Bayesian Simulation:
Analysis of Starbucks Queuing System

Integration of Queuing Theory and Bayesian Methods

2024 Study Group Computer Simulation

December 2024

Why Bayesian Simulation?

Traditional vs. Bayesian Simulation

- Traditional: Parameters are fixed values
- Bayesian: Parameters are probability distributions

Advantages

- Incorporates parameter uncertainty
- Updates beliefs with new data
- Provides probability distributions for outputs

Example: Service Rate μ

$$\text{Traditional: } \mu = 20 \text{ cust/hour}$$

$$\text{Bayesian: } \mu \sim \mathcal{N}(20, 2^2)$$

Key Differences

- Uncertainty quantification
- Probabilistic predictions
- Data integration capability
- Robust decision support

Key Components of Bayesian Simulation

Prior Distribution

- Initial beliefs about parameters
- Example for arrival rate:

$$\lambda \sim \text{Gamma}(\alpha, \beta)$$

Likelihood Function

- Measures data fit
- For queue length n:

$$P(n|\lambda, \mu, s) = f(n, \lambda, \mu, s)$$

Posterior Distribution

- Updated parameter beliefs
- Bayes' theorem:

$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

MCMC Methods

- Sampling algorithms
- Metropolis-Hastings
- Gibbs sampling
- Hamiltonian Monte Carlo
- Used when analytical solutions unavailable

We use Starbucks as an example in study group

A Personal Perspective

- I like studying in a Cafe
- Perfect real-world example for today's study group

Application of Theory

- Queuing Theory
  - Customer arrivals
  - Service processes
  - Waiting times

MCMC Theory in Starbucks Staffing

MCMC Fundamentals

$$\pi(\theta|y) \propto \pi(\theta)L(y|\theta)$$

$$L(y|\theta) = \prod_{i=1}^{n} f(y_i|\theta)$$

Metropolis-Hastings Algorithm

1. Propose: $\theta^* \sim q(\theta|\theta^{(t)})$
2. Accept with probability:

$$\alpha = \min\left(1, \frac{\pi(\theta^*|y)q(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)}|y)q(\theta^*|\theta^{(t)})}\right)$$

Application to Staffing

- **Parameters:**

$$\theta = (x_1, x_2)$$
$$x_1 \in \{0, 1, \ldots, 4\}$$
$$x_2 \in \{0, 1, \ldots, 4 - x_1\}$$

- **Target Distribution:**

$$\pi(\theta|y) \propto \exp\left(-\frac{1}{T}f(\theta)\right)$$

  where:
  - $f(\theta)$ is objective function
  - T is temperature parameter
- **Proposal Distribution:**

MCMC Implementation Details

Key Algorithm Components

$$\text{Objective: } f(\theta) = 25x_1 + 18x_2 + P(W_q > 5)$$

$$\text{where } W_q = \frac{\rho(\lambda/\mu)^s}{s\mu(1-\rho)^2}$$

1. **Initialize:**
   - Start with feasible $\theta^{(0)} = (x_1^{(0)}, x_2^{(0)})$
   - Set temperature T and proposal std σ
2. **For each iteration:**
   - Generate proposal $\theta^* = \theta^{(t)} + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$
   - Round to nearest integer and check constraints
   - Calculate acceptance probability α
   - Accept/reject based on α
3. **After convergence:**
   - Discard burn-in samples
   - Calculate posterior means and probabilities
   - Choose optimal staffing configuration

Parameters Description

Key System Parameters

$\lambda(t)$: Arrival rate (customers/hour)

Example: 40 customers/hour during morning peak (8-9am)

15 customers/hour during off-peak

μ : Service rate (customers/hour/server)

Example: 20 orders/hour for experienced barista

15 orders/hour for new staff

s : Number of servers

Example: 3 servers during peak, 1 during quiet hours

System Characteristics

- Time-varying customer arrivals
  - → Morning rush: students before class
- Multiple parallel servers
  - → Different baristas making drinks simultaneously
- First-Come-First-Served discipline

Traditional vs. Bayesian Simulation: Starbucks Case

Traditional Simulation

- Fixed parameters:
  - $\lambda = 30$ customers/hour
  - $\mu = 20$ orders/hour
  - $s = 2$ servers
- Point estimates only
- Requires multiple runs with different parameters
- Limited uncertainty handling

