

MEM6810 Engineering Systems Modeling and Simulation



工程系统建模与仿真

Theory Analysis

Lecture 7: Output Analysis I: Single Model

SHEN Haihui 沈海辉

Sino-US Global Logistics Institute
Shanghai Jiao Tong University

 shenhaihui.github.io/teaching/mem6810f
 shenhaihui@sjtu.edu.cn

Spring 2025 (full-time)



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

董浩云智能制造与服务管理研究院
CY TUNG Institute of Intelligent Manufacturing and Service Management
(中美物流研究院)
(Sino-US Global Logistics Institute)



- 1 Introduction
 - ▶ Types of Simulations
- 2 Point and Interval Estimations
 - ▶ Basics
 - ▶ Specified Precision
 - ▶ Example
- 3 Terminating Simulation
 - ▶ Discrete Outputs
 - ▶ Continuous Outputs
- 4 Steady-State Simulation
 - ▶ Initialization Bias
 - ▶ Intelligent Initialization
 - ▶ Warm-up Period Deletion
 - ▶ Estimation with Multiple Replications
 - ▶ Estimation with Single Replication

- 1 Introduction
 - ▶ Types of Simulations
- 2 Point and Interval Estimations
 - ▶ Basics
 - ▶ Specified Precision
 - ▶ Example
- 3 Terminating Simulation
 - ▶ Discrete Outputs
 - ▶ Continuous Outputs
- 4 Steady-State Simulation
 - ▶ Initialization Bias
 - ▶ Intelligent Initialization
 - ▶ Warm-up Period Deletion
 - ▶ Estimation with Multiple Replications
 - ▶ Estimation with Single Replication



- Output analysis is the examination of data generated by a simulation.
- Output analysis is needed because the output data from a stochastic simulation exhibits random variability.
- Suppose the true performance of the *simulated system* is θ .
 - The result from a set of simulation runs will be an estimator $\hat{\theta}$.
 - The precision of the estimator $\hat{\theta}$ can be measured by the **standard error** of $\hat{\theta}$ (i.e., estimator's standard deviation) or the width of a confidence interval for θ .
- The purpose of the statistical analysis:
 - Estimate the true performance θ .
 - Control the estimation precision.
- Types of simulation with regard to output analysis:
 - terminating vs. nonterminating.



- A **terminating simulation** is one that runs for some well-defined time duration T_E .
 - E is a specified event (or set of events) that stops each simulation run (replication).
 - Simulation starts at time 0 under well-specified initial conditions, and ends at the stopping time T_E .
 - T_E can be either a deterministic or random variable.
- Example: A bank opens at 9 AM (*time 0*) with no customers present and 8 of the 11 tellers working (*initial conditions*), and closes at 5 PM (*time $T_E = 8$ hours*).
 - $E = \{8 \text{ hours of simulated time have elapsed}\}$.
- It actually stops service when the last customer who entered before 5 PM has been served.
 - $E = \{\text{at least 8 hours of simulated time have elapsed and the system is empty}\} \implies T_E$ is a random variable.



- A **nonterminating simulation** is one that runs continuously and without a natural event E to stop the simulation run.
 - Initial conditions are defined by the analyst, but its effect *fades away* as simulation time increases.
 - Stopping time is conceptually infinite, and in practice it is determined by the analyst with certain statistical precision.
- Examples: Production line that runs 24/7, hospital emergency rooms, continuously operating computer networks, etc.
- For a simulation model that is run in a nonterminating way and *has a steady-state (stationary) distribution*:
 - The objective is often to study the long-run, or steady-state, behavior of a system, which is not influenced by the initial conditions.
 - Such nonterminating simulation is also called *steady-state simulation*.

- 1 Introduction
 - ▶ Types of Simulations
- 2 Point and Interval Estimations
 - ▶ Basics
 - ▶ Specified Precision
 - ▶ Example
- 3 Terminating Simulation
 - ▶ Discrete Outputs
 - ▶ Continuous Outputs
- 4 Steady-State Simulation
 - ▶ Initialization Bias
 - ▶ Intelligent Initialization
 - ▶ Warm-up Period Deletion
 - ▶ Estimation with Multiple Replications
 - ▶ Estimation with Single Replication

- Suppose we want to estimate $\theta = \mathbb{E}[X]$ based on the iid sample $\{X_1, \dots, X_n\}$.
- The point estimator (a single random variable) is

$$\hat{\theta} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

- How good is this estimator?
 - Unbiased: $\mathbb{E}[\hat{\theta}] = \theta$.
 - Consistent: $\hat{\theta} \xrightarrow{a.s.} \theta$, as $n \rightarrow \infty$.[†]
- Point estimator says *nothing* about the estimation error for finite sample size n .
 - Small estimation error means high estimation precision.

