Class 7

~~Arrivals: Forecasting, and some loose ends~~
Service Times; Phase-type Distributions

Arrivals: Review

- Poisson processes: review;
- Forecasting arrivals;
- The Offered Load.

Defining, Modelling and Designing Service Times

- What is "Service-Time"? via Empirical analysis of face-to-face, telephone services; hospitals, ...
- Service time is a Statistical Distribution: lognormal, exponential.
- Service time is a Process: Phase-type distributions.
- Beyond Means and Beyond CV's.
- Stochastic Ordering.
- Subtleties.

Laws of Congestion: Old and New
The 0-th Law for (The) *Causes of Operational* Queues :
Scarce Resources and Synchronization Gaps (in DS-Project Networks);
The First Law of *Conservation* :
Little's Law for Customers, Service-providers and Managers.
Little's Law for the Offered Load (Utilization Profiles).
The Second Law of Completely *Random Arrivals* :
Levy/Watanabe Axioms of Randomness;
The Law of Poisson-Counting (Law of Rare Events);
The Law of Independent Memoryless (Exponential) Inter-arrivals;
The Brownian-Law of Rescaling & Centering Arrivals;
The Laws of Decomposition-Superposition.
The Third Law of *Human Service-durations* :
The Law of Phase-types for the Durations of Human Upaced Services;
The Empirical Law of Exponential/Log-Normal Durations.
The Fourth Law of *Sampling* :
Random Sampling: Wolff's PASTA = Poisson Arrivals See Time Averages;
Biased Sampling: Costs of Randomness; (Coefficient of Variation, or Form Factor).

Recitation 7. Statistical analysis of an arrival process.

0

Service Engineering

Class 7

Service Times (Durations, Processes)

Why Significant? eg. +1 second of 1000 agents costs $500K yearly.

Why Interesting? Must accurately
Model, Estimate, Predict, Analyze, Design:

- Resolution: Sec's (phone)? min's (email)? hr's (hospital)
- Parameter, Distribution (Static) or Process (Dynamic)?
- Does it include after-call work?
- Does it include interruptions?
  - Whisper time, hold time, phones during face-to-face,...
- Does is account for return services?

How affected by covariates? How affects performance?

- Experience and Skill of agents (Learning Curve)
- Type of Customer: Service Type, VIP Status
- Time-of-Day: Congestion-Level
- Human Factor: Incentives, pending workload, fatigue
- Heavy-Traffic: The ED and QED Operational Regimes (later)

How to calculate **Offered-Load**? (towards Staffing)

Contents: **Service Times; Phase-Type Durations.**

- Service duration = **Statistical Distribution**:
  - **Empirical**: Histogram, CDF, Hazard Rates (Later);
  - **Parametric**: **LogNormal, Exponential**, Others.
- Empirical Introduction, mainly via DataMOCCA.
- Motivating Examples.
  - Designing an IVR/VRU.
  - Pooling a Service Network.
  - Long-term Care of the Elderly.
- Sample size.
- **What is Service Time (Duration)?**
  A complex answer to a "simple" question:
  - Single vs. multiple visits.
  - After-Call Work (ACW); Utilization Profiles.
  - Time- vs. State-dependency.
  - Incentives (Call Center, Hospital)
  - Averages do not tell the whole story: the need for **Distributions**.
- Stochastic Ordering (of distributions).
- Service = **Stochastic Process**: **Phase-type** MJP.
- "Sufficient Statistics" in Heavy Traffic: ED, QED (later)
- Offered-Load (Work)

1

Parametric Distribution of Service Times

Most common parametric distributions in service systems are Exponential and Lognormal.

Exponential Distribution:

Density: $f(x) = \lambda e^{-\lambda x}, \ x \geq 0,$

Mean: $\mathrm{E}[X] = \lambda^{-1},$

Variance: $Var(X) = \lambda^{-2},$

Coefficient of Variance: $C_v = \frac{SDV(X)}{\mathrm{E}[X]} = 1,$

Median: $\lambda^{-1} ln2.$

An important property of the exponential distribution is that it is memoryless. This means that if a random variable T is exponentially distributed, its conditional probability obeys
$\Pr(T > s + t \mid T > s) = \Pr(T > t)$ for all $s, t \geq 0$.

Lognormal Distribution:

Definition: X is a lognormal random variable if ln(X) is normally distributed with mean μ and variance σ^2.

Density: $f(x) = \frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{(lnx-\mu)^2}{2\sigma^2}}, \ x \geq 0,$

Mean: $\mathrm{E}[X] = e^{\mu + \sigma^2/2},$

Variance: $Var(X) = e^{\mu + \sigma^2/2}(e^{\sigma^2} - 1),$

Coefficient of Variance: $C_v = \sqrt{e^{\sigma^2} - 1},$

Note that CV does not depend on μ. For small σ ($\sigma < 0.5$),one can use $CV \approx \sigma$.

Median: e^{μ}.

Local Municipalities Service Time

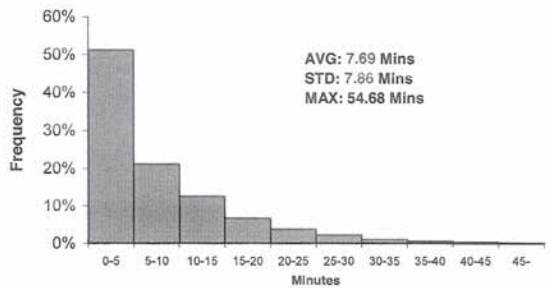
Department	Station No.	Total Customers	Avg. Arrival Rate (1/Hr)	Avg. Service Time (Mins)	STD (Mins)	Maximal Service Time (Mins)	Utilization	Avg. Waiting Time (Mins)
Water	N/A	187	1.8 ± 0.2	8.87 ± 1.0	8.15	54.68	13.3%	4.76
Tellers	N/A	1328	12.6 ± 0.5	8.82 ± 0.4	8.55	49.37	30.8%	7.73
Cashier	N/A	757	7.2 ± 0.4	6.64 ± 0.4	6.94	29.95	79.7%	3.89
Manager	N/A	190	1.8 ± 0.2	7.99 ± 1.0	8.44	38.97	24.1%	9.16
Discounts	N/A	317	3.0 ± 0.3	4.59 ± 0.4	4.54	36.72	23.1%	3.65

Water	1	57	N/A	7.80 ± 1.70	7.61	31.28	6.5%	N/A
	2	130	N/A	9.34 ± 1.20	8.37	54.68	19.3%	N/A
Tellers	3	336	N/A	9.04 ± 0.80	8.93	49.05	48.2%	N/A
	4	208	N/A	9.93 ± 1.00	8.82	49.12	33.0%	N/A
	5	417	N/A	8.97 ± 0.70	8.55	49.37	59.4%	N/A
	6	144	N/A	9.53 ± 1.20	8.75	41.70	21.8%	N/A
	7	156	N/A	8.03 ± 1.10	7.96	35.27	19.8%	N/A
	8	67	N/A	3.74 ± 0.70	3.58	21.03	4.0%	N/A
Cashier	9	757	N/A	6.64 ± 0.40	6.94	29.95	79.7%	N/A
Manager	10	190	N/A	1.99 ± 1.00	8.44	38.97	24.1%	N/A
Discounts	11	317	N/A	4.59 ± 0.40	4.54	36.72	23.1%	N/A

*Service time ranges given with 90% confidence.

Service Time Histogram – Overall:

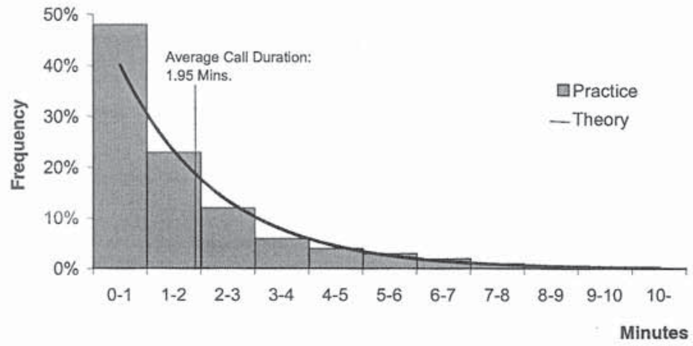
Range	Frequency
0-5	51.3
5-10	21.1
10-15	12.6
15-20	6.7
20-25	3.8
25-30	2.3
30-35	1.1
35-40	0.6
40-45	0.3
45-	0.2



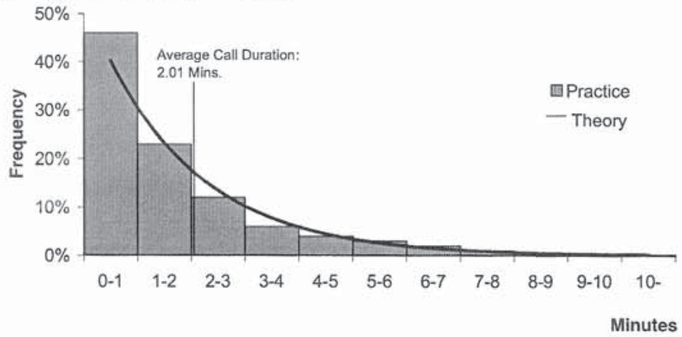
2

Service Times: Exponential (Phone Calls)

Call-Duration Frequency - North:



Call-Duration Frequency – Central:



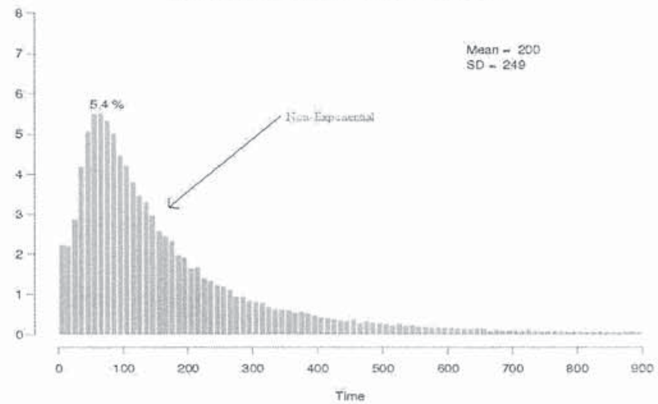
Q: How to recognize "Exponential" when you "see" one?

A: Geometric Approximation

LogNormal Distribution

Empirically prevalent in call centers (overall, service types, individual agents), but yet **no** theoretical explanation.

Israeli Bank. Nov-Dec.



Good in **statistical** models
(eg. regression of log(service-time)).

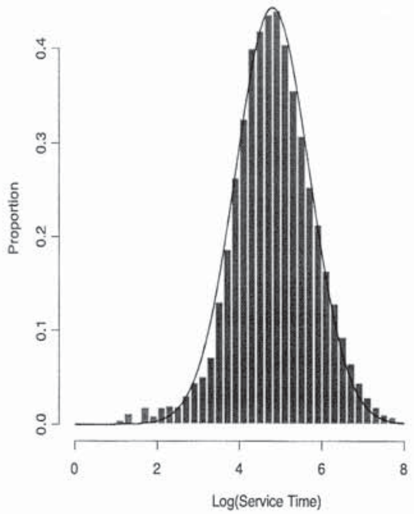
Not so good for **queueing** models
(which typically "prefer" Exponential durations).

The practical good news for service time distribution in **queueing** models
CV is more important, if tail of distribution is similar one can use exponential distribution.

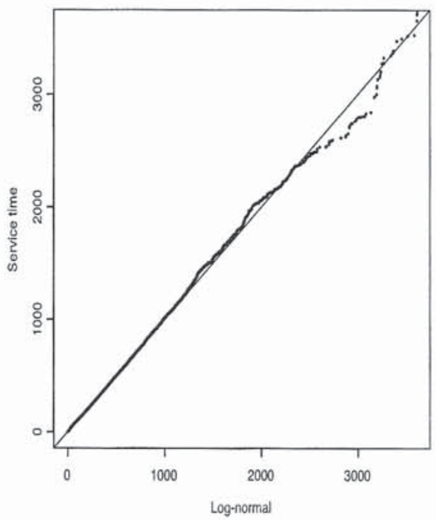
3

Validating LogNormality of Service Times

Log(Service Times).



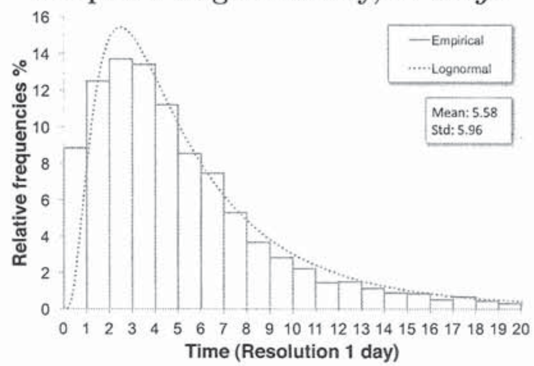
QQ Plot



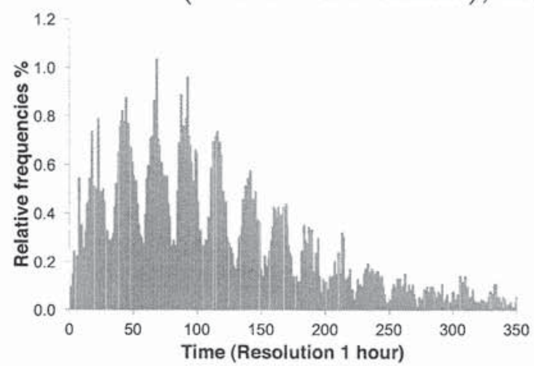
QQ Plots will be reviewed at the Recitation

Beyond Standard Distributions: Service Times at Hospital

Length-of-Stay (LOS) at the Internal Wards of a Hospital: LogNormality, in *Days*



Length-of-Stay (LOS) at the Internal Wards of a Hospital: Mixture (of Skewed-Normals), in *Hours*



4

Service Times: Mixture Model

A mixture model represents the presence of sub-populations within an overall population.

