Verification & Validation of Modeling & Simulation based on Testing for Uniformity

April 11, 2019

Dr. Shannon Shelburne
Joint Research and Development, Inc.

Overview

Research Question: Find a way to compare modeling and simulation (M&S) and Live-Fire (LF) results

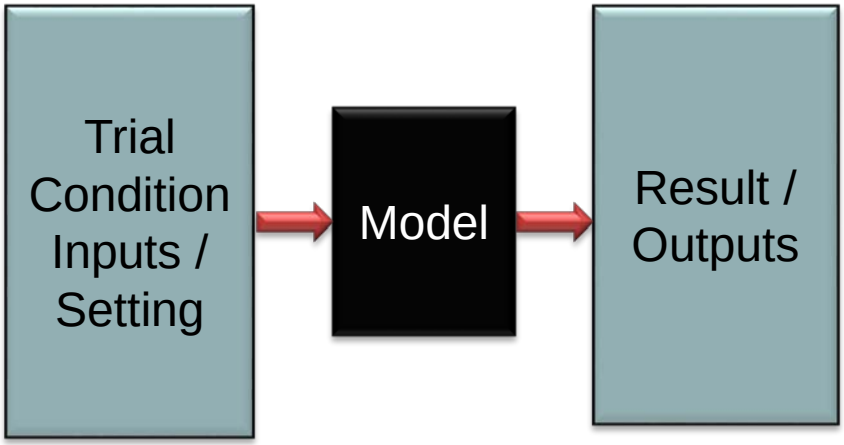
Objective: Summarize research to compare based on testing for uniformity

- Provide an overview of the problem
- Review applicable combined probability and goodness-of-fit tests
- Present the results of a power study
- Provide results based on an example situation

M&S vs. LF

M&S

- Relatively Quick and Cost Effective
  - Multiple Trials
  - Multiple Trial Conditions
- Results Based on Assumptions



LF

- Typically Time Consuming and Costly
  - Very Limited Trials
  - Very Limited Trial Conditions
- Results Based on Real Conditions

Photo Credits:
https://www.atc.army.mil/directorates/Firepower.html
https://www.atc.army.mil/directorates/Surv_Leth.html

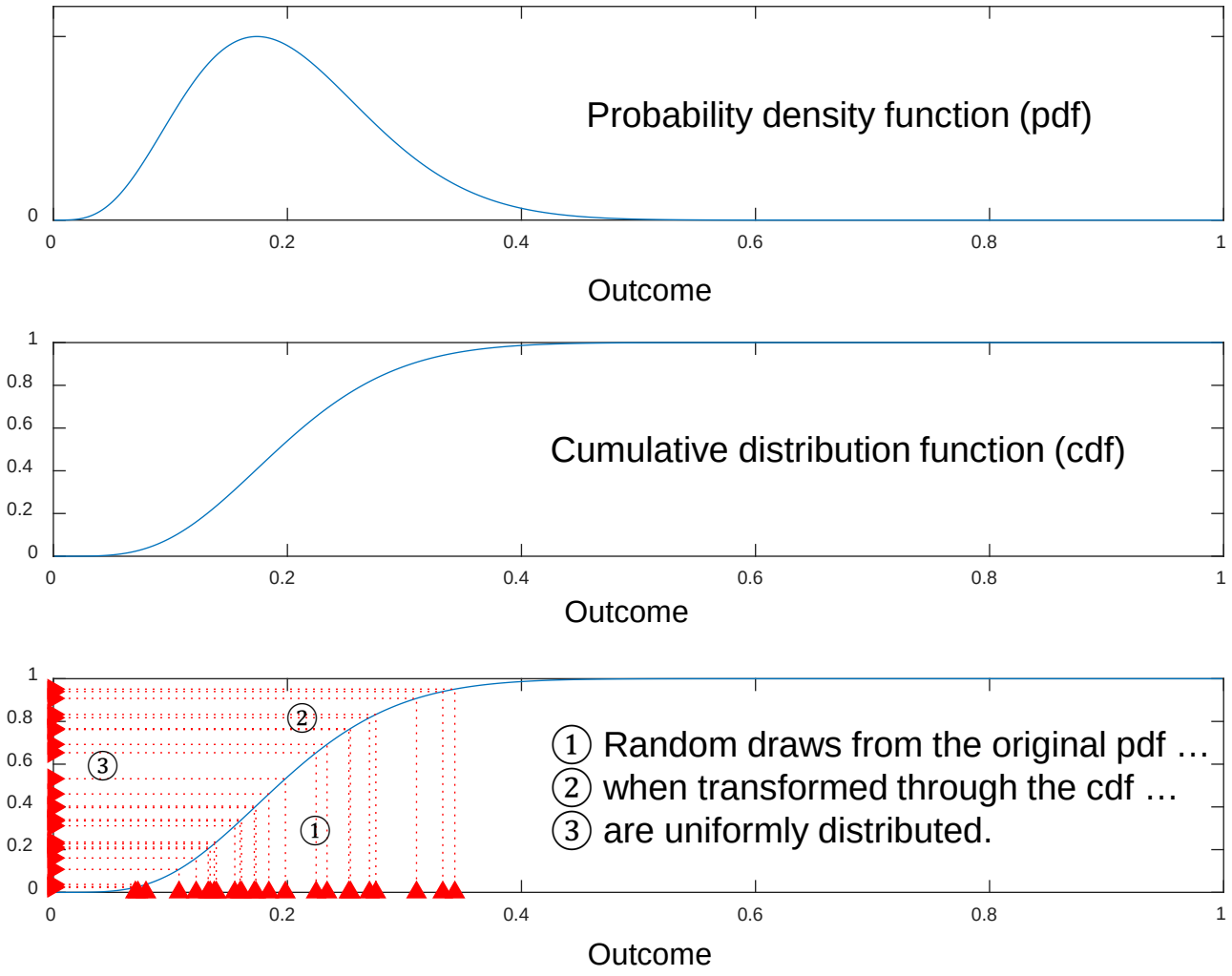
Question: Do the LF results and the M&S outcomes agree?

