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

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

Tel-Aviv, 16th Israeli Engineering and Management Conference

March 24, 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.
- Validate arrival model via data analysis.
- Plug arrival model into relevant queueing models.
- Derive asymptotic results relevant for real-life staffing problems and provide practical guidelines.
- Validate queueing model via numerical experiments.

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., 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.

Whitt W.

Staffing a call center with uncertain arrival rate and absenteeism. 2006.

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:

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

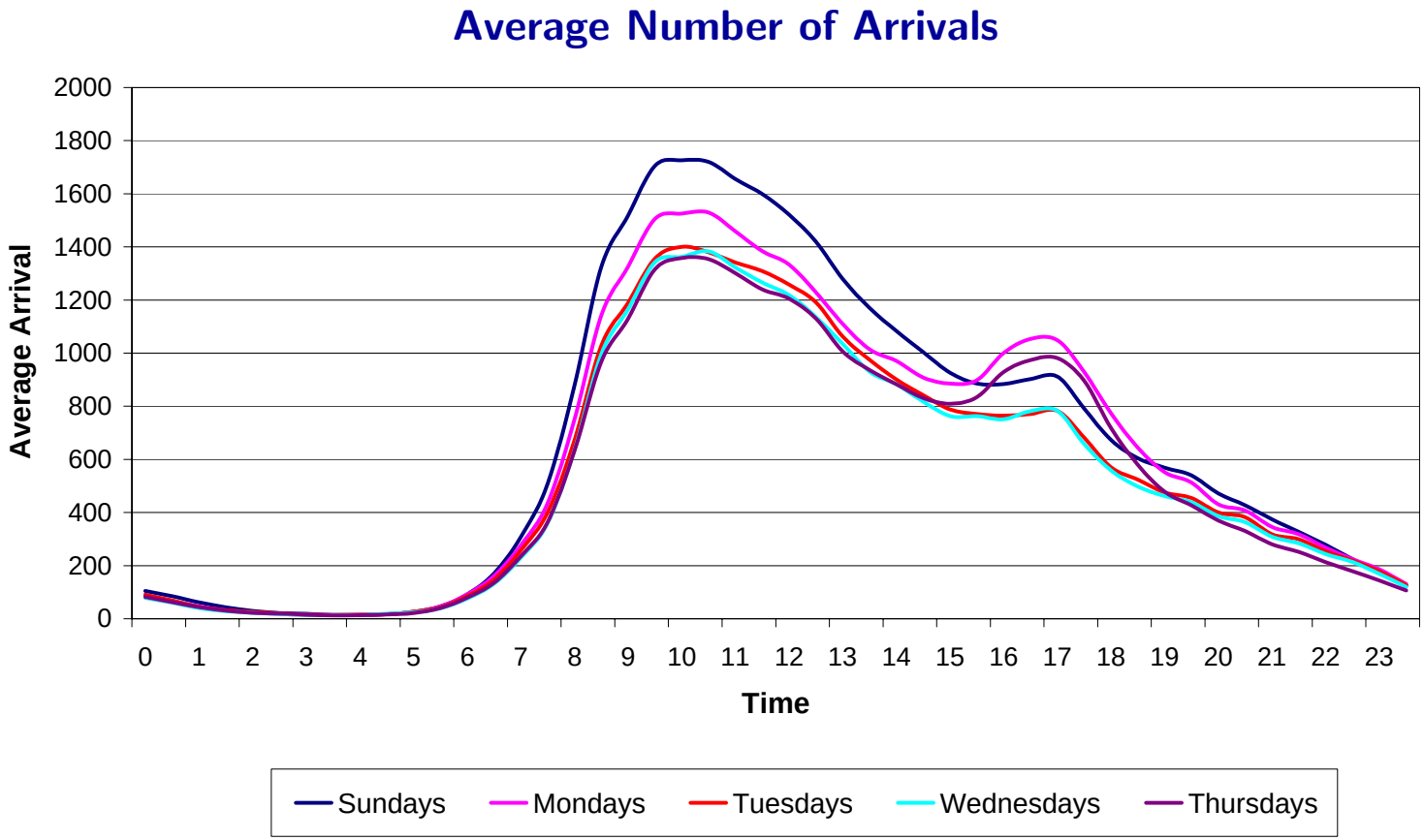
where X is a random variable with zero mean and finite variance.

