# From Monte Carlo to MCMC

(BUGS is based on MCMC)

- Monte Carlo methods providing i.i.d samples (independent identically distributed)
  - In practice: with standard distributions, random number generators available in statistical software, e.g. in R: rbinom, rbeta, rgamma…
  - If non-standard, do-it-yourself:
  - Inverting cumulative distribution function
  - Rejection sampling
  - Importance sampling
  - In Bayesian inference: posterior distribution is our target distribution in all cases below.

# Inverting cumulative distribution function

- If target density $\pi(\theta)$ has a cumulative distribution function F (kertymäfunktio)

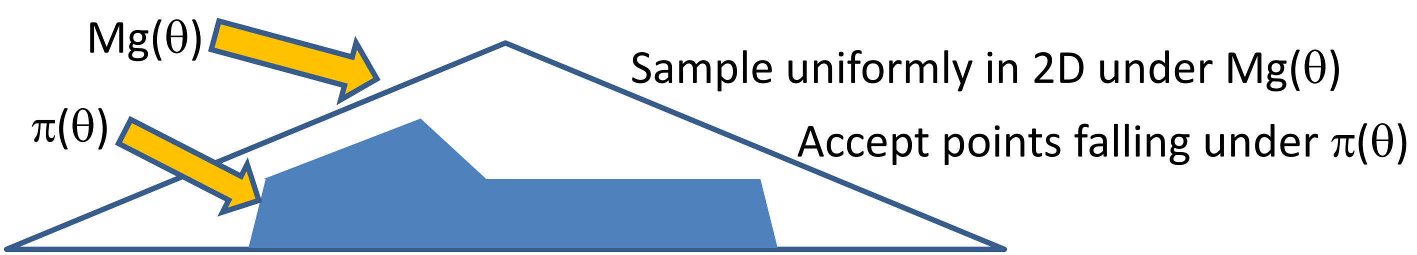
  $$F(\theta') = P(\theta<\theta') = \int_{-\infty}^{\theta'} \pi(\theta)d\theta = u(\theta')$$

  which can be inverted for solving $\theta' = F^{-1}(u)$, then we can generate $u \sim U(0,1)$ and evaluate $\theta'=F^{-1}(u)$.

  Resulting variables will be distributed as the target density.

# Rejection sampling

- Target density is some $\pi(\theta)$.
- Choose instrumental density $g(\theta)$, and some constant M so that $\pi(\theta)/(Mg(\theta)) \le 1$
- Instrumental density should be *easy to sample*, and have the same support as $\pi(\theta)$.
- Algorithm:
  - Step 1. Generate random value from density g.
  - Step 2. Accept this with probability $\pi(\theta)/(Mg(\theta))$.
  - Repeat until enough large sample was obtained.



# Importance sampling

- Target density is $\pi(\theta)$.
- Choose instrumental density $g(\theta)$, easy to sample, same support as $\pi(\theta)$.
- Use weighted sample in calculations, e.g. for mean:

$$E_\pi(\theta) = \int \theta \pi(\theta) d\theta = \int \left[ \theta \frac{\pi(\theta)}{g(\theta)} \right] g(\theta) d\theta =$$

$$= E_g(\theta \frac{\pi(\theta)}{g(\theta)}) \approx \frac{1}{K} \sum_{k=1}^{K} \theta_k \frac{\pi(\theta_k)}{g(\theta_k)}$$

# Rejection sampling from $\pi(\theta,X)$, ABC-method

- Target density is $\pi(\theta|X)$, for some data X.
- Use method of composition to sample $\theta$ from $\pi(\theta)$, then X from $\pi(X|\theta)$.
- Accept only those samples of X & $\theta$, where X equals the observed data X.
- This produces exactly the conditional probability according to Bayes theorem
- ABC = Approximate Bayesian Computation:
  - When data X have continuous variables, use $[X-\varepsilon, X+\varepsilon]$
  - ABC useful when likelihood function cannot be written in analytically closed from, but if we can only simulate X.

