# The Bayesian Idea

- **Data model**: Given parameter $\theta$ the data, $X$, is assumed to be distributed as

$$X \sim \pi(x|\theta)$$

- Parameter $\theta$ of interest is unknown.
- A priory knowledge is summarised in terms of **prior density**

$$\pi(\theta)$$

- Conditional on the observed data $x$, we obtain the **posterior**

$$\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta).$$

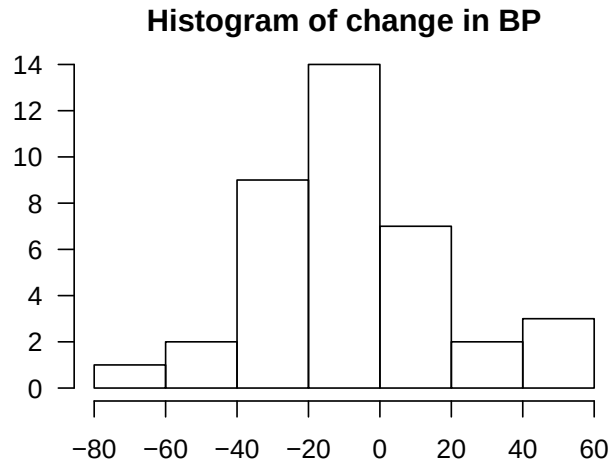
- All **conclusions** regarding $\theta$ are based on the posterior.

# Example Blood pressure

**Setup**: A group of $n = 38$ (imaginary) patients with high blood pressure are given a new drug.

Let $x_i$ denote the change in blood pressure ($x_i < 0$ is good).

**Data**:




**Data model**:

$$x_1, \ldots, x_{62} \overset{iid}{\sim} \mathcal{N}(\mu, \tau).$$

**Question**: Is there a positive effect of the drug? I.e. is $\mu < 0$?

# Bayesian analysis of blood pressure

**Prior**: (assume $\tau$ known)

$$\pi(\mu) = \mathcal{N}(\mu_0, \tau_0)$$

**Posterior**:

$$\pi(\mu|\mathbf{x}) \propto \pi(\mathbf{x}|\mu)\pi(\mu) \propto \mathcal{N}\left(\frac{n\tau\bar{x} + \tau_0\mu_0}{n\tau + \tau_0}, n\tau + \tau_0\right).$$

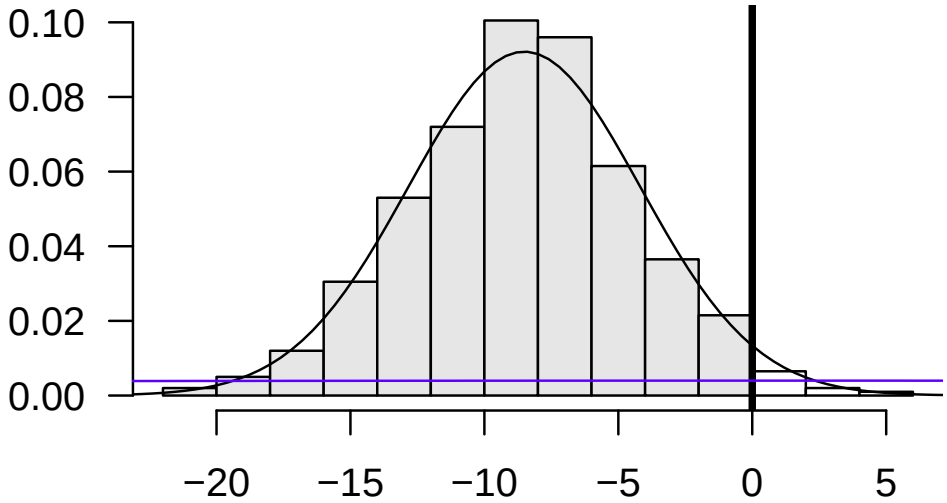
**Question**: What is the posterior probability that $\mu < 0$, i.e. that the (mean) effect of the drug is "good"? In maths terms:

$$P(\mu < 0|\mathbf{x})?$$

**Answer**: $P(\mu < 0|\mathbf{x}) = 0.975$ (Looked up in a table)

# Monte Carlo: Simulating an answer

- Notice that $P(\mu < 0|\mathbf{x}) = \mathbb{E}\Big[\mathbb{1}[\mu < 0]|\mathbf{x}\Big]$
- Simulate $\mu^{(1)}, \mu^{(2)}, \ldots, \mu^{(1000} \sim \pi(\mu|\mathbf{x}) = \mathcal{N}(\ldots, \ldots)$. Histogram:



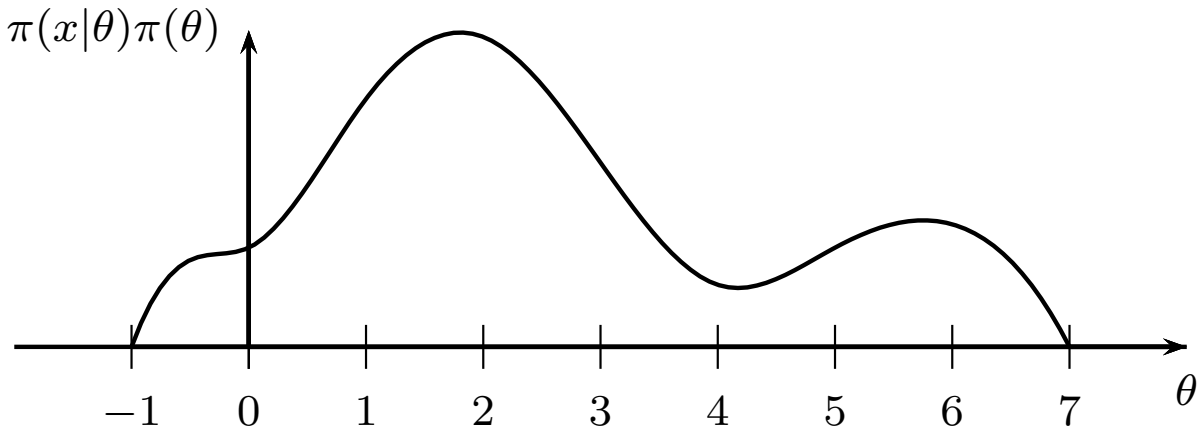

- A (Monte Carlo) estimate of $P(\mu < 0|\mathbf{x})$ is now given by

$$\frac{1}{1000}\sum_{i=1}^{1000} \mathbb{1}[\mu^{(i)} < 0] = 0.981$$

# New problem

- The posterior is $\pi(\theta|x) \propto \pi(x|\theta)\pi(\theta)$.
- In general $\pi(x|\theta)\pi(\theta)$ is *not* a normalised pf/pdf.

