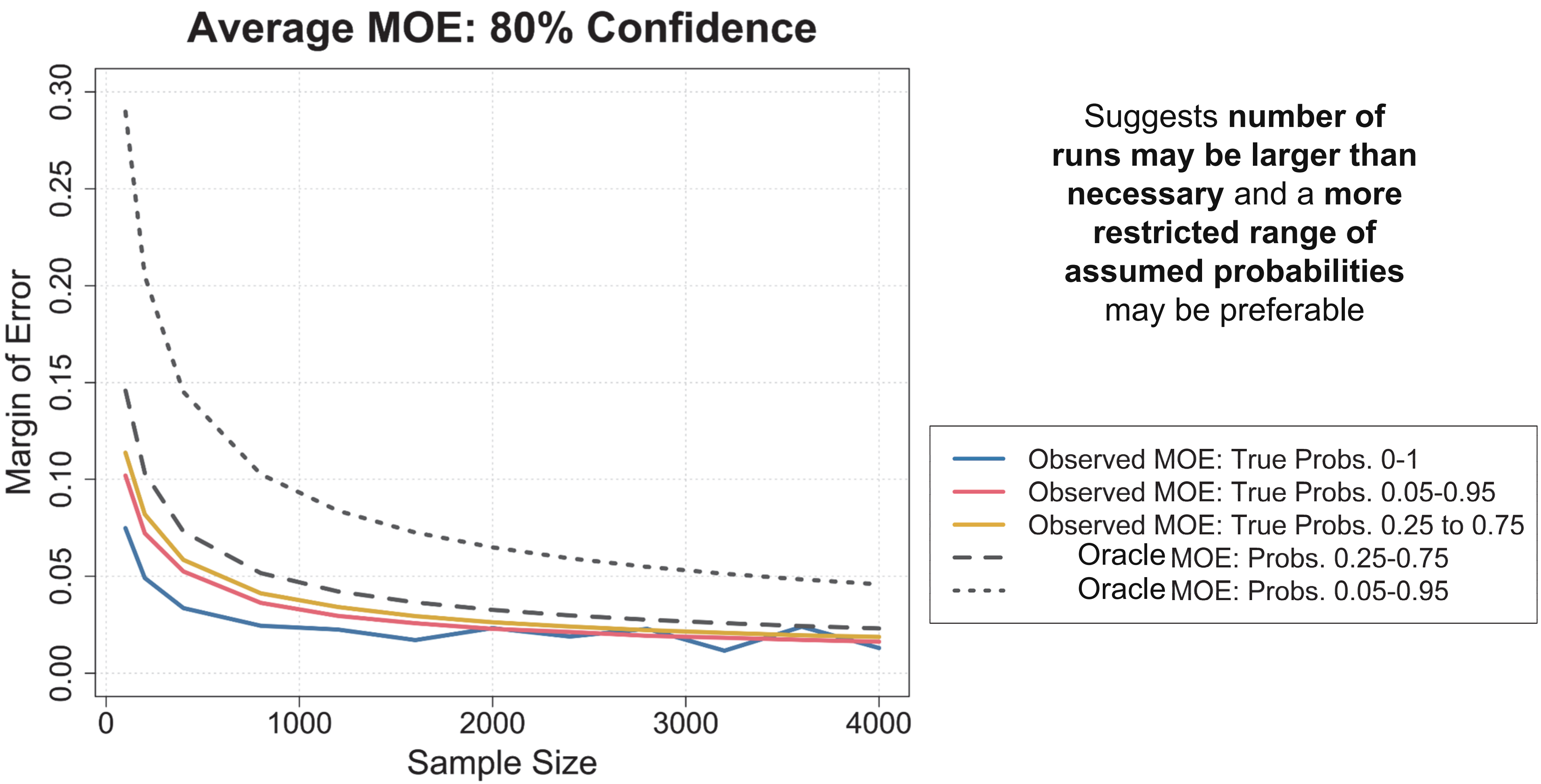
Average Oracle Margins of Error Are Still Relatively Rarely Exceeded When Assumed Underlying Probabilities Are Wrong



more **extreme probabilities** (closer to 0 and 1) lead to **more violations**

MOE: Margin of Error

Average Observed Margins of Error are Typically Smaller Than Average Oracle Bounds, Even When Assumptions are Incorrect



Suggests **number of runs may be larger than necessary** and a **more restricted range of assumed probabilities** may be preferable

MOE: Margin of Error

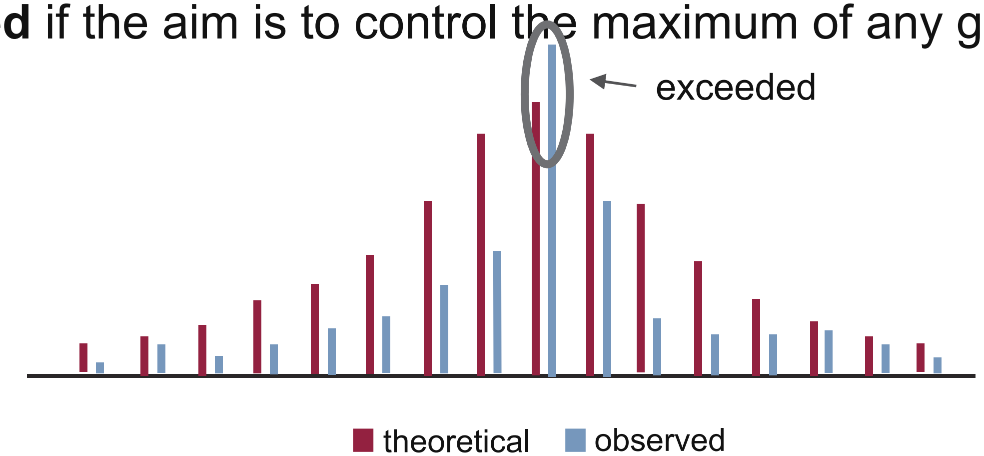
Results Suggest Theory Generally Performs Well in Practice

Approach can **be used to select a number of runs** for a certain design, before collecting data, to achieve a certain level of statistical precision

- In particular, assumed probabilities of 0.25-0.75 seem to work well when focus is on the average margin of error

Choosing a sample size is a **balance between risk aversion** and what is a **feasible number of runs**

- Additional caution is needed for certain scenarios
- **Oracle maximum MOE was violated far more frequently than the oracle average MOE across the design space**, so **extra caution is needed** if the aim is to control the maximum of any given point



MOE: Margin of Error

Backup Slides

References

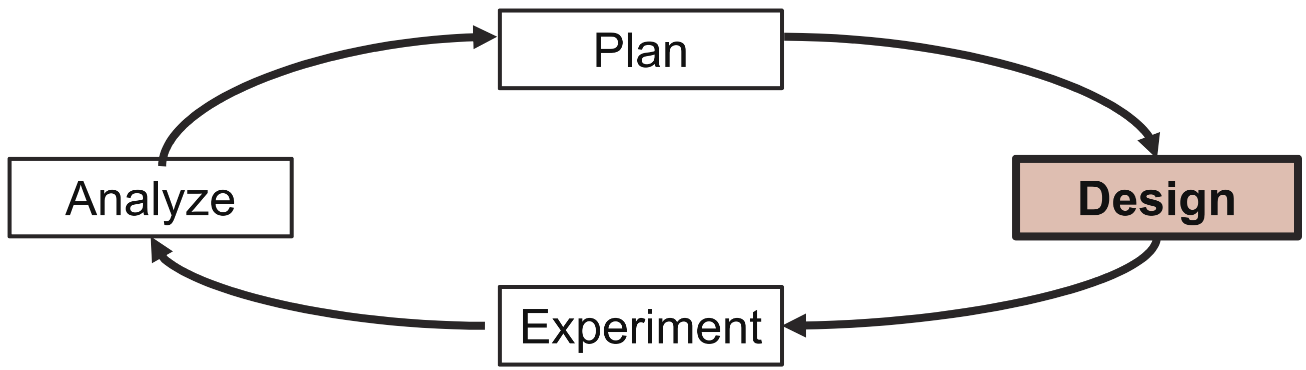
Joseph, V. Roshan, Evren Gul, and Shan Ba. “Maximum Projection Designs for Computer Experiments.” *Biometrika* 102 no. 2 (2015): 371—80. https://doi.org/10.1093/biomet/asv002.