Finite mixture: Given a finite set of probability density functions $f_1(x), \dots, f_n(x)$, and weights $w_1, \dots, w_n$ such that $w_i \geq 0$ and $\Sigma_{i=1}^{n} w_i = 1$, the mixture distribution can be represented by writing the density, f, as a sum (which is a convex combination):

$$f(x) = \sum_{i=1}^{n} w_i\, f_i(x).$$

Moments:

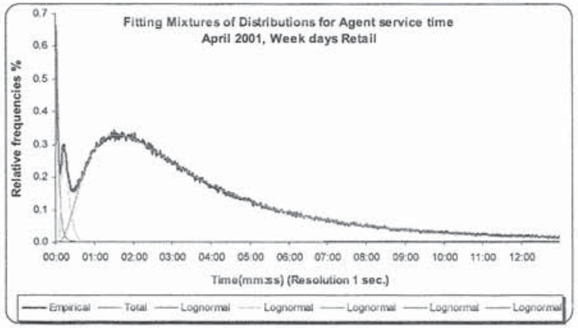
$$\mathrm{E}[X] = \mu = \sum_{i=1}^{n} w_i \mu_i,$$

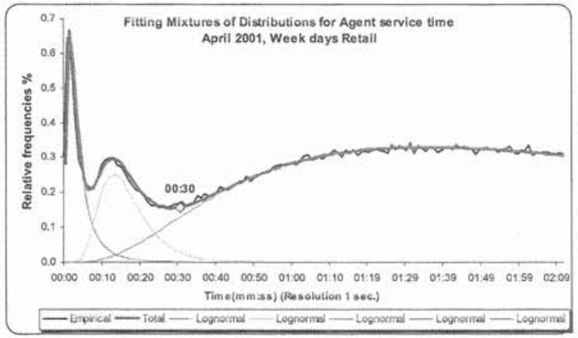
$$\mathrm{E}[(X-\mu)^2] = \sigma^2 = \sum_{i=1}^{n} w_i(\mu_i^2 + \sigma_i^2) - \mu^2.$$

Example: Service time distribution in a call center, Length-of-stay in Maternity Wards.

Service Times: Mixture

LOS in Call Center as a mixture of LogNormals





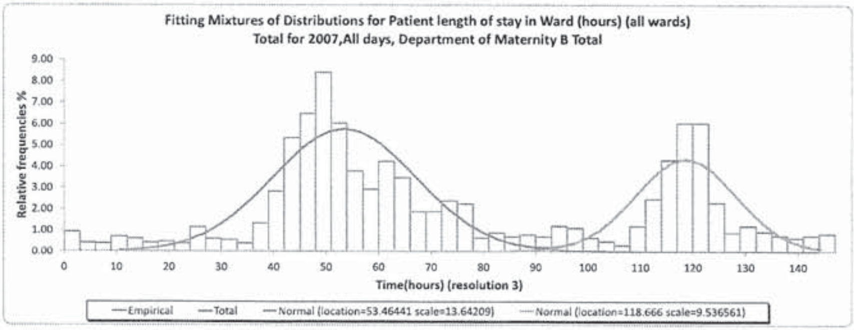
Going to the previous Excel Sheet, to view the corresponding Tables (by scrolling it down), one notes that the main component has weight 91% in the mixture – its role in the chart is to fit the part beyond 30 seconds, which it does very well.

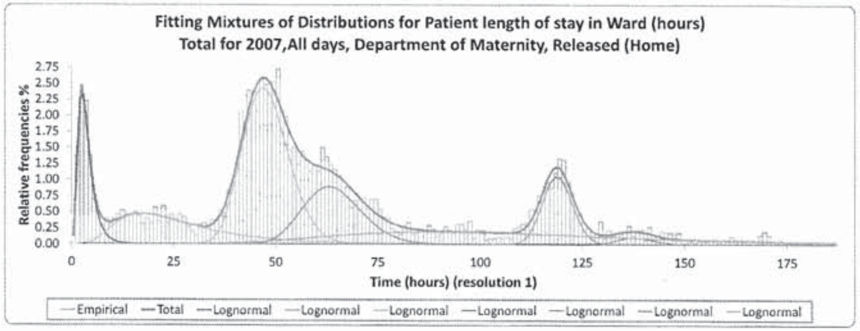
Parameter Estimates	
Components	Mixing Proportions (%)
1. Lognormal	3.08
2. Lognormal	3.58
3. Lognormal	91.14
4. Lognormal	1.86
5. Lognormal	0.34

5

Service Times: Mixture

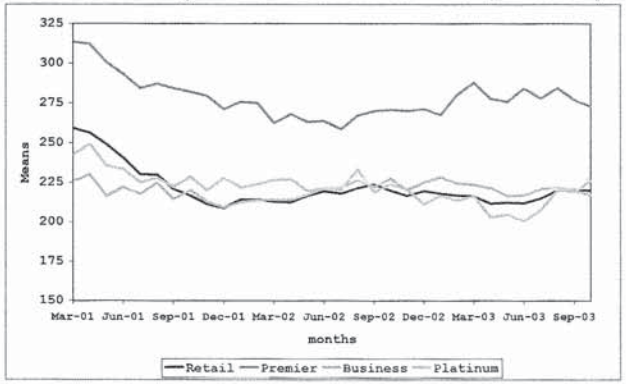
A Mixture Distribution for LOS in Maternity Wards





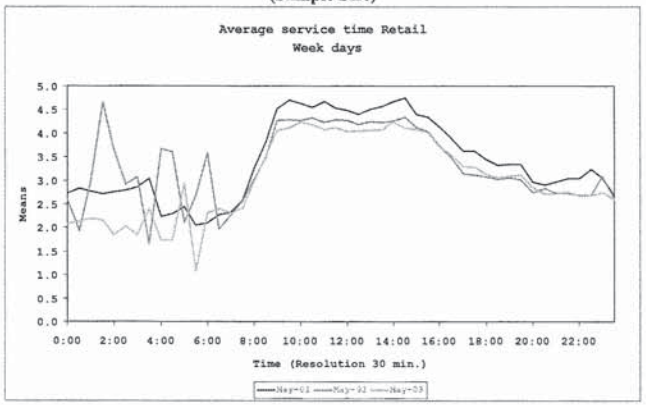
Service Times: Trends and Stability

USBank Average Customer Service Time, Weekdays



USBank Average Customer Service Time, Telesales

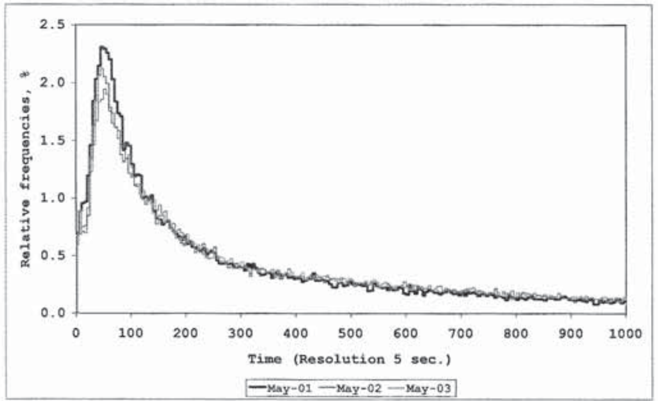
US Bank: Dynamics of average customer service time for Retail calls (Sample Size)

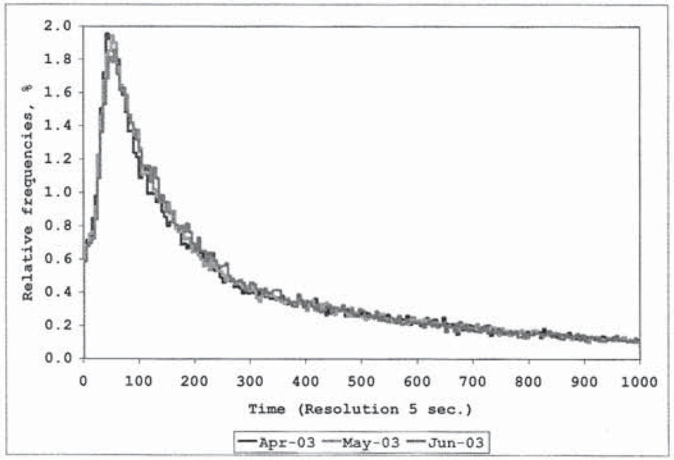


6

Service Times: 5 Sec's Resolution

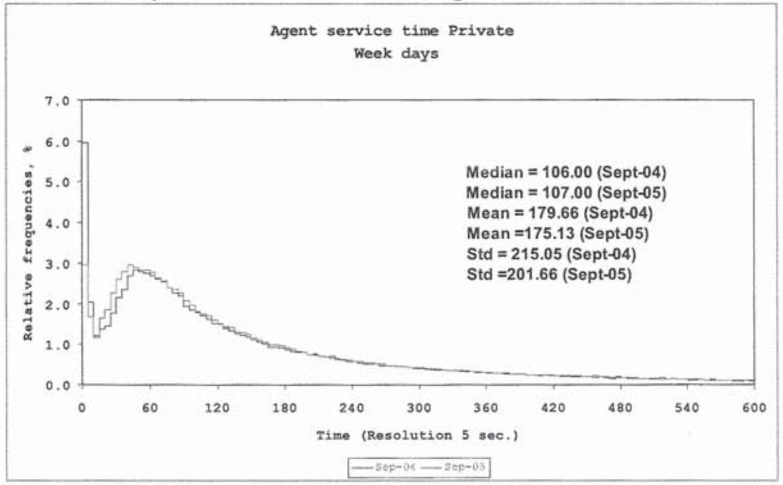
USBank. Service-Time Histograms for Telesales (MOCCA)





Service Times in Israeli Telecom

IL Telecom: Dynamics of the distribution of agent service time for Private calls



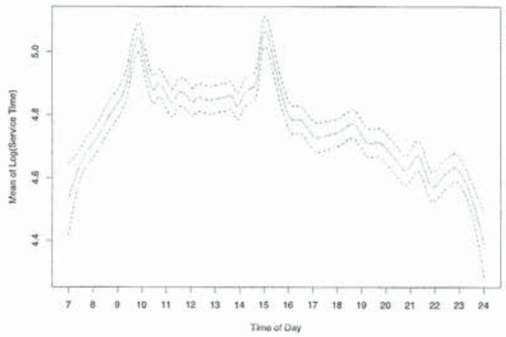
- Overall pattern seems close to LogNormal (except for the very short service times);
- Histograms of different months are **very** similar;
- Reason for short service durations unknown here.

7

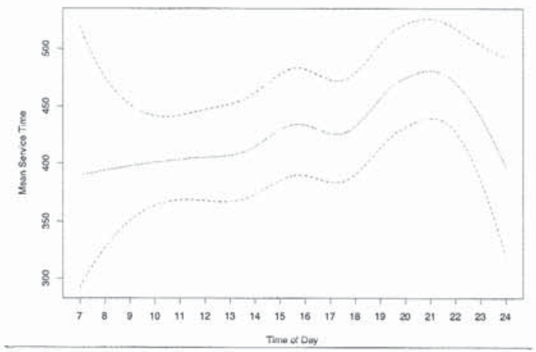
Service Times: The Human Factor, or Why Longest During Peak Loads?

Mean Service Time vs. Time-of-Day

Regular Service (PS)

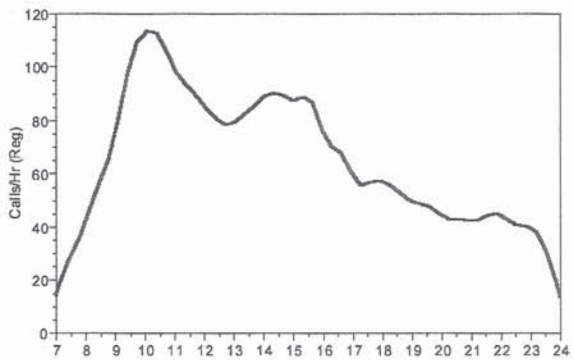


Internet Consulting (IN)



Service Times vs. Arrival Rates

Regular Service (PS): Arrival Rate



At **10:00 & 15:00**: longest services and peak arrival rates?

Possible Reasons:

1. *Services are longer* during congestion since customers start with complaints.)
2. *Agents are slower* at times of peak loads.
3. Customers that arrive during peak hours require, for some reason, *longer service*.
4. An additional (*human*) reason will be provided after we study *customers' impatience*.

