Queueing Models with Uncertain Arrival Rate: Theory and Validation

Shimrit Maman, Rafael
Avishai Mandelbaum, Faculty of Industrial Engineering and Management, Technion
Sergey Zeltyn, IBM Research Lab, Haifa

Stockholm, APS Conference

July 7, 2010

Motivation

Standard assumption in service system modeling: arrival process is Poisson with known parameters.
Emergency departments and call centers: known arrival rates for each basic interval (say, one hour in EDs, 15 min in CCs).

Application of standard approach to basic interval (say, next Tuesday, 9am-10am):

- Derive Poisson parameters from historical data and some forecasting procedure.
- Plug parameters into a queueing model (Erlang-C, Erlang-A, Queueing Network, Skills-Based Routing models, ...).
- Set staffing levels according to model and service constraint (e.g., 80% of CC customers answered within 30 sec).

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 a relevant queueing model (M/M/n+G).
- Derive asymptotic results relevant for real-life staffing problems, assess quality of approximations.
- Validate arrival model via data analysis. Develop parameter estimation methods.
- Check model scalability with respect to the basic interval length (ongoing).

Related Work

Steckley, Henderson and Mehrotra

Forecast Errors in Service Systems. 2007.

Koole and Jongbloed

Managing uncertainty in call centers using poisson mixtures. 2001.

Halfin and Whitt

Heavy-traffic Limits for Queues with many Exponential Servers. 1981.

Zeltyn and Mandelbaum

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

Feldman, Mandelbaum, Massey and Whitt

Staffing of Time-Varying Queues to Achieve Time-Stable Performance. 2008

Kim, Lee and Sung

A shifted Gamma Distribution Model for Long-Range Dependent Internet Traffic. 2003.

Queue with Overdispersed Arrival Rate: 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: time that a customer is willing to wait in queue.

Random Poisson Arrival Rate:

$$M = \lambda + \lambda^c X, \quad 1/2 \leq c \leq 1,$$

where X is a random variable with zero mean and finite variance.

- $c = 1/2$: Conventional variability $\sim$ QED staffing regime.
- $1/2 < c < 1$: Moderate variability $\sim$ **QED-c regime**.
- $c = 1$: Extreme variability $\sim$ Efficiency-Driven regime.

QED-c Regime: Fixed Arrival Rate

QED-c staffing rule:

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

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 $P\{W_q > 0\}$, $P\{Ab\}$, $E[W_q]$ and $E[V]$, where W_q - waiting time, V - offered wait.

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 $M = \lambda + \lambda^c \mu^{1-c} X, \quad c \in (1/2, 1)$, $E[X] = 0$, 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: $$P_{M,n}\{W_q > 0\} \sim 1 - F(\beta).$$

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

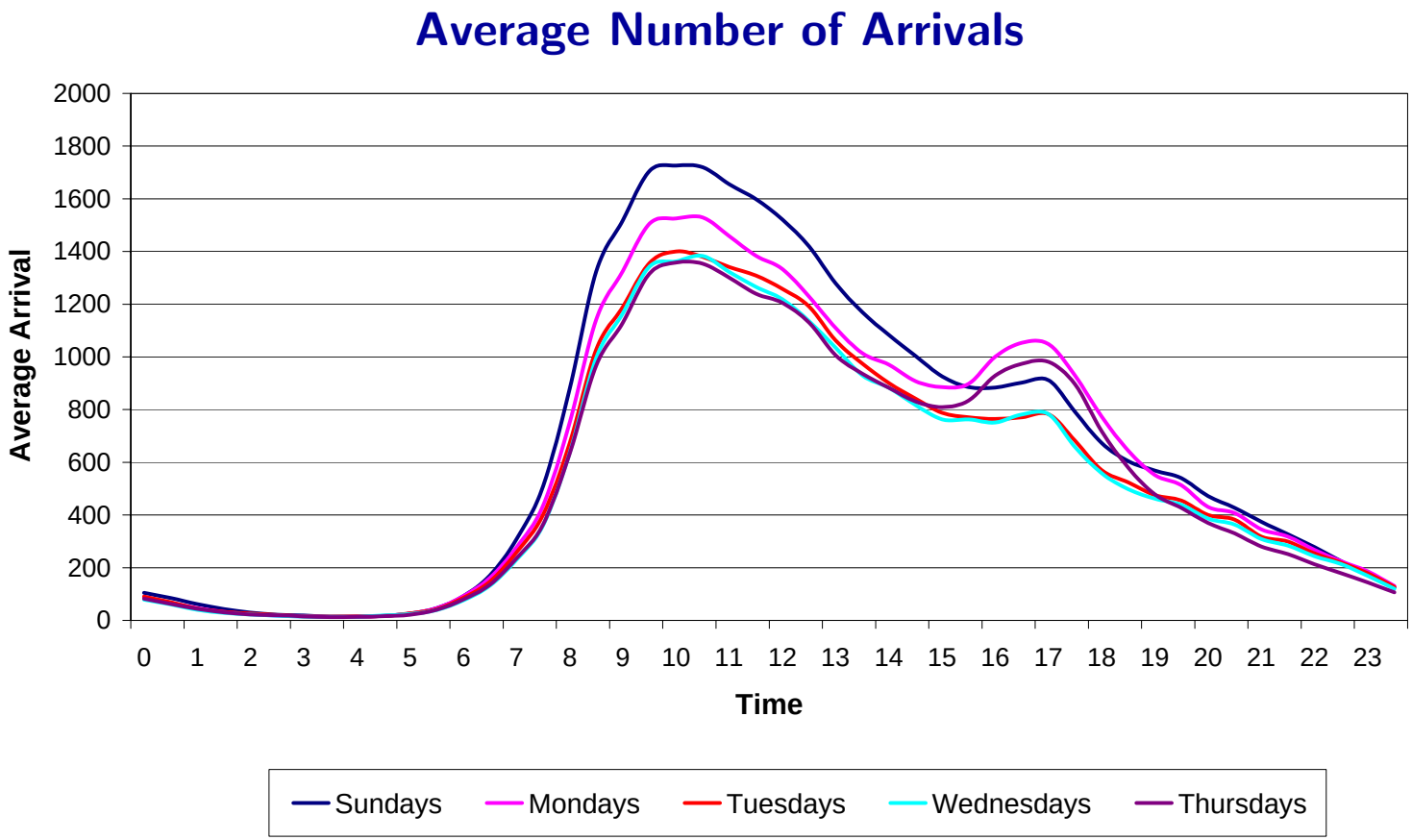
c. Average waiting time and offered wait:
$$E_{M,n}[W_q] \sim E_{M,n}[V] \sim \frac{E[X-\beta]_+}{n^{1-c} \cdot g_0},$$ where g_0=patience density at the origin.

