

Uncertainty in the Demand for Service: The Case of Call Centers and Emergency Departments

Shimrit Maman

Technion - Israel Institute of Technology

March 16, 2009

Advisors: Prof. Avishai Mandelbaum and Dr. Sergey Zeltyn

Outline

- 1 Introduction
 - Motivation
 - Research Outline
 - Related Work
 - Model Definition
- 2 Case Studies
 - Financial Call Center
 - Emergency Department
- 3 Theoretical Results
 - QED-c Regime
 - ED Regime
- 4 Time-Varying Queues
- 5 Future Research

Motivation

Standard assumption in service system modeling: arrival process is Poisson with known parameters.

Example of call centers: known arrival rates for each basic interval (say, half-hour).

Motivation

Standard assumption in service system modeling: arrival process is Poisson with known parameters.

Example of call centers: known arrival rates for each basic interval (say, half-hour).

Application of standard approach to basic interval (say, next Tuesday, 9am-9:30am):

- Derive Poisson parameters from historical data.
- Plug parameters into a queueing model ($M|M|n$, $M|M|n + M$, Skills-Based Routing models, ...).
- Set staffing levels according to this model and service-level agreement.

Motivation

Standard assumption in service system modeling: arrival process is Poisson with known parameters.

Example of call centers: known arrival rates for each basic interval (say, half-hour).

Application of standard approach to basic interval (say, next Tuesday, 9am-9:30am):

- Derive Poisson parameters from historical data.
- Plug parameters into a queueing model ($M|M|n$, $M|M|n + M$, Skills-Based Routing models, ...).
- Set staffing levels according to this model and service-level agreement.

Is standard Poisson assumption valid? As a rule it is not, one observes larger variability of the arrival process than the one expected from the Poisson hypothesis.

Research Outline

- Design model for overdispersed arrival rate.
- Plug arrival model into $M|M|n + G$ queueing model.
- Derive asymptotic results relevant for real-life staffing problems.
- Validate our approach via analysis of real data.

Related Work



Henderson S.

Input model uncertainty: Why do we care and what should we do about it?. 2003.



Steckley S., Henderson S. and Mehrotra V.

Forecast errors in service systems. 2007.



Koole G. and Jongbloed G.

Managing uncertainty in call centers using poisson mixtures. 2001.



Halfin S. and Whitt W.

Heavy-traffic limits for queues with many exponential servers. 1981.



Zeltyn S. and Mandelbaum A.

Call centers with impatient customers: Exact analysis and many-server asymptotics of the $M|M|n + G$ queue. 2005.



Feldman Z., Mandelbaum A., Massey W. A. and Whitt W.

Staffing of time-varying queues to achieve time-stable performance. 2007.

Model Definition

The $M^?|M|n + G$ Queue:

- λ - **Expected** arrival rate of a Poisson arrival process.
- μ - Exponential service rate.
- n service agents.
- G - Patience distribution. Assume that the patience density exists at the origin and its value g_0 is strictly positive.

Model Definition

The $M^?|M|n + G$ Queue:

- λ - **Expected** arrival rate of a Poisson arrival process.
- μ - Exponential service rate.
- n service agents.
- G - Patience distribution. Assume that the patience density exists at the origin and its value g_0 is strictly positive.

Random Arrival Rate: Let X be a random variable with cdf F , $E[X] = 0$, and finite $\sigma(X)$. Assume that the arrival rate varies from interval to interval in an i.i.d. fashion:

$$\Lambda = \lambda + \lambda^c X, \quad c \leq 1,$$

Model Definition

The $M^?|M|n + G$ Queue:

- λ - **Expected** arrival rate of a Poisson arrival process.
- μ - Exponential service rate.
- n service agents.
- G - Patience distribution. Assume that the patience density exists at the origin and its value g_0 is strictly positive.

Random Arrival Rate: Let X be a random variable with cdf F , $E[X] = 0$, and finite $\sigma(X)$. Assume that the arrival rate varies from interval to interval in an i.i.d. fashion:

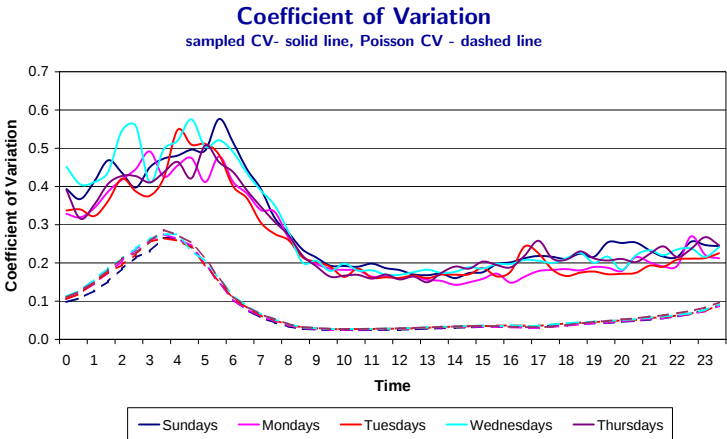
$$\Lambda = \lambda + \lambda^c X, \quad c \leq 1,$$

- $c \leq 1/2$: Conventional variability $\sim$ QED staffing regime.
- $1/2 < c < 1$: Moderate variability $\sim$ QED-c regime (**new**).
- $c = 1$: Extreme variability $\sim$ ED regime.

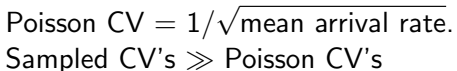
Financial Call Center: Data Description

- Israeli Bank.
- Arrival counts to the Retail queue are studied.
- 263 regular weekdays ranging between April 2007 and April 2008.
- Holidays with different daily patterns are excluded.
- Each day is divided into 48 half-hour intervals.