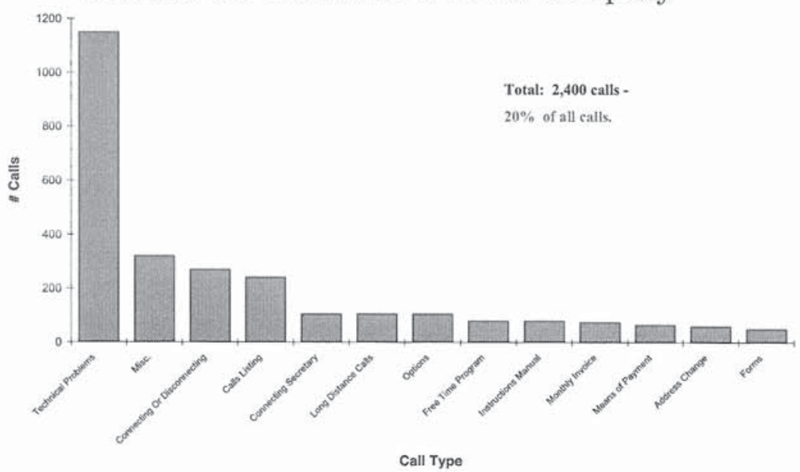
8

Service Time $\neq$ Contact-Time

Common (Often Too Common):

- Customers routed for additional services (vs. "First-Time-Resolution");
- Servers interrupt face-to-face service with a phone-call (vs. the increasingly prevalent "Central Call Center");
- Agents place customers on hold, eg. technical consultation with veterans;
- Agents can be engaged in non-phone activities, eg. ACW Time (After-Call Work).

Reasons for Redials in a Cable Company



Calculating (Mean) Service Time

First approach:
Sum up components of the "service time", then add related activities of servers.

Second approach (Avoids Ambiguities):
Fix a time interval (eg. a shift).

$$\text{Mean Service Time} = \frac{\text{Available Time - } \textbf{Idle Time}}{\text{Number of Calls}},$$

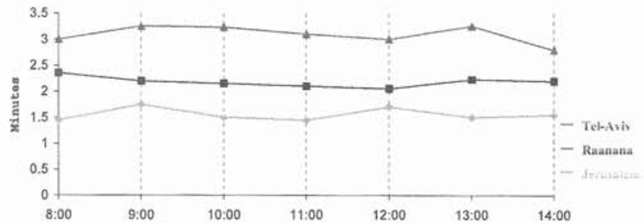
where
Available Time = # Agents × Interval Duration,
and
Idle Time is summed over all agents.

9

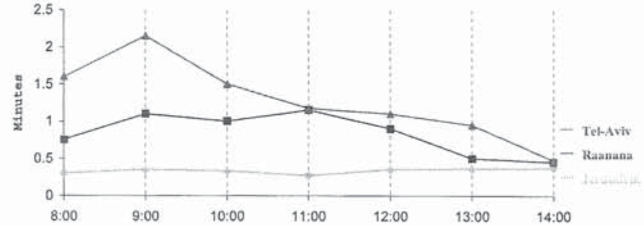
Israel Electric Company: 3 Centers

Service Performance

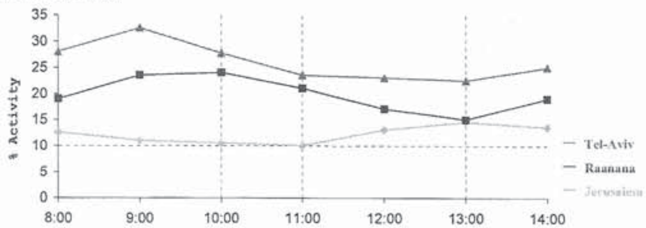
Service Time – Average:



Waiting Time – Average:



% Abandonment:

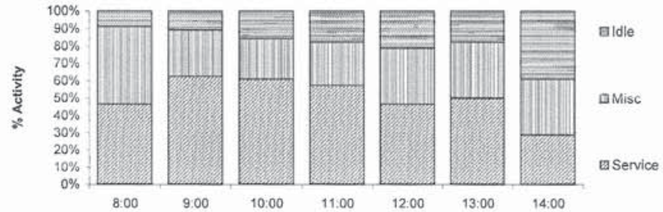


Israel Electric Company: 3 Centers

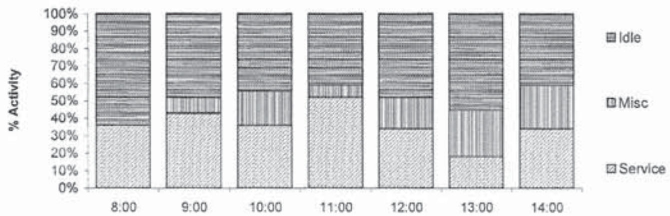
What is "Service Time" ?

Utilization Profile in 3 Call Centers Doing the Same Thing

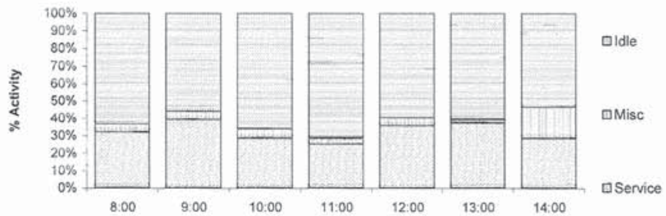
Tel-Aviv:



Raanana:

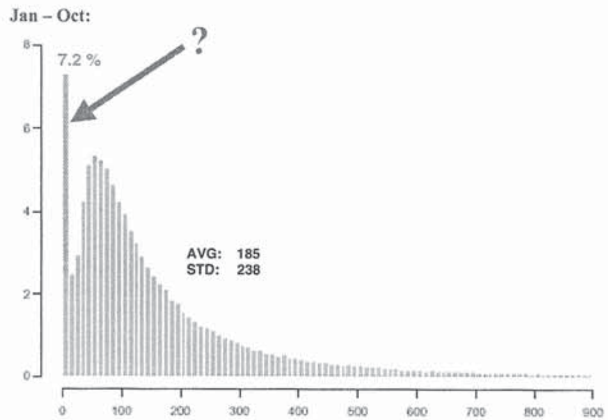


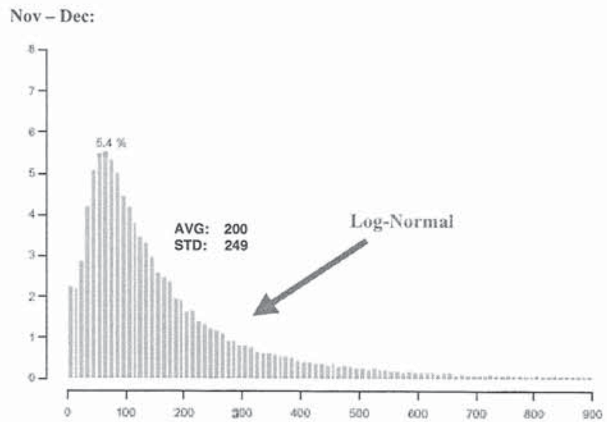
Jerusalem:



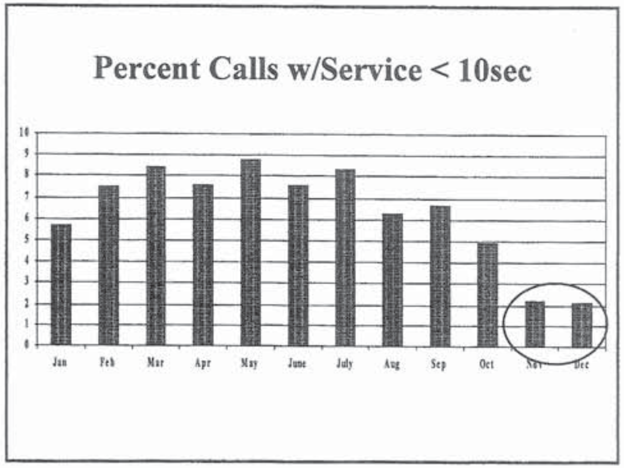
10

Beyond Data Averages
Short Service Times





Short Service Times: Time Series



Short Service Times: Individuals

Table 52: Number of calls handled by an agent

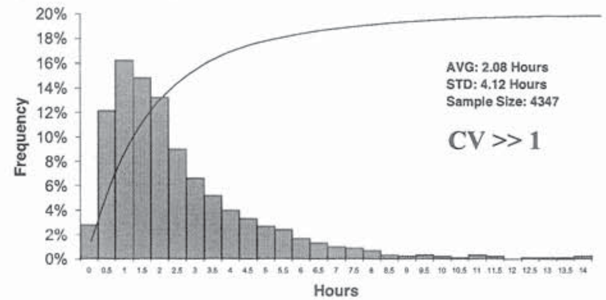
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
AVI	0	0	0	1117	2208	2019	2789	2710	1417	2026	2523	2395
AVNI	1493	1736	642	539	1786	2219	2092	2392	1156	1888	1988	2136
BASCH	999	1164	1708	1155	982	906	858	2185	1973	1055	1326	1242
BENSION	1283	1135	0	1053	1108	1016	1682	1298	1076	1303	1546	1176
DARMON	309	515	633	519	577	436	309	370	297	194	425	128
DORIT	696	1047	0	811	546	862	750	2228	1319	1384	1640	1605
ELI	387	508	777	447	560	436	395	458	416	363	502	352
GELBER	333	143	510	427	859	281	386	332	67	179	165	269
GILI	668	614	1155	803	1108	974	418	0	355	456	412	298
KAZAV	1995	1693	1240	1451	1731	2251	1737	1168	729	1570	1047	2038
MEIR	0	0	0	0	0	0	127	344	318	280	406	454
MORIAH	1360	1223	1591	1351	1866	1980	2416	2152	1526	1940	1793	515
PINHAS	79	40	359	244	31	311	422	241	143	105	51	63
ROTH	0	0	397	1292	1928	1967	1831	1749	1625	1914	1458	1038
SHARON	1985	1674	2780	1938	2563	2657	2537	2875	1803	1935	2532	2140
STEREN	0	1043	2294	1516	2163	2231	1423	2455	1672	709	2375	2568
TOVA	1923	1679	1562	1059	1464	1389	1890	1811	1361	1971	941	0
VICKY	895	0	0	0	1006	1378	1415	1674	1472	1582	1641	1990
YIFAT	1312	1901	1745	1305	1464	1076	780	90	1137	1315	0	0
YITZ	1771	1791	1402	1203	1355	1367	1009	69	705	1743	2420	2353
ZOHARI	891	1144	1398	1148	1479	1450	980	1494	1423	1359	1504	1094
Z2ARIE	0	0	0	0	0	0	0	56	225	315	432	534
Z2ELINOR	0	0	0	0	0	0	0	45	352	288	222	310
Z2EYAL	0	0	0	0	0	0	0	95	331	428	579	618
Z2IFAT	0	0	0	0	0	0	0	94	260	314	215	0
Z2LIOR	0	0	0	0	0	0	0	84	250	136	126	138
Z2NIRIT	0	0	0	0	0	0	0	116	327	474	387	545
Z2OFERZ	0	0	0	0	0	0	0	71	311	260	242	334
Z2SPIEGEL	0	0	0	0	0	0	0	71	311	260	153	322

Table 53: Number of calls with short service time

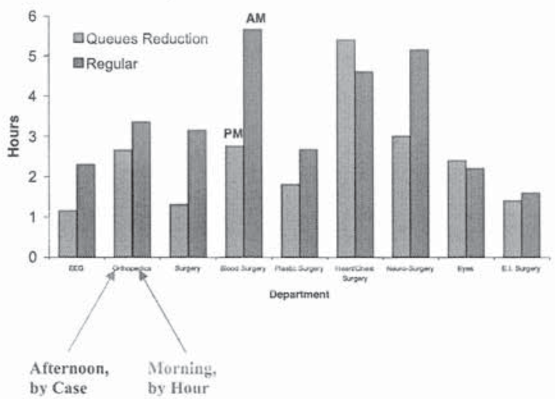
	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
MORIAH	233	230	356	290	614	695	865	597	490	455	4	1
AVI	0	0	0	47	111	144	295	221	121	76	35	26
AVNI	11	13	4	5	6	25	16	18	4	8	8	11
DARMON	2	11	8	9	10	7	1	0	1	1	0	0
ELI	9	7	10	12	22	18	15	4	8	3	6	5
KAZAV	57	40	48	44	48	63	40	27	15	18	4	6
MEIR	0	0	0	0	0	0	1	8	3	1	2	1
PINHAS	3	0	58	25	4	14	11	6	8	1	0	0
ROTH	0	0	10	10	36	21	43	25	32	31	3	6
SHARON	58	49	86	52	67	78	66	63	38	23	43	49
TOVA	52	163	269	132	231	193	100	109	207	190	6	0
ZOHARI	4	8	12	22	17	20	9	14	5	7	10	7

Service Times: The Human Factor, or Even "Doctors" Can Manage

Operations Time Histogram:



Operations Time - Morning vs. Afternoon:



Ethical?

Even Doctors Can Manage!

