NYU

Introduction to Data Science

Center for Data Science
Iddo Drori, Spring 2019

NYU

Bayesian Inference

Bayes Rule

- Point distributions

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

$$P(B|A) = P(A|B)P(B) / P(A)$$

Bayes Rule

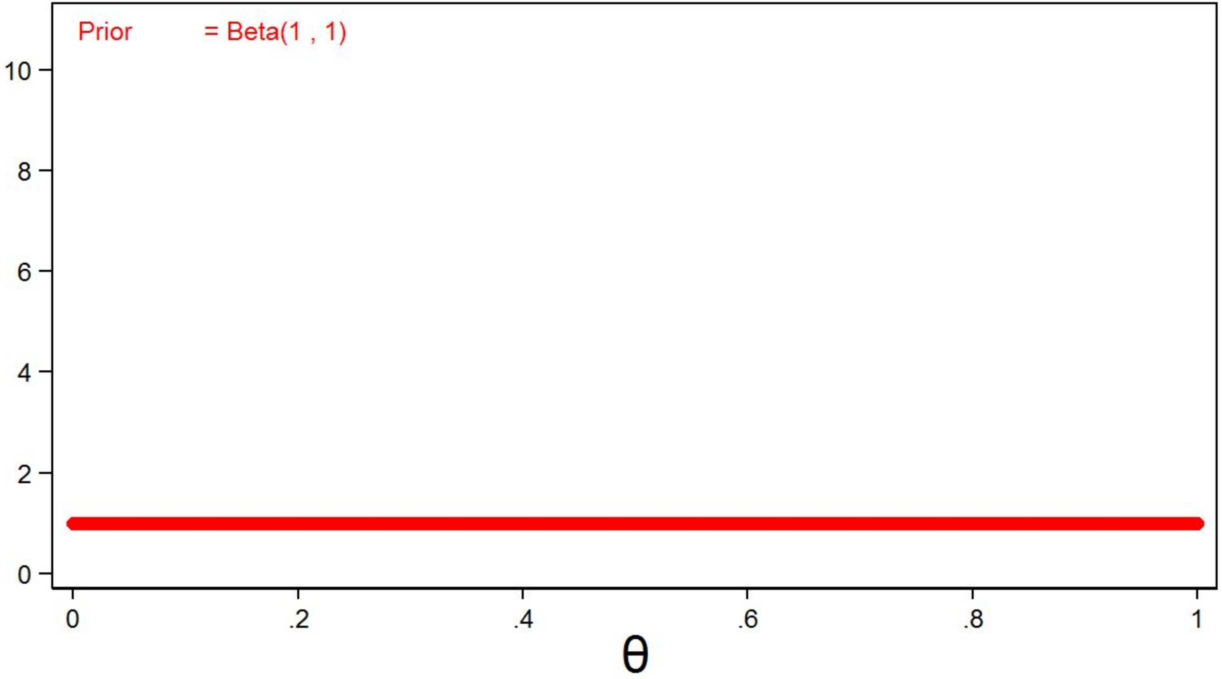
- Probability distributions

Frequency

- Flip coin multiple times
- Count number of time coin lands heads out of total flips

Uninformative Beta Prior

Uniform [0,1]



Source: stata.com

Prior Knowledge

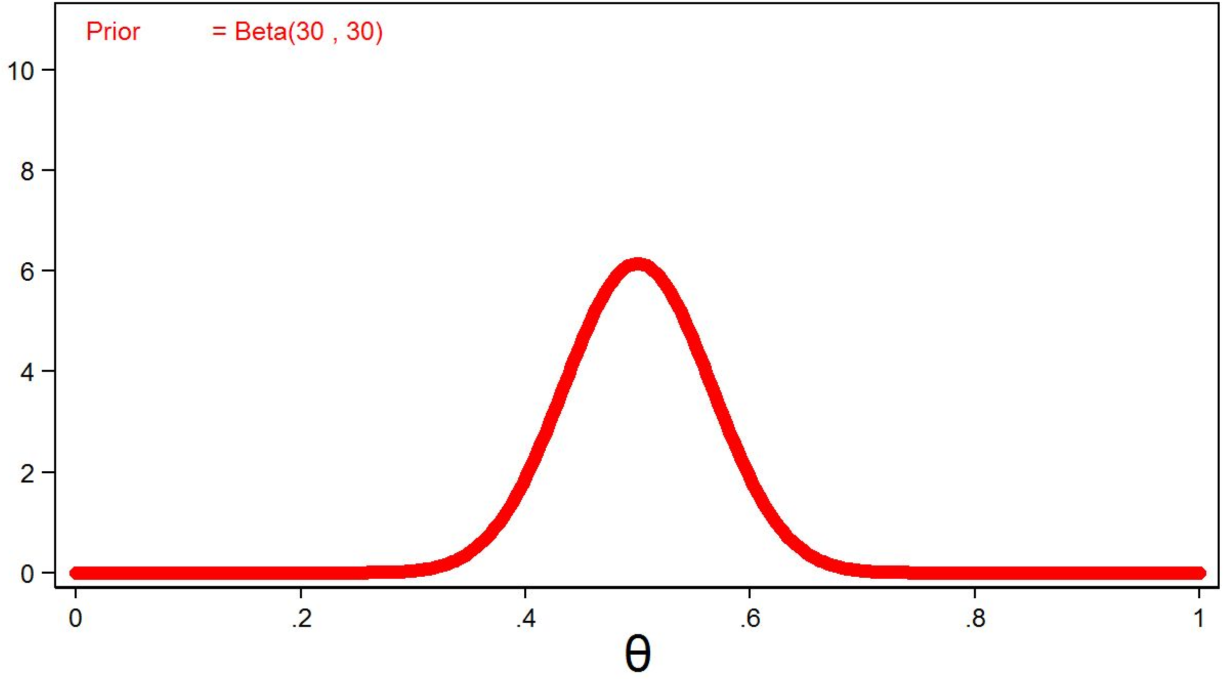
- Prior knowledge that coin is nearly fair
  Prior represents expert knowledge

- Rather than estimate single value for parameter, obtain distribution over possible values

- Use Bayes rule to update posterior after observing data given prior.