Dilemma: How to show agreement between a set of a 1,000 simulated values and only 1 LF result?

Key:

- Determine percentile for each LF result
- Make use of statistical property that:
  - Under the assumption that LF and the Model share the same distribution for any given shot condition
  - These percentiles are distributed U(0,1)

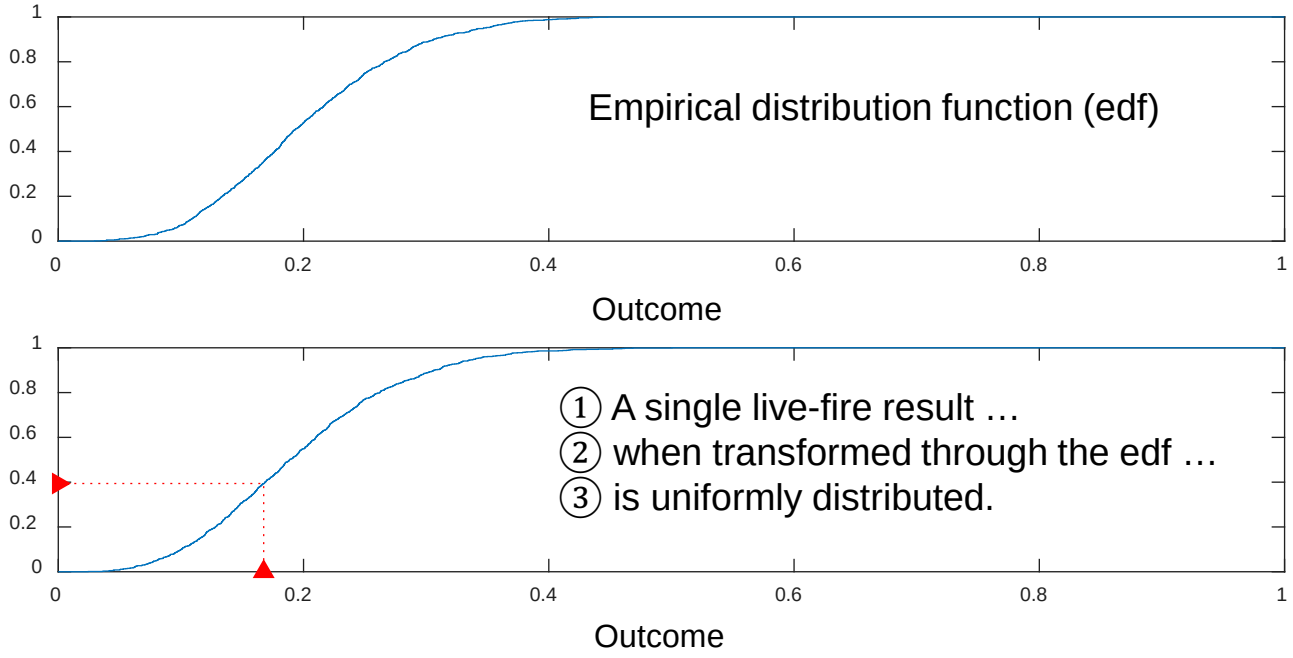
For example,


Probability density function (pdf)
0
0.2
0.4
0.6
0.8
1
Outcome
Cumulative distribution function (cdf)
Outcome
① Random draws from the original pdf …
② when transformed through the cdf …
③ are uniformly distributed.
Outcome

In practice,

(1) We don't know the true CDF, but we can estimate it from an empirical distribution function based upon many model replications

(2) We only have one live-fire result



Given a Live-Fire Result of w*, and a large number (M) of model outputs (w_1, w_2, …, w_M), the transformed value, called "p", is

$$p = \text{“number of } w_i \le w^* \text{”} / M$$

i.e., the proportion of times that the Model Output is at or below the Live-Fire Result

Strategy

Shot Cond'n	Model Output	Live-Fire Result	Percentile
1	$\{X_{1,1}, X_{1,2}, \ldots, X_{1,1000}\}$	LF_1	p_1
2	$\{X_{2,1}, X_{2,2}, \ldots, X_{2,1000}\}$	LF_2	p_2
⋮	⋮	⋮	⋮
N	$\{X_{N,1}, X_{N,2}, \ldots, X_{N,1000}\}$	LF_N	p_N

Conduct a test of uniformity on the entire set of percentiles as a collective test for agreement between LF and the Model

Under the null hypothesis,

1) For Test Condition (TC) #1, the Model Output and the single Live-Fire Result combine to yield a value p_1 from a uniform distribution

2) For TC #2, we get p_2, also from a uniform distribution

.
.
.

