Bayesian Analysis Monte Carlo-Markov Chains

September 17, 2019

Alejandro Ortega, FCAS, CFA, MAAA

Antitrust Notice

- The Casualty Actuarial Society is committed to adhering strictly to the letter and spirit of the antitrust laws. Seminars conducted under the auspices of the CAS are designed solely to provide a forum for the expression of various points of view on topics described in the programs or agendas for such meetings.
- Under no circumstances shall CAS seminars be used as a means for competing companies or firms to reach any understanding – expressed or implied – that restricts competition or in any way impairs the ability of members to exercise independent business judgment regarding matters affecting competition.
- It is the responsibility of all seminar participants to be aware of antitrust regulations, to prevent any written or verbal discussions that appear to violate these laws, and to adhere in every respect to the CAS antitrust compliance policy.

Overview

- Chocolate Chips – MLE & Bootstrap
- Chocolate Chips – a Bayesian Analysis
- More Complex Model – Bayesian Analysis
- Metropolis Hastings
- Gibbs Sampling
- Tests

Chocolate Chips – MLE & Bootstrap

- Curious to know how many chocolate chips are in a cookie
- Estimate of the Mean (# of Chips in a Cookie)
- Distribution of the Mean
- Distribution of Chips in Cookies
- Draw 6 cookies

5, 5, 7, 10, 10, 11

Chocolate Chips – MLE & Bootstrap

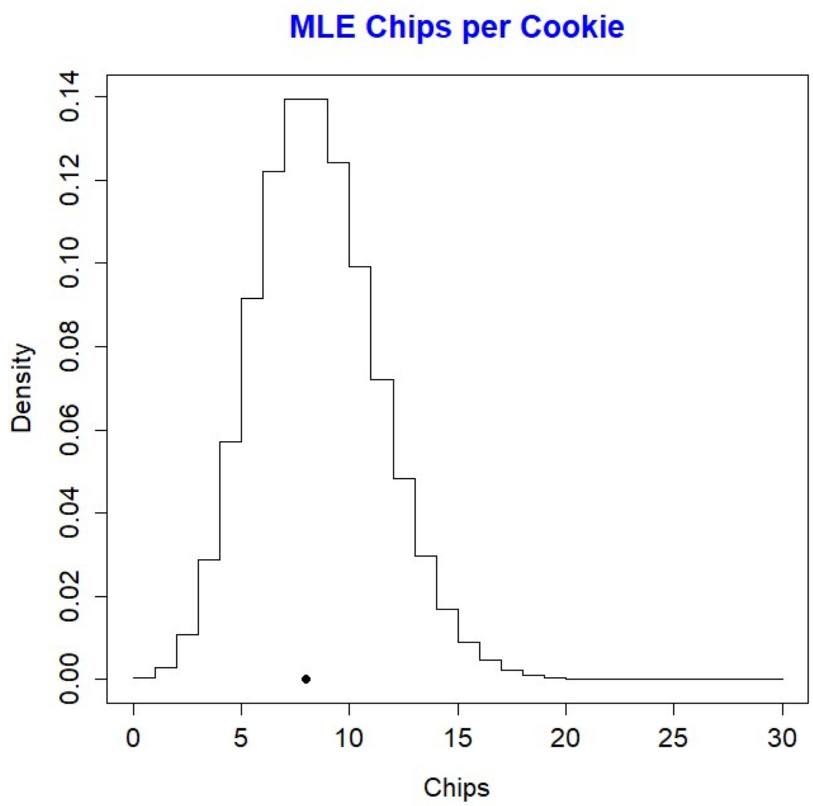
5, 5, 7, 10, 10, 11

- Assume they follow a Poisson Distribution
- The Maximum Likelihood Estimator is the Sample Mean

$$\mu = \bar{x} = \frac{5 + 5 + 7 + 10 + 10 + 11}{6} = 8.00$$

Chocolate Chips – MLE & Bootstrap

5, 5, 7, 10, 10, 11



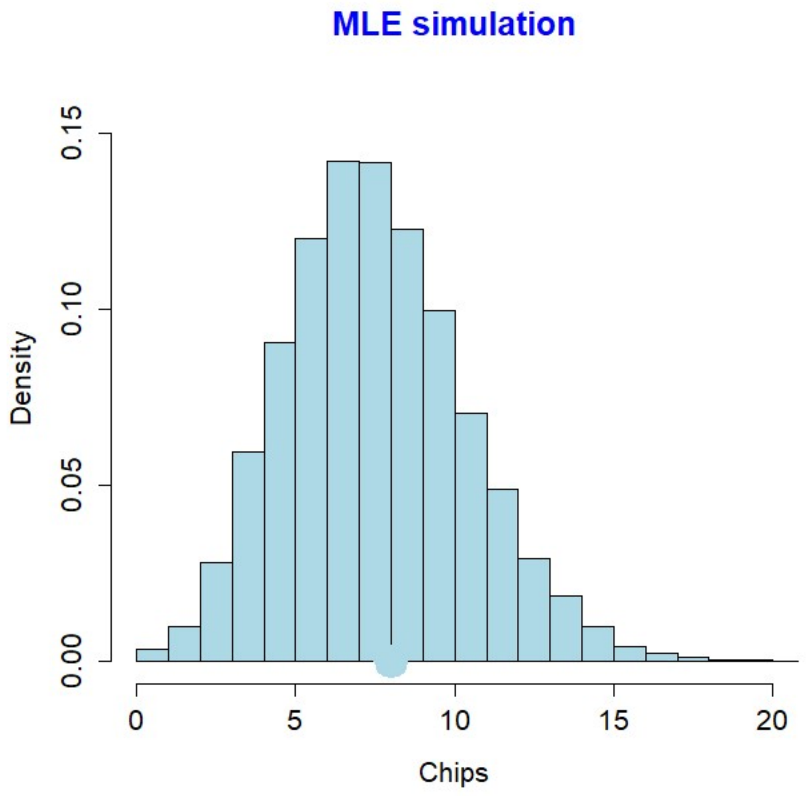
- Does not consider Parameter Risk for λ
- Assumes we estimated λ perfectly
- The Distribution here only has Process Risk

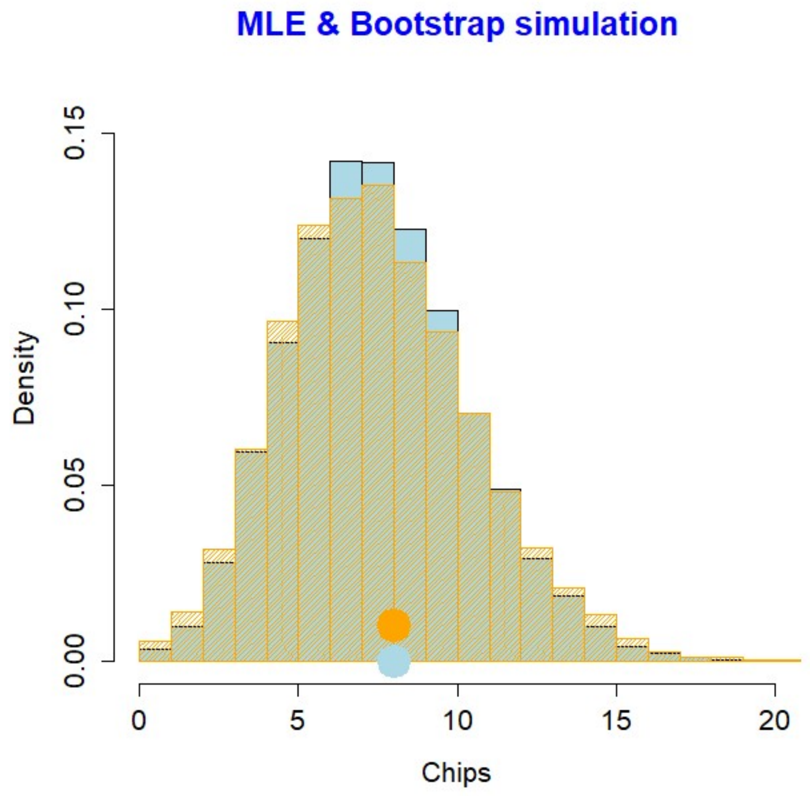
Chocolate Chips – MLE & Bootstrap

5, 5, 7, 10, 10, 11

- To add Parameter Risk, we can **Bootstrap**
- We draw 6 “cookies”, from our set above, with replacement
- Calculate a sample mean
- Draw one cookie from Poisson with this sample mean
- Repeat 20,000 times

