A Workship on Bayesian Statistics: From Applications to the General Theory and From Theory to Applications

(A statistical inferential method named after Thomas Bayes)

Nabendu Pal
Department of Mathematics
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Session-I: Basics of Bayesian Theory: From Applications to the General Theory
Session-II: Challenges in Bayesian Statistics: From Gibbs Sampling to MCMC Algorithm with Applications

Session-I: Basics of Bayesian Theory: From Applications to the General Theory

Bayes' Theorem

The idea of Bayes' Theorem (1763) is very simple.

- Let A be an event (i.e., $A \subset \mathcal{S}$ = sample space).
  We know $P(A)$ and $P(A^c)$.

- B is another event for which we know

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \implies P(B^c|A) = 1 - P(B|A)$$

$$P(B|A^c) = \frac{P(B \cap A^c)}{P(A^c)} \implies P(B^c|A^c) = 1 - P(B|A^c)$$

Bayes' Rule

- But we are interested in $(P(A|B)$ which can be expressed as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A \cap B)}{P(A \cap B) + P(A^c \cap B)}$$
$$= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

: Bayes' Theorem in simplest form (Bayes' Rule)

Application

- A simple application:
  A = A person truly has COVID
  B = A person tests positive by a quick test

  In the larger society we have a fairly good idea about $P(A)$ and $P(A^c)$.
  [For example, $P(A)$ may be 5% or 20%]

Application

The company making the quick test must submit to the government two key pieces of info.

$$P(B|A) = P(\text{ Tests positive } | \text{ Has COVID }) = \text{called the sensitivity of the quick test method}$$

$$P(B^c|A^c) = P(\text{ Tests negative } | \text{ No COVID }) = \text{called the specificity of the quick test method}$$

Note

- A company marketing a quick test method obtains the sensitivity and specificity values after a long research and development (R & D) process in the Lab using the DNA / RNA analysis of the test subjects' body fluid.
- Usually,

$$\text{sensitivity} \geq 0.95$$

$$\text{specificity} \geq 0.90$$

Sensitivity and Specificity

$A =$ A person has COVID $= H_0$ (say)
$B =$ Quick test confirms COVID $=$ Retain H_0

$$\begin{aligned}\text{Sensitivity} &= P(B|A) = P(+|+)\\ &= 1 - P(B^c|A)\\ &= 1 - P(\text{Tests Negative}|\text{Has COVID})\\ &= 1 - \text{Probability of Type-I Error}\end{aligned}$$

$$\begin{aligned}\text{Specificity} &= P(B^c|A^c) = P(-|-)\\ &= 1 - P(B|A^c)\\ &= 1 - P(\text{Tests Positive}|\text{No COVID})\\ &= 1 - \text{Probability of Type-II Error}\end{aligned}$$

Sensitivity and Specificity

However, we want to know

$$\begin{aligned} P(A|B) &= P(\text{Has COVID}|\text{Tests Positive}) \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

If $\pi = P(A) =$ proportion of people who truly have COVID,

$$P(A|B) = \frac{\pi * \text{Sensitivity}}{\pi * \text{Sensitivity} + (1-\pi)(1-\text{Specificity})}$$

Sensitivity and Specificity

A = A person has COVID
B = Quick test confirms COVID

[Sensitivity $= 0.95$, Specificity $= 0.90$]

$$\text{If } \pi = 0.05 \implies P(A|B) = 0.3333 = 33.33\%$$
$$[\text{i.e., } P(A^c|B) = 0.6667 = 66.67\%]$$

$$\text{If } \pi = 0.20 \implies P(A|B) = 0.7037 = 70.37\%$$
$$[\text{i.e., } P(A^c|B) = 0.2963 = 29.63\%]$$

Even when $\pi = 0.50$ (i.e, every other person is infected),
$P(A|B) \sim 0.90$.

[Moral of the story: Don't Panic!]

Bayesian Framework

In Bayes' Theorem: A = True reality

B = What we see / perceive

We know $\longrightarrow$ $P(A)$ & $P(A^c)$: Prior Information

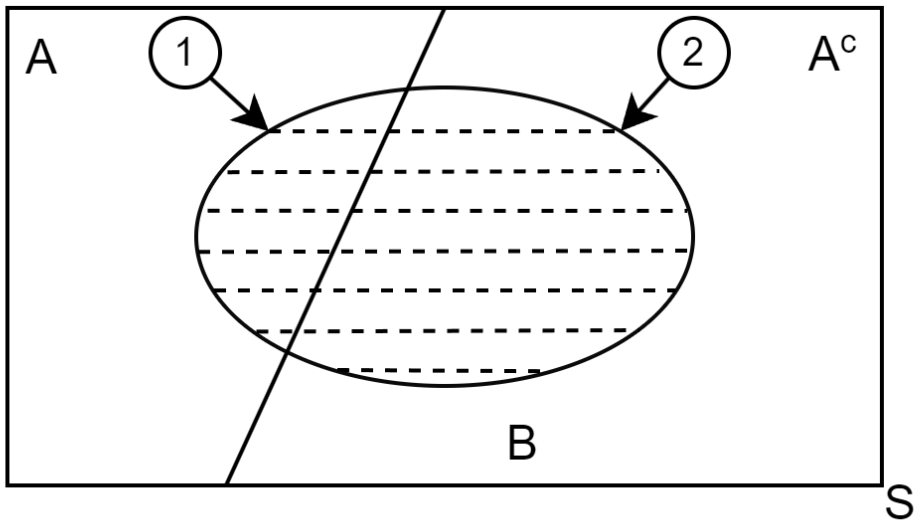
and $\longrightarrow$ $P(B|A)$, $P(B|A^c)$: Conditional information of what we see given the reality (prior)

But we want to know

$$P(A|B) = P(\text{ True reality } | \text{ What we see })$$

Conditional Information

Q. So, what does the Bayes' Theorem or Rule say?
A. It is all about updating the prior probabilities when some conditional information in terms of the happening of a special event is known.
We know $P(A)$, $P(A^c)$, $P(B|A)$, $P(B|A^c)$



We want to know $P(A|B)$ and $P(A^c|B)$
NOTE: $\{A, A^c\}$ is a partition of $\mathcal{S}$

Generalization

Generalization of the simplest form:

Suppose we have a general partition of $\mathcal{S}$ as

$$\mathcal{S} = \{A_1 \cup A_2 \cup \dots\}$$

such that $A_i \cap A_j = \emptyset$ for $i \neq j$

NOTE: The partition may be finite / infinite

We know $\{P(A_i),\ i = 1, 2, 3, \dots\} \longleftarrow$ prior probabilities

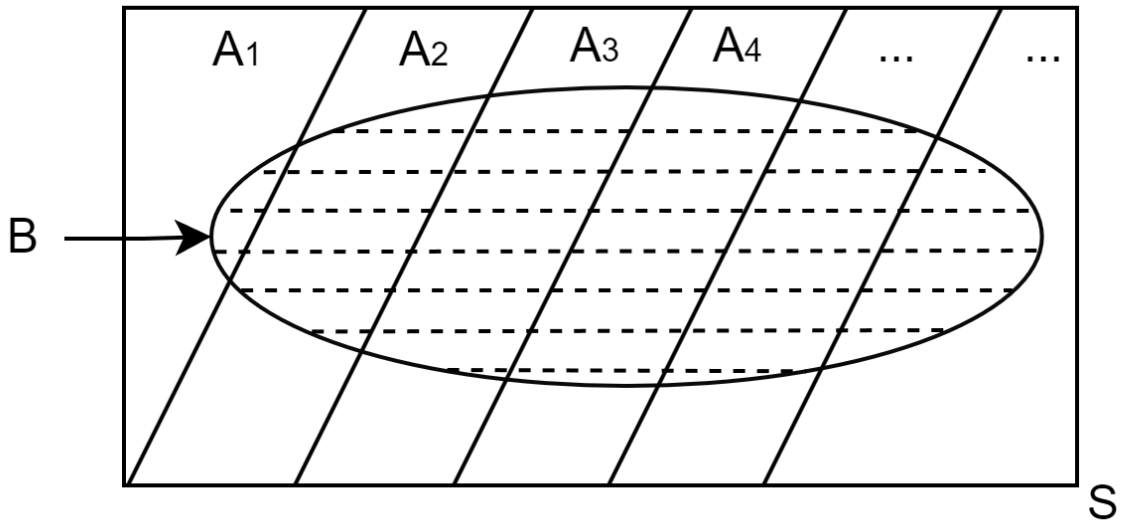
For a special even B, we know
$\{P(B|A_i),\ i = 1, 2, 3, \dots\} \longleftarrow$ conditional probabilities

Generalization

Then, we can express $P(A_k|B)$ for any k $(k = 1, 2, 3, \dots)$ as

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \cdots}$$

$\longrightarrow$ posterior probability



$\{A_1, A_2, A_3, \dots\}$ is a partition of $\mathcal{S}$

Another Application (US 2018 Health Insurance)

Age distribution of Americans in 2018

$$A_1 = \text{Age under 18} \qquad P(A_1) = 22.8\%$$
$$A_2 = \text{Age 18-64} \qquad P(A_2) = 61.4\%$$
$$A_3 = \text{Age above 64} \qquad P(A_3) = 15.8\%$$

B = event of not having health insurance

$$P(B|A_1) = 5.1\% \qquad P(B|A_2) = 12.4\% \qquad P(B|A_3) = 1.1\%$$

A person has been selected at random

$$P(\text{Age above 64}|\text{No health insurance}) = P(A_3|B)$$
$$= 0.0194 = 1.94\% \approx 2\%$$

Classical Case

Now we expand the idea of Bayes' Theorem in a classical statistical set-up as follows:

$$\theta = \text{parameter}$$

$$\mathbf{X} = \text{data}$$

$$\Omega = \text{parameter space}$$

$$f(\mathbf{x}|\theta), \theta \in \Omega : \underline{\text{MODEL}}$$

In a non-Bayesian set-up we use the Likelihood principle to draw inferences about θ or some $\tau(\theta)$ (some known function of θ)

Bayesian Case

In a Bayesian set-up,

we think that $\theta \sim \pi(\theta)$: prior distribution (to be known) $\longrightarrow P(A)$

and, $\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$: conditional distribution (similar to $P(B|A)$)

θ = state of the nature = the true reality

$\mathbf{X}$ = what we see / perceive

After observing $\mathbf{X} = \mathbf{x}$, we want to know
the distribution of $(\theta|\mathbf{X} = \mathbf{x}) \longrightarrow$ Posterior Distribution

Posterior Distribution

The posterior distribution, denoted by $\pi(\theta|\mathbf{x})$, is

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int\limits_{\Omega} f(\mathbf{x}|\theta)\pi(\theta)d\theta}$$

In Bayesian statistics, we use this posterior distribution to draw inferences about θ.

$*$ One may use the posterior mean, or
posterior median, or
posterior mode

to estimate θ.

Bayesian Estimators

- The most common Bayes' estimator of θ is

$$\begin{aligned}\text{the } \underline{\text{posterior mean}} &= \int_{\Omega} \theta\, \pi(\theta|\mathbf{x}) d\theta \\ &= \frac{\int_{\Omega} \theta f(\mathbf{x}|\theta)\pi(\theta) d\theta}{\int_{\Omega} f(\mathbf{x}|\theta)\pi(\theta) d\theta} \\ &= \text{a function of } \mathbf{x} \text{ (data)}\end{aligned}$$

- But the main criticism against Bayesian statistics is- "How do we know the prior $\pi(\theta)$?"

Bayesian Estimators

- But there are ways to get around this criticism.
  - NO INFORMATION ABOUT θ IS INFORMATION ITSELF
  - This has given rise to Bayesian statistics with non-informative priors and Emperical Bayes statistics
- Let us take a simple example as follows:
  5 tosses of a new coin: $\{T, T, T, H, T\}$, $\theta = P(H)$,
  Estimate θ

Bayesian Estimators

X =the number of H's $\sim$ Bino$(5, \theta)$

$\hat{\theta}_{\text{MLE}} = X/n \longrightarrow 1/5$ with $X = 1,\ n = 5.$

In fact $\hat{\theta}_{\text{MLE}}$ can take the value

$$0 \text{ w.p. } 0.03125$$
$$1 \text{ w.p. } 0.03125$$
$$(\text{assuming } \theta = 0.5)$$

but we know that $\theta \in (0, 1)$, possibly near 0.5.