[†] Assume $\mathbb{E}[|X|] < \infty$, or, a stronger condition, $\text{Var}(X) < \infty$.



- If $X \sim \mathcal{N}(\theta, \sigma^2)$, then

$$\sqrt{n} \left(\frac{\hat{\theta} - \theta}{\sigma} \right) \sim \mathcal{N}(0, 1).$$

- If X follows arbitrary distribution and $\sigma^2 = \text{Var}(X) \in (0, \infty)$, then by Central Limit Theorem,

$$\sqrt{n} \left(\frac{\hat{\theta} - \theta}{\sigma} \right) \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{as } n \rightarrow \infty. \quad (1)$$

- $\sqrt{n} \left(\frac{\hat{\theta} - \theta}{\sigma} \right) \sim \mathcal{N}(0, 1)$ approximately when n is large.
- σ^2 is typically unknown, and we substitute it by the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2.$$



- If $X \sim \mathcal{N}(\theta, \sigma^2)$, then

$$\sqrt{n} \left(\frac{\hat{\theta} - \theta}{S} \right) \sim t_{n-1}, \quad (2)$$

where t_p denotes t distribution with p degrees of freedom.

- If X follows arbitrary distribution and $\sigma^2 = \text{Var}(X) \in (0, \infty)$, then by Equation (1) and the fact that $\frac{\sigma}{S} \xrightarrow{a.s.} 1$ as $n \rightarrow \infty$,

$$\sqrt{n} \left(\frac{\hat{\theta} - \theta}{S} \right) \xrightarrow{d} \mathcal{N}(0, 1), \quad \text{as } n \rightarrow \infty. \quad (3)$$

- $\sqrt{n} \left(\frac{\hat{\theta} - \theta}{S} \right) \sim \mathcal{N}(0, 1)$ approximately when n is large.
- Results (2) and (3) are the basis of the confidence interval estimation for θ .

- If $X \sim \mathcal{N}(\theta, \sigma^2)$, where θ and σ are unknown, then a $1 - \alpha$ confidence interval (CI) for θ is

$$\left[\hat{\theta} - t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}, \hat{\theta} + t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}} \right], \quad (4)$$

where $t_{n-1, 1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of t_{n-1} distribution.

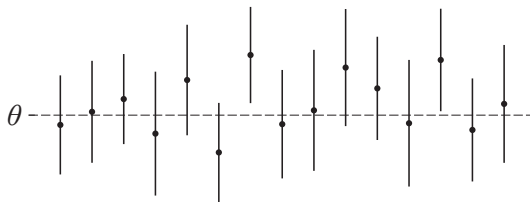
Proof.

$$\begin{aligned} & \mathbb{P} \left\{ \theta \in \left[\hat{\theta} - t_{n-1, 1-\alpha/2} S/\sqrt{n}, \hat{\theta} + t_{n-1, 1-\alpha/2} S/\sqrt{n} \right] \right\} \\ &= \mathbb{P} \left\{ |\theta - \hat{\theta}| \leq t_{n-1, 1-\alpha/2} S/\sqrt{n} \right\} \\ &= \mathbb{P} \left\{ \left| \frac{\theta - \hat{\theta}}{S/\sqrt{n}} \right| \leq t_{n-1, 1-\alpha/2} \right\} = 1 - \alpha, \end{aligned}$$

where the last equality is due to (2) and the symmetry of t distribution.



- The interpretation of CI:
 - If one constructs a very large number of independent $1 - \alpha$ CIs, each based on n observations, the proportion of CIs that actually contain (cover) θ should be $1 - \alpha$.



- **Caution:** There is nothing probabilistic about a single CI **after** the data have been observed and the interval's endpoints have been given numerical values, e.g., $[1.1, 2.4]$.
- Try it out! <http://www.rossmanchance.com/applets/ConfSim.html>

- If X follows arbitrary distribution, θ and $\sigma^2 = \text{Var}(X)$ are unknown, and $0 < \sigma^2 < \infty$, then an approximate $1 - \alpha$ CI for θ with large n is

$$\left[\hat{\theta} - z_{1-\alpha/2} \frac{S}{\sqrt{n}}, \hat{\theta} + z_{1-\alpha/2} \frac{S}{\sqrt{n}} \right], \quad (5)$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of $\mathcal{N}(0, 1)$.

- The proof is similar as before by using (3), instead of (2).
- The interpretation is the same as before.
- In practice, people may also use (4) as approximate CI even when X does not follow a normal distribution.
 - Both (4) and (5) are approximation for finite n when X is non-normal.
 - $t_{n-1, 1-\alpha/2} > z_{1-\alpha/2}$, so CI (4) will be wider than CI (5).
 - CI (4) generally has coverage closer to the desired level $1 - \alpha$.
 - $t_{n-1, 1-\alpha/2} \rightarrow z_{1-\alpha/2}$ as $n \rightarrow \infty$.