Chocolate Chips – MLE & Bootstrap





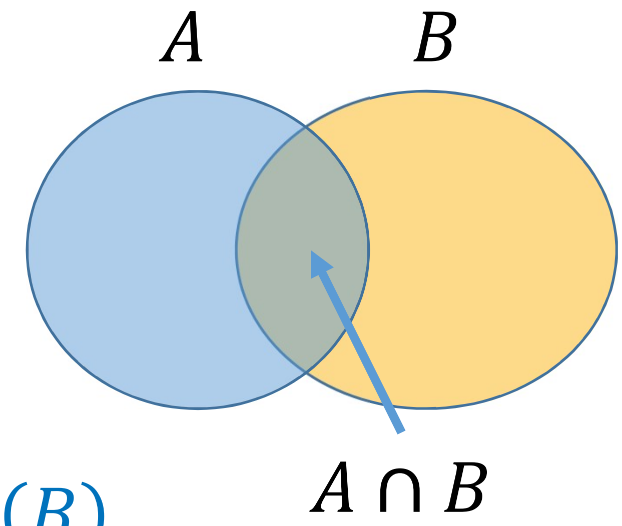
Chocolate Chips – a Bayesian Analysis

- Bootstrap was useful in adding Parameter Risk
- Bayesian Analysis provides another way to do this
- Also allows us to consider Expert Opinion

Chocolate Chips – a Bayesian Analysis

- Bayes Theorem

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

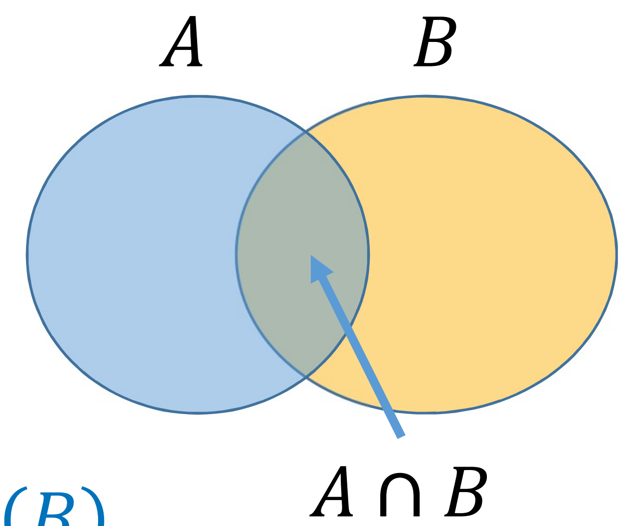


$$P(A \cap B) = P(A|B) \cdot P(B)$$
$$= P(B|A) \cdot P(A)$$

Chocolate Chips – a Bayesian Analysis

- Bayes Theorem

$$P(A|B) = \frac{\boldsymbol{P(A \cap B)}}{P(B)}$$



$$P(A \cap B) = P(A|B) \cdot P(B)$$
$$= \boldsymbol{P(B|A) \cdot P(A)}$$

$$\boldsymbol{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

Chocolate Chips – a Bayesian Analysis

$$\boldsymbol{P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}}$$

- Using Continuous Density notation $f(\cdot)$
- Let y be a set of data
- Let θ be a collection of parameters

$$\boldsymbol{f(\theta|y) = \frac{f(y|\theta) \cdot f(\theta)}{f(y)}}$$

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{\boldsymbol{f(y|\theta)} \cdot f(\theta)}{f(y)}$$

- $f(y|\theta)$ – is the Likelihood (LL)
  - This is the density of our data, given a set of parameters
  - The set of parameters, $\hat{\theta}$ that maximizes the Likelihood are called the Maximum Likelihood Estimators - MLE
  - The MLE is used often to find a “best” set of parameters

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{f(y|\theta) \cdot \boldsymbol{f(\theta)}}{f(y)}$$

- $f(\theta)$ – is the Prior Distribution of θ
  - This is our Opinion on θ **before** we have collected data
  - This can be an **informed** opinion, or an **uninformed** opinion
  - In the latter case, we typically select a large variance

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{f(y|\theta) \cdot f(\theta)}{\boldsymbol{f(y)}}$$

- $f(y)$ – is the integral of the numerator

$$f(y) = \int \boldsymbol{f(y|\theta)} \cdot \boldsymbol{f(\theta)}\, d\theta$$

- Often – Calculate the numerator, and then integrate to get the denominator
- Allows us to drop constants in the numerator

Chocolate Chips – a Bayesian Analysis

$$\boldsymbol{f(\theta|y) = \frac{f(y|\theta) \cdot f(\theta)}{f(y)}}$$

- $f(\theta|y)$ – is the **Posterior Distribution** of θ
- $f(y|\theta)$ – is the **Likelihood**
- $f(\theta)$ – is the **Prior Distribution** of θ
- $f(y)$ – is the **Normalizing Constant**

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{f(y|\theta) \cdot f(\theta)}{f(y)}$$

- Let's return to our Chocolate Chip problem
- y are the data points (the 6 cookies)
- $y = \{5,5,7,10,10,11\}$
- θ is the parameter of the Poisson distribution
- We will now estimate a distribution for $\theta|y$
- Then, we can estimate a distribution for y

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{f(y|\theta) \cdot \boldsymbol{f(\theta)}}{f(y)}$$

- We need a **Prior** Distribution for $\boldsymbol{\theta}$
- We are not confident in our prior, so select a wide variance
- Select **Gamma**, with mean $\mu = 10$, and $\sigma = 4$
- $\alpha = \frac{\mu^2}{\sigma^2} = 6.25\,;\beta = \frac{\mu}{\sigma^2} = 0.625$

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot \theta^{\alpha-1} \cdot e^{-\beta\theta}$$

$$f(\theta) \propto \theta^{5.25} \cdot e^{-0.625\theta}$$

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{\boldsymbol{f(y|\theta)} \cdot f(\theta)}{f(y)}$$

5, 5, 7, 10, 10, 11

- We need the Poisson Likelihood:

$$f(y_i|\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}$$

$$f(y|\theta) = \prod_{i=1}^{6} f(y_i|\theta) = \prod_{i=1}^{6} \frac{e^{-\theta}\theta^{y_i}}{y_i!}$$

$$\propto e^{-6\theta}\theta^{\sum y_i} = \boldsymbol{e^{-6\theta}\theta^{48}}$$

Chocolate Chips – a Bayesian Analysis

$$f(\theta|y) = \frac{\boldsymbol{f(y|\theta) \cdot f(\theta)}}{f(y)}$$

$$f(y|\theta) \cdot f(\theta) \propto \left[e^{-6\theta}\theta^{48}\right] \times \left[\theta^{5.25} \cdot e^{-0.625\theta}\right]$$

$$= \theta^{\mathbf{53.25}} \cdot e^{\mathbf{-6.625}\theta}$$

- This is the Gamma Distribution:

$$f(\theta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \cdot \theta^{\boldsymbol{\alpha-1}} \cdot e^{\boldsymbol{-\beta}\theta}$$