Informative Beta Prior Distribution

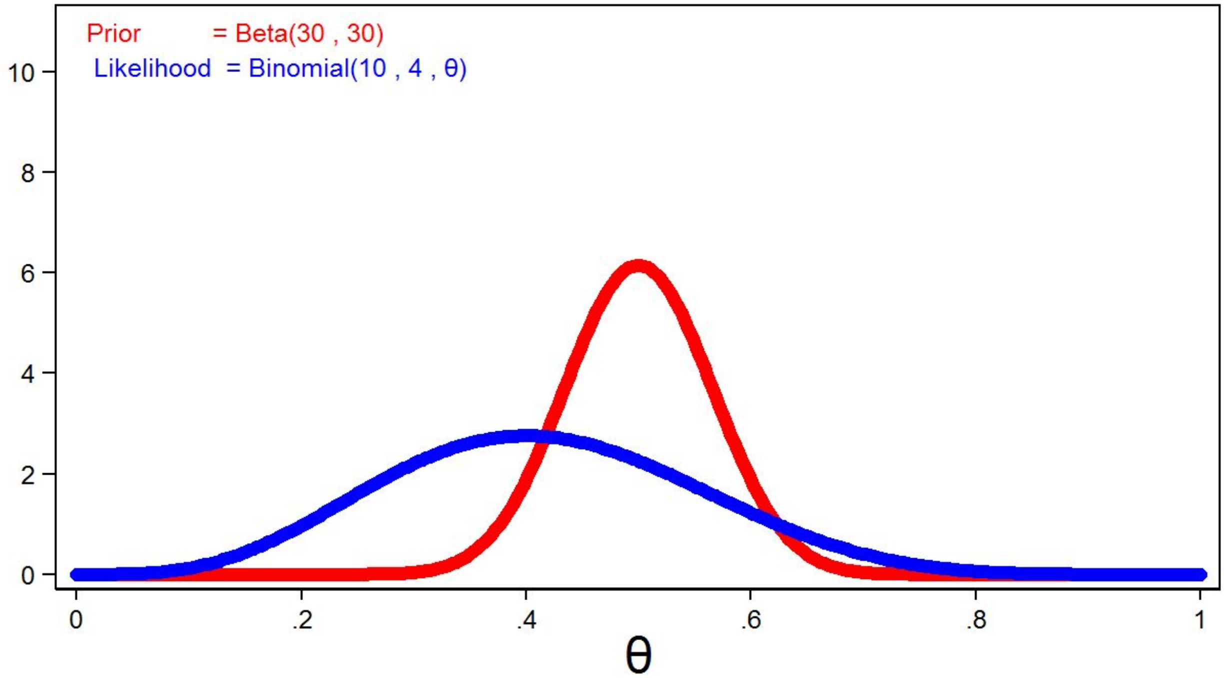
$p(\theta)$



Source: stata.com

Binomial Likelihood and Beta Prior

$p(\theta)$ $p(y|\theta)$

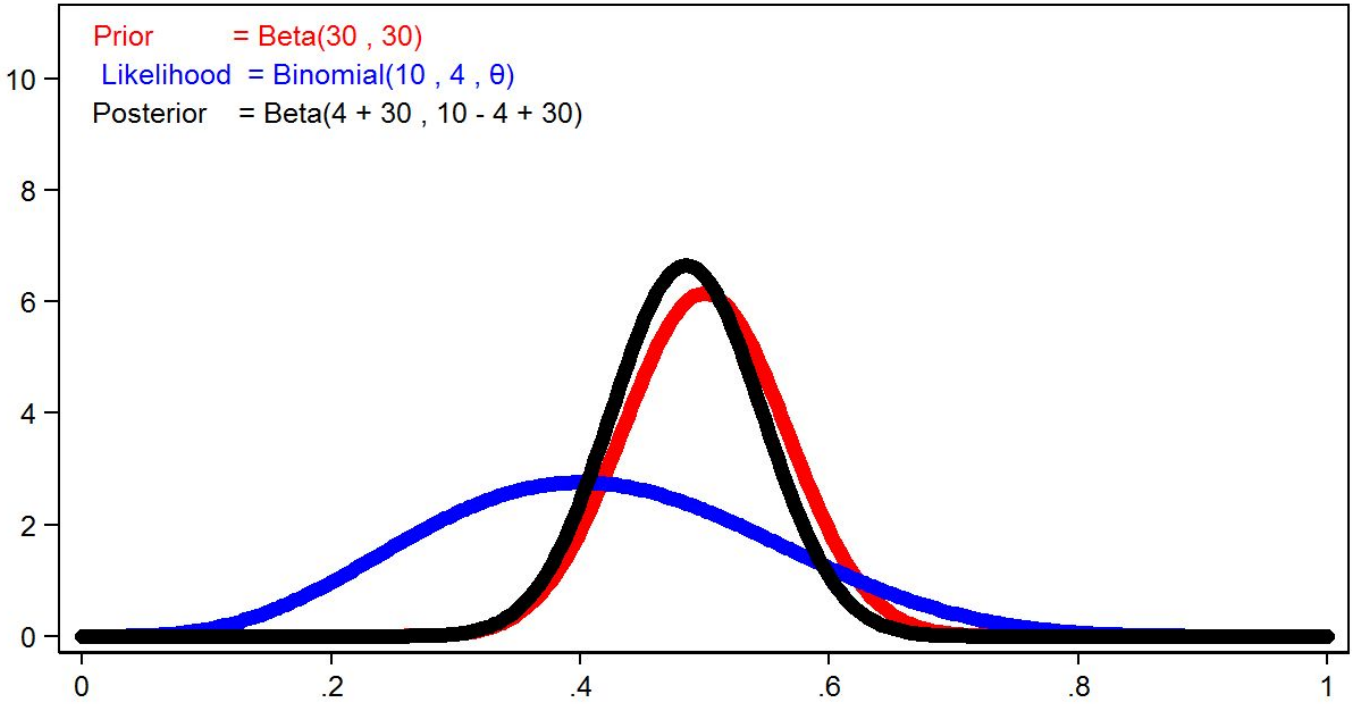


Observe 4 heads out of 10 coin flips: Binomial likelihood function

Source: stata.com

Update Belief based on Experiment Results

$$Posterior \propto Prior \times Likelihood$$

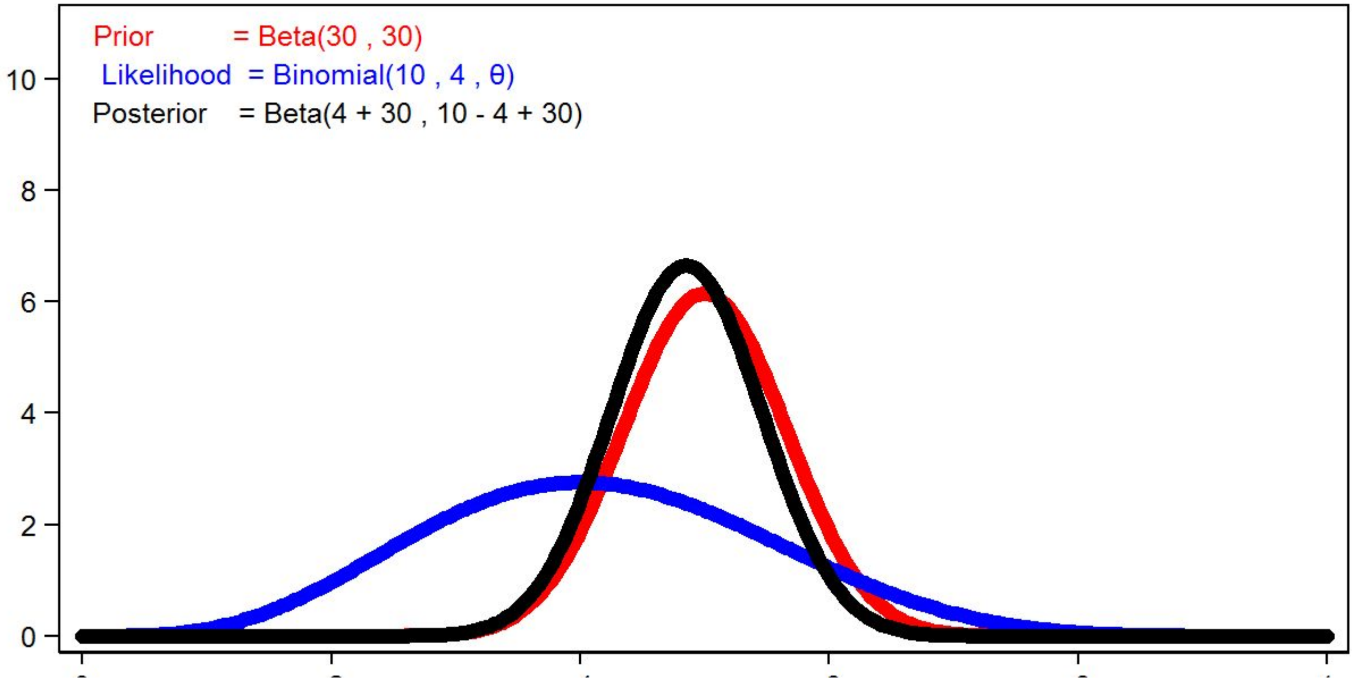


$$p(\theta|y) \propto p(\theta)p(y|\theta)$$

Source: stata.com

Update Belief based on Experiment Result

$$p(\theta|y) = Beta(\alpha, \beta) x Binomial(n, \theta) = Beta(y + \alpha, n - y + \beta)$$