N) For TC #N, we get p_N, also from a uniform distribution

How to Test

Collectively test the values p_1, p_2, …, p_N for uniformity

- Option 1: Combined Probability Test methods
- Option 2: Goodness-of-Fit (GOF) methods for the uniform distribution

Combined Probability Tests

- H_0: The distribution of Model Outcomes is the same as the distribution of Live-Fire Results, for each test condition

- H_a: The distribution of Model Outcomes is different from the distribution of Live-Fire Results, for each test condition

Popular Combined Probability Tests

- Fisher's combined probability

Additional Combined Probability Tests

- Maximum P
- Minimum P
- Sum Z
- Sum P
- Logit
- Mean P
- Mean Z

R: metap package

GOF Testing

- In general:

  H_0: Data come from a ***specified*** distribution

  H_a: Data do not come from the ***specific*** distribution

- We are interested in:

  H_0: Data come from the U(0,1) distribution

  H_a: Data do not come from the U(0,1) distribution

Popular GOF Methods

- Anderson-Darling (AD)
- Chi-Square (C2)
- Kolmogorov-Smirnov (KS)

Additional GOF Methods

- Cramér-von Mises (CVM)
- Neyman/Neyman-Barton (NB)
- Dudewicz-van der Meulen (DVDM)
- Quesenberry-Miller (QM)

R: AD and CVM using the goftest package; C2 and KS using the EnvStats package; others using the uniftest package

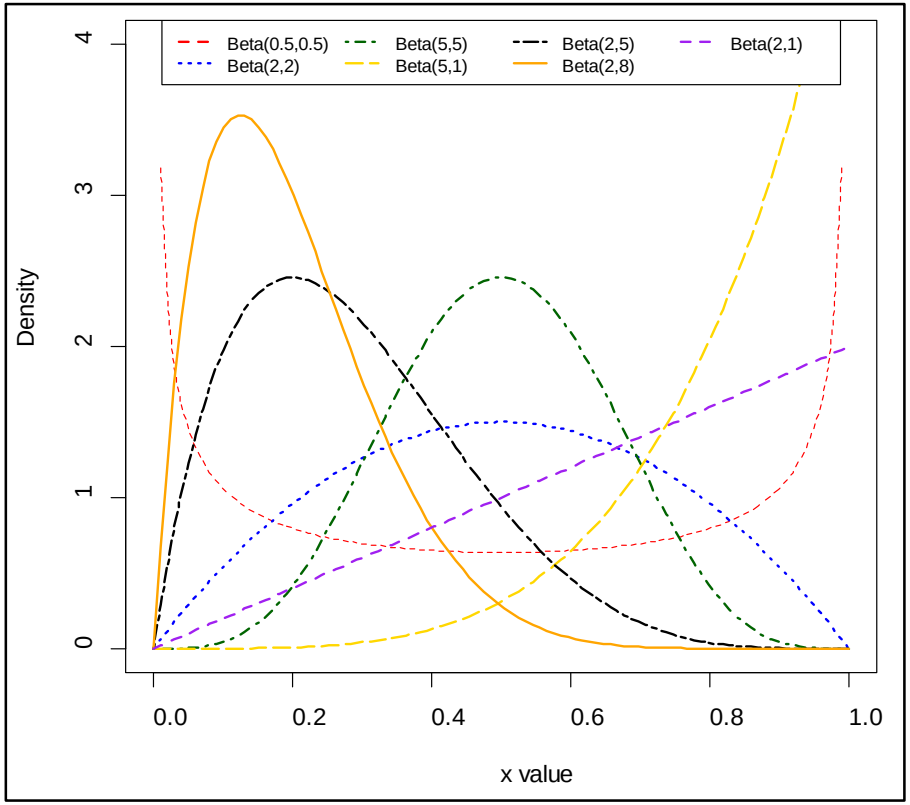
Power Study Assumptions

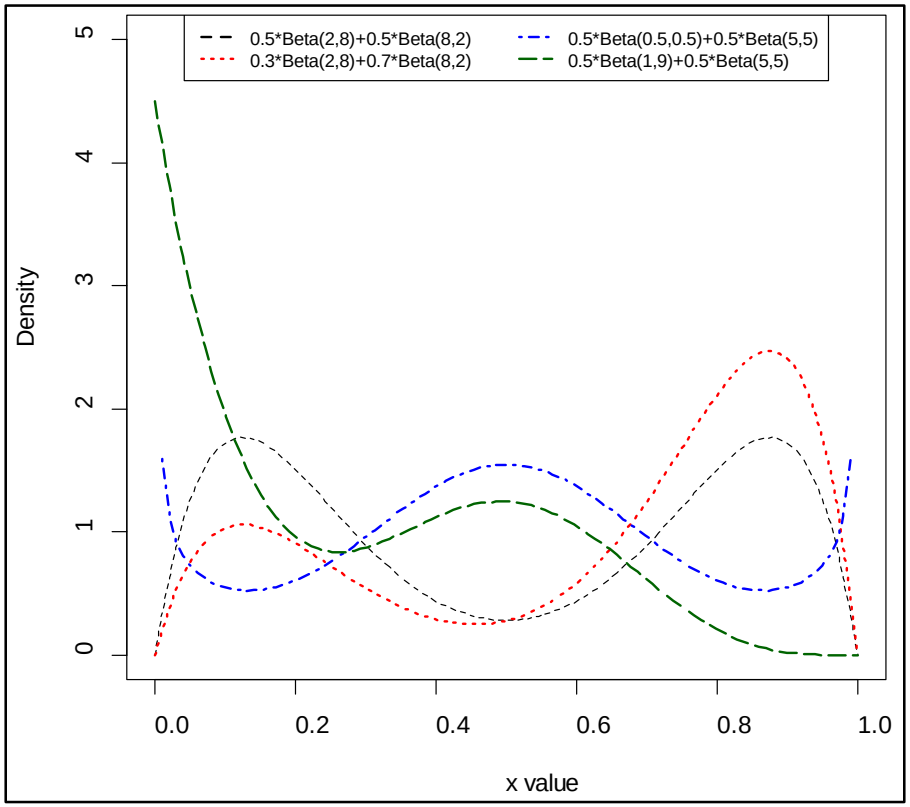
- Sample sizes examined were 2, 3, 4, 5, 10, 12, 15, 20, 25, and 30
- For each alternative hypothesis, 10,000 datasets of the given sample size were randomly generated based on the true distribution
- For methods without closed-form critical value tables, 1,000 replicates were used to determine the *p*-value
- Target type I error rates (α) were 5%, 10%, and 20%