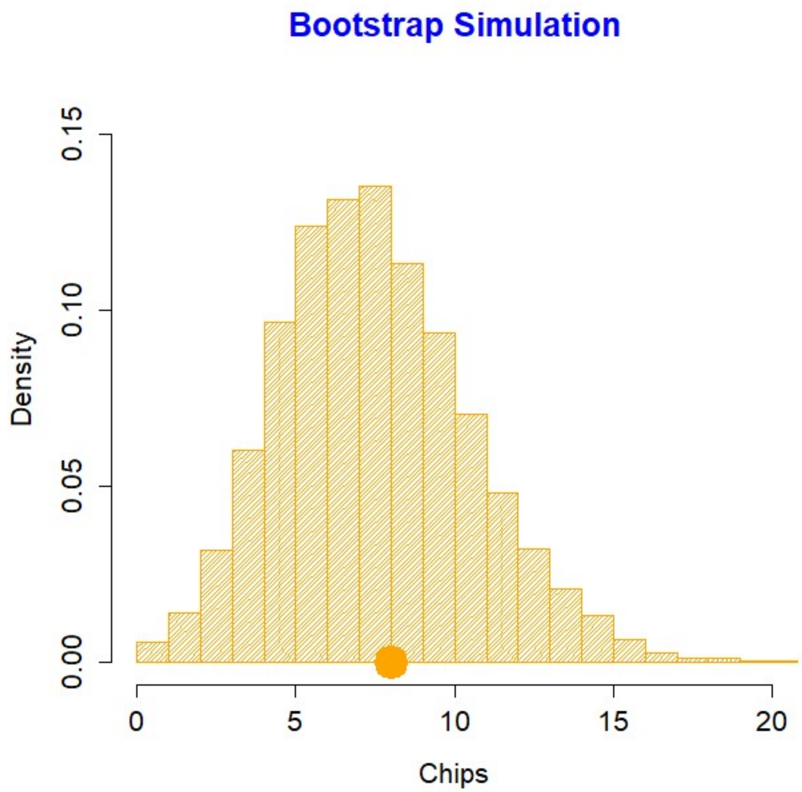
$$\alpha = 54.25$$
$$\beta = 6.625$$

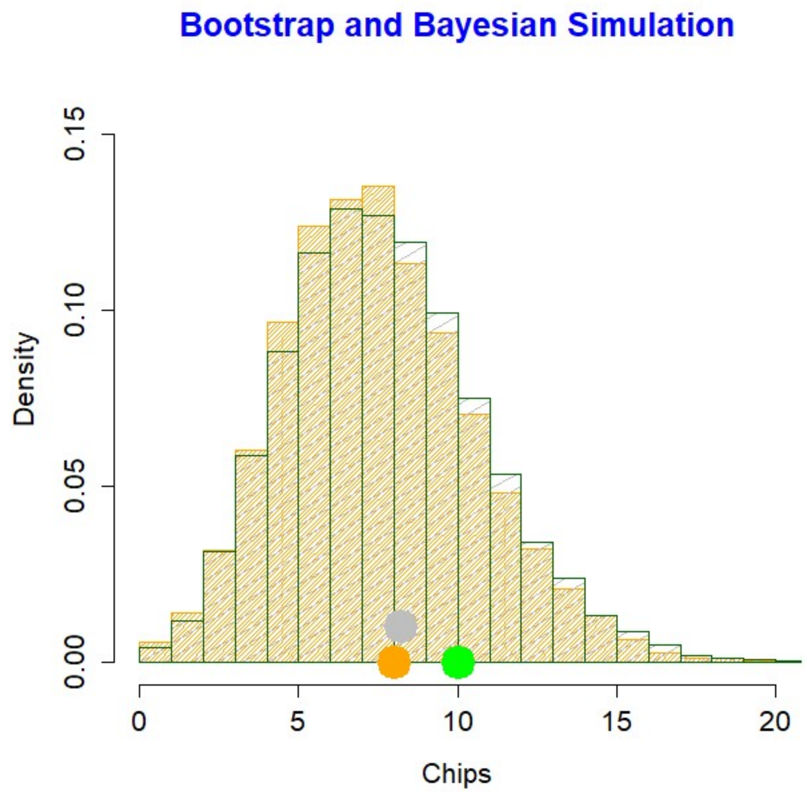
$$\text{Mean} = \frac{\alpha}{\beta} = 8.19 \quad \text{Std. Dev} = \sqrt{\frac{\alpha}{\beta^2}} = 1.11$$

Chocolate Chips – a Bayesian Analysis

- We Calculated the Likelihood Formula, $f(y|\theta)$
- Selected a Prior, $f(\theta)$
- Calculated the Posterior, $f(\theta|y)$
- Draw 20,000 samples from the posterior for θ
- For each θ, draw 1 cookie

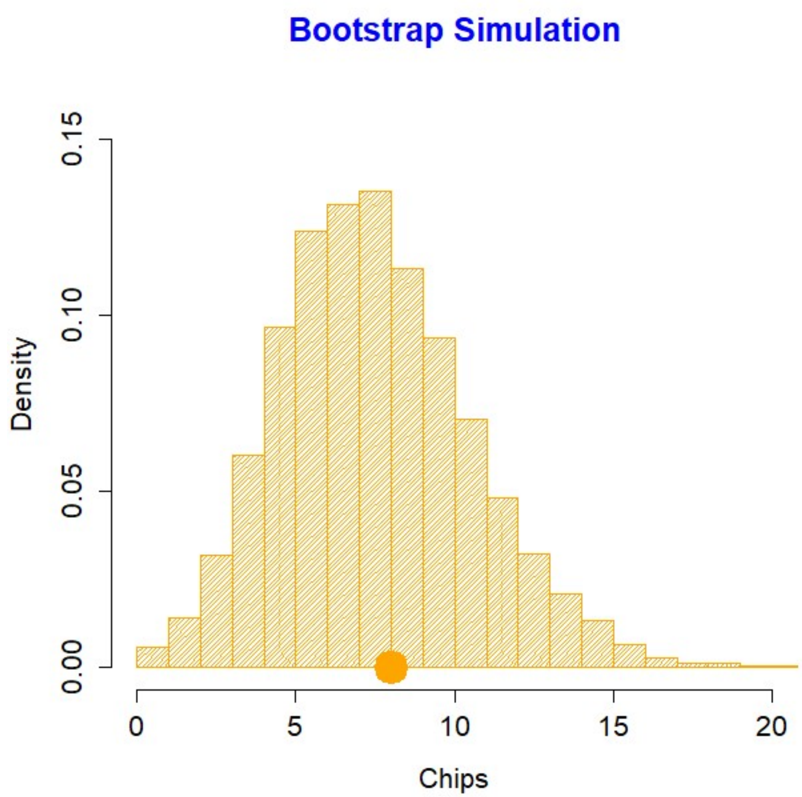
Chocolate Chips – a Bayesian Analysis

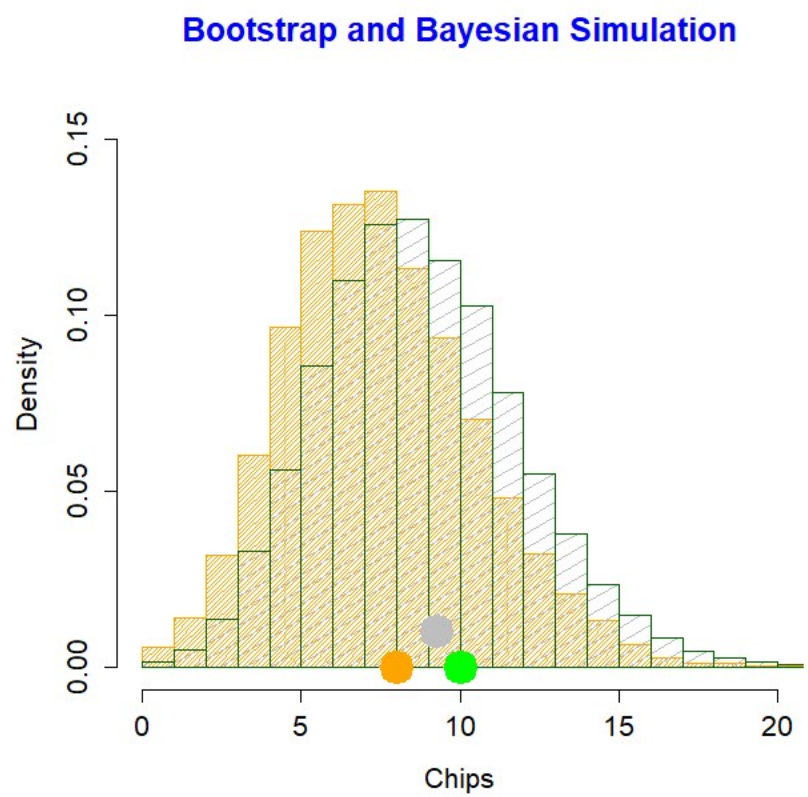




Chocolate Chips – a Bayesian Analysis

- Prior Distribution of Mean of Chips to have:
- Mean= 10, Stdev= 1





Bootstrap vs Bayesian

- Both Bootstrap and Bayesian gave us predictive distribution of the # of Chips in a Cookie
- Bayesian – allowed us to consider our Prior, Expert Opinion

Bayesian Model

- The cookie problem had a relatively simple Posterior Distribution – we lucked out, and it was the Gamma Distribution
- In insurance modeling, we'll have many more parameters, and solving the integral to determine the normalizing constant is intractable

Bayesian Model

- Take a 4x4 Triangle
- Incremental Losses, X_{ij}, are Over Dispersed Poisson, with fixed (but unknown) dispersion parameter ϕ
- 4 Row parameters: α_i ; $i \in 1..4$
- 3 Column Parameters β_j ; $j \in 2..4$
- $\beta_1 = 1$; fixed
- Mean of each cell is: $\alpha_i \cdot \beta_j$
- Variance $= \alpha_i \cdot \beta_j \cdot \phi$
- α_i, β_j have Gamma Priors

