Patience Estimation and Phase-Type Service Times

- Part 1: Patience Estimation (p.2-10)
- Part 2: Phase-Type Service Times (p.11-17)
Patience Estimation

• In class we saw,

  o if we have a \textit{M/M/s+M} model in steady state,
    • M/M/s queue with customer abandonment with service time \sim \text{exp i.i.d.} and patience time \sim \text{exp}(\theta) i.i.d
    • Also called \textit{Erlang A} model

  o \( \text{P\{Abandon\}} = \theta \ E[\text{Wait}] \)
    • \( P(\text{Abandon}) = \frac{\text{Abandonment rate}}{\lambda} = \frac{\sum_{i=1}^{\infty} P(L_q=i)i\theta}{\lambda} = \frac{E[L_q]\theta}{\lambda} = \theta E[\text{Wait}] \)

  o How to estimate \( \theta \)?
Patience Estimation (2)

• In class we saw (in Erlang A models),
  • $P\{\text{Abandon} \mid \text{Wait} > 0\} = \theta E[\text{Wait} \mid \text{Wait} > 0]$ is also true
    
    o One quick way to check:
      • $P\{\text{Abandon}\} = P\{\text{Abandon} \mid \text{Wait} > 0\}P\{\text{Wait} > 0\} + P\{\text{Abandon} \mid \text{Wait} = 0\}P\{\text{Wait} = 0\}$
      • $E[\text{Wait}] = E[\text{Wait} \mid \text{Wait} > 0]P\{\text{Wait} > 0\} + E[\text{Wait} \mid \text{Wait} = 0]P\{\text{Wait} = 0\}$

• Empirical experience suggests that calculating average patience based on conditioning on $[\text{Wait}>0]$ proves to be more reliable (theoretically no difference)

• Data based on all customers are less relevant since customers that received service immediately do not add any information.
Patience Estimation (3)

- Example) **Estimating Average Patience:**

| Interval | Number of Agents | Calls per Interval | P\{Ab\}   | AHT | E\{Wait | P\{Ab | W>0\} | P\{Ab | W>0\} |
|----------|------------------|--------------------|------------|-----|---------|---------|---------|---------|
| 1        | 6                | 117                | 0.19116932 | 180 | 27.145914 | 0.452453965 |
| 2        | 8                | 140                | 0.12394227 | 180 | 23.321377 | 0.388634027 |
| 3        | 4                | 100                | 0.33184335 | 180 | 33.182468 | 0.553013382 |
| 4        | 10               | 180                | 0.11541903 | 180 | 21.336991 | 0.355551921 |
| 5        | 12               | 200                | 0.0759018  | 180 | 18.941371 | 0.315697006 |
| 6        | 13               | 200                | 0.04916909 | 180 | 17.536591 | 0.292449535 |
| 7        | 14               | 225                | 0.05553867 | 180 | 17.234851 | 0.287137236 |

- \(P\{\text{Wait}>0\}\) and \(E\{\text{Wait}\}\) can be computed by the method on page 3

<table>
<thead>
<tr>
<th>Interval</th>
<th>P{Wait&gt;0}</th>
<th>E{Wait}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.42252</td>
<td>11.46960</td>
</tr>
<tr>
<td>2</td>
<td>0.31892</td>
<td>7.43760</td>
</tr>
<tr>
<td>3</td>
<td>0.60006</td>
<td>19.91160</td>
</tr>
<tr>
<td>4</td>
<td>0.32462</td>
<td>6.92640</td>
</tr>
<tr>
<td>5</td>
<td>0.24043</td>
<td>4.55400</td>
</tr>
<tr>
<td>6</td>
<td>0.16813</td>
<td>2.94840</td>
</tr>
<tr>
<td>7</td>
<td>0.19342</td>
<td>3.33360</td>
</tr>
</tbody>
</table>
Patience Estimation (4)

- **Estimating Average Patience**: Three Alternative Methods

  - Recall we assume patience time \( \sim \exp(\theta) \) i.i.d

1. Censored Sampling
2. Regression
3. Using the Erlang-A model
   - apply using the 4CallCenters software

- Why use these methods instead of estimating hazard rate?
  - Experience shows that the methods are good (e.g., easier to compute, lower variance)
Patience Estimation (5)

