INSTITUTE FOR DEFENSE ANALYSES

DATAWorks 2024: Simulation Insights on Power Analysis with Binary Responses--From SNR Methods to 'skprJMP'

Robert Atkins, Project Leader

Tyler Morgan-Wall
Curtis Miller

April 2024

IDA Product ID 3002093

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 BD-9-2299(37), "Shipboard Air Defense Systems," 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
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Adam Miller, Dr. Dhruv Patel, Dr. George Khoury, Dr. John Haman, and Dr. Madalina Colci O'Hara from the Operational Evaluation Division.

For more information:
Dr. Robert Atkins, Project Leader
ratkins@ida.org • 703-933-6546

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

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Product ID 3002093

DATAWorks 2024: Simulation Insights on Power Analysis with Binary Responses--From SNR Methods to 'skprJMP'

Robert Atkins, Project Leader

Tyler Morgan-Wall
Curtis Miller

Executive Summary

Design of Experiments (DoE) provides a solid analytical basis for test planning because it includes tools to characterize the design space, maximizes the value of each run or mission, and links the planning of a test to the data analysis that follows the test. Importantly, DoE allows testers to characterize and minimize risk for a test by calculating statistical power for a particular design and analysis method, providing decision makers with a quantitative assessment of the likelihood of an experiment discovering an effect if one exists. Because of all these benefits, the Government favors test designs that use DoE.

However, the most common power calculation software packages used in the Department of Defense do not support logistic regression, an important and widely used data analysis method used to analyze designs with a binomial response. To address this capability gap, IDA issued guidance in 2014 that recommended the use of the signal-to-noise ratio (SNR) approximation method to calculate power for these types of designs. Since then, developments in statistical software and improvements in computational power have made more accurate and reliable methods available for this purpose.

This briefing describes the results of a simulation study assessing the accuracy and reliability of the SNR method. The analysis shows that the SNR method does not produce accurate or reliable power estimates when compared to modern methods, and its use can result in both oversized and undersized tests.

This briefing then describes two software products developed at IDA that provide a more accurate and robust Monte Carlo-based method to compute power for binomial responses: "skpr" and "skprJMP." IDA released the R package "skpr" in 2017, and since then the package has supported developmental and operational test planning across dozens of programs, and has become a core piece of analytical test infrastructure for many organizations. However, some organizations have been unable to use "skpr" due to strict information technology constraints that disallow the installation of the R programming language. The new "skprJMP" add-in implements a subset of the capabilities provided by skpr entirely within JMP, allowing these organizations to calculate power for these designs. The presentation gives an overview of the features and

capabilities of these two software packages, and shows how users can download them and learn more about their use.

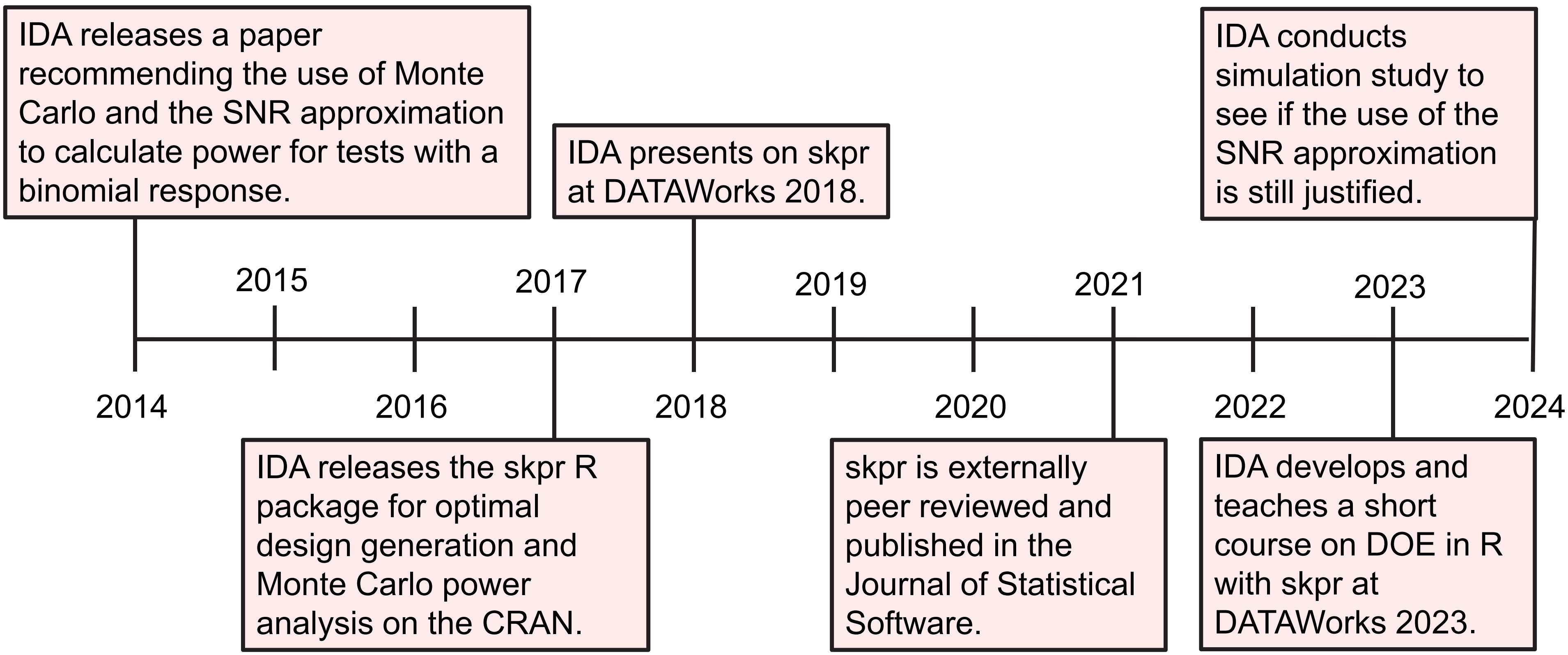
DATAWORKS 2024: Simulation Insights on Power Analysis with Binary Responses–From SNR Methods to 'skprJMP'

Tyler Morgan-Wall
Curtis Miller
Rob Atkins

April 16th, 2024

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

Background



SNR: Signal-to-noise Ratio; CRAN: Comprehensive R Archive Network; DOE: Design of Experiments;

Problem Statements

- In 2024, we now have a user-friendly and accurate tool (skpr) for computing Monte Carlo power: given this existing capability, is the SNR approximation still a reasonable method to estimate power for tests with a binomial response?

Short Answer: No.

- If not, can we bring that easy Monte Carlo power for organizations using JMP?

Short Answer: Yes!

SNR: Signal-to-noise Ratio

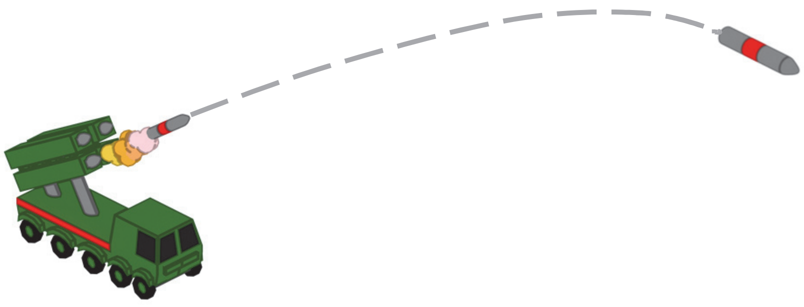
Outline

- How do we estimate power for a binomial response?
- Overview of the SNR approximation method
- Overview of the Monte Carlo method
- Analysis assessing the accuracy of the SNR approximation
- skpr and skprJMP

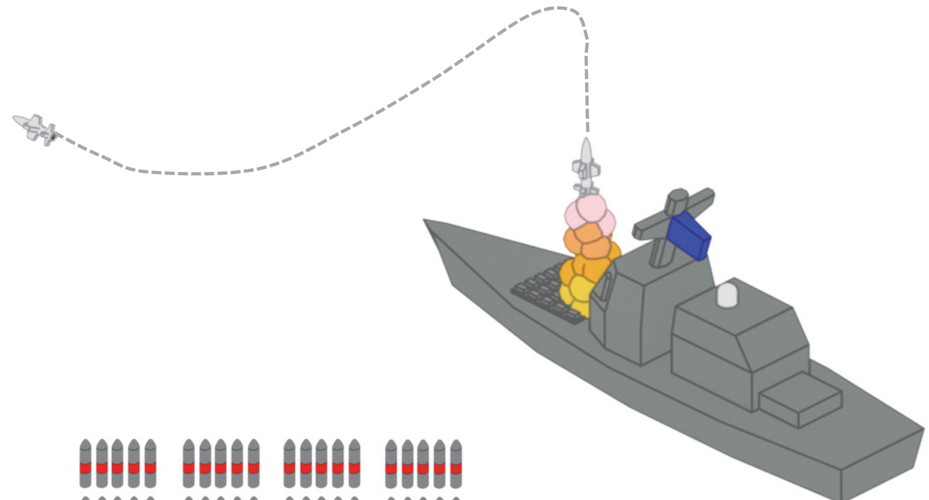
SNR: Signal-to-noise Ratio

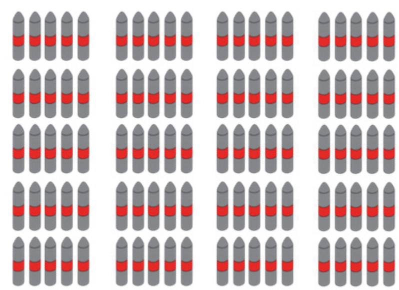
Background: Why is it difficult to calculate power when the response variable is a probability?