- For CI (4), the half-length under $1 - \alpha$ confidence level is

$$H = t_{n-1, 1-\alpha/2} \frac{S}{\sqrt{n}}.$$

- For CI (5), the half-length under $1 - \alpha$ confidence level is

$$H = z_{1-\alpha/2} \frac{S}{\sqrt{n}}.$$

- Half-length H presents the precision (or error) of the estimation for θ .
- We want H to be small enough for our decision making, say, $H \leq \epsilon$, under $1 - \alpha$ confidence level.

- Usually we take an initial sample of size n_0 to get an estimate of σ^2 , say S_0^2 .
 - Assume that the estimate of σ^2 will not change (appreciably) from S_0^2 as the sample size increases.

- For CI (4), an approximate expression for the total sample size required to make $H \leq \epsilon$ is given by

$$n^* = \min \left\{ n \geq n_0 : t_{n-1, 1-\alpha/2} \frac{S_0}{\sqrt{n}} \leq \epsilon \right\}.$$

- For CI (5), an approximate expression is given by

$$n^* = \min \left\{ n \geq n_0 : z_{1-\alpha/2} \frac{S_0}{\sqrt{n}} \leq \epsilon \right\} = \left\lceil \left(\frac{z_{1-\alpha/2} S_0}{\epsilon} \right)^2 \right\rceil. \quad (6)$$

- For simplicity, people sometimes use (6), regardless of the distribution of X .
- Take $n^* - n_0$ additional sample points, or start over and take a sample of size n^* , to form the $1 - \alpha$ CI (with new S).

- Suppose an iid sample is taken and the values are as follows:

| | | | | | | | | | |
|---------|--------|--------|--------|-------|-------|--------|-------|-------|--------|
| 79.919 | 3.081 | 0.062 | 1.961 | 5.845 | 0.941 | 0.878 | 3.371 | 2.157 | 7.579 |
| 3.027 | 6.505 | 0.021 | 0.013 | 0.123 | 0.624 | 5.380 | 3.148 | 7.078 | 23.960 |
| 6.769 | 59.899 | 1.192 | 34.760 | 5.009 | 0.590 | 1.928 | 0.300 | 0.002 | 0.543 |
| 18.387 | 0.141 | 43.565 | 24.420 | 0.433 | 7.004 | 31.764 | 1.005 | 1.147 | 0.219 |
| 144.695 | 2.663 | 17.967 | 0.091 | 9.003 | 3.217 | 14.382 | 1.008 | 2.336 | 4.562 |

- Construct a 95% CI and a 99% CI for $\theta = \mathbb{E}[X]$.

$n = 50$, $\hat{\theta} = \bar{X} = 11.894$, $S = 24.953$. We use CI (4) and get $t_{49, 0.975} = 2.010$, $t_{49, 0.995} = 2.680$. Then,

95% CI: $11.894 \pm 2.010 \times \frac{24.953}{\sqrt{50}} = 11.894 \pm 7.093 = [4.801, 18.987]$;

99% CI: $11.894 \pm 2.680 \times \frac{24.953}{\sqrt{50}} = 11.894 \pm 9.457 = [2.437, 21.351]$.

- Want to make half-length $H \leq 2$ under 95% confidence level.

We use (6) and get $z_{0.975} = 1.960$, $S_0 = 24.953$, $\epsilon = 2$. Then,

$$n^* = \left\lceil \left(\frac{1.960 \times 24.953}{2} \right)^2 \right\rceil = \lceil 597.995 \rceil = 598.$$

Take $598 - 50 = 548$ additional sample points.



- 1 Introduction
 - ▶ Types of Simulations
- 2 Point and Interval Estimations
 - ▶ Basics
 - ▶ Specified Precision
 - ▶ Example
- 3 Terminating Simulation
 - ▶ Discrete Outputs
 - ▶ Continuous Outputs
- 4 Steady-State Simulation
 - ▶ Initialization Bias
 - ▶ Intelligent Initialization
 - ▶ Warm-up Period Deletion
 - ▶ Estimation with Multiple Replications
 - ▶ Estimation with Single Replication



Terminating Simulation

- A terminating simulation runs over the time interval $[0, T_E]$ and produce observations (outputs).
- Discrete outputs: $\{Y_1, Y_2, \dots, Y_n\}$.
 - n may be deterministic or random depending on how T_E is specified.
 - E.g., waiting time of all customers.
 - A common goal is to estimate $\theta := \mathbb{E}[\frac{1}{n} \sum_{i=1}^n Y_i]$, e.g., the *expectation* of the *average* waiting time.
- Continuous outputs: $\{Y(t) : 0 \leq t \leq T_E\}$.
 - T_E may be deterministic or random.
 - E.g., number of customer in the waiting line at time t , $0 \leq t \leq T_E$.
 - A common goal is to estimate $\theta := \mathbb{E}[\frac{1}{T_E} \int_0^{T_E} Y(t)dt]$, e.g., the *expectation* of the *average* waiting line length.
- In general, independent replications (runs) are used, each with a different random number stream.

