Implementation

Approximate Bayesian Computation (ABC)

Conchi Ausín and Mike Wiper
Department of Statistics
Universidad Carlos III de Madrid

Master in Business Administration and Quantitative Methods
Master in Mathematical Engineering

Objective

How can we carry out Bayesian inference without the likelihood function?

The problem

- The likelihood function is a fundamental part of Bayes theorem.

$$f(\theta|\mathbf{x}) \propto l(\theta|\mathbf{x})f(\theta)$$

The problem

- The likelihood function is a fundamental part of Bayes theorem.

$$f(\theta|\mathbf{x}) \propto l(\theta|\mathbf{x})f(\theta)$$

- We need this even to do MCMC.

The problem

- The likelihood function is a fundamental part of Bayes theorem.

$$f(\theta|\mathbf{x}) \propto l(\theta|\mathbf{x})f(\theta)$$

- We need this even to do MCMC.
- In many modern problems, we cannot even write down the likelihood function.

The problem

- The likelihood function is a fundamental part of Bayes theorem.

$$f(\theta|\mathbf{x}) \propto l(\theta|\mathbf{x})f(\theta)$$

- We need this even to do MCMC.
- In many modern problems, we cannot even write down the likelihood function.
- Do we have to give up?

The problem

- The likelihood function is a fundamental part of Bayes theorem.

$$f(\theta|\mathbf{x}) \propto l(\theta|\mathbf{x})f(\theta)$$

- We need this even to do MCMC.
- In many modern problems, we cannot even write down the likelihood function.
- Do we have to give up?
- No! We can try likelihood free methods?

The idea: discrete data

Suppose that we have a discrete variable X and parameter θ. Assume we don't know $f(\mathbf{x}|\theta)$, but that it is easy to sample.

How can we sample from $f(\theta|\mathbf{x})$?

The idea: discrete data

Suppose that we have a discrete variable X and parameter θ. Assume we don't know $f(\mathbf{x}|\theta)$, but that it is easy to sample.

How can we sample from $f(\theta|\mathbf{x})$?

Repeat:

- Sample θ from the prior distribution, $f(\theta)$
- Sample $\mathbf{y} \sim f(\cdot|\theta)$.

Then the pairs $(\mathbf{y}, \theta)$ are a sample from the joint distribution.

The idea: discrete data

Suppose that we have a discrete variable X and parameter θ. Assume we don't know $f(\mathbf{x}|\theta)$, but that it is easy to sample.

How can we sample from $f(\theta|\mathbf{x})$?

Repeat:

- Sample θ from the prior distribution, $f(\theta)$
- Sample $\mathbf{y} \sim f(\cdot|\theta)$.

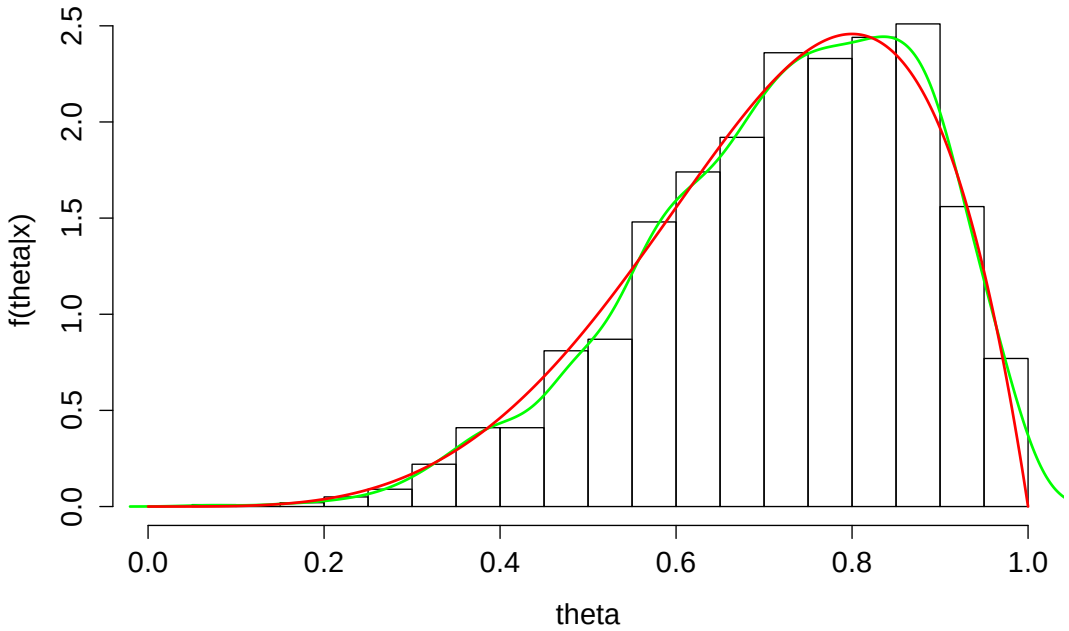
Then the pairs $(\mathbf{y}, \theta)$ are a sample from the joint distribution.