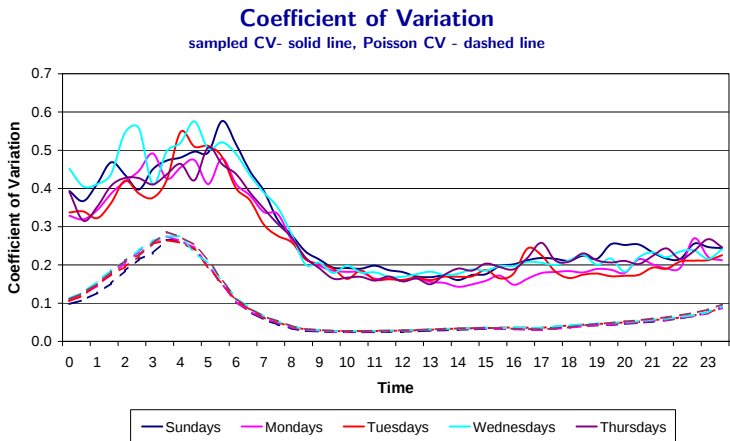
Financial Call Center: Over-Dispersion Phenomenon



sampled CV- solid line, Poisson CV - dashed line



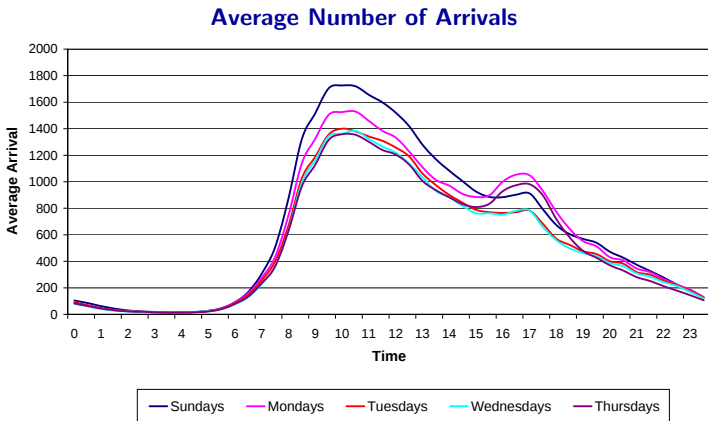
Financial Call Center: Over-Dispersion Phenomenon



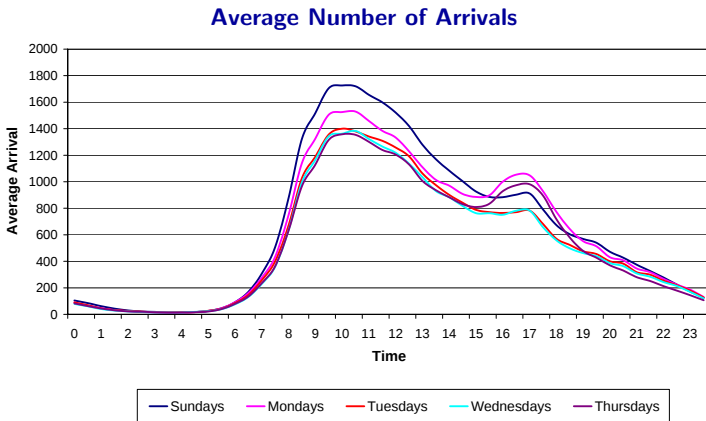
Poisson CV = $1/\sqrt{\text{mean arrival rate}}$.

Sampled CV's $\gg$ Poisson CV's $\Rightarrow$ **Over-Dispersion**

Financial Call Center: Arrival Rates



Financial Call Center: Arrival Rates



(1) Sundays;

(2) Mondays;

(3) Tuesdays and Wednesdays;

(4) Thursdays;

Financial Call Center:

Relation between Mean and Standard Deviation

Consider a Poisson mixture variable Y with random rate $\Lambda = \lambda + \lambda^c \cdot X$, where $E[X] = 0$, finite $\sigma(X) > 0$ and $1/2 < c \leq 1$. Then,

$$\text{Var}(Y) = \lambda^{2c} \cdot \text{Var}(X) + \lambda + \lambda^c \cdot E(X)$$

and

$$\lim_{\lambda \rightarrow \infty} (\ln(\sigma(Y)) - c \ln(\lambda)) = \ln(\sigma(X)).$$

Financial Call Center:

Relation between Mean and Standard Deviation

Consider a Poisson mixture variable Y with random rate $\Lambda = \lambda + \lambda^c \cdot X$, where $E[X] = 0$, finite $\sigma(X) > 0$ and $1/2 < c \leq 1$. Then,

$$\text{Var}(Y) = \lambda^{2c} \cdot \text{Var}(X) + \lambda + \lambda^c \cdot E(X)$$

and

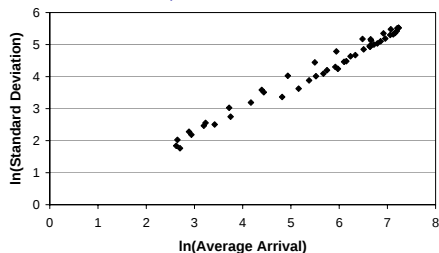
$$\lim_{\lambda \rightarrow \infty} (\ln(\sigma(Y)) - c \ln(\lambda)) = \ln(\sigma(X)).$$

Therefore, for large λ ,

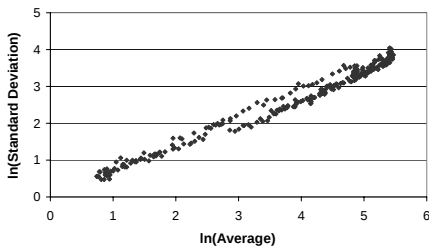
$$\ln(\sigma(Y)) \approx c \cdot \ln(\lambda) + \ln(\sigma(X)).$$

