

# Posterior simulation

- So far we have been able to get posterior inference “directly” from the known distribution
- Sometime the distribution is not in known form
- Sometimes it is not convenient to sample from the known form, even if we have it
  - E.g., posterior prediction
  - E.g., multi-parameter models → conditioning
- Simulation is just a way to get random samples from a desired distribution

# Basic methods

- Direct methods

- Inverse CDF

- Requires that  $f(x)$  is normalized

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

- Have (or can compute)  $F^{-1}(x)$

- Rejection

$$F(x) = \Pr(X \leq x) = \int_{-\infty}^x f(x)dx$$

- Typically use for discrete distributions

- “Indirect”

- Rejection sampling

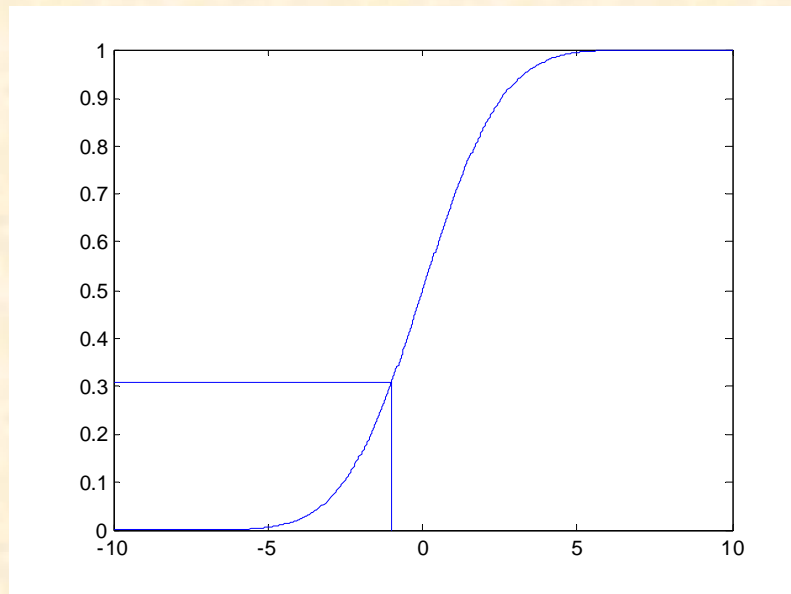
- MCMC methods

# Posterior simulation

- Simulations can be used to describe behavior of random variables
  - Use for sketching posterior distributions
- Useful software allows draws to be made from common distributions
- Helpful to know methods for sampling random variables

# Inverse CDF

- Also called rather grandly *probability integral transform approach*
- If you know the cdf then generate a  $U(0,1)$  and invert the CDF



# Generating an Exponential

- $f_Y(y) = \lambda e^{-\lambda y} \quad 0 < \lambda, 0 < y$

$$F_Y(y) = \int_0^{\infty} \lambda e^{-\lambda y} dy = 1 - e^{-\lambda y}$$

- Generate  $u \sim U(0, 1) \Leftrightarrow (1-u) \sim U(0,1)$

$$y = \frac{-\ln(u)}{\lambda}$$

# Exercise 5

- Simulate 1000 values from an exponential with parameter  $\lambda = 0.56$ 
  - Note: ln in Excel is ln() not log()
- Compute the mean, median, and 95% credibility interval. Plot the distribution (scatter plot)

# Standard Distributions

- Most computer packages will draw uniform, normal, Poisson, beta, gamma deviates
  - Textbooks and internet good sources for algorithms
- Binomial
  - Draw  $u \sim U(0,1)$ : Assign  $y = \begin{cases} 0 & u > p \\ 1 & u < p \end{cases}$   
then  $y \sim B(1, p)$
  - Sum of  $n$  independent  $B(1, p)$  draws is  $B(n, p)$
  - [binomial\\_sim2.xls](#)

- For a  $k$ -cell Multinomial( $n, \{p_i\}$ ) compute  $c_i = \sum_{h=1}^i p_h$  the cumulative probabilities. Construct the partition:

$(0, c_1)$	1
$(c_1, c_2)$	2
$\vdots$	$\vdots$
$(c_{k-1}, 1)$	$k$

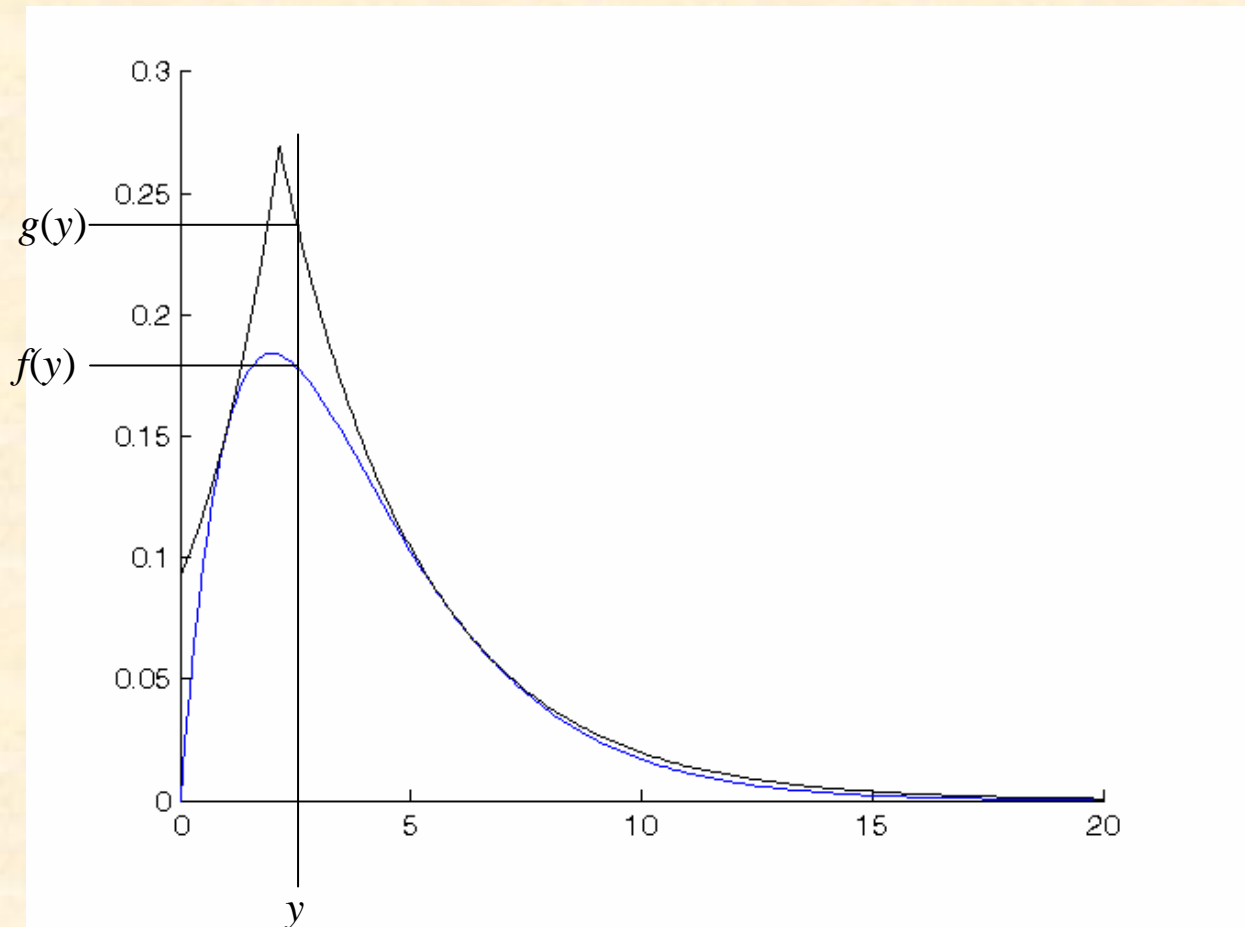
- Draw  $u \sim U(0, 1)$   $n$  times and tally the number of times  $u$  falls in each segment