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

Financial Call Center: Data Description

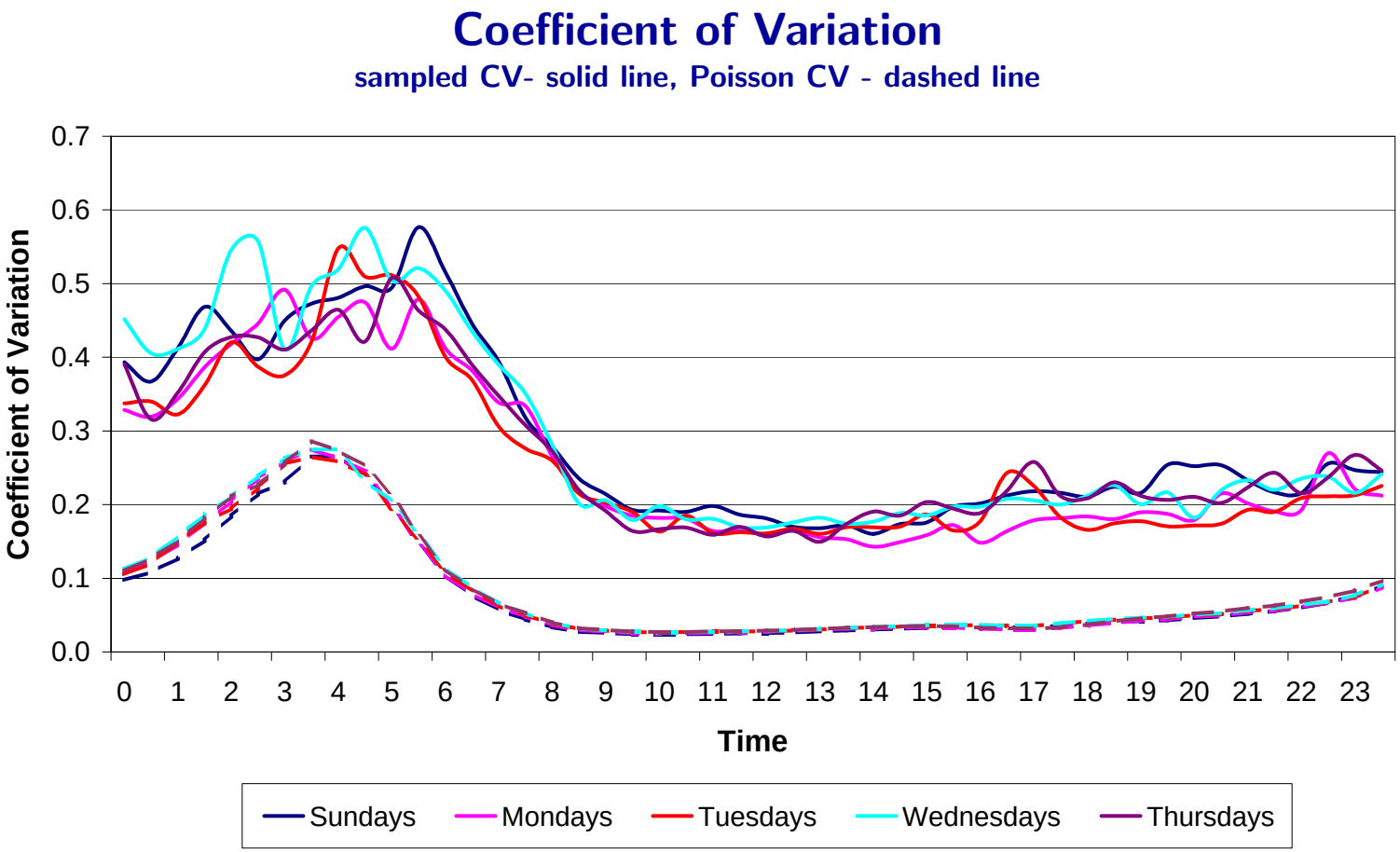
- Israeli bank.
- Arrival counts to the Retail queue are studied.
- 263 regular weekdays ranging between April 2007 and April 2008.
- Holidays with different daily patterns are excluded.
- Each day is divided into 48 half-hour intervals (or 288 five-minute intervals).