More extreme probabilities can be estimated with better precision than probabilities closer to 0.5



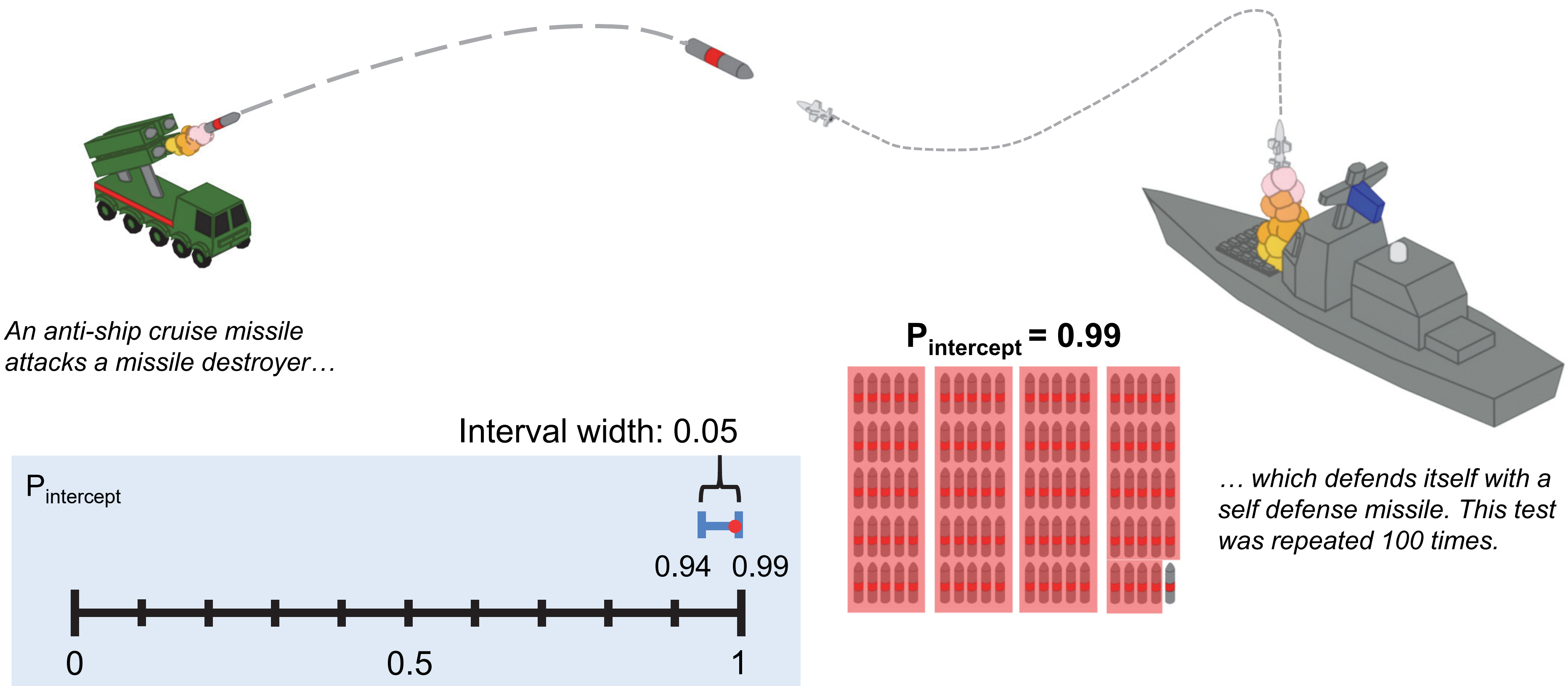
An anti-ship cruise missile attacks a missile destroyer…



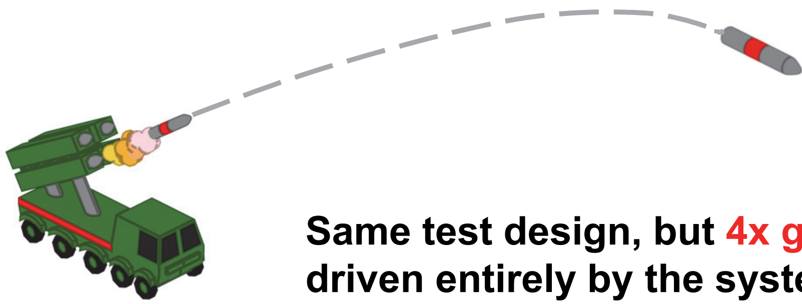


… which defends itself with a self defense missile. This test was repeated 100 times.

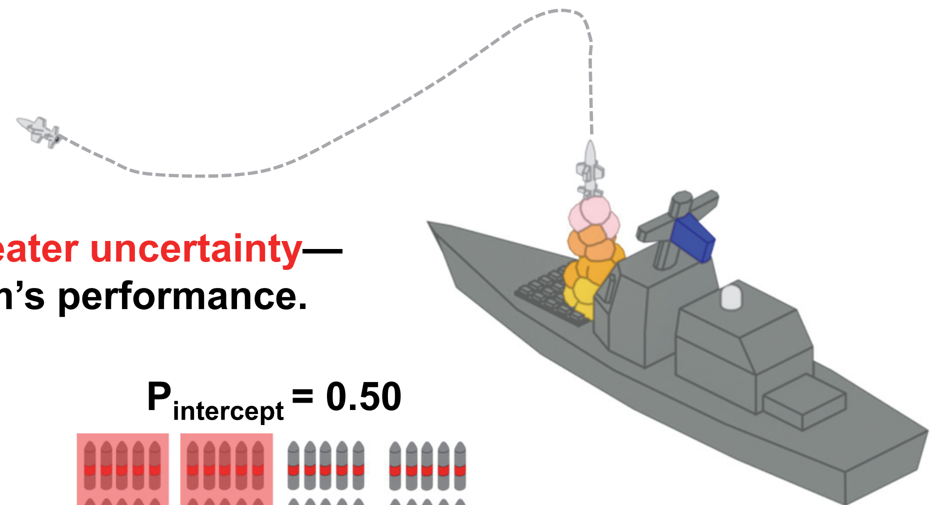
More extreme probabilities can be estimated with better precision than probabilities closer to 0.5



More extreme probabilities can be estimated with better precision than probabilities closer to 0.5

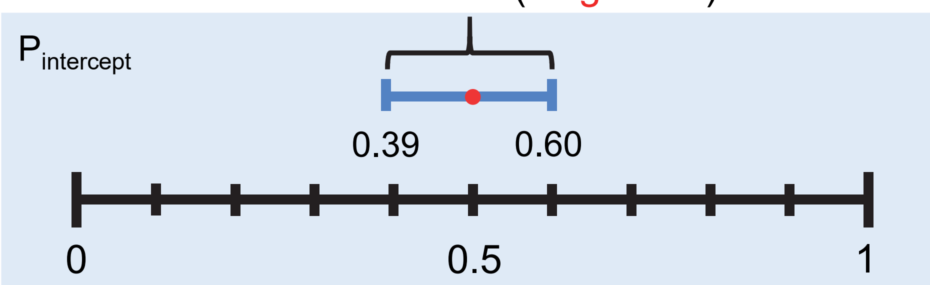


Same test design, but 4x greater uncertainty—driven entirely by the system's performance.

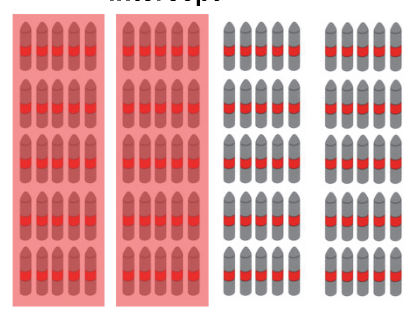


An anti-ship cruise missile attacks a missile destroyer…

Interval width: 0.21 (4x greater)



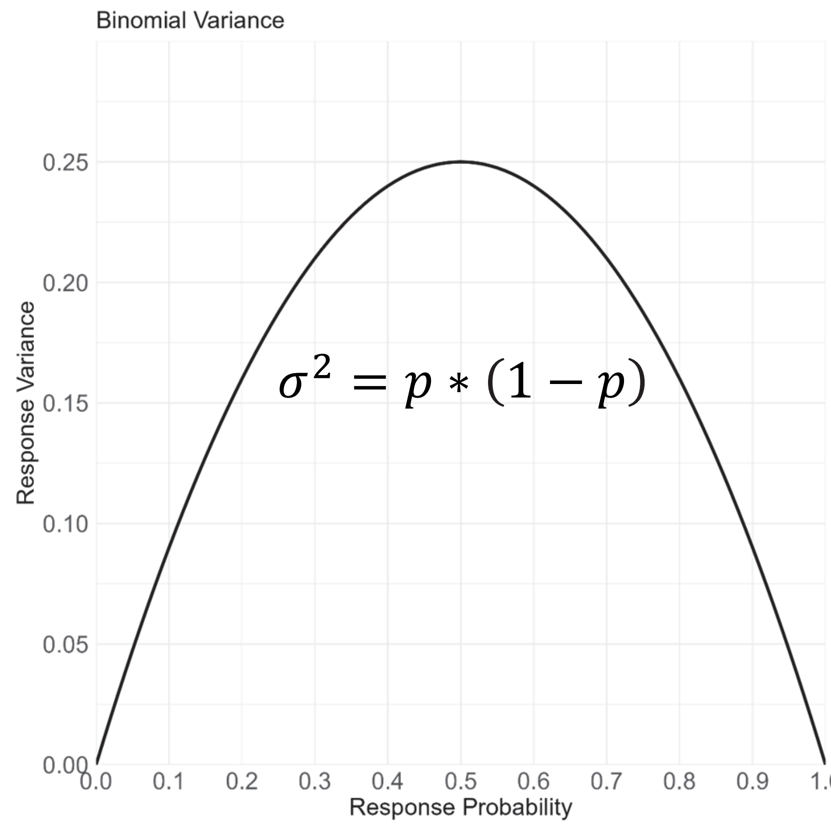
$P_{intercept}$ = 0.50



… which defends itself with a self defense missile. This test was repeated 100 times.