- In Bayesian statistics, we take a prior $\pi(\theta)$ over $\Omega = (0, 1)$ as Beta(a, b) distribution.
- The posterior distribution of $\theta | X = x$ is
$$\text{Beta}(x + a, n - x + b)$$
- $a,\ b \longrightarrow$ prior parameters $=$ hyper parameters

Bayesian Estimators

- So, the posterior mean (the most common Bayes estimator)

$$\longrightarrow \hat{\theta}_B = \frac{x + a}{n + a + b} \in (0, 1)$$

In the above example with $n = 5$ and $X = 1$,

$$\hat{\theta}_B = \frac{1 + a}{5 + a + b} \quad \text{But we do not know a and b.}$$

If $a = b = 1$, $\implies$ Beta(a, b) = Uniform$(0, 1)$ distribution

$$\implies \hat{\theta}_B = 2/7$$

Bayesian Estimators

If $a = b = 0.5 \implies \text{Beta}(0.5, 0.5) = \text{Jeffreys prior}$
$\implies \hat{\theta}_B = 1.5/6 = 0.25$

In the Emperical Bayes method, $\hat{\theta}_B = \frac{x+\hat{a}}{n+\hat{a}+\hat{b}}$

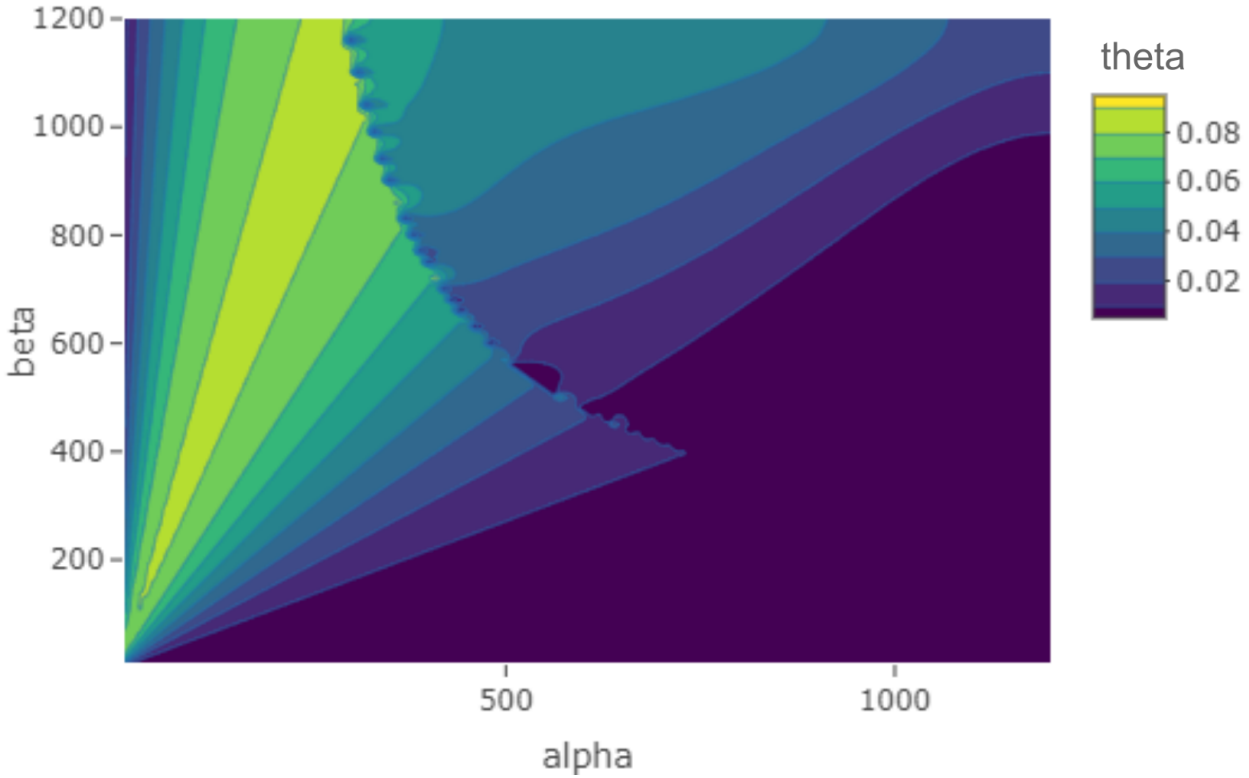
where $\hat{a}$ and $\hat{b}$ are called Type-II maximum Likelihood estimators of a and b found by maximizing the marginal distribution of X

$$
\begin{aligned}
m(x|a,b) &= \int_0^1 f(x|\theta)\pi(\theta)d\theta \\
&= \int_0^1 \left\{ \binom{n}{x} \theta^x (1-\theta)^{n-x} \theta^{a-1} (1-\theta)^{b-1} / B(a,b) \right\} d\theta \\
&= \binom{n}{x} B(X+a, n-x+b)/B(a,b),
\end{aligned}
$$

w.r.t. $a > 0$ and $b > 0$

Empirical Bayes estimator $\hat{\theta}_B$

For $x = 1$ and $n = 5$, the marginal is maximized (via the Nelder-Mead algorithm) at $\alpha = 297.357$, $\beta = 1173.786$, giving $m(x|a, b) = 0.111$ and $\hat{\theta}_B = \frac{x+\hat{a}}{n+\hat{a}+\hat{b}} = \frac{298.357}{1178.786} \approx 0.253$.



Q. How do we know the prior $\pi(\theta)$?
A. If we have some specific knowledge about the distribution of θ, then make use of it (This is called Informative Prior).

However, if we don't have any knowledge about the distribution of θ, then that is also some knowledge, and this gives rise to the concept of Noninformative Prior.

There is a lengthy discussion on prior selection presented in James O. Berger's book - 'Statistical Decision Theory & Bayesian Analysis'.

A prior $\pi(\theta)$ is called a conjugate prior if the posterior distribution $\pi(\theta|\mathbf{x})$ belongs to the same family as $\pi(\theta)$.