- Within-replication data vs. across-replication data:

| Replication | Within-Rep Data (each row) | Across-Rep Data |
|-------------|-----------------------------------|---|
| 1 | $Y_{11}, Y_{12}, \dots, Y_{1n_1}$ | $\bar{Y}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i}$ |
| 2 | $Y_{21}, Y_{22}, \dots, Y_{2n_2}$ | $\bar{Y}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} Y_{2i}$ |
| \vdots | \vdots | \vdots |
| R | $Y_{R1}, Y_{R2}, \dots, Y_{Rn_R}$ | $\bar{Y}_R = \frac{1}{n_R} \sum_{i=1}^{n_R} Y_{Ri}$ |

- Across-rep data are independent and identically distributed, when *same* initial conditions and *different* random number streams are used.
- Within-rep data are typically **neither independent nor identically distributed**:
 - waiting times of successive customers are heavily correlated;
 - waiting times during the peak hours are longer than off-peak hours, so they're not identically distributed.
- Use **across-rep** data to do point/interval estimation!



- Example: What is the expectation of average waiting time for customers during $[0, T_E]$?
 - Use $\{\bar{Y}_1, \dots, \bar{Y}_R\}$ as an iid sample of size R .

- Point estimator:

$$\bar{Y} = \frac{1}{R} \sum_{r=1}^R \bar{Y}_r.$$

- $1 - \alpha$ CI using (4):

$$\left[\bar{Y} - t_{R-1, 1-\alpha/2} \frac{S}{\sqrt{R}}, \bar{Y} + t_{R-1, 1-\alpha/2} \frac{S}{\sqrt{R}} \right],$$

where $S^2 = \frac{1}{R-1} \sum_{r=1}^R (\bar{Y}_r - \bar{Y})^2$.

- Necessary number of replications for specified precision $H \leq \epsilon$ under $1 - \alpha$ confidence level, can be computed using (6).

- Within-replication data vs. across-replication data:

| Replication | Within-Rep Data | Across-Rep Data |
|-------------|--------------------------------------|--|
| 1 | $\{Y_1(t) : 0 \leq t \leq T_{E_1}\}$ | $\tilde{Y}_1 = \frac{1}{T_{E_1}} \int_0^{T_{E_1}} Y_1(t) dt$ |
| 2 | $\{Y_2(t) : 0 \leq t \leq T_{E_2}\}$ | $\tilde{Y}_2 = \frac{1}{T_{E_2}} \int_0^{T_{E_2}} Y_2(t) dt$ |
| \vdots | \vdots | \vdots |
| R | $\{Y_R(t) : 0 \leq t \leq T_{E_R}\}$ | $\tilde{Y}_R = \frac{1}{T_{E_R}} \int_0^{T_{E_R}} Y_R(t) dt$ |

- Across-rep data are independent and identically distributed, when *same* initial conditions and *different* random number streams are used.
- Example: What is the expectation of the average waiting line length during $[0, T_E]$?
 - Use $\{\tilde{Y}_1, \dots, \tilde{Y}_R\}$ as an iid sample of size R , and the rest is similar as before.

- 1 Introduction
 - ▶ Types of Simulations
- 2 Point and Interval Estimations
 - ▶ Basics
 - ▶ Specified Precision
 - ▶ Example
- 3 Terminating Simulation
 - ▶ Discrete Outputs
 - ▶ Continuous Outputs
- 4 Steady-State Simulation
 - ▶ Initialization Bias
 - ▶ Intelligent Initialization
 - ▶ Warm-up Period Deletion
 - ▶ Estimation with Multiple Replications
 - ▶ Estimation with Single Replication

Steady-State Simulation

- Consider a single run of a simulation model whose purpose is to estimate a steady-state, or long-run, performance measure of the system.
 - Theoretically speaking, such steady-state performance measure has nothing to do with initial conditions.
- Discrete outputs: $\{Y_1, Y_2, \dots\}$.
 - A common goal is to estimate $\phi := \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i$.
- Continuous outputs: $\{Y(t) : t \geq 0\}$.
 - A common goal is to estimate $\phi := \lim_{T_E \rightarrow \infty} \frac{1}{T_E} \int_0^{T_E} Y(t) dt$.
- However, we cannot simulate a system “to infinity” but must stop somewhere.
 - The simulation run length (n or T_E) is a design choice instead of inherently determined by the nature of the problem.