Bayesian Model

$$f(X|\alpha,\beta,\phi) = \prod_{(i,j)\in\Delta} \frac{e^{-(\alpha_i\beta_j)/\phi}\left(\frac{\alpha_i\beta_j}{\phi}\right)^{\frac{X_{ij}}{\phi}}}{(X_{ij}/\phi)!}$$

$$f(\alpha,\beta) \propto \prod_{i=1}^{4} \alpha_i^{a_i-1} e^{-b_i\alpha_i} \times \prod_{j=2}^{4} \beta_j^{c_j-1} e^{-d_j\beta_j}$$

$$\alpha_i \sim Gamma(a_i,b_i)\,;\,\beta_j \sim Gamma(c_j,d_j)$$

Bayesian Model

- We would be able to calculate the Posterior Distribution to within a Normalizing Constant
- We would **not** be able to integrate it – it's intractable
- $g(\alpha, \beta, \phi) \propto f(\alpha, \beta | \phi, X)$
- We can find the **ratio** of the density for any set of parameters to any other set of parameters; but we **don't** have the actual density
- There is an algorithm that allows us to **sample** from this distribution – Metropolis Hastings

Metropolis-Hastings

- Markov Chain is a mapping where the probability of the next state is dependent only on the current state
- A continuous version, with a single parameter could be written as a probability density

$$q(\theta_i|\theta_{i-1})$$

Metropolis-Hastings

- If we have, posterior to within a constant:

$$g(\theta) \propto f(\theta|y)$$

- Generating Function $q(\theta_i|\theta_{i-1})$
- The following algorithm will (in the limit) be samples from the distribution with density $f(\theta|y)$

Metropolis-Hastings

- 1. Select an initial θ_0
- 2. Draw θ^* from $q(\theta^*|\theta_{i-1})$
- 3. Calculate α

$$\alpha = \frac{g(\theta^*)/q(\theta^*|\theta_{i-1})}{g(\theta_{i-1})/q(\theta_{i-1}|\theta^*)}$$

- 4. Draw $u \in Unif(0,1)$
- 5. If $u < \alpha$ then $\theta_i = \theta^*$; otherwise $\theta_i = \theta_{i-1}$

Repeat Steps 2-5 many times

Metropolis-Hastings

- In the cookie, example we calculated a Posterior:

$$f(\theta|y) \propto g(\theta) = e^{-6.625\theta} \cdot \theta^{53.25}$$

- Generating Function is Uniform with width 2
- $\theta_0 = 10$

$$q(\theta^*|\theta_0 = 10) = \begin{Bmatrix} 0.5 & : \theta^* \in [9,11] \\ 0 & : \text{else} \end{Bmatrix}$$

Metropolis-Hastings

$$g(\theta) = e^{-6.625\theta} \cdot \theta^{53.25}$$

- $\theta_0 = 10$

$$q(\theta^*|\theta_0 = 10) = \begin{cases} 0.5 & : \theta^* \in [9,11] \\ 0 & : \text{else} \end{cases}$$

$$\theta^* = 9.72$$

$$g(\theta^* = 9.72) = 4.235 \cdot 10^{24}$$

$$g(\theta_0 = 10) = 3.006 \cdot 10^{24}$$

$$q(\theta^*|\theta_0) = q(\theta_0|\theta^*) = 0.5$$

$$\alpha = \frac{g(\theta^*)/q(\theta^*|\theta_0)}{g(\theta_0)/q(\theta_0|\theta^*)} = \frac{g(\theta^*)}{g(\theta_0)} = \frac{4.235}{3.006} = 1.409$$

Since $\alpha > 1$, $\boldsymbol{\theta_1 = \theta^* = 9.72}$

Metropolis-Hastings

$$g(\theta) = e^{-6.625\theta} \cdot \theta^{53.25}$$

- $\theta = \{10, 9.72\}$

$$q(\theta^* | \boldsymbol{\theta_1} = \mathbf{9.72}) = \begin{Bmatrix} 0.5 & : \theta^* \in [8.72, 10.72] \\ 0 & : \text{else} \end{Bmatrix}$$

$$\theta^* = 10.18$$

$$g(\theta^* = 10.18) = 2.359 \cdot 10^{24}$$

$$g(\theta_1 = 9.72) = 4.235 \cdot 10^{24}$$

$$\alpha = \frac{g(\theta^*)}{g(\theta_1)} = \frac{2.359}{4.235} = 0.557$$

$$u \sim Unif(0,1); u = 0.779$$

$$u < \alpha: \ 0.779 < 0.557 \ \ FALSE, \text{do not move}$$

$$\boldsymbol{\theta_2} = \boldsymbol{\theta_1} = \mathbf{9.72}$$

Metropolis-Hastings

$$g(\theta) = e^{-6.625\theta} \cdot \theta^{53.25}$$

- $\theta = \{10, 9.72, 9.72\}$

$$q(\theta^* | \boldsymbol{\theta_2} = \mathbf{9.72}) = \begin{Bmatrix} 0.5 & : \theta^* \in [8.72, 10.72] \\ 0 & : \text{else} \end{Bmatrix}$$

$$\theta^* = 10.49$$

$$g(\theta^* = 10.49) = 1.494 \cdot 10^{24}$$

$$g(\theta_2 = 9.72) = 4.235 \cdot 10^{24}$$

$$\alpha = \frac{g(\theta^*)}{g(\theta_2)} = \frac{1.494}{4.235} = 0.353$$

$$u \sim Unif(0,1); u = 0.119$$

$$u < \alpha: \ 0.119 < 0.353 \ \ TRUE, \text{accept proposal}$$

$$\boldsymbol{\theta_3 = \theta^* = 10.49}$$

Metropolis-Hastings

$$g(\theta) = e^{-6.625\theta} \cdot \theta^{53.25}$$

- $\theta = \{10;\ 9.72;\ 9.72; 10.49\}$

$$q(\theta^*|\theta_3 = 10.49) = \begin{Bmatrix} 0.5 & : \theta^* \in [9.49, 11.49] \\ 0 & : \text{else} \end{Bmatrix}$$

$$\theta^* = 9.63$$

$$g(\theta_2 = 9.63) = 4.684 \cdot 10^{24}$$

$$g(\theta^* = 10.49) = 1.494 \cdot 10^{24}$$

$$\alpha = \frac{g(\theta^*)}{g(\theta_3)} = \frac{4.684}{1.494} = 3.135$$

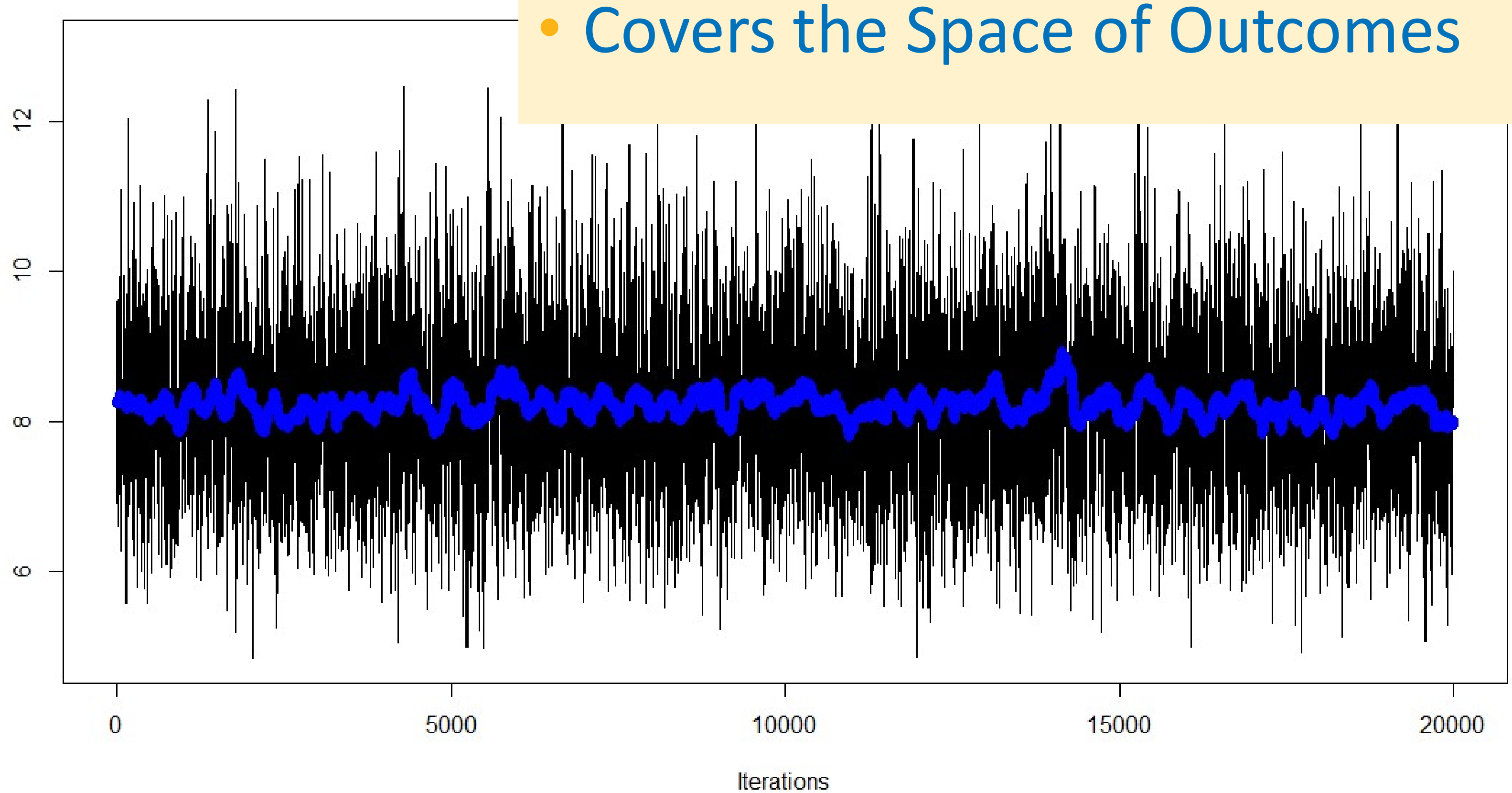
Since $\alpha > 1$, $\boldsymbol{\theta_4 = \theta^* = 9.63}$

