DataMOCCA

Data MOdels for Call Center Analysis

Project Collaborators:

Technion:	Paul Feigin, Avi Mandelbaum Technion SEElab: Valery Trofimov, Ella Nadjharov, Igor Gavako, Katya Kutsy, Polyna Khudyakov, Shimrit Maman, Pablo Liberman Students (PhD, MSc, BSc), RAs
Wharton:	Larry Brown, Noah Gans, Haipeng Shen (N. Carolina), Students, Wharton Financial Institutions Center
Companies:	U.S. Bank, Israeli Telecom, 2 Israeli Banks, Israeli Hospitals, ...

The SEE Center - Project DataMOCCA

Goal: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing, displaying and interacting with transaction-based data.

The SEE Center - Project DataMOCCA

Goal: Designing and Implementing a (universal) data-base/data-repository and interface for storing, retrieving, analyzing, displaying and interacting with transaction-based data.

Enable the Study of:

- Customers (Callers, Patients)	Waiting, Abandonment, Returns
- Servers (Agents, Nurses)	Service Duration, Activity Profile
- Managers (System)	Loads, Queue Lengths, Trends

DataMOCCA History: The Data Challenge

- Queueing Research lead to Service Operations (Early 90s)
- Services started with Call Centers which, in turn, created data-needs
- Queueing Theory had to expand to Queueing Science: Fascinating
- WFM was Erlang-C based, but customers abandon! (Im)Patience?
- (Im)Patience censored hence Call-by-Call data required: 4-5 years saga
- Finally Data: a small call center in a small IL bank (15 agents, 4 service types, 350K calls per year)
- Technion Stat. Lab, guided by Queueing Science: Descriptive Analysis
- Building blocks (Arrivals, Services, (Im)Patiece): even more Fascinating

DataMOCCA: System Components

1. Clean Databases: Operational histories of **individual** customers and servers (mostly with IDs).
   - In Call Centers: from IVR to Exit;
   - In Hospitals: from ED to Exit (or just ED).

2. **SEEStat: Online** GUI (friendly, flexible, powerful)
   - Queueing-Science perspective;
   - Operational data (vs. financial, contents or clinical);
   - Flexible customization (e.g. seconds to months);

3. Tools:
   - Online statistics (survival analysis, mixtures, smoothing);
   - Dynamic Graphs (flow-charts, work-flows)
   - Simulators (CC, ED; data-driven).

Current Databases

1. U.S. Bank (**PUBLIC**): 220M calls, 40M agent-calls, 1000 agents, 2.5 years, 7-40GB.
2. Israeli Banks:
   - Small (**PUBLIC**): 350K calls, 15 agents, 1 year. Started it all in 1999 (JASA), now "romancing" again (Medium, with 300 agents);
   - Large (ongoing): 500 agents, 1.5 years, 3-8GB.
3. Israeli Telecom (ongoing): 800 agents, 3.5 years; 5-55GB.
4. Israeli Hospitals:
   - Six ED's (to be made **PUBLIC**);
   - Large (ongoing): 1000 beds, 45 medical units, 75,000 patients hospitalized yearly, 4 years, 7GB.
5. Website (pilot).

DataMOCCA: Future

- Operational (ACD) data with Business (CRM) data, Contents/Medical
- Contact Centers: IVR, Chats, Emails; Websites
- Daily update (as opposed to montly DVDs)
- Web-access (Research; Applications, e.g. CC/ED Simulation; Teaching)
- Nurture Research, for example
  - Skills-Based Routing: Control, staffing, design, online; HRM
  - The Human Factor: Service-anatomy, agents learning, incentives
- Hospitals (OCR: with IBM, Haifa hospital): Operational, Human-Factors, Medical & Financial data; RFIDs for flow-tracking

DataMOCCA Interface: SEEStat

- Daily / monthly / yearly reports & flow-charts for a complete operational view.

- Graphs and tables, in customized resolutions (month, days, hours, minutes, seconds) for a variety of (pre-designed) operational measures (arrival rates, abandonment counts, service- and wait-time distribution, utilization profiles,).

- Graphs and tables for new user-defined measures.

- Direct access to the raw (cleaned) data: export, import.

- Online Statistics: Survival Analysis, Mixtures, Smoothing, Graphics.

Data-Based Research: Must (?) & Fun

- Contrast with "EmpOM": Industry / Company / Survey Data (Social Sciences)

- Converge to: Measure, Model, Validate, Experiment, Refine (Physics, Biology, ...) - The Scientific Paradigm

- Prerequisites: OR/OM, (Marketing) for Design; Computer Science, Information Systems, Statistics for Implementation

- Outcomes: Relevance, Credibility, Interest; Pilot (eg. Healthcare, Web). Moreover,

  Teaching: Class, Homework (Experimental Data Analysis); Cases.

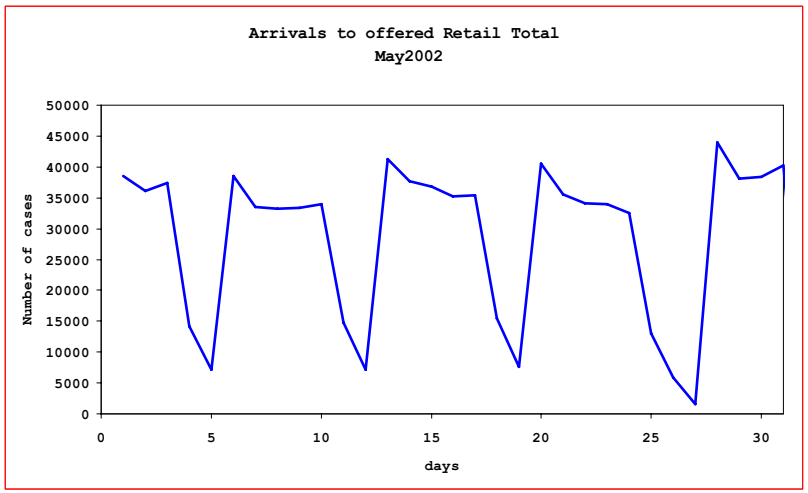
  Research: Test (Queueing) Theory / Laws, Stimulate New Models / Theory.

  Practice: OM Tools (Scenario Analysis), Mktg (Trends, Benchmarking).