Financial Call Center: Fitting Regression Model

Tue-Wed, 30 min resolution

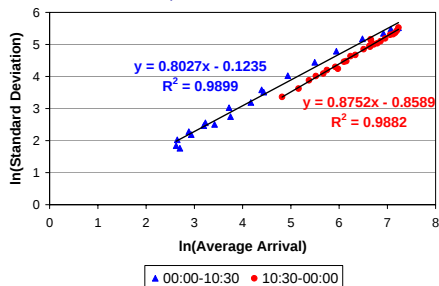


Tue-Wed, 5 min resolution

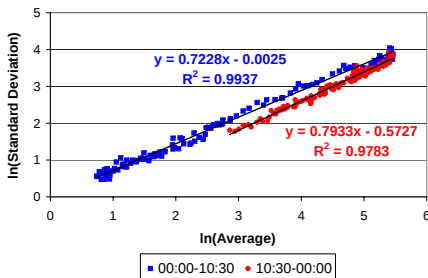


Financial Call Center: Fitting Regression Model

Tue-Wed, 30 min resolution



Tue-Wed, 5 min resolution



Results:

- Two clusters exist: midnight-10:30am and 10:30am-midnight.
- Very good fit ($R^2 > 0.97$).
- Significant linear relations for different weekdays and time-resolution (5-30 min):

$$\ln(\sigma(Y)) = c \cdot \ln(\lambda) + \ln(\sigma(X)).$$

Financial Call Center: Gamma Poisson Mixture Model

(Jongbloed and Koole ['01])

Assume a prior Gamma distribution for the arrival rate

$$\Lambda \stackrel{d}{=} \text{Gamma}(a, b),$$

and denote $E[\Lambda] \triangleq \lambda$.

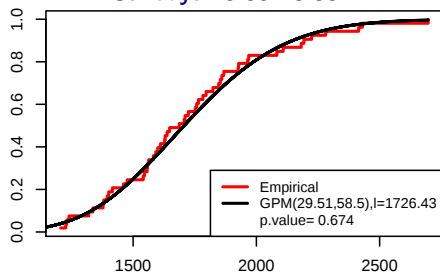
Then, the distribution of $Y \stackrel{d}{=} \text{Poisson}(\Lambda)$ is Negative Binomial.

- ① Maximum likelihood estimators of a and b .
- ② Goodness of fit test including FDR control method to correct the multiple comparisons.

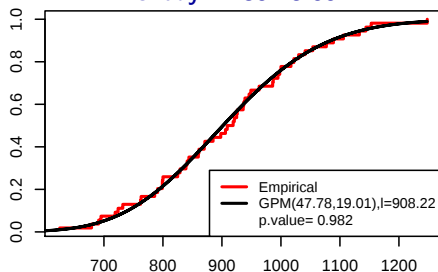
Financial Call Center: Gamma Poisson Mixture Model

$$H_0 : \Lambda_\lambda \stackrel{d}{=} \text{Gamma}(a_\lambda, b_\lambda)$$

Sundays 10:00-10:30



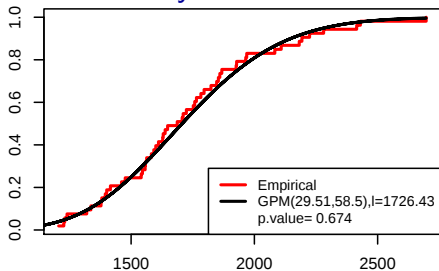
Monday 14:30-15:00



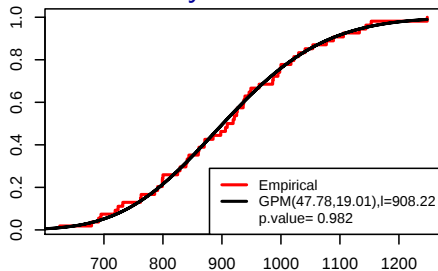
Financial Call Center: Gamma Poisson Mixture Model

$$H_0 : \Lambda_\lambda \stackrel{d}{=} \text{Gamma}(a_\lambda, b_\lambda)$$

Sundays 10:00-10:30



Monday 14:30-15:00



Results:

- Very good fit.
- Only 13 hypotheses are rejected (out of 192).

Financial Call Center: Outline of Additional Results

$$\lambda + \lambda^c \cdot X_\lambda \stackrel{d}{=} \text{Gamma}(a_\lambda, b_\lambda)$$

Under Gamma assumption and convergence of $\sigma(X_\lambda)$ to a constant $\sigma(X)$:

- Relation between our main model and Gamma Poisson mixture model is established.

Significant linear relations:

$$\ln(b_\lambda) = (2c - 1) \cdot \ln(\lambda) + \ln(\sigma^2(X)).$$

- Distribution of X is derived. It is **asymptotically normal** given $\lambda \rightarrow \infty$:

$$X_\lambda \xrightarrow{\mathcal{D}} \text{Norm}(0, \sigma^2(X))$$

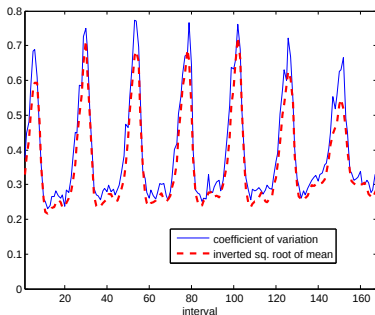
Emergency Department: Data Description

- Israeli Emergency Department.
- 194 weeks between from January 2004 till October 2007 (five war weeks are excluded from data).
- The analysis is performed using two resolutions: hourly arrival rates (168 intervals in a week) and three-hour arrival rates (56 intervals in a week).
- Holidays are not excluded (results with excluded holidays are similar).

