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

Shimrit Maman, Technion
Avishai Mandelbaum, Technion
Sergey Zeltyn, IBM Research Lab, Haifa

ORSIS Annual Meeting, Herzliya, May 10, 2009

Contents of Talk

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:

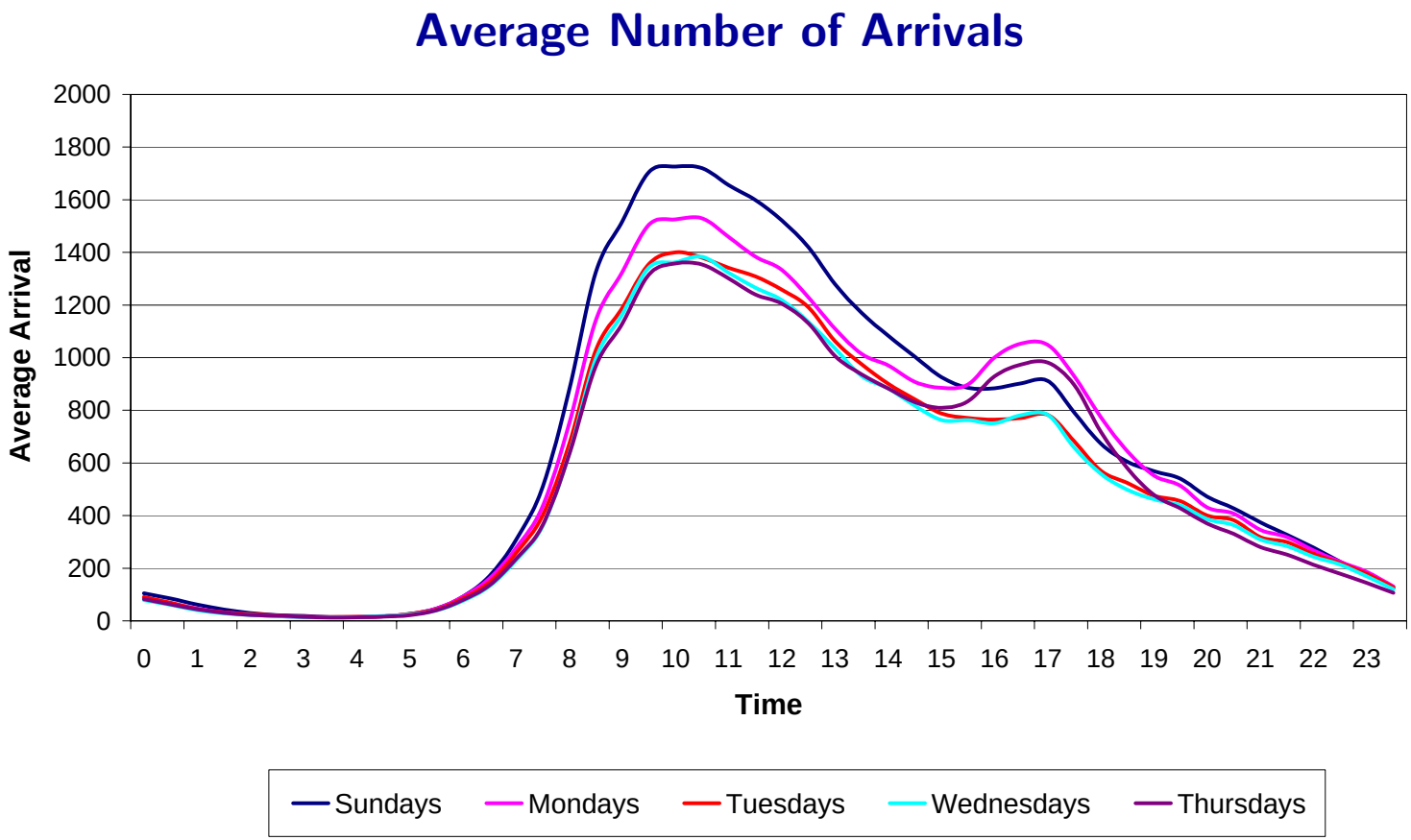
$$\Lambda = \lambda + \lambda^c X, \quad c \leq 1,$$