Binomial responses have non-constant variance

- For binomial responses, the variance is not constant and depends on the value of the response. The variance peaks at probability 0.50 and goes to zero as the probability approaches either 0.0 or 1.0.
- Non-constant variance violates one of the main assumptions of linear regression.
- Tests with binomial responses are usually fit with logistic regression, a method which has no easy analytic solution for power.

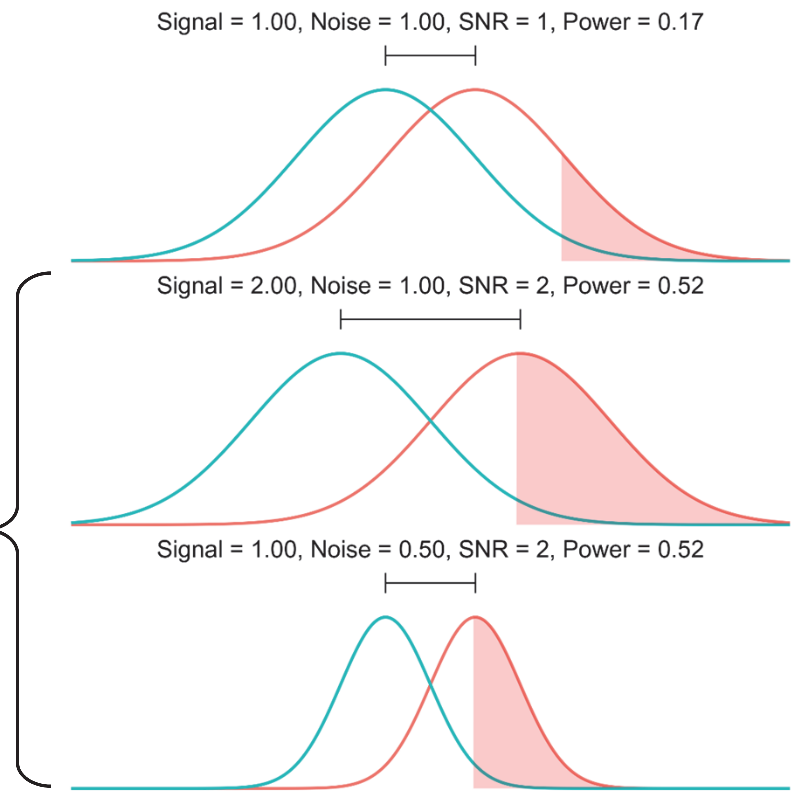


SNR: Signal to noise ratio

Power depends on both the signal (effect size) and noise (variance)

- Power tells us the probability that our design is able to detect an effect of a specific size.
- We generally use signal-to-noise ratios when calculating power for designs with a linear response, as statistical power depends on their ratio (δ/σ) and not on the specific values for either.
- For linear models, these calculations assume constant variance—the noise term does not depend on the value of the response.

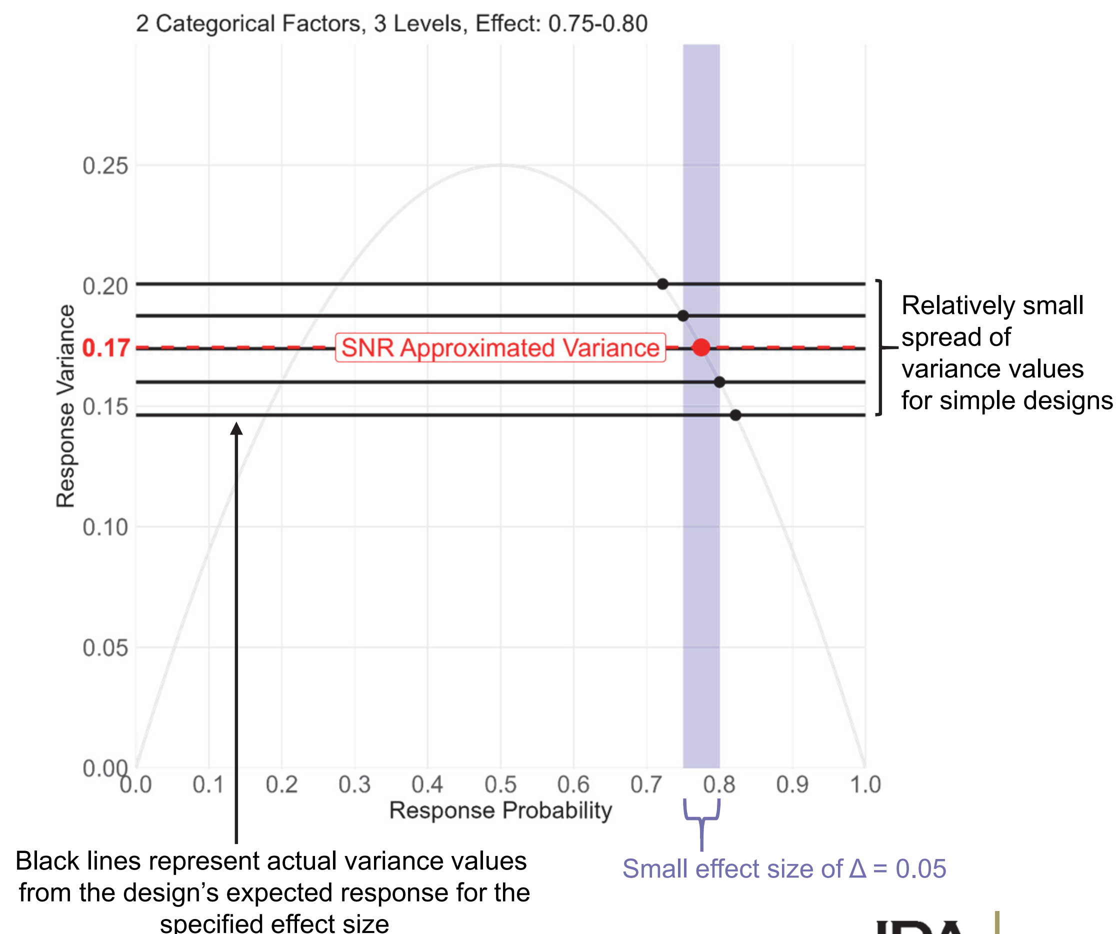
Different Signal
Different Noise
Same SNR
Same Power