Now, reject all sampled pairs such that $\mathbf{y} \neq \mathbf{x}$. Then, the values of θ that remain come from the posterior distribution $f(\theta|\mathbf{x})$.

Example

Consider Bernoulli trials, $P(X = 1|\theta) = \theta$ with a uniform prior $\theta \sim U(0, 1)$.

Suppose we observe data $\mathbf{x} = (1, 1, 0, 1, 1)^T$. Then, we know theoretically that $\theta|\mathbf{x} \sim Beta(5, 2)$. Let's do this using ABC.



Computational time 5s

Sufficient statistics

A problem with the previous approach is that if the sample size is large, it may take a very large number of iterations to generate artificial samples ($\mathbf{y} = \mathbf{x}$). How can we avoid this?

Sufficient statistics

A problem with the previous approach is that if the sample size is large, it may take a very large number of iterations to generate artificial samples ($\mathbf{y} = \mathbf{x}$). How can we avoid this?

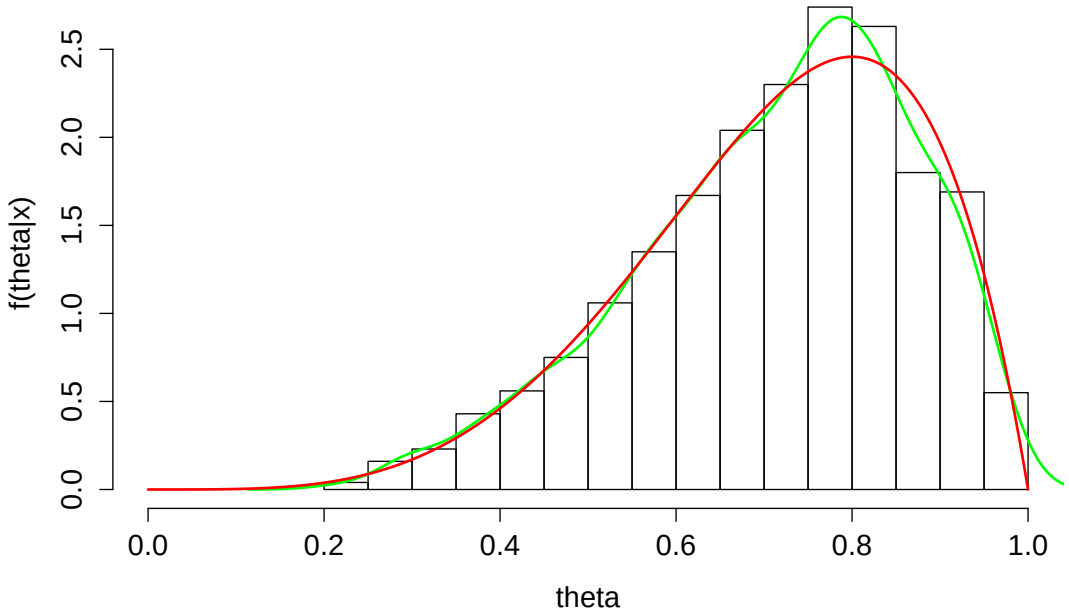
Remember that if a sufficient statistic exists, then $f(\theta|s(\mathbf{x})) = f(\theta|\mathbf{x})$. Therefore, we can use the same idea as previously, but accepting samples $\mathbf{y}$ such that $s(\mathbf{y}) = s(\mathbf{x})$.

Sufficient statistics

A problem with the previous approach is that if the sample size is large, it may take a very large number of iterations to generate artificial samples ($\mathbf{y} = \mathbf{x}$). How can we avoid this?

Remember that if a sufficient statistic exists, then $f(\theta|s(\mathbf{x})) = f(\theta|\mathbf{x})$. Therefore, we can use the same idea as previously, but accepting samples $\mathbf{y}$ such that $s(\mathbf{y}) = s(\mathbf{x})$.

