INSTITUTE FOR DEFENSE ANALYSES

Introduction to Bayesian Analysis

Heather M. Wojton, Project Leader

Keyla Pagan-Rivera
John T. Haman
Rebecca M. Medlin

August 2021

Approved for Public Release.
Distribution Unlimited.

IDA Document NS D-20484

Log: H 2020-000510/2

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

Approved for public release; distribution is unlimited.

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 229990, "Test Science Research and Applications," for the Office of the Director, Operational Test and Evaluation. 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 Mr. Robert R. Soule and consisted of Dr. Kelly M. Avery, Dr. Leonard D. Wilkins, Mr. Curtis G. Miller, and Dr. Thomas H. Johnson from the Operational Evaluation Division.

For more information:
Heather M. Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Rigorous Analysis | Trusted Expertise | Service to the Nation

Approved for public release; distribution is unlimited.

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-20484

Introduction to Bayesian Analysis

Heather M. Wojton, Project Leader

Keyla Pagan-Rivera
John T. Haman
Rebecca M. Medlin

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Executive Summary

As operational testing becomes increasingly integrated, and research questions become more difficult to answer, IDA's Test Science team has found Bayesian models to be a powerful set of data analysis methods. Analysts and decision-makers should understand the differences between this approach and the conventional way of analyzing data. It is also important to recognize when an analysis could benefit from the inclusion of prior information (what we know about a system's performance before testing) and to know the proper way to incorporate it.

To apply Bayesian methods, analysts need to comprehend important technical aspects of this approach, and they need to know how to properly use statistical software. The intent of this course is to demonstrate the essentials of Bayesian modeling with practical examples, rather than slog through all the technicalities. Still, some details of the course are somewhat technical, so the presenters have assumed a basic familiarity with conventional statistics and R software.

Course presenters recognize that analyzing operational test data is often complicated, and data problems are difficult to put into a single methodological box. To this end, this course provides students with a sample of the most common techniques in Bayesian applied statistics: single-parameter models and linear regression models.

This course covers these topics in three sections:

- *Bayesian Thinking*: As the title suggests, this section introduces the concept of Bayesian thinking. The section starts with a motivational example that showcases the differences between Bayesian and frequentist probability. Then, it covers some historical details of Bayesian statistics, including how analysts have been combining data when evaluating military systems for some time. Following this introduction, the presenters provide more details about the elements that comprise Bayes' theorem: the prior, likelihood, and posterior distributions. Next, they use two examples to illustrate how to apply Bayes' theorem, giving details about the role of the prior distribution. This section ends with a summary of the benefits and limitations of Bayesian statistics and a comparison of Bayesian to frequentist statistics.

- *Single-Parameter Models*: This section familiarizes students with some technical aspects of Bayesian methods, using single-parameter models and conjugate priors. The case studies in this section present various ways of incorporating prior information, show how to mathematically derive the posterior distribution, and describe different ways of coding the Bayesian models using statistical software. Students also learn about model assessment and some considerations to keep in mind when choosing the prior distribution.
- *Linear Regression Models*: Simple and multiple linear regression models are the focus of this section. Throughout the section, students learn how to construct prior densities for the parameters of the regression model. Students also learn how to determine whether the model fits the data well. For example, they may use posterior predictive checks to assess model fit, or they may transform a variable to improve model fit. These topics are illustrated through two notional examples, mathematical details, visualizations, and software implementations.

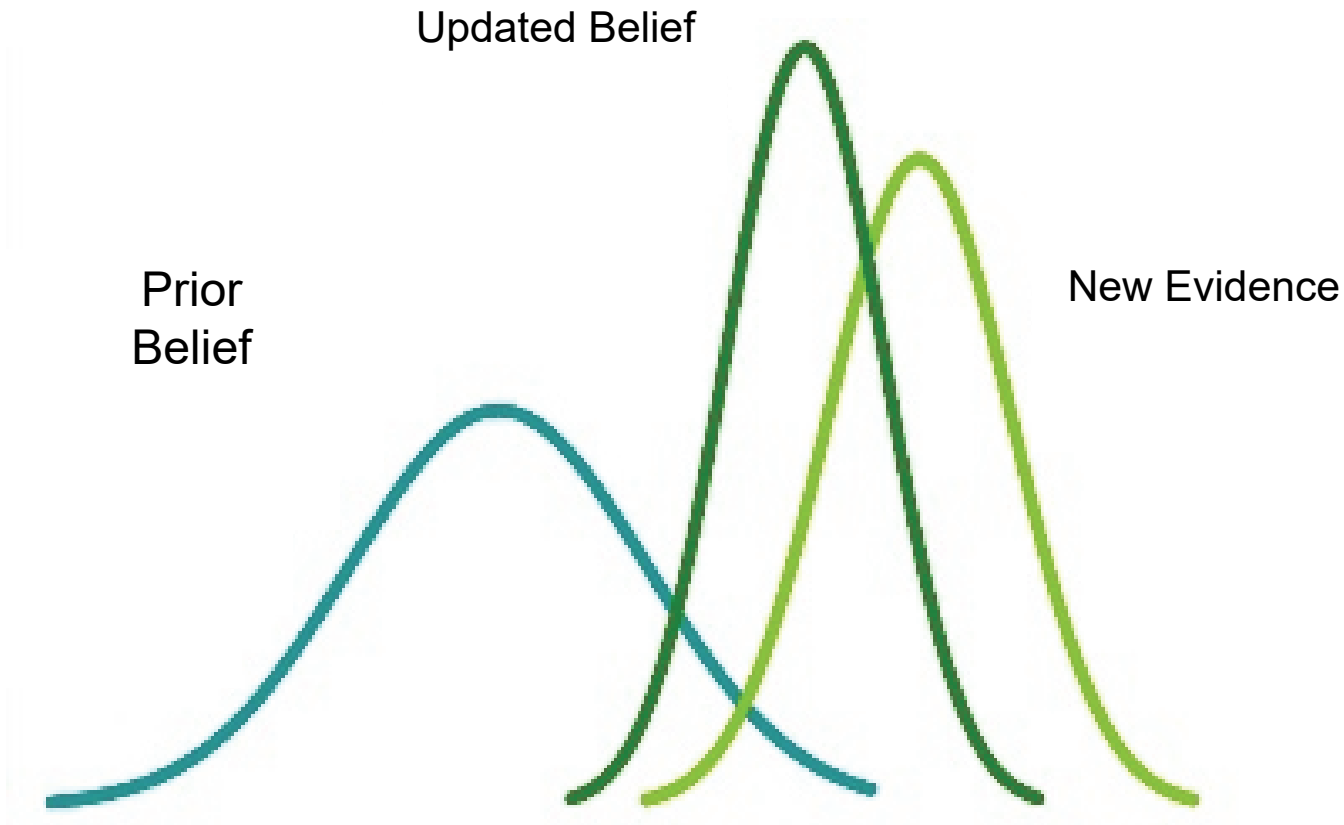
Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Introduction to Bayesian Analysis

Section I – Bayesian Thinking



Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

“Everybody is a Bayesian. It's just that some know it, and some don't.” – Trivellore Raghunathan

Approved for public release; distribution is unlimited.

Prior knowledge could affect your belief about some outcomes

- Experiment 1
  - A fine musician, specializing in classical works, tells us that he can distinguish whether Haydn or Mozart composed a classical song. Small excerpts of the compositions of both authors are selected at random and played for the musician to identify. The musician makes 10 correct guesses in 10 trials.

- Experiment 2
  - The guy next to you at the bar says he can correctly guess in a coin toss what face of the coin will fall down. Again, after 10 trials the man correctly guesses the outcomes of the 10 throws.

Approved for public release; distribution is unlimited.

Frequentist vs. Bayesian thinking

- Frequentist statistical analysis
  - You have the same confidence in the musician's ability to identify composers as in the bar guy's ability to predict coin tosses. In both cases, there were 10 successes in 10 trials.
- Bayesian statistical analysis
  - Presumably, you are inclined to have more confidence in the musician's claim than the guy at the bar's claim. Post-analysis, the credibility of both claims will have increased, though the musician will continue to have more credibility than the bar guy.

Using frequentist statistics for some jobs and Bayesian statistics for others does not mean you have to sign up for a lifetime of using only one tool!

Approved for public release; distribution is unlimited.

Goals of this training

- Introduce Bayesian statistics and the similarities and differences between Bayesian statistics and classical statistics
- Explain the methodologies behind Bayesian models and their implementation using software
- Demonstrate some ways of intelligently combining information
- Describe ways to communicate the results from a Bayesian data analysis

Approved for public release; distribution is unlimited.

Statistics is more than summarizing data

Operational test reports summarize the findings of a test, but simply tabulating operational test data is not sufficient.

We should also strive to:

- Generalize from operational test to operational employment
- Generalize from sample of operators to population of operators
- Infer performance in untested conditions

To accomplish these things we need to build statistical **models**.

Approved for public release; distribution is unlimited.

Bayesian methods can be used in operational testing and evaluation

- Operational tests can be complex, expensive, and time consuming, but they are our best tool to understand a system's performance
- Most of the time, data from tests are analyzed independently of any prior test data or subject matter expert knowledge
- Analysts could use Bayesian statistics to wisely incorporate available information when analyzing operational test data

Approved for public release; distribution is unlimited.

Motivation for using ALL information is not new

Military Operations Research Society / International Test and Evaluation Association

Joint Mini-Symposium: How Much Testing Is Enough?, 1994

National Research Council Studies

Statistics, Testing, and Defense Acquisition, 1998

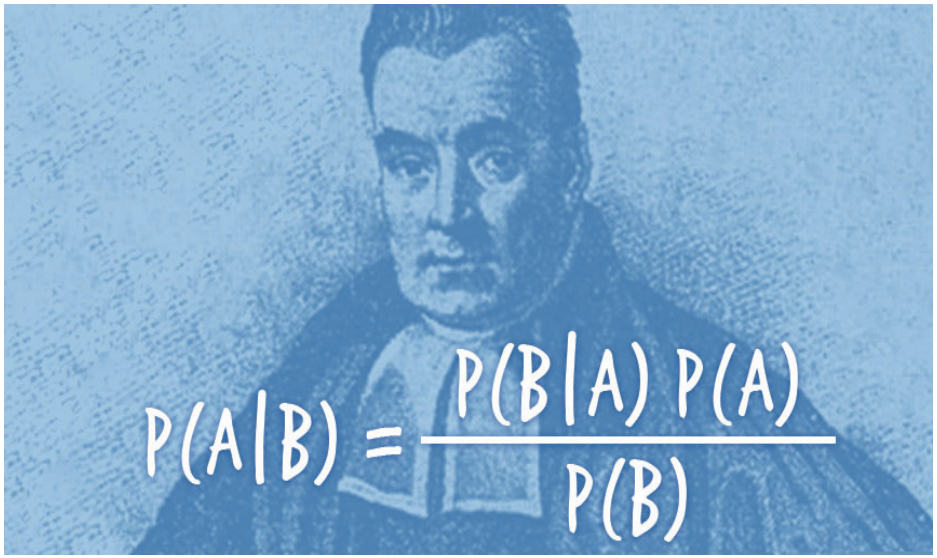
Improved Operational Testing and Evaluation, 2004

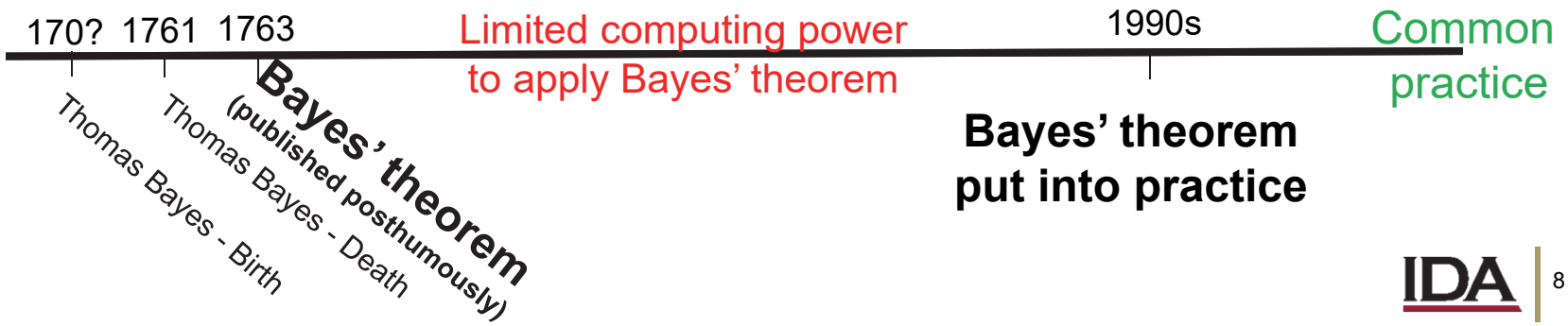
The idea of combining all sources of information has been discussed in the defense community for some time.

Approved for public release; distribution is unlimited.

Bayesian methods have been around for centuries

English statistician, philosopher, and clergyman Thomas Bayes formulated a way to calculate the likelihood of an event based on prior knowledge.





Approved for public release; distribution is unlimited.

Bayesian statistics in a nutshell

Prior or available information from subject matter expert, legacy system, developmental testing, other

Observed data plus model for the data (for example, Binomial or Normal)

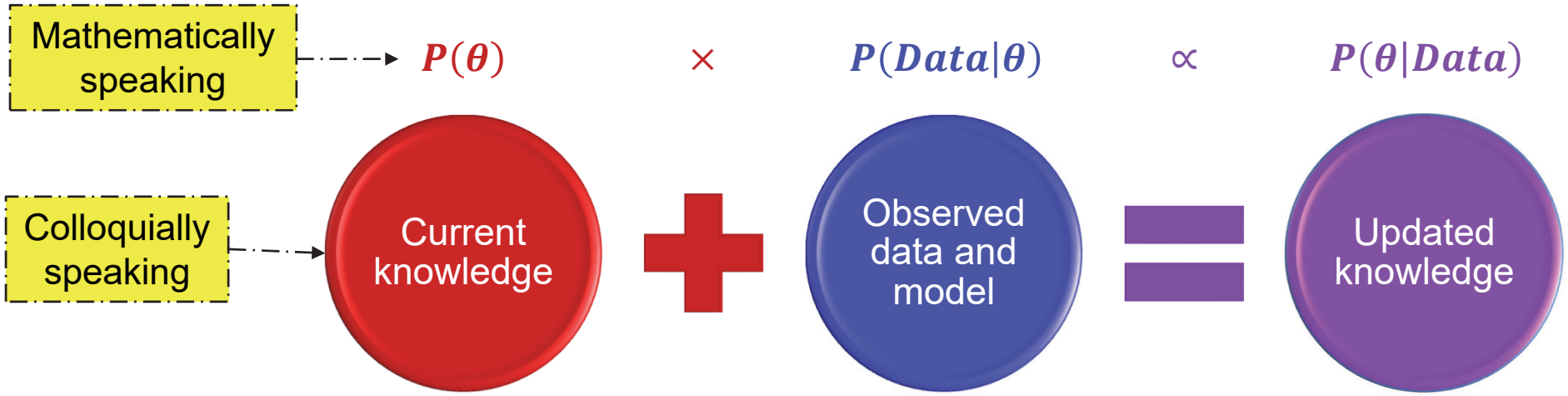
Make inferences based on this (for example, point estimates and intervals)

In order to combine the prior information with the observed data, we need to understand the military system and statistics.

Approved for public release; distribution is unlimited.

A few more details…

- Bayes' theorem: $\boldsymbol{P(\theta|Data) \propto P(\theta)P(Data|\theta)}$
- Analysts can decide how much weight to put into the prior distribution (vague or informative prior)
- At minimum, analysts should have an idea of the possible values the prior might take



θ could be any measure of system performance (for example, reliability or success rate).

Approved for public release; distribution is unlimited.

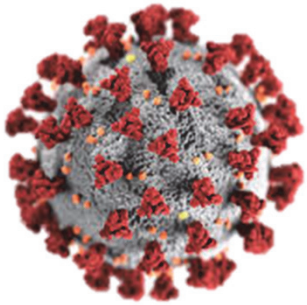
What is the probability that someone has COVID-19?

Suppose someone has been taking precautions but still has COVID-like symptoms.

Before getting tested, they know that the probability of having COVID-19, $P(C^+)$, is 0.05.*

They also know that the test is not 100% perfect. In fact:

- The probability of having a positive result given that the person has COVID-19, $P(T^+|C^+)$, is 0.95.*
- The probability of having a negative test result given that the person in fact does not have COVID-19, $P(T^-|C^-)$, is 0.98.*



* See the slide notes for more information about these numbers.

Approved for public release; distribution is unlimited.

Let's collect data and use the information we have to find out!

Now assume that this person's result came back positive.

We can update our knowledge based on this new information:

$$P(C^+|T^+) \propto P(T^+|C^+)\, P(C^+)$$

Note that the equation uses $\propto$. We need to normalize the result so it is a probability.*

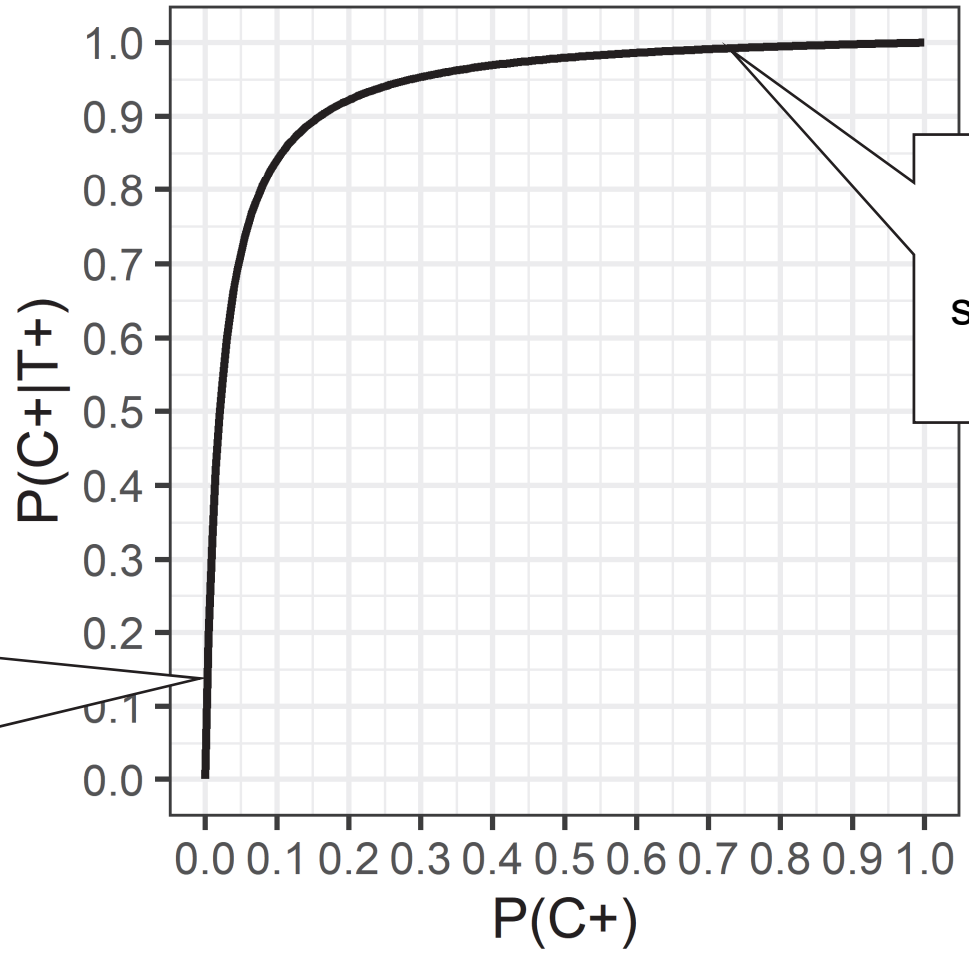
$$P(C^+|T^+) = \frac{P(C^+)P(T^+|C^+)}{P(C^+)P(T^+|C^+)+P(C^-)P(T^+|C^-)} = \frac{0.05\times0.95}{0.05\times0.95+0.95\times0.02} = \mathbf{0.71}$$

The chances of having COVID-19 increased after testing positive.

* In future examples, we work with distributional functions (for example, Normal or Binomial). In those cases, the distribution will "take care" of the normalization.

Approved for public release; distribution is unlimited.

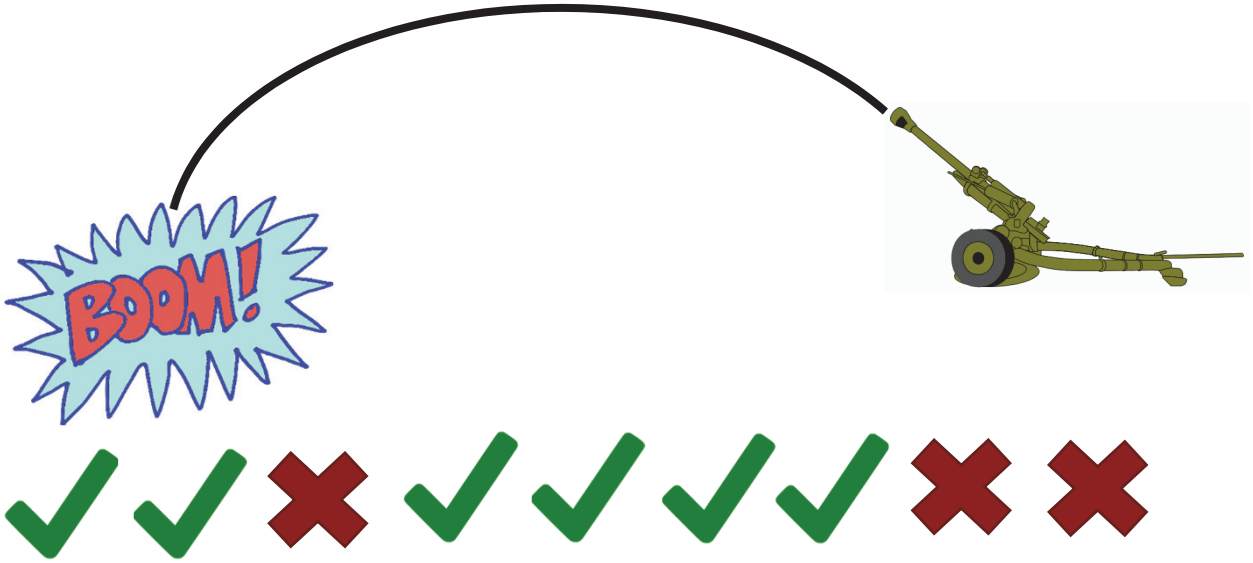
$P(C^+|T^+)$ depends on $P(C^+)$



If COVID-19 is very uncommon, the detections are not credible

Test credibility increases substantially with prevalence

Incorporating historical data (notional example)



Assume the legacy system test results were:

5 ✓

1 ✗

Frequentist approach:

Uses current test data

Probability of a successful launch = 6/9 = 0.67

95% confidence interval is (0.36, 0.98)

Bayesian approach:

Incorporates historical data

Probability of a successful launch = 0.73

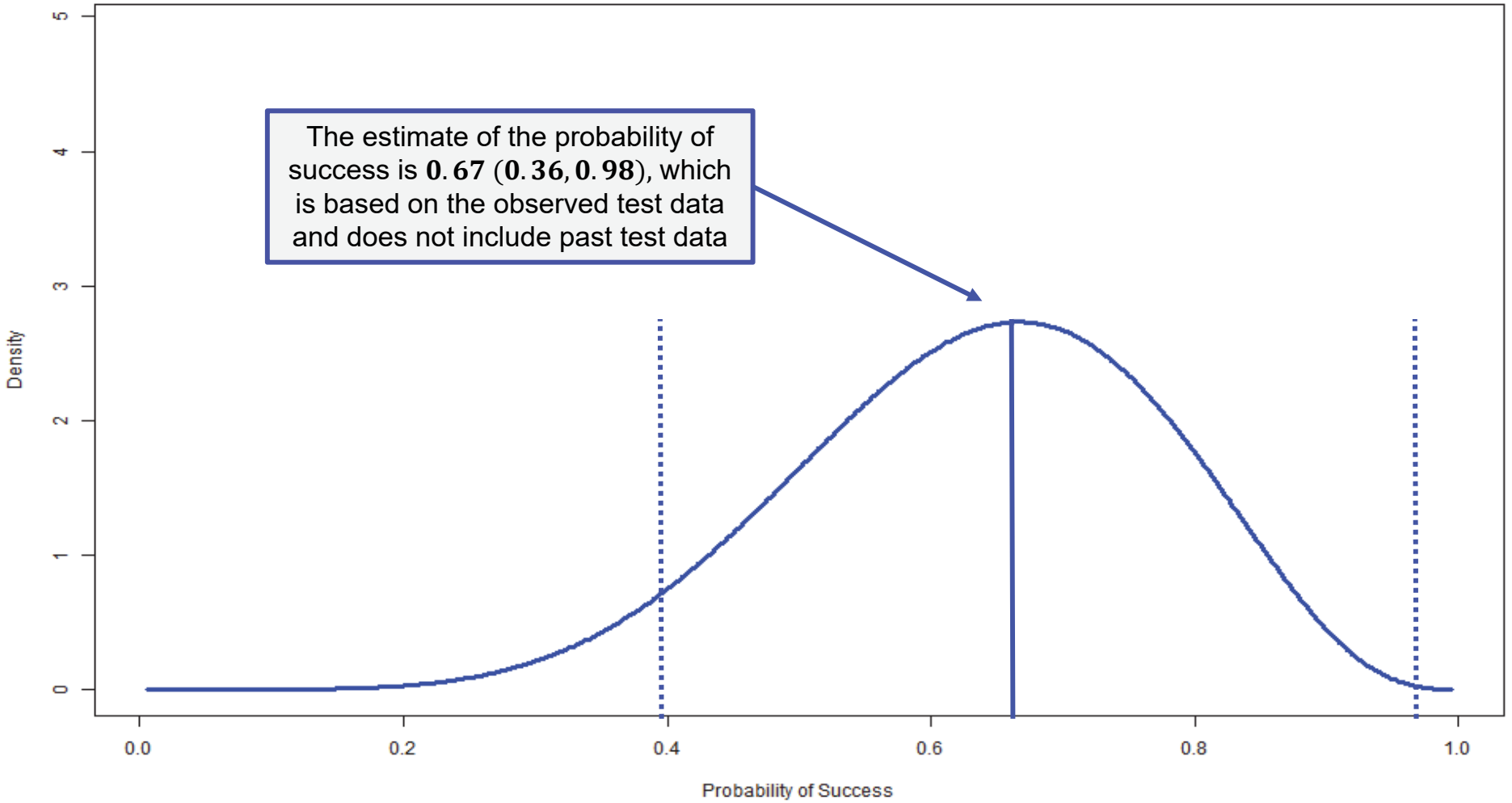
95% confidence interval is (0.49, 0.92)

Example from https://testscience.org/characterize-system/test-evaluation-analyses/bayesian-credible-intervals/

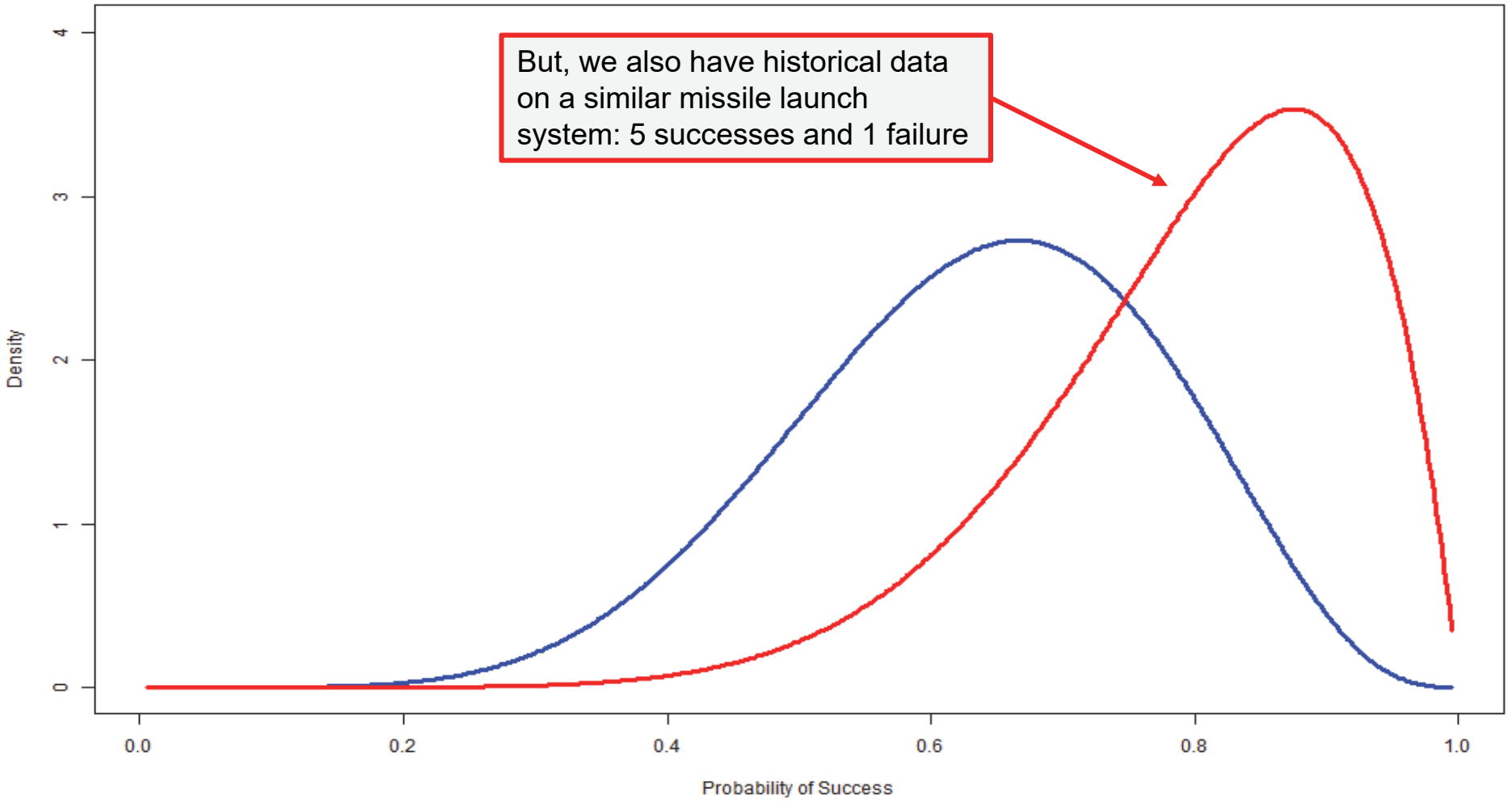
Assuming equal weight for prior and observed data. We discuss more realistic options later.

Approved for public release; distribution is unlimited.

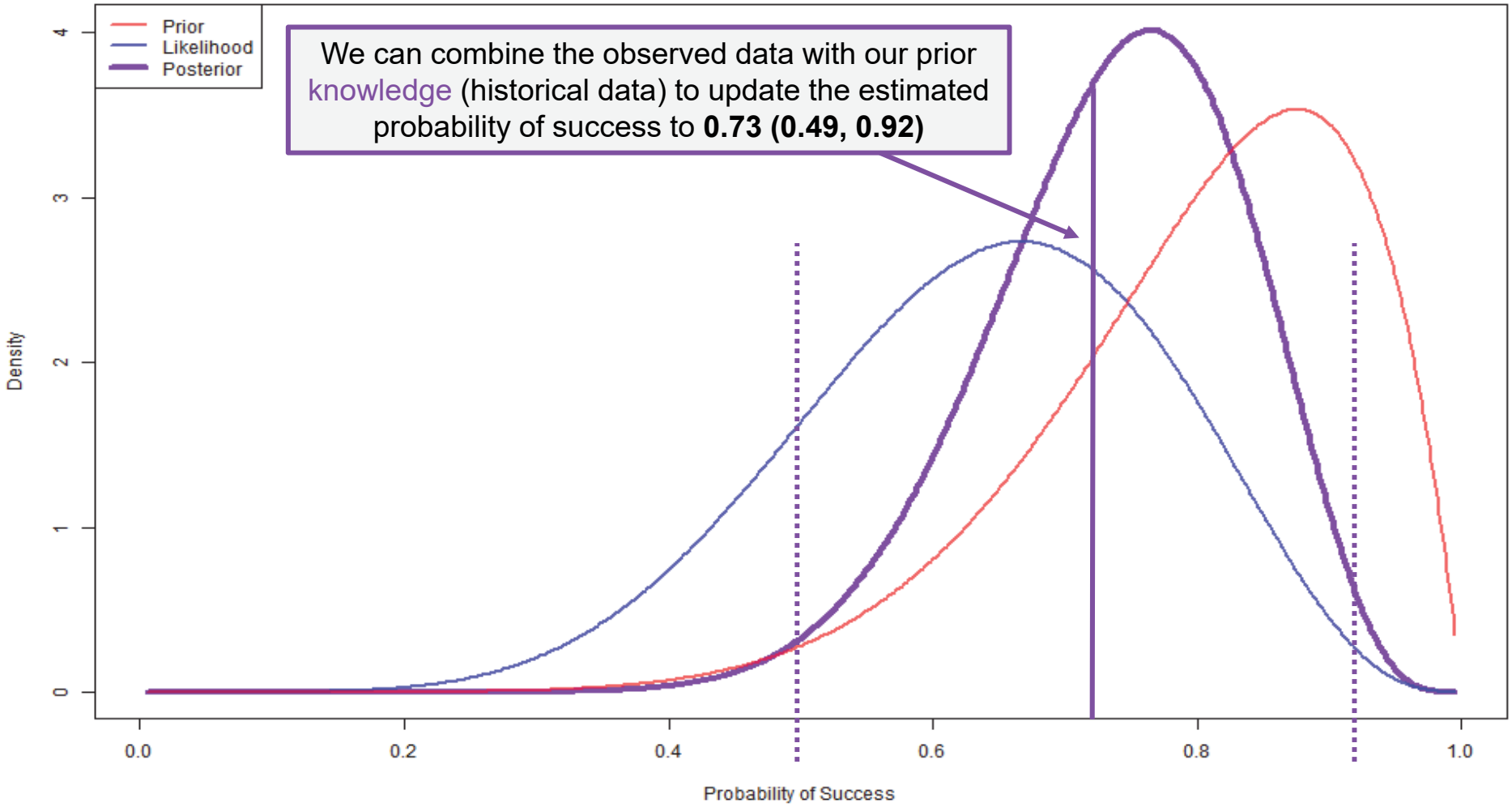
Bayesian analysis of missile launch system: Likelihood



Bayesian analysis of missile launch system: Prior



Bayesian analysis of missile launch system: Posterior



Tool to compute credible intervals and obtain plot is from https://test-science.shinyapps.io/BayesianBinomialCIs/

Approved for public release; distribution is unlimited.

Smaller prior effective sample size puts less weight on the prior data

Credible Level:

0 0.95

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Number of Successes:

6

Number of Trials:

9

Choice of Prior Distribution:

- ○ Uniform Prior
- ○ Jeffreys Prior
- ◉ Enter Prior Based on Historical Data

Prior Number of Successes:

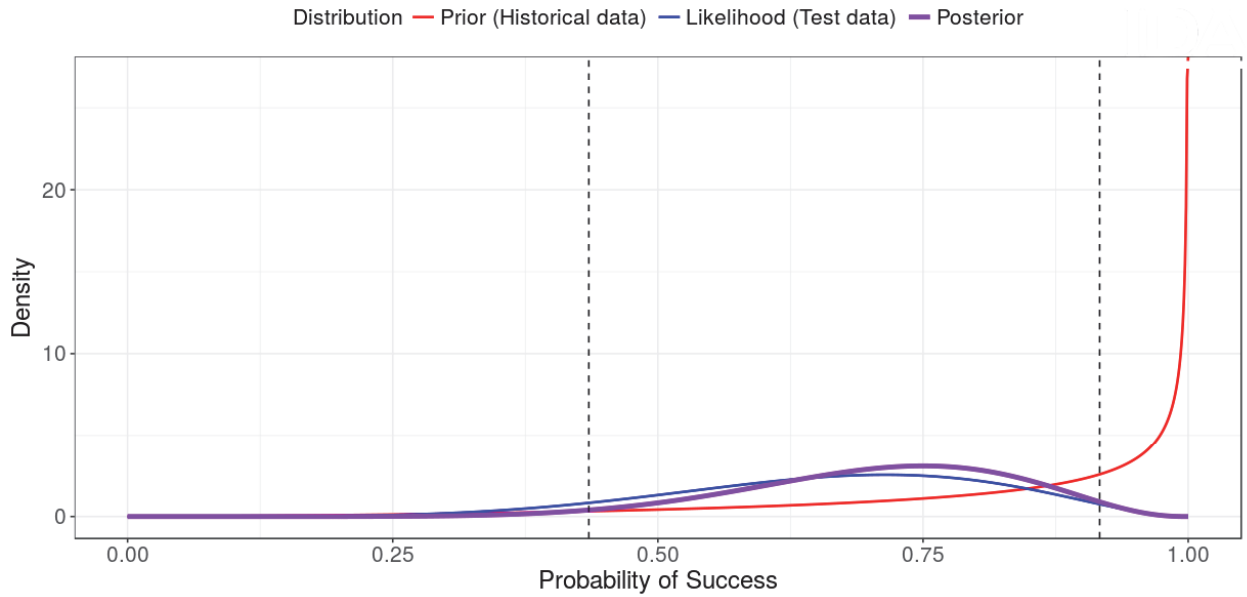
5

Prior Number of Trials:

6

Effective Sample Size of Prior Data:

3



The 95% posterior central interval is [0.435, 0.917]. This means that there is a 95% chance that the true probability of success lies between 0.435 and 0.917. The mean of the posterior distribution is 0.708. The maximum likelihood estimate, representing the point estimate from the test data, is 0.667.

Approved for public release; distribution is unlimited.

Larger prior effective sample size puts more weight on the prior data

Credible Level:

0 0.95

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Number of Successes:

6

Number of Trials:

9

Choice of Prior Distribution:

- ○ Uniform Prior
- ○ Jeffreys Prior
- ◉ Enter Prior Based on Historical Data

Prior Number of Successes:

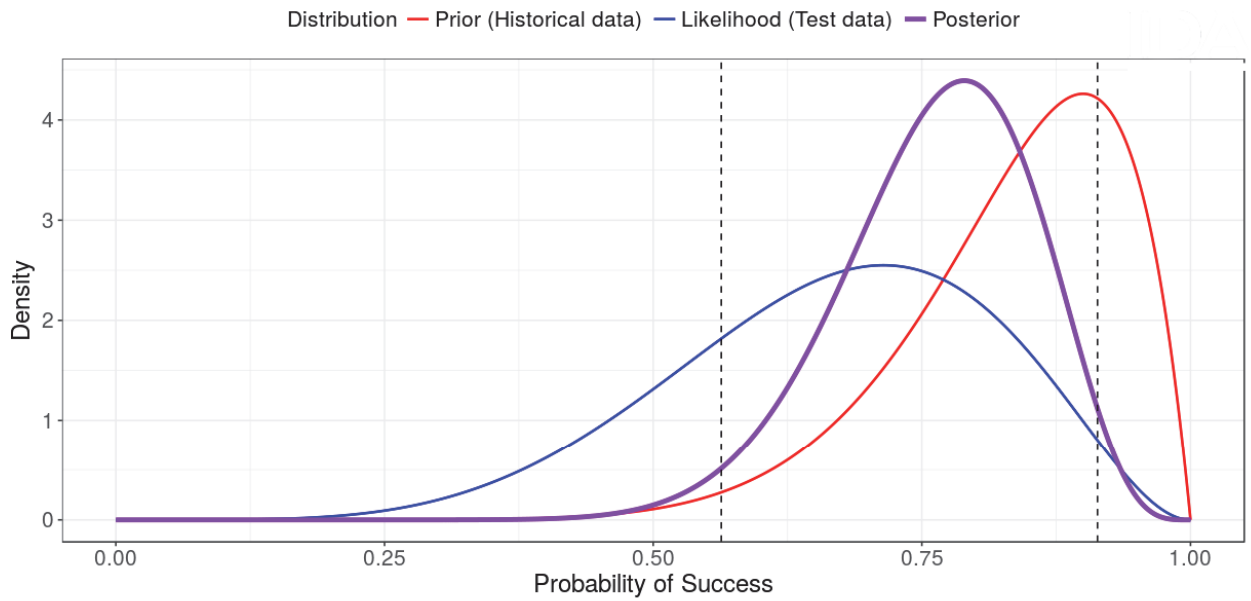
5

Prior Number of Trials:

6

Effective Sample Size of Prior Data:

12



The 95% posterior central interval is [0.563, 0.913]. This means that there is a 95% chance that the true probability of success lies between 0.563 and 0.913. The mean of the posterior distribution is 0.762. The maximum likelihood estimate, representing the point estimate from the test data, is 0.667.

Approved for public release; distribution is unlimited.

A prior is always a choice, and there is no best prior for any real data analysis

Non-informative (Vague) Priors

- A non-informative prior is a prior that provides little information relative to the experiment*
- Useful for real data analysis, because they remove most subjectivity

Informative Priors

- Informative priors contain our beliefs before looking at the data
- Most useful for sequential analysis and regression
- Sensitivity analysis recommended

Weakly Informative Priors

- Some benefits from both sides
- Common practice

Default recommendation

* Box and Tiao (1973)

Approved for public release; distribution is unlimited.

Bayesian updating: We can summarize the data at the end of the test, or after any run

We could collect n data points, analyze the data, and use the posterior distribution of the analysis as the prior for the next data point(s). That is:

$$P(\theta|y_1, \ldots, y_n) \propto P(y_n|\theta) \times P(\theta|y_1, \ldots, y_{n-1})$$

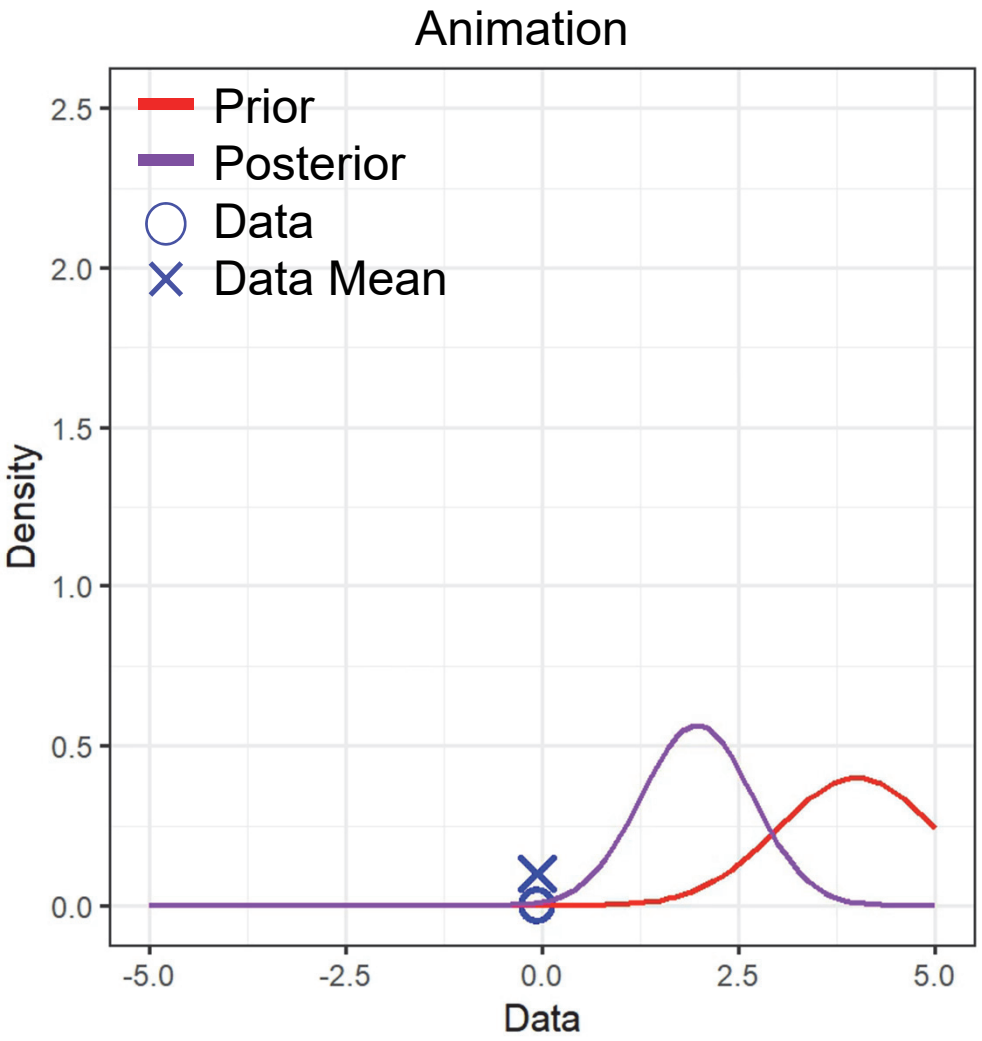
$$\propto P(y_n|\theta) \times P(y_{n-1}|\theta) \times P(\theta|y_1, \ldots, y_{n-2})$$

$$\propto P(y_n|\theta) \times P(y_{n-1}|\theta) \times \cdots \times P(y_1|\theta)P(\theta)$$

This means that in sequential analysis, there is no difference between analyzing the data in chunks versus all at once.