Human factors: Learning and Forgetting

Service Time Trends of Individual Servers

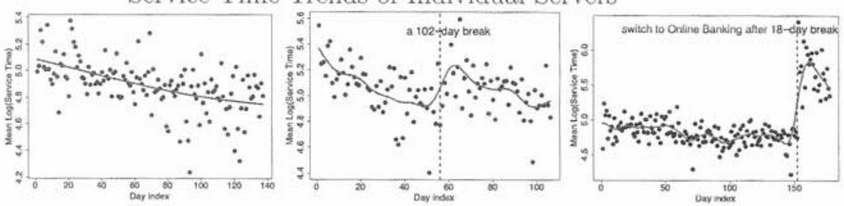


Figure 5: Long-term trend of daily average service time

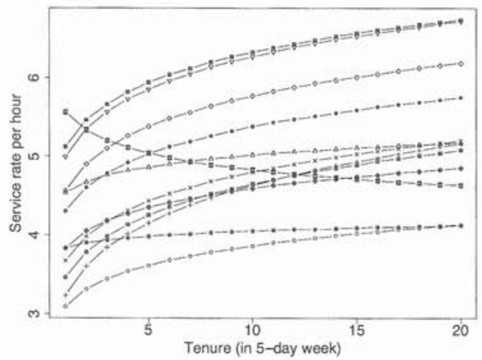


Figure 6: Daily learning curves of the 12 agents at site S

Classic learning models:

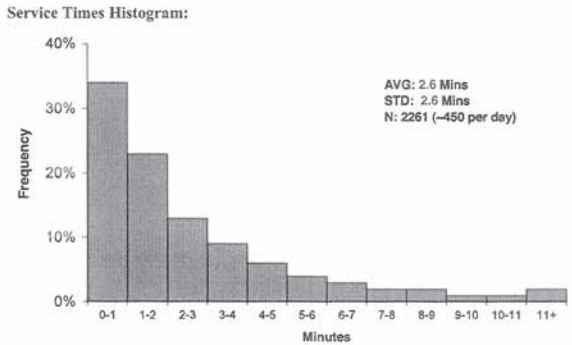
Assuming service times are lognormal distributed. y_{jk} - the service time of the kth call during the jth day. n_j - the total number of calls served by this agent during the jth day. Define $z_{jk} = log(y_{jk})$. Then, the basic learning model is:

$$z_{jk} = a + blog(j) + \epsilon_{jk},\ \ \epsilon_{jk} \sim N(0, \sigma_j^2)$$

Service Times: from Exponential to Phase-Type

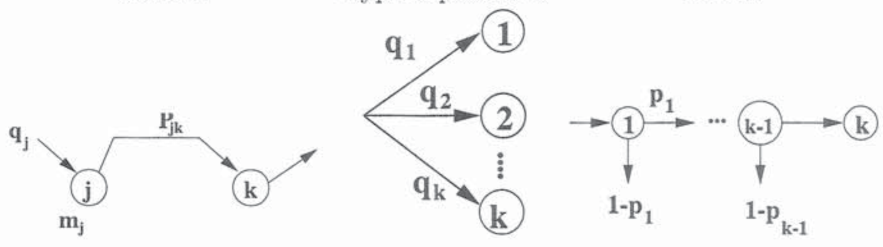
Static Model: Exponential Duration

Face-to-Face Services in a Government Office



Dynamic Model: Phase-Type Duration

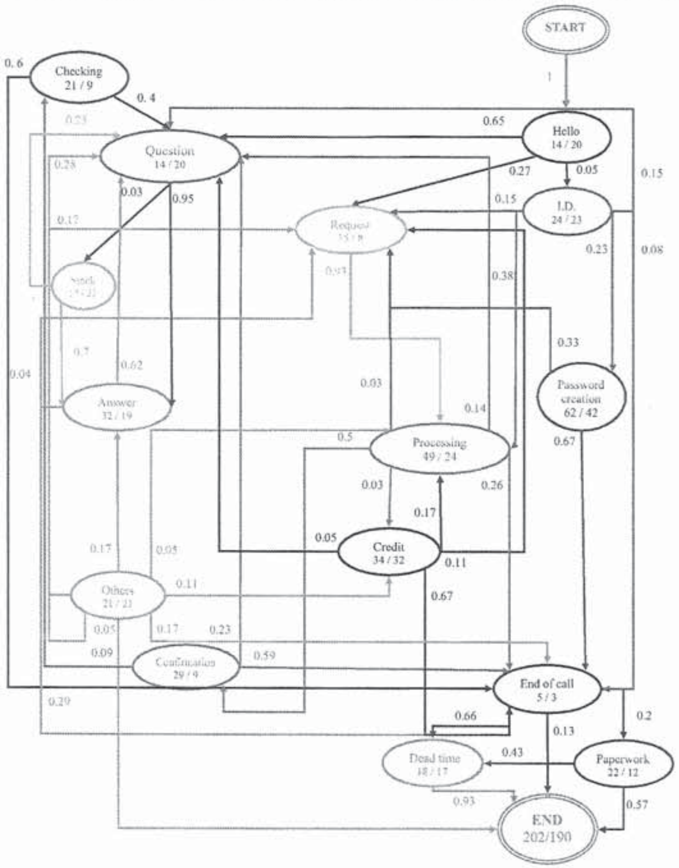
General Hyperexponential Coxian



12

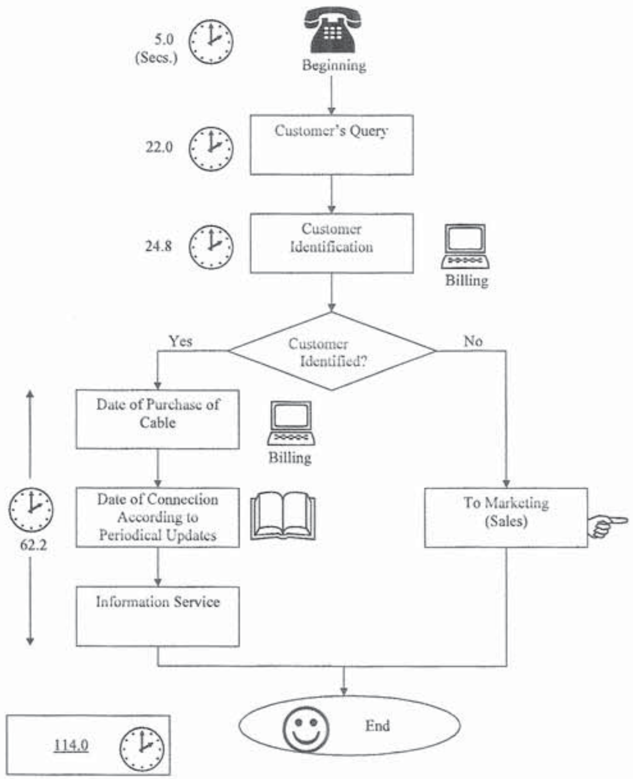
Phase-Type Model of a Telephone Call

Figure 20: **Phase-Type Model of a Telephone Call (# within phases: Mean/STD)**



Service Times: Phase-Type Model

Late Connections



? Where does human-service start / end (recall 144)?
"Average" picture.

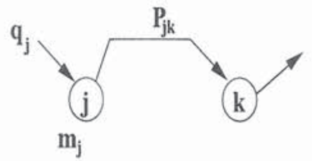
13

Phase-Type Service Times (Durations).

Service-Time = a sequence/collection of tasks, of an *exponential* duration.
There are K types of tasks, indexed by $k = 1, \dots, K$.

m_k = expected duration of task k; $\quad m = (m_k)$
q_k = % of services in which k is first; $\quad q = (q_k)$
P_{jk} = % of incidences in which task j is immediately followed by k. $\quad P = [P_{jk}]$

$1 - \sum_{\ell=1}^{K} P_{k\ell}$ = probability to end service at k.



Fact: service = *finite* number of tasks $\Leftrightarrow \exists [I-P]^{-1}$
Indeed, $[I-P]^{-1}_{jk}$ = expected number of "visits to k", given j was first.
$(q[I-P]^{-1})_k$ = expected number of "visits to k").

As will be articulated below, service-time duration is *Phase-type* (PH).
(Assuming independence among task-durations.)

Definition. Phase-type distribution = absorption time of a finite-space continuous-time Markov chain, with a single absorbing state.

Formally: $X = \{X_t, t \geq 0\}$ Markov on states $\{1, 2, \dots, K, \Delta\}$, with infinitesimal generator

$$Q = \begin{array}{c} 1 \\ \vdots \\ K \\ \Delta \end{array} \begin{bmatrix} & R & r \\ & & \\ 0 & \dots 0 & 0 \end{bmatrix}$$

- Δ absorbing (since $q_{\Delta\Delta} = 0$)
- $r = -R1$ (since $Q1 = 0$)
- $1, \dots K$ transient $\Leftrightarrow \exists R^{-1}$ (fact)

and initial distribution (of X_0) is given by $(q_1, \dots, q_k, 0) = (q, 0)$.

Recall:
$$P\{X_t = k\} = \sum_j q_j [\exp(tR)]_{jk} = q[\exp(tR)]_k$$

Define: $\quad T = \inf\{t > 0 : X_t = \Delta\}$ has phase-type distribution, say $F_T(\cdot)$.
Claim: $\quad F_T(t) = 1 - qe^{tR}1, \ t \geq 0$.

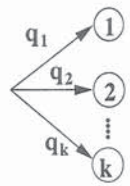
Proof. $P(T > t) = P\{X_t \neq \Delta\} = \sum_k q(e^{tR})_k = qe^{tR}1$.

Parameters:

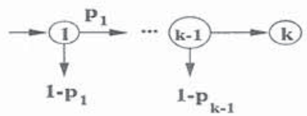
density	$f_T(t) = qe^{Rt}r$
Laplace transform	$\int_0^\infty e^{-xt} F_T(dt) = q[xI - R]^{-1} r$
nth moment	$\int_0^\infty t^n F_T(dt) = (-1)^n n!\, qR^{-n}1$
(mean $= -qR^{-1}1$)	

Special Cases:

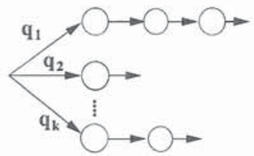
- Exponential (μ) : $R = [-\mu]$ and $q = 1$.
- Erlang: $\to$ [1] $\to$ [2] $\to$ [K] $\quad$ iid tasks / phases $\left(C^2(T) = \frac{1}{K}\right)$.
- Generalized Erlang: exponential phases in series (tandem) $(C^2 < 1)$.
- Hyperexponential: K tasks in parallel (mixture) $(C^2 > 1)$.



- Coxian: K phases; end at phase k with probability p_k.



- Minimum of exponential random variables is exponential.
- Max of exponential random variables is phase-type: e.g., $X_i \sim \exp(1)$ iid. This easily implies that $E(\max X_i) = \sum_i \frac{1}{i}$, Var $(\max X_i) = \sum_i \frac{1}{i^2}$ bounded!
- Erlang mixtures:



14

Importance of Phase-type distributions.

- Empirical + wishful thinking: homogeneous human tasks are exponential.
- Richness: the family of phase-type distributions is dense among all distributions on $[0,\infty)$. For every non-negative distribution G, there exists a sequence of phase-type distributions $F_n \ni F_n \Rightarrow G$.
  (In particular, we can guarantee convergence of any finite number of moments.)

  Dense subfamilies: Coxian, Erlang mixtures.

  For Erlang mixtures, this can be explained by the following two facts:

  1. The family of discrete distributions is dense.

  2. Constants can be approximated by Erlang distributions. Therefore, discrete distributions can be approximated by Erlang mixtures.

- Modelling, via the *method* of *phases*. For example, consider M/PH/1 queue (see HW).

  M/PH/1: state-space is (i,k) $\quad$ (i = number in queue; k = phase of service) or 0; e.g., $0 \xrightarrow{\lambda q_k} (1,k)$.

Representation directly in terms of (q, P, m).

Denote here $R = [I-P]^{-1}$ (as in Mandelbaum & Reiman).
Average work content $\quad E(T) = qRm \qquad (= \sum_j q_j R_{jk} m_k)$.

$$\text{Moments:} \quad E(T^n) = n!\, q(RM)^n q\,, \qquad \text{where } M = \begin{bmatrix} m_1 & & 0 \\ & \ddots & \\ 0 & & m_K \end{bmatrix}$$

$$\frac{E(T^2)}{2(E(T))^2} = \frac{1+C^2(T)}{2} = \frac{q(RM)^2 1}{(qRM1)^2}$$

Pooling Services: Municipality Service System

Current state:

Service Station	No. of Work Stations	Arrivals per Hour	Occupancy	Average Waiting Time [Minutes]	Average Queue Length
Collection - Front Office	4	17.50	0.48	28.08	8.19
Collection - Immigrants	2	8.43	0.76	30.92	4.34
Collection - Back Office	6	3.20	0.13	13.45	0.72
Cashier	2	22.80	0.70	9.73	3.70
Assessment - Front Office	4	11.73	0.50	15.58	3.05
Assessment - Back Office	2	22.80	0.05	11.72	0.07
Land registry office	2-4	2.00			

Recommended staffing in overloaded periods (using model):

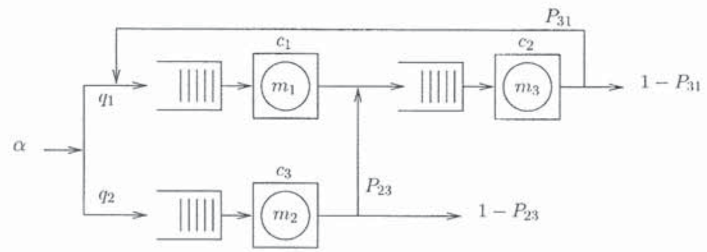
Service Station	Arrivals per Hour	Service Time Time	Recommend No. of Clercks	Percent of Customers Served	Average Waiting Time [Minutes]	Probability of Waiting	Spaces for Waiting
Collection - Front Office	23.40	6.98	6	45	2.1-4.8	0-0.06	4-6
Collection - Immigrants	4.50	14	3	35	7.1-16.2	0.006-0.1	3-5
Collection - Back Office	11.80	12	4	59	7.3-16.8	0.007-0.27	6-9
Cashier	31.40	3.5	3	61	3-6.6	0-0.37	7-10
Assessment - Front Office	16.00	10.9	6	48	3.53-8.1	0-0.09	5-7
Assessment - Back Office	0.60	18.18	2	9	10-22.8	0.015-0.027	2

Total of 24 clerks.