SNR: Signal-to-Noise Ratio

The idea behind the SNR approximation

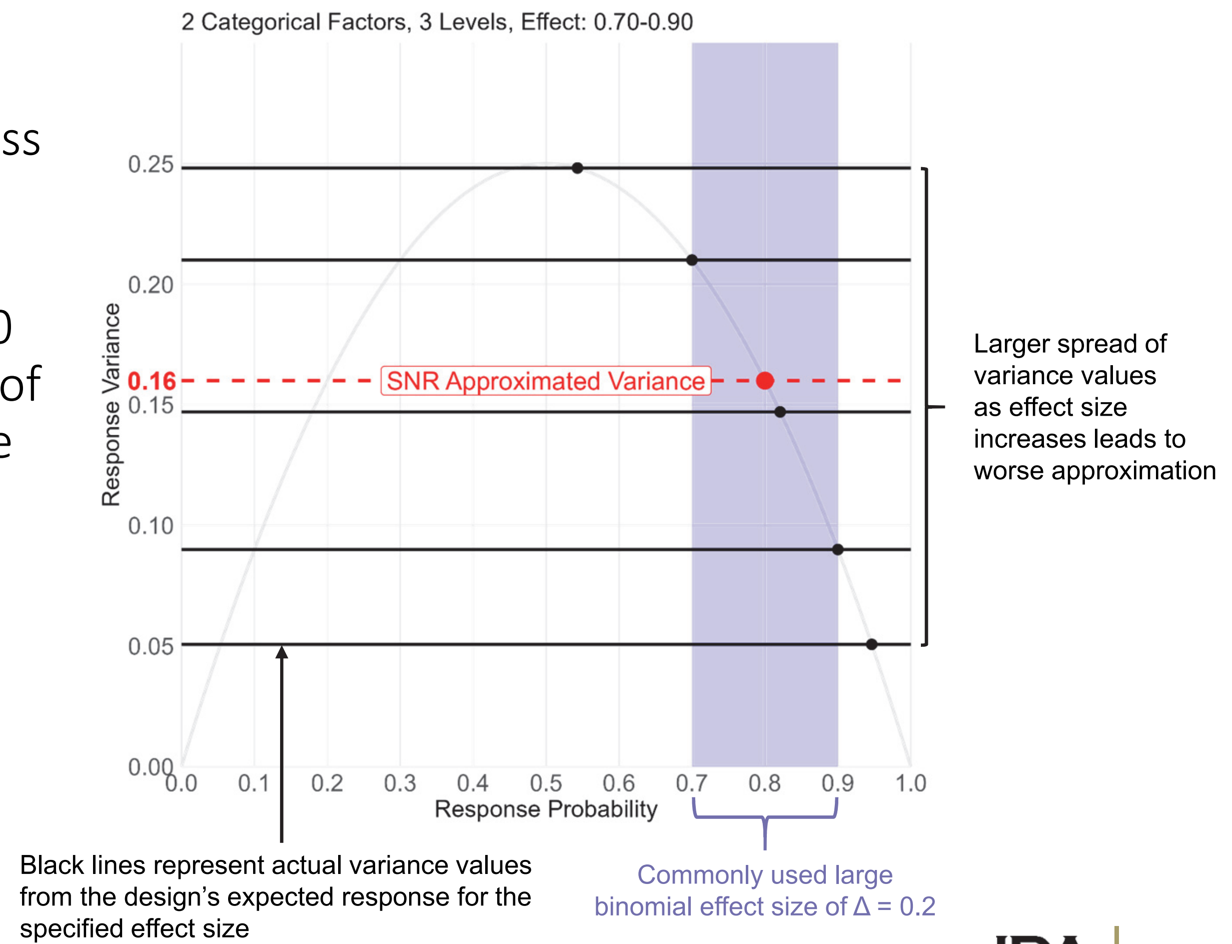
- The SNR approximation is a rough solution for the non-constant variance problem.
- The approximation uses the mid-point of the high and low probabilities to calculate a single variance value for the model.
- For simple designs with small effect sizes, this approximation appears reasonable—the difference between the approximated variance and the actual variance is relatively small and the variance values are evenly clustered around the approximation.



SNR: Signal to noise ratio

What is the SNR approximation?

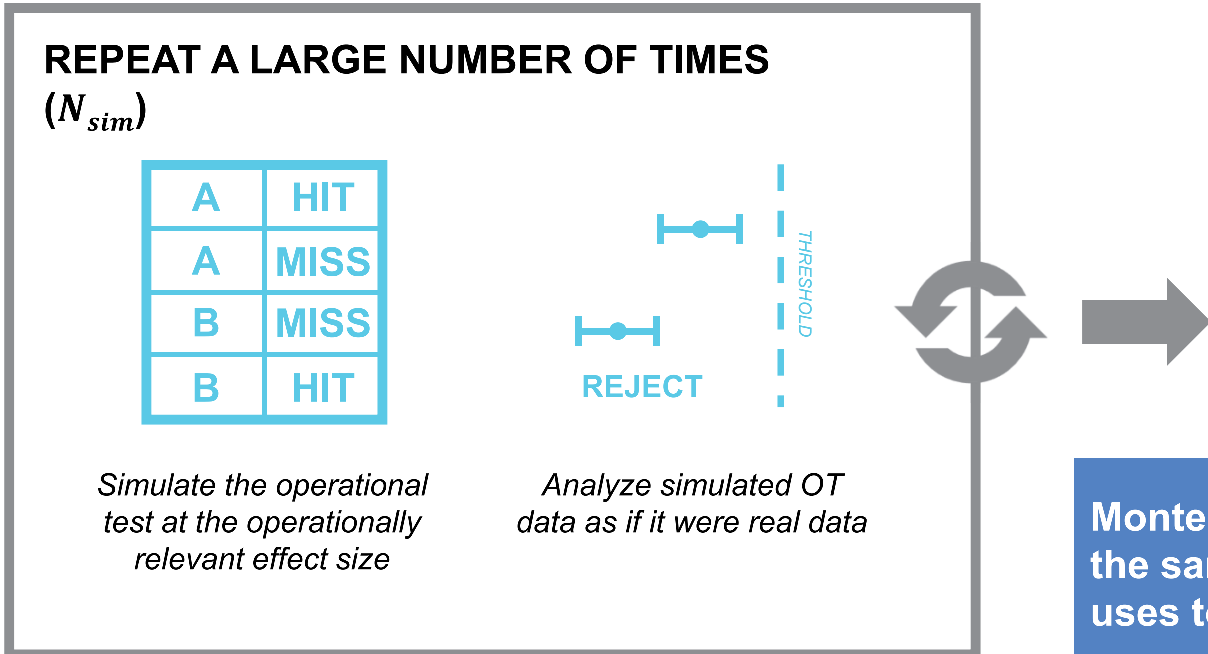
- This approximate value becomes less accurate for complex designs and more extreme probabilities.
- Example: An effect size of 0.70-0.90 leads to an approximated variance of 0.16, but the actual design variance ranges from 0.05 to 0.25.



SNR: Signal to noise ratio

Primary question: Is the SNR approximation an accurate method for calculating power?

Monte Carlo power computation obtains exact power numbers



$$\text{Power} = \frac{\text{Rejections}}{N_{sim}}$$

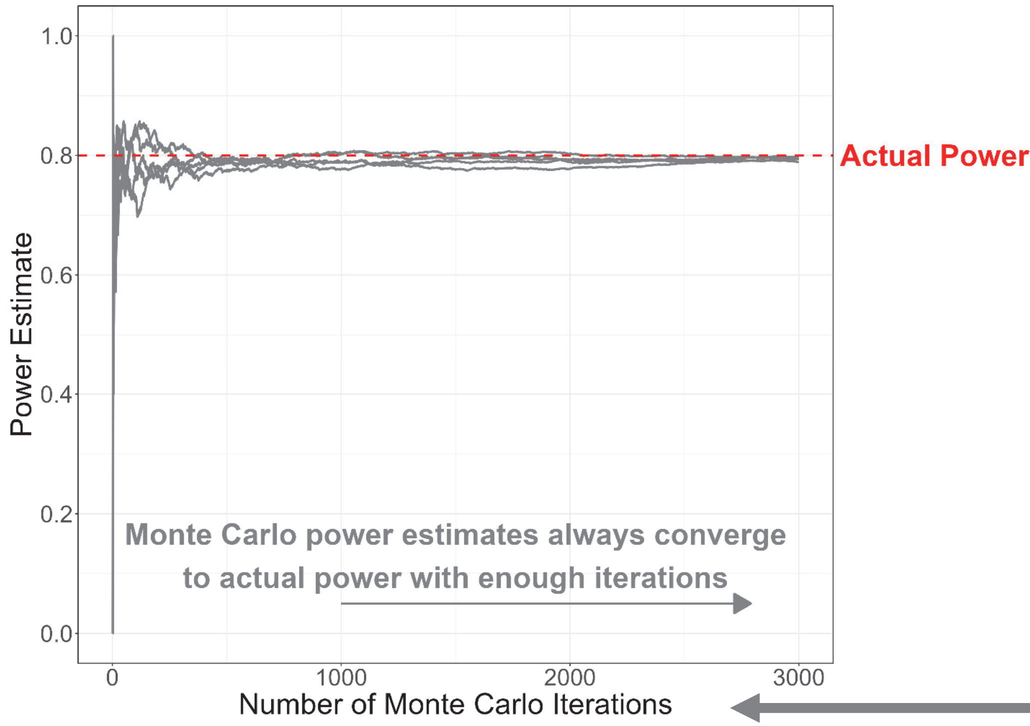
Count the number of times the effect is detected

Monte Carlo power is generated using the same analytical methods the tester uses to analyze the data.

This is the analytical equivalent of using production-representative test articles in realistic conditions for IOT&E.

OT: Operational Test; IOT&E: Initial Operational Test and Evaluation

Monte Carlo computation converges to the correct power for any sample size and test configuration



Monte Carlo power studies are the gold standard for evaluating methods both in academic and practitioner settings