# Rejection Sampling

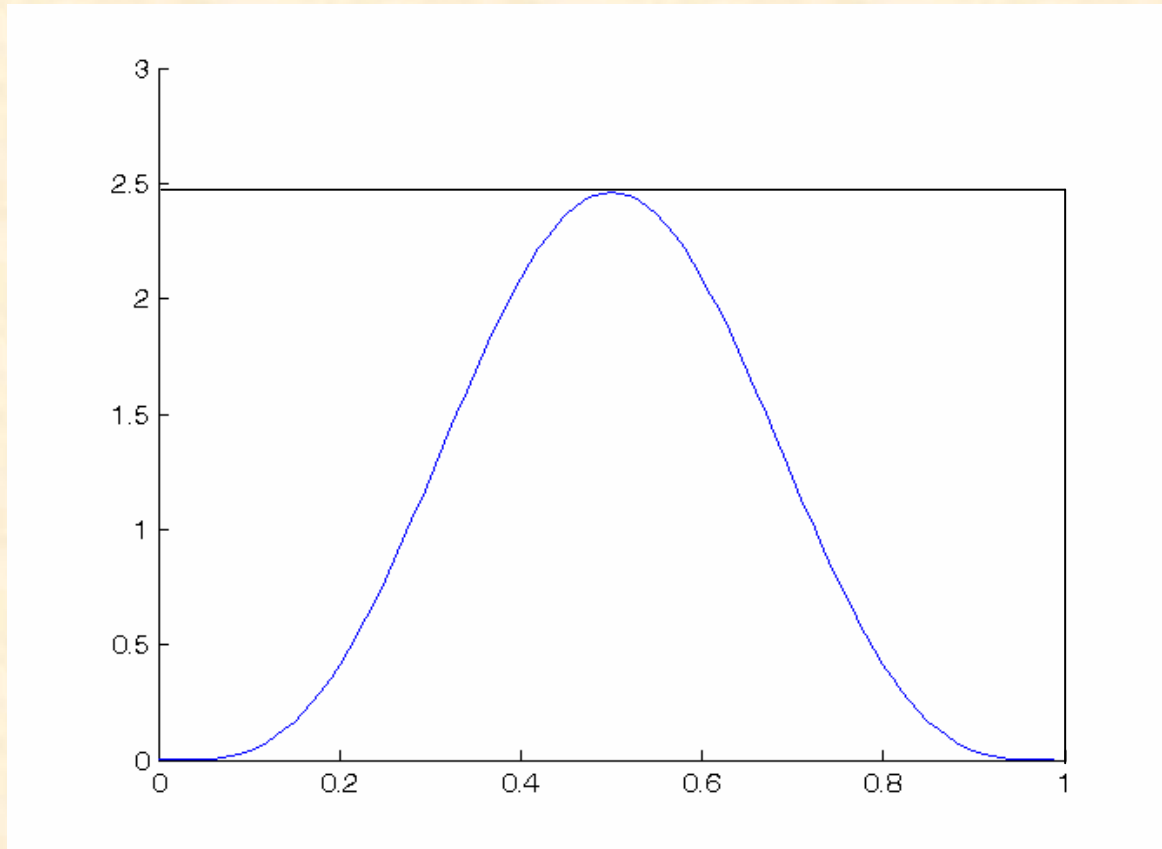
- Common in Bayesian inference for a posterior kernel to be specified
- If it is difficult to find the exact distribution several methods allow us to draw from the exact distribution using just the kernel.
- *Rejection sampling* allows us to draw a random sample from the exact target distribution
  - There is an efficiency cost

# Rejection Sampling

- Make use of another density  $g(y)$  that is easy to sample from that contains our target density  $f(y)$ 
  - Envelope
- Requires knowledge of  $f(y)$  that we use in constructing  $g(y)$
- Easy to construct an envelope if the target density is log-concave
  - Second derivative negative if scalar
  - Matrix of second partials of the log-density must be negative definite



- Generate a variate  $y$  from  $g(\cdot)$
- Accept with probability  $f(y)/g(y)$



- Often one can use a uniform envelope
  - Just need to find the mode
  - But this can be very inefficient

# Exercise 6

- Use a uniform envelope to sample ( $n=1000$ ) from a exponential with parameter  $=0.56$  and compute summary statistics
- Compare to earlier results (direct simulation)
- See if you can find a more efficient envelope distribution  $g(x)$ !

# Markov Chains

- Rejection sampling most useful when we have just one distribution to sample from
  - Usually interested in sampling from an often high-dimension posterior
- What if we are not smart enough to construct an envelope?
- Alternative is to take a non-random sample by constructing a Markov chain

# McMC

- Markov chain Monte Carlo
- Simulation based evaluation of posterior distributions
- Use dependent sequences of random variables  $\{X_t\}$ 
  - Usual positive autocorrelation means that the effective sample size is smaller than dimension of sequence

# First-order Markov chain

- $\{X_t\}$  is a first-order Markov chain if

$$\Pr(X_{t+1} | X_1, X_2, \dots, X_t) = \Pr(X_{t+1} | X_t)$$

- Corresponding to the *transition kernel*  $\Pr(\cdot | \cdot)$  is a unique stationary distribution satisfying:

$$\phi(X)dx = \lim_{t \rightarrow \infty} \Pr(X_{t+1} \in (x, x + dx] | X_t)$$

- MCMC involves constructing a Markov chain for which the stationary distribution is the target posterior distribution