In our example, consider $\sum_{i=1}^{5} x_i = 3$ and accept samples such that $\sum_{i=1}^{5} y_i = 3$.



Computational time 0.1s

Continuous data

When $\mathbf{X}$ is continuous, for any θ, the time taken to generate $\mathbf{y} = \mathbf{x}$ (or $s(\mathbf{y}) = s(\mathbf{x})$) will be infinite. What can we do?

Continuous data

When $\mathbf{X}$ is continuous, for any θ, the time taken to generate $\mathbf{y} = \mathbf{x}$ (or $s(\mathbf{y}) = s(\mathbf{x})$) will be infinite. What can we do?

Accept samples not too far away from $\mathbf{x}$.

Continuous data

When $\mathbf{X}$ is continuous, for any θ, the time taken to generate $\mathbf{y} = \mathbf{x}$ (or $s(\mathbf{y}) = s(\mathbf{x})$) will be infinite. What can we do?

Accept samples not too far away from $\mathbf{x}$.

Defining a distance measure, $||\cdot||$ and a tolerance, ϵ, we accept samples such that $||\mathbf{y} - \mathbf{x}|| < \epsilon$ or $||s(\mathbf{y}) - s(\mathbf{x})|| < \epsilon$.

Continuous data

When $\mathbf{X}$ is continuous, for any θ, the time taken to generate $\mathbf{y} = \mathbf{x}$ (or $s(\mathbf{y}) = s(\mathbf{x})$) will be infinite. What can we do?

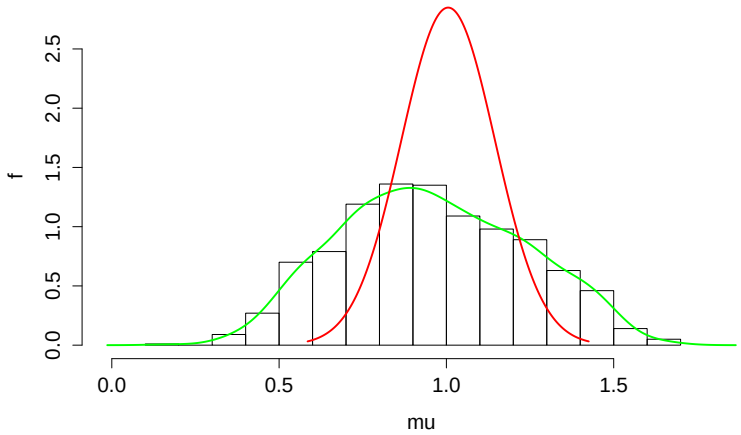
Accept samples not too far away from $\mathbf{x}$.

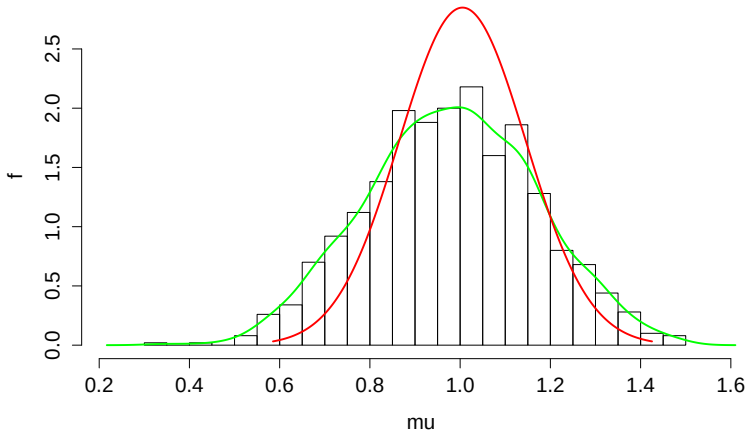
Defining a distance measure, $||\cdot||$ and a tolerance, ϵ, we accept samples such that $||\mathbf{y} - \mathbf{x}|| < \epsilon$ or $||s(\mathbf{y}) - s(\mathbf{x})|| < \epsilon$.