Beta distribution is a conjugate prior for binomial likelihood function since posterior distribution belongs to same family as prior distribution

Source: stata.com

Conjugate Priors

- Closed form representation of posterior
- Posterior and prior have same algebraic form as function of parameter

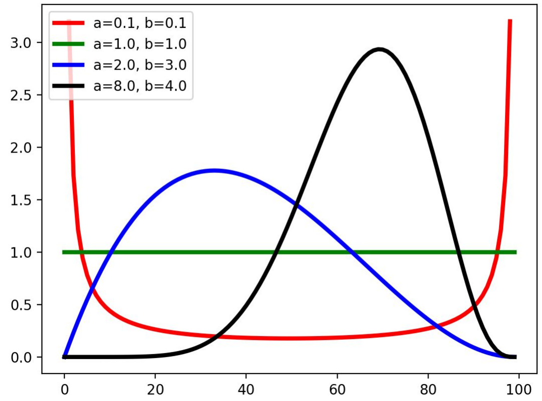
Beta Distribution

Defined over [0,1].
Conjugate prior for Bernoulli, binomial, geometric distributions.

$$Beta(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

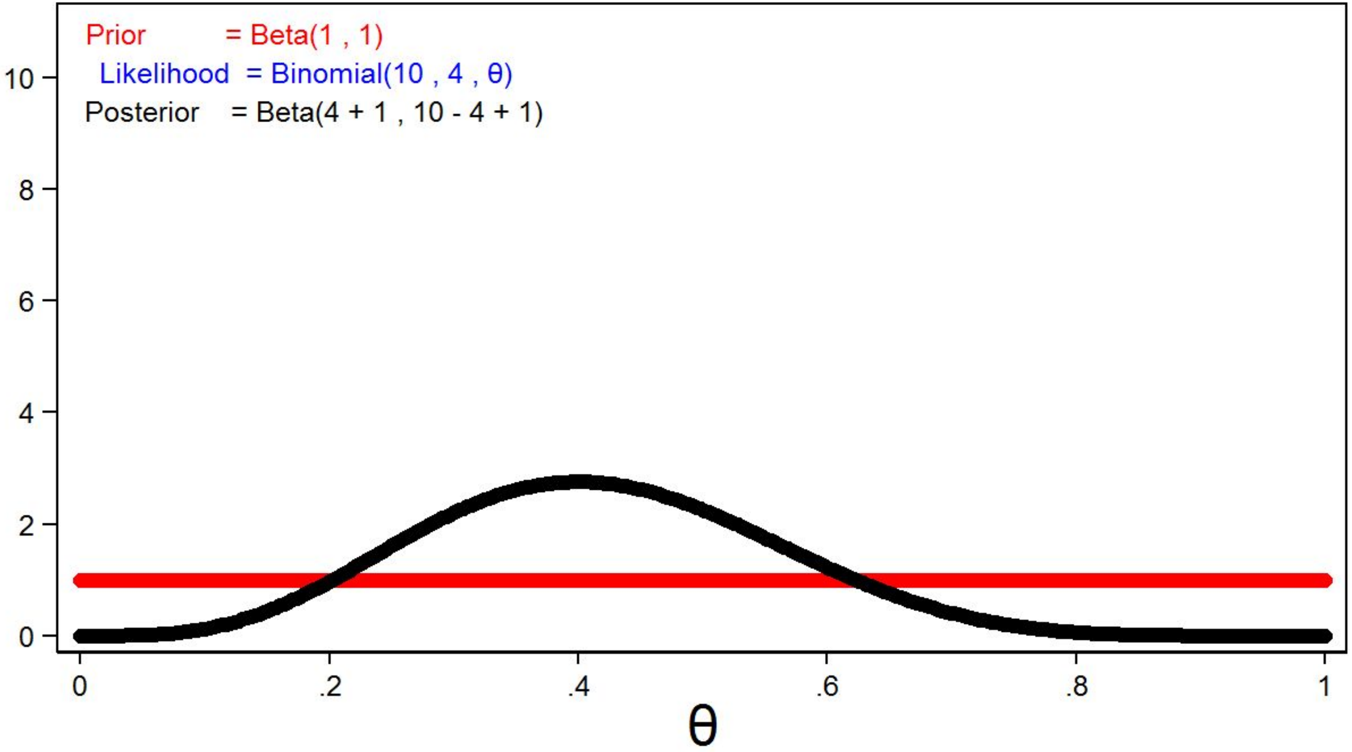
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import beta