Loeppky, Jason L., Jerome Sacks, and William J. Welch. “Choosing the sample size of a computer experiment: A practical guide.” *Technometrics* 51 no. 4 (2014): 366—76.

Naval Undersea Warfare Center. “Weapons Analysis Facility (WAF) Technical Engineering Services.” Last modified June 12, 2014. https://www.navsea.navy.mil/Portals/103/Documents/NUWC_Newport/ReadingRoom/Code85_WAF.pdf.

Wojton, Heather M., Kelly M. Avery, Han G. Yi, and Curtis G. Miller. Space filling designs for modeling & simulation validation. Paper, Alexandria, Virginia: Institute for Defense Analyses (2021).

Goal: Determine Sample Size in Test Planning for Simulations with Binary Outcomes



Approach:

- We studied *a priori* upper bounds for predicted probability confidence interval margins of error estimated with linear logistic regression
- We evaluated how well this works in practice using a simulation study

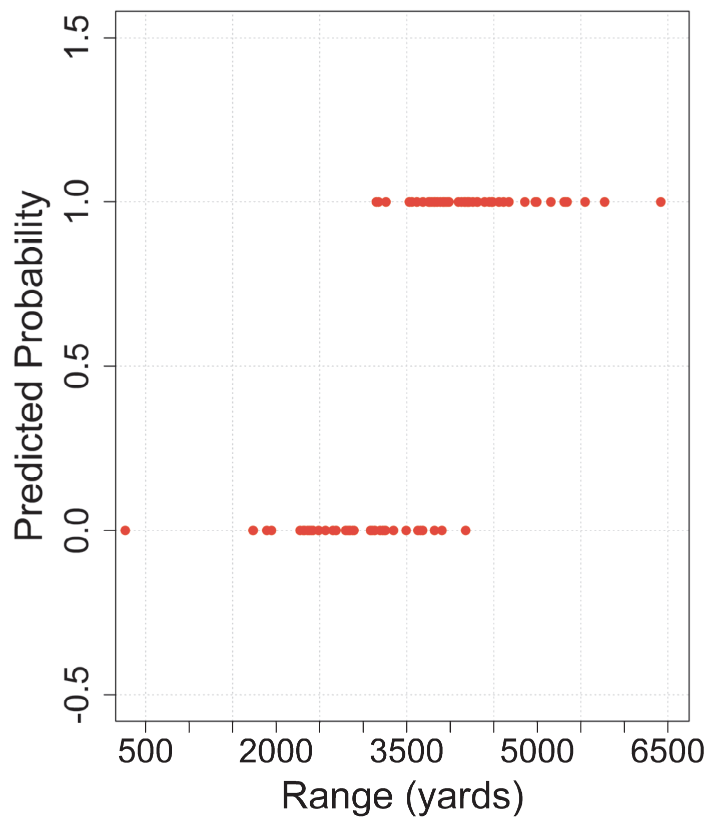
Key Findings:

- Upper bound estimate appears to work well in simulations in many cases
- Approach may have use in selecting sample sizes in practice

Logistic Models are Used to Study Torpedo Hits and Misses

A typical linear regression assumes the response is normally distributed

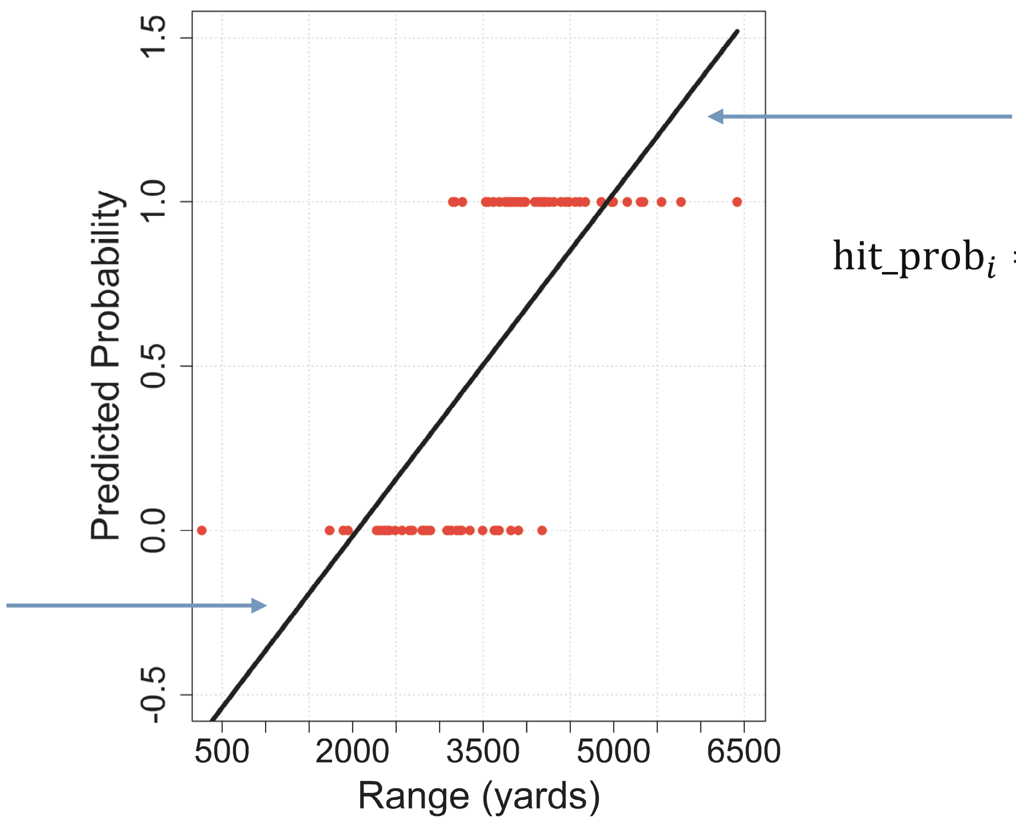
When we have a **binary response** (e.g., hit or miss), we need to use a *logistic regression* model instead



Logistic Models are Used to Study Torpedo Hits and Misses

A typical linear regression assumes the response is normally distributed

When we have a **binary response** (e.g., hit or miss), we need to use a *logistic regression* model instead

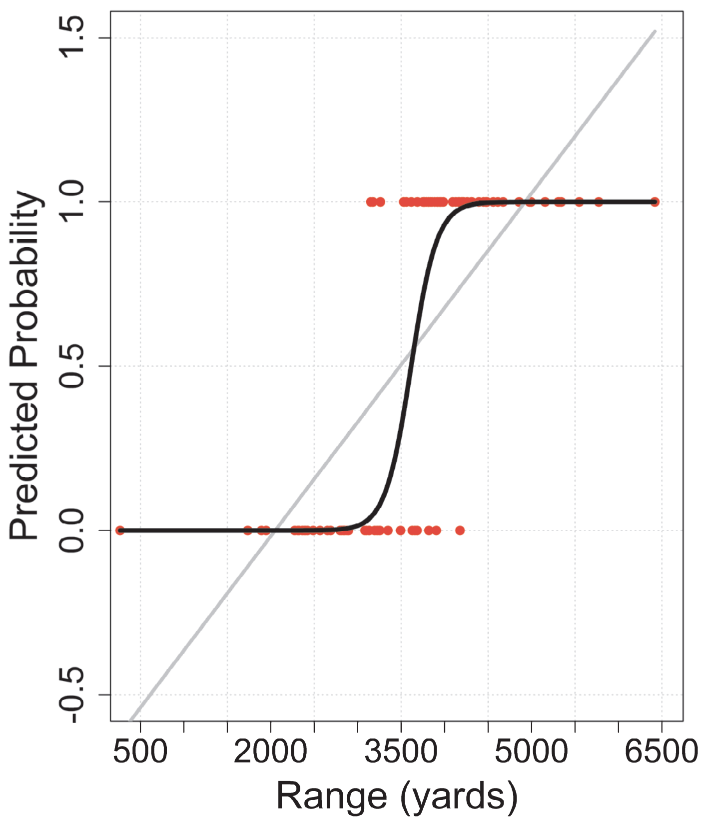


$$\text{hit\_prob}_i = \beta_0 + \beta_1 \text{range}_i$$

Logistic Models are Used to Study Torpedo Hits and Misses

A typical linear regression assumes the response is normally distributed

When we have a **binary response** (e.g., hit or miss), we need to use a *logistic regression* model instead



