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

Shimrit Maman

Technion - Israel Institute of Technology

January 28, 2009

Advisors: Prof. Avishai Mandelbaum and Dr. Sergey Zeltyn

Outline

Motivation

A standard assumption in the modeling of a service system postulates that the arrival process is a Poisson process with known parameters. For example, the prevalent approach in call centers is to assume known arrival rates for each basic interval (say, half-hour).

However, as a rule, call centers data contradicts this assumption and shows a larger variability of the arrival process than the one expected from the Poisson hypothesis.

We explain this over-dispersion by the natural uncertainty of the arrival rate.

Literature Review

Henderson S.

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

Whitt W.

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

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.

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.

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 exist 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 exist 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 day to day in an i.i.d. fashion:

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

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 exist 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 day to day in an i.i.d. fashion:

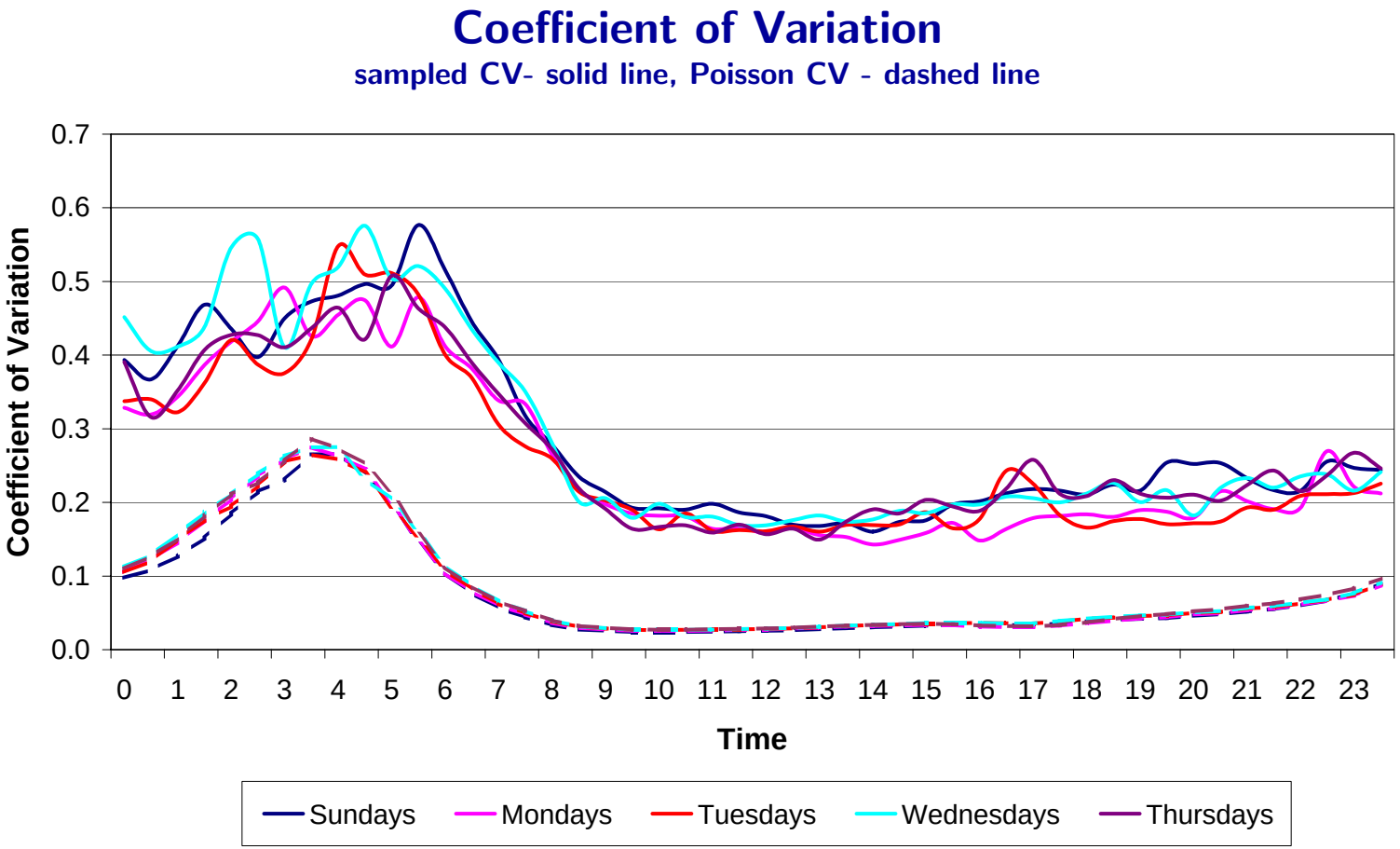
$$M = \lambda + \lambda^c X, \quad c \leq 1,$$

- $c \leq 1/2$: Conventional variability.
- $1/2 < c < 1$: Moderate variability.
- $c = 1$: Extreme variability.

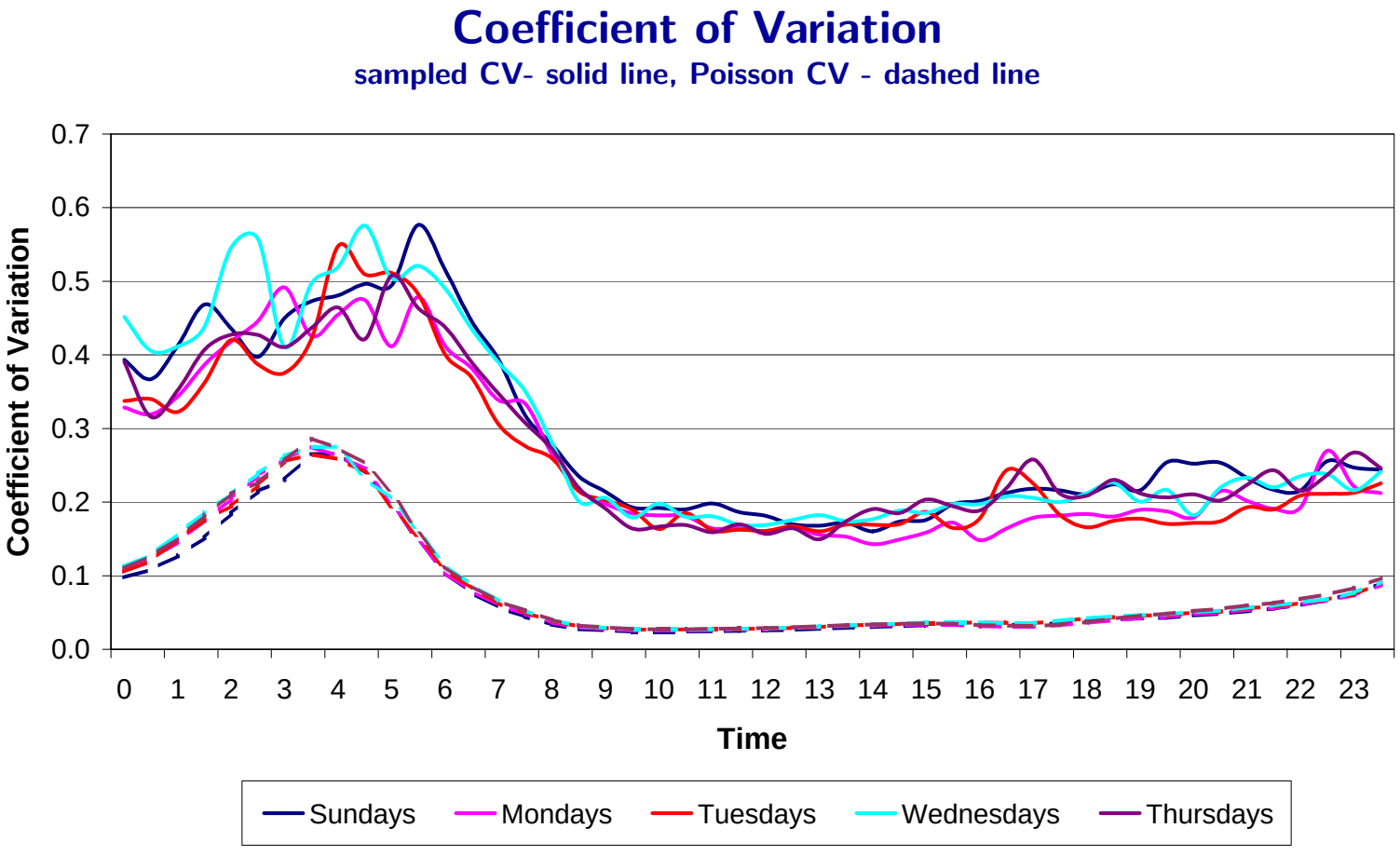
The Case of a Financial Call Center

- Our study focuses on the arrival counts to the Retail queue.
- We consider 263 regular weekdays ranging between April 2007 and April 2008.
- Holidays which exhibit different daily patterns are excluded from our analysis.
- Each day is divided into 48 half-hour intervals.