Approved for public release; distribution is unlimited.

Bayesian updating in action

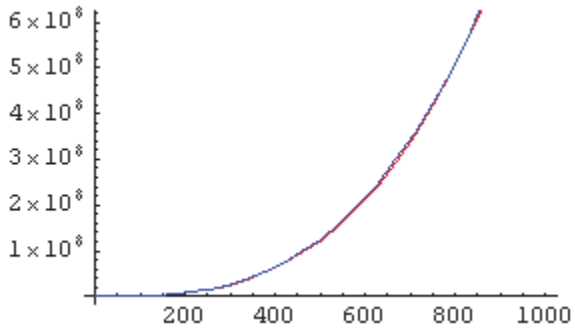


- As new data arrive, the posterior distribution can be updated for each observation
- The prior does not match the data, but the data eventually overpower it
- This is Bayesian "sequential learning"

Approved for public release; distribution is unlimited.

Why use Bayesian statistics?

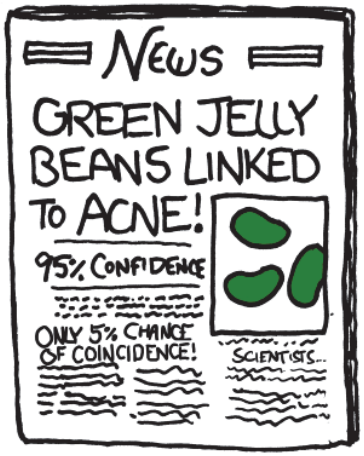
Coherent and widely applicable method



No need for asymptotic assumptions to justify methods

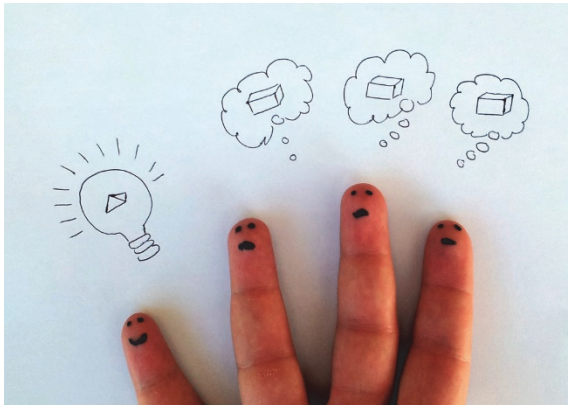
Works in cases when frequentist methods fail

Why use Bayesian statistics?



Bayesian results typically not affected by multiple comparisons

Tests can be stopped once results are conclusive!



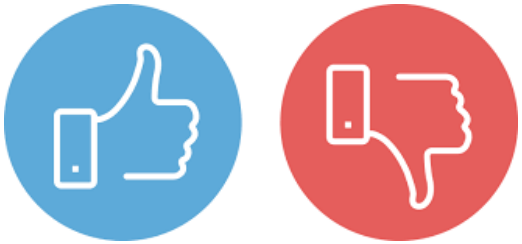
Interpretations are easy

Approved for public release; distribution is unlimited.

Bayesian statistics also has disadvantages

Can be computationally intensive (especially for complicated models)

Learning curve might be steep



Priors can be criticized

```
parameters {
  real beta0;
  real beta1;
  real<lower=0> sigma2;
}
model {
  // priors
  beta0 ~ normal(0, 30);
  beta1 ~ normal(0, 15);
  sigma2 ~ inv_gamma(0.1, 0.1);
  // Likelihood
  time_spent ~ normal(beta0 + beta1 * dista, sqrt(sigma2));
}
```

Frequentist methods are available in standard software, whereas some Bayesian analyses need to be coded from scratch

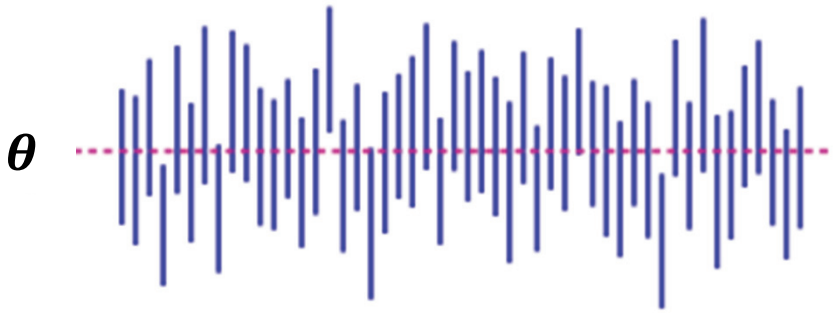
Approved for public release; distribution is unlimited.

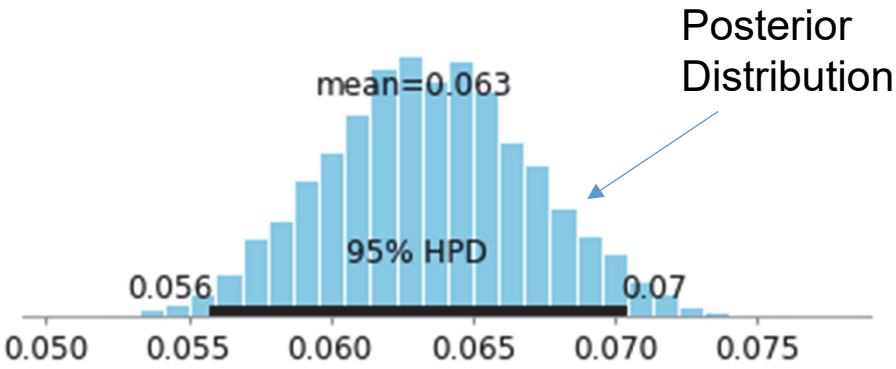
Frequentist vs. Bayesian statistics

	Frequentist	Bayesian
Probabilities are:	Long-term frequencies	Degrees of belief
Inference based on:	Sampling distribution	Posterior distribution
Parameters are:	Fixed (but unknown)	Random
Intervals are:	Random	Fixed
Modeling goal:	Maximize likelihood (typically)	Estimate *entire* posterior distribution

Approved for public release; distribution is unlimited.

Confidence and credible intervals express a range of plausible values for a parameter or effect





Confidence (Frequentist) Interval:

Under repeated sampling, a 95% confidence interval will cover the parameter θ 95% of the time

Credible (Bayesian) Interval:

A 95% credible interval contains the parameter θ with probability 95%

HPD = Highest Posterior Density

Approved for public release; distribution is unlimited.

When and how to use Bayesian statistics?

It depends…

Test Science thinks analysts should be pragmatic and weigh the advantages and disadvantages of Bayesian analysis. But here's some rules of thumb:

- Consider using frequentist methods when the problem can be solved with standard methods and there is no prior information
- Consider using Bayesian approaches when there is good prior information available, or when a frequentist approach doesn't cut it

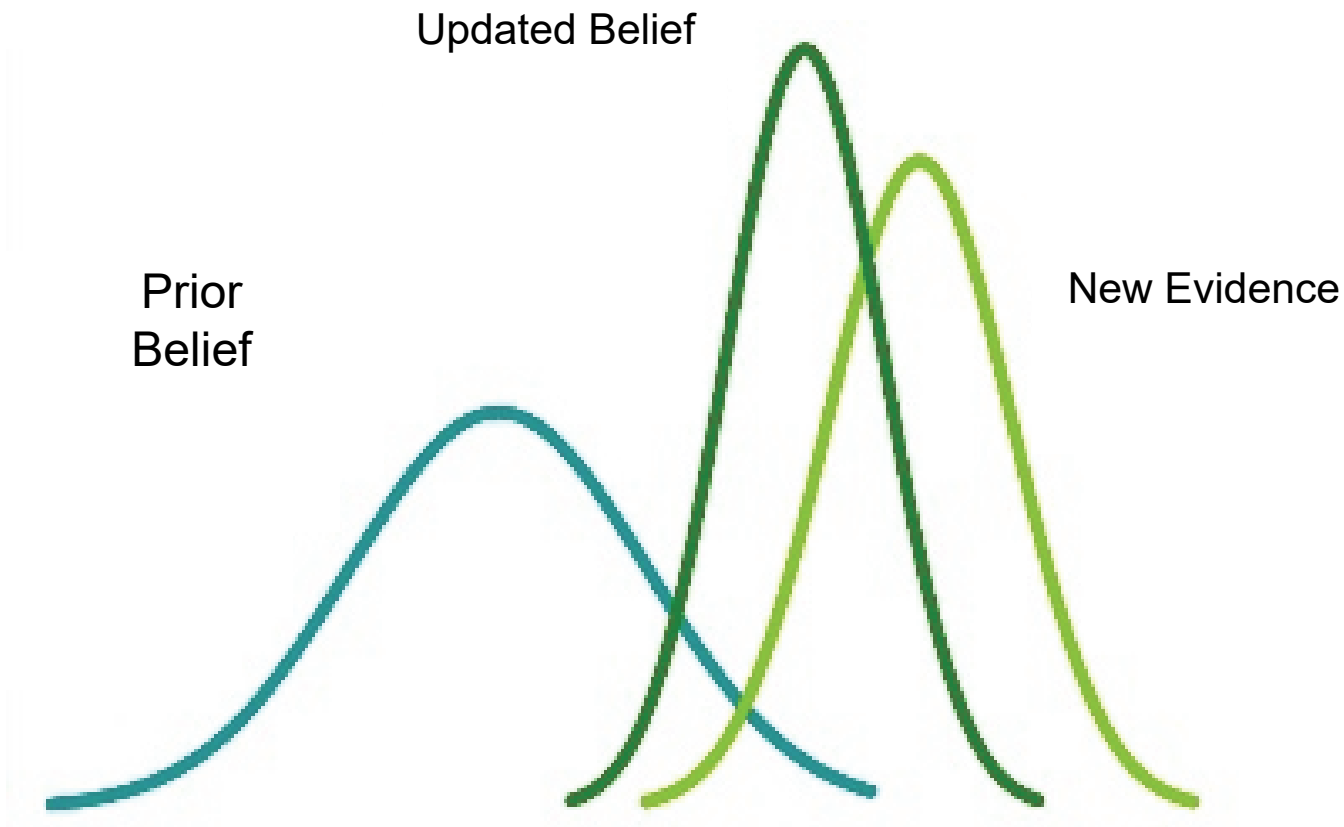
Approved for public release; distribution is unlimited.

Bayesian analysis is a different (but principled) way of analyzing data that offers numerous advantages to both the researcher and the decision-maker.

Approved for public release; distribution is unlimited.

Introduction to Bayesian Analysis

Section II – Single-Parameter Models



Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

"If I'm doing an experiment to save the world, I better use my prior." – Andrew Gelman

Approved for public release; distribution is unlimited.

Conjugate Priors

Approved for public release; distribution is unlimited.

Conjugate priors guarantee posterior is same distribution as the prior

- Determined by the data likelihood
- Justified by invariance reasoning
- Highly tractable

Toy Example:

- $P(y|\mu) = Normal(\mu, 1)$ — Likelihood
- $P(\mu) = Normal(\mu_0, 1)$ — Conjugate prior
- $P(\mu|y) = Normal\left(\frac{\mu_0+y}{2}, \frac{1}{2}\right)$ — Posterior is the same "shape" as the prior!

Approved for public release; distribution is unlimited.

This table shows common conjugate models

Likelihood	Parameter	Prior	Posterior
$Binomial(\boldsymbol{n}, \boldsymbol{\theta})$	$0 \leq \theta \leq 1$	$Beta(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$Beta(\alpha', \beta')$ $\alpha' = \alpha + y$ $\beta' = \beta + n - y$
$Poisson(\lambda)$	$\lambda > 0$	$Gamma(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$Gamma(\alpha', \beta')$ $\alpha' = \alpha + n$ $\beta' = \beta + \sum t$
$Exponential(\lambda)$	$\lambda > 0$	$Gamma(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$Gamma(\alpha', \beta')$ $\alpha' = \alpha + n$ $\beta' = \beta + \sum t$

This Wikipedia page includes more examples of conjugate distributions: https://en.wikipedia.org/wiki/Conjugate_prior

Approved for public release; distribution is unlimited.

Incorporating Legacy Data into Evaluation: Beta-Binomial

Approved for public release; distribution is unlimited.

Conjugate priors make incorporating prior information straightforward

- Remember the missile example, where the outcome of interest was "successful launch" or "failed launch."
- The set of all runs follows a Binomial distribution with $n - y$ failures, y successes, and a probability of success θ.

$$P(y|\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y} \propto \theta^{y}(1-\theta)^{n-y}$$

- The probability of success is a continuous quantity bounded by 0 and 1. Therefore, we could use a Beta distribution as the prior for θ.

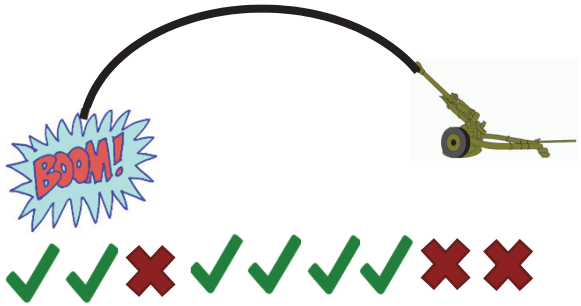
$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- The posterior distribution is then a Beta distribution.

$$P(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} = \theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}$$

Approved for public release; distribution is unlimited.

Estimate the probability of success using the prior data and keep in mind the implications



Legacy system test results:
5 ✓
1 ✗

$$P(y|\theta) = \binom{9}{6}\theta^6(1-\theta)^{9-6} \propto \theta^6(1-\theta)^3$$

$$P(\theta) \propto \theta^{5-1}(1-\theta)^{1-1}$$

$$P(\theta|y) \propto \theta^{11-1}(1-\theta)^{4-1}$$

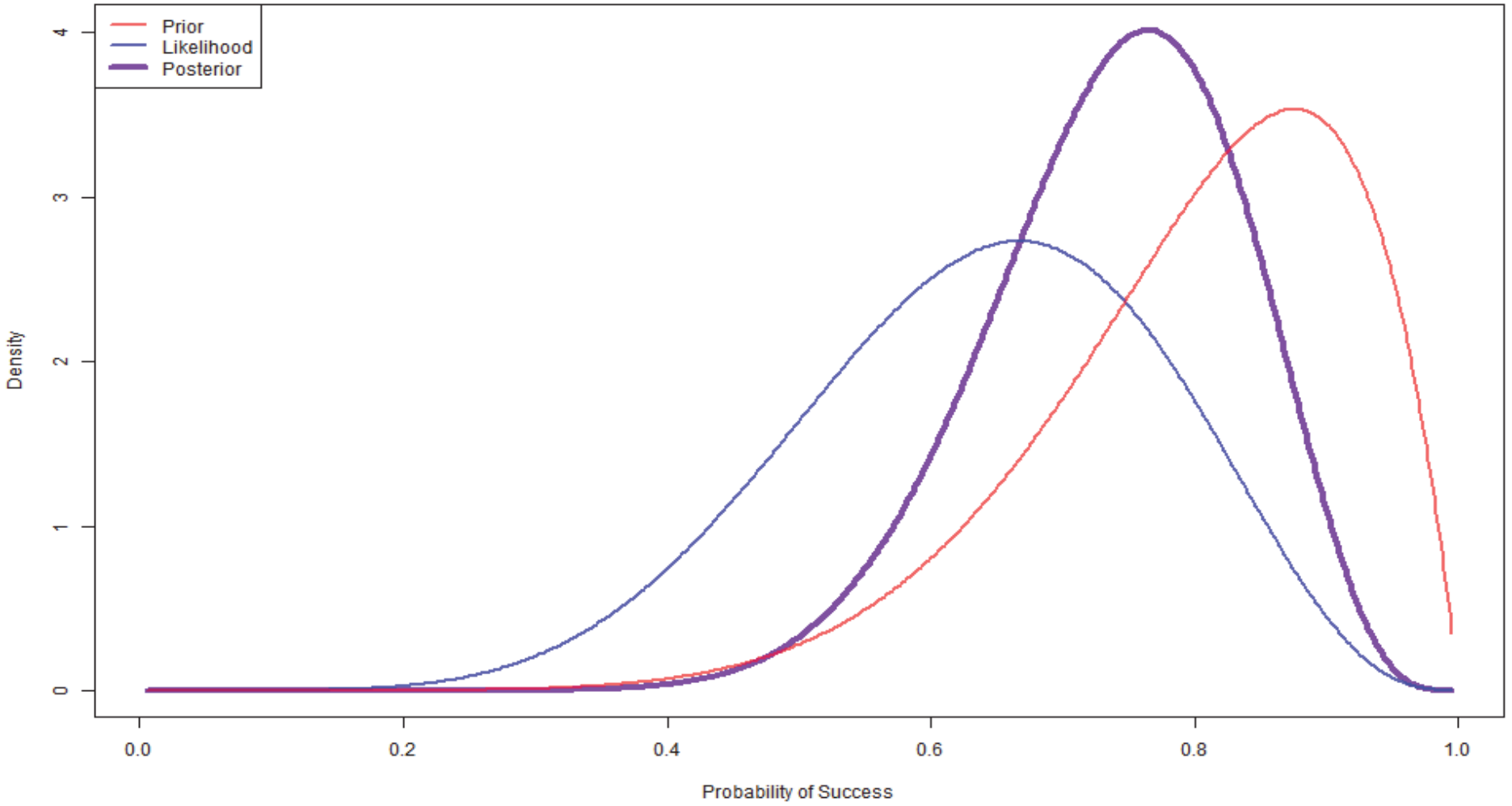
Be careful with the weight on the prior distribution.

Example is from https://testscience.org/characterize-system/test-evaluation-analyses/bayesian-credible-intervals/

Tool to compute credible intervals is from https://test-science.shinyapps.io/BayesianBinomialCIs/

Approved for public release; distribution is unlimited.

Conjugate priors helped us obtain results from a Beta distribution in our missile launch system example



Tool to compute credible intervals and obtain plot is from https://test-science.shinyapps.io/BayesianBinomialCIs/

Approved for public release; distribution is unlimited.

Incorporating all available information leads to less uncertainty in our estimation

Frequentist:

Point estimate is $\hat{\theta} = \frac{y}{n} = \frac{6}{9} = 0.67$, and the confidence interval is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = (0.36, 0.98)^*$$

Confidence Interval Interpretation: *If the test were repeated an infinite number of times and we constructed a confidence interval each time, then 95% of the confidence intervals will contain the true probability of a successful launch* (θ)

Bayesian:

Point estimate and credible interval for $\hat{\theta}$ are computed using the posterior distribution (mean, median, quantiles). The posterior mean and median are 0.73 and 0.74, respectively.

Credible Interval Interpretation: *The probability that* θ *is in the interval of* $(0.49, 0.92)$ *is* 0.95

* Using the central limit theorem. If we use the Wilson Score, the interval is (0.35, 0.98).

Approved for public release; distribution is unlimited.

Do we use the posterior mean or median?

We have the entire posterior distribution, but how do you make an estimate? It depends on how you measure error:

Type of Loss	Loss Equation	Best* Estimate
Squared error loss	$(\theta - \hat{\theta})^2$	Posterior mean
Absolute error loss	$\lvert\theta - \hat{\theta}\rvert$	Posterior median

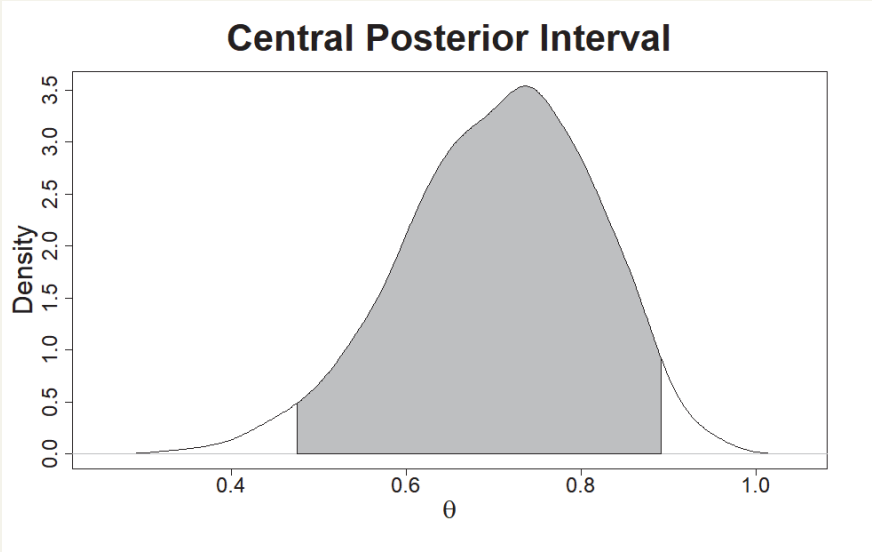
Practical advice: For skewed posterior distributions, we favor the median. For symmetric distributions, the difference is negligible.

* Best in terms of minimizing error.

Approved for public release; distribution is unlimited.

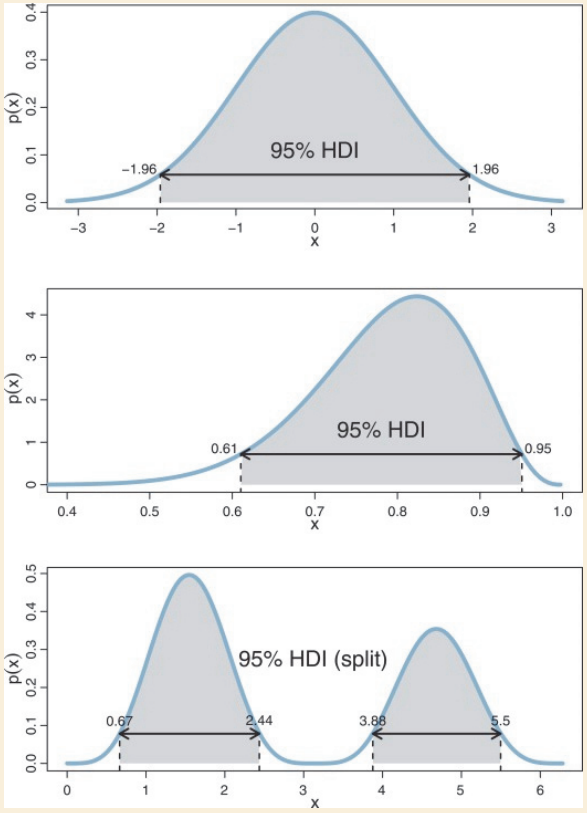
There is more than one way to summarize the posterior uncertainty in the parameter of interest

Central Posterior Intervals



- Compute using density quantiles
- Same probability on each tail
- Invariance property (one-to-one transformations of θ)

Highest Posterior Density (HPD) Intervals



- Density is smaller outside of the interval
- Not necessarily connected
- More accurate than central intervals

Note: For symmetric and unimodal distributions, both intervals will be the same.

HDI image source: Kruschke, John. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2014.

Approved for public release; distribution is unlimited.

Implementation using R

Approved for public release; distribution is unlimited.

https://mc-stan.org/

What is Stan?

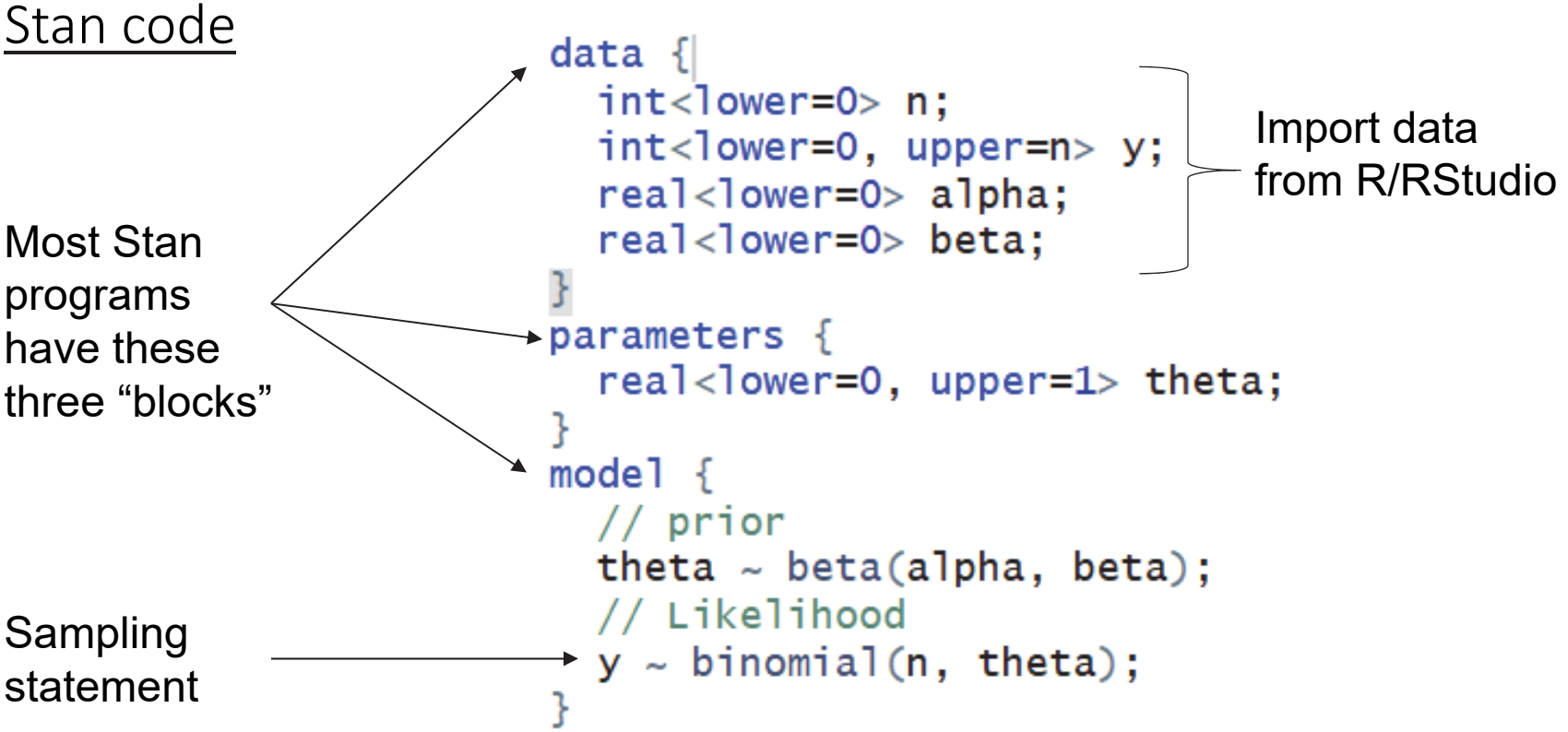
- Stan is a Bayesian model specification (and compilation) language
- Started at Columbia University in 2012
- Open source
- Inferential tasks are generic
- Access to Bayesian statistical models

```
data {
  int<lower=0> n;
  int<lower=0, upper=n> y;
  real<lower=0> alpha;
  real<lower=0> beta;
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  // prior
  theta ~ beta(alpha, beta);
  // Likelihood
  y ~ binomial(n, theta);
}
```


Approved for public release; distribution is unlimited.

We could specify our Beta-Binomial model using Stan

Stan code



We use R to run our model

R code

Data to be imported to Stan

Details for the model fit

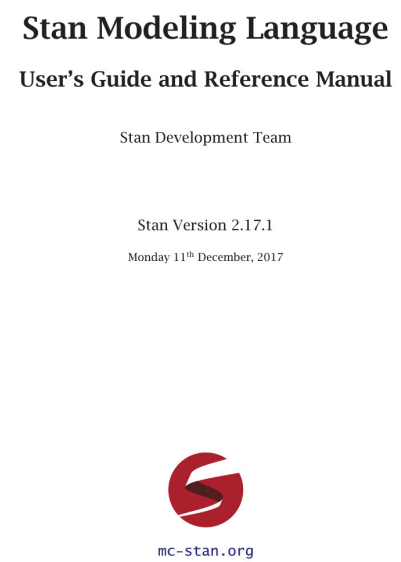
```
library(rstan)
set.seed(1638)
data_bb <- list(n = obs_suc + obs_fail, y = obs_suc,
                alpha = pri_suc, beta = pri_fail)

fit_bb <- stan(file = paste0(path, "R code/beta_binom.stan"),
               data = data_bb,
               pars = "theta", chains = 4, iter = 5000,
               warmup = 1000,
               init = list(list("theta" = 0.01),
                           list("theta" = 0.3),
                           list("theta" = 0.7),
                           list("theta" = 0.99)))
```


Approved for public release; distribution is unlimited.

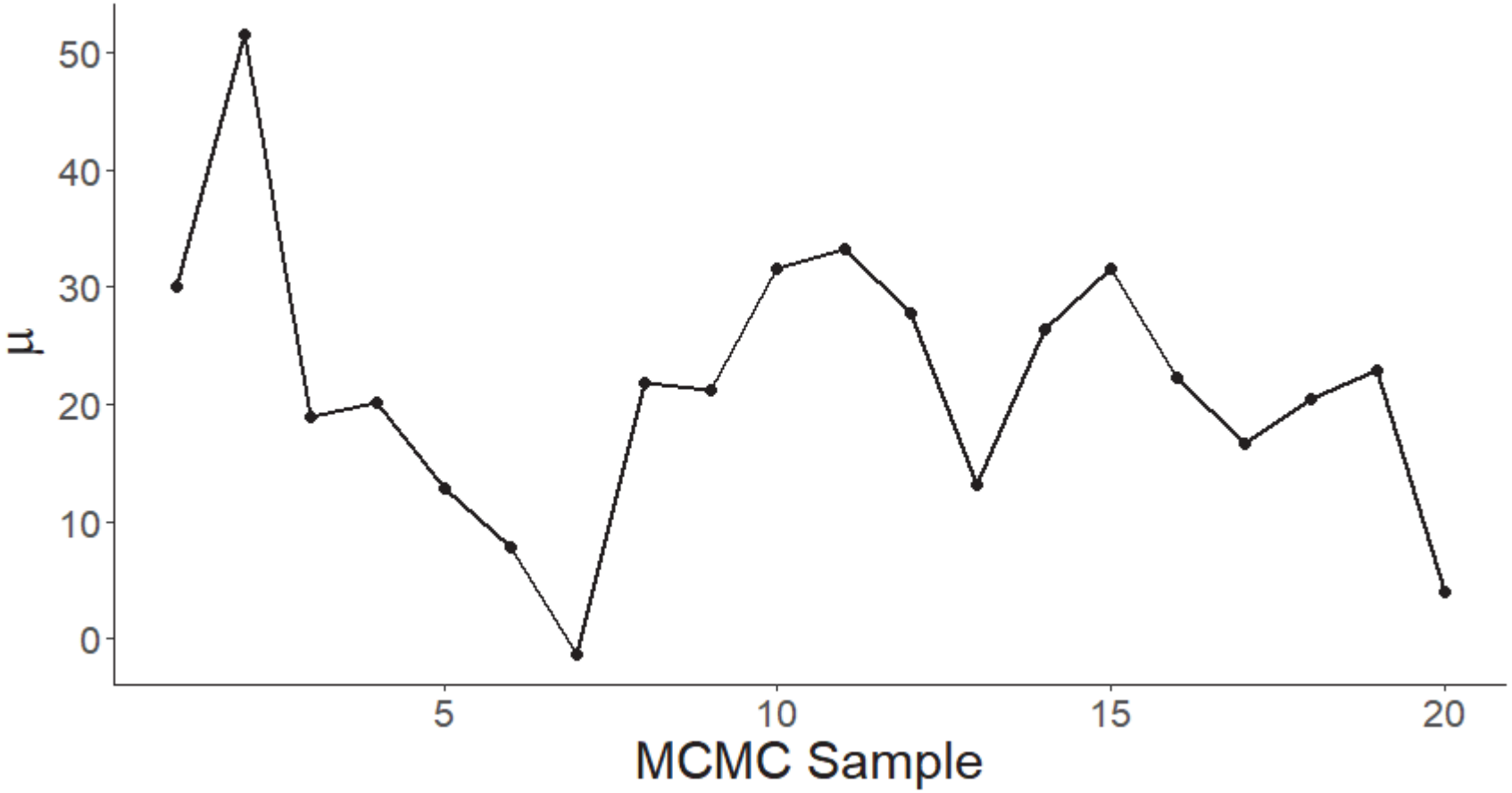
Resources

- Tutorials
- Videos
- Reference manual
- Case studies



Approved for public release; distribution is unlimited.

Markov Chain Monte Carlo is a method that allows us to sample from the posterior distribution



Approved for public release; distribution is unlimited.

First, look at the estimated posterior distribution

Approved for public release; distribution is unlimited.

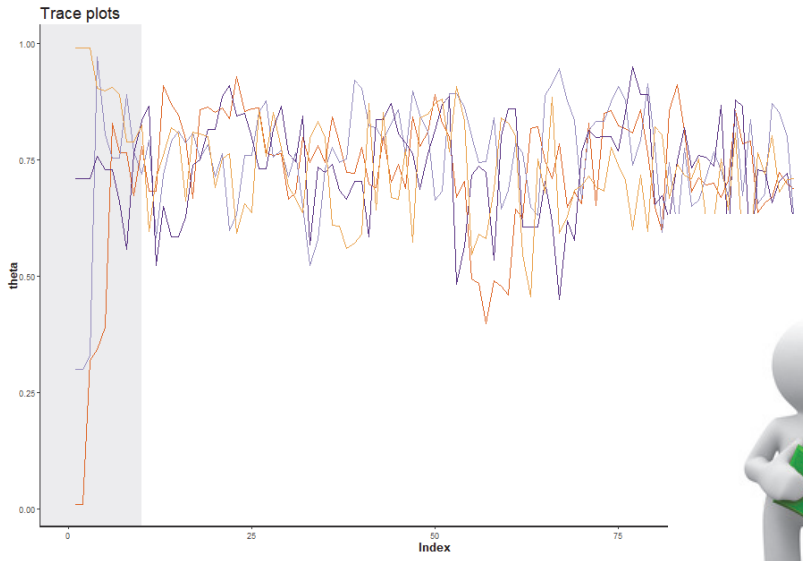
We need to make sure the models are consistent with the data

Convergence Assessment

- Trace plots
- Brooks-Gelman-Rubin

Efficiency Assessment

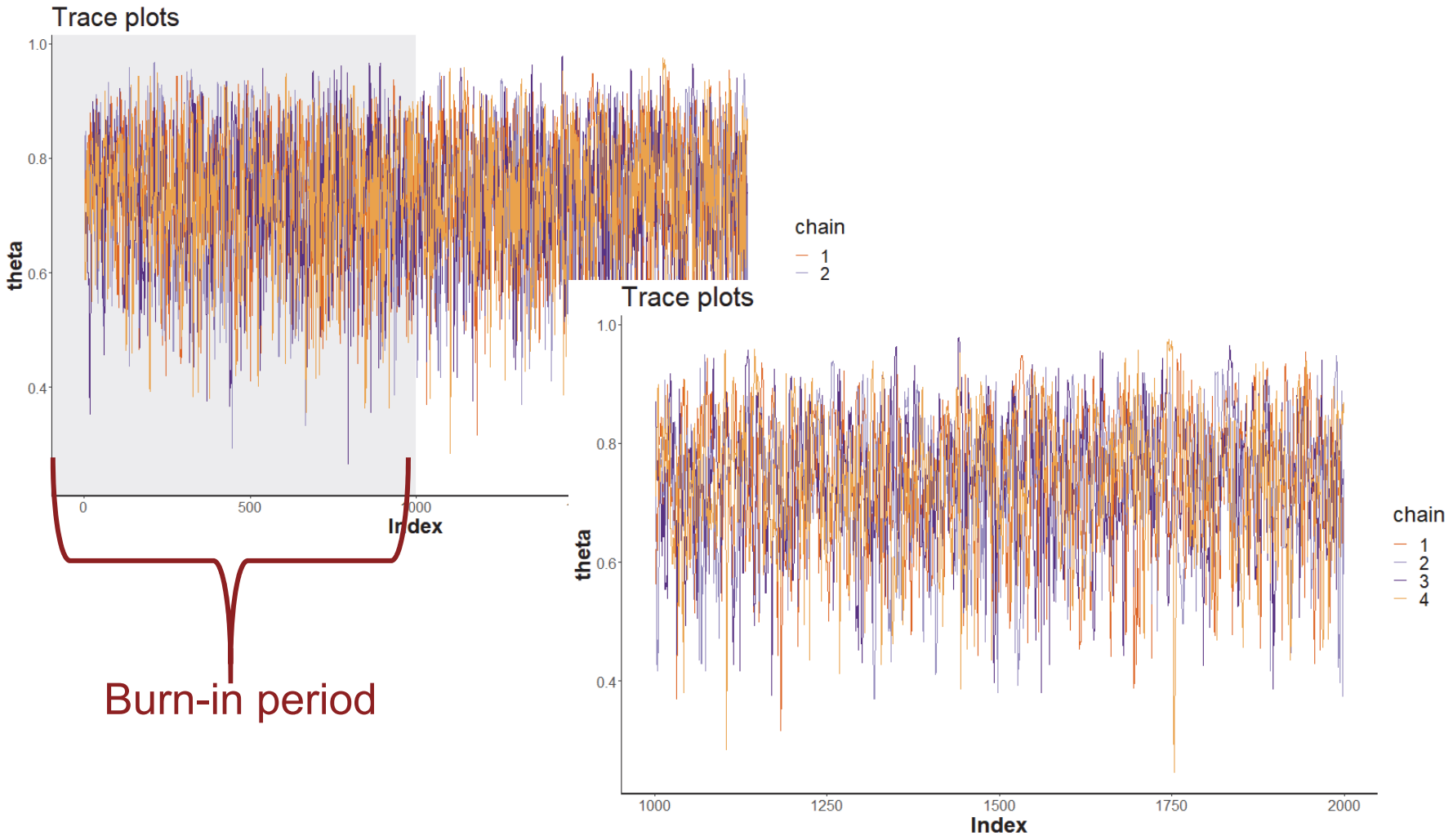
- Markov Chain errors
- Effective sample size



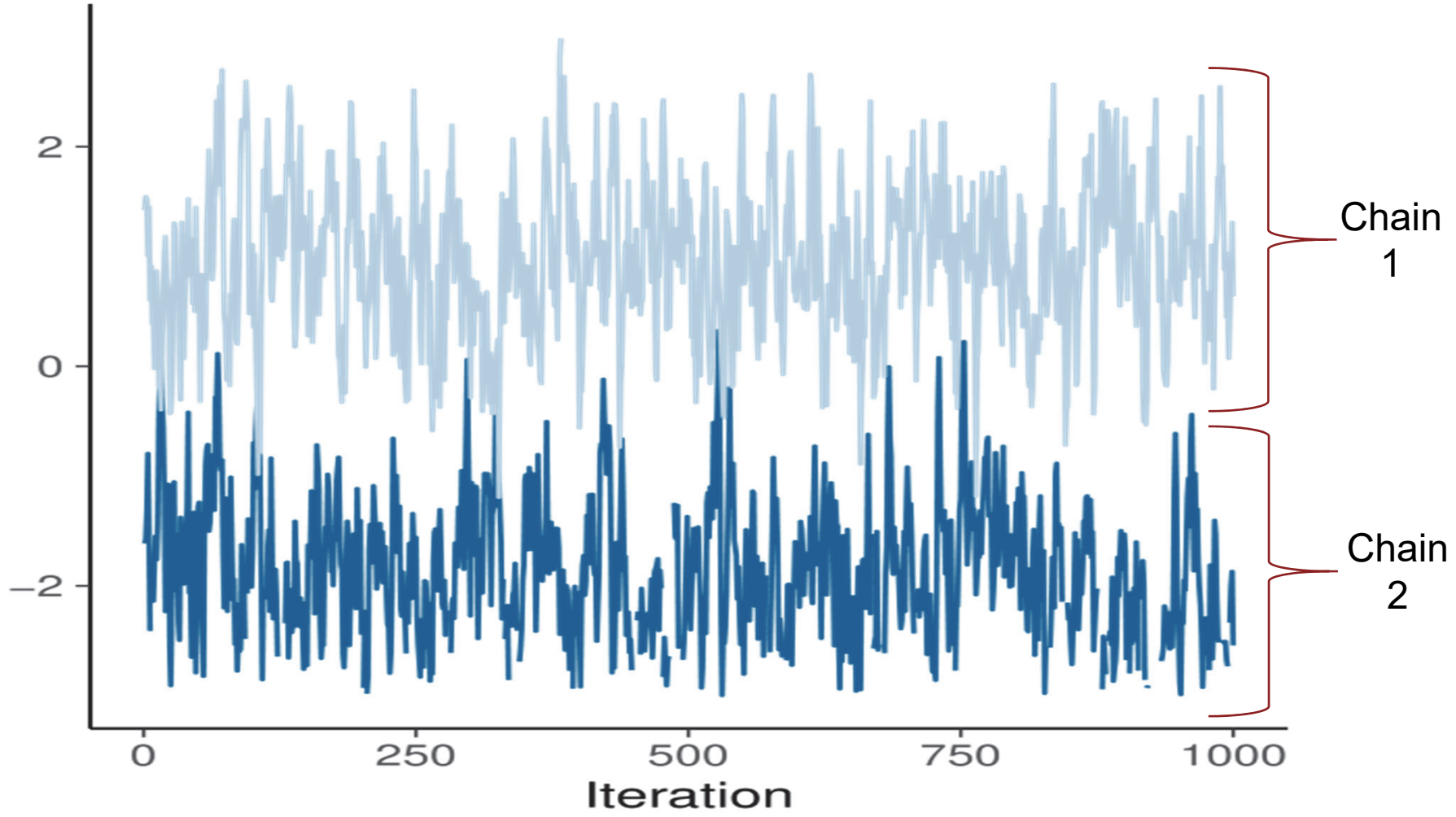
Note that these are just a few examples of simple descriptive statistics and plots. (Quick check, no inferential methods.)

Approved for public release; distribution is unlimited.