$$\text{hit\_prob}_i = \boxed{\eta}(\beta_0 + \beta_1 \text{range}_i)$$

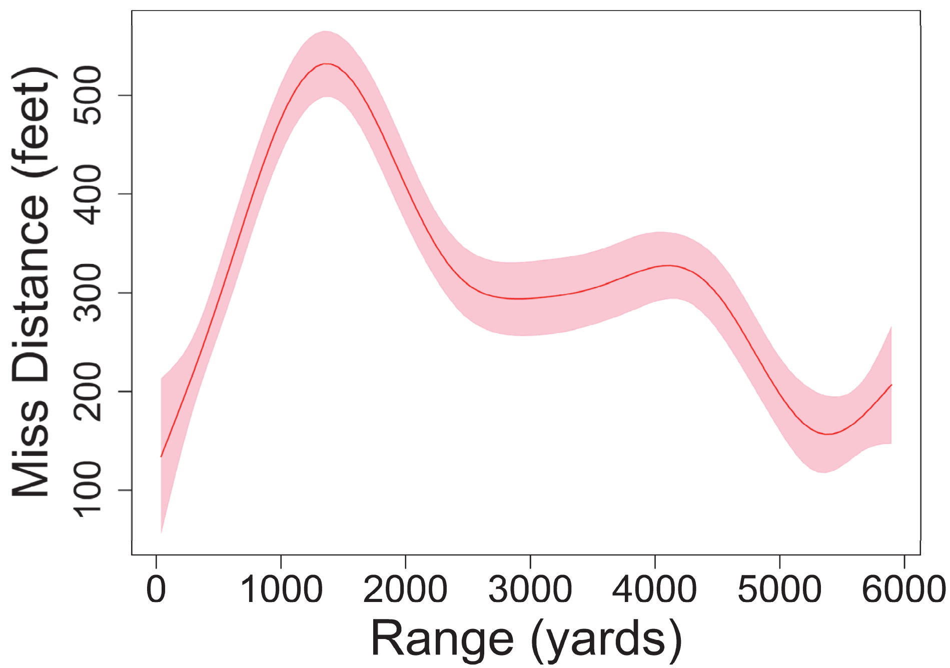
$$\eta = \frac{1}{1 + e^{-x}}$$

Next Steps are to Apply a Similar Approach to Generalized Additive Models

Generalized additive models (GAMs) are a more flexible class of models that are generally better suited for meta-modeling[1]

- Ideally, an ability to base sample size on confidence in GAMs may lead to the most efficient M&S for meta-model construction and use