The Case of a Financial Call Center

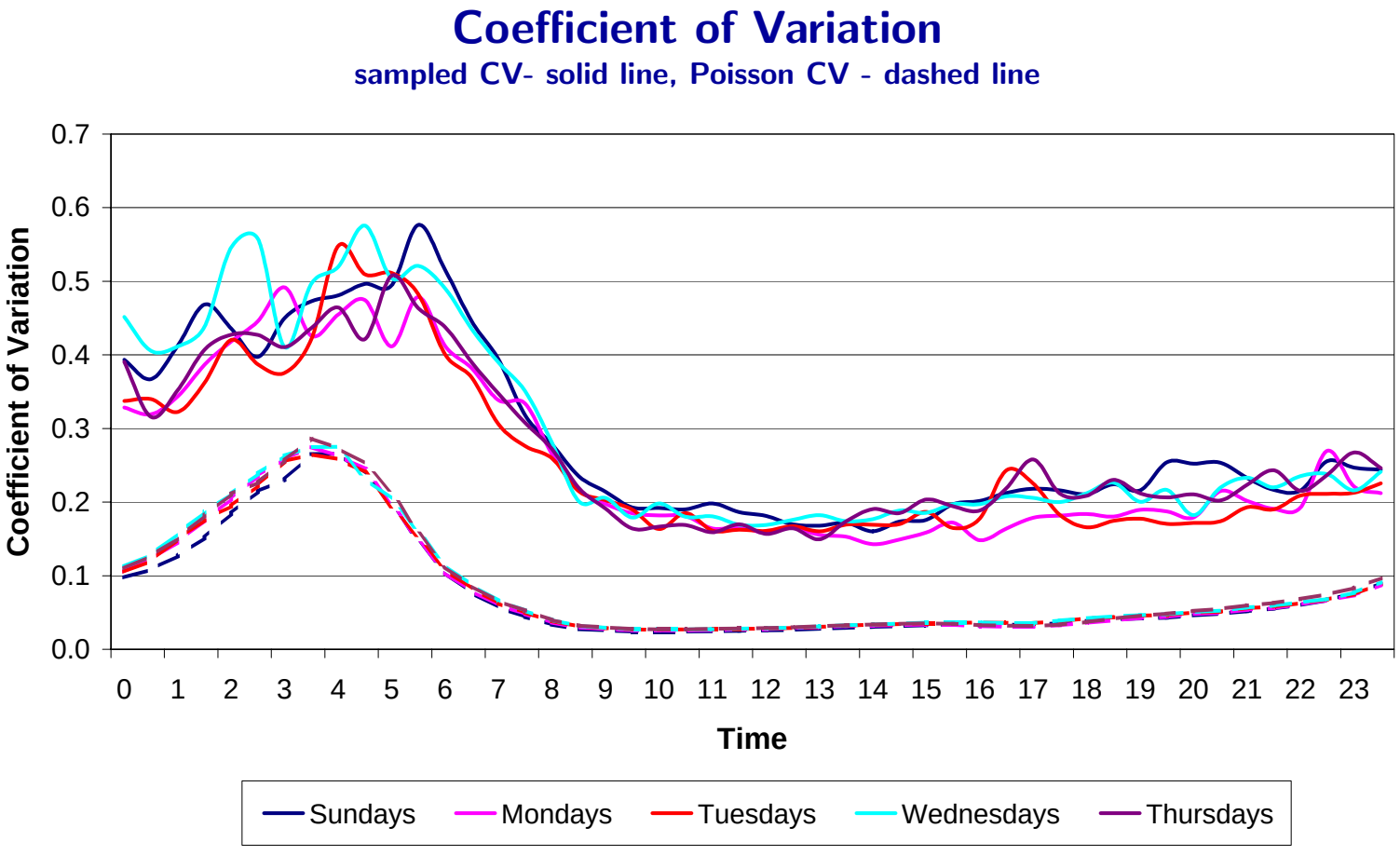


The Case of a Financial Call Center



Sampled CV's $\gg$ CV's of Poisson variables with fixed arrival rates

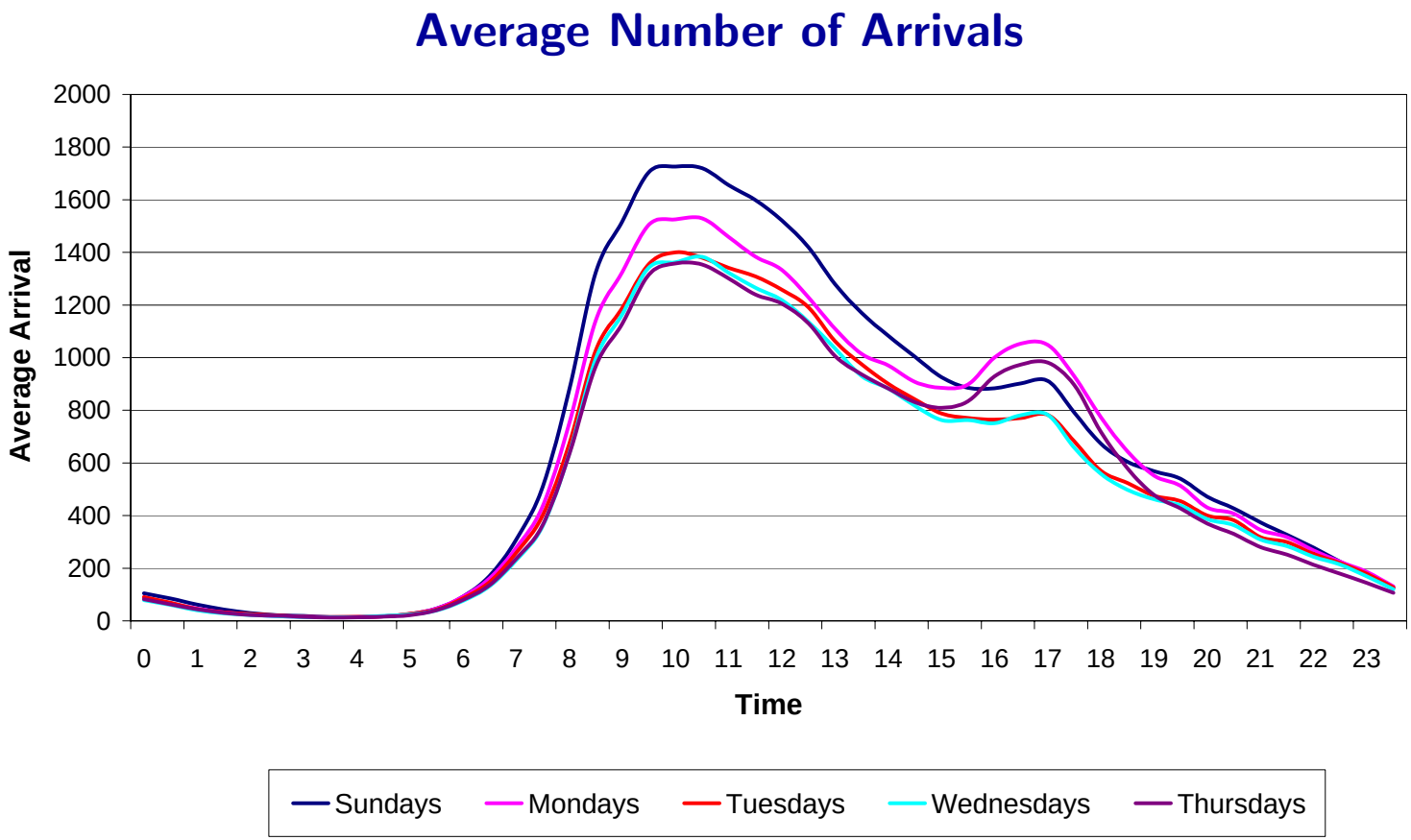
The Case of a Financial Call Center



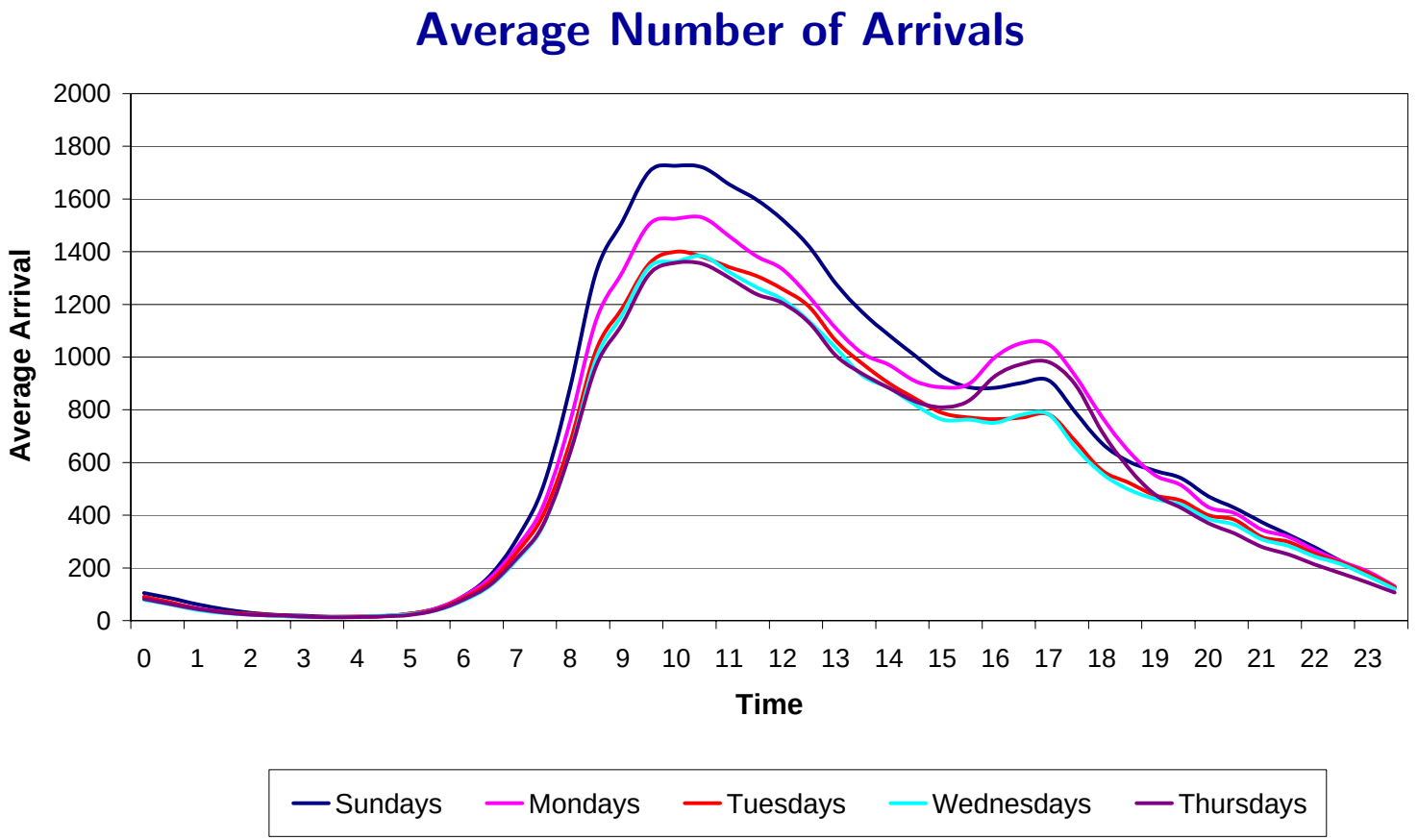
Sampled CV's $\gg$ CV's of Poisson variables with fixed arrival rates

$\Rightarrow$**Over-Dispersion**

The Case of a Financial Call Center



The Case of a Financial Call Center



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

The Case of a Financial Call Center

Basic definitions and notation:

- λ_{id} - The expectation of arrival volume in the i^{th} interval for day type d, $i = 1, 2, \ldots, 48$ and $d = 1, 2, 3, 4$.
- σ_{id} - The std of arrival volume for the i^{th} interval for day type d.
- c_d - The uncertainty coefficient for day type d.
- $\hat{\lambda}_{id}$ - The average call volume in the i^{th} interval over all days of type d.
- $\hat{\sigma}_{id}$ - The sampled standard deviation of call volumes in the i^{th} interval over all days of type d.

The Case of a Financial Call Center

1. Relation between λ_{id} and σ_{id}:

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

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

Given $\lambda \to \infty$

$$\sigma(Y) \;\sim\; \lambda^c \cdot \sigma(X),^1$$

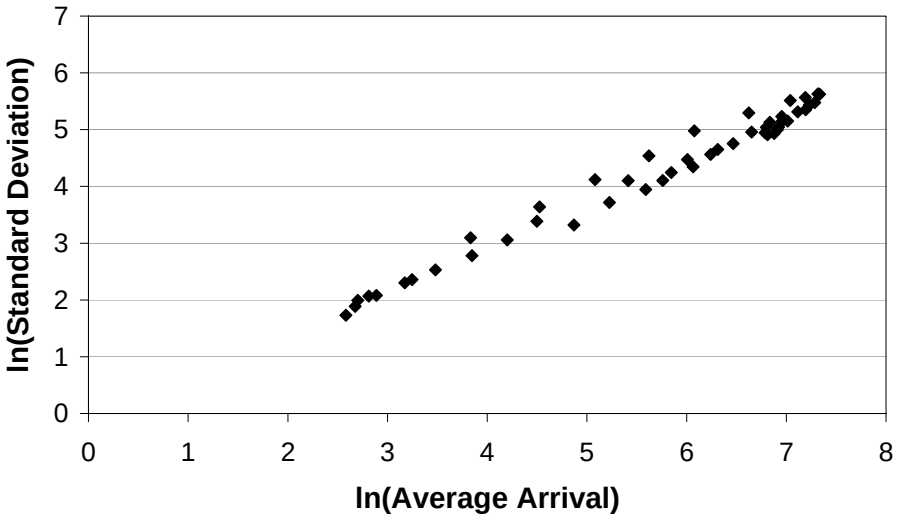
and

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

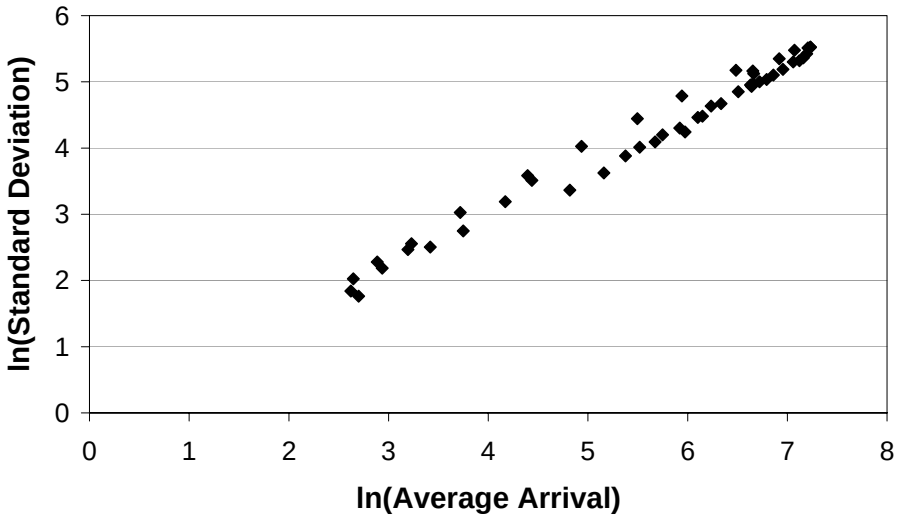
[1] $f(\lambda) \sim g(\lambda)$ denotes that $\lim_{\lambda \to \infty} f(\lambda)/g(\lambda) = 1$.

The Case of a Financial Call Center

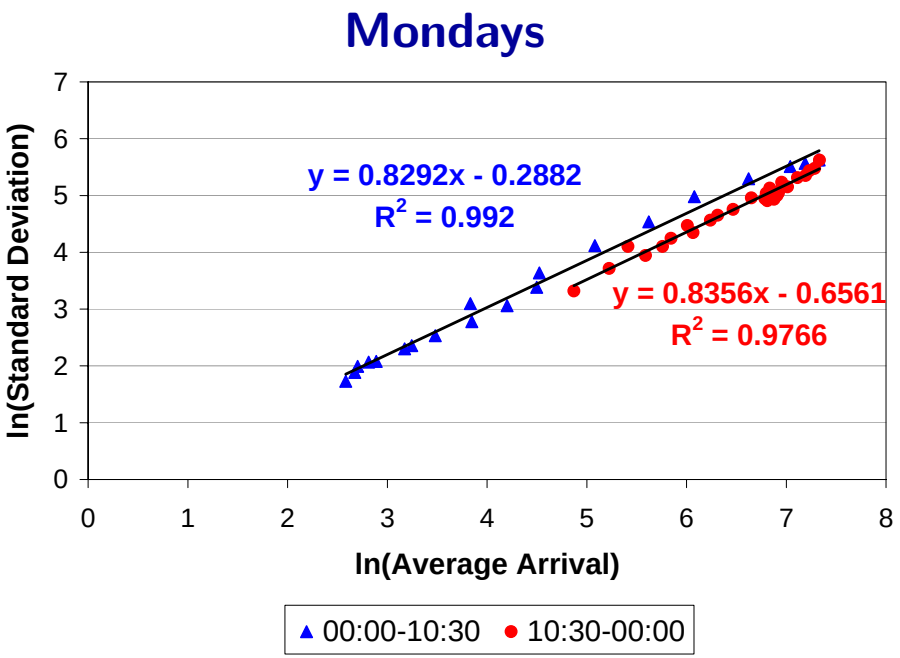
Mondays

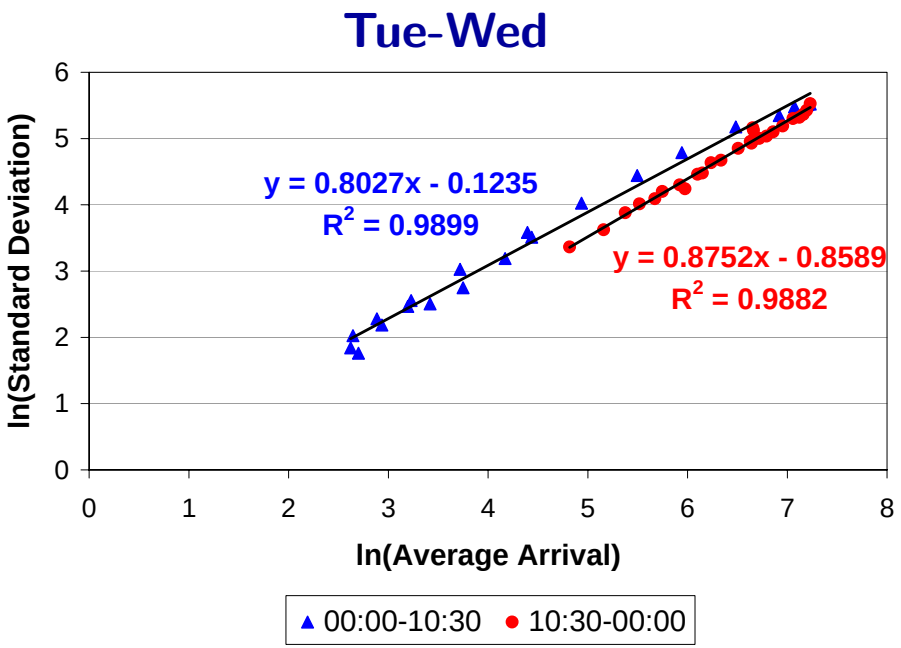


Tue-Wed



The Case of a Financial Call Center





Results:

- Two clusters exists. Denote $j(i) = 1$ for $i = 1, 2, \ldots, 21$, and $j(i) = 1$ for $i = 22, 23, \ldots, 48$.
- Very good fit ($R^2 > 0.97$).
- Significant linear relations:

$$\ln(\sigma_{id}) = c_{dj(i)} \cdot \ln(\lambda_{id}) + \ln(\sigma(X_{dj(i)})) \quad \forall\ d, i$$

The Case of a Financial Call Center

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

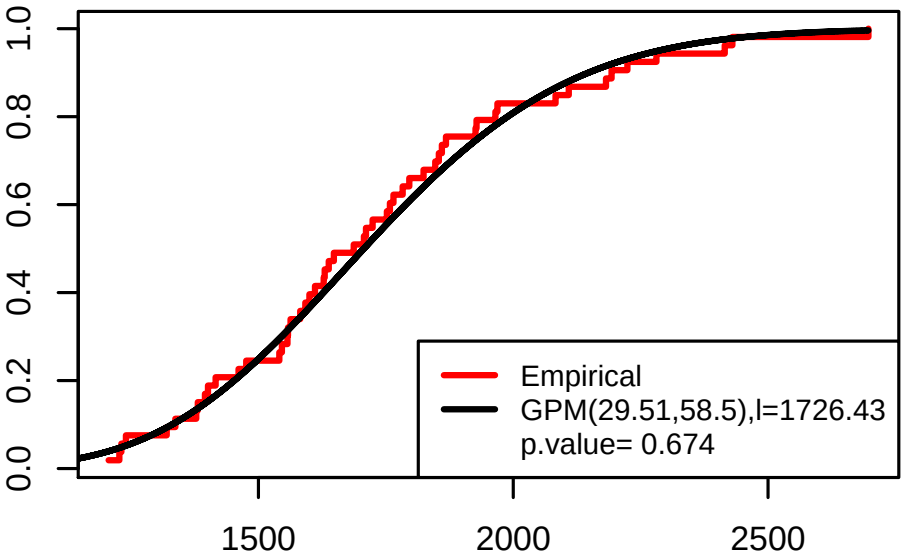
Assume a prior Gamma distribution for the arrival rate $\lambda + \lambda^c X \stackrel{d}{=} Gamma(a, b)$. Then, the distribution of Y is Negative Binomial.

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

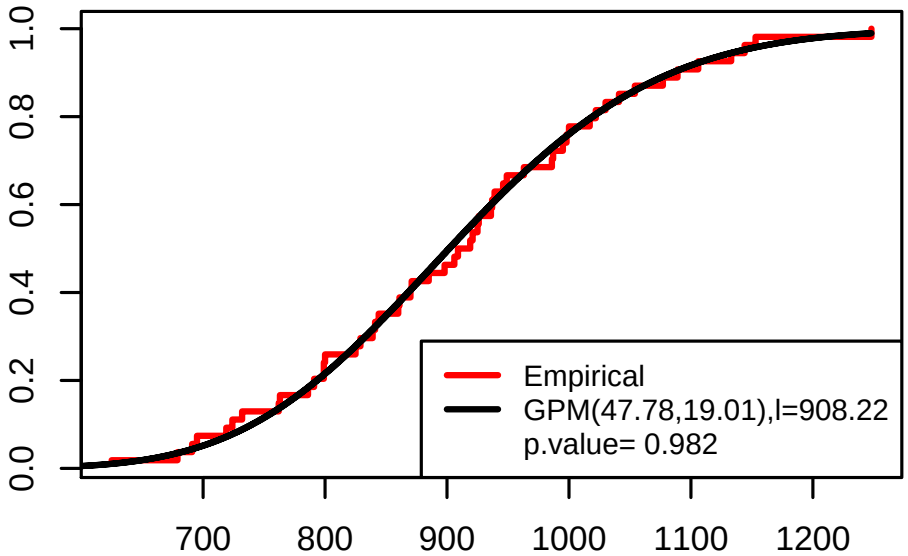
The Case of a Financial Call Center

$$\mathrm{H}_{0,id}: \; M_{id} \overset{d}{=} Gamma(\hat{a}_{id}, \hat{b}_{id})$$