- Censored Sampling

  - Compute the average patience by

    \[ \text{Average Patience} = \frac{E[Wait | Wait > 0]}{P\{Abandon | Wait > 0\}} \]

<table>
<thead>
<tr>
<th>Interval</th>
<th>Average Patience (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>59.99707396</td>
</tr>
<tr>
<td>2</td>
<td>60.00858232</td>
</tr>
<tr>
<td>3</td>
<td>60.00301092</td>
</tr>
<tr>
<td>4</td>
<td>60.01090063</td>
</tr>
<tr>
<td>5</td>
<td>59.99857661</td>
</tr>
<tr>
<td>6</td>
<td>59.96450294</td>
</tr>
<tr>
<td>7</td>
<td>60.02304417</td>
</tr>
</tbody>
</table>
Patience Estimation (6)

- Regression
  - Fit a simple regression to the data, taking into account only the customers that had to wait [EXCEL is used]
Patience Estimation (7)

- Regression - continued

![Graph showing the relationship between \( P(\text{Abandon}|\text{Wait}>0) \) vs. \( E(\text{Wait}|\text{Wait}>0) \)]

and we get

\[
P(\text{Abandon}|\text{Wait} > 0) = 4.35 \cdot 10^{-5} + 0.01666E(\text{Wait}|\text{Wait} > 0)
\approx 0.01666E(\text{Wait}|\text{Wait} > 0)
\]

\[\Rightarrow \text{Average Patience} = \frac{1}{0.01666} = 60.00817 \text{ sec}\]
Patience Estimation (8)

- Using the Erlang-A model
  - the 4CallCenters software
Patience Estimation (9)

• Comparison of the three methods

<table>
<thead>
<tr>
<th>Interval</th>
<th>Number of Agents</th>
<th>Call per Interval</th>
<th>Censored Sampling</th>
<th>Regression</th>
<th>4CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>117</td>
<td>59.99707</td>
<td>60.00817199</td>
<td>62</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>140</td>
<td>60.00858</td>
<td>60.00817199</td>
<td>69</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>100</td>
<td>60.00301</td>
<td>60.00817199</td>
<td>63</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>180</td>
<td>60.01090</td>
<td>60.00817199</td>
<td>74</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>200</td>
<td>59.99858</td>
<td>60.00817199</td>
<td>48</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>200</td>
<td>59.96450</td>
<td>60.00817199</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>225</td>
<td>60.02304</td>
<td>60.00817199</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>60.00079</td>
<td>60.00817199</td>
<td>58.054216</td>
</tr>
</tbody>
</table>
Phase-Type Service Times

• Recall
  o Phase-type service time: a sequence/collection of tasks, of an exponential duration.

• Example
  o A city council clerk handles incoming mails. The arrival of documents to the clerk is a Poisson process with constant arrival rate of 5 documents per hour.
  o The handling time of a document is divided into the two stages: reading time and processing time. The average reading time is equal to 3 minutes. It was found that 30% of the documents require long processing time (5 minutes, on average), 60% require short processing time (2 minutes, on average) and the rest of the documents are sent to a waste basket immediately after reading.
  o Assumptions
    • Reading, short/long processing times are exponentially distributed.
    • The stochastic components of the system (interarrival times, service times, switches between states) are independent.
Phase-Type Service Times (2)

• **Q1.** What is the mean and the variance of the service time?

• **Mean:**

\[
E[\text{service time}] = 3 + 0.6 \cdot 2 + 0.3 \cdot 5 = 5.7 \text{ min.}
\]
Phase-Type Service Times (3)

- Variance:

  Reading time is independent of processing time, therefore

  \[ \text{Var[service time]} = \text{Var[reading time]} + \text{Var[processing time]}, \]

  \[ \text{Var[reading time]} = \frac{1}{(1/3)^2} = 9. \]

  The processing time \( X \) can be represented in the following form:

  \[ X = \begin{cases} 
  \exp\left(\frac{1}{5}\right), & p = 0.3 \\
  \exp\left(\frac{1}{2}\right), & p = 0.6 \\
  0, & p = 0.1 
  \end{cases} \]

  Note that the second moment of an exponential random variable \( Y \sim \exp(\lambda) \) is given by:

  \[ \mathbb{E}[Y^2] = (\mathbb{E}Y)^2 + \text{Var}[Y] = \frac{1}{\lambda^2} + \frac{1}{\lambda^2} = \frac{2}{\lambda^2} \]

