Class 7 (14/12/2005)

Arrivals: Some Loose Ends
Service Times; Phase-type Distributions

Arrivals: Review

- Poisson processes (Scaling a Bernoulli Process);
- Brownian Motions (Scaling and Centering);
- A unifying (axiomatic) framework: Levy Processes.

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.

Recitation 7

- Log-Normal models for Call Center Service Times;
- Phase-Type Services - An Example;

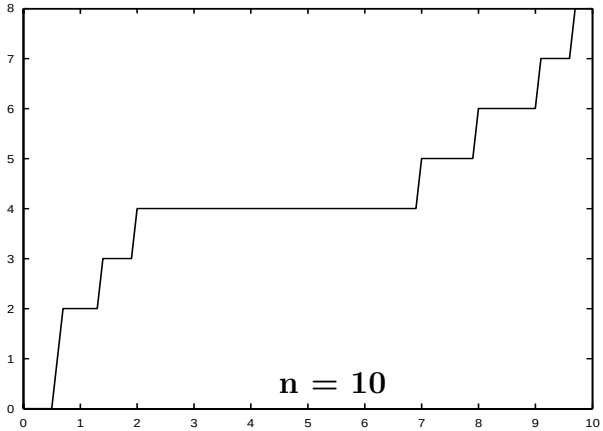
Laws of Congestion: Old and New

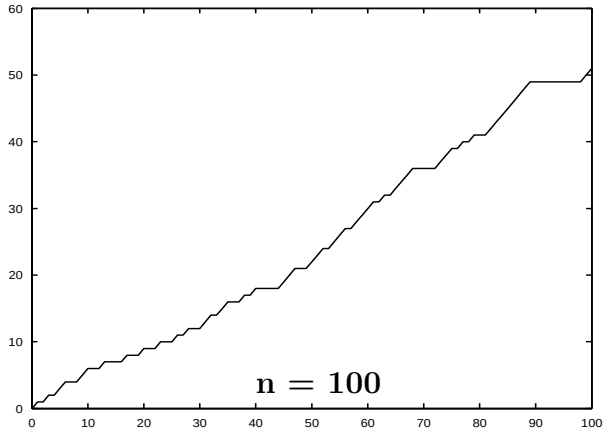
The 0-th Law for (The) *Causes of Operational* Queues :
Scarce Resources and Synchronization Gaps (in DS-Project Networks);
The First Law of *Conservation* :
Little's Law for Customers, Service-providers and Managers.
Little's Law for the Offered Load (Utilization Profiles).
The Second Law of Completely *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).

