7.7. The Inspection Paradox

Given a renewal process \( \{N(t), t \geq 0\} \) with interarrival times \( \{X_i, i \geq 1\} \), the length of the current cycle,

\[
X_{N(t)+1} = S_{N(t)+1} - S_{N(t)}
\]

tend to be longer than \( X_i \), the length of an ordinary cycle.

Precisely speaking, \( X_{N(t)+1} \) is stochastically greater than \( X_i \), which means

\[
P(X_{N(t)+1} > x) \geq P(X_i > x), \quad \text{for all } x \geq 0.
\]

Heuristic Explanation of the Inspection Paradox

Suppose we pick a time \( t \) uniformly in the range \([0, T]\), and then select the cycle that contains \( t \).

- The list of possible cycles to select is \( X_1, X_2, \ldots, X_{N(T)+1} \)
- These cycles are not equally likely to be selected. The longer the cycle, the greater the chance.

\[
P(X_i \text{ is selected}) = \frac{X_i}{T}, \quad \text{for } 1 \leq i \leq N(T)
\]

- So the expected length of the selected cycle \( X_{N(t)+1} \) is roughly

\[
\sum_{i=1}^{N(T)} \frac{X_i}{T} = \frac{\sum_{i=1}^{N(T)} X_i^2}{E[X_i]} \geq E[X_i] \quad \text{as } T \to \infty.
\]

- Last time we have shown that if \( F \) is non-lattice,

\[
\lim_{t \to \infty} E[Y(t)] = \lim_{t \to \infty} E[A(t)] = \frac{E[X_i^2]}{2E[X_i]},
\]

Since \( X_{N(t)+1} = A(t) + Y(t) \),

\[
\lim_{t \to \infty} E[X_{N(t)+1}] = \frac{E[X_i^2]}{E[X_i]}
\]

Example: Bus Waiting Time

- Passengers arrive at a bus station at Poisson rate \( \lambda \)
- Buses arrive one after another according to a renewal process with interarrival times \( X_i, i \geq 1 \), independent of the arrival of customers.
- If \( X_i = 10 \text{min} \) is deterministic, then on average, a passenger has to wait 5 min.
- If \( X_i \) is random with mean 10 min, then for a passenger arrive at the bus station at time \( t \), the amount of time to wait is \( Y(t) \), the residual life of the bus arrival process. We know that

\[
E[Y(t)] \to \frac{E[X_i^2]}{2E[X_i]} = \frac{E[X_i]}{2} = 5 \text{ min}.
\]

Passengers on average will wait longer than half of the average interarrival time of buses.
Example: Crowded Buses
- Passengers arrive at a bus station at Poisson rate $\lambda$.
- Empty buses arrive one after another according to a renewal process with interarrival times $\{X_i \mid i \geq 1\}$, independent of the arrival of customers, and $\mathbb{E}[X_i] = \mu$.
- Each bus departs practically immediately carrying all passengers waiting in line.
- Let $M_i$ be the number of passengers on the $i$-th bus. Note that given $X_i$, $M_i \sim \text{Poisson}(\lambda X_i)$ and hence $\mathbb{E}[M_i] = \mathbb{E}[\mathbb{E}[M_i|X_i]] = \mathbb{E}[\lambda X_i] = \lambda \mu$.
- If you arrive at the station at time $t$, you will get on the $(N(t) + 1)$st bus with $M_{N(t)+1}$ passengers.
- Is $\mathbb{E}[M_{N(t)+1}] = \mathbb{E}[M_i] = \lambda \mu$?

No. Given $X_{N(t)+1}$, $M_{N(t)+1} \sim \text{Poisson}(\lambda X_{N(t)+1})$.

$$
\mathbb{E}[M_{N(t)+1}] = \mathbb{E}[\mathbb{E}[M_{N(t)+1}|X_{N(t)+1}]] \\
= \mathbb{E}[\lambda X_{N(t)+1}] = \frac{\mathbb{E}[X_i^2]}{\mathbb{E}[X_i]} \geq \lambda \mathbb{E}[X_i]
$$

### Limiting Distribution of $X_{N(t)+1}$

If the distribution $F$ of the interarrival times is non-lattice, we can use an alternating renewal process argument to determine

$$
G(x) = \lim_{t \to \infty} P(X_{N(t)+1} \leq x).
$$

We say the renewal process is ON at time $t$ iff $X_{N(t)+1} \leq x$, and OFF otherwise. Thus in the $i$th cycle,

the length of ON time is

$$
\begin{cases} 
X_i & \text{if } X_i \leq x, \\
0 & \text{otherwise}
\end{cases}
$$

and hence

$$
G(x) = \lim_{t \to \infty} P(X_{N(t)+1} \leq x) = \frac{\mathbb{E}[\text{On time in a cycle}]}{\mathbb{E}[\text{cycle time}]} \\
= \frac{\mathbb{E}[X_i 1\{X_i \leq x\}]}{\mathbb{E}[X_i]} = \frac{\int_0^x zf(z)dz}{\mu}
$$

In fact $G(x) = -\frac{x(1-F(x))}{\mu} + F_e(x) < F_e(x)$. 