# Stationary Distributions

- Major problem in Markov chain theory is to determine if an invariant distribution  $\phi(X)$  exists and conditions under which iterations of the transition kernel converge to  $\phi(X)$
- MCMC reverse this problem – given  $\phi(X)$  find a transition kernel that converges to  $\phi(X)$

# McMC

- Create a Markov process whose stationary distribution is the target distribution
  - Run the simulation long enough so that the distribution of current draws is close to the target
  - Need to decide how long
- Many ways to construct these Markov chains
- Need to check for convergence

# Example

- 2 states – at  $t$  make a decision about state at  $t+1$ .  
This depends on the state at  $t$

- $\mathbf{p}_t = (p_{1t} \quad p_{2t})$   $p_{it} = \text{Pr}(\text{in state } i \text{ at time } t)$

- *Transition matrix:*  $\Psi = \begin{pmatrix} \psi_{11} & 1 - \psi_{11} \\ \psi_{21} & 1 - \psi_{21} \end{pmatrix}$

$$(p_{1t} \quad p_{2t}) \times \begin{pmatrix} \psi_{11} & 1 - \psi_{11} \\ \psi_{21} & 1 - \psi_{21} \end{pmatrix} = (p_{1t+1} \quad p_{2t+1})$$

# Example

- 2-stage transition:  $\mathbf{p}_{t+2} = \mathbf{p}_t \Psi \Psi$
- $n$ -stage transition:  $\mathbf{p}_{t+n} = \mathbf{p}_t \Psi^n$  ( $\Psi^n = \prod_{i=1}^n \Psi$ )

but as  $n \rightarrow \infty$   $\mathbf{p}_{t+n} \rightarrow \begin{pmatrix} \frac{\psi_{21}}{\psi_{21} + \psi_{12}} & \frac{\psi_{12}}{\psi_{21} + \psi_{12}} \end{pmatrix}$

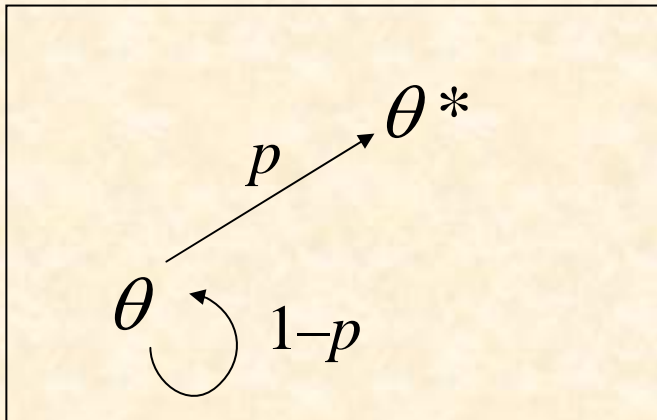
- If Markov chain ‘well-behaved’ final state is the same regardless of where you start
  - *Stationary distribution*

# When Does it Work?

- (1) The sequence must be a Markov chain with a unique stationary distribution
  - Chain must be irreducible, aperiodic, not transient
  - Except for trivial conditions any random walk is aperiodic and not transient
  - Irreducible if there is a non-zero probability of reaching any state from any other state
- (2) The unique stationary distribution must be the target distribution

# Metropolis-Hastings

- The Metropolis-Hastings algorithm ( and Gibbs sampling as a special case) satisfy these properties provided the chain that is generated is irreducible
- Currently at  $\theta$  - propose a new value  $\theta^*$  that is drawn from a candidate generating distribution  $J(\theta^*|\theta)$



- Accept the proposal with

$$p = \frac{f(\theta^*)J(\theta|\theta^*)}{f(\theta)J(\theta^*|\theta)}$$

# Metropolis-Hastings

- Acceptance probability has two parts:

$$(1) \frac{f(\theta^*)}{f(\theta)} \text{ - model} \quad (2) \frac{J(\theta | \theta^*)}{J(\theta^* | \theta)} \text{ - candidate generation}$$

- Useful for posterior simulation as only the kernel is needed

# Example

- Excel demo [mh.xls](#)
  - Binomial kernel
  - Logit normal jumping distribution
- Issues:
  - Specification of jumping (proposal) distribution
  - Starting values
  - Convergence diagnostics
- Python implementation [mh.py](#)

# Gibbs Sampling

- Used in multidimensional applications
- *Alternating conditional sampling*
- Suppose  $\theta$  has  $k$  components  $(\theta_1, \dots, \theta_k)$
- At each iteration, cycle through the components drawing  $\theta_i$  from  $p(\theta_i | \theta_{i-})$ 
  - $\theta_{i-}$  is all the components of  $\theta$  except  $\theta_i$
- $p(\theta_i | \theta_{i-}) =$  Full conditional distribution for  $\theta_i$ 
  - Will denote this later as  $[\theta_i | \cdot]$