# *Multidimensional* posteriors

- Practical problems nearly always have many unknown parameters $\theta_1, \theta_2, \ldots, \theta_n$
- Target is: $\pi(\theta_1, \theta_2, \ldots, \theta_n \mid \text{data})$
- Multivariate distributions can be handled in a sequence of univariate distributions, e.g.:

$$\pi(\theta_1, \theta_2, \theta_3) = \pi(\theta_3 \mid \theta_1, \theta_2)\, \pi(\theta_2 \mid \theta_1)\, \pi(\theta_1)$$

- Useful method for simulating n-dimensional posteriors: Markov chain Monte Carlo (BUGS is based on this)

# Monte Carlo Markov chain

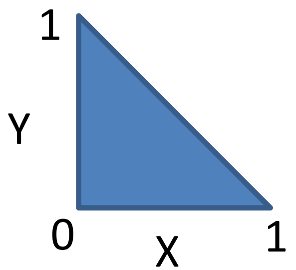
**Innovation:**

- Construct a sampler that works as a Markov chain, for which stationary distribution exists, and this stationary distribution is the same as our target distribution.
- This can be done even without knowing normalizing constant of the posterior – so we only need to be able to evaluate:

  Posterior $\propto$ prior $\times$ likelihood
- Mathematical proofs left for advanced courses...

# Special case: Gibbs sampler

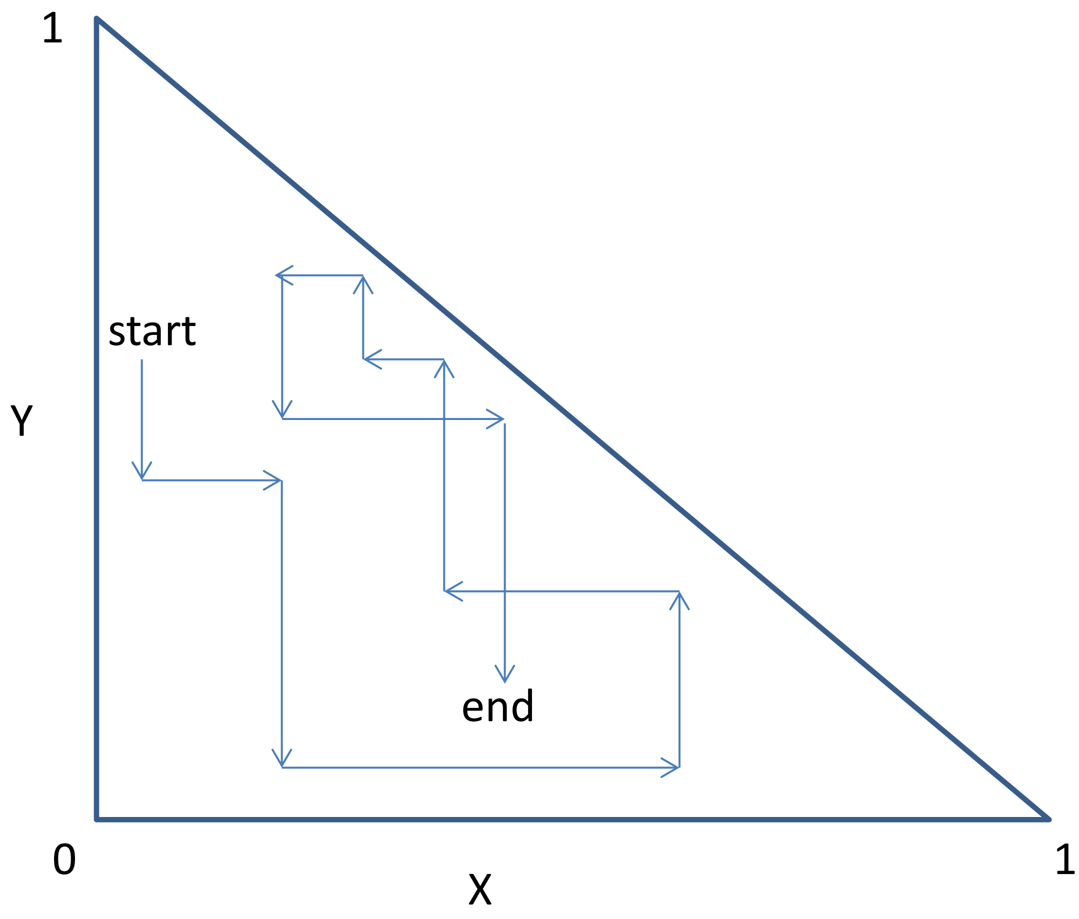
- Gibbs sampling in 2D
  - Example: uniform distribition in a triangle.