Assume the following is a plot of $\pi(x|\theta)\pi(\theta)$:




**Question**: What is the probability $P(\theta > 5|x)$?

**Half answer**: Simulations of $\pi(\theta|x)$ could answer this.

**Problem**: Density is not (proportional to) any well-known distribution.

**Solution**: Simulation — in particular Markov chain based simulation.

# Markov chain

## Definition: Markov chain

A sequence of random variables $(X^{(0)}, X^{(1)}, X^{(2)}, \ldots)$ is called a *homogeneous* Markov chain if

$$\begin{aligned} &P(X^{(t+1)} \in A | X^{(0)} = x^{(0)}, X^{(1)} = x^{(1)}, \ldots, X^{(t)} = x^{(t)}) \\ &\quad = P(X^{(t+1)} \in A | X^{(t)} = x^{(t)}) \\ &\quad = P(x^{(t)}, A), \end{aligned}$$

where $P(x, A)$ is the **Transition kernel**.

## Definition: $n$-step transition kernel

$$P^n(x, A) = P(X^{(t+n)} \in A | X^{(t)} = x).$$

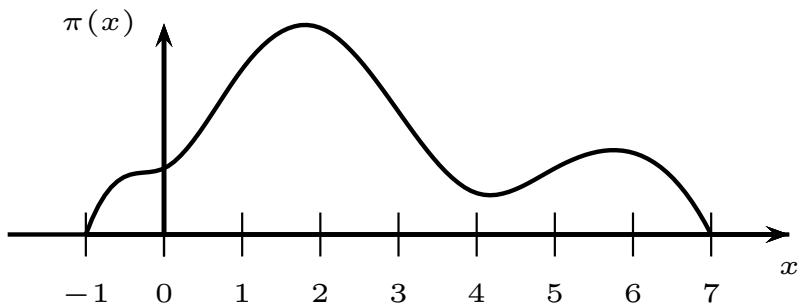
# Simulating more answers

The Gibbs sampler relies on generating samples from known well-known distributions.

We now consider the situation, where want to sample from distribution which is not standard. We restrict attention to distributions specified by a density.

Let $\pi(x)$ be out *Target density*, i.e. the density we want to sample from.

As an example we want generate a sample from this density:




**Aim**: We want to generate a Markov chain $(X^{(0)}, X^{(1)}, X^{(2)}, \ldots)$, so that $X^{(t)}$ is approximately a sample from a distribution with density $\pi(x)$.

**Notation**: In the following $\Pi(A) = \int_A \pi(x)dx$.

# Accept-Reject Algorithm

Let $\pi(x)$ be out *Target density*, i.e. the density we want to sample from.

## Accept-Reject Algorithm

Choose initial value $x^{(0)}$.
For $t = 1, 2, \ldots, T$

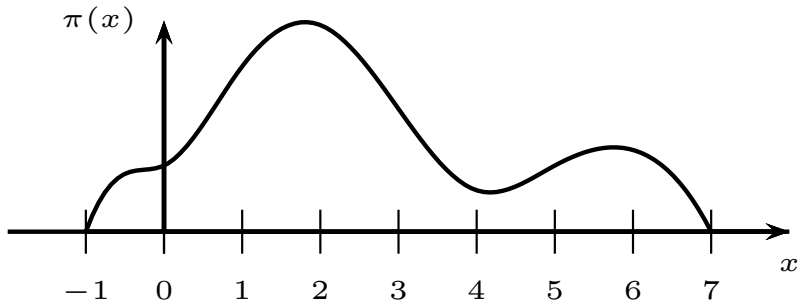
1. Generate **Proposal**: $y \sim q(x^{(t-1)}, y)$.
2. Accept proposal with probability: $a(x^{(t-1)}, y)$
   otherwise reject it.
3. If **accepting**: $x^{(t)} = y$
4. If **rejecting**: $x^{(t)} = x^{(t-1)}$

This algorithm generate a realisation of a time homogeneous Markov chain.

How do we choose $q(x, y)$ and $a(x, y)$ so that the unique invariant distribution of the resulting Markov chain is given by $\pi(x)$?

# Example: The Metropolis Algorithm

We want to generate a sample from a distribution with this density:



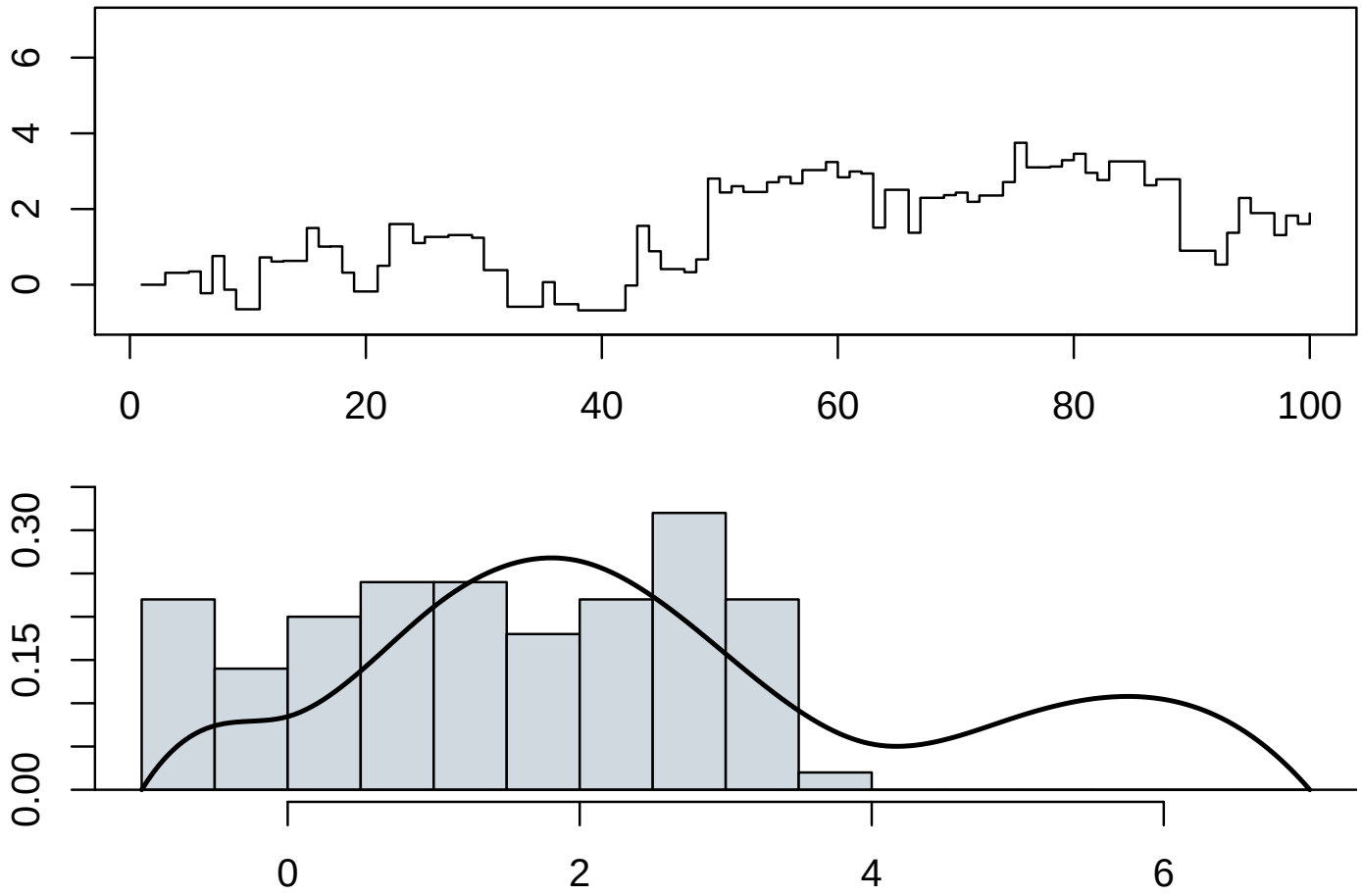

**Proposal distribution** is normal, centred at current value, and with precision $\tau_p$:

$$q(x, y) = \sqrt{\frac{\tau_p}{2\pi}} \exp(-\frac{1}{2}\tau_p(y - x)^2)$$

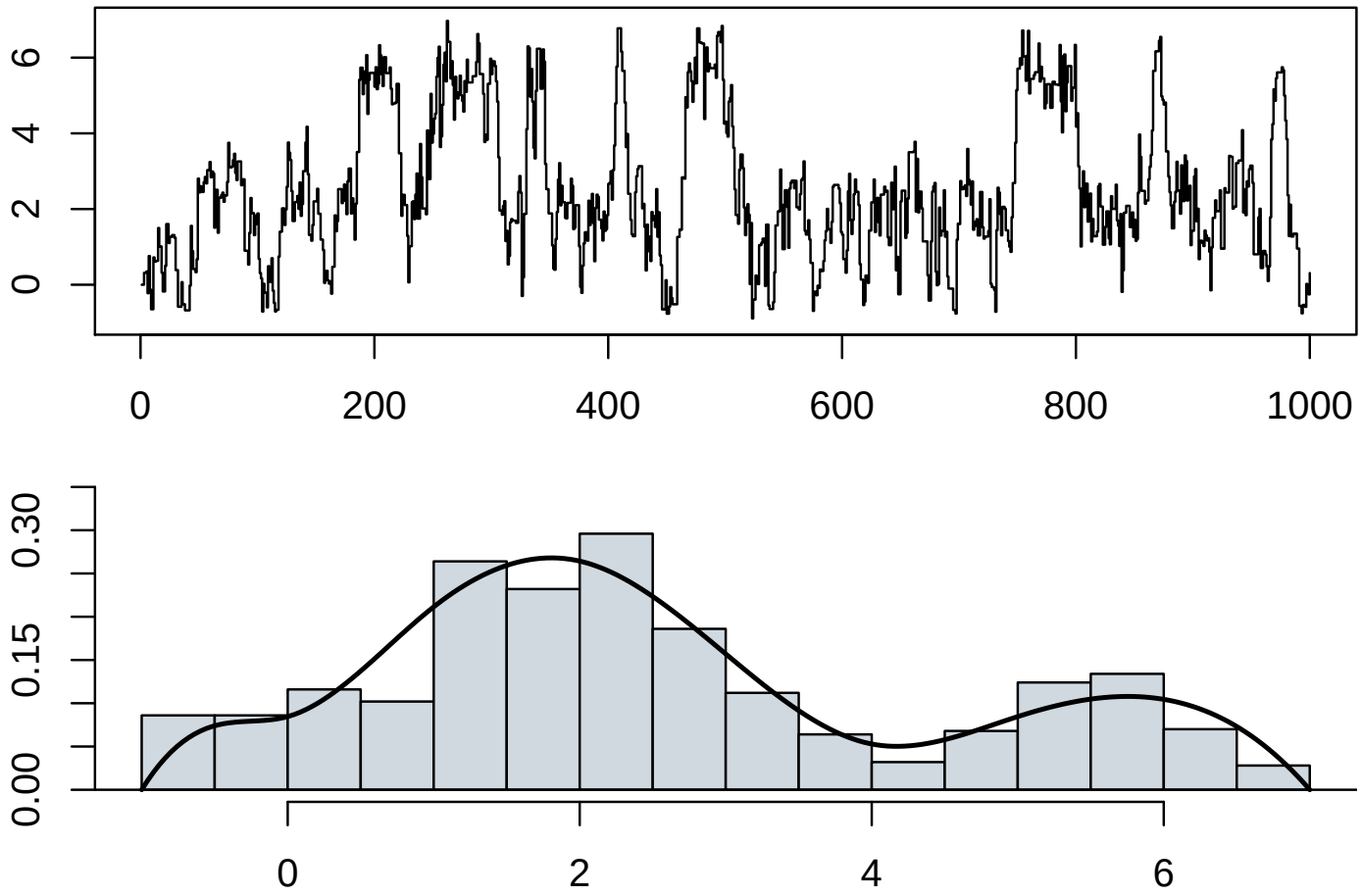
Let the **acceptance probability be**

$$a(x, y) = \min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}$$

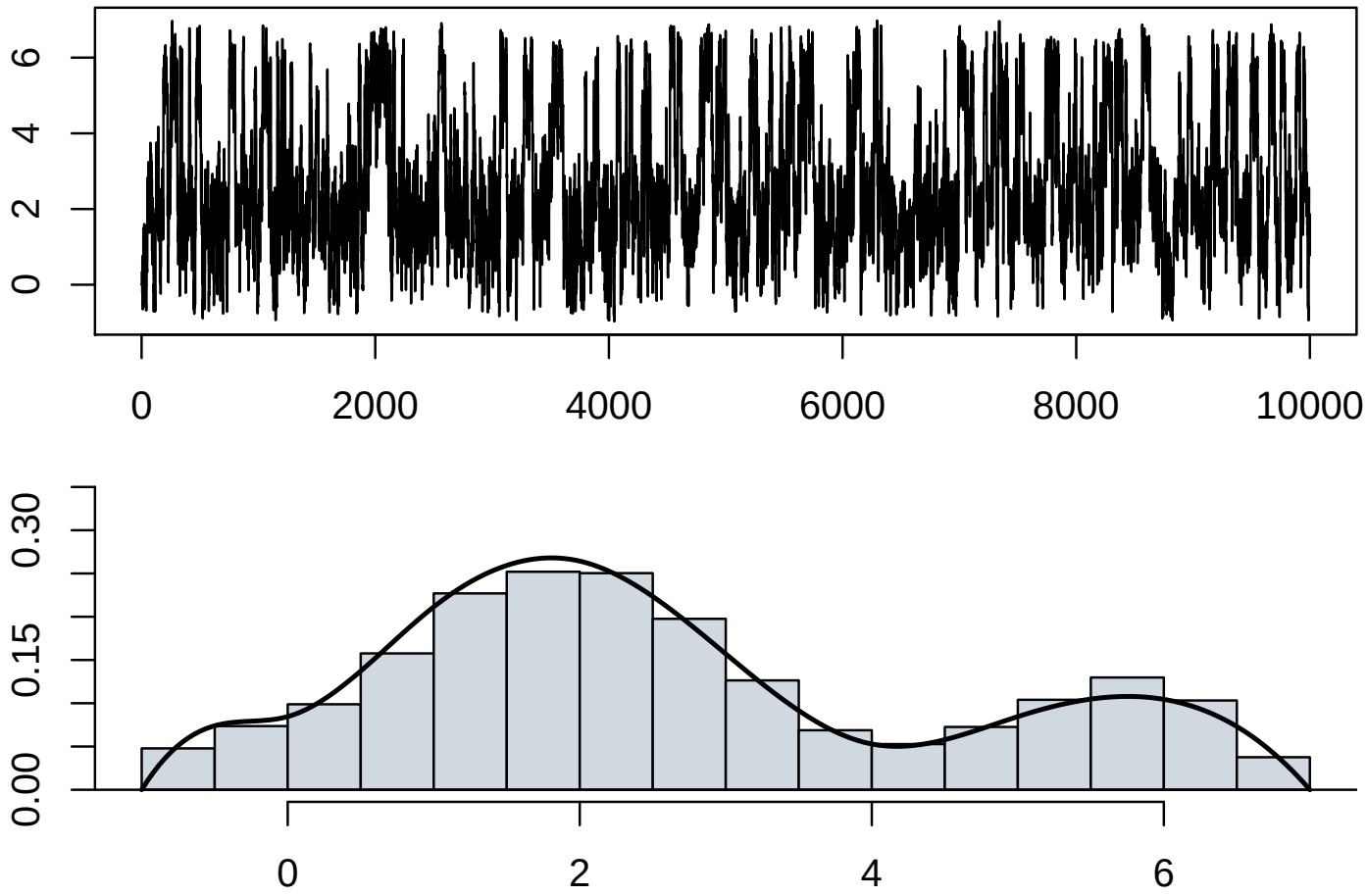
# Results: 100 iterations



# Results: 1000 iterations



# Results: 10000 iterations



# The Metropolis-Hastings algorithm

How to choose $q(x, y)$ and $a(x, y)$?

One choice leads to the Metropolis-Hasting algorithm. The user specifies a proposal kernel $q(x, y)$. The algorithm then "automatically" chooses the correct acceptance probability.

## Metropolis-Hastings algorithm

- Choose any proposal kernel $q(x, y)$
- Define the *Hastings ratio*

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)},$$

  where $H(x, y) = \infty$ if $\pi(x)a(x, y) = 0$.
- The acceptance probability is

$$a(x, y) = \min\{1, H(x, y)\}.$$

# The Metropolis algorithm

A special case of the MH-algorithm is when the proposal kernel is symmetric:

$$q(x, y) = q(y, x)$$

In this case the Hastings-ratio simplifies to