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

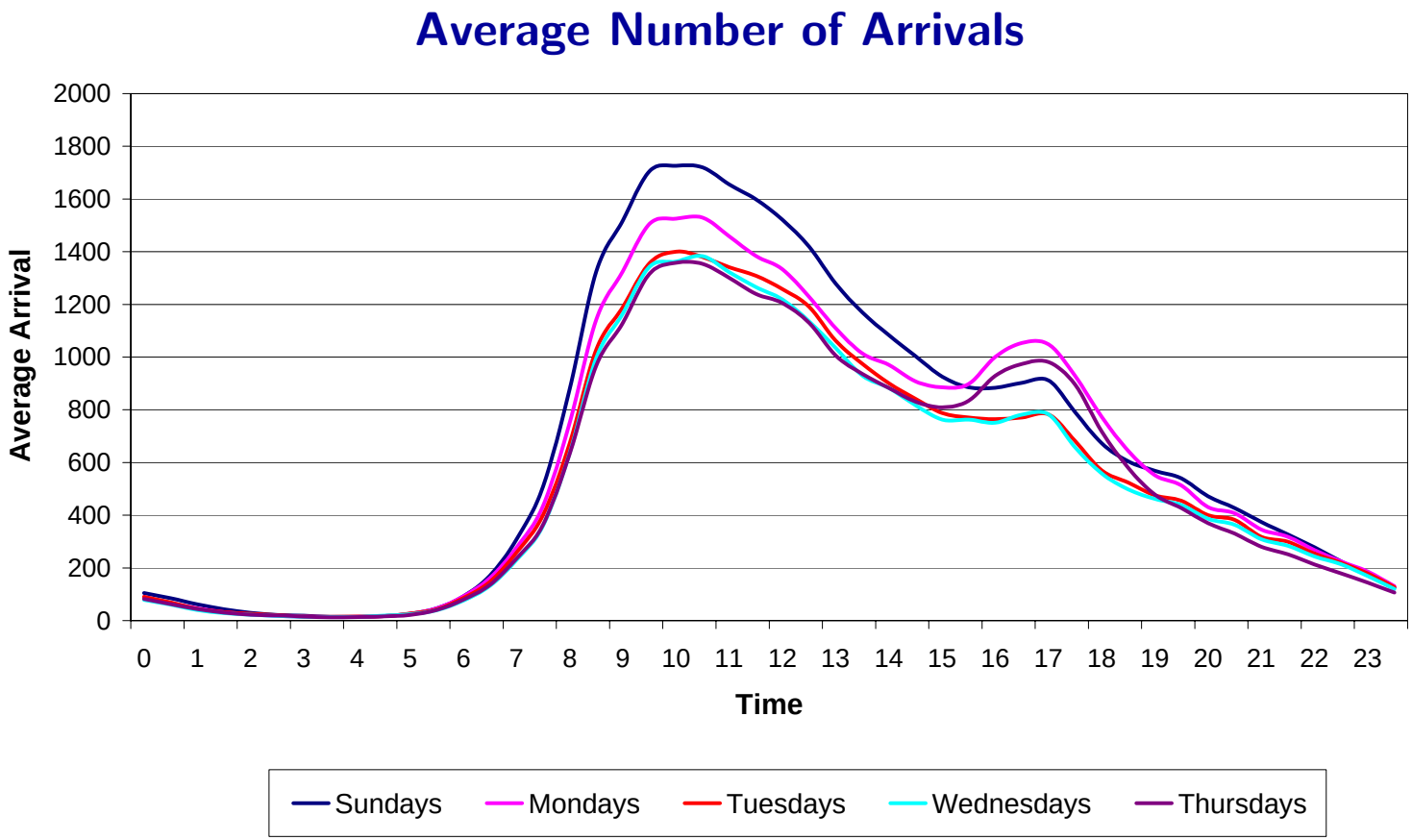
Financial Call Center: Data Description

- 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: Arrival Rates