x = np.linspace(0, 1, 100)
aa = [0.1, 1., 2., 8.]
bb = [0.1, 1., 3., 4.]
props = ['r-', 'g-', 'b-', 'k-']
for a, b, p in zip(aa, bb, props):
    y = beta.pdf(x, a, b)
    pl.plot(y, p, lw=3, label='a=%.1f, b=%.1f' % (a, b))
plt.legend(loc='upper left')
plt.show()
```



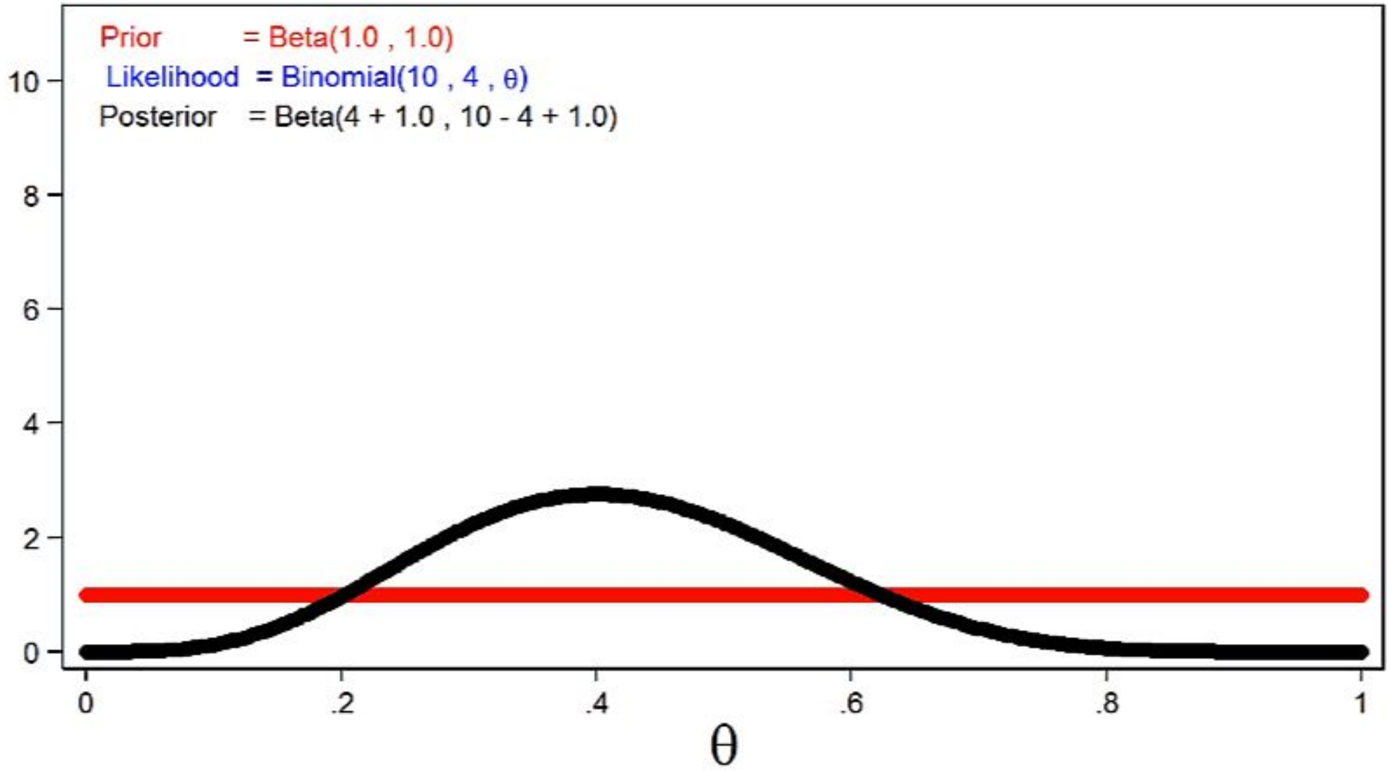
Posterior for Beta(1,1) Prior

$$p(\theta|y) = Beta(\alpha, \beta) x Binomial(n, \theta) = Beta(y + \alpha, n - y + \beta)$$



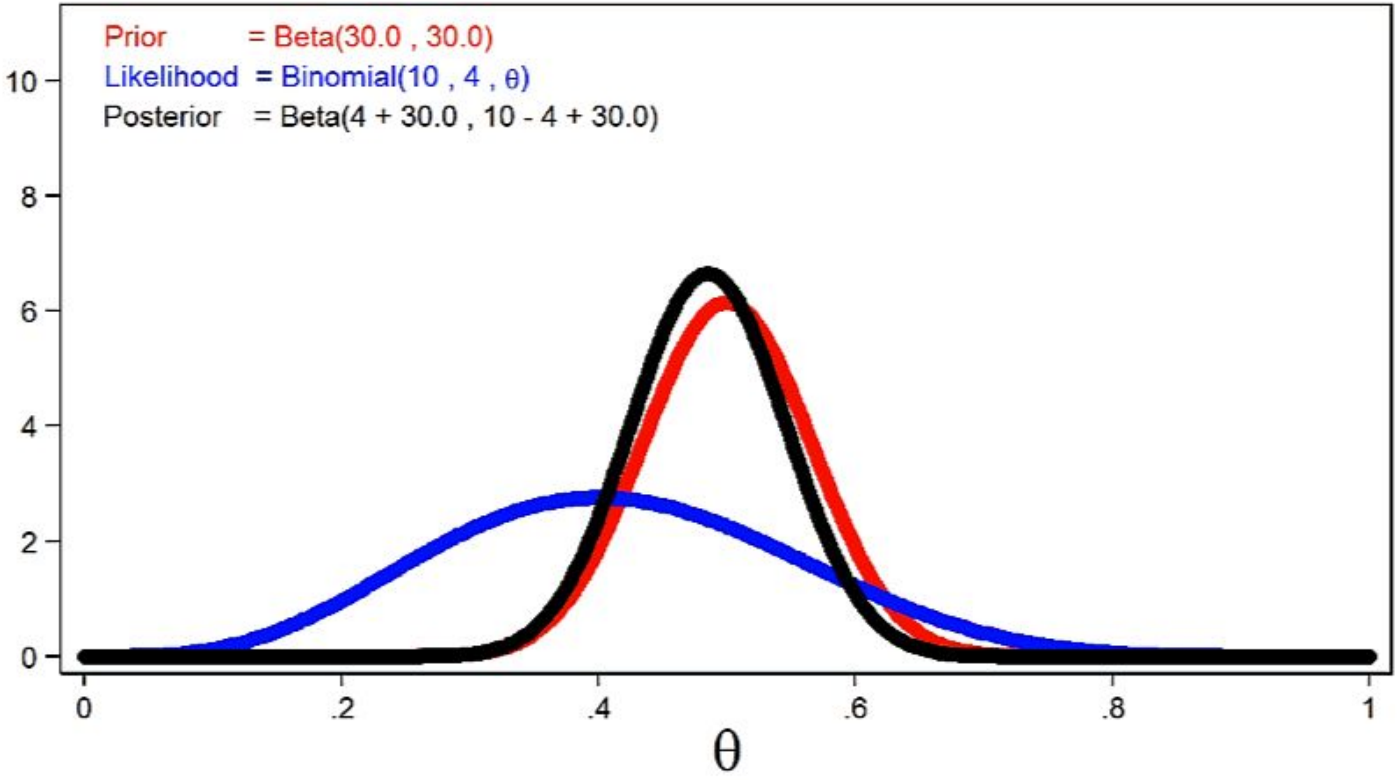
Source: stata.com

Effect of Informative Prior Distributions on Posterior Distributions



Source: stata.com

Effect of Larger Sample Sizes on Posterior Distribution



Source: stata.com

Posterior Distribution

- Bayesian approach
- Prior information encoded as distribution over possible parameter values
- Use Bayes rule to update posterior based on observations

Maximum A Posteriori (MAP) Estimation

- From distribution to single parameter value
- Choose value which is most probable given observed data and prior belief

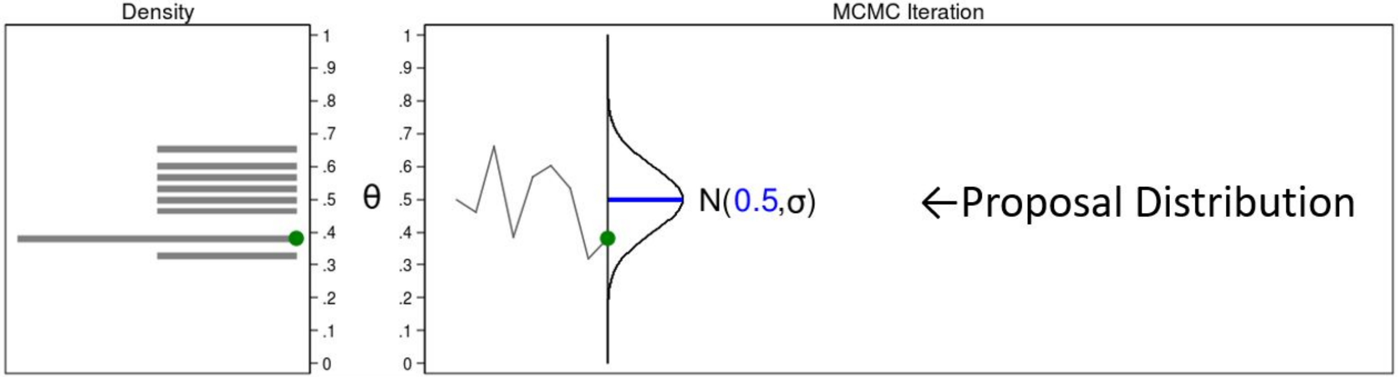
$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(\theta / D)$$
$$= \underset{\theta}{\operatorname{argmax}}\, P(D/\theta)P(\theta)$$

Markov Chain Monte Carlo (MCMC)

Motivation

- Goal: estimate posterior distribution of parameter θ, which is probability that coin flip results in heads.
- Prior distribution is uninformative Beta distribution with parameters (1,1).