$$H(x, y) = \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)} = \frac{\pi(y)}{\pi(x)}.$$

**Example**: The most common example, is when the proposal is normal distributed with $x$ as the mean value, and $\tau_p$ as the precision:

$$q(x, y) = \sqrt{\frac{\tau_p}{2\pi}} \exp\left(-\frac{1}{2}\tau_p(y - x)^2\right).$$

Clearly, $q(x, y) = q(y, x)$.

# Invariance

## Definition: Invariant density

Markov chain with Transition kernel $P(x, A)$ has **invariant density** $\pi(x)$, if for all $A \subseteq \Omega$

$$\int_{\Omega} \pi(X) P(x, A) dx = \int_{A} \pi(x) dx$$

## Theorem

The Metropolis-Hastings algorithm produces a time homogeneous Markov chain with $\pi$ as an invariant density.

# Irreducible

## Definition: Irreducible Markov chain

Assume Markov chain has invariant distribution with density $\pi(x)$.
Then the Markov chain is **Irreducible** if for all $x \in \Omega$ and $A \subseteq \Omega$ (where $\Pi(A) > 0$) there exists $t$, so that

$$P^t(x, A) > 0$$

## Theorem

An irreducible Markov chain has a unique invariant distribution.

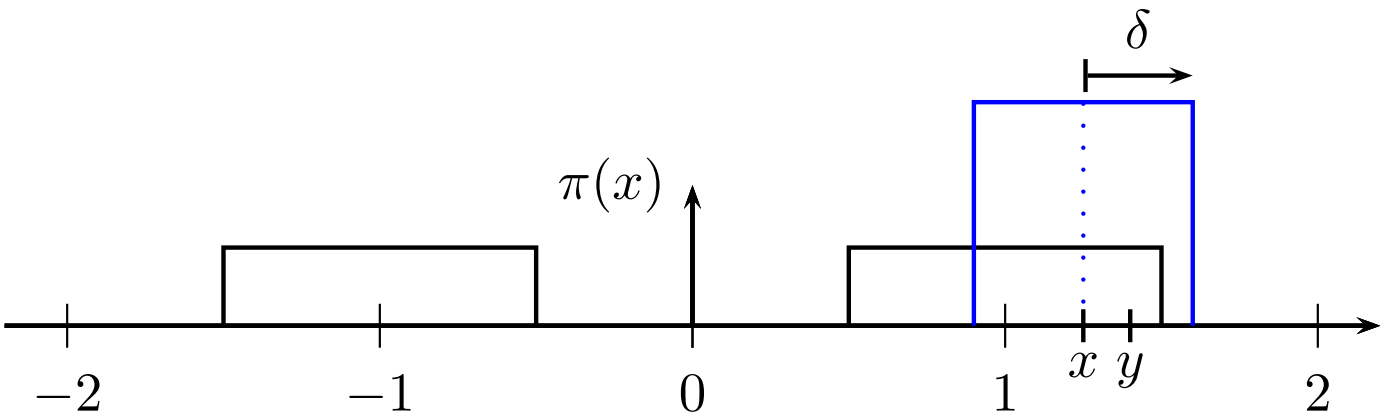
## Theorem

If $q(x, y) > 0$ (for all $x$) whenever $\pi(y) > 0$ then the MH algorithm produces an irreducible Markov chain.

