INSTITUTE FOR DEFENSE ANALYSES

Estimating System Reliability from Heterogeneous Data

Caleb Browning
Laura Freeman
Alyson Wilson
Kassandra Fronczyk
Rebecca Dickinson

June 2015

Approved for public release;
distribution is unlimited.

IDA Document D-5515
Log: H 15-000546

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication

The reliability of complex systems is often best measured by multiple task-level reliabilities. In this paper we describe the challenge of estimating the reliability of a complex system using the information from task-level reliability measures. We propose an experimental design approach for assessing the impact of different usage profiles on overall system reliability. We consider appropriate models for accounting for correlation between heterogeneous task-level reliabilities and for integrating data from multiple tests.

Copyright Notice

Estimating System Reliability from Heterogeneous Data

Caleb Browning, Laura Freeman, Alyson Wilson*,
Kassandra Fronczyk, Rebecca Dickinson

Mathematical Methods in Reliability
Tokyo, Japan
June 2, 2015

alyson_wilson@ncsu.edu

Reliability in the DoD Context

Reliability : the ability of an item to perform **a required function**, under given **environmental** and **operating conditions** and for a stated **period of time** (ISO 8402, International Standard: Quality Vocabulary, 1986)

- **Operational mission reliability**
  - Most complex defense systems serve more than one required function (e.g., ships may provide transportation, defense, self-protection, etc.)
  - Multiple operating environments: desert, littoral, mountain, etc.
  - Operating conditions vary depending on mission
  - Requirements typically specify a fixed time period
- **An additional considerations in operational mission reliability**
  - Diverse population of system operators: crew-caused failures are still failures in a defense context.
- **Concept of operations – Essential for defining operational mission reliability**
  - Defines standard mission length
  - Provides a breakdown the expected activities during a mission
  - Can change over time as operational missions evolve

Motivating Example: Paladin Integrated Management (PIM)

- **The M109 Family of Vehicles (PIM) consists of two vehicles: the Self-Propelled Howitzers (SPH) and Carrier, Ammunition, Tracked (CAT) resupply vehicles.**
  - The M109 FoV SPH is the focus of this case study because of its two distinct functions. Additionally, the self-propelled 155 mm howitzer is designed to improve reliability over the legacy howitzer fleet.

- **PIM Mission - Field Artillery units employ the M109 FoV to destroy, defeat, or disrupt the enemy by providing integrated, massed, and precision indirect fire effects in support of maneuver units conducting unified land operations.**
  - In other words – must **move** with the unit and conduct fire missions **(shoot)**

PIM Operational Mission Reliability

- **System requirement**
  - Probability 0.75 of completing an 18-hour combat mission
  - Standard translation using an exponential distribution is a mean time between system aborts of 62.6 hours:

$$\int_{18}^{\infty} \frac{1}{62.6} \exp\left(-\frac{t}{62.6}\right) dt = 0.75$$

- **System Abort (SA) failures versus Essential Function Failures (EFF)**
  - System aborts are a subset of essential function failures
  - Very limited system abort data (3 out of 23 for one test)

PIM Operational Mission Reliability

- **Concept of Operations: Operational Mode Summary/Mission Profile (OMS/MP)**
  - During early testing an 18 hour-combat mission was specified as *drive 17.4 miles and shoot 223 rounds* (12.8 rounds/mile)
  - Prior to limited user testing the OMS/MP was updated to *drive 58.8 miles and shoot 104 rounds* (1.78 rounds/mile)

- **The requirement and OMS/MP highlight an issue**
  - How do we best measure PIM reliability for two distinct functions (driving and shooting)?

PIM Self Propelled Howitzer Data

- **Developmental testing focused on reliability from a driving and shooting perspective.**
  - – Lacking 18-hour mission context, hours were not recorded
- **Operational testing collected hours data and testing was conducted in 18-hour mission cycles.**
- **Test-Fix-Test Approach**
- **Data limitations**
  - – Usage rates are confounded with system changes
  - – Developmental testing focused on rounds/miles
  - – Operational testing in mission context

Test Phase	Vehicle	Number of Essential Function Failures	Cumulative Miles	Cumulative Hours	Cumulative Rounds fired	Ratio Rounds/Miles
Developmental Test 1	Vehicle 1	24	66.4		555	8.36
	Vehicle 2	21	67.3		445	6.61
Developmental Test 2	Vehicle 1	21	316.9		680	2.15
	Vehicle 2	22	254		743	2.93
Limited User Test	Vehicle 1	9	431.5	109.9	624	1.45
	Vehicle 2	16	432.6	112.8	623	1.44