$$\pi(x,y) = 2 \times \mathbf{1}_{\{y<1-x,\,0<x<1,\,0<y<1\}}(x,y)$$

  - Sample this using Gibbs

# Gibbs sampler visually



# Gibbs sampling in 2D

- Remember product rule:

  **$\pi(x,y) = \pi(x|y)\pi(y) = \pi(y|x)\pi(x)$**

  - Solve the marginal density $\pi(x)$:

$$\pi(x) = \int_0^1 \pi(x,y)dy$$

$$= \int_0^1 2 \times \mathbf{1}_{\{y<1-x,\, 0<x<1,\, 0<y<1\}}(x,y)dy = \int_0^{1-x} 2\, dy = 2(1-x)$$

  - Then solve: $\pi(y|x)=\pi(x,y)/\pi(x)$

# Gibbs sampling in 2D

- Solve the conditional density:

$$\pi(y \mid x) = \frac{\pi(x,y)}{\pi(x)} = \frac{2 \times \mathbf{1}_{\{y<1-x,\, 0<x<1,\, 0<y<1\}}(x,y)}{2(1-x)}$$

$$= \frac{1}{1-x}\mathbf{1}_{\{y<1-x,\, 0<y<1\}}(y) \quad = \quad U(0,1-x)$$

  - Note: above it would suffice to recognize $\pi(y|x)$ up to a constant term, so that solving $\pi(x)$ is not necessary.
  - Similarly, get $\pi(x|y) = U(0,1-y)$.

# Gibbs sampling in 2D

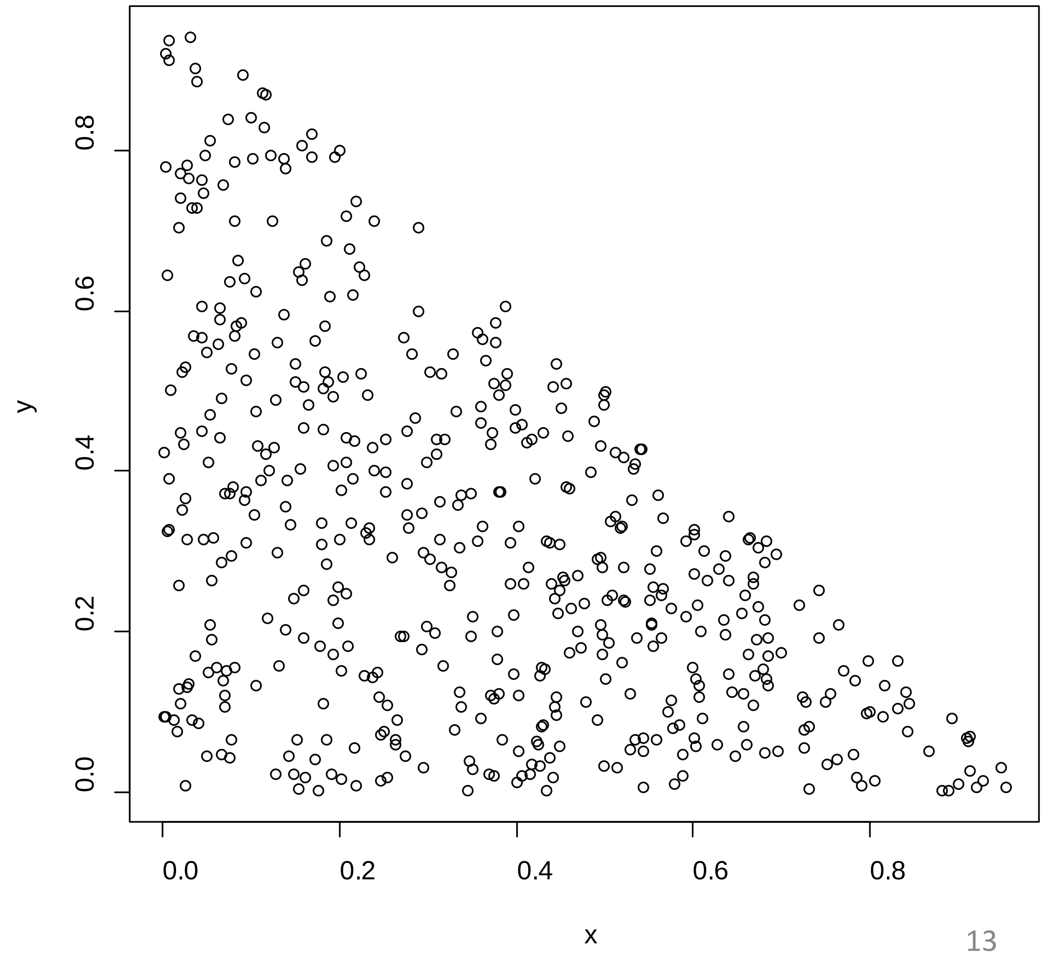
- Starting from the joint density $\pi(x,y)$, we have obtained two important conditional densities: $\pi(x|y)$ and $\pi(y|x)$     (aka 'full conditionals')
  - Gibbs algorithm is then:
  - (1) start from $x^0, y^0$. Set k=1.
  - (2) sample $x^k$ from $\pi(x|y^{k-1})$
  - (3) sample $y^k$ from $\pi(y|x^k)$. Set k=k+1.
  - (4) go to (2) until sufficiently large sample.