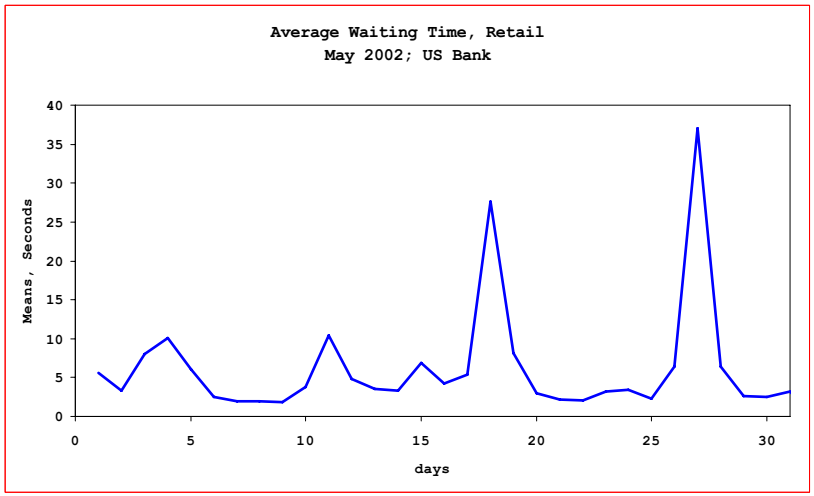
Little's Law, $L = \lambda \cdot W$

US Bank: Retail calls, May 2002

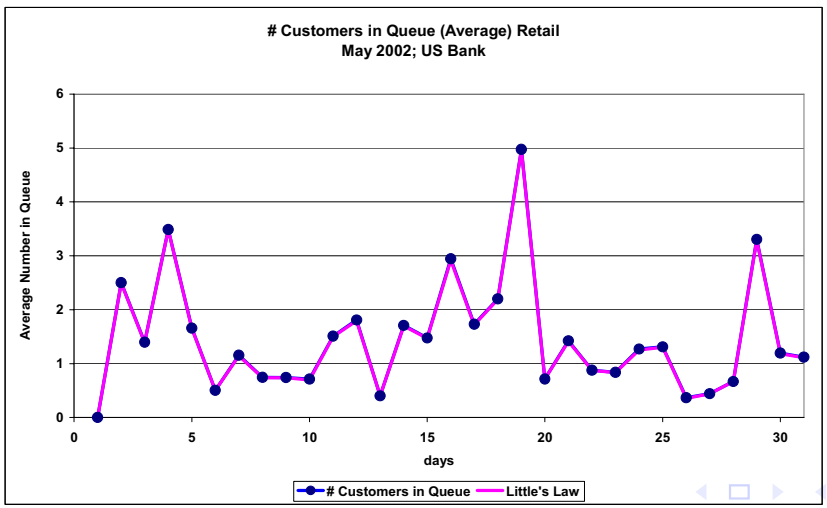
λ, Throughput Rate



W, Average Waiting Time

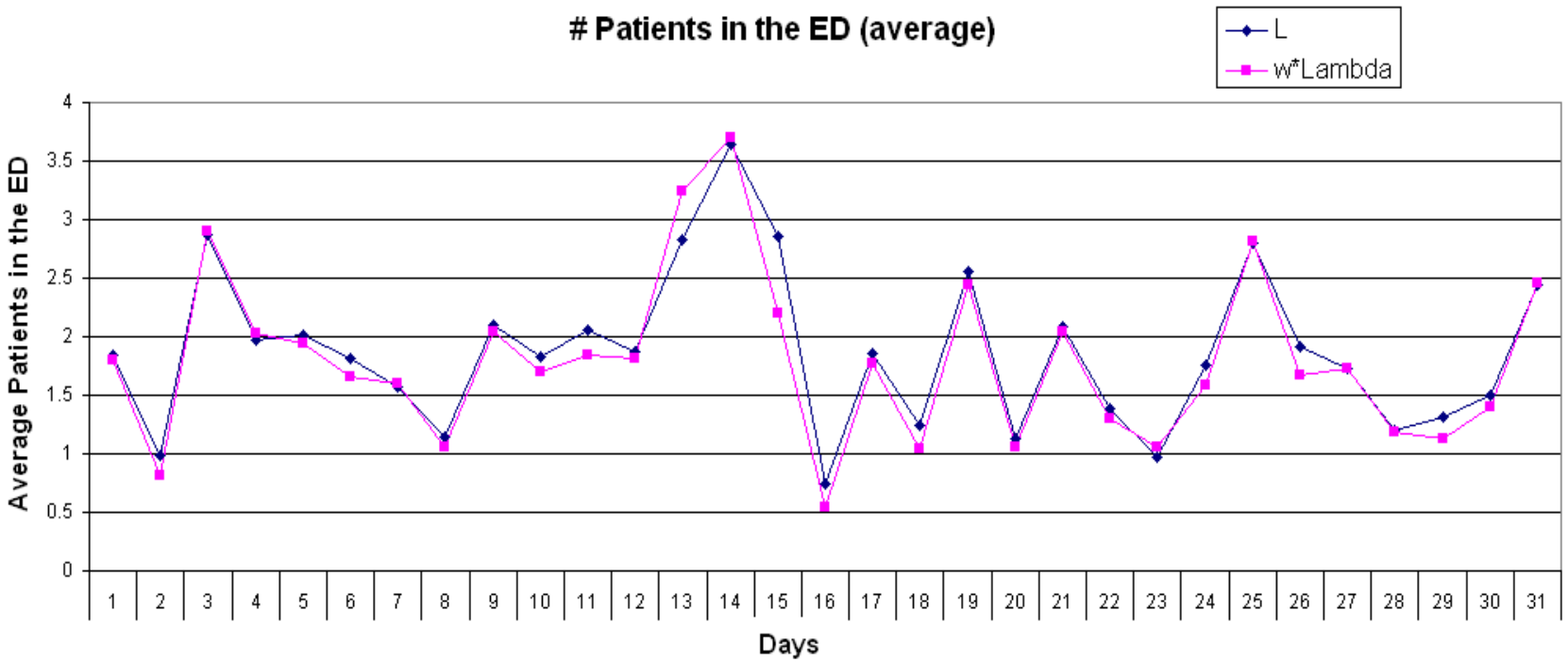


L, Average Queue Length



Little's Law, $L = \lambda \cdot W$

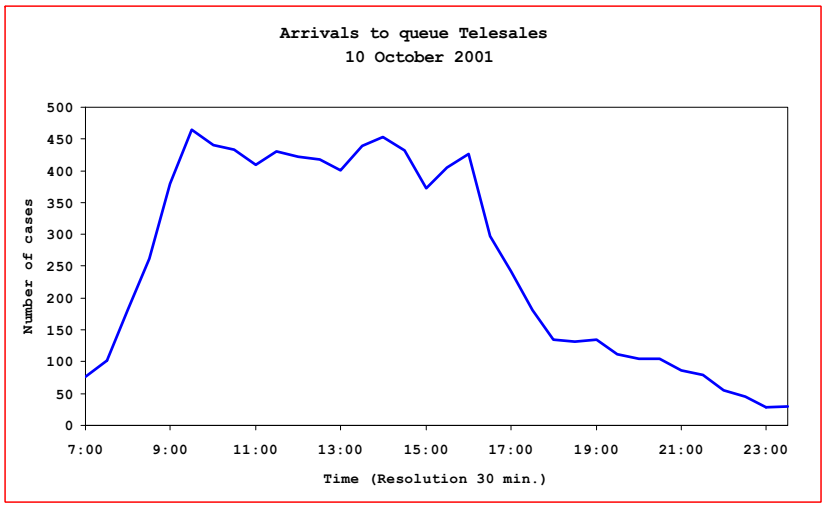
Israeli ED, October 1999, Day Resolution



