SLIDES ONLY

NO SCRIPT PROVIDED

INSTITUTE FOR DEFENSE ANALYSES

Improving Test Efficiency: A Bayesian Assurance Case Study

Rebecca M. Medlin, Project Leader

August 2023

IDA Document NS 3000024

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, "CRP Statistics WorkGroup." 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 Keyla Pagan-Rivera, Dr. John T. Haman, and Mr. Addison D. Adams 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

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS 3000024

Improving Test Efficiency:
A Bayesian Assurance Case Study

Rebecca M. Medlin, Project Leader

Executive Summary

Research staff members from the Institute for Defense Analyses (IDA) were invited to present this talk on Bayesian Assurance Testing at the August 2023 Joint Statistical Meetings, the largest gathering of statisticians and data scientists held in North America. The talk is part of a session sponsored by the American Statistical Association's Section on Statistics in Defense and National Security.

The Department of Defense develops and acquires some of the world's most advanced, sophisticated, and expensive systems. As new technologies emerge and are incorporated into systems, the test and evaluation community faces the challenge of ensuring that these systems undergo adequate and efficient testing prior to operational use. In this presentation, we highlight how the use of Bayesian methods can help the test and evaluation community take steps toward increasing test efficiency; specifically in the area of planning for the test and evaluation of a system's reliability.

In this presentation, we address two questions. First, how should we plan a test to evaluate system reliability? Second, in planning, can we improve test efficiency?

The first question seeks to address the best approach or methodology for planning a test that specifically targets the evaluation of system reliability. It prompts consideration of the key factors, techniques, and considerations involved in designing a test plan for assessing reliability. The follow-on question explores the possibility of improving test efficiency during the planning phase. It invites discussion on strategies, methodologies, or approaches that can be employed to enhance the efficiency of the testing process without compromising the reliability of the results.

Understanding system reliability is an important component in evaluating suitability. A challenge that we face in test planning is scoping an operational test that is long enough (miles, flight hours, cycles, etc.) to adequately evaluate system reliability. The length of a test is informed by the reliability requirement, the true reliability of the system, and the risk we are willing to accept in making an incorrect statement about the system's reliability.

Traditionally, we rely on reliability demonstration tests, which use only the data from the test to assess whether the reliability-related quantity of interest meets or exceeds the requirement. Such tests often require lengthy testing periods, which we can't necessarily accommodate due to things like time and resources.

To address the dilemma of lengthy testing periods, the presentation recommends considering reliability assurance tests. These tests utilize supplementary data and information, including reliability models, prior test results, expert judgment, and knowledge of environmental conditions, to reduce the required amount of testing. Bayesian test plans use supplementary data and information; hence, we refer to them as reliability assurance tests.

To improve test planning, we propose the use of Bayesian methods to incorporate supplementary data and reduce testing duration. Furthermore, we recommend Bayesian methods be employed in the analysis phase to better quantify uncertainty. We find that when using Bayesian methods for test planning, we can scope smaller tests; and using Bayesian methods in analysis results in a more precise estimate of reliability, improving uncertainty quantification.

Improving Test Efficiency:
A Bayesian Assurance Case Study

Rebecca Medlin

Joint Statistical Meetings | Toronto, Canada |August 2023

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

How should we plan a test to evaluate system reliability?

And, in planning, can we increase test efficiency?

Bottom Line Up Front

Understanding system reliability is an important component in evaluating suitability (reliability, availability, maintainability).

A challenge that we face in test planning is scoping a test that is long enough (miles, flight hours, cycles, etc.) to adequately evaluate system reliability.

- Traditionally, we plan a reliability demonstration test, which uses only the data from the test to evaluate system reliability. This approach often requires lengthy testing periods.
- The test length is informed by the reliability requirement, the true reliability of the system, and the risk we are willing to accept in making an incorrect statement about the system's reliability.

In response to this dilemma, we should also consider using Bayesian methods.

- In test planning, we should consider Bayesian methods to incorporate supplementary data and information to reduce the required testing duration.
- In analysis, we should consider Bayesian methods to incorporate supplementary data and information to better quantify uncertainty.

When might an Assurance Testing approach work the best?

- Longer previous test periods (e.g., a production verification testing)
- Minimal system changes expected between testing (e.g., the past test and the test we are planning for)
- Testing adequate for evaluating effectiveness is not long enough to support reliability demonstration test (e.g., only need 20 missions to evaluate effectiveness, but that doesn't equate to enough test hours for reliability)

Outline

- Terminology
- Traditional Planning Approach
- Bayesian Planning Approach
- Example
- Thoughts on Selecting a Planning Approach
- Future Work

Terminology

Reliability – Probability that the system performs its intended function under operating conditions, for a specified period of time.

Reliability Requirement – Level of reliability, determined necessary by the user, that the system is expected to achieve. The mean time between failure (MTBF) is the most common metric used to define reliability requirements in test and evaluation.

True Reliability of the System – True but unknown reliability of the system.

Consumer Risk (probability of false positive) – Concluding the system meets the requirement when the true reliability is below the requirement.

Producer Risk (probability of false negative) – Concluding the system fails the requirement when the true reliability is above the requirement.