Power Study Alternative Hypotheses

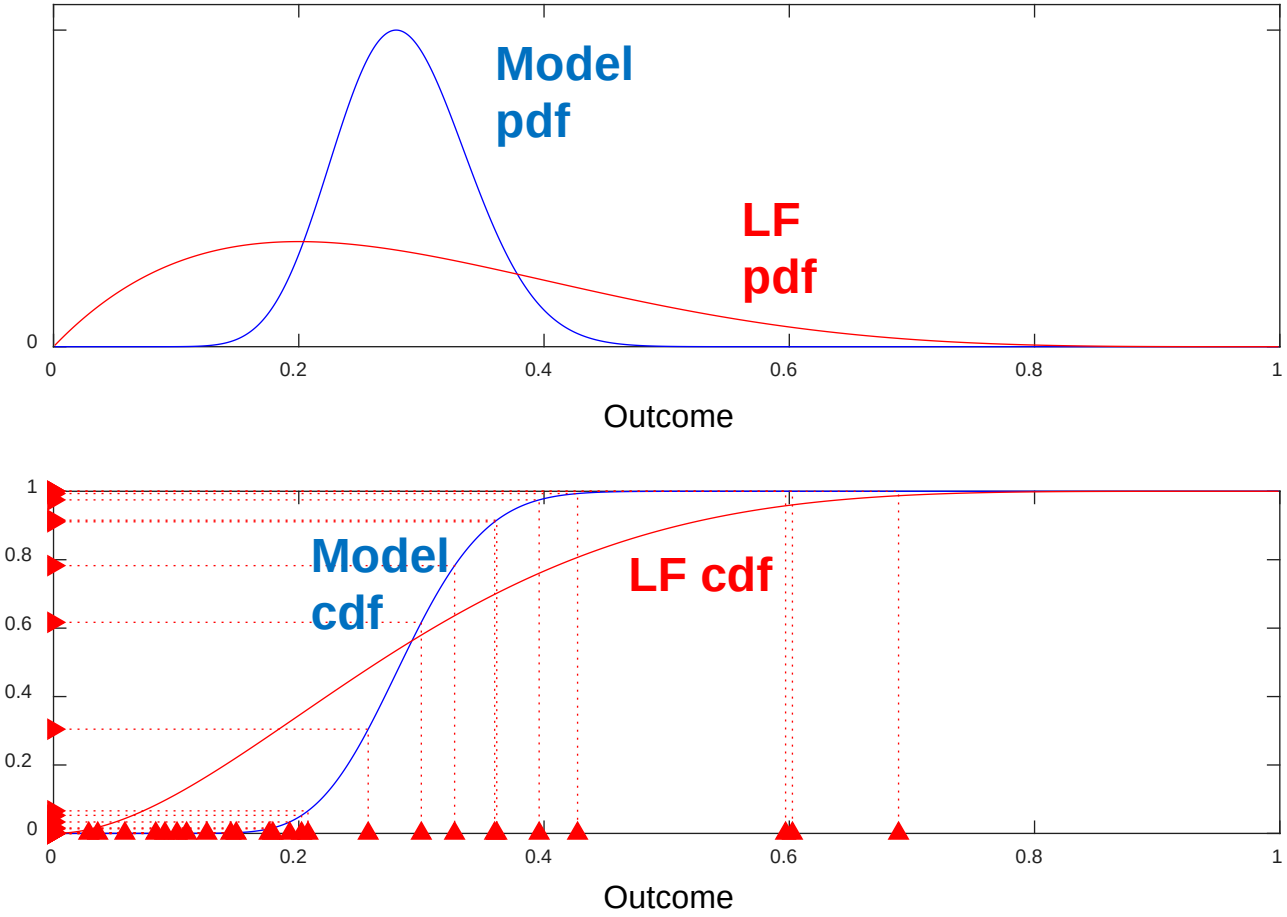
Case 1: Beta(0.5, 0.5), data clumped in the tails
Case 2: Beta(2, 2) and Beta(5, 5), data clumped in the middle
Case 3: Beta(2, 5) and Beta(2, 8), data clumped at one extreme
Case 4: Beta(2, 1), linearly increasing trend
Case 5: Beta(5, 1), exponentially increasing trend