Metropolis-Hastings

- Want a Result Like this
- Blue Line is 200 point average

Acceptance Rate: 52%

- Stationary
- Covers the Space of Outcomes

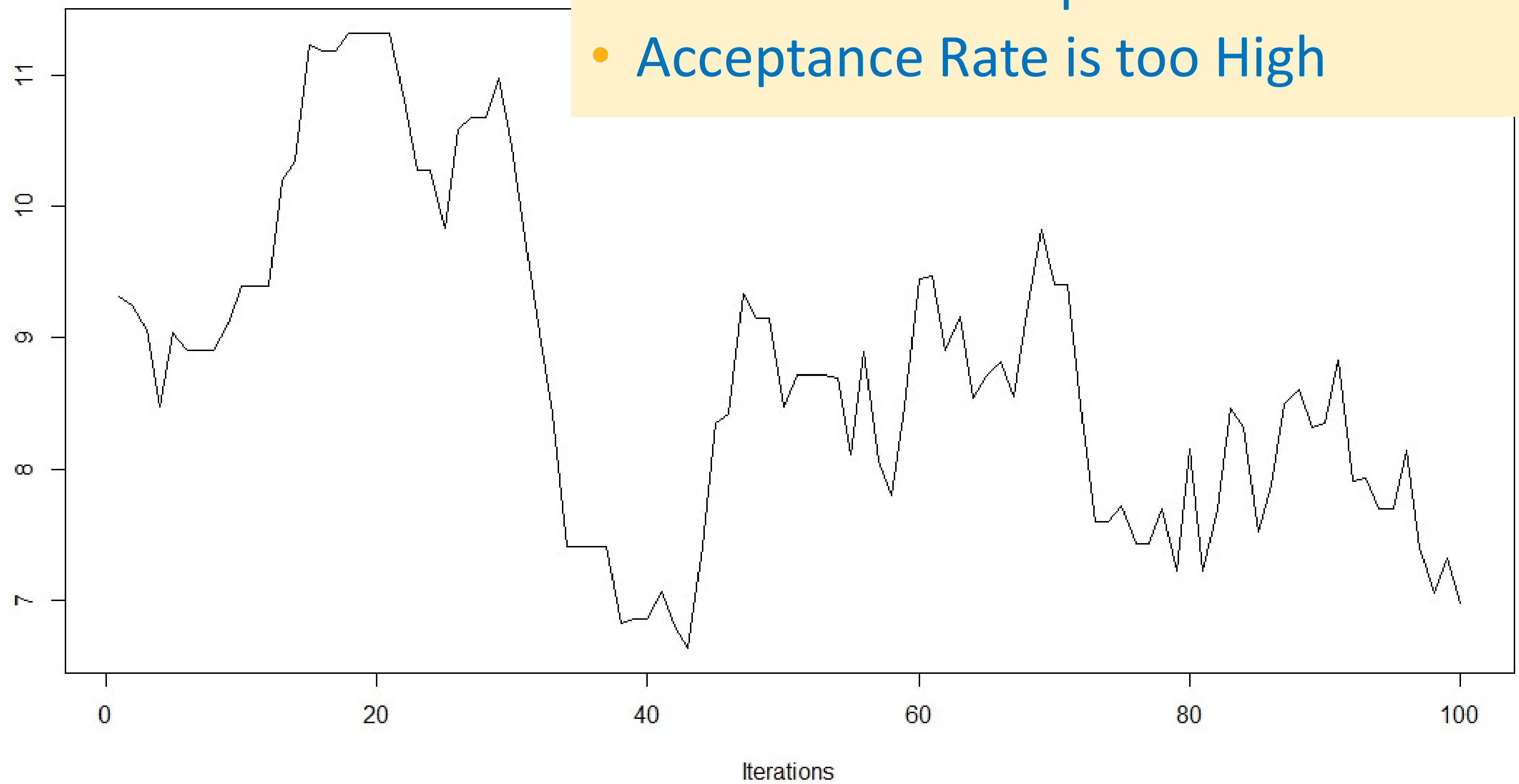


Metropolis-Hastings

- Iterate 100 times

Acceptance Rate: 79%

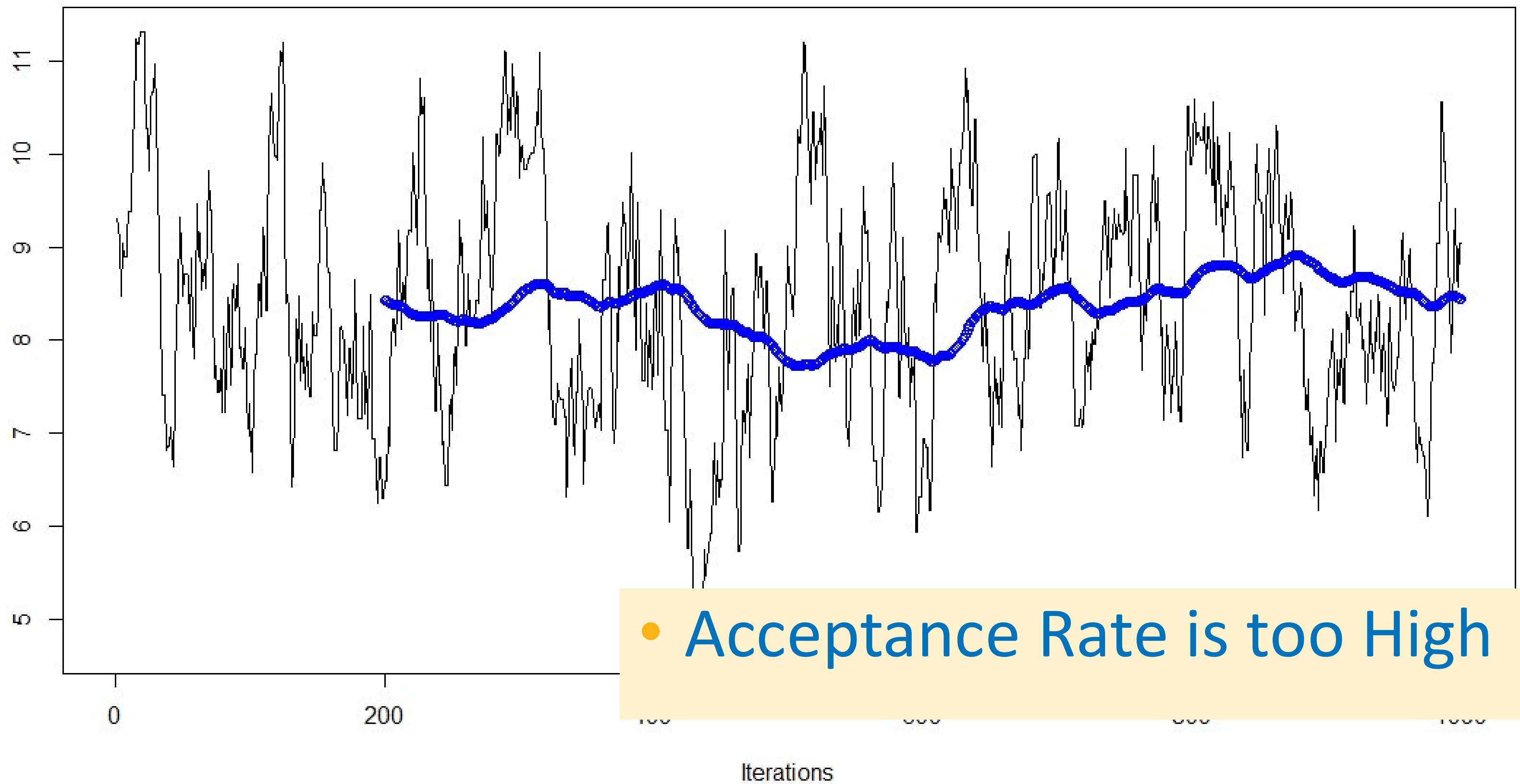
- Too Short
- Does not Cover Space
- Acceptance Rate is too High