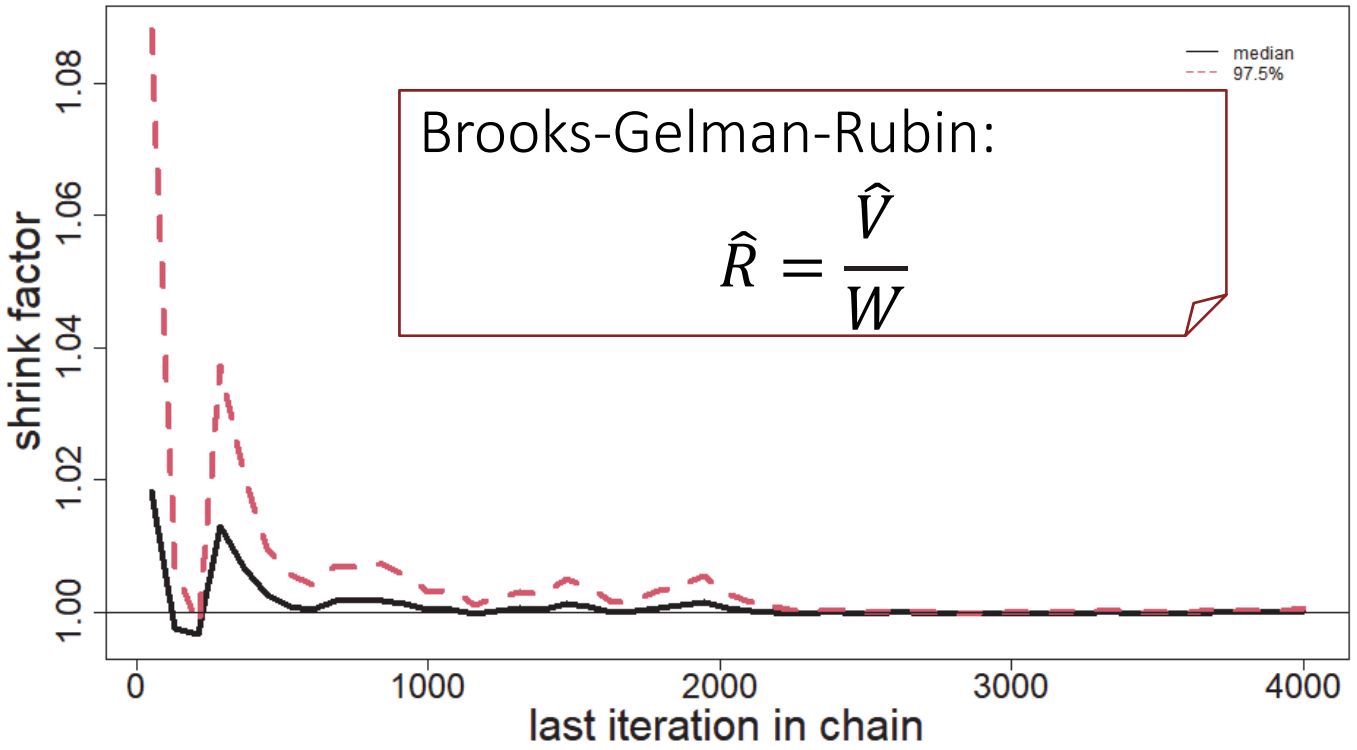
We can perform a qualitative assessment of the convergence by looking at the trace plots



We want to avoid trace plots where the chains do not overlap or mix well



There are also quantitative methods for convergence assessment



Brooks-Gelman-Rubin:

$$\hat{R} = \frac{\hat{V}}{W}$$

$\hat{R}$ compares the variability within each chain's output to the variability of the pooled samples of all chains.

One way to assess efficiency of an MCMC sampler is by computing the posterior effective sample size

```
Inference for the input samples (4 chains: each with iter=4000; warmup=0):

        mean se_mean    sd    2.5%     50%  97.5% n_eff  Rhat
theta  0.731   0.001 0.111   0.489   0.742   0.916  5974 1.001
lp__  -9.226   0.010 0.764 -11.326 -8.936 -8.699  6089 1.000

For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

$$n_{eff} = \frac{ML}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

The posterior effective sample size of the chain is the number of independent MCMC samples.

MCMC = Markov Chain Monte Carlo; n_eff = effective sample size; sd = standard deviation; se_mean = standard error of the mean

Approved for public release; distribution is unlimited.

Another way to assess the performance of an MCMC sampler is by computing the MCSE

- MCSE is a measure of computer simulation error
  - If we had performed infinitely many runs, the simulation error would be 0
  - In simple models, it is very easy to get MCSE near 0
- Measures the amount of uncertainty in the posterior mean

$$MCSE(\bar{\theta}) = \frac{s}{\sqrt{L}}$$

- For non-independent samples, use the posterior effective sample size instead of the number of iterations (L)

MCMC = Markov Chain Monte Carlo; MCSE = Monte Carlo Standard Error

Approved for public release; distribution is unlimited.

Implementation using R and Stan

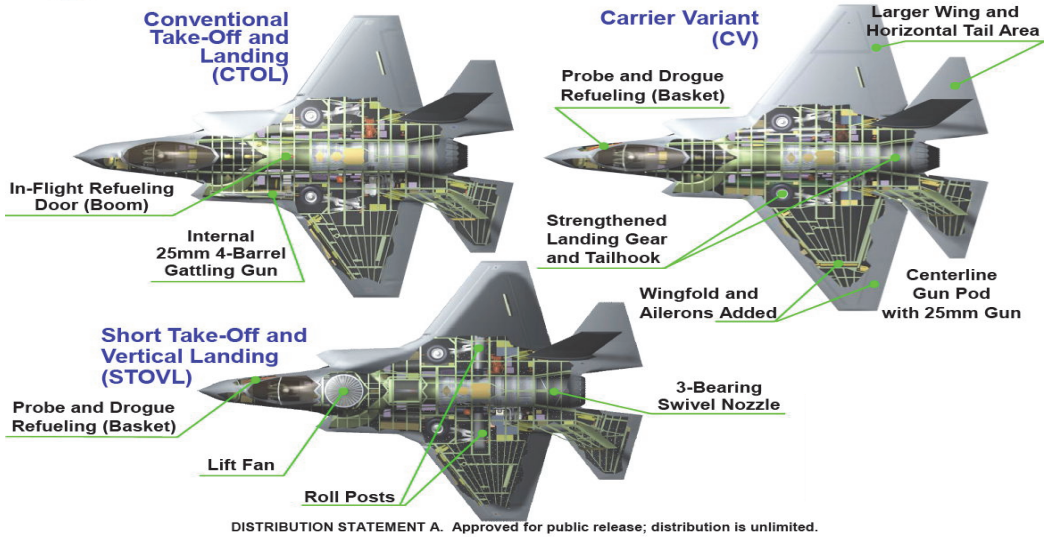
Approved for public release; distribution is unlimited.

Incorporating Engineers' Intuition into Evaluation: Exponential-Gamma

Approved for public release; distribution is unlimited.

There are ways to incorporate a subject matter expert's knowledge into our analyses

- The F-35 cost per flying hour estimate is affected by fleet reliability (for example, maintenance costs)
- At the beginning of the program, the only available data are engineer estimates of reliability
- Traditional methods compute the mean flight hours between repair (MFHBR) as $\text{MFHBR} = \frac{Total\ Flight\ Hours}{Failures} = \frac{FH}{N}$



https://dataworks.testscience.org/wp-content/uploads/sites/4/formidable/7/BramMedlin_DATAWorks19.pdf

Approved for public release; distribution is unlimited.

Challenges and possible solutions

- Traditional methods struggle with reliability estimation if $n = 0$ (no failures)

- What happens if the engineer estimate is 990, the flight hours to date are 40,000, and there have been 2 observed failures?
  - Should we use the traditional method and estimate MFHBR = 20,000 hours and ignore the subject matter expert's intuition?
  - Average the engineer estimate and the traditional estimate?
  - Put more weight on one source of information?

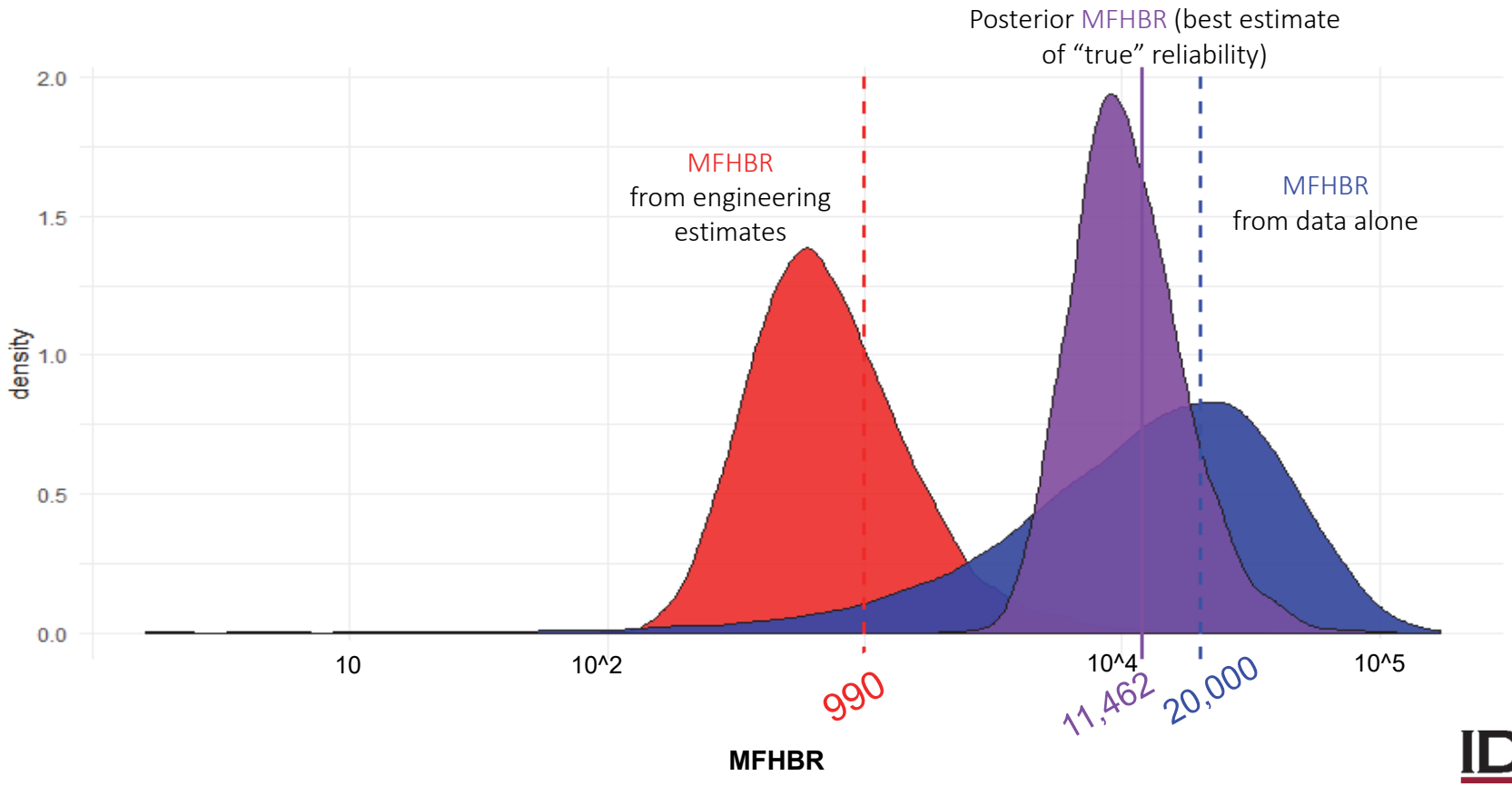
- Bayesian analysis combines the engineering estimates and the actual failure data that are available

Approved for public release; distribution is unlimited.

Bayesian statistics combine "prior" knowledge with observed data to produce an estimate

Example for Component X:

- Engineering Estimate MFHBR = 990 hours (red "prior" below)
- Flight Hours to Date: 40,000 hours
- Observed 2 Failures
- Traditional Methods Estimate: MFHBR = 40,000 / 2 = 20,000 hours (blue "likelihood" below)



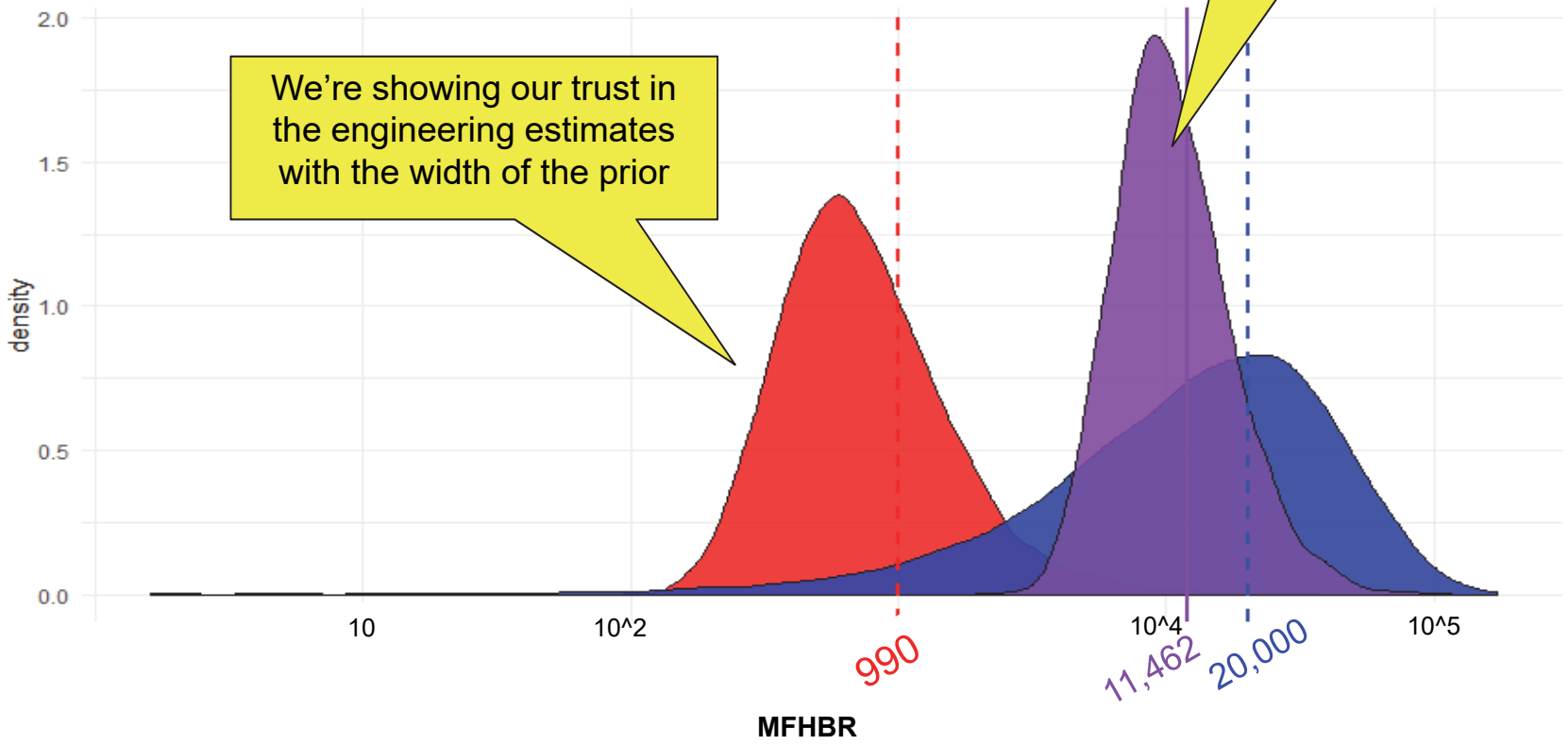
Bayesian statistics combine “prior” knowledge with observed data to produce an estimate

Example for Component X:

- Engineering Estimate MFHBR = 990 hours (red “prior” below)
- Flight Hours to Date: 40,000 hours
- Observed 2 Failures
- Traditional Methods Estimate: MFHBR = 40,000 / 2 = 20,000 hours (blue “likelihood” below)

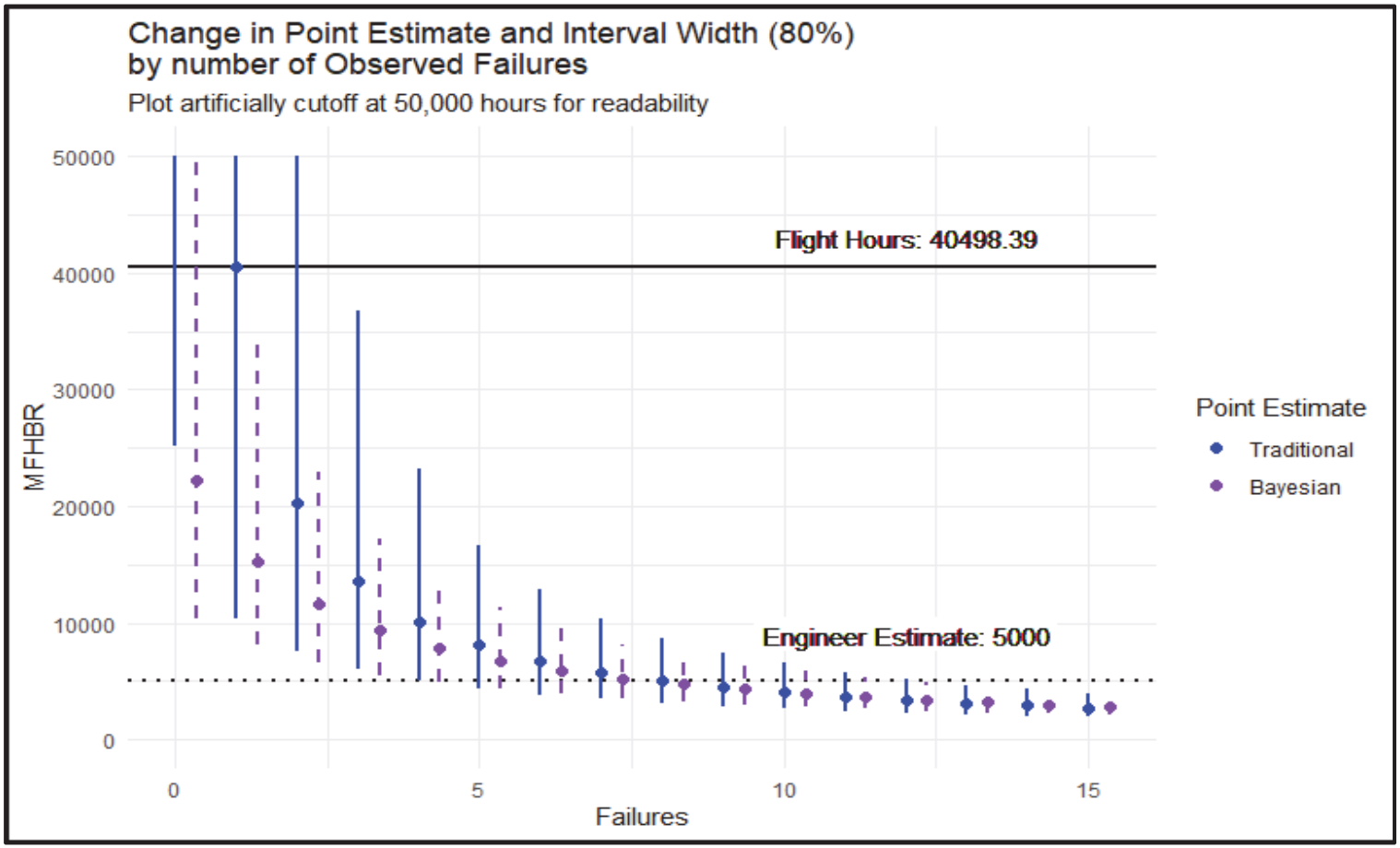
The final estimate is influenced significantly by the subject matter expert estimate because
1. Few data (failures) exist
2. We chose a narrow distribution for the prior

We’re showing our trust in the engineering estimates with the width of the prior



Approved for public release; distribution is unlimited.

A robust methodology for many cases



Approved for public release; distribution is unlimited.

How did we obtain those results?

- Suppose we have observed n independent failure times, say, $t_1, t_2, \ldots t_n$, which follow an Exponential distribution

$$P(\sum t|\lambda) \propto \lambda^n \, Exponential(\lambda \sum t), \qquad t \geq 0, \qquad \lambda > 0$$

where the exponential rate parameter λ is estimated by

$$\frac{n}{\sum_i t_i} = \frac{Total\ number\ failures}{Total\ test\ time} = \frac{1}{MFHBR}$$

- The engineer estimate of the MFHBR is a value greater than 0 and we would like to control the amount of weight we put on the prior. Therefore, we could use a Gamma distribution to represent the prior information.

$$P(\lambda) \propto \lambda^{\alpha-1} \, Exponential(-\beta\lambda)$$

Approved for public release; distribution is unlimited.

Choose the prior parameters based on subject matter expert's inputs

- For λ to have a Gamma prior distribution, the MFHBR must have an inverse Gamma prior distribution

$$P(MFHBR) \propto MFHBR^{-(\alpha+1)}\, Exponential(-\beta/MFHBR)$$

- We can use the engineer estimate as the mean of the prior distribution

$$Mean = \frac{\beta}{\alpha - 1} = MFHBR_{eng\ est}$$

- A similar process gives us an estimate of the prior standard deviation
- Using the prior mean and standard deviation, we can determine parameters for the inverse Gamma prior distribution (α, β)

For more details, see the 2018 IDA memo "Estimating JSF Component Reliability."

Approved for public release; distribution is unlimited.

Obtaining the posterior distribution is not too complicated, and we reduce uncertainty

By combining the failure and time data with the engineer estimates, we obtain the posterior distribution

$$P(\sum t|\lambda) \propto \lambda^n \, Exponential(\lambda \sum t)$$
$$P(\lambda) \propto \lambda^{\alpha-1} \, Exponential(-\beta\lambda)$$
$$P(\lambda|\sum t) \propto \lambda^{\alpha-1+n} \, Exponential\big(\lambda(\sum t + \beta)\big)$$

which is a Gamma distribution with:

$$\alpha' = \alpha + n$$
$$\beta' = \beta + \sum t$$

For our example with 2 failures and 40,000 flight hours we have:

Method	MFHBR Estimate and 80% Interval
Bayesian	10,056 (5,773–19,846)
Frequentist	20,000 (7,516–75,215)

Approved for public release; distribution is unlimited.

We must use the ENTIRE posterior distribution of λ in order to make inference about $MFHBR$

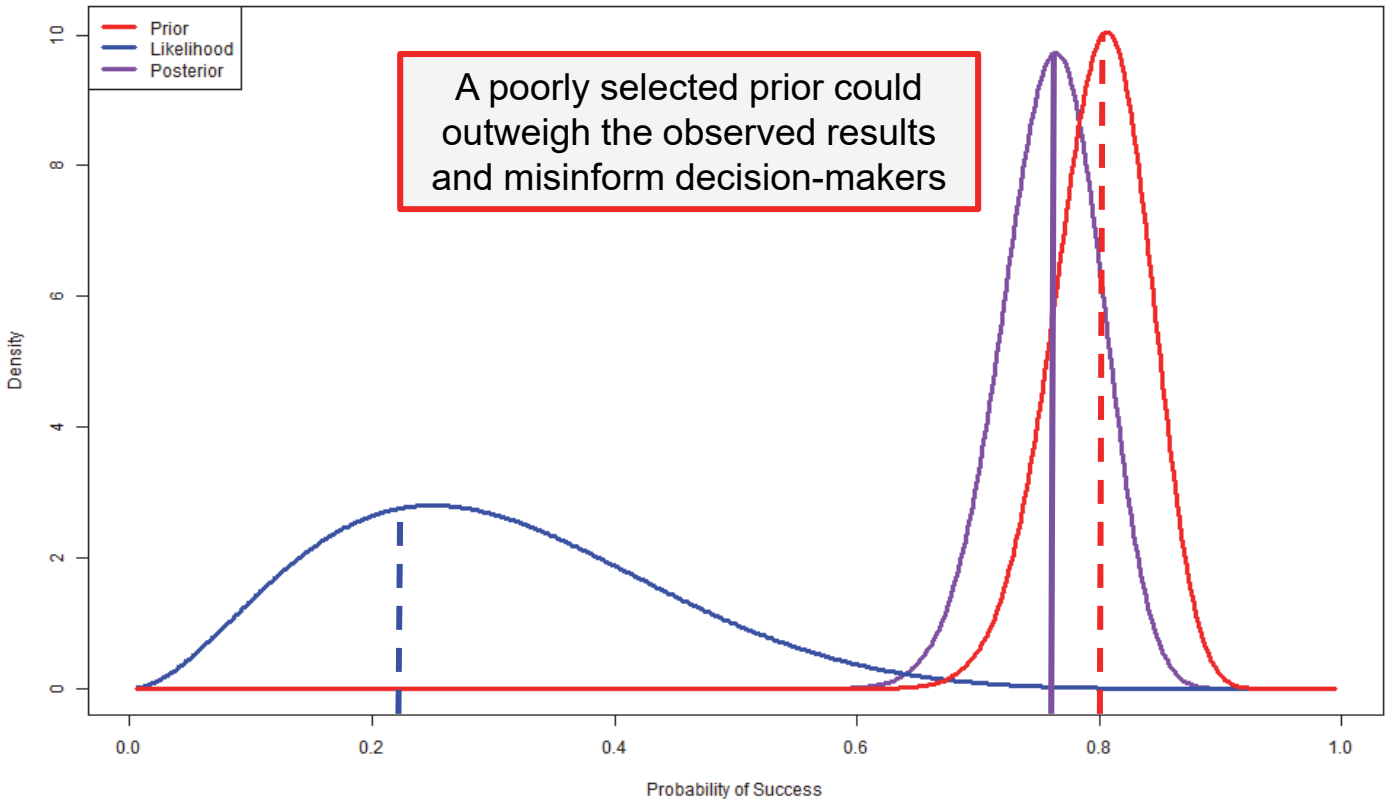
- The posterior samples can be simulated for any transformation $f(\lambda)$ of λ
- In our case this transformation is $\frac{1}{\lambda} = MFHBR$
- First, we simulate samples from the posterior $P(\lambda|\sum t)$
- Then, we transform **each** draw and use those transformed draws to represent draws from $P(MFHBR|\sum t)$
- Note that we are *NOT* obtaining a point estimate $(\hat{\lambda})$ and transforming this point estimate; rather, we are using the *ENTIRE* distribution to obtain our point estimate for $MFHBR$

Approved for public release; distribution is unlimited.

Some notes on prior distributions

Approved for public release; distribution is unlimited.

Prior distribution is the key to Bayesian inference but choosing the prior is not trivial



Requirement = 80% probability of success

Always put careful thought into the prior – naively specified priors can lead to misleading results.

Approved for public release; distribution is unlimited.

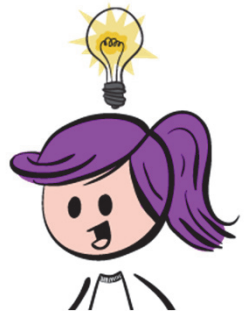
Prior distribution considerations

- Choose a prior based on the available information:
  - Centered at a specific value and with equivalent prior sample size
  - Draw a few distributions and see which one represents your prior beliefs (center, spread/intervals)

- Keep in mind how much your posterior is influenced by the choice of the prior (sensitivity analysis)

- If the prior density is improper, make sure the posterior distribution is proper

- Remember: There is no single "correct" prior in practice!



A proper density is a distribution that integrates to one.

Approved for public release; distribution is unlimited.

What if I don't have any prior information?

- Flat Prior:
  - Works best when parameter range is finite
  - Example: $\pi(\theta) = 1$
- Vague Prior
  - Large variability
  - Example: $\theta \sim Normal(0, 100)$
- Compound Prior
  - Avoid sensitivity to prior variance by assigning a distribution
  - Example: $\theta \sim Normal(0, \sigma_0)$
    $\sigma_0 \sim Half - Cauchy$

Commonly lead to similar results

When no direct information about θ is available, a good approach to prior selection is through the marginal distribution of the data

$$m(y) = \int p(y|\theta)\pi(\theta)d\theta$$

$m(y)$ is what we expect the data to look like

Approved for public release; distribution is unlimited.

Simulating from the marginal distribution can calibrate priors

Unlike conventional models, Bayesian models can generate data before any data have been collected.

We can use the generated data to select priors based on whether the data look like what we expect.

Example:

1. I will use a Normal model to evaluate the test data
2. I need to know if $\mu \sim Normal(2, 2), \sigma \sim Exponential(1)$ is a good prior

We can use simulation to calibrate the priors.

Approved for public release; distribution is unlimited.

Implementation using R

Approved for public release; distribution is unlimited.

Incorporating Legacy Data into Plan: Exponential

Paladin Integrated Management (PIM) self-propelled howitzer

Approved for public release; distribution is unlimited.

We can incorporate legacy data to develop a reliability test plan

We need to determine the test size (hours) for operational testing and the number of failures allowed before we declare whether the PIM howitzer meets or does not meet its 64-hour MTBF reliability threshold. There are two errors we can make:

- Test is passed when the vehicle reliability is actually below threshold (consumer risk)
- Test is failed when the reliability is actually above threshold (producer risk)

Traditional DoD demonstration tests are classical hypothesis tests, which use only data from the current test to assess whether reliability requirements have been met.

- ✓ Fix the risk of passing the test given that the true reliability is less than the threshold (consumer risk)
- ✓ Effectively ignore the risk of failing the test given that the true reliability is greater than a threshold (producer risk)
- ✓ Find the minimum test size around a fixed number of failures
- ✓ Often require an exorbitant amount of testing

Bayesian assurance testing leverages information from various sources in an attempt to reduce the amount of testing required to meet a requirement.[1]

- ✓ Fixes the risk that the true reliability is less than the threshold given that the test is passed (consumer risk)
- ✓ Fixes the risk that the true reliability is greater than a threshold given that the test is failed (producer risk)
- ✓ Finds the minimum test around a fixed number of failures
- ✓ By incorporating all available information, typically requires less testing

1. If the information we are looking to incorporate is poor – say, for example, developmental testing suggests poor system reliability – then incorporating this information will buy us no advantages and will not shorten the length of a test.

DoD = Department of Defense; MTBF = Mean Time Between Failures

Approved for public release; distribution is unlimited.

What is the maximum number of failures, c, permitted for a successful test of length T?

Traditional Risk Criteria

$$Consumer's\ Risk$$

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

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

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

Bayesian Posterior Risk Criteria

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

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

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