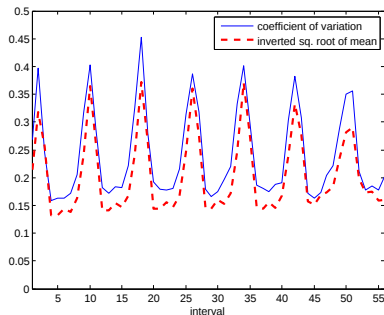
Emergency Department: Over-Dispersion Phenomenon

Coefficient of Variation

One-hour resolution



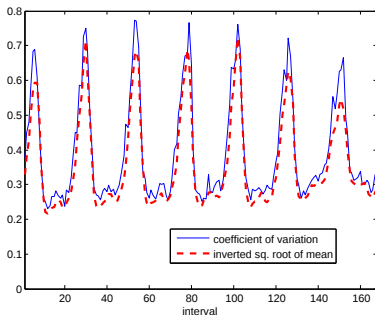
Three-hour resolution



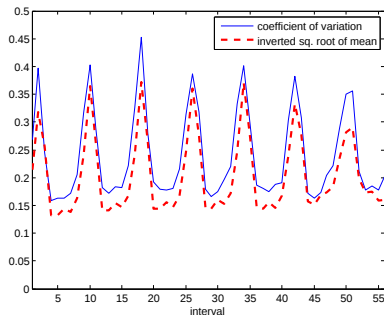
Emergency Department: Over-Dispersion Phenomenon

Coefficient of Variation

One-hour resolution



Three-hour resolution



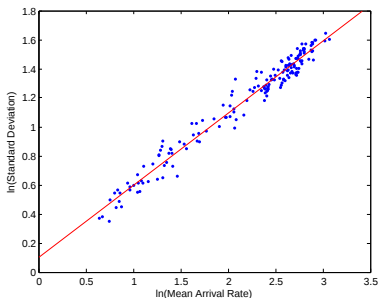
- Moderate over-dispersion.
- $c = 1/2$ seems to be reasonable assumption for hourly resolution.

Emergency Department: Fitting Regression Model

$\ln(\sigma)$ versus $\ln(\lambda)$ plots

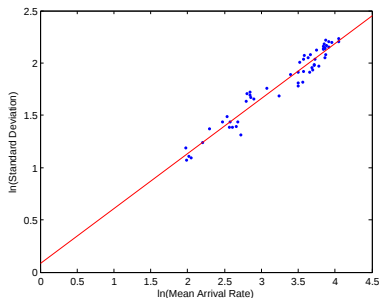
One-hour resolution

$$y = 0.497x + 0.102, R^2 = 0.968$$



Three-hour resolution

$$y = 0.527x + 0.087, R^2 = 0.947$$

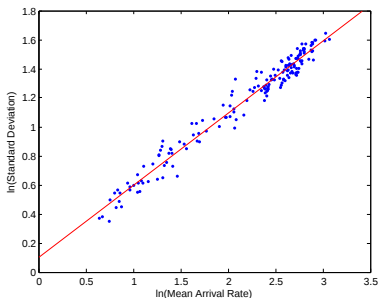


Emergency Department: Fitting Regression Model

$\ln(\sigma)$ versus $\ln(\lambda)$ plots

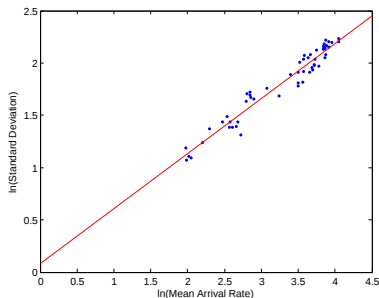
One-hour resolution

$$y = 0.497x + 0.102, R^2 = 0.968$$



Three-hour resolution

$$y = 0.527x + 0.087, R^2 = 0.947$$



- A linear pattern with the slope that is very close to 0.5 (derivation of asymptotic relation is based on $c > 1/2$).
- Overdispersion is observed at daily level $\Rightarrow$ scaling problem should be studied. (Dependence of c on the interval length.)

QED- c Regime: Fixed Arrival Rate

QED- c staffing rule:

$$n = \frac{\lambda}{\mu} + \beta \left(\frac{\lambda}{\mu} \right)^c + o(\sqrt{\lambda}), \quad \beta \in \mathbb{R}, \quad c \in (1/2, 1).$$

QED- c Regime: Fixed Arrival Rate

QED- c staffing rule:

$$n = \frac{\lambda}{\mu} + \beta \left(\frac{\lambda}{\mu} \right)^c + o(\sqrt{\lambda}), \quad \beta \in \mathbb{R}, \quad c \in (1/2, 1).$$

Assume an $M|M|n + G$ queue with **fixed arrival rate** λ .

Take λ to ∞ :

- $\beta > 0$: Over-staffing.
- $\beta < 0$: Under-staffing.

QED- c Regime: Fixed Arrival Rate

QED- c staffing rule:

$$n = \frac{\lambda}{\mu} + \beta \left(\frac{\lambda}{\mu} \right)^c + o(\sqrt{\lambda}), \quad \beta \in \mathbb{R}, \quad c \in (1/2, 1).$$

Assume an $M|M|n + G$ queue with **fixed arrival rate** λ .

Take λ to ∞ :

- $\beta > 0$: Over-staffing.
- $\beta < 0$: Under-staffing.

For both cases we provide asymptotically equivalent expressions (or bounds) for $P\{W_q > 0\}$, $P\{Ab\}$ and $E[V]$, where W_q - waiting time, V - offered wait (wait given infinite patience).

Proofs: based on $M|M|n + G$ building blocks from Zeltyn and Mandelbaum['05], carried out via the Laplace Method for asymptotic calculation of integrals.

