Stochastic Models: Parameter Estimation

Lecture 4 – Estimation for Poisson, M/M/1, and M/G/1

Sebastian Müller

Lecture 4

What We Cover Today

- Poisson process: likelihood with right-censoring, rate MLE, Fisher information, asymptotic CI.
- M/M/1: estimating λ and μ under common observation schemes; plug-in performance.
- M/G/1: estimating λ and service moments; Pollaczek–Khinchine plug-in; uncertainty.
- Practical issues: sufficiency, censoring at the observation horizon, near-critical sensitivity.

Setup and Inter-arrival Times

Consider a homogeneous Poisson process of rate $\lambda > 0$ observed on $[0, T]$. Let $N(T)$ be the number of events by time T, with event times $0 < t_1 < \cdots < t_{N(T)} \leq T$.

Define inter-arrivals:

$$T_1 = t_1, \qquad T_i = t_i - t_{i-1}, \ i = 2, \ldots, n.$$

The last interval is right-censored at T:

$$T_{n+1} > T - t_n.$$

For $i = 1, \ldots, n$, $T_i \sim \mathrm{Exp}(\lambda)$ with density $f(T_i \mid \lambda) = \lambda e^{-\lambda T_i}$.
The censored contribution is $\mathbb{P}(T_{n+1} > T - t_n \mid \lambda) = e^{-\lambda(T - t_n)}$.

Likelihood and Sufficiency

The likelihood is

$$L(\lambda \mid \text{data}) = \prod_{i=1}^{n} \left(\lambda e^{-\lambda T_i}\right) \cdot e^{-\lambda(T-t_n)} = \lambda^n e^{-\lambda t_n}\, e^{-\lambda(T-t_n)}.$$

We get the simplified likelihood

$$L(\lambda \mid n, T) = \lambda^n e^{-\lambda T}.$$

Note

The likelihood depends only on (n, T), not on the exact event times t_i: $N(T)$ is a sufficient statistic for λ.

MLE and Properties

Log-likelihood: $\ell(\lambda) = n \log \lambda - \lambda T$ (concave for $\lambda > 0$).

$$\frac{\partial \ell}{\partial \lambda} = \frac{n}{\lambda} - T = 0 \;\Rightarrow\; \hat{\lambda} = \frac{n}{T} = \frac{N(T)}{T}.$$

- Unbiased: $\mathbb{E}[\hat{\lambda}] = \lambda$.
- Efficient: attains Cramér–Rao; no unbiased estimator has smaller variance.
- Consistent: $\hat{\lambda} \xrightarrow{p} \lambda$ as $T \to \infty$.

Fisher Information and CI

Score: $S(\lambda) = \partial \ell / \partial \lambda = n/\lambda - T$. Hessian: $H(\lambda) = -n/\lambda^2$.

$$I(\lambda) = -\mathbb{E}\big[H(\lambda)\big] = \frac{\mathbb{E}[n]}{\lambda^2} = \frac{\lambda T}{\lambda^2} = \frac{T}{\lambda}.$$

Cramér–Rao: $\mathrm{Var}(\hat{\lambda}) \geq 1/I(\lambda) = \lambda/T$. In fact,

$$\mathrm{Var}\left(\frac{N(T)}{T}\right) = \frac{\lambda}{T}.$$

Asymptotic normality: $\sqrt{T}(\hat{\lambda} - \lambda) \Rightarrow \mathcal{N}(0, \lambda)$. 95% CI:

$$\hat{\lambda} \pm 1.96 \sqrt{\frac{\hat{\lambda}}{T}}.$$

Why Event Times Drop Out

- $N(T)$ is sufficient (factorization theorem).
- Conditionally on $N(T) = n$, the order statistics $t_1, \ldots, t_n$ are those of i.i.d. $\mathrm{Unif}(0, T)$ draws.
- Their joint density is $\frac{n!}{T^n}$, independent of λ, and cancels in the likelihood.

Estimating M/M/1 Parameters

Goal: estimate arrival rate λ and service rate μ from observations of an M/M/1 queue over $[0, T]$.

A) Interarrival & service times observed

- $A_1, \ldots, A_{n_A} \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$ (interarrivals)
- $S_1, \ldots, S_{n_S} \overset{\text{iid}}{\sim} \text{Exp}(\mu)$ (service times)

$$\boxed{\hat{\lambda} = \frac{n_A}{\sum A_i}, \quad \hat{\mu} = \frac{n_S}{\sum S_i}}$$

MLEs from independent exponential samples.