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

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

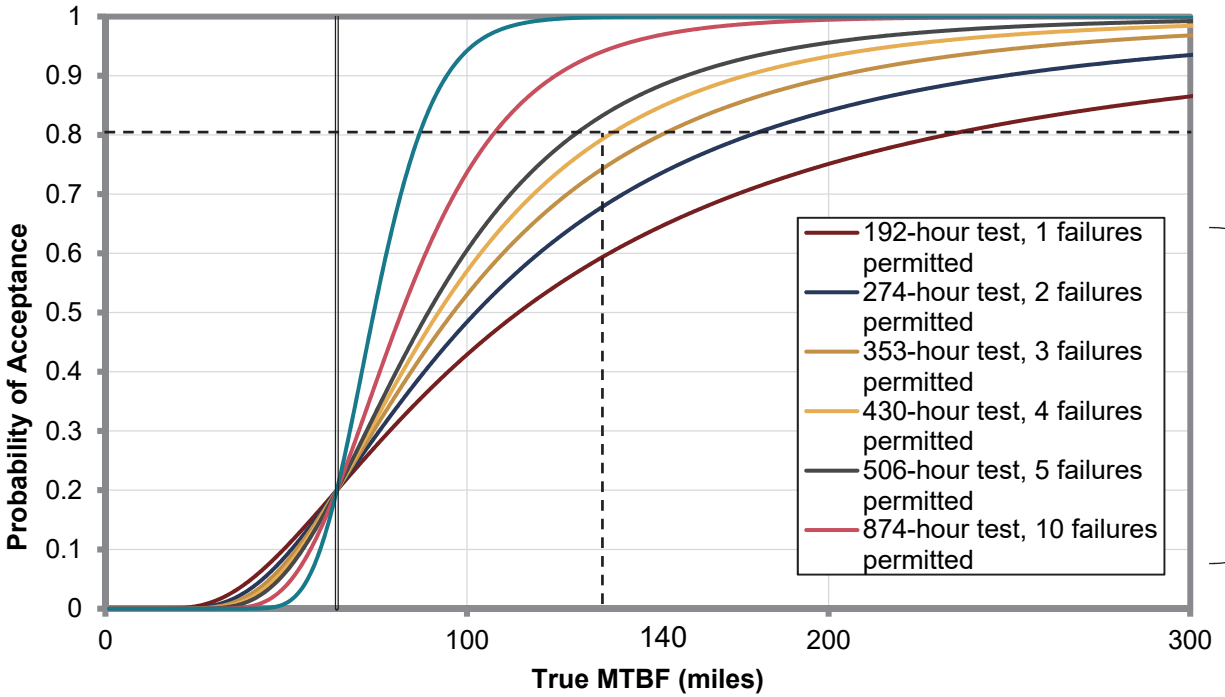
Reference: Michael S. Hamada et al., *Bayesian Reliability,* 2008, Chapter 10.

Approved for public release; distribution is unlimited.

Evaluating a mission-based threshold with traditional methods requires a very long test

Traditional Operating Characteristic Curves are based on demonstrating 64-hour MTBF.

- Using optimistic assumptions about true howitzer reliability, a minimum of 430 hours of operational testing are required to evaluate the howitzer's reliability with $\alpha = 0.2$ and probability of acceptance = 0.80



We need to find the combination of test length, T, and maximum allowed failures, c, that satisfies

$$P(y \le c|\lambda) = \sum_{y=0}^{c} \frac{\left(\frac{T}{64}\right)^{y} \exp\left(-\left(\frac{T}{64}\right)\right)}{y!} \le .2$$

Approved for public release; distribution is unlimited.

We can use available data to construct our test plan

Likelihood Distribution

T = 400 Hours of Testing
N = 2 Failures

Available Data, x, from Previous Test Event

$P(x|\lambda) \sim Exponential(\lambda)$

Prior Distribution

Non-Informative Prior

$P(\lambda) \sim Gamma(\alpha = .001, \beta = .001)$

Posterior Distribution

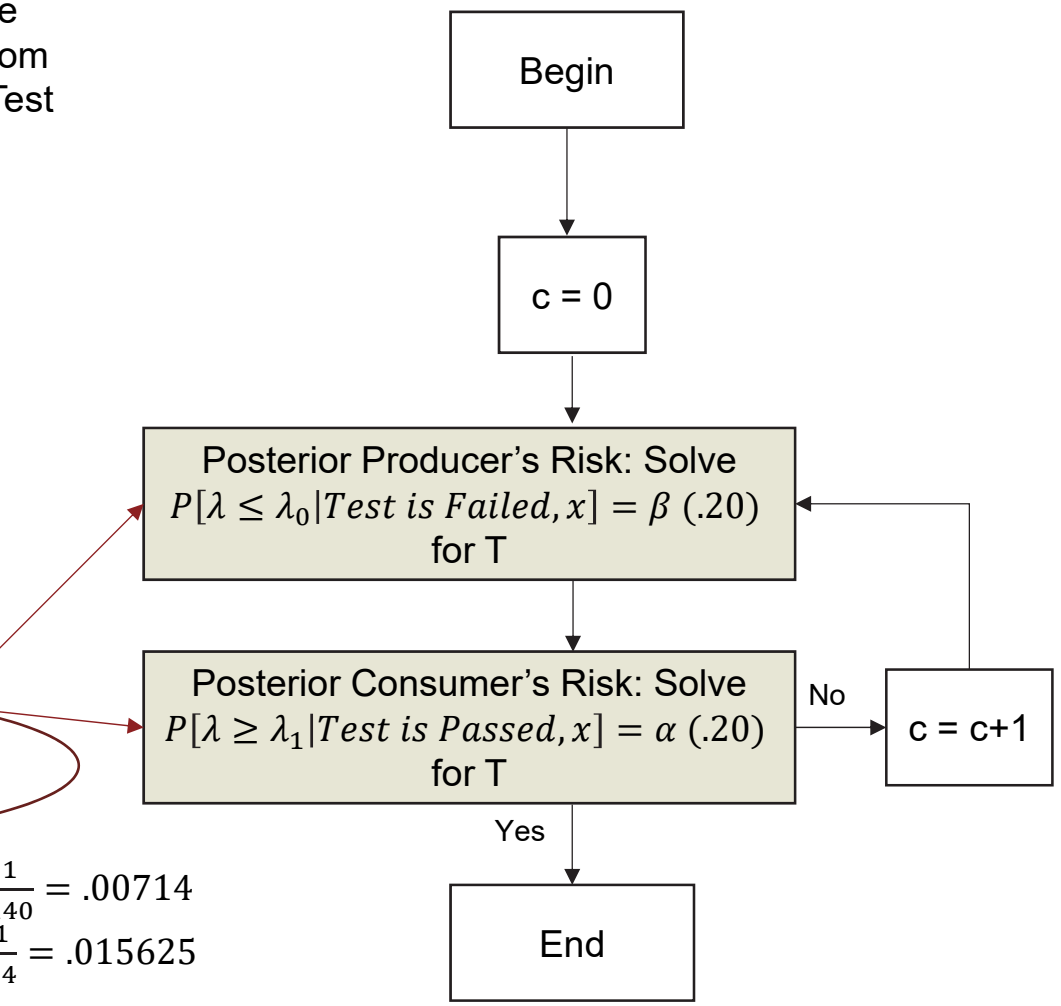
$P(\lambda|x) \sim Gamma(\alpha` = \alpha + N, \beta` = \beta + T)$

Inputs based on growth curve test objective and requirement

$\lambda_0 = \frac{1}{140} = .00714$

$\lambda_1 = \frac{1}{64} = .015625$

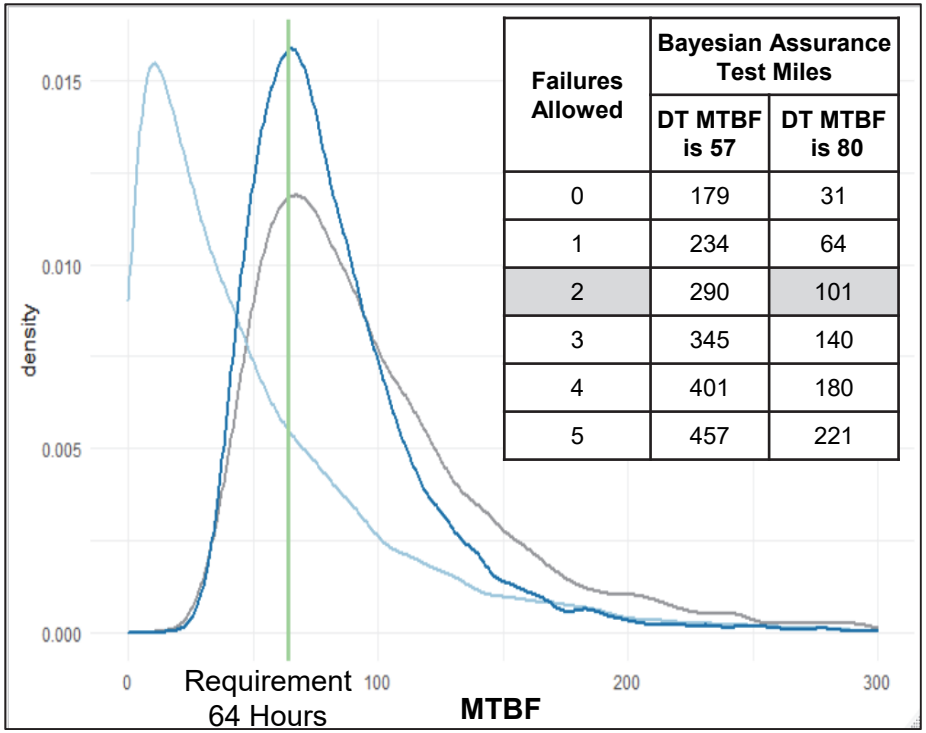
Bayesian Test Plan Algorithm



Approved for public release; distribution is unlimited.

Bayesian assurance testing offers a more powerful and efficient way to assess reliability compared to traditional methods

Bayesian assessment of DT-OT data to assess a 64-mile MMBOMF



Failures Allowed	Bayesian Assurance Test Miles	
	DT MTBF is 57	DT MTBF is 80
0	179	31
1	234	64
2	290	101
3	345	140
4	401	180
5	457	221

DT Reliability — *OT Observed* (101 miles; 2 failures) — OT Predicted Reliability

When **DT and OT testing** are carried out under **similar conditions**, incorporating DT data into the OT reliability assessment may be reasonable.

- **Bayesian assurance methods** provide a structured way to leverage DT
- Similar to Operational Characteristic Curve assumptions, chart on the left assumes **the howitzer** attains a true reliability of 80 hours
- Permits a more powerful assessment of reliability with fewer miles compared to traditional methods that only consider OT data

DT = Developmental Test; MMBOMF = Mean Miles Between Operational Mission Failures; OT = Operational Test

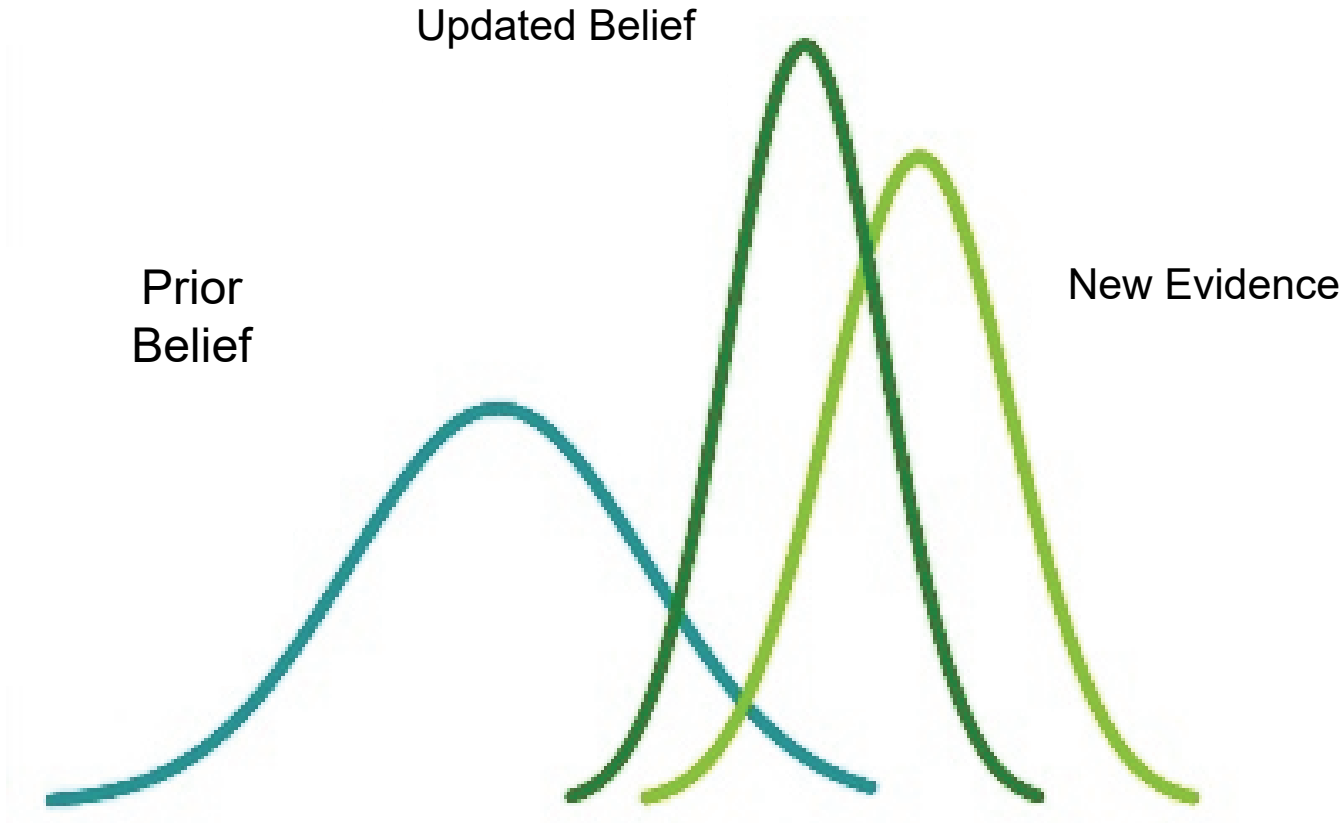
Approved for public release; distribution is unlimited.

Bayesian statistics can provide more informed estimates and can decrease uncertainty, when used properly.

Approved for public release; distribution is unlimited.

Introduction to Bayesian Analysis

Section III – Linear Regression Models



Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

"All models are wrong, but some are useful."
– George Box

Approved for public release; distribution is unlimited.

Simple Linear Regression

Approved for public release; distribution is unlimited.

We can assess the effect of some factor on the outcome of interest

shutterstock.com • 1603321879

Approved for public release; distribution is unlimited.

A common way to analyze the data is to use frequentist statistics

- The regression model we use to analyze the data is

$$mpg = \beta_0 + \beta_1 wt + \varepsilon, \qquad \varepsilon \sim Normal(0, \sigma^2)$$

- Check the model assumptions (stats 101)

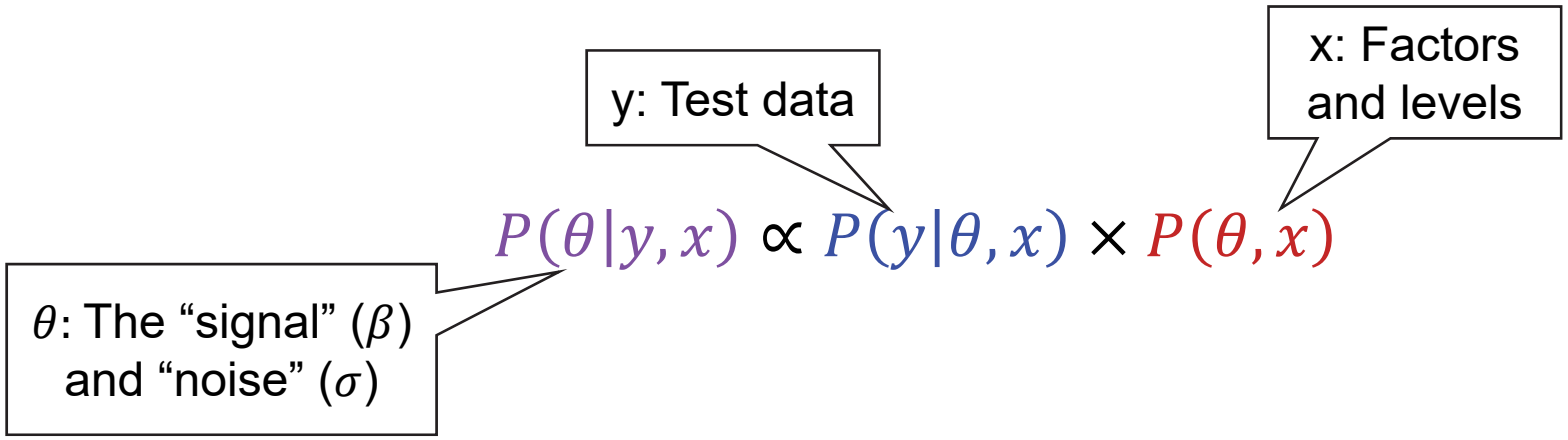
- Obtain parameter estimates and intervals

$\boldsymbol{P(Data|\theta)}$

Observed data and model

Approved for public release; distribution is unlimited.

Bayes' theorem for regression*



Typical assumption: Regression parameters θ do not depend on factors x, so we can simplify

$$P(\theta|y,x) \propto P(y|\theta,x) \times P(\theta)$$

In our example:

$$P(\beta_0,\beta_1,\sigma^2|mpg,wt)$$

$$\propto P(mpg|\beta_0,\beta_1,\sigma^2,wt) \times P(\beta_0) \times P(\beta_1) \times P(\sigma^2)$$

* Note that we add an x to the usual Bayes' theorem because regression depends on covariates.

Approved for public release; distribution is unlimited.

Bayesian analyses use prior distributions to incorporate what we know about the data

- Suppose we have data from a previous test event where slightly different protocols and vehicles were used
- The results from these previous data will help us build the priors for our model
- Increased variability allows us to put less weight on these informative priors

$$\beta_0 \sim Normal\left(\mu_{\beta_0} = 40.8, \sigma_{\beta_0} = 1.63 * 5\right)$$

$$\beta_1 \sim Normal\left(\mu_{\beta_1} = -6.28, \sigma_{\beta_1} = 0.42 * 5\right)$$

$$\sigma^2 \sim Inverse\ Gamma(6.1, 81.08)$$

Approved for public release; distribution is unlimited.

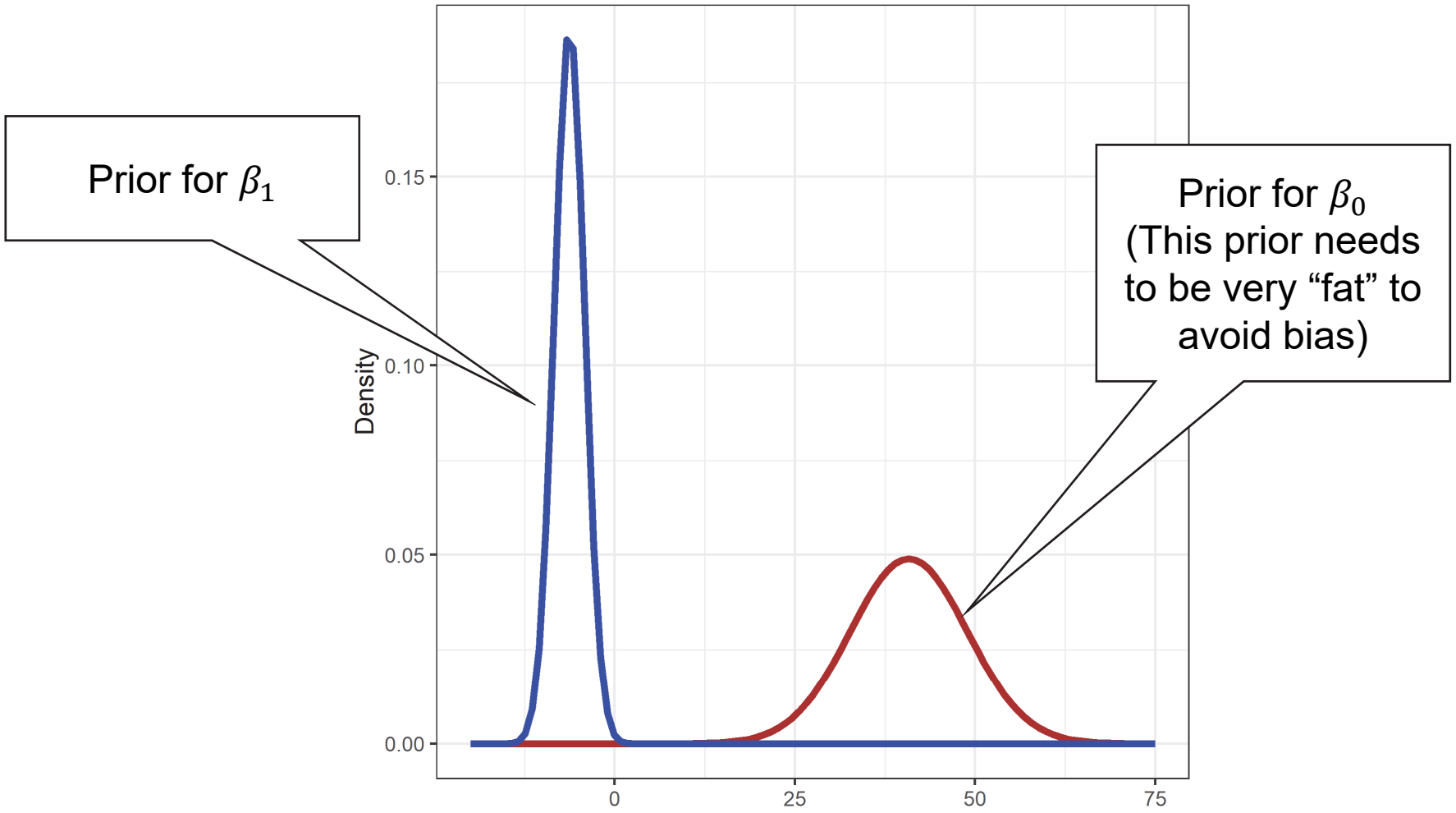
Incorporating prior information in new regression models is not trivial

No general rules for how to do this (I'm sorry …). There might be two types of information:

- Prior information on the regression intercept (β_0)
  - **Be careful:** It could add bias to the overall regression model
- Prior information on the individual factor effects (β_1, β_2, etc.)
  - Less dangerous

Approved for public release; distribution is unlimited.

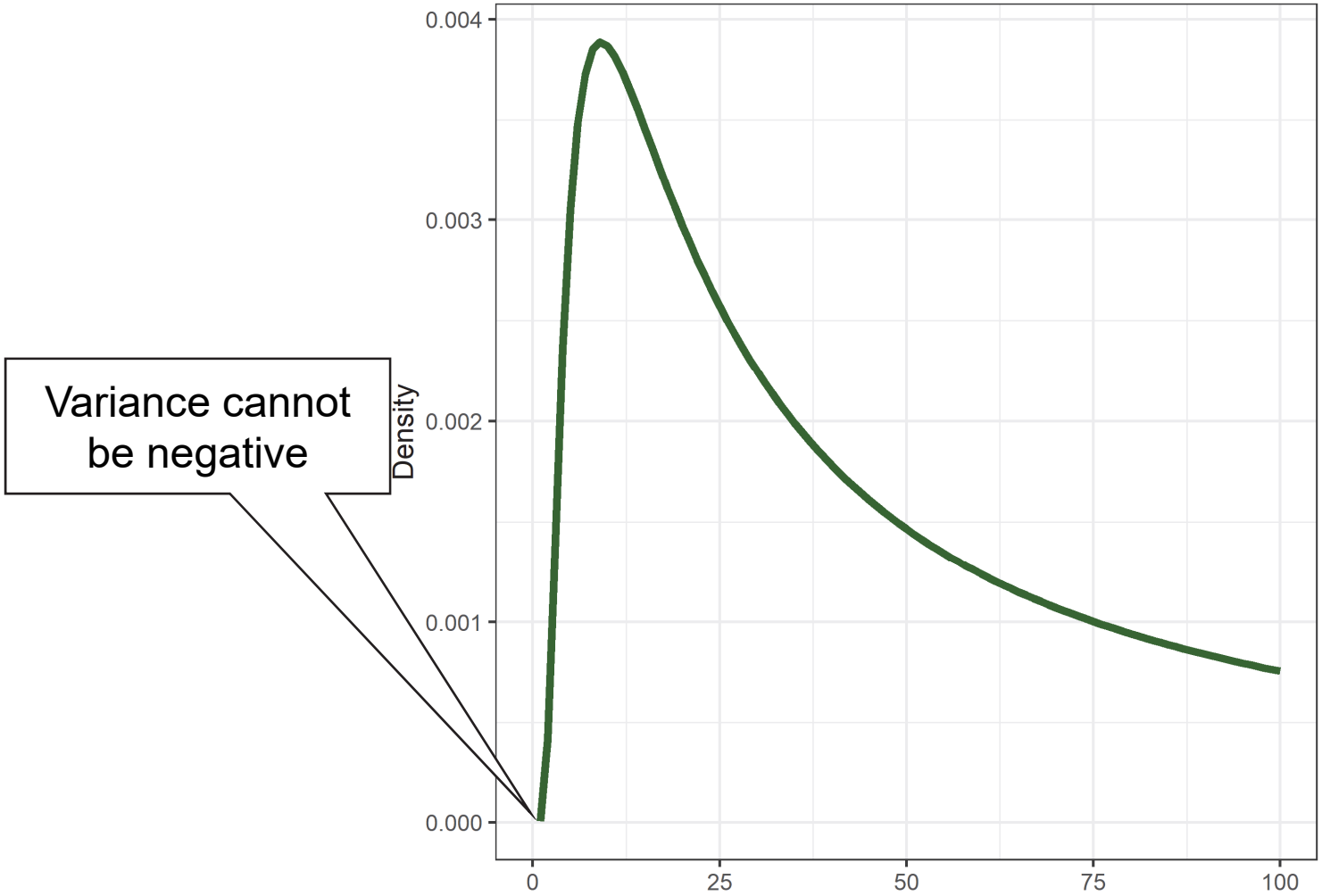
Illustration of the prior distributions for the regression coefficients



Approved for public release; distribution is unlimited.

The prior distribution for σ^2 (residual variance)

Remember: The prior doesn't have to be perfect to be useful!



The prior penalizes very small and very large variance.

Approved for public release; distribution is unlimited.

Complicated posterior distributions cannot be sampled directly!

$$P(\beta_0|\beta_1,\sigma^2,mpg) \propto P(mpg|\beta_0,\beta_1,\sigma^2) \times P(\beta_0) \times P(\beta_1) \times P(\sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{\frac{[mpg_i - (\beta_0 + \beta_1 wt_i)]^2}{2\sigma^2}\right\}$$

$$\times \frac{1}{\sqrt{2\pi 1.63 \times 5}} \exp\left\{\frac{[\beta_0 - 40.8]^2}{2 * (1.63 \times 5)^2}\right\} \times \frac{1}{\sqrt{2\pi 0.42 \times 5}} \exp\left\{\frac{[\beta_1 - (-6.28)]^2}{2 * (0.42 \times 5)^2}\right\}$$

$$\times \frac{81.08^{6.1}}{\Gamma(6.1)} (\sigma^2)^{-(6.1+1)} \exp(-81.08/\sigma^2)$$

$$\propto \exp\left\{\sum \frac{[mpg_i - (\beta_0 + \beta_1 wt_i)]^2}{2\sigma^2}\right\} \times \exp\left\{\frac{\beta_0^2}{132.85}\right\}$$

$P(\theta|Data)$

The posterior distribution for each parameter does not have a closed form (no conjugacy). Therefore, we need to simulate values from the posterior distribution via MCMC.

Note that this is **one** posterior distribution; MCMC = Markov Chain Monte Carlo

Approved for public release; distribution is unlimited.

There are situations in which the results* from both analyses are similar

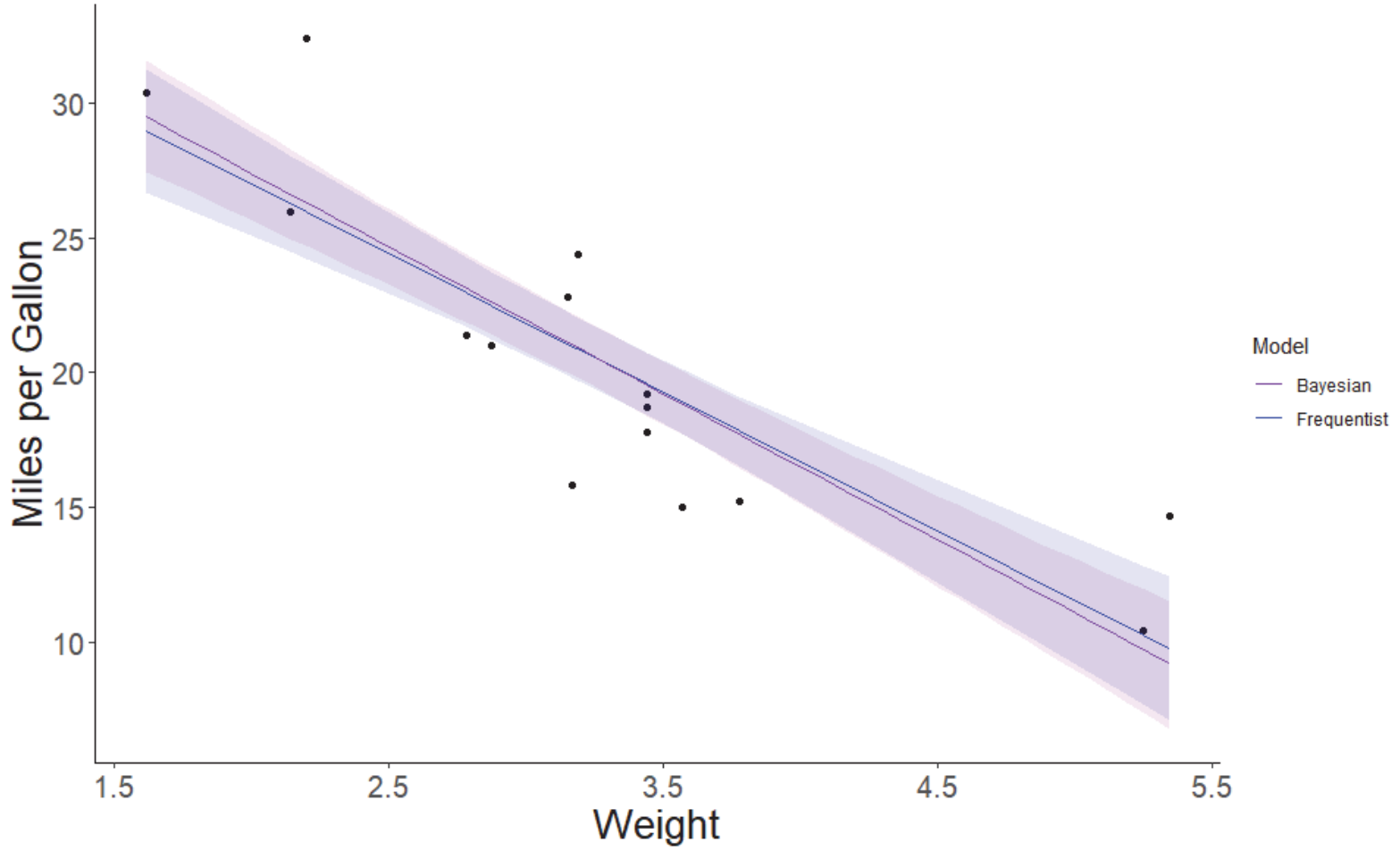
Parameter	Bayesian**	Frequentist
β_0	38.22 (34.64, 41.81)	37.32 (33.27, 41.36)
β_1	-5.42 (-6.45, -4.39)	-5.15 (-6.33, -3.98)

* 80% confidence and credible intervals

** Parameter estimates based on the posterior median

Approved for public release; distribution is unlimited.

We might want to show the results in an operational context



Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Implementation using R

Approved for public release; distribution is unlimited.

Posterior predictive checks

If our model is a good fit, then we should be able to use it to generate data that look a lot like the data we observed.

Approved for public release; distribution is unlimited.

Posterior *predictive* distributions are used to make inferences about data (not parameters)

We know that

$$P(\theta|y,x) \propto P(y|\theta,x) \times P(\theta)$$

But we want $P(y^{rep}|y,x)$, the distribution of new data given current data (assuming we can replicate the experiment with the same factors)

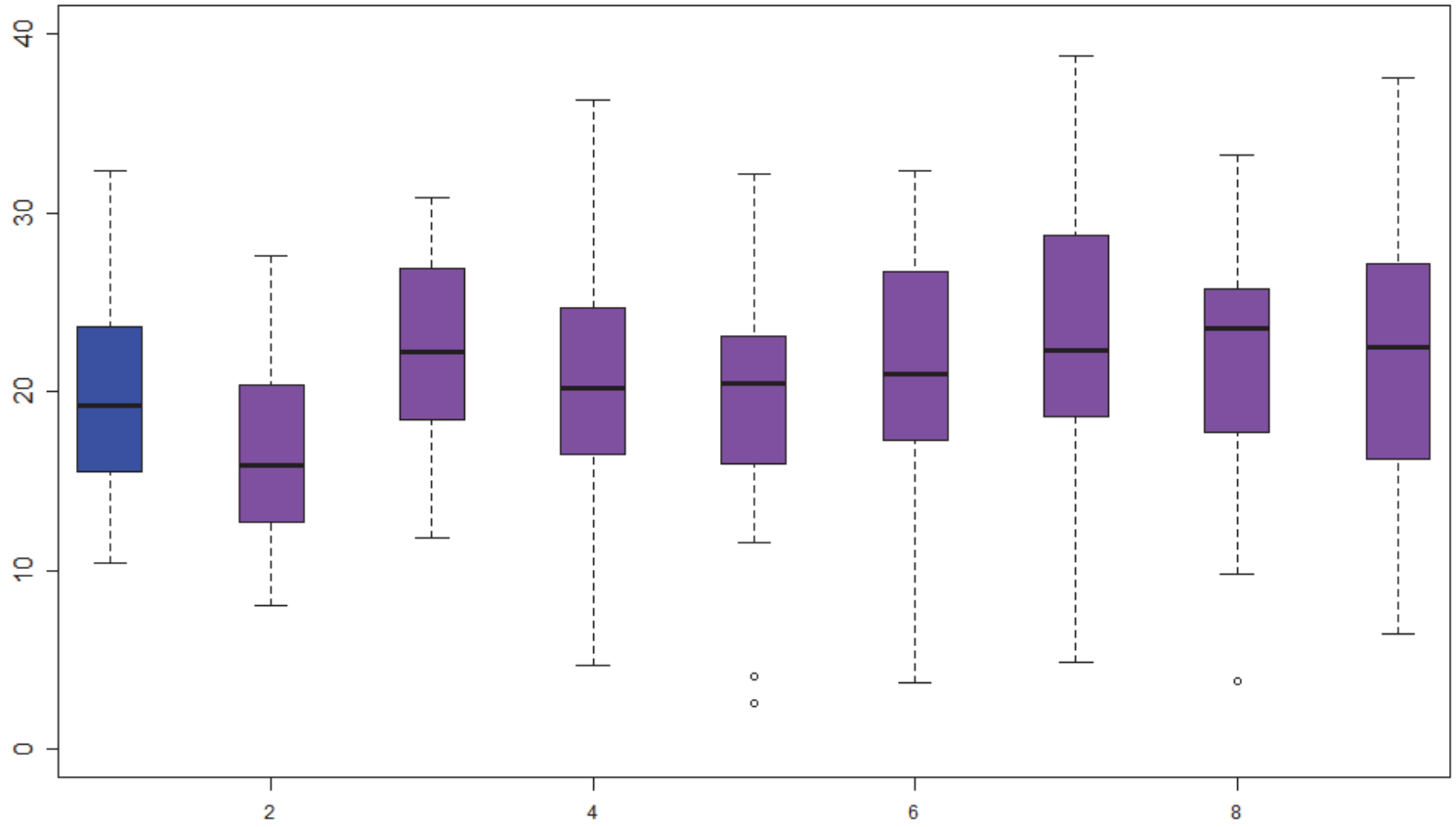
Luckily, it is rather easy to get this:

$$P(y^{rep}|y,x) = \int P(y^{rep},\theta|y,x)d\theta$$

$$= \int P(y^{rep}|\theta,x)P(\theta|y,x)d\theta$$

Approved for public release; distribution is unlimited.

Remember to assess convergence and then make sure that the model fits the data well

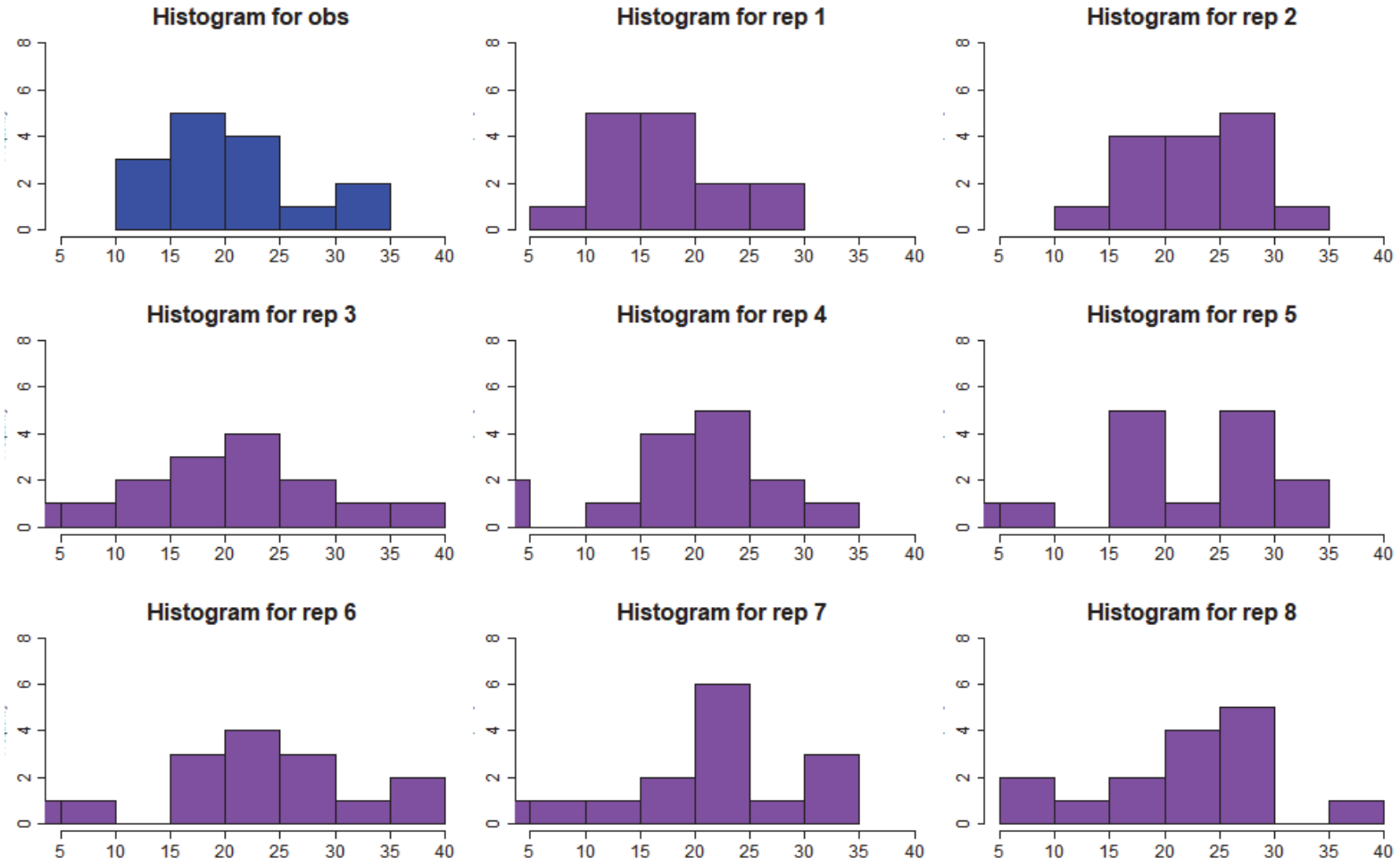


Blue = observed data

Purple = replicated data ($\boldsymbol{y^{rep}}$)

Approved for public release; distribution is unlimited.

We can also compare the observed and replicated data using histograms



Blue = observed data

Purple = replicated data ($\boldsymbol{y^{rep}}$)

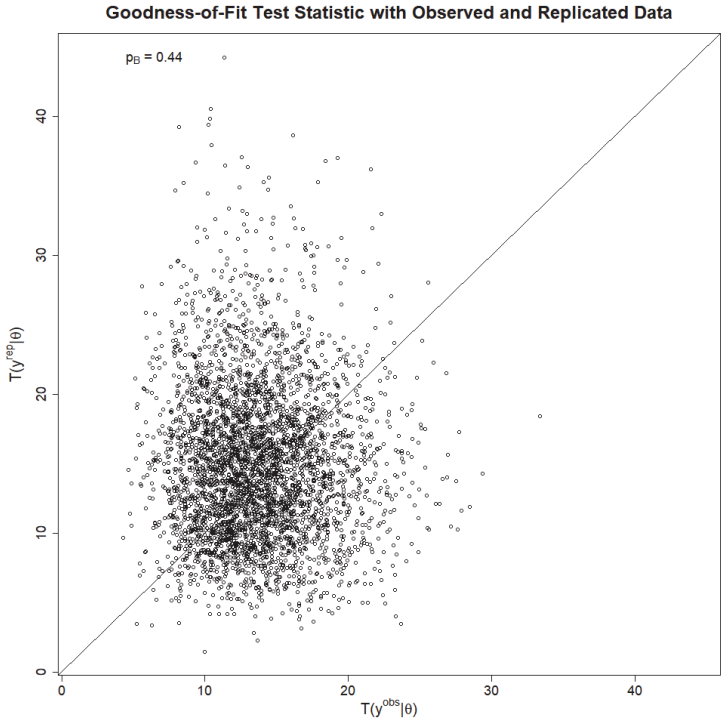
Approved for public release; distribution is unlimited.

The posterior predictive p-value is a quantitative way to assess model fit

Bayesian p-values

$$p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y)$$

where $T(y, \theta)$ is a test statistic (for example, goodness-of-fit, quantile, etc.)



Approved for public release; distribution is unlimited.

Implementation using R

Approved for public release; distribution is unlimited.

Multiple Linear Regression

Approved for public release; distribution is unlimited.

Analyze the joint effect of all operational test factors at once with multiple regression

Separate signal from noise (smooth the data)

Predict important outcomes

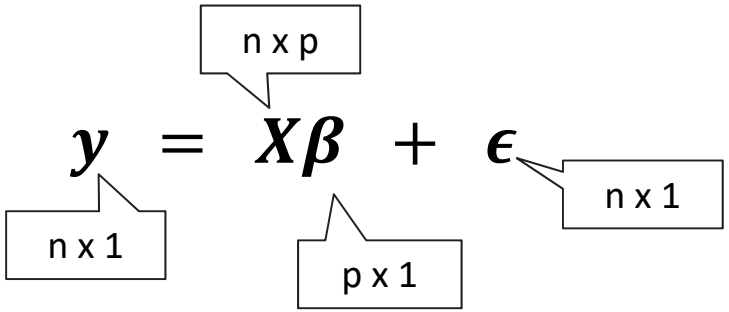
Handle categorical and continuous variables

Determine effects of factors on outcomes

Quantify noise in outcomes

Approved for public release; distribution is unlimited.

The general linear model – beyond simple linear regression



- $\boldsymbol{y}$ is a vector (or a one-column matrix) of n observations on our response variable
- $\boldsymbol{X}$ is an $n \times p$ matrix of observations on $p - 1$ factors
- $\boldsymbol{\beta}$ is a $p \times 1$ matrix of unknown parameters
- $\boldsymbol{\epsilon}$ is a vector of n observation-specific deviations from the expected value

Approved for public release; distribution is unlimited.

General linear model – assumptions

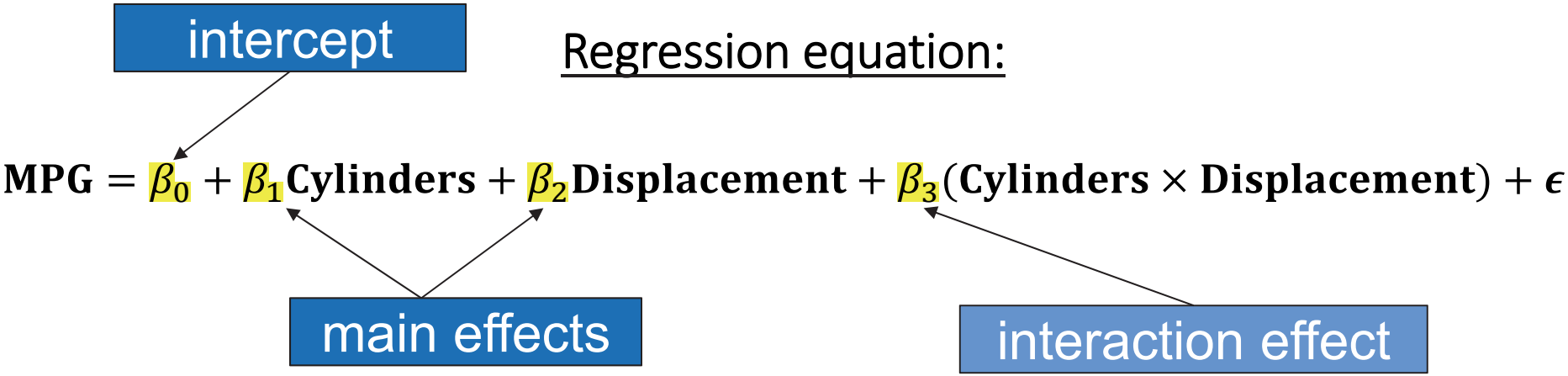
For estimation to be valid, we only need these four assumptions.

- Homoskedasticity – constant variance about any value of the regression function (e.g., deviation for each unit has the same variance)
- Independence – errors are statistically independent
- Linearity – the expected values (means) are linear functions of the parameters
- Existence – finite mean and variance

Always check these assumptions with **residual plots** and **posterior predictive checks**.

Approved for public release; distribution is unlimited.

General linear model – example



$$\begin{bmatrix} MPG_1 \\ MPG_2 \\ MPG_3 \\ MPG_4 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \end{bmatrix}$$

$$\boldsymbol{MPG = X\beta + \epsilon}$$

Approved for public release; distribution is unlimited.

It's better to have research goals in mind *before* you analyze the data

1974 was about the time of the oil crisis, so we can assume that mpg is the outcome variable.

Possible research questions:

- Do the data explain which factors affect mpg?
  - To what extent?
- Do we need all the variables to explain mpg?
- Is a linear relationship suitable?
- If we make predictions, can we make them with confidence?

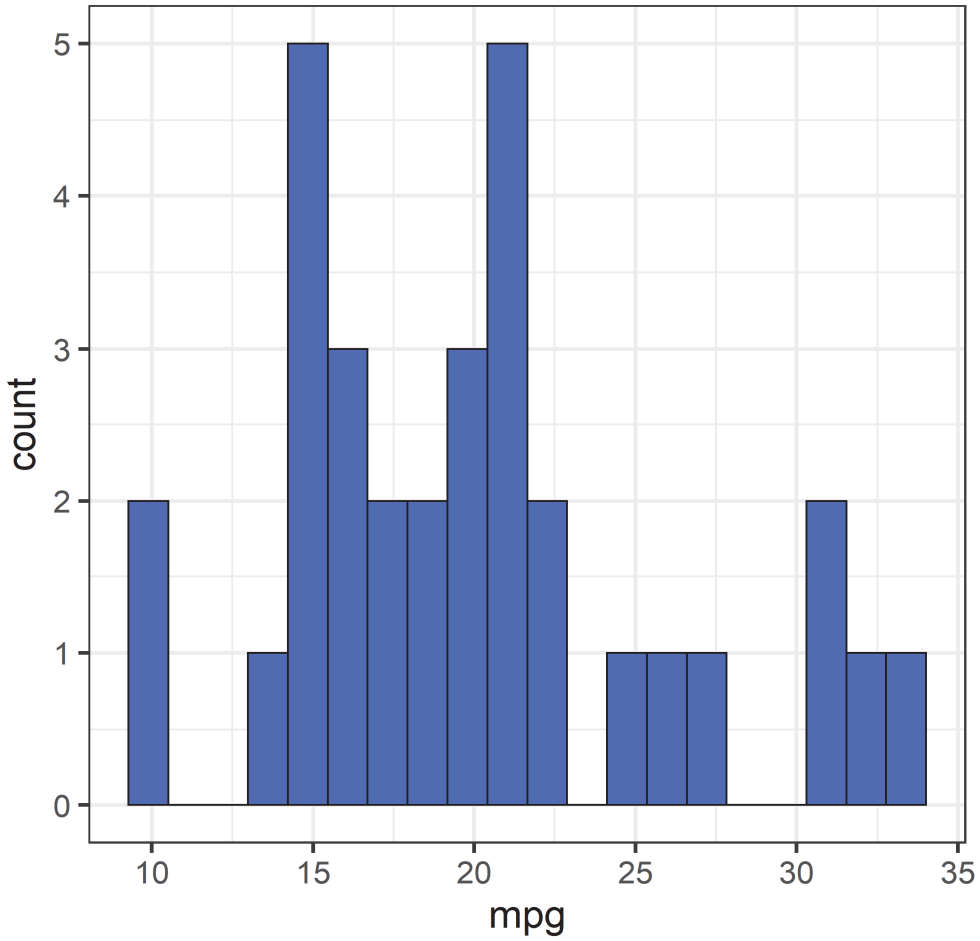
Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

A Little Bit about Data Exploration

Approved for public release; distribution is unlimited.

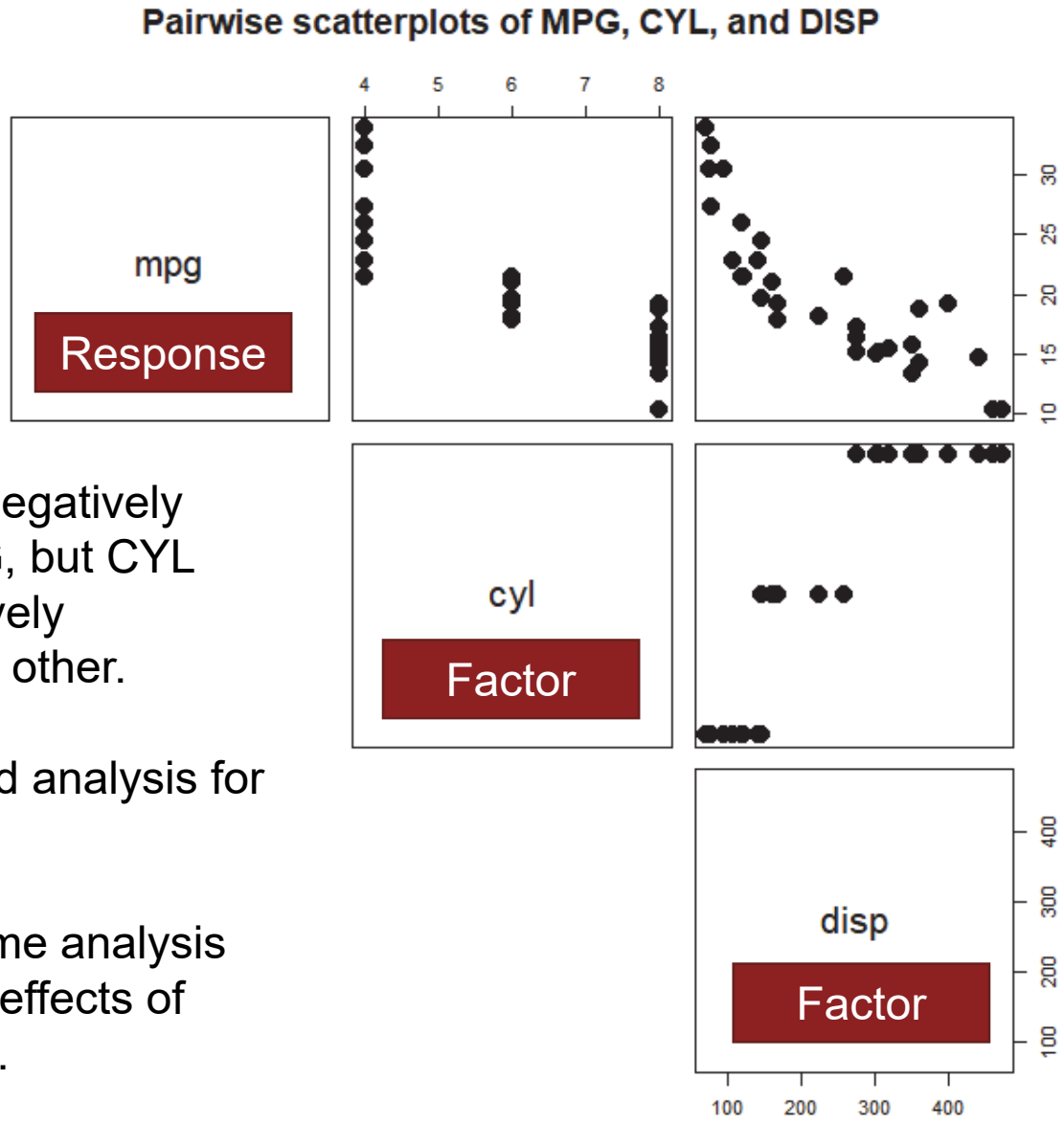
MPG does not look normally distributed, but this is not a problem



Remember: For regression to work well, we want the *residuals* to look normal-ish, not the raw data!

Approved for public release; distribution is unlimited.

	Explanation
mpg	Miles per gallon
cyl	Number of cylinders
disp	Displacement (cu. in.)



- CYL and DISP are negatively correlated with MPG, but CYL and DISP are positively correlated with each other.
- Regression is a good analysis for this type of data.
- One-variable-at-a-time analysis would overstate the effects of each factor on MPG.

Approved for public release; distribution is unlimited.

For example, notice how the coefficients change once we add another variable (least squares with `lm`)