- The run length in steady-state simulation needs to be *carefully chosen*, with several considerations:
 - bias that is due to artificial or arbitrary initial conditions;
 - can be severe if run length is too short
 - generally decreases as run length increases
 - the desired precision of the point estimator;
 - measured by the **standard error** (i.e., estimator's standard deviation) or confidence interval half-width
 - budget constraints on the time available to execute the simulation.

- The effect of the initial condition is persistent and typically does not vanish after a *finite* time.
 - Unless the initial condition is specified as the “stationary distribution” of the system, which is unknown in general.
- **Fact 1:** Estimators based on a finite-time simulation (finite n or T_E) are biased:

$$\mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n Y_i \right] \neq \phi, \quad \mathbb{E} \left[\frac{1}{T_E} \int_0^{T_E} Y(t) dt \right] \neq \phi.$$

- **Fact 2:** The bias cannot be replicated away:

$$\lim_{R \rightarrow \infty} \frac{1}{R} \sum_{r=1}^R \bar{Y}_r \neq \phi, \quad \lim_{R \rightarrow \infty} \frac{1}{R} \sum_{r=1}^R \tilde{Y}_r \neq \phi.$$

- With more replications, we get a more “precise” estimate of an incorrect value.
 - The confidence interval is narrower but it is centered at an incorrect position.

- Example of $M/M/1$ queue: <https://xiaoweiz.shinyapps.io/MM1queue>
 - If $\lambda < \mu$, the system is stable and the **steady-state expectation** (or **long-run average**) of waiting time is

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n Y_i = \frac{\lambda/\mu}{\mu - \lambda}.$$

- Choosing different initial conditions (in this example, number of customers in station, also known as initial state) gives different looks of sample paths (over finite time period).
- Methods to reduce initialization bias:
 - intelligent initialization;
 - warm-up period deletion;
 - low-bias estimator (advanced topic).

- Initialize the simulation in a state that is more representative of long-run conditions.
- If the system exists, collect data on it and use these data to specify more nearly typical initial conditions:
 - fit a probability distribution to describe the initial state;
 - or, simply use the sample mean as a representative.
- If the system can be simplified enough to make it analytically solvable, e.g. queueing models, use the theoretical solution to initialize the simulation.
 - Solve the simplified model to find the stationary distribution or most likely conditions (e.g., the expected number of customers in a station).
 - This is another important value of those analytically solvable queueing models.

- The impact of the initial condition gradually vanishes as the run length increases.
- So we divide a simulation run into two periods:
 - warm-up period: from time 0 to time T_0 ;
 - data-collection period: from time T_0 to T_E .

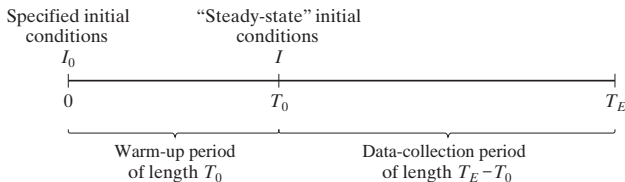


Figure: Warm-up Period Deletion (from [Banks et al. \(2010\)](#))

- T_0 should be sufficiently large so that at time T_0 the impact of the initial condition is very weak and the system behaves approximately as in the steady state.

- To determine T_0
 - There are no widely accepted and proven techniques.
 - Plots are often used.
- The raw output data plot from a single simulation run is usually too fluctuating to detect the trend. – **not helpful**
- Instead of directly plotting raw output data, we usually use some smoother plots to see when the curve “stabilizes”:
 - cumulative average (累积均值); – **OK**
 - ensemble average (总体均值). – **recommended**

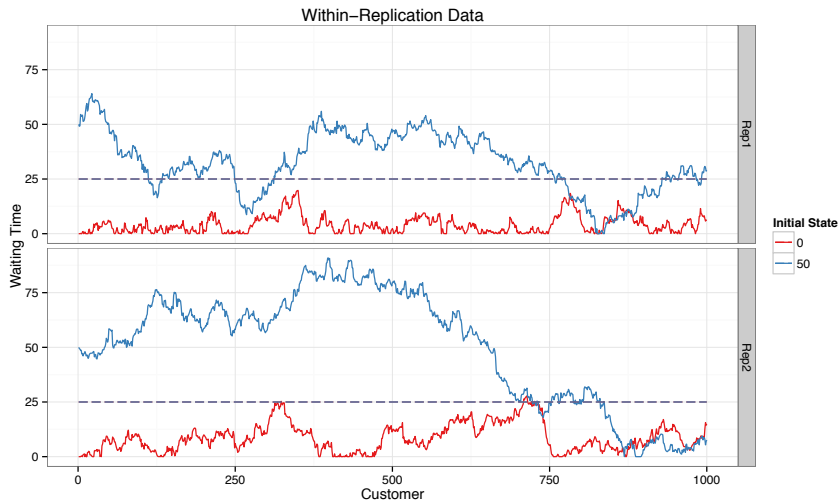


Figure: Raw Output of Waiting Time of Each Customer in $M/M/1$ Queue with $\lambda = 0.962$ and $\mu = 1$ (from [ZHANG Xiaowei](#))