  Then

  \[ \mathbb{E}[X^2] = 0.3 \cdot \mathbb{E}\left[\exp\left(\frac{1}{5}\right)^2\right] + 0.6 \cdot \mathbb{E}\left[\exp\left(\frac{1}{2}\right)^2\right] = 0.3 \cdot 50 + 0.6 \cdot 8 = 19.8 \]

  \[ \text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}X)^2 = 12.51 \]

  \[ \text{Var[service time]} = 9 + 12.51 = 21.51 \]
Phase-Type Service Times (4)

• [Another method] We saw in class:

    Average work content \( E(T) = qRm \) \((= \sum_j q_j R_{jk} m_k)\).

    Moments: \( E(T^n) = n! q(RM)^n q \), where \( M = \begin{bmatrix} m_1 & 0 \\ \vdots & \ddots \\ 0 & \cdots & m_K \end{bmatrix} \)

• where \( m_k \) = expected duration of task \( k \);
  \( q_k \) = % of services in which \( k \) is first;
  \( P_{jk} \) = % of incidences in which task \( j \) is immediately followed by \( k \).  \( P = [P_{jk}] \)
  \( R = [I - P]^{-1} \)

\[
M = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 5 \end{bmatrix}, \quad q = [1 \ 0 \ 0], \quad P = \begin{bmatrix} 0 & 0.6 & 0.3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}
\]

\[
E[T] = qRM = 5.7
\]
\[
E[T^2] = q(RM)^2 = 54
\]
\[
Var[T] = E[T^2] - (E[T])^2 = 21.51
\]
Phase-Type Service Times (5)

- **Q2.** If two documents or more are on the clerk's desk (including the document he is working on), any new arriving documents are sent to his assistant. What is the fraction of documents handled by the clerk's assistant?

- **Define a stochastic process:**
  
  In order to describe the activities of the clerk, define a stochastic process
  \[ X = \{X(t), \ t \geq 0\} \]
  by
  \[ X(t) = (X_1(t), X_2(t)), \ t \geq 0, \]
  
  where
  \[ X_1(t) = I, \text{ if the server is Idle at time } t; \]
  \[ X_1(t) = R, \text{ if the server is Reading a document at time } t; \]
  \[ X_1(t) = L, \text{ if the server is engaged in a Long processing at } t; \]
  \[ X_1(t) = S, \text{ if the server is engaged in a Short processing at } t; \]
  \[ X_2(t) = 1, \text{ if there is a document awaiting for treatment at time } t; \]
  \[ X_2(t) = 0, \text{ if there is no document in the queue at } t. \]

  Under certain assumptions, it is feasible to represent \( X = \{X(t), \ t \geq 0\} \) as a Markov Jump Process on the state space

  \[ S = \{(I, 0), (R, 0), (R, 1), (S, 0), (S, 1), (L, 0), (L, 1)\}. \]
Phase-Type Service Times (6)

- Transitions-rate diagram
Phase-Type Service Times (7)

- Steady-state equations

\[
\begin{align*}
5\pi_{I0} & = 2\pi_{R0} + 30\pi_{S0} + 12\pi_{L0} \\
20\pi_{R1} & = 5\pi_{R0} \\
35\pi_{S0} & = 12\pi_{R0} \\
17\pi_{L0} & = 6\pi_{R0} \\
30\pi_{S1} & = 12\pi_{R1} + 5\pi_{S0} \\
12\pi_{L1} & = 6\pi_{R1} + 5\pi_{L0} \\
\pi_{I0} + \pi_{R0} + \pi_{S0} + \pi_{L0} + \pi_{R1} + \pi_{S1} + \pi_{L1} & = 1
\end{align*}
\]

\[
\begin{align*}
\pi_{I0} & = 0.582 \\
\pi_{R0} & = 0.176 \\
\pi_{S0} & = 0.060 \\
\pi_{L0} & = 0.062 \\
\pi_{R1} & = 0.044 \\
\pi_{S1} & = 0.028 \\
\pi_{L1} & = 0.048
\end{align*}
\]

- Fraction of documents handled by the clerk’s assistant

The clerk’s assistant handles 12% of the documents since

\[
\pi_{R1} + \pi_{S1} + \pi_{L1} = 0.12 \text{ (PASTA)}
\]