Sundays 10:00-10:30



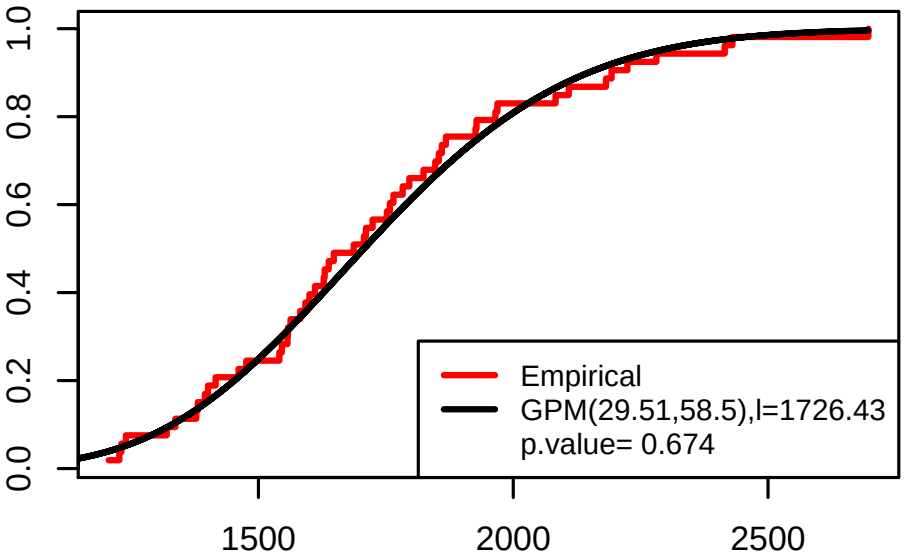
Monday 14:30-15:00



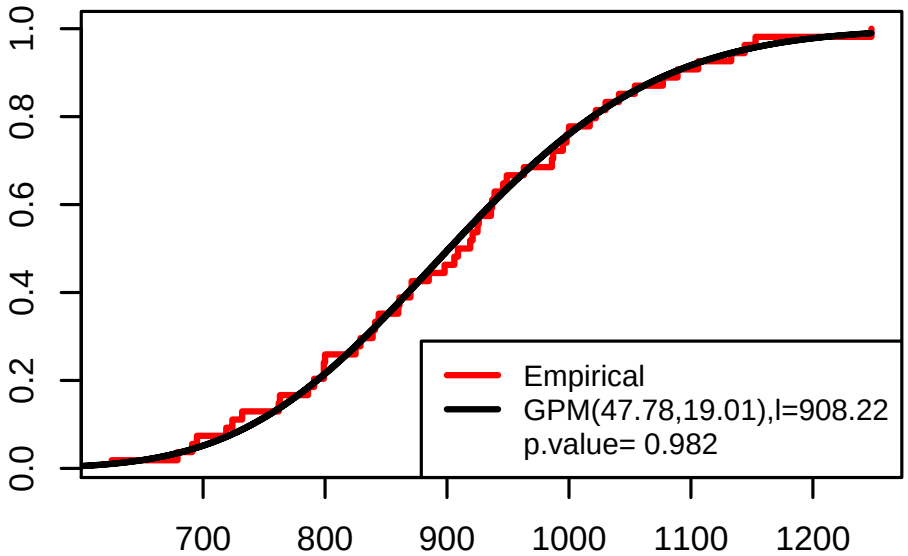
The Case of a Financial Call Center

$$\mathrm{H}_{0,id}: \; M_{id} \overset{d}{=} Gamma(\hat{a}_{id}, \hat{b}_{id})$$

Sundays 10:00-10:30



Monday 14:30-15:00



Results:

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

The Case of a Financial Call Center

3. Relation between our main model and Gamma Poisson mixture model:

Let $M = \lambda + \lambda^c X \stackrel{d}{=} Gamma(a, b)$. Then,

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

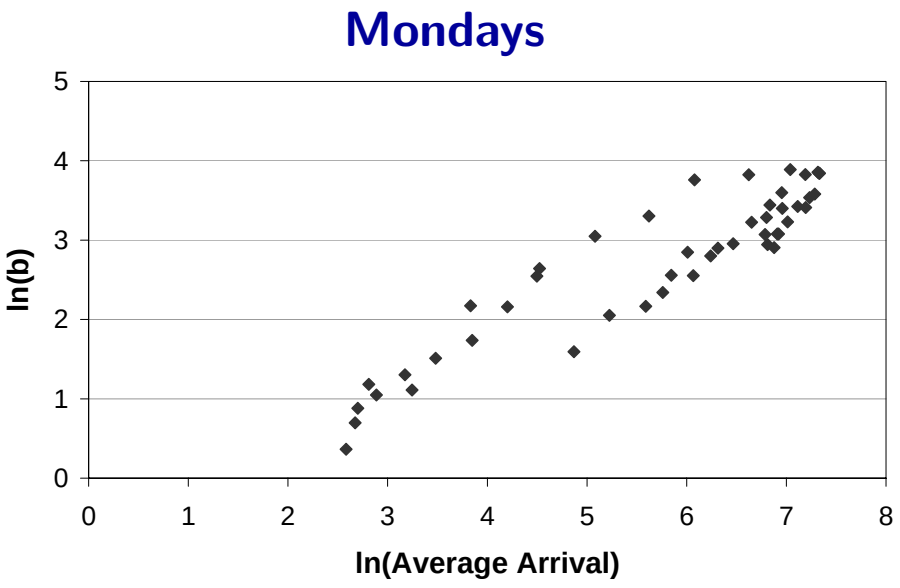
We derive the following relations

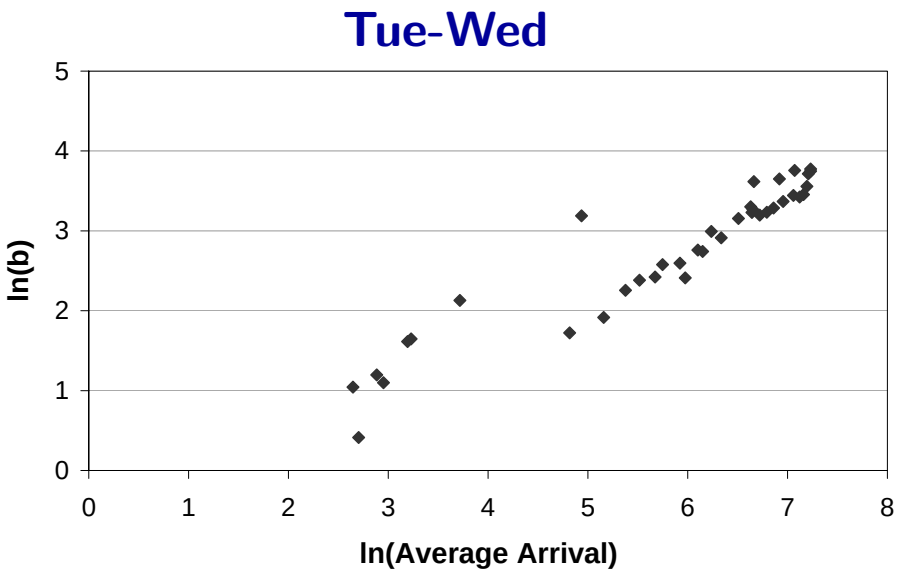
$$\begin{aligned} b &= \sigma^2(X) \cdot \lambda^{2c-1}, \\ a &= \sigma^{-2}(X) \cdot \lambda^{2-2c}. \end{aligned}$$

and conclude that

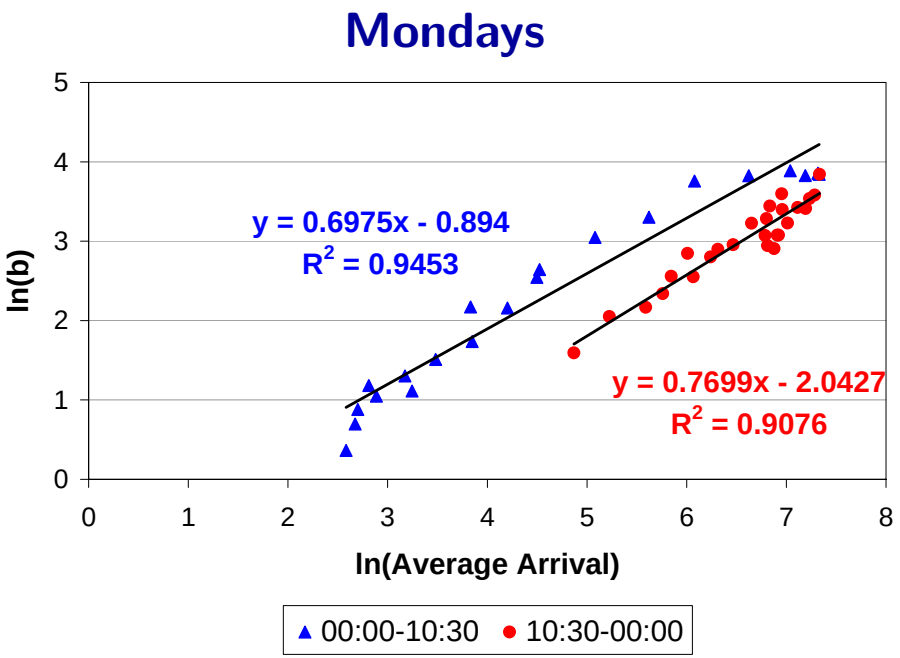
$$\ln(b) = (2c-1) \cdot \ln(\lambda) + \ln(\sigma^2(X)).$$

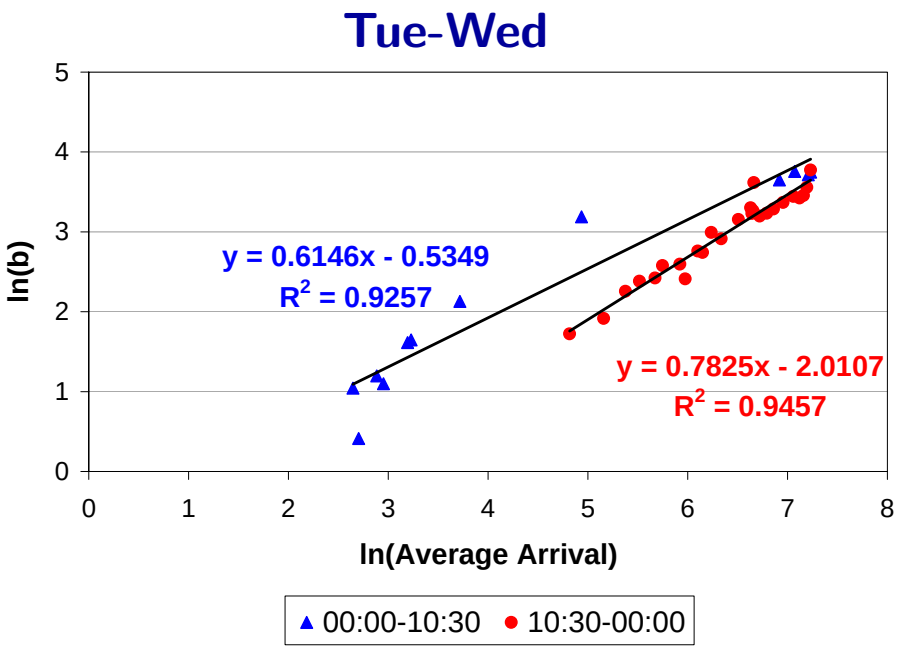
The Case of a Financial Call Center





The Case of a Financial Call Center





Results:

- Two clusters.
- Very good fit.
- Significant linear relations:

$$\ln(b_{id}) = c_{dj(i)} \cdot \ln(\lambda_{id}) + \ln(\sigma^2(X_{dj(i)})) \quad \forall\ d, i$$

The Case of a Financial Call Center

4. Asymptotic distribution of X:

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

The Case of a Financial Call Center

4. Asymptotic distribution of X:

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

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

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 $Norm(0,1)$.

The Case of a Financial Call Center

4. Asymptotic distribution of X:

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

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 $X/\sigma(X)$ converges in distribution to a standard normal distributed variable.

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

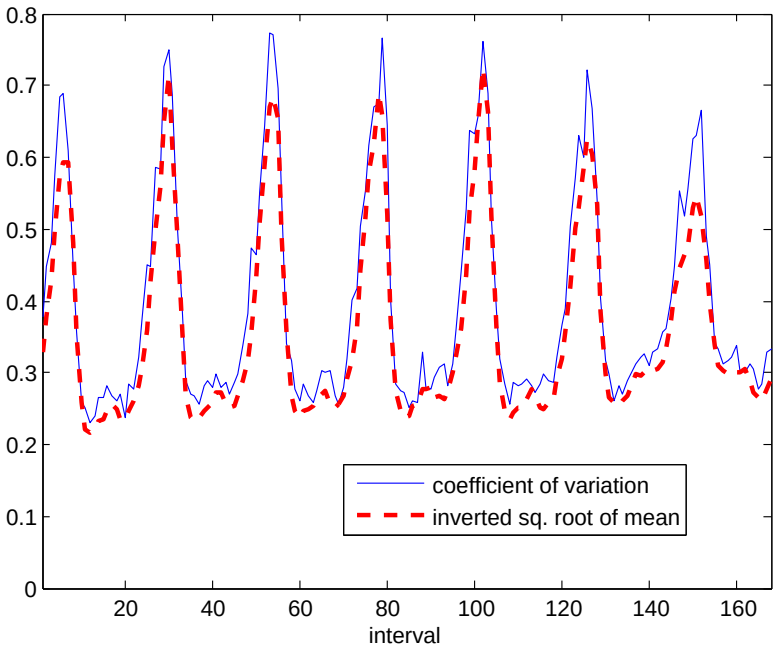
The Case of an Emergency Department

- Consider 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).
- Meanwhile, we do not clean our data (Jewish holidays are not excluded as they should be).