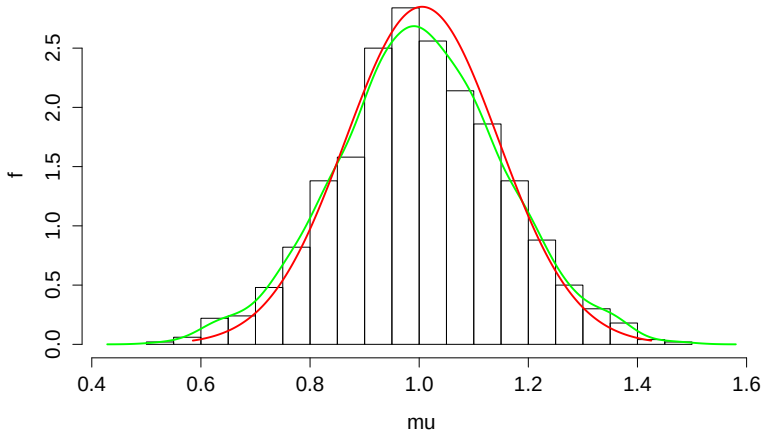
In practice, we don't usually fix ϵ but instead accept just a certain proportion of the sampled values (e.g. 5%, 1%, 0.5%) with the smallest differences from the real data. If the estimated posterior looks much the same, we might consider the approximation to be reasonable.

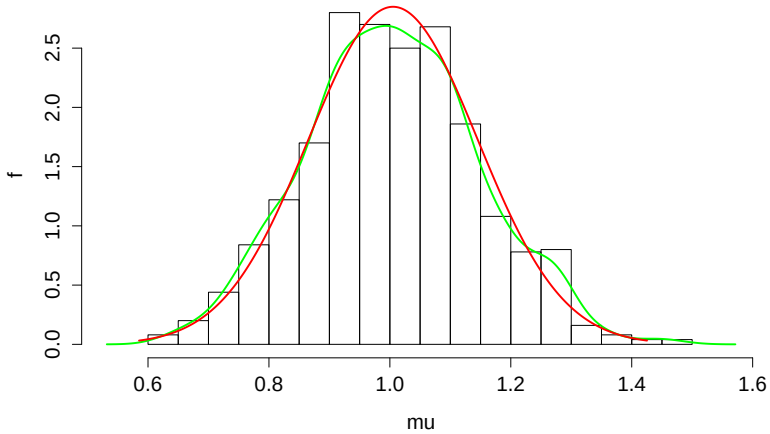
Example

Consider data generated from a normal($\mu = 1$,1) distribution and a $N(0, 1)$ prior for μ. The fitted curves are from an accepted sample of size 1000 generated by accepting 20%, 10%, 5% and 1% of the sampled values.





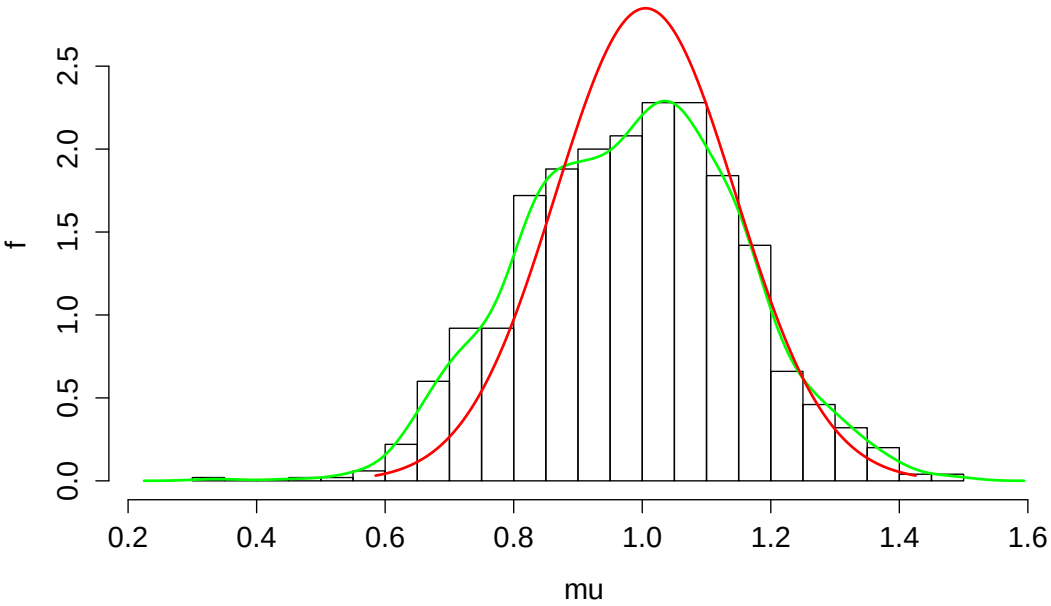




Insufficient statistics

In complex problems, sufficient statistics are not usually available. We generally use statistics that we hope are close to sufficient. As $\epsilon \to 0$, we are then approximating $f(\theta|s(\mathbf{x}))$.