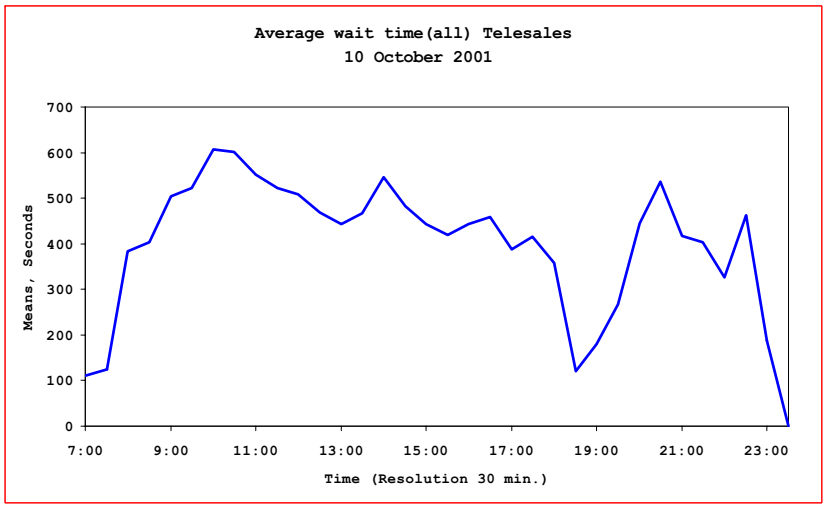
Little's Law, $L = \lambda \cdot W$

US Bank: Telesales Calls, October 10, 2001

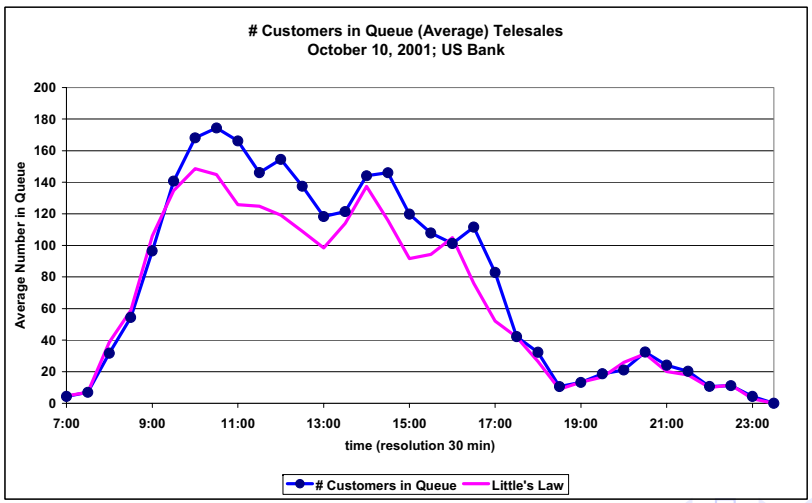
λ, Throughput Rate



W, Average Waiting Time

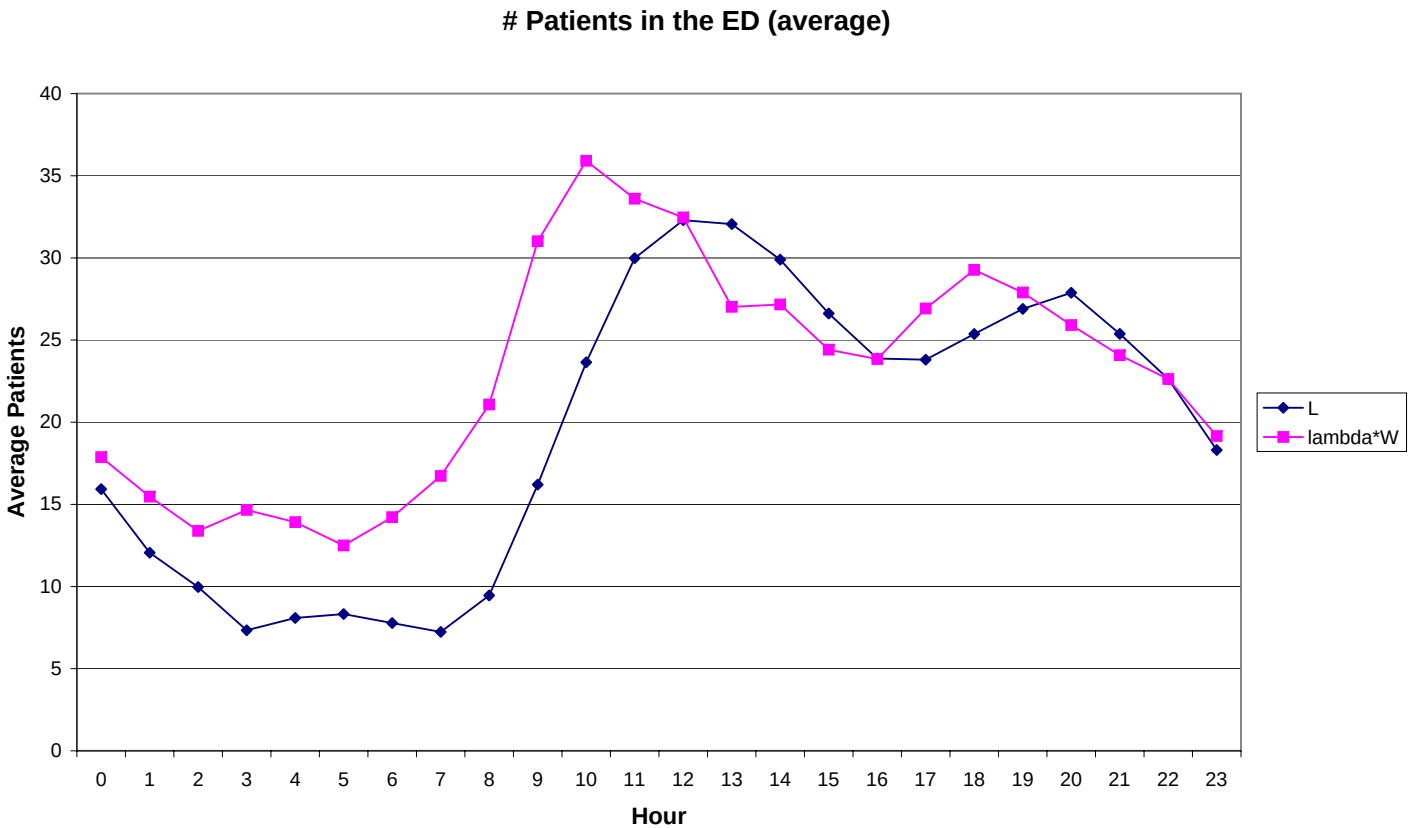


L, Average Queue Length



Little's Law, $L = \lambda \cdot W$

Israeli ED, Hour Resolution



Workload and Offered-Load

- **Workload**: Stochastic process, representing the amount of work present at time t, under the assumptions of **infinitely many resources** (service commences immediately upon arrival).
- **Offered-Load**: Function of time $t \geq 0$, representing the average of the workload at time t.