1.

Pooling Services: Municipality Queueing Network (Server's Perspective)

Figure 3 **A Specialized Model with Task Repetition and Feedback**



Station 1 - Collection;
Station 2 - Assessment;
Station 3 - Cashier.

Pooling Services: Municipality Service Times per Service Position

Dept.	Server ID	Service Time Avg. (Min)	Std. Deviation	Utilization %	Service Time Max. (Min)	Total Services
Collection - Front Office	1	7.55 ± 0.68	7.96	37	79.32	370
	2	5.42 ± 0.33	6.27	68	105.20	951
	3	6.51 ± 0.50	6.94	44	63.33	510
	4	8.41 ± 0.75	8.90	42	58.15	377
Collection - Immigrants	5	11.59 ± 0.80	10.88	76	74.60	493
	6	10.32 ± 0.52	8.98	78	50.87	569
Collection - Back Office	7	10.80 ± 1.98	12.82	16	93.73	114
	8	9.07 ± 3.56	11.50	3	52.07	28
	9	18.32 ± 4.90	20.34	10	113.57	47
	10	23.39 ± 5.52	17.75	9	63.77	28
	11	11.99 ± 3.16	14.75	9	70.30	59
	12	16.73 ± 2.34	16.08	28	88.68	128
Cashier	13	2.51 ± 0.21	4.92	48	52.18	1460
	14	3.86 ± 0.18	4.16	72	46.92	1416
Assessment - Front Office	15	13.74 ± 1.07	12.02	62	69.68	340
	16	10.88 ± 0.92	10.60	52	87.92	363
	17	6.66 ± 0.50	6.68	42	49.93	473
	18	11.22 ± 1.30	13.81	45	100.60	302
Assessment - Back Office	19	19.29 ± 5.64	19.99	8	78.27	34
	20	12.2 ± 3.86	8.47	3	29.28	13
Total		7.24 ± 0.10	9.10			8075

- 90% confidence intervals
- 7364 distinct customers

Recall: Exponential $\Rightarrow$ $E = \sigma$ **(i.e. CV=1)**

16

Pooling Services: Municipality Activity Pareto

Rank Service-Types by "Effort".

	Service Type	Avg. Time (Min)	Transactions (% of Total)	Time Allocated (% of Total)	Cumulative (% of Effort)
1	Tax Query	7.25	29.6	34	34
2	Cashier Payment	4.4	42.8	26.3	60.3
3	Title Transfer	12.1	5.5	10.6	70.9
4	Water Query	5.6	8.3	7.35	78.25
5	Owner Change	17.3	1.5	4.2	82.45
6	Title Deed Verification	7.2	3.4	3.9	86.35
7	Waivers & Discounts	12.4	1.4	2.8	89.15
8	Water Disconnection	15.6	1.1	2.6	91.75
9	Discount Application	13.7	0.8	1.8	93.55
10	Update	10.4	1.1	1.8	95.35
11	Information	8.1	1.3	1.7	97.05
12	Measuring Device	5.9	1	0.9	97.95
13	Measurement Req.	12.5	0.4	0.8	98.75
14	Payment Schedule	6.3	0.7	0.7	99.45
15	Account Change	3.8	0.7	0.4	99.85
16	Cash Transfer – Rebate	2.3	0.26	0.1	99.95
17	Water Account Change	1.8	0.14	0.05	100

4 service-types require 80% of effort

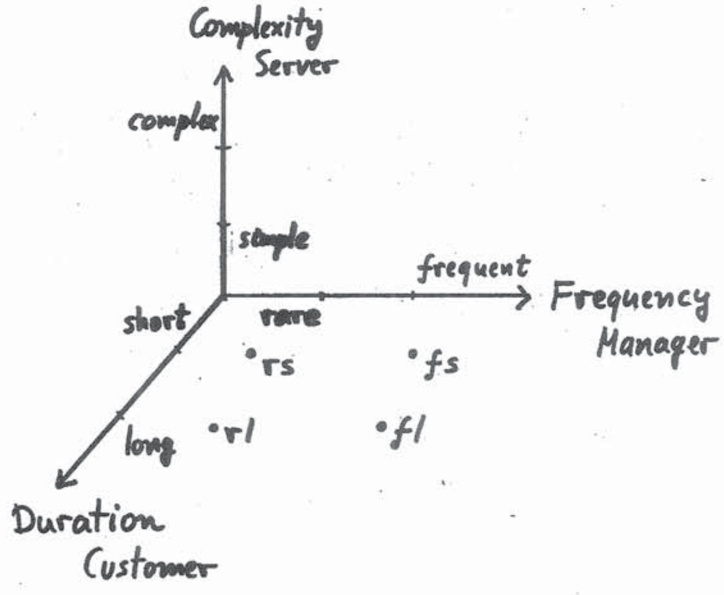
+ space constraints + poor service level

⇒ Redesign network as a single-station

Specialized vs. Flexible; Pooling

Pooling Services: Services Classification

A Classification of Service Tasks



Improvement efforts F C L → R S S

17

Pooling Services: Municipality Server Recommendation

Recommendations: Flexible clerks for all activities. Change from Figure 3 to Figure 4.

Figure 3 A Specialized Model with Task Repetition and Feedback

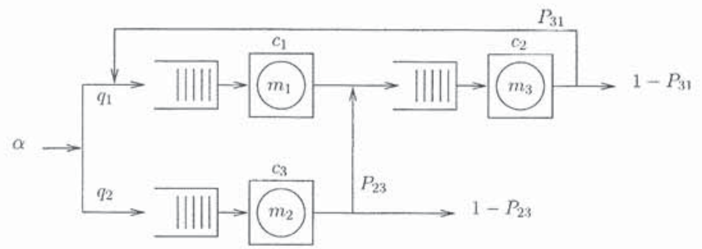
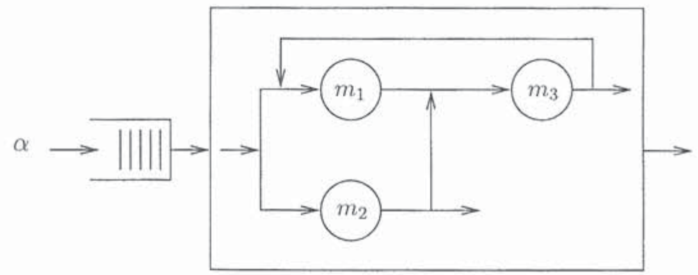


Figure 4 The Flexible Model, under Complete Pooling, that Corresponds to Figure 3



Pooling Services: Municipality Server Recommendation

State of system under recommendations:

Hour	Arrival Rate	Staffing	Occupancy	Waiting Room Size	Average Waiting Time [Minutes]	% Waiting More then 10 Min
7:30-8:30	36.3	(7) 8	(69) 60	8-12	3.20	4.7
8:30-9:30	79.4	(13) 14	(82) 76	14-22	3.10	3.6
9:30-10:30	87.4	(15) 15	(78) 78	16-24	3.05	3.9
10:30-11:30	85.4	(14) 15	(81) 76	15-22	2.85	3
11:30-12:30	64.5	(11) 12	(78) 72	12-18	3.00	3.7
12:30-1:30	24.5	(6) 7	(54) 46	6-8	2.70	2.7
1:30-2:30	24.2	(6) 7	(54) 46	5-8	2.70	2.6
2:30-3:30	30.6	(7) 8	(58) 51	6-9	2.71	2.2
3:30-4:30	11.3	(4) 5	(34) 30	4-5	2.65	3.4

- Number of work-stations: 15.
- Staffing change over time between 5 and 15.
- Guidance on matching available agent to waiting customer.
- Standardization of services and work procedures.
- Turnover clerks to achieve high occupancy.
- Possible separation between Russian speaking clerks.

18

On Pooling in Queueing Networks

Avishai Mandelbaum • Martin I. Reiman
Faculty of Industrial Engineering and Management, Technion, Haifa, Israel
Bell Labs, Lucent Technologies, Murray Hill, New Jersey 07974

We view each station in a Jackson network as a queue of tasks, of a particular type, which are to be processed by the associated *specialized* server. A complete pooling of queues, into a single queue, and servers, into a single server, gives rise to an $M/PH/1$ queue, where the server is *flexible* in the sense that it processes all tasks. We assess the value of complete pooling by comparing the steady-state mean sojourn times of these two systems. The main insight from our analysis is that care must be used in pooling. Sometimes pooling helps, sometimes it hurts, and its effect (good or bad) can be unbounded. Also discussed briefly are alternative pooling scenarios, for example complete pooling of only queues which results in an $M/PH/S$ system, or partial pooling which can be devastating enough to turn a stable Jackson network into an unstable Bramson network. We conclude with some possible future research directions.
(*Service Facility Design; Flexible Server; Specialized Server; Service Operations, Efficiency, Stability, Economics of Scale*)

1. Introduction

A fundamental problem in the design and management of stochastic service systems is that of pooling, namely the replacement of several ingredients by a functionally equivalent single ingredient. We analyze the pooling phenomenon within the framework of queueing networks where in our case, as will be explained momentarily, it can take one of three forms: pooling queues (the demand), pooling tasks (the process) or pooling servers (the resources). Here we consider pooling queues and servers simultaneously, but keep the task structure intact, and we provide an efficiency index (5) to determine when such pooling is or is not advantageous.

Our models are described in terms of customers who seek *service* provided by *servers*. Service amounts to a collection of *tasks*, of which there are a finite number of *types*. Two main models are considered: in the first *specialized* model, each task type has a server and a queue dedicated to it. For example, Figure 1 exhibits a queueing network in which every customer requires a service that constitutes three tasks, and the tasks are carried out successively, each by its own specialized server. Customers arrive at rate α, average task durations are m_k and servers' capacities are c_k. In the second *flexible* model, servers are capable of handling all tasks and they collectively attend to a single queue of services. For example, Figure 2 exhibits such a model, which arises through pooling the tandem network from Figure 1: customers arrive at rate α, seeking the same three-task service as before; they all join a single queue, which is now attended by a single flexible server of capacity $\Sigma_k c_k$.

Customer arrivals are assumed Poisson and task durations exponential. (We comment on these distributional assumptions in the Addendum.) As articulated in §2, we allow a service to consist of a random sequence of tasks in a way that the service duration has a phase-type distribution (a phase corresponds to a task). The specialized (unpooled) model turns out to be a Jackson network (Jackson 1957), as in Figure 3, and the flexible (pooled) architecture is modeled by an $M/PH/1$ system (Neuts 1981) as in Figure 4.

In addition to the above two main models, we also consider briefly alternative designs of pooling. For example, Figure 5 depicts the network from Figure 1, with its queues pooled into a single queue and the servers made flexible while still maintaining their individual identities (see §5.3). Figure 6 depicts partial pooling of only queues and servers 1 and 2 (see §5.4). Figure 7 depicts a split of the service so that a customer, upon completion of a task, rejoins the queue (see §5.5), and additional designs are possible as well. A common feature of our models is that service is unaltered. For example, in Figures 1, 2, 5, 6, and 7, service always consists of tasks 1, 2, and 3 in succession.

1.1. Motivation

The present research arose from an analysis of a service network consisting of several specialized departments. The network was redesigned as a pooled single department, which was still responsible for the same services, but whose servers were flexible enough to process all tasks. In trying to analyze this transition, we found that prevalent pooling models failed to cover our *network* scenario.

Our models provide a new simple framework that helps in assessing the effects on pooling of *utilization*, *variability*, and service *design*. While this is not aimed as a review paper, our framework also relates, as it happens, rather disparate concepts and results, for example (Bramson 1994, Jackson 1957, Klimov 1974, Neuts 1981, Smith and Whitt 1981, and Tcha and Pliska 1977). We believe that the usefulness of the framework goes beyond the original motivating applications, pertaining to the design of telephone call centers (Brigandi et al. 1994), evaluation of communication networks (Smith

Figure 1 A Specialized Model with Tasks Attended by Specialized Servers

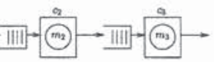


Figure 2 A Flexible Model with Complete Pooling into a Single Queue and a Single Flexible Server

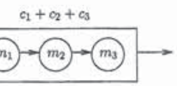


Figure 3 A Specialized Model with Task Repetition and Feedback

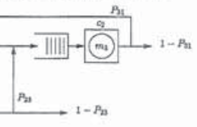


Figure 4 The Flexible Model, under Complete Pooling, that Corresponds to Figure 3

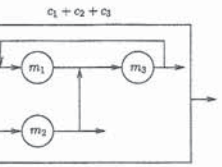


Figure 5 Complete Pooling of Queues Only (Servers Are Made Flexible but Maintain Individual Identities.)

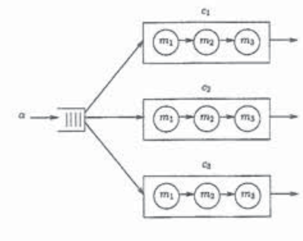


Figure 6 Partial Pooling

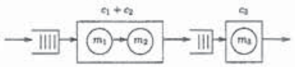
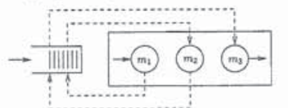


Figure 7 Splitting Services (Each Task Returns to the End of the Queue.)