QED-*c* Regime: Fixed Arrival Rate

Square-Root Staffing versus QED-*c* staffing

$R = \frac{\lambda}{\mu}$	β	SRS ¹ ($c = 1/2$)	QED- <i>c</i> Staffing					
			$c = 0.6$		$c = 0.75$		$c = 0.9$	
100	0.5	105	108	(+3%)	116	(+10%)	132	(+25%)
	1	110	116	(+5%)	132	(+20%)	163	(+48%)
	1.5	115	124	(+8%)	147	(+28%)	195	(+69%)
500	0.5	511	521	(+2%)	553	(+8%)	634	(+24%)
	1	522	542	(+4%)	606	(+16%)	769	(+47%)
	1.5	534	562	(+5%)	659	(+23%)	903	(+69%)
1000	0.5	1016	1032	(+2%)	1089	(+7%)	1251	(+23%)
	1	1032	1063	(+3%)	1178	(+14%)	1501	(+46%)
	1.5	1047	1095	(+5%)	1267	(+21%)	1752	(+67%)

¹Square-Root-Staffing: $n = R + \beta\sqrt{R}$

QED- c Regime: Random Arrival Rate

Theorem

Assume random arrival rate $\Lambda = \lambda + \lambda^c \mu^{1-c} X$, $c \in (1/2, 1)$, $E[X] = 0$, finite $\sigma(X) > 0$, and staffing according to the QED- c staffing rule with the corresponding c . Then, as $\lambda \rightarrow \infty$,

- a. Delay probability: $P_{\Lambda,n}\{W_q > 0\} \sim 1 - F(\beta)^a$
- b. Abandonment probability: $P_{\Lambda,n}\{Ab\} \sim \frac{E[X - \beta]_+}{n^{1-c}}$
- c. Average offered waiting time: $E_{\Lambda,n}[V] \sim \frac{E[X - \beta]_+}{n^{1-c} \cdot g_0}$

^a $f(\lambda) \sim g(\lambda)$ denotes that $\lim_{\lambda \rightarrow \infty} f(\lambda)/g(\lambda) = 1$.

Proofs: based on conditioning on values of X and results for QED- c staffing rule.

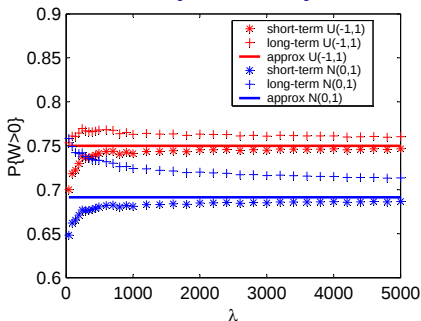
QED- c Regime: Numerical Experiments

Examples: Consider two distributions of X

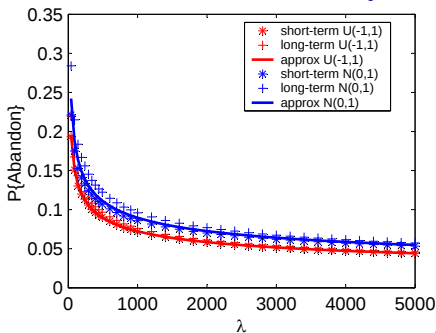
- Uniform distribution on $[-1,1]$,
- Standard Normal distribution.

$$\beta = -0.5, \quad c = 0.7$$

Delay Probability



Abandonment Probability



QED- c Regime: Practical Guidelines

- Determine "uncertainty coefficient c " via regression analysis.
- Check if Gamma Poisson mixture model is reasonable.
- Assume that X is asymptotically normal, calculate standard deviation from regression model.
- Apply our QED- c asymptotic results in order to determine appropriate staffing.

Constraint Satisfaction Problem

Formulation of the Problem:

Define the **optimal staffing level** by

$$n_{\lambda}^* = \operatorname{argmin}_n \{P_{\Lambda,n}\{W_q > 0\} \leq \alpha\}.$$

Constraint Satisfaction Problem

Formulation of the Problem:

Define the **optimal staffing level** by

$$n_{\lambda}^* = \operatorname{argmin}_n \{P_{\Lambda,n}\{W_q > 0\} \leq \alpha\}.$$

The staffing level n_{λ} is called **asymptotically feasible** if

$$\limsup_{\lambda \rightarrow \infty} P_{\Lambda,n_{\lambda}}\{W_q > 0\} \leq \alpha.$$

Constraint Satisfaction Problem

Formulation of the Problem:

Define the **optimal staffing level** by

$$n_{\lambda}^* = \operatorname{argmin}_n \{P_{\Lambda,n}\{W_q > 0\} \leq \alpha\}.$$

The staffing level n_{λ} is called **asymptotically feasible** if

$$\limsup_{\lambda \rightarrow \infty} P_{\Lambda,n_{\lambda}}\{W_q > 0\} \leq \alpha.$$

In addition, n_{λ} is **asymptotically optimal** if

$$|n_{\lambda}^* - n_{\lambda}| = o(f(\lambda)),$$

$f(\lambda)$ is defined separately for every special case..

ED Regime: $c = 1$, Discrete Random Arrival Rate

Assume $\Lambda = \lambda X$, where X is a discrete random variable which takes values $x_1 > x_2 > \dots > x_I > 0$, with probabilities $p_1, p_2, \dots, p_I$, respectively. In addition, let $E[X] = 1$, $\sigma(X) < \infty$ and $\lambda \rightarrow \infty$.