10000+ simulations can be run in under a minute on standard consumer laptops

Topic #1: Assessing the accuracy and reliability of the SNR approximation

SNR approximation is a tool that can be assessed like any other system

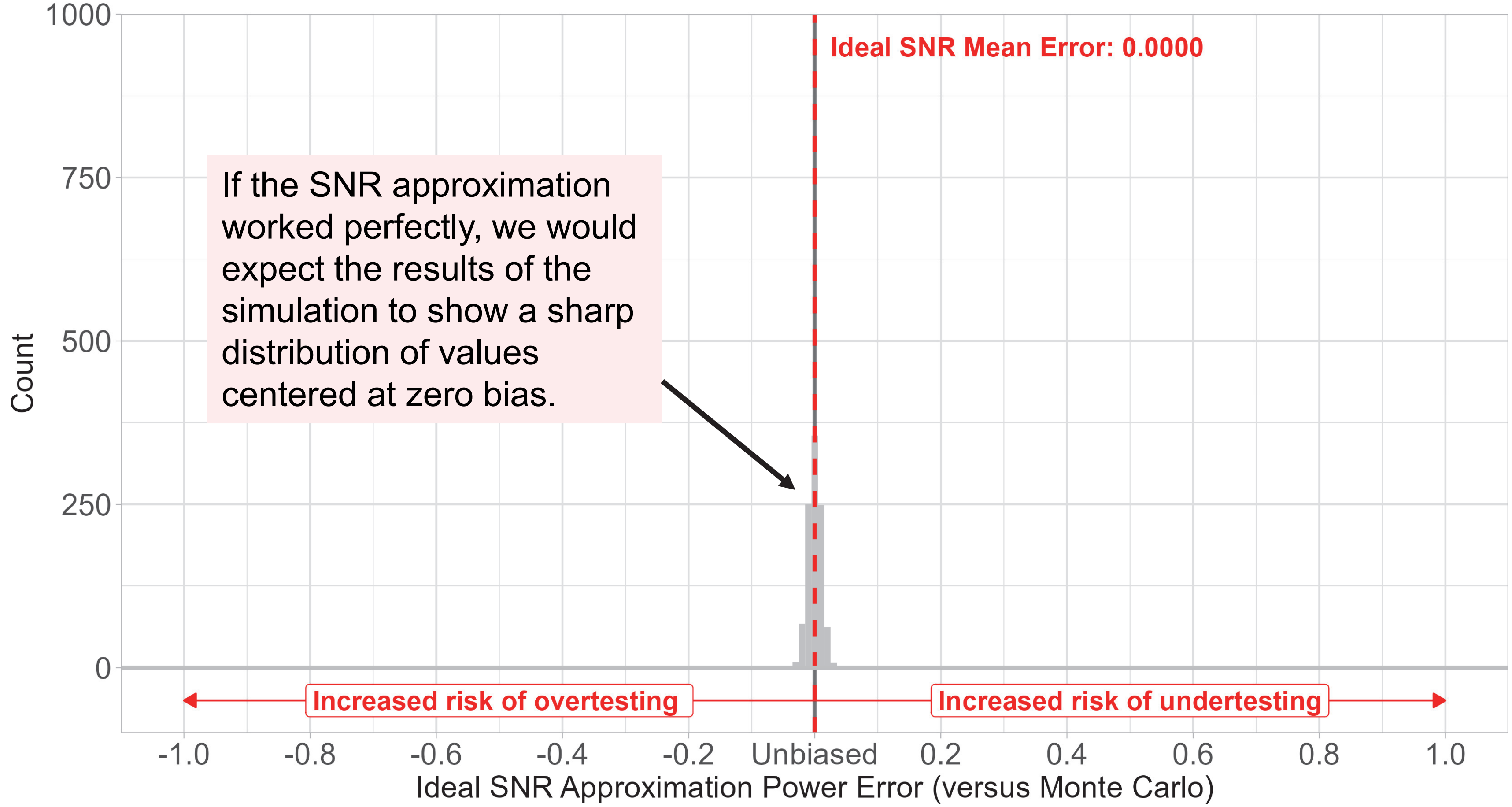
- How accurate is the SNR approximation? How effective is it as a tool to generate power values?
- We performed a simulation study across a wide range of designs comparing the accuracy of the SNR approximation to the “gold standard” Monte Carlo result.
  - Study is a full factorial design
  - Response: Bias in power as compared to the exact Monte Carlo result for each design and model term
  - 480 different designs included in study

SNR Approximation Study Input Factors

Input	Levels
Runs	20, 100
Effect Size	**Wide**: 0.4-0.6, 0.7-0.9, 0.79-0.99 **Narrow**: 0.45-0.55, 0.8-0.9, 0.89-0.99
Design Factors	3 continuous 3 categorical (3 levels) 3 categorical (5 levels) 3 continuous, 3 categorical (3 levels) 3 continuous, 3 categorical (5 levels)
Type-I Error	0.01, 0.20
Optimality	D, I
Model	Main effects only, main effects + interactions

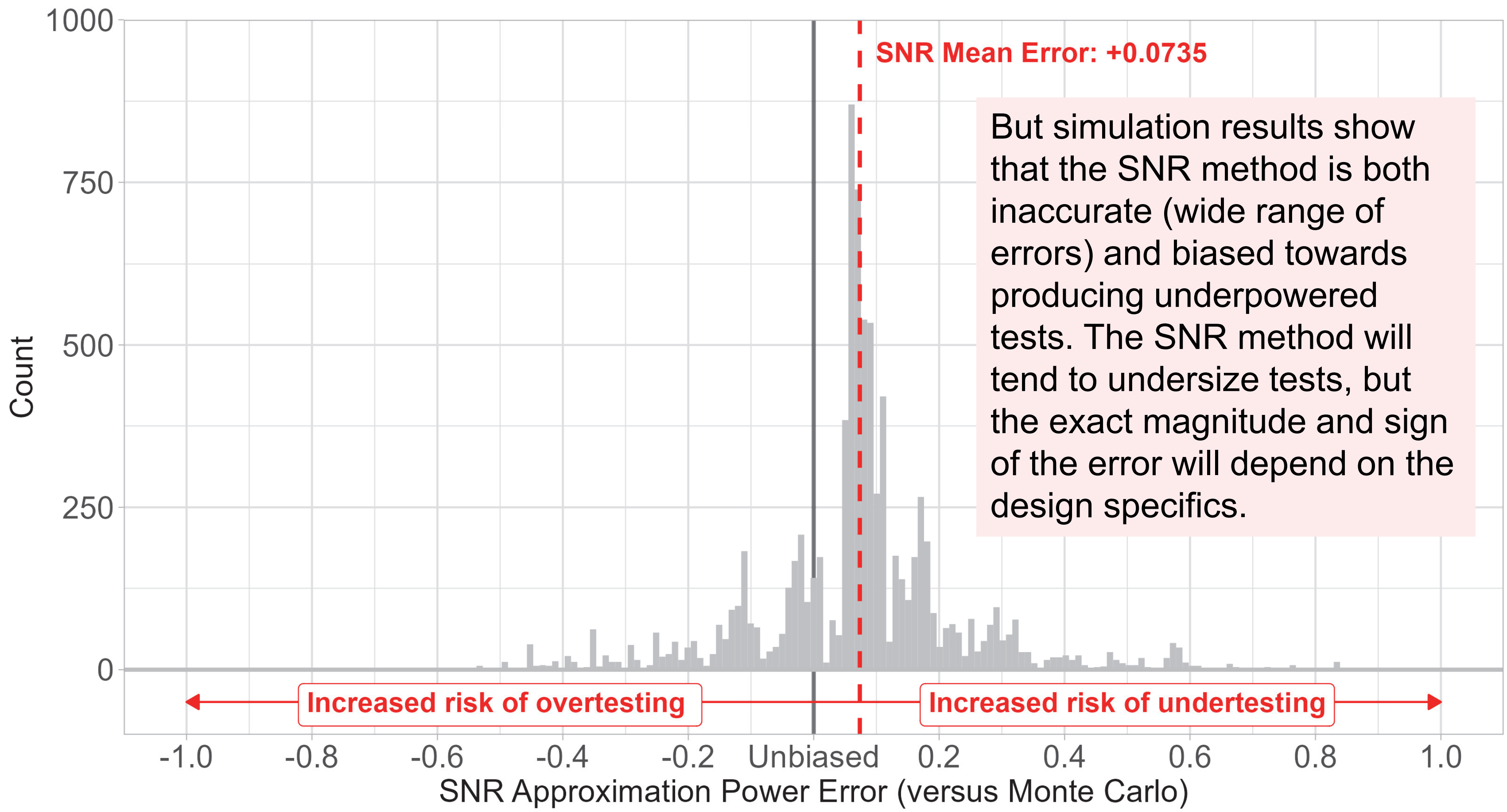
SNR: Signal-to-Noise Ratio; N: Number of simulations; DOE: Design of Experiments