The Offered-Load, $R(t)$, determines staffing level via c-staffing ($c = 0.5$ is conventional square-root staffing):

$$N(t) = R(t) + \beta \cdot [R(t)]^c$$

Offered-Load Representations (or Time-Varying Little)

For the $M_t/GI/N_t + GI$ queue, the **offered-load** $R = \{R(t),\, t \geq 0\}$, has the following representations:

$$R(t) = E[L(t)] = \int_{-\infty}^{t} \lambda(u) \cdot P(S > t-u)du = E\Big[A(t) - A(t-S)\Big] =$$

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

where
$A = \{A(t),\, t \geq 0\}$ is the Arrival process;
S is a generic service time;
S_e is a generic excess (residual) service.

In stationary models, where $\lambda(t) \equiv \lambda$, the offered-load $R(t)$ is the familiar $\lambda \cdot E[S]$ (or λ/μ), measured in Erlangs.

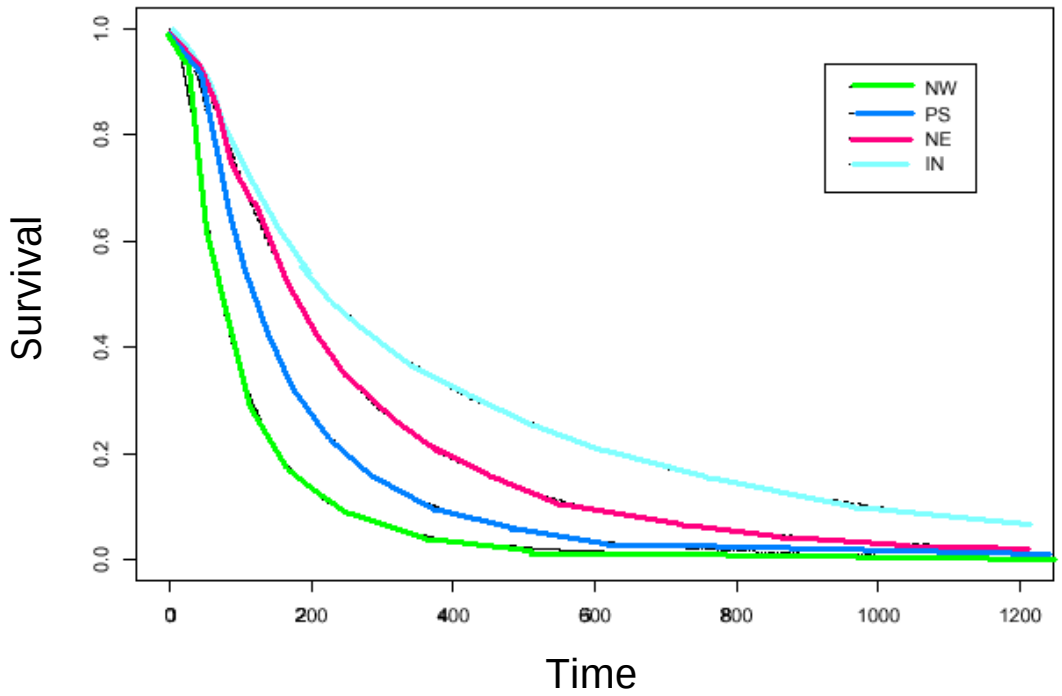
Imputing Service Times of Abandoning Customers

In calculating the offered-load, one must account for service-times of abandoning customers.

A prevalent assumptions is that service times and (im)patience times are independent. Experience suggests that this assumption is often violated.
For example, it is not unreasonable that customers who anticipate longer service times, will be willing to wait more for service before abandoning.

Service Times: Stochastic Order

Small Israeli Bank: Survival Functions by Type



Service times stochastic order: $S_{NW} \overset{st}{<} S_{PS} \overset{st}{<} S_{NE} \overset{st}{<} S_{IN}$

Patience times stochastic order: $\tau_{NW} \overset{st}{<} \tau_{IN} \overset{st}{<} \tau_{PS} \overset{st}{<} \tau_{NE}$

Relationship Between Service-Time and (Im)Patience

Ongoing research (w/ M. Reich, Y. Ritov) develops a procedure for calculating the function $E(S|\tau = w)$:

1. Introduce $g(w) = E(S|\tau > W = w)$, which is the mean service time of those who waited exactly w units of time **and were served**. Then calculate g via the non-linear regression:

$$S_i = g(W_i) + \varepsilon_i \ ,$$

   where i indexing served customers.

2. Calculate $E(S|\tau = w)$ via the (established) relation

$$E(S|\tau = w) = g(w) - \frac{g'(w)}{h_\tau(w)} \ ,$$

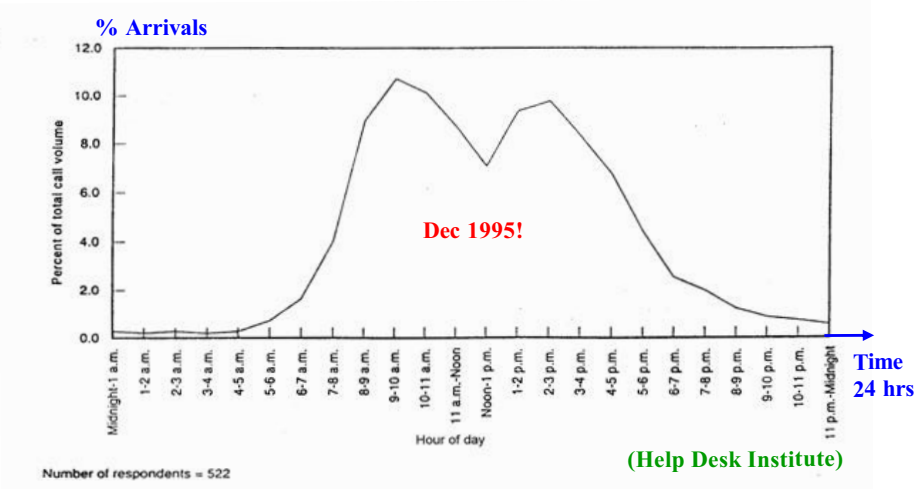
   where $h_\tau(w)$ is the hazard-rate function of (im)patience, to be estimated via un-censoring.