- Cumulative average (累积均值): For one replication, say, replication 1, plot the average from time 0 up to now.
 - Discrete outputs: Plot $\bar{Y}_1(n) = \frac{1}{n} \sum_{i=1}^n Y_{1i}$ with respect to n ;
 - Continuous outputs: Plot $\tilde{Y}_1(T) = \frac{1}{T} \int_0^T Y_1(t) dt$ with respect to T .
- It can be plotted for each replication, so we usually detect different warm-up period durations from different replications.
- The cumulative plot is usually conservative, i.e., the warm-up period it detects is longer than necessary.
 - It retains all of the data including the warm-up period, so the bias needs more time to diminish.

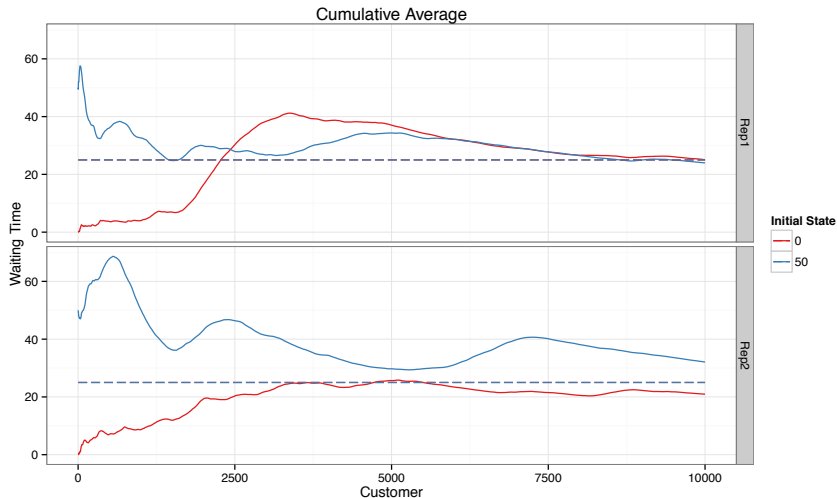


Figure: Cumulative Average Waiting Time of Customers in $M/M/1$ Queue with $\lambda = 0.962$ and $\mu = 1$ (from [ZHANG Xiaowei](#))



- Ensemble average (总体均值): For multiple replications $1, \dots, R$, compute the average across replications and make the plot.
 - Discrete outputs: Plot $\bar{Y}(n) = \frac{1}{R} \sum_{r=1}^R Y_{nr}$ with respect to n ;
 - Continuous outputs: *Divide the raw data of replication r into small batches*, e.g., $\{Y_r(t) : (j-1)m \leq t < jm\}$, $j = 1, 2, \dots$; plot $\tilde{Y}(j) = \frac{1}{R} \sum_{r=1}^R \left[\frac{1}{m} \int_{(j-1)m}^{jm} Y_r(t) dt \right]$ with respect to j .
- We detect one warm-up period duration for multiple replications.
- Some variations are smoothed out by averaging across multiple replications.
 - This leads to more accurate detection of warm-up period.

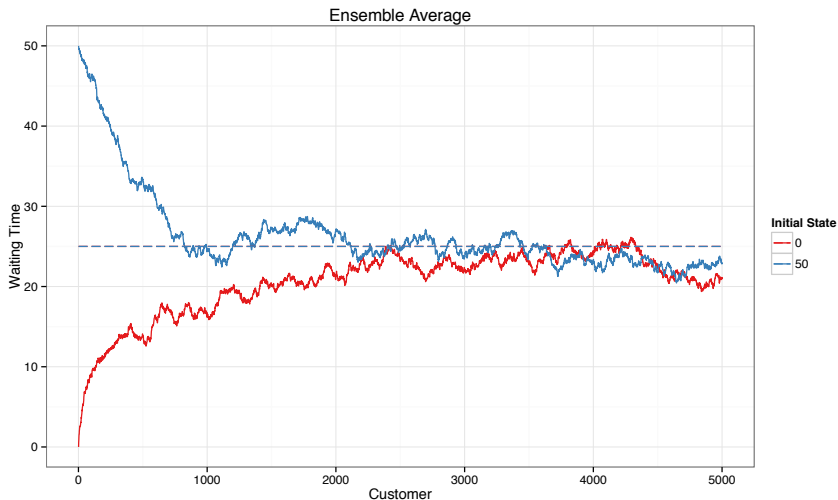


Figure: Ensemble Average Waiting Time of n -th Customer in $M/M/1$ Queue with $\lambda = 0.962$ and $\mu = 1$ (from [ZHANG Xiaowei](#))



- When first starting to detect the warm-up period, a run length and number of replications will have to be guessed.
 - Increase the number of replications if the ensemble averages are not smooth enough.
 - Increase the run length if the ensemble averages do not stabilize.
- Since each ensemble average is the sample mean of iid observations across R replications, a confidence interval can be placed around each point.
 - Use them to judge whether or not the plot is precise enough to decide that the bias has vanished.
 - This is the preferred method to determine a deletion point.