- Use binomial likelihood function to quantify data: 4 heads / 10 coin flips.
- Use MCMC with M-H algorithm to generate sample from posterior distribution of θ. Use sample to estimate mean and confidence intervals of posterior distribution

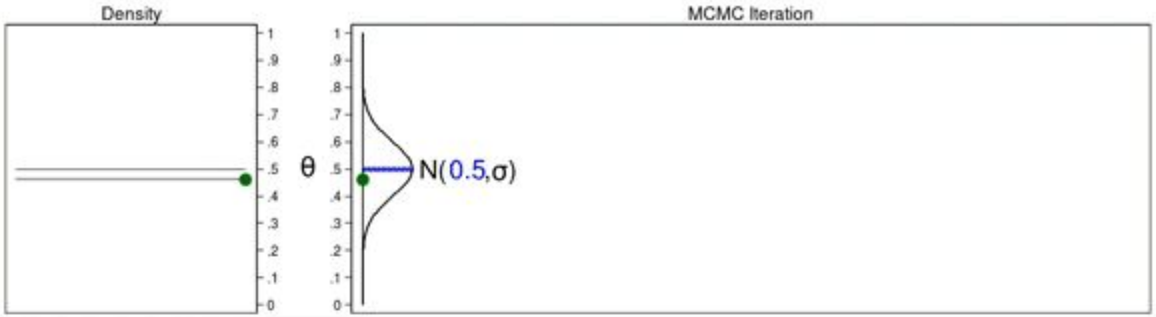
Proposal Distribution, Trace Plot, Density Plot



Draw $\theta_t \sim \text{Normal}(0.5, \sigma) = 0.381$

Source: stata.com

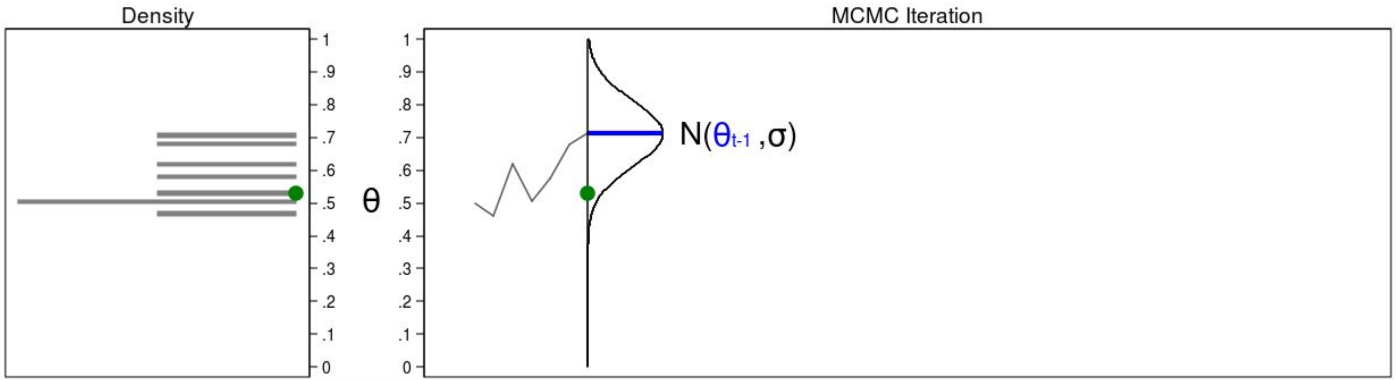
Monte Carlo Trace Plot



Draw $\theta_t \sim \text{Normal}(0.5,\sigma) = 0.460$

Source: stata.com

Markov Chain Monte Carlo Trace Plot

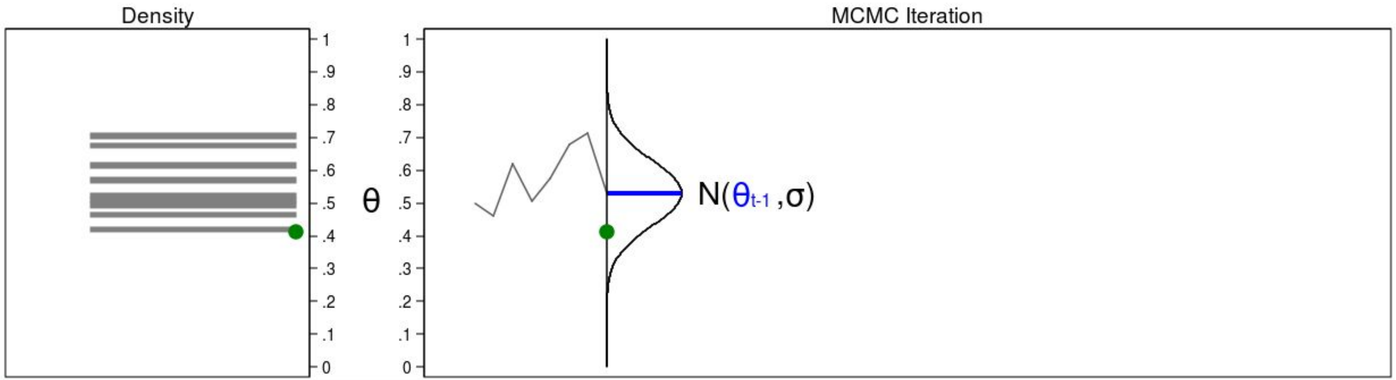


Draw $\theta_t \sim \text{Normal}(\theta_{t-1},\sigma)$

$\text{Normal}(0.712,\sigma) = 0.530$

Source: stata.com

Markov Chain Monte Carlo Trace Plot

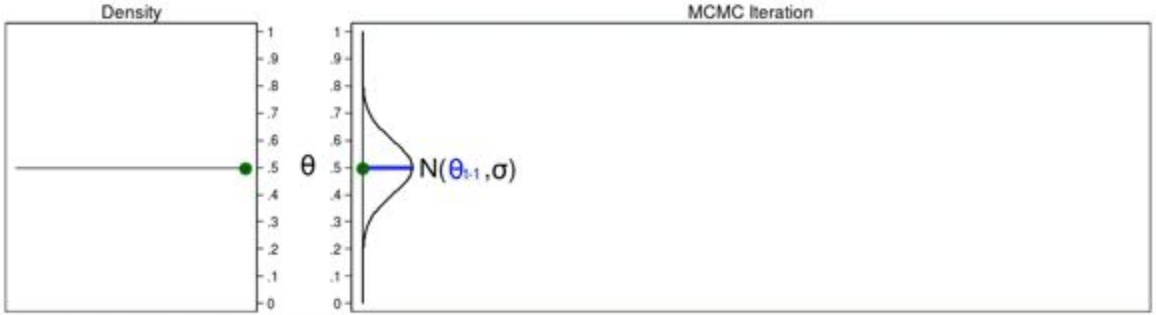


$$\text{Draw } \theta_t \sim \text{Normal}(\theta_{t-1}, \sigma)$$

$$\text{Normal}(0.530, \sigma) = 0.411$$

Source: stata.com

Markov Chain Monte Carlo Trace Plot



Draw $\theta_t \sim \text{Normal}(\theta_{t-1}, \sigma)$

$\text{Normal}(0.500, \sigma) = 0.497$

Source: stata.com

Markov Chain Monte Carlo

- Proposal distribution is changing with each iteration.
- Trace plot with random walk pattern, variability is not same over all iterations.

- Problem: resulting density plot does not look like proposal distribution, or a posterior distribution.