```
Call:
lm(formula = mpg ~ cyl, data = dat)


Coefficients:
(Intercept)         cyl6         cyl8
     26.664       -6.921      -11.564
```

```
Call:
lm(formula = mpg ~ cyl + disp, data = dat)


Coefficients:
(Intercept)         cyl6         cyl8         disp
   29.53477     -4.78585     -4.79209     -0.02731
```

It's challenging to assess factor effects when variables are highly correlated! (This is one reason we stress good experimental designs.)

Approved for public release; distribution is unlimited.

Fitting Bayesian Models to Data

Approved for public release; distribution is unlimited.

Bayesian priors for regression models

Informative priors: $\beta_j \sim Normal(0, \sigma_\beta^2), \sigma^2 \sim Inverse - Gamma(a, b)$

Priors centered at 0 are **more conservative** than least squares

- Scale experimental design variable to make priors easier to set
- Do not use informative priors for β_0 without very good reason (this biases overall model)

Non-informative priors:

- "Standard"* non-informative prior for regression is $p(\beta) \propto 1$ and $\log(\sigma) \propto 1$

Data model:

$$y_i \sim Normal\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_i, \sigma^2\right)$$

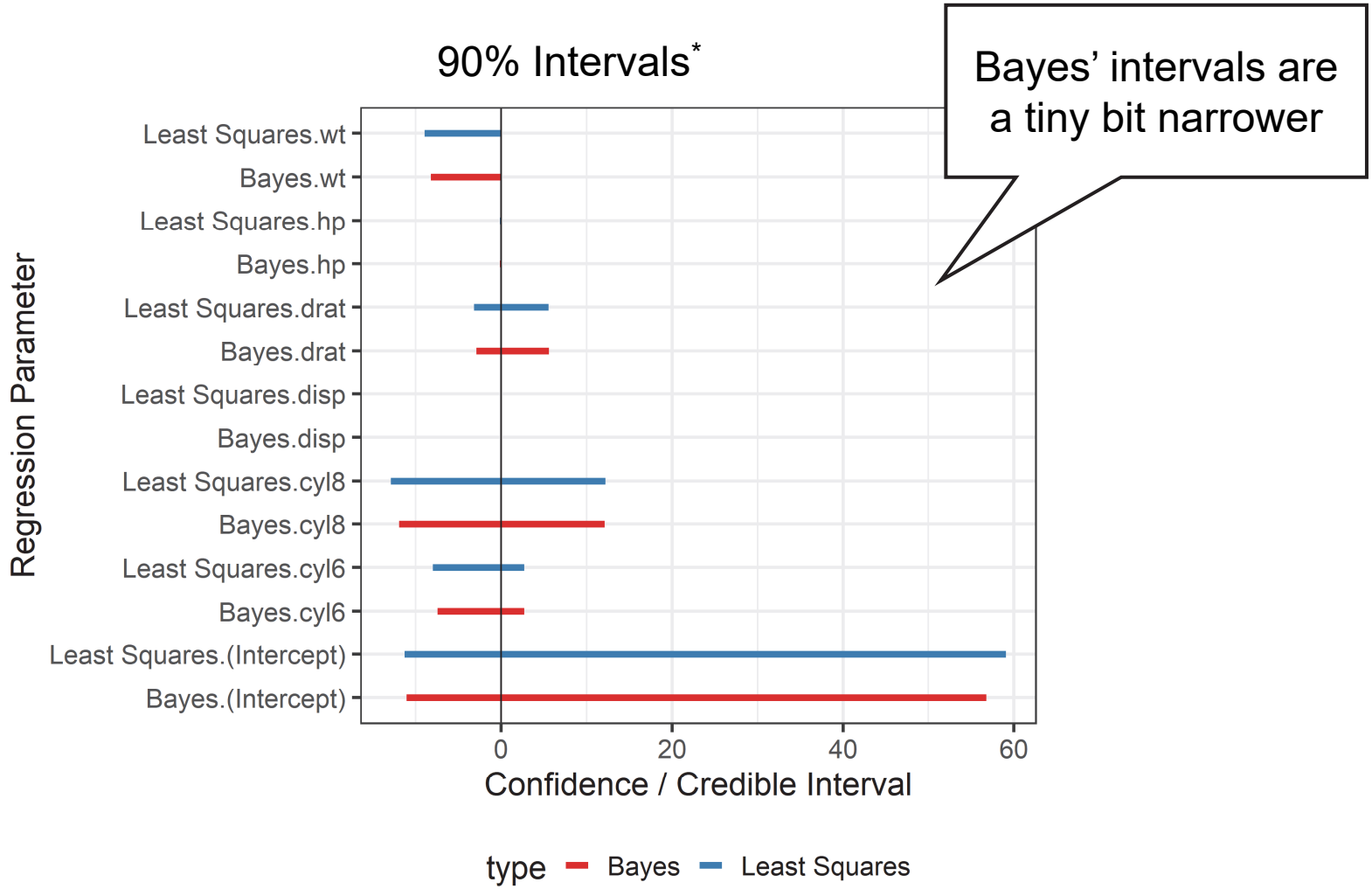
Or, using matrix notation,

$$\boldsymbol{y} \sim Normal(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$$

"Signal" "Noise"

* Bayesian Data Analysis, p. 355

Bayesian analysis with Stan's default priors is similar to least squares



* Only a few regression parameters are plotted for ease of visualization.

Approved for public release; distribution is unlimited.

What does Stan use for default priors?

Stan uses "weakly informative" data-driven priors by default

$$\beta_i \sim Normal\left(0, 2.5 \times \frac{\mathrm{sd}(y)}{\mathrm{sd}(x_i)}\right)$$

$$\sigma \sim Exponential\left(\mathrm{sd}(y)\right)$$

$$\beta_0 \sim Normal\left(\bar{y}, 2.5 \times \mathrm{sd}(y)\right)$$

You could use the defaults in the absence of good priors*

* Default priors have a greater effect in logistic regression and a relatively small effect in linear regression.

Approved for public release; distribution is unlimited.

Variable selection – Do I include `qsec`? `vs`? `wt`? `hp`?

Difficult and controversial!

- Some statisticians: Reduce the number of variables as much as possible!
- Others: Parsimony is the enemy of predictive performance!

Approved for public release; distribution is unlimited.

Shrinkage is one way to handle variable selection

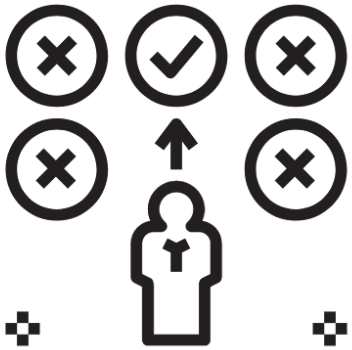
"Shrinkage" is the idea that large, but weak, estimates should be nudged towards zero.

A slightly biased estimate can be more accurate than an unbiased one

Reducing effect sizes means results are more likely to replicate

Shrinkage often helps us avoid computer problems (numerical instability)

Perform model selection through estimation



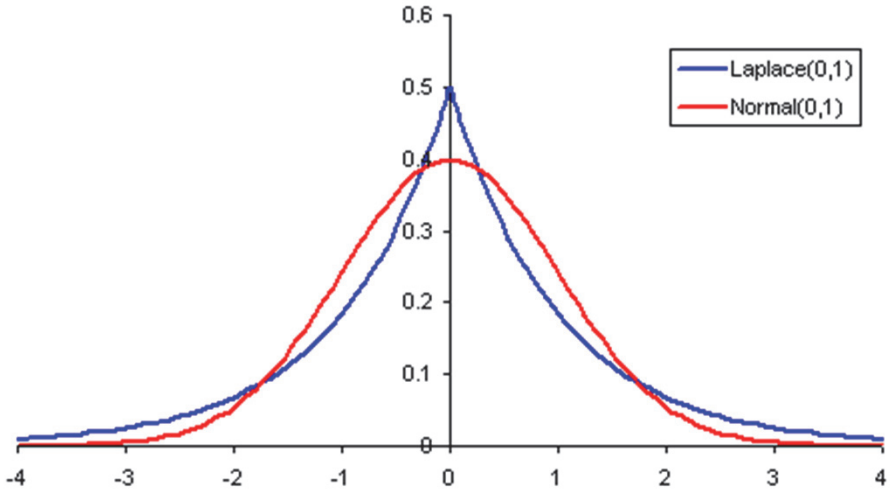
Why a Bayesian approach to shrinkage?

Credible intervals

Frequentist models don’t produce confidence intervals when effects are shrunk!

Approved for public release; distribution is unlimited.

Bayesian shrinkage model



Commonly, we use Laplace priors to "shrink" factor effects towards 0.

In frequentist statistics, the analogous procedure is called LASSO.

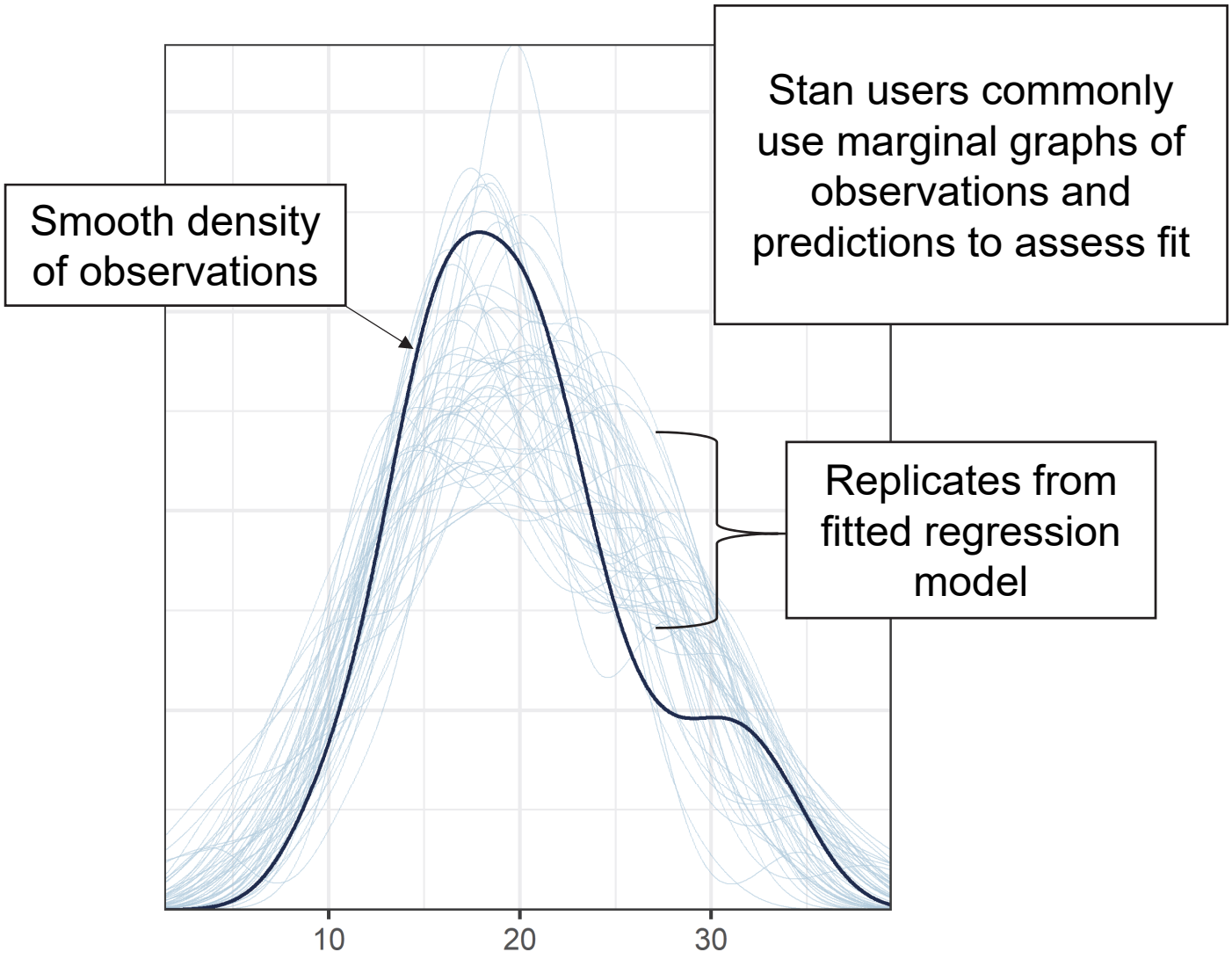
LASSO = Least Absolute Shrinkage and Selection Operator

Approved for public release; distribution is unlimited.

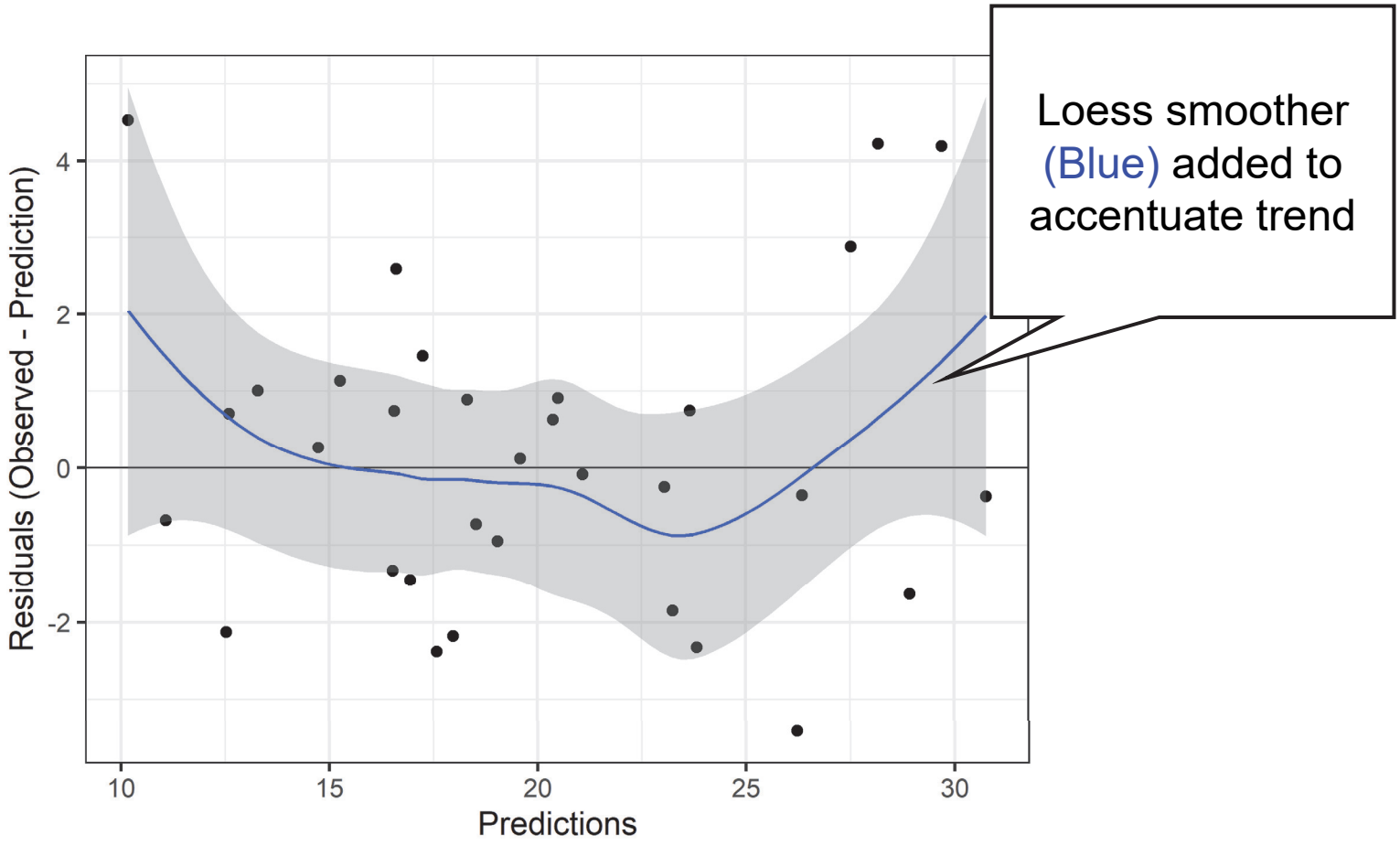
Fitting the mpg model in Stan

Approved for public release; distribution is unlimited.

Use graphical methods to validate Bayesian models



Use graphical methods to validate Bayesian models



Not a good model fit!

Don’t worry, the first model usually isn’t very good.

Use corrective procedures to ensure the results are reliable

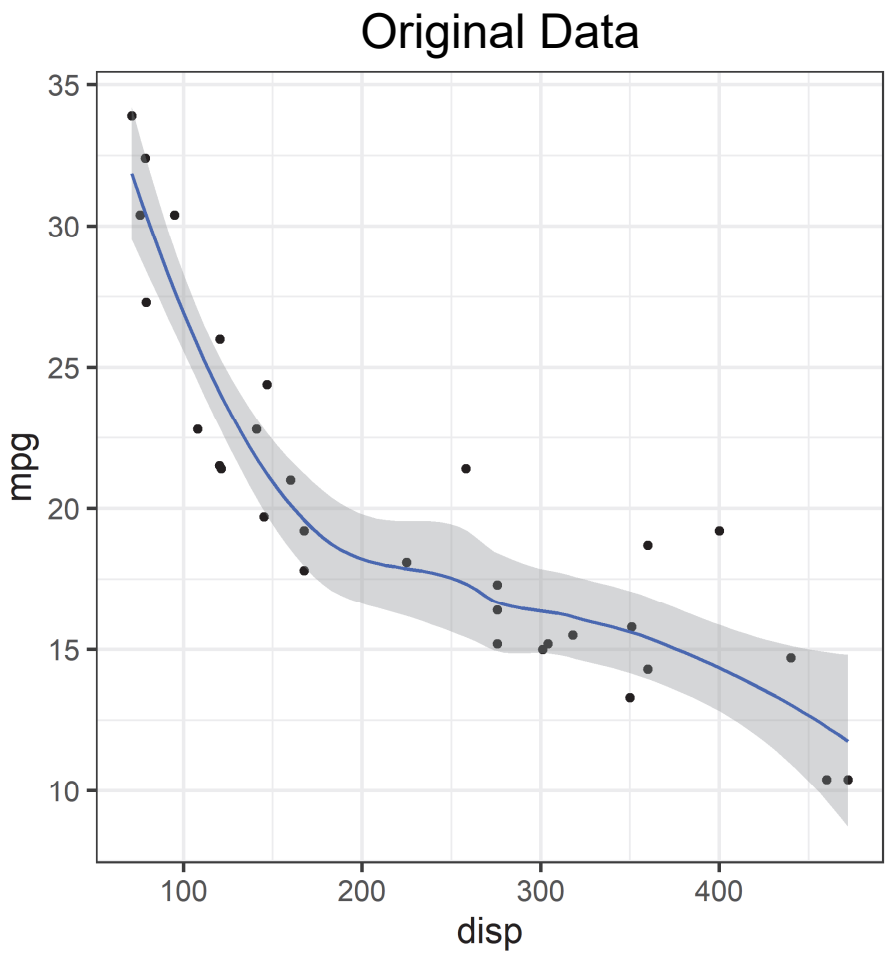
- Variable transformations
- Change of model (for example, linear model to lognormal)
- Splines (rather automatic; recommended)

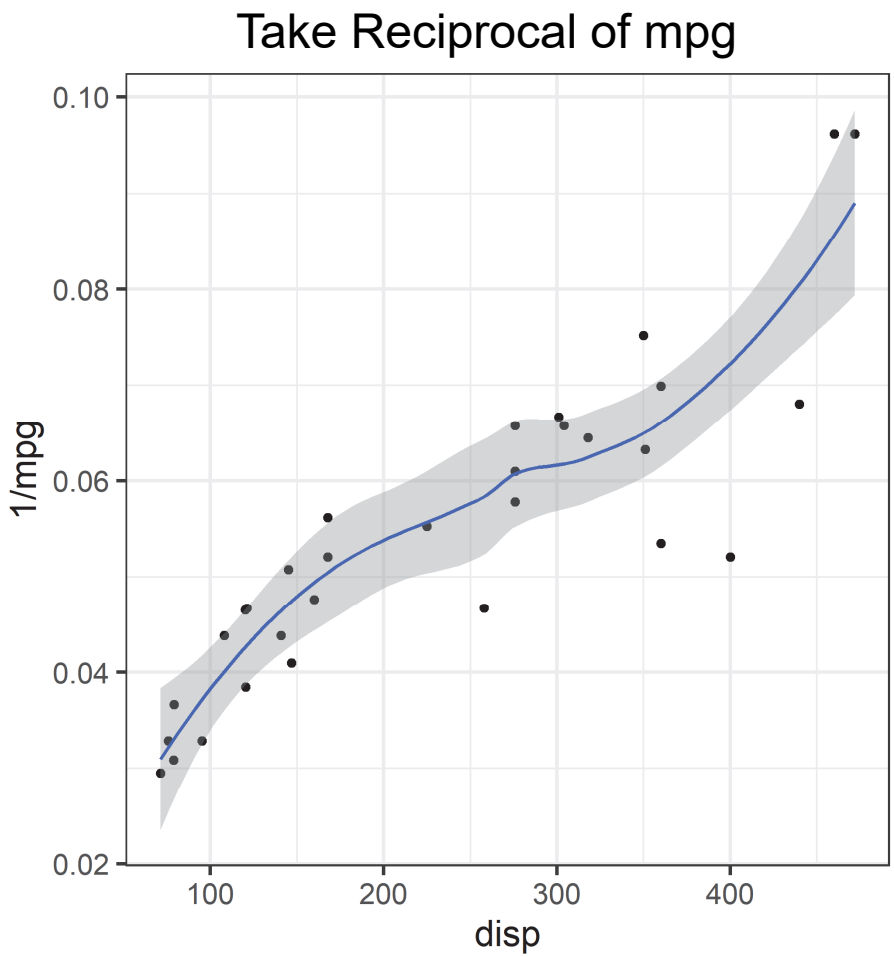
It's much harder to get a good fit when working with continuous variables.

Note that these procedures could also be used in frequentist analyses.

Approved for public release; distribution is unlimited.

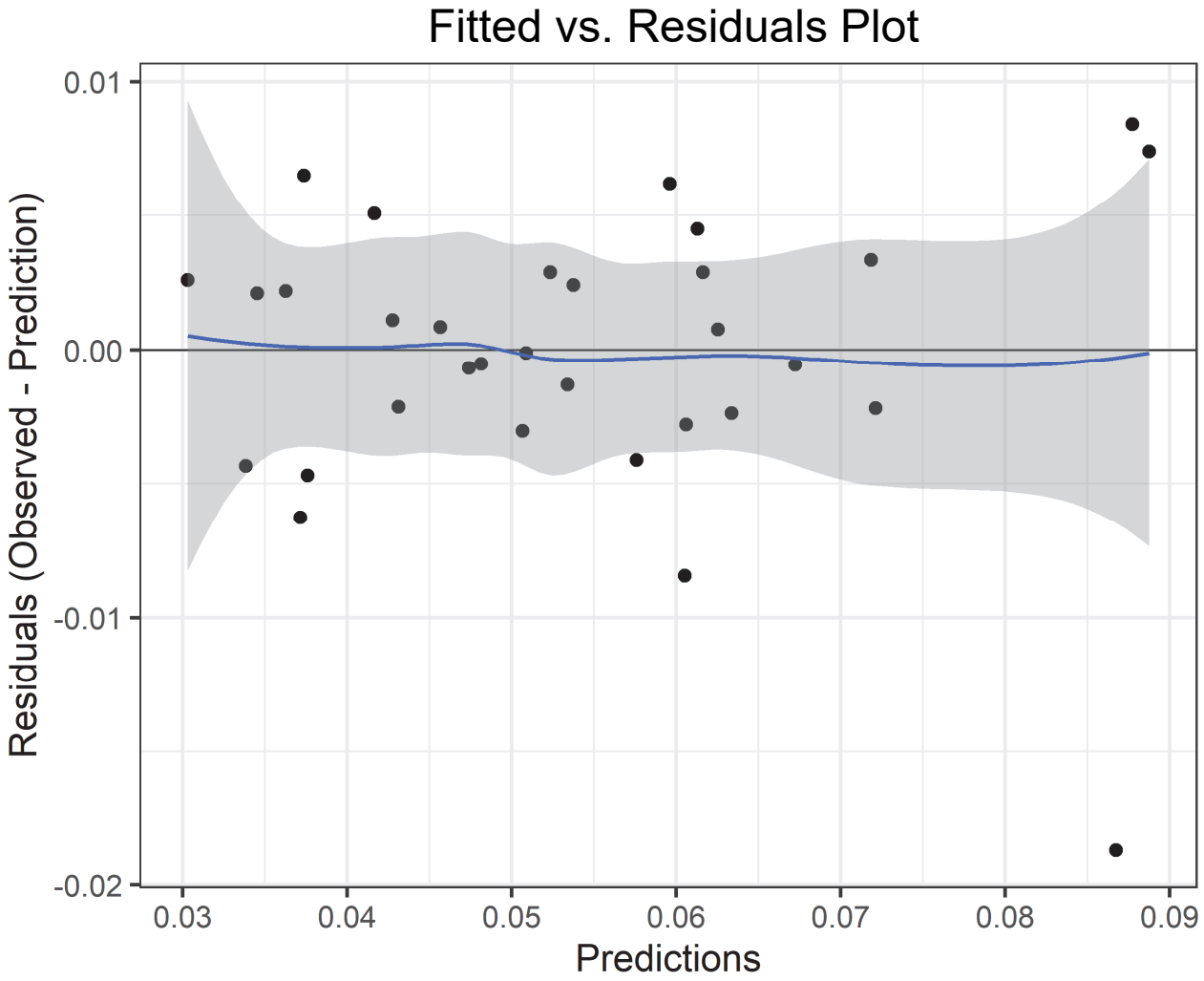
Inverting mpg makes for a more linear relationship with some variables





Approved for public release; distribution is unlimited.

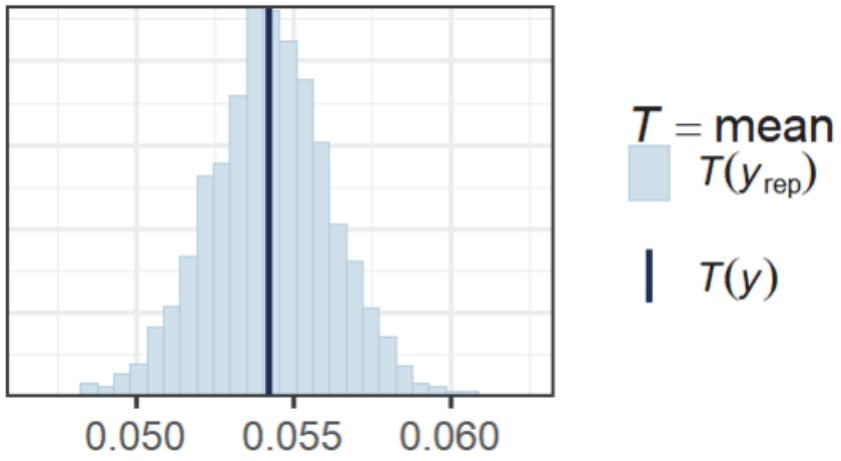
Graphical checks validate the model



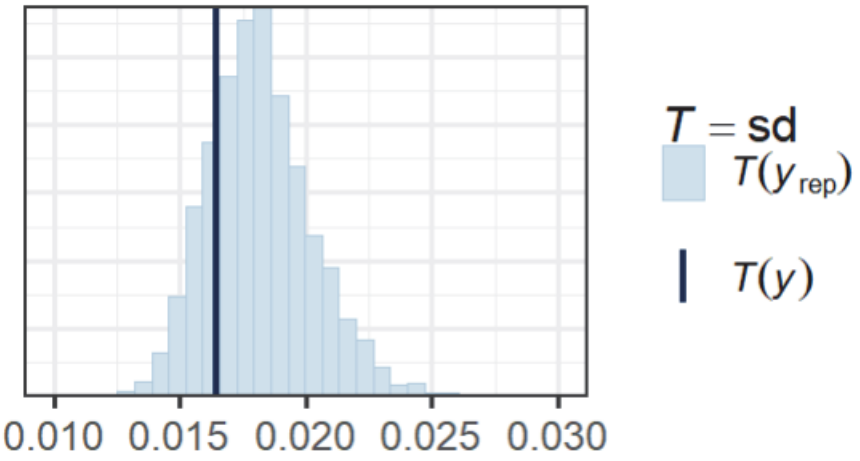
Much better!

Approved for public release; distribution is unlimited.

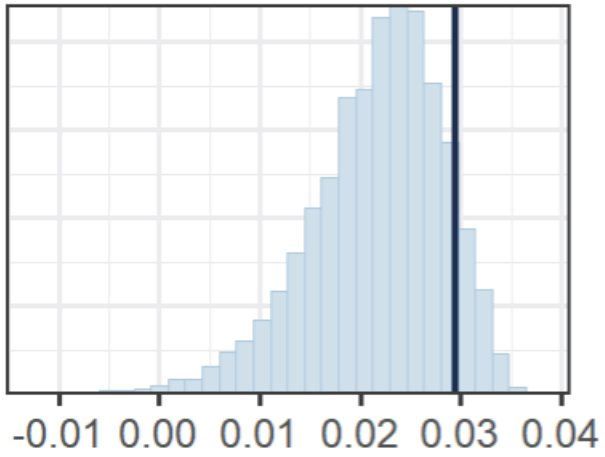
Posterior p-values can be graphed for easier model checking



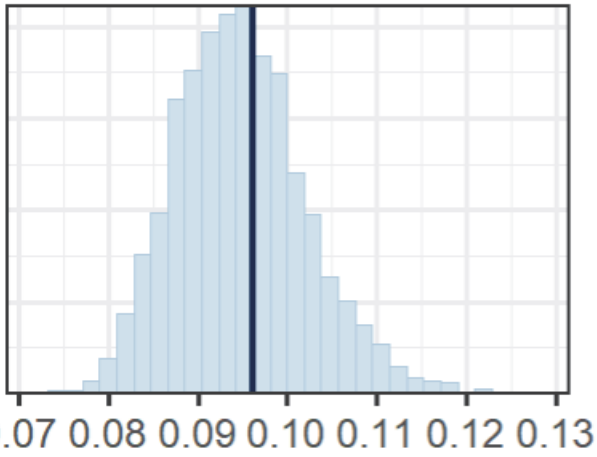
$T = \text{mean}$
$T(y_{rep})$
$T(y)$



$T = \text{sd}$
$T(y_{rep})$
$T(y)$



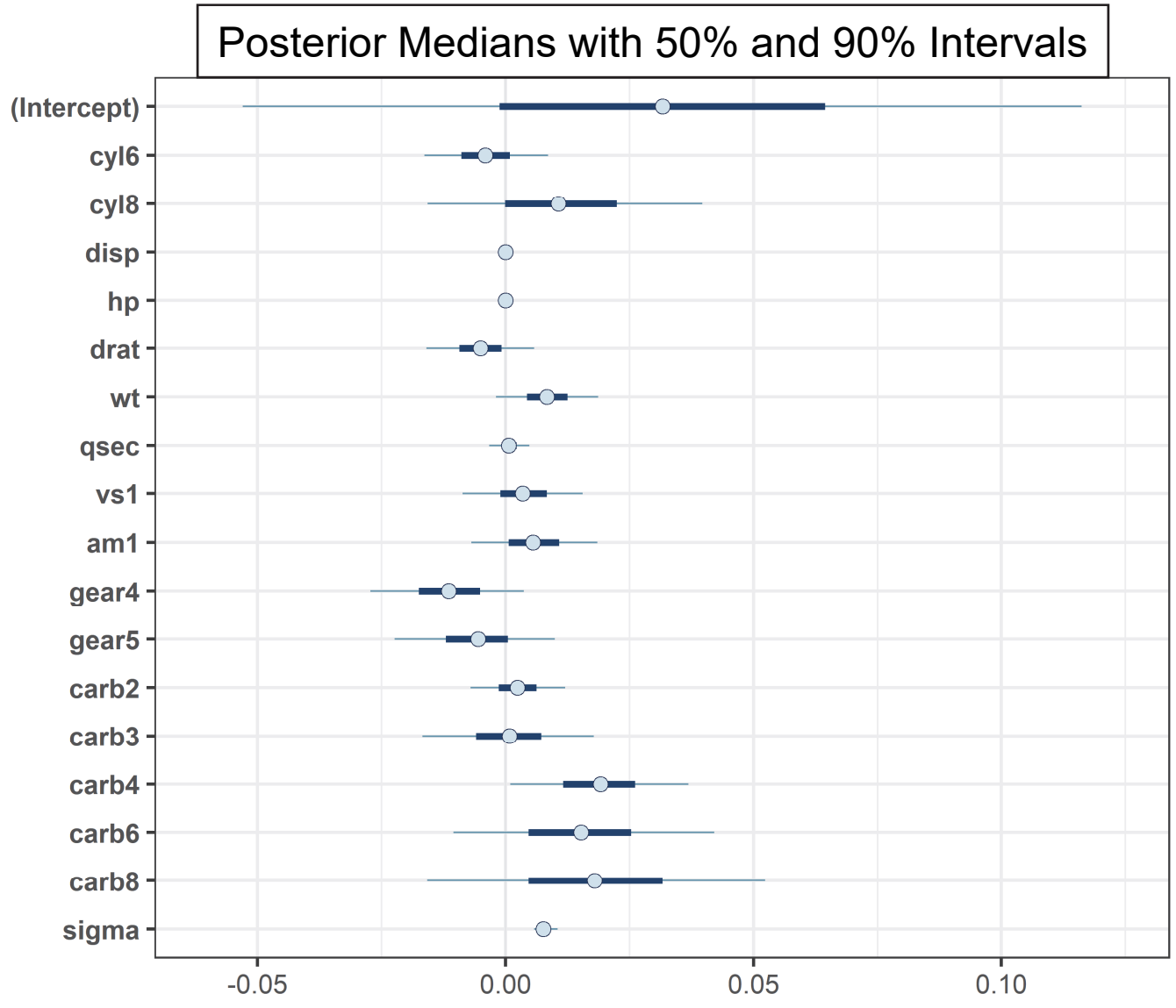
$T = \text{min}$
$T(y_{rep})$
$T(y)$



$T = \text{max}$
$T(y_{rep})$
$T(y)$

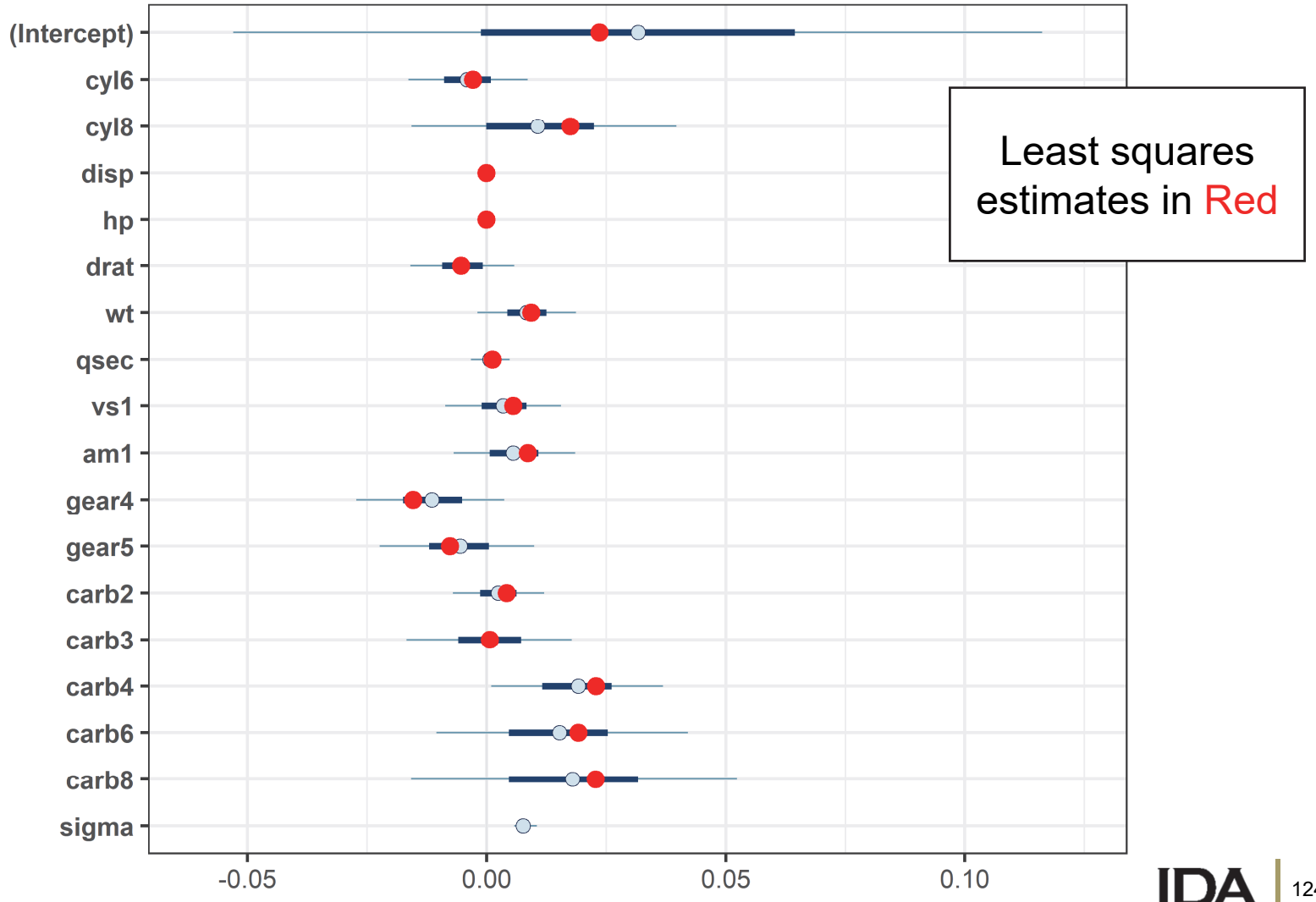
Approved for public release; distribution is unlimited.

Estimated regression coefficients for the 1/mpg model



Approved for public release; distribution is unlimited.

Bayesian estimates (blue) are “shrunk” toward zero



Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Predictions

Approved for public release; distribution is unlimited.

It's easy to get mpg predictions from the 1/mpg model

There are two ingredients:

1. The posterior predictive distribution (discussed earlier)
2. The transformation on the entire distribution

Approved for public release; distribution is unlimited.

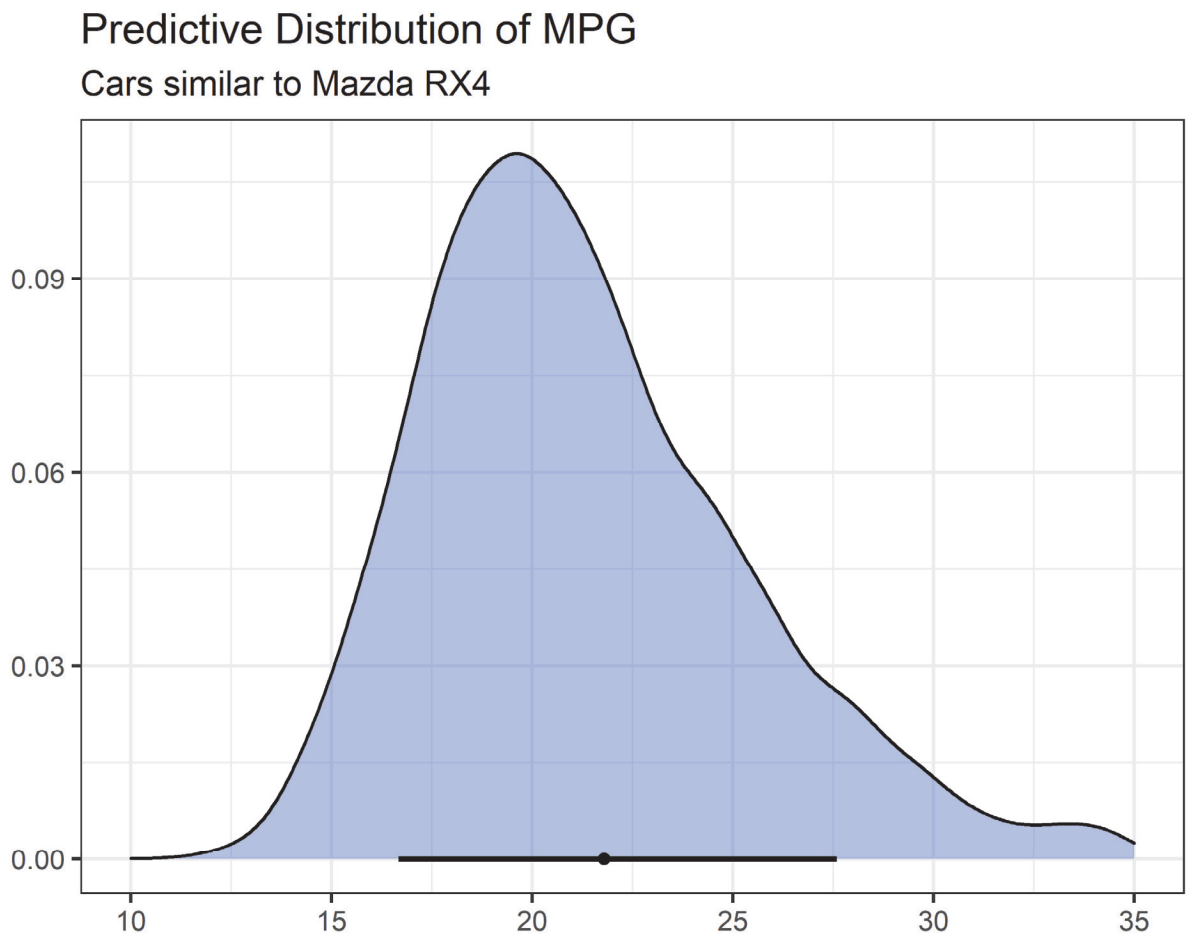
We can use the posterior *predictive* distribution and transform the data back to original data scale

In our case, we have samples from $P\left(\frac{1}{mpg^{rep}}\middle|mpg\right)$, but we want $P(mpg^{rep}|mpg)$.

Translating back is just a matter of applying the reciprocal function to each of the samples from $P\left(\frac{1}{mpg^{rep}}\middle|mpg\right)$.

Approved for public release; distribution is unlimited.

Use the posterior predictive distribution to summarize prediction error



Summarizing predictions with Stan

Approved for public release; distribution is unlimited.

What do you do when Stan is not cooperating?

Two tricks to make Stan cooperate:

1. Increase `adapt_delta`
2. Increase `max_treedepth`

For more information about Stan warnings, see https://mc-stan.org/misc/warnings.html.

Approved for public release; distribution is unlimited.

Bayesian analysis extends to regression, and it offers several technical advantages to researchers:

- Simply obtain predictions after transformation
- Calculate intervals after applying shrinkage
- Apply a range of effective model criticism tools

Approved for public release; distribution is unlimited.

Bayesian statistics is another tool in your statistical analysis toolbox

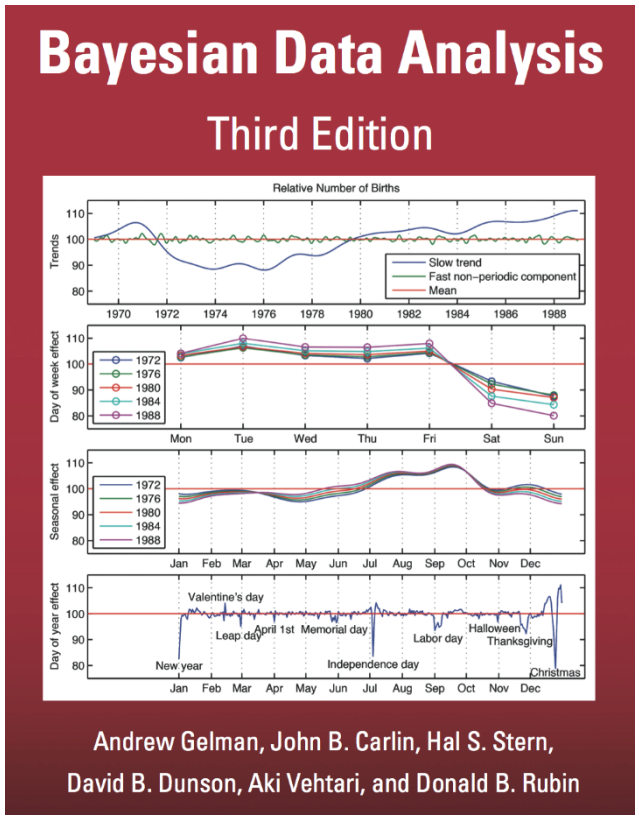
Although the goal of Bayesian and frequentist statistics is to answer a research question, the analysis of the data and interpretation of the results differ between them.

There are multiple ways to properly incorporate prior information into a Bayesian model.

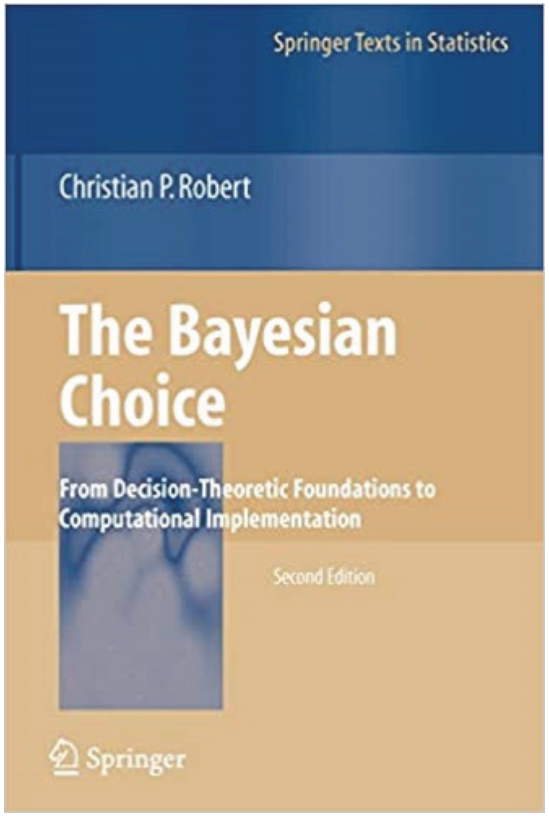
As with frequentist statistics, once you fit your statistical model, make sure your model fits the data well.

Approved for public release; distribution is unlimited.

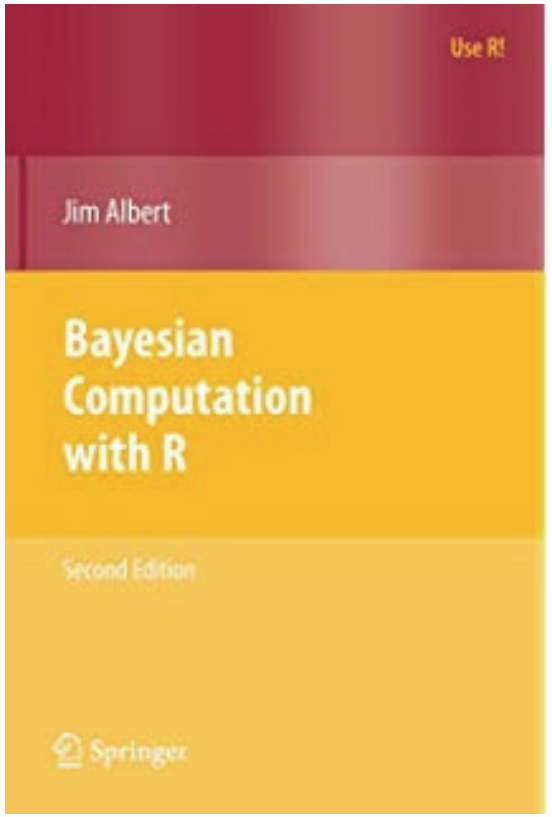
Great references



Application



Theory



Programming

Approved for public release; distribution is unlimited.

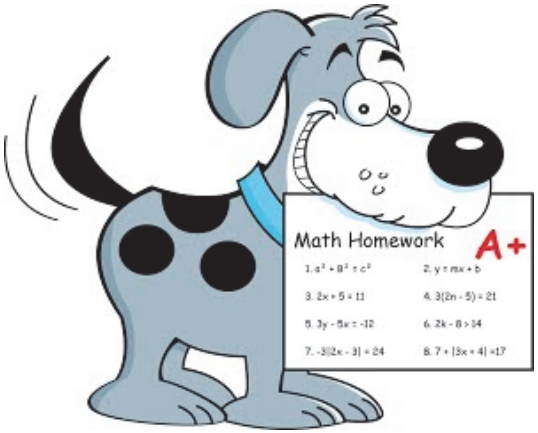
Thank you!

Contact Info:

Dr. Keyla Pagán-Rivera – kpaganri@ida.org

Dr. John Haman – jhaman@ida.org

Dr. Rebecca Medlin – rmedlin@ida.org



Resources:

- https://testscience.org/
- Past trainings – OED SharePoint site
- Future trainings – check https://test-science.shinyapps.io/TSTraining/

Approved for public release; distribution is unlimited.

Backup

Approved for public release; distribution is unlimited.

Details on computing the posterior predictive p-value

Steps:

- Use the posterior distribution to simulate a vector of θ
- Obtain the joint posterior distribution by drawing y^{rep} from the sampling distribution using the simulated vector of θ
- Estimate the Bayesian p-value

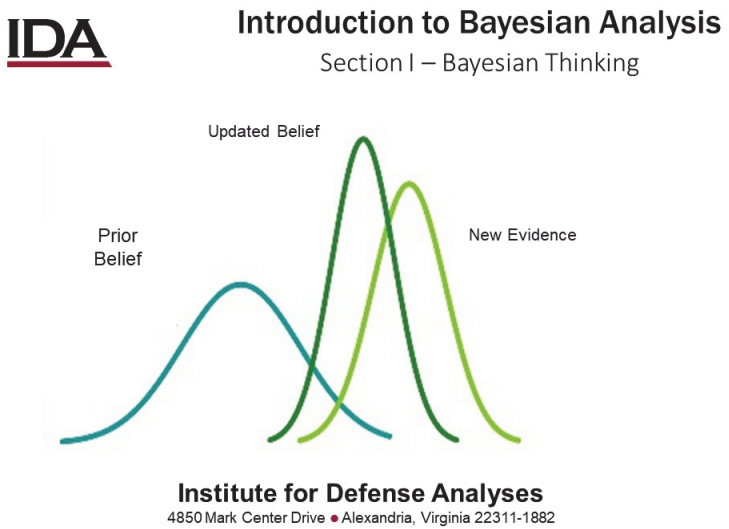
$$p_B \approx \frac{1}{L}\sum_{l=1}^{L} I_{\{T(y^{rep,l},\theta^l)>T(y,\theta^l)\}}$$

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Appendix A
Bayesian Thinking Annotated Briefing



Slide notes:

Bayesian statistics is a mathematical tool that allows us to update our current beliefs in light of new evidence. It allows us to properly use all available data. I have some prior belief, which can change as new information is gathered.

"Everybody is a Bayesian. It's just that some know it, and some don't." – Trivellore Raghunathan

Slide notes:

For this class, we assume that people are familiar with frequentist statistics. Therefore, this section focuses on introducing the reasoning behind Bayesian statistics and describing how Bayesian statistics compares with frequentist statistics.

IDA | 1

Prior knowledge could affect your belief about some outcomes

- Experiment 1
  - A fine musician, specializing in classical works, tells us that he can distinguish whether Haydn or Mozart composed a classical song. Small excerpts of the compositions of both authors are selected at random and played for the musician to identify. The musician makes 10 correct guesses in 10 trials.

- Experiment 2
  - The guy next to you at the bar says he can correctly guess in a coin toss what face of the coin will fall down. Again, after 10 trials the man correctly guesses the outcomes of the 10 throws.

IDA | 2

Frequentist vs. Bayesian thinking

- Frequentist statistical analysis
  - You have the same confidence in the musician's ability to identify composers as in the bar guy's ability to predict coin tosses. In both cases, there were 10 successes in 10 trials.
- Bayesian statistical analysis
  - Presumably, you are inclined to have more confidence in the musician's claim than the guy at the bar's claim. Post-analysis, the credibility of both claims will have increased, though the musician will continue to have more credibility than the bar guy.

Using frequentist statistics for some jobs and Bayesian statistics for others does not mean you have to sign up for a lifetime of using only one tool!

IDA 3

Goals of this training

- Introduce Bayesian statistics and the similarities and differences between Bayesian statistics and classical statistics
- Explain the methodologies behind Bayesian models and their implementation using software
- Demonstrate some ways of intelligently combining information
- Describe ways to communicate the results from a Bayesian data analysis

IDA 4

Statistics is more than summarizing data

Operational test reports summarize the findings of a test, but simply tabulating operational test data is not sufficient.

We should also strive to:

- Generalize from operational test to operational employment
- Generalize from sample of operators to population of operators
- Infer performance in untested conditions

To accomplish these things we need to build statistical **models**.

IDA 5

Bayesian methods can be used in operational testing and evaluation

- Operational tests can be complex, expensive, and time consuming, but they are our best tool to understand a system's performance
- Most of the time, data from tests are analyzed independently of any prior test data or subject matter expert knowledge
- Analysts could use Bayesian statistics to wisely incorporate available information when analyzing operational test data

IDA 6

Approved for public release; distribution is unlimited.

Motivation for using ALL information is not new

Military Operations Research Society / International Test and Evaluation Association

Joint Mini-Symposium: How Much Testing Is Enough?, 1994

National Research Council Studies

Statistics, Testing, and Defense Acquisition, 1998

Improved Operational Testing and Evaluation, 2004

The idea of combining all sources of information has been discussed in the defense community for some time.

IDA | 7

Specific notes from these references:

Military Operations Research Society / International Test and Evaluation Association

Joint Mini-Symposium: How Much Testing Is Enough?, **1994**

- Extensive discussion on the use of analytical techniques, such as simple aggregation, meta-analysis, Bayes
- ian statistics, and non-parametric statistics; several examples of each were provided as potential means for the meaningful pooling of information.

National Research Council Studies

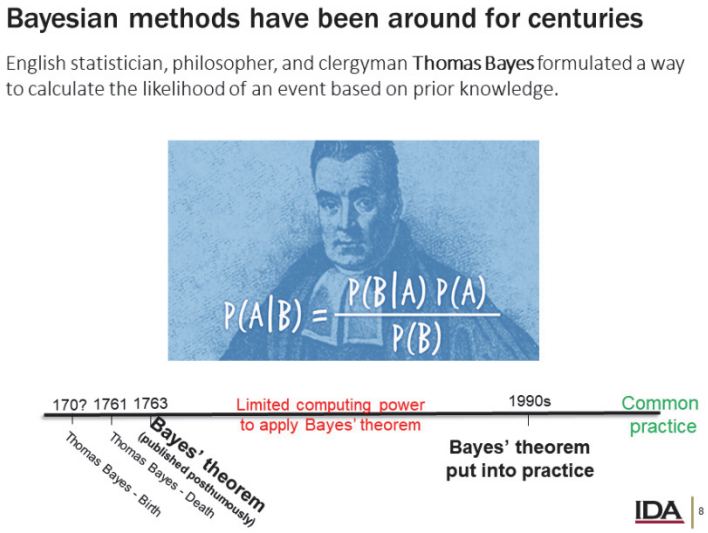
Statistics, Testing, and Defense Acquisition, **1998**

- Emphasizes that all relevant information be examined for possible use in both the design and evaluation of operational tests.
- **State-of-the-art statistical methods for combining information** should be used, when appropriate, to make tests and their associated evaluations as cost efficient as possible.

Improved Operational Testing and Evaluation, **2004**

- Focuses specifically on methods of combining information for the Stryker family of vehicles.

Approved for public release; distribution is unlimited.



Slide notes:

Before the 20th century, statistics and probability were highly commingled.

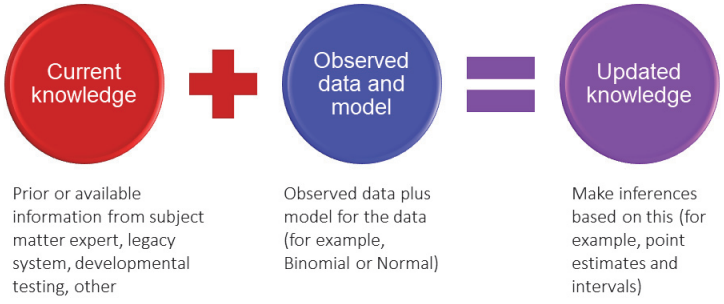
- Frequentist view gained traction.
  - Defines probability through frequencies of events
  - Does not accept inverse probability
- Bayesian analysis is based entirely on the rules of probability theory.
  - Uncertainty and evidence are represented with probabilities
  - Probabilities are updated in light of new information

Fisher may have been the first to use the term "Bayesian" to discriminate between Bayesian statistics and probability theory.

Thomas Bayes (170?–1761)

- Considered the problem of *inverse probability*.
- The equation in the image is known as Bayes' theorem. We explain each term in the next few slides but as a motivational example, let's consider the following:
  - Suppose A represents the true outcome of an event – for example, having COVID-19.
  - Suppose that B represents the result of a test – for example, a COVID-19 test result (PCR or other test). These tests could have false positive results and false negative results.
  - We are interested in computing the probability of a person having COVID-19 given the results of the PCR/antibody/other test, but we cannot compute this probability directly.
  - Bayes' theorem allows us to estimate the conditional probability, $P(A|B)$by using the information we have: the likelihood, the prior distribution, and the marginal distribution.
  - Note that $P(B)$ is the normalizing constant that makes this equation a probability. Many times, we omit it from the equation and instead write $P(A|B) \propto P(B|A)P(A)$. The $\propto$ symbol reads "proportional to."

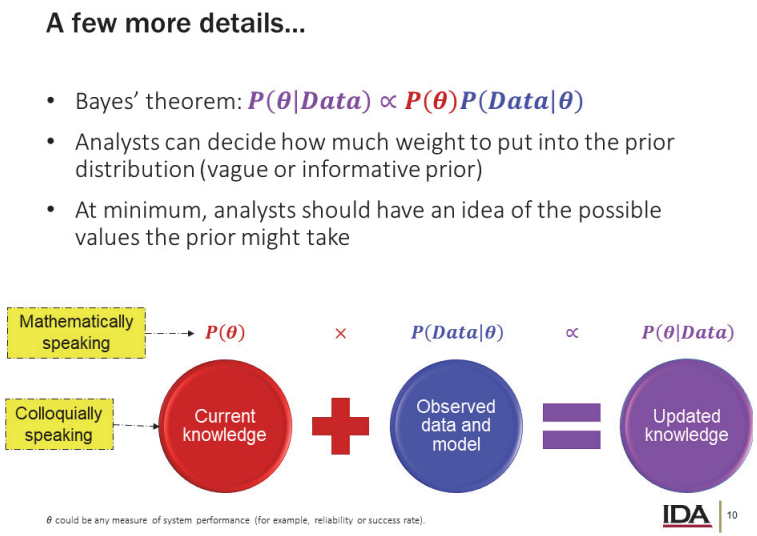
Bayesian statistics in a nutshell



Slide notes:

Colloquially speaking, we can talk about incorporating or "adding" the prior information. There could be multiple sources of variability as well.

Approved for public release; distribution is unlimited.



Slide notes:

Assume we want to estimate the probability of a positive result. We know that the true unobserved probability of a positive test (θ) is between 0 and 1. We might want to use the COVID-19 prevalence on a similar population to build our prior distribution. We could also have a conversation with an SME (for example, an epidemiologist or virologist) and use their knowledge to construct the prior distribution. There has to be a conversation among all interested parties regarding how much weight either prior information will have in the prior distribution. However, we do know the parameter is bounded by 0 and 1.

Let's remember the equation we saw 2 slides ago: $\boldsymbol{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$

- $\boldsymbol{P(A|B)}$ is the posterior distribution or updated knowledge.
- $\boldsymbol{P(B|A)}$ is the likelihood (observed data and model).
- $\boldsymbol{P(A)}$ is the prior distribution or current knowledge.
- $\boldsymbol{P(B)}$ is the marginal distribution (distribution of an event across conditions – that is, without conditioning on another event).
- This equation is commonly written as $\boldsymbol{P(A|B) \propto P(B|A)P(A)}$.
- To be more specific, we say that A is the parameter we want to estimate, and B is the data we observe.

For the COVID-19 example:

- $P(\theta|\ Data)$ is the probability of having COVID-19 given the test results.
- $P(\theta)$ is the prior knowledge – we could estimate this by using the COVID-19 prevalence in the population (say, prevalence in Maryland, Virginia, and Washington, DC). This prior distribution could also represent the subject matter expert's (SME's) knowledge about the parameter of interest.
- $P(Data|\theta)$ is the probability of a test result given the truth.

Approved for public release; distribution is unlimited.

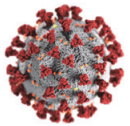
What is the probability that someone has COVID-19?

Suppose someone has been taking precautions but still has COVID-like symptoms.

Before getting tested, they know that the probability of having COVID-19, $P(C^+)$, is 0.05.*

They also know that the test is not 100% perfect. In fact:

- The probability of having a positive result given that the person has COVID-19, $P(T^+|C^+)$, is 0.95.*
- The probability of having a negative test result given that the person in fact does not have COVID-19, $P(T^-|C^-)$, is 0.98.*



* See the slide notes for more information about these numbers.

IDA | 11

Slide notes:

Regarding these numbers:

- The probability of having COVID-19 in places like Arlington and Alexandria, Virginia, or Washington DC is between 0.05 and 0.06. For simplicity, we chose 0.05.
- The FDA guidance for COVID-19 tests varies depending, among other things, on the type of test. For this example, we chose the guidance from the "Molecular Diagnostic Template for Commercial Manufacturers" (updated July 28, 2020). This guidance offers a lower bound but the actual test might have a higher number.

Note that the probability of

- A false positive, $P(T^+|C^-)$ is $1 - P(T^-|C^-) = 0.02$
- A false negative, $P(T^-|C^+)$ is $1 - P(T^+|C^+) = 0.05$
- Not having COVID-19, $P(C^-)$ is $1 - P(C^+) = 0.95$

Approved for public release; distribution is unlimited.

Let's collect data and use the information we have to find out!

Now assume that this person's result came back positive.

We can update our knowledge based on this new information:

$$P(C^+|T^+) \propto P(T^+|C^+)\, P(C^+)$$

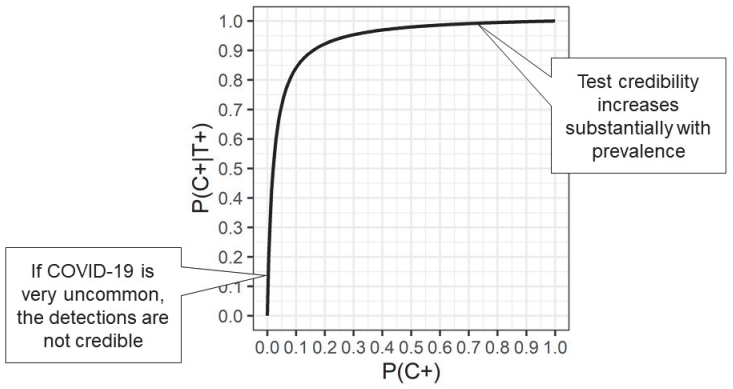
Note that the equation uses $\propto$. We need to normalize the result so it is a probability.*

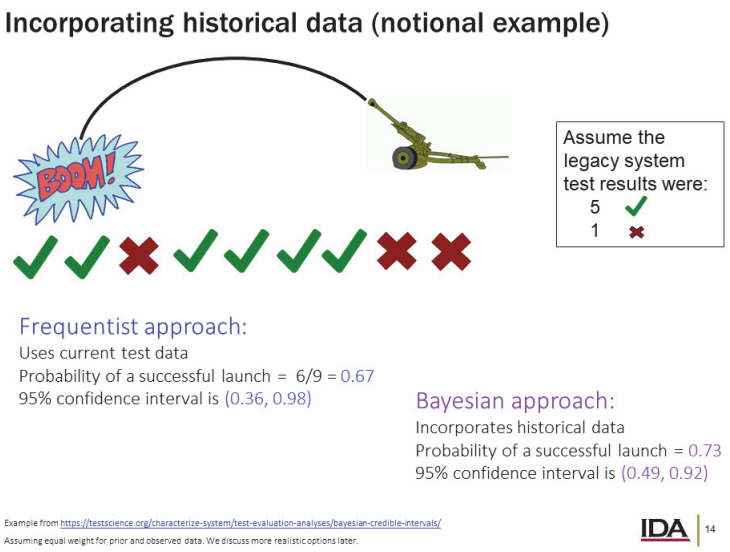
$$P(C^+|T^+) = \frac{P(C^+)P(T^+|C^+)}{P(C^+)P(T^+|C^+)+P(C^-)P(T^+|C^-)} = \frac{0.05\times0.95}{0.05\times0.95+0.95\times0.02} = \mathbf{0.71}$$

The chances of having COVID-19 increased after testing positive.

* In future examples, we work with distributional functions (for example, Normal or Binomial). In those cases, the distribution will "take care" of the normalization.

$P(C^+|T^+)$ depends on $P(C^+)$





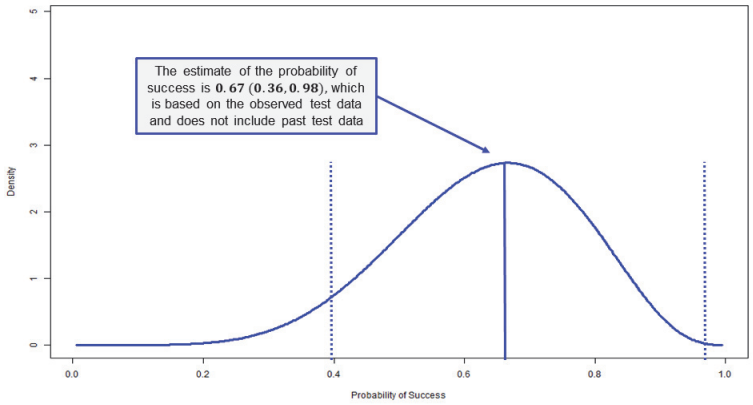
Slide notes:

Suppose we have a missile launch system, and we are interested in the probability of obtaining a successful launch. We observe 6 successful launches and 3 failed launches. Additionally, we have historical data about a similar missile launch system: 5 successes and 1 failure.

- Frequentist approach computes the point estimate for the probability of a successful launch as $\frac{6}{9} = 0.67$. The 95% confidence interval is *(0.36, 0.98)*.
- Bayesian approach incorporates the historical data and obtains a point estimate of *0.73*. The 95% credible interval is *(0.49, 0.92)*.
- Obtaining the Bayesian results is not as simple as adding the successes and failures from both tests. More about this in the next section.
- The uncertainty around the Bayesian estimate is smaller than the uncertainty around the frequentist estimate.

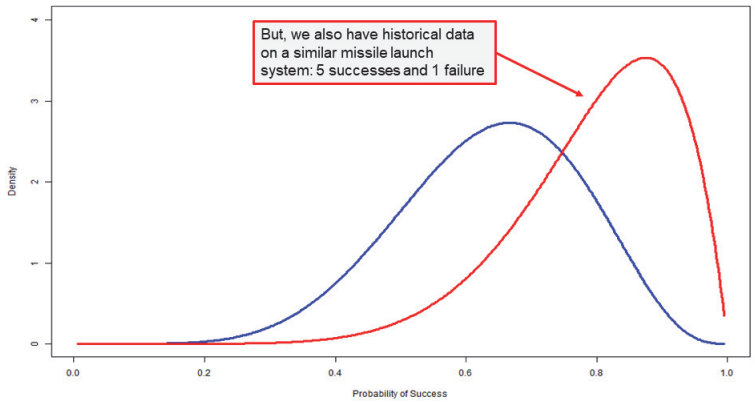
The next section discusses how we obtained the Bayesian results and ways to modify the weight of the prior data. For this example, we are assuming equal weight.

Bayesian analysis of missile launch system: Likelihood



IDA 15

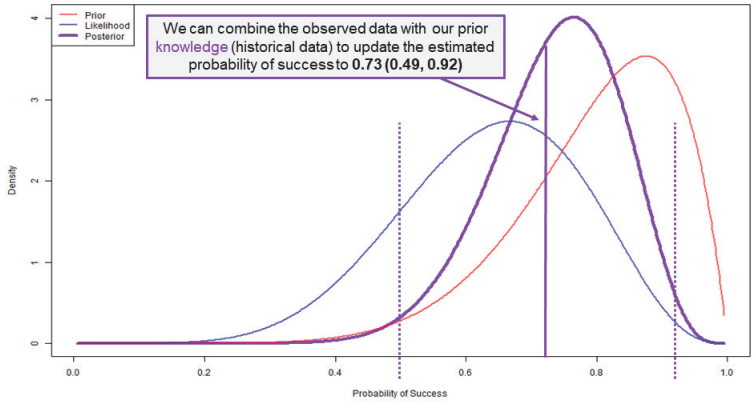
Bayesian analysis of missile launch system: Prior



IDA 16

Approved for public release; distribution is unlimited.

Bayesian analysis of missile launch system: Posterior

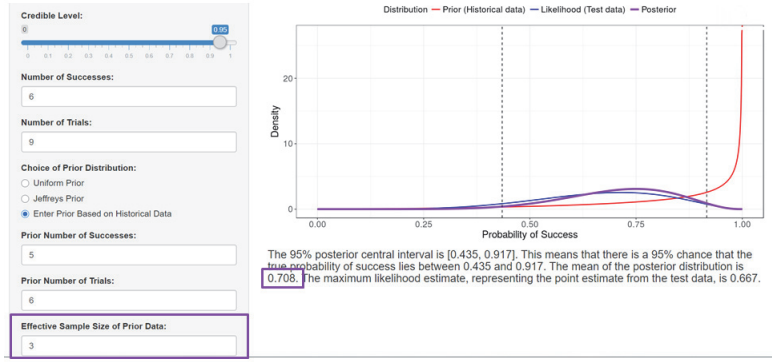


Tool to compute credible intervals and obtain plot is from https://test-science.shinyapps.io/BayesianBinomialCIs/

Slide notes:

We explain how we got this in the next section. Although for this simple example it looks like we are just adding the numbers, the process involves conjugate distributions.

Smaller prior effective sample size puts less weight on the prior data



Slide notes:

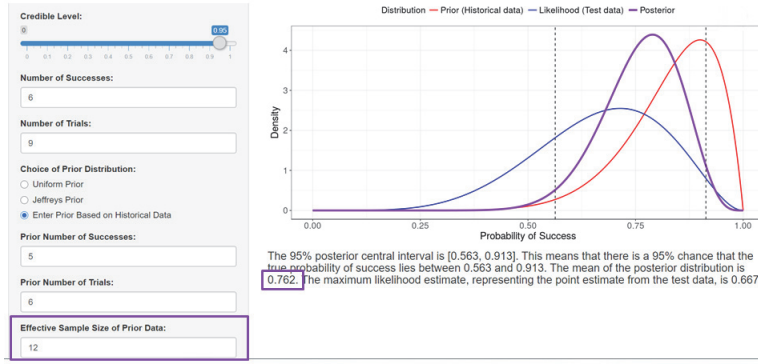
The prior effective sample size is one way of controlling how much weight we are putting on the prior.

According to Wiesenfarth and Calderazzo (2018), we could think about the prior effective sample size as if the prior came from historical data with a certain number of observations.

Note that our Beta-Binomial example makes it easy to use the prior effective sample size. For more complicated models, we might want to explore the distribution of the prior, and adjust our beliefs using the **variability** of such prior distributions.

Approved for public release; distribution is unlimited.

Larger prior effective sample size puts more weight on the prior data



A prior is always a choice, and there is no best prior for any real data analysis

Non-informative (Vague) Priors

- A non-informative prior is a prior that provides little information relative to the experiment*
- Useful for real data analysis, because they remove most subjectivity

Informative Priors

- Informative priors contain our beliefs before looking at the data
- Most useful for sequential analysis and regression
- Sensitivity analysis recommended

Weakly Informative Priors

- Some benefits from both sides
- Common practice

Default recommendation

* Box and Tiao (1973)

Bayesian updating: We can summarize the data at the end of the test, or after any run

We could collect n data points, analyze the data, and use the posterior distribution of the analysis as the prior for the next data point(s). That is:

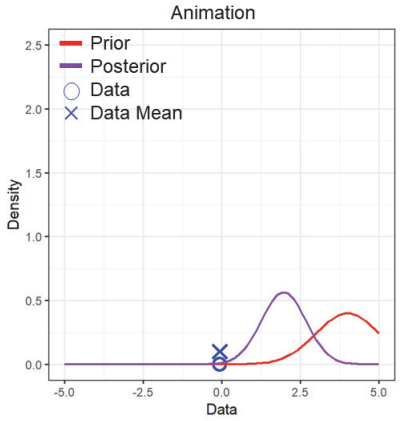
$$P(\theta|y_1, \dots, y_n) \propto P(y_n|\theta) \times P(\theta|y_1, \dots, y_{n-1})$$

$$\propto P(y_n|\theta) \times P(y_{n-1}|\theta) \times P(\theta|y_1, \dots, y_{n-2})$$

$$\propto P(y_n|\theta) \times P(y_{n-1}|\theta) \times \cdots \times P(y_1|\theta)P(\theta)$$

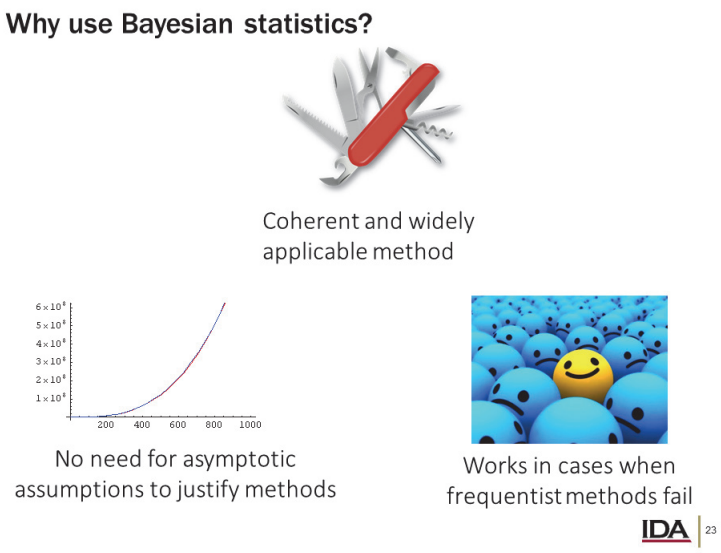
This means that in sequential analysis, there is no difference between analyzing the data in chunks versus all at once.

Bayesian updating in action



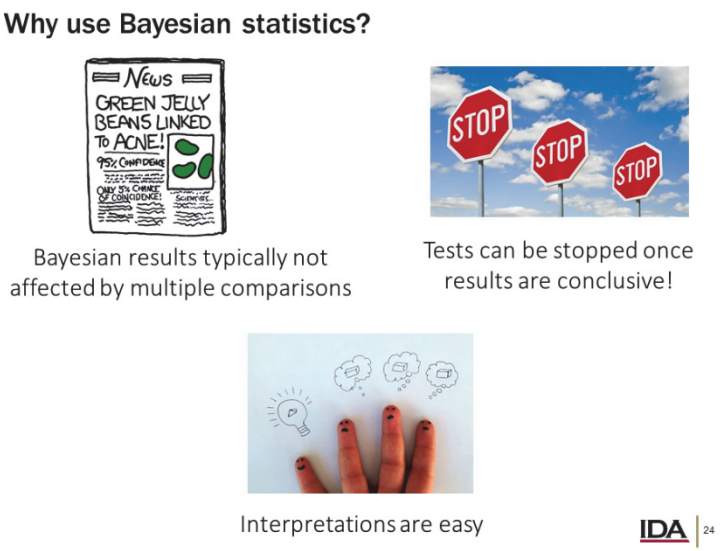
- As new data arrive, the posterior distribution can be updated for each observation
- The prior does not match the data, but the data eventually overpower it
- This is Bayesian "sequential learning"

Approved for public release; distribution is unlimited.



Slide notes:

- Methods can work when frequentist methods fail
  - Frequentist methods may fail if we have lots of parameters and not so much data. Bayes can work, if you choose the right model and priors.
  - Can handle complicated models in a "not so complicated way" and can avoid unrealistic approximations (for example, reliability estimates of 0 or 1)
  - Sometimes it is not trivial to obtain confidence intervals (for example, frequentist methods rely on delta method), but Bayesian intervals are obtained the same way intervals from simple models are obtained
- Coherent and widely applicable method
  - Once you know how to solve one problem, you can solve pretty much everything else (even if there is no prior info). It can be done with our friend R/RStudio. No need for special ways of obtaining intervals (for example, no need for delta method).
  - Formal framework to combine prior information
  - Straightforward way to produce parameter estimates and intervals
  - Computation of inference typically is pretty easy when compared to calculating test statistics using frequentist approach
- Replaces all arguments with one argument
  - Rather than thinking about assumptions, confidence levels, design, etc., worry about the prior and how the data sources are related
  - Inference is based on posterior samples



Slide notes:

- There are different ways of seeing the problem of multiple comparisons. Andrew Gelman talks (among other things) about multiple comparisons on his blog (https://statmodeling.stat.columbia.edu/). Regarding multiple comparisons, he says that the problem with p-hacking is not that you have to adjust your p-value, but that you choose which subset of comparisons to make. Instead, he suggests we use a hierarchical model to analyze all the comparisons at once.
- Interpretations are easy.
  - For example, a Bayesian interval can be interpreted as the probability of containing the parameter of interest.
  - We can compute the probability that $\theta > 0.5$ by just computing mean($\boldsymbol{\theta} > 0.5$), where $\boldsymbol{\theta}$ is the posterior distribution for the parameter.
- All inference is based on the posterior. Note that in the missile launch example, we used a Binomial distribution to model the observed data. Suppose we changed our plan. Instead of having 9 runs, we launched a missile until we had X number of successful launches. In this scenario, we would have used a negative Binomial distribution. For the frequentist approach, the way of computing the estimate and the intervals would have changed. For the Bayesian approach, we would have changed the likelihood distribution but the estimates and intervals would have been computed in the same way as with the Binomial.

Bayesian statistics also has disadvantages

Can be computationally intensive (especially for complicated models)

Learning curve might be steep

Priors can be criticized

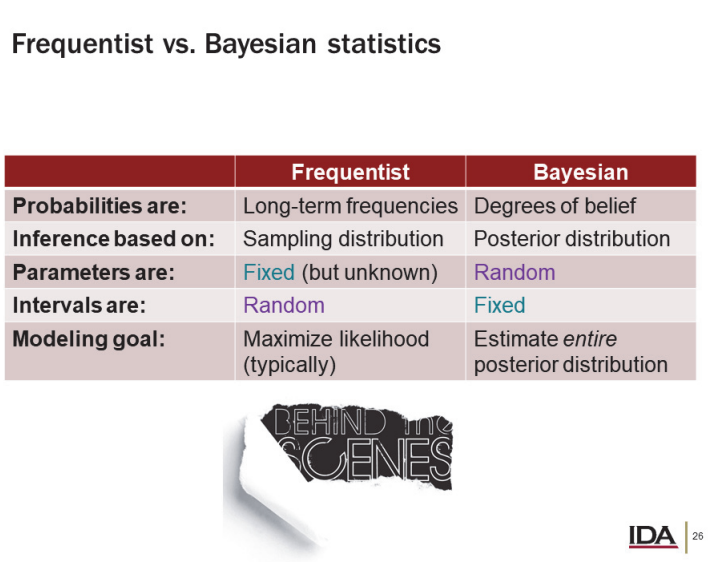
```
parameters {
  real beta0;
  real beta1;
  real<lower=0> sigma2;
}
model {
  // priors
  beta0 ~ normal(0, 30);
  beta1 ~ normal(0, 15);
  sigma2 ~ inv_gamma(0.1, 0.1);
  // Likelihood
  time_spent ~ normal(beta0 + beta1 * dista, sqrt(sigma2));
}
```

Frequentist methods are available in standard software, whereas some Bayesian analyses need to be coded from scratch

Slide notes:

- Can be computationally intensive. This is especially important when working with complicated models like mixed models, hierarchical models, and simulations. It might take some time for the model to run and it might need lots of computer memory.
- Learning curve might be steep and Bayesian statistics can be easy to misuse. It might be difficult to switch from a fully frequentist point of view to a Bayesian one. If the analyst doesn't fully understand Bayesian statistics, they could do things like pooling all data without differentiating the data sources.
- Choosing priors and Bayesian view of probability have been criticized over the years for being subjective.
- Frequentist methods are available in standard software, whereas for many Bayesian analyses the user needs to know how to code. Even if you know how to code in R, you might need to understand JAGS, Stan, or other software that will allow you to fit your Bayesian model (this includes specifying your priors rather than using the default options). The rstanarm package is available for analyzing common models with common priors.

Approved for public release; distribution is unlimited.



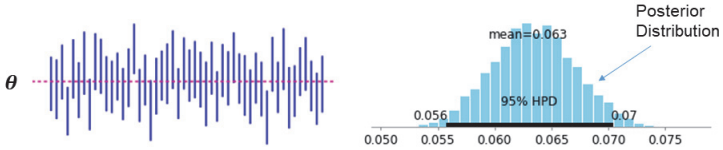
Slide notes:

Our reports usually include a point estimate, an interval, and an interpretation in an operational context. Behind the scenes, we need to understand that these numbers are based on different theories and, therefore, their interpretations are not the same.

- Probabilities
  - Frequentist: Suppose we can repeat the same experiment many, many, many times under the same conditions and collect the same number of samples.
  - Bayesian: How much trust do I have in this result? There are no other hypothetical experiments; there are only the data at hand.
- Inference
  - Sampling distribution: This is the distribution of estimates we could observe if we repeated the experiment many times.
  - Posterior distribution: Sampling from the posterior regardless of its form.
- Parameters
  - Frequentist: The parameters are true/unobserved fixed quantities.
  - Bayesian: Parameters have a distribution, which we adjust depending on the data observed.
- Intervals
  - Frequentist: Interval bounds are random, because if we conduct the experiment again, we would get a different interval.
  - Bayesian: The bounds are fixed quantities. The posterior distribution is fixed (the parameter is random, but its distribution is not), so any intervals based on the posterior are also fixed.
- Modeling
  - Frequentist: Maximize the likelihood to obtain estimates.
  - Bayesian: Estimate the entire posterior distribution to obtain estimates.

Approved for public release; distribution is unlimited.

Confidence and credible intervals express a range of plausible values for a parameter or effect



Confidence (Frequentist) Interval:	Credible (Bayesian) Interval:
Under repeated sampling, a 95% confidence interval will cover the parameter θ 95% of the time	A 95% credible interval contains the parameter θ with probability 95%

HPD = Highest Posterior Density

IDA 27

When and how to use Bayesian statistics?

It depends…

Test Science thinks analysts should be pragmatic and weigh the advantages and disadvantages of Bayesian analysis. But here's some rules of thumb:

- Consider using frequentist methods when the problem can be solved with standard methods and there is no prior information
- Consider using Bayesian approaches when there is good prior information available, or when a frequentist approach doesn't cut it

IDA 28

Slide notes:

Incorporating the prior information depends on multiple factors. For example, how much trust do we put in that prior information? What type of prior information do we have at hand? Subject matter expert intuition? Data from a test conducted 10 years ago? Data from a test conducted on a production-representative system with a similar protocol to the new test, and during the same acquisition timeline?

Approved for public release; distribution is unlimited.

Bayesian analysis is a different (but principled) way of analyzing data that offers numerous advantages to both the researcher and the decision-maker.

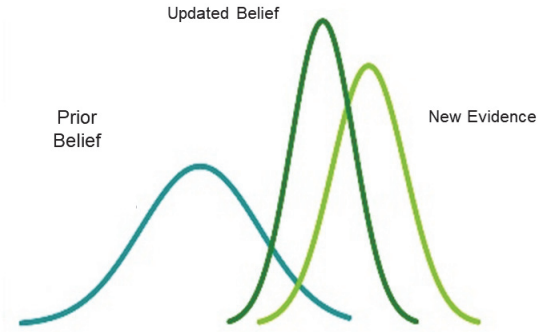
Approved for public release; distribution is unlimited.

Appendix B
Single-Parameter Annotated Briefing

IDA

Introduction to Bayesian Analysis

Section II – Single-Parameter Models



Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

"If I'm doing an experiment to save the world, I better use my prior." – Andrew Gelman

Slide notes:

In this section, we show some ways of combining our prior knowledge with the observed data. The examples we use are simple and therefore limited to specific scenarios. However, they serve as the base for the more complex models we will see in the next section.

Conjugate Priors

Approved for public release; distribution is unlimited.

Conjugate priors guarantee posterior is same distribution as the prior

- Determined by the data likelihood
- Justified by invariance reasoning
- Highly tractable

Toy Example:

- $P(y|\mu) = Normal(\mu, 1)$ — Likelihood
- $P(\mu) = Normal(\mu_0, 1)$ — Conjugate prior
- $P(\mu|y) = Normal\left(\frac{\mu_0+y}{2}, \frac{1}{2}\right)$ — Posterior is the same "shape" as the prior!

IDA 32

This table shows common conjugate models

Likelihood	Parameter	Prior	Posterior
$Binomial(\boldsymbol{n}, \boldsymbol{\theta})$	$0 \leq \theta \leq 1$	$Beta(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$Beta(\alpha', \beta')$ $\alpha' = \alpha + y$ $\beta' = \beta + n - y$
$Poisson(\lambda)$	$\lambda > 0$	$Gamma(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$Gamma(\alpha', \beta')$ $\alpha' = \alpha + n$ $\beta' = \beta + \sum t$
$Exponential(\lambda)$	$\lambda > 0$	$Gamma(\alpha, \beta)$ $\alpha > 0, \beta > 0$	$Gamma(\alpha', \beta')$ $\alpha' = \alpha + n$ $\beta' = \beta + \sum t$

This Wikipedia page includes more examples of conjugate distributions: https://en.wikipedia.org/wiki/Conjugate_prior IDA 33

Slide notes:

A conjugate prior makes sure the posterior is the same distribution as the prior.

- It is determined by the data likelihood.
- It is justified by invariance reasoning: When data modify the prior, the information is limited, so they should change *only* the prior parameters, not the structure of the prior.
- It is highly tractable: Posteriors are relatively simple math formulas from known distributions, and therefore, obtaining posterior samples is not too complicated.

Approved for public release; distribution is unlimited.

Incorporating Legacy Data into Evaluation: Beta-Binomial

IDA | 34

Conjugate priors make incorporating prior information straightforward

- Remember the missile example, where the outcome of interest was "successful launch" or "failed launch."
- The set of all runs follows a Binomial distribution with $n - y$ failures, y successes, and a probability of success θ.

$$P(y|\theta) = \binom{n}{y}\theta^{y}(1-\theta)^{n-y} \propto \theta^{y}(1-\theta)^{n-y}$$

- The probability of success is a continuous quantity bounded by 0 and 1. Therefore, we could use a Beta distribution as the prior for θ.

$$P(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

- The posterior distribution is then a Beta distribution.

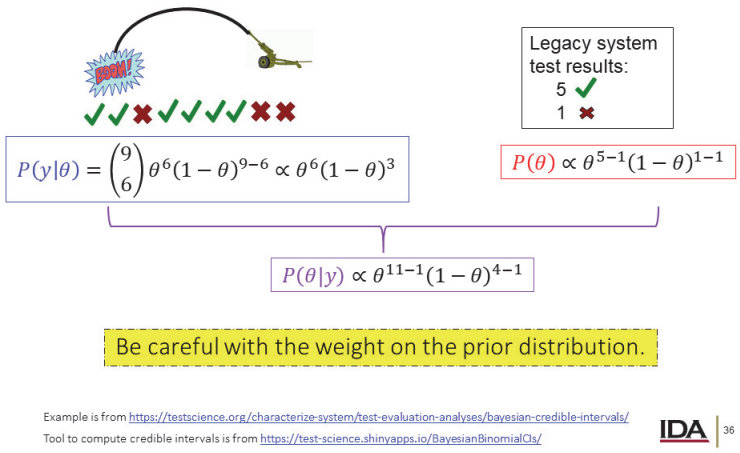
$$P(\theta|y) \propto \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} = \theta^{\alpha^{*}-1}(1-\theta)^{\beta^{*}-1}$$

IDA | 35

Slide notes:

- Note that y is our "data" from the previous section.
- Each of the test runs can be seen as a Bernoulli trial.
- We can work with the kernel of the density. That is, we don't need to keep the constants, just the terms that involve the parameter of interest (in this case, θ).
- We could think of α as the prior number of successes and β as the prior number of failures, assuming we place equal weight on the prior and observed data.

Approved for public release; distribution is unlimited.



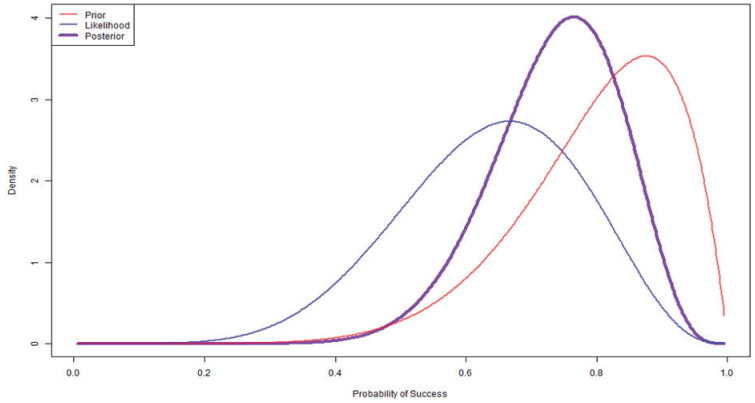
Slide notes:

Although this is a good "textbook" example, it does not always translate to real life, so be careful.

- In this example, we are combining the data as if the prior and observed data have the same weight. This is not common in real life. We saw in Section 1 how the weight changes the posterior results.
- If the previous data have many observations (as compared to the new data), we might be putting too much weight on the prior. That is, the larger sample size in the historical data could drive the results.
- We will talk more about priors later in this section.

If the data collection had been different, we would just need to change the likelihood to a Negative Binomial (for example) and rethink the prior. The actual method of computing the point estimate and credible intervals would not change. This is not true for frequentist statistics.

Conjugate priors helped us obtain results from a Beta distribution in our missile launch system example



Tool to compute credible intervals and obtain plot is from https://test-science.shinyapps.io/BayesianBinomialCIs/

IDA | 37

Incorporating all available information leads to less uncertainty in our estimation

Frequentist:

Point estimate is $\hat{\theta} = \frac{y}{n} = \frac{6}{9} = 0.67$, and the confidence interval is

$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} = (0.36, 0.98)^*$

Confidence Interval Interpretation: *If the test were repeated an infinite number of times and we constructed a confidence interval each time, then 95% of the confidence intervals will contain the true probability of a successful launch* (θ)

Bayesian:

Point estimate and credible interval for $\hat{\theta}$ are computed using the posterior distribution (mean, median, quantiles). The posterior mean and median are 0.73 and 0.74, respectively.

Credible Interval Interpretation: *The probability that* θ *is in the interval of* $(0.49, 0.92)$ *is* 0.95

* Using the central limit theorem. If we use the Wilson Score, the interval is (0.35, 0.98).

IDA | 38

Slide notes:

Even though the textbook examples for binomial intervals use the central limit theorem, to be fair we might want to use the Wilson Score or an exact method.

Do we use the posterior mean or median?

We have the entire posterior distribution, but how do you make an estimate? It depends on how you measure error:

Type of Loss	Loss Equation	Best* Estimate
Squared error loss	$(\theta - \hat{\theta})^2$	Posterior mean
Absolute error loss	$\lvert\theta - \hat{\theta}\rvert$	Posterior median

Practical advice: For skewed posterior distributions, we favor the median. For symmetric distributions, the difference is negligible.

* Best in terms of minimizing error.

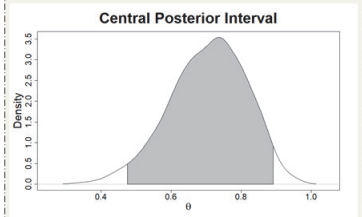
IDA | 39

Slide notes:

Loss functions can be seen as the "loss" or "cost" of estimating a quantity of interest, given the true value of such a quantity. That is, how different/bad/far is your estimate from the true value?

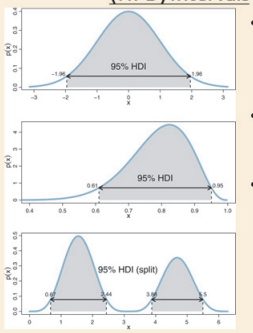
There is more than one way to summarize the posterior uncertainty in the parameter of interest

Central Posterior Intervals



- Compute using density quantiles
- Same probability on each tail
- Invariance property (one-to-one transformations of θ)

Highest Posterior Density (HPD) Intervals



- Density is smaller outside of the interval
- Not necessarily connected
- More accurate than central intervals

Note: For symmetric and unimodal distributions, both intervals will be the same.

HDI image source: Kruschke, John. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press, 2014.

IDA | 40

Slide notes:

Gelman et al. mention in their book (*Bayesian Data Analysis, 2nd Ed.*) that many times we prefer the central posterior interval because it is invariant to transformations, has a direct interpretation (posterior $\frac{\alpha}{2}$ *and* $1 - \frac{\alpha}{2}$ quantiles, same amount in both tails), and is easy to compute.

Other people like the HPD because it is usually a tighter interval, and it's better when we have bimodal or highly skewed distributions. That is, every point inside the interval has a higher density than the points outside the interval. This is only true for central posterior intervals when the interval is symmetric and unimodal.

Slide notes:

See the R script.

Implementation using R

IDA 41

Slide notes:

In cases where the posterior distribution is not the result of a conjugate prior, we need to rely on other methods to sample from the posterior.

Stan is a computer language for writing models and drawing samples from those models.

We talk more about the process/algorithm behind Stan later in this section. For now, let's focus on understanding what Stan is and its code structure.

https://mc-stan.org/

IDA 42

What is Stan?

- Stan is a Bayesian model specification (and compilation) language
- Started at Columbia University in 2012
- Open source
- Inferential tasks are generic
- Access to Bayesian statistical models

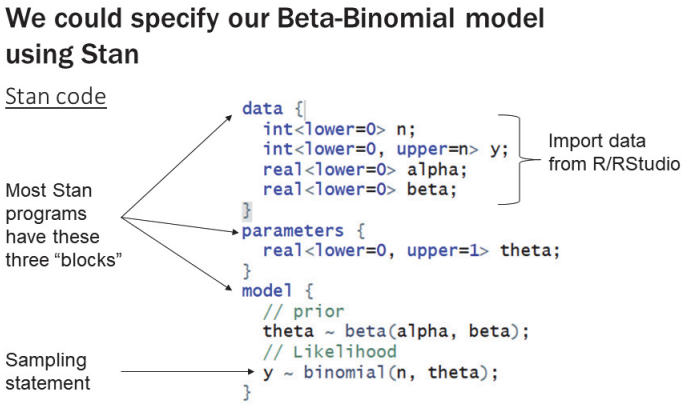
```
data {
  int<lower=0> n;
  int<lower=0, upper=n> y;
  real<lower=0> alpha;
  real<lower=0> beta;
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  // prior
  theta ~ beta(alpha, beta);
  // Likelihood
  y ~ binomial(n, theta);
}
```

IDA | 43

Slide notes:

- Stan is a Bayesian model specification (and compilation) language
  - Model assumptions are specified at the beginning of the model process
  - Stan is a language specifically for writing models
  - Model building is made transparent in Stan
- Started at Columbia University in 2012
  - Under active development
- Open source
  - Available as an R package (for example, rstan)
- Inferential tasks are generic
  - Uncertainty quantification is done by calculating summary statistics from posterior samples
  - That samples are drawn from the model (the model is treated as a random number generator) makes uncertainty quantification very straightforward, so Stan makes inferential tasks generic
- Access to Bayesian statistical models
  - Flexibility to fit a large array of models

Approved for public release; distribution is unlimited.



Slide notes:

For this conjugate model example, we could obtain the posterior distribution without using more complicated code like Stan. If we wanted to use Stan, we could use the code on this slide.

Code details:

- We define the data, parameters, and model.
  - What data are important to the model?
  - What parameters do we care about?
  - What are the relationships?
  - Each block of code is within { } and each line ends with ;
- Note that n and y (number of runs and number of successes) are integers, but alpha, beta, and theta can be any real number.
- The information inside the <> are the constraints. We are using lower and upper bounds.
- Inside the model block, we define our prior as $\theta \sim Beta(\alpha, \beta)$, and our likelihood as $y \sim Binomial(n, \theta)$
- Commented lines start with //
- You'd need to save this model in a .stan file.

We use R to run our model

R code

Data to be imported to Stan

Details for the model fit

