Service Engineering - Recitation 10

Patience Estimation and Phase-Type Service Times

Patience Estimation

- In class we saw,
  - if we have a **M/M/s+M** model in steady state,
    - M/M/s queue with customer abandonment with service time ~ exp i.i.d. and patience time ~ exp(**θ**) i.i.d
    - Also called **Erlang A** model
  - $\mathbf{P\{Abandon\} = \theta\ E[Wait]}$
    - $P(Abandon) = \frac{Abandonment\ rate}{\lambda} = \frac{\sum_{i=1}^{\infty} P(L_q=i) i\theta}{\lambda} = \frac{E[L_q]\theta}{\lambda} = \theta E[Wait]$
  - How to estimate **θ**?

Patience Estimation (2)

- In class we saw (in **Erlang A** models),
  - $\mathbf{P\{Abandon \mid Wait > 0\} = \theta E[Wait|Wait > 0]}$ is also true

  - One quick way to check:
    - $\mathrm{P\{Abandon\} = P\{Abandon \mid Wait > 0\}P\{Wait > 0\} + P\{Abandon \mid Wait = 0\}P\{Wait = 0\}}$
    - $\mathrm{E[Wait] = E[Wait|Wait > 0]P\{Wait > 0\} + E[Wait|Wait = 0]P\{Wait = 0\}}$

- Empirical experience suggests that calculating average patience based on conditioning on [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[W|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(Wait>0) and E{Wait} can be computed by the method on page 3

Interval	P{Wait>0}	E[Wait]
1	0.42252	11.46960
2	0.31892	7.43760
3	0.60006	19.91160
4	0.32462	6.92640
5	0.24043	4.55400
6	0.16813	2.94840
7	0.19342	3.33360

Patience Estimation (4)

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

  - Recall we assume patience time ~ exp(θ) 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

$$Average\ Patience = \frac{E[Wait \mid Wait > 0]}{P\{Abandon \mid Wait > 0\}}$$

Interval	Average Patiece (sec)
1	59.99707396
2	60.00858232
3	60.00301092
4	60.01090063
5	59.99857661
6	59.96450294
7	60.02304417

Patience Estimation (6)

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

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999999563
R Square	**0.999999126**
Adjusted R Square	0.999998951
Standard Error	9.9184E-05
Observations	7

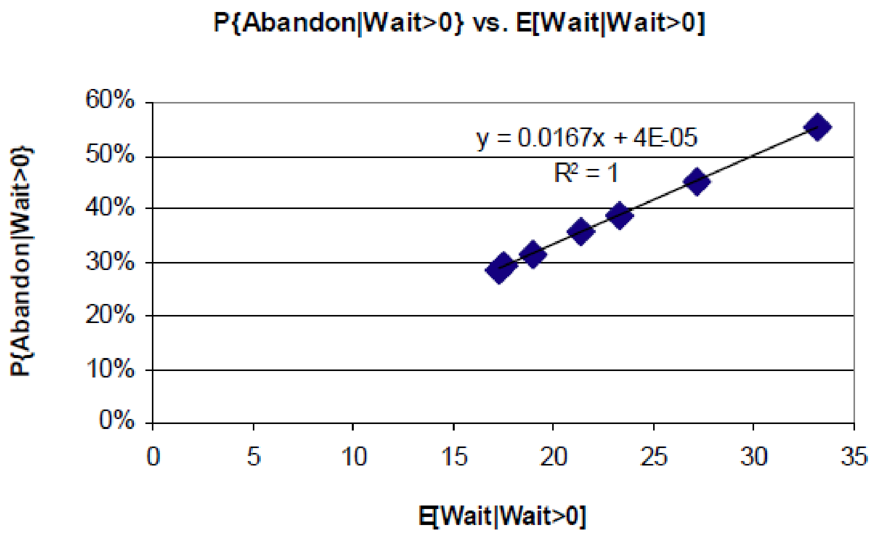
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	0.056246452	0.056246452	5717573.586	**2.42814E-16**
Residual	5	4.91873E-08	9.83747E-09		
Total	6	0.056246501			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	**4.35077E-05**	0.000162388	0.267924422	0.79945905	-0.000373924	0.000460939	-0.000373924	0.000460939
X Variable 1	**0.016664397**	6.96921E-06	2391.144827	2.42814E-16	0.016646482	0.016682312	0.016646482	0.016682312

Patience Estimation (7)