- Solution: improve sample keeping proposed values of θ more likely under posterior distribution and discarding less likely values.

- Problem: difficult to accept or reject proposed values of θ based on posterior distribution since we don't know functional form of posterior distribution

MCMC with Metropolis Hastings

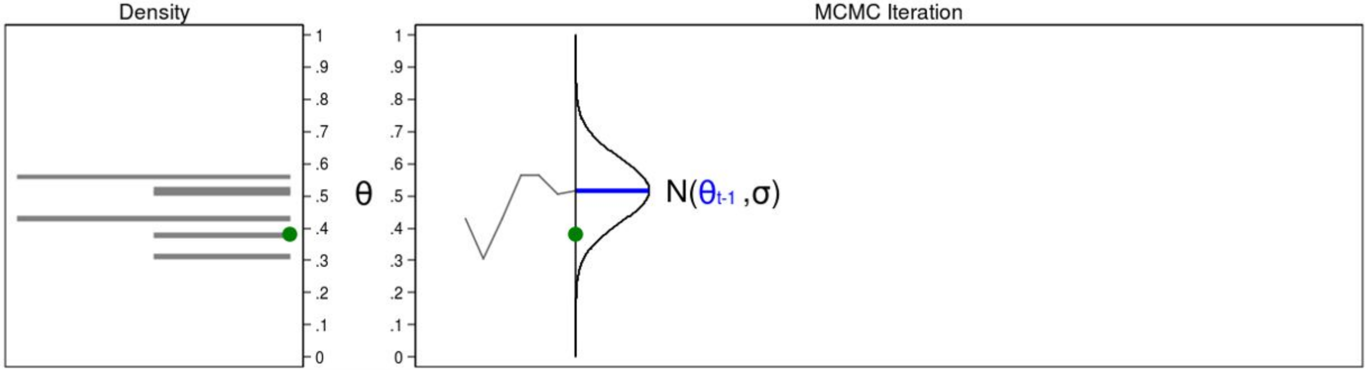
Metropolis Hastings Algorithm

- Decide which proposed values of θ to accept or reject even when we don't know the functional form of posterior distribution
- Compute ratio: $r(\theta_{new}, \theta_{t-1}) = \dfrac{Posterior(\theta_{new})}{Posterior(\theta_{t-1})}$
- Compute accept probability in [0,1]:

$$\alpha\ (\theta_{new}, \theta_{t-1}) = min(r(\theta_{new}, \theta_{t-1}), 1)$$

- Draw u~Uniform [0,1]: if u < α then accept new value

MCMC-MH

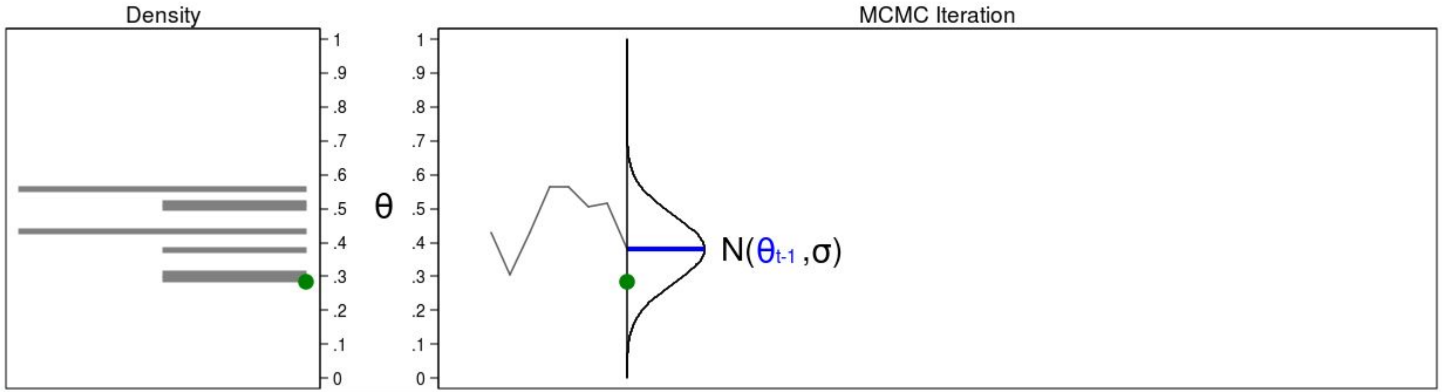


Step 1: $r(\theta_{new}, \theta_{t-1}) = \dfrac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \dfrac{\text{Beta}(1,1,0.380) \times \text{Binomial}(10,4,0.380)}{\text{Beta}(1,1,0.517) \times \text{Binomial}(10,4,0.517)} = 1.307$

Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{1.307, 1\} = 1.000$

Source: stata.com

MCMC-MH



Step 1: $r(\theta_{new}, \theta_{t-1}) = \dfrac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \dfrac{\text{Beta}(1,1,0.286) \times \text{Binomial}(10,4,\ 0.286)}{\text{Beta}(1,1,0.380) \times \text{Binomial}(10,4,\ 0.380)} = 0.747$

Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.747, 1\} = 0.747$

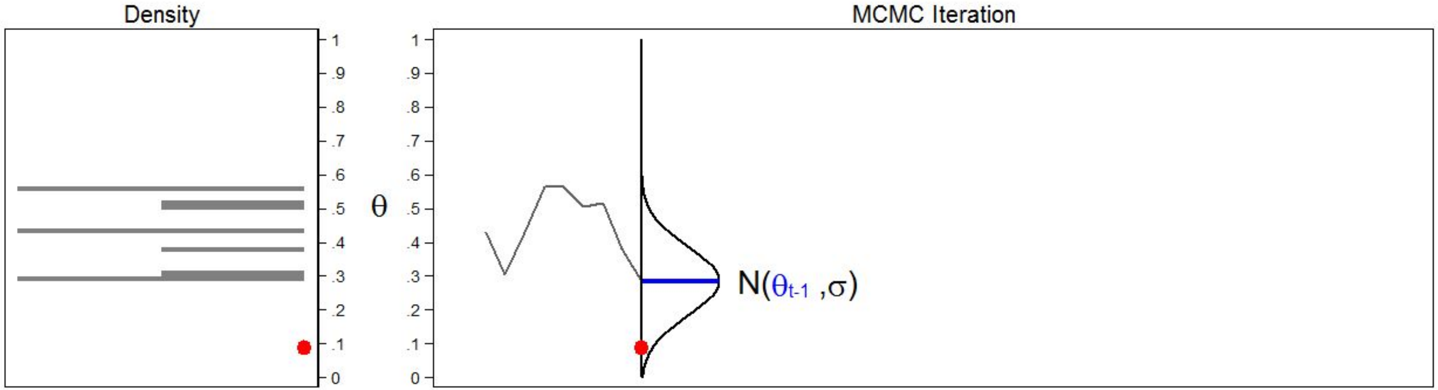
Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.094$

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1}) \rightarrow$ If $0.094 < 0.747$ Then $\theta_t = \theta_{new} = 0.286$

Otherwise $\theta_t = \theta_{t-1} = 0.380$

Source: stata.com

MCMC-MH



Step 1: $r(\theta_{new}, \theta_{t-1}) = \dfrac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \dfrac{\text{Beta}(1,1,0.088) \times \text{Binomial}(10,4,0.088)}{\text{Beta}(1,1,0.286) \times \text{Binomial}(10,4,0.286)} = 0.039$

Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.039, 1\} = 0.039$

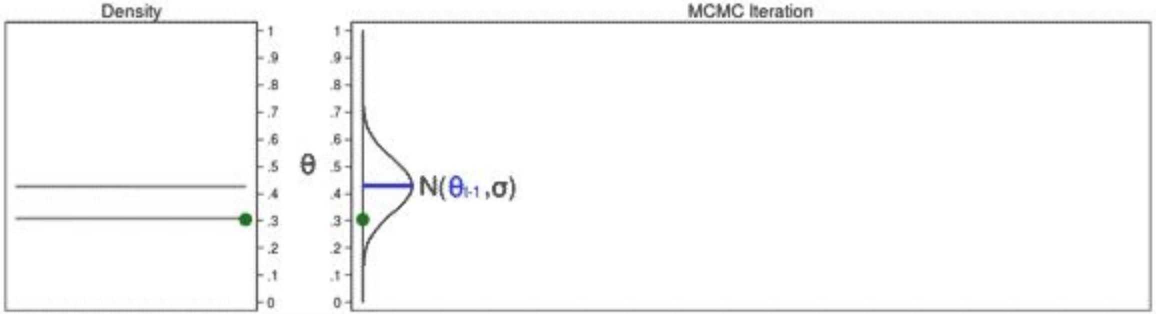
Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.247$

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1}) \rightarrow$ If $0.247 < 0.039$ Then $\theta_t = \theta_{new} = 0.088$

Otherwise $\theta_t = \theta_{t-1} = 0.286$

Source: stata.com

MCMC-MH



Step 1: $r(\theta_{new}, \theta_{t-1}) = \dfrac{\text{Posterior}(\theta_{new})}{\text{Posterior}(\theta_{t-1})} = \dfrac{\text{Beta}(1,1,0.306) \times \text{Binomial}(10,4,0.306)}{\text{Beta}(1,1,0.429) \times \text{Binomial}(10,4,0.429)} = 0.834$