Questions of Interest

Mission activities may be appropriately measured with different metrics

- Different activities (moving, shooting, idling) may be best measured in different units (miles, rounds, hours)
- Motivated by PIM limited data problem, but useful in other complex systems

System versus Mission Reliability

- Mission reliability depends on the use of individual systems across operational missions
- For a given analysis, how do we quantify mission reliability taking into account the range of operational missions?
  - PIM focus on usage rates (miles driven, rounds fired) – could be extended to include environmental and operator considerations

Structuring the Problem

	Activity 1	Activity 2	Activity 3
Subsystem 1	$f_{11}(\lambda_{11k})$	$f_{12}(\lambda_{12k})$	$f_{13}(\lambda_{13k})$
Subsystem 2	$f_{21}(\lambda_{21k})$	$f_{22}(\lambda_{22k})$	$f_{23}(\lambda_{23k})$
Subsystem 3	$f_{31}(\lambda_{31k})$	$f_{32}(\lambda_{32k})$	$f_{33}(\lambda_{33k})$

Miles

Shots/Rounds

Hours

Simulated Data

- As a starting point to address this problem, we simulated data based on the PIM reliability problem
- Simulated data allows us to answer the question for an ideal data collection case before addressing PIM’s data limitations

Mission Type	Miles	Rounds	Hours (Miles)	Hours (Rounds)	Hours(Idle)	Hours Total
Current OMS/MP	58.5	104	7.7	4.4	5.9	18
Midpoint	38	164	5.0	7.0	6.0	18
Low Operational Tempo	17.4	104	2.3	4.4	11.3	18
High Operational Tempo	58.5	223	7.7	9.5	0.8	18
Original OMS/MP	17.4	223	2.3	9.5	6.2	18

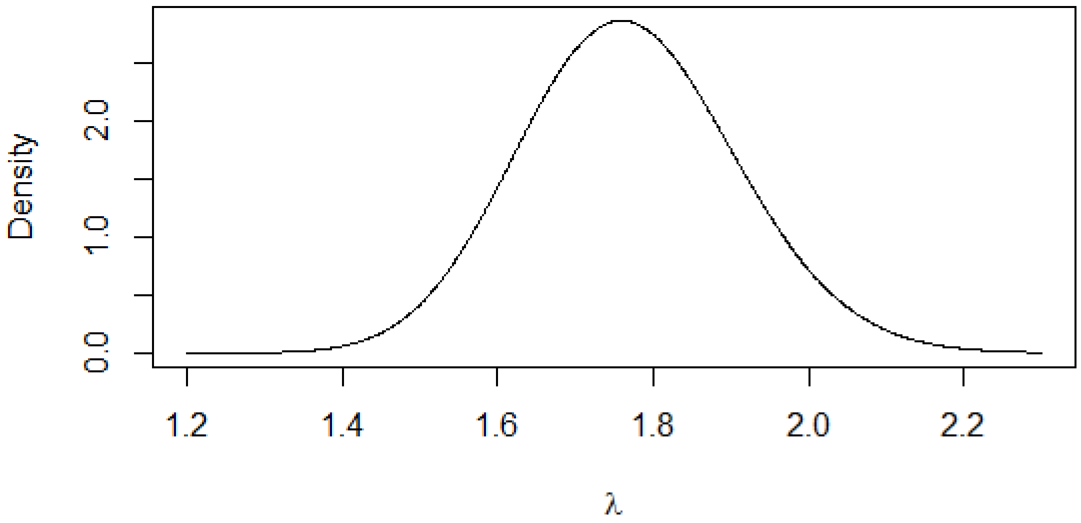
Simulated Data

- We simulated five missions.
- All lifetime distributions are exponential.
- Within each activity, there are three subsystems that might fail.
- We track miles, rounds, hours.
- Mission 4 (higher op-tempo) has failure rates multiplied by 1.5.

	Move	**Shoot**	**Idle**
Drive	$\lambda = 1.05$	$\lambda = 1.40$	$\lambda = 0.35$
Gun	$\lambda = 0.175$	$\lambda = 0.175$	$\lambda = 0.35$
Other	$\lambda = 0.175$	$\lambda = 0.175$	$\lambda = 0.35$

Model 1:
A Bayesian Version of the Standard DoD Solution

- Model all observations as Exponential(λ) using a diffuse prior.
- We may be able to develop an informative prior using data from previous tests.