- Regression - continued



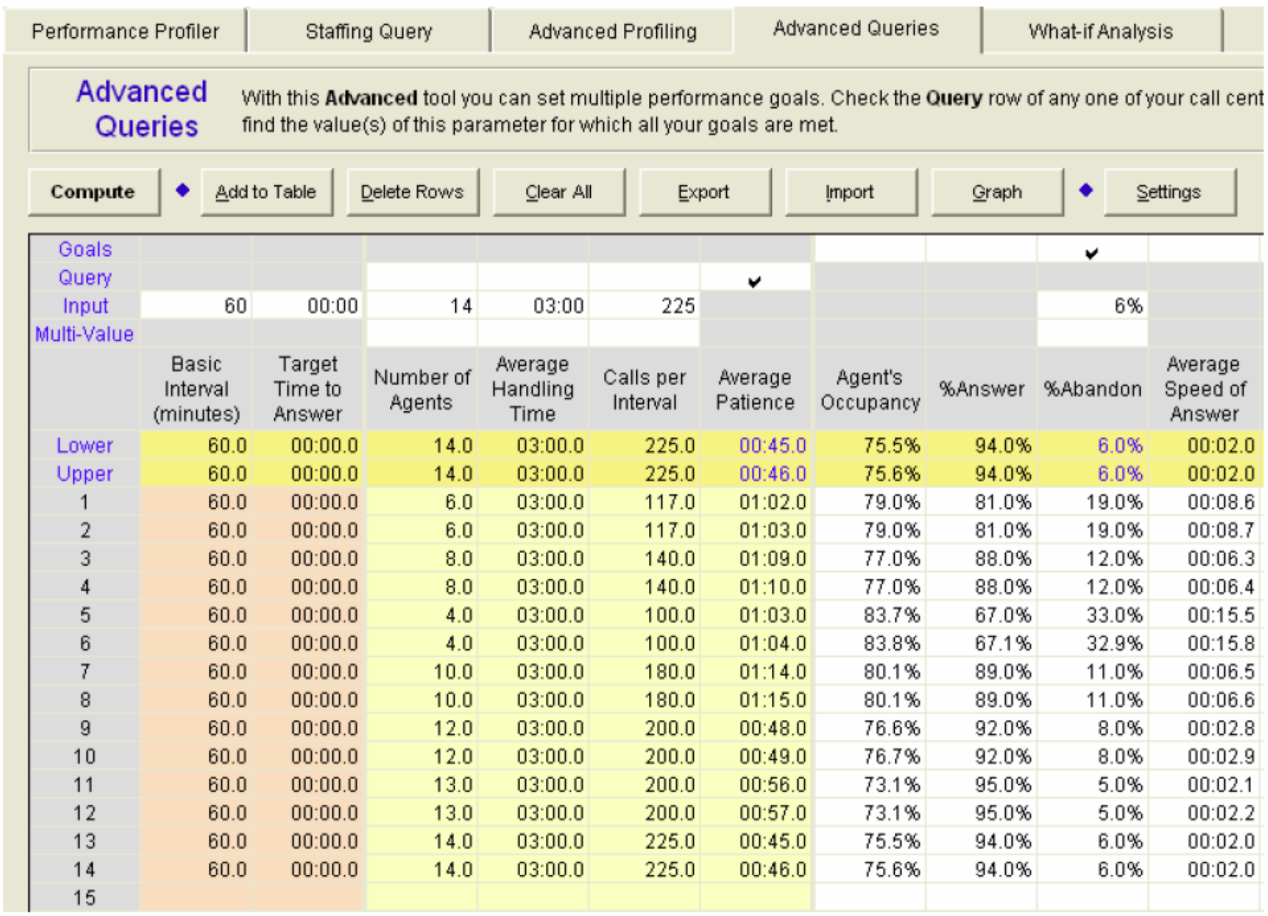
and we get

$$P\{Abandon \mid Wait > 0\} = 4.35 \cdot 10^{-5} + 0.01666 E[Wait \mid Wait > 0]$$
$$\approx 0.01666 E[Wait \mid Wait > 0]$$

$$\Rightarrow \quad Average\ Patience = \frac{1}{0.01666} = 60.00817 \text{ sec}$$

Patience Estimation (8)

- Using the Erlang-A model
  - o the 4CallCenters software



Performance Profiler | Staffing Query | Advanced Profiling | Advanced Queries | What-if Analysis

Advanced Queries

With this **Advanced** tool you can set multiple performance goals. Check the **Query** row of any one of your call cent
find the value(s) of this parameter for which all your goals are met.