The ideal distribution of bias values if the SNR approximation worked



SNR: Signal-to-Noise Ratio

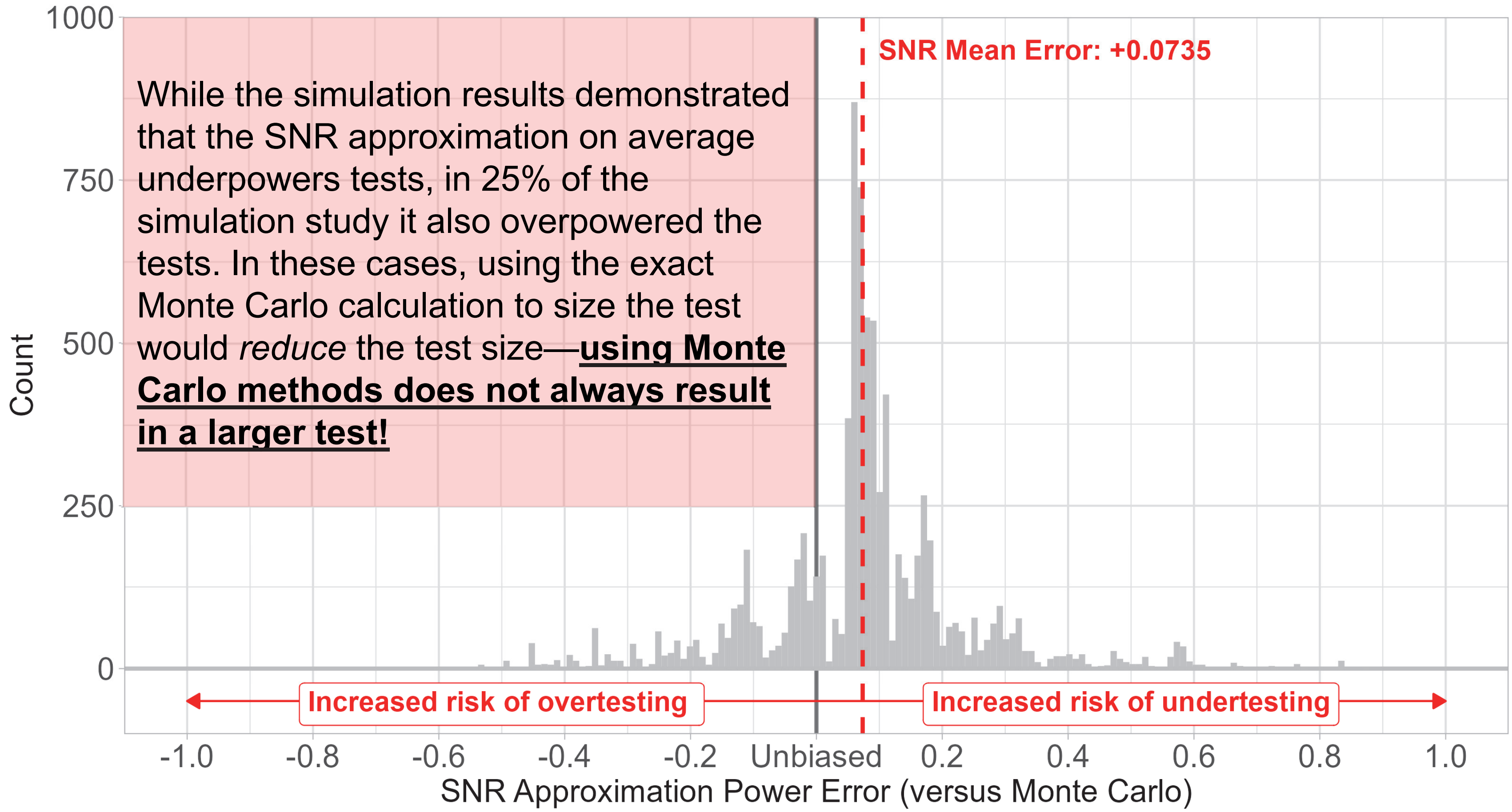
Simulation results show the SNR approximation is inaccurate and biased



But simulation results show that the SNR method is both inaccurate (wide range of errors) and biased towards producing underpowered tests. The SNR method will tend to undersize tests, but the exact magnitude and sign of the error will depend on the design specifics.

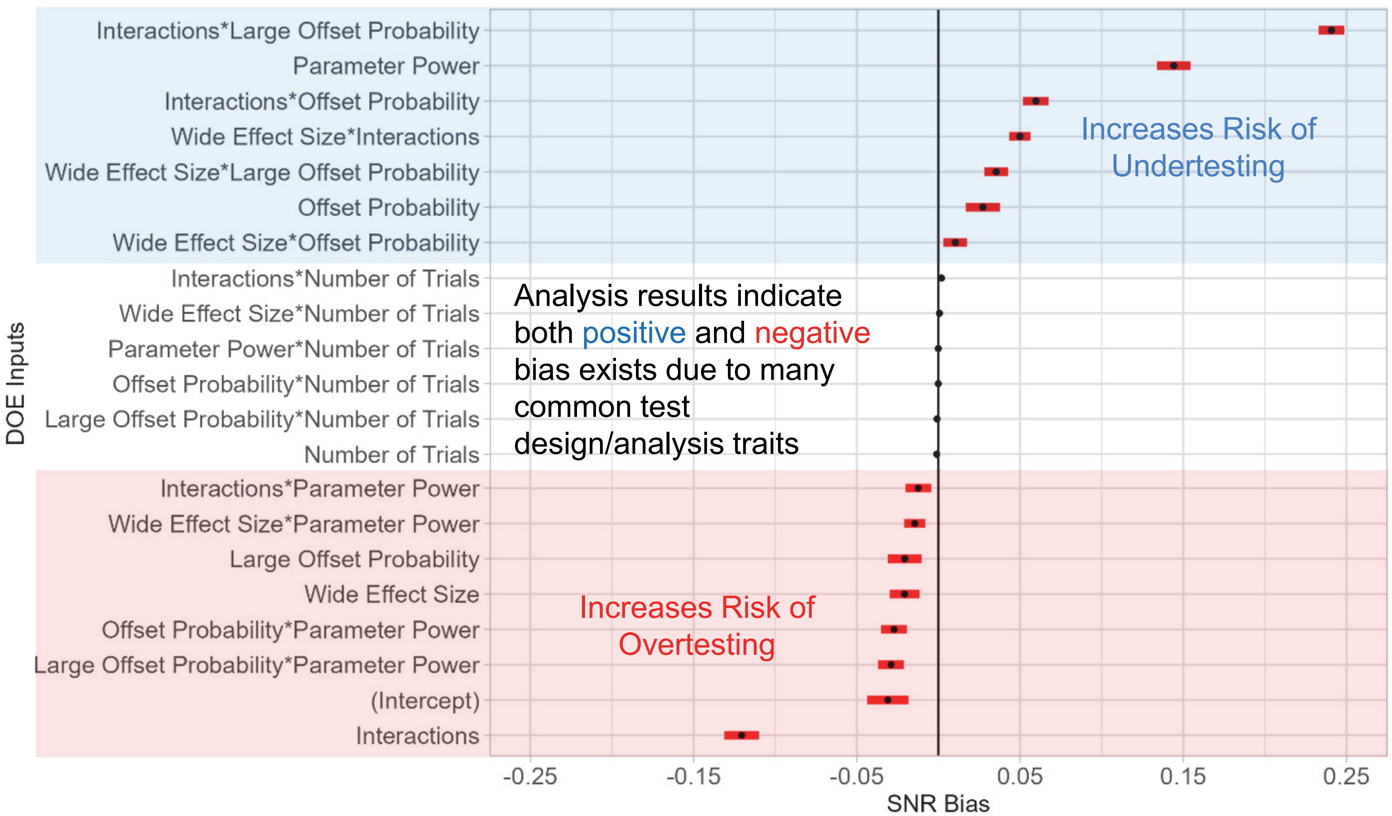
SNR: Signal-to-Noise Ratio

The SNR approximation can also waste test resources



SNR: Signal-to-Noise Ratio

SNR bias is complex and can be attributed to a wide range of factors

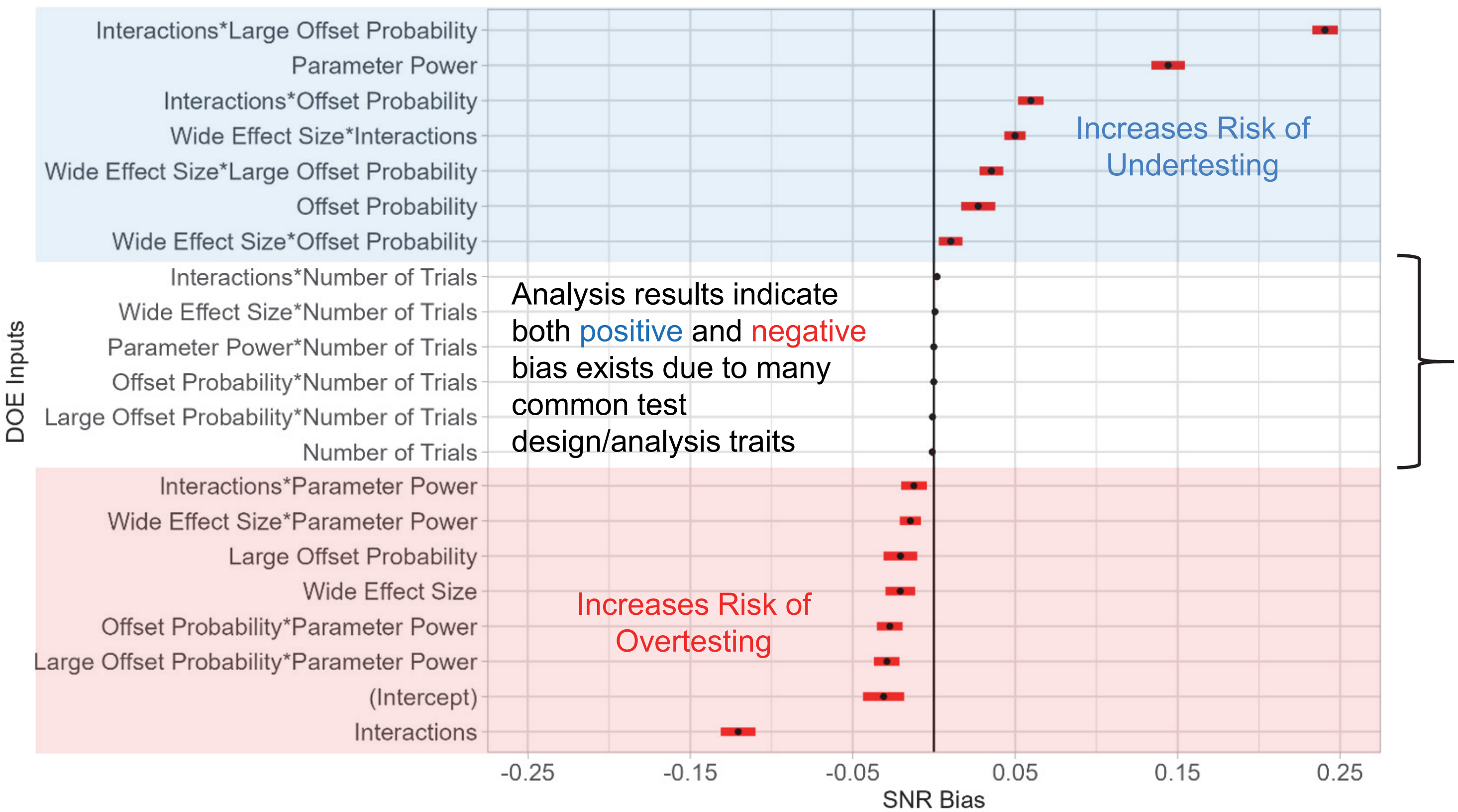


The only term which had no effect on bias is the number of trials, which means the SNR ***approximation bias does not disappear with increasing design size***, and thus you cannot expect the bias to disappear even if the design is relatively large.