Financial Call Center: Arrival Rates



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

Financial Call Center: Overdispersion Phenomenon



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

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

Moderate and Extreme Variability: Relation between Mean and Standard Deviation of Arrival Rate

Number of arrivals during a basic interval (say, Tue, 9-9:30am): Poisson Y with random rate $M = \lambda + \lambda^c \cdot X$, where $E(X) = 0$, standard deviation $\sigma(X) > 0$ and $1/2 < c \leq 1$. Then,

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

Proof:

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

$$\sigma(Y)/\lambda^c = \sqrt{\sigma^2(X) + \lambda^{1-2c}},$$

and

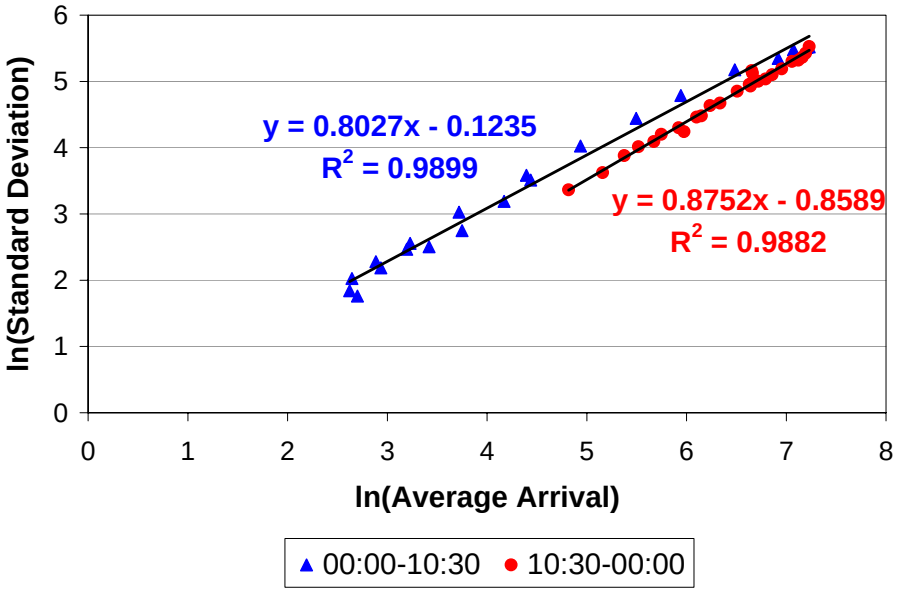
$$\lim_{\lambda \to \infty} (\ln(\sigma(Y)) - c \ln(\lambda)) = \ln(\sigma(X)).$$

Therefore, for large λ,

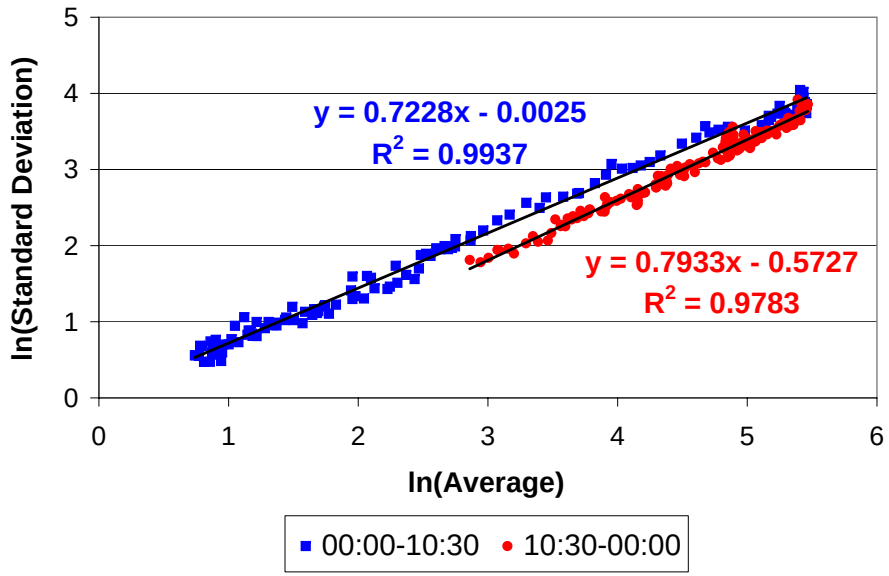
$$\ln(\sigma(Y)) \approx c \cdot \ln(\lambda) + \ln(\sigma(X)).$$