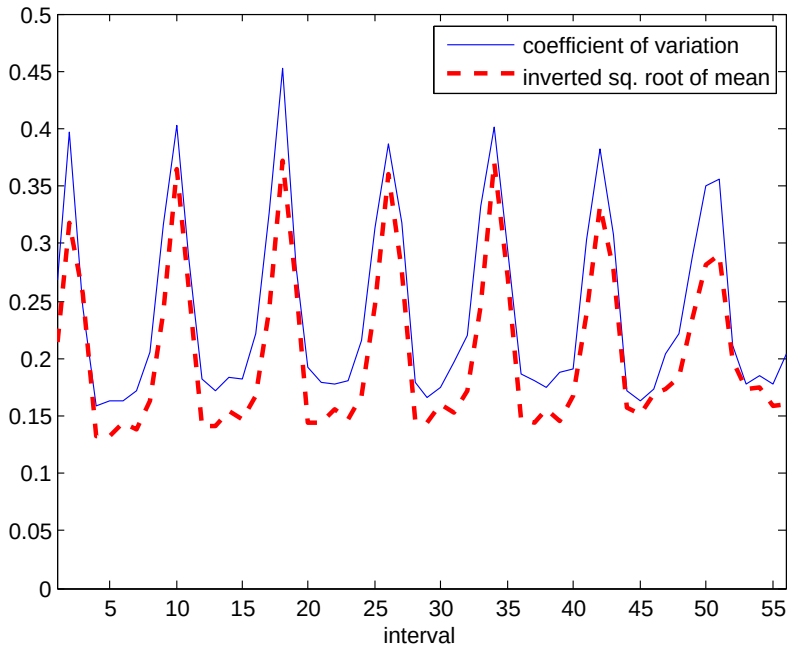
The Case of an Emergency Department

Coefficient of Variation

One-hour resolution



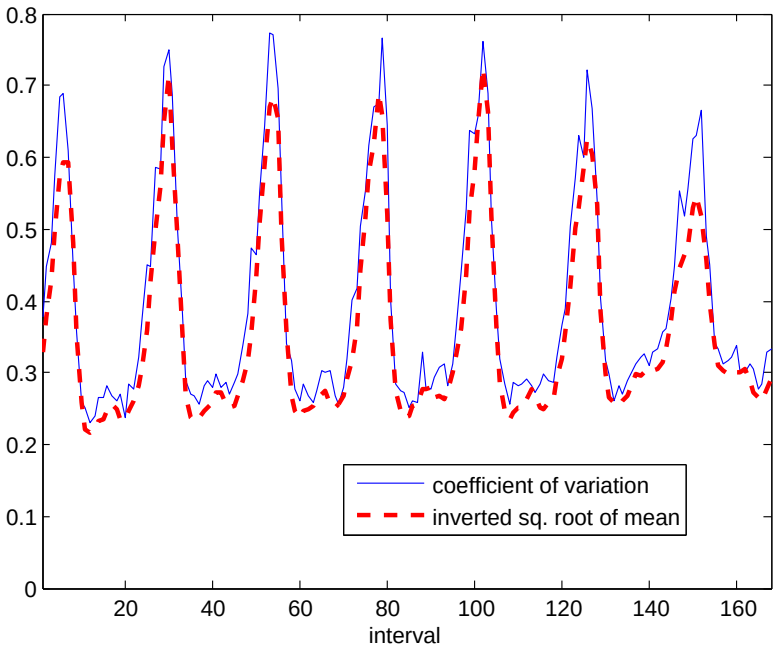
Three-hour resolution



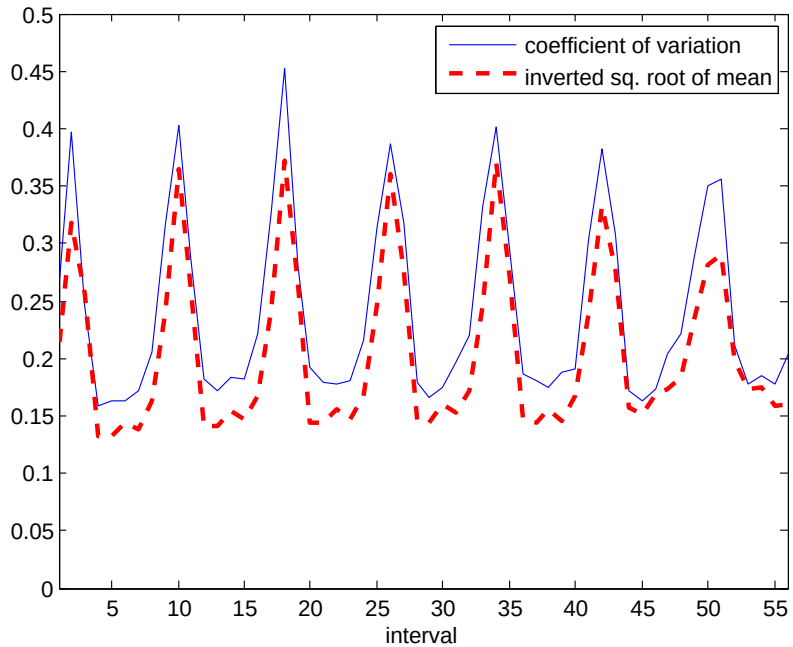
The Case of an Emergency Department

Coefficient of Variation

One-hour resolution



Three-hour resolution



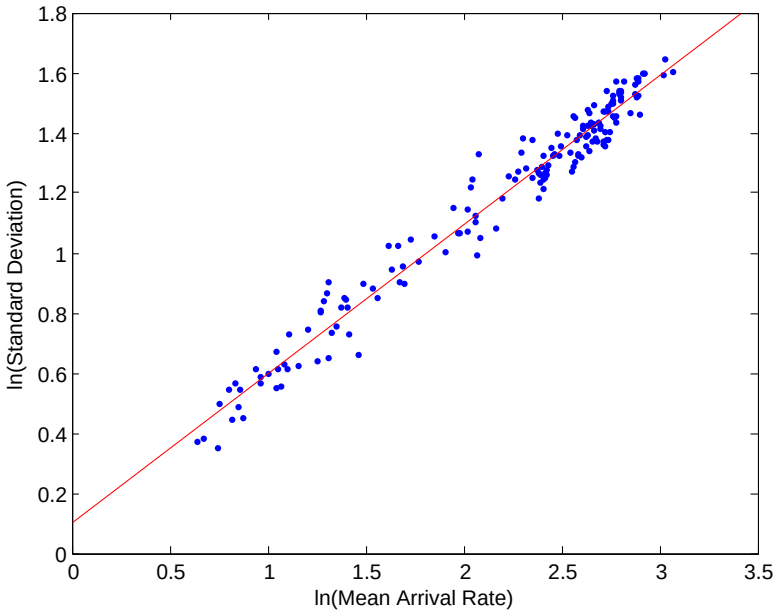
- In contrast to the call center study, inverted square root of mean is relatively close to CV's.
- Peaks at graphs correspond to night periods with small arrival rates.

The Case of an Emergency Department

ln(σ) versus ln(λ) plots

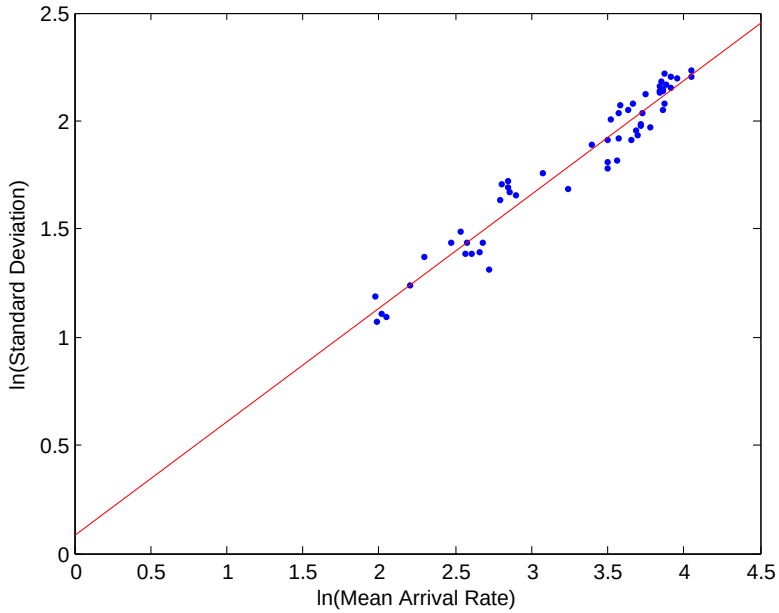
One-hour resolution

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



Three-hour resolution

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

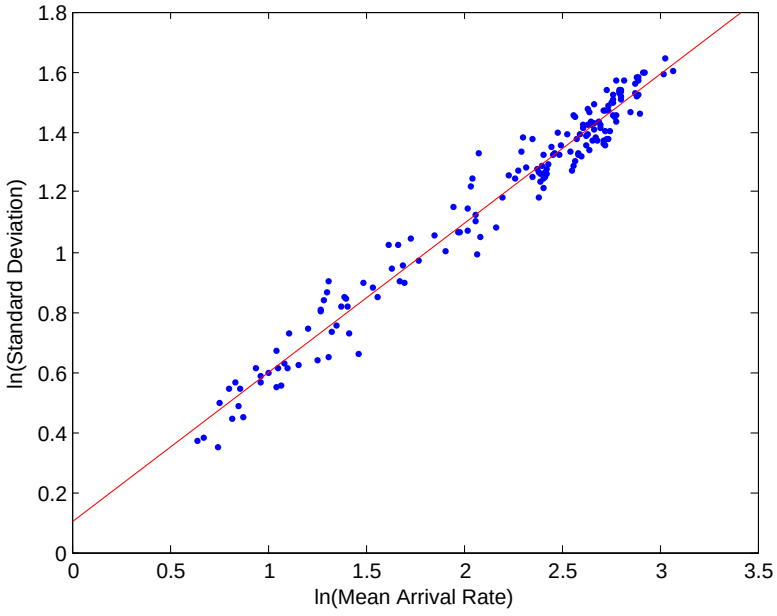


The Case of an Emergency Department

ln(σ) versus ln(λ) plots

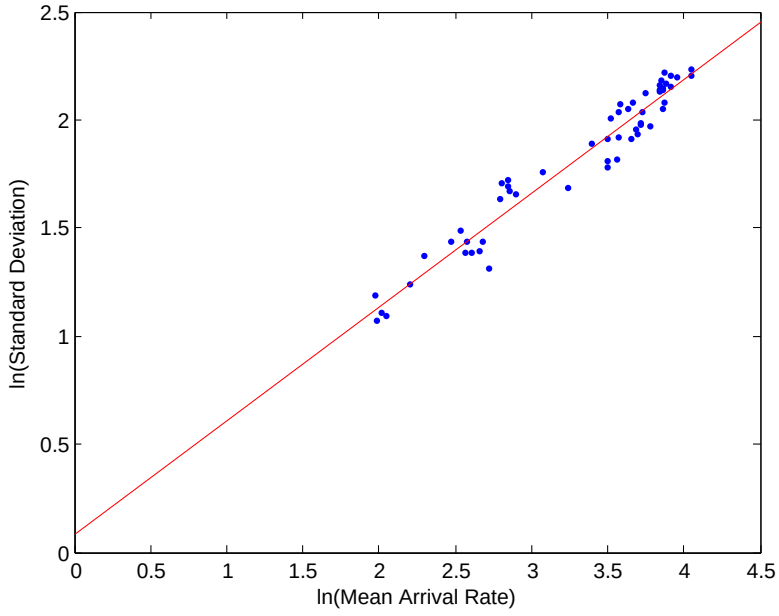
One-hour resolution

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



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 the asymptotic linear relation is based on $c > 1/2$. It is unclear how it works for c that is close to $1/2$ and relatively small λ.

Practical Guidelines

- Determine "uncertainty coefficient c" via regression analysis.
- Check if Gamma model is reasonable.
- Calculate X distribution (asymptotic analysis).
- Apply our QED-c results in order to determine appropriate staffing.

Queues Performances in a Random Environment

Assume a set of D days. At each day an arrival rate, m, is independently generated. Let Y be a system performance measure, and denote by Y_M the corresponding random performance measure in the random environment.

Denote:

- arr_j - Number of arrivals on day j
- y_j - The value of the performance measure Y_M, on day j.

Queues Performances in a Random Environment

Assume a set of D days. At each day an arrival rate, m, is independently generated. Let Y be a system performance measure, and denote by Y_M the corresponding random performance measure in the random environment.

Denote:

- arr_j - Number of arrivals on day j
- y_j - The value of the performance measure Y_M, on day j.

If the performance measure is related to **system** performance, e.g. offered load and queue length, by SLLN

$$\bar{Y} = \frac{1}{D}\sum_{j=1}^{D} y_j \xrightarrow{D\uparrow\infty} E[Y_M].$$

Queues Performances in a Random Environment

We classify performance measures that relate to the **customers**, e.g. delay probability and waiting time, into two classes:

Queues Performances in a Random Environment

We classify performance measures that relate to the **customers**, e.g. delay probability and waiting time, into two classes:

- **Short-Term Performance Measure:** What will happen tomorrow?

Queues Performances in a Random Environment

We classify performance measures that relate to the **customers**, e.g. delay probability and waiting time, into two classes:

- **Short-Term Performance Measure:** What will happen tomorrow?

$$\bar{Y} = \frac{1}{D}\sum_{j=1}^{D} y_j \overset{D\uparrow\infty}{\longrightarrow} E[Y_M].$$

Queues Performances in a Random Environment

We classify performance measures that relate to the **customers**, e.g. delay probability and waiting time, into two classes:

- **Short-Term Performance Measure:** What will happen tomorrow?

$$\bar{Y} = \frac{1}{D}\sum_{j=1}^{D} y_j \overset{D\uparrow\infty}{\longrightarrow} E[Y_M].$$

- **Long-Term Performance Measure:** What would be the performances in the long-run?

Queues Performances in a Random Environment

We classify performance measures that relate to the **customers**, e.g. delay probability and waiting time, into two classes:

- **Short-Term Performance Measure:** What will happen tomorrow?

$$\bar{Y} = \frac{1}{D}\sum_{j=1}^{D} y_j \overset{D\uparrow\infty}{\longrightarrow} E[Y_M].$$

- **Long-Term Performance Measure:** What would be the performances in the long-run?

$$\widetilde{Y} = \frac{\sum_{j=1}^{D} arr_j \cdot y_j}{\sum_{j=1}^{D} arr_j} \overset{D\uparrow\infty}{\longrightarrow} \frac{E[M \cdot Y_M]}{E[M]} > \bar{Y}.$$

Queues Performances in a Random Environment

Lemma

Assume that the system is in steady state. The long-term value and the short-term value of Y_M are asymptotically equivalent for all $c < 1$.

$$\lim_{\lambda \to \infty} \frac{E[M \cdot Y_M]}{E[M]} = \lim_{\lambda \to \infty} E[Y_M].$$

Queues Performances in a Random Environment

Lemma

Assume that the system is in steady state. The long-term value and the short-term value of Y_M are asymptotically equivalent for all $c < 1$.

$$\lim_{\lambda \to \infty} \frac{E[M \cdot Y_M]}{E[M]} = \lim_{\lambda \to \infty} E[Y_M].$$

Remark: Note that the above does not hold for $c = 1$. In this work we focus on long-term performance measures. The techniques of solutions, when considering the short-term performance measures, are similar.

QED-c Regime

QED-c staffing rule:

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

QED-c Regime

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.

QED-c Regime

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 provided asymptotically equivalent expressions (or bounds) for $P\{W_q > 0\}$, $P\{Ab|V > 0\}$ and $E[V|V > 0]$. Calculations, based on building blocks from Zeltyn['05], are carried out via the Laplace Method.

Square-Root Staffing versus QED-c staffing

$\mathbf{R} = \lambda/\mu$	β	SRS[2]	QED-c Staffing					
			$\mathbf{c = 0.6}$		$\mathbf{c = 0.75}$		$\mathbf{c = 0.9}$	
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%)
2000	0.5	2022	2048	(+1%)	2150	(+6%)	2468	(+22%)
	1	2045	2096	(+2%)	2300	(+12%)	2936	(+43%)
	1.5	2067	2143	(+4%)	2449	(+18%)	3403	(+65%)

[2] Square-Root-Staffing

QED-c Regime

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: $$E_{M,n}[W_q] \sim \frac{E[X-\beta]_+}{n^{1-c} \cdot g_0}$$

Remark: For simplifying the formulae we take $M = \lambda + \mu^{1-c} \cdot \lambda^c X$ instead of $M = \lambda + \lambda^c X$.

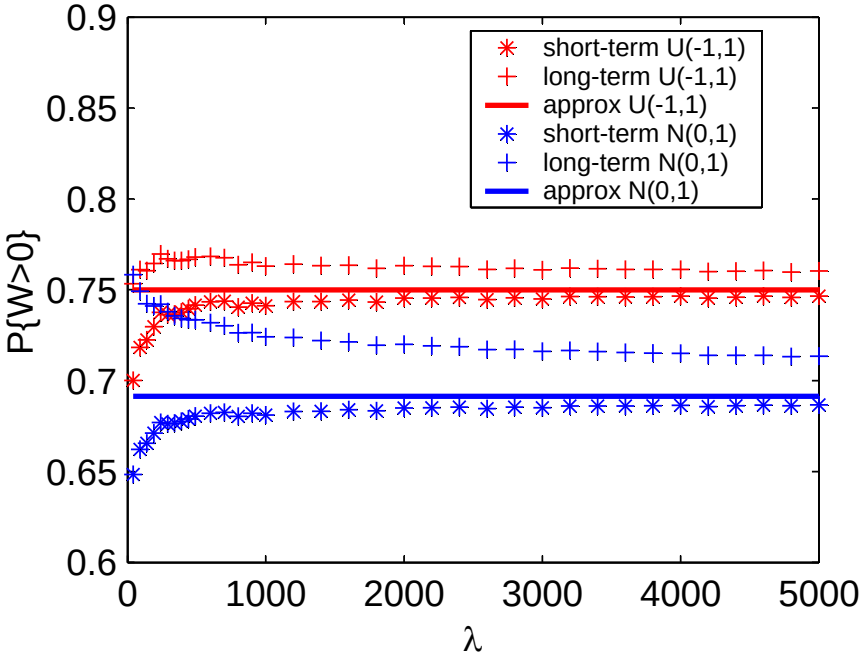
QED-c Regime

Examples: Consider two distributions of X

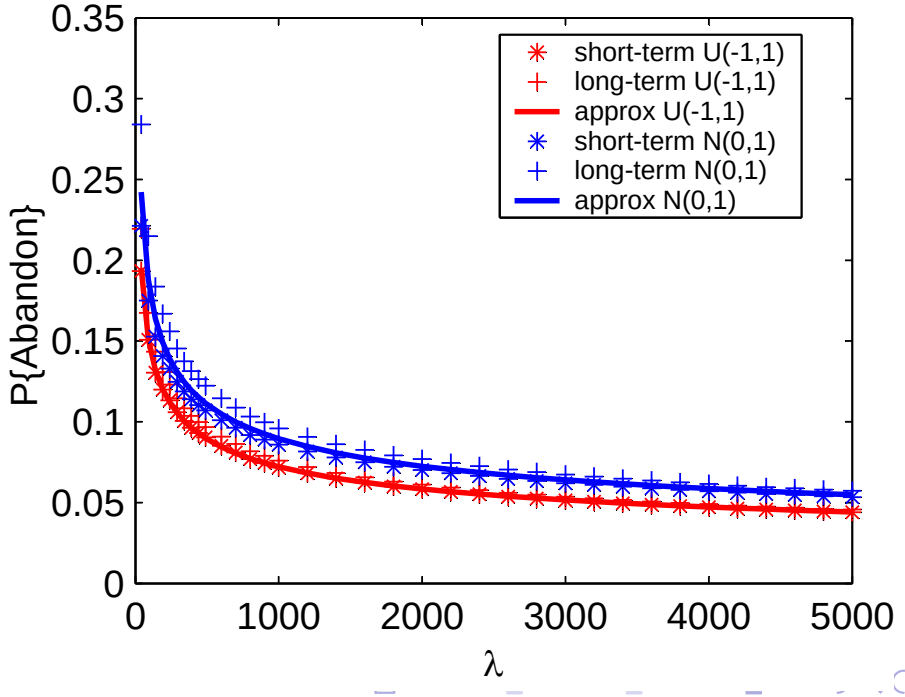
- Uniform distribution on [-1,1],
- Standard Normal distribution.