- Only possible for the “real” data in operational testing, since we need all observations in common units.

Posterior Predictive Checks and DIC

- These are methods for model checking and selection.

- The idea of posterior predictive checks are to
  - Define y^{rep} to be “replicated” data—data that could have been observed if the same experiment was repeated with the same model and the same (unknown) value of θ that produced y.
  - Generate replicated data sets using draws from the posterior distribution of θ.
  - Compute values of a test statistic $T(y^{rep})$
  - Plot the values of the test statistic; see how these compare to what we actually observed in the data T(y)
  - Poor agreement suggests that the model may not have captured all of the “features of interest”

Posterior Predictive Checks and DIC

- The DIC is a penalized likelihood criterion used to compare models.

- Define the deviance as

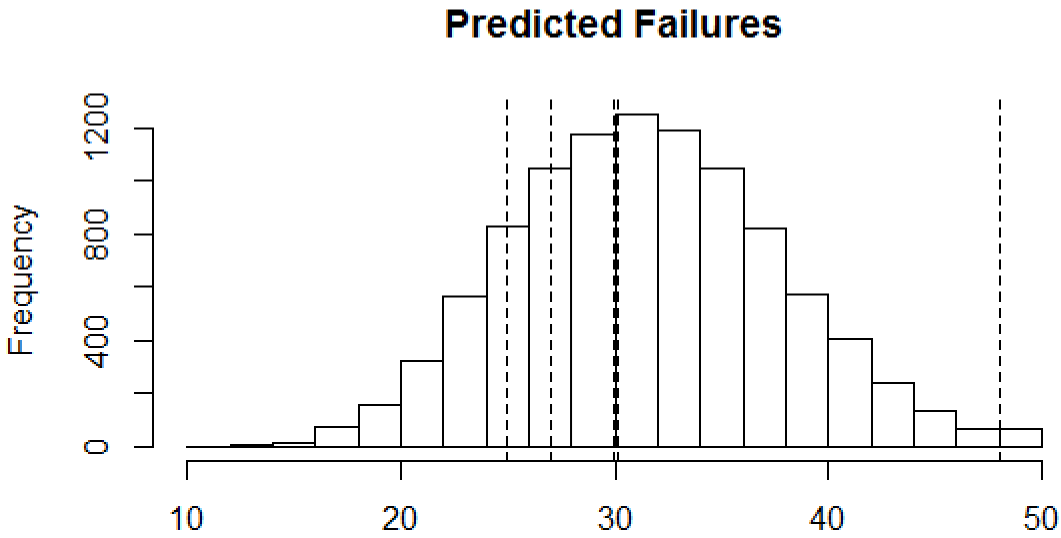
$$D(y,\theta) = -2\log f(y|\theta)$$

- The DIC is the sum of the average deviance (averaged over the posterior distribution of θ) and a penalty term p_D

$$p_D = D_{avg} - D\left(y, \hat{\theta}(y)\right)$$

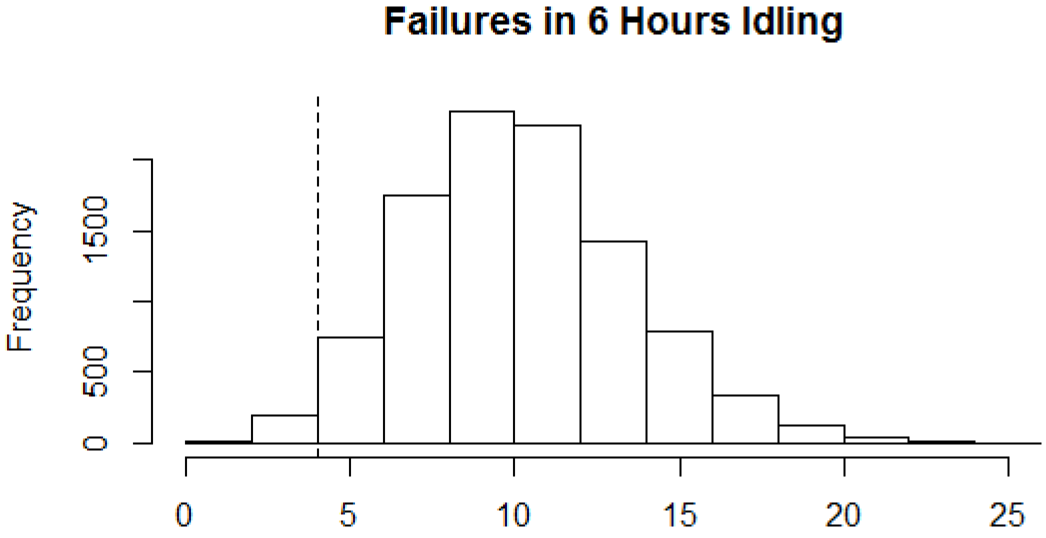
where $\hat{\theta}$ is an estimate of θ, often the posterior mean.