0025-1909/98/4407/0971$05.00

19

Example: Phase-Type Service Times

Reference: **"Length of Stay of Elderly People in Institutional Long-Term Care"**, Xie, Chaussalet & Millard, 2005.

Operational significance:

- "Most common causes of delay in **discharge from hospital** are patients awaiting placement in a nursing or residential home and awaiting assessment of their needs."
- Significant **costs** associated with maintaining elderly people in care homes, hence relevant to "government agencies (funding, planers), insurance companies, and purchasers and providers of care."

Elderly people go through three states, after being admitted to long-term care:

- **Residential** home care **(R)**;
- **Nursing** home care **(N)**;
- **Discharge** state **(D)**.

Goal: Estimate the **sojourn time in long-term care**, both **duration** and **structure**.

Data: "Paths" of 889 patients, some censored:

- **392 patients:** R → D (219 censored);
- **451 patient:** N → D (156 censored);
- **46 patients:** R → N → D.

The states **R** and **N** are aggregated: Service time in each is modeled by a *Coxian* (Phase-Type) distribution.

Summary: The above approach is potentially useful in other service contexts. For example, estimating **duration** and **structure** of

- *Telephone or face-to-face services*, in which case data censoring is not important since observations are complete; aggregation is significant, balancing complexity against goodness-of-fit.
- *Customers' Impatience*, in which case censoring is very important to account for (as will be explained in due time).

20

J. R. Statist. Soc. A (2005)
168, *Part* 1, *pp.* 51–61

A continuous time Markov model for the length of stay of elderly people in institutional long-term care

H. Xie, T. J. Chaussalet and P. H. Millard

University of Westminster, London, UK

[Received January 2003. Final revision January 2004]

Summary. The paper develops a Markov model in continuous time for the length of stay of elderly people moving within and between residential home care and nursing home care. A procedure to determine the structure of the model and to estimate parameters by maximum likelihood is presented. The modelling approach was applied to 4 years' placement data from the social services department of a London borough. The results in this London borough suggest that, for residential home care, a single-exponential distribution with mean 923 days is adequate to provide a good description of the pattern of the length of stay, whereas, for nursing home care, a mixed exponential distribution with means 59 days (short stay) and 784 days (long stay) is required, and that 64% of admissions to nursing home care will become long-stay residents. The implications of these findings and the advantages of the proposed modelling approach in the general context of long-term care are discussed.

Keywords: Length-of-stay modelling; Long-term care; Markov model; Survival

1. Introduction

In the UK, the National Audit Office has recently reported that the most common causes of delay in discharges from hospital are patients awaiting placement in a nursing or residential home and awaiting assessment of their needs (National Audit Office, 2003). Under the 1990 National Health Service and Community Care Act and the Care Standard Act 2000, local authorities in Great Britain are responsible for the placement and finance of adults in publicly funded residential and nursing home care that conforms to national standards. Discharge to long-term care is a central component of plans for acute hospital care and the demand for long-term care is expected to increase substantially as the population ages (Wittenberg *et al.*, 2001). In England, already 1 in 5 people aged 85 years or over live in a long-term care institution (Laiho, 2001). In addition, the UK Government is planning to fine local authorities for failing to provide vacancies in residential and nursing home care for hospital discharges. Therefore, it is important for both health authorities and local authorities to have a sound understanding of the patterns of the length of stay (LOS) and movements of residents in long-term care.

A recent survey showed that nearly 70% of the residents in residential and nursing homes were publicly funded and were there permanently (Netten *et al.*, 2001). In earlier research, we found that older people who are placed in nursing homes are more likely to have complex problems. Factors such as being male, immobile, dependent in feeding, urine incontinent, having open wounds and taking multiple drugs are associated with nursing home care placements, whereas older people who are admitted to residential home care are likely to be more independent (Xie

Address for correspondence: T. J. Chaussalet, Department of Mathematics, Cavendish School of Computer Science, University of Westminster, 9–18 Euston Centre, London, NW1 3ET, UK.
E-mail: chausst@wmin.ac.uk

0964–1998/05/168051

et al., 2002). Therefore, we would expect differences in the pattern of LOS in residential and nursing home care.

Research in the UK shows that the mortality rate for residents in nursing home care is particularly high in the first few months and then gradually levels out (Smith and Lowther, 1976; Bebbington *et al.*, 2001; Rothera *et al.*, 2002). This observation supports the notion of phases in residents' stay in care homes. In the context of hospital geriatric departments, Harrison and Millard (1991) and Taylor *et al.* (1998, 2000) have shown that, despite the great heterogeneity between individuals (Millard, 1988), compartmental and Markov models, which divide patients' LOSs into short-stay and long-stay phases, capture successfully the behaviour of patients' LOSs. Similar results for residential and nursing home care can be expected.

We model the flow of elderly residents within and between residential and nursing home care by using a continuous time Markov model, in which residents' stay in care homes is modelled as a two-phase process: short stay and long stay. First, we describe the model that we propose and present a procedure for determining the model structure and estimating parameters by the method of maximum likelihood. We also show and discuss results that are obtained from fitting the model to a real data set.

2. A model for movement of elderly people in residential and nursing home care

The proposed conceptual model for the movement of elderly people in residential and nursing care facilities is depicted in Fig. 1. In this model, elderly people can be admitted into residential home care or nursing home care directly, either from the community or following discharge from hospital. In each type of care, residents start their stay in the short-stay phase and either leave care after a short period of time or continue their stay to become long-stay residents. People in residential home care can move to nursing home care if their conditions deteriorate to such an extent that residential home care is no longer adequate. In this paper, we consider only those residents who require local authority funding, and we exclude residents whose admissions are meant to provide short respites for their carers. This restriction is imposed because most local authorities have means of determining suitable care placements for applicants requiring public funds; therefore, these admissions will better reflect residents' physical conditions and needs. Movements from nursing home care to residential home care rarely occur among residents who are supported by local authority funds (Bebbington *et al.*, 2001) and are not modelled.

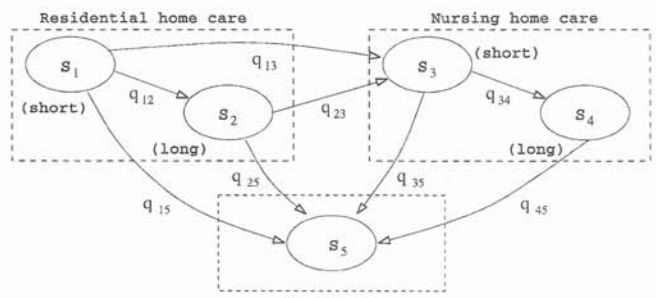


Fig. 1. Markov model for movements of elderly people in residential and nursing home care

21

Discharges from institutional long-term care are considered permanent. They occur predominately by death and, although a small number of residents are discharged to the community or hospital, they are not expected to return to institutional long-term care. Discharges to the community are rare for local-authority-funded residents, and those to hospital usually mean terminal care (Bebbington *et al.*, 2001).

We construct a continuous time Markov model of the flow of elderly people within and between residential and nursing home care. The phases in each type of care and the discharge state form the system states. Given the Markov model that is described in Fig. 1, the generator matrix **Q** is written as

$$\mathbf{Q}=\left(\begin{array}{cc:cc:c} q_{11} & q_{12} & q_{13} & 0 & q_{15} \\ 0 & q_{22} & q_{23} & 0 & q_{25} \\ \hdashline 0 & 0 & q_{33} & q_{34} & q_{35} \\ 0 & 0 & 0 & q_{44} & q_{45} \\ \hdashline 0 & 0 & 0 & 0 & 0 \end{array}\right), \qquad (1)$$

where q_{ij} is the instantaneous transition rate between state i and state j $(i \neq j)$, and the elements in the main diagonal are defined such that row sums are 0, i.e. $q_{ii} = -\Sigma_{j \neq i}\, q_{ij}$.

3. Maximum likelihood estimation of model parameters

The actual states of the Markov model are not observable. We can only observe which type of care a person is in. For example, at any time, we observe that a person is in residential home care but we do not know whether she or he is in a short-stay (S_1) or long-stay (S_2) state. This is an aggregated Markov process, i.e. a Markov process in which system states are aggregated into a number of classes (Fredkin and Rice, 1986). There are three classes in the model that is outlined in Fig. 1, namely residential home care, nursing home care and discharge (denoted by $\mathcal{R}$, $\mathcal{N}$ and $\mathcal{D}$ respectively). We partition the matrix **Q** according to the class structure of the model, i.e.

$$\mathbf{Q}=\begin{pmatrix} \mathbf{Q}_{\mathcal{R}\mathcal{R}} & \mathbf{Q}_{\mathcal{R}\mathcal{N}} & \mathbf{Q}_{\mathcal{R}\mathcal{D}} \\ \mathbf{0} & \mathbf{Q}_{\mathcal{N}\mathcal{N}} & \mathbf{Q}_{\mathcal{N}\mathcal{D}} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad (2)$$

where the submatrices correspond to those delimited by broken lines in equation (1) and the subscripts represent system classes. For instance, $\mathbf{Q}_{\mathcal{R}\mathcal{N}}$ is the submatrix of transition rates from states in $\mathcal{R}$ to states in $\mathcal{N}$, and $\mathbf{Q}_{\mathcal{R}\mathcal{R}}$ that of transition rates between states within $\mathcal{R}$.

The theory of aggregated Markov processes has been motivated by and applied to the modelling of ion channels in neurophysiological applications (Colquhoun and Hawkes, 1981, 1982; Fredkin *et al.*, 1985). Generalization and parameter estimation have been investigated by various researchers, including Ball and Sansom (1989), Fredkin and Rice (1986) and Qin *et al.* (1997). We adapt and modify the approach that was taken by these researchers to suit our modelling needs and to deal with the existence of an absorbing state and censored observations.

3.1. Distribution of sojourn time in a class

Calculating the first-passage time (Cox and Miller, 1965) leads to the probability density function (PDF) of the sojourn time in a class, say class $\mathcal{R}$ (Colquhoun and Hawkes, 1981)

$$f_{\mathcal{R}}(t) = -\phi_{\mathcal{R}}^{\mathrm{T}} \exp(\mathbf{Q}_{\mathcal{R}\mathcal{R}}t)\mathbf{Q}_{\mathcal{R}\mathcal{R}}\mathbf{1}_{\mathcal{R}}, \qquad (3)$$

Table 2. Determination of the number of states in $\mathcal{R}$ and $\mathcal{N}$

Number of states	*Results for residential home care*		*Results for nursing home care*	
	AIC	*BIC*	*AIC*	*BIC*
1	3430.651	3434.733	4879.295	4883.504
2	3433.142	3445.388	4774.788	4787.414
3	3437.142	3457.553	4778.792	4799.835

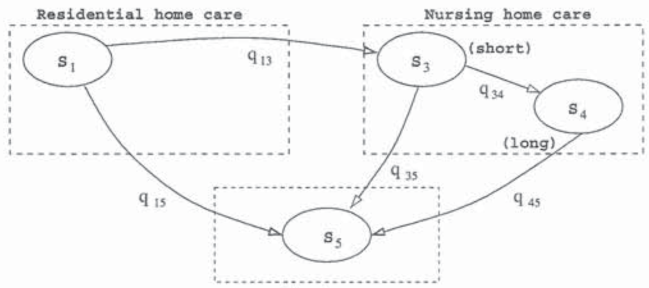


Fig. 2. Structure of the Markov model for the Merton data set

(Fig. 2). The second-stage Markov model fitting procedure converged quickly with the starting-point proposed in Section 3.3. One-dimensional views of the log-likelihood surface along all parameter axes suggested that the maximum was well defined and that the log-likelihood surface was relatively quadratic near the maximum. For each type of care, the close agreement between the survivor curve that was derived from the estimated matrix **Q** (see equation (5)) and the survivor curve that was estimated by the Kaplan–Meier estimator (Kaplan and Meier, 1958) indicates that the Markov model provides a good fit to the data (Fig. 3). This is confirmed by the probability plots (Fig. 4).

4.3. Results

The estimated parameters for the Markov model are summarized in Table 3. These results give interesting insights into the survival patterns of elderly people in institutional long-term care in the London Borough of Merton. A single state provides a good fit to the LOS pattern in residential home care ($\mathcal{R}$), thus indicating a constant rate of departure from $\mathcal{R}$. The average LOS for $\mathcal{R}$ is estimated by $1/(q_{13}+q_{15})$, i.e. 923 days (about 2.5 years). On leaving $\mathcal{R}$, about 79% of the residents will be discharged (permanently) and 21% of them will transfer to nursing home care ($\mathcal{N}$). Two distinctive states are observed in $\mathcal{N}$: a short-stay state with an average LOS of 59 days and a long-stay state with an average LOS of 784 days (about 2.1 years). The rate of discharge from the short-stay state is about five times that from the long-stay state. This agrees with empirical observations that initial mortality is higher for the first few months following admission to nursing care (Smith and Lowther, 1976; Bebbington *et al.*, 2001; Rothera *et al.*,

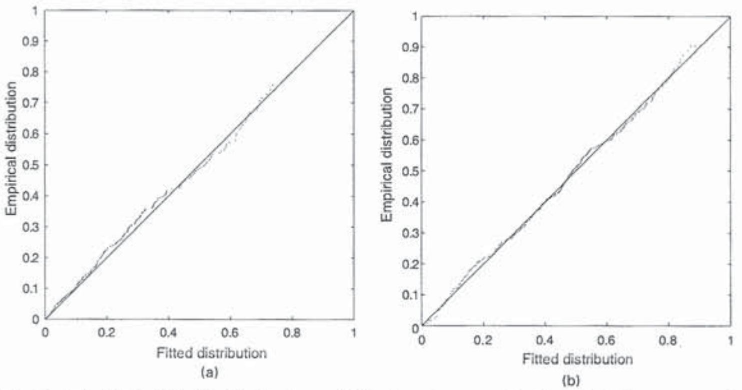


Fig. 4. Probability (*P–P*-) plot of the Markov model fitted survivor curves for (a) residential home care and (b) nursing home care for the Merton data set

Table 3. Estimated parameters for the Merton data set

Parameter	*Estimate*	*Standard error*	*95% confidence interval*
q_{13}	0.000228	0.000034	(0.000162, 0.000293)
q_{15}	0.000855	0.000065	(0.000728, 0.000983)
q_{34}	0.010874	0.002961	(0.005071, 0.016677)
q_{35}	0.006138	0.000793	(0.004584, 0.007692)
q_{45}	0.001275	0.000135	(0.001010, 0.001540)

older people who have been placed in $\mathcal{R}$ by the local authority, 50% will stay more than 21 months, 25% will live longer than 3.5 years and 10% will be there after 5.7 years. Of those who have been placed in $\mathcal{N}$, 50% will stay for more than 8 months, 25% will live longer than 2.1 years and 10% will still be there 4.1 years after they have been admitted.