Metropolis-Hastings

- Iterate 1,000 times
- Blue Line is 200 point rolling average

Acceptance Rate: 82%



- Acceptance Rate is too High

Metropolis-Hastings

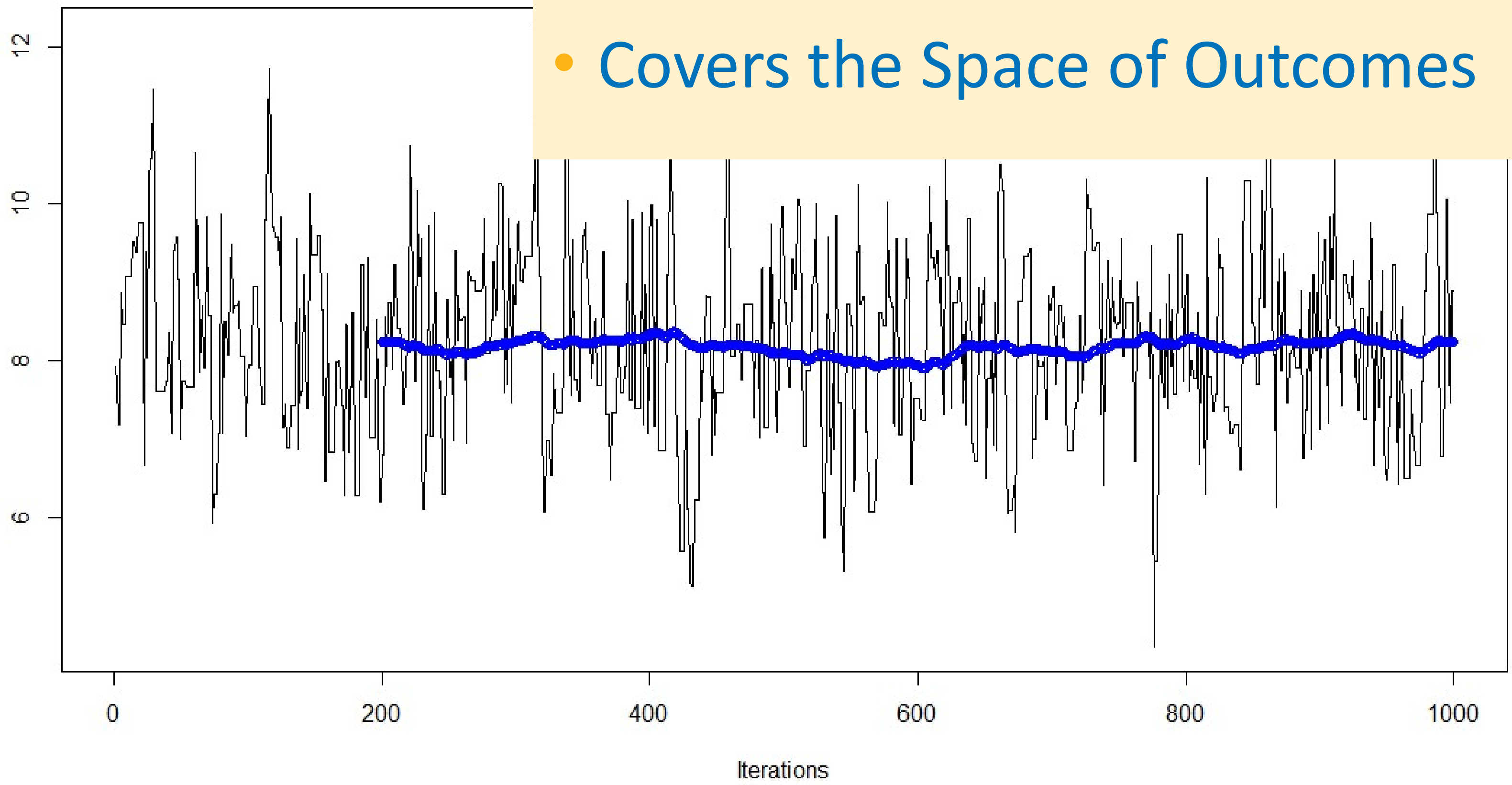
- Want Acceptance Rate between 23% and 50%
- A High Acceptance Rate will result in the Generating Distribution
- To Decrease Acceptance – Increase Variance of Generating Function
- Switch from $Uniform(\theta_{i-1} \pm 1)$
- To $Uniform(\theta_{i-1} \pm 3)$

Metropolis-Hastings

Acceptance Rate: 55%

- 1,000 Iterations; Generating $\sim Unif \pm 3$
- Blue Line is 200 point average

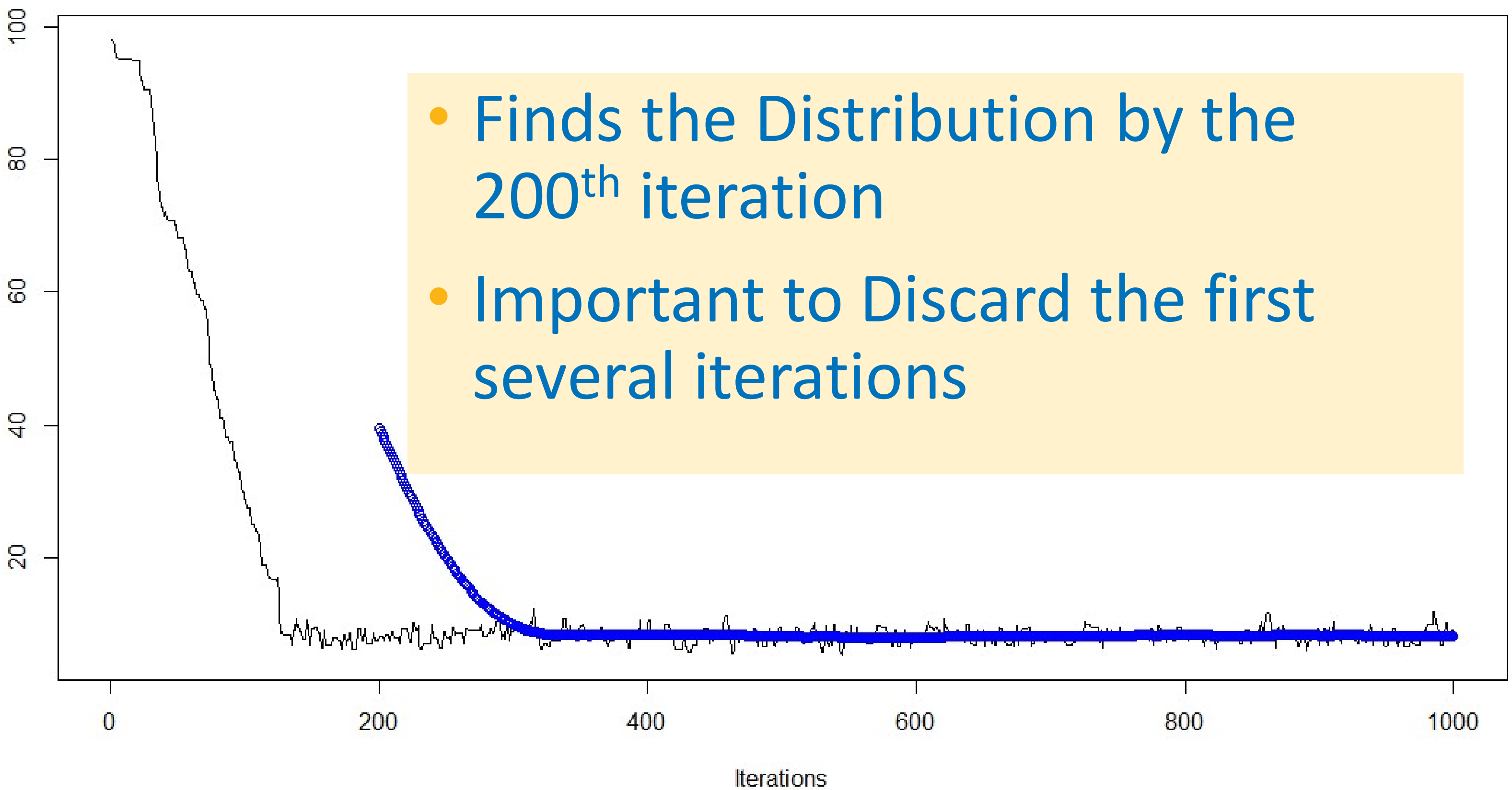
- Stationary
- Covers the Space of Outcomes