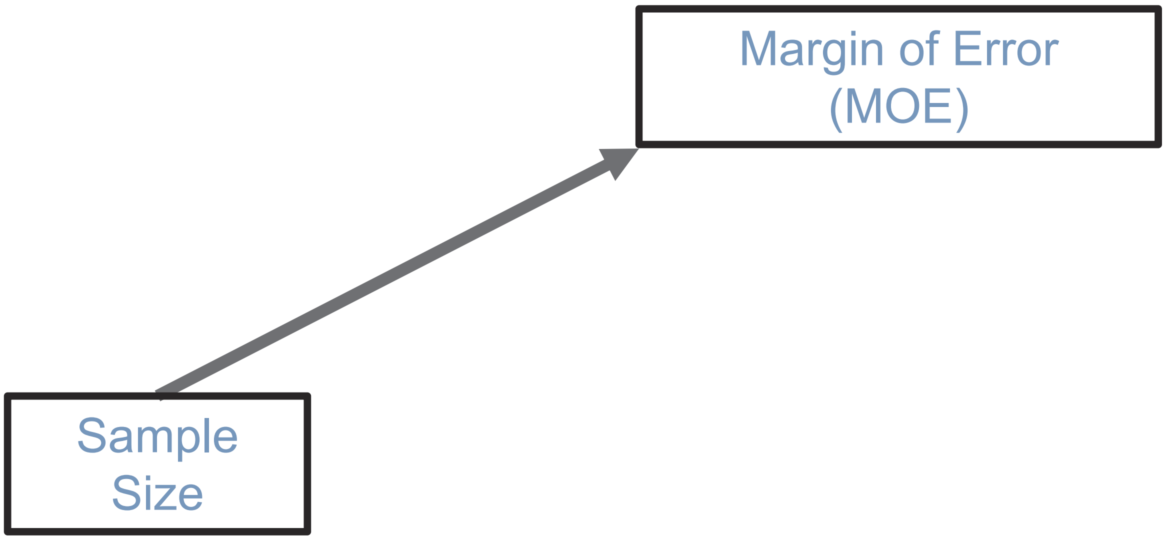
Notional Example



M&S: Modeling and Simulation

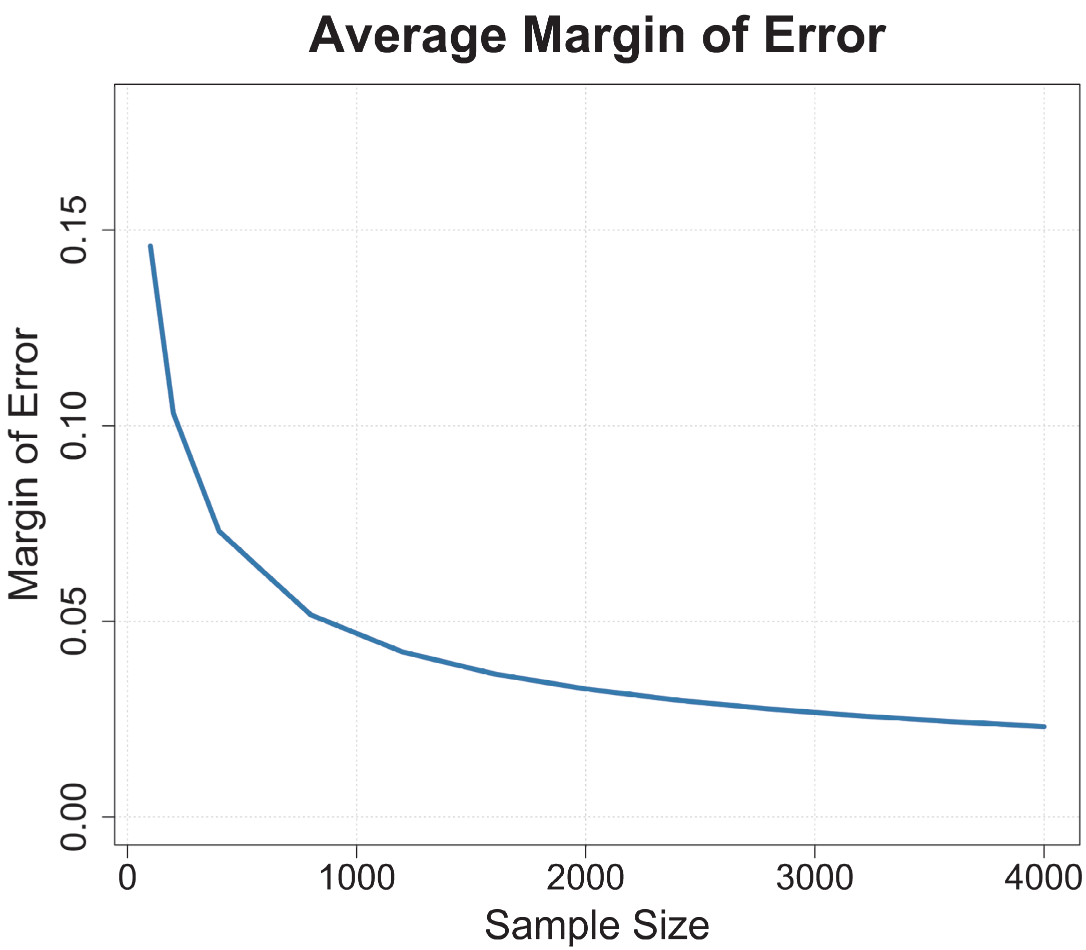
[1] Haman & Miller, 2023

Factors That Affect Theoretical Margin of Error

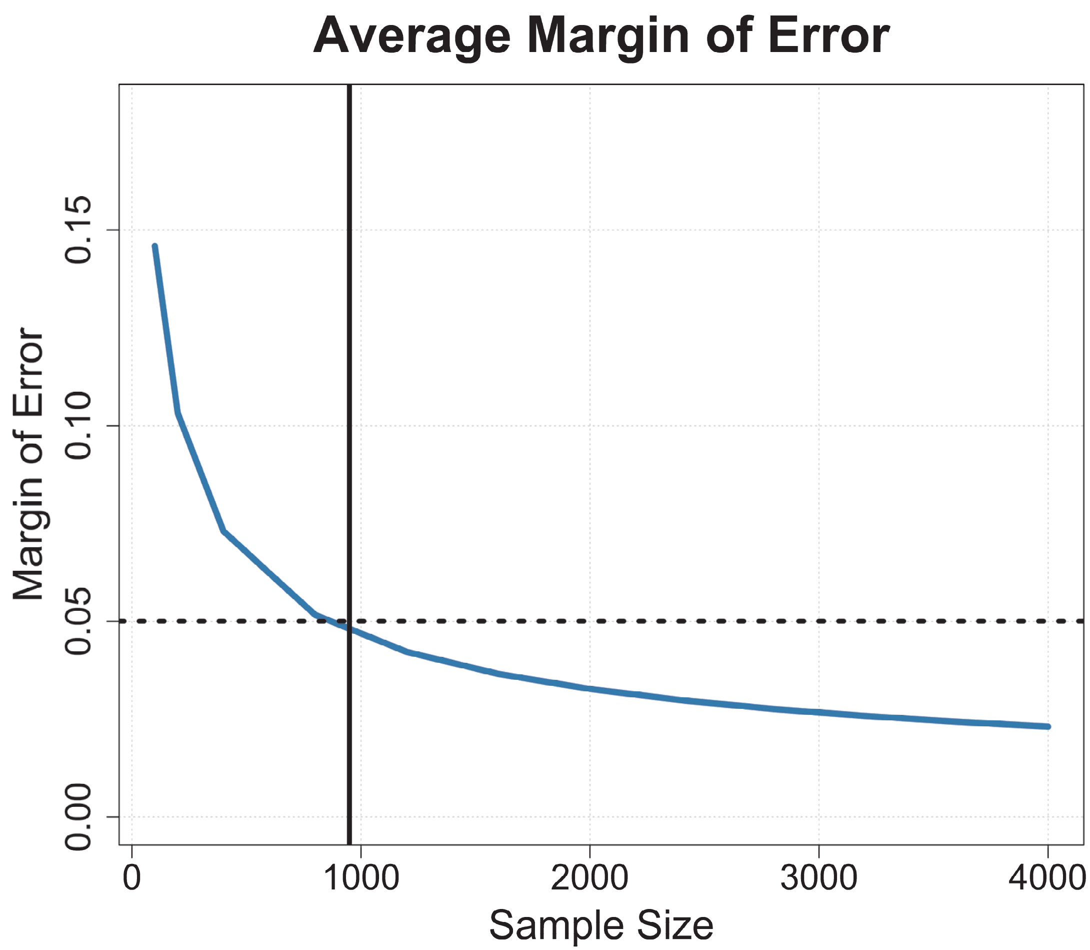


sample size ↑,
MOE ↓

As Sample Size Increases, Margin of Error Decreases

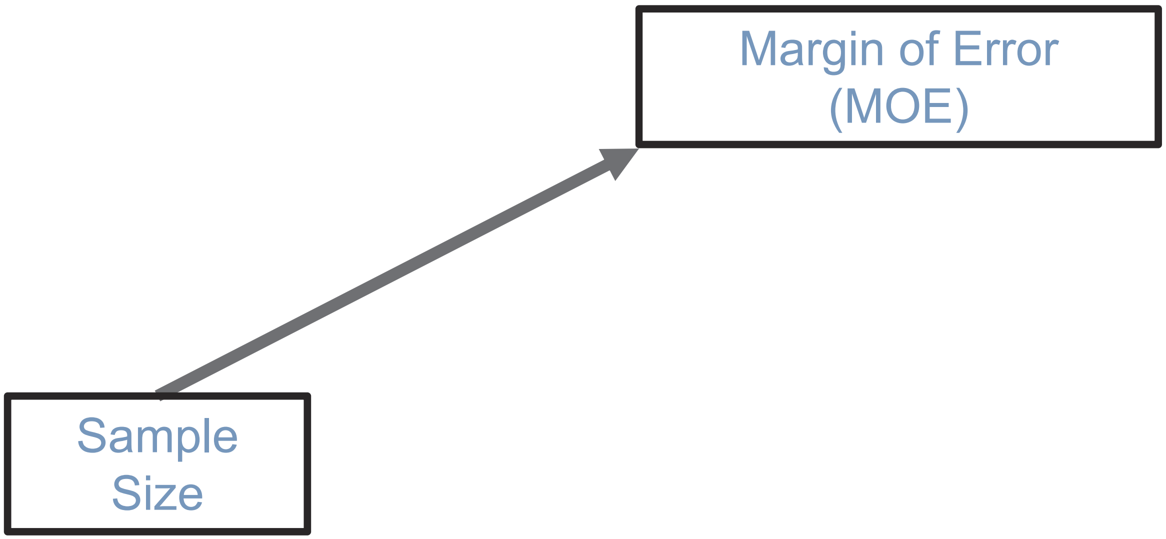


As Sample Size Increases, Margin of Error Decreases



We want to determine a sample size for a selected margin of error

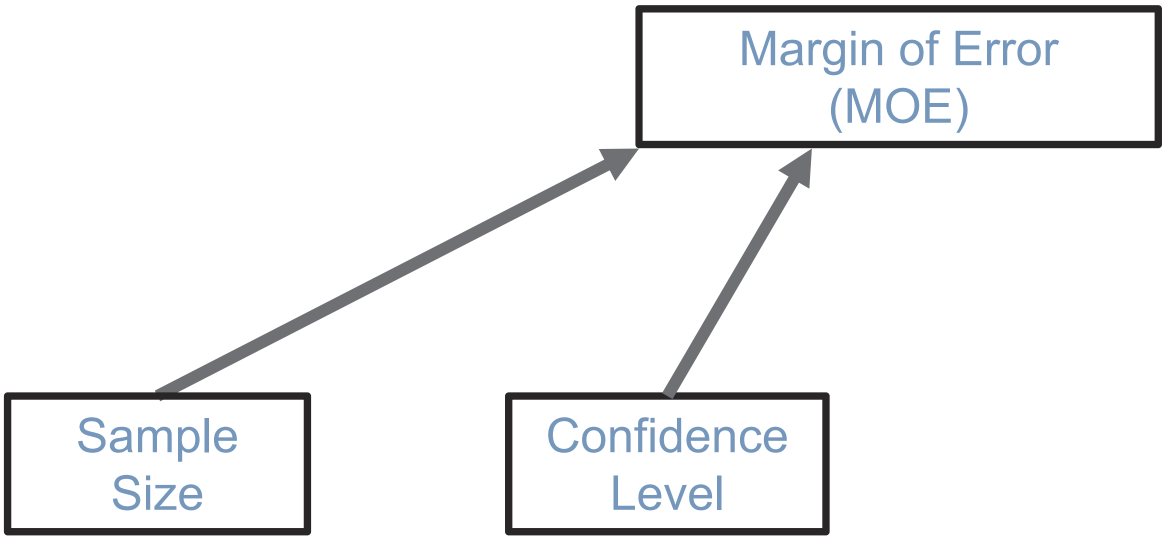
Factors That Affect Theoretical Margin of Error