5. Discussion

We have built a continuous time Markov model which captures the flow of elderly people within and between residential and nursing home care. Using the framework of aggregated Markov processes, we derived a procedure for fitting the model to observed data. By modelling the system of long-term care as a whole, we captured the movements between facilities and estimated parameters by using the overall joint likelihood function. Using a real data set we showed that the LOS in residential home care can be approximated by a single-exponential distribution with mean 923 days, whereas in nursing home care a mixed exponential distribution with short-stay mean 59 days and long-stay mean 784 days is needed to provide a good fit. About 21% of residential home care vacancies were created by transfers to nursing home care and 64% of all admissions to nursing home care will become long-stay residents. In nursing home care, the

mortality rate in the short-stay state is about five times that in the long-stay state. Thus, the model quantifies the large heterogeneity in mortality rates that is widely observed in nursing home care.

Extensive research in the UK has been conducted to identify the characteristics that are associated with differences in survival patterns in long-term care. This research has mainly focused on identifying risk factors that are associated with mortality, e.g. Bebbington *et al.* (2001), Dale *et al.* (2001) and Rothera *et al.* (2002). From the point of view of individual elderly people, their doctors and social workers, the identification of risk factors that are associated with transfer, early death and long-term survival is of considerable importance. But, for planning, care managers and budget holders need to know the overall pattern of LOS in long-term care. Our model complements other research in providing a full picture of the overall behaviour of LOS in residential and nursing home care.

Methods that explicitly model the survival time (or the LOS in care) of elderly people have consistently shown that a mixture of exponentials gives a good fit to observed LOS data (Harrison and Millard, 1991; McClean and Millard, 1993; Taylor *et al.*, 1998, 2000). Struthers (1963) first reported that LOS in a hospital geriatric department in Southampton followed a combination of two exponential curves: one had a 'half-life' of 2 months and the other had a half-life of 2 years. A mixed exponential distribution implies that a proportion of elderly people in residential and nursing home care will live substantially longer than the mean and the longer their stay the longer their expected further stay will be. A large proportion of older people who have been placed by the Merton Social Service Department in residential and nursing home care will stay substantially longer than their expected LOS, 2.5 years and 1.5 years respectively. In residential home care, 25% will live longer than 3.5 years and 10% will live longer than 5.7 years; in nursing home care, 25% will live longer than 2.1 years and 10% will live longer than 4.1 years. This means that short-term decisions to increase the number of permanent admissions to residential and nursing home care will have serious long-term financial and organizational consequences. Such action will result in, as time passes, a reduction in the places that are available for new admissions since the number of beds occupied by residents admitted in earlier years increases.

The model that we have developed in this paper could help planning authorities to understand the overall pattern of usage of resources for elderly people in their catchment area. Our model can be extended to cope with possible differences in survival pattern between nursing care residents who are admitted directly and those who are transferred from residential care, although we did not find significant evidence to suggest that such differences existed in the data set that we used. Further work is needed to confirm our findings and to extend the model to take into account the attributes of elderly people, e.g. their age, gender and physical and mental conditions.

Given the importance of having vacancies in long-term care to run acute hospitals efficiently and the significant costs that are associated with maintaining elderly people in care homes, the findings of this paper should be of great interest to Government departments, insurance companies, health and social services planners, and purchasers and providers of residential and nursing home care.

Acknowledgements

We thank Ms Teresa Temple, Mr Peter Crowther and the late Mr Terry Bucher from the Housing and Social Services Department of the London Borough of Merton for providing the data. This work was partially supported by the Peel Trust and by the Engineering and Physical Sciences Research Council (grant GR/R86430/01).

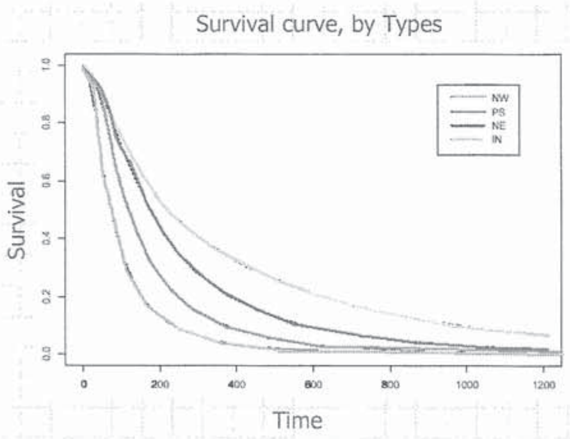
23

Comparing Service Durations

First: Means, Standard Deviations, Medians

	Overall	Regular service	New customers	Internet	Stock
Mean	188	181	111	381	269
SD	240	207	154	485	320
Med	114	117	64	196	169

Then: Distributions (Stochastic Order?)



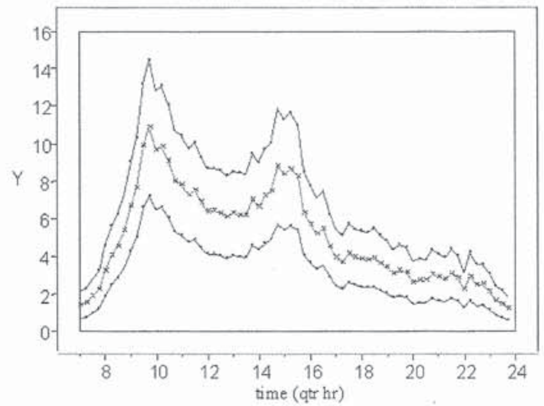
Workload (Offered-Load)

Stationary System Workload (Offered Load):
$R = \lambda \times E[S]$
"minutes" of work (=service) that arrive per "minute".

Example: $\lambda = 3000$ calls/hour; $E[S] = 3$ min.
Consistent time-units, eg. $\lambda = 3000/60 = 50$ calls/min.
Workload $R = 50 \cdot 3 = 150$ min of work per min.
(If time-units hours? hence Workload in **Erlangs**.)

Non-Stationary System Workload (Offered Load):
Use R(t) from $M_t/G/\infty$ queue: $R(t) = E[\lambda(t - S_e)] \times E[S]$

Prediction of Workload: Small Israeli Bank



24

Root Cause Analysis of Emergency Department Crowding and Ambulance Diversion in Massachusetts

A report submitted by the Boston University
Program for the Management of Variability in Health Care Delivery
under a grant from the
Massachusetts Department of Public Health

October, 2002

Emergency Room Diversion Study: Analysis and Findings

Phase I

Phase I of these investigations involved formulation of a conceptual model that would permit data collection and analysis germane to the problem of ambulance diversion. As preparation for this study, a wide range of relevant medical publications, policy statements and commissioned studies were reviewed. This was followed by personal interviews with representatives in government, hospital administration, public health and the Emergency Medicine community. Information was gathered from throughout Massachusetts and from other key states. Particular attention was given to experience in areas where crowding is particularly severe including metropolitan Boston, San Francisco, Los Angeles and the states of Arizona and Florida. Overall, numerous potential root causes of diversion had been articulated both in the medical literature and lay press, but empirical data to support them were lacking. Available research tended to be descriptive, documenting the extent of crowding without clear delineation of its sources. Various solutions had been proposed and implemented, all without consistent benefit. A partial summary of this analysis has been previously released by the Massachusetts Health Policy Forum of Brandeis University.

An operations management perspective suggested straightforward input-throughput-output analysis. Hospital utilization data provided by the Division of Health Care Finance and Policy was therefore reviewed alongside diversion data provided by regional EMS providers. Analysis of this information revealed the likely operation of mechanisms both internal and external to emergency departments. In addition to simple supply/demand imbalances for emergency care, diversion and utilization patterns suggested bottlenecks and backlogs related to the competition of emergency and non-emergency patients for similar resources. The interrelationships of hospital services then became the focus of attention and patient care pathways were explored with administrators from the two study hospitals.

Two paradigms for the quantitative study of interrelationships among hospital departments were considered. The first involved an analytical approach wherein each relationship was identified, its stochastic character estimated, and appropriate

mathematical models applied. The second involved a simulation approach, wherein stochastic relationships were embedded into computer software that translated real patient flow inputs into utilization and capacity information. Computer simulation was ultimately selected as the route of choice because of its scalability and adaptability.

Phase II

Data Collection/Analysis Effort:

The study was performed at two hospitals in Massachusetts: Hospital A, a large tertiary academic hospital, and Hospital B, a medium-sized acute care community hospital. The following data were collected:

- 42 days of information covering:
- 6000+ admissions
- 8000+ ED visits
- 2000+ staffing/capacity data points
- 300,000+ patient movement/status data points

In order to analyze the relationship between diversion status and other factors within the hospital environment all measures were split into observations at one hour increments. The study period of 42 days, with 24 hours per day, yielded a total of 1008 full sets of observations. The analysis required collection of patient flow data well beyond the usual capabilities of contemporary hospital information systems.

Point-biserial coefficients of correlation, with diversion status as the binary variable, were examined against a variety of factors. Comparisons when using full hours of diversion versus partial hours as the "true" condition did not reveal significant differences, so partial diversion hours were evaluated as the "true" binary throughout the analysis for the sake of consistency.

It is important to note that in the real world the decisions to commence or cease diversion status are, but their nature, highly subjective. Because the purpose of the study was to examine the root causes of diversion, we did not approach the task from the standpoint of critiquing or attempting to influence this inherent operational subjectivity. As a result, any such analysis is itself subjective to a certain degree.

Because both hospitals straddled EMS regional borders and diversion rules vary by region, each hospital's data was used for the sake of determining diversion status rather than using centralized EMS data. Also, all diversions were considered equally rather than separately analyzing the factors related to each individual diversion type.

Patterns of diversion were also examined as averages across the hours of the day and the days of the week in order to ascertain relevant hour of the day and day of the week patterns. This data analysis was performed separately for each of the hospitals.

25

Hospital A:

Diversion Pattern "Hospital A - Diversion Minutes by Hour"

- There were a total of 22 episodes of diversion which started and ended within the study, with an average length of 814 minutes. There was one episode that began prior to the study and ended after the study began and so is not included in this calculation, nor in any calculations which involve the beginning of diversion episodes.
- The hourly diversion pattern shows diversion is highest in the evening hours, settles back down during the early morning hours, and then stays steady until the mid to late afternoon (see Fig. 1).
- The goal of the project was to determine the drivers which create this pattern.

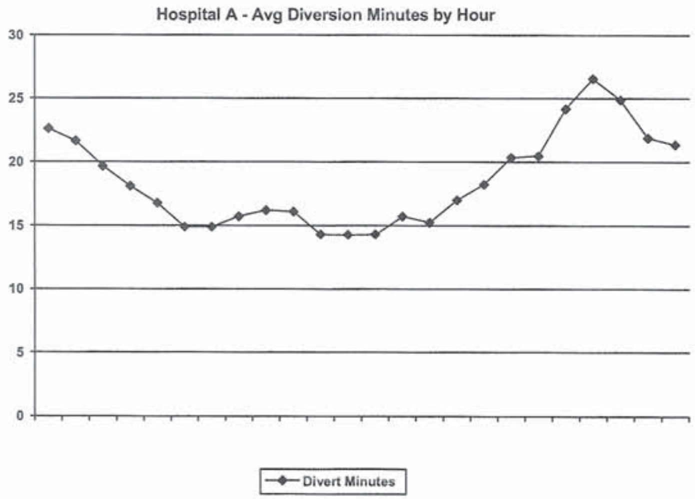


Fig. 1

The following 3 hypotheses were tested to determine the drivers of diversions:

1. ED arrival rate is too high, leading to diversion when the ED becomes full.
2. ED processing of patients is too slow, causing backups that lead to diversion
3. ED arrival and processing rates are fine, but there are not enough beds in the hospital to accommodate the admissions.