Financial Call Center: Fitting Regression Model

Tue-Wed, 30 min resolution



Tue-Wed, 5 min resolution



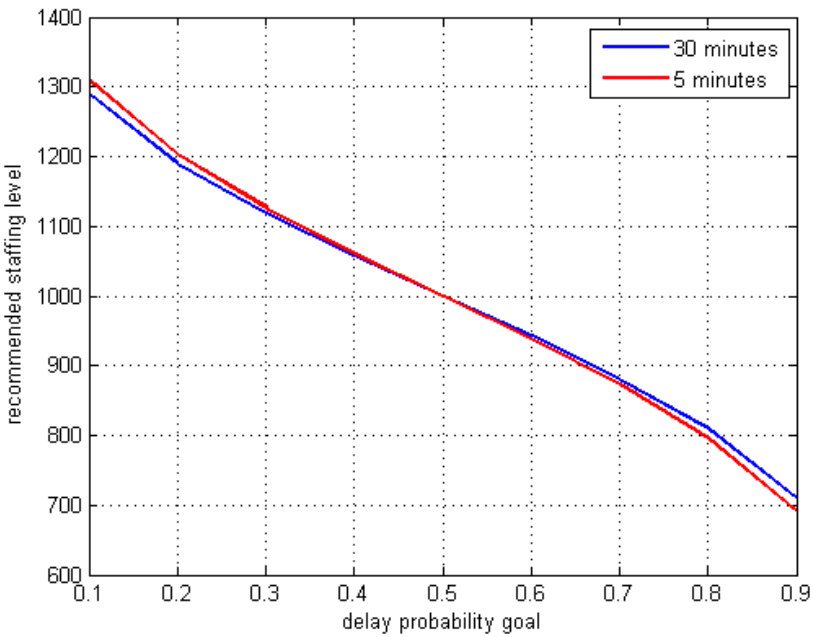
Results:

- Two clusters exists: 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):

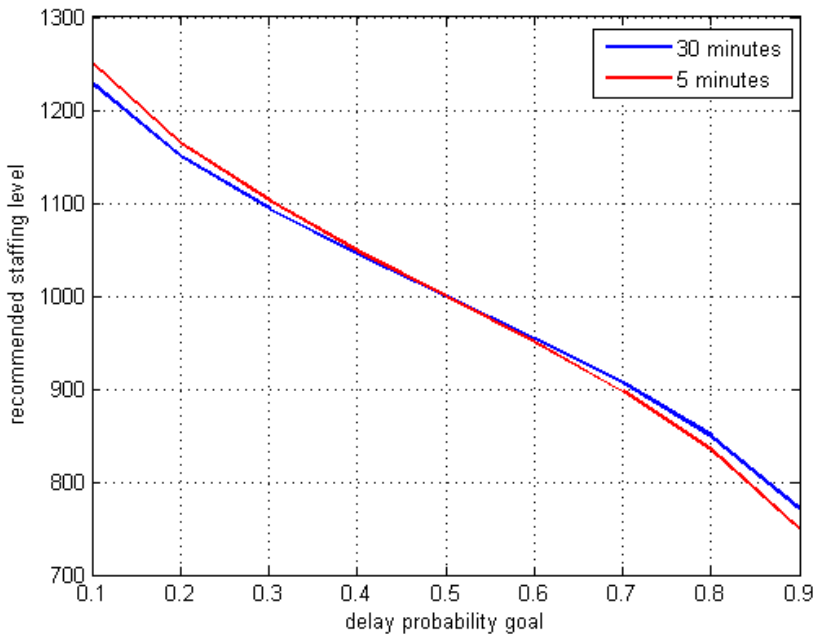
$$\ln(\sigma(Y)) = c \cdot \ln(\lambda) + \ln(\sigma(X)).$$