# Irreducible: Example

Consider target density

$$\pi(x) = \frac{1}{2}1\left[|x+1| \leq \frac{1}{2}\right] + \frac{1}{2}1\left[|x-1| \leq \frac{1}{2}\right]$$




Consider proposal density

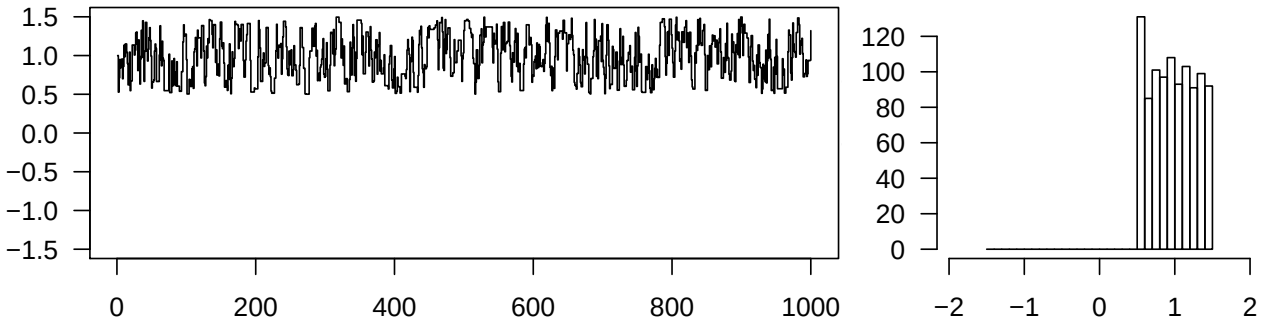
$$q(x,y) = \frac{1}{2\delta}1\left[|x-y| \leq \delta\right]$$

Notice that the proposal density is symmetric in $x$ and $y$, i.e. $q(x,y) = q(y,x)$.

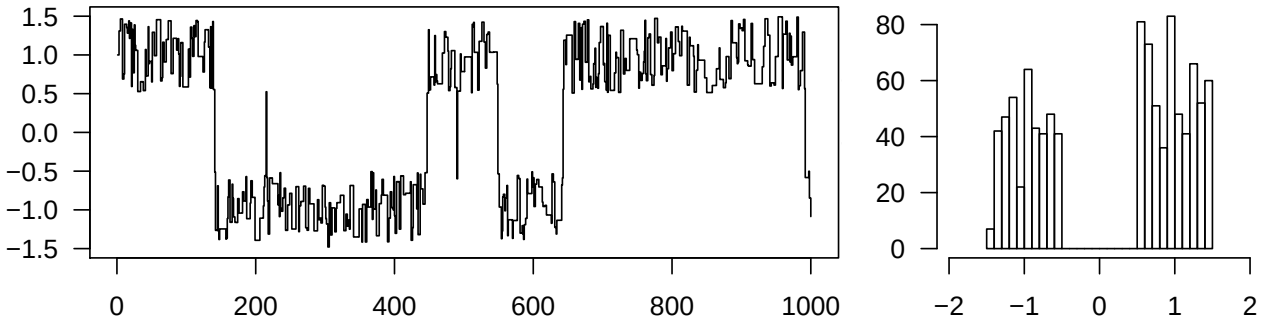
**Notice**: If $\delta$ is too small the Markov chain becomes reducible.

# Irreducible: Example — *cont.*

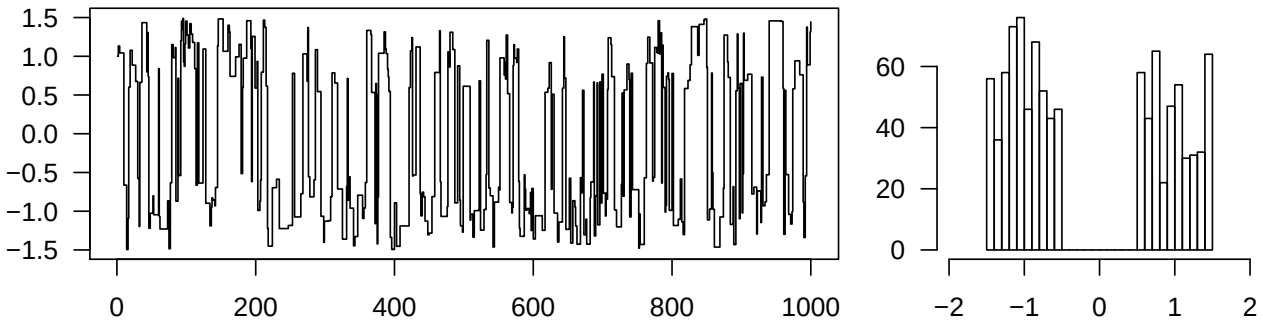
$\delta = 0.5$




$\delta = 1.2$




$\delta = 2.0$



# A Strong Law of Large Numbers

## Theorem: Strong law of large number for Markov chains

Assume Markov chain is **irreducible** with $\pi(x)$ as the unique invariant density.
Assume $h : \Omega \to \mathbf{R}$, so that the **mean** $\mu = \int h(x)\pi(x)dx$ **exists**.
For any $m \geq 0$ define **sample mean**