  - These samples are no longer i.i.d.

# Gibbs sampler

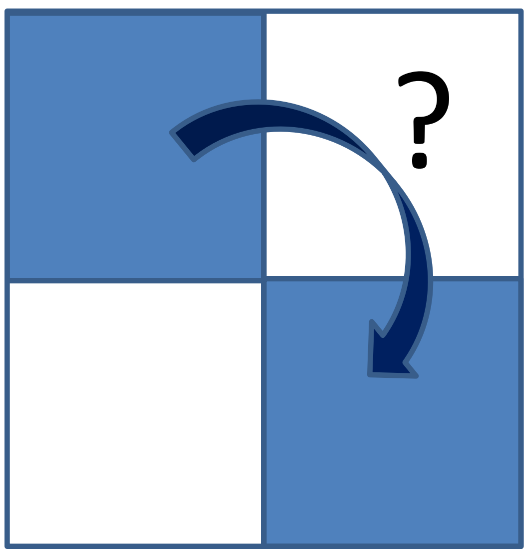
- **In R, you could type:**

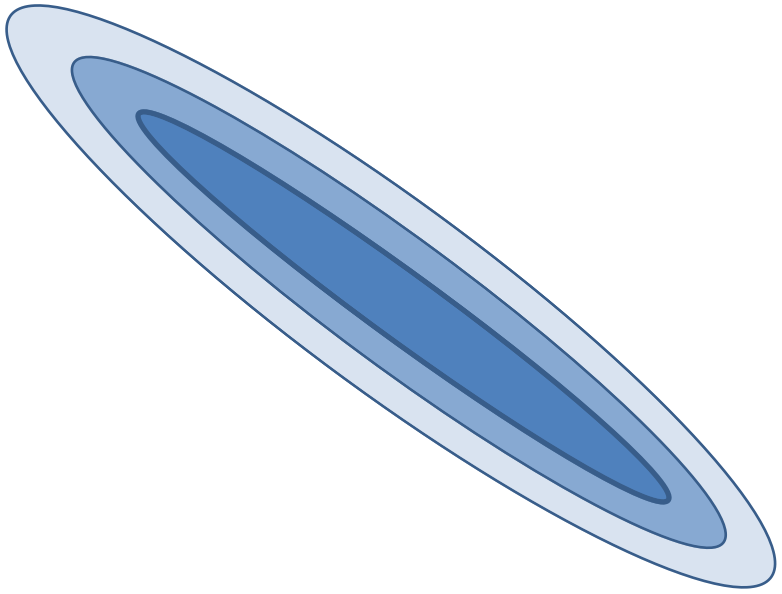
```
x <- numeric()
y <- numeric()
x[1] <- 0.5
y[1] <- 0.4
for(i in 2:500){
    y[i] <- runif(1,0,1-x[i-1])
    x[i] <- runif(1,0,1-y[i])
    }
plot(x,y)
```