sample size ↑,
MOE ↓

What we want to determine based on the other factors

Factors That Affect Theoretical Margin of Error

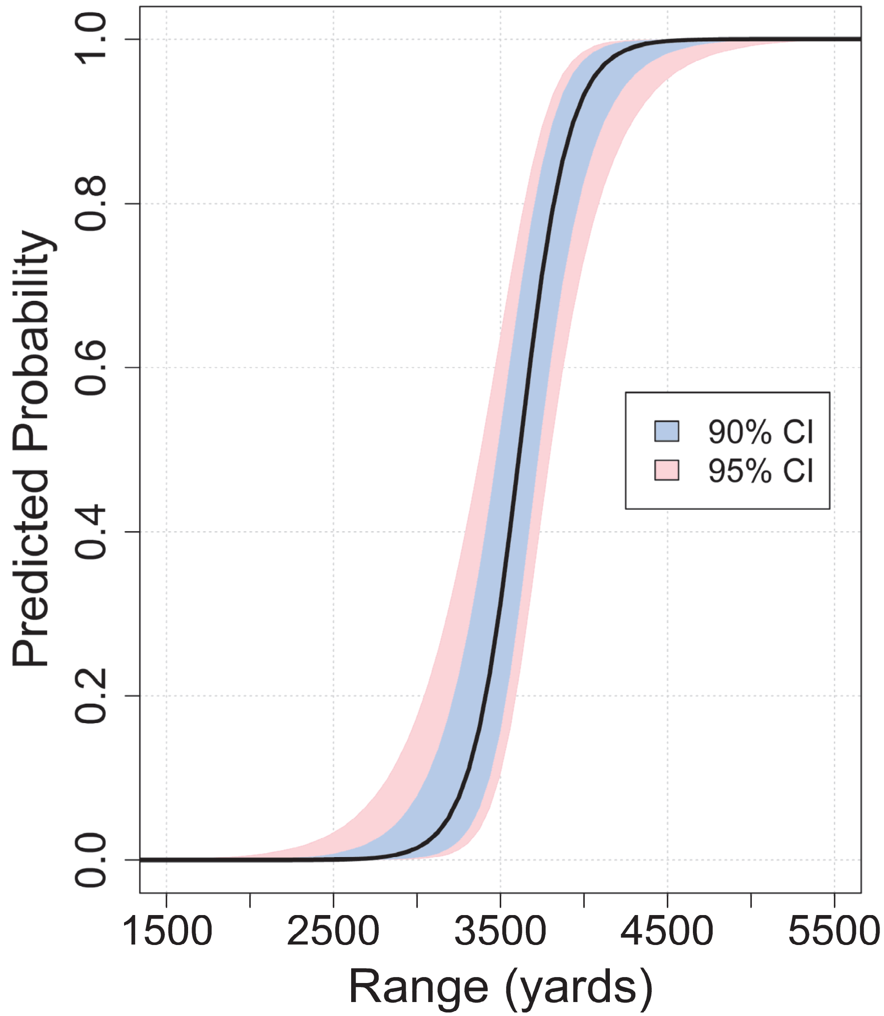


sample size ↑,
MOE ↓

confidence ↑,
MOE ↑

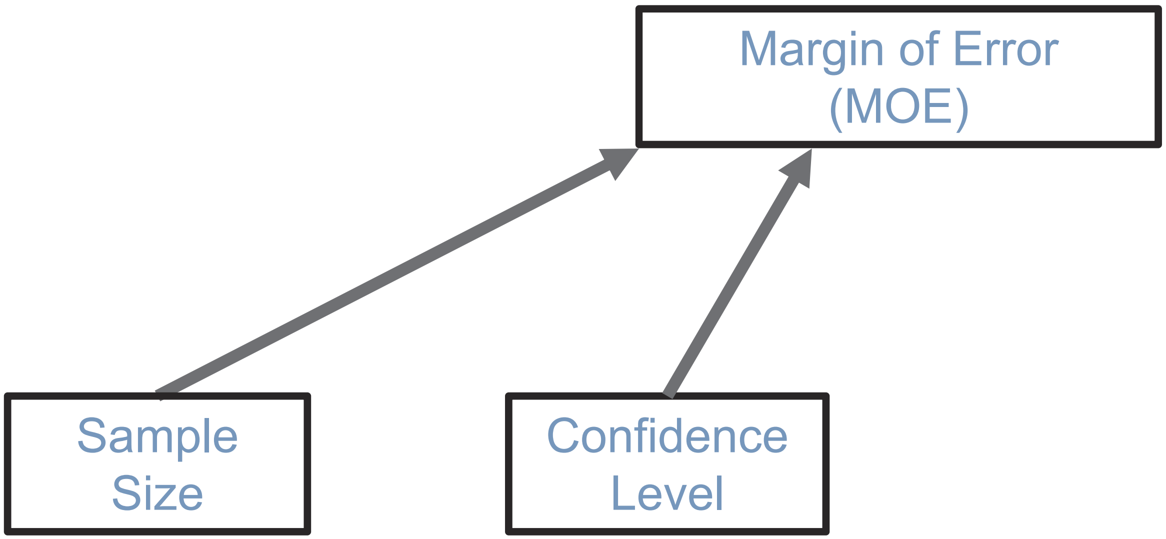
What we want to determine based on the other factors

As Confidence Level Increases, Margin of Error Increases



CI: Confidence Interval

Factors That Affect Theoretical Margin of Error



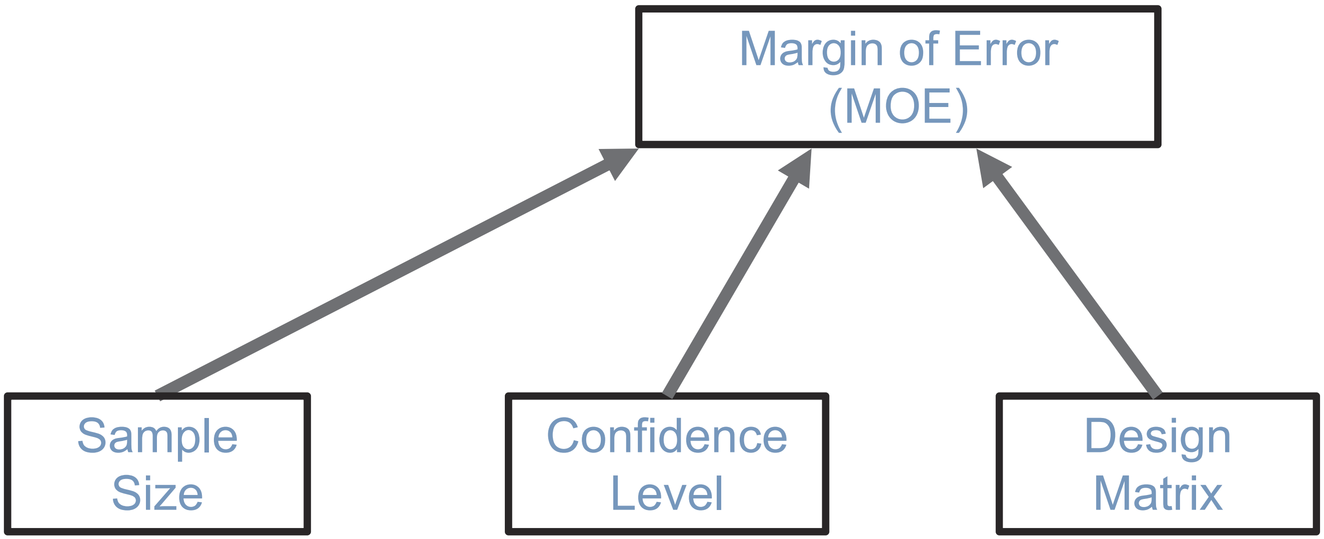
sample size $\uparrow$,
MOE $\downarrow$