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

Financial Call Center: Data Description

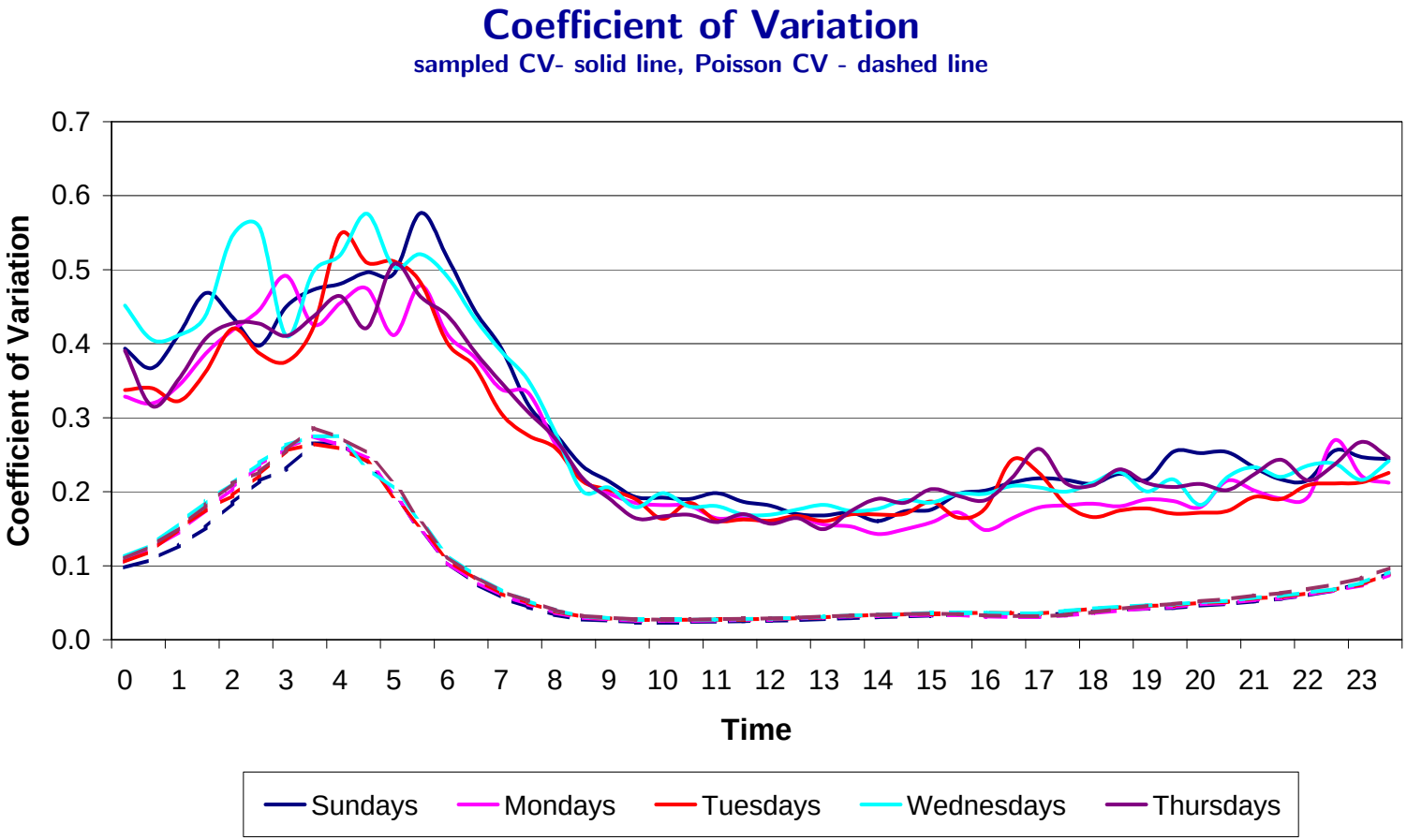
- Israeli bank.
- Arrival counts to the Retail queue are studied.
- 263 regular weekdays ranging between April 2007 and April 2008.
- Holidays with different daily patterns are excluded.
- Each day is divided into 48 half-hour intervals.