The following is constructed with a sample of 1000 after accepting 0.005% of the sampled values.

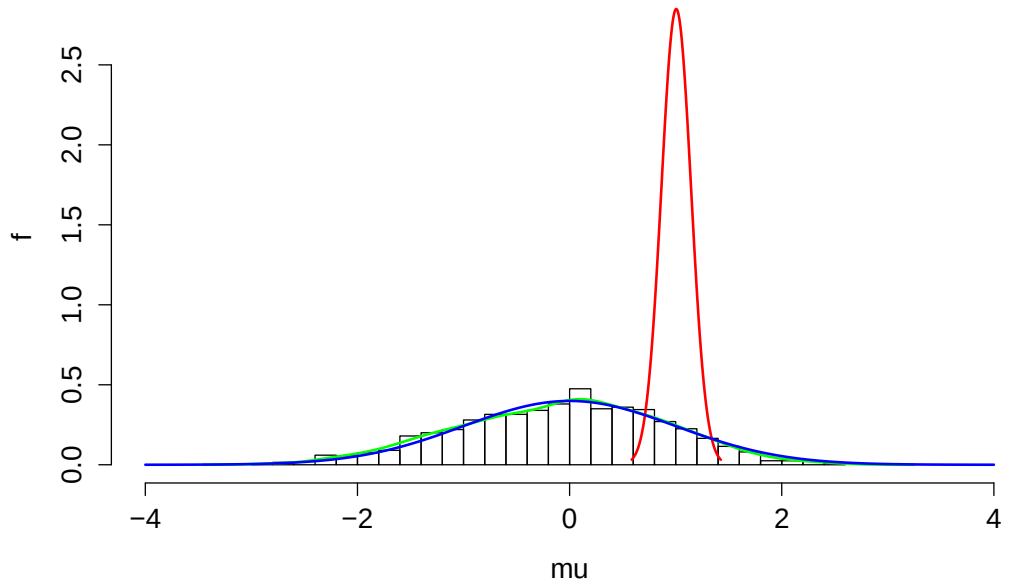


The median works well here as the sample mean $\bar{x} = 1.025$ and median $\tilde{x} = 1.012$ are very similar.

Very insufficient statistics

If the statistic is not useful, the estimated posterior can be very poor.

Suppose we use the sample variance instead of the sample median.

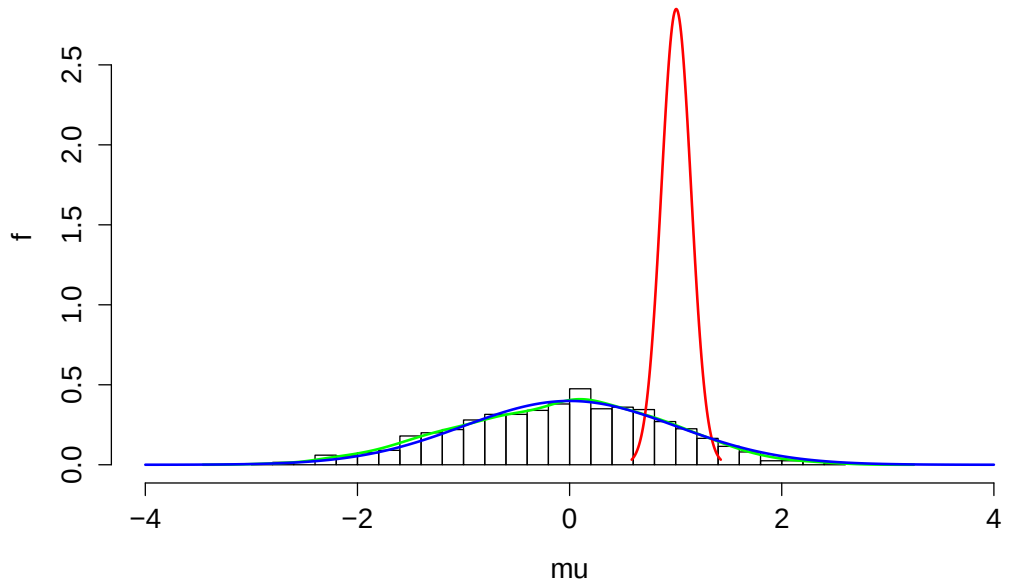


Even accepting 0.001% of the samples, the estimated curve approaches the prior. Why?

Very insufficient statistics

If the statistic is not useful, the estimated posterior can be very poor.