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

Delay Probability



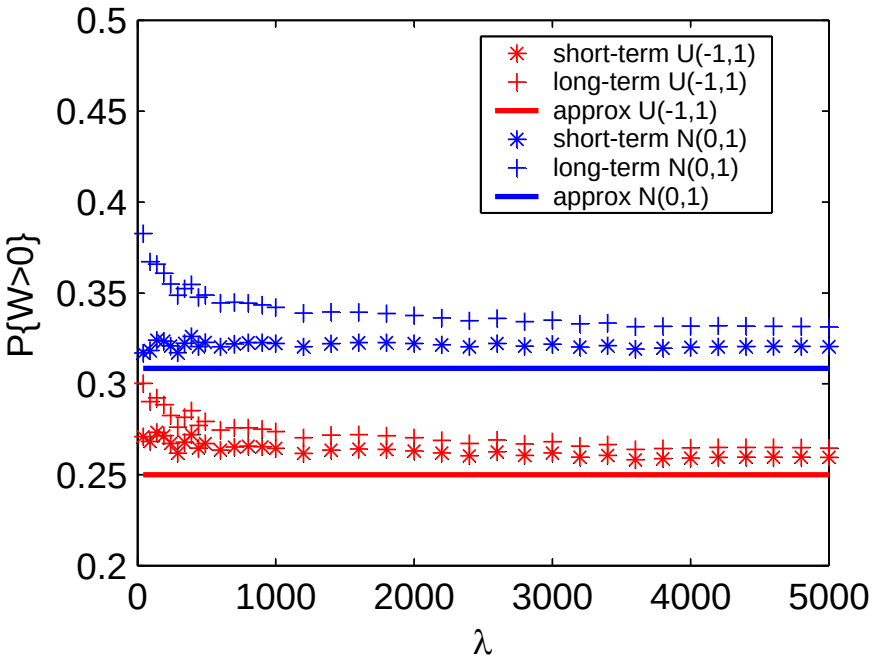
Abandonment Probability



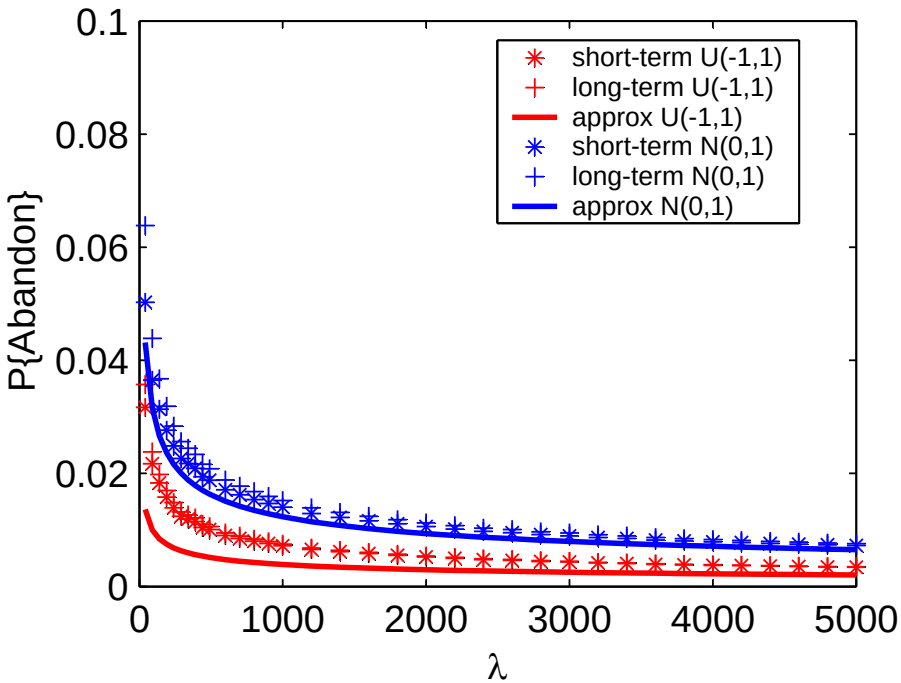
QED-c Regime

$$(2)\ \beta = 0.5,\ c = 0.6$$

Delay Probability



Abandonment Probability



Constraint Satisfaction

Formulation of the Problem:

Define the **optimal staffing level** by

$$n_\lambda^* = \operatorname{argmin}_n \left\{ \widetilde{P}_{M,n}\{W_q > 0\} \leq \alpha \right\}.$$

Constraint Satisfaction

Formulation of the Problem:

Define the **optimal staffing level** by

$$n_\lambda^* = \operatorname{argmin}_n \left\{ \widetilde{P}_{M,n}\{W_q > 0\} \leq \alpha \right\}.$$

The staffing level n_λ is called **asymptotically feasible** if

$$\limsup_{\lambda \to \infty} \widetilde{P}_{M,n_\lambda}\{W_q > 0\} \leq \alpha.$$

Constraint Satisfaction

Formulation of the Problem:

Define the **optimal staffing level** by

$$n_\lambda^* = \operatorname{argmin}_n \left\{ \widetilde{P}_{M,n}\{W_q > 0\} \leq \alpha \right\}.$$

The staffing level n_λ is called **asymptotically feasible** if

$$\limsup_{\lambda \to \infty} \widetilde{P}_{M,n_\lambda}\{W_q > 0\} \leq \alpha.$$

In addition, n_λ is **asymptotically optimal** if

$$|n_\lambda^* - n_\lambda| = o(f(\lambda)).$$

QED Regime

$c = 1/2$

Assume $M = \lambda + \sqrt{\mu\lambda} \cdot X$, $E[X] = 0$ and finite $\sigma(X) > 0$. Let $\lambda \to \infty$.

Theorem (Staffing)

a. The optimal staffing level satisfies

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

where β^* is the unique solution of the equation

$$\alpha = E[\alpha(\beta - X)]$$

with respect to the unknown β.
$\alpha(\cdot)$ is the Garnett function

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

QED Regime

Theorem (Staffing). Continued.

b. Introduce the staffing level

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

Then the staffing level n_β^* is both *asymptotically feasible* and *asymptotically optimal* $(f(\lambda) = \sqrt{\lambda})$

Proof Outline

Given X=x,

$$n_\beta = \frac{\lambda}{\mu} + \beta \cdot \sqrt{\frac{\lambda}{\mu}} = \frac{\lambda + \sqrt{\mu\lambda} \cdot x}{\mu} + b(\beta, x) \cdot \sqrt{\frac{\lambda + \sqrt{\mu\lambda} \cdot x}{\mu}}$$

$$\Rightarrow \quad \text{as } \lambda \to \infty, \qquad b(\beta, x) \sim \beta - x.$$

QED Regime

Theorem (Performance Measures)

Under the square-root staffing level, as $\lambda \to \infty$,

a. $$P_{M,n^*_\beta}\{Ab\} \sim \frac{1}{\sqrt{n}} \cdot E(\gamma^*_X).$$ ▸ γ^*_X

b. $$\rho_{n^*_\beta} = 1 - \frac{\beta^*}{\sqrt{n}} + o\left(\frac{1}{\sqrt{n}}\right).$$

c. $$\frac{P_{M,n^*_\beta}\{Ab\}}{E_{M,n^*_\beta}[W_q]} \sim g_0.$$

QED Regime

$c < 1/2$

Assume $\Lambda = \lambda + \mu^{1-c}\lambda^c \cdot X, \quad c < 1/2$, $E[X] = 0$ and finite $\sigma(X) > 0$. Let $\lambda \to \infty$.

Lemma

Assume the square-root staffing level

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \quad -\infty < \beta < \infty$$

Then, asymptotically, the random part of the arrival rate does not affect the system's performances. Namely, the queue is asymptotically equivalent to the $M|M|n + G$ queue with deterministic arrival rate λ, for all $c < 1/2$.
The $M|M|n+G$ queue was analyzed comprehensively by Zeltyn['05].

QED Regime

$c < 1/2$

Assume $\Lambda = \lambda + \mu^{1-c}\lambda^c \cdot X, \quad c < 1/2$, $E[X] = 0$ and finite $\sigma(X) > 0$. Let $\lambda \to \infty$.

Lemma

Assume the square-root staffing level

$$n = \frac{\lambda}{\mu} + \beta\sqrt{\frac{\lambda}{\mu}} + o(\sqrt{\lambda}), \quad -\infty < \beta < \infty$$

Then, asymptotically, the random part of the arrival rate does not affect the system's performances. Namely, the queue is asymptotically equivalent to the $M|M|n + G$ queue with deterministic arrival rate λ, for all $c < 1/2$.
The $M|M|n+G$ queue was analyzed comprehensively by Zeltyn['05].

ED Regime

$c = 1$, Discrete Random Arrival Rate

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

ED Regime

$c = 1$, Discrete Random Arrival Rate

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

Let

$$\kappa = \operatorname*{argmin}_s \left\{ \sum_{i=1}^{s} x_i p_i \geq \alpha \right\}.$$

Assume that the inequality is **strict**.

Define

$$\alpha^* \triangleq \frac{\alpha - \sum_{i=1}^{\kappa-1} x_i p_i}{x_\kappa p_\kappa}.$$

ED Regime

Theorem (Staffing)

a. The optimal staffing level satisfies

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

here β^* is the unique solution of the equation

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

b. Introduce the staffing level

$$n^*_\beta = \left\lceil \frac{\lambda x_\kappa}{\mu} + \beta^* \sqrt{\frac{\lambda x_\kappa}{\mu}} \right\rceil.$$

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

▸ Proof

ED Regime

What if $\alpha = \sum_{i=1}^{\kappa} x_i p_i$ **?**

ED Regime

What if $\alpha = \sum_{i=1}^{\kappa} x_i p_i$ **?**

$$n = \frac{\lambda x_{\kappa+1}}{\mu} + \beta \cdot h(\lambda),$$

ED Regime

What if $\alpha = \sum_{i=1}^{\kappa} x_i p_i$ **?**

$$n = \frac{\lambda x_{\kappa+1}}{\mu} + \beta \cdot h(\lambda),$$

1. $\frac{h(\lambda)}{\sqrt{\lambda}} \rightarrow$ constant

ED Regime

What if $\alpha = \sum_{i=1}^{\kappa} x_i p_i$ **?**

$$n = \frac{\lambda x_{\kappa+1}}{\mu} + \beta \cdot h(\lambda),$$

1. $\dfrac{h(\lambda)}{\sqrt{\lambda}} \rightarrow \text{constant} \Rightarrow \widetilde{P}_{M,n}\{W_q > 0\} > \alpha. \quad (\beta \in \mathbb{R})$

ED Regime

What if $\alpha = \sum_{i=1}^{\kappa} x_i p_i$ **?**

$$n = \frac{\lambda x_{\kappa+1}}{\mu} + \beta \cdot h(\lambda),$$

1. $\frac{h(\lambda)}{\sqrt{\lambda}} \rightarrow \text{constant} \Rightarrow \widetilde{P}_{M,n}\{W_q > 0\} > \alpha. \quad (\beta \in \mathbb{R})$
2. $\frac{h(\lambda)}{\sqrt{\lambda}} \rightarrow \infty$

ED Regime

What if $\alpha = \sum_{i=1}^{\kappa} x_i p_i$ **?**

$$n = \frac{\lambda x_{\kappa+1}}{\mu} + \beta \cdot h(\lambda),$$

1. $\frac{h(\lambda)}{\sqrt{\lambda}} \rightarrow \text{constant} \Rightarrow \widetilde{P}_{M,n}\{W_q > 0\} > \alpha. \quad (\beta \in \mathbb{R})$
2. $\frac{h(\lambda)}{\sqrt{\lambda}} \rightarrow \infty \Rightarrow \widetilde{P}_{M,n}\{W_q > 0\} = \alpha. \quad (\beta > 0)$

ED Regime

Theorem (Performance Measures)

Under the staffing level n^*_β, as $\lambda \to \infty$,

a. The long-term abandonment probability:

$$\widetilde{P}_{M,n^*_\beta}\{Ab\} \sim \sum_{i=1}^{\kappa-1} p_i(x_i - x_\kappa).$$

b. Mean server's utilization:

$$E_{M,n^*_\beta}[U] \sim \sum_{i=1}^{\kappa} p_i + \sum_{i=\kappa+1}^{I} p_i \cdot \frac{x_i}{x_\kappa}.$$

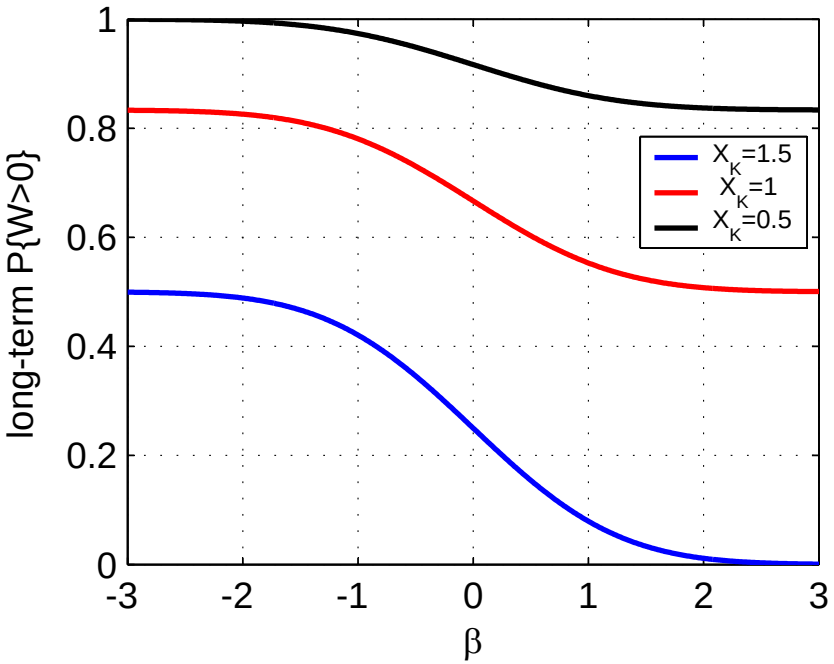
c. Assume that for each $i < \kappa$ the equation $G(y) = 1 - \frac{x_\kappa}{x_i}$ has a unique solution y_i^*, and $g(y_i^*) > 0$. Then, the long-term average waiting time

$$\widetilde{E}_{M,n^*_\beta}[W_q] \sim \sum_{i=1}^{\kappa-1} \left[x_i p_i \cdot \int_0^{y_i^*} \bar{G}(u)du \right].$$