Financial Call Center: Arrival Rates



(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, 9-10am): 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,

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

Proof:

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

and

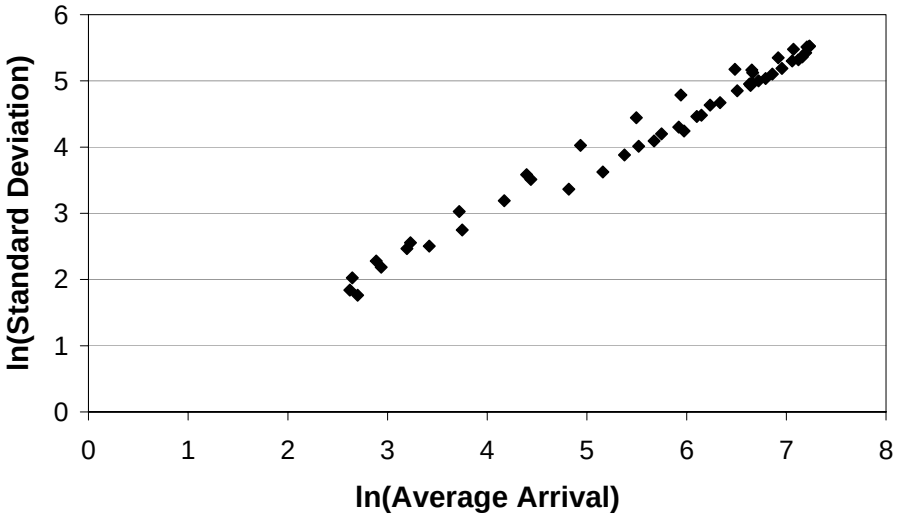
$$\lim_{\lambda \to \infty} (\ln(\sigma(Y)) - c \ln(\lambda)) = \ln(\sigma(X)).$$

Therefore, for large λ,

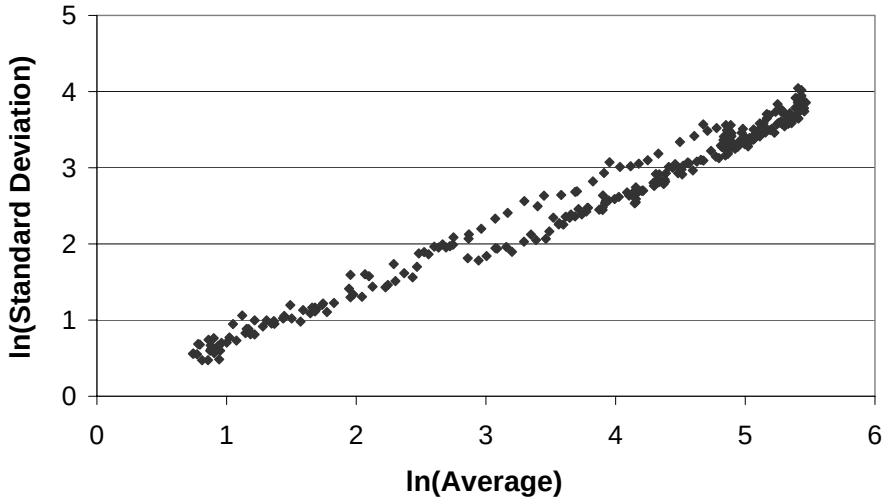
$$\ln(\sigma(Y)) \ \sim \ c \cdot \ln(\lambda) + \ln(\sigma(X)).$$

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):

