Chapter 8 Variability and Waiting Time Problems

- A Call Center Example
- Arrival Process and Service Variability
- Predicting Waiting Times
- Waiting Line Management

### 8.1 The Operation of a Typical Call Center

<table>
<thead>
<tr>
<th>Call Center</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answered Calls</td>
</tr>
</tbody>
</table>

- At peak, 80% of calls dialed received a busy signal.
- Customers getting through had to wait on average 10 min.
- Extra telephone expense per day for waiting was $25,000.

<table>
<thead>
<tr>
<th>Lost throughput</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holding cost</td>
</tr>
<tr>
<td>Lost goodwill</td>
</tr>
<tr>
<td>Lost throughput (abandoned)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cost of capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost per customer</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Revenue</th>
</tr>
</thead>
</table>
A “Perfect or Somewhat Odd” Call Center

<table>
<thead>
<tr>
<th>Patient</th>
<th>Arrival Time</th>
<th>Service Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>50</td>
<td>4</td>
</tr>
<tr>
<td>12</td>
<td>55</td>
<td>4</td>
</tr>
</tbody>
</table>

A More Realistic Service Process

<table>
<thead>
<tr>
<th>Patient</th>
<th>Arrival Time</th>
<th>Service Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>25</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>30</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>36</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>45</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>51</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>55</td>
<td>3</td>
</tr>
</tbody>
</table>
Observations

Most customers have to wait, although, on average, there is plenty of capacity in the call center.

The call center is unable to provide consistent service quality.

If customers abandon calls after long waits, the call center loses customer goodwill and revenue.
8.2 Variability: Where does it come from?

Input
- Unpredicted Volume swings
- Random arrivals (randomness is the rule, not the exception)
- Incoming quality
- Product Mix

Buffer

Processing

Tasks
- Inherent variation
- Lack of SOPs
- Quality (scrap / rework)

Resources
- Breakdowns / Maintenance
- Operator absence
- Set-up times

Routes
- Variable routing
- Dedicated machines

Ignoring Variability Leads to Problems

Random arrivals and varying demands are common in services. In the presence of variability, one cannot estimate the process performance based on averages.

Q: Why does variability not average out over time?

A: You cannot inventory services — Capacity can never run ahead of demand.
8.3 Analyzing an Arrival Process

<table>
<thead>
<tr>
<th>Call</th>
<th>Arrival Time, ATi</th>
<th>Interarrival Time, IAi = ATi+1 - ATi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6:00:29</td>
<td>00:23</td>
</tr>
<tr>
<td>2</td>
<td>6:00:52</td>
<td>01:24</td>
</tr>
<tr>
<td>3</td>
<td>6:02:16</td>
<td>00:34</td>
</tr>
<tr>
<td>4</td>
<td>6:02:50</td>
<td>02:24</td>
</tr>
<tr>
<td>5</td>
<td>6:05:14</td>
<td>00:36</td>
</tr>
<tr>
<td>6</td>
<td>6:05:56</td>
<td>00:36</td>
</tr>
<tr>
<td>7</td>
<td>6:06:28</td>
<td>00:38</td>
</tr>
</tbody>
</table>

Coefficient of Variation = \frac{\text{Standard Deviation}}{\text{Mean}}

Seasonality over the Course of a Day

An arrival process is not stationary if the average number of arrivals in any given time interval is not fixed over the entire time period.
### Exponential Distribution

Customers arriving independently from each other follow exponential inter-arrival times. ⇒ Random (Poisson) arrivals

\[ f(t) = \frac{1}{a} e^{-\frac{t}{a}} \]

\( a \) = average inter-arrival time

---

### How to Analyze a Demand/Arrival Process

- **Stationary Arrivals?**
  - **YES**
  - Random Arrivals?
    - **YES**
      - Compute \( a \): average interarrival time
      - \( CV_a = 1 \)
      - All results of chapters 8 and 9 apply
    - **NO**
      - Break arrival process up into smaller time intervals
      - \( CV_a \) = St.dev. of interarrival times/\( a \)
      - All results of chapter 8 apply
      - Results of chapter 9 do not apply, require simulation or more advanced tools

\[ \text{mean} = a \quad CV_a = \frac{\text{Standard deviation of interarrival time}}{\text{Average interarrival time}} \]
8.4 Service Times in Call Center

![Histogram showing call durations in seconds with mean, standard deviation, and median values.]

mean = \( p \)  
\[ \text{CV}_p = \frac{\text{Standard deviation of activity time}}{\text{Average activity time}} \]

8.5 Predicting Average Waiting Time: One Server

![Diagram illustrating the flow of inventory and service times with labels for waiting time, service time, and flow time.]