Suppose we use the sample variance instead of the sample median.



Even accepting 0.001% of the samples, the estimated curve approaches the prior. Why?

The sample variance provides zero information about the mean.

A more complicated problem

- We have an $M(\lambda)/U(\beta_1, \beta_2)/1$ queueing system.

A more complicated problem

- We have an $M(\lambda)/U(\beta_1, \beta_2)/1$ queueing system.
- We can do inference if we observe both interarrival and service times but here, we only observe departure times.

A more complicated problem

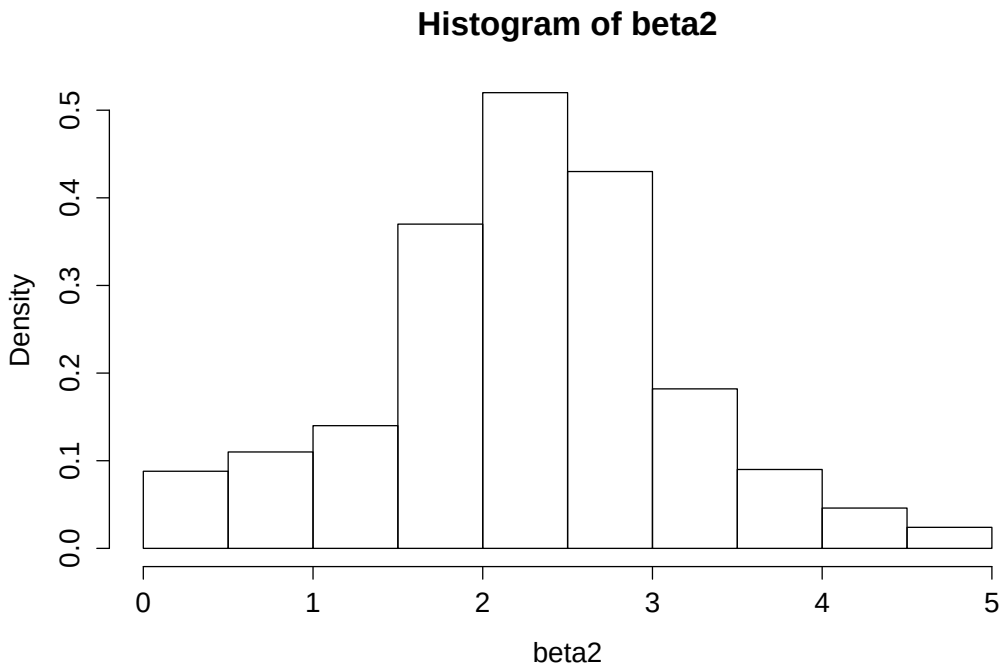
- We have an $M(\lambda)/U(\beta_1, \beta_2)/1$ queueing system.
- We can do inference if we observe both interarrival and service times but here, we only observe departure times.
- The likelihood is intractable.

A more complicated problem

- We have an $M(\lambda)/U(\beta_1, \beta_2)/1$ queueing system.
- We can do inference if we observe both interarrival and service times but here, we only observe departure times.
- The likelihood is intractable.
- It is easy to simulate departure times though:
  1. For $i = 1, ..., n$, simulate $u_i \sim Ex(\lambda)$ and $s_i \sim U(\beta_1, \beta_2)$.
  2. Set $a_i = \sum_{j=1}^{i} u_i$ to be the arrival times.
  3. Define the departure times ($d_0 = 0$) and $d_i = s_i + \max\{a_i, d_{i-1}\}$

Example

- Generate 5 departure times from $M(1)/U(0,3)/1$.
- Suppose we fix $\lambda = 1$ and $\beta_1 = 0$ and assume $\beta_2 \sim U(0,5)$.
- Simulate values of β_2 from the prior and use Euclidean distance between the observed and simulated departure times for selection.
- Accept the smallest 1% of values from a sample of size 1000000.



Problems and extensions

- If the prior is diffuse, we will reject lots of values and inference will be very inefficient.
  - MCMC ABC. Ideas to use proposal distributions instead of the prior to generate candidate values. These can be accepted using e.g. MCMC steps.
- How to choose sensible statistics?
  - Some work on automatic selections but generally problem specific.
- Correcting for $\epsilon > 0$ by post processing.
  - Idea is to fit a (regression) model to the accepted pairs $(\theta, \mathbf{x})$ and deduce $f(\theta|\mathbf{x})$.
- Easy implementation
  - Two R packages `ABC` and `EasyABC` are available.