Model 1 Posterior Predictive Checks



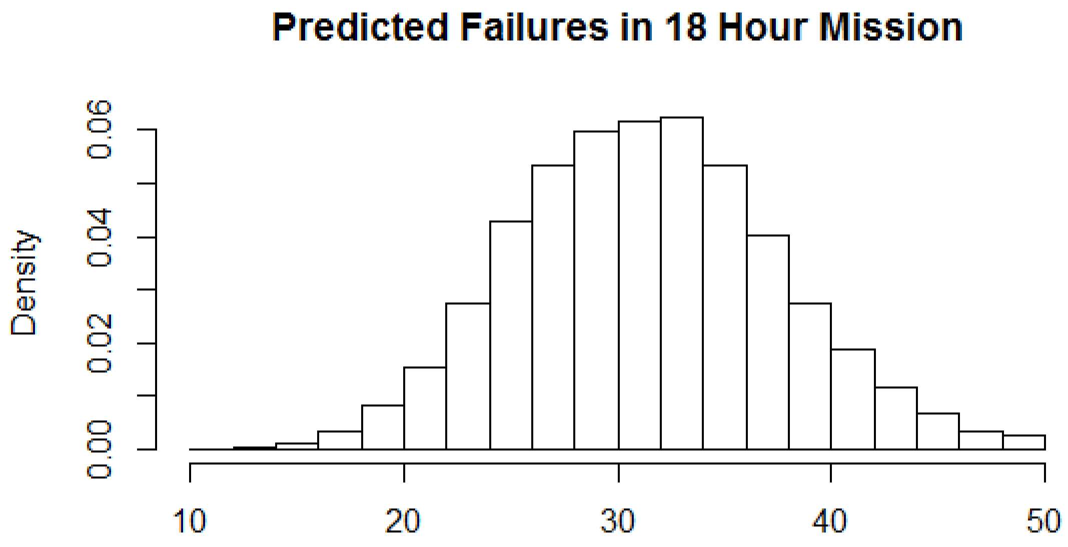
Evidence not all missions have same failure rate

Evidence not all activities have same failure rate (here, from Mission 1)