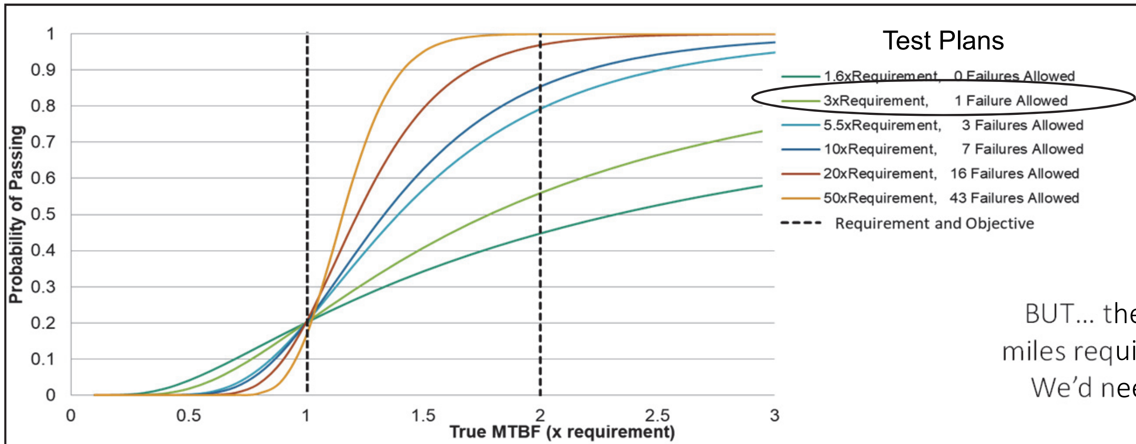
Reliability Test Plan (T,c) – A test plan specifies a time (T) and a maximum number of failures (c) permitted for a successful test.

A Traditional Planning Approach – Reliability Demonstration Test

A traditional reliability demonstration test is a classical hypothesis test, which uses only the data from the test to assess whether the reliability quantity of interest (e.g. MTBF) meets or exceeds the requirement.

Operating Characteristics Curves are a useful tool for determining and comparing Test Plans (T,c).

- Often sized with only consumer risk in mind – we do not want to accept a system with reliability less than the requirement



Requirement: 1,000 miles MTBF

Selected Test Plan: 3,000 miles, 1 failure

One-sided Confidence Interval: (1,002, -)

BUT… the probability of actually observing 1 failure or fewer in 3,000 miles requires a true reliability of at least 3x the requirement (not likely). We'd need about 20,000 miles to have a high probability of passing.

MTBF: Mean time between failure

How should we plan a test to evaluate system reliability? And, in planning, can we increase test efficiency?

There is actually nothing wrong with using a demonstration test to plan, and scoping a test this way might actually work best for many situations.

But…

What if the test length specified isn't possible?

What if the true reliability required isn't possible?

What if we have additional information, like prior knowledge of the demonstrated system reliability?

Perhaps, we are most interested in assuring that a certain level of reliability or quality is achieved and maintained.

Bayesian Planning Approach – Reliability Assurance Test

A reliability assurance test is one that uses additional supplementary data and information to reduce the required amount of testing.

Additional data and information may include appropriate reliability models, earlier test results on the same or similar systems, expert judgment regarding performance, knowledge of the environmental conditions under which the systems are used, benchmark design information on similar systems, prior knowledge of possible failures, etc.

Bayesian test plans use supplementary data and information, hence we refer to them as reliability assurance tests.

Before we solve for a Bayesian Test Plan, let's talk about the Risk Criteria. There is a difference…

Traditional Risk Criteria

Traditional Consumer Risk

Probability of passing the test when $\lambda = \lambda_1$

$$= P\left(Test\ is\ Passed \middle| \lambda = 1/MTBF_{Req}\right)$$

$$= P(y \le c|\lambda) = \sum_{y=0}^{c} \frac{\lambda^y e^{-\lambda}}{y!} \le \alpha$$

We choose c to be the largest non-negative integer that satisfies this inequality.

Bayesian Posterior Risk Criteria

Posterior Consumer Risk: If the test is passed then the consumer desires a maximum probability α that $\lambda \ge \lambda_1$.

$$= P(\lambda \ge \lambda_1 | Test\ is\ Passed, \boldsymbol{x})$$

$$\approx \frac{\sum_{j=1}^{N}\left[\sum_{y=0}^{c} \frac{\left(\lambda^{(j)}T\right)^y \exp\left(-\lambda^{(j)}T\right)}{y!}\right] I\left(\lambda^{(j)} \ge \lambda_1\right)}{\sum_{j=1}^{N}\left[\sum_{y=0}^{c} \frac{(\lambda^{(j)}T)^y \exp(-\lambda^{(j)}T)}{y!}\right]} \le \alpha$$

Posterior Producer Risk: If the test is failed, then the producer desires a maximum probability β *that* $\lambda \le \lambda_0$.

$$= P(\lambda \le \lambda_0 | Test\ is\ Failed, \boldsymbol{x})$$

$$\approx \frac{\sum_{j=1}^{N}\left[1 - \sum_{y=0}^{c} \frac{\left(\lambda^{(j)}T\right)^y \exp\left(-\lambda^{(j)}T\right)}{y!}\right] I\left(\lambda^{(j)} \le \lambda_0\right)}{\sum_{j=1}^{N}\left[1 - \sum_{y=0}^{c} \frac{(\lambda^{(j)}T)^y \exp(-\lambda^{(j)}T)}{y!}\right]} \le \beta$$

Where x is available data, $\lambda^{(j)}$ are the posterior predictive draws, and $\lambda_0 < \lambda_1$

Reference: Michael S. Hamada et al., *Bayesian Reliability,* 2008, Chapter 10.