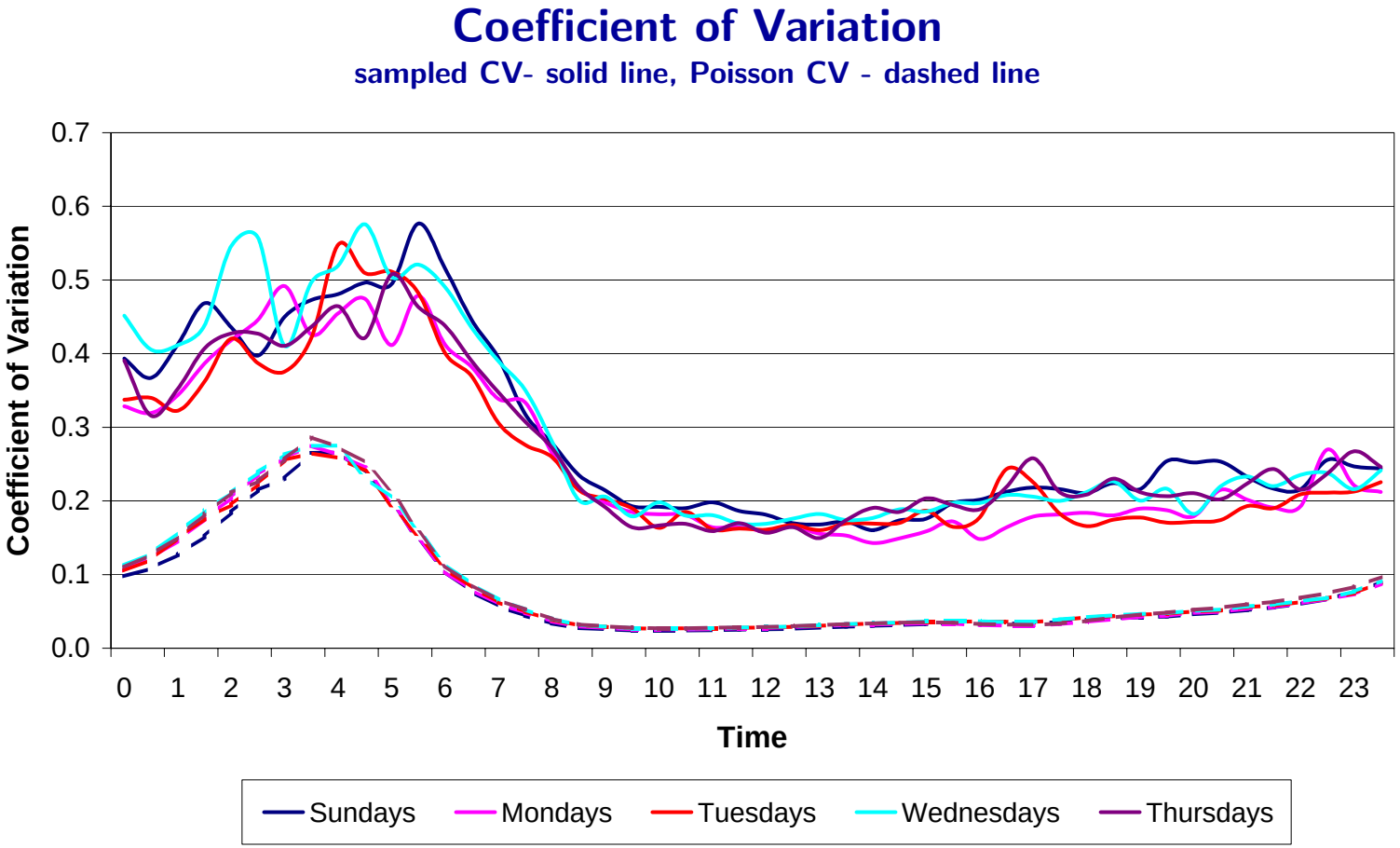


Financial Call Center: Arrival Rates



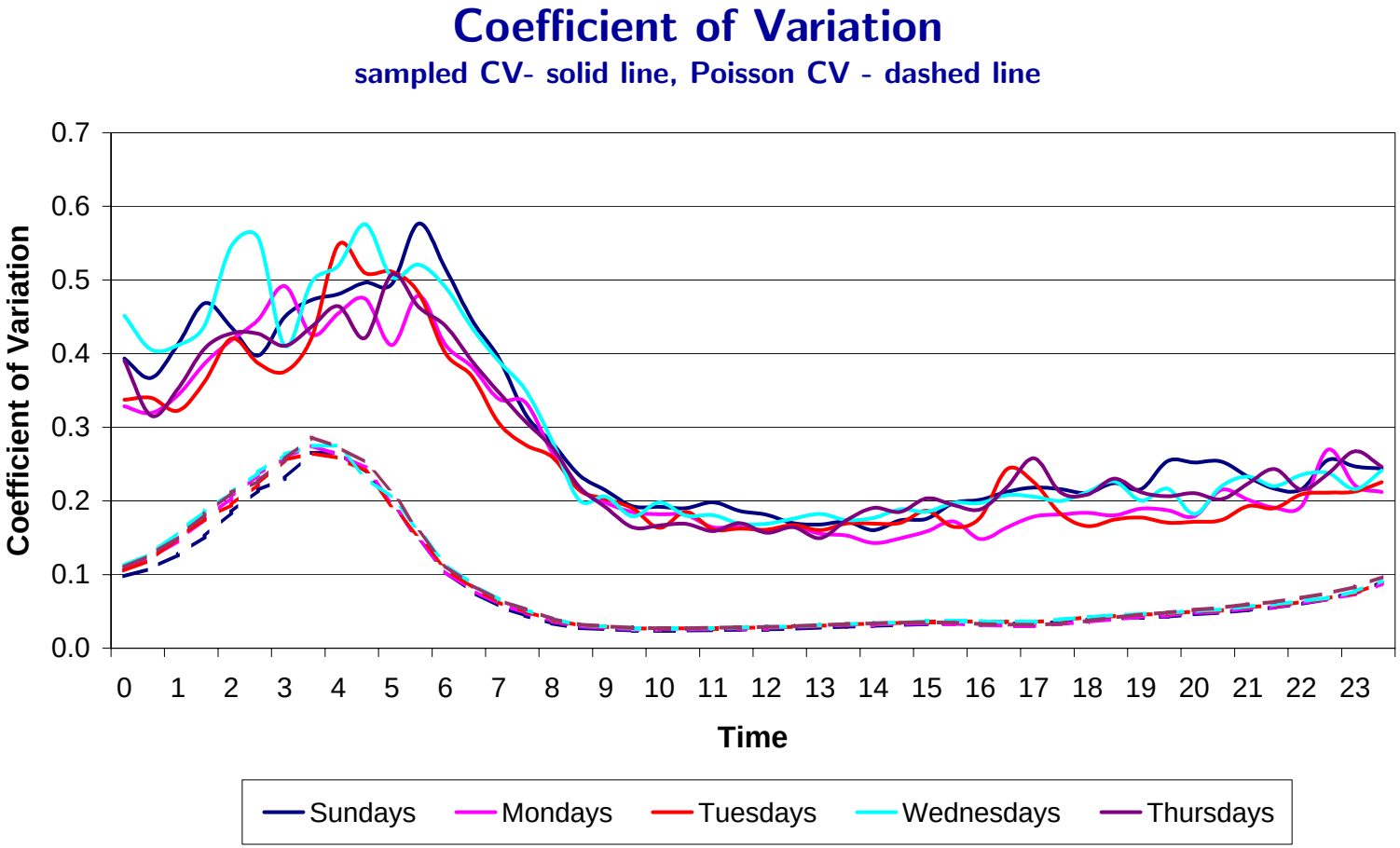
(1) Sundays;
(2) Mondays;
(3) Tuesdays and Wednesdays;
(4) Thursdays.