Theorem

a. The optimal staffing level satisfies

$$n^* = \frac{\lambda x_k}{\mu} + \beta^* \sqrt{\frac{\lambda x_k}{\mu}} + o(\sqrt{\lambda});$$

$$\text{where } k = \operatorname{argmin}_s \left\{ \sum_{i=1}^s x_i p_i \geq \alpha \right\}; \quad \alpha^* \triangleq \frac{\alpha - \sum_{i=1}^{k-1} x_i p_i}{x_k p_k};$$

and β^* is the unique solution of the equation

$$\alpha^* = \left[1 + \sqrt{\frac{g_0}{\mu}} \cdot \frac{h(\beta \sqrt{\mu/g_0})}{h(-\beta)} \right]^{-1}.$$

ED Regime: Staffing

Theorem

b. Introduce the staffing level

$$n_{\beta}^* = \left\lceil \frac{\lambda x_k}{\mu} + \beta^* \sqrt{\frac{\lambda x_k}{\mu}} \right\rceil.$$

Then the staffing level n_{β}^* is both *asymptotically feasible*, as well as *asymptotically optimal* with $f(\lambda) = \sqrt{\lambda}$.

ED Regime: Performance Measures

Theorem

Under the staffing level n_β^* , as $\lambda \rightarrow \infty$,

- a.** The abandonment probability:

$$P_{\Lambda, n_\beta^*} \{Ab\} \sim \sum_{i=1}^{k-1} p_i (x_i - x_k).$$

- b.** Mean server's utilization:

$$E_{\Lambda, n_\beta^*} [U] \sim \sum_{i=1}^k p_i + \sum_{i=k+1}^l p_i \cdot \frac{x_i}{x_k}.$$

- c.** Assume that for each $i < k$ the equation $G(y) = 1 - \frac{x_k}{x_i}$ has a unique solution y_i^* , and $g(y_i^*) > 0$. Then, the average waiting time

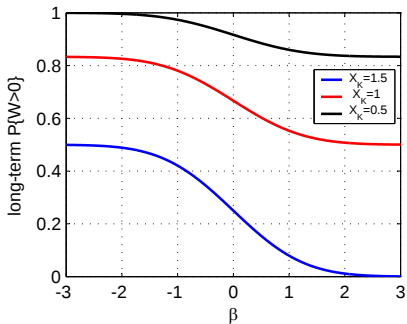
$$E_{\Lambda, n_\beta^*} [W_q] \sim \sum_{i=1}^{k-1} \left[x_i p_i \cdot \int_0^{y_i^*} \bar{G}(u) du \right].$$

ED Regime: Numerical Experiment

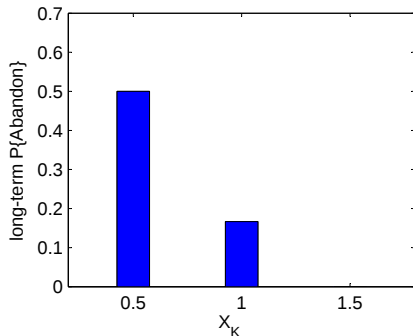
Example: X takes the values 1.5, 1 and 0.5, with probability 1/3 for each.

$$n = \left\lceil \frac{\lambda x_k}{\mu} + \beta \sqrt{\frac{\lambda x_k}{\mu}} \right\rceil$$

Delay Probability



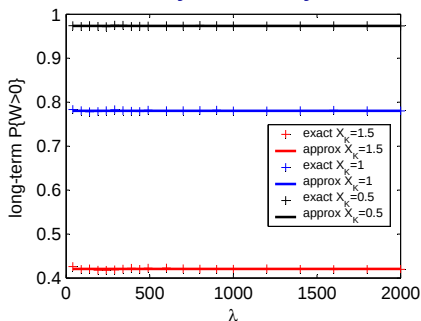
Abandonment Probability



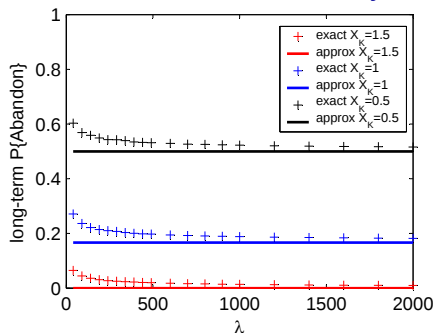
ED Regime: Numerical Experiment

$$\beta = -1$$

Delay Probability



Abandonment Probability



Outline of Additional Results

- **Queueing Theory.** Asymptotic performance measures derived and constraint satisfaction problems solved for:
 - QED regime ($c = 1/2$).
 - ED regime ($c = 1$), continuous distribution of X .
- **Numerical Experiments.** Very good fit between asymptotic results and the exact ones.

Time - Varying Queues

- **Iterative Staffing Algorithm (ISA)**, a simulation code developed by Feldman et al. [’07] with the features of random arrival rate in the time-varying $M|M|n + G$ queue.
- **Goal:** determine time-dependent staffing levels aiming to achieve a time-stable delay probability.

Time - Varying Queues

- **Iterative Staffing Algorithm (ISA)**, a simulation code developed by Feldman et al. [’07] with the features of random arrival rate in the time-varying $M|M|n + G$ queue.
- **Goal:** determine time-dependent staffing levels aiming to achieve a time-stable delay probability.
- **Example:** $c = 1$, Discrete Random Arrival Rate

① $\Lambda(t) = \lambda(t) \cdot X$, where

$$\lambda(t) = 100 - 20 \cdot \cos(t), \quad \text{and} \quad X = \begin{cases} 1.5 & w.p. \ 1/3 \\ 1 & w.p. \ 1/3 \\ 0.5 & w.p. \ 1/3 \end{cases}$$

- ② Service time and patience are distributed exponentially with mean 1 ($\mu = \theta = 1$).