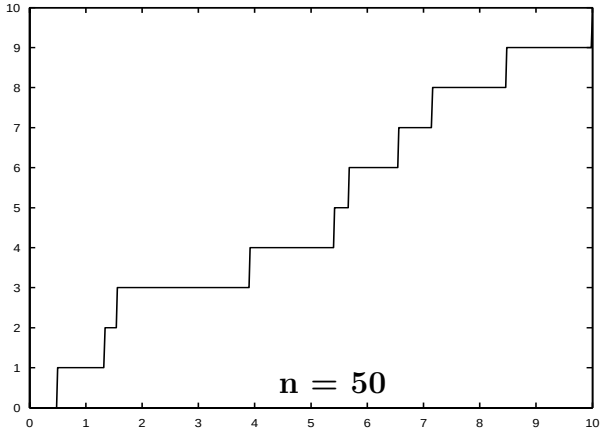
Bernoulli ⇒ Poisson

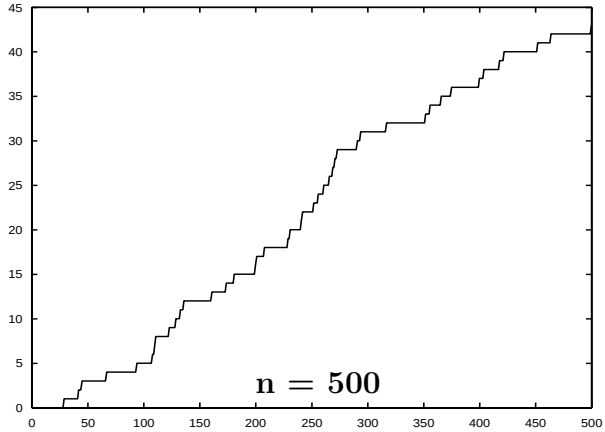
$$\forall n,\ iid \qquad \Delta_k^n = \begin{cases} 1 & p_n = \frac{\lambda}{n} \\ \text{wp} & \\ 0 & q_n = 1 - p_n \end{cases} \quad , \quad S^n(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k^n \ , \quad t \geq 0 \, .$$

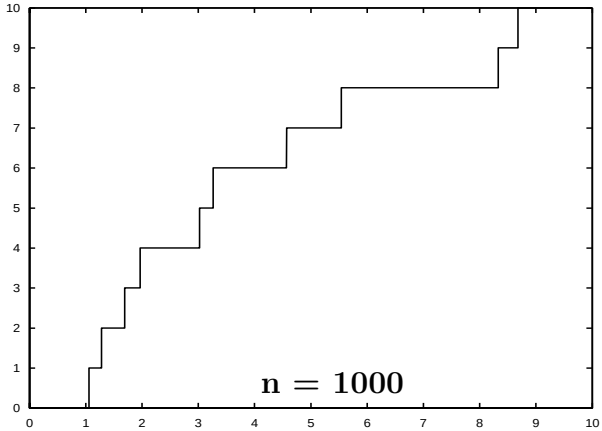
Rare Events: $S^n(nt) \sim$ Binomial $(\lfloor nt \rfloor,\ \frac{\lambda}{n}) \Longrightarrow$ Poisson $(\lambda t),\ t \geq 0$.

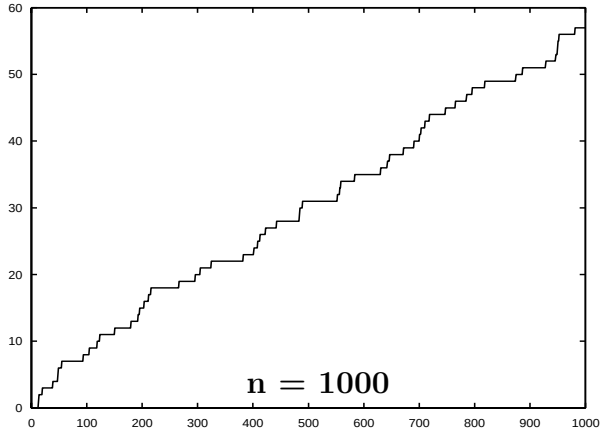












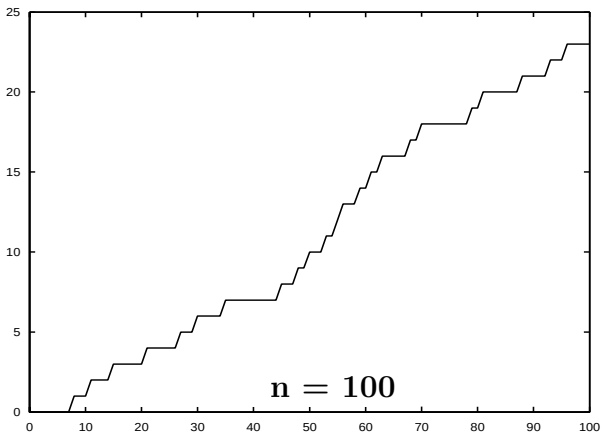
$S^n(nt),\ 0 \leq t \leq 10;\ \lambda = 1.$