Financial Call Center: Over-Dispersion Phenomenon



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

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: 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,

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

and

$$\lim_{\lambda \to \infty} \left(\ln(\sigma(Y)) - c \ln(\lambda) \right) \;=\; \ln(\sigma(X)).$$

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,

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

and

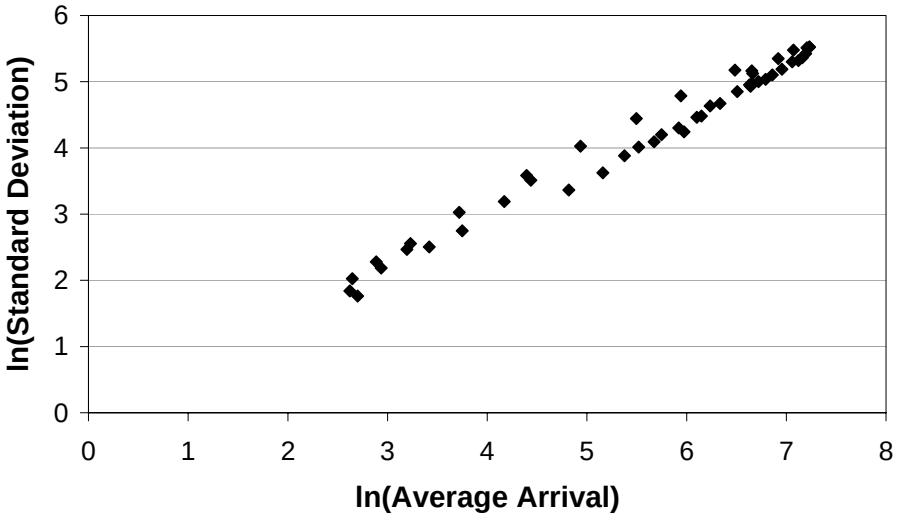
$$\lim_{\lambda \to \infty} \left(\ln(\sigma(Y)) - c \ln(\lambda) \right) \;=\; \ln(\sigma(X)).$$

Therefore, for large λ,

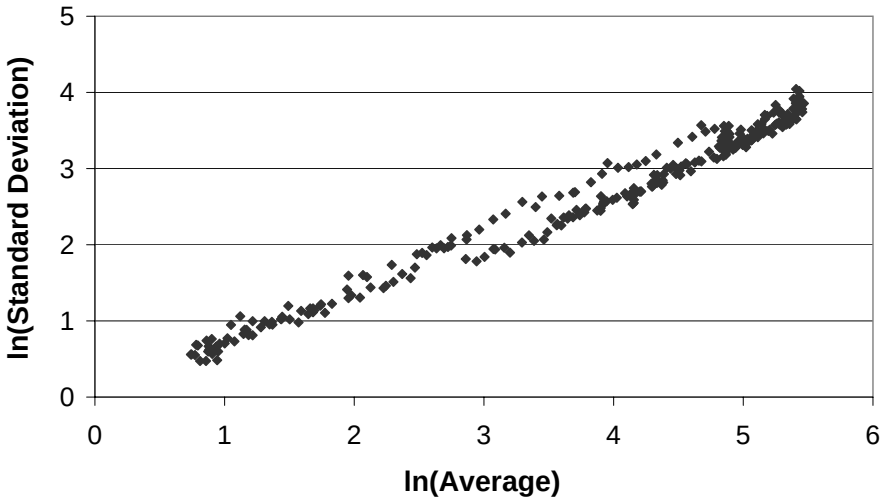
$$\mathbf{\ln(\sigma(Y)) \;\approx\; c \cdot \ln(\lambda) + \ln(\sigma(X)).}$$