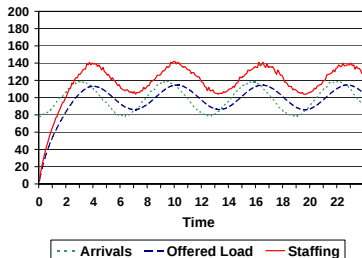
Time - Varying Queues

Arrivals, Offered Load and Staffing Level

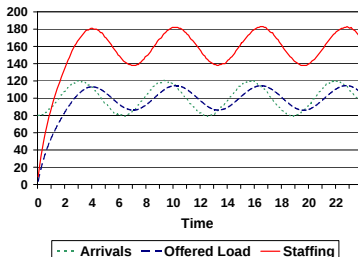
$$R_t = E \left[\int_{t-S}^t \lambda(u) du \right]$$

$$R_t^X = E \left[E \left[\int_{t-S}^t (\lambda(u) \cdot X) du \right] \right] = R_t.$$

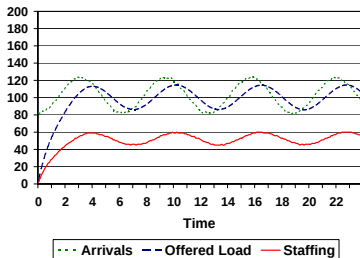
Target $\alpha=0.5$



Target $\alpha=0.1$

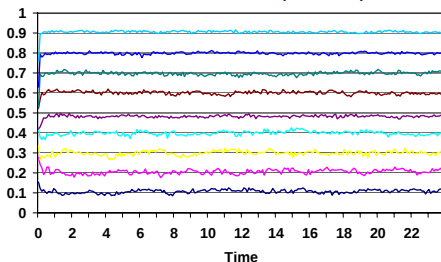


Target $\alpha=0.9$

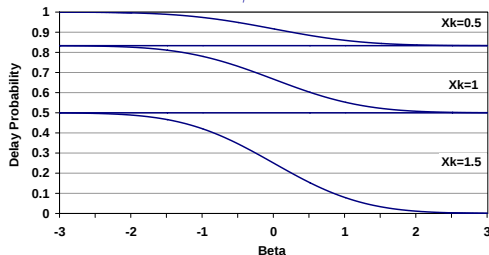


Time - Varying Queues

Delay Probability (Target)



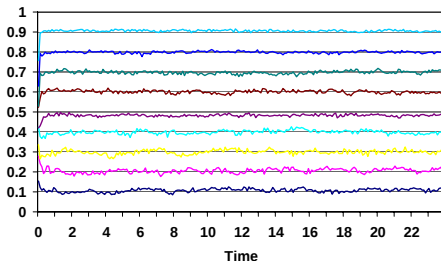
Theoretical Delay Probability vs. X_k and β



— Theoretical

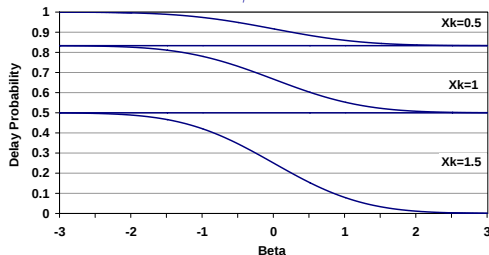
Time - Varying Queues

Delay Probability (Target)



Stable delay probability

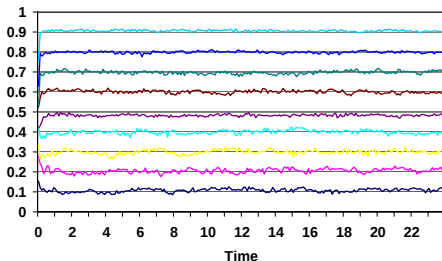
Theoretical Delay Probability vs. x_k and β



— Theoretical

Time - Varying Queues

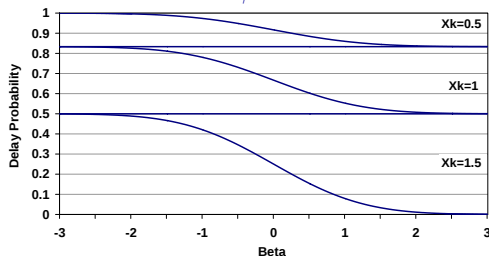
Delay Probability (Target)



$$n = R \cdot x_k + \beta \sqrt{R \cdot x_k}, \quad R = \lambda / \mu$$

Stable delay probability

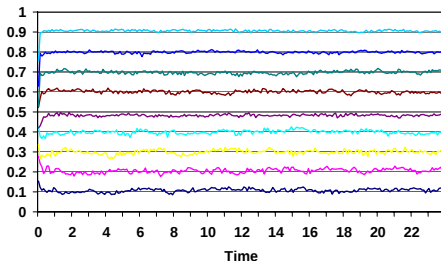
Theoretical Delay Probability vs. x_k and β



— Theoretical

Time - Varying Queues

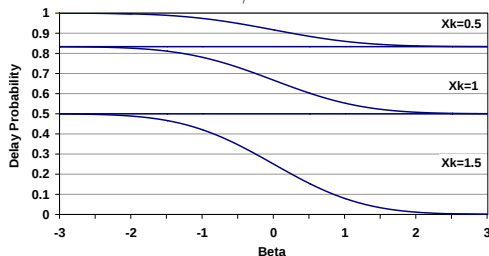
Delay Probability (Target)



$$n = R \cdot x_k + \beta \sqrt{R \cdot x_k}, \quad R = \lambda / \mu$$

$$n_t = R_t \cdot x_k + \beta_t \sqrt{R_t \cdot x_k} \quad ?$$

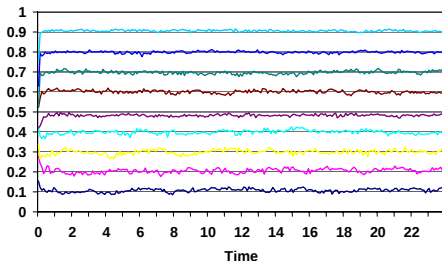
Theoretical Delay Probability vs. x_k and β



— Theoretical

Time - Varying Queues

Delay Probability (Target)