Metropolis-Hastings

Acceptance Rate: 53%

- 1,000 Iterations; Generating $\sim Unif \pm 3$
- Choose a ridiculous starting point $\theta_0 = 100$

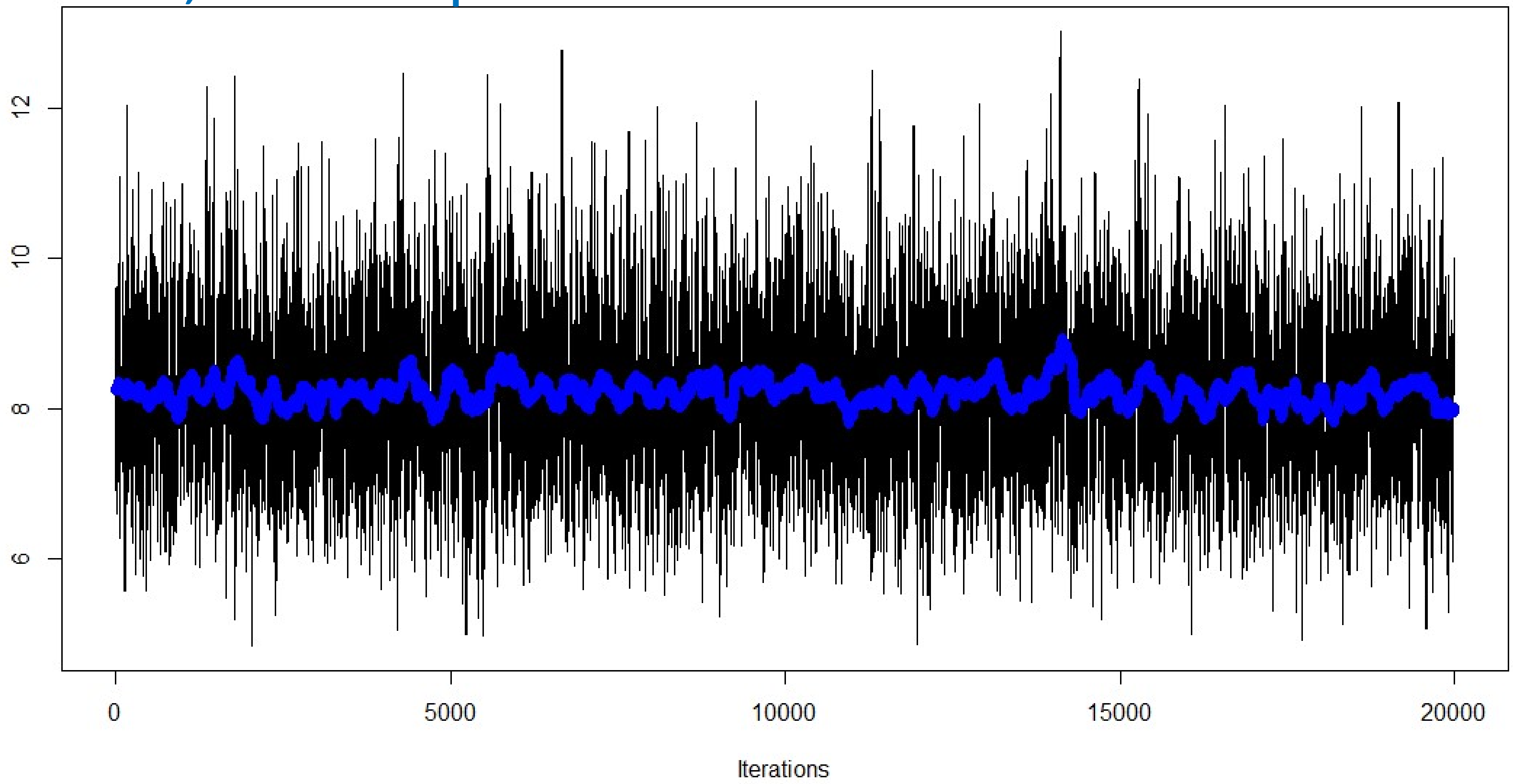


- Finds the Distribution by the 200th iteration
- Important to Discard the first several iterations

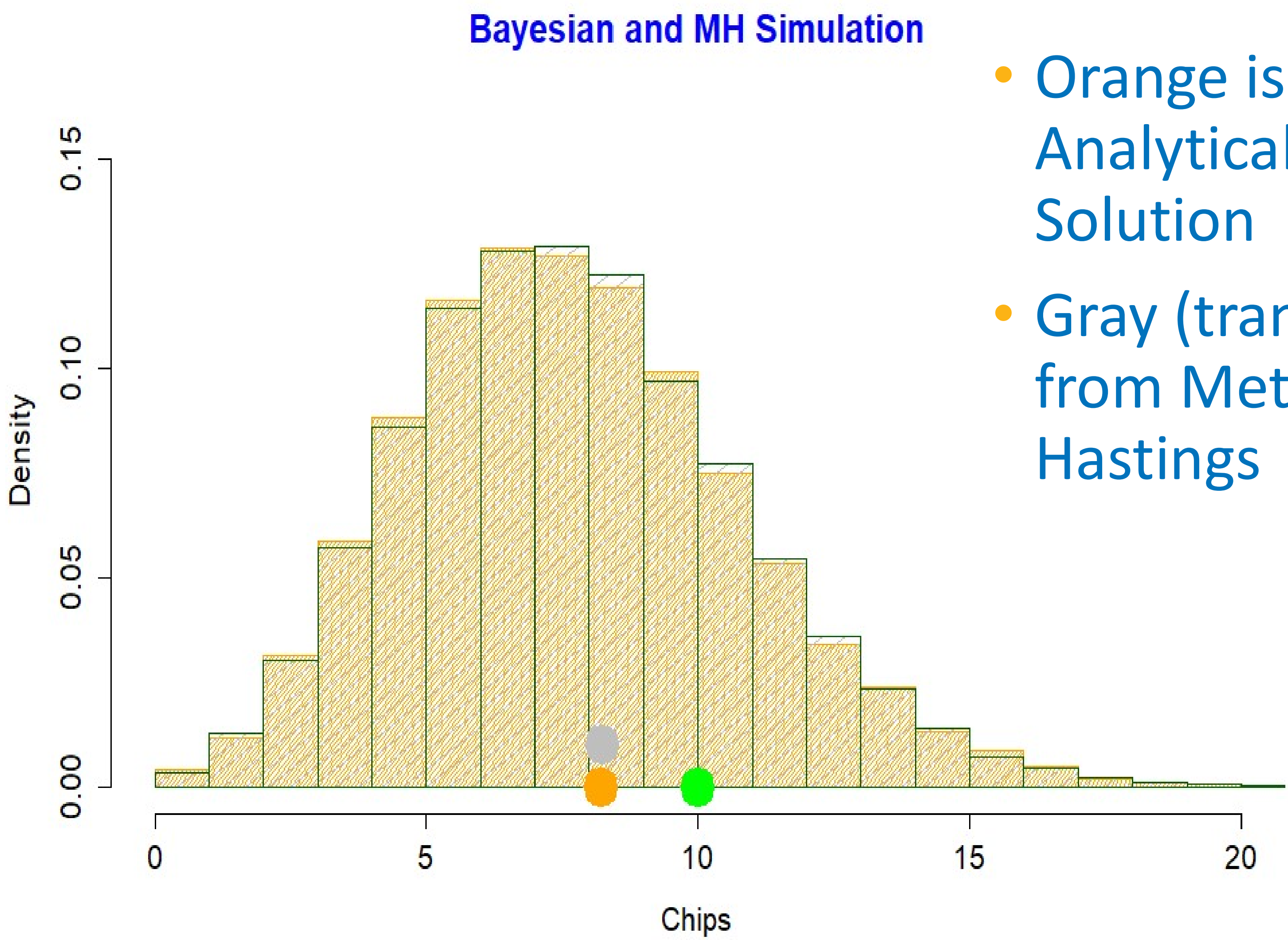
Metropolis-Hastings

Acceptance Rate: 52%

- First 1,000 Iterations Discarded; Generating $\sim Unif \pm 3$
- 20,000 samples



Metropolis-Hastings



- Orange is using the Analytical Bayesian Solution
- Gray (transparent) is from Metropolis Hastings

