



Exceptional service in the national interest

## EXPLORING THE ROBUSTNESS OF BAYESIAN ADAPTIVE DESIGN OF EXPERIMENTS

Daniel Ries

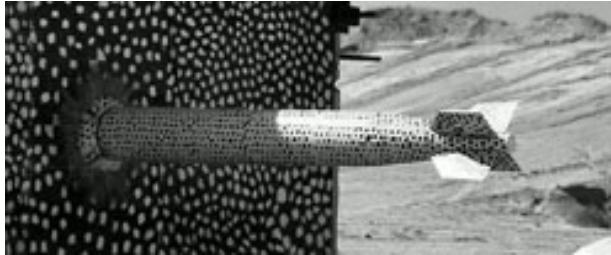
DATAWorks 2022

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.



# Motivation

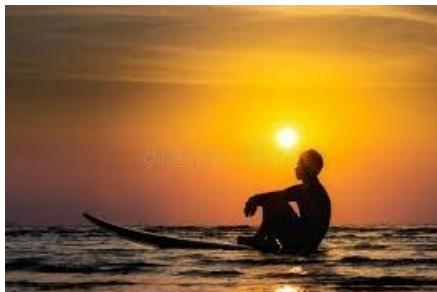
Because of the cost and complexity of our experiments, we want to run the **minimum** number of tests to walk away with the data needed





# Bayesian Adaptive Design of Experiments (BADE)

Traditional DOEx = Static



wait until test is complete to perform statistical analysis

- Statistical guarantees like power are pre-planned

Applies to BADE too!

Bayesian Adaptive DOEx = Dynamic



continuously update statistical analysis as new data comes, allowing **early stopping**

- Use collected data and “play out” the different ways the remaining data could emerge. Make decisions based on the relative frequencies of these “played out” scenarios

If you know after 15 tests that there's a 99% chance you would conclude X at the end 20-test series, why shouldn't you stop and save resources?

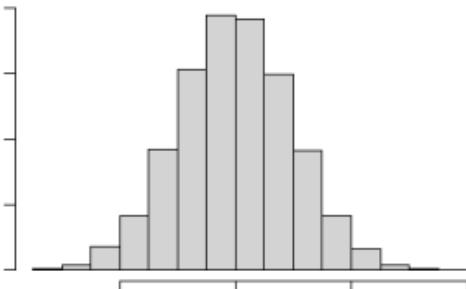
BADE can give you this probability!

# Assessing the Robustness of BADE

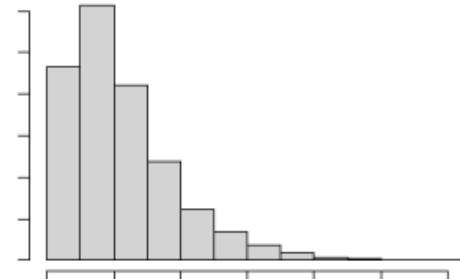
However, defense applications are **high-consequence** by nature, so we want to make sure stopping early is optimal decision under varying conditions

Therefore, we need to assess **robustness** to *model* and *prior*

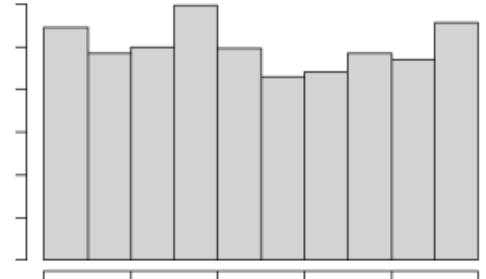
What if your model expects data that looks like:



But your data ends up looking like:



OR



Will BADE early stopping decisions still be correct?



# Model<sup>1</sup> and Robustness Simulation Study

**Model:**

$$Z_i \sim N(\mu, \sigma^2), i = 1, \dots, n_t$$

$$p(\mu, \sigma^2) \propto \frac{1}{\sigma^2}$$

**Quantity of Interest:**

$$\phi = P(2 < Z_i < 5)$$

**Question:**

Is  $\phi > 0.8$ ?