Note: The expression *Test is Failed* means that the number of observed failures is larger than the maximum number of allowed failures. That is, $y > c$. Similarly, *Test is Passed* means that $y \le c$.

Steps to generate a Bayesian Assurance Test Plan

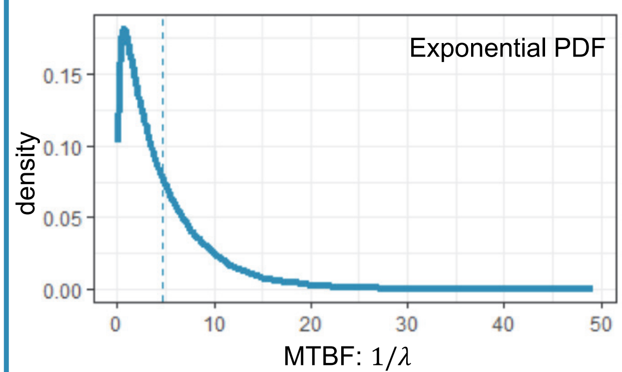
Likelihood: $P(h|\lambda) \sim$ Exponential $(\lambda = F/H)$
Conjugate Prior: $P(\lambda) \sim$ Gamma (α, β)
Posterior: $P(\lambda|h) \sim$ Gamma $(\alpha` = \alpha + F, \beta` = \beta + H)$

Step 1

Observed Test Data*

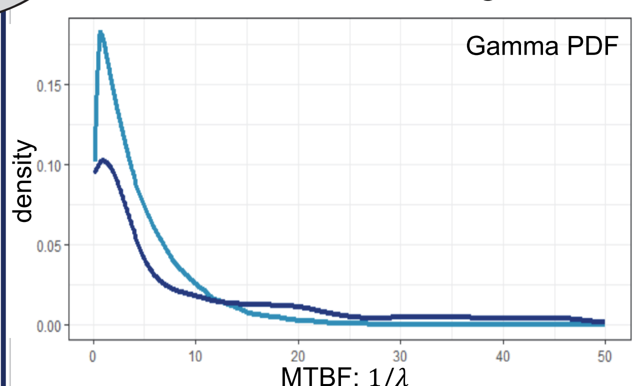
Hours: 300
Failures: 65
MTBF: 4.6

Supplementary Data: Past Test



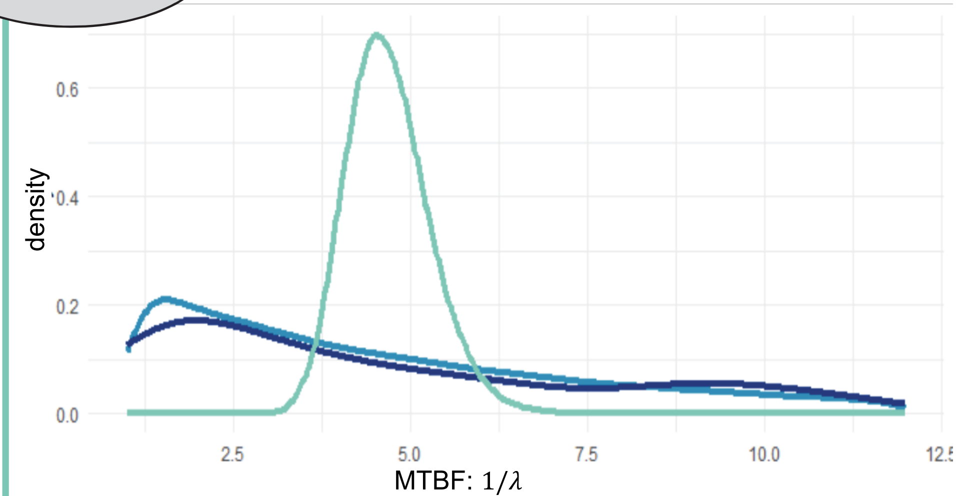
Step 2

Prior: Non-Informative gamma



Step 3

Posterior of MTBF (1/ $P(\lambda|h)$)



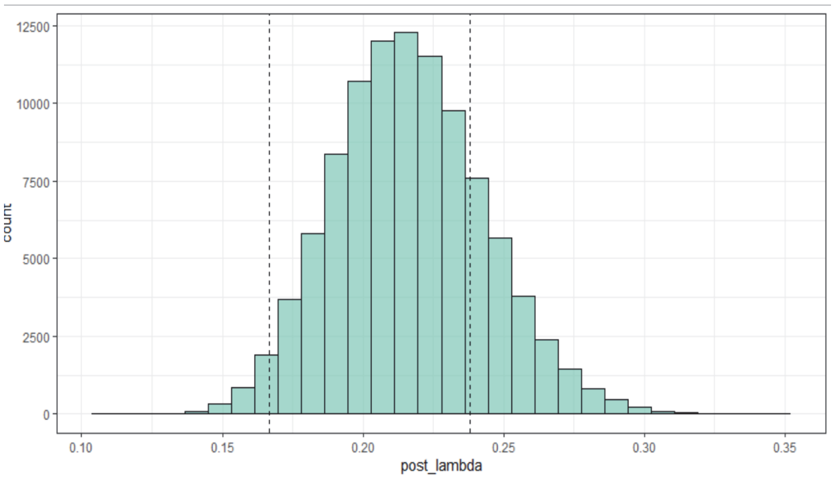
- The posterior is our belief statement about the true value of MTBF
- We use the **posterior** to solve for a test plan (T,c)

* All data are notional
MTBF: Mean time between failure

Steps to generate a Bayesian Assurance Test Plan (cont.)