Estimating M/M/1 Parameters

B) Calendar-time observation on $[0, T]$

- $N_A(T)$: number of arrivals
- $C(T)$: number of completions
- $B(T) = \int_0^T \mathbf{1}\{Q(t) \geq 1\}\, dt$: *total busy time*

$$\boxed{\hat{\lambda} = \frac{N_A(T)}{T}, \quad \hat{\mu} = \frac{C(T)}{B(T)}}$$

Note

In B), during busy periods, completions form a Poisson process of rate μ. Thus, $C(T) \sim \text{Poisson}(\mu B(T)) \implies \hat{\mu} = C(T)/B(T)$ is the MLE.

Estimating M/M/1 Parameters

C) Only queue length process $Q(t)$ observed

- Full trajectory or event times $\rightarrow$ CTMC birth-death inference
- Contains all info from B (and more) — sufficient for MLE + model validation

$$\hat{\lambda} = \frac{\text{up transitions}}{T}, \hat{\mu} = \frac{\text{down transitions}}{\text{total busy time}}$$

Note

CTMC likelihood needed to prove that they are MLE, efficient, and optimal.

M/M/1 Estimation: Observation Types — Full Comparison

Aspect	**A: $A_i + S_i$ (paired)**	**B: Aggregates**	**C: $Q(t)$ trajectory**
Data observed	Interarrivals A_i, service times S_i	$N_A(T)$, $C(T)$, $B(T)$	Full $Q(t)$
Can reconstruct $Q(t)$?	*Yes* (FIFO + pairing)	*No*	*Yes* (given)
Data volume & storage	High (per event)	*Low* (3 numbers)	High (per-event)
Privacy risk	High	*Low*	High
Real-world availability	*Rare*	*Common*	Rare
MLE $(\hat{\lambda}, \hat{\mu})$?	Yes	Yes	Yes
Validate Exp. interarrivals?	Yes	Yes (via $N_A(T)$)	Yes
Validate Exp. services in context?	*Yes* (via reconstructed $Q(t)$)	*No*	*Yes*
Test non-Markovianity?	*Yes* (via $Q(t)$)	*No*	*Yes*
Estimate $\rho = \lambda/\mu$?	Yes	Yes	Yes
Statistical power (theory)	*Highest*	Sufficient	*Highest*
Practical utility	*Low*	*High*	Medium

Note

B is the practical winner: minimal data, full MLE, widely available. **A and C are theoretically strongest** — but require rare, high-resolution data.

Sampling Variability and Confidence Intervals

A: Individual $A_i \sim \textbf{Exp}(\lambda)$, $S_i \sim \textbf{Exp}(\mu)$

- n_A interarrivals, n_S services observed
- MLEs (from i.i.d. exponentials):

$$\hat{\lambda} = \frac{n_A}{\sum A_i}, \quad \hat{\mu} = \frac{n_S}{\sum S_i}$$

$$\sqrt{n_A}(\hat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda^2), \quad \sqrt{n_S}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \mu^2)$$

$$\text{95\% CI}: \ \hat{\lambda} \pm 1.96\frac{\hat{\lambda}}{\sqrt{n_A}}, \qquad \hat{\mu} \pm 1.96\frac{\hat{\mu}}{\sqrt{n_S}}$$

Sampling Variability and Confidence Intervals

B: Aggregates on $[0, T]$

- $N_A(T)$, $C(T)$, $B(T) =$ busy time
- Asymptotic variance (from Poisson/CTMC):

$$\boxed{\mathrm{Var}(\hat{\lambda}) \approx \frac{\lambda}{T}, \qquad \mathrm{Var}(\hat{\mu}) \approx \frac{\mu}{B(T)}}$$

$$\boxed{95\%\ \mathrm{CI}: \ \hat{\lambda} \pm 1.96\sqrt{\frac{\hat{\lambda}}{T}}, \qquad \hat{\mu} \pm 1.96\sqrt{\frac{\hat{\mu}}{B(T)}}}$$

Sampling Variability and CIs

C: Full $Q(t)$ trajectory

- Extract $N_A(T)$, $C(T)$, $B(T) \rightarrow$ **same as B**
- **Asymptotic** CIs (identical to B)

$$\hat{\lambda} \pm 1.96\sqrt{\frac{\hat{\lambda}}{T}}, \quad \hat{\mu} \pm 1.96\sqrt{\frac{\hat{\mu}}{B(T)}}$$

Note

A: Exact. **B & C**: Asymptotic (large T, $\rho < 1$), but enables *model validation*.