**Update stopping rule:**

$$PP = P_{Y|X}(Y: P(\phi > .8 | X, Y) > .9)$$

$Z = (X, Y)$ ,  $X$  is observed,  $Y$  is unobserved

**Stop early if at any point:**

PP>.95: stop testing and conclude  $P(\phi > 0.8) > 0.9$

PP<.05: stop testing and conclude  $P(\phi > 0.8) < 0.9$

**Simulated data family:**

Normal, Gamma, Uniform

**Total sample size:**

50,100

**Outliers:**

none, 1 extreme

**Effect size** (true  $\phi$  for simulated data):

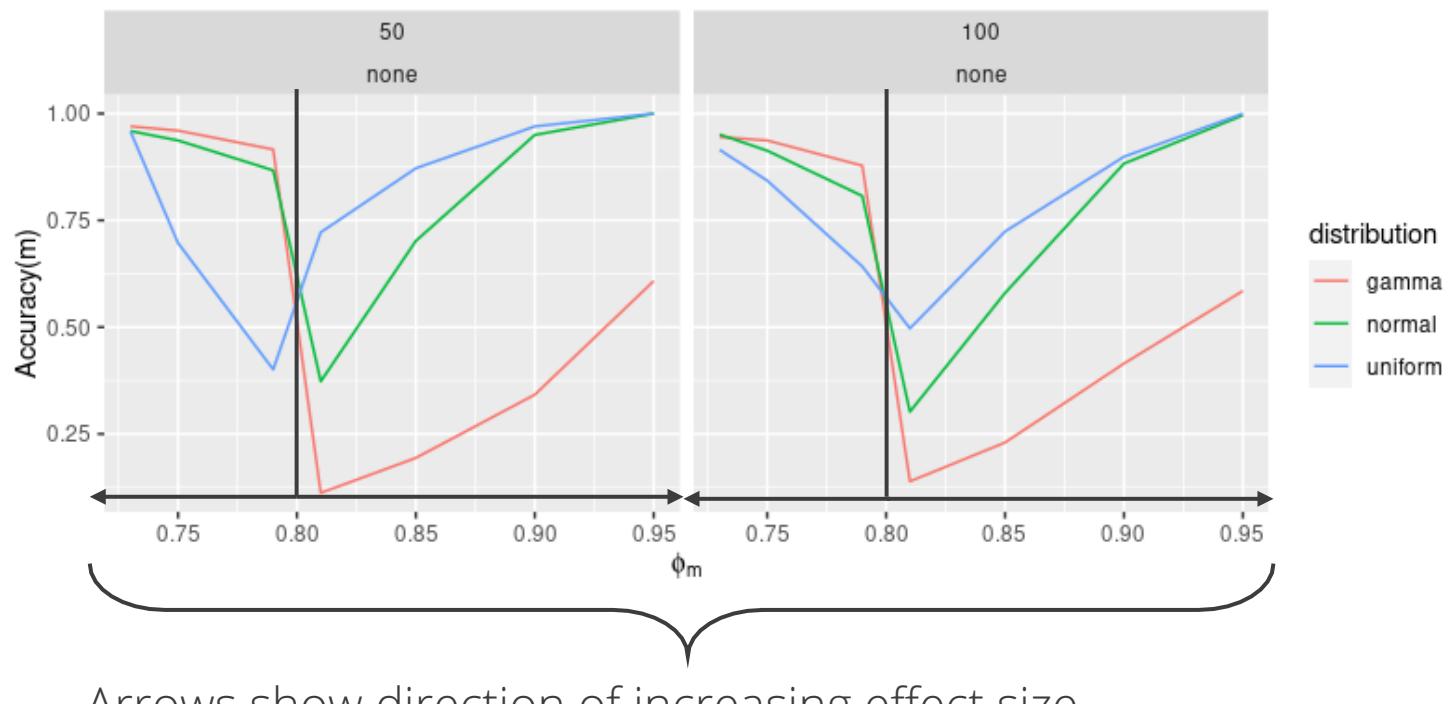
.73,.75,.79,.81,.85,.9,.95

\*Recompute PP after every additional 5 observations, starting at  $n_0 = 10$  observations

# Results

- When there is a large effect size, BADE is generally reliable
  - Varies by simulated data distribution
  - Planned sample size doesn't affect much
    - This shouldn't be surprising
- When the effect size is low early stopping decisions are inaccurate

Proportion of times the early stopping decision was “correct” for different simulated data sets



Arrows show direction of increasing effect size

**Takeaway:** Early stopping using BADE is justifiable when the true “effect size” is large, otherwise there is potential for unreliable decisions



Thank you for listening!

Visit my poster later for more details!  
[dries@sandia.gov](mailto:dries@sandia.gov)