Case 6: mixture distributions, representing multimodal data



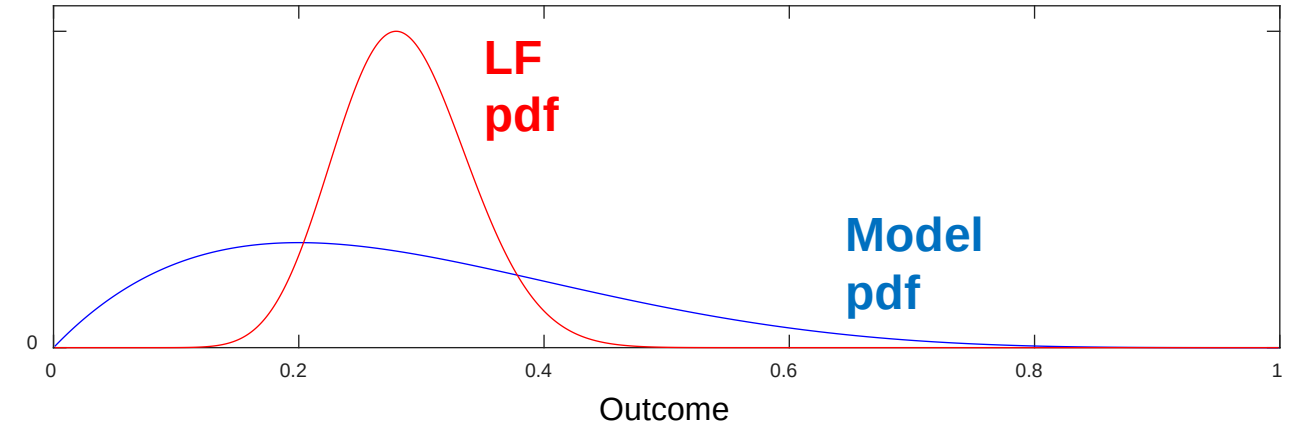


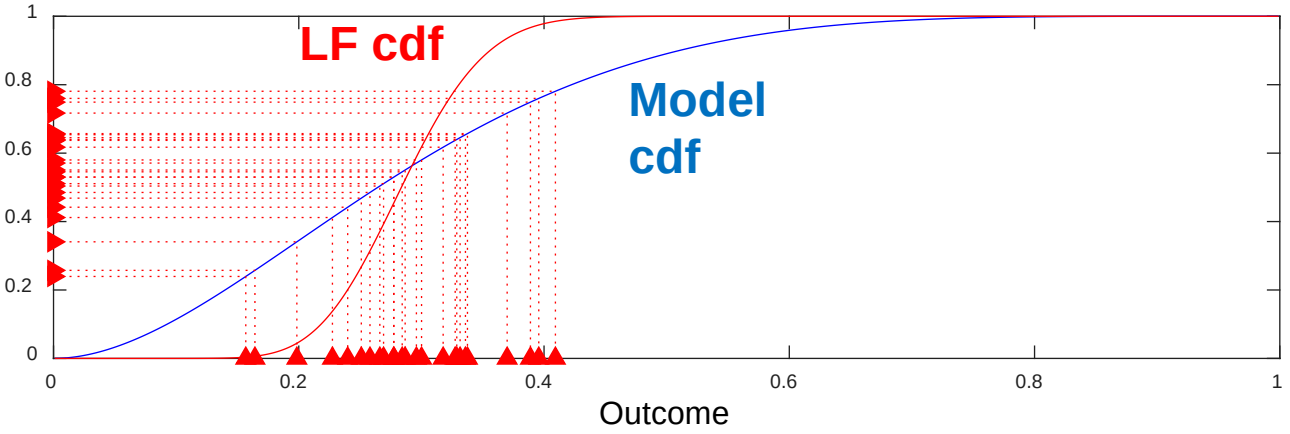
Case 1: Although the means are the same, the Model has less variability than LF



Evidence: p-values tend toward the extremes of zero and one, away from the middle

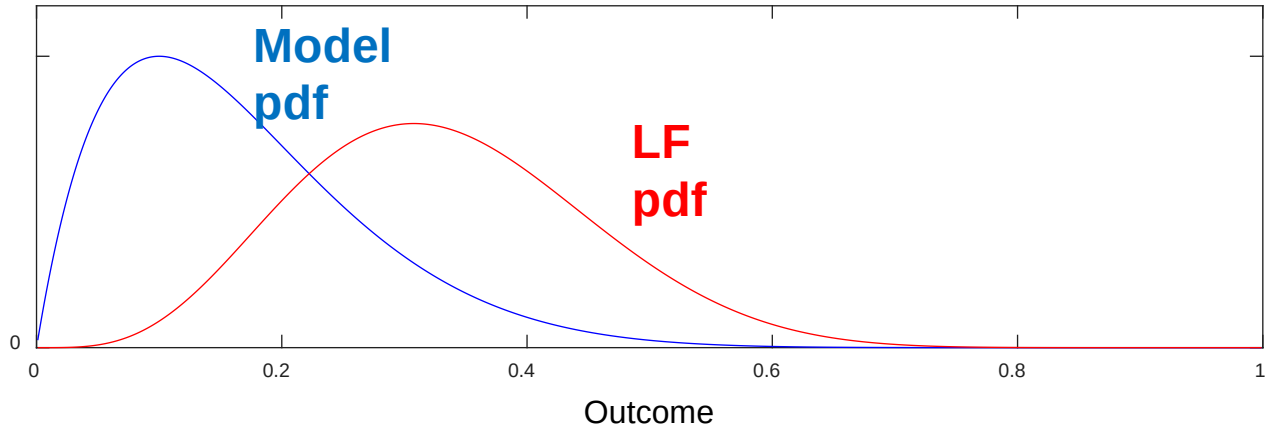
Case 2: Although the means are the same, the Model has more variability than LF