Step 4

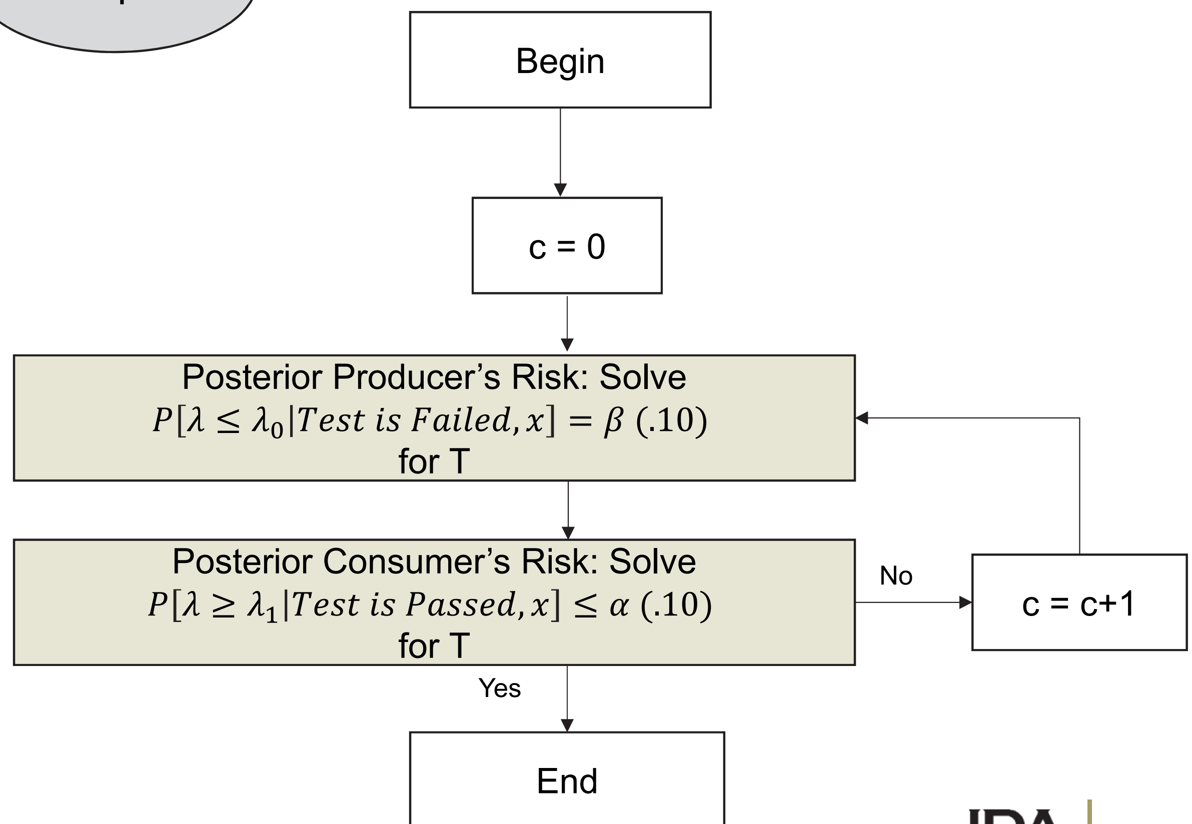
Select acceptable and rejectable values
for producer and consumer risk