The accuracy of the SNR approximation depends on the specifics of each test and is unknown without running a Monte Carlo simulation.

SNR: Signal-to-Noise Ratio

SNR bias is complex and can be attributed to a wide range of factors



The only term which had no effect on bias is the number of trials, which means the SNR ***approximation bias does not disappear with increasing design size***, and thus you cannot expect the bias to disappear even if the design is relatively large.

The accuracy of the SNR approximation depends on the specifics of each test and is unknown without running a Monte Carlo simulation.

Takeaway: The SNR approximation is an unreliable method for computing power.

SNR: Signal-to-Noise Ratio

So what other tools can be used?

Easy Monte Carlo power calculation interface provided by externally-vetted tool "skpr"

- The **skpr** package is a peer-reviewed R library designated a "core" DOE package by the Comprehensive R Archive Network (CRAN) and published in the Journal of Statistical Software.
- This package has been used/cited by:
  - DoD test organizations and OTAs
  - Academic research institutions
  - Pharmaceutical companies
  - And more!
- <u>The skpr package is free!</u>

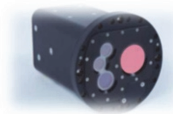
<u>DoD Test Organizations using skpr</u>

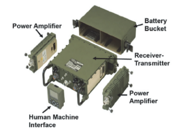
<u>Research institutions using skpr</u>

DOE: Design of Experiments; DoD: Department of Defense; OTA: Operational Test Authority

Many DoD programs have successfully used skpr for DT/OT test planning

Examples:
(not exhaustive)


Advanced Threat Warner



HMS


Stryker ATGM

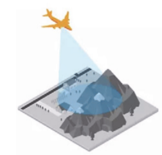

CSASS


CIRCM


NGSW


ENVG-B


IVAS


ALT NAV


Active Protection System


CMWS


Advanced Threat Infrared Countermeasures


UH-60V


PrSM


ER GMLRS


Apache gun accuracy and APSI


Manpack


FMTV


TacDS

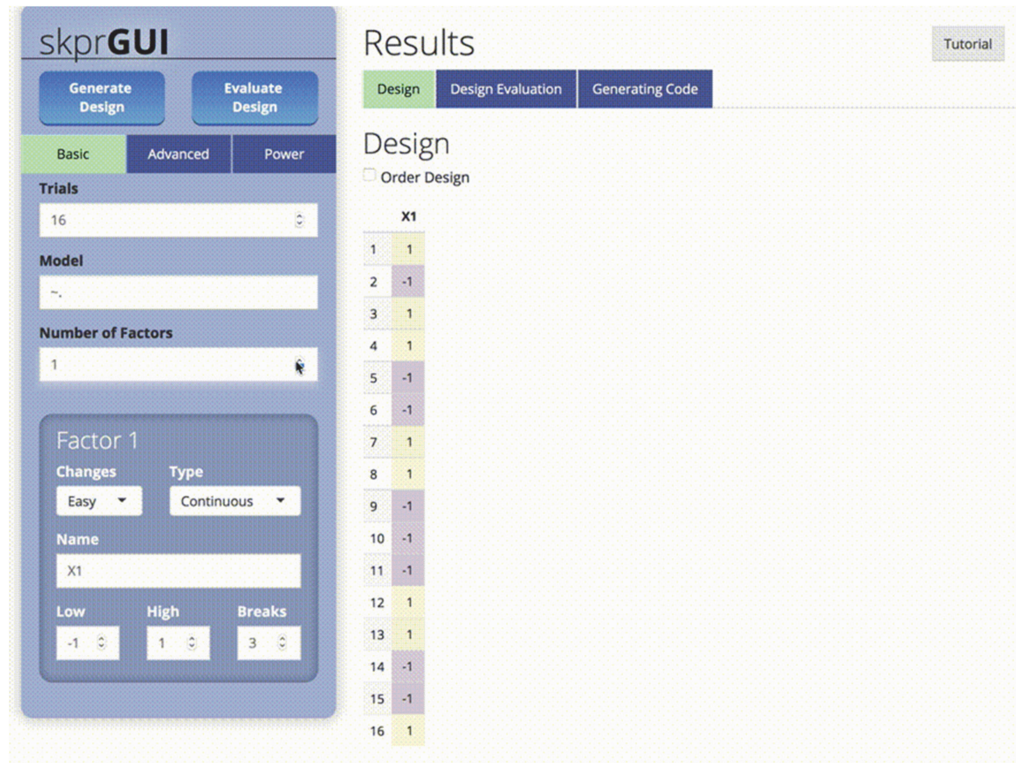
IPPS-A


TUAS (SHADOW)

IDA

No code required: skpr has a user-friendly graphical user interface

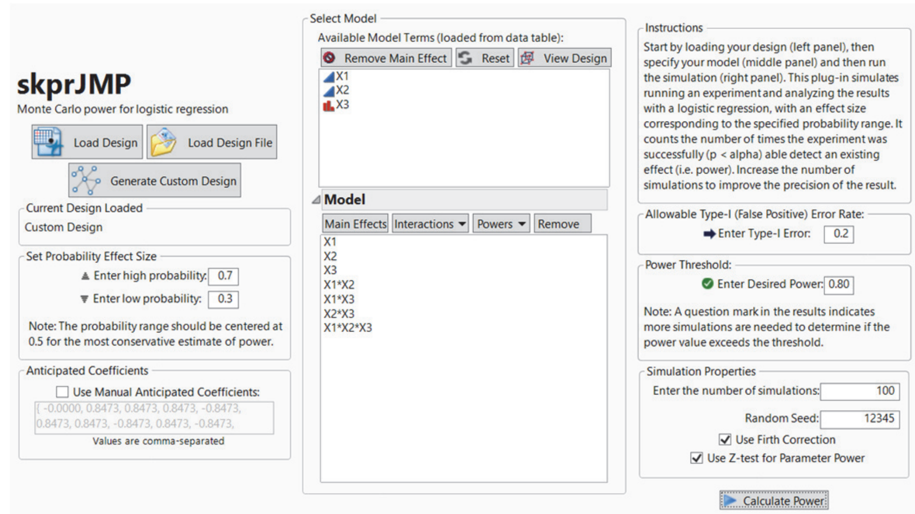
- While one of the main benefits of skpr is scripting and automating complex DoE analyses in code, the package also provides a web-based graphical user interface (GUI) for non-R users.
- Previous versions of the GUI only supported a limited number of factors—as of skpr version 1.6.2, **the GUI now supports an unlimited number of input factors.**

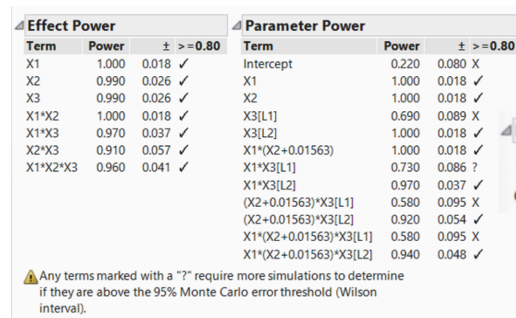


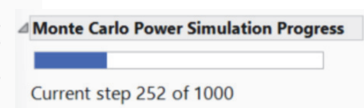
DoE: Design of Experiments;

For JMP users: check out skprJMP!

- Key Features:
  - **Design Import and Generation**: skprJMP allows analysts to **import an existing design** from their filesystem or JMP session as well as **use JMP's custom design interface** to generate designs directly within the add-in.
  - **Model Specification**: Includes a **JMP-style interface for specifying and building a model**, providing users with a point-and-click interface to specify interactions and add and remove model terms.
  - **Reproducibility**: The user can **specify a random seed** to ensure that they can reproduce the results of any single power analysis at a future date.
  - **Effect Size Calculation**: skprJMP **automates the calculation** of binomial effect sizes using a user-friendly probability range as the input.
  - **Firth Correction Support**: Logistic regression can be degraded from a type of degeneracy called "separation," where the model fails to converge for certain arrangements of data. skprJMP provides an **option to analyze the simulations using a Firth correction**, which fixes this issue.







Find me at the poster session for a live demo of both skpr and skprJMP!