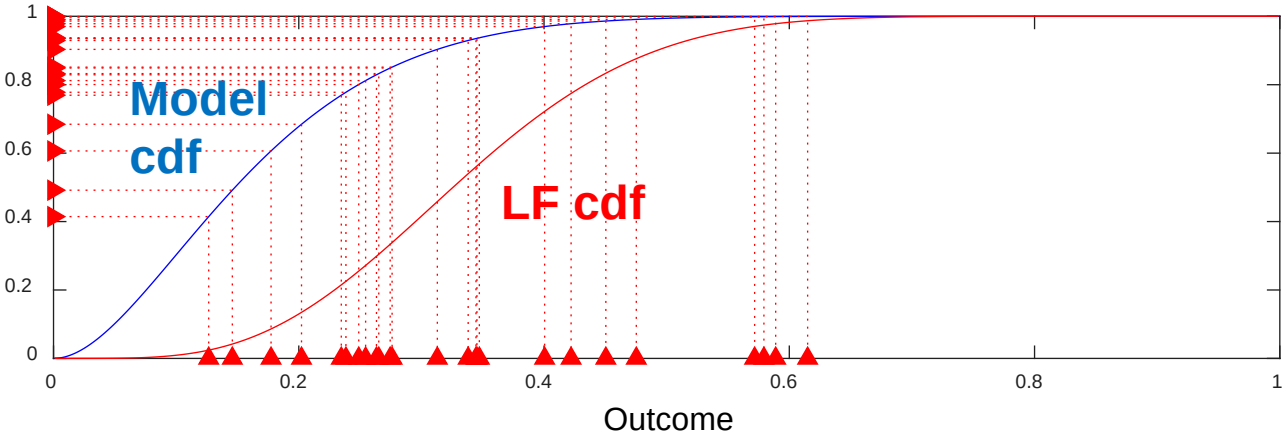




Evidence: p-values tend toward the middle, away from the extremes of zero and one

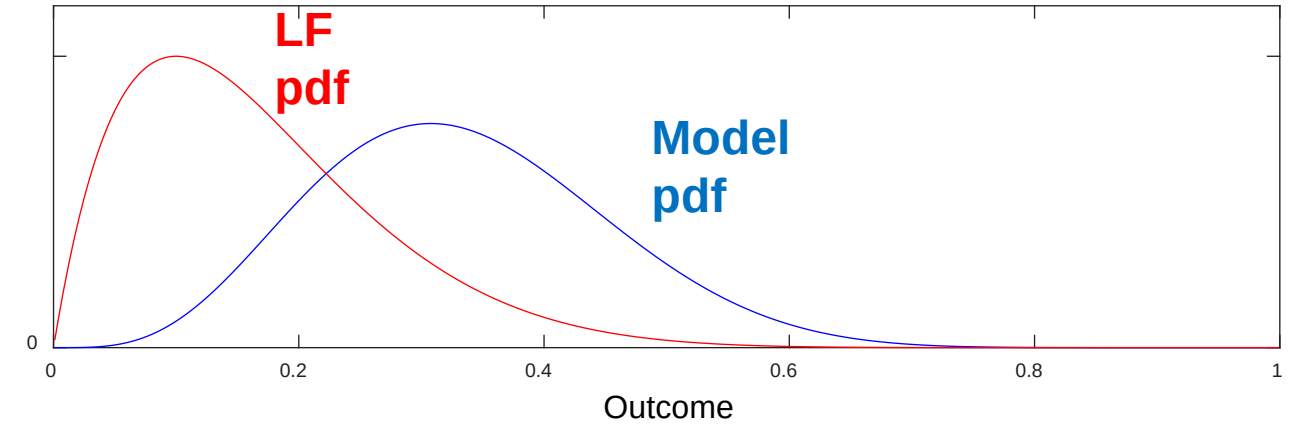
Cases 3, 4, and 5: The Model tends to underpredict the Results

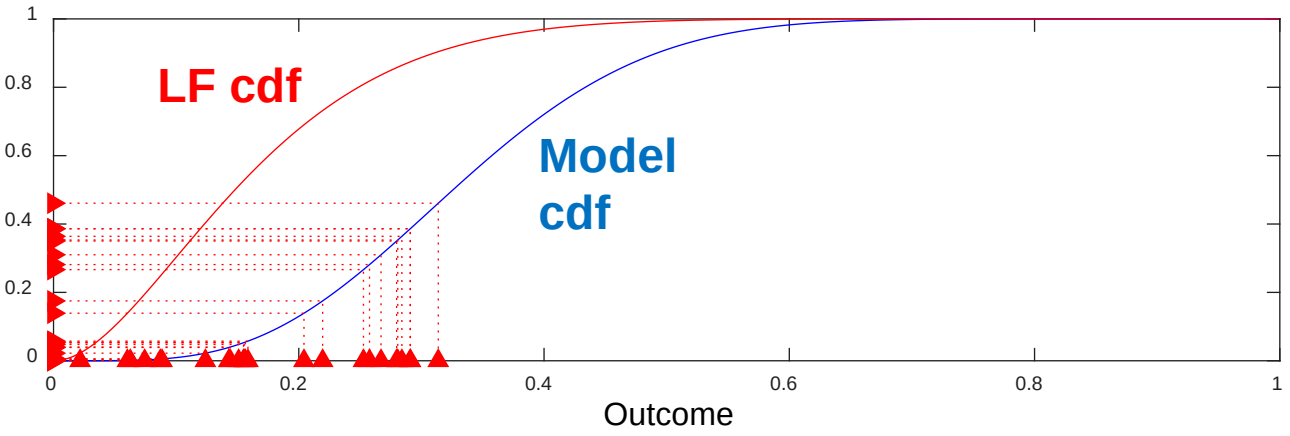




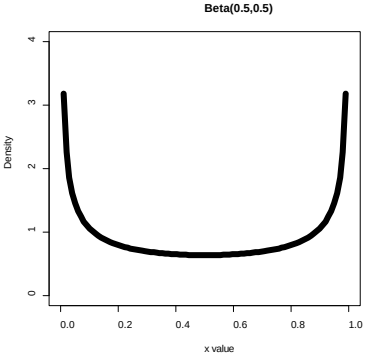
Evidence: p-values tend toward one, away from zero