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

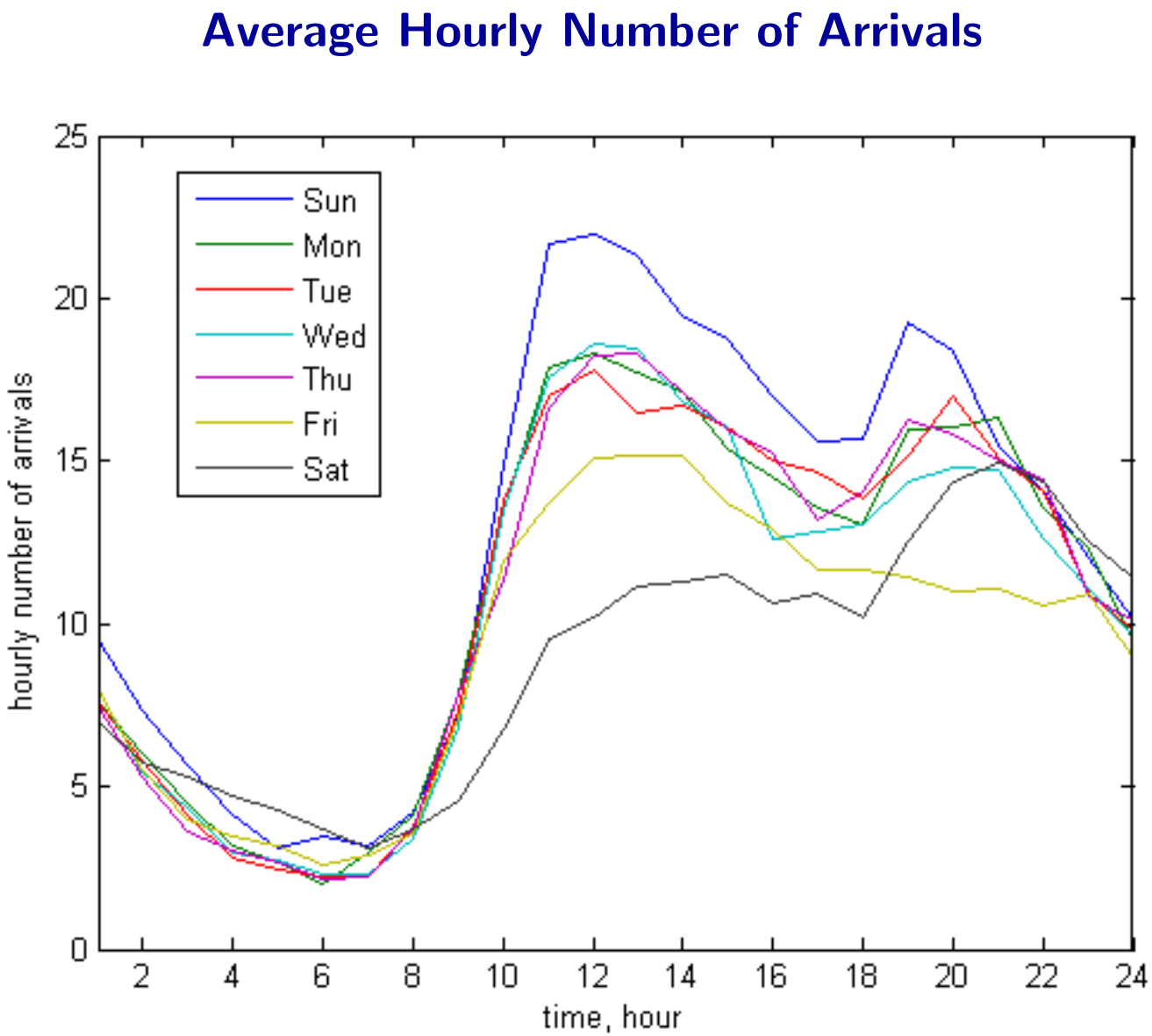
Financial Call Center: Outline of Additional Results

- Good fit of a well-known Gamma mixture model (Jongbloed and Koole ['01]) to data of Financial Call Center.
- Relation between our main model and Gamma Poisson mixture model is established.
- Distribution of X is derived under Gamma assumption: it is **asymptotically normal** given a large arrival rate.

Emergency Department: Data Description

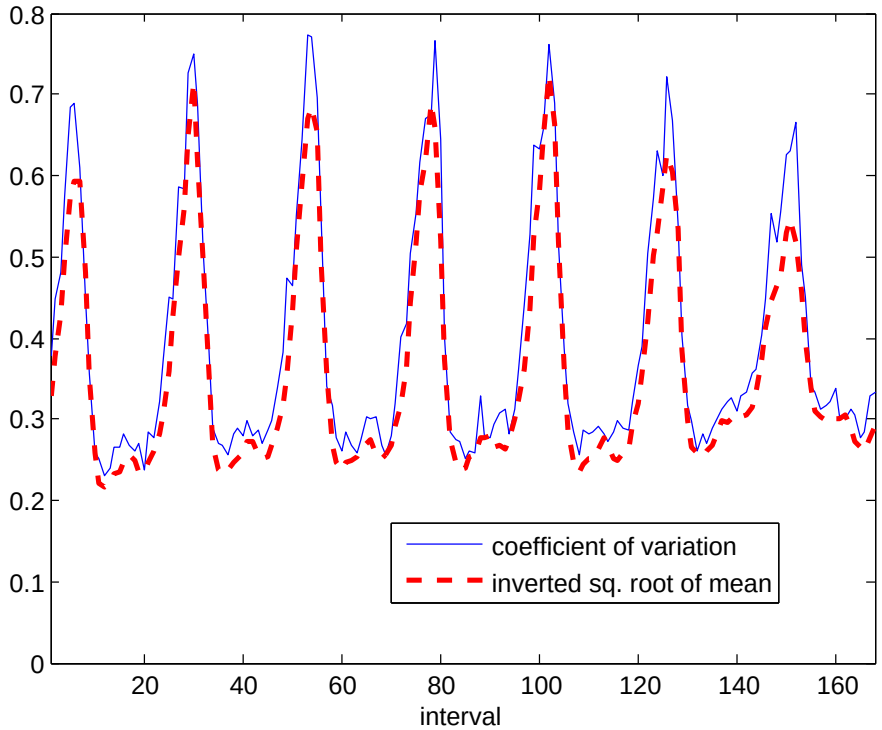
- Rambam hospital Emergency Department.
- 194 weeks between from January 2004 till October 2007 (five war weeks are excluded from data).
- The analysis is performed using two resolutions: hourly arrival rates (168 intervals in a week) and three-hour arrival rates (56 intervals in a week).