What we want to determine based on the other factors

confidence $\uparrow$,
MOE $\uparrow$

Known, specified in formula

Factors That Affect Theoretical Margin of Error



sample size ↑,
MOE ↓

What we want to determine based on the other factors

confidence ↑,
MOE ↑

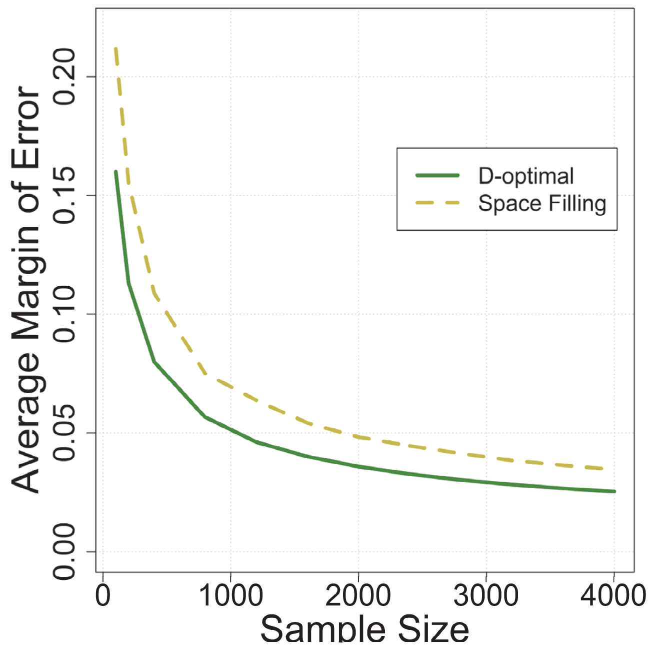
Known, specified in formula

complicated

Many Aspects of the Design Affect the Margin of Error

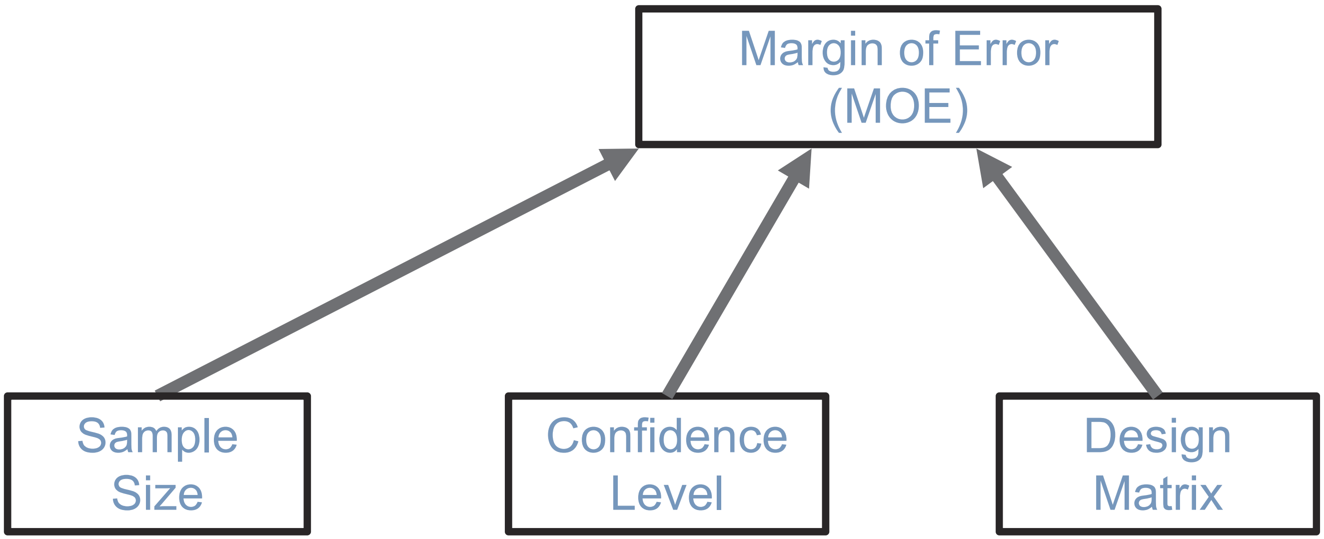
The theoretical margin of error is affected by design aspects such as:

- Number of categorical factors
- Type of design
- Sample size



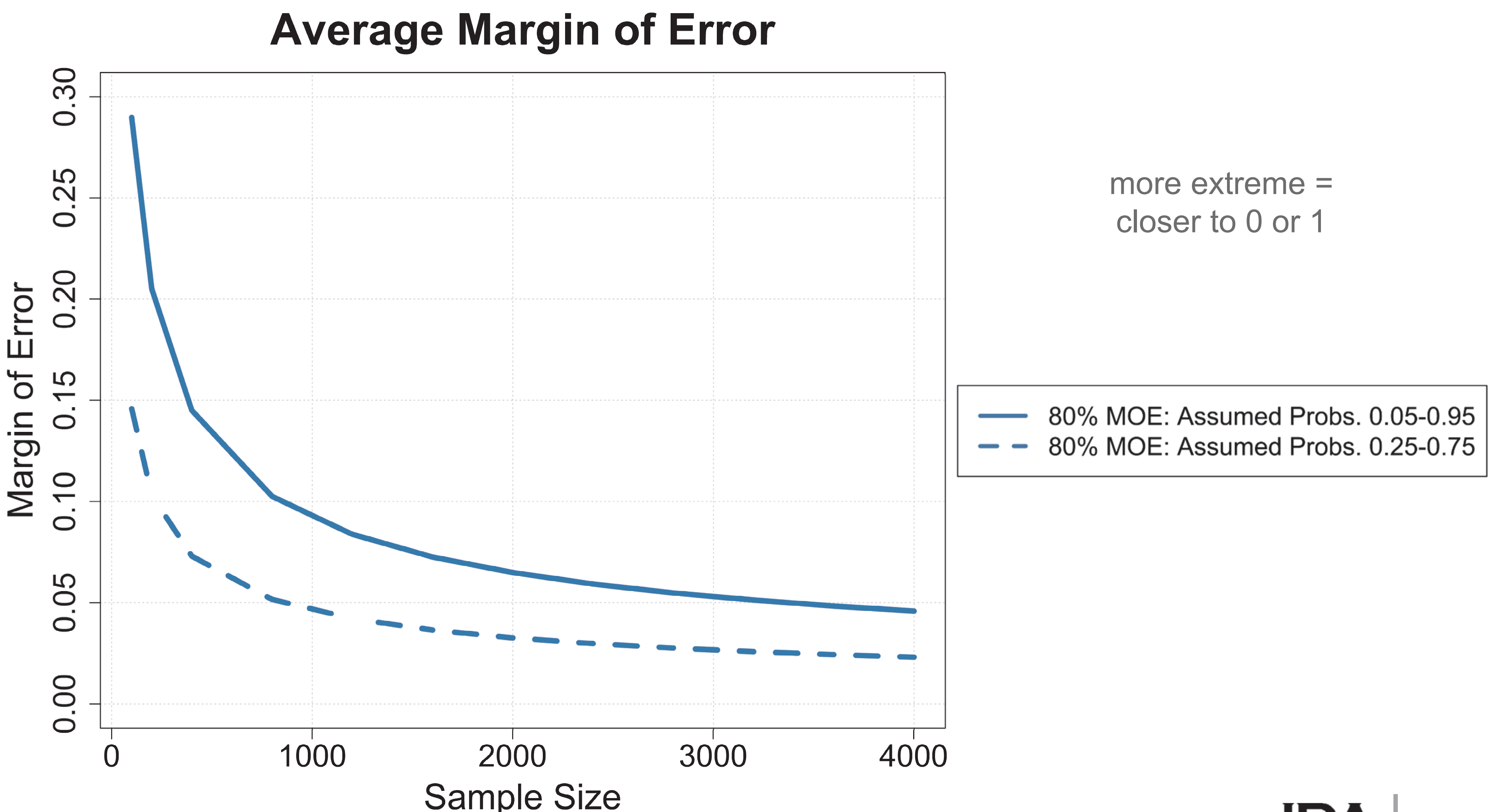
Ultimately, design is chosen based on project aims and the design matrix is considered fixed

Factors That Affect Theoretical Margin of Error



Sample Size	Confidence Level	Design Matrix
sample size ↑, MOE ↓	confidence ↑, MOE ↑	complicated
What we want to determine based on the other factors	**Known, specified in formula**	**Known*, specified in formula**
		*for a given sample size and design type