Cases 3, 4 and 5: The model tends to overpredict the Results



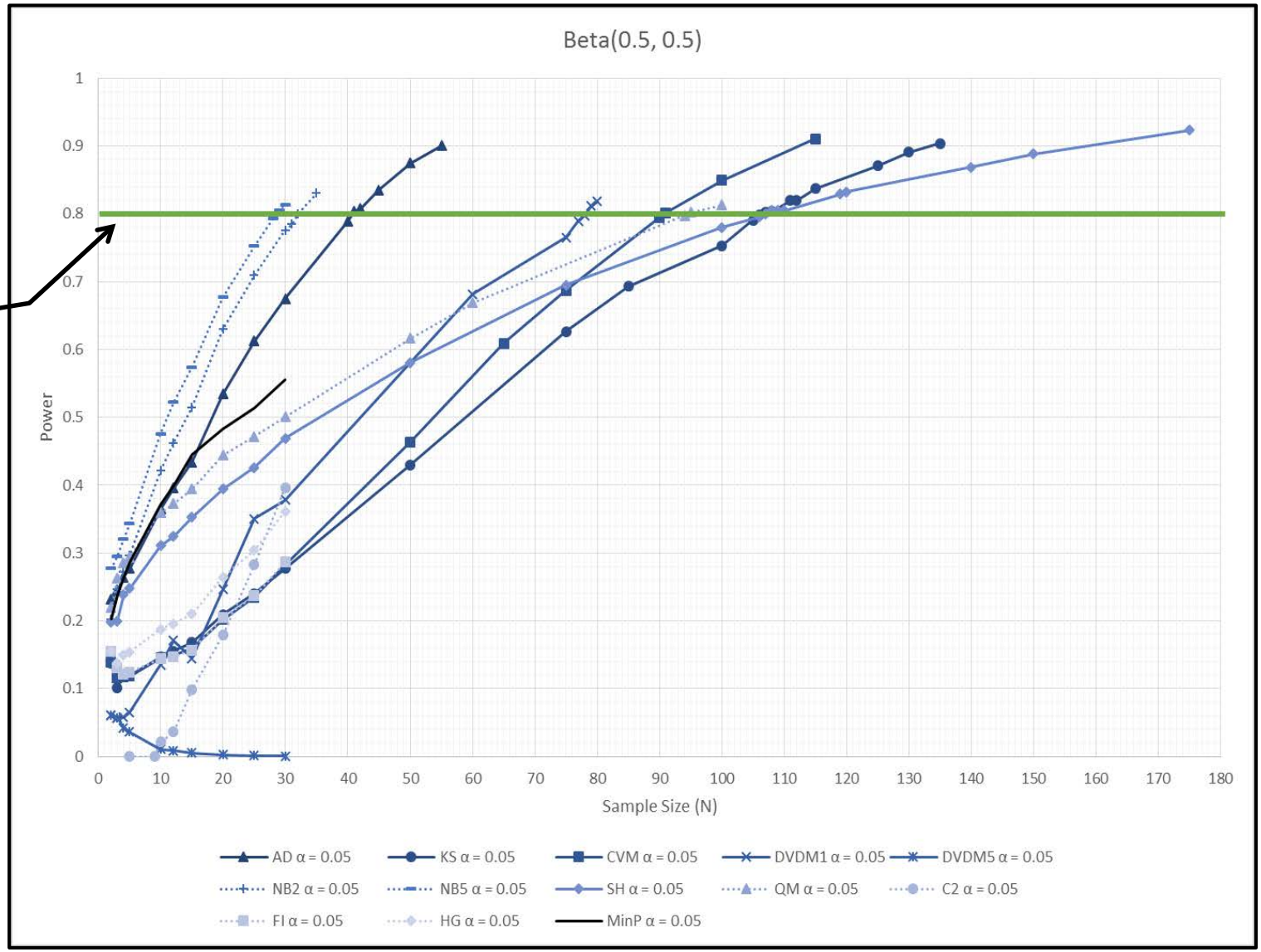


Evidence: p-values tend toward zero, away from one



Power Study Results: Case 1, α = 5%

80% power



Sample Size Needed to Achieve 80% Power @ 5% alpha

True Distribution	AD	KS	CVM	DVDM5	DVDM1	NB5	NB2	QM	MeanZ*
Case 1	41	107	91		79	29	32	95	
Case 2	84	136	109	28	44	63	43	115	
Case 2	23	30	25	8	15	17	11	24	
Case 3	11	10	10	9	15	16	12	15	5
Case 3	6	5	5	6	8	9	7	8	2
Case 4	23	26	22	31	34	37	26	49	
Case 5	5	5	5	7	9	8	6	7	
Case 6a	109	107	111		45	45	106	73	
Case 6a	37	31	35	135	30	29	39	44	
Case 6b	77	85	83	64	82	40	59	57	
Case 6b	29	30	31	22	26	31	30	32	20

Smallest Sample Size(s)

Not Recommended

* Combined probability methods only shown when at least as good as GOF test

GOF Method Comparison

True Distribution	Recommended Method
Case 1: Peak at both extremes (Bathtub curve)	NB, AD
Case 2: Peak in the middle (symmetric)	DVDM, NB2
Case 3: Skewed right (or left)	Mean Z
Case 4: Linearly increasing (or decreasing)	AD, CVM
Case 5: Exponential increasing (or decreasing)	Any
Case 6a: Bimodal	QM, NB5, DVDM1, KS
Case 6b: Peak at extremes and middle	Mean Z