# Gibbs sampler

- **Jumping around? Possible problems.**



# Gibbs sampling Binomial model

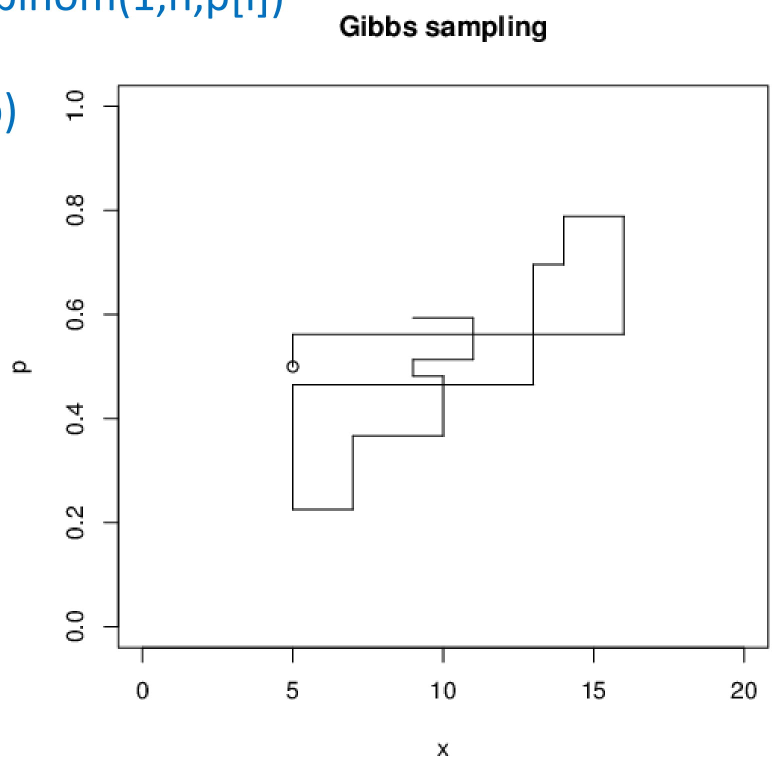
- **”conditional on N”**
  - **Joint distribution $\pi(\theta,X|N)$** can be expressed either as $\pi(X|\theta,N)\pi(\theta|N)$ or $\pi(\theta|X,N)\pi(X|N)$.
  - From the first, we recognize **$\pi(X|\theta,N)=\text{Bin}(N,\theta)$**
  - With e.g. uniform prior **$\pi(\theta|N)=\pi(\theta)=U(0,1)$,** we would know **$\pi(\theta|X,N) = \text{Beta}(X+1,N-X+1)$.**