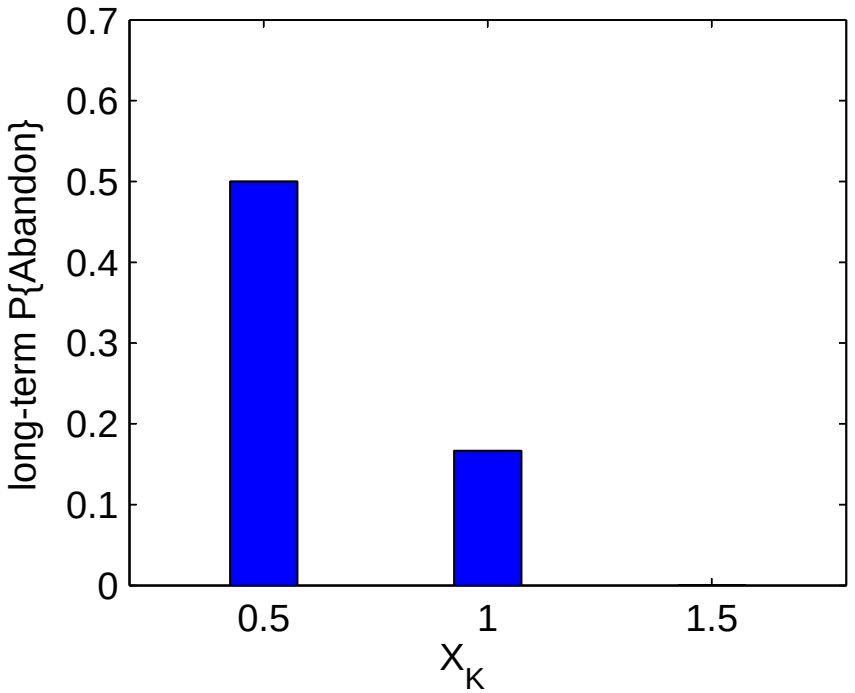
ED Regime

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

Delay Probability



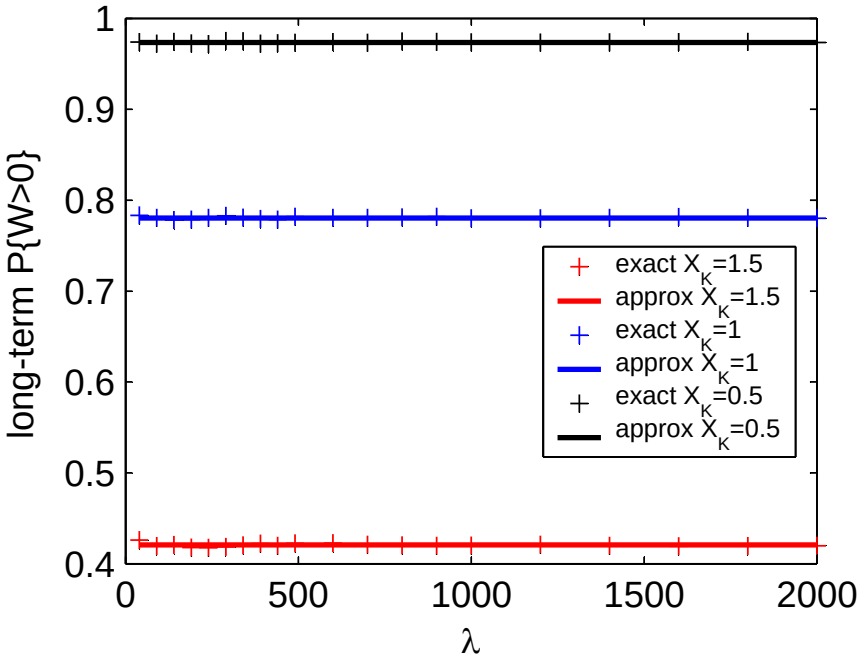
Abandonment Probability



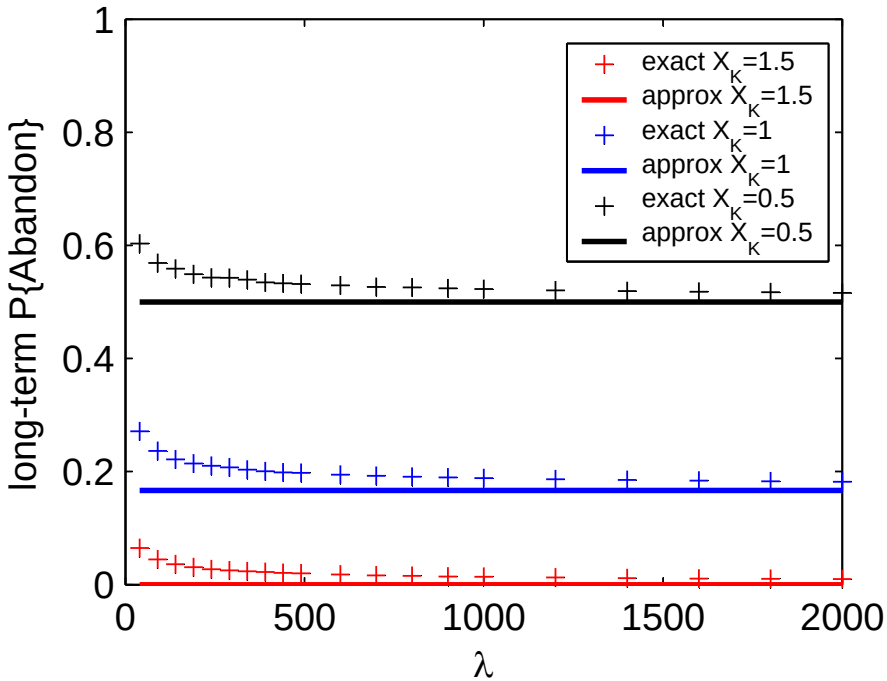
ED Regime

$$\beta = -1$$

Delay Probability



Abandonment Probability



ED Regime

$c = 1$, Continuous Random Arrival Rate

Assume $M = \lambda X$, where X is a continuous random variable with strictly continuous cdf F over the distribution support. Let $E[X] = 1$ and $\sigma(X) < \infty$. Assume $\lambda \to \infty$.

Theorem (Staffing)

a. The optimal staffing level satisfies

$$n^* = \frac{\lambda}{\mu} \cdot \delta^* + o(\lambda), \quad \delta^* \in supp(f),$$

where δ^* is the unique solution of the equation

$$\alpha = \int_{\delta}^{\infty} x dF(x).$$

b. Introduce the staffing level $n_\delta^* = \left\lceil \frac{\lambda \delta^*}{\mu} \right\rceil$.

Then the staffing level n_δ^* is both *asymptotically feasible*, as well as *asymptotically optimal* ($f(\lambda) = \lambda$).

ED Regime

Theorem (Performance Measures)

Under the staffing level n_δ^*, as $\lambda \to \infty$,

a. The short-term delay probability:

$$\bar{P}_{M,n_\delta^*}\{W_q > 0\} \sim \bar{F}(\delta).$$

b. The long-term abandonment probability:

$$\widetilde{P}_{M,n_\delta^*}\{Ab\} \sim (E[X|X > \delta] - \delta) \cdot \bar{F}(\delta).$$

c. The short-term abandonment probability:

$$\bar{P}_{M,n_\delta^*}\{Ab\} \sim E[1 - \delta/X|X > \delta] \cdot \bar{F}(\delta).$$

ED Regime

Theorem (Performance Measures). Continued.

d. Mean server's utilization:

$$E_{M,n_\delta^*}[U] \sim \frac{1}{\delta} \cdot E[X|X<\delta] \cdot F(\delta) + \bar{F}(\delta).$$

e. Assume that for each $x > \delta$ the equation $G(y_x) = 1 - \frac{\delta}{x}$ has a unique solution y_x^*, and $g(y_x^*) > 0$. Then, the long-term average waiting time

$$\widetilde{E}_{M,n_\delta^*}[W_q] \sim \int_\delta^\infty \left(x \cdot \int_0^{y_x^*} \bar{G}(u)du \right) dF(x).$$

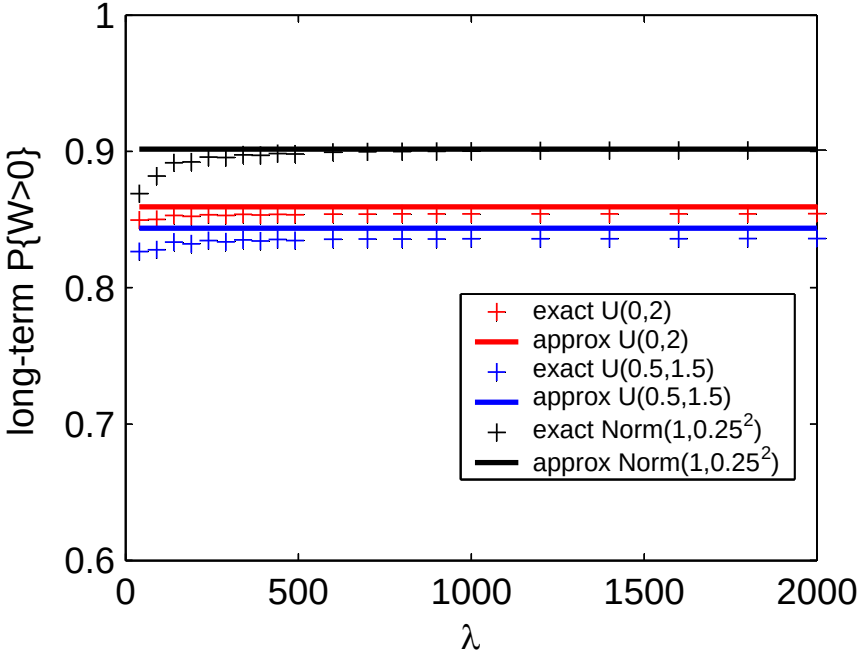
ED Regime

Examples: Consider three distributions of X

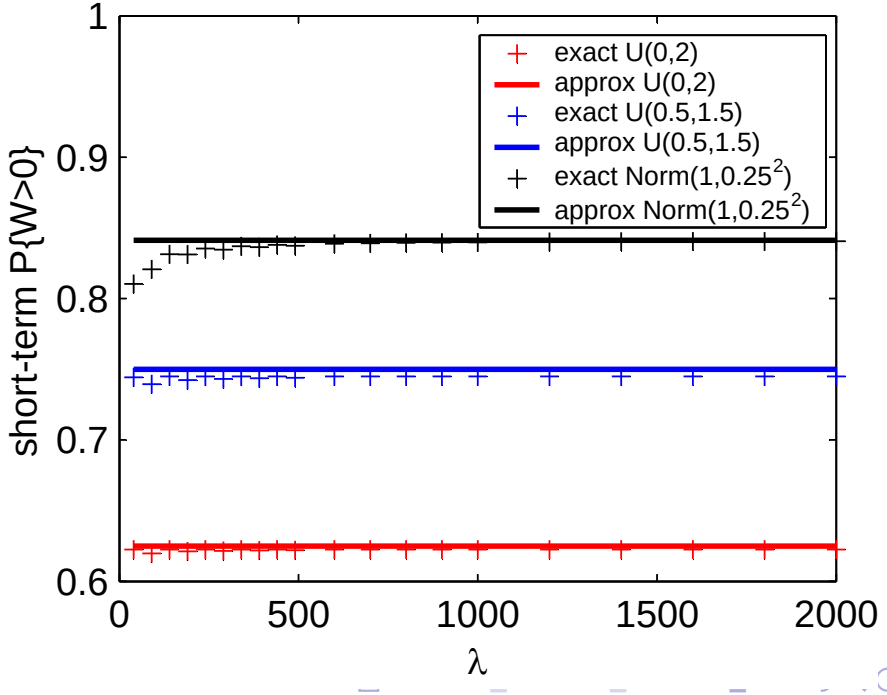
- Uniform distribution on [0,2].
- Uniform distribution on [0.5,1.5].
- Normal distribution with mean=1 and std=0.25.

$$\delta = 0.75$$

Long-Term $P\{W_q > 0\}$



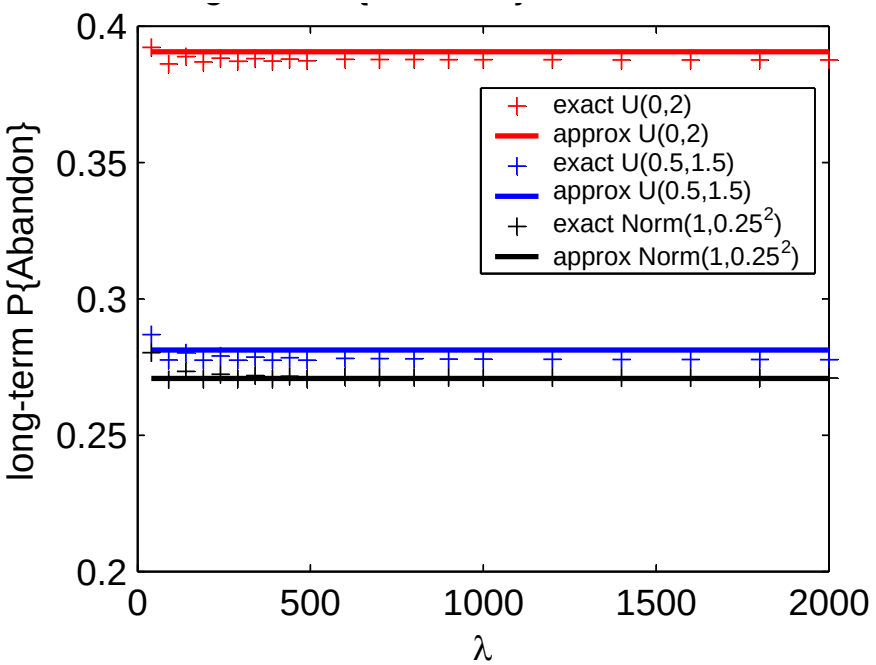
Short-Term $P\{W_q > 0\}$



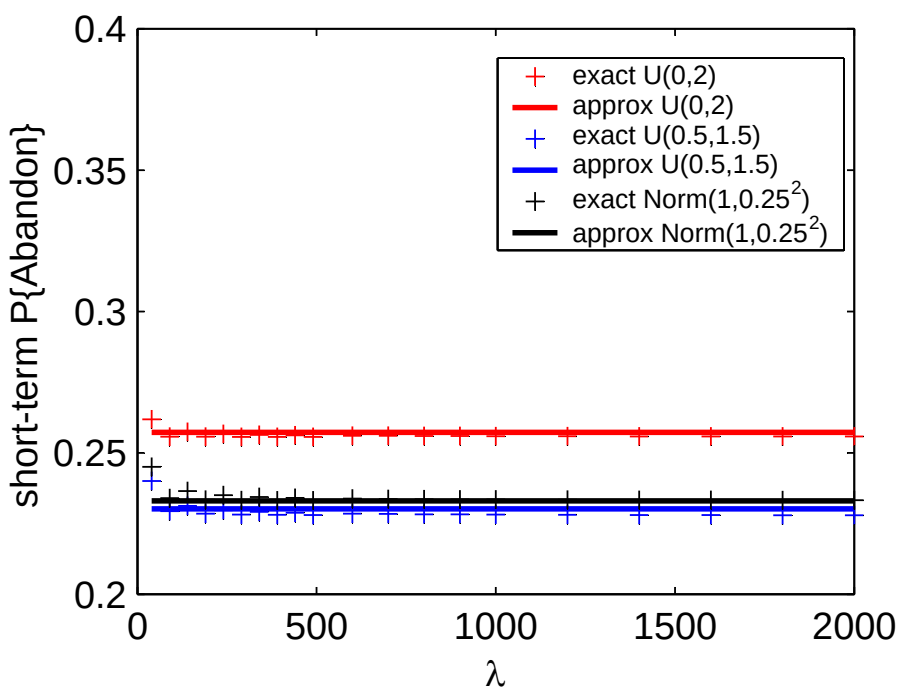
ED Regime

$$\delta = 0.75$$

Long-Term $P\{Ab\}$



Short-Term $P\{Ab\}$



Time - Varying Queues

- Based on the ISA (Iterative Staffing Algorithm), a simulation code developed by Feldman['04] with the features of random arrival rate in the time varying $M/M/n+G$ queue.
- The goal is to determine time-dependence staffing levels aiming to achieve a given constant-over-time long- term delay delay probability, α.

Time - Varying Queues

- Based on the ISA (Iterative Staffing Algorithm), a simulation code developed by Feldman['04] with the features of random arrival rate in the time varying $M/M/n+G$ queue.
- The goal is to determine time-dependence staffing levels aiming to achieve a given constant-over-time long- term delay delay probability, α.
- Example 1: $c = 1$, Discrete Random Arrival Rate
  1. $M_t = \lambda_t \cdot X$, where

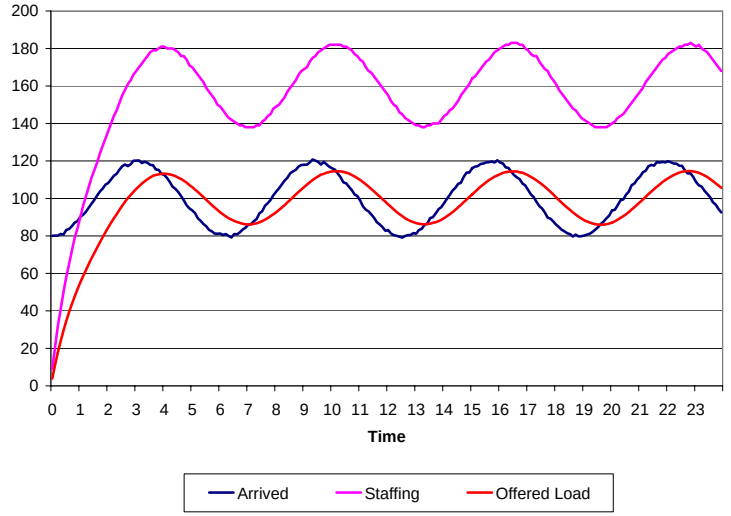
$$\lambda_t = 100 - 20 \cdot \cos(t), \quad \text{and} \quad X = \begin{cases} 1.5 & w.p.\ 1/3 \\ 1 & w.p.\ 1/3 \\ 0.5 & w.p.\ 1/3 \end{cases}$$