$$\hat{\mu}_n = \frac{1}{n+1} \sum_{t=m}^{m+n} h(x^{(t)})$$

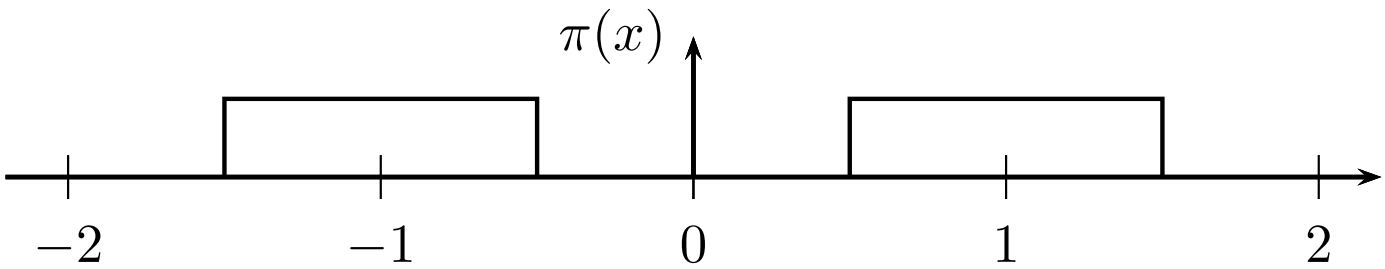
Then there exists a set $C \subseteq \Omega$ with $\Pi(C) = 1$ so that for all $x \in C$

$$P(\hat{\mu}_n \to \mu \text{ as } n \to \infty | X^{(0)} = x) = 1$$

The estimate $\hat{\mu}_n$ is a so-called Markov chain Monte Carlo (MCMC) estimate of $\mathbb{E}[h(X)]$.

# Law of Large Numbers: Example

- **Setup**: Assume $X$ is distributed as before:

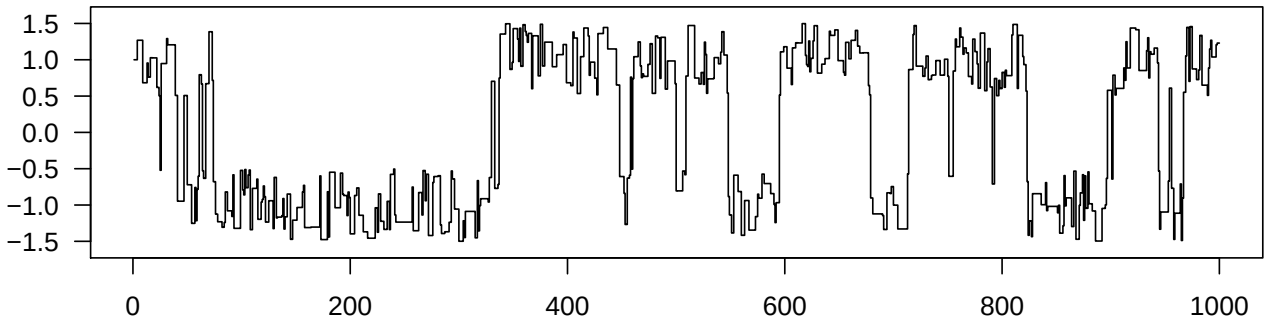



- **Question**: What is the probability $P(X \geq 0)$?
- Notice that $P(X \geq \frac{1}{2}) = \mathbb{E}\left[\mathbb{1}[X \geq 0]\right]$.
- Accordingly $h(x) = \mathbb{1}[X \geq 0]$
- **Solution**: Generate Markov chain $x^{(1)}, x^{(2)}, \ldots, x^{(1000)}$ with proposal as before.
- An estimate for $P(X \geq 0)$ is then

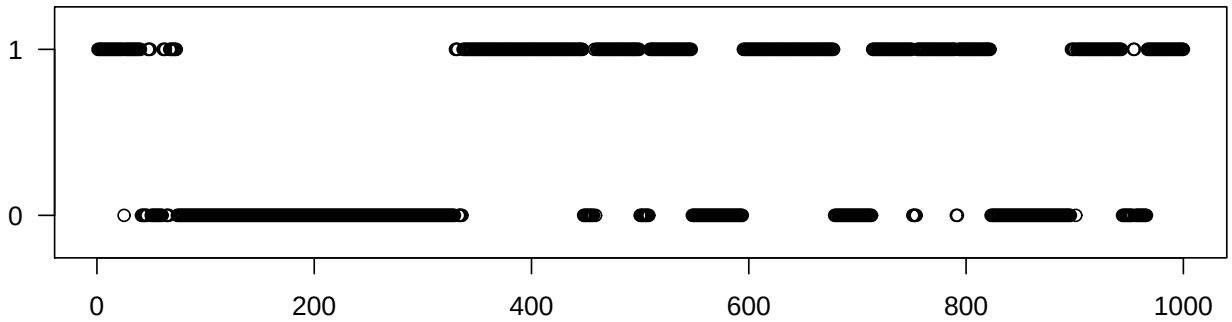
$$\hat{\mu}_{1000} = \frac{1}{1000} \sum_{i=1}^{1000} \mathbb{1}[x^{(i)} \geq 0]$$

# Law of Large Numbers: Example *cont.*

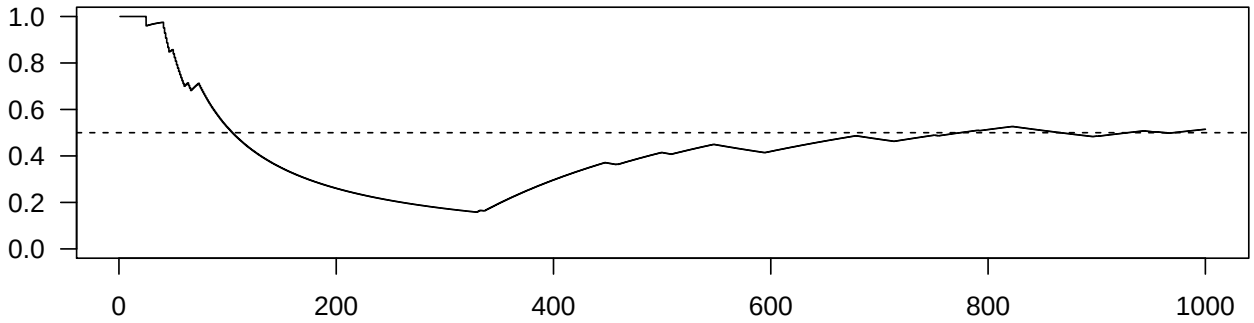
Plot of $x^{(t)}$




Plot of $h(x^{(t)})$




Plot of $\hat{\mu}_t$



# Periodicity

## Definition: Periodicity and Aperiodicity

An irreducible Markov chain is **Periodic** if there exists partition $\Omega = A_0 \cup A_1 \cup A_2 \cup \cdots \cup A_k$, where $k \geq 2$, $\Pi(A_0) = 0$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, so that

- $x \in A_1 \Rightarrow P(x, A_2) = 1$
- $x \in A_2 \Rightarrow P(x, A_3) = 1$
- $\vdots$
- $x \in A_k \Rightarrow P(x, A_1) = 1$

If Markov chain is not periodic, then is it called **Aperiodic**.