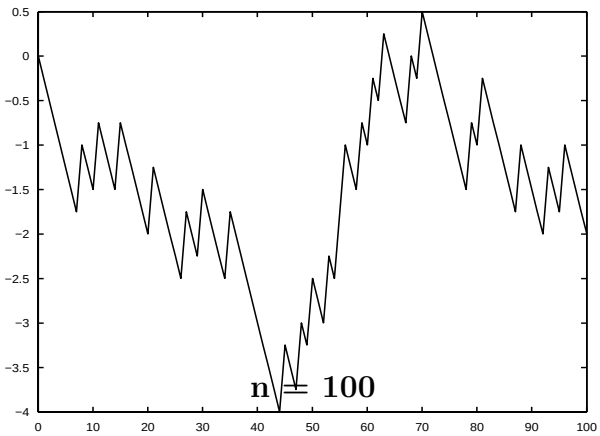
$S^n(t),\ 0 \leq t \leq n;\ \lambda = 50.$

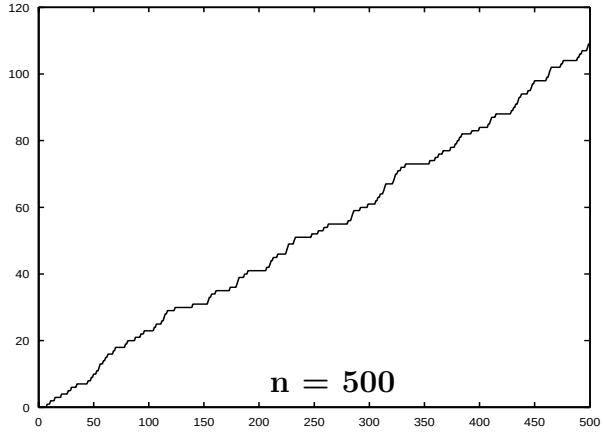
Bernoulli $\Rightarrow$ **Brownian** (Review)

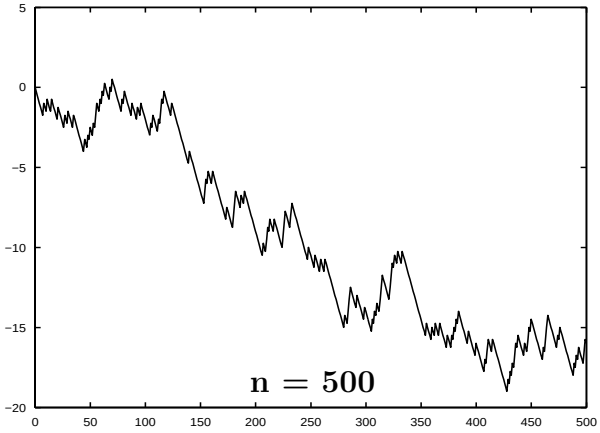
$$iid \qquad \Delta_k = \begin{cases} 1 & p = 0.25 \\ \text{wp} & \\ 0 & q \end{cases} \quad , \quad S(t) = \sum_{k=1}^{\lfloor t \rfloor} \Delta_k^n \ , \quad t \geq 0 \, .$$

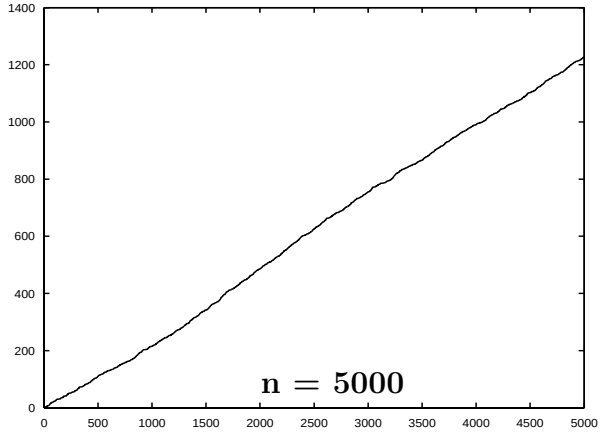
FSLLN: $\frac{1}{n} S(nt) \xrightarrow{wp1} p \cdot t$; FCLT: $\sqrt{n}[\frac{1}{n} S(nt) - p \cdot t] \stackrel{d}{\Rightarrow} BM(0, pq)$