Conclusions

- The SNR approximation provides a poor estimate of power and using it increases the risk of under-testing *and* over-testing.
- For accurate power assessments, use Monte Carlo methods to calculate power whenever there is no exact, analytic method available.
- If you use R, you can use the skpr package (available on CRAN) to both generate optimal designs and evaluate them with accurate Monte Carlo methods.
- If you use JMP, check out the skprJMP add-in (available on testscience.org) for easy Monte Carlo power estimates for probabilistic responses.

OTA: Operational Test Authority

Back-up Slides

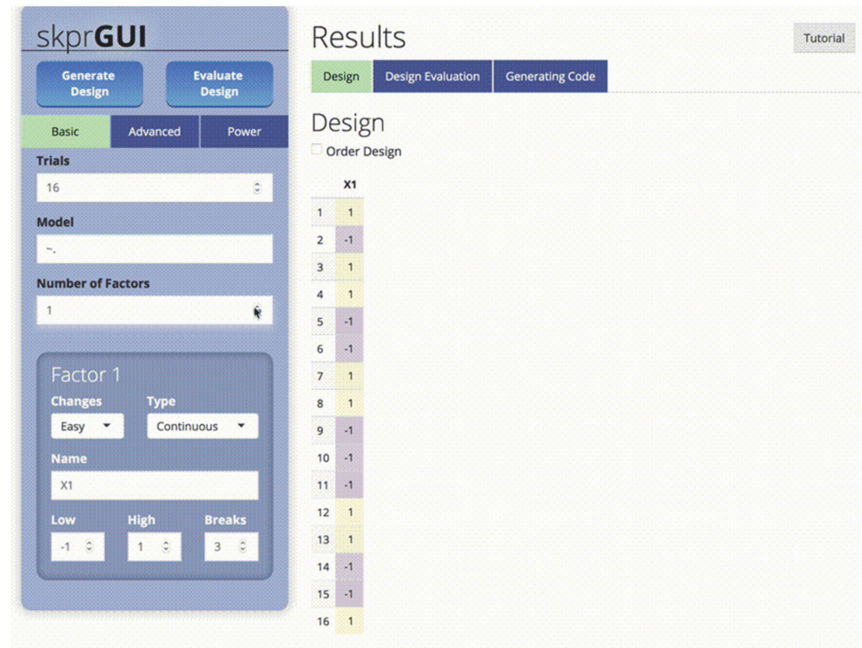
Power methodology must be linked to analysis methodology

- The underlying issue is the mismatch between the power calculation methodology and the analysis methodology: calculating power assuming the use of a linear model—like the SNR approximation does—and then using a different model during the actual data analysis may result in wrong power values being used to size the test. This risk can lead to wasted test resources if the test is overpowered or being unable to detect significant performance issues if the test is underpowered.
- Monte Carlo methods do not suffer from this risk: the same tools are used for both the power analysis and the data analysis.
- This is the analytical equivalent of requiring production-representative test articles in realistic conditions for IOT&E: we want our power numbers to reflect the exact methods and process used in the final analysis.

IOT&E: Initial Operational Test and Evaluation;

Getting started with skpr: Learning resources and courses

- DATAWorks 2023 workshop: "Design of Experiments in R with the skpr Package"
  - Video recording (IDA's YouTube channel): www.youtube.com/watch?v=vMyHZTNCsKA
  - Course materials: https://github.com/tylermorganwall/skpr_dataworks_workshop/
  - Course website: https://tylermorganwall.github.io/skpr_dataworks_workshop/
- Journal of Statistical Software Paper: "Optimal Design Generation and Power Evaluation in R: The skpr Package"
  - https://www.jstatsoft.org/article/view/v099i01



Simulation Study Design Power

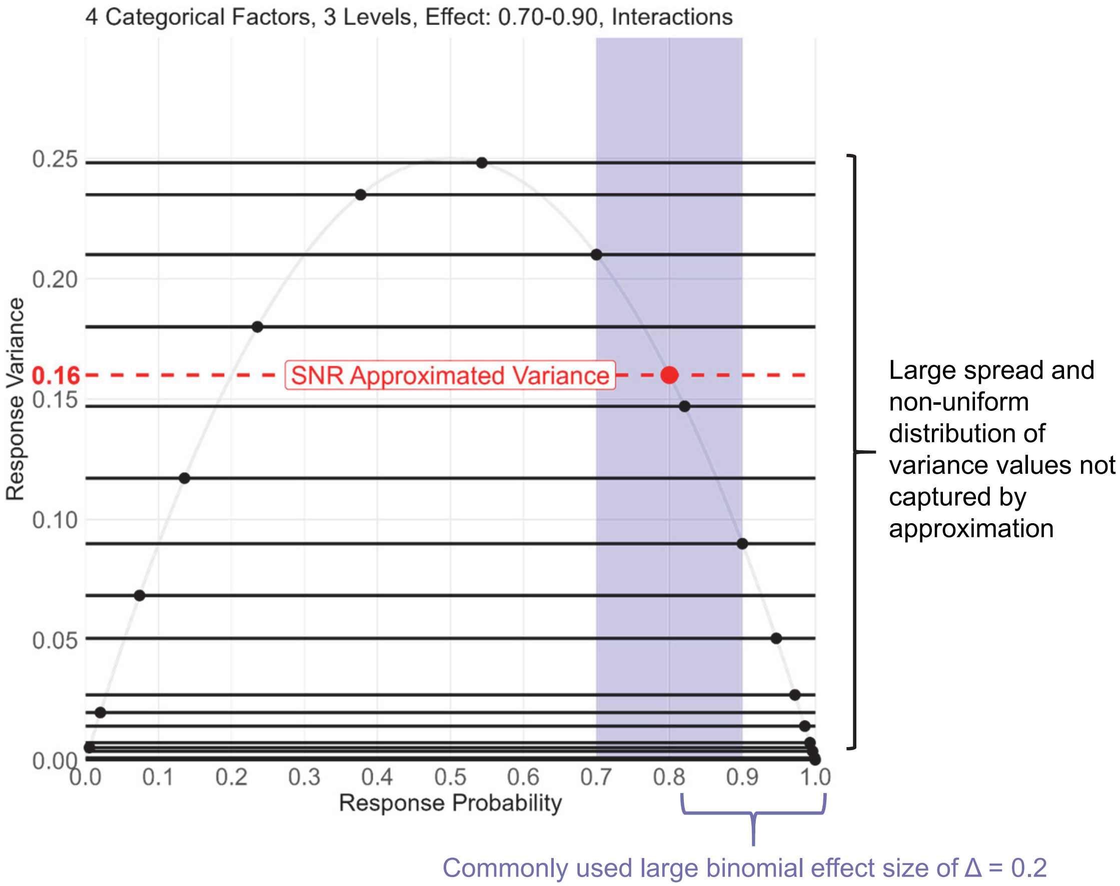
- Full factorial design
- Continuous response:
  Bias in power as compared to the exact Monte Carlo result
- N = 480
- Type-I error: 0.10
- Power = 90% power (main effects + interactions) model to detect a power bias of 0.06.

Input	Levels
Runs	20, 100
Effect Size	**Wide**: 0.4-0.6, 0.7-0.9, 0.79-0.99 **Narrow**: 0.45-0.55, 0.8-0.9, 0.89-0.99
Design Factors	3 continuous 3 categorical (3 levels) 3 categorical (5 levels) 3 continuous, 3 categorical (3 levels) 3 continuous, 3 categorical (5 levels)
Type-I Error	0.01, 0.20
Optimality	D, I
Model	Main effects only, main effects + interactions

SNR: Signal-to-Noise Ratio; N: Number of simulations

Moderate design/model complexity further degrades approximation

- Minor to moderate design complexity (here, 4 categorical factors with 3 levels and interaction terms) results in an even worse approximation.
- Example: 1000x difference between lowest variance (0.00025) and highest value (0.25), all being approximated by a single variance of 0.16.
- Non-uniform clustering of variance values at edges of probability space not captured by approximation.
- <u>Almost all test designs in the DoD test community fall into the "complex" category by this standard.</u>



GLM: Generalized Linear Models

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
XX/04/2024	Final	

4. TITLE AND SUBTITLE

DATAWorks 2024: Simulation Insights on Power Analysis with Binary Responses--From SNR Methods to 'skprJMP'

5a. CONTRACT NUMBER
HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Tyler Morgan-Wall
Curtis Miller
Robert Atkins

5d. PROJECT NUMBER
BD-9-2299(37)

5e. TASK NUMBER
229937

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

3002093

3002093

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington D.C. 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Robert W Atkins

14. ABSTRACT

Logistic regression is a commonly-used method for analyzing tests with probabilistic responses in the test community, yet calculating power for these tests has historically been challenging. This difficulty prompted the development of methods based on signal-to-noise ratio (SNR) approximations over the last decade, tailored to address the intricacies of logistic regression's binary outcomes. However, advancements and improvements in statistical software and computational power have reduced the need for such approximate methods. Our research presents a detailed simulation study that compares SNR-based power estimates with those derived from exact Monte Carlo simulations, highlighting the inadequacies of SNR approximations. To address these shortcomings, we will discuss improvements in the open-source R package "skpr" as well as present "skprJMP," a new plug-in that offers more accurate and reliable power calculations for logistic regression analyses.

15. SUBJECT TERMS

operational testing; logistic regression; statistical power; signal-to-noise ratio; monte carlo

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Robert W Atkins
Unclassified	Unclassified	Unclassified	SAR	TBD	**19b. TELEPHONE NUMBER** *(Include area code)* 703-933-6546