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

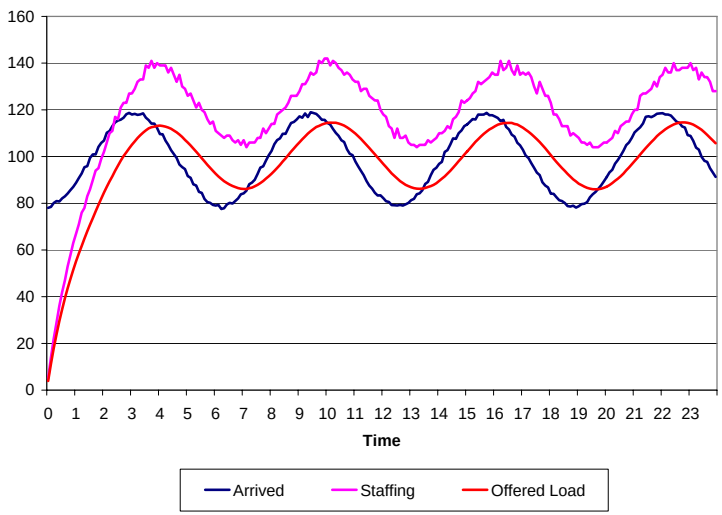
Time - Varying Queues

Arrivals, Offered Load and Staffing Level

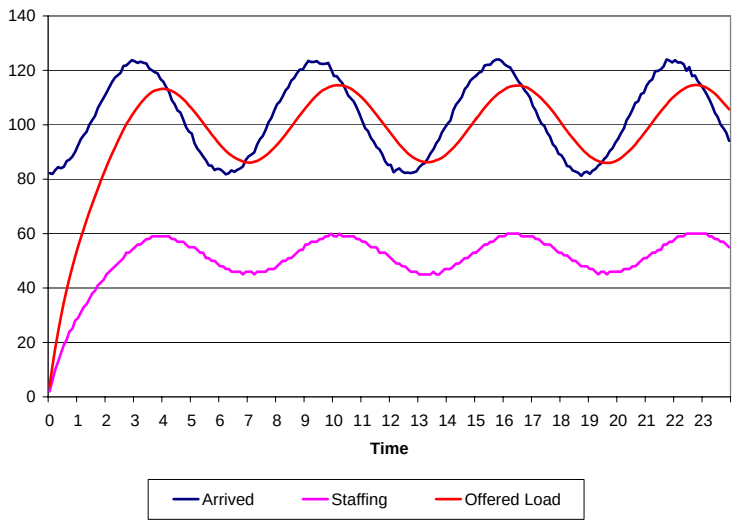
Target α=0.1



Target α=0.5



Target α=0.9

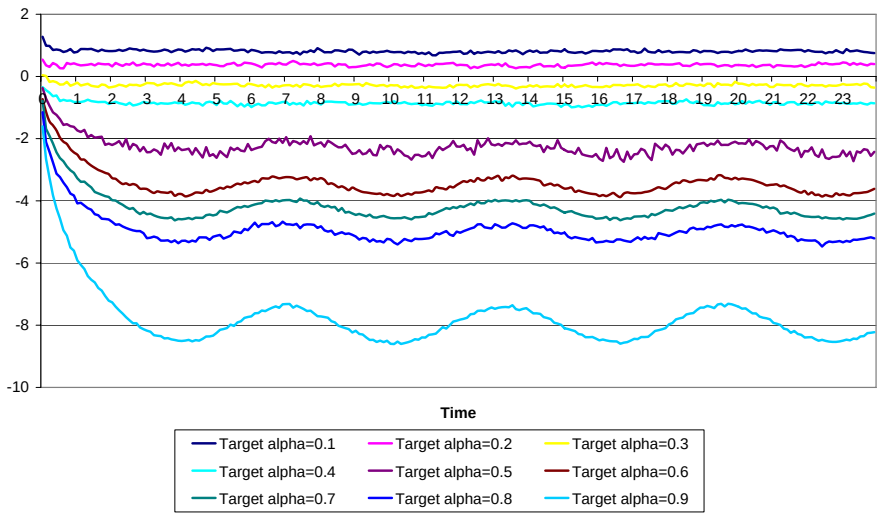


Time - Varying Queues

- Theoretical staffing level (homogenous arrival rate) depends both on X and β:

$$n = \frac{\lambda x_\kappa}{\mu} + \beta\sqrt{\frac{\lambda x_\kappa}{\mu}}.$$

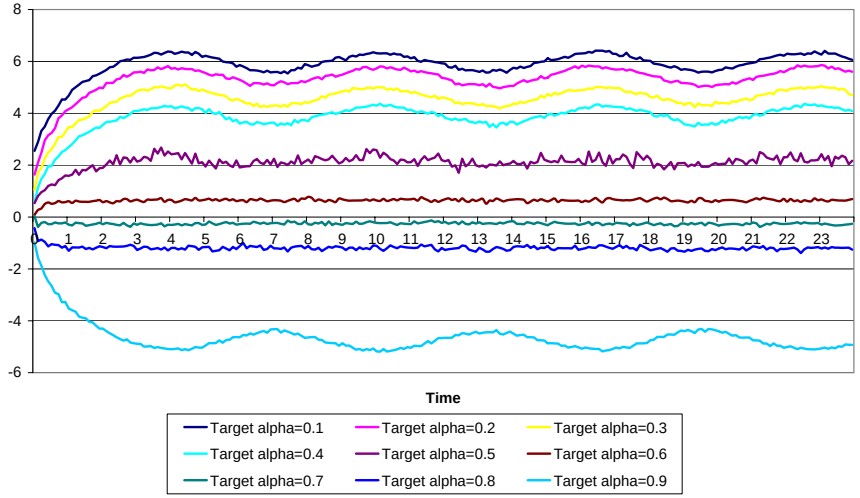
β **for** $x_\kappa = 1.5$



$\Rightarrow$ β **is stabilized for target** $\alpha = 0.1, 0.2, 0.3$ **and** 0.4**.**

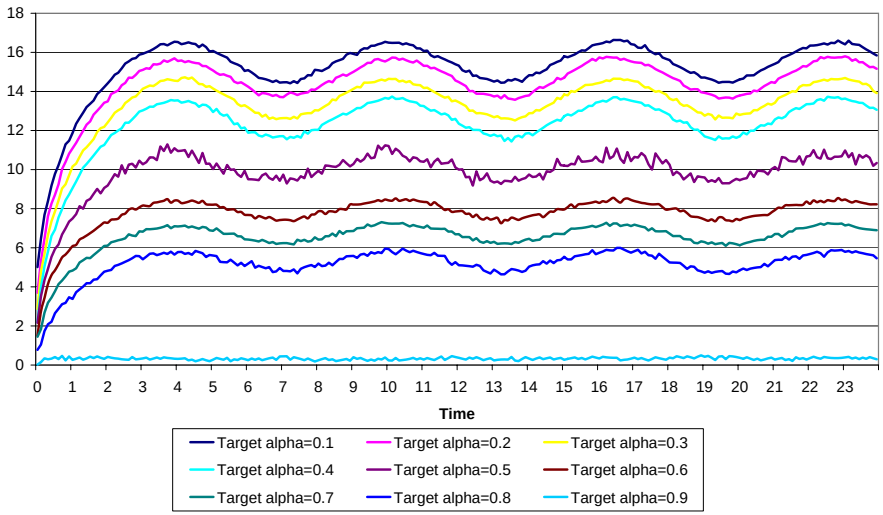
Time - Varying Queues

β for $x_\kappa = 1$



$\Rightarrow$ β **is stabilized for target** $\alpha = 0.6, 0.7$ **and** 0.8**.**

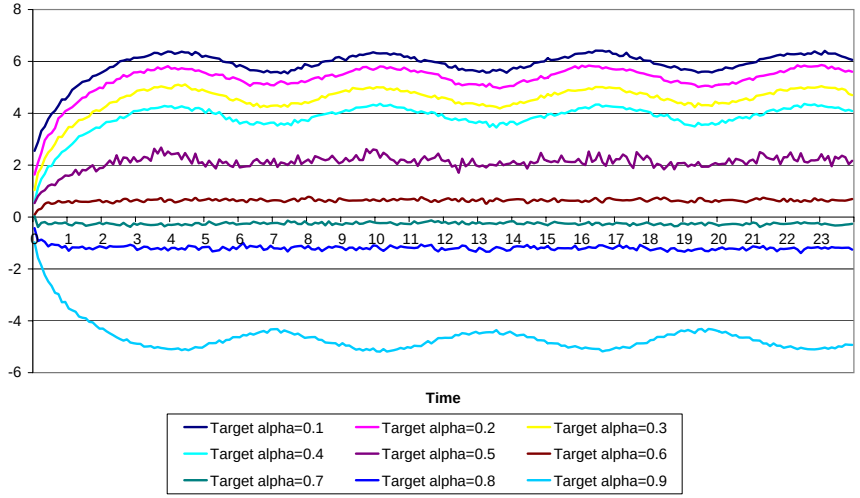
β for $x_\kappa = 0.5$



$\Rightarrow$ β **is stabilized for target** $\alpha = 0.9$

Time - Varying Queues

β **for** $x_\kappa = 1$

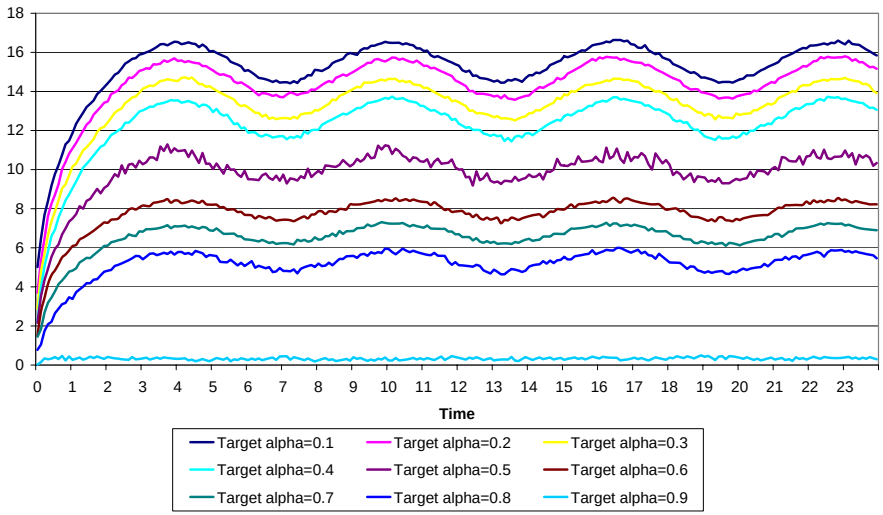


$\Rightarrow$ β **is stabilized for target** $\alpha = 0.6, 0.7$ **and** 0.8**.**

What about target $\alpha = 0.5$?

Extreme values (-2.3, 2.1 & 10) and noisy

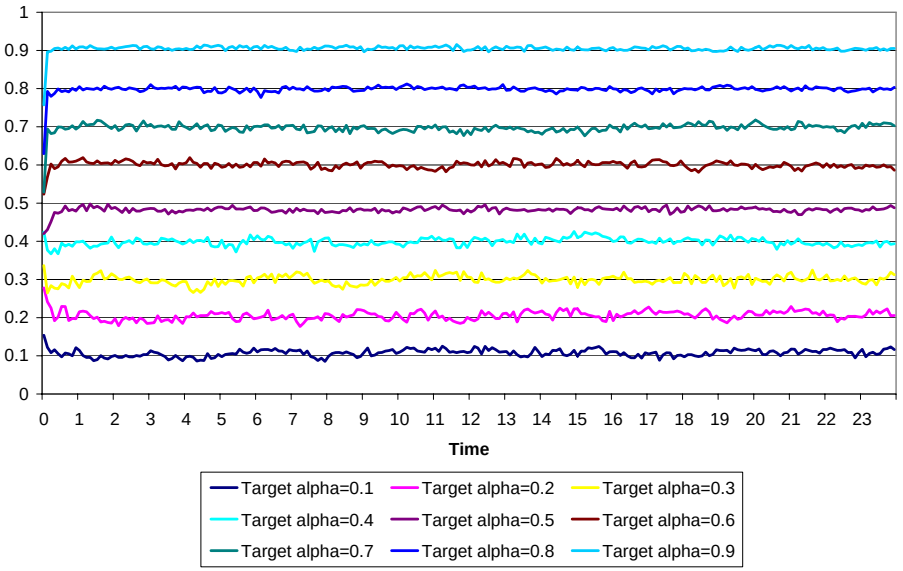
β **for** $x_\kappa = 0.5$



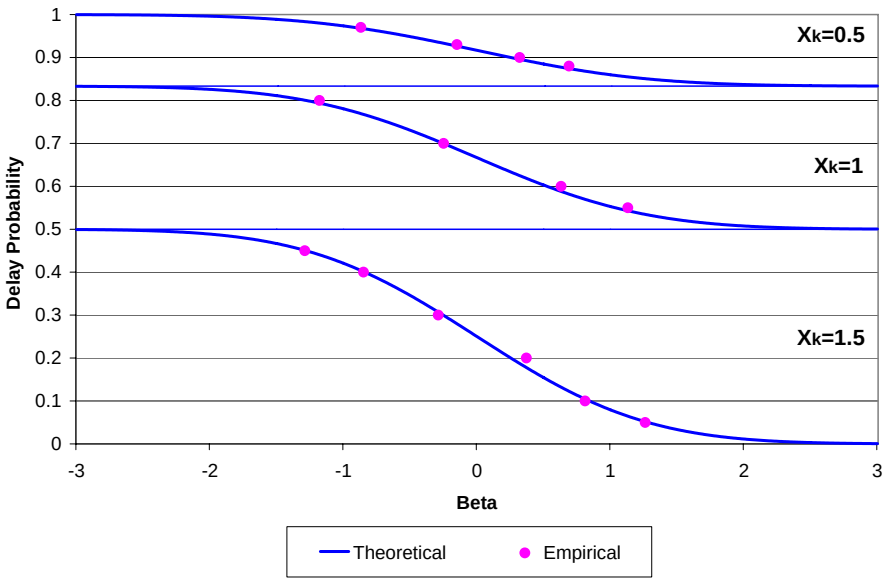
$\Rightarrow$ β **is stabilized for target** $\alpha = 0.9$

Time - Varying Queues

Delay Probability (Target)