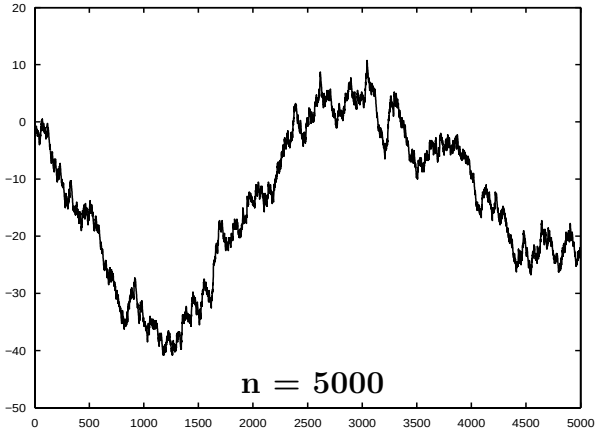












$S(t), \ 0 \leq t \leq n$ $\qquad$ $S(t) - pt, \ 0 \leq t \leq n$

Stochastic-Process Limits

An Introduction to Stochastic-Process Limits And their Application to Queues

Ward Whitt

AT&T Labs - Research
The Shannon Laboratory
Florham Park, New Jersey

Draft
June 13, 2001

1.1.4. Making an Interesting Game

We have digressed from our original game of chance to consider the statistical regularity observed in the plots, which of course really is our main interest. But now let us return for a moment to the game of chance.

A gambling house cannot afford to make the game fair. The gambling house needs to charge a fee greater than the expected payoff in order to make a profit. What would be a good fee for the gambling house to charge?

From the perspective of the gambling house, one might think the larger the fee the better, but the players presumably have the choice of whether or not to play. If the gambling house charges too much, few players will want to play. The fee should be large enough for the gambling house to make money, but small enough so that potential players will want to play. We take that to mean that the individual players should have a good chance of winning.

One might think that those objectives are inconsistent, but they are not. The key to achieving those objectives is the realization that *the player and the gambling house experience the game in different time scales.* An individual player might contemplate playing the game 100 times on a single day, while the gambling house might offer the game to hundreds or thousands of players on each of many consecutive days.

Thus, the player might evaluate his experience by the possible outcomes from about 100 plays of the game, while the gambling house might evaluate its experience by the possible outcomes from something like $10^4 - 10^6$ plays of the game. What we need, then, is a fee close enough to \$0.50 that the player has a good chance of winning in 100 plays, while the gambling house receives a good reliable return over $10^4 - 10^6$ games.

A reasonable fee might be \$0.51, giving the gambling house a 1 cent or 2% advantage on each play. (Gambling houses actually tend to take more, which shows the appeal of gambling despite the odds.) To see how the \$0.51 fee works, let us consider the possible experiences of the player and the gambling house. In Figure 1.9 we plot six independent realizations of a player's position during 100 plays of the game when there is a fee of \$0.51 for each play. The game looks pretty interesting for the player from Figure 1.9. The player has a reasonable chance of winning. Indeed, the player wins in plots 3 and 5, and finishes about even in plot 2. How do things look for the gambling house?