Staffing Recommendations: Comparison between Different Resolutions

Tue-Wed, midnight-10:30am



Tue-Wed, 10:30am-midnight



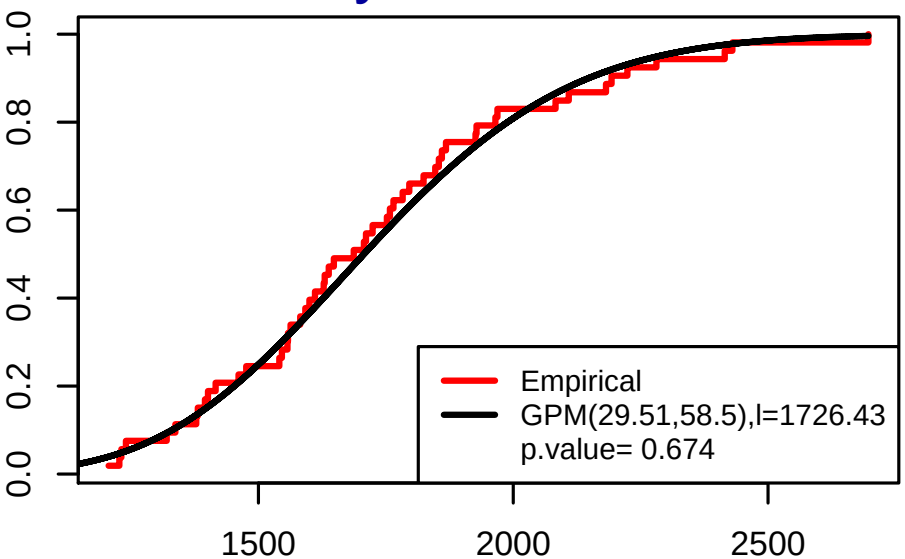
- Expected arrival rate 1000, service rate 1.
- X assumed normal.
- Very small differences for 0.3-0.7 delay probability goals.

Fitting Gamma Poisson Mixture Model to the Data

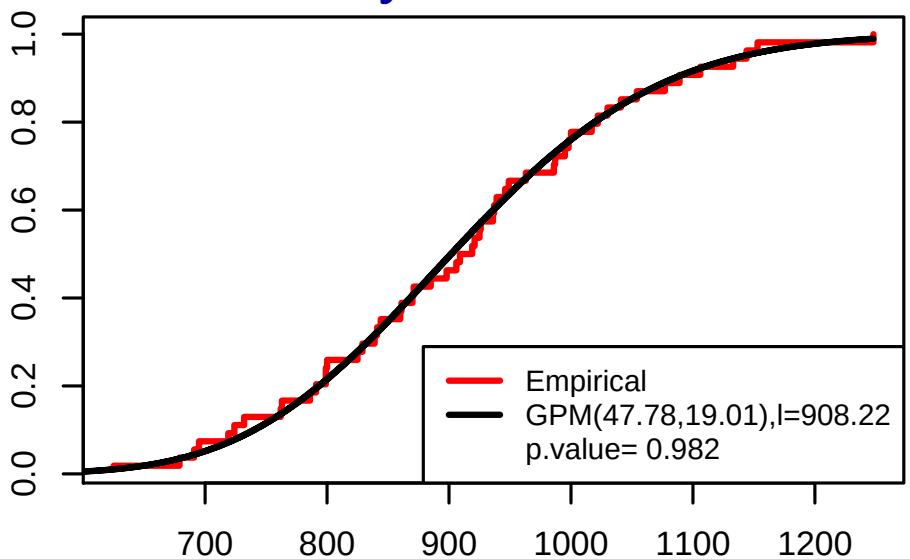
Assume Gamma distribution for the arrival rate $M = Gamma(a, b)$ (Koole and Jongbloed). Check this hypothesis for different intervals.

1. Maximum likelihood estimators of a and b.
2. Goodness of fit test including FDR (False Discovery Rate) control method to correct the multiple comparisons.

Sundays 10:00-10:30



Monday 14:30-15:00



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

Relation between Main Model and Gamma Poisson Mixture Model

Let $M = \lambda + \lambda^c X_\lambda \stackrel{d}{=} Gamma(a_\lambda, b_\lambda)$. Then,

$$E[M] = ab = \lambda; \quad Var(M) = \lambda b \quad and \quad Var(X) = \lambda^{1-2c} \cdot b.$$

If $\sigma(X_\lambda) \to \sigma(X), \quad \lambda \to \infty,$

$$\begin{aligned} b_\lambda &\sim \sigma^2(X) \cdot \lambda^{2c-1}, \quad \lambda \to \infty, \\ a_\lambda &\sim \sigma^{-2}(X) \cdot \lambda^{2-2c}, \quad \lambda \to \infty, \end{aligned}$$

and

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

Asymptotic Distribution of X

$$X_\lambda = \frac{M-\lambda}{\lambda^c} = \frac{M-ab}{(ab)^c} = \frac{M-E[M]}{\sigma(M)/\sigma(X_\lambda)}$$

Let $W_a \overset{d}{=} \dfrac{Gamma(a,b)-ab}{b\sqrt{a}} = \dfrac{Gamma(a,1)-a}{\sqrt{a}}$.