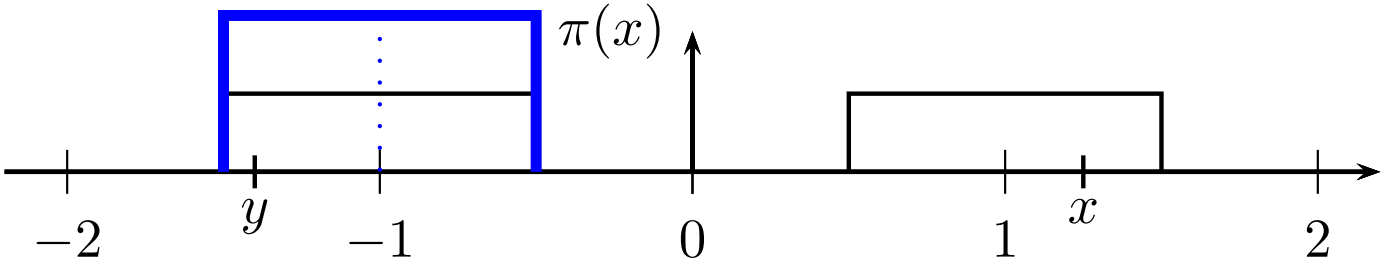
## Theorem

If $P(X^{(t+1)} = X^{(t)}) > 0$ then the MH algorithm produces a aperiodic Markov chain.

# Periodicity: Example

Consider again target density

$$\pi(x) = \frac{1}{2} 1\left[|x+1| \leq \frac{1}{2}\right] + \frac{1}{2} 1\left[|x-1| \leq \frac{1}{2}\right]$$




Consider proposal density

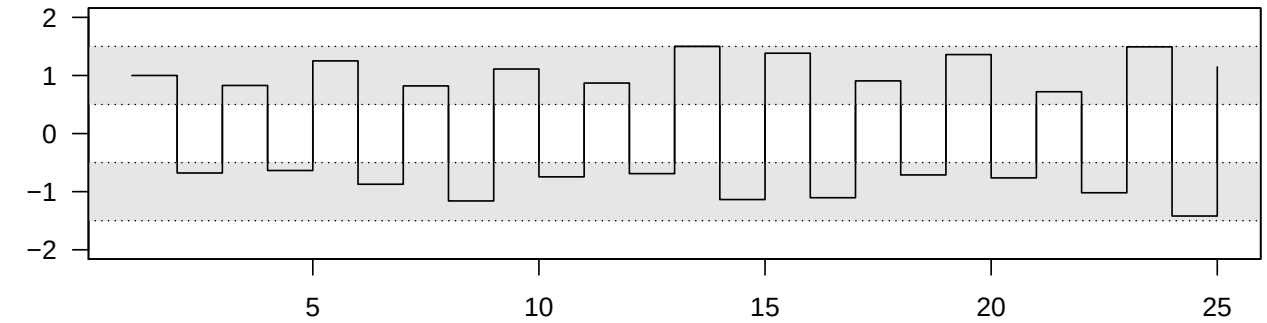
$$q(x, y) = 1\left[\left|x + \mathsf{sign}(x)\right| \leq \frac{1}{2}\right]$$

Accordingly:

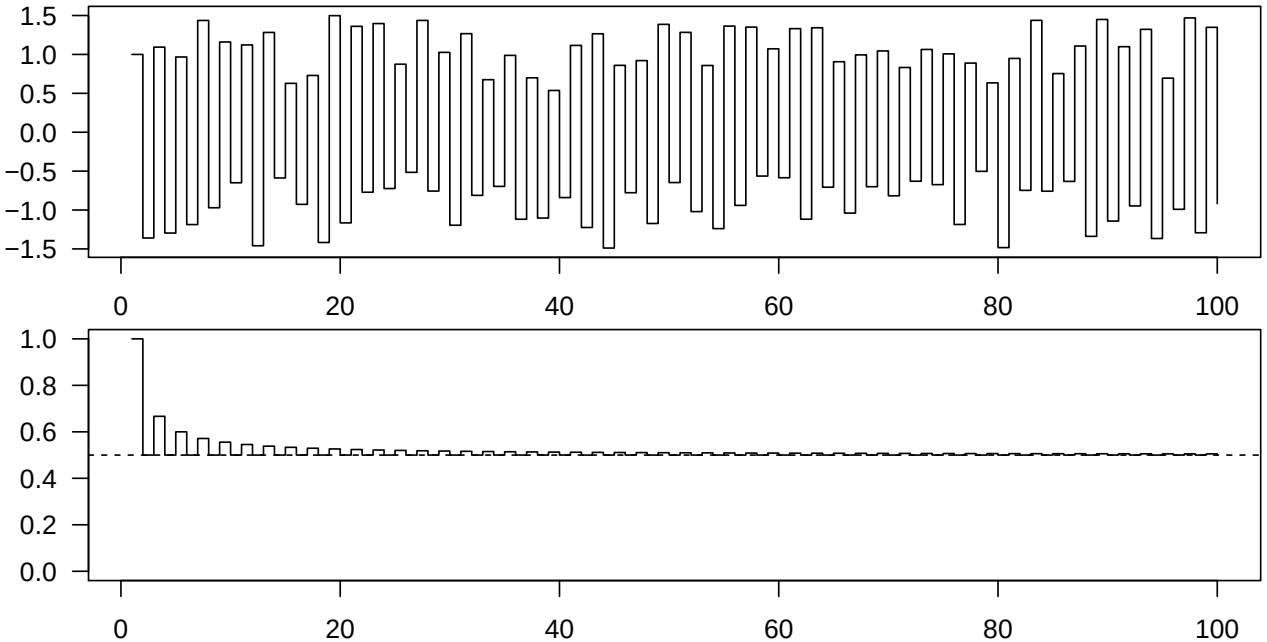
- If $x > 0$ then $y \sim \mathsf{Unif}([-1.5, -0.5])$.
- If $x < 0$ then $y \sim \mathsf{Unif}([0.5, 1.5])$.