For example:

(i) $X|\theta \sim \text{Bino}(n, \theta)$

$\theta \sim \pi(\theta) = \text{Beta}(a, b)$

$\longrightarrow$ Posterior: $\pi(\theta|x) = \text{Beta}(x + a, n - x + b)$

(ii) $X|\theta \sim \text{N}(\theta, \sigma^2)$

$\theta \sim \pi(\theta) = \text{N}(\mu, \tau^2)$

$\longrightarrow$ Posterior: $\pi(\theta|x) = \text{N}(\eta, \dfrac{\sigma^2\tau^2}{\sigma^2 + \tau^2})$

where $\eta = \left(\dfrac{\sigma^2}{\sigma^2 + \tau^2}\right)\mu + \left(\dfrac{\tau^2}{\sigma^2 + \tau^2}\right)x$

A few words about noninformative prior

(i) Suppose θ is a location paramter, i.e., $X|\theta \sim f(x|\theta) = f(x-\theta)$. When we have no idea about the distribution of θ, then

$$P_\pi(\theta \in (a,b)) = P_\pi(\theta \in (a+\epsilon, b+\epsilon)) \forall a,\ b,\ \epsilon$$

$$\text{i.e.,} \int_a^b \pi(\theta)d\theta = \int_{a+\epsilon}^{b+\epsilon} \pi(\theta)d\theta$$

$$\text{i.e.,} \int_a^b \pi(\theta)d\theta = \int_a^b \pi(\theta-\epsilon)d\theta\ \forall a,\ b,\ \epsilon$$

$$\iff \pi(\theta) = \pi(\theta-\epsilon) \forall \epsilon \in \Omega = \mathbb{R}$$

$$\iff \pi(\theta) = \pi(0) = \text{constant}$$

i,e, $\theta \sim \pi(\theta) =$ constant on Ω.

(ii) Similarly, if θ is a scale parameter, i.e., $X|\theta \sim f(x|\theta) = \frac{1}{\theta} f(\frac{x}{\theta}),\ \theta > 0$, then $\theta \sim \pi(\theta) = (\text{costant}/\theta)$

(iii) In general. for any parameter θ, we use Jeffreys' noninformative prior defined as

$$\pi_J(\theta) = \sqrt{I(\theta)}, \text{ where}$$

$I(\theta) =$ Fisher Information $= E\left(\{\nabla\theta(ln)f(\mathbf{X}|\theta)\}^2\right)$

For a multidimensional parameter vector $\boldsymbol{\theta}$, we have

$$\pi_J(\boldsymbol{\theta}) = |I(\boldsymbol{\theta})|^{\frac{1}{2}}$$

Note: A prior $\pi(\theta)$ may not have a finite integral, i.e., we may have $\int \pi(\theta)d\theta = \infty$, but what we need is that the posterior should be a probability distribution in order to draw inferences.

Today Bayesian Statistics is a highly developed area which is used in every branch of science,

from agriculture to forestry
from sociology to psychology
from Data mining to AI,
just to name a few.

Thanks!

Session-II: Challenges in Bayesian Statistics:
From Gibbs Sampling to MCMC Algorithm
with Applications

Nabendu Pal
Department of Mathematics
University of Louisiana at Lafayette
Lafayette, Louisiana, USA

Some remarks

- The main challenge in Bayesian Statistics is the computations
- Quite often we find ourselves with intractable marginal distribution $m(\mathbf{x}) = \int_\Omega f(\mathbf{x}|\theta)\pi(\theta)d\theta$, and / or posterior distribution $\pi(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta)/m(\mathbf{x})$.
- Several computational methods are available to deal with such issues.

Let us look at a simple problem where compuations are easy.

We are going to look at the recent Interstate Highway casualty figures on I-10 over the Atchafalaya Basin.

Figure: Highway I-10

This Atchafalaya Basin bridge is a major bottleneck in the southern US vital transportation link, and it often gets disrupted due to traffic accidents. The highway patrol is always on edge about this issue.

Figure: Atchafalaya Basin Bridge

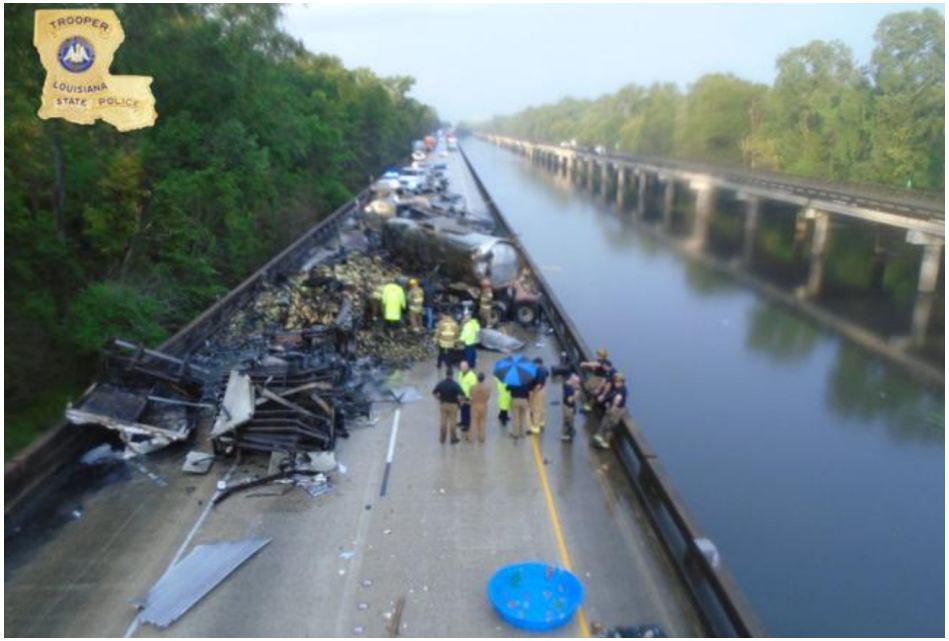


Figure: Traffic accident on the Atchafalaya Basin Bridge, photo by the Louisiana State Police.

We are going to see the casualty figures per year from 2014 to 2021 (over 8 years).