Finally, extend the above to calculate the distribution of S, given w, which is then used to impute service-times for calculating the offered-load.

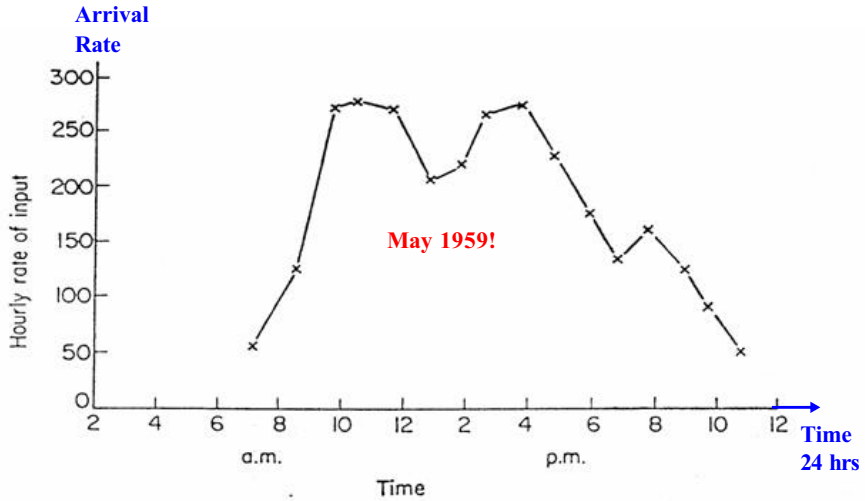
Daily Arrivals to Service: Time-Inhomogeneous (Poisson?)

Intraday Arrival-Rates (per hour) to Call Centers

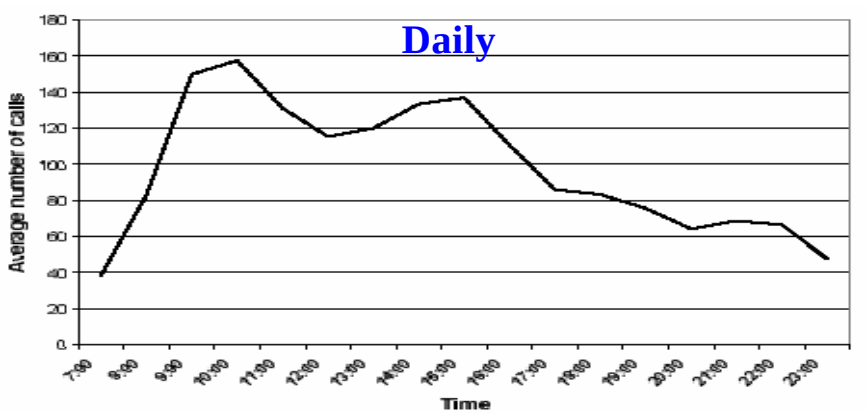
December 1995 (700 U.S. Helpdesks)



May 1959 (England)



November 1999 (Israel)

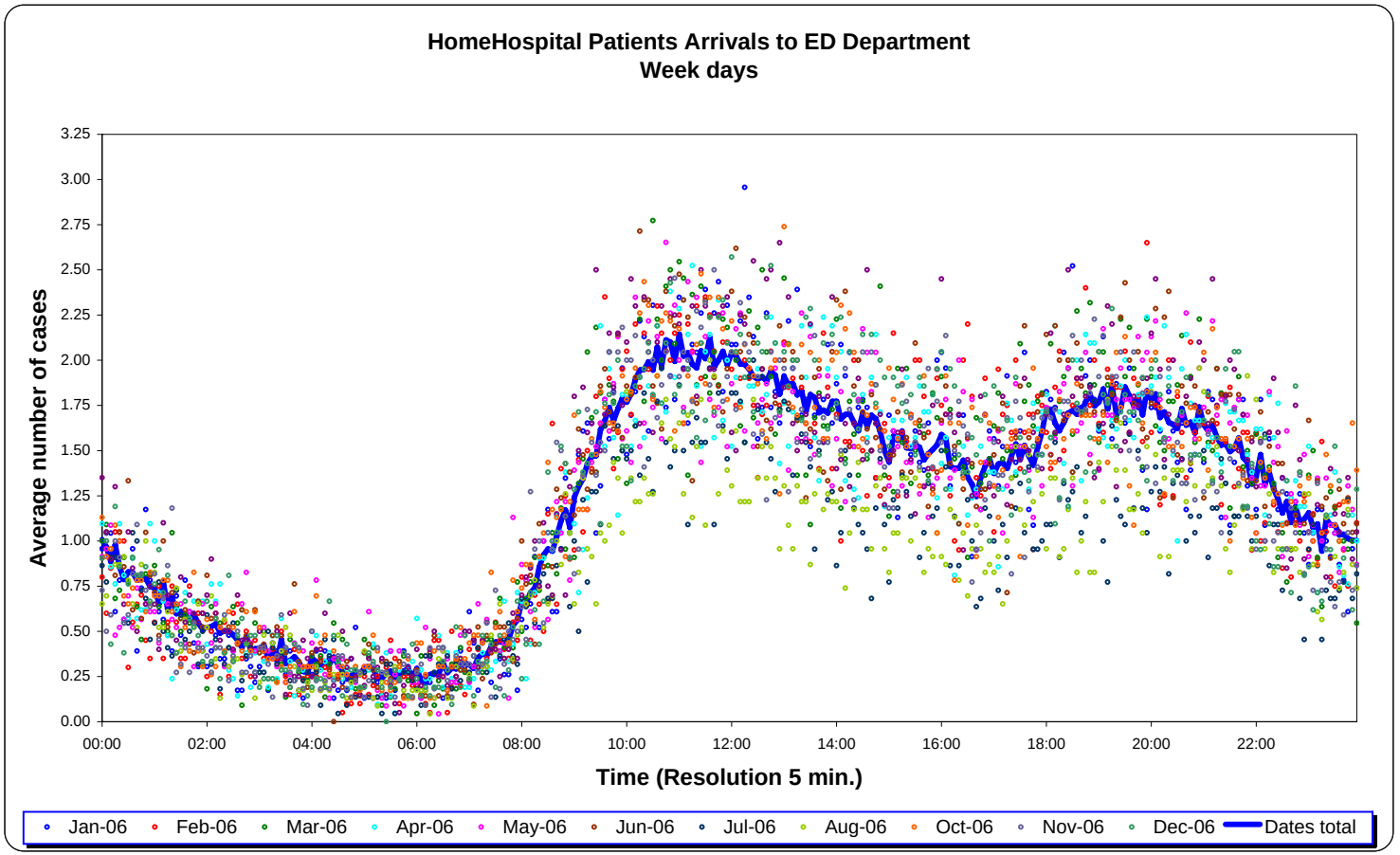


Observation:

Peak Loads at **10:00 & 15:00**

Arrivals to an Emergency Department (ED)

Large Israeli ED, 2006

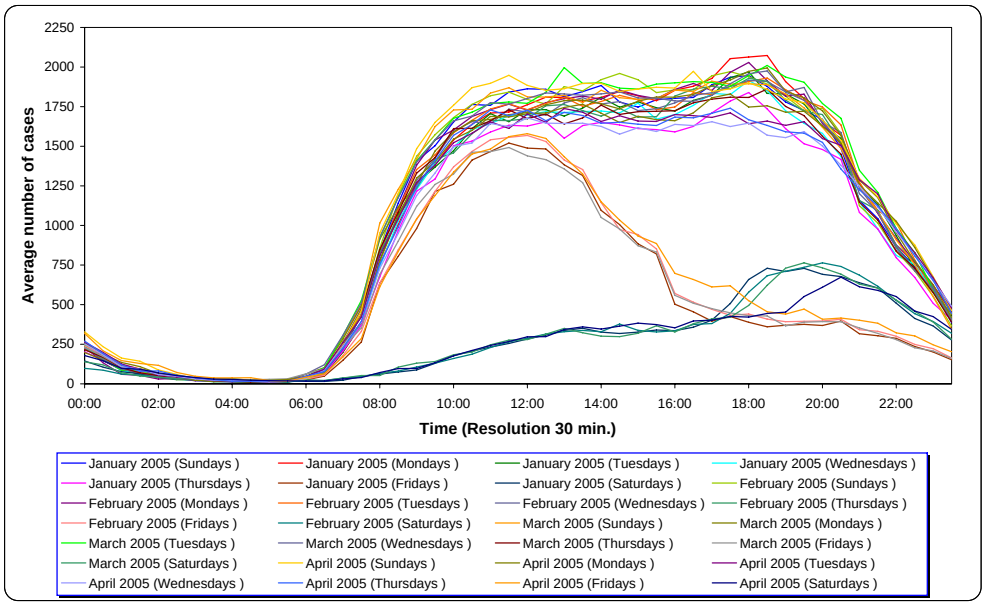


- **Second** peak at 19:00 (vs. 15:00 in call centers).
- How much stochastic variability ?

Intraday Arrival Rates: Does a Day have a Shape ?

Arrival Patterns, Israeli Telecom

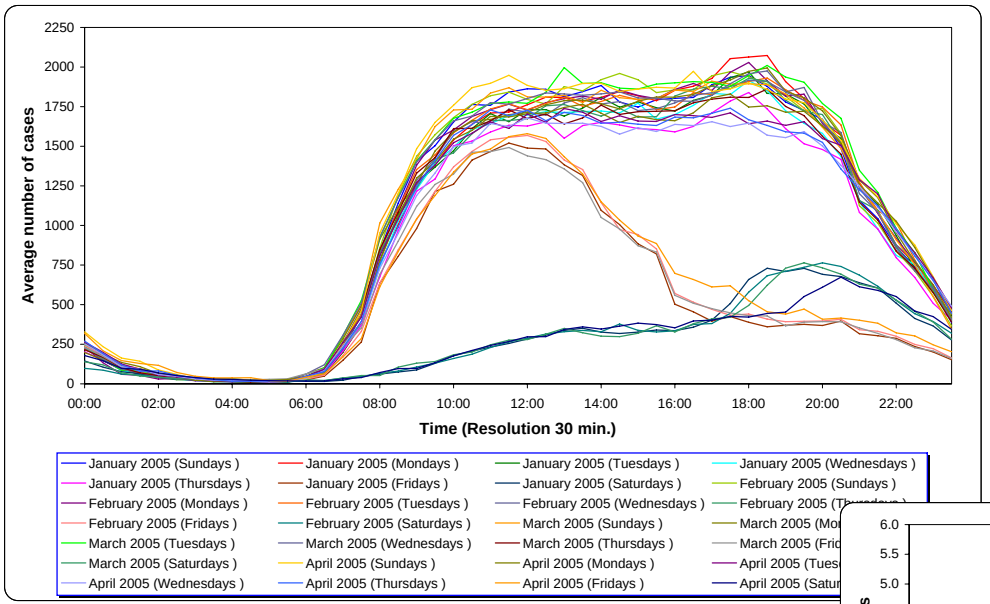
Arrivals, Avg. Weekdays/1-4/2005



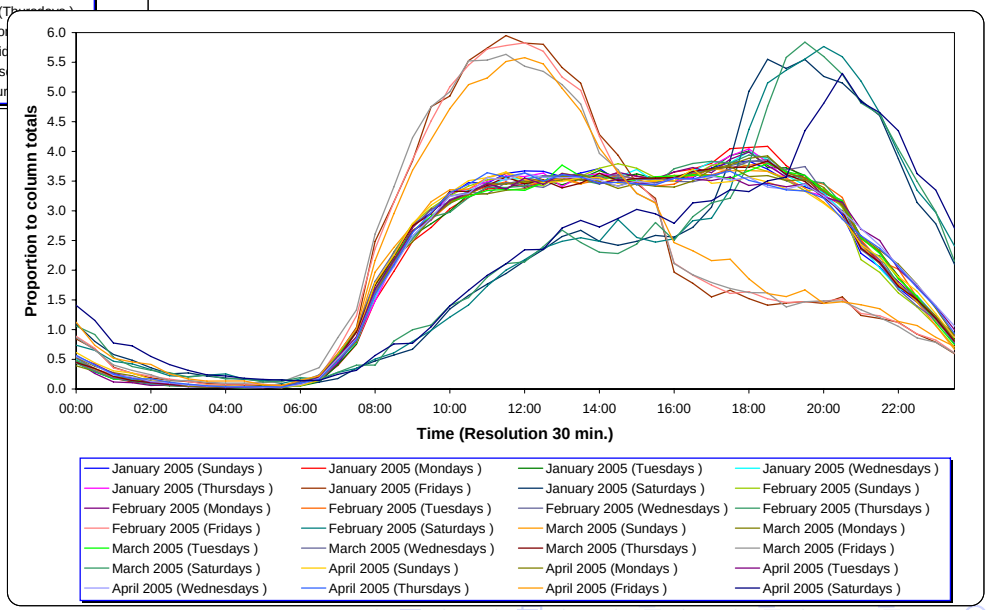
Intraday Arrival Rates: Does a Day have a Shape ?