$$\lambda_0 = \frac{1}{6} = 0.16 \qquad \lambda_1 = \frac{1}{4.2} = 0.24$$

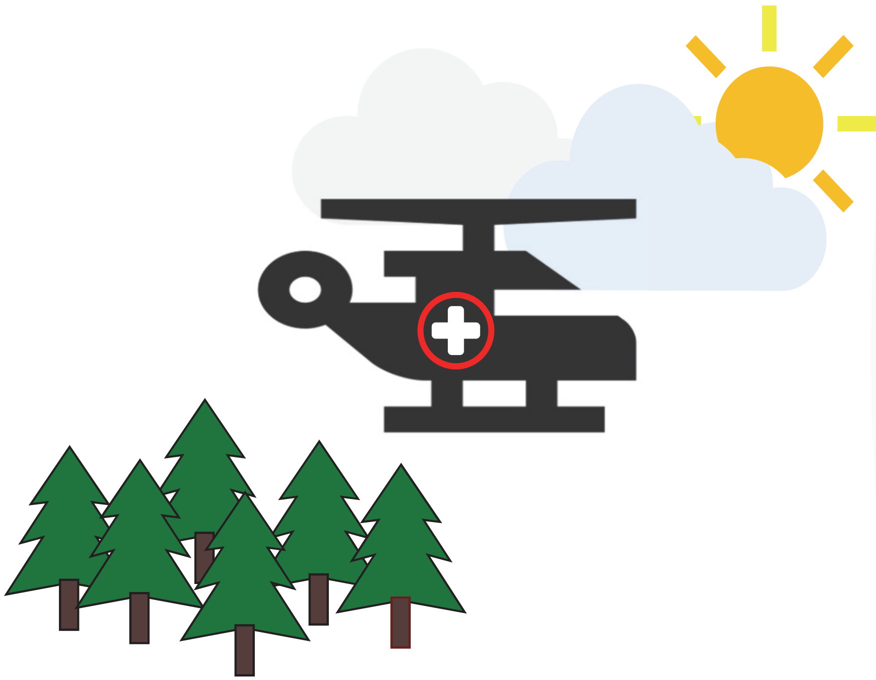
Step 5

Bayesian Test Plan Algorithm to solve for T and c



MTBF: Mean time between failure

Example: Testing a new variant of a helicopter



Testing a new variant of a helicopter reliability requirement: 4.2 hours MTBF

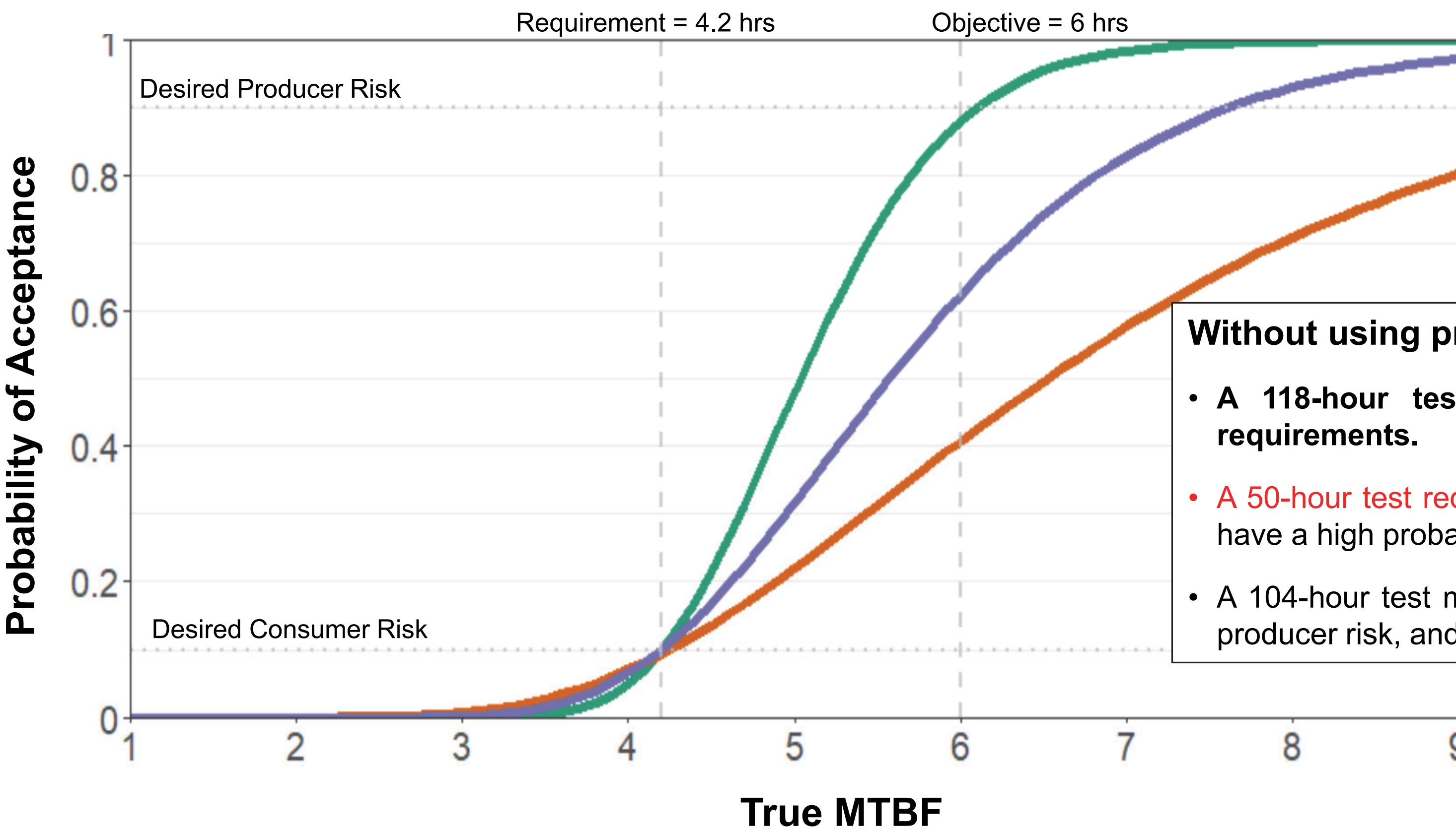
- Prior testing on base variant includes 300 hours and 65 failures*
- Minimal system changes between 2 variants
- Only 20 missions needed for evaluating effectiveness, including training amounts to about 50 hours for test.

Is 50 hours sufficient for assuring system reliability?