Atchafalaya Basin bridge casualty per year

2014	2015	2016	2017	2018	2019	2020	2021
3	3	3	2	0	3	3	2

Q. For any arbitrary year (after 2021), what is the predicted number of casualties?

A. A classical statistical point of view dictates us to look at the average $\bar{x} = 2.25$.
Hence a predicted value of ≈ 2.
A parametric approach would be to assume that
$X =$ yearly casualty $\sim$Poisson(θ), $\theta > 0$.
$\hat{\theta}_{\text{MLE}} = \bar{x} = 2.25$

[Actually Poisson gives a good fit with Goodness of Fit (GoF)
p-value $\approx 19\%$.]
But, remember that the GoF Test itself is asymptotic and we have only $n = 8$.

In fact, for Poisson(2.25),

$$P(x=0) = 0.1054$$
$$P(x=1) = 0.2371$$
$$P(x=2) = 0.2668$$
$$P(x=3) = 0.2001$$
$$P(x=4) = 0.1126$$

$P(x=2) = 0.2668$ is the mode, so, $(\hat{X}_{\text{future}}|\text{past}) = 2$.

Or one can look at an approximate CI as $[\bar{x} \pm (\text{constant} \cdot \sqrt{\bar{x}})]$

$= [2.25 \pm c(1.5)]$

- The Bayesian viewpoint is different for prediction
- Let the given data: $\mathbf{X}|\theta \sim f(\mathbf{x}|\theta)$ ["Model"]
- Let $\theta \sim \pi(\theta)$ ["Prior"]
- We get $(\theta|\mathbf{X} = \mathbf{x}) \sim \pi(\theta|\mathbf{x})$ ["Posterior"] = Updated info about θ given $\mathbf{X} = \mathbf{x}$.
- Interested in some future observation Y such that $Y|\theta \sim g(y|\theta)$
- So, we construct the distribution of Y given $\mathbf{X} = \mathbf{x}$ as

$$h(y|\mathbf{x}) = \int_{\Omega} g(y|\theta) \cdot \pi(\theta|\mathbf{x}) d\theta$$

We call $h(y|\mathbf{x})$ as the predictive distribution of Y given $\mathbf{X} = \mathbf{x}$.

$g(y|\theta) \cdot \pi(\theta|\mathbf{x})$ is the mixing of the distribution of Y with the updated information about θ after observing $\mathbf{X} = \mathbf{x}$.

So, let us implement this predictive inference in the Bayesian set up.

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\theta),\ \theta > 0$$

Sufficient statistic: $X = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\theta)$

i.e., $X|\theta \sim f(x|\theta) = e^{-n\theta}\frac{(n\theta)^x}{x!},\ \theta > 0.$

Usual conjugate prior: $\pi_c(\theta|\alpha,\beta) = \text{Gamma}(\alpha,\beta)$
$$= \frac{1}{\Gamma(\alpha)\beta^\alpha} \cdot e^{-\theta/\beta} \cdot \theta^{\alpha-1}$$

Noninformative Jeffreys' prior: $\pi_J(\theta) = \sqrt{I(\theta)} \propto \theta^{-1/2}$

[Note, $\pi_c(\theta|\alpha,\beta) \underset{\alpha=1/2,\ \beta\to\infty}{\longrightarrow} \pi_J(\theta)$]

Priors

Jeffreys' $\pi_J(\theta)$:

Marginal of $X \sim m_J(x) = (n^x/x!)\Gamma(x+1/2)n^{-(x+1/2)}$

Posterior: $\theta|x \sim \pi_J(\theta|x) = \text{Gamma}(\alpha_* = x+1/2, \beta_* = 1/n)$

Conjugate $\pi_c(\theta|\alpha,\beta)$:

Marginal of $x \sim m_c(x|\alpha,\beta) = (n^x/x!)\Gamma(x+\alpha)(n+1/\beta)^{-(x+\alpha)}$

Posterior: $\theta|x \sim \pi_c(\theta|x,\alpha,\beta) = \text{Gamma}(\alpha_* = x+\alpha, \beta_* = 1/(n+1/\beta))$

If we use the conjugate Gamma(α, β) prior, then the question is - how to choose the hyperparameters α, β?

In this situation, one can follow the Empirical Bayes' approach to estimate α & β from the data by maximizing the marginal

$$\begin{aligned} m_c(x|\alpha, \beta) &= (n^x/x!)\Gamma(x+\alpha)(n+1/\beta)^{-(x+\alpha)} \\ &\propto m_c^*(x|\alpha, \beta) \text{ (say)} \end{aligned}$$

Note: $(n^x/x!)$ is constant for the given data.

Now, back to the prediction of a future observation:

$$Y|\theta(= X_{(n+1)}|\theta) \sim \text{Poisson}(\theta) = e^{-\theta}\frac{\theta^y}{y!} = g(y|\theta).$$

Predictive distribution of Y:
(i) Under $\pi_J(\theta) \longrightarrow h_J(y|x) = \int_0^\infty g(y|\theta)\pi_J(\theta|x)d\theta$
(ii) Under $\pi_c(\theta|\alpha,\beta) \longrightarrow h_c(y|x,\alpha,\beta) = \int_0^\infty g(y|\theta)\pi_c(\theta|x,\alpha,\beta)d\theta$

$$h_J(y|x) = \frac{\Gamma(y+x+1/2)n^{x+1/2}}{\Gamma(x+1/2)y!(n+1)^{y+x+1/2}}$$

$$h_c(y|x,\alpha,\beta) = \frac{\Gamma(y+x+\alpha)(n+1/\beta)^{x+\alpha}}{\Gamma(x+\alpha)y!(n+1+1/\beta)^{y+x+\alpha}}$$

[Note, $h_c(y|x,\alpha,\beta) \underset{\alpha=1/2,\ \beta\to\infty}{\longrightarrow} h_J(y|x)$]
But, for the conjuigate prior case, we should use
$\alpha \approx \hat{\alpha} = \ldots$ & $\beta \approx \hat{\beta} = \ldots$
which are obtained through the empirical Bayes approach.

More on computations

In our application with a real dataset, we have seen the need to maximize a distribution and / or finding a suitable expectation.
In a complex application, where the distribution (either marginal or posterior) is complicated, we have to use approximation.

Think of approximating

$$E(G(W)) = \int G(w)p(w)dw < \infty, \text{ where } W \sim p(w).$$

If we can generate iid $W_i \sim p(w)$, then

$$E(G(W)) \simeq \frac{1}{N}\sum_{i=1}^{N} G(W_i). \text{ [Monte Carlo Method]}$$

The Monte-Carlo method depends on the Strong Law of Large Numbers.