Emergency Department: Arrival Rates

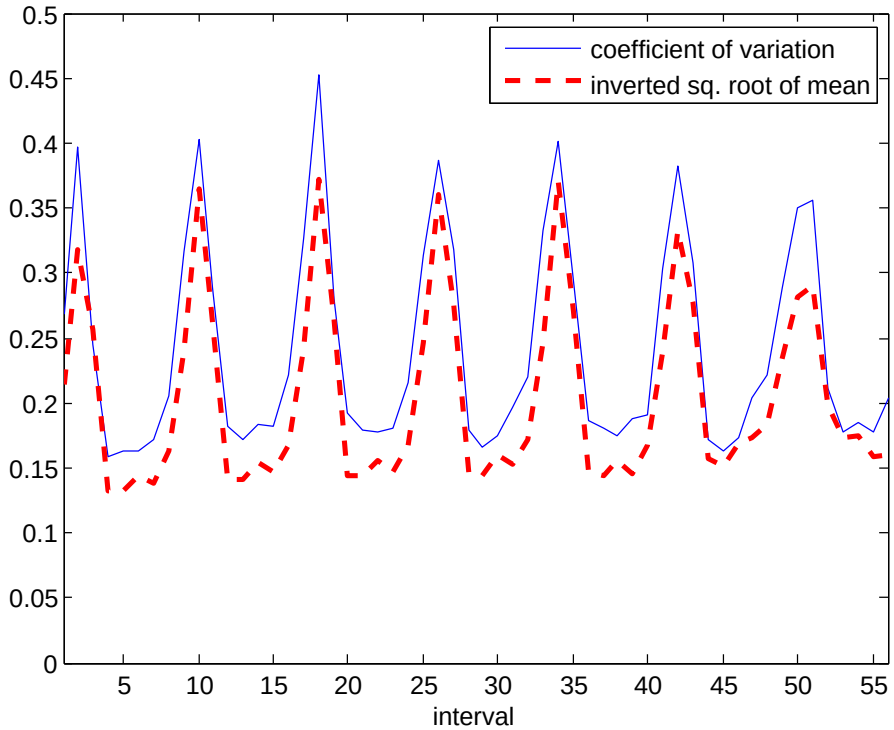


Emergency Department: Over-Dispersion Phenomenon

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

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 $M = \lambda + \lambda^c \mu^{1-c} X$, $c \in (1/2, 1)$, $E[X] = 0$, 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: $$E_{M,n}[W_q] \sim \frac{E[X-\beta]_+}{n^{1-c} \cdot g_0}.$$

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

QED-c Regime: Practical Guidelines

- Determine "uncertainty coefficient" c via regression analysis.
- If $1/2 < c < 1$, assume that X is asymptotically normal, calculate standard deviation from regression model.
- Apply our QED-c (or QED, Efficiency-Driven) 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 ($c = 1/2$).
  - Efficiency-Driven regime ($c = 1$), discrete and continuous distribution of X.

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

- **Iterative Staffing Algorithm (ISA)**, a simulation code developed by Feldman['04] 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., 2007).
- **Scaling problem:** dependence of c on the basic interval duration.
- **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).

Thank You