Design and Inference of a Call Center with an Answering Machine (IVR)

Polina Khudyakov

Joint work with Prof. Avishai Mandelbaum, Prof. Malka Gorfine and Prof. Paul Feigin

June 29, 2010
Outline

- Design of a Call Center with an IVR
- Customer Patience Analysis
- Service Time Analysis
Background

- Halfin and Whitt (1981) \((M/M/S)\)
- Massey and Wallace (2004) \((M/M/S/N)\)
- Garnet, Mandelbaum and Reiman (2002) \((M/M/S+M)\)
- Srinivasan, Talim and Wang (2002) \((Call\ center\ with\ an\ IVR)\)
Customer interaction with a call center (Model I)

M/M/S/N+M
Customer interaction with a call center (Model II)
Model II (with an IVR)

Model description

- $N$ - number of trunk lines
- $\text{Poisson}(\lambda)$ - arrival process
- $p$ - probability to request agent’s service
- $S$ - number of agents
- $\exp(\theta)$ - IVR service time
- $\exp(\mu)$ - agent’s service time
- $\exp(\delta)$ - customer patience time
Schematic model

Model II (with an IVR)

"IVR"
N servers

1
2
3
N

θ
1-p

"Agents"
S servers

1
2
S

μ

δ

λ

p
Closed Jackson Network

- Brandt, Brandt, Spahl and Weber (1997)

\[
\begin{align*}
\text{N servers} & \quad \text{1} \quad \exp(\theta) \quad \text{1-p} \quad \exp(\lambda) \\
\text{S servers} & \quad \text{2} \quad \text{p} \quad \cdots \quad \exp(\mu)/\exp(\delta) \\
\text{1 server} & \quad \text{3} \quad \cdots \quad \text{1-p} \quad \text{p}
\end{align*}
\]
Stationary Probabilities

\[ \pi(i,j) = \begin{cases} 
\pi_0 \frac{1}{i!} \left( \frac{\lambda}{\theta} \right)^i \frac{1}{j!} \left( \frac{\lambda p}{\mu} \right)^j, & j \leq S, \ 0 \leq i + j \leq N; \\
\pi_0 \frac{1}{i!} \left( \frac{\lambda}{\theta} \right)^i \frac{1}{S!} \left( \frac{\lambda p}{\mu} \right)^S \frac{(\lambda p)^{j-S}}{\prod_{k=1}^{j}(S\mu + k\delta)}, & j \geq S, \ 0 \leq i + j \leq N; \\
0, & \text{otherwise,}
\end{cases} \]

where

\[ \pi_0 = \left[ \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \frac{1}{i!} \left( \frac{\lambda}{\theta} \right)^i \frac{1}{S!} \left( \frac{\lambda p}{\mu} \right)^S \frac{(\lambda p)^{j-S}}{\prod_{k=1}^{j}(S\mu + k\delta)} \right. \\
+ \left. \sum_{i+j \leq N, j \leq S-1} \frac{1}{i!} \left( \frac{\lambda}{\theta} \right)^i \frac{1}{j!} \left( \frac{\lambda p}{\mu} \right)^j \right]^{-1}. \]
Exact Performance Measures

\[ P(W > 0) = \sum_{i=0}^{N-S} \sum_{j=S}^{N-i} \chi(i+j,j), \]

where \( \chi(k,j), 0 \leq j < k \leq N \), is the probability that the system is in state \((k,j)\), given that a call (among the \(k-j\) customers) is about to complete its IVR service:

\[ \chi(k,j) = \frac{(k-j) \pi(k-j,j)}{\sum_{l=0}^{N} \sum_{m=0}^{l} (l-m) \pi(l-m,m)} . \]

\[ P(Ab|W > 0) = \frac{\sum_{i=0}^{N-S} \sum_{j=S+1}^{N} \pi(i,j)(j-S)d}{\sum_{i=0}^{N-S} \sum_{j=S+1}^{N} \pi(i,j)(S\mu + (j-S)d)}, \]

\[ E[W|W > 0] = \frac{1}{d} P(Ab|W > 0). \]
The domain for asymptotic analysis: QED