Plug-in Performance: Estimating $L = \frac{\rho}{1-\rho}$

M/M/1 Performance Metrics

$$L = \frac{\rho}{1-\rho}, \quad L_q = \frac{\rho^2}{1-\rho}, \quad W = \frac{1}{\mu-\lambda}, \quad W_q = \frac{\lambda}{\mu(\mu-\lambda)}$$

with $\rho = \lambda/\mu < 1$.

Plug-in Estimator

Replace $(\lambda, \mu) \to (\hat{\lambda}, \hat{\mu})$:

$$\hat{L} = \frac{\hat{\rho}}{1-\hat{\rho}}, \quad \hat{\rho} = \frac{\hat{\lambda}}{\hat{\mu}}$$

Delta Method for Uncertainty in $\hat{L}$

Asymptotic Variance

Let $g(\lambda, \mu) = \dfrac{\lambda/\mu}{1 - \lambda/\mu} = \dfrac{\lambda}{\mu - \lambda}$. Using the gradient

$$\nabla g = \left(\frac{\partial g}{\partial \lambda}, \frac{\partial g}{\partial \mu} \right) = \left(\frac{\mu}{(\mu - \lambda)^2}, \frac{-\lambda}{(\mu - \lambda)^2} \right).$$

the variances $\mathrm{Var}(\hat{\lambda}) \approx \lambda/T$, $\mathrm{Var}(\hat{\mu}) \approx \mu/B(T)$ (assuming independence), the Delta method gives:

$$\boxed{\mathrm{Var}(\hat{L}) \approx \frac{1}{(\mu - \lambda)^4} \left[\frac{\mu^2 \lambda}{T} + \frac{\lambda^2 \mu}{B(T)} \right]}$$

Replace unknowns by hats for a plug-in estimate. Near $\rho \to 1$, the factor $\left(\frac{\mu}{\mu - \lambda}\right)^4$ inflates uncertainty.

$$\boxed{\text{95\% CI (delta)} : \ \hat{L} \pm 1.96 \sqrt{\mathrm{Var}(\hat{L})}}$$

Plug-in Performance: Numerical Example

Example

$T = 1000$, $N_A = 800$, $C = 790$, $B(T) = 900$ $\hat{\lambda} = 0.8$,
$\hat{\mu} = 790/900 \approx 0.878$, $\hat{\rho} \approx 0.911$ $\hat{L} \approx 10.24$
Delta variance with ∇g above and $\mathrm{Var}(\hat{\lambda}), \mathrm{Var}(\hat{\mu})$ from B):

$$\widehat{\mathrm{Var}}(\hat{L}) \approx \left(\frac{\hat{\mu}}{\hat{\mu} - \hat{\lambda}}\right)^4 \left(\frac{\hat{\lambda}}{\hat{\mu}^2 T} + \frac{\hat{\lambda}^2}{\hat{\mu}\, B(T)}\right).$$

Numerically, this yields a *large* standard error (SE ≈ 5.8) and a wide CI, reflecting near-criticality ($\hat{\rho} \approx 0.91$).

Note

Near criticality ($\hat{\rho} \to 1^-$): $\hat{L} \to \infty$, variance explodes. Always check $\hat{\rho} < 1$ and report uncertainty!

M/G/1: Model

M/G/1 Queue

- Arrivals: Poisson process rate λ
- Service: i.i.d. with distribution G, mean $m_1 = \mathbb{E}[S]$, second moment $m_2 = \mathbb{E}[S^2]$
- Stability: $\rho = \lambda m_1 < 1$

Pollaczek-Khinchine Formula

$$\boxed{\mathbb{E}[W_q] = \frac{\lambda m_2}{2(1-\rho)} = \frac{\lambda m_2}{2(1-\lambda m_1)}}$$

M/G/1: Plug-in Estimators

Data on $[0, T]$

- $N_A(T)$: number of arrivals
- $S_1, \ldots, S_{n_S}$: *observed service times* (assume all completed)

Plug-in Estimators

$$\hat{\lambda} = \frac{N_A(T)}{T}, \quad \hat{m}_1 = \overline{S} = \frac{1}{n_S} \sum S_i,$$
$$\hat{m}_2 = \overline{S^2} = \frac{1}{n_S} \sum S_i^2, \quad \hat{\rho} = \hat{\lambda} \cdot \hat{m}_1$$

$$\hat{W}_q^{\mathrm{PK}} = \frac{\hat{\lambda}\, \hat{m}_2}{2\left(1 - \hat{\rho}\right)}$$

Uncertainty Propagation: Delta Method

Let $g(\lambda, m_1, m_2) = \frac{\lambda m_2}{2(1-\lambda m_1)}$.