To see how the gambling house fares, we should look at the net payoffs over a much larger number of games. Hence, in Figures 1.10 and 1.11 we plot

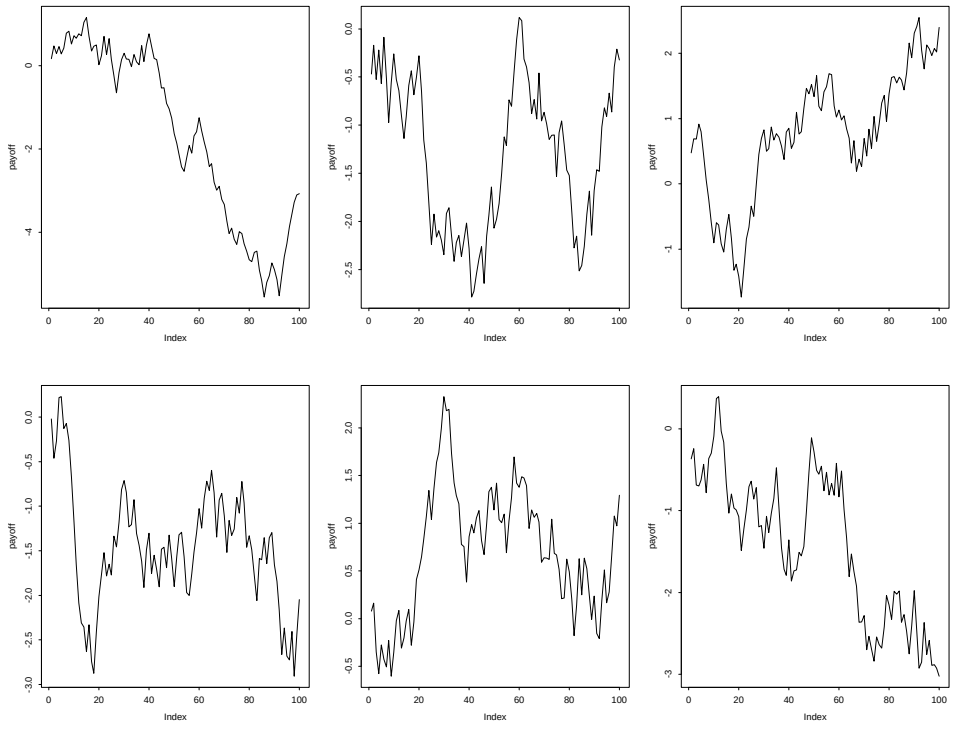


Figure 1.9: Six possible realizations of the first 100 net payoffs, positions of the random walk $\{S_k - 0.51k : k \geq 0\}$, with steps U_k uniformly distributed in the interval $[0, 1]$ and a fee of \$0.51.