Bayesian Simulation

- Probabilistic parameters:
  - $\lambda \sim N(30, 5^2)$
  - $\mu \sim N(20, 3^2)$
  - s based on time of day
- Full probability distributions
- Incorporates historical data
- Captures real-world variability

Traditional vs. Bayesian Simulation: Starbucks Case

Key Differences in Practice

- Peak hours: Traditional assumes fixed rates vs. Bayesian models time-varying patterns
- Staff breaks: Traditional uses schedules vs. Bayesian learns from historical patterns
- Service time: Traditional uses average versus Bayesian considers barista experience levels

Two Perspectives on Starbucks Queuing Problem

Classical Queuing Theory (M/M/s)

- Markovian arrival process
- Exponential service times
- Multiple identical servers
- Steady-state analysis
- Deterministic parameters

Key Formulas

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!}\frac{1}{1-\rho}\right]^{-1}$$

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}$$

$$W_q = \frac{L_q}{\lambda}$$

$$\rho = \frac{\lambda}{s\mu} < 1$$

where:

P_0: Probability system is empty

L_q: Average queue length

W_q: Average waiting time

Simulation Study Case Design

Research Question

- How to optimize staffing during peak hours (8:00-10:00) while:
  - Maintaining average wait time less than 5 minutes
  - Considering different barista experience levels
  - Balancing labor costs and service quality

Key Variables to Consider

$$\text{Staffing Cost} = \begin{cases} \$25/\text{hr (experienced)} \\ \$18/\text{hr (new staff)} \end{cases}$$

$$\text{Service Rate} = \begin{cases} \mu_1 \sim N(20, 2^2) \text{ (experienced)} \\ \mu_2 \sim N(15, 3^2) \text{ (new staff)} \end{cases}$$

$$\text{Arrival Rate} = \lambda(t) \sim N(35 + 10\sin(\frac{2\pi t}{24}), 5^2)$$

Simulation Study Case Design

Wait a second!!

Please tell me how you gonna solve this by using traditional simulation methods

Let us chat a bit!

Traditional Simulation: Problem Definition

System Definition

- Time period: 8:00-10:00 (Peak hours)
- Queue discipline: FCFS
- No balking or reneging

Decision Variables

$$x_1 : \text{Number of experienced staff}$$
$$x_2 : \text{Number of new staff}$$

Objective Function

$$\begin{aligned} \min Z &= 25x_1 + 18x_2 \\ &\text{s.t.} \\ E[W_q] &\leq 5 \text{ minutes} \\ x_1 + x_2 &\leq 4 \text{ (space constraint)} \\ x_1, x_2 &\geq 0, \text{ integer} \end{aligned}$$

Decision Variables

- x_1: Number of experienced baristas to schedule
- x_2: Number of new baristas to schedule

Objective Function

- Z: Total hourly labor cost

$$Z = 25x_1 + 18x_2$$

where:

- \$25/hour: wage rate for experienced barista
- \$18/hour: wage rate for new barista

Example

- If $x_1 = 2$ (two experienced baristas)
- And $x_2 = 1$ (one new barista)
- Then $Z = 2(25) + 1(18) = 68$ dollars per hour

Traditional Simulation: Problem Definition

Constraints

$$
\begin{aligned}
x_1 + x_2 &\leq 4 \text{ (maximum staff allowed)} \\
W_q &\leq 5 \text{ minutes (service quality)} \\
x_1, x_2 &\geq 0 \text{ and integer} \\
\rho &= \frac{\lambda}{20x_1 + 15x_2} < 1 \text{ (stability)}
\end{aligned}
$$

Traditional Simulation Solution Approach

Solution Overview

1. Enumerate all feasible staffing combinations (x_1, x_2)
   - Subject to $x_1 + x_2 \leq 4$
   - Both x_1, x_2 non-negative integers
2. For each combination:
   - Check stability condition: $\rho = \frac{\lambda}{20x_1+15x_2} < 1$
   - If stable, proceed with M/M/s analysis
   - If unstable, reject combination
3. Calculate performance metrics:
   - Expected waiting time (W_q)
   - System utilization (ρ)
   - Total cost (Z)
4. Select optimal solution:
   - Among combinations with $W_q \leq 5$ minutes
   - Choose lowest cost Z