Metropolis-Hastings

- 1. Select an initial θ_0
- 2. Draw θ^* from $q(\theta^*|\theta_{i-1})$
- 3. Calculate α

$$\alpha = \frac{\boldsymbol{g}(\boldsymbol{\theta}^*)/\boldsymbol{q}(\boldsymbol{\theta}^*|\boldsymbol{\theta}_{\boldsymbol{i-1}})}{\boldsymbol{g}(\boldsymbol{\theta}_{\boldsymbol{i-1}})/\boldsymbol{q}(\boldsymbol{\theta}_{\boldsymbol{i-1}}|\boldsymbol{\theta}^*)}$$

- 4. Draw $u \in Unif(0,1)$
- 5. If $u < \alpha$ then $\theta_i = \theta^*$; otherwise $\theta_i = \theta_{i-1}$

Repeat Steps 2-5 many times

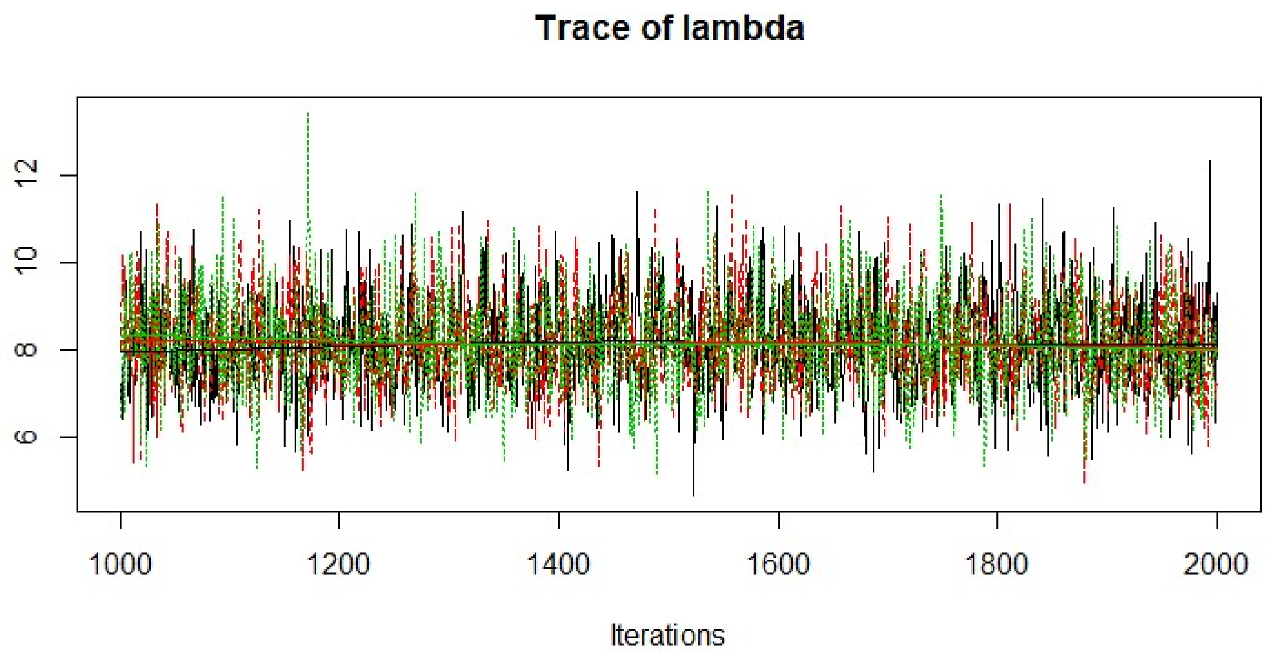
Diagnostics

- Gelman Diagnostic
- Auto Correlation
- Effective Size

Diagnostics

- Gelman Diagnostic
- Run multiple Markov Chains
- Compare the variance within each chain to the variance in other chains

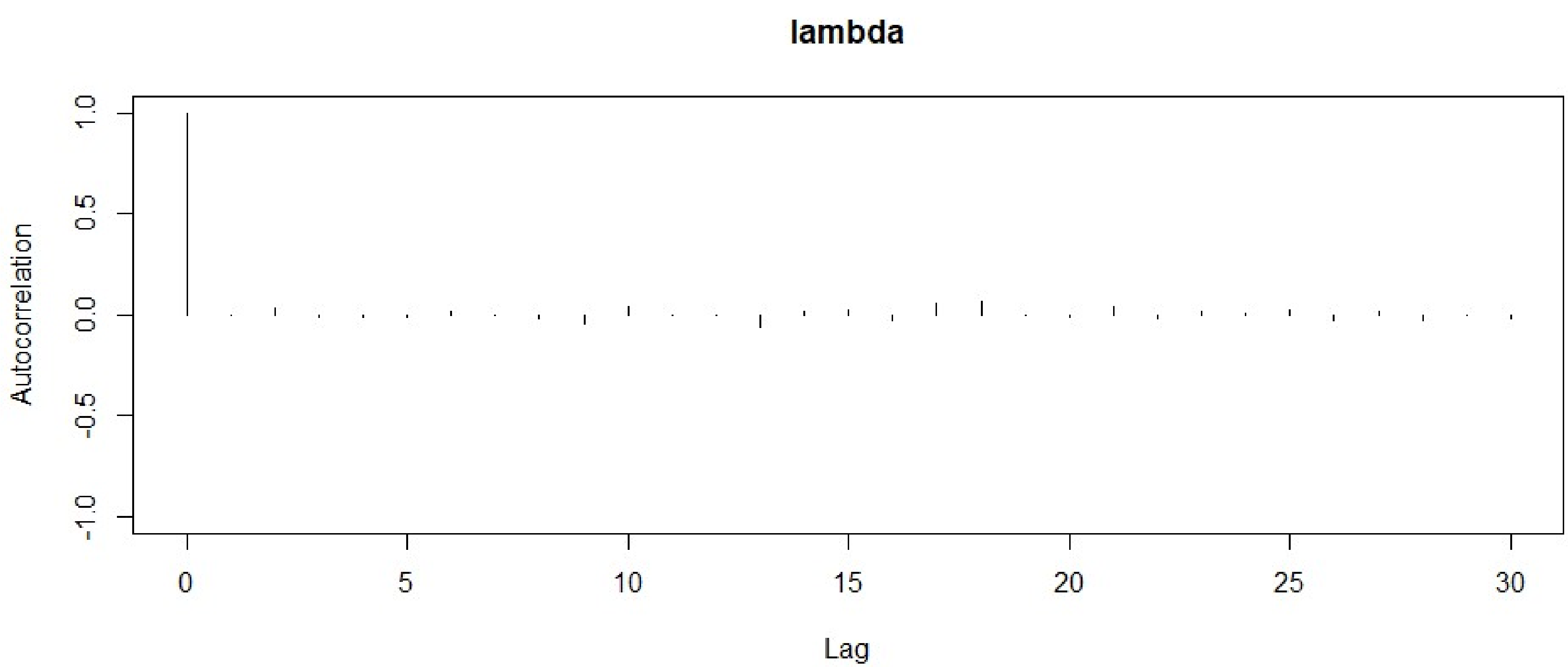
Diagnostics – Gelman



- 3 Chains each of length 1,000
- Gelman Diagnostic is 1.000

Diagnostics - Autocorrelation

- High Autocorrelation, reduces the information in the Markov Chain
- This dataset has low Autocorrelation



Diagnostics - Autocorrelation

- High Autocorrelation, reduces the information in the Markov Chain
- Data can be “thinned”
- Eg. Only use every 20 points from the Markov Chain if the Autocorrelation is small at lag 20 and beyond

Diagnostics – Effective Size

- Effective Size
- If there is Autocorrelation in the Markov Chain, it does not have the same amount of information as the same number of points drawn from the target distribution
- Effective Size tells you how much information is in your markov chain in terms of if it was actual pulls from the target distribution

Gibbs Sampling

- $g(\alpha, \beta\ ) \propto f(\alpha, \beta|y)$
- Calculate
- $g(\alpha|\beta) \propto f(\alpha|\beta, y)$
- Treat β as a constant; and drop constant terms
- $g(\beta|\alpha)$
- When sampling, do the MH algorithm, assuming $\beta$is known; and draw a sample from α
- Then use, this α as a constant, and use MH to draw a sample from β
- Repeat

Summary

- Bayesian Analysis – allows for a **prior** opinion on parameters
- Create a $g(\tilde{\theta}) \propto f(\tilde{\theta}|\tilde{y})$, **posterior** distribution
- Use Metropolis-Hastings to **sample** from the posterior distribution
- Use Gibbs Sampling if you have more than one parameter
- Use Diagnostics to test convergence
- You have a sample of the posterior of **all** the parameters
- Simulate Data, using draws from the sampled parameters

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, Virginia 22203

www.casact.org