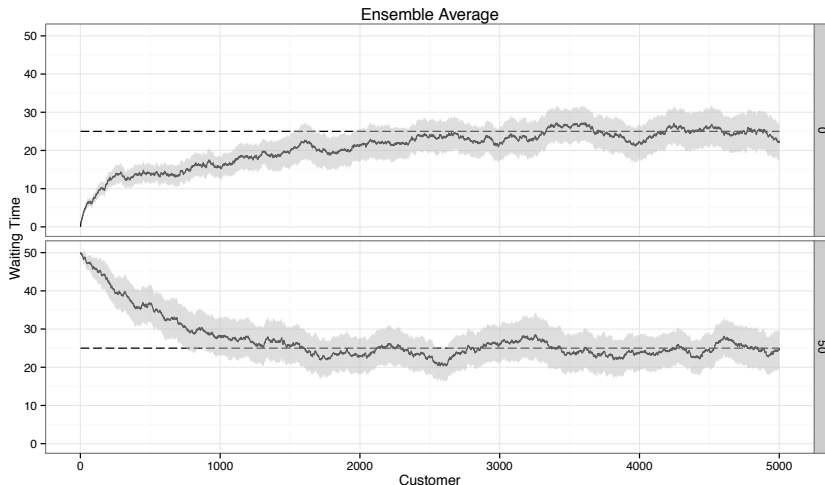


Figure: Ensemble Average Waiting Time and 95% CI of n -th Customer in $M/M/1$ Queue with $\lambda = 0.962$ and $\mu = 1$ (from [ZHANG Xiaowei](#))

- Cumulative averages become less variable as more data are averaged.
 - So the right side of the curve is always smoother than the left side.
- Cumulative averages tend to converge more slowly to long-run performance than ensemble averages do.
 - Because cumulative averages contain all observations including the most biased ones from the very beginning.
 - Cumulative averages should be used *only if* ensemble averages can not be computed, such as when only a single replication is possible.
- Different performance measures could approach steady state with different speed.

- **Idea**: Make multiple replications (long enough), remove warm-up period for each one, and then work as if we were in a terminating simulation.
- **Caution**: Make sure that initialization bias in the point estimator has been reduced to a negligible level.
 - Otherwise the estimation can be misleading.
- **Note**: Initialization bias is not affected by the number of replications.
 - It is affected by deleting more data (i.e. increasing T_0) or extending the run length (i.e. increasing T_E).
 - Increasing the number of replications could produce narrower interval around the “wrong point”.

- Discrete outputs:

- Suppose we decide to delete first d observations of the total n observations in a replication.[†]
- The across-replication data from R replications are

$$\bar{Y}_1 = \frac{1}{n-d} \sum_{i=d+1}^n Y_{1i}, \dots, \bar{Y}_R = \frac{1}{n-d} \sum_{i=d+1}^n Y_{Ri}.$$

- Continuous outputs:

- Suppose we decide to delete data in $[0, T_0]$ period and only use those in $[T_0, T_E]$ in a replication.
- The across-replication data from R replications are

$$\tilde{Y}_1 = \frac{1}{T_E - T_0} \int_{T_0}^{T_E} Y_1(t) dt, \dots, \tilde{Y}_R = \frac{1}{T_E - T_0} \int_{T_0}^{T_E} Y_R(t) dt.$$

[†] d and n may vary between different replications, in which case they are replaced by d_r and n_r , respectively.



- Similar as in terminating simulation, the across-replication data are iid.
 - So we can use them to compute the point estimator, CI, and necessary number of replications for specified precision, in the same way as before.
- Unlike terminating simulation, the above mentioned estimators are biased for finite n or T_E .
 - The bias is negligible if d and n , or T_0 and T_E , are sufficiently large.
- A rough rule for relationship between d and n , or T_0 and T_E :

$$(n - d) \geq 10d, \quad (T_E - T_0) \geq 10T_0.$$

- Suppose analysis indicates that $R - R_0$ additional replications are needed after the initial number R_0 , in order to meet the desired precision.
- An alternative to increasing replications is to increase run length T_E within each replication.
 - Increase run length T_E in the same proportion (R/R_0) to a new run length $(R/R_0)T_E$.
 - More data will be deleted, from time 0 to time $(R/R_0)T_0$.
 - More data will be used to compute the estimate, from time $(R/R_0)T_0$ to time $(R/R_0)T_E$.
 - The total amount of simulation effort is the same as if we had simply increased the number of replications.