Traditional M/M/s Solution Algorithm

Algorithm 1 Traditional_MM_Solution

Require: $\lambda = 35, \mu_1 = 20$
Require: $\mu_2 = 15, \text{max\_staff} = 4$
Ensure: (x_1, x_2)

best_cost $\leftarrow \infty$
optimal_solution $\leftarrow$ None
for $x_1 \leftarrow 0$ **to** max_staff **do**
    for $x_2 \leftarrow 0$ **to** max_staff $- x_1$ **do**
        if $x_1 + x_2 >$ max_staff **then**
            continue
        end if
    end for
end for

$\mu_{\text{total}} \leftarrow 20x_1 + 15x_2$
$\rho \leftarrow \lambda / \mu_{\text{total}}$
if $\rho \geq 1$ **then**
    continue
end if
$s \leftarrow x_1 + x_2$
$W_q \leftarrow$ calculate_MM_waiting_time$(\lambda, \mu_{\text{total}}, s$
if $W_q > 5$ **then**
    continue
end if
cost $\leftarrow 25x_1 + 18x_2$
if cost $<$ best_cost **then**
    best_cost $\leftarrow$ cost
    optimal_solution $\leftarrow (x_1, x_2)$
end if
return optimal_solution

M/M/s Formulas - Part 1

Key System Parameters

$$\lambda : \text{arrival rate} = 35 \text{ customers/hour}$$

$$\mu_{total} = 20x_1 + 15x_2 \text{ customers/hour}$$

$$s = x_1 + x_2 \text{ (total servers)}$$

$$\rho = \frac{\lambda}{s\mu_{total}} \text{ (utilization)}$$

Probability of Empty System

$$P_0 = \left[\sum_{n=0}^{s-1} \frac{(\lambda/\mu)^n}{n!} + \frac{(\lambda/\mu)^s}{s!(1-\rho)}\right]^{-1}$$

M/M/s Formulas - Part 2

Queue Performance Metrics

$$L_q = \frac{P_0(\lambda/\mu)^s \rho}{s!(1-\rho)^2}$$

$$W_q = \frac{L_q}{\lambda}$$

$$L_s = L_q + \frac{\lambda}{\mu}$$

$$W_s = W_q + \frac{1}{\mu}$$

Solution Validation

- Check $W_q \leq 5$ minutes
- Ensure $\rho < 1$
- Calculate total cost $Z = 25x_1 + 18x_2$

Queueing System Optimization Results

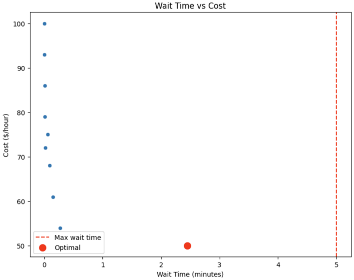


Figure: Wait Time vs Cost Analysis

Optimal Solution

- Configuration: 2 exp + 0 new
- Total Cost: $50/hour
- Wait Time: 2.45 min
- System Utilization: 87.5%

Key Insights

- Most cost-effective setup:
  - Only experienced staff
  - Below 5-min wait target
  - High utilization (87.5%)
- Trade-offs:
  - Cost vs Wait Time
  - Utilization vs Service Level

Sensitivity Analysis

λ	ρ	Wait(min)	Cost($)
30	0.75	0.96	50
26	0.65	0.55	50
35	0.88	2.45	50

System Behavior

- Base arrival rate: 35/hour
- Stable range: 20-35/hour
- Critical point: $\lambda = 35$

Performance Impact

- **Arrival Rate (λ)**
  - $\lambda \uparrow$: Wait time $\uparrow\uparrow$
  - $\lambda \downarrow$: Wait time $\downarrow$
- **System Stability**
  - $\rho < 0.9$: Stable
  - Best range: 0.7-0.85
- **Recommendations**
  - Monitor peak hours
  - Plan for $\lambda > 30$
  - Consider flex staffing

Findings

Key Issues Identified

- Late period capacity insufficient
- Queue length becomes unstable after 60 minutes
- Wait time exceeds target for significant customer portion