* All data are notional

MTBF: Mean time between failure

Test Length Scoped Using Operating Characteristic Curves



Test Length	Permitted Failures
50 hrs	7
104 hrs	18
240 hrs	47

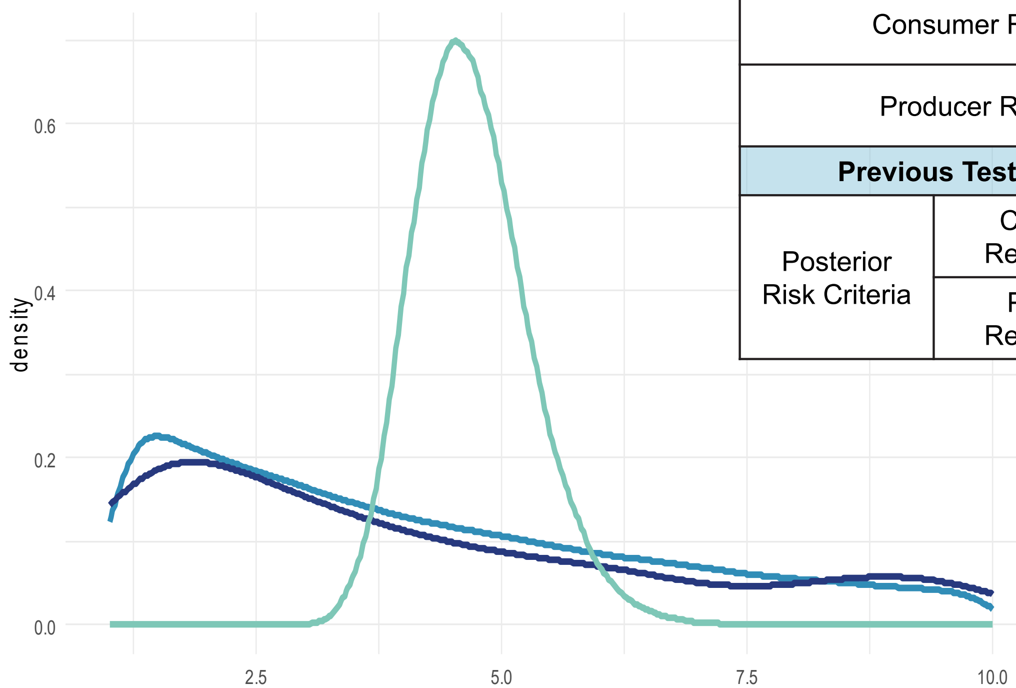
Without using prior data…

- **A 118-hour test is the shortest satisfying all risk requirements.**
- A 50-hour test requires a true MTBF of more than 9 hrs. to have a high probability of demonstrating the requirement.
- A 104-hour test meets consumer risk, comes very close to producer risk, and requires a true MTBF of 7.5.

MTBF: Mean time between failure

We can use a Bayesian Assurance Test to Leverage Previous Data For Test Planning And Analysis

Criteria and Assumptions			
Consumer Risk		0.1	Probability of incorrectly concluding requirement is met
Producer Risk		0.1	Probability of incorrectly concluding requirement is not met
Previous Test Data*		**4.61 hrs.**	**65 failures in 300 hrs. from base variant**
Posterior Risk Criteria	Consumer Requirement	4.2 hrs.	Minimal acceptable reliability
	Producer Requirement	6 hrs.	Desired reliability



Likelihood is Exponential ($\lambda = \frac{failures=65}{hours=300}$).
Gamma failure rate prior (α =.001, β =.001); no degradation in reliability.
Posterior failure rate for determining test length is Gamma(α =.001 + $failures$, β=.001+$Hours$).

* All data are notional
MTBF: Mean time between failure

Leveraging Prior Data, A 49-Hour Test With 18 Allowable Maintenance Actions Suffices For Reliability Assessment

Using Prior Data* (65 failures, 300 hours)…

- A 50-hour test will yield an MTBF estimate that provides a sufficient determination of reliability.
- 8 or fewer failures will constitute meeting the requirement.
- Resulting producer risk: 0.01
- Resulting consumer risk: 0.09

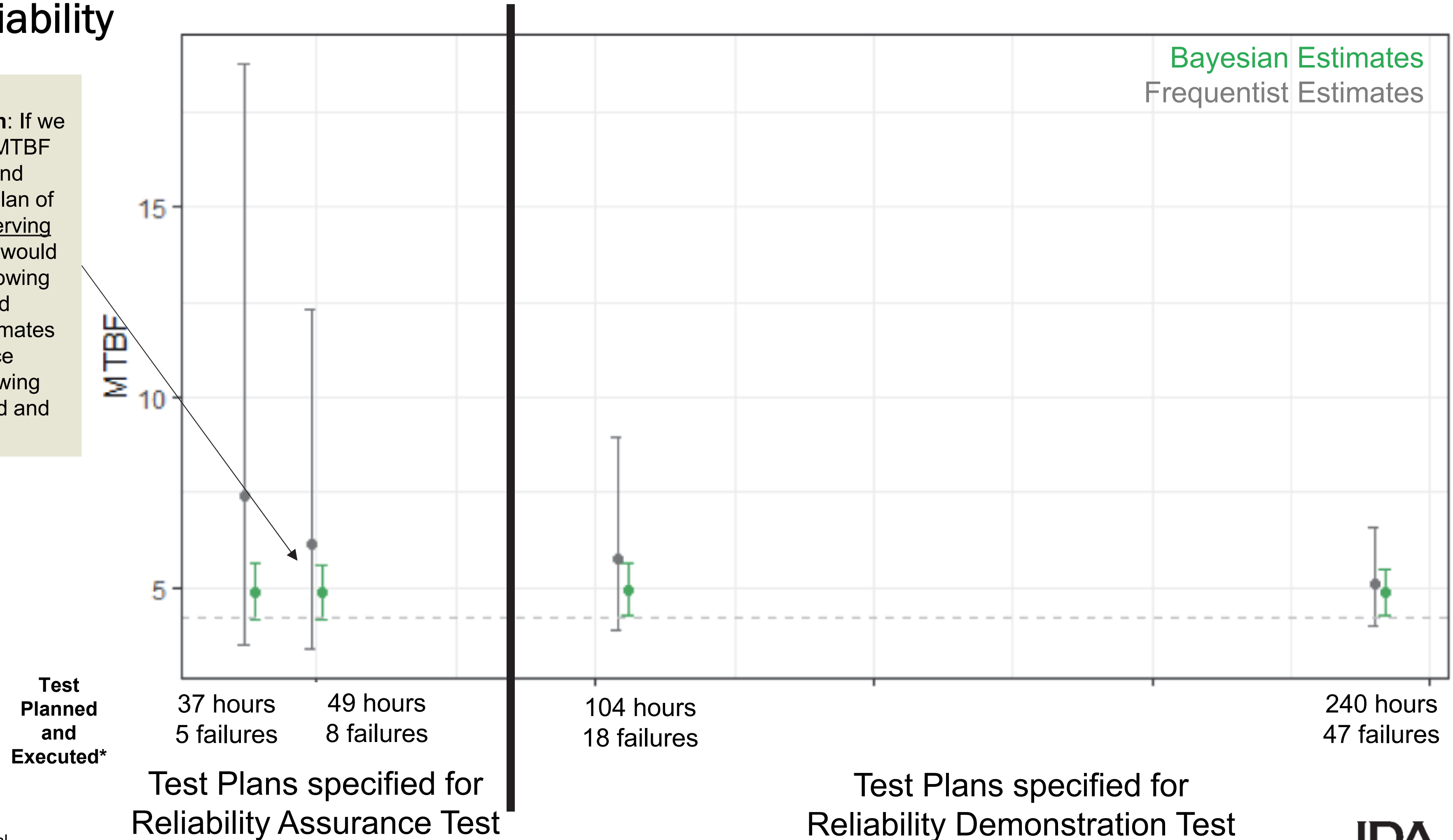
Test Plans Meeting Criteria	
Test Duration	Permitted Failures
37	5
41	6
45	7
49	8
53	9
57	10
61	11
69	12
.	.
115	25