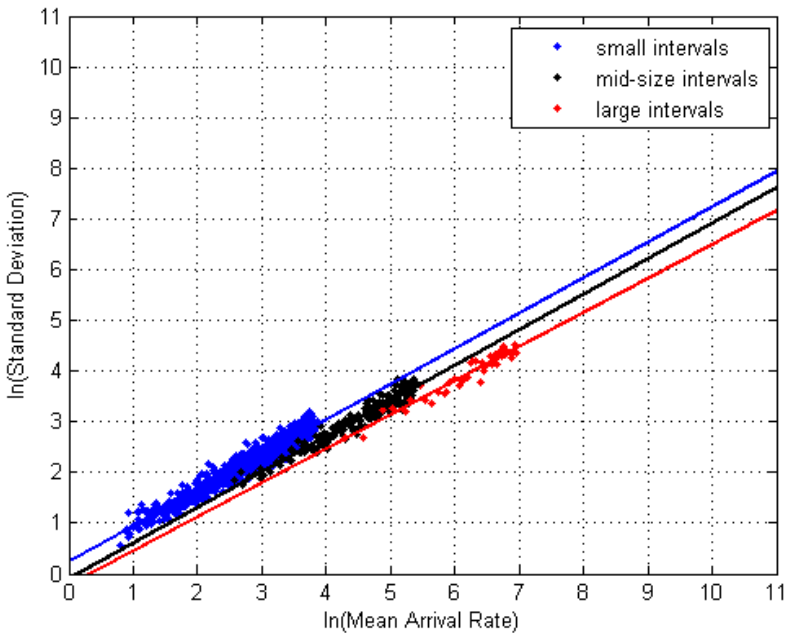
Then $\lim_{a\to\infty} \mathrm{MGF}_a(t) = e^{t^2/2}$, $t \in \mathbb{R}$, and this limit is the moment generating function of the standard normal distribution $Norm(0,1)$.

Conclusion: As $\lambda \to \infty$ (equivalent to $a \to \infty$), the random variable $X_\lambda/\sigma(X_\lambda)$ converges in distribution to a standard normal distributed variable.

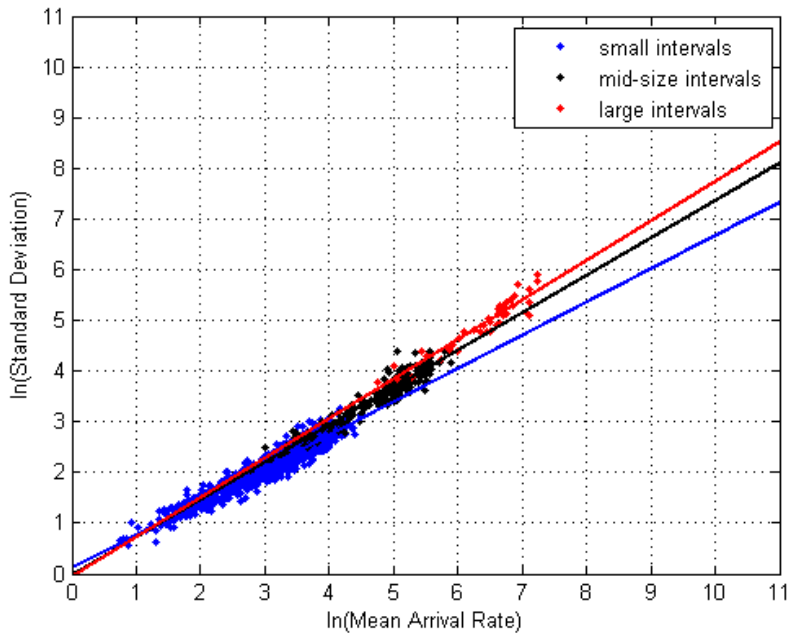
$$\frac{X_\lambda}{\sigma(X_\lambda)} \xrightarrow{\mathcal{D}} Norm(0,1).$$

Simulation Experiments with Underlying Fractional Gamma Motion

Hurst Parameter H=0.5



Hurst Parameter H=0.9



- $H = 0.5$ - classical Gamma process with iid increments.
- $H = 0.9$ - process with positively correlated increments.
- Three time resolutions considered.
- Phenomena that are similar to real-data ones were observed for large H and not too large arrival rates.

Outline of Additional Results

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

- **Numerical Experiments.** Very good fit between asymptotic results and the exact ones (simulation).

- **Time-varying Expected Arrival Rate.** Enhance ISA algorithm developed by Feldman with the features of random arrival rate in the $M_t/M/n + G$ queue.
  Goal: determine time-dependent staffing levels aiming to achieve a time-stable delay probability.