DIC = 119.19

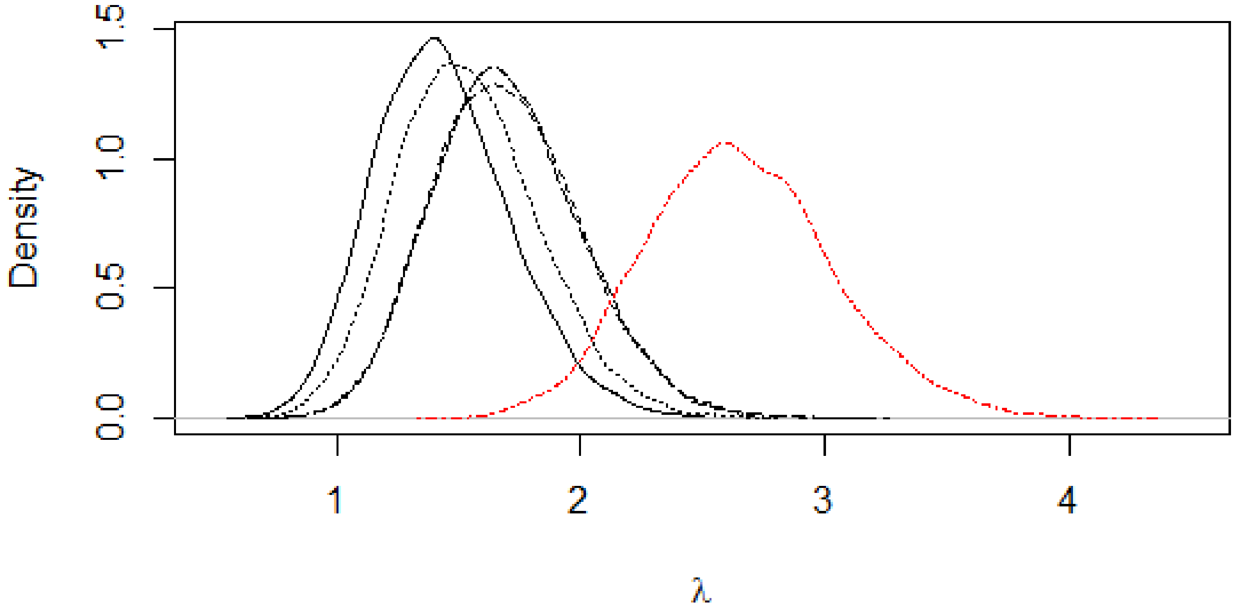
Model 1 Mission Reliability



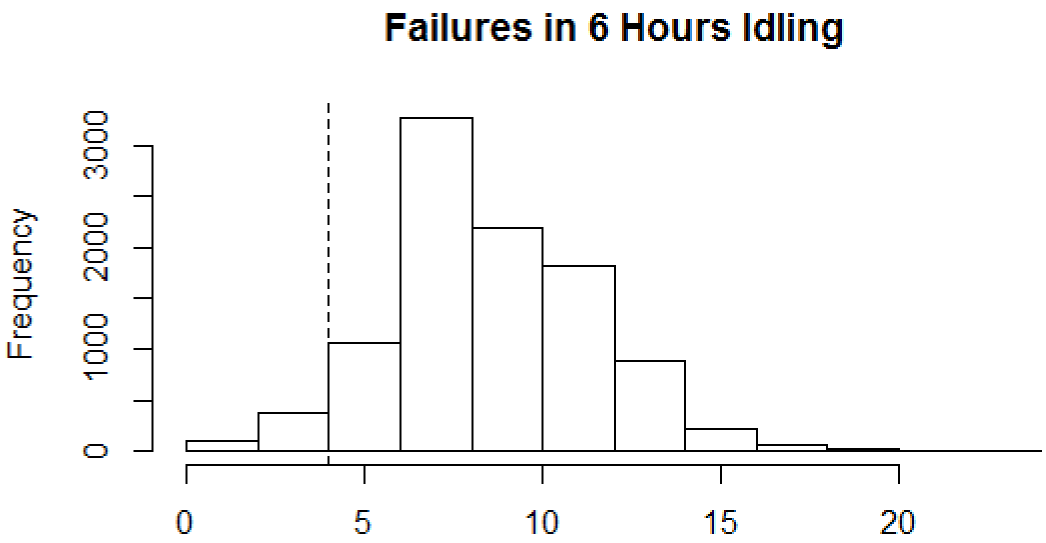
With this model, there is no way to account for variation in mission: a mission is simply 18 hours of operation.

Model 2: Accounting for Mission Differences

- Model each mission as having its own failure rate.
- Use a hierarchical prior for the mission failure rates.
- Develop an informative prior using data from developmental testing or use a diffuse prior.



Model 2 Posterior Predictive Checks

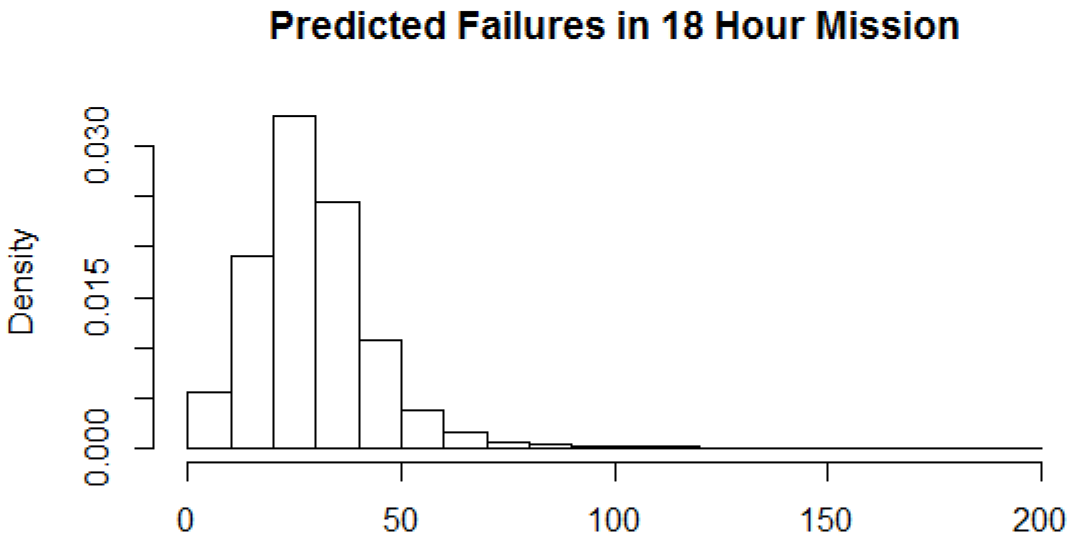


Evidence not all activities have same failure rate (here, from Mission 1)

DIC = 118.23

Lower value is better, but this difference is likely too small to be meaningful.