* All data are notional

TBF: Mean time between failure

Using Bayesian Methods For Analysis Supports a More Precise Estimate Of Reliability

Chart Interpretation: If we observed an MTBF of 4.6 hours and select a test plan of 49 hours observing 8 failures, we would report the following frequentist and Bayesian estimates and confidence intervals (showing both two-sided and one-sided).



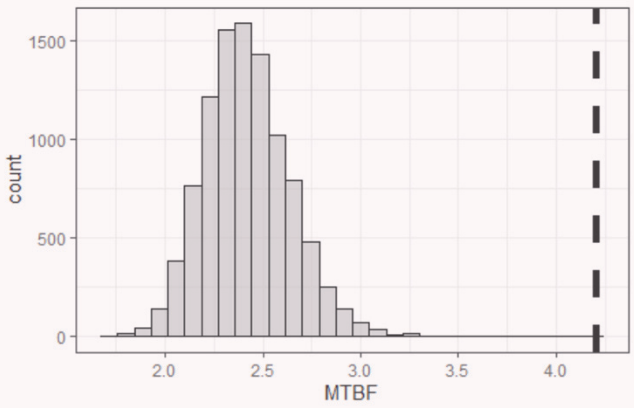
* All data are notional

MTBF: Mean time between failure

Three cases we have to consider when planning for a Bayesian Assurance Test

Reliability Requirement: MTBF 4.2*

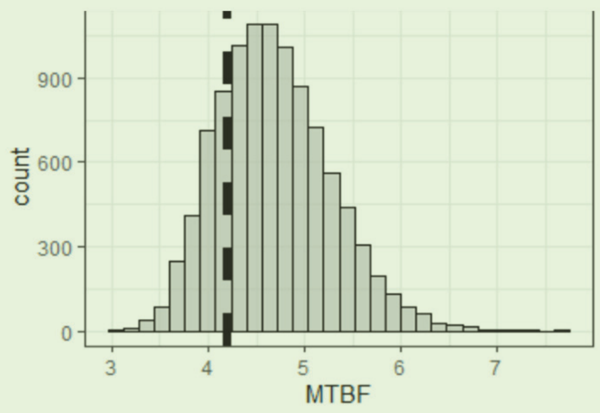
Case 1: Previous Test Suggests
Poor Reliability



Math Outcome:
Lots of Testing

Nothing really gained by leveraging the previous test data alone, we must think about the prior and other sources of information.

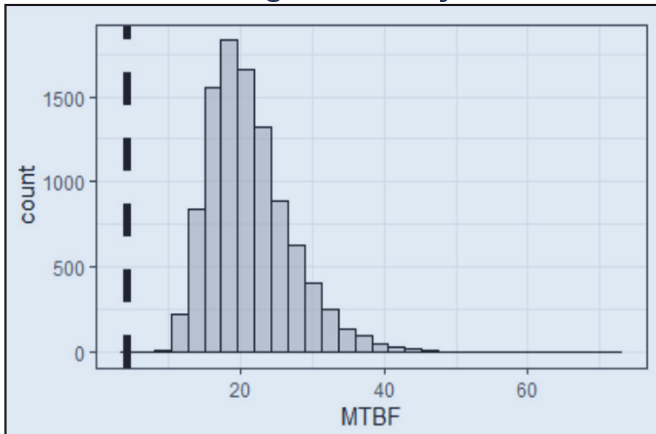
Case 2: Previous Test Suggests
"Good" Reliability



Math Outcome:
Reasonable Test Length

Leverage previous data and test to assure

Case 3: Previous Test Suggests
High Reliability



Math Outcome:
No Testing

We still want to test. In this case, a traditional approach, assuming a much higher true MTBF than the requirement will yield a reasonable test length answer.