Strong Law of Large Numbers (SLLN):

$V_1, V_2, \ldots, V_N$ iid $\sim q(v)$ (pdf / pmf) with finite mean μ_V. Then

$$\sum_{i=1}^{n} V_i/N \longrightarrow \mu_V \text{ almost surely as } N \longrightarrow \infty$$

Q. How do we generate $W_i \sim p(w)$ (known)?

A. For most of the common probability distributions, we have random value generators

Let $F(\cdot)$ be the cdf of the pdf/pmf $p(w)$.
Define $W^* = F(W)$ where $W \sim p(w)$.
Then $W^* \sim \text{Uniform}(0, 1)$.

So, first generate W^*, and then get $W = F^{-1}(W^*)$.

Caution: This works fine as long as we have a tractable F so that we can find F^{-1}. Otherwise, it is not possible (or, difficult) to generate $W \sim p(w)$.

When it is difficult to generate "data" from $p(w)$, we can follow "importance sampling" as follows:

$$E(G(W)) = \int G(w)p(w)dw \simeq \frac{1}{N}\sum_{1}^{N} G(W_i),\ W_i \sim p(w)$$

$$= \int G(w)\frac{p(w)}{q(w)}q(w)dw,$$

where $q(w)$ is an easier distribution to work with

$$= \int G_q^*(w)q(w)dw,\ G_q^*(w) = G(w)\frac{p(w)}{q(w)}$$

$$\simeq \frac{1}{N}\sum_{i=1}^{N} G_q^*(W_i),\ W_i \sim q(w)$$

Note: A major concern is the variance of $(\sum_1^N G_q^*(W_i)/N)$ depending on $q(\cdot)$.

Gibbs Sampling

Another way to deal with $E(G(W)), W \sim p(w)$, is through the Gibbs sampler where we work through a pair (W, V). We can generate a sample from $p(w)$ by sampling from the conditional distributions $p_*(w|v)$ and $p_{**}(v|w)$ which are easier to deal with.

Gibbs sequence:

$$V_0', W_0', V_1', W_1', V_2', W_2', \ldots, V_R', W_R', \ldots$$

$$w_j' \sim p_*(w|v_j'),\ v_{j+1}' \sim p_* * (v|w_j')$$

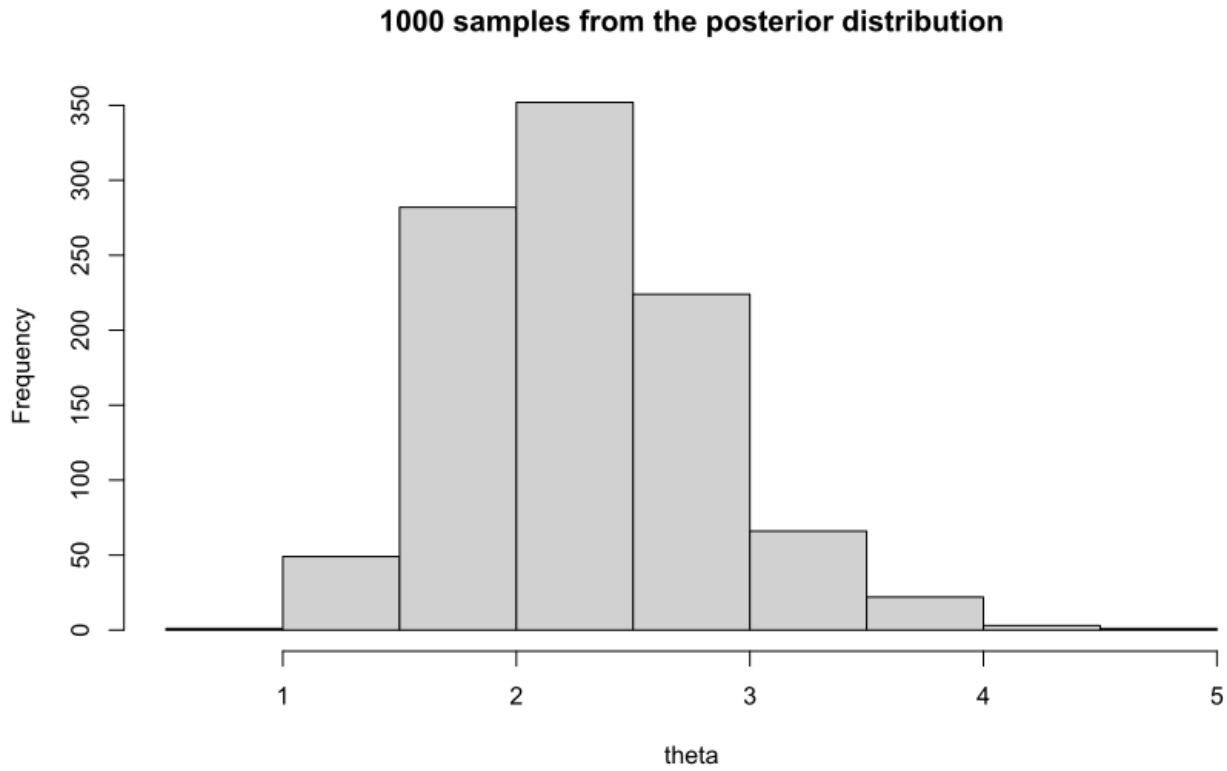
[The initial value $V_0' = v_0'$ is specified.]

Remarks:

- Under fairly general conditions, as $k \to \infty$, the distribution of W'_k is approximately that of W.
- This is a simple form of MCMC.
- Starting with N different seed values of V'_0 we can obtain N different values of W'_k ($\approx W$), thereby giving us an effective sample of size N from $p(w)$. [Often, $k = 100$ is good enough.]