### Proof of the Inspection Paradox

For $s > x$,

$$
P(X_{N(t)+1} > x|S_{N(t)} = t-s, N(t) = i) = 1 \geq P(X_i > x)
$$

For $s < x$,

$$
P(X_{N(t)+1} > x|S_{N(t)} = t-s, N(t) = i) = P(X_{i+1} > x|S_i = t-s) \\
= P(X_{i+1} > x|X_i+1 > s) \\
= \frac{P(X_{i+1} > x, X_i+1 > s)}{P(X_{i+1} > s)} \\
\geq \frac{P(X_{i+1} > x)}{P(X_{i+1} > s)} \\
\geq P(X_{i+1} > x) = P(X_i > x)
$$

Thus $P(X_{N(t)+1} > x|S_{N(t)} = t-s, N(t) = i) \geq P(X_i > x)$ for all $N(t)$ and $S_{N(t)}$. The claim is validated.

### Chapter 8 Queueing Models

A queueing model consists "customers" arriving to receive some service and then depart. The mechanisms involved are

- input mechanism: the arrival pattern of customers in time
- queueing mechanism: the number of servers, order of the service
- service mechanism: the time to serve one or a batch of customers

We consider queueing models that follow the most common rule of service: first come, first served.
Common Queueing Processes

It is often reasonable to assume
- the interarrival times of customers are i.i.d. (the arrival of customers follows a renewal process),
- the service times for customers are i.i.d. and are independent of the arrival of customers.

Notation: $M =$ memoryless, or Markov, $G =$ General
- $M/M/1$: Poisson arrival, service time $\sim \text{Exp}(\mu)$, 1 server
  - a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv \mu$
- $M/M/\infty$: Poisson arrival, service time $\sim \text{Exp}(\mu)$, $\infty$ servers
  - a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv j \mu$
- $M/M/k$: Poisson arrival, service time $\sim \text{Exp}(\mu)$, $k$ servers
  - a birth and death process with birth rates $\lambda_j \equiv \lambda$, and death rates $\mu_j \equiv \min(j, k) \mu$

Quantities of Interest for Queueing Models

Let
- $X(t) =$ number of customers in the system at time $t$
- $Q(t) =$ number of customers waiting in queue at time $t$

Assume that $\{X(t), t \geq 0\}$ and $\{Q(t), t \geq 0\}$ has a stationary distribution.
- $L =$ the average number of customers in the system
  $$L = \lim_{t \to \infty} \frac{\int_0^t X(t) \, dt}{t};$$
- $L_Q =$ the average number of customers waiting in queue (not being served);
  $$Q = \lim_{t \to \infty} \frac{\int_0^t Q(t) \, dt}{t};$$
- $W =$ the average amount of time, including the time waiting in queue and service time, a customer spends in the system;
- $W_Q =$ the average amount of time a customer spends waiting in queue (not being served).

Little's Formula

Let
- $N(t) =$ number of customers enter the system at or before time $t$

We define $\lambda_a$ be the arrival rate of entering customers,
  $$\lambda_a = \lim_{t \to \infty} \frac{N(t)}{t}$$

**Little's Formula:**

$$L = \lambda_a W$$
$$L_Q = \lambda_a W_Q$$
Cost Identity

Many of interesting and useful relationships between quantities in Queueing models can be obtained by using the cost identity. Imagine that entering customers are forced to pay money (according to some rule) to the system. We would then have the following basic cost identity:

average rate at which the system earns

\[ = \lambda_a \times \text{average amount an entering customer pays} \]

Proof. Let \( R(t) \) be the amount of money the system has earned by time \( t \). Then we have

average rate at which the system earns

\[ = \lim_{t \to \infty} \frac{R(t)}{t} = \lim_{t \to \infty} \frac{N(t)R(t)}{tN(t)} = \lambda_a \lim_{t \to \infty} \frac{R(t)}{N(t)} \]

\[ = \lambda_a \times \text{average amount an entering customer pays}, \]

provided that the limits exist.

Proof of Little's Formula

To prove \( L = \lambda aW \):

- we use the payment rule:
  - each customer pays $1 per unit time while in the system.
  - the average amount customers pay = \( W \), the average waiting time of customers.
  - the amount of money the system earns during the time interval \( (t, t + \Delta t) \) is \( X(t)\Delta t \), where \( X(t) \) is the number of customers in the system at time \( t \),
  - and the rate the system earns is thus
  \[ \lim_{t \to \infty} \frac{\int_0^t X(s) ds}{t} = L, \]

the formula follows from the cost identity.

To prove \( L_Q = \lambda aW_Q \), we use the payment rule:

- each customer pays $1 per unit time while in queue.

The argument is similar.