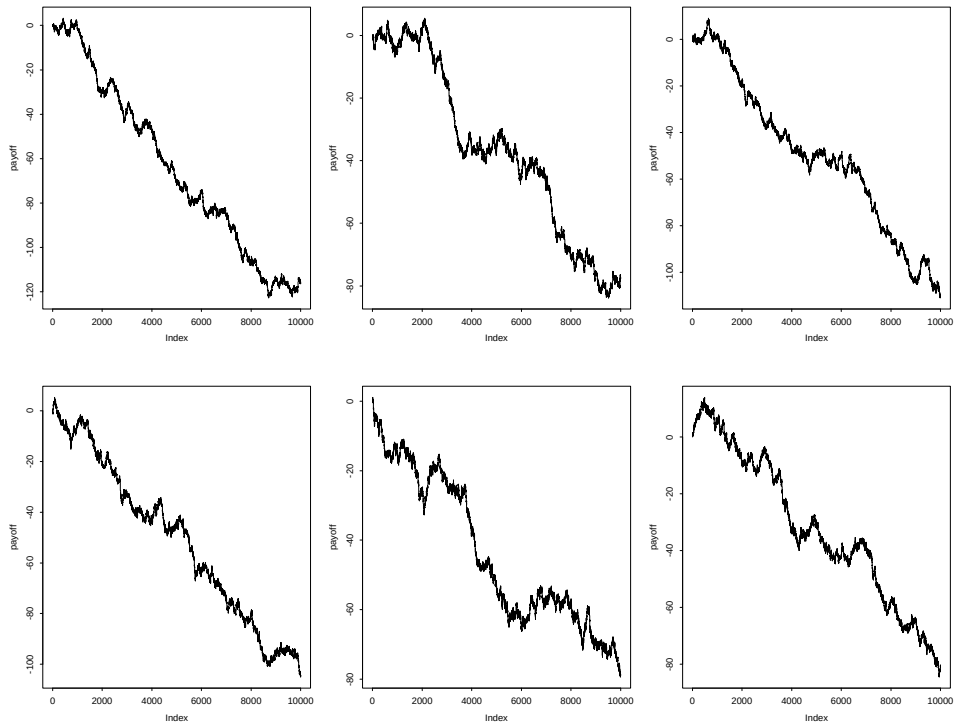


Figure 1.10: Possible realizations of the first 10^4 net payoffs (steps of the random walk $\{S_k - 0.51k : k \geq 0\}$ with steps U_k uniformly distributed in the interval $[0, 1]$.

six independent realizations of a player's position during 10^4 and 10^6 plays of the game. As before, we let the plotter automatically do the scaling, so that the units on the vertical axes change from plot to plot. But that does not alter the conclusions. In these larger time scales, we see that the player consistently loses money, so that a profit for the gambling house becomes essentially a sure thing. When we increase the number of plays to 10^6, there is little randomness left. That is shown in Figure 1.11. Further repetitions of the experiment confirm these observations. We again see the regularity associated with a macroscopic view of uncertainty.

Above we picked a candidate fee out of the air. We could instead be more systematic. For example, we might seek the largest fee such that the player satisfies some criteria indicating a good experience. Letting the fee for each game be f, we might want to constrain the probability p that a

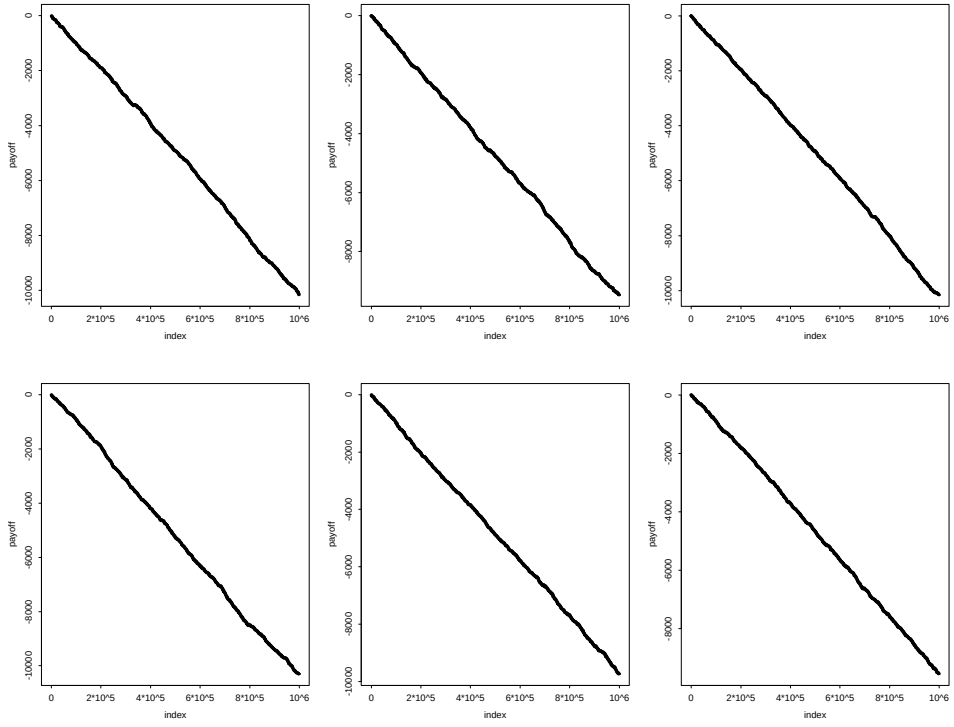


Figure 1.11: Possible realizations of the first 10^6 net payoffs (steps of the random walk $\{S_k - 0.51k : k \geq 0\}$ with steps U_k uniformly distributed in the interval $[0, 1]$.

player wins at least a certain amount w, i.e., by requiring that

$$P(S_{100} - f(100) \geq w) \geq p \ .$$

Given such a formulation, we can determine the optimal fee f, i.e., the maximum fee f such that the constraint is satisfied, which is attained when the probability just equals p.