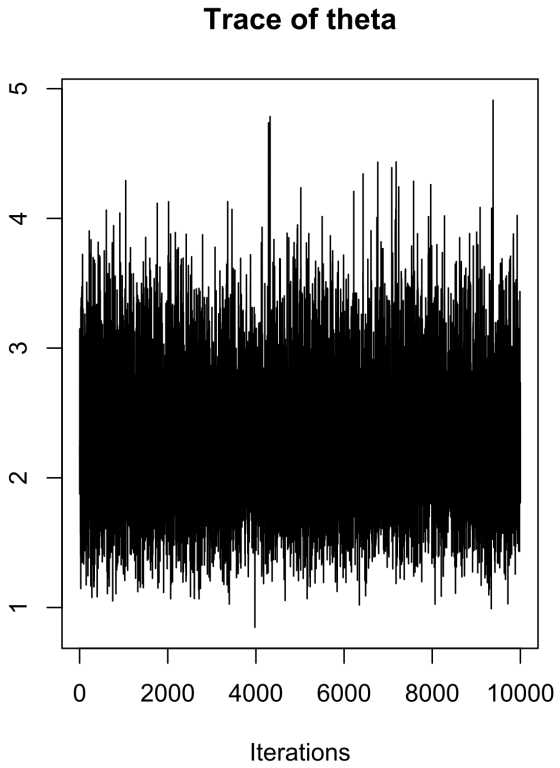
Gibbs example

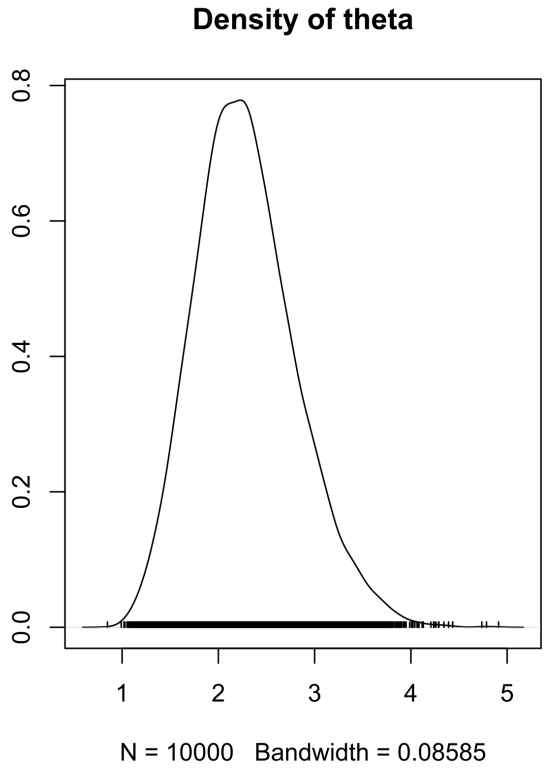
We will model Atchafalaya Basin bridge casualty dataset (x=18, n=8) with the Poisson distribution and the Jeffereys' prior. The histogram below shows 1000 samples generated from the posterior distribution after convergence.



Gibbs example

Below is the full trace of the Gibbs sampling procedure from the posterior distribution.





The functionality of an MCMC algorithm depends on the theory of Ergodic Chain (a specialized form of a Markov Chain) which converges to a stationary distribution.

Another MCMC method is the Metropolis Algorithm (and further refined Metropolis-Hastings Algorithm).
Quite often, we know the structure of $p(w)$ where $W \sim p(w)$, except the normalizing constant (which may not be easy to obtain), i.e.,

$$p(w) \propto h^*(w) \text{ [i.e., } p(w) = ch^*(w)]$$

Knowing $h^*(w)$ is good enough to simulate from $p(w)$.

Simulating from $p(w)$

- Start with $W = w_0 = w_{\text{current}}$
- Generate $W = w_{\text{new}}$ in the neighborhood of w_0

$$W \sim N(w_{\text{current}}, \delta),\ \delta = \text{known}$$

  - Look at the likelihood ratio $\Lambda = \frac{p(w_{\text{new}})}{p(w_{\text{current}})} = \frac{h^*(w_{\text{new}})}{h^*(w_{\text{current}})}$
- If $\Lambda > 1$, then update $W = w_{\text{new}} = w_1$
- If $\Lambda < 1$, then $w_1 = w_{\text{new}}$ w.p. Λ and $w_1 = w_0$ w.p. $(1 - \Lambda)$
- Use $W = w_1 = w_{\text{current}}$
- Repeat this a large number of times (say, $k = 10^3$ times) to converge.

So, to mimic a posterior distribution

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d\theta},$$

all we need is to look at

$$\Lambda = \frac{\pi(\theta_{\text{new}}|x)}{\pi(\theta_{\text{current}}|x)} = \frac{f(x|\theta_{\text{new}})\pi(\theta_{\text{new}})}{f(x|\theta_{\text{current}})\pi(\theta_{\text{current}})};$$

and move from θ_{current} to θ_{new} w.p. $\min(1, \Lambda)$
and stay at θ_{current} w.p. $\{1\text{-}\min(1, \Lambda)\}$.

Metropolis-Hastings Algorithm generalizes the above by injecting an extra asymmetric transition probability.