Step 2: Acceptance probability $\alpha(\theta_{new}, \theta_{t-1}) = \min\{r(\theta_{new}, \theta_{t-1}), 1\} = \min\{0.834, 1\} = 0.834$

Step 3: Draw $u \sim \text{Uniform}(0,1) = 0.617$

Step 4: If $u < \alpha(\theta_{new}, \theta_{t-1}) \rightarrow$ If $0.617 < 0.834$ Then $\theta_t = \theta_{new} = 0.306$

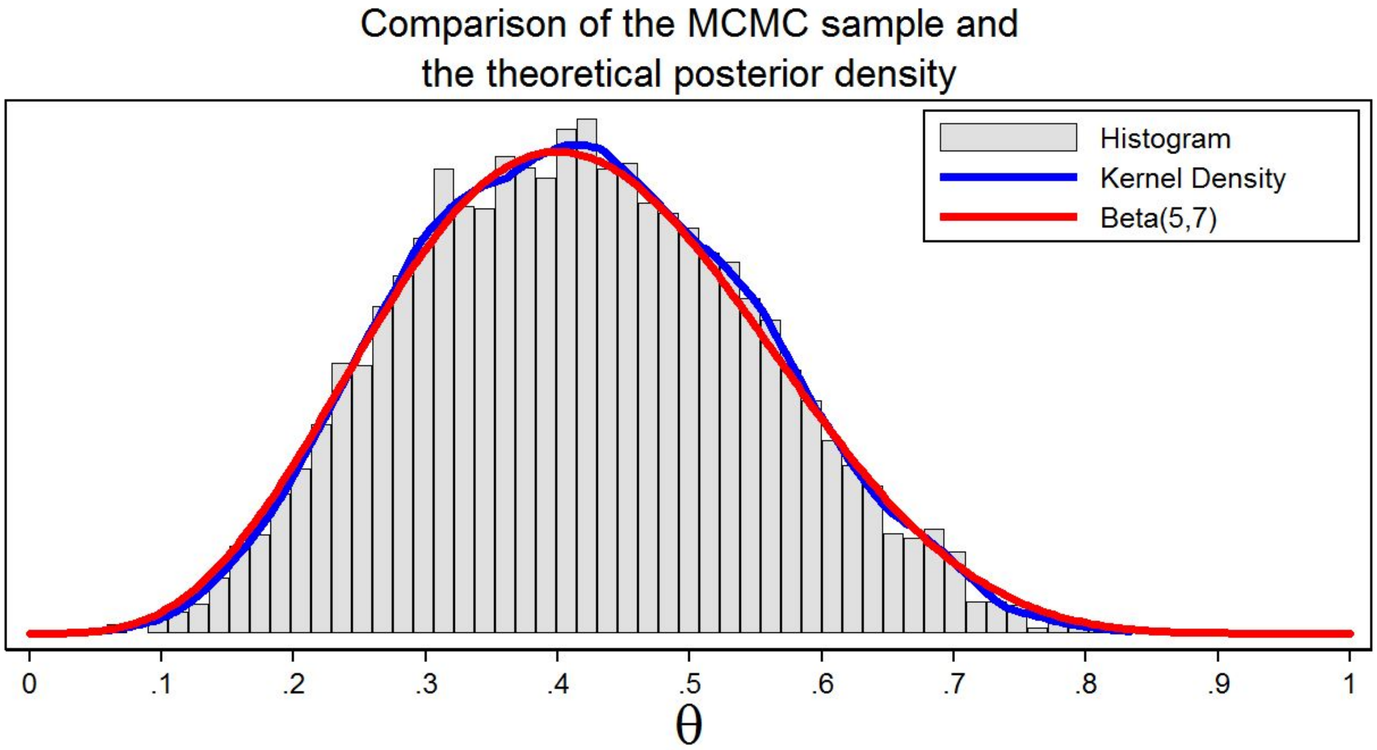
Otherwise $\theta_t = \theta_{t-1} = 0.429$

Source: stata.com

MCMC-MH

- Proposal distribution changes with most iterations
- Trace plot does not exhibit random walk pattern observed using MCMC
- Density is useful distribution
- Use sample to estimate mean or median of posterior distribution, 95% credible interval, probability that θ falls within arbitrary interval

Sample Posterior Distribution with MCMC-MH

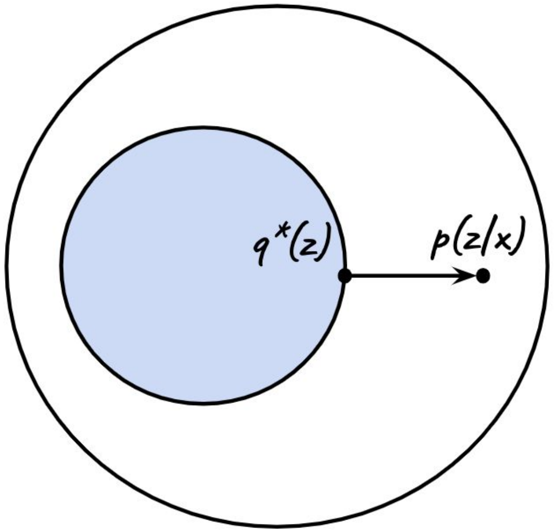


Source: stata.com

Variational Bayes

Optimization

- Approximate posterior



$$q^*(z) = \underset{q}{\operatorname{argmin}}\, KL\big(q(z)||p(z|x)\big)$$

KL Divergence

- KL divergence (asymmetric)

$$KL\big(q(z)||p(z|x)\big) = \int q(z)\log\frac{q(z)}{p(z|x)}dz = \int q(z)\log\frac{q(z)p(x)}{p(z,x)}dz = \log p(x) - \int q(z)\log\frac{p(z,x)}{q(z)}dz$$

$$p(z,x) = p(z|x)p(x) \qquad \log\frac{1}{x} = -\log x$$

Evidence Lower Bound (ELBO)

- KL is non-negative

$$KL\big(q(z)||p(z|x)\big) = \log p(x) - \int q(z) \log \frac{p(z,x)}{q(z)} dz$$

$$\log p(x) \geq \int q(z) \log \frac{p(z,x)}{q(z)} dz$$

evidence lower bound (ELBO)