Flow Time = \( T_q + p \)
The Waiting Time Formula

\[ \text{utilization} = \frac{\text{flow rate}}{\text{capacity}} = \frac{1}{a} = \frac{1}{p} = \frac{p}{a} < 100\% \]

\[ \text{Time in queue} = \text{Activity Time} \times \left( \frac{\text{utilization}}{1 - \text{utilization}} \right) \left( \frac{CV_a^2 + CV_p^2}{2} \right) \]

Reducing average waiting time does not guarantee customer satisfaction.

A small percentage of customers may experience long waits and complain bitterly.

**Solution**: service guarantee and/or service recovery
8.6 Multiple, Parallel Resources with One Queue

![Diagram of multiple, parallel resources with one queue]

- **Inflow** waiting $I_q$
- **Flow rate**
- **Outflow**
- **Entry to system**
- **Begin Service**
- **Departure**
- **Service**
- **Time in queue** $T_q$
- **Service Time** $p$
- **Flow Time** $T = T_q + p$

---

**Waiting Time for Multiple, Parallel Resources**

Under the assumption that

\[
\text{Utilization} = \frac{\text{Flow rate}}{\text{Capacity}} = \frac{1}{m \times (1/\text{activity time})} = \frac{1/a}{m/p} = \frac{p}{a \times m} < 1
\]

we approximate the average waiting time as

\[
\text{Time in queue} = \left(\frac{\text{Activity time}}{m}\right) \times \left(\frac{\text{utilization} \sqrt{2(m+1)}}{1-\text{utilization}}\right) \times \frac{CV_a^2 + CV_p^2}{2}
\]
The Power of Pooling

**Independent Resources**
\[ 2x(m=1) \]

**Pooled Resources**
\[ (m=2) \]

Pooling benefits are lower if queues are not truly independent

---

8.7 Service Levels in Waiting Systems

Fraction of customers who have to wait x seconds or less

Waiting times for those customers who do not get served immediately

Fraction of customers who get served without waiting at all

Waiting time [seconds]

90% of calls had to wait 25 seconds or less

Target Wait Time (TWT) depends on your market position and the importance of incoming calls for your business

Service Level = Probability(Waiting Time ≤ TWT)
Staffing & Incoming Calls over the Course of a Day

8.10 Priority Rules in Waiting Time Systems

- First-Come-First-Serve: easy to implement + perceived fairness
- Shortest Processing Time Rule: Minimizes average waiting time
- Sequence based on importance: emergency or profitable customers

A: 9 minutes  B: 10 minutes  C: 4 minutes  D: 8 minutes

Total wait time: 9+19+23=51 min

Total wait time: 4+13+21=38 min
8.11 Reducing Variability

Reduce Arrival Variability
- appointment/reservation: how to handle late arrivals or no-shows
- encourage customers to avoid peak hours.

Reduce Service Time Variability
- training and technology
- limit service selection
- reduce customer involvement

Actual Wait Time vs. Perceived Wait Time

Perceived Wait Time
- Amount of time customers believe they have waited prior to receiving service.
- Has a greater effect on customer satisfaction than actual waiting time

![Graph: Actual Wait Time vs. Perceived Wait Time]

Satisfaction

Perceived Wait
Factors Affecting Perceived Wait Times

Server-Related Factors
- Passive vs. active waits
- Unfair vs. fair waits
- Uncomfortable vs. comfortable waits
- Unexplained vs. explained waits
- Unproductive vs. Productive waits

Customer-Related Factors
- Solo versus group waits
- Waits for more valuable versus less valuable services
- Customer’s own tolerance

Suggestions for Managing Queues

1. Determine an acceptable waiting time for your customers
2. Try to divert your customer’s attention when waiting
3. Inform your customers of what to expect
4. Keep employees not serving the customers out of sight
5. Segment customers
6. Train your servers to be friendly
7. Encourage customers to come during the slack periods
8. Take a long-term perspective (and redesign the system)
Balancing Efficiency with Responsiveness

- Variability is the norm, not the exception
  - understand where it comes from and eliminate what you can
- Variability leads to waiting times although utilization<100%

Operations benefit from flexibility in capacity
- Demand can exhibit seasonality → Time varying capacity
- Pooling resources can reduce waiting times

Managing customers’ perceived wait times