Financial Call Center: Fitting Regression Model

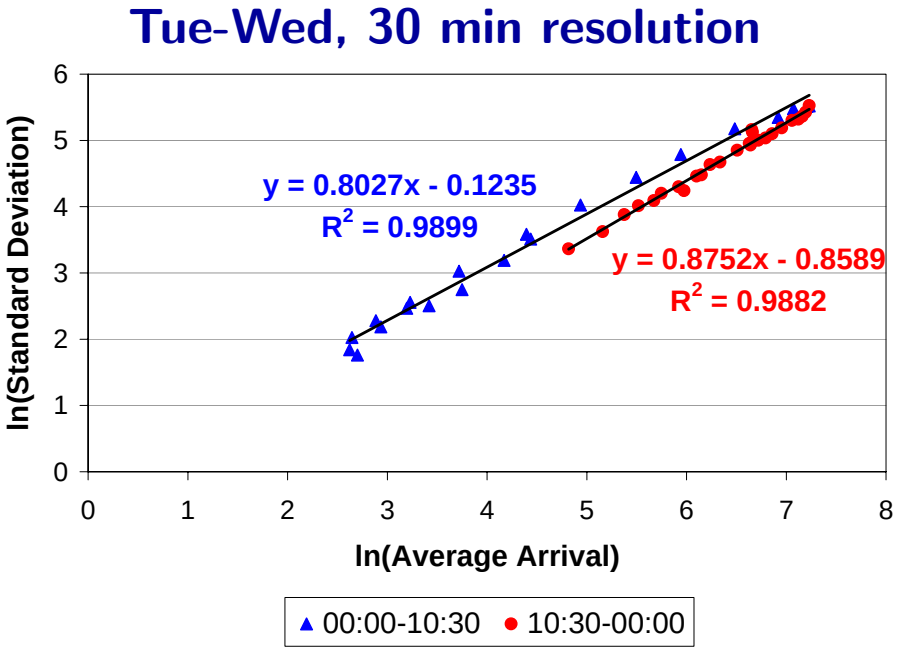
Tue-Wed, 30 min resolution

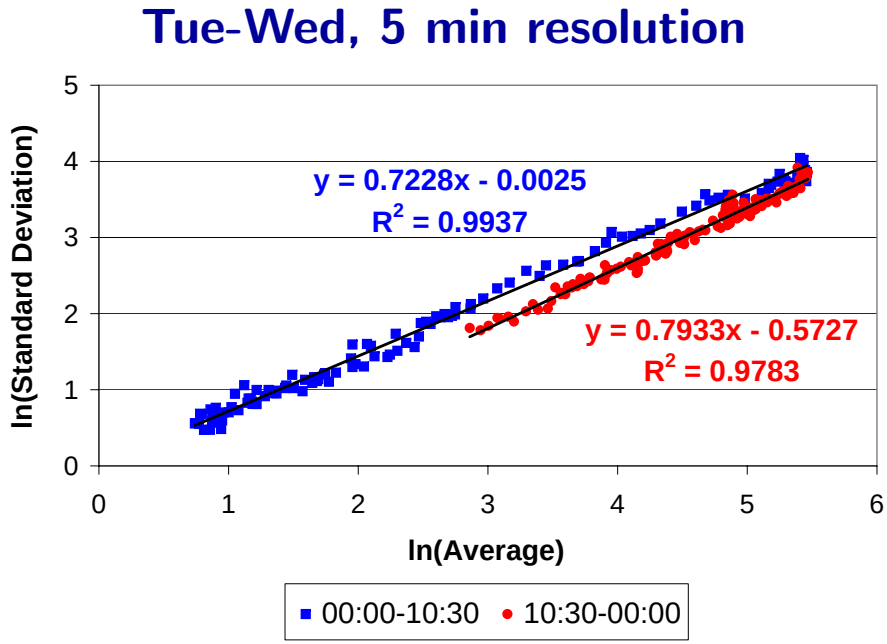


Tue-Wed, 5 min resolution



Financial Call Center: Fitting Regression Model





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: Outline of Additional Results

- **Fitting a Gamma Poisson mixture model to the data:** (Jongbloed and Koole ['01])

  Assume a prior Gamma distribution for the arrival rate $\mathbf{\Lambda \stackrel{d}{=} Gamma(a, b)}$. Then, the distribution of $Y \stackrel{d}{=} Poiss(\Lambda)$ is Negative Binomial.

Financial Call Center: Outline of Additional Results

- **Fitting a Gamma Poisson mixture model to the data:** (Jongbloed and Koole ['01])

  Assume a prior Gamma distribution for the arrival rate $\mathbf{\Lambda \stackrel{d}{=} Gamma(a, b)}$. Then, the distribution of $Y \stackrel{d}{=} Poiss(\Lambda)$ is Negative Binomial.

- Good fit of Gamma Poisson mixture model to data of the Financial Call Center for most intervals.

Financial Call Center: Outline of Additional Results

- **Fitting a Gamma Poisson mixture model to the data:** (Jongbloed and Koole ['01])

  Assume a prior Gamma distribution for the arrival rate $\mathbf{\Lambda \stackrel{d}{=} Gamma(a, b)}$. Then, the distribution of $Y \stackrel{d}{=} Poiss(\Lambda)$ is Negative Binomial.

- Good fit of Gamma Poisson mixture model to data of the Financial Call Center for most intervals.

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

Financial Call Center: Outline of Additional Results

- **Fitting a Gamma Poisson mixture model to the data:** (Jongbloed and Koole ['01])

  Assume a prior Gamma distribution for the arrival rate $\mathbf{\Lambda \stackrel{d}{=} Gamma(a, b)}$. Then, the distribution of $Y \stackrel{d}{=} Poiss(\Lambda)$ is Negative Binomial.

- Good fit of Gamma Poisson mixture model to data of the Financial Call Center for most intervals.

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

- The distribution of X is derived under Gamma assumption: it is **asymptotically normal**, given $\lambda \longrightarrow \infty$.