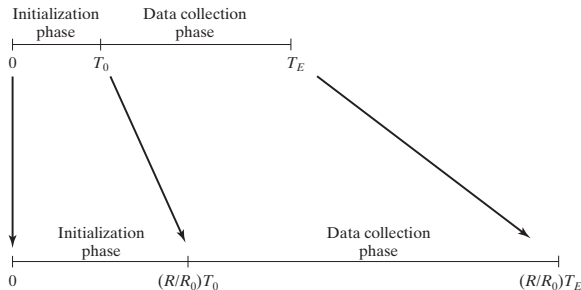


Figure: Increasing Run Length to Achieve Specified Precision (from [Banks et al. \(2010\)](#))

- Advantage: Any residual bias in the point estimator would be further reduced.
- Disadvantage: It is necessary to have saved the state of the model at time T_E and to be able to continue the running.
 - Otherwise, the simulations would have to be re-run from time 0, which could be time consuming for a complex model.

- A disadvantage of the replication method is that the warm-up period must be deleted on each replication.
 - This can become very costly in terms of computation time especially when the model warms up very slowly.
 - E.g., $M/M/1$ queue with utilization close to 1.
- This suggests that we could use one single, (very) long replication for estimation, so that only one warm-up period is deleted.
- Besides, it is also possible that we are in a situation where only the data from one long replication are available.

- Point estimator: Sample mean after the warm-up period deletion

$$\bar{Y} = \frac{1}{n-d} \sum_{i=d+1}^n Y_i, \quad \tilde{Y} = \frac{1}{T_E - T_0} \int_{T_0}^{T_E} Y(t) dt.$$

- The disadvantage of the single-replication design arises when we try to estimate the variance of the above estimators, because of
 - the strong but unknown dependence among Y_1, Y_2, \dots, Y_n ;
 - the non-identical distribution of Y_1, Y_2, \dots, Y_n ;
 - and the integral form of \tilde{Y} .

- **Caution:** It is tempting to compute

$$S^2 = \frac{1}{n-d-1} \sum_{i=d+1}^n (Y_i - \bar{Y})^2,$$

and use $S^2/(n-d)$ to estimate $\text{Var}(\bar{Y})$. However, such estimation will be **terrible**, since Y_1, Y_2, \dots, Y_n are **neither independent nor identically distributed**.

- The CI based on $S^2/(n-d)$ would also be misleading.
- Example: Suppose Y_{d+1}, \dots, Y_n are identically distributed but positive correlated (which is common for waiting time), then

$$\mathbb{E}[S^2/(n-d)] < \text{Var}(\bar{Y}).$$

- The constructed CI using $S^2/(n-d)$ will be narrower than the actual valid one.

- Batch Means (批均值) Method:

- Divide the output data from one replication (after deleting warm-up period) into k **large batches**, and compute the batch means.

- Discrete outputs: $\bar{Y}_j = \frac{1}{m} \sum_{i=(j-1)m+1}^{jm} Y_{i+d}, j = 1, \dots, k.$

$$\underbrace{Y_1, \dots, Y_d}_{\text{deleted}}, \underbrace{Y_{d+1}, \dots, Y_{d+m}}_{\text{Batch 1: } \bar{Y}_1}, \underbrace{Y_{d+m+1}, \dots, Y_{d+2m}}_{\text{Batch 2: } \bar{Y}_2}, \dots, \underbrace{Y_{d+(k-1)m+1}, \dots, Y_{d+km}}_{\text{Batch } k: \bar{Y}_k}$$

- Continuous outputs: $\tilde{Y}_j = \frac{1}{m} \int_{T_0+(j-1)m}^{T_0+jm} Y(t)dt, j = 1, \dots, k.$
- Treat the means of these batches **as if** they were independent.

- Why it works?

- The correlation between two observations decreases as they are farther apart.
- If the batch size is sufficiently large,
 - most of the observations in a batch will be approximately independent of those in other batches;
 - only those near the end of the batches are significantly correlated.

- Strictly speaking, the batch means are not independent.
- However, if the batch size is sufficiently large, successive batch means will be approximately independent.
- Unfortunately, there is no widely accepted and relatively simple method for choosing an acceptable batch size m (or equivalently, choosing a number of batches k).
- Some general guidelines:
 - In most applications, it is suggested to let $10 \leq k \leq 30$, according to [Schmeiser \(1982\)](#).
 - If the run length is to be increased to attain a specified precision, it is suggested to allow both m and k to grow.