# Gibbs Sampling

- In iteration  $t$  element  $\theta_i$  is updated conditional on the latest components of  $\theta$  which are iteration  $t$  values if they have already been updated, iteration  $t-1$  values otherwise
- For many problems the full conditional distributions can be found as known distributions
  - If not use M-H

# Example

- Closed population mark-recapture, model  $p(t)$ 
  - Capture probabilities same for all animals
  - Capture probabilities vary from sample-to-sample

$$[X | U, \{p_i\}] = \frac{(U + u.)!}{u.!U!} \prod_{i=1}^t p_i^{n_i} (1 - p_i)^{U+u.-n_i}$$

- Meadow voles:  $t = 5$ ,  $u. = 54$ ,  $n = (27, 23, 26, 22, 23)$



# Solutions

- Direct sampling from posterior distribution via rejection sampling or M-H
- Derive full conditional distribution and use Gibbs sampling
  - Direct sampling from each if known forms
  - M-H or rejection sampling if not

Likelihood

$$[X | U, \{p_i\}] = \frac{(U + u.)!}{U!} \prod_{i=1}^k p_i^{n_i} (1 - p_i)^{U + u. - n_i}$$

$$[p_i] = p_i^{\alpha_i - 1} (1 - p_i)^{\beta_i - 1} \quad [U] = c$$

Posterior

$$[U, \{p_i\} | X] \propto [X | U, \{p_i\}][\{p_i\}][U]$$

$$\propto \frac{(U + u.)!}{U!} \prod_{i=1}^k p_i^{n_i + \alpha_i - 1} (1 - p_i)^{U + u. - n_i + \beta_i - 1}$$

## Full conditional distribution - $p_i$

$$\frac{(U + u.)!}{U!} \prod_{i=1}^k p_i^{n_i + \alpha_i - 1} (1 - p_i)^{U + u. - n_i + \beta_i - 1}$$

$$[p_i | \cdot] \propto p_i^{n_i + \alpha_i - 1} (1 - p_i)^{U + u. - n_i + \beta_i - 1}$$

$$\propto \text{Beta}(n_i + \alpha_i, U + u. - n_i + \beta_i)$$

# Full conditional distribution - U

$$\frac{(U + u.)!}{U!} \prod_{i=1}^k p_i^{n_i + \alpha_i - 1} (1 - p_i)^{U + u. - n_i + \beta_i - 1}$$

$$[U | .] \propto \frac{(U + u.)!}{U!} \prod_{i=1}^k (1 - p_i)^{U + u. - n_i + \beta_i - 1} \propto \frac{(U + u.)!}{U!} \pi_0^U$$

$$NB(u. + 1, 1 - \pi_0)$$

$$\pi_0 = \prod_{i=1}^k (1 - p_i)$$

# So sampling is by

- Given initial values for  $p$  and  $U$

- Sample  $p_i$  from

$$Beta(n_i + \alpha_i, U_i + u. - n_i + \beta_i)$$

- Sample  $U$  from

$$NB(u. + 1, 1 - \pi_0)$$

– where

$$\pi_0 = \prod_{i=1}^k (1 - p_i)$$

- Implementation in [Mt.py](#) (uniform prior on  $p$ 's )

# Exercise 7

- Code a Gibbs's sampler for the CMR data just described under the assumption of constant  $p$ 's across periods (individuals, etc.)
  - Assume uniform prior on  $U$
  - Beta(1,1) prior on  $p$

- Hint:

$$[X | U, p] = \frac{(U + u.)!}{U!} p^{n.} (1 - p)^{k(U + u.) - n.}$$

$$n. = \sum_{i=1}^k n_i$$

# Issues

- Starting values
- Proposal (jumping) distributions
- Convergence

# Starting values

- Start with reasonable (good) values!
  - Moment or MLE estimates
  - Approximations
  - Limiting values (e.g., lower bound of  $U$  is  $u$ )
  - Random values from priors

# Jumping distributions

- Symmetric vs. asymmetric
  - Asymmetric sometimes converges faster
- Variance
  - Too low: not enough mixing
  - Too high: locks in on extreme values
  - This is where ‘adaptive tuning’ comes in

# Convergence

- Examine trace
- Diagnostic statistics
  - Geweke
  - Autocorrelation
  - Others (see CODA)
- Burn in:
  - Theoretically not needed for GS
  - In general, good rule is to throw out first 1/2