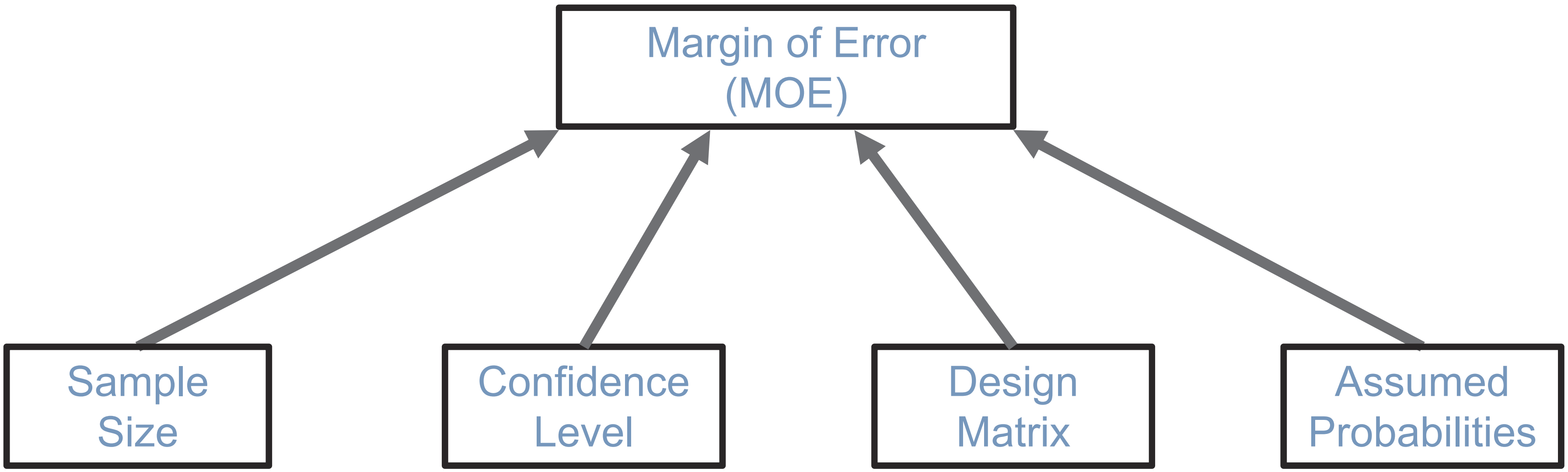
As Range of Assumed Probabilities Get More Extreme, Margin of Error Increases



more extreme =
closer to 0 or 1

MOE: Margin of Error

Factors That Affect Theoretical Margin of Error

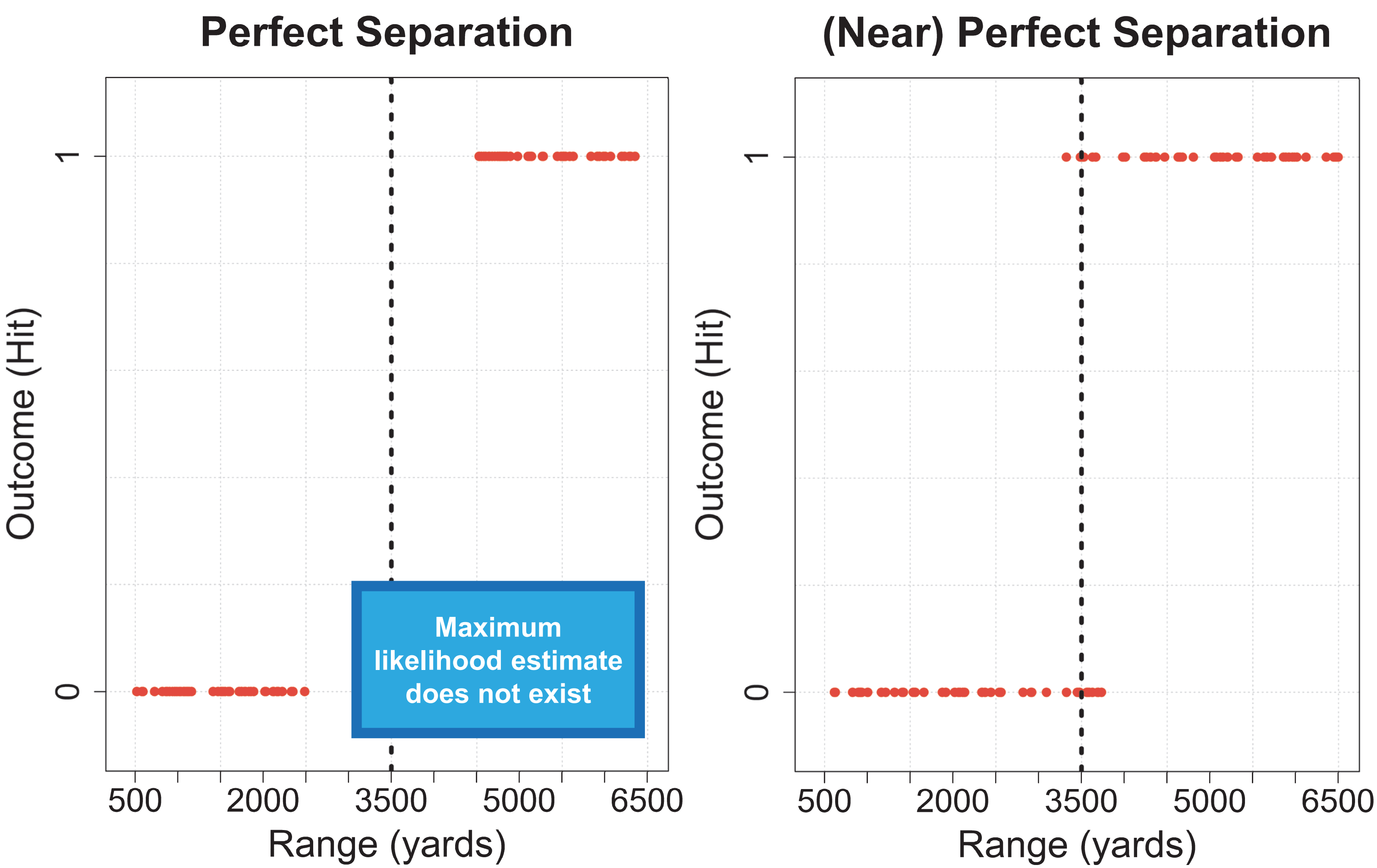


Sample Size	Confidence Level	Design Matrix	Assumed Probabilities
sample size ↑, MOE ↓	confidence ↑, MOE ↑	complicated	probabilities ↑ ↓, MOE ↑
What we want to determine based on the other factors	**Known, specified in formula**	**Known*, specified in formula**	**Unknown, specified in formula**