Gradient

$$\nabla g^T = \left(\frac{m_2}{2(1-\rho)^2}, \frac{\lambda^2 m_2}{2(1-\rho)^2}, \frac{\lambda}{2(1-\rho)} \right)$$

Asymptotic Covariance of Estimators

$$\text{Cov} \begin{pmatrix} \hat{\lambda} \\ \hat{m}_1 \\ \hat{m}_2 \end{pmatrix} \approx \begin{pmatrix} \frac{\lambda}{T} & 0 & 0 \\ 0 & \frac{\text{Var}(S)}{n_S} & \frac{\text{Cov}(S, S^2)}{n_S} \\ 0 & \frac{\text{Cov}(S, S^2)}{n_S} & \frac{\text{Var}(S^2)}{n_S} \end{pmatrix}$$

Delta Method Variance

$$\boxed{\text{Var}\left(\hat{W}_q^{\text{PK}}\right) \approx (\nabla g)^\top \hat{\Sigma} (\nabla g)}$$

Bootstrap: Input vs System

Caution

Do not naively resample inter-arrivals A_i and services S_i *independently* to bootstrap queue performance. Queueing *creates dependence* via the workload process; you must simulate dynamics or use regenerative structure.

Two Valid Paths

- **Input bootstrap (PK plug-in)**: Resample service times S_i to propagate uncertainty in (m_1, m_2); treat λ parametrically (e.g., Poisson). This yields a CI for the *PK plug-in* $\hat{W}_q^{\mathrm{PK}}$, *not* for the true queue output.
- **System bootstrap (regenerative)**: Partition the trajectory into *busy cycles* (idle-to-idle), treat cycles as i.i.d., and resample cycles to form a CI for mean waiting time (ratio-of-means).

Note

Bottom line: Input bootstrap quantifies *parameter* uncertainty in PK; for *system* uncertainty, use regenerative or block bootstrap and preserve dependence.

Input Bootstrap (PK Plug-in): Setup

Target

Estimate uncertainty in the *PK plug-in* $\hat{W}_q^{\mathrm{PK}} = \dfrac{\hat{\lambda}\,\hat{m}_2}{2(1-\hat{\rho})}$ arising from finite samples of the inputs (λ, m_1, m_2).

Data and Assumptions

On $[0, T]$: observe arrivals (count $N_A(T)$) and service samples $S_1, \ldots, S_{n_S}$ (i.i.d.). Assume Poisson arrivals and i.i.d. services; $\rho = \lambda m_1 < 1$.

Notation

$\hat{\lambda} = N_A(T)/T, \quad \hat{m}_1 = \overline{S}, \quad \hat{m}_2 = \overline{S^2}, \quad \hat{\rho} = \hat{\lambda}\,\hat{m}_1.$

Input Bootstrap (PK Plug-In): Algorithm

Basic (service-only) bootstrap

1. Resample $S_1^*, \ldots, S_{n_S}^*$ with replacement from $\{S_i\}$.
2. Compute $\hat{m}_1^*, \hat{m}_2^*$ from the resample; set $\hat{\lambda}^* = \hat{\lambda}$ (fixed exposure).
3. Form $(\hat{W}_q^{\mathrm{PK}})^* = \dfrac{\hat{\lambda}^*\, \hat{m}_2^*}{2\left(1 - \hat{\lambda}^*\, \hat{m}_1^*\right)}$.
4. Repeat B times; take percentile CI of $\{(\hat{W}_q^{\mathrm{PK}})_b^*\}$.

Including arrival uncertainty

Optionally, draw $N_A^*(T) \sim \mathrm{Poisson}(\hat{\lambda} T)$ and set $\hat{\lambda}^* = N_A^*(T)/T$. This treats arrival sampling variability consistently with Poisson exposure.

Remarks

This CI reflects *input* uncertainty only (PK formula). It is **not** a CI for the true queue mean under finite horizon — use regenerative/bootstrap of busy cycles for that.