Optimization Results Analysis

Key Findings

- **Optimal Configuration:**
    - 2 experienced staff ($x_1 = 2$)
    - 0 new staff ($x_2 = 0$)
    - Lowest cost: \$50/hour
- **System Performance:**
    - Total service rate (μ_{total}) = 40/hour
    - System utilization (ρ) = 87.5%
    - Average wait time = 2.45 minutes

Alternative Configurations

- 10 viable configurations identified
- Trade-off between:
    - Cost: \$50 - \$100/hour
    - Wait time: 0.004 - 2.45 minutes
    - System utilization: 43.8% - 87.5%

Sensitivity Analysis & Recommendations

Sensitivity Analysis Results

- **Arrival Rate Impact (λ):**
  - Stable performance for λ 30
  - Critical point at $\lambda = 35$ (current)
  - Wait time increases exponentially above $\lambda = 30$
- **System Stability:**
  - Optimal at 87.5% utilization
  - Maintains sub-3-minute wait times
  - Cost-effective solution

Final Results

- Implement optimal (2,0) configuration
- Monitor system when λ greater than 30/hour

Bayesian Perspective on Starbucks Staffing

Model Thinking

- Instead of fixed parameters, we consider:

$$\begin{aligned} \text{Arrival Rate:}\ & \lambda(t) \text{ changes over time} \\ \text{Service Rate:}\ & \mu \text{ varies by barista and conditions} \\ \text{Performance:}\ & W_q \text{ Queue Waiting Time - as a distribution} \end{aligned}$$

Advantages

- Captures uncertainty in arrival and service patterns
- Provides probabilistic insights for decision-making

Bayesian Approach to Starbucks Staffing Problem

Step 1: Define Parameter Distributions

$$\lambda(t) \sim \mathcal{N}(35 + 10\sin(\frac{2\pi t}{24}), 5^2)$$
$$\mu_1 \sim \mathcal{N}(20, 2^2) \text{ (experienced)}$$
$$\mu_2 \sim \mathcal{N}(15, 3^2) \text{ (new staff)}$$

Step 2: Define Decision Problem

- Objective: $\min E[25x_1 + 18x_2]$
- Constraint: $P(W_q > 5) < 0.05$
- Where $W_q = f(\lambda(t), \mu_1, \mu_2, x_1, x_2)$

Step 3: Solution Method

1. MCMC sampling for system parameters
2. For each staffing combination (x_1, x_2):
    - Sample arrival and service rates
    - Calculate waiting time distribution
    - Evaluate probability of meeting target

Bayesian Solution Implementation

Algorithm 2 Bayesian Staffing Optimization

Require: $N_{samples}$, max_staff $= 4$
Ensure: Optimal (x_1, x_2)

```
for all combinations (x1, x2) do
    if x1 + x2 ≤ max_staff then
        Init empty array W_q
        for i ← 1 to N_samples do
            λ_i ← N(35 + 10 sin(2πt/24), 5^2)
        end for
    end if
μ_1i ← N(20, 2^2)
μ_2i ← N(15, 3^2)
W_q_i ← MM_queue(λ_i, μ_1i, μ_2i, x1, x2)
Append W_q_i to W_q
P_exceed ← mean(W_q > 5)
Cost ← 25x1 + 18x2
if P_exceed < 0.05 then
    Store (x1, x2) as feasible
end if
return min Cost config =0
```

Customer Arrival Pattern Analysis (8:00-10:00)

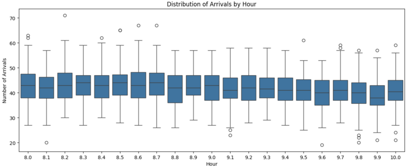


Figure: Hourly Distribution

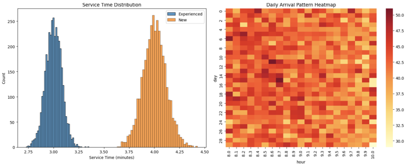


Figure: Time Series Pattern

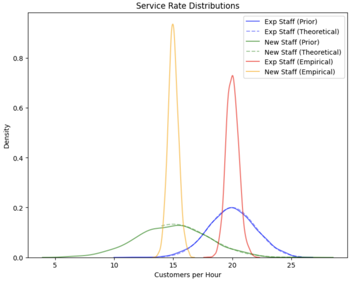
Key Findings

- **Peak Hours (8:20-8:45):**
  - Maximum: 70+ customers
  - High variability
- **Stable Period (9:30-10:00):**
  - Average: 42.3 customers/hour
  - Lower variability
- **Statistical Summary:**