```
library(rstan)
set.seed(1638)
data_bb <- list(n = obs_suc + obs_fail, y = obs_suc,
                alpha = pri_suc, beta = pri_fail)

fit_bb <- stan(file = paste0(path, "R code/beta_binom.stan"),
               data = data_bb,
               pars = "theta", chains = 4, iter = 5000,
               warmup = 1000,
               init = list(list("theta" = 0.01),
                           list("theta" = 0.3),
                           list("theta" = 0.7),
                           list("theta" = 0.99)))
```

IDA 45

Slide notes:

R code details:

- Load the R package (install it if you haven't yet).
- (Optional) Set a seed so you can reproduce your results. Otherwise, there might be some sampling variability – minimal but there. You could also do this inside the stan() function with the seed option.
- Create a list with your data. Note that this is NOT an R dataframe. You could've also created a list with the initial values for the chains.
- Run your model. In this case, we are using the rstan package with the stan() function.
  - You'd need to specify the path of the model file and the lis with the data. Everything else could be left to the Stan defaults.
  - pars, chains, iter, and warmup are parameters to follow (ir this case, theta), the number of chains, the number of iterations, and the warmup or burn-in period, respectively.
  - You could let Stan set the initial chain values or you could specify them yourself (recommended for complex models by using the init option. Note that we are running 4 chains and, therefore, we need 4 initial values (one per chain).

Resources

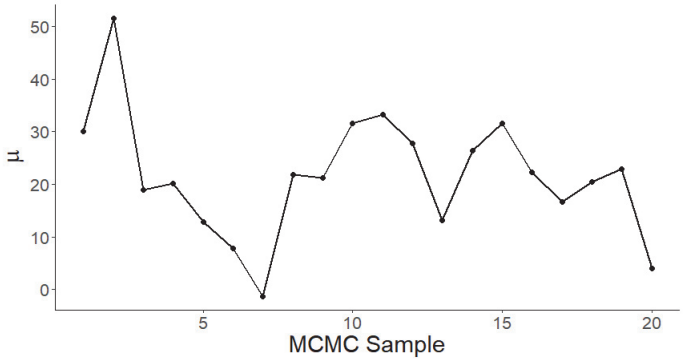
- Tutorials
- Videos
- Reference manual
- Case studies

Slide notes:

There are many references available on Stan's website, mc-stan.org. Notably, some people have called Stan's reference manual not just a good manual, but one of the best applied statistics textbooks you can read. And it's free!

Approved for public release; distribution is unlimited.

Markov Chain Monte Carlo is a method that allows us to sample from the posterior distribution



IDA 47

Slide notes:

- Many times, we are working with posterior distributions that are not the result of conjugate priors; rather, the posterior does not have a closed form (for example, the posterior does not follow a Normal or a Beta distribution).
- Direct sampling from those more complicated posterior distributions is not possible.
- Markov Chain Monte Carlo (MCMC) is a way of sampling from the posterior distribution.
- There are different algorithms to implement MCMC.
  - Stan uses Hamiltonian Monte Carlo, which explores th entire space of all the parameters at each step.
  - JAGS uses different methods (Gibbs, Metropolis, slice sampling, etc.), which update one scalar parameter spa at a time.
  - You could also program your own MCMC.
- This slide shows an algorithm used to sample from the posterior using MCMC.
  - See how the random samples are covering the space of the posterior distribution.
  - The current point/sample only depends on the previous sample.
  - Eventually, there will be more points close to the paramet estimate.
  - Videos of Metropolis implementation are here: https://www.youtube.com/watch?v=zL2lg_Nfi80 and https://www.youtube.com/watch?v=4I6TaYo9j_Y
- Imagine we only take the first 10 posterior samples.
  - What would be the posterior average or median?
  - What would be the credible interval?
  - How much variability would we have?
- Now imagine that we get rid of the first few steps – the ones prior to convergence. How would things change?
  - A conservative choice given by Gelman et al. in their book (Bayesian Data Analysis, 2nd Ed.) is to discard the first half of the iterations.
  - These discarded iterations are known as the "burn-in period."

First, look at the estimated posterior distribution

IDA 48

Slide notes:

- We need to check the model output before we make any inference.
- We should discard the iterations before the chains' convergence (burn-in period) and make inference with the remaining samples.
- This will be more relevant when we start working with more complicated models, such as when we do not have a conjugate prior or when more parameters are involved in the model. In our example, the chains converged pretty quickly.

We need to make sure the models are consistent with the data

Convergence Assessment

- Trace plots
- Brooks-Gelman-Rubin

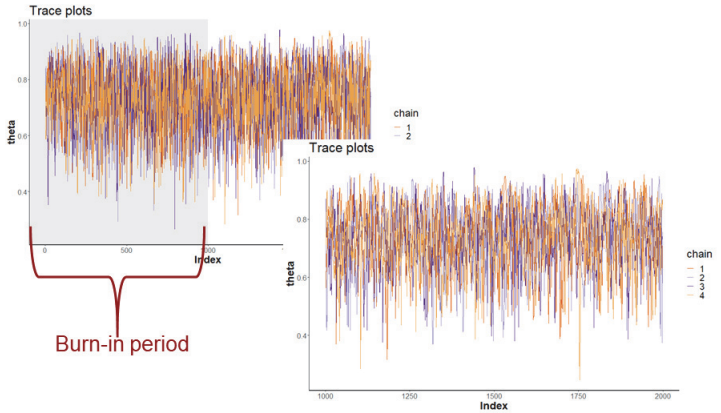
Efficiency Assessment

- Markov Chain errors
- Effective sample size

Note that these are just a few examples of simple descriptive statistics and plots. (Quick check, no inferential methods.)

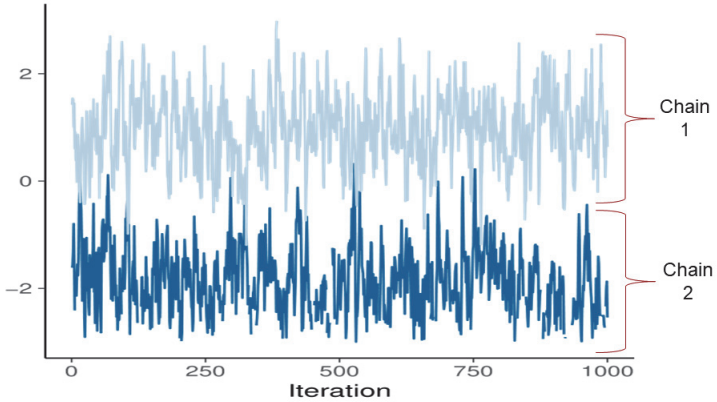
IDA 49

We can perform a qualitative assessment of the convergence by looking at the trace plots



IDA 50

We want to avoid trace plots where the chains do not overlap or mix well

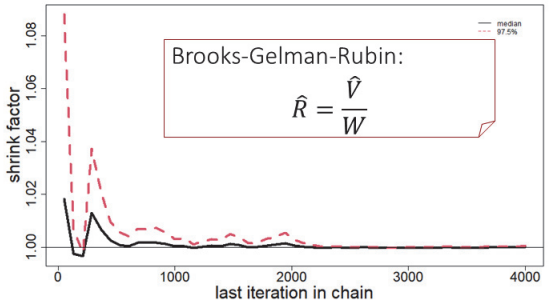


IDA 51

Slide notes:

- A chain is a sequence of samples from the posterior distribution. In these figures we have 4 chains.
- Use more than one chain with different starting points to evaluate trace plots.
- Look for chain "mixing."
- If we have a multimodal posterior, your trace plots will not look like this. However, most of the time this trace plot assessment will work.
- The burn-in period in this example is mainly for illustration purposes – our model converged quickly so we could have had a shorter burn-in.

There are also quantitative methods for convergence assessment



$\hat{R}$ compares the variability within each chain's output to the variability of the pooled samples of all chains.

IDA 52

Slide notes:

- $\hat{R}$ was proposed in 1992 (Gelman-Rubin) but corrected and generalized in 1998 (Brooks-Gelman).
  - The correction is $\hat{R} = \frac{d+3}{d+1}\frac{\hat{V}}{W}$, where d is the degrees of freedom.
  - As the number of Markov Chain Monte Carlo samples increases, *d* tends to be large and it doesn't affect the origi $\hat{R}$.
- $\hat{V}$ is a weighted average of the between- and within-chain variability, and W is the within-chain variability
- If $\hat{R}$ is not near to 1, we have reason to think that convergence has not been met and that more runs could improve inference.
- How close? Some references say $\hat{R}$ should be < 1.2, and others say 1.1, but they agree it depends on the problem and how precise we want to be with our inferences.
- Requires 2 or 3 parallel chains.

Approved for public release; distribution is unlimited.

One way to assess efficiency of an MCMC sampler is by computing the posterior effective sample size

```
Inference for the input samples (4 chains: each with iter=4000; warmup=0):

       mean se_mean    sd    2.5%     50%  97.5% n_eff  Rhat
theta 0.731   0.001 0.111   0.489   0.742   0.916  5974 1.001
lp__ -9.226   0.010 0.764 -11.326 -8.936 -8.699  6089 1.000

For each parameter, n_eff is a crude measure of effective sample size,
and Rhat is the potential scale reduction factor on split chains (at
convergence, Rhat=1).
```

$$n_{eff} = \frac{ML}{1 + 2\sum_{t=1}^{\infty} \rho_t}$$

The posterior effective sample size of the chain is the number of independent MCMC samples.

MCMC = Markov Chain Monte Carlo; n_eff = effective sample size; sd = standard deviation; se_mean = standard error of the mean

Slide notes:

$n_{eff} = \frac{ML}{1+2\sum_{t=1}^{T} \rho_t}$, where M is the number of chains, L is the number of samples within each chain, and ρ is the autocorrelation at lag t.

- The number of posterior effective sample sizes (ESSs) that you need depends on the estimate you want to obtain.
  - For areas with small density (for example, the highest posterior density interval limits), a large ESS is recommended.
  - If we want to know about estimates from areas with more density (for example, posterior median), a smaller ESS mi be adequate.
- For some chains that have negative autocorrelations, the ESS could be larger than the total number of MCMC samples.
- Given that the MCMC samples are not independent, n_{eff} is used to estimate the standard error as opposed to using the total number of samples.

Approved for public release; distribution is unlimited.

Another way to assess the performance of an MCMC sampler is by computing the MCSE

- MCSE is a measure of computer simulation error
  - If we had performed infinitely many runs, the simulation error would be 0
  - In simple models, it is very easy to get MCSE near 0
- Measures the amount of uncertainty in the posterior mean

$$MCSE(\bar{\theta}) = \frac{s}{\sqrt{L}}$$

- For non-independent samples, use the posterior effective sample size instead of the number of iterations (L)

MCMC = Markov Chain Monte Carlo; MCSE = Monte Carlo Standard Error

IDA | 54

Slide notes:

- Measures the amount of uncertainty in the posterior mean, $MCSE(\bar{\theta}) = \frac{s}{\sqrt{L}}$, where $\bar{\theta} = \frac{\sum_{l=1}^{L} \theta^l}{L}$ (average of all θ samples), s is the sample standard deviation, and L is the total number of iterations/samples.
- MCMC generally are not independent, and the MCSE will be higher than that of an independent sample.
- One way to calculate the MCSE with autocorrelated samples is to use the posterior effective sample size instead of the sample size.
- The measures of effective sample size and MCSE suggest how stable and accurate the chain is.
- A rule of thumb is that an MCMC sampler should be run long enough after convergence for the MC error to be less than 1/20 as large as the estimated posterior standard deviation of the parameter.

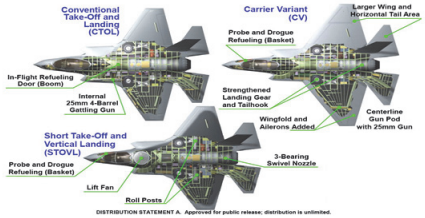
Approved for public release; distribution is unlimited.

Implementation using R and Stan

Incorporating Engineers' Intuition into Evaluation: Exponential-Gamma

There are ways to incorporate a subject matter expert's knowledge into our analyses

- The F-35 cost per flying hour estimate is affected by fleet reliability (for example, maintenance costs)
- At the beginning of the program, the only available data are engineer estimates of reliability
- Traditional methods compute the mean flight hours between repair (MFHBR) as $\text{MFHBR} = \frac{Total\ Flight\ Hours}{Failures} = \frac{FH}{N}$



https://dataworks.testscience.org/wp-content/uploads/sites/4/formidable/7/BramMedlin_DATAWorks19.pdf

Challenges and possible solutions

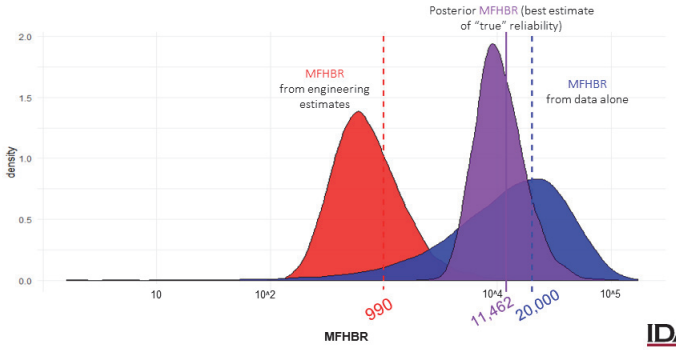
- Traditional methods struggle with reliability estimation if $n = 0$ (no failures)
- What happens if the engineer estimate is 990, the flight hours to date are 40,000, and there have been 2 observed failures?
  - Should we use the traditional method and estimate MFHBR = 20,000 hours and ignore the subject matter expert's intuition?
  - Average the engineer estimate and the traditional estimate?
  - Put more weight on one source of information?
- Bayesian analysis combines the engineering estimates and the actual failure data that are available

Approved for public release; distribution is unlimited.

Bayesian statistics combine "prior" knowledge with observed data to produce an estimate

Example for Component X:

- Engineering Estimate MFHBR = 990 hours (red "prior" below)
- Flight Hours to Date: 40,000 hours
- Observed 2 Failures
- Traditional Methods Estimate: MFHBR = 40,000 / 2 = 20,000 hours (blue "likelihood" below)



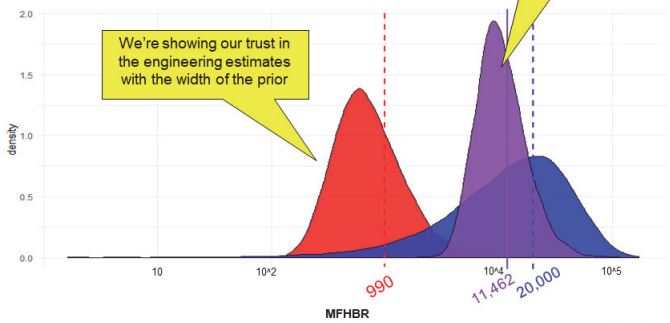
IDA 59

Bayesian statistics combine "prior" knowledge with observed data to produce an estimate

Example for Component X:

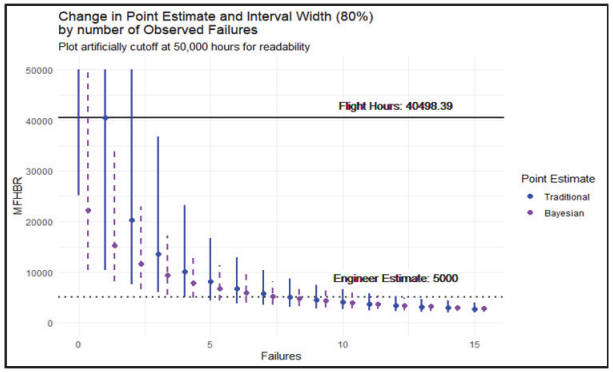
- Engineering Estimate MFHBR = 990 hours (red "prior" below)
- Flight Hours to Date: 40,000 hours
- Observed 2 Failures
- Traditional Methods Estimate: MFHBR = 40,000 / 2 = 20,000 hours (blue "likelihood" below)

The final estimate is influenced significantly by the subject matter expert estimate because
1. Few data (failures) exist
2. We chose a narrow distribution for the prior



IDA 60

A robust methodology for many cases



IDA 61

How did we obtain those results?

- Suppose we have observed n independent failure times, say, $t_1, t_2, \ldots t_n$, which follow an Exponential distribution

$$P(\sum t|\lambda) \propto \lambda^n \, Exponential(\lambda \sum t), \qquad t \geq 0, \qquad \lambda > 0$$

where the exponential rate parameter λ is estimated by

$$\frac{n}{\sum_i t_i} = \frac{Total\ number\ failures}{Total\ test\ time} = \frac{1}{MFHBR}$$

- The engineer estimate of the MFHBR is a value greater than 0 and we would like to control the amount of weight we put on the prior. Therefore, we could use a Gamma distribution to represent the prior information.

$$P(\lambda) \propto \lambda^{\alpha-1} \, Exponential(-\beta\lambda)$$

IDA 62

Slide notes:

- Bayesian method appropriately moves MFHBR estimate toward the traditional result as the available data increase.
- The approach also automatically handles cases where N = 0 (something not satisfactorily handled with traditional approaches).
- As usual, be careful with the weight on the prior information. If the engineer estimate is way off and we put lots of weight in this prior, it will drive the posterior results.

Choose the prior parameters based on subject matter expert's inputs

- For λ to have a Gamma prior distribution, the MFHBR must have an inverse Gamma prior distribution

$$P(MFHBR) \propto MFHBR^{-(\alpha+1)}\, Exponential(-\beta/MFHBR)$$

- We can use the engineer estimate as the mean of the prior distribution

$$Mean = \frac{\beta}{\alpha - 1} = MFHBR_{eng\ est}$$

- A similar process gives us an estimate of the prior standard deviation
- Using the prior mean and standard deviation, we can determine parameters for the inverse Gamma prior distribution (α, β)

For more details, see the 2018 IDA memo "Estimating JSF Component Reliability."

IDA | 63

Obtaining the posterior distribution is not too complicated, and we reduce uncertainty

By combining the failure and time data with the engineer estimates, we obtain the posterior distribution

$$P(\sum t|\lambda) \propto \lambda^n\, Exponential(\lambda \sum t)$$
$$P(\lambda) \propto \lambda^{\alpha-1}\, Exponential(-\beta\lambda)$$
$$P(\lambda|\sum t) \propto \lambda^{\alpha-1+n}\, Exponential\left(\lambda(\sum t + \beta)\right)$$

which is a Gamma distribution with:

$$\alpha' = \alpha + n$$
$$\beta' = \beta + \sum t$$

For our example with 2 failures and 40,000 flight hours we have:

Method	MFHBR Estimate and 80% Interval
Bayesian	10,056 (5,773–19,846)
Frequentist	20,000 (7,516–75,215)

IDA | 64

Slide notes:

Using the estimated α and β values, we can create our prior distribution for λ.

Approved for public release; distribution is unlimited.

We must use the ENTIRE posterior distribution of λ in order to make inference about $MFHBR$

- The posterior samples can be simulated for any transformation $f(\lambda)$ of λ
- In our case this transformation is $\frac{1}{\lambda} = MFHBR$
- First, we simulate samples from the posterior $P(\lambda|\sum t)$
- Then, we transform each draw and use those transformed draws to represent draws from $P(MFHBR|\sum t)$
- Note that we are *NOT* obtaining a point estimate $(\hat{\lambda})$ and transforming this point estimate; rather, we are using the *ENTIRE* distribution to obtain our point estimate for $MFHBR$

IDA 65

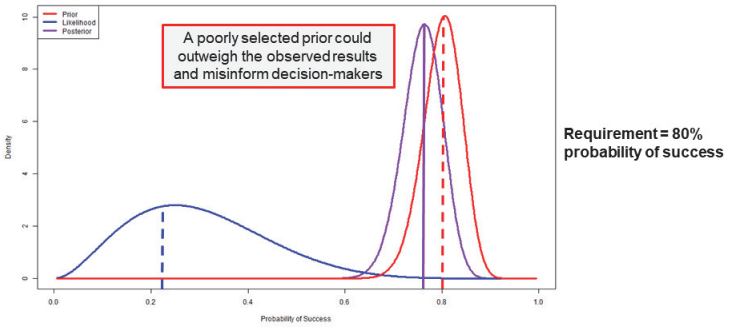
Some notes on prior distributions

Slide notes:

- Priors allow us to consider multiple sources of data and thereby use all available information in our analysis.
- Models that use conjugate priors are easy to implement but sometimes we need more flexible models (next section).
- Bayesian statistics are useful when we do not have lots of data (help us avoid unrealistic results when there are 0 failures).

IDA 66

Prior distribution is the key to Bayesian inference but choosing the prior is not trivial



Always put careful thought into the prior – naively specified priors can lead to misleading results.

IDA | 67

Slide notes:

- Prior distribution is the key to Bayesian inference; its selection might be the most important step in the process.
- There is no such thing as *the* prior distribution, except in some very special situations.
- Influence of the prior can be negligible, moderate, or enormous – again, this depends on how much weight we put on such a prior.
- Bayesian estimates interpolate between prior and frequentist estimates.
- This can be problematic if the prior is chosen to be a threshold.
- A requirement is NOT a good choice for the prior. It could be used in situations where the subject matter expert or the prior data suggest otherwise.

Prior distribution considerations

- Choose a prior based on the available information:
  - Centered at a specific value and with equivalent prior sample size
  - Draw a few distributions and see which one represents your prior beliefs (center, spread/intervals)
- Keep in mind how much your posterior is influenced by the choice of the prior (sensitivity analysis)
- If the prior density is improper, make sure the posterior distribution is proper
- Remember: There is no single "correct" prior in practice!

A proper density is a distribution that integrates to one.

IDA | 68

What if I don't have any prior information?

- Flat Prior:
  - Works best when parameter range is finite
  - Example: $\pi(\theta) = 1$
- Vague Prior
  - Large variability
  - Example: $\theta \sim Normal(0, 100)$
- Compound Prior
  - Avoid sensitivity to prior variance by assigning a distribution
  - Example: $\theta \sim Normal(0, \sigma_0)$
    $\sigma_0 \sim Half - Cauchy$

Commonly lead to similar results

When no direct information about θ is available, a good approach to prior selection is through the marginal distribution of the data

$$m(y) = \int p(y|\theta)\pi(\theta)d\theta$$

$m(y)$ is what we expect the data to look like

IDA | 69

Slide notes:

Flat Prior: $\pi(\theta) = 1$

- This prior may be improper
- Flat is not non-informative but often yields the same results as a frequentist analysis
- **Not** flat if θ is transformed
- Makes mathematical sense, but not good for computations

Vague Prior

- Uses a very large variance in the prior distribution
- Some examples: $\pi(\theta) \sim Beta(1,1)$ or $(\theta) \sim Normal(\mu = 0, \sigma = 100)$

Compound Prior

- Suppose we need a distribution for a real-valued θ
- $\theta \sim N(0,100)$ is one option; the results may be sensitive to the value *100*, and we don't know the variance.
- Take $\theta|a \sim N(0, a)$, *with* $a \sim X_1^2$
- Now the prior is $\pi(\theta) = \int \pi(\theta|a)\pi(a)da$
- We marginalized out the variance of the Normal distribution, leaving a prior with fatter tails.
- This is the most common type of prior now. When in doubt about a hyper-parameter, give it a distribution that expresses our uncertainty.

Simulating from the marginal distribution can calibrate priors

Unlike conventional models, Bayesian models can generate data before any data have been collected.

We can use the generated data to select priors based on whether the data look like what we expect.

Example:

1. I will use a Normal model to evaluate the test data
2. I need to know if $\mu \sim Normal(2, 2), \sigma \sim Exponential(1)$ is a good prior

We can use simulation to calibrate the priors.

Implementation using R

Incorporating Legacy Data into Plan: Exponential

Paladin Integrated Management (PIM) self-propelled howitzer

We can incorporate legacy data to develop a reliability test plan

We need to determine the test size (hours) for operational testing and the number of failures allowed before we declare whether the PIM howitzer meets or does not meet its 64-hour MTBF reliability threshold. There are two errors we can make:

- Test is passed when the vehicle reliability is actually below threshold (consumer risk)
- Test is failed when the reliability is actually above threshold (producer risk)

Traditional DoD demonstration tests are classical hypothesis tests, which use only data from the current test to assess whether reliability requirements have been met.

- ✓ Fix the risk of passing the test given that the true reliability is less than the threshold (consumer risk)
- ✓ Effectively ignore the risk of failing the test given that the true reliability is greater than a threshold (producer risk)
- ✓ Find the minimum test size around a fixed number of failures
- ✓ Often require an exorbitant amount of testing

Bayesian assurance testing leverages information from various sources in an attempt to reduce the amount of testing required to meet a requirement.[1]

- ✓ Fixes the risk that the true reliability is less than the threshold given that the test is passed (consumer risk)
- ✓ Fixes the risk that the true reliability is greater than a threshold given that the test is failed (producer risk)
- ✓ Finds the minimum test around a fixed number of failures
- ✓ By incorporating all available information, typically requires less testing

1. If the information we are looking to incorporate is poor – say, for example, developmental testing suggests poor system reliability – then incorporating this information will buy us no advantages and will not shorten the length of a test.

DoD = Department of Defense; MTBF = Mean Time Between Failures

Approved for public release; distribution is unlimited.

What is the maximum number of failures, c, permitted for a successful test of length T?

Traditional Risk Criteria

$$Consumer's\ Risk = P(Test\ is\ Passed | \lambda = T/MTBF_{Req})$$

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

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

Bayesian Posterior Risk Criteria

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

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

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

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

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.

We can use available data to construct our test plan

Likelihood Distribution

T = 400 Hours of Testing
N = 2 Failures

Available Data, x, from Previous Test Event

$P(x|\lambda) \sim Exponential(\lambda)$

Prior Distribution

Non-Informative Prior

$P(\lambda) \sim Gamma(\alpha = .001, \beta = .001)$

Posterior Distribution

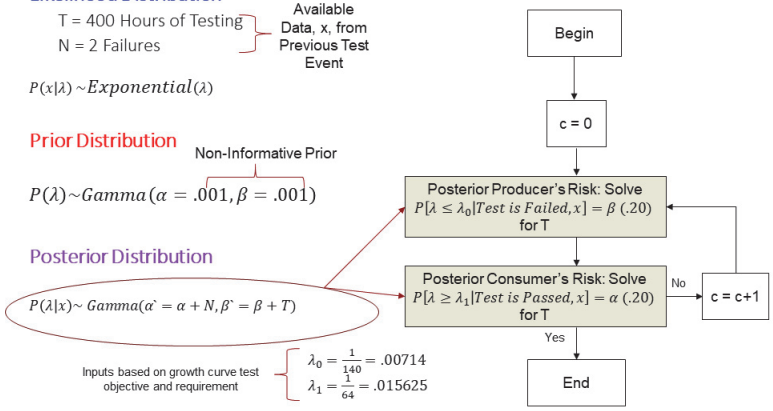
$P(\lambda|x) \sim Gamma(\alpha` = \alpha + N, \beta` = \beta + T)$

Inputs based on growth curve test objective and requirement

$\lambda_0 = \frac{1}{140} = .00714$

$\lambda_1 = \frac{1}{64} = .015625$

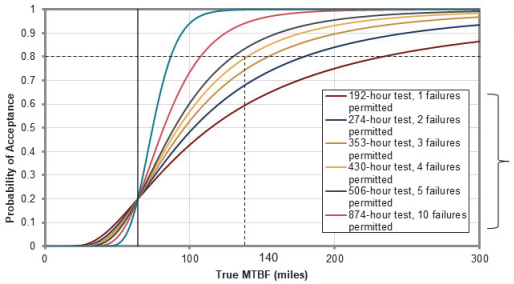
Bayesian Test Plan Algorithm



Evaluating a mission-based threshold with traditional methods requires a very long test

Traditional Operating Characteristic Curves are based on demonstrating 64-hour MTBF.

- Using optimistic assumptions about true **howitzer** reliability, a minimum of **430 hours** of operational testing are required to evaluate **the howitzer's** reliability with $\alpha = 0.2$ and probability of acceptance = 0.80

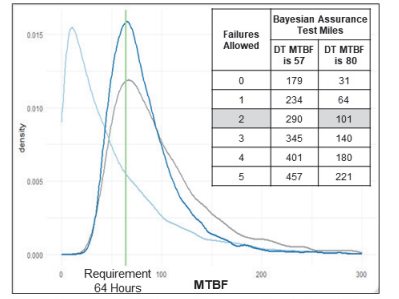


We need to find the combination of test length, T, and maximum allowed failures, c, that satisfies

$$P(y \le c|\lambda) = \sum_{y=0}^{c} \frac{\left(\frac{T}{64}\right)^y \exp\left(-\left(\frac{T}{64}\right)\right)}{y!} \le .2$$

Bayesian assurance testing offers a more powerful and efficient way to assess reliability compared to traditional methods

Bayesian assessment of DT-OT data to assess a 64-mile MMBOMF



Failures Allowed	Bayesian Assurance Test Miles	
	DT MTBF is 57	DT MTBF is 80
0	179	31
1	234	64
2	290	101
3	345	140
4	401	180
5	457	221

When **DT and OT testing** are carried out under **similar conditions**, incorporating DT data into the OT reliability assessment may be reasonable.

- **Bayesian assurance methods** provide a structured way to leverage DT
- Similar to Operational Characteristic Curve assumptions, chart on the left assumes **the howitzer** attains a true reliability of 80 hours
- Permits a more powerful assessment of reliability with fewer miles compared to traditional methods that only consider OT data

DT = Developmental Test; MMBOMF = Mean Miles Between Operational Mission Failures; OT = Operational Test

Approved for public release; distribution is unlimited.

Bayesian statistics can provide more informed estimates and can decrease uncertainty, when used properly.

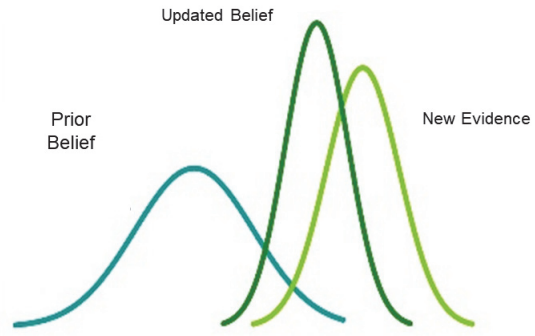
Approved for public release; distribution is unlimited.

Appendix C
Linear Regression Models Annotated Briefing

IDA

Introduction to Bayesian Analysis

Section III – Linear Regression Models



Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

Slide notes:

In this section, we talk about using regression to assess the effect of a factor (or combination of factors) on the response variable.

We also cover some ways to evaluate model fit.

"All models are wrong, but some are useful."
– George Box

Simple Linear Regression

We can assess the effect of some factor on the outcome of interest

IDA 81

Slide notes:

Suppose that we are interested in estimating how a vehicle's weight (wt; 1,000 lb) affects its performance, as measured by mileage per U.S. gallon (mpg).

A common way to analyze the data is to use frequentist statistics

- The regression model we use to analyze the data is

$$mpg = \beta_0 + \beta_1 wt + \varepsilon, \qquad \varepsilon \sim Normal(0, \sigma^2)$$

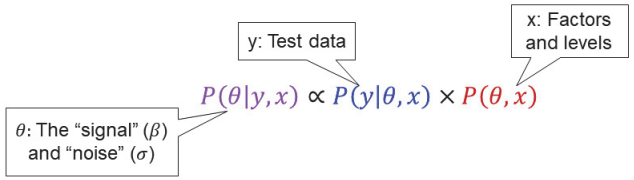
- Check the model assumptions (stats 101)
- Obtain parameter estimates and intervals

IDA 82

Slide notes:

Remember from the stats 101 training: *mpg* is our outcome variable, *wt* is the factor we are controlling, β_0 is the intercept, β_1 is the slope, and ε is the error term.

Bayes' theorem for regression*



Typical assumption: Regression parameters θ do not depend on factors x, so we can simplify

$$P(\theta|y,x) \propto P(y|\theta,x) \times P(\theta)$$

In our example:

$$P(\beta_0, \beta_1, \sigma^2|mpg, wt)$$

$$\propto P(mpg|\beta_0, \beta_1, \sigma^2, wt) \times P(\beta_0) \times P(\beta_1) \times P(\sigma^2)$$

* Note that we add an x to the usual Bayes' theorem because regression depends on covariates.

Approved for public release; distribution is unlimited.

Bayesian analyses use prior distributions to incorporate what we know about the data

- Suppose we have data from a previous test event where slightly different protocols and vehicles were used
- The results from these previous data will help us build the priors for our model
- Increased variability allows us to put less weight on these informative priors

$$\beta_0 \sim Normal\left(\mu_{\beta_0} = 40.8, \sigma_{\beta_0} = 1.63 * 5\right)$$

$$\beta_1 \sim Normal\left(\mu_{\beta_1} = -6.28, \sigma_{\beta_1} = 0.42 * 5\right)$$

$$\sigma^2 \sim Inverse\ Gamma(6.1, 81.08)$$

IDA | 84

Slide notes:

- We want to leverage the prior information but we do not want our prior overpowering the observed data.
  - For example, assume we saw in a previous analysis that as the weight increases, the mileage per gallon will decrease by 6.28. The standard deviation (sd) of this parameter estimate was 0.42. Now, we are centering our prior on the same mean prior but we are increasing the variability.
  - When incorporating these results, we multiply the sd of the results by 5 so there is more variability in the prior.
  - This value of 5 is a baseless number chosen for the sake of this example. In reality, we should make sure our prior distribution matches our belief about the prior information.
- A Normal prior distribution for the βs and an Inverse Gamma for σ^2 are common choices, but choose whichever works the best for your problem.
- We could add or reduce the variability depending on how much confidence we have in the prior information. Keep in mind that this could affect the results.
- You should also keep in mind the parameterization of the statistical software. For example, Stan uses the standard deviations for the Normal distribution parameterization (same as R), but other software use precisions $\left(\frac{1}{\sigma^2}\right)$.

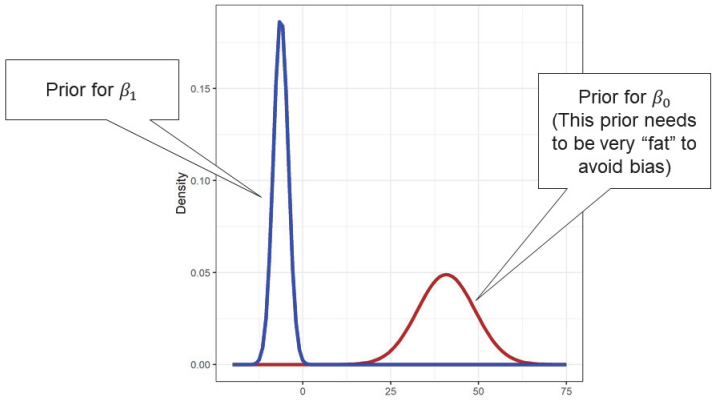
Incorporating prior information in new regression models is not trivial

No general rules for how to do this (I'm sorry …). There might be two types of information:

- Prior information on the regression intercept (β_0)
  - **Be careful:** It could add bias to the overall regression model
- Prior information on the individual factor effects (β_1, β_2, etc.)
  - Less dangerous

IDA 85

Illustration of the prior distributions for the regression coefficients



IDA 86

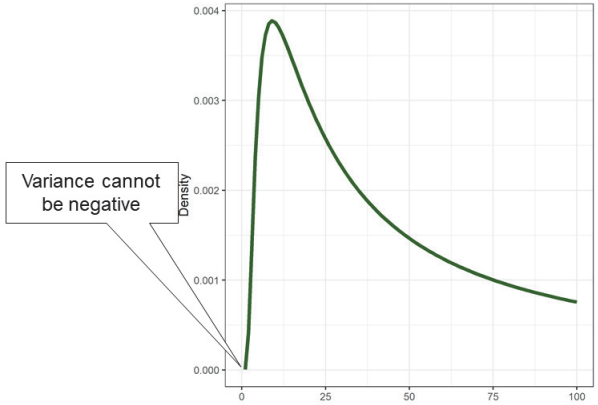
Slide notes:

Note that we have the density distribution with mean equal to the subject matter expert estimate, but with enough variability so that the prior does not dictate the posterior.

We could plot a few distributions and decide which one better reflects the weight we want to put in the prior distribution.

The prior distribution for σ^2 (residual variance)

Remember: The prior doesn't have to be perfect to be useful!



The prior penalizes very small and very large variance.

IDA 87

Complicated posterior distributions cannot be sampled directly!

$$P(\beta_0|\beta_1,\sigma^2,mpg) \propto P(mpg|\beta_0,\beta_1,\sigma^2) \times P(\beta_0) \times P(\beta_1) \times P(\sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{\frac{[mpg_i - (\beta_0 + \beta_1 wt_i)]^2}{2\sigma^2}\right\}$$

$$\times \frac{1}{\sqrt{2\pi 1.63 \times 5}} \exp\left\{\frac{[\beta_0 - 40.8]^2}{2 * (1.63 \times 5)^2}\right\} \times \frac{1}{\sqrt{2\pi 0.42 \times 5}} \exp\left\{\frac{[\beta_1 - (-6.28)]^2}{2 * (0.42 \times 5)^2}\right\}$$

$$\times \frac{81.08^{6.1}}{\Gamma(6.1)} (\sigma^2)^{-(6.1+1)} \exp(-81.08/\sigma^2)$$

$$\propto \exp\left\{\sum \frac{[mpg_i - (\beta_0 + \beta_1 wt_i)]^2}{2\sigma^2}\right\} \times \exp\left\{\frac{\beta_0^2}{132.85}\right\}$$

$P(\theta|Data)$

Updated knowledge

The posterior distribution for each parameter does not have a closed form (no conjugacy). Therefore, we need to simulate values from the posterior distribution via MCMC.

Note that this is **one** posterior distribution; MCMC = Markov Chain Monte Carlo

IDA 88

There are situations in which the results* from both analyses are similar

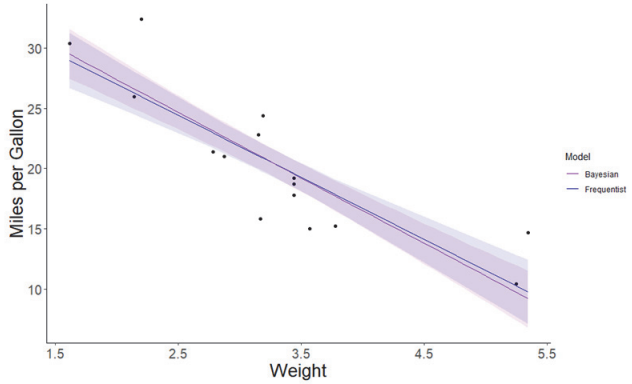
Parameter	Bayesian**	Frequentist
β_0	38.22 (34.64, 41.81)	37.32 (33.27, 41.36)
β_1	-5.42 (-6.45, -4.39)	-5.15 (-6.33, -3.98)

* 80% confidence and credible intervals
** Parameter estimates based on the posterior median

Slide notes:

Results could be similar, especially in simple examples like this one.

We might want to show the results in an operational context



Approved for public release; distribution is unlimited.

Implementation using R

Posterior predictive checks

If our model is a good fit, then we should be able to use it to generate data that look a lot like the data we observed.

Approved for public release; distribution is unlimited.

Posterior *predictive* distributions are used to make inferences about data (not parameters)

We know that

$$P(\theta|y,x) \propto P(y|\theta,x) \times P(\theta)$$

But we want $P(y^{rep}|y,x)$, the distribution of new data given current data (assuming we can replicate the experiment with the same factors)

Luckily, it is rather easy to get this:

$$P(y^{rep}|y,x) = \int P(y^{rep},\theta|y,x)d\theta$$

$$= \int P(y^{rep}|\theta,x)P(\theta|y,x)d\theta$$

IDA | 93

Slide notes:

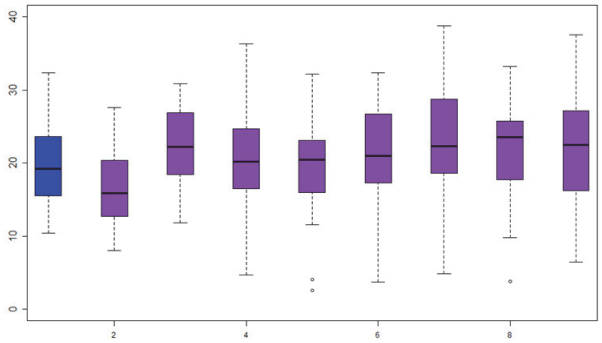
The mathematics is very complicated, but the computation is very simple.

The mathematics suggest that we can sample from the posterior, then "push" those posterior samples back through the likelihood to get the posterior predictive distribution.

This is the general principle in Bayesian analysis: We do not condition on what we do not know. Since we do know y, we can condition on it, but since we do not know theta, we integrate it out.

Approved for public release; distribution is unlimited.

Remember to assess convergence and then make sure that the model fits the data well



Blue = observed data
Purple = replicated data ($\boldsymbol{y^{rep}}$)

Slide notes:

Remember to assess convergence by evaluating trace plots, Gelman-Rubin diagnosis, etc.

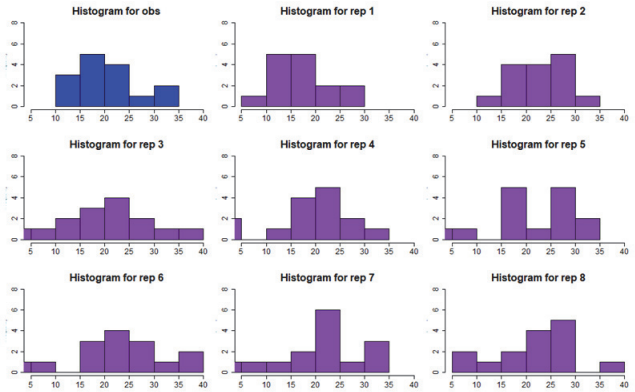
Then, we can use posterior predictive checks to visually evaluate whether there are any model-data discrepancies.

- Sample some data from the posterior predictive distribution. Each replicated dataset ($\boldsymbol{y^{rep}}$) will have the same number of observations as the original/observed dataset.
- Are the replicated data consistent with the observed data? (Hopefully, the answer is yes.)
- Visual inspection is a good start but it might not be enough.

In this example, we have 15 replicated observations per dataset.

Approved for public release; distribution is unlimited.

We can also compare the observed and replicated data using histograms



Blue = observed data

Purple = replicated data (y^{rep})

Slide notes:

If we see discrepancies in the histograms, then we have an indication that the model might not be fitting the data well.

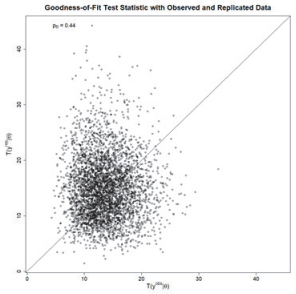
Approved for public release; distribution is unlimited.

The posterior predictive p-value is a quantitative way to assess model fit

Bayesian p-values

$$p_B = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y)$$

where $T(y, \theta)$ is a test statistic (for example, goodness-of-fit, quantile, etc.)



Slide notes:

- Once again, visualizations are a good start but we should leverage other options to assess model fit.
- The posterior predictive p-value, or Bayesian p-value, is a way to compare the observed data with the posterior fit.
- Gelman et al. define Bayesian p-value as "the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity: p_B … "
- A Bayesian p-value close to 0.5 suggests good model fit, whereas a Bayesian p-value close to 0 or 1 indicates bad fit.
- The proportion of points above the diagonal line is about the same as the proportion of points below the line.

Implementation using R

Multiple Linear Regression

Approved for public release; distribution is unlimited.

Analyze the joint effect of all operational test factors at once with multiple regression

Separate signal from noise (smooth the data)

Predict important outcomes

Handle categorical and continuous variables

Determine effects of factors on outcomes

Quantify noise in outcomes

IDA | 99

The general linear model – beyond simple linear regression

$$\underset{n\times 1}{\boldsymbol{y}} = \underset{n\times p}{\boldsymbol{X}}\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$$

- $\boldsymbol{y}$ is a vector (or a one-column matrix) of n observations on our response variable
- $\boldsymbol{X}$ is an $n \times p$ matrix of observations on $p - 1$ factors
- $\boldsymbol{\beta}$ is a $p \times 1$ matrix of unknown parameters
- $\boldsymbol{\epsilon}$ is a vector of n observation-specific deviations from the expected value

IDA | 100

Slide notes:

Icons are copyrighted by The Noun Project.

General linear model – assumptions

For estimation to be valid, we only need these four assumptions.

- **Homoskedasticity**– constant variance about any value of the regression function (e.g., deviation for each unit has the same variance)
- **Independence**– errors are statistically independent
- **Linearity**– the expected values (means) are linear functions of the parameters
- **Existence**– finite mean and variance

Always check these assumptions with **residual plots** and **posterior predictive checks**.

IDA 101

General linear model – example

intercept

Regression equation:

$$\text{MPG} = \beta_0 + \beta_1 \text{Cylinders} + \beta_2 \text{Displacement} + \beta_3 (\text{Cylinders} \times \text{Displacement}) + \epsilon$$

main effects

interaction effect

$$\begin{bmatrix} MPG_1 \\ MPG_2 \\ MPG_3 \\ MPG_4 \\ \vdots \end{bmatrix} = \begin{bmatrix} 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ 1 & \text{Cyl} & \text{Disp} & \text{Cyl} \times \text{Disp} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \end{bmatrix}$$

$$\boldsymbol{MPG = X\beta + \epsilon}$$

IDA 102

Slide notes:

Normality of errors is also a common assumption in many applied statistics textbooks, but in practice it is too hard to validate normality, and it is (at least according to Gelman and Hill's textbook) the least important regression assumption.

Approved for public release; distribution is unlimited.

It's better to have research goals in mind *before* you analyze the data

1974 was about the time of the oil crisis, so we can assume that mpg is the outcome variable.

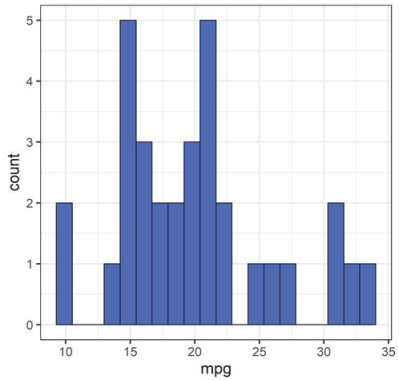
Possible research questions:

- Do the data explain which factors affect mpg?
  - To what extent?
- Do we need all the variables to explain mpg?
- Is a linear relationship suitable?
- If we make predictions, can we make them with confidence?

A Little Bit about Data Exploration

Approved for public release; distribution is unlimited.

MPG does not look normally distributed, but this is not a problem



Remember: For regression to work well, we want the *residuals* to look normal-ish, not the raw data!

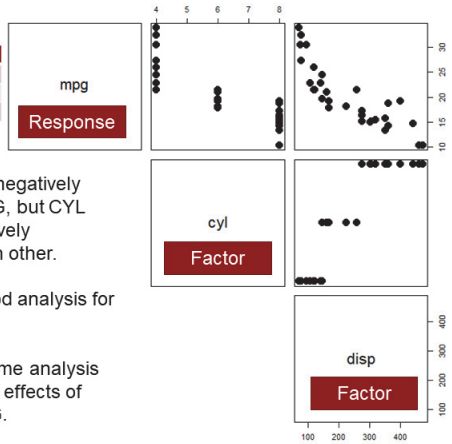
IDA 105

Slide notes:

We want the residuals to look like a bell shape, but this is very hard to attain in practice, so we will be content if they just "sort of" look like a bell shape.

Pairwise scatterplots of MPG, CYL, and DISP

	Explanation
mpg	Miles per gallon
cyl	Number of cylinders
disp	Displacement (cu. in.)



- CYL and DISP are negatively correlated with MPG, but CYL and DISP are positively correlated with each other.
- Regression is a good analysis for this type of data.
- One-variable-at-a-time analysis would overstate the effects of each factor on MPG.

IDA 106

For example, notice how the coefficients change once we add another variable (least squares with `lm`)