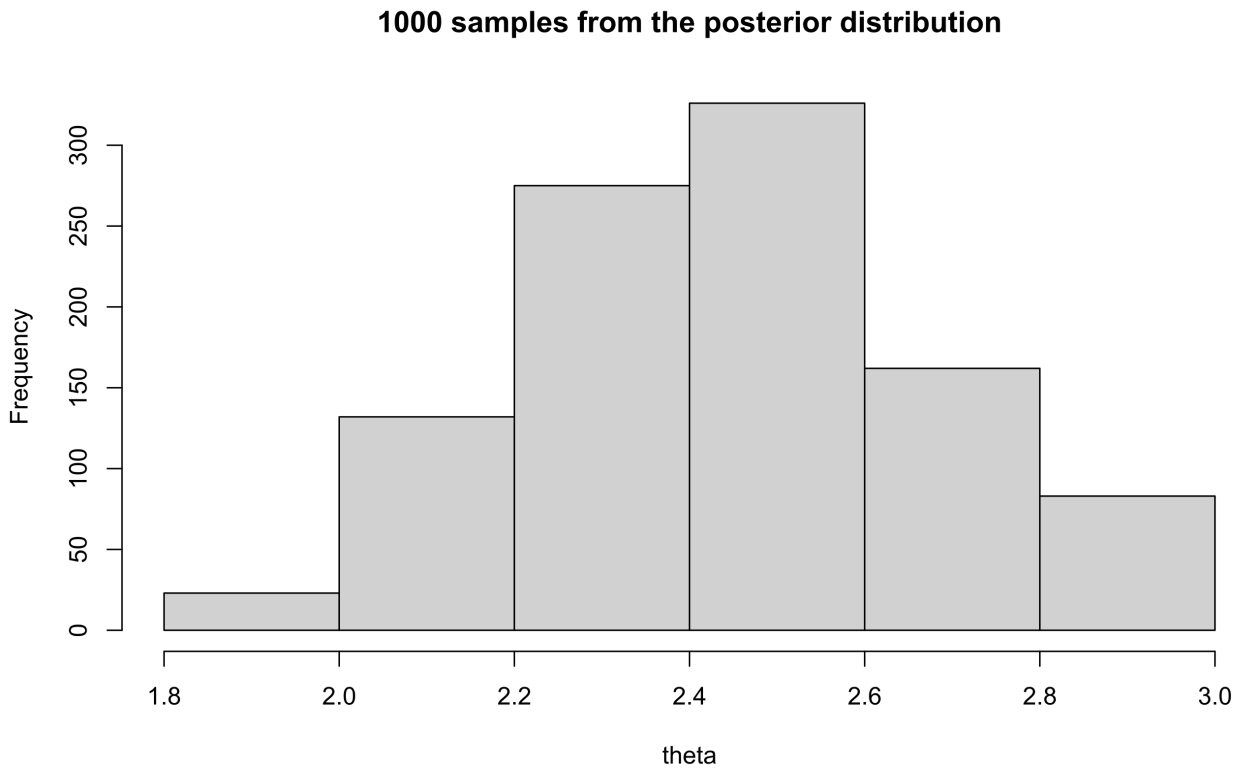
Look at

$$\alpha_* = \min\left\{1, \frac{p(w_{\text{new}})q(w_{\text{current}}|w_{\text{new}})}{p(w_{\text{current}})q(w_{\text{new}}|w_{\text{current}})}\right\}$$

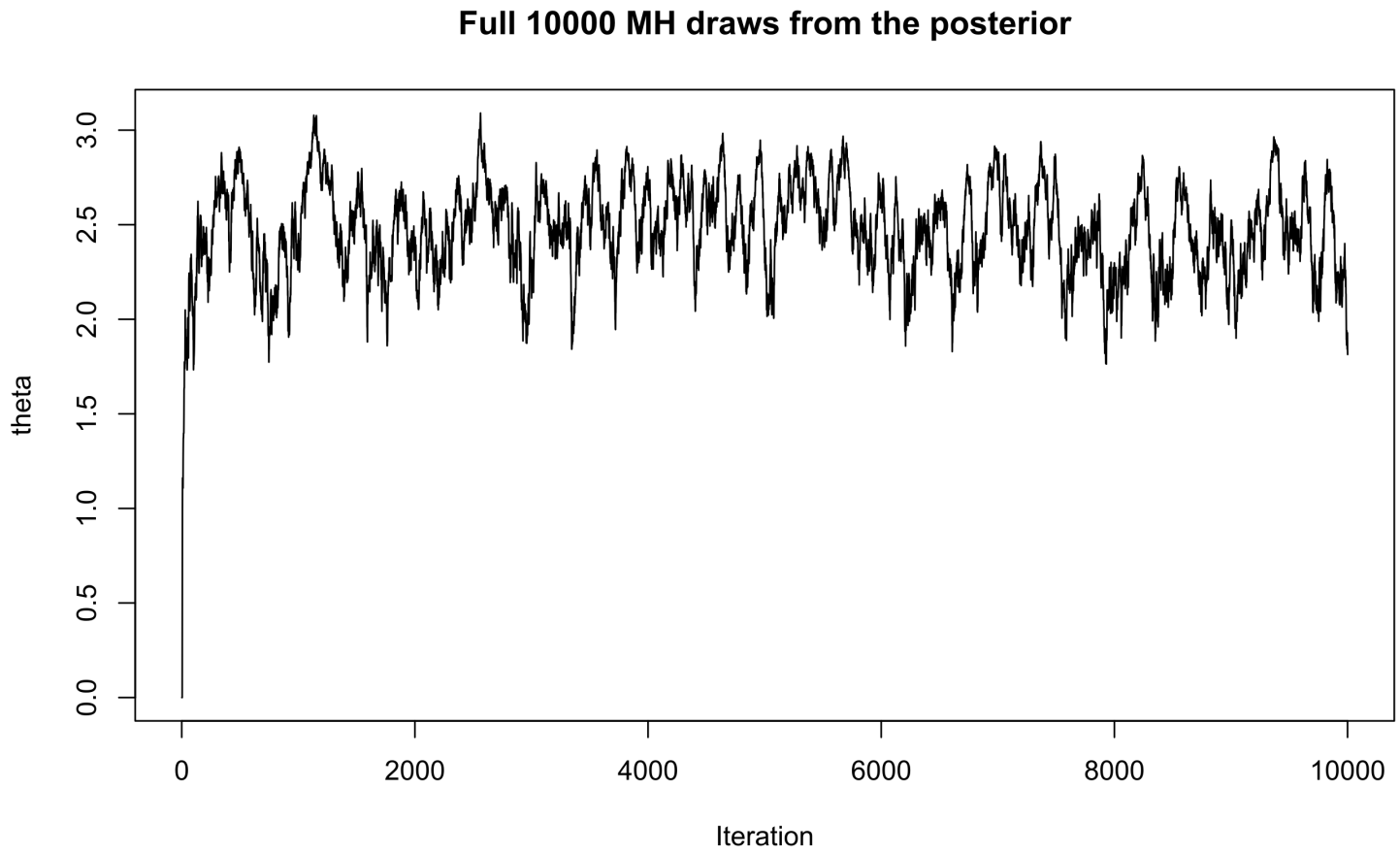
then go from w_{current} to w_{new} w.p α_*
and stay at w_{current} w.p. $(1 - \alpha_*)$.

Metropolis example

Using the same highway accident data and the initial value ($\theta = 1$) as for Gibbs, we generated the posterior with a Metropolis algorithm. The histogram below shows the final 1000 draws from the posterior.



Metropolis example



A word of caution: MCMC Algorithms are like a blackbox. What really goes on inside is beyond our control, but we just hope that it works.

Thanks!

Acknowledgement

I would like to thank Mr. Daniel Fuller, doctoral student, Clarkson University, New York, for his time and patience to prepare these slides and the computations.