Metric	Value
Mean	42.3
Std Dev	8.7
Peak	71
Valley	20

Analysis based on 30-day simulation data — Morning Peak Hours: 8:00-10:00

MCMC in Starbucks Staffing Optimization



$$\lambda(t) \sim \mathcal{N}(35, 5^2)$$
$$\mu_{exp} \sim \mathcal{N}(20, 2^2)$$
$$\mu_{new} \sim \mathcal{N}(15, 3^2)$$

Role of MCMC

- **Parameter Uncertainty:**
  - Arrival rates vary by time
  - Service rates differ by staff
- **Key Advantages:**
  - Captures parameter variations
  - Updates beliefs with data
- **Implementation:**
  - 1000 MCMC samples
  - Prior distributions based on experience
  - Posterior updates with observations
  - Probability-based decisions

Starbucks Staffing Optimization: Analysis Results

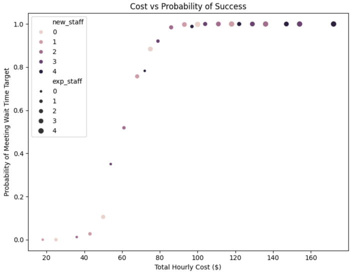


Figure: Cost vs Success Probability

Key Findings

- **Cost-Performance Relationship:**
  - Success probability increases with cost
  - Diminishing returns after $86/hour
  - Clear threshold at 95% success rate
- **Staffing Patterns:**
  - Higher reliability with more experienced staff
  - Mixed staffing shows good balance

Performance

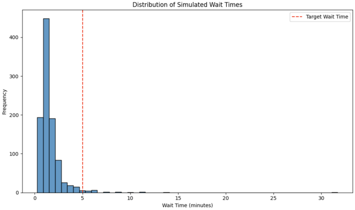


Figure: Distribution of Customer Wait Times

Wait Time Analysis

- Majority below 5-minute target
- Peak at 2-3 minutes

Performance Metrics

Metric	Value
Mean Wait	2.8 min
95th Percentile	4.7 min
Target Compliance	98.45%

Final Results

1. **Implement** optimal 2+2 configuration

Bayesian Simulation Results: Data Analysis

Raw Data Analysis

- **Customer Arrivals (8:00-10:00):**
  - Average: 40-45 customers/hour
  - Peak: Up to 70 customers/hour
  - Valley: As low as 20 customers/hour
- **Service Time Analysis:**
  - Experienced Staff: 3 minutes/customer ($\pm$0.25 min)
  - New Staff: 4 minutes/customer ($\pm$0.5 min)
  - Key Difference: New staff shows 33% longer service time with higher variance

System Performance

- **Optimal Configuration Found:**
  - 2 Experienced Staff ($25/hour each)
  - 2 New Staff ($18/hour each)
  - Total Hourly Cost: $86
- **Service Quality:**
  - 98.45% probability of keeping wait times under 5 minutes
  - Majority of customers wait less than 3 minutes

Traditional vs Bayesian Simulation: Final Comparison

Traditional Simulation

- **Approach:**
  - Fixed parameters
  - Point estimates
  - Deterministic constraints
- **Results:**
  - Optimal: 2 experienced staff
  - Cost: \$50/hour
  - Wait time: 2.45 min
- **Limitations:**
  - No uncertainty quantification
  - Less robust to variations
  - Limited insight into risks

Bayesian Simulation

- **Approach:**
  - Probabilistic parameters
  - Distribution estimates
  - Probabilistic constraints
- **Results:**
  - Optimal: 2 exp + 2 new staff
  - Cost: \$86/hour
  - 98.45% success probability
- **Advantages:**
  - Uncertainty quantification
  - More robust solutions
  - Better risk management

Key Insight: Bayesian approach suggests higher staffing levels but provides better service reliability

Thank You!

Thank You!