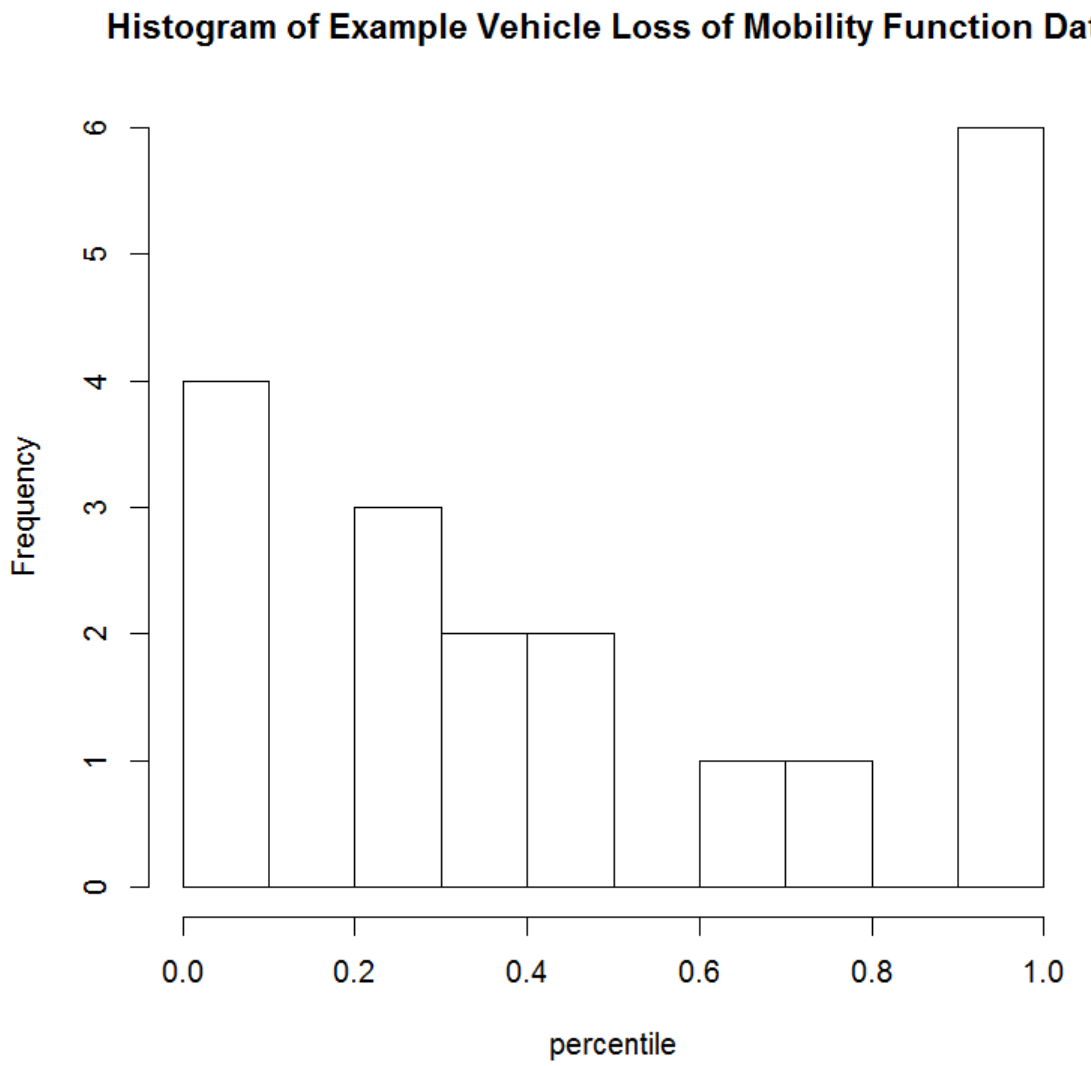
Example

- Suppose from 19 LF Results and associated Models Outputs, we observe the following percentiles:

- 0.01, 0.02, 0.06, 0.07, 0.27, 0.30, 0.30, 0.35, 0.39, 0.43, 0.48, 0.66, 0.71, 0.93, 0.93, 0.96, 0.99, 0.99, 0.99

- Do these pass the test for U(0,1)?

Example Data



Uniform?

Method	P-value[1]
AD	0.076
C2	0.810
CVM	0.344
KS[2]	0.201
NB5	0.003
NB2	0.041
DVDM5	0.990
DVDM1	<0.001
QM	0.041

Method	P-value
FCP	0.217
Logit	0.768
Max P	0.826
Min P	0.174
Mean P	0.606
Mean Z	0.6618
Sum P	0.606
Sum Z	0.731

[1]Based on 100,000 MC for determining p-value for methods without closed form critical value tables
[2]Ties in the data affect the method

Conclusions

- No one method is “best” for all situations
  - C2 GOF test suffers when sample size is small
  - NB, DVDM, and CVM GOF tests show strong power against multiple alternatives
  - MeanZ has stronger power than the GOF methods in some cases

- In practice:
  - If available, use information from similar models to understand the historical trends (ex. models tend to be less variable than LF)
  - Use the sample size tables to help plan testing
  - Graph current data

This Work Directed and Funded By:

Methodology Division, ISMED
Army Evaluation Center
APG, MD 21005
Contract Number: W91CRB-14-D-0028-003

Contact Information

Shannon Shelburne
sshelburne@jrad.us
540.288.3132 x213