*for a given sample size and design type

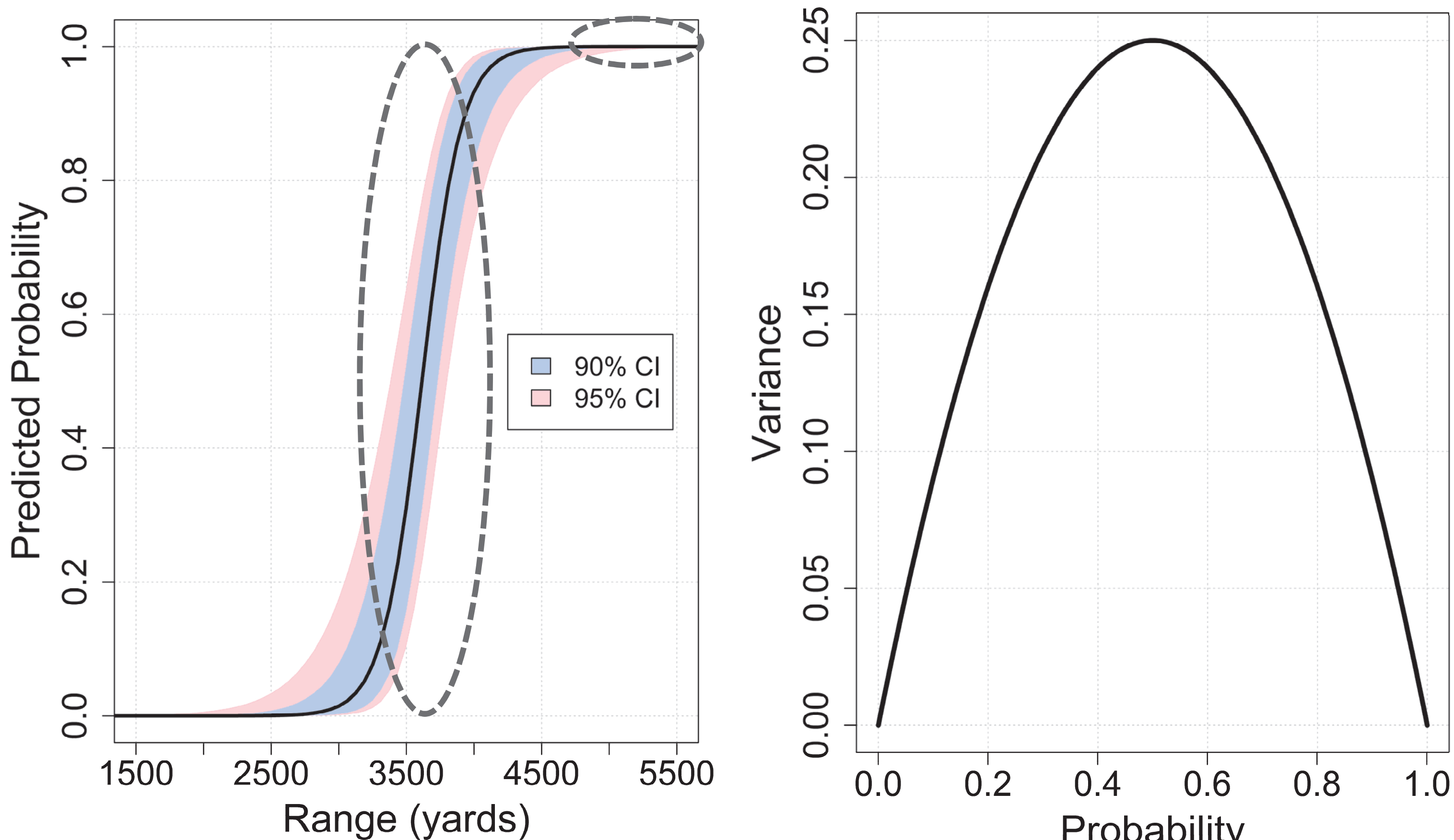
Theoretical Margin of Error Accounts for Issues from Extreme Probabilities

Logistic regression has issues when all probabilities are very close to 0 and 1



Theoretical Margin of Error Accounts for Probabilities Near Middle

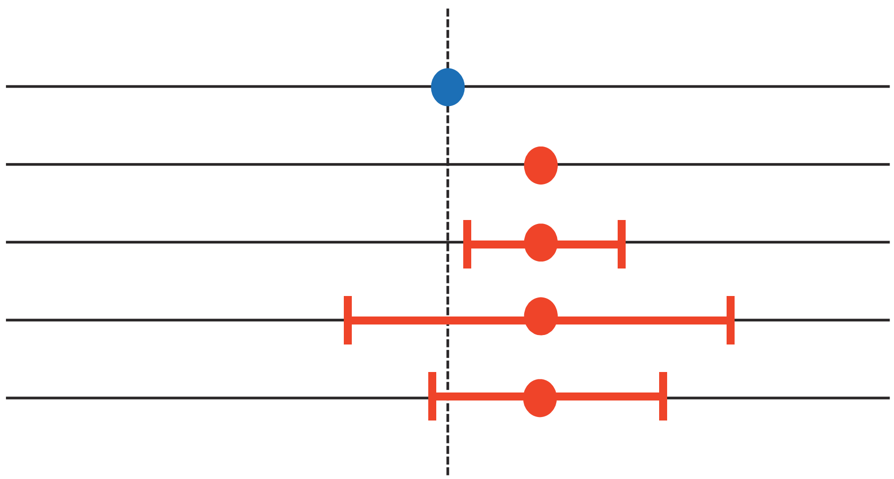
The theoretical formula accounts for the fact that probabilities near the middle have the widest margins of error



CI: Confidence Interval

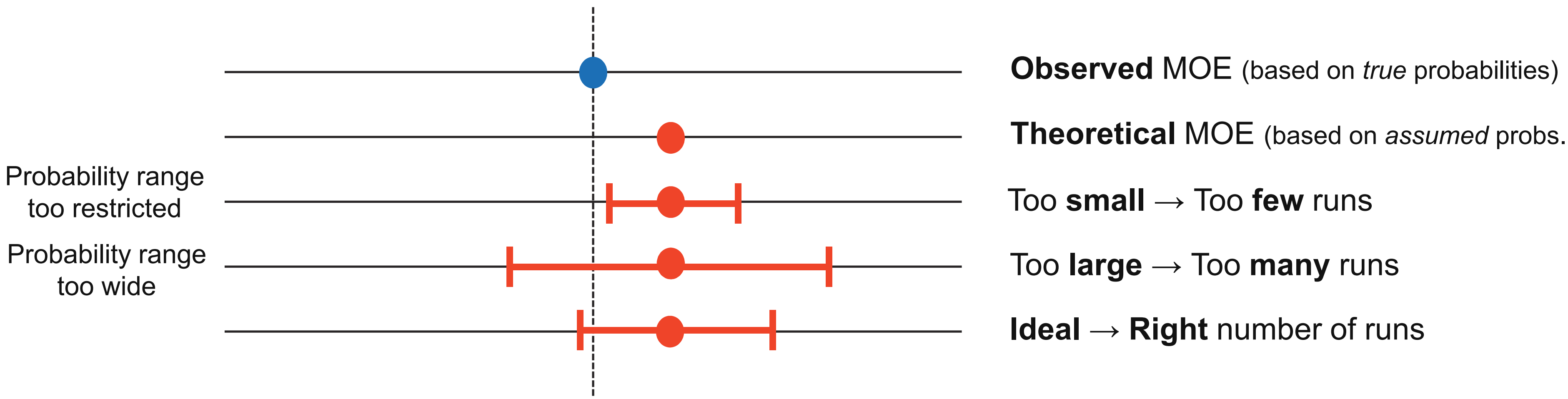
A Simulation Study Can Be Used to Evaluate the Effect of Different Assumed and True Probabilities

If we expect the probabilities to range fairly close to 0 and 1, the theoretical margin of error gets extremely large



A Simulation Study Can Be Used to Evaluate the Effect of Different Assumed and True Probabilities

If we expect the probabilities to range fairly close to 0 and 1, the theoretical margin of error gets extremely large



MOE: Margin of Error

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
08-XX-2024

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Determining the necessary number of runs in computer simulations with binary outcomes

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Kelly Duffy
Curtis G. Miller

5d. PROJECT NUMBER
C9082

5e. TASK NUMBER
C9082

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER
3002814

3002814

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
IDA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Statement A. Approved for public release: distribution is unlimited.

13. SUPPLEMENTARY NOTES
Project Leader: Rebecca M. Medlin

14. ABSTRACT
How many success-or-failure observations should we collect from a computer simulation? Often, researchers use space-filling design of experiments when planning modeling and simulation (M&S) studies. We are not satisfied with existing guidance on justifying the number of runs when developing these designs, either because the guidance is insufficiently justified, does not provide an unambiguous answer, or is not based on optimizing a statistical measure of merit. Analysts should use confidence interval margin of error as the statistical measure of merit for M&S studies intended to characterize overall M&S behavioral trends. Unfortunately, the margin of error for studies involving factors and success-or-failure (or binary) outcomes requires knowing model parameters when using logistic regression. We explore how an upper bound on the margin of error, needing less information about the statistical model we need to estimate, can assist in sample size planning. While the upper bound needs further theoretical refinement, simulation studies suggest the upper bound may provide a means of justifying M&S study sample sizes with a statistical measure of merit.

15. SUBJECT TERMS
Design of Experiments, Modeling and Simulation, Statistics, Space-Filling Designs

16. SECURITY CLASSIFICATION OF: a. REPORT	b. ABSTRACT	c. THIS PAGE	17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR	25TBD	Rebecca M. Medlin
					19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6731