- Model I (M/M/S/N+M)

\[
(i) \lim_{\lambda \to \infty} N - S = \eta \sqrt{\frac{\lambda}{\mu}}, \quad 0 < \eta < \infty;
\]

\[
(ii) \lim_{\lambda \to \infty} S = \frac{\lambda}{\mu} + \beta \sqrt{\frac{\lambda}{\mu}}, \quad -\infty < \beta < \infty.
\]

- Model II (Call Center with an IVR)

\[
(i) \lim_{\lambda \to \infty} N - S = \frac{\lambda}{\theta} + \eta \sqrt{\frac{\lambda}{\theta}}, \quad -\infty < \eta < \infty;
\]

\[
(ii) \lim_{\lambda \to \infty} S = \frac{\lambda p}{\mu} + \beta \sqrt{\frac{\lambda p}{\mu}}, \quad -\infty < \beta < \infty.
\]
QED Approximations (Model I)

**Theorem:**

Let the variables $\lambda$, $S$ and $N$ tend to $\infty$ simultaneously and satisfy QED conditions, where $\mu$ is fixed. Then the asymptotic behavior of the system is described in terms of the following performance measures:

- the probability $P(W > 0)$ that a served customer waits after the IVR:

$$
\lim_{\lambda \to \infty} P(W > 0) = \left( 1 + \frac{\sqrt{\frac{\delta}{\mu}} \Phi(\beta) \varphi(\beta \sqrt{\frac{\mu}{\delta}})}{\varphi(\beta) \left[ \Phi(\eta \sqrt{\frac{\delta}{\mu}} + \beta \sqrt{\frac{\mu}{\delta}}) - \Phi(\beta \sqrt{\frac{\mu}{\delta}}) \right]} \right)^{-1},
$$

- the probability of abandonment, given waiting:

$$
\lim_{\lambda \to \infty} \sqrt{S} P(Ab|W > 0) = \frac{\sqrt{\frac{\delta}{\mu}} \varphi(\beta \sqrt{\frac{\mu}{\delta}}) - \beta \left[ \Phi(\eta \sqrt{\frac{\delta}{\mu}} + \beta \sqrt{\frac{\mu}{\delta}}) - \Phi(\beta \sqrt{\frac{\mu}{\delta}}) \right]}{\Phi(\eta \sqrt{\frac{\delta}{\mu}} + \beta \sqrt{\frac{\mu}{\delta}}) - \Phi(\beta \sqrt{\frac{\mu}{\delta}})}.
$$
Theorem:

Let the variables $\lambda$, $S$ and $N$ tend to $\infty$ simultaneously and satisfy the QED conditions, where $\mu, p, \theta, \delta$ are fixed. Then the asymptotic behavior of our system, is captured by the following performance measures:

- the probability $P(W > 0)$ that a served customer waits after the IVR:

  $$
  \lim_{\lambda \to \infty} P(W > 0) = \left(1 + \frac{\gamma}{\xi_1 - \xi_2}\right)^{-1};
  $$

- the probability of abandonment, given waiting:

  $$
  \lim_{\lambda \to \infty} \sqrt{SP(\text{Ab}|W > 0)} = \frac{\sqrt{\frac{\delta}{\mu}} \varphi(\beta \sqrt{\frac{\mu}{\delta}}) \Phi(\eta) - \beta \int_{\beta \sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta \sqrt{\frac{\mu}{\delta}} - t) \sqrt{\frac{p\theta}{\mu}}) \varphi(t)dt}{\int_{\beta \sqrt{\frac{\mu}{\delta}}}^{\infty} \Phi(\eta + (\beta \sqrt{\frac{\mu}{\delta}} - t) \sqrt{\frac{p\theta}{\mu}}) \varphi(t)dt}.
  $$

where

- $\xi_1 = \sqrt{\frac{\mu}{\delta}} \varphi(\beta) \int_{-\infty}^{\eta} \Phi\left((\eta - t) \sqrt{\frac{\delta}{p\theta}} + \beta \sqrt{\frac{\mu}{\delta}}\right) \varphi(t)dt$,

- $\xi_2 = \sqrt{\frac{\mu}{\delta}} \varphi(\beta) \Phi(\beta \sqrt{\frac{\mu}{\delta}}) \varphi(\eta)$,

- $\gamma = \int_{-\infty}^{\beta} \Phi\left(\eta + (\beta - t) \sqrt{\frac{p\theta}{\mu}}\right) \varphi(t)dt$. 