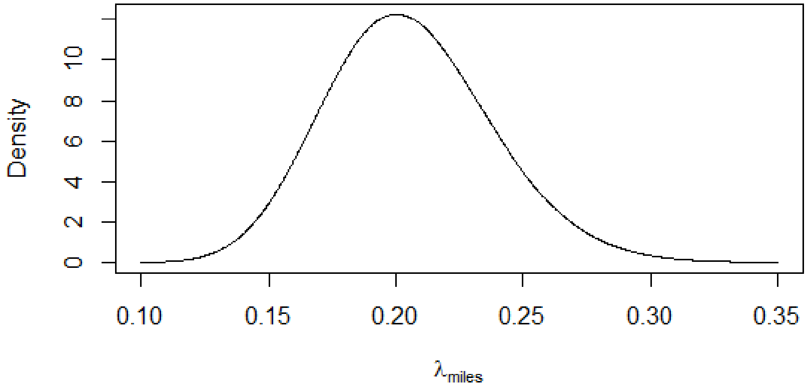
Model 2 Mission Reliability

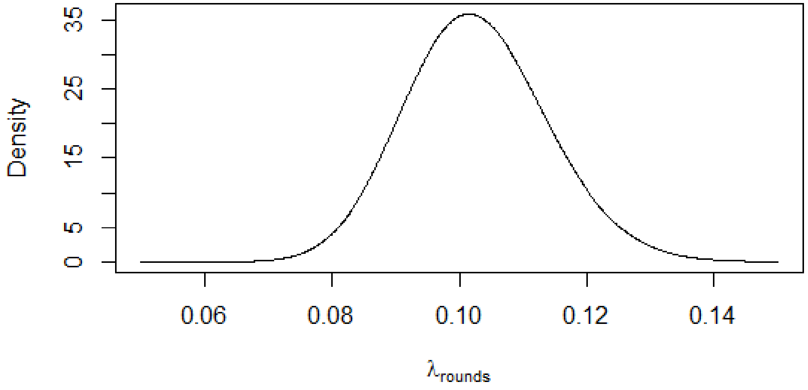


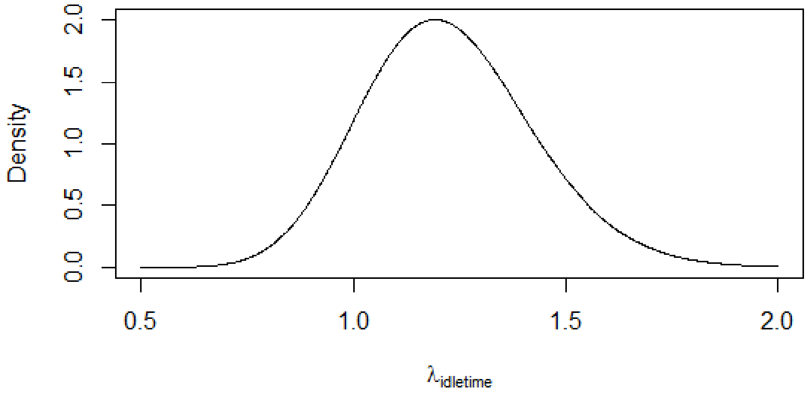
- Posterior median failure rate is similar, but more variability reflected.
- Still have no way to account for differences in missions.
- Without an informative prior, the predictions can be wide – five missions leaves uncertainty in the predictive distribution for $_{|}$.

Model 3:
Accounting for Activity Differences

- Model each activity as Exponential(λ_i) with a diffuse prior.