Compute | Add to Table | Delete Rows | Clear All | Export | Import | Graph | Settings

	Basic Interval (minutes)	Target Time to Answer	Number of Agents	Average Handling Time	Calls per Interval	Average Patience	Agent's Occupancy	%Answer	%Abandon	Average Speed of Answer
Goals									✓	
Query						✓				
Input	60	00:00	14	03:00	225				6%	
Multi-Value										
Lower	60.0	00:00.0	14.0	03:00.0	225.0	00:45.0	75.5%	94.0%	6.0%	00:02.0
Upper	60.0	00:00.0	14.0	03:00.0	225.0	00:46.0	75.6%	94.0%	6.0%	00:02.0
1	60.0	00:00.0	6.0	03:00.0	117.0	01:02.0	79.0%	81.0%	19.0%	00:08.6
2	60.0	00:00.0	6.0	03:00.0	117.0	01:03.0	79.0%	81.0%	19.0%	00:08.7
3	60.0	00:00.0	8.0	03:00.0	140.0	01:09.0	77.0%	88.0%	12.0%	00:06.3
4	60.0	00:00.0	8.0	03:00.0	140.0	01:10.0	77.0%	88.0%	12.0%	00:06.4
5	60.0	00:00.0	4.0	03:00.0	100.0	01:03.0	83.7%	67.0%	33.0%	00:15.5
6	60.0	00:00.0	4.0	03:00.0	100.0	01:04.0	83.8%	67.1%	32.9%	00:15.8
7	60.0	00:00.0	10.0	03:00.0	180.0	01:14.0	80.1%	89.0%	11.0%	00:06.5
8	60.0	00:00.0	10.0	03:00.0	180.0	01:15.0	80.1%	89.0%	11.0%	00:06.6
9	60.0	00:00.0	12.0	03:00.0	200.0	00:48.0	76.6%	92.0%	8.0%	00:02.8
10	60.0	00:00.0	12.0	03:00.0	200.0	00:49.0	76.7%	92.0%	8.0%	00:02.9
11	60.0	00:00.0	13.0	03:00.0	200.0	00:56.0	73.1%	95.0%	5.0%	00:02.1
12	60.0	00:00.0	13.0	03:00.0	200.0	00:57.0	73.1%	95.0%	5.0%	00:02.2
13	60.0	00:00.0	14.0	03:00.0	225.0	00:45.0	75.5%	94.0%	6.0%	00:02.0
14	60.0	00:00.0	14.0	03:00.0	225.0	00:46.0	75.6%	94.0%	6.0%	00:02.0
15										

Patience Estimation (9)

- Comparison of the three methods

Interval	Number of Agents	Call per Interval	Censored Sampling	Regression	4CC
1	6	117	59.99707	60.00817199	62
2	8	140	60.00858	60.00817199	69
3	4	100	60.00301	60.00817199	63
4	10	180	60.01090	60.00817199	74
5	12	200	59.99858	60.00817199	48
6	13	200	59.96450	60.00817199	56
7	14	225	60.02304	60.00817199	45
			60.00079	60.00817199	58.054216

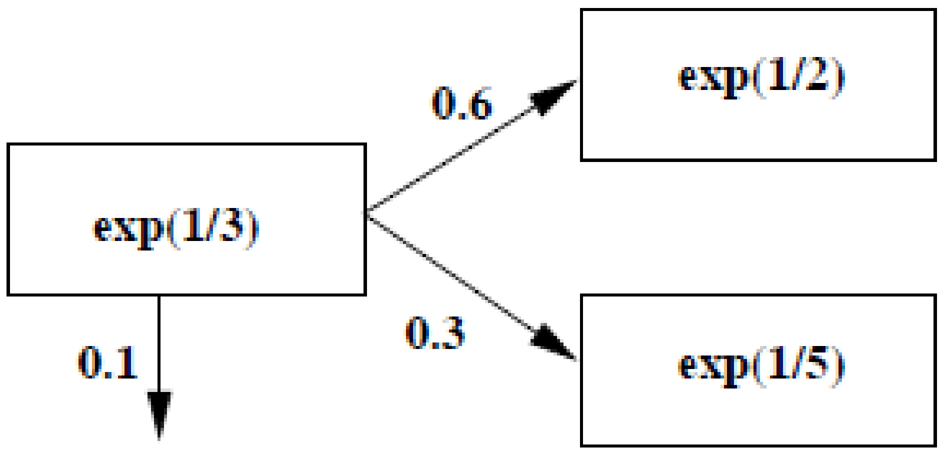
Phase-Type Service Times

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

- Example
  - 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.
  - 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.
  - 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: $\mathrm{E}[\text{service time}] = 3 + 0.6 \cdot 2 + 0.3 \cdot 5 = 5.7 \text{ min.}$

Phase-Type Service Times (3)

- Variance:

Reading time is independent of processing time, therefore

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

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

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:

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

Then

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

$$\mathrm{Var}[X] = \mathrm{E}[X^2] - (\mathrm{E}X)^2 = 12.51$$

$$\mathrm{Var}[\text{service time}] = 9 + 12.51 = 21.51$$

Phase-Type Service Times (4)

- [Another method] We saw in class:

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

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

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

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

$\rightarrow$ $E[T] = qRM = 5.7$

$\rightarrow$ $E[T^2] = q(RM)^2 = 54$

$\rightarrow$ $Var[T] = E[T^2] - (E[T])^2 = 21.51$

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

$X_1(t) = I$, if the server is Idle at time t;

$X_1(t) = R$, if the server is Reading a document at time t;

$X_1(t) = L$, if the server is engaged in a Long processing at t;

$X_1(t) = S$, if the server is engaged in a Short processing at t;

$X_2(t) = 1$, if there is a document awaiting for treatment at time t;

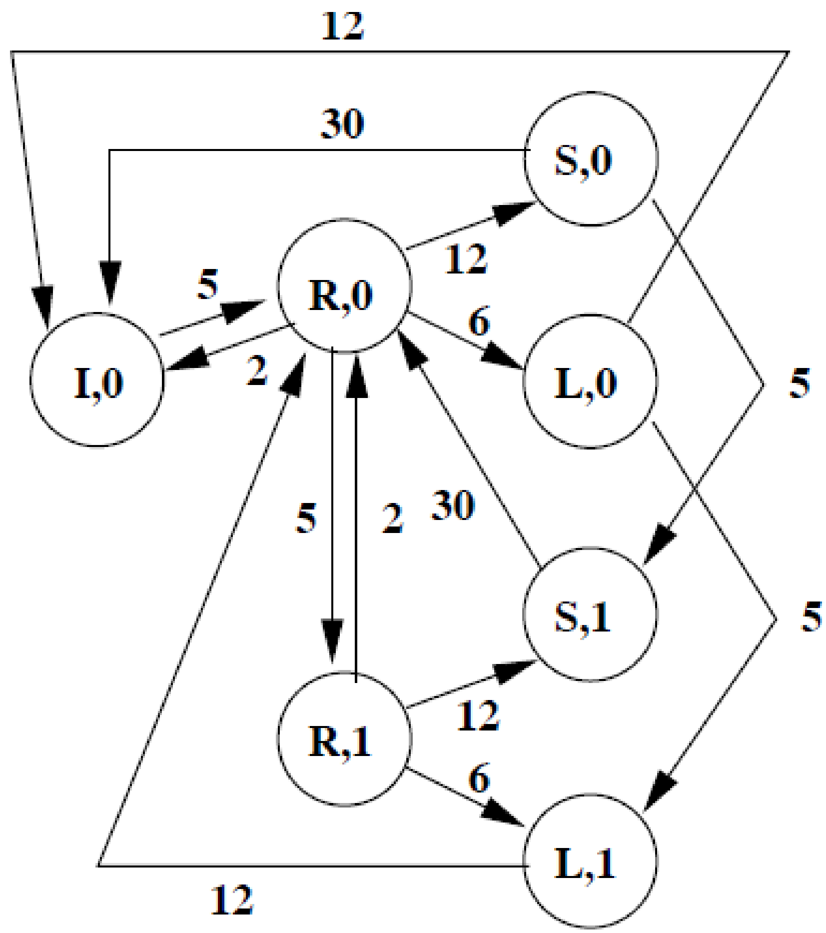
$X_2(t) = 0$, if there is no document in the queue at t.

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

$$\begin{cases} 5\pi_{I0} = 2\pi_{R0} + 30\pi_{S0} + 12\pi_{L0} \\ 20\pi_{R1} = 5\pi_{R0} \\ 35\pi_{S0} = 12\pi_{R0} \\ 17\pi_{L0} = 6\pi_{R0} \\ 30\pi_{S1} = 12\pi_{R1} + 5\pi_{S0} \\ 12\pi_{L1} = 6\pi_{R1} + 5\pi_{L0} \\ \pi_{I0} + \pi_{R0} + \pi_{S0} + \pi_{L0} + \pi_{R1} + \pi_{S1} + \pi_{L1} = 1 \end{cases} \Longrightarrow \begin{cases} \pi_{I0} = 0.582 \\ \pi_{R0} = 0.176 \\ \pi_{S0} = 0.060 \\ \pi_{L0} = 0.062 \\ \pi_{R1} = 0.044 \\ \pi_{S1} = 0.028 \\ \pi_{L1} = 0.048 \end{cases}$$

- Fraction of documents handled by the clerk's assistant

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

$$\pi_{R1} + \pi_{S1} + \pi_{L1} \;=\; 0.12 \quad \text{(PASTA)}$$