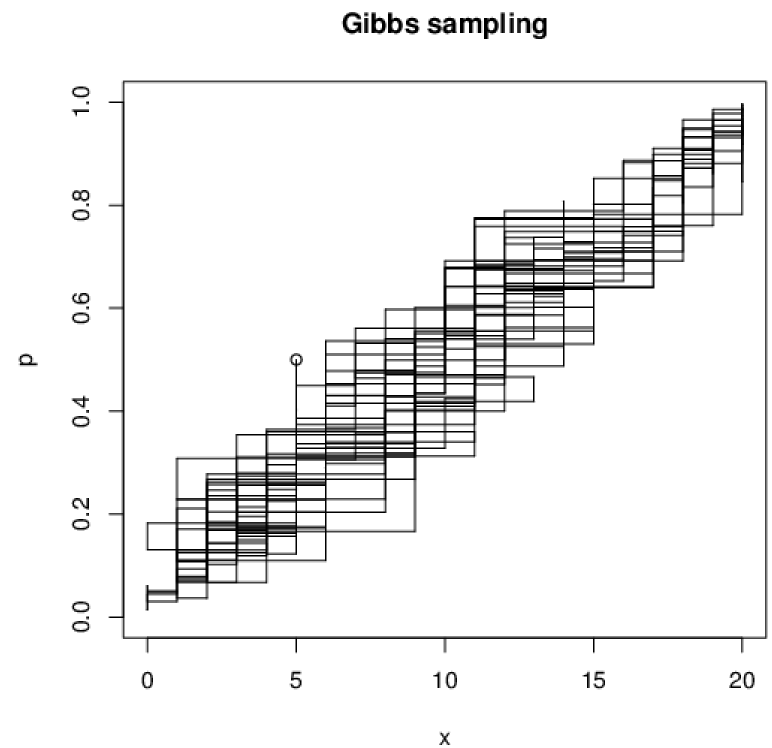
  - **This gives $\pi(\theta|X)$ and $\pi(X|\theta)$ needed for Gibbs.**
  - Gibbs will produce the same joint distribution for $\theta$,X as with the method of composition.
  - Note: for this one-parameter inference (when X is fixed data) Gibbs is not needed, but could be used to obtain predictive distribution of X.

# Gibbs sampler

- **Binomial model, ”conditional on N”, in R:**

```
n<-20; p <- numeric(); x<- numeric()
p[1] <- 0.5; x[1] <- 10   # initial values
for(i in 2:1000){
p[i] <- rbeta(1,x[i-1]+1,n-x[i-1]+1)
x[i] <- rbinom(1,n,p[i])
}
plot(x,p)
```



# Gibbs and 2D-normal density

- **2D normal density:**

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

- Marg. densities $\pi(x)$ and $\pi(y)$ are both N(0,1)
- Joint density function is:

$$\pi(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

# Gibbs and 2D-normal density

- **2D normal density:**

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \sim N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right)$$

- Conditional density $\pi(y|x)=\pi(x,y)/\pi(x)$ is:

$$\pi(y \mid x) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(\rho x - y)^2\right) = N(\rho x, 1-\rho^2)$$

# Gibbs and 2D-normal density

- Gibbs would then be sampling repeatedly from:

  - **$\pi(y|x) = N(\rho x, 1-\rho^2)$**
  - **$\pi(x|y) = N(\rho y, 1-\rho^2)$**

  - This can mix slowly if X & Y heavily correlated.

  - **General remark about Gibbs:** full conditionals need to be solved from the correct joint distribution. Not any $\pi(y|x)$ and $\pi(x|y)$ will constitute a proper joint distribution $\pi(y,x)$. E.g. sampling from y ~ N(x,1) and x ~ N(y,1) does not converge anywhere.

# Metropolis-Hastings

- **This is a very general purpose sampler**

- **The core is: 'proposal distribution' and 'acceptance probability'.**

- **At each iteration:**
  - Random draw is obtained from proposal density $Q(\theta^* | \theta^{i-1})$, which can depend on previous iteration.
  - Simply, it could be $U(\theta^{i-1} - L/2 , \theta^{i-1} + L/2)$.