Arrival Patterns, Israeli Telecom

Arrivals, Avg. Weekdays/1-4/2005



Percent to Total, 30 min.



A (Common) Model for Call Arrivals

Whitt (99'), Brown et. al. (05'), Gans et. al. (09'), and others:

Doubly-stochastic (Cox, Mixed) Poisson with instantaneous rate

$$\mathbf{\Lambda(t) = \lambda(t) \cdot X} \ ,$$

where $\int_0^T \lambda(t)dt = 1$.

- $\boldsymbol{\lambda(t)}$ = "Shape" of weekday [**Predictable variability**]
- **X** = Total # arrivals [**Unpredictable variability**]

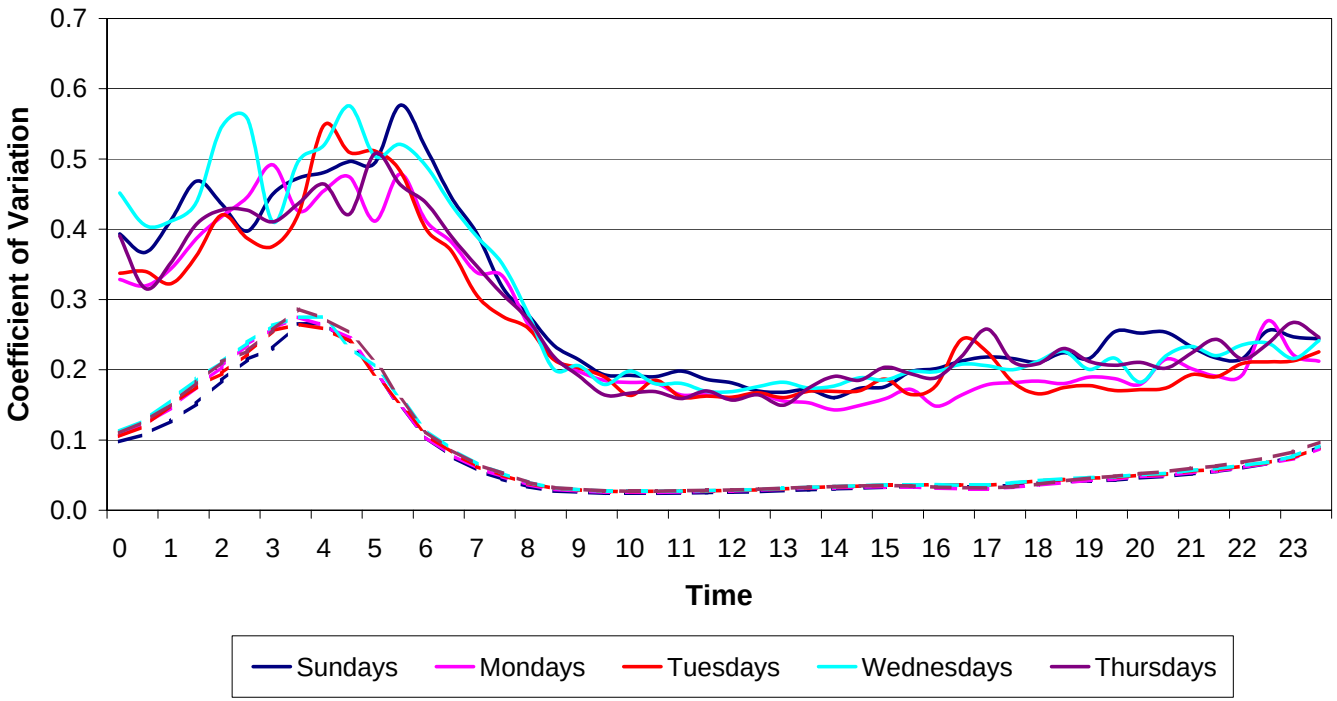
w/ Maman & Zeltyn (09'):
Above assumes **"too-much" stochastic variability!**

Over-Dispersion (Relative to Poisson), Maman et al. ('09)

Israeli-Bank Call-Center

Arrival Counts - Coefficient of Variation (CV), per 30 min.

Sampled CV - solid line, **Poisson** CV - dashed line

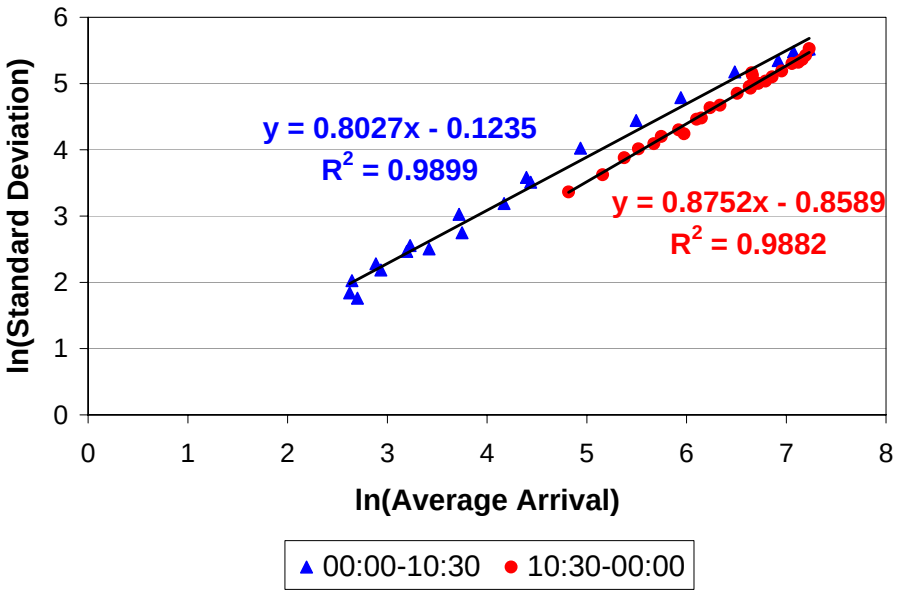


- 263 regular **days**, 4/2007 - 3/2008.
- Poisson CV $= 1/\sqrt{\text{mean arrival-rate}}$.
- Sampled CV's $\gg$ Poisson CV's $\Rightarrow$ **Over-Dispersion**.