* All data are notional
MTBF: Mean time between failure

Ideas for Future Work

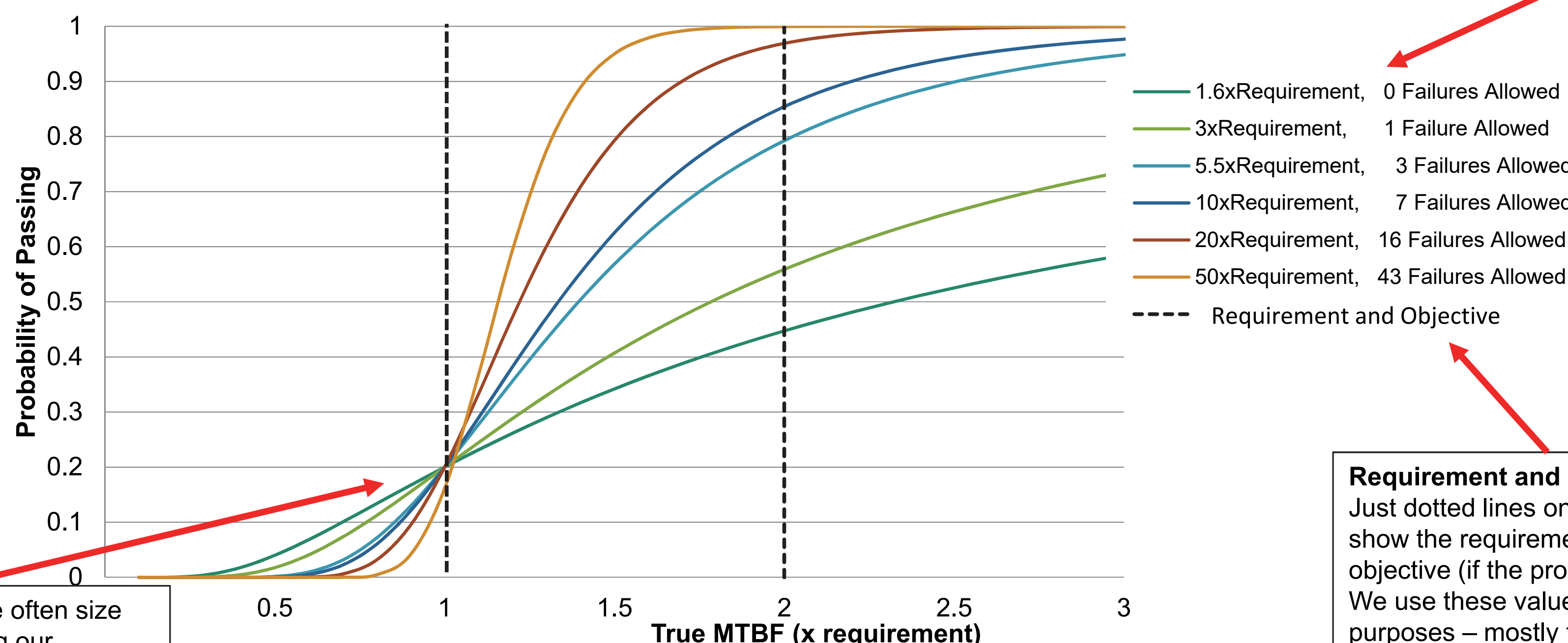
- Develop an application that can be shared with the test community
- Compile case studies examples and develop “guidance” / training to help educate the test community

Back Up Slides

The Operating Characteristic Curve

Probability of Passing – For a given test length curve, what is the probability that you see c or fewer failures (and therefore pass the test)? We need true reliability of the system to be high – especially for a short test – in order to have a high probability of seeing c or fewer failures. We think of Probability of Acceptance like power and aim for 80% when possible.

Test Length for give number of c failures. Basically, if you completed the x-hour test and saw c or fewer failures you would demonstrate your requirement with a lower confidence bound ≥ to requirement



Requirement and Objective Just dotted lines on the plot to show the requirement and objective (if the program has one). We use these values for planning purposes – mostly to set the consumer risk.

Consumer Risk - We often size the test by associating our requirement with our consumer risk of 20%. If a system enters test with a True Reliability equal to the Requirement – we only have a 20% probability of acceptance. We don't want to accept a bad system.

TRUE MTBF – is the true but unknown reliability of the system. We need the system to enter test with a high reliability (designed into the system) in order to have a high probability of acceptance. A lot of time programs will have a threshold and objective requirement – we want the contractor to design reliability into the system that is higher than the requirement.

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

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
XX-08-2023	Final	xxxx xxxx-Aug 2023

4. TITLE AND SUBTITLE
Improving Test Efficiency: A Bayesian Assurance Case Study

5a. CONTRACT NUMBER
Seperate Contract

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Medlin, Rebecca, M

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
3000024

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

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

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

12. DISTRIBUTION/AVAILABILITY STATEMENT
This OED Draft 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: Rebecca M. Medlin

14. ABSTRACT
To improve test planning for evaluating system reliability, we propose the use of Bayesian methods to incorporate supplementary data and reduce testing duration. Furthermore, we recommend Bayesian methods be employed in the analysis phase to better quantify uncertainty. We find that when using Bayesian Methods for test planning we can scope smaller tests and using Bayesian methods in analysis results in a more precise estimate of reliability – improving uncertainty quantification.

15. SUBJECT TERMS
Bayesian analysis
Reliability Analysis
Test Planning

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	26	Rebecca M. Medlin
					19b. TELEPHONE NUMBER *(Include area code)* 703-845-6731