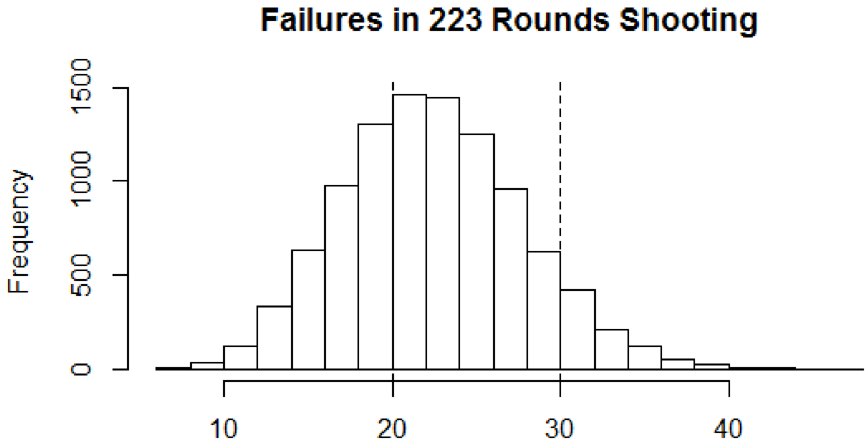






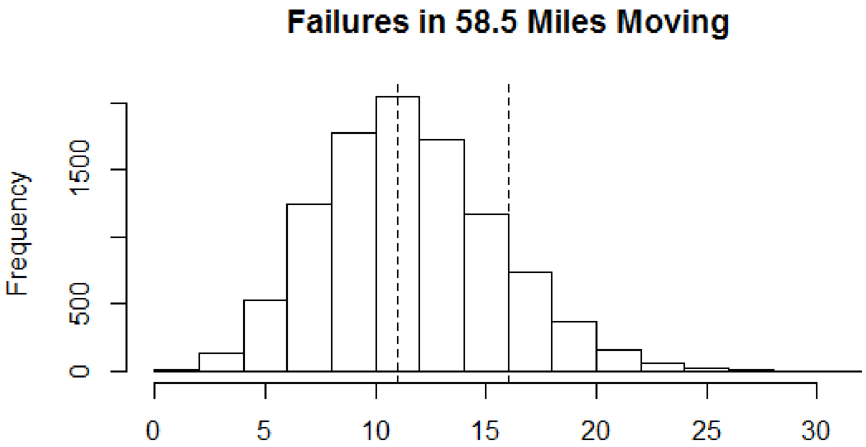
DIC = 118.69

Model 3 Posterior Predictive Checks

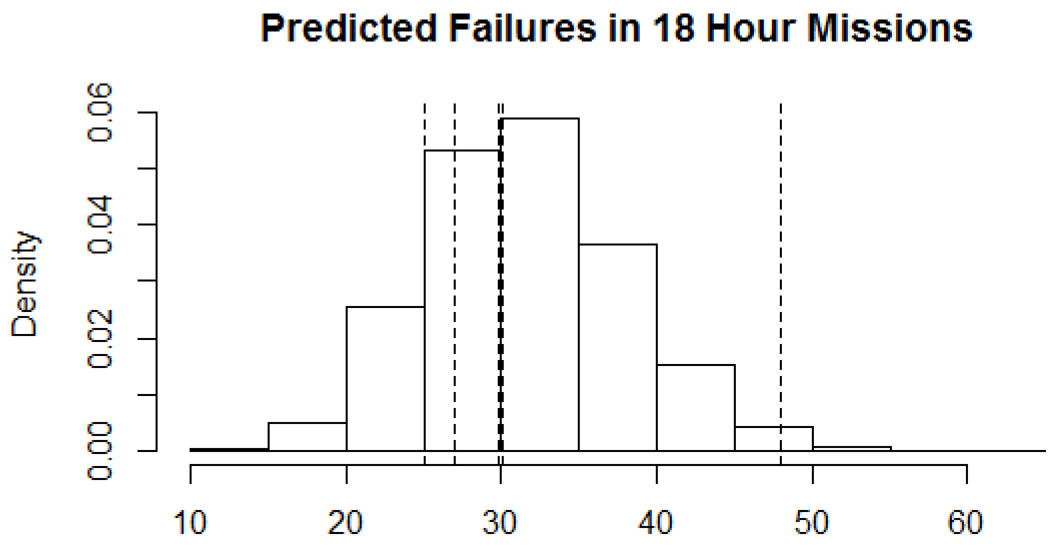


Missions 4 and 5

Missions 1 and 5

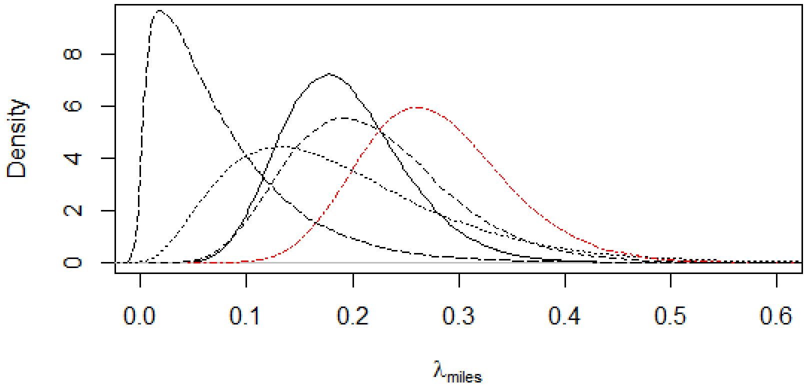


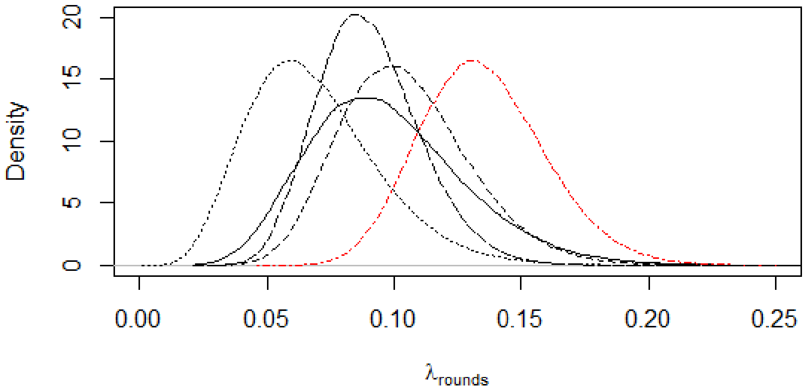
Model 3 Mission Reliability

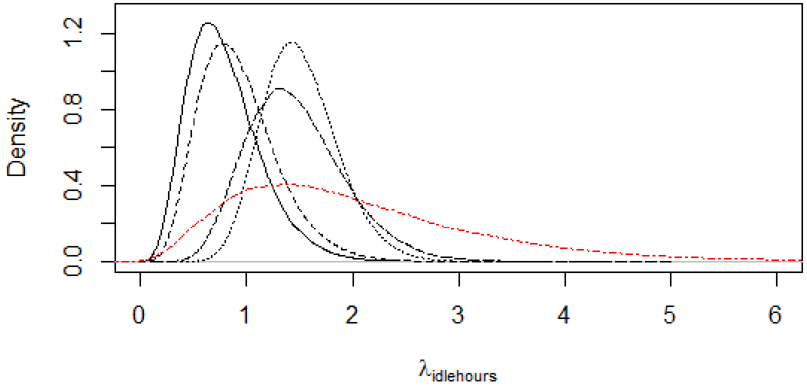


- "Representative" missions generated using Dirichlet distribution. For our simulated data, we can write missions in terms of hours per activity.
- In general, we determine how to simulate from a "distribution" of missions.
- Only about 1.5% of missions have as many failures as Mission 4.

Model 4:
Missions and Activity Based Analysis







DIC = 130.82

Lower value is better. More parameters than are supported by the data.

Looking Forward (1)

- **How do we use data from DT, LUT, and the test-fix-test paradigm?**
  - DT is not necessarily “operationally realistic.” It may, however, give us estimates of failure rates for specific activities.
  - It then becomes a modeling question about how we understand the changes in failure rates due to changes in operational realism and fixes to the system.

Experimental Design Approach

The simulated data used an “experimental design” approach to generating data

- Five missions followed a factorial design with center point layout
- No replication
- No controlling for order effects
  - It probably isn’t reasonable to assume that the later missions are not impacted by the earlier missions, especially in the case of crew-induced failure modes. Does that suggest we should start easy and progress to hard or randomly select, etc.?
- Limited data (5 missions with limited failures) provides low statistical power to test for mission effects in system reliability analysis.

Experimental Design Approach

Why experimental design then?

- Ensures coverage of operational mission usage
- If failure rates change dramatically by mission, then we have a chance to detect this change
  - It is not unreasonable to assume that “operational tempo” might impact failure rates

Looking Forward (2)

- **What’s a smart way to design the sequence of tests to let us understand mission reliability?**
  - PIM is a relatively simple mixing of three primary activities to make a mission
  - Ships may consist of dozens of tasks using dozens of system functions to complete very different looking missions
  - Can we expand this analysis to address complex systems more holistically than simply converting to hours?

- **Can we adapt assurance testing ideas to plan the OT?**
  - Is there a better breakdown for covering missions than a simple experimental design approach?