As noted at the outset, when we consider making the game interesting, we might well conclude that a uniform payoff distribution for each play is boring. We might want to have the possibility of much larger positive and/or negative payoffs on one play. It is easy to devise more interesting games with different payoff distributions, but the statistical regularity associated with large numbers observed above tends to be the same. Readers are invited to make their own games and look at the net payoffs for 10^j plays for various values of j.

An extreme case that is often attractive is to have, like a lottery, some small chance of a very large payoff. However, with independent trials, as determined by successive spins of the spinner, the gambling house faces the danger of having to make too many large payoffs. Such large losses are avoided in lotteries by not letting the game be based on independent trials. In a lottery only a few prizes are awarded (and possibly shared) so that the people running the lottery are guaranteed a positive return. However, an insurance company cannot control the outcomes so tightly, so that careful analysis of the possible outcomes is necessary; e.g., see Embrechts, Klüppelberg and Mikosch (1997). We too will be interested in the possibility of exceptionally large values in random events.

1.2. Stochastic-Process Limits

The plots we have looked at indicate that there is statistical regularity associated with large n, i.e., with large sample sizes. We now want to understand *why* we see what we see, and what we will see in other related situations. For that purpose, we turn to probability theory; see Ross (1993) and Feller (1968) for introductions.

1.2.1. A Probability Model

We can use probability theory to explain what we have seen in the random walk plots. The first step is to introduce an appropriate mathematical model: Assuming that our random number generator is working properly

Mathematical Framework: Levy Processes

Discrete-time: Random Walk

$$\begin{aligned} S(n) &= \Delta_1 + \cdots + \Delta_n, \quad n \geq 0, \quad \text{where } \Delta_1, \Delta_2, \ldots, \text{i.i.d. r.v.} \\ S(0) &= 0\ . \end{aligned}$$

Properties:

1. $S(m+n) - S(m) \stackrel{d}{=} S(n) - S(0) \quad \forall\, m, n \geq 0$ ($\stackrel{d}{=}$ same distribution)
2. $S(m_1) - S(0),\ S(m_2) - S(m_1),\ S(m_3) - S(m_2), \ldots$ independent $\forall\, m_1 \leq m_2 \leq \cdots$

$S = \{S(n), \quad n \geq 0\}$ has **stationary** (1) and **independent** (2) increments.

The continuous-time analogue is a

Levy process A stochastic process $X = \{X_t,\ t \geq 0\}$ is a Levy process if

(0) $X(0) \equiv 0$ (for simplicity);

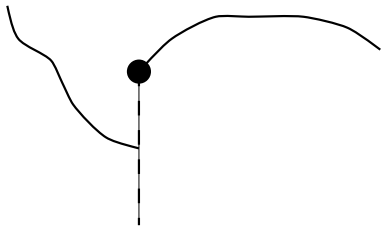
(1) X has *stationary* increments, that is
$X(t+\tau) - X(t) \stackrel{d}{=} X(\tau) \quad \forall\, t,\ \tau \geq 0$;

(2) X has *independent* increments, that is
$X(t+\tau) - X(t)$ independent of $\{X(s),\ s \leq t\}, \quad \forall\, t,\ \tau \geq 0$;

equivalently, $X(t_1),\ X(t_2) - X(t_1),\ X(t_3) - X(t_2) \ldots$ independent $\forall\, t_1 \leq t_2 \leq \cdots$

(*Technical*) (3) X is continuous in probability: $\lim_{t \to 0} P\{|X_t| > \epsilon\} = 0, \quad \forall \epsilon > 0$.