Polina Khudyakov (Technion, Israel)
Illustration of the $P(W > 0)$ approximation

- **A small-size call center**
  - $\lambda = 10$, $N = 50$, $\theta = \mu = p = 1$, $\delta = 10$

- **A mid-size call center**
  - $\lambda = 50$, $N = 150$, $\theta = \mu = p = 1$, $\delta = 10$

- **A mid-size call center**
  - $\lambda = 80$, $N = 200$, $\theta = \mu = p = 1$, $\delta = 10$

- **A large call center**
  - $\lambda = 500$, $N = 1500$, $\theta = \mu = p = 1$, $\delta = 10$
Data Description

- Service type - Retail (80% of total customers)
- April 12, 2001 (an ordinary week day)
- Observed time period is from 07:00 till 18:00
- About 600 agents per shift
True vs. Approximated Performance Measures

P(W>0) and its approximation

- exact
- approx (model with an IVR)
- approx (M/M/S/N+M)
Customer Patience Analysis

- **Customer patience** is the willingness of a customer to wait before being served
- Customers receive the service before they lose their patience
- Censored data
- Survival analysis
Customer Patience Analysis (cont.)

- Does customers’ individual features influence on their waiting behavior?
  - observed $\Rightarrow$ covariates
  - unobserved $\Rightarrow$ frailty
Customer Patience Analysis (cont.)

- Does customers’ individual features influence on their waiting behavior?
  - observed $\Rightarrow$ covariates
  - unobserved $\Rightarrow$ frailty

- How customers’ experience with the system affect of their waiting behavior?
Customer Patience Analysis (cont.)

- Does customers’ individual features influence on their waiting behavior?
  - observed ⇒ covariates
  - unobserved ⇒ frailty

- How customers’ experience with the system affect of their waiting behavior?

- Data selection

![Diagram showing customer satisfaction over time with 1 month interval between observations](image-url)
The Model

- Conditional hazard function

\[ \lambda_{ij}(t | w_i) = \lambda_{0j}(t) w_i e^{\beta^T Z_{ij}} \quad i = 1, \ldots, n \quad j = 1, \ldots, m_i \]

- \( \lambda_{0j}(t) \) is an unspecified baseline hazard function of call \( j \)
- \( \beta \) is a \( p \)-dimensional vector of unknown regression coefficients
- \( w_i \) are i.i.d. random variables with known density \( f(w) \equiv f(w; \theta) \)
- \( \theta \) is an unknown vector of parameters
Estimation Procedure

- Gorfine, Zucker and Hsu (Biometrika, 2006)

\[
L = \prod_{i=1}^{n} \prod_{j=1}^{m_i} \left\{ \lambda_{0j} \left( T_{ij} \right) e^{\beta^T Z_{ij}} \right\} \delta_{ij} \prod_{i=1}^{n} \int_{0}^{\infty} w^{N_i(\tau)} e^{-wH_i(\tau)} f(w) dw
\]

**Step 1.**
Given the value of \( \gamma = (\theta, \beta^T) \) estimate \( \left\{ \Lambda_{0j}(\tau) \right\}_{j=1}^{m} \) by using

\[
\Delta \hat{\Lambda}_{0j}(\tau_{jk}) = \left( \sum_{i=1}^{n} dN_{ij}(\tau_{jk}) \right) / \left( \sum_{i=1}^{n} h_{ij}(\tau_{jk}) Y_{ij}(\tau_{jk}) \right)
\]

**Step 2.**
Given value of \( \left\{ \hat{\Lambda}_{0j} \right\}_{j=1}^{m} \), estimate \( \gamma \) by using the score equations \( U(\gamma, \left\{ \hat{\Lambda}_{0j} \right\}_{j=1}^{m}) = 0 \)

**Step 3.**
Repeat Steps 1 and 2 until convergence is reached with respect to \( \left\{ \hat{\Lambda}_{0j} \right\}_{j=1}^{m} \) and \( \hat{\gamma} \).
Data Analysis

- The sample consists of \( n = 50,000 \)
- Each customer did at most 5 calls
- Standard errors are based on 150 bootstrap samples
- Estimates of the model parameters:

<table>
<thead>
<tr>
<th></th>
<th>( \hat{\theta} )</th>
<th>( \hat{\beta}_1 )</th>
<th>( \hat{\beta}_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.9973</td>
<td>-0.3006</td>
<td>-0.1211</td>
</tr>
<tr>
<td>SD</td>
<td>0.1767</td>
<td>0.1046</td>
<td>0.0567</td>
</tr>
</tbody>
</table>
Test for Equality of Baseline Hazard Functions

- \( H_0 : \Lambda_{01}(t) = \Lambda_{02}(t) = \Lambda_0(t) \quad \text{for all} \quad 0 < t \leq \tau \)

Our test statistic:

\[
S_n(\tau) = \frac{1}{\sqrt{n}} \int_0^\tau \hat{W}_n(s) \frac{\bar{Y}_1(s) \bar{Y}_2(s)}{\bar{Y}_1(s) + \bar{Y}_2(s)} \left\{ d\hat{\Lambda}_{01}(s) - d\hat{\Lambda}_{02}(s) \right\}
\]

- \( \bar{Y}_j(s) = \sum_{i=1}^n \hat{h}_{ij}(s) Y_{ij}(s) \)

- \( \hat{W}_n(s) \) a nonnegative weight function
Data analysis

- Results of the 10 tests
- The sample consists from $n = 50,000$ customers:

<table>
<thead>
<tr>
<th></th>
<th>1 vs. 2</th>
<th>1 vs. 3</th>
<th>1 vs. 4</th>
<th>1 vs. 5</th>
<th>2 vs. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_n(250)$</td>
<td>-0.464</td>
<td>-0.771</td>
<td>-0.051</td>
<td>-0.048</td>
<td>0.027</td>
</tr>
<tr>
<td>$\hat{\sigma}_i(250)$</td>
<td>0.039</td>
<td>0.199</td>
<td>0.018</td>
<td>0.016</td>
<td>0.027</td>
</tr>
<tr>
<td>$S_n(250)/\hat{\sigma}_i(250)$</td>
<td>-11.915</td>
<td>-3.871</td>
<td>-2.884</td>
<td>-3.019</td>
<td>1.024</td>
</tr>
<tr>
<td>$p-value$</td>
<td>&lt;0.001</td>
<td>&lt;0.001</td>
<td>0.002</td>
<td>0.001</td>
<td>0.847</td>
</tr>
<tr>
<td>$FDR p-value$</td>
<td>&lt;0.001</td>
<td>0.003</td>
<td>0.042</td>
<td>0.003</td>
<td>1.000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2 vs. 4</th>
<th>2 vs. 5</th>
<th>3 vs. 4</th>
<th>3 vs. 5</th>
<th>4 vs. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_n(250)$</td>
<td>0.058</td>
<td>-0.029</td>
<td>-0.014</td>
<td>-0.030</td>
<td>-0.019</td>
</tr>
<tr>
<td>$\hat{\sigma}_i(250)$</td>
<td>0.163</td>
<td>0.016</td>
<td>0.016</td>
<td>0.015</td>
<td>0.012</td>
</tr>
<tr>
<td>$S_n(250)/\hat{\sigma}_i(250)$</td>
<td>0.355</td>
<td>-1.841</td>
<td>-0.907</td>
<td>-2.051</td>
<td>-1.493</td>
</tr>
<tr>
<td>$p-value$</td>
<td>0.639</td>
<td>0.033</td>
<td>0.182</td>
<td>0.020</td>
<td>0.068</td>
</tr>
<tr>
<td>$FDR p-value$</td>
<td>1.000</td>
<td>0.096</td>
<td>1.000</td>
<td>0.042</td>
<td>0.422</td>
</tr>
</tbody>
</table>
95% Pointwise Confidence Intervals

- 1-st call vs. 2-nd call
- 2-nd call vs. 3-nd call
1-st call vs. 5-th call

Cumulative Baseline Hazard Function

- 1-st call
- 5-th call

p-value = 0.00127
Phase Type Distribution for Agents’ Service Time
Fitting the Agents’ Service Time

![Graph showing service time distribution with data and phase type models: Log-Normal.](image)