# Periodicity: Example *cont.*

The Markov chain is clearly periodic (and irreducible):

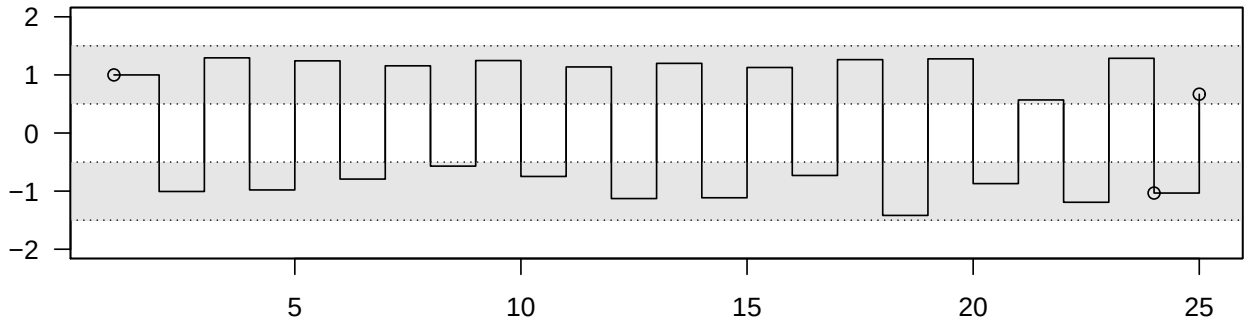



Law of large numbers still "works" (since the Markov chain is irreducible):



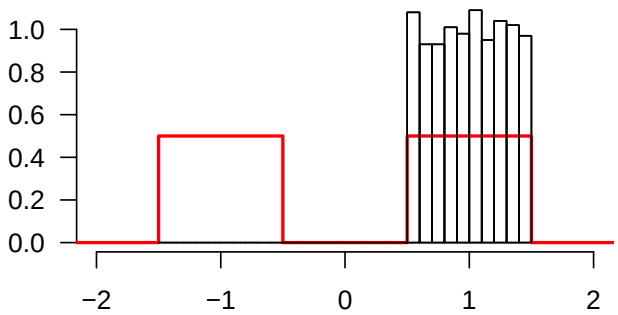
# Periodicity: Example *cont.*

The Markov chain is clearly periodic (and irreducible):

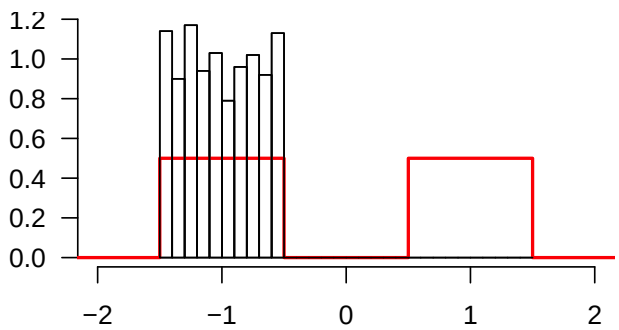



**However**, the chain clearly does *not* converge:

1000 replicates of $x^{(25)}$ when $x^{(1)} = 1$




1000 replicates of $x^{(24)}$ when $x^{(1)} = 1$



# Markov chain convergence theorem

## Theorem: Markov chain convergence theorem

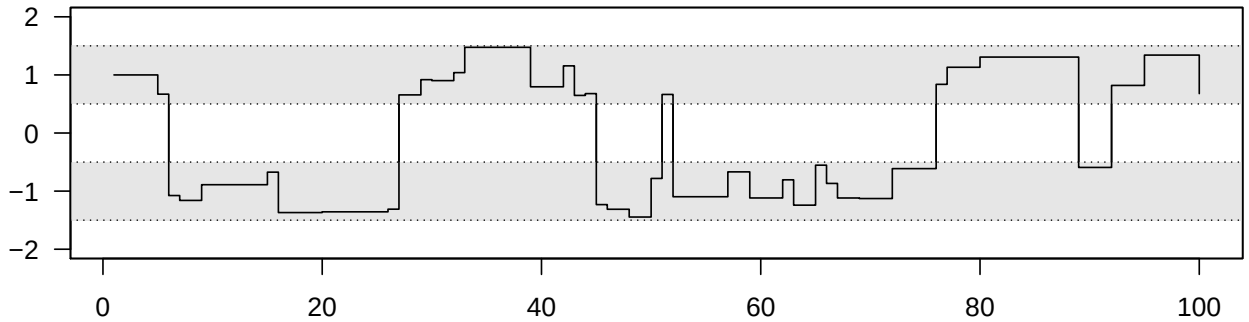
For an irreducible and aperiodic Markov chain with invariant distribution $\pi(x)$, there exists $C \subseteq \Omega$, so that $\Pi(C) = 1$ and for all $x \in C$ and $A \subseteq \Omega$

$$P(X^{(t)} \in A | X^{(0)}) \to \Pi(A) \quad \text{as } t \to \infty$$

If the Markov chain is Harris recurrent, then $C = \Omega$.

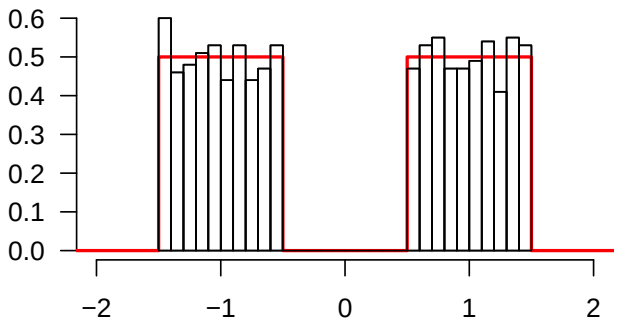
# Convergence: Example

Use the irreducible and aperiodic chain from earlier:

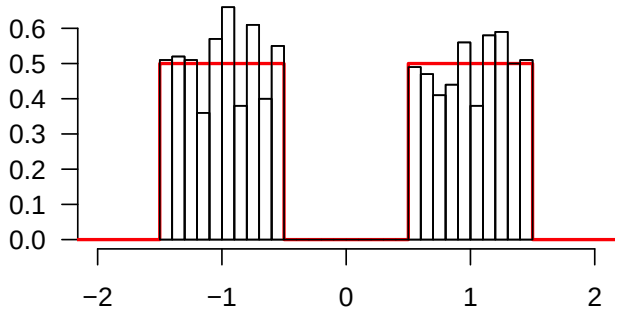



The chain clearly *does* converge:

1000 replicates of $x^{(25)}$ when $x^{(1)} = 1$




1000 replicates of $x^{(24)}$ when $x^{(1)} = 1$



# Convergence: Example *cont.*

1000 replicates of $x^{(i)}$ when $x^{(1)} = 1$ for $i = 1, \ldots, 12$ ($\delta = 2$):