Future Research Challenges

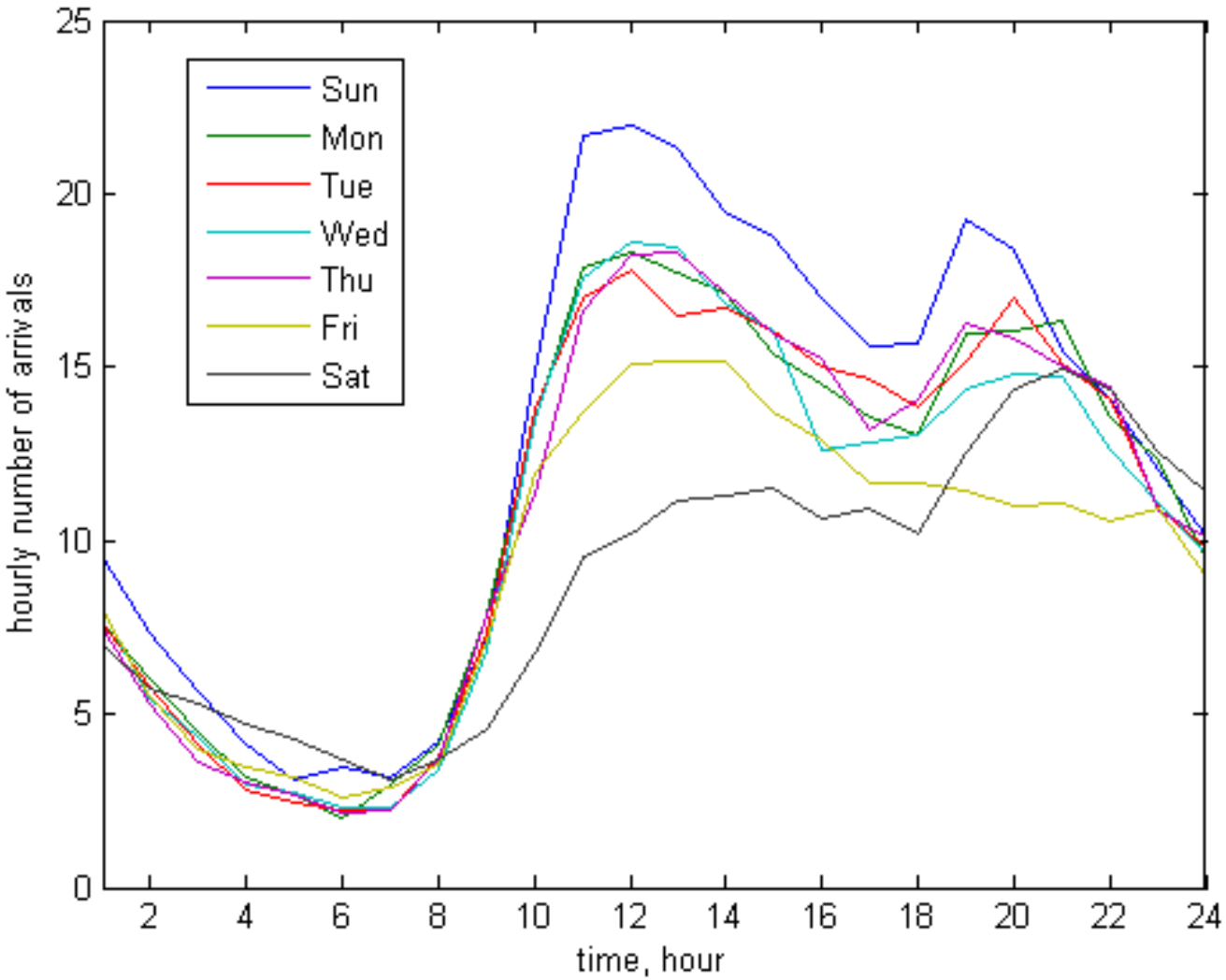
- Incorporating **forecasting errors** into our model (in the spirit of Steckley et al., 2007).
- **Scaling problem:** dependence of c on the basic interval duration, exploring underlying Gamma model.
- **Time-varying queueing models:** achieving time-stable performance measures (probability to abandon,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

Emergency Department: Data Description

- Rambam hospital 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).