(*Convention*) (4) X has sample paths that are **R**ight-**C**ontinuous with **L**eft **L**imits (RCLL).



The Distribution of a Levy Process. (Probabilistic Characterization.)

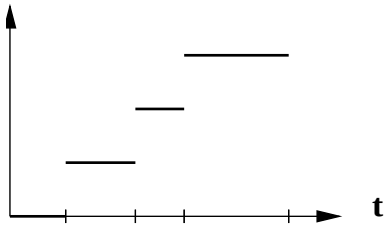
The *finite-dimensional distributions* are determined by marginals:

$$\begin{array}{lllll} X(t_1),\, X(t_2),\, X(t_3), \ldots \Leftrightarrow & X(t_3) - X(t_2), & X(t_2) - X(t_1), & X(t_1) - X(0), \ldots & \text{independent} \\ & X(t_3 - t_2) & X(t_2 - t_1) & X(t_1), \ldots & \text{stationary} \end{array}$$

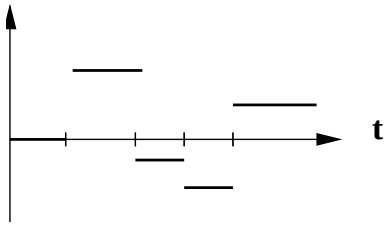
In fact, they are determined by $X(1)$!

Modeller's Dream (from "qualitative" to "quantitative")

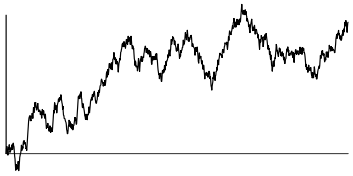
1. A Levy *counting* process is *Poisson*
   (Cinlar, pg. 71)



2. A Levy *jump* process is *Compound Poisson*
   (Cinlar pg. 91)
   changes state in jumps and jumps finitely
   in finite times.



3. A Levy *continuous* process is *Brownian Motion*
   (Breiman pg. 248)
   has continuous sample paths.



The "emergence" of the parameters:

Suppose $\exists\ m(t) = EX(t),\ \ t \geq 0$. Then

$$\begin{aligned} m(s+t) &= E[X(t+s) - X(t)] + EX(t) = m(s) + m(t), \quad \forall\, s,t \geq 0 \\ \Rightarrow m(t) &= \mu \cdot t \ \text{ for some } \ \mu\ . \end{aligned}$$

Suppose $\exists\ V(t) = \text{Var}\ X(t),\ t \geq 0$. Then

$$V(s+t) = V(s) + V(t),\ \forall\, s,t \Rightarrow V(t) = \sigma^2 t, \ \text{ for some } \ \sigma \geq 0\ .$$

Final Practical Characterizations

- *Poisson* process with parameter λ (Poisson(λ)): Levy and Counting;
  $X_t \stackrel{d}{=}$ Poisson $(\lambda t),\ \ t \geq 0$.
- *Compound Poisson*: $X_t = \sum_{k=1}^{A_t} \Delta_k, \quad t \geq 0$, where

  $A = \{A_t,\ t \geq 0\}$ is Poisson (λ); $\Delta = \{\Delta_1, \Delta_2, \ldots\}$ iid (distribution F); A and Δ independent.

- *Brownian* motion, with parameters μ, σ^2 (BM(μ, σ^2)): Levy continuous sample paths; $X_t \sim N(\mu t, \sigma^2 t),\ t \geq 0$.

  $\mu = 0,\ \sigma = 1 \Rightarrow$ *standard* BM (SBM).

  $X \stackrel{d}{=} \text{BM}(\mu, \sigma^2) \Rightarrow X_t = \mu