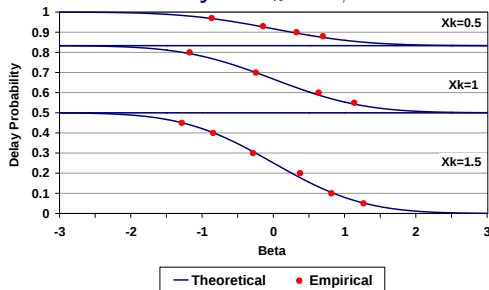
Stable delay probability

$$n = R \cdot x_k + \beta \sqrt{R \cdot x_k}, \quad R = \lambda / \mu$$

$$n_t = R_t \cdot x_k + \beta_t \sqrt{R_t \cdot x_k} \quad ?$$

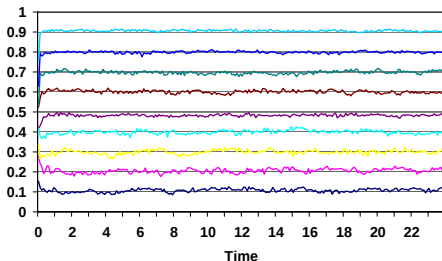
$$\text{Stable } \beta_t = \frac{n_t - R_t \cdot x_k}{\sqrt{R_t \cdot x_k}} = \beta$$

Theoretical and Empirical Delay Probability vs. x_k and β



Time - Varying Queues

Delay Probability (Target)



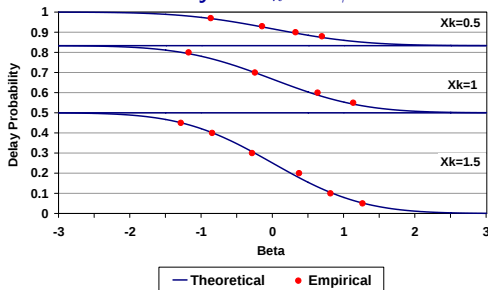
$$n = R \cdot x_k + \beta \sqrt{R \cdot x_k}, \quad R = \lambda / \mu$$

$$n_t = R_t \cdot x_k + \beta \sqrt{R_t \cdot x_k} \quad \checkmark$$

$$\text{Stable } \beta_t = \frac{n_t - R_t \cdot x_k}{\sqrt{R_t \cdot x_k}} = \beta$$

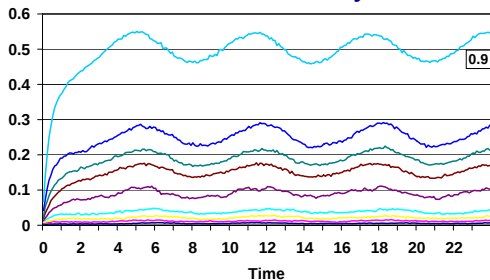
Stable delay probability

Theoretical and Empirical Delay Probability vs. x_k and β

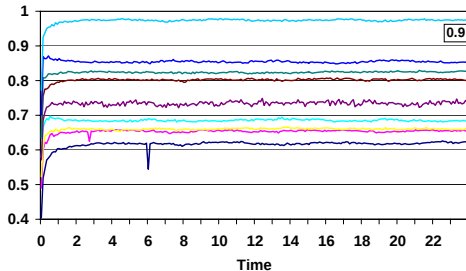


Time - Varying Queues

Abandonment Probability

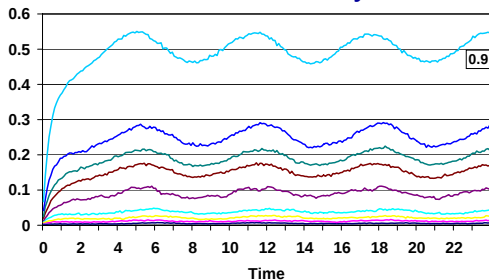


Servers' Utilization



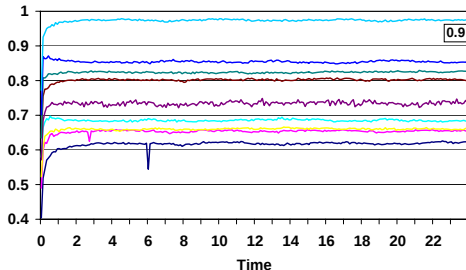
Time - Varying Queues

Abandonment Probability



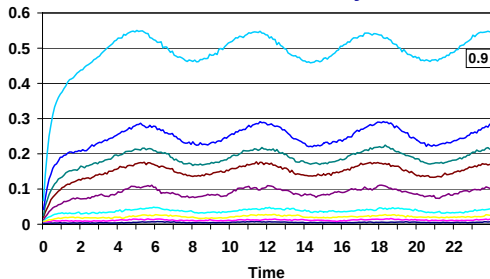
For large values of the target delay probability the abandon probability become less time-stable.

Servers' Utilization



Time - Varying Queues

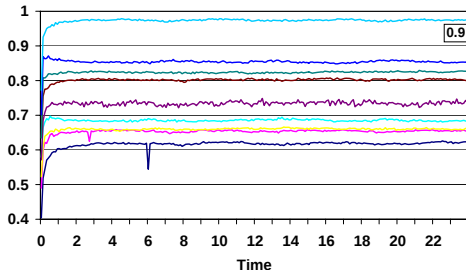
Abandonment Probability



For large values of the target delay probability the abandon probability become less time-stable.

Servers utilizations are found to be quite stable.

Servers' Utilization



Future Research Challenges

- Incorporating **forecasting errors** into our model (in the spirit of Steckley et al., 2007).
- **Scaling problem:** dependence of c on the basic interval duration.
- **ISA:** achieving time-stable performance measures (probability to abandon, average wait).
- Validation of $M^?|M|n+M$ (or $M^?|M|n+G$) model in call center environment (and probably other service systems).

Thank You