System Bootstrap (Regenerative): Details

Cycle identification

Detect idle-to-idle cycles. For each cycle k: total waiting $W_k = \sum_{i \in \mathcal{C}_k} W_{q,i}$, number of jobs $N_k = |\mathcal{C}_k|$.

Estimator and CI

Point: $\hat{W}_q = \dfrac{\sum_k W_k}{\sum_k N_k}$. Bootstrap: resample cycles $\mathcal{C}_k$ with replacement to build $\{\hat{W}_q^{*(b)}\}$ and a percentile CI.

Pros/Cons

Pros: Preserves dependence; valid system-level uncertainty. **Cons**: Needs many complete cycles; sensitive near criticality (long cycles).

Parametric Options for Service Distribution

Assume $S \sim$ Gamma(α, β)

$$m_1 = \frac{\alpha}{\beta}, \quad m_2 = \frac{\alpha(\alpha+1)}{\beta^2}$$

MLE:

$$\hat{\alpha}, \hat{\beta} \quad \rightarrow \quad \hat{m}_1 = \frac{\hat{\alpha}}{\hat{\beta}}, \; \hat{m}_2 = \frac{\hat{\alpha}(\hat{\alpha}+1)}{\hat{\beta}^2}$$

Plug into PK formula.

Lognormal, Weibull, etc.

Same idea: estimate parameters $\rightarrow$ compute $\hat{m}_1, \hat{m}_2$.

Numerical Example: M/G/1

Example

$T = 1000$, $N_A(T) = 800$, $n_S = 790$ services observed Sample: $\overline{S} = 1.125$, $\overline{S^2} = 1.406$

$$\hat{\lambda} = 0.8, \ \hat{m}_1 = 1.125, \ \hat{m}_2 = 1.406, \ \hat{\rho} = 0.9$$

$$\widehat{W_q} = \frac{0.8 \cdot 1.406}{2(1 - 0.9)} = \frac{1.1248}{0.2} = 5.624$$

Bootstrap CI (B=1000): $[4.1, 7.8]$ **Delta CI** (approx): $[4.3, 7.0]$

Note

Near $\hat{\rho} = 0.9$, small changes in $\hat{\rho} \rightarrow$ large changes in $\hat{W}_q$. Uncertainty explodes as $\rho \rightarrow 1$.

Case Study: M/G/1 with Uniform Service

Scenario

Customer-support chat where each ticket requires a bounded handling time due to SLA constraints and triage tools.

Model

Arrivals: Poisson(λ). Service: $S \sim \mathrm{Unif}(a, b)$, independent of arrivals.

$$m_1 = \mathbb{E}[S] = \tfrac{a+b}{2}, \quad m_2 = \mathbb{E}[S^2] = \tfrac{a^2+ab+b^2}{3}, \quad \rho = \lambda m_1 < 1$$

PK plug-in for the mean waiting time:

$$\hat{W}_q = \frac{\hat{\lambda}\,\hat{m}_2}{2(1-\hat{\rho})}.$$

Example

$a = 0.5$, $b = 1.5$ minutes $\Rightarrow m_1 = 1.0$, $m_2 = \frac{0.25+0.75+2.25}{3} = 1.0833\overline{3}$; with $\lambda = 0.5$/min, $\rho = 0.5$.

Uniform Case: Dataset and Tasks

Dataset schema (per job)

arrival_time, service_time, start_service_time, completion_time, wait_time, system_time, queue_len_at_arrival

Tasks

- Estimate $\hat{\lambda} = N(T)/T$ from arrivals in $[0, T]$.
- Estimate $\hat{m}_1, \hat{m}_2$ from service_time samples.
- Compute $\hat{W}_q$ and compare to empirical mean wait.
- Build a 95% CI using delta method and via bootstrap; discuss differences.
- Stress test near criticality by increasing λ or b; observe sensitivity.

Key Takeaways

- Poisson rate: $\hat{\lambda} = N(T)/T$; sufficient statistic $N(T)$; exact variance λ/T.
- M/M/1: $\hat{\lambda}$ from counts over time; $\hat{\mu}$ from completions over busy time or from service samples.
- M/G/1: estimate λ and service moments; plug into PK; use delta or bootstrap for CIs.
- Always check stability ($\hat{\rho} < 1$) and report uncertainty, especially near critical regimes.