# Metropolis-Hastings

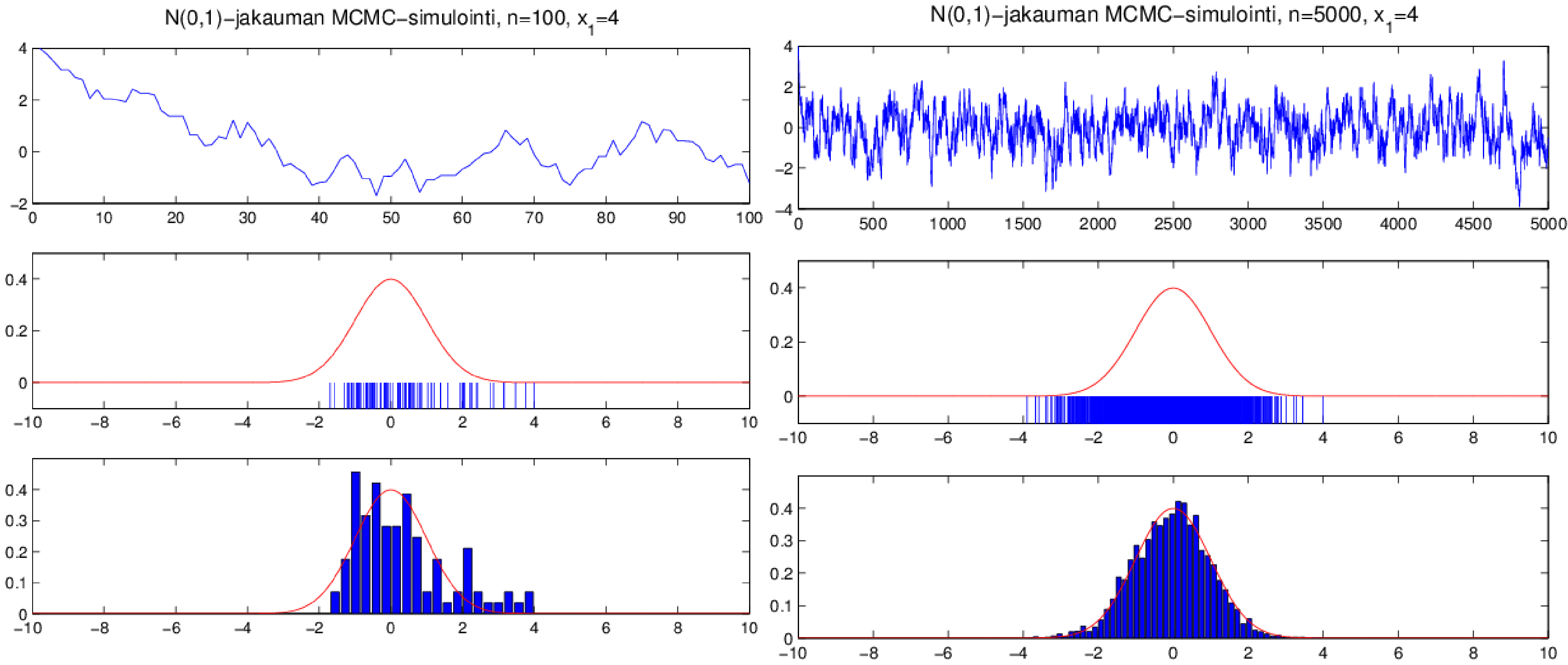
- **At each iteration:**
  - **Proposal is accepted with probability**

$$r = \min\left( \frac{\pi(\theta^* \mid \text{data}) Q(\theta^{i-1} \mid \theta^*)}{\pi(\theta^{i-1} \mid \text{data}) Q(\theta^* \mid \theta^{i-1})}, 1 \right)$$

  - **Note how little we need to solve about π(θ|data)!**
    - **Normalizing constant cancels out from the ratio.**
    - **Enough to be able to evaluate prior and likelihood terms.**
    - **Proposals too far → accepted rarely → slow sampler**
    - **Proposals too near → small moves → slow sampler**
    - **Acceptance probability ideally about 20%-40%**
  - **Gibbs sampler is a special case of MH-sampler**
    - **In Gibbs, the acceptance probability is 1.**
    - **Block sampling also possible.**

# Metropolis-Hastings

- **Sampling from N(0,1), using MH-algorithm:**



# MCMC convergence

- **Remember to monitor for convergence!**
  - Chain is only approaching the target density, when iterating a long time, k→∞.
  - Convergence **can be very slow** in some cases.
  - Autocorrelations between iterations are then large → makes sense to take a thinned sample.
  - Systematic patterns, trends, sticking, indicate problems.

  - Pay attention to starting values! Try different values in different MCMC chains. (discard burn-in period).

# MCMC convergence

- Can only diagnose poor convergence, but cannot fully prove a good one! (e.g. multimodal densities).