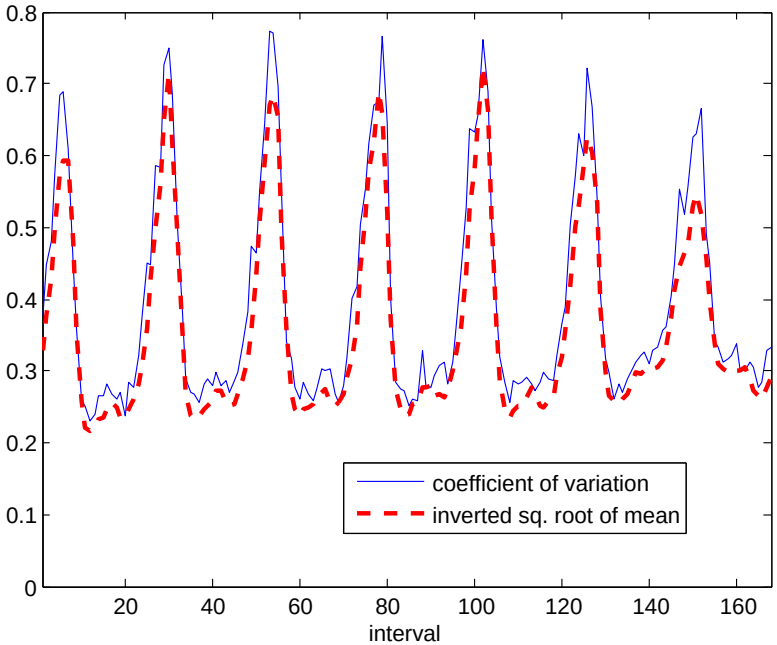
Emergency Department: Data Description

- 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).

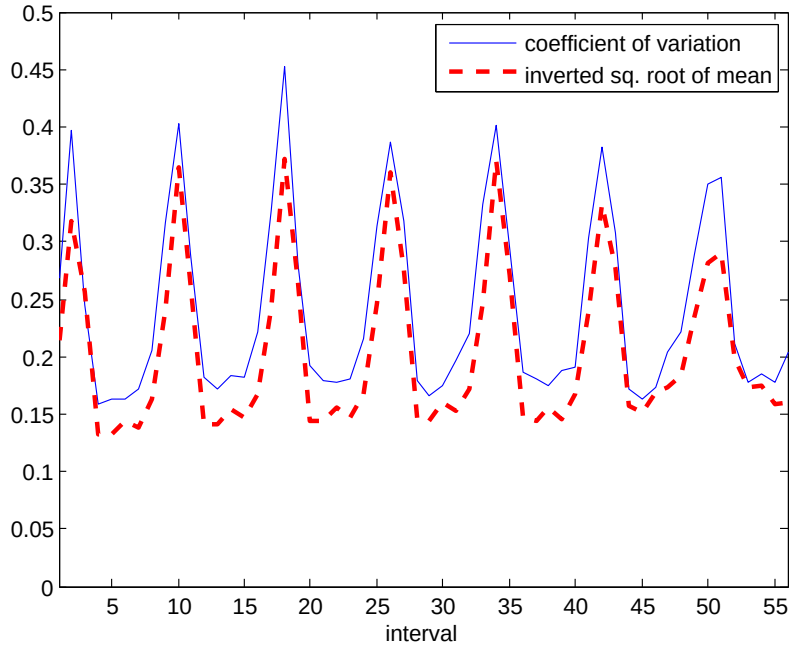
Emergency Department: Over-Dispersion Phenomenon