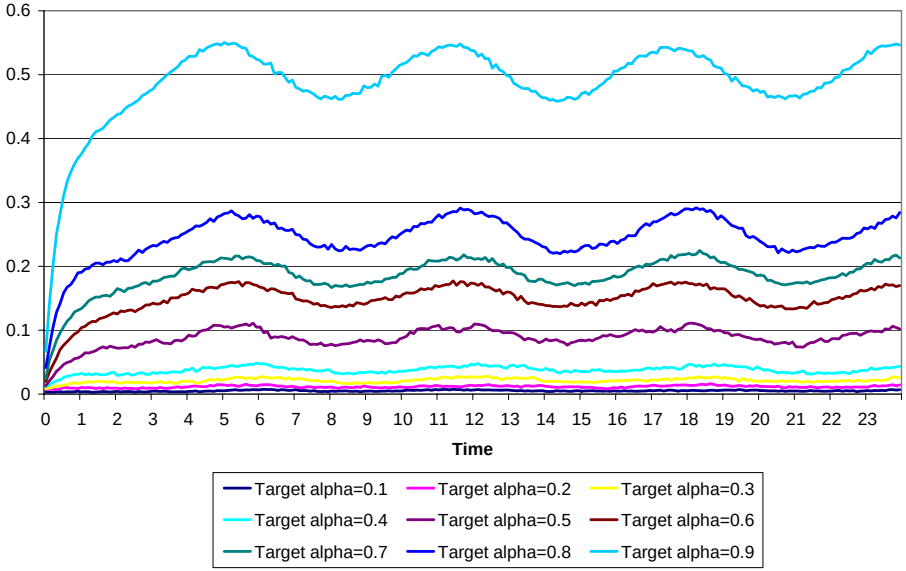


Theoretical and Empirical Delay Probability vs. x_κ and β

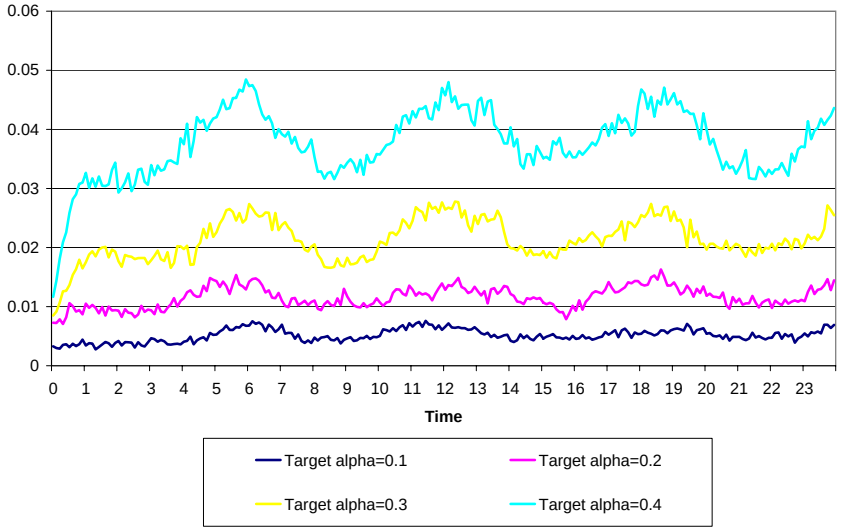


Time - Varying Queues

Abandonment Probability

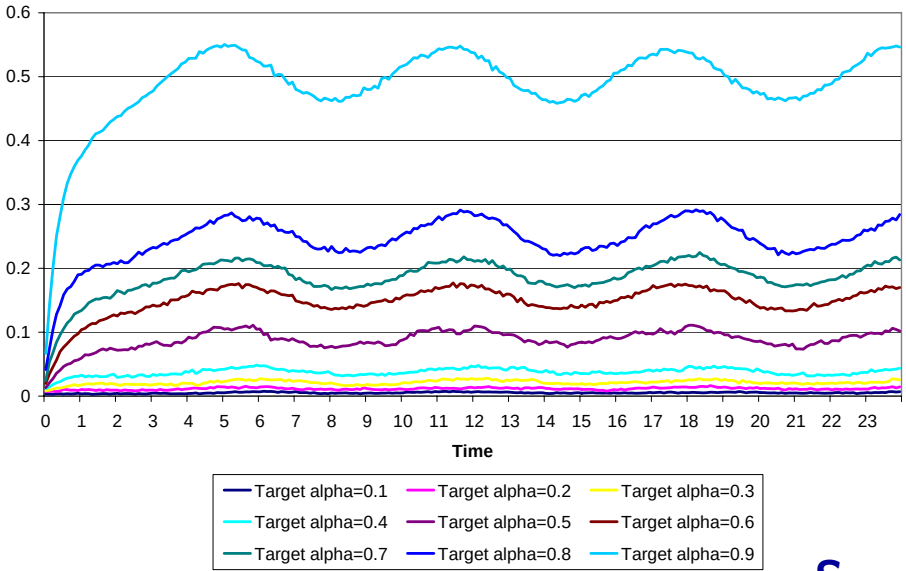


Abandonment Probability - Low α's

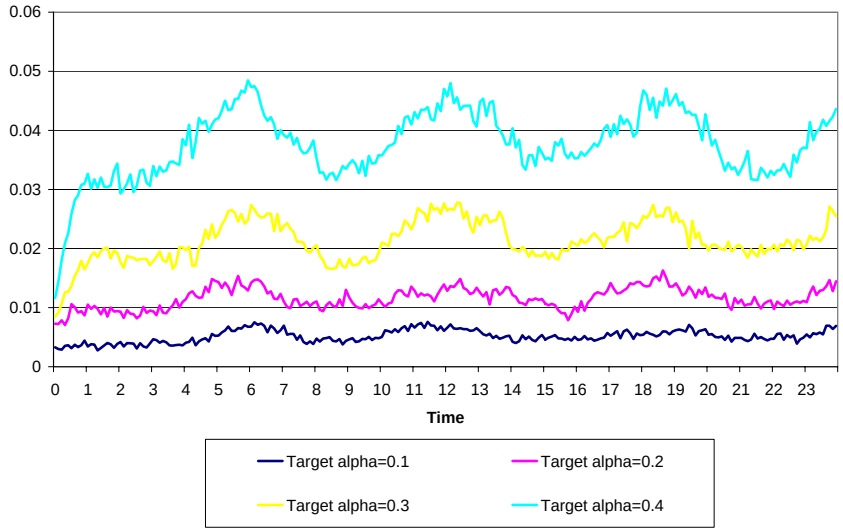


Time - Varying Queues

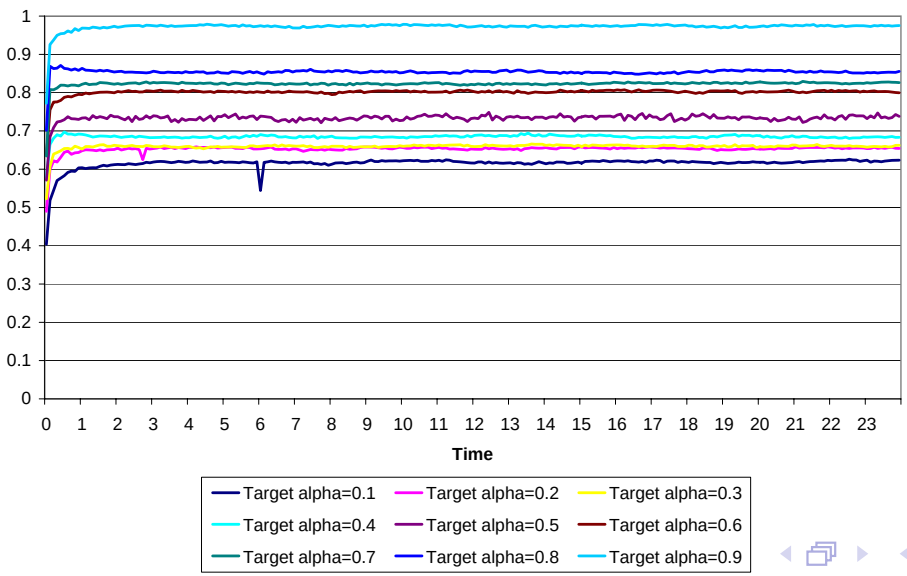
Abandonment Probability



Abandonment Probability - Low α's



Servers' Utilization



Time - Varying Queues

- Example 2: $c = 1$, Continuous Random Arrival Rate
  1. $M_t = \lambda_t \cdot X$, where

$$\lambda_t = 100 - 20 \cdot \cos(t), \quad \text{and} \quad X \sim \text{Uni}[0.5, 1.5].$$

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

Time - Varying Queues

- Example 2: $c = 1$, Continuous Random Arrival Rate
  1. $M_t = \lambda_t \cdot X$, where

$$\lambda_t = 100 - 20 \cdot \cos(t), \quad \text{and} \quad X \sim \text{Uni}[0.5, 1.5].$$

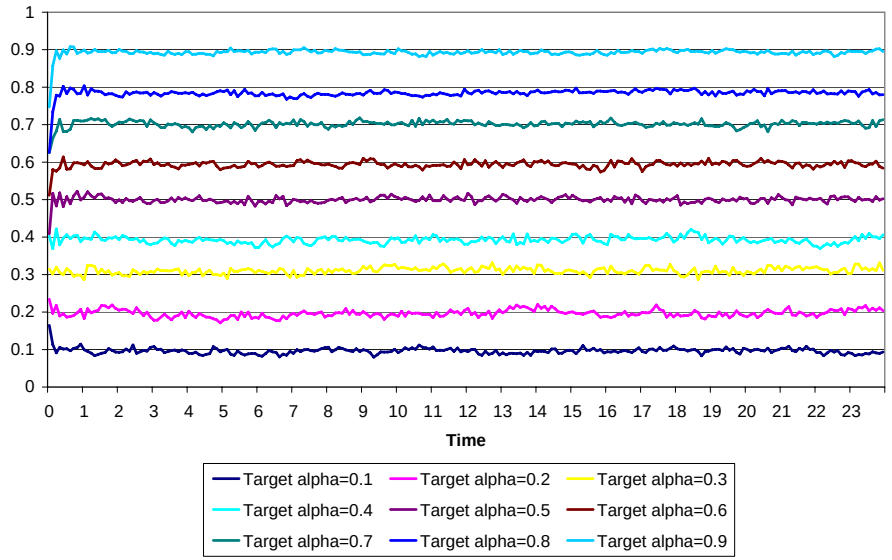
  2. Service time and patience are distributed exponentially with mean 1 ($\mu = \theta = 1$).

- Theoretical staffing level (homogenous arrival rate) depends on the QOS parameter δ:

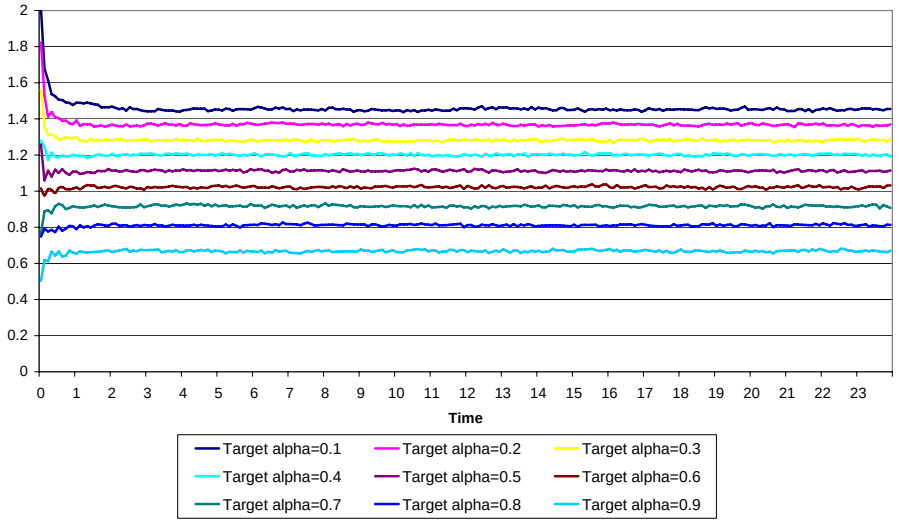
$$n = \frac{\lambda}{\mu} \cdot \delta$$

Time - Varying Queues

Delay Probability (Target)

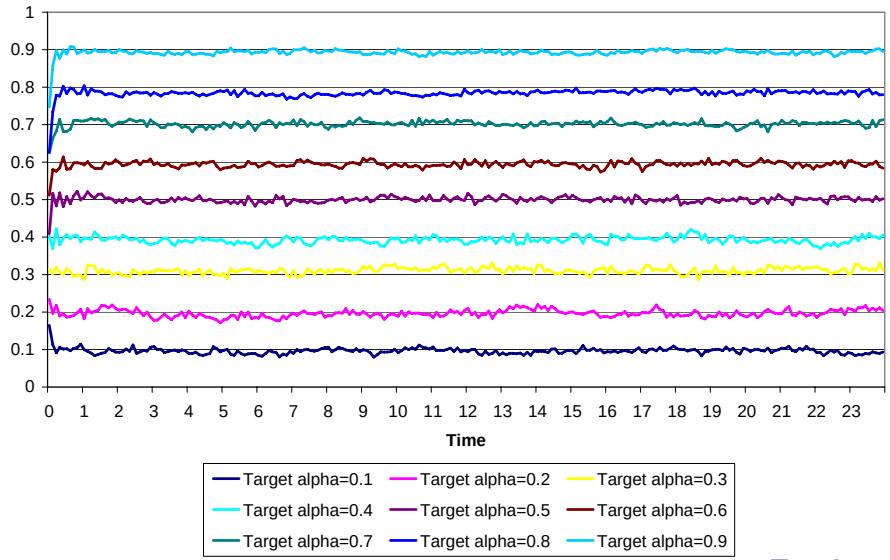


QOS δ

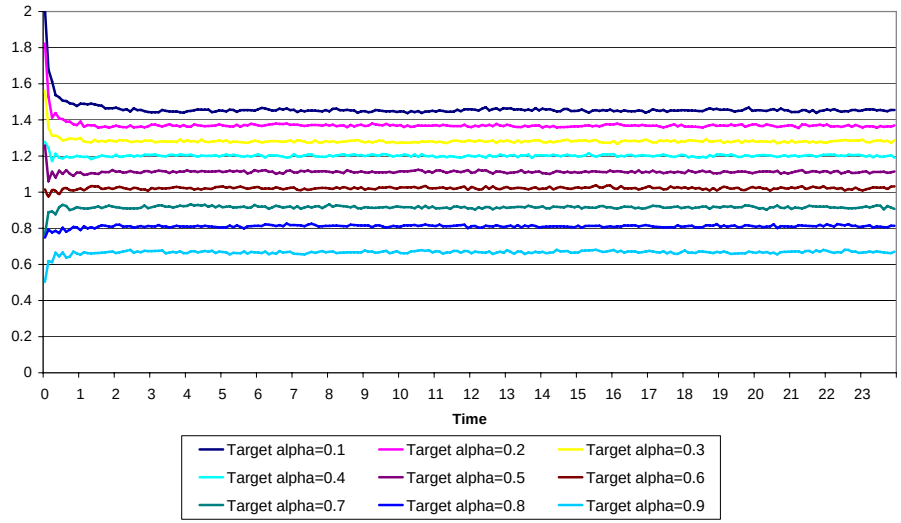


Time - Varying Queues

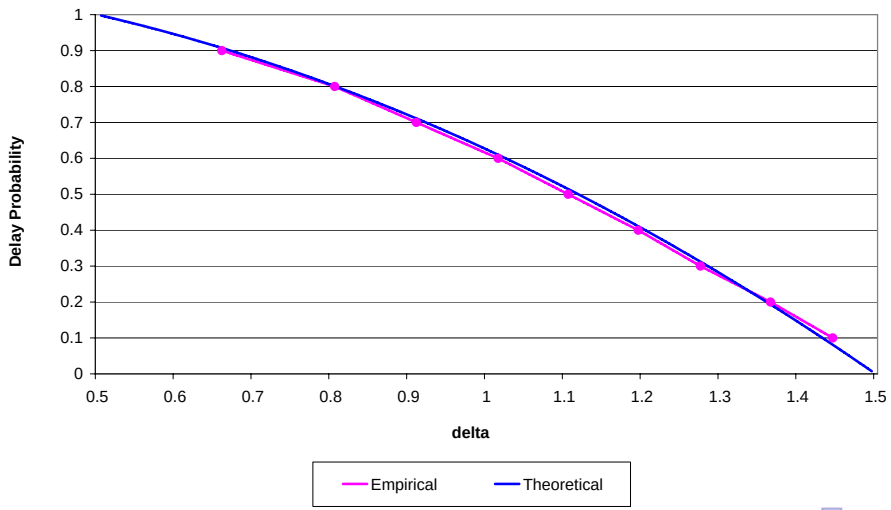
Delay Probability (Target)



QOS δ



Delay Probability vs. δ



Thank You

Appendix 1

$c = 1/2, \quad \gamma_x^*$

Denote

$$\begin{aligned} \beta_x^* &= \beta^* - x, \\ \hat{\beta}_x^* &= \beta_x^* \sqrt{\frac{\mu}{g_0}} = (\beta^* - x) \cdot \sqrt{\frac{\mu}{g_0}}. \end{aligned}$$

Then,

$$\gamma_x^* = \alpha \cdot \beta_x^* \cdot \left[\frac{h(\hat{\beta}_x^*)}{\hat{\beta}_x^*} - 1 \right].$$

Back

Appendix 2

Proof Outline ($c = 1$, Discrete Random Arrival Rate)

1. $$n = \frac{\lambda x_\kappa}{\mu} + \beta\sqrt{\frac{\lambda x_\kappa}{\mu}} + o(\sqrt{\lambda}) = \frac{\lambda x_i}{\mu} \cdot \frac{x_\kappa}{x_i} + o(\lambda)$$
   $\Rightarrow$ for $i < \kappa$ $(x_i > x_\kappa)$ we are in the ED regime, for $i = \kappa$ in the QED regime, and for $i > \kappa$ $(x_i < x_\kappa)$ in the QD regime.

2. $$\lim_{\lambda\to\infty} \widetilde{P}_{M,n}\{W_q > 0\} = \lim_{\lambda\to\infty} E[XP_{M,n}\{W_q > 0\}]$$
   $$= E \lim_{\lambda\to\infty} \{XP_{M,n}\{W_q > 0\}\} = \sum_{i=1}^{\kappa-1} x_i p_i + x_\kappa p_\kappa \cdot \alpha^*.$$

3. For a certain value x_s of X, the asymptotic long-term delay probability is bounded in the open interval $A_s = \left(\sum_{i=1}^{s-1} x_i p_i,\ \sum_{i=1}^{s} x_i p_i\right)$. $\Rightarrow$ There is a unique s which for $\alpha \in A_s$.

Appendix 2

Proof Outline ($c = 1$, Discrete Random Arrival Rate)

4. For some $\epsilon > 0$ define

$$n_1 \triangleq \left\lceil \frac{\lambda x_\kappa}{\mu} + (\beta^* - \epsilon)\sqrt{\frac{\lambda x_\kappa}{\mu}} \right\rceil ; \quad n_2 \triangleq \left\lfloor \frac{\lambda x_\kappa}{\mu} + (\beta^* + \epsilon)\sqrt{\frac{\lambda x_\kappa}{\mu}} \right\rfloor .$$

- Garnett function is strictly decreasing in β

  $\Rightarrow \widetilde{P}_{M,n_1}\{W_q > 0\} \to \alpha + \tau_1$; $\widetilde{P}_{M,n_2}\{W_q > 0\} \to \alpha - \tau_2$.

  $\Rightarrow$ For large λ

  $\widetilde{P}_{M,n_1}\{W_q > 0\} > \alpha + \tau_1/2$; $\widetilde{P}_{M,n_2}\{W_q > 0\} < \alpha - \tau_2/2$.

- Delay probability is monotonically decreasing in number of servers

  $\Rightarrow n_1 \leq n^* \leq n_2$.

- $\epsilon > 0$ is arbitrary $\Rightarrow$ **q.e.d**.

Back