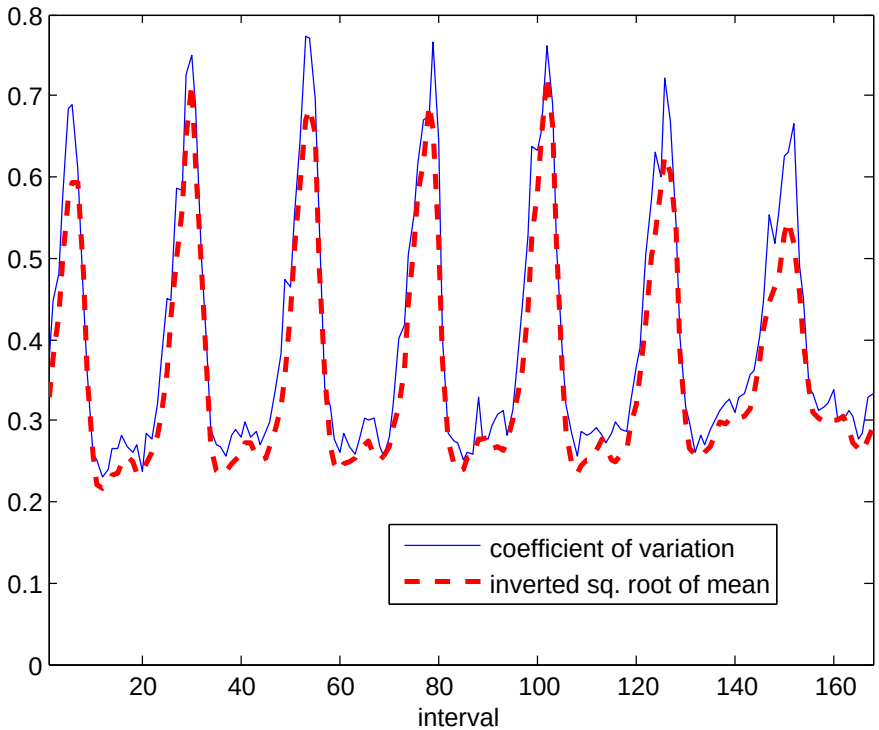
Emergency Department: Arrival Rates

Average Hourly Number of Arrivals

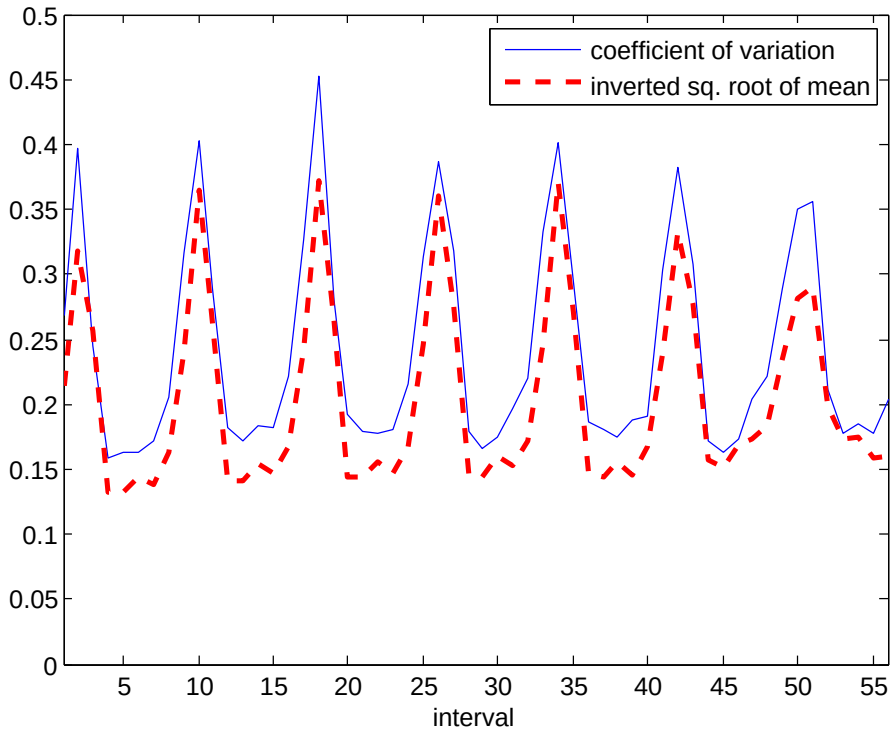


Emergency Department: Over-Dispersion Phenomenon

One-hour resolution



Three-hour resolution



- Moderate over-dispersion.
- Conventional variability ($c = 1/2$) seems to be a reasonable assumption for hourly resolution.

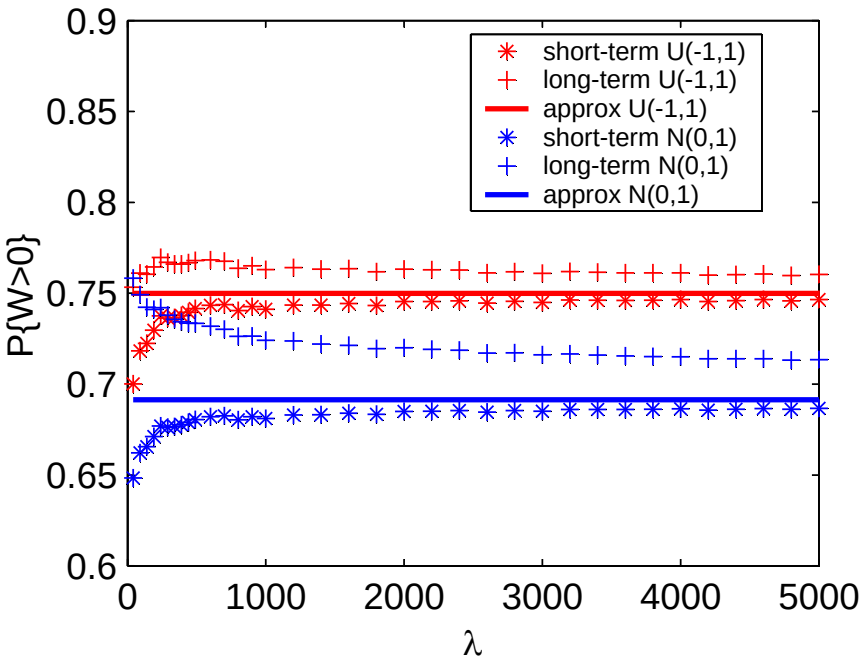
QED-c Regime: Numerical Experiments

Examples: Consider two distributions of X

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

$$(1)\ \beta = -0.5,\ c = 0.7$$

Delay Probability



Abandonment Probability