Coefficient of Variation

One-hour resolution



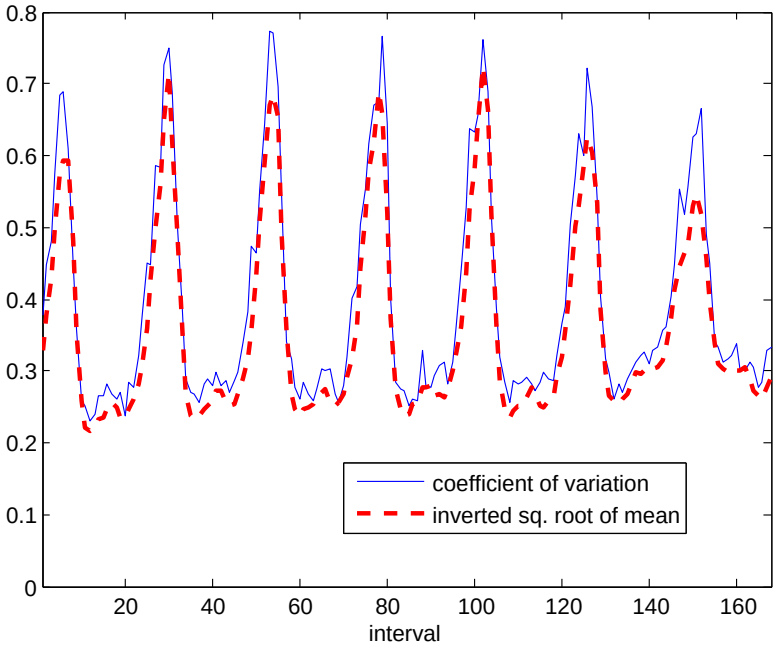
Three-hour resolution



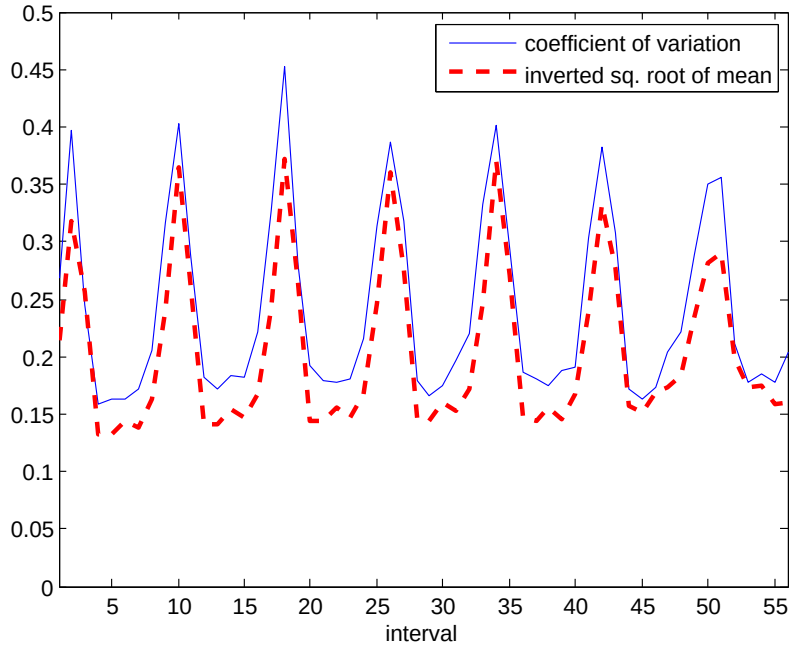
Emergency Department: Over-Dispersion Phenomenon

Coefficient of Variation

One-hour resolution



Three-hour resolution

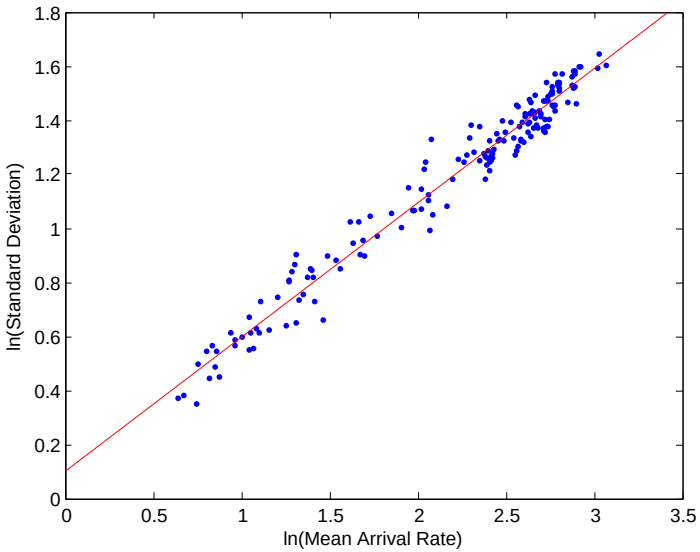


- Moderate over-dispersion.
- Overdispersion is observed at daily level $\Rightarrow$ scaling problem should be studied. (Dependence of c on the interval length.)
- $c = 1/2$ seems to be reasonable assumption for hourly resolution.

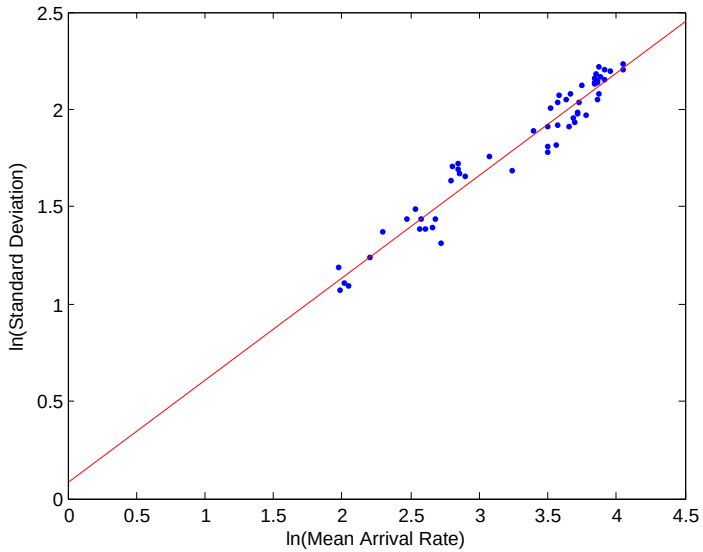
Emergency Department: Fitting Regression Model

Linear Regression: $\ln(\sigma(Y)) = c \cdot \ln(\lambda) + \ln(\sigma(X))$

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 while derivation of the asymptotic relation is based on $c > 1/2$.

Emergency Department: Fitting Regression Model

Non-Linear Regression: $\ln(\sigma(Y)) = 0.5 \cdot \ln(\lambda^{2c}\sigma^2(X) + \lambda)$

	One-hour resolution		Three-hour resolution	
	$\hat{c}$	$\hat{\sigma}(X)$	$\hat{c}$	$\hat{\sigma}(X)$
Linear Regression	0.497	1.108	0.527	1.087
Non-Linear Regression	0.481	0.476	0.595	0.466

- Estimates for c are close but the estimates for $\sigma(X)$ are significantly higher in the case of linear regression.
- The two regression curves are almost identical.