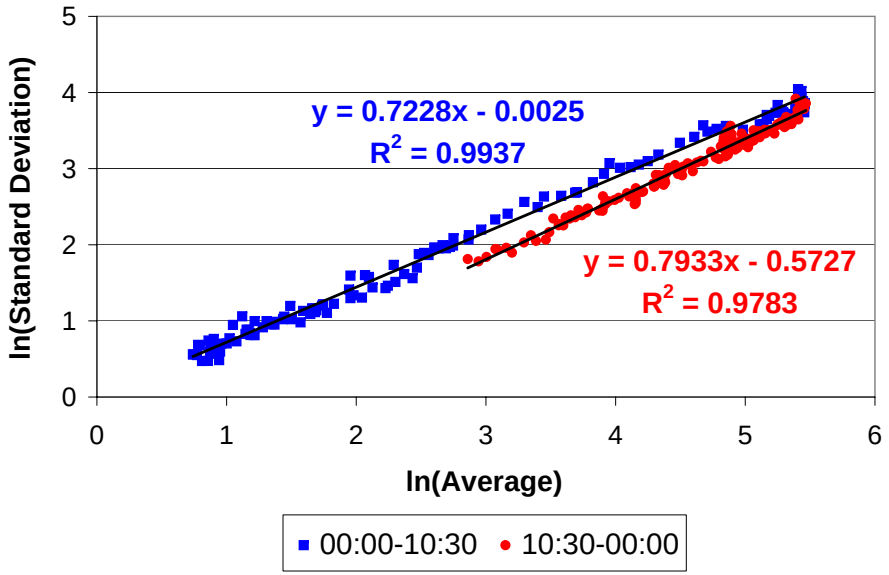
Over-Dispersion: Fitting a Regression Model

ln(STD) vs. ln(AVG)

Tue-Wed, 30 min resolution



Tue-Wed, 5 min resolution



Significant linear relations (Aldor & Feigin):

$$\ln(\mathrm{STD}) = c \cdot \ln(\mathrm{AVG}) + a$$

Over-Dispersion: Random Arrival-Rate Model

The **linear relation** between ln(STD) and ln(AVG) motivates the following model:

Arrivals distributed **Poisson with a Random Rate**

$$\Lambda = \lambda + \lambda^{c} \cdot X, \quad 0 \leq c \leq 1 ;$$

- X is a random-variable with $E[X] = 0$, capturing the magnitude of **stochastic deviation** from mean arrival-rate.
- c determines **scale-order** of the over-dispersion:
  $c = 1$, proportional to λ;
  $c = 0$, Poisson-level, same as $0 \leq c \leq 1/2$.

In **call centers**, over-dispersion (per 30 min.) is of order λ^{c}, $c \approx 0.8 - 0.85$.

Over-Dispersion: Distribution of X ?

- Fitting a **Gamma Poisson** mixture model to the data:
  Assume a (conjugate) prior Gamma distribution for the arrival rate $\Lambda \stackrel{d}{=} Gamma(a, b)$.
  Then, $Y \stackrel{d}{=} Poiss(\Lambda)$ is Negative Binomial.

- Very good fit of the Gamma Poisson mixture model, to data of the Israeli Call Center, for the majority of time intervals .

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

- Distribution of X derived, under the Gamma prior assumption: X **is asymptotically normal**, as $\lambda \to \infty$.

Over-Dispersion: The QED-c Regime

QED-c Staffing: Under offered-load $\mathbf{R = \lambda \cdot E[S]}$,

$$\mathbf{n \;=\; R \;+\; \beta \cdot R^{c}\,, \quad 0.5 < c < 1}$$

Performance measures:

a. Delay probability: $P\{W_q > 0\} \;\sim\; 1 - F(\beta)$

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

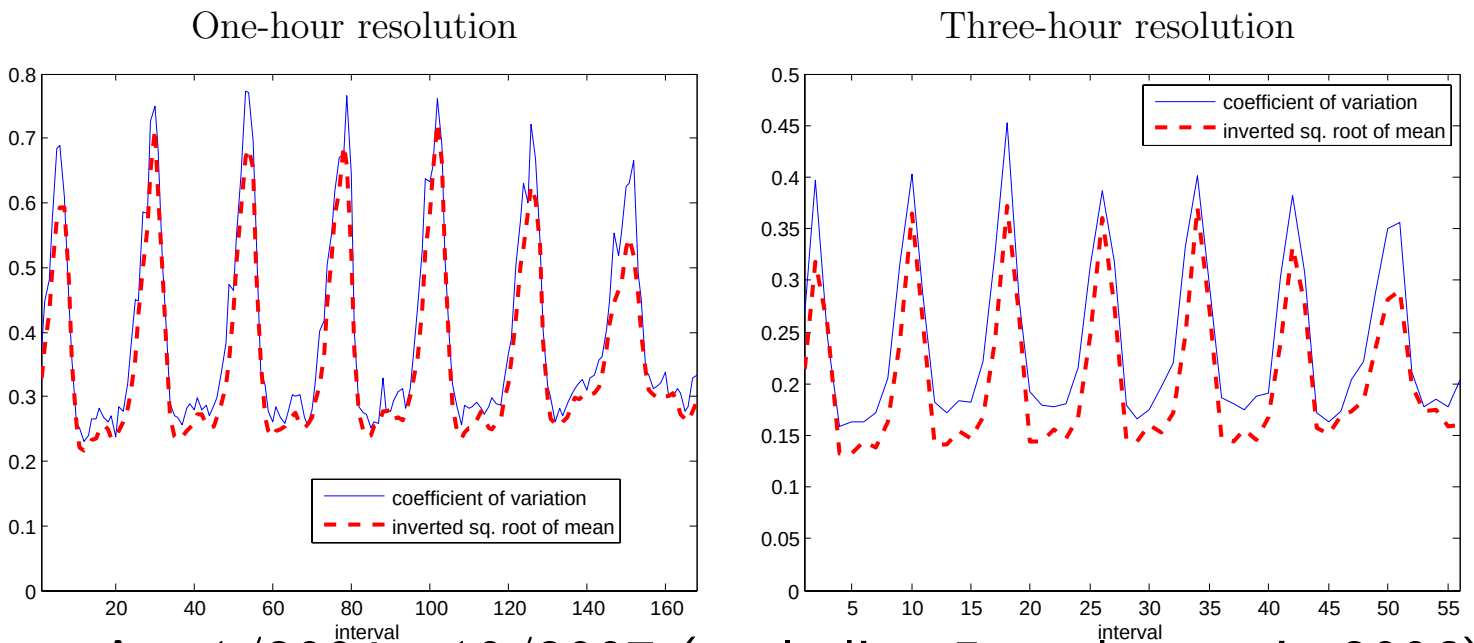
c. Average offered wait: $E[V] \;\sim\; \dfrac{E[X-\beta]_+}{n^{1-c} \cdot g_0}$

d. Average actual wait: $E_{\Lambda,n}[W] \;\sim\; E_{\Lambda,n}[V]$

Over-Dispersion: The Case of ED's

Israeli-Hospital Emergency-Department

Arrival Counts - Coefficient of Variation, per 1-hr. & 3-hr.



- 194 **weeks**, 1/2004 - 10/2007 (excluding 5 weeks war in 2006).
- Moderate over-dispersion: **c = 0.5** reasonable for hourly resolution.
- **ED beds in conventional QED** (**Less** var. than call centers ! ?).