```
Call:
lm(formula = mpg ~ cyl, data = dat)

Coefficients:
(Intercept)         cyl6         cyl8
     26.664       -6.921      -11.564
```

```
Call:
lm(formula = mpg ~ cyl + disp, data = dat)

Coefficients:
(Intercept)         cyl6         cyl8         disp
   29.53477     -4.78585     -4.79209     -0.02731
```

It's challenging to assess factor effects when variables are highly correlated! (This is one reason we stress good experimental designs.)

Fitting Bayesian Models to Data

Bayesian priors for regression models

Informative priors: $\beta_j \sim Normal(0, \sigma_\beta^2), \sigma^2 \sim Inverse - Gamma(a, b)$

Priors centered at 0 are **more conservative** than least squares

- Scale experimental design variable to make priors easier to set
- Do not use informative priors for β_0 without very good reason (this biases overall model)

Non-informative priors:

- "Standard"* non-informative prior for regression is $p(\beta) \propto 1$ and $\log(\sigma) \propto 1$

Data model:

$$y_i \sim Normal\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_i, \sigma^2\right)$$

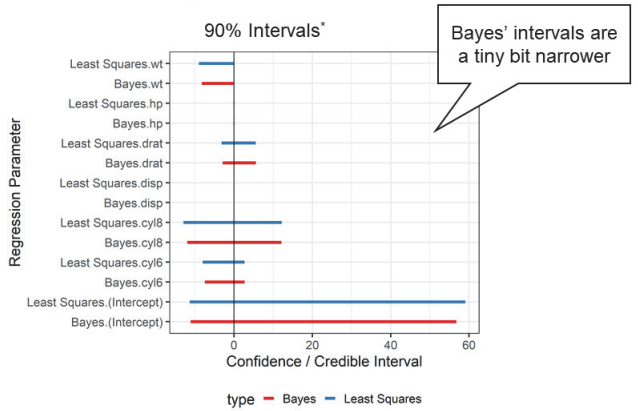
Or, using matrix notation,

$$\boldsymbol{y} \sim Normal(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I})$$

"Signal" "Noise"

* Bayesian Data Analysis, p. 355

Bayesian analysis with Stan's default priors is similar to least squares



* Only a few regression parameters are plotted for ease of visualization.

What does Stan use for default priors?

Stan uses "weakly informative" data-driven priors by default

$$\beta_i \sim Normal\left(0, 2.5 \times \frac{\mathrm{sd}(y)}{\mathrm{sd}(x_i)}\right)$$

$$\sigma \sim Exponential\left(\mathrm{sd}(y)\right)$$

$$\beta_0 \sim Normal\left(\bar{y}, 2.5 \times \mathrm{sd}(y)\right)$$

You could use the defaults in the absence of good priors*

* Default priors have a greater effect in logistic regression and a relatively small effect in linear regression.

Variable selection – Do I include `qsec`? `vs`? `wt`? `hp`?

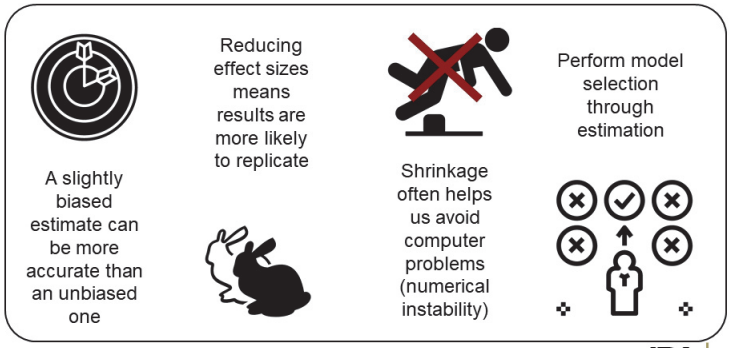
Difficult and controversial!

- Some statisticians: Reduce the number of variables as much as possible!
- Others: Parsimony is the enemy of predictive performance!

Approved for public release; distribution is unlimited.

Shrinkage is one way to handle variable selection

"Shrinkage" is the idea that large, but weak, estimates should be nudged towards zero.



IDA 113

Why a Bayesian approach to shrinkage?

Credible intervals

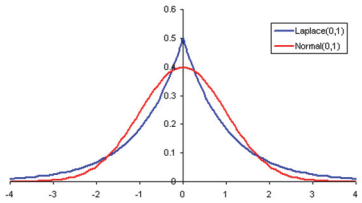
Frequentist models don't produce confidence intervals when effects are shrunk!

IDA 114

Slide notes:

- 3D Replication by H Y P E R M O R G E N from the Noun Project
- Tripping by Luis Prado from the Noun Project
- Accuracy Time by Graphic Engineer from the Noun Project
- Selection by Nithinan Tatah from the Noun Project

Bayesian shrinkage model



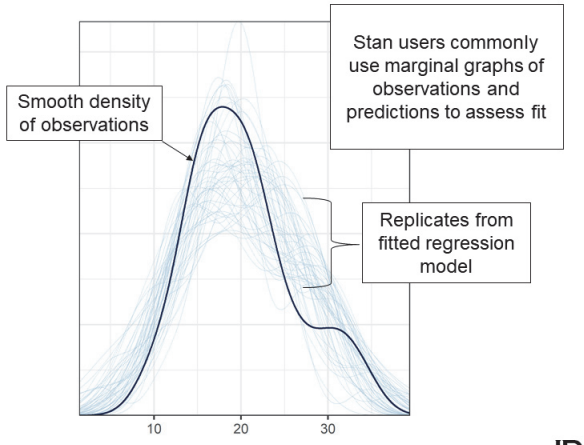
Commonly, we use Laplace priors to "shrink" factor effects towards 0.

In frequentist statistics, the analogous procedure is called LASSO.

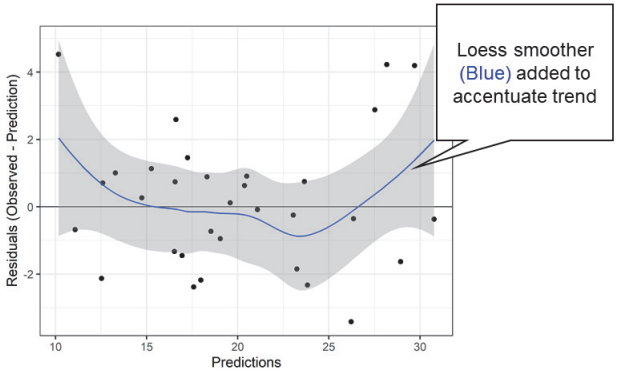
LASSO = Least Absolute Shrinkage and Selection Operator

Fitting the mpg model in Stan

Use graphical methods to validate Bayesian models



Use graphical methods to validate Bayesian models



Not a good model fit!

Don't worry, the first model usually isn't very good.

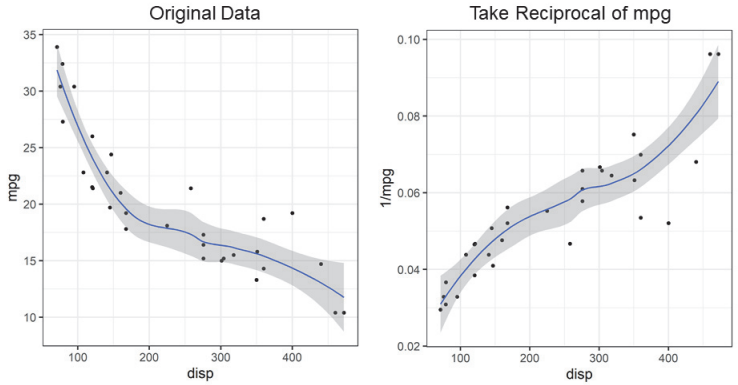
Use corrective procedures to ensure the results are reliable

- Variable transformations
- Change of model (for example, linear model to lognormal)
- Splines (rather automatic; recommended)

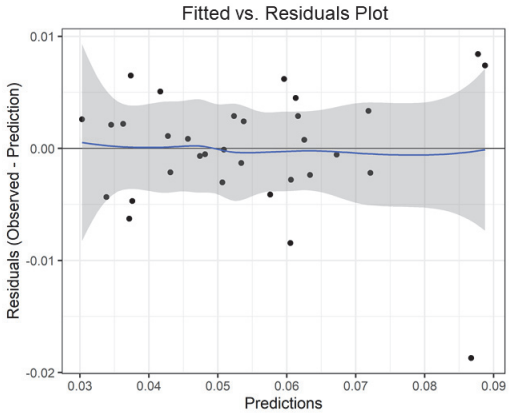
It's much harder to get a good fit when working with continuous variables.

Note that these procedures could also be used in frequentist analyses.

Inverting mpg makes for a more linear relationship with some variables

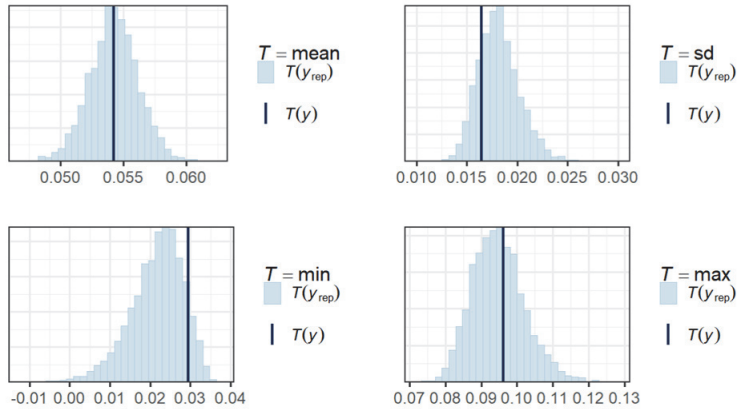


Graphical checks validate the model



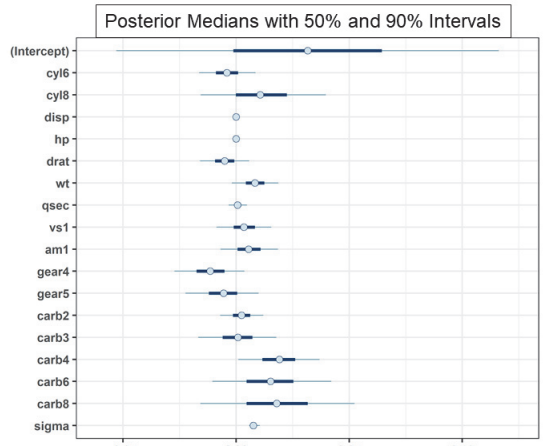
Much better!

Posterior p-values can be graphed for easier model checking

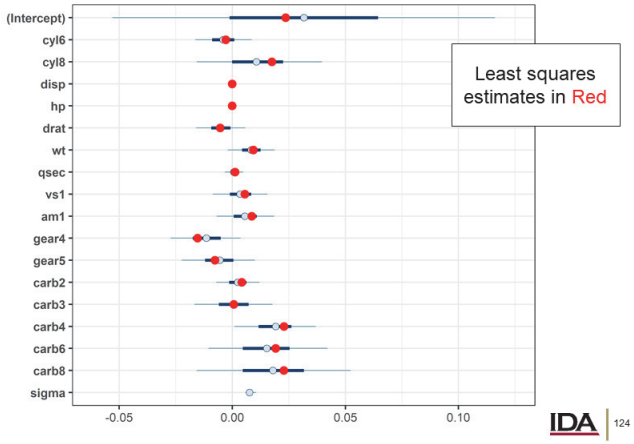


Approved for public release; distribution is unlimited.

Estimated regression coefficients for the 1/mpg model



Bayesian estimates (blue) are "shrunk" toward zero



Predictions

It's easy to get mpg predictions from the 1/mpg model

There are two ingredients:

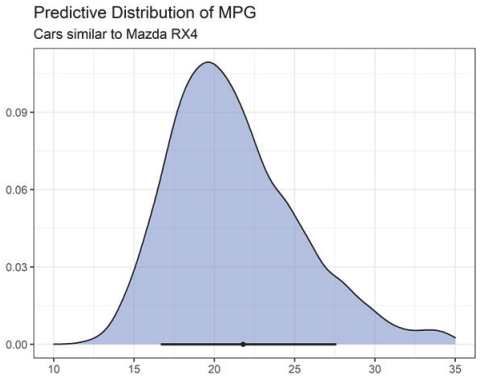
1. The posterior predictive distribution (discussed earlier)
2. The transformation on the entire distribution

We can use the posterior *predictive* distribution and transform the data back to original data scale

In our case, we have samples from $P\left(\frac{1}{mpg^{rep}}\middle|mpg\right)$, but we want $P(mpg^{rep}|mpg)$.

Translating back is just a matter of applying the reciprocal function to each of the samples from $P\left(\frac{1}{mpg^{rep}}\middle|mpg\right)$.

Use the posterior predictive distribution to summarize prediction error



Summarizing predictions with Stan

What do you do when Stan is not cooperating?

Two tricks to make Stan cooperate:

1. Increase `adapt_delta`
2. Increase `max_treedepth`

For more information about Stan warnings, see https://mc-stan.org/misc/warnings.html.

Approved for public release; distribution is unlimited.

Bayesian analysis extends to regression, and it offers several technical advantages to researchers:

- Simply obtain predictions after transformation
- Calculate intervals after applying shrinkage
- Apply a range of effective model criticism tools

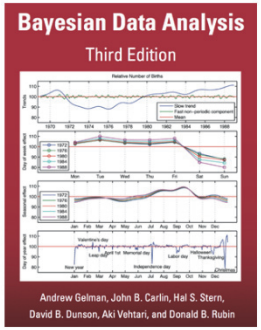
Bayesian statistics is another tool in your statistical analysis toolbox

Although the goal of Bayesian and frequentist statistics is to answer a research question, the analysis of the data and interpretation of the results differ between them.

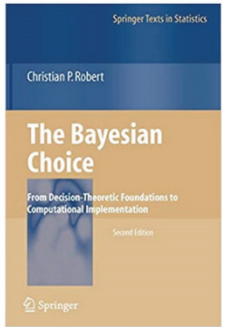
There are multiple ways to properly incorporate prior information into a Bayesian model.

As with frequentist statistics, once you fit your statistical model, make sure your model fits the data well.

Great references



Application

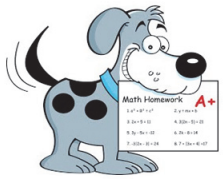


Theory

Programming

Thank you!

Contact Info:
Dr. Keyla Pagán-Rivera – kpaganri@ida.org
Dr. John Haman – jhaman@ida.org
Dr. Rebecca Medlin – rmedlin@ida.org



Resources:

- https://testscience.org/
- Past trainings – OED SharePoint site
- Future trainings – check https://test-science.shinyapps.io/TSTraining/

Backup

Details on computing the posterior predictive p-value

Steps:

- Use the posterior distribution to simulate a vector of θ
- Obtain the joint posterior distribution by drawing y^{rep} from the sampling distribution using the simulated vector of θ
- Estimate the Bayesian p-value

$$p_B \approx \frac{1}{L}\sum_{l=1}^{L} I_{\{T(y^{rep,l},\theta^l)>T(y,\theta^l)\}}$$

Slide notes:

L is the length of your Markov Chain Monte Carlo sample (the number of iterations in the chain). Test statistics are the mean and the standard deviation.

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Appendix D
R and Stan Code

```
#####################################
#### R code for Bayes 101 course ####
############## Section 2 ##############
#####################################

##### Beta-Binomial #####
### Data ###
pri_suc <- 5
pri_fail <- 1

obs_suc <- 6
obs_fail <- 3

### Bayesian (sampling from posterior) ###
set.seed(805)
post <- rbeta(1000, pri_suc + obs_suc, pri_fail + obs_fail)
mean(post)
median(post)
```

```
### Central Posterior Interval ###
#Compute the central posterior interval directly from posterior Beta
qbeta(c(0.025, 0.5, 0.975), pri_suc + obs_suc, pri_fail + obs_fail)

#Compute the central posterior interval from the MCMC samples
quantile(post, c(0.025, 0.5, 0.975))

### Highest Posterior Density Interval ###
library(HDInterval)
hdi(post)

### Plot the intervals ###
#Plot the distribution with the central posterior interval
lo_q <- quantile(post, c(0.025))
up_q <- quantile(post, c(0.975))

lo <- min(which(density(post)$x > lo_q))
up <- max(which(density(post)$x < up_q))

par(mar = c(5, 4, 4, 2) + 1.5)
plot(density(post), xlab = expression(theta),
     main = "Central Posterior Interval", cex.main = 3.5, cex.lab = 2.5,
     cex.axis = 2)
with(density(post), polygon(x = c(x[c(lo, lo:up, up)]),
                            y = c(0, y[lo:up], 0),
                            col = "gray"))
```

```
#Plot the distribution with the HDI
lo_q_hdi <- hdi(post)["lower"]
up_q_hdi <- hdi(post)["upper"]

lo_hdi <- min(which(density(post)$x > lo_q_hdi))
up_hdi <- max(which(density(post)$x < up_q_hdi))

plot(density(post), xlab = expression(theta),
     main = "Highest Posterior Density Interval", cex.main = 3.5,
     cex.lab = 2.5, cex.axis = 2)
with(density(post), polygon(x = c(x[c(lo_hdi, lo_hdi:up_hdi, up_hdi)]),
                            y = c(0, y[lo_hdi:up_hdi], 0),
                            col = "gray"))
dev.off()

################################################################################
############################### Back to our slides ##############################
################################################################################

### STAN model ###
library(rstan)
set.seed(1638)

### Data ###
data_bb <- list(n = obs_suc + obs_fail, y = obs_suc,
```

```
                alpha = pri_suc, beta = pri_fail)
### Stan model ###
fit_bb <- stan(file = "Section 2/beta_binom.stan",
               data = data_bb, pars = "theta", chains = 4,
               iter = 5000,  warmup = 1000,
               init = list(list("theta" = 0.01), list("theta" = 0.3),
                           list("theta" = 0.7), list("theta" = 0.99)))

### Central posterior interval ###
print(fit_bb, probs = c(0.025, 0.5, 0.975))
#MCSE is se_mean column. It can be obtained by dividing sd/sqrt(n_eff)
#http://www.stat.columbia.edu/~gelman/book/software.pdf


### Highest posterior density interval ###
# Right now, Stan doesn't have an easy way to compute the HPD intervals. We
# could save the posterior as an MCMC list and use the hdi() function.
library(coda)
th <- extract(fit_bb) #extract the posterior theta values
#Split the th object into 4 mcmc lists (one per chain)
mc_th <- mcmc.list(as.mcmc(matrix(th$theta[1:4000], nrow = 4000)),
                   as.mcmc(matrix(th$theta[4001:8000], nrow = 4000)),
                   as.mcmc(matrix(th$theta[8001:12000], nrow = 4000)),
                   as.mcmc(matrix(th$theta[12001:16000], nrow = 4000)))
hdi(mc_th)
```

```
### Trace plots ###
#With the burn-in/warmup period
plot(fit_bb, plotfun = "trace", inc_warmup = TRUE) +
  ggtitle("Trace plots") +
  xlab("Index") +
  theme(text = element_text(size = 20), axis.title = element_text(size = 24),
        legend.text = element_text(size = 20),
        legend.title = element_text(size = 24),
        plot.title = element_text(size = 30))

#Without the burn-in/warmup period
plot(fit_bb, plotfun = "trace", inc_warmup = FALSE) +
  ggtitle("Trace plots") +
  xlab("Index") +
  theme(text = element_text(size = 20), axis.title = element_text(size = 24),
        legend.text = element_text(size = 20),
        legend.title = element_text(size = 24),
        plot.title = element_text(size = 30))

### Posterior distribution plot ###
plot(fit_bb, plotfun = "hist", bins = 20) +
  ggtitle("Posterior distributions")

### Gelman and Rubin plot/diagn ###
### Stan doesn't have a nice way to obtain the GR diagnostics so once again, we
### need the list() workaround we used for the HDI
```

```
gelman.diag(mc_th)
par(mar = c(5, 4, 4, 2) + 1.5)
gelman.plot(mc_th, cex.lab = 2.5, cex.axis = 2, lwd = 4)
dev.off()

### We could also compare the posterior estimates for theta with a Beta(11, 4)
### density
th <- extract(fit_bb) #extract the theta values
hist(th$theta, freq = FALSE,
     main = "Posterior samples (hist.) vs Beta distribution (red line)",
     xlab = expression(theta))
ax <- seq(0, 1, length.out = length(th$theta))
lines(ax, dbeta(ax, pri_suc + obs_suc, pri_fail + obs_fail), col = 2)

### Frequentist results
library(binom)
binom.confint(x = 6, n = 9, conf.level = 0.95)

################################################################################
############################### Back to our slides ##############################
################################################################################

##### Exponential-Gamma #####
### Data
n_fail <- 2
t_hours <- 40000
```

```
### Bayesian (sampling from posterior)
#Prior
eng <- 990
p <- 1.5
a <- 2 + (1/p^2)
b <- (a-1) * eng

#posterior
a_post <- a + n_fail
b_post <- b + t_hours
set.seed(1044)
post <- rgamma(1000, a_post, rate=b_post)

mean(post) #posterior mean of lambda
median(post) #posterior median of lambda

mean(1/post) #posterior MFHBR mean
#Posterior median and 80% interval
MFHBR_quant <- quantile(1/post, prob=c(0.1, 0.5, 0.9))

###Frequentist
MFHBR_freq <- t_hours / n_fail
#Interval
alpha <- 0.2
lower_ci <- (2 * t_hours) / qchisq(alpha/2, 2*n_fail + 2, lower.tail = FALSE)
```

```
upper_ci <- (2 * t_hours) / qchisq(alpha/2, 2*n_fail, lower.tail = TRUE)

#Compare results
rbind(Bayes = MFHBR_quant, Freq = c(lower_ci, MFHBR_freq, upper_ci))

################################################################################
############################### Back to our slides ##############################
################################################################################

# Selecting a prior by marginal distribution

## Set up priors
mu_prior <- rnorm(1000, 2, 2)
sigma_prior <- rexp(1000, 1)
## m(y) is this density:
y <- rnorm(1000, mu_prior, sigma_prior)

## graph m(y). This is what we expect the data to look like BEFORE we do the test.
plot(density(y))

## We can also compute summary statistics of m(y)
summary(y)
## Now we ask ourselves if this is what we expect the test data to look like
## (roughly). If not, then we can modify the prior parameters until the data
## look like what are expected. We can also increase/decrease the variability
## to reflect our confidence in the prior distribution.
```

```
////////////////////////////////////////////
///// Beta-Binomial example for Bayes 101 /////
////////////////////////////////////////////
data {
  int<lower=0> n;
  int<lower=0, upper=n> y;
  real<lower=0> alpha;
  real<lower=0> beta;
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  // prior
  theta ~ beta(alpha, beta);
  // Likelihood
  y ~ binomial(n, theta);
}

########################################
##### R code for Bayes 101 course #####
############## Section 3 ###############
########################################
```

```
##### Simple linear model #####
### Read data ###
load("Section 3/slr_data.RData")

### Fit the SLR Model in Stan ###
library(rstan)
#data and vector for credible intervals domain
wt_cred = seq(min(slr_data$wt), max(slr_data$wt), length.out = 50)

data_slr_P1 <- list(N = nrow(slr_data), mpg = slr_data$mpg, wt = slr_data$wt,
                    K = length(wt_cred), wt_cred = wt_cred)
#fit model
fit_slr <- stan(file = "Section 3/SLR Stan model - P1.stan",
                data = data_slr_P1, seed = 1120,
                pars = c("beta0", "beta1", "sigma2", "cred_int"),
                chains = 4, iter = 5000, warmup = 1000)
#visualize posterior samples
plot(fit_slr, plotfun = "trace", nrow = 3, inc_warmup = FALSE,
     pars = c("beta0", "beta1", "sigma2")) +
  ggtitle("Trace plots")
plot(fit_slr, plotfun = "hist",
     pars = c("beta0", "beta1", "sigma2")) +
  ggtitle("Posterior distributions")
#parameter estimates and credible intervals
print(fit_slr, probs = c(0.10, 0.50, 0.90),
      pars = c("beta0", "beta1", "sigma2"))
```

```
### Credible Intervals for operational results ###
#extract MCMC samples
cred_int <- extract(fit_slr)$cred_int
class(cred_int) #matrix with rows = posterior samples; columns = 50 distances

#Obtain credible intervals and posterior medians
CrI0 <- apply(cred_int, 2, function(x) quantile(x, c(0.1, 0.5, 0.9)))
CrI <- data.frame(t(CrI0))
CrI$wt_cred <- seq(min(slr_data$wt), max(slr_data$wt),
                   length.out = 50) #Credible Interval domain

### Add the frequentist confidence interval
# Freq model (assume we have checked the assumptions were met)
freq_model <- lm(slr_data$mpg ~ slr_data$wt)
summary(freq_model)
confint(freq_model, level = 0.8) #parameters interval

library(ciTools)
slr_data <- add_ci(slr_data, freq_model, alpha = 0.2,
                   names = c("freq_10", "freq_90"),
                   yhatName = "freq_est")
#Plot
cl <- c("Frequentist" = "blue", "Bayesian" = "purple") #for the manual legend

ggplot(data = slr_data, aes(x = wt, y = mpg)) +
```

```
  geom_point() +
  geom_line(aes(y = freq_est, color = "Frequentist")) +
  geom_ribbon(aes(ymin = freq_10, ymax = freq_90), fill = "blue", alpha = 0.1) +
  geom_line(data = CrI, aes(x = wt_cred, y = X50., color = "Bayesian")) +
  geom_ribbon(data = CrI, aes(x = wt_cred, ymin = X10., ymax = X90.),
              fill = "purple", alpha = 0.1, inherit.aes = FALSE) +
  labs(x = "Weight", y = "Miles per Gallon", color = "Model") +
  scale_color_manual(values = cl) +
  theme_classic() +
  theme(axis.text = element_text(size = 15),
        axis.title = element_text(size = 20))



################################################################################
############################### Back to our slides ##############################
################################################################################

#Restart R Session

### Read data and load libraries (again) ###
load("Section 3/slr_data.RData")
library(rstan)

data_slr_P2 <- list(N = nrow(slr_data), mpg = slr_data$mpg, wt = slr_data$wt)

fit_slr_P2 <- stan(file = "Section 3/SLR Stan model - P2.stan",
```

```
                    data = data_slr_P2, seed = 1120,
                    pars = c("beta0", "beta1", "sigma2", "mpg_rep",
                             "chi_obs", "chi_rep", "p_chi"),
                    chains = 4, iter = 2000, warmup = 1000)

### Graphical Posterior Predictive Checks ###

#obtain 8 replicated datasets (time_rep or y^rep) each with 15 observations
#using the posterior distribution. We have a matrix with
#rows = the number of MCMC samples, and columns = the number of observations in
#the original example. We want to sample a few (8 in our case) replicated
#datasets and compare those replicated datasets with the observed data.
mpg_rep <- extract(fit_slr_P2)$mpg_rep
set.seed(1252)
mpg_rep_sample <- mpg_rep[sample(1:nrow(mpg_rep), 8), ]

### boxplot with predicted quantities
par(mar = c(5, 4.7, 4, 2))
plot(x = c(1:9), y = seq(0, 40, length.out = 9), type = "n", xlab = "",
     ylab = "", yaxt='n')
boxplot(slr_data$mpg, col = "blue", xlab = "Dataset",
        ylab = "Time", add = TRUE, at = 1, cex.axis = 1.3, cex.lab = 2)
for(i in 1:8){
  boxplot(mpg_rep_sample[i,], add = TRUE, xlab = "",
          col = "purple", at = i + 1, axes = FALSE)
}
```

```
dev.off()
### histogram with predictive quantities
par(mfrow = c(3, 3), mar = c(5, 4, 4, 2) - 1)
hist(slr_data$mpg, col = "blue", xlab = "", xlim = c(5, 40), ylim = c(0, 8),
     main = "Histogram for obs", cex.main = 2, cex.axis = 1.5)
for(i in 1:8){
  hist(mpg_rep_sample[i,], xlab = "", col = "purple", cex.main = 2,
       main = paste0("Histogram for rep ", i), cex.axis = 1.5,
       xlim = c(5, 40), ylim = c(0, 8))
}
dev.off() #reset graphics parameters


### Bayes p-value using the Goodness-of-Fit test ###
chi_obs <- extract(fit_slr_P2)$chi_obs
chi_rep <- extract(fit_slr_P2)$chi_rep
(p_chi <- round(mean(chi_obs > chi_rep), 2))

par(mar = c(5, 5.2, 4, 2))
plot(chi_obs, chi_rep,
     xlim = c(min(chi_obs, chi_rep),
              max(chi_obs, chi_rep)),
     ylim = c(min(chi_obs, chi_rep),
              max(chi_obs, chi_rep)),
     xlab = expression(paste("T(y"^{"obs"}, "|", theta, ")")),
     ylab = expression(paste("T(y"^{"rep"}, "|", theta, ")")),
```

```
     main = "Goodness-of-Fit Test Statistic with Observed and Replicated Data",
     cex.main = 2, cex.lab = 1.7, cex.axis = 1.5)
text(x = min(chi_obs, chi_rep) + 5, y = max(chi_obs, chi_rep), cex = 1.5,
     substitute(paste("p"["B"], " = ", p_chi), list(p_chi=p_chi)))
abline(0,1)
```

```
################################################################################
############################### Back to our slides ###############################
################################################################################

### Bayesian Modeling ###

# Using a shrinkage prior
# Note: several shrinkage priors are available in rstanarm.
fit_lasso <- stan_glm(mpg ~ ., data = dat,
                      prior = laplace(autoscale = TRUE),
                      seed = 1,
                      adapt_delta = 0.999,
                      open_progress = FALSE)

## Investigate the model interactively with shinystan
shinystan::launch_shinystan(fit_lasso)
```

```
########################################
## Then we do some model checking ... ##
########################################


## The figure on slide 130 (and also available in shinystan)
pp_check(fit_lasso)

## The figure on slide 131
d1 <- data.frame(fitted = fitted(fit_lasso),
                 resid = residuals(fit_lasso))

ggplot(d1, aes(x = fitted, y = resid)) +
  geom_point(size = 3) +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth() +
  xlab("Predictions") +
  ylab("Residuals (Observed - Prediction)")

####################################
## This model seems to be okay... ##
####################################

## Shrinkage prior on gpm. This model may be a better representation of the data.
fit_gpm <- stan_glm(1 / mpg ~ ., data = dat,
                    prior = laplace(autoscale = TRUE),
```

```
                    seed = 1,
                    adapt_delta = 0.9999, # this computational adjustment may be necessary to ensure good fit
                    open_progress = FALSE)

## View a summary of the model
print(fit_gpm, digits = 5) # use lots of sigfigs because the coefficients are tiny on the 1/y scale

## The figure on slide 133
## May differ slightly because of simulation error
d2 <- data.frame(fitted = fitted(fit_gpm),
                 resid = residuals(fit_gpm))

ggplot(d2, aes(x = fitted, y = resid)) +
  geom_point(size = 3) +
  geom_hline(aes(yintercept = 0)) +
  geom_smooth() +
  xlab("Predictions") +
  ylab("Residuals (Observed - Prediction)")

## The figure on slide 135
plot(fit_gpm)

################################################################################
############################### Back to our slides ##############################
################################################################################
```

```
### Bayesian Predictions ###

## In this final section, we make summary predictions

## Point at which to make a prediction
xnew <- dat[1, -1]

## Correct for the reciprocal we took earlier
## This takes reciprocal of _entire_ posterior predictive distribution
pp_xnew <- 1 / posterior_predict(fit_gpm, newdata = xnew)

## Plot the results
## May differ slightly because of simulation error
ggplot(as.data.frame(pp_xnew), aes(x = pp_xnew)) +
  ggtitle("Predictive Distribution of MPG", "Cars similar to Mazda RX4") +
  scale_x_continuous(n.breaks = 6, limits = c(10, 35)) +
  xlab("") +
  ylab("") +
  geom_density(fill = "royalblue2", size = 1, alpha = 0.4) +
  geom_errorbarh(aes(xmin = quantile(c(pp_xnew), 0.1),
                     xmax = quantile(c(pp_xnew), 0.9), y = 0), height = 0, size = 1.5) +
  geom_point(aes(x = mean(pp_xnew), y = 0), size = 3, color = "black")

//////////////////////////////////////////
///// SLR model for Bayes 101 - Part I /////
//////////////////////////////////////////
```

```
data {
  int<lower=0> N; //define number of observations
  vector[N] wt; //define vector of observed weights
  vector[N] mpg; //define vector of observed miles per gallon
  int<lower=0> K;
  vector[K] wt_cred; //define vector of weights for credible intervals
}
parameters {
  real beta0;
  real beta1;
  real<lower=0> sigma2;
}
transformed parameters {
  vector[K] cred_int = beta0 + beta1 * wt_cred; //credible interval data
}
model {
  // priors
  beta0 ~ normal(40.8, 1.63*5);
  beta1 ~ normal(-6.28, 0.42*5);
  sigma2 ~ inv_gamma(6.1, 81.08);
  // Likelihood
  mpg ~ normal(beta0 + beta1 * wt, sqrt(sigma2)); //Stan used sd, not variance
}
```

```
//////////////////////////////////////////
///// SLR model for Bayes 101 - Part 2 /////
//////////////////////////////////////////

data {
  int<lower=0> N; //define number of observations
  vector[N] wt; //define vector of observed weight
  vector[N] mpg;
}
parameters {
  real beta0;
  real beta1;
  real<lower=0> sigma2;
}
model {
  // priors
  beta0 ~ normal(40.8, 1.63*5);
  beta1 ~ normal(-6.28, 0.42*5);
  sigma2 ~ inv_gamma(6.1, 81.08);
  // Likelihood
  mpg ~ normal(beta0 + beta1 * wt, sqrt(sigma2));
}
generated quantities {
  real mpg_rep[N];
  real chi_obs0[N];
  real chi_rep0[N];
```

```
  real chi_obs;
  real chi_rep;
  int<lower=0, upper=1> p_chi;

  mpg_rep = normal_rng(beta0 + beta1 * wt, sqrt(sigma2));

  for (i in 1:N) {
    chi_obs0[i] = (mpg[i] - (beta0 + beta1 * wt[i]))^2 / sigma2;
    chi_rep0[i] = (mpg_rep[i] - (beta0 + beta1 * wt[i]))^2 / sigma2;
  }
  chi_obs = sum(chi_obs0);
  chi_rep = sum(chi_rep0);
  p_chi = (chi_obs >= chi_rep);
}
```

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
August 2021	Final	

4. TITLE AND SUBTITLE

Introduction to Bayesian Analysis

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Keyla Pagan-Rivera (OED); John T. Haman (OED); Rebecca M. Medlin (OED)

5d. PROJECT NUMBER

BD-09-2299(90)

5e. TASK NUMBER

229990

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

NS-D-20484

H 2020-000510/2

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

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES

Heather M. Wojton

14. ABSTRACT

As operational testing becomes increasingly integrated and research questions become more difficult to answer, IDA's Test Science team has found Bayesian models to be powerful data analysis methods. Analysts and decision-makers should understand the differences between this approach and the conventional way of analyzing data. It is also important to recognize when an analysis could benefit from the inclusion of prior information—what we already know about a system's performance—and to understand the proper way to incorporate that information. To apply Bayesian methods, analysts need to comprehend some technical aspects of this approach and know how to properly use appropriate statistical software. In this course, students learn the intuition behind Bayesian statistics, the mathematical details of posterior distributions, how to fit simple Bayesian models using computer software, and how to assess model fit.

15. SUBJECT TERMS

Bayesian; Defense Testing; Analyses

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Heather M. Wojton
Unclassified	Unclassified	Unclassified		253	19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6811

Approved for public release; distribution is unlimited.