QED-c Regime: Fixed Arrival Rate

QED-c staffing rule:

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

QED-c Regime: Fixed Arrival Rate

QED-c staffing rule:

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

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:

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

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: Random Arrival Rate

Theorem

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

a. Delay probability: $P_{\Lambda,n}\{W_q > 0\} \sim 1 - F(\beta)$.[a]

b. Abandonment probability: $P_{\Lambda,n}\{Ab\} \sim \dfrac{E[X-\beta]_+}{n^{1-c}}$.

c. Average offered waiting time: $E_{\Lambda,n}[V] \sim \dfrac{E[X-\beta]_+}{n^{1-c} \cdot g_0}$.

[a] $f(\lambda) \sim g(\lambda)$ denotes that $\lim_{\lambda\to\infty} f(\lambda)/g(\lambda) = 1$.

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

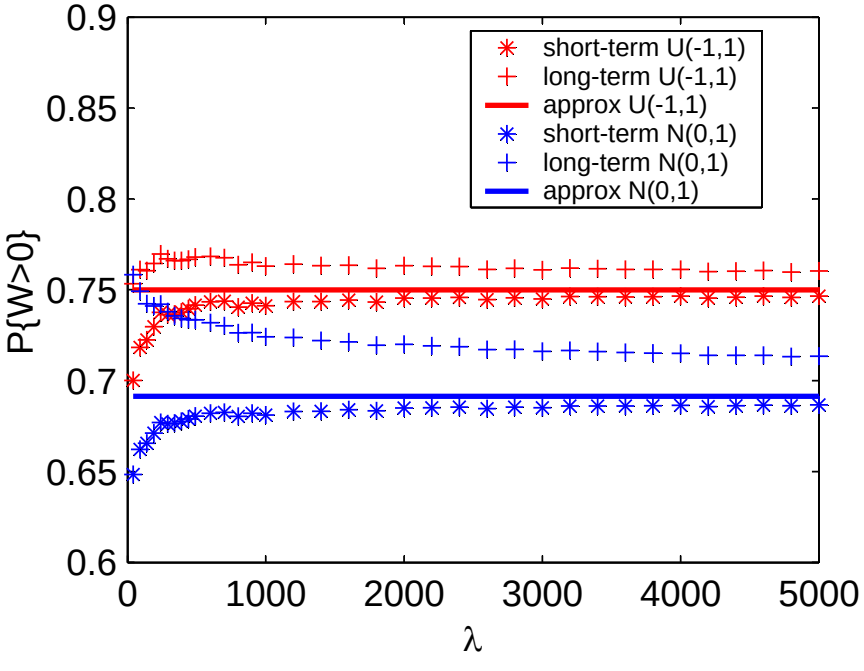
QED-c Regime: Numerical Experiments

Examples: Consider two distributions of X

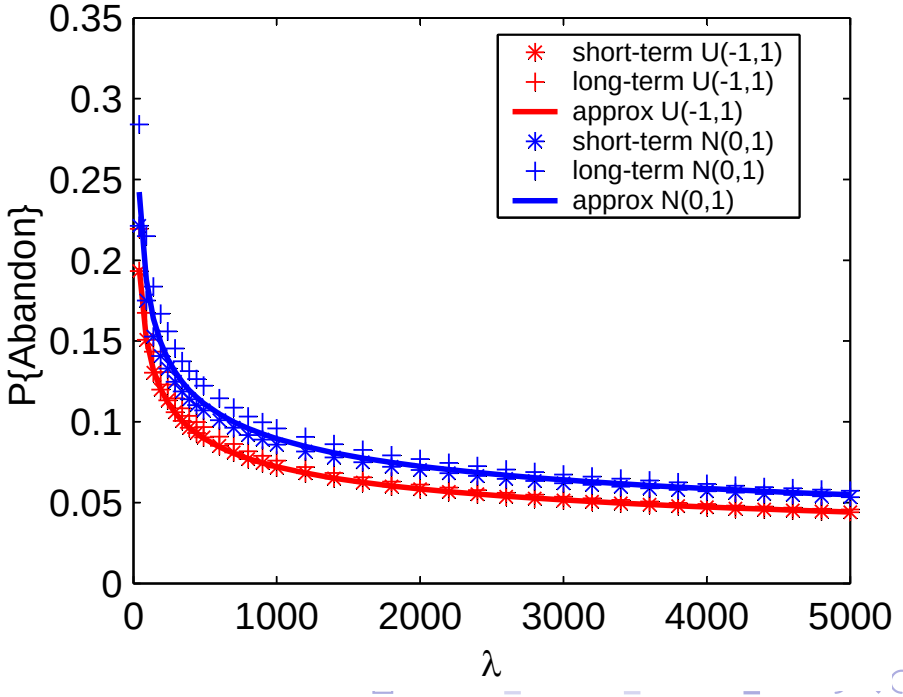
- Uniform distribution on [-1,1],
- Standard Normal distribution.

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

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.

Outline of Additional Results

- **Queueing Theory.** Asymptotic performance measures derived and constraint satisfaction problems solved for:
  - QED regime ($c = 1/2$).
  - ED regime ($c = 1$), discrete and continuous distribution of X.

Outline of Additional Results

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

Outline of Additional Results

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

Future Research Challenges

- Incorporating **forecasting errors** into our model (in the spirit of Steckley et al. ['07]).
- **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