There are seven sets of data (see Fig. 2), each representing a different view of arrivals into the ED. The "Arrivals_0" category only includes new arrivals from the hour in question. Each subsequent category, from "Arrivals_1" to "Arrivals_6" includes one more hour's worth added to the total. In other words, "Arrivals_1" includes arrivals from the current hour added to the arrivals from the previous hour, "Arrivals_2" includes all of "Arrivals_1" plus the arrivals from two hours ago, and so on. This is what accounts for the stacked shape as each additional hour is layered on top. Because average length of stay was 340 minutes, 6 hours is used as the maximum lag. Correlation coefficients from each of these cumulatives to Avg Diversion Minutes by hour are as follows:

Arrivals_0 = -0.073
Arrivals_1 = 0.001
Arrivals_2 = 0.078
Arrivals_3 = 0.165
Arrivals_4 = 0.259
Arrivals_5 = 0.359
Arrivals_6 = 0.460

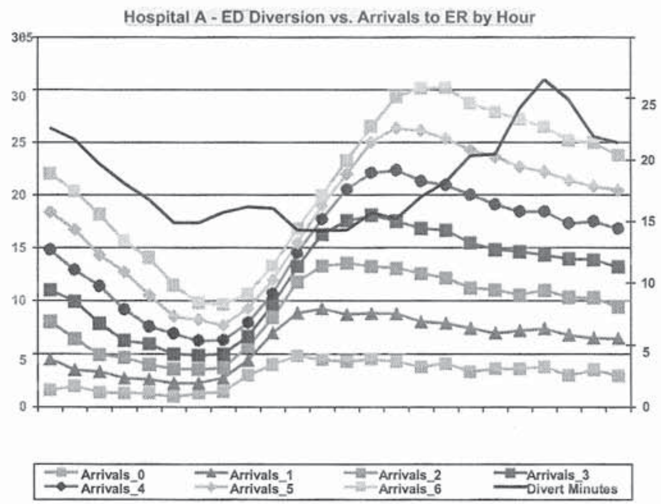


Fig. 2

There is also a possible corollary to hypothesis #1, that overall ED census is a driver of diversion. When counting the non-boarding census and comparing it to diversion status, however, the resulting point-biserial coefficient (r = -0.051) makes clear that this potential explanation should be rejected as well.

21

again points towards examining hospital capacity as the primary determinate of diversion.

Census/Admissions/Discharges: Hospital B

The overall relationship between inpatient census and ED boarders in Hospital B was similar to that of Hospital A. However, detailed analysis of admission sources in Hospital B is not presented because scheduled demand played a far smaller role than that observed in Hospital A.

During the study period, there were 1,158 weekday unscheduled admissions (average: 38.6/day) and 208 weekday scheduled admissions (average: 6.9/day). This suggests very little operational flexibility in controlling the variability or timing of scheduled arrivals. This likely reflects a fundamental difference between most community hospitals and larger academic centers.

Hospital B Conclusions:

The findings at Hospital B are consistent with and reinforce those at Hospital A. Specifically, there was no evidence that ED process times were temporally or mechanistically related to ED diversion while the relationship between ED arrival rate and diversion was weak. Instead, the data suggest that factors outside of the ED that combine to increase boarders and limit ED capacity are more important.

Phase II Summary:

Detailed flow analysis in two very different types of hospitals yielded similar findings with respect to the root cause of emergency department crowding and ambulance diversion. Neither increased patient inflow nor increased process time could be strongly related to diversion status. Instead, diversion was seen as an outflow problem, with busy emergency departments crowding as patients await transfer to crowded inpatient services. This problem is exacerbated in hospitals with large volumes of scheduled admissions, since these necessarily compete for the same resources. The "collision" of scheduled and unscheduled patient flows results in diversion patterns that are specific and reproducible. Because scheduled patient flows are theoretically controllable, better understanding of this phenomenon may suggest means of decreasing diversion. If the experience here may be generalized, we conclude that institutions with small (or uncontrollable) scheduled patient flows will require addition of resources *on the inpatient side* if diversion is to be substantially reduced.

27

Winter 2009A
(See also Winter 2007A+B, Spring 2007B)

שאלה 4 – תיאוריה ושימושיה: עומס מוצע (Offered-Load)

חלק 1. ייצוגים.

במערכת שירות $M_t|G|\infty$, המופע הוא פואסוני לא הומוגני בזמן עם קצב מופע $\{\lambda(t), t\ge 0\}$, ומשכי השירות S הם בלתי-תלויים, שווי התפלגות, ובלתי תלויים במופע. נסמן ב- L(t) את **מספר האנשים במערכת** בזמן t ויהי R(t) **ממוצע** מספרם: $R(t)=EL(t)$.

ניתן להראות ש- $R(t)$ נתון על ידי כל אחד משלושת הביטויים הבאים:

$$R(t)=\int_{-\infty}^{t}\lambda(u)P(S>t-u)du=E\left(\int_{t-S}^{t}\lambda(u)du\right)=E(S)\cdot E(\lambda(t-S_e))$$

S_e מסמן את "זמן שארית השירות" כפי שהוגדר בכיתה.
(בייצוגים הנ"ל, $\lambda(t)=0$ עבור t שליליים.)

4.1.1 נניח שמשך השירות S הוא **דטרמיניסטי** D: $S\equiv D$. הוכיחו את השוויון בין שלושת ההצגות של $R(t)$ למקרה זה.

<u>תשובה</u>:

S_e מתפלג אחיד על [0,D]. $S_e\overset{d}{=}U(0,D)$.

$$\text{Representation }3=D\cdot E(\lambda(t-S_e))=D\cdot\int_0^D\frac{1}{D}\cdot\lambda(t-x)dx=\int_0^D\lambda(t-x)dx\underset{u=t-x}{=}\int_{t-D}^{t}\lambda(u)du=\text{Rep. }2$$

$$\text{Rep. }1=\int_{-\infty}^{t}\lambda(u)1_{\{u>t-D\}}du=\int_{t-D}^{t}\lambda(u)du=\text{Rep. }2$$

4.1.2 הסבירו כיצד ניתן **להרחיב** את המקרה הדטרמיניסטי ב- 4.1.1 הנ"ל למקרה של S **המקבל מספר סופי של ערכים**, דהיינו כיצד מוכיחים שאכן:

$$\int_{-\infty}^{t}\lambda(u)P(S>t-u)du=E\left(\int_{t-S}^{t}\lambda(u)du\right)=E(S)\cdot E(\lambda(t-S_e))$$

הערה: מאחר וניתן לקרב כל משתנה מקרי, בדיוק רב כרצונכם, על ידי משתנה מקרי המקבל מספר **סופי** של ערכים (מ"מ דיסקרטי), סעיף זה מוביל להוכחת שלושת הזהויות הנ"ל למקרה של S כללי.

<u>תשובה</u>:

נניח p_i w.p. $S=D_i$, מ"מ דיסקרטי.

על סמך עמידה מוטה, $\frac{p_iD_i}{E(S)}$ w.p. $S_e=U_i$, כאשר $U_i\overset{d}{=}U(0,D_i)$.

לכן,

$$\text{Representation }2=\sum_i p_i\int_{t-D_i}^{t}\lambda(u)du\underset{\lambda_i=\lambda\cdot p_i}{=}\sum_i\int_{t-D_i}^{t}\lambda_i(u)du\underset{4.1.1}{=}\sum_i D_iE(\lambda_i(t-U_i))=$$

$$=E(S)\sum_i\frac{p_iD_i}{E(S)}E(\lambda(t-U_i))=E(S)E(\lambda(t-S_e))=\text{Representation }3$$

$$\text{Representation }2=\sum_i p_i\int_{t-D_i}^{t}\lambda(u)du=\sum_i p_i\int_{-\infty}^{t}\lambda(u)1_{\{u>t-D_i\}}du=$$

$$=\int_{-\infty}^{t}\lambda(u)\sum_i p_iP(D_i>t-u)du=\text{Representation }1$$

28

חלק 2. שימושים.

בכתה הוסבר ש- R(t), כפי שהוגדר לעיל, משמש כהגדרת **העומס המוצע (Offered Load)** למערכת שירות בזמן t, כאשר המופע אליה ומשכי השירות בה הם כמתואר בחלק 1. (כזכור, R(t) נמדד ב- Erlang).

4.2.1 הסבירו בקצרה מדוע אכן **ההגדרה מתאימה**. עשו זאת בהסתמך על הייצוג הרביעי הבא של R(t):

נסמן – $A(t)$ – מספר המופעים הכולל עד זמן t (כולל).

$E[A(t)] = \Lambda(t)$ - תוחלת מספר המופעים המצטבר עד זמן t.

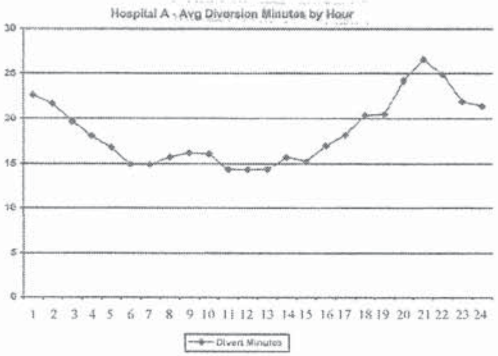
אזי, מאחת ההצגות של R(t) נובע שמתקיים: $R(t) = E\left[\Lambda(t) - \Lambda(t-S)\right]$.

תשובה:

R(t) = כמות העבודה הנמצאת במערכת בזמן t; עבודה נמדדת ביחידות זמן שירות, או לחילופין מספר יחידות זמן-שרת הנדרשות לטיפולה. לכן $R(t) = E\left(L(t)\right)$.

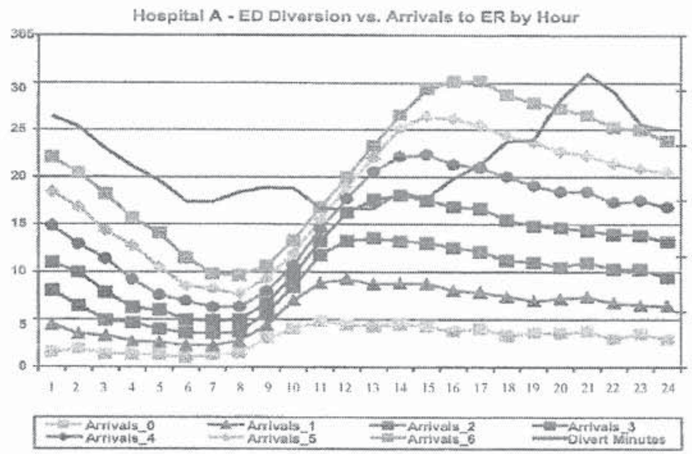
הביטוי $R(t) = E\left[\Lambda(t) - \Lambda(t-S)\right]$ ממחיש יפה את העובדה שכמות העבודה בזמן t כוללת עבודה שהגיעה לפני כן, בהתאם לתהליך ההגעות הממוצע $\Lambda(t)$.

4.2.2 הגרף הבא מתאר את תופעת ה- Ambulance Diversion (הסטת אמבולנסים) שתוארה בכיתה (ונפוצה בארה"ב). למשל ערך הגרף בשעה 21:00 הוא אחוז הפעמים בתקופה הנמדדת שהשעה 20:00-21:00 הוכרזה כשעה שבה המיון חסום לכניסת חולים חדשים דרך אמבולנסים (במקרה שלנו 27%).



הגרף הבא מנסה לקשר בין שעת שיא הסטת האמבולנסים לשיא העומס במיון. הסבירו את הגרף. כיצד ניתן היה לשפר את ההתאמה בין הזמנים בהם מתרחשים השיאים?

שימו לב: תוחלת זמן השהיה במיון הנדון היא 6 שעות.



תשובה:

הגרף Arrivals_0 מתאר את כמות החולים הממוצעת לשעה שהגיעו למיון לאורך היום, עם שיא הגעות ב- 11 בבוקר. הגרף Arrival_1 בשעה t הוא כמות ההגעות בשעה t + כמות ההגעות בשעה t-1, כלומר הוא סוכם את ההגעות בשעתיים האחרונות. וכו'... כלומר הגרף Arrivals_6 סוכם את ההגעות ב- 6 השעות האחרונות ע"י הנוסחה: $Arrivals\_6(t) = \Lambda(t) + \Lambda(t-1) + ... + \Lambda(t-6)$.

לכן גרף 6 מחשב בפועל את העומס המוצע (R(t)) לפי הקירוב הבא:

$R(t) = E[\Lambda(t) - \Lambda(t-S)] \approx \Lambda(t) - \Lambda(t - E[S]) = \Lambda(t) + \Lambda(t-1) + ... + \Lambda(t-6)$. בעזרתו מראים ששיא <u>העומס</u> מתקבל בשעה 15-17, כלומר הוא מוזז בזמן לעומת שיא ההגעות שקורה ב- 10-11.

אך עדיין יש פער זמן לא מוסבר בין שיא העומס הנ"ל לשיא ההסטות (שמתרחש בשעה 21). הפרש זה יכול לנבוע, למשל, מאחד ההסברים הבאים (או משילובם): א. האמד ל- R(t) לא מספיק מדויק. במקרה זה אנו ממליצים על השיפור הבא: כדאי להשתמש בנוסחת העומס המוצע המדויקת, או באחד הקירובים היותר מדויקים שנלמדו בכיתה.

ב. הסטת האמבולנסים, כפי שהוסבר בכיתה, בדרך כלל מופעלת כאשר מחלקות האשפוז חסומות, וזה קורה בשעה יותר מאוחרת, לעומת שעת שיא העומס במיון.

ג. יתכן שפער הזמנים מצביע על כך שהפעלת נוהל הסטת האמבולנסים מופעל מאוחר מדי, כלומר רק לאחר שרואים עומס בלתי סביר במיון, מתחילים להפעיל אותו ועד שהוא מגיע לאפקטיביות מירבית עובר זמן ניכר.

29