JSC Safety & Mission Assurance Data Analysis Overview

Roger L. Boyer
Analysis Branch Chief

and

Bruce C. Reistle
Data Lead

December 14, 2010

Why Data Analysis?

There is data and there is data. What one needs may not be what one has. Therefore, one needs to adjust or improvise. This is data analysis.

- First, what question(s) are you trying to answer?
- Can you answer them directly or indirectly?
- Is this a statistical or a probabilistic question?
  - Statistical issues imply substantial data (i.e. we've seen this many times before)
  - Probabilistic issues imply the lack of data (i.e. hardly seen it at all or hope to never see it)
- Given answers to these questions, guides the analyst down different paths to the solution

Topics

These slides describe the data analysis methods that are used to determine inputs for probabilistic risk models supporting the Space Shuttle Program. Other applications can follow a similar path probably using different data sources. Statistical approaches are different and not addressed here. Topics included here:

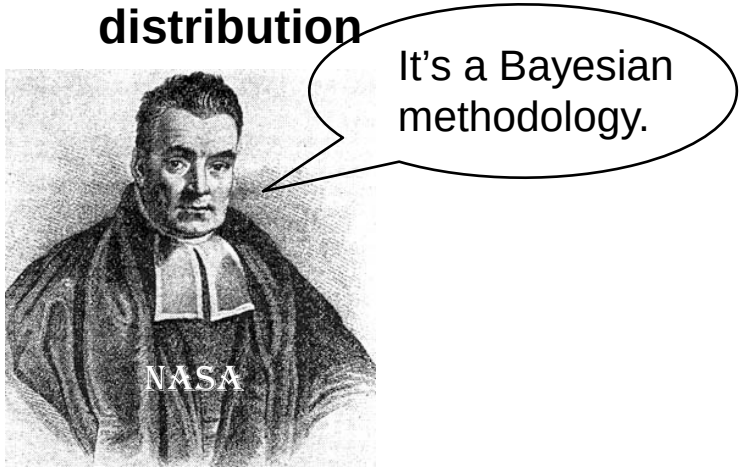
- **Prior Distribution**
- **Likelihood Data**
- **Bayesian Updating**
- **Uncertainty and Error**

Note:
This is a high-level discussion and is not intended to be a tutorial.

High Level View

Obtaining data for risk models is a process where you start with your initial best estimate of a failure rate (or probability), then make adjustments to that estimate as new information becomes available.

- The initial best estimate is called the **prior distribution**
- The new information is called the **likelihood data**
- The adjusted estimate is called the **posterior distribution**



Reverend Thomas Bayes

High Level View—Cont'd

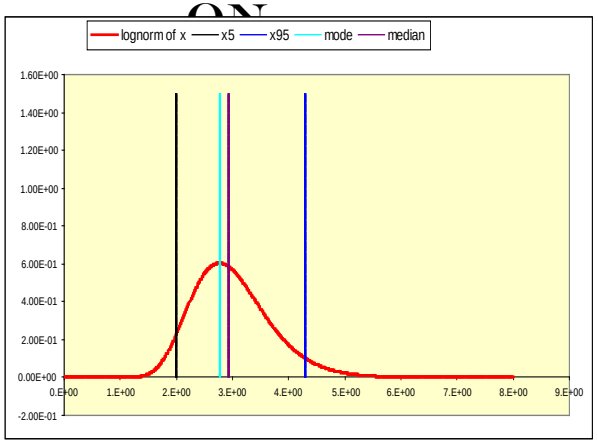
Bayesian Updating refers to the process of using Bayes' Theorem to combine the prior with the likelihood to get the posterior.

PRIOR → LIKELIHOOD → POSTERIOR

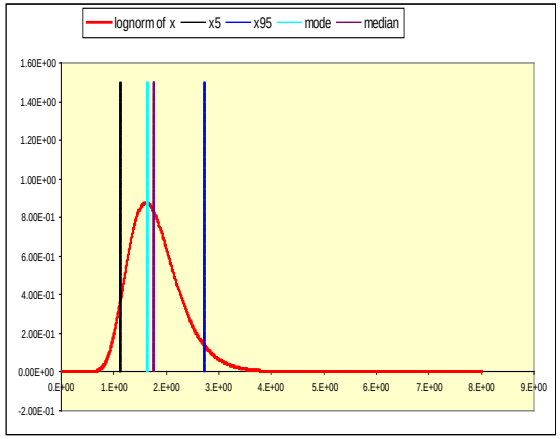
INITIAL FAILURE RATE DISTRIBUTION

NUMBER OF FAILURES ALONG WITH EXPOSURE TIME OR DEMANDS

UPDATED DISTRIBUTION



Prior Distribution



Posterior Distribution

Prior Distribution

The **prior distribution** can be determined by

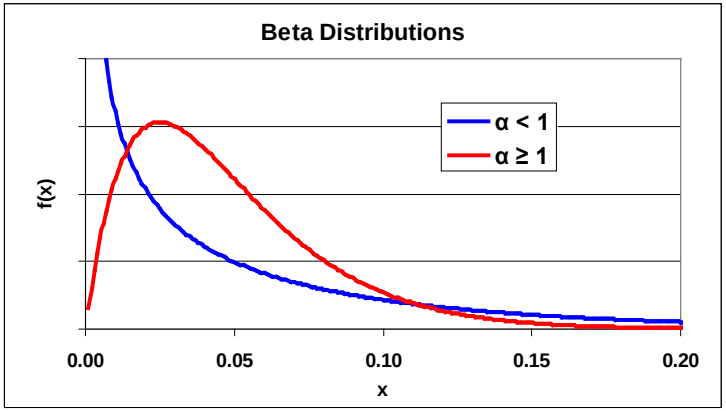
- Historical data—Shuttle data, Soyuz, Air Force, etc.
- Expert elicitation—A formal and rigorous process with a panel of experts
- Vendor estimates—Boeing, Honeywell, etc. based on testing, history, or analysis
- Parts-count analysis—MIL Std, Relex, PRISM, etc.
- Surrogate data—NPRD, EPRD, NUCLARR, etc.
- Rule of thumb—Based on component type, e.g., electrical, mechanical, etc.

In general, demand-based priors are modeled using a **beta** distribution and rate-based priors are modeled using a **gamma or lognormal** distribution.

Prior Distribution—Cont'd

Beta—Demand Based

- Is the most common choice for modeling probabilities
- It is bounded on $(0, 1)$ so there is no chance of selecting a probability > 1.0
- Has parameters a and b
- Given a failures and b successes:
  - The mean is $\mu = \frac{a}{a+b}$
  - The variance of the mean is $\sigma^2 = \frac{ab}{(a+b)^3}$

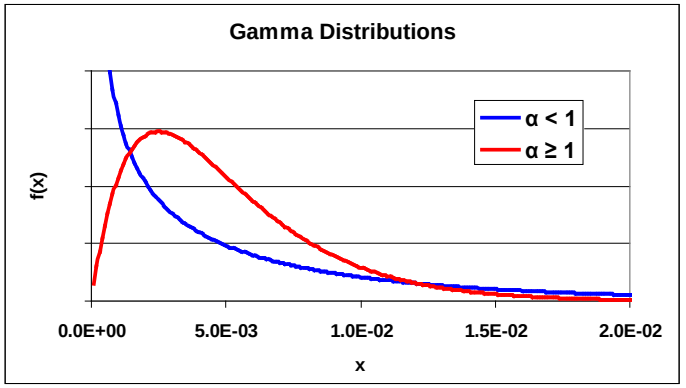


The beta distribution is an obvious choice for modeling random variables that are constrained by zero and one. That does not mean it is always the "correct" distribution but it is usually the default demand-based distribution unless there is additional information available that suggests otherwise.

Prior Distribution—Cont'd

Gamma—Rate Based

- Has parameters α and β
- When $\alpha > 1$, it takes a similar shape to the lognormal distribution
- It is the conjugate of the Poisson distribution
- Given α failures and β operating time:
  - The mean is $\mu = \frac{\alpha}{\beta}$
  - The variance of the mean is $\sigma^2 = \frac{\alpha}{\beta^2}$



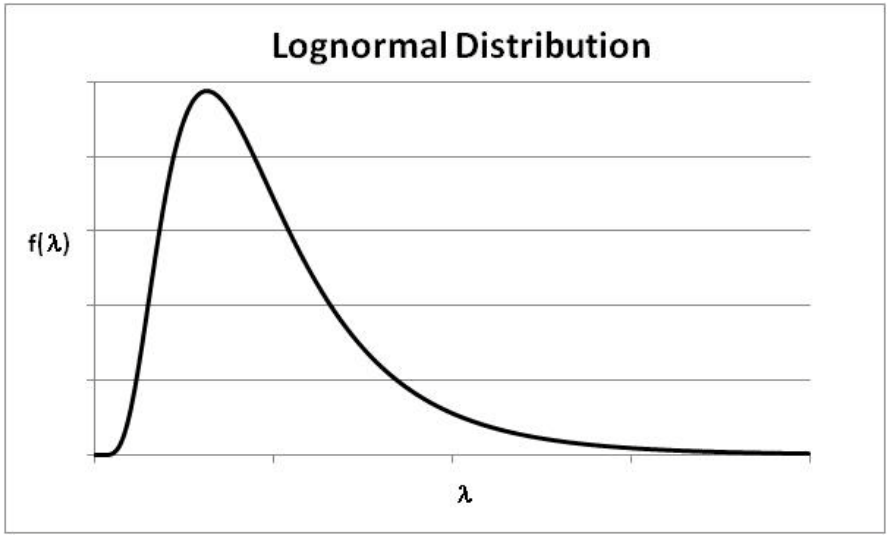
The gamma distribution is frequently used as a waiting time distribution (i.e., time until death). The gamma distribution is selected because of its shape (similar to a lognormal) and because it is the conjugate to the Poisson.

Prior Distribution—Cont'd

Lognormal—Rate Based

- Is the traditional distribution for modeling failure rates in PRAs
- Any positive value is possible—from zero to infinity
- The density is focused on the left but its fat tail allows for larger values
- It is often described by the mean and Error Factor (EF)
  - $EF = \sqrt{\frac{95th}{5th}}$

All lognormal distributions have the shame general shape.

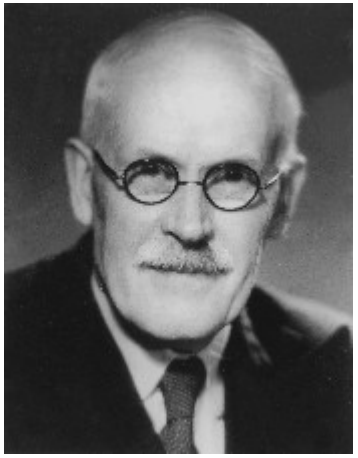


Prior Distribution—Jeffreys Priors

Sometimes there is insufficient information to form an informed prior distribution. Also, sometimes data sources are zero-failure sources. In these cases a Jeffreys Noninformative Prior may be used.

The Jeffreys Prior for the **beta distribution** is: $\text{Beta}(a = 1/2, b=1/2)$
The Jeffreys Prior for the **gamma distribution** is: $\text{Gamma}(a = 1/2, b = 0)$

Note: It might appear that these priors simply assume "half a failure." The derivation of the Jeffreys Prior makes no such assumptions and the appearance of "half a failure" is *coincidental.*



Sir Harold Jeffreys
(1891 – 1989)

Note: "Jeffreys Prior" is not typically expressed as a possessive whereas "Bayes' Theorem" is.

Prior Distribution—Cont'd

How do you **form a prior from multiple data sources**?

Current methodology—at a very high level—involves getting an overall mean and an overall variance.

Likelihood Data

Likelihood data

- Data that is collected after forming the prior
- Usually comes directly from the system being modeled

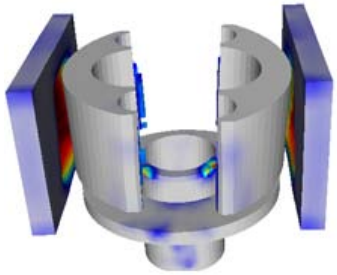
Typically, likelihood data arrives in the form of **failures** over a given **exposure**.

Failures

- Typically based on screening corrective action reports
- Can be assigned as **partial failures** based on redesigns or “fixes”

Exposure

- “Exposure” refers to usage
- Can be rate based, e.g., per hour, per flow rate, etc.
- Can be demand based, e.g., per detonation, per attempt, per impulse, etc.



Frangible Nut
(Demand Based)

Likelihood Data—Cont'd

Likelihood data must have an underlying **distribution**.

Binomial—Demand-Based

- Is a discrete distribution
- Counts events over a fixed number of demands
- Assumes each attempt has a constant failure probability
- Assumes the attempts are independent
- Assumes no wearout
- Is the conjugate of the beta distribution

Poisson—Rate-Based

- Is a discrete distribution
- Counts events over a fixed time period
- Assumes each time interval has a constant failure rate
- Assumes the time intervals are independent
- Assumes no batch arrivals, no wearout
- Is the conjugate of the gamma distribution

Conjugate Pairs

Some combinations of prior and likelihood distributions result in posteriors with known distribution types. These combinations are called conjugate pairs. Conjugate pairs are easy to update.

The **beta-binomial** and **gamma-Poisson** are the most commonly used conjugate pairs.

Conjugate Pairs

Prior	Likelihood	Posterior
uniform	Bernoulli	beta
beta	Bernoulli	beta
gamma	**Poisson**	**gamma**
normal	normal	normal
gamma	exponential	gamma
beta	**binomial**	**beta**
Pareto	uniform	Pareto
beta	negative binomial	beta
gamma	normal	gamma
inverse gamma	exponential	inverse gamma
Dirichlet	multinomial	Dirichlet

Types of Uncertainty

Aleatory uncertainty is the inherent (irreducible) uncertainty in the time to failure of a given item. A typical assumption of aleatory uncertainty is that all rate-based items have exponential distributions with parameter λ and all demand-based items are Bernoulli trials (i.e., binomial) with parameter p.

The **epistemic uncertainty** is what is captured by expressing the uncertainty of λ and p (e.g., lognormal, gamma, beta, etc.).

Example—If we knew λ with certainty (we never do) all that would remain would be aleatory uncertainty (Figure 1). Applying epistemic uncertainty to the parameter results in a range of possible exponential distributions (Figure 2).

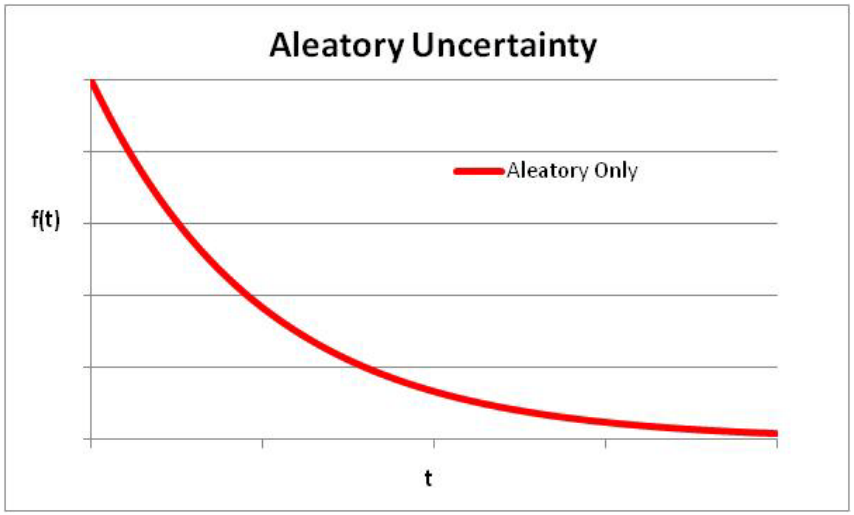


Figure 1

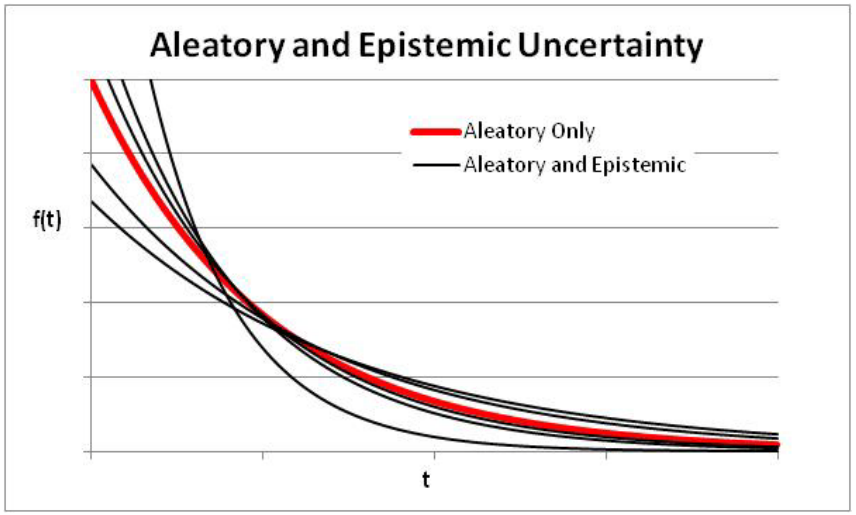


Figure 2

Sources of Error

Possible sources of error include:

- Prior not representative
  - Poorly chosen
  - Calculated incorrectly
  - Incorrect failure count or exposure time
- Likelihood error
  - Not modeled correctly (e.g., poor distribution choice)
  - Incorrect failure count or exposure time
  - Data not representative of system being modeled
- Posterior error
  - Calculated incorrectly
  - Approximation error
    - Error due to moment matching
    - Error due to not moment matching

References

Probabilistic Risk Assessment Procedures Guide for NASA Managers and Practitioners

– http://www.hq.nasa.gov/office/codeq/doctree/praguide.pdf

NUREG/CR-6823 Handbook of Parameter Estimation for Probabilistic Risk Assessment

– http://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr6823/

Bayesian Inference for NASA Probabilistic Risk and Reliability Analysis

– http://www.hq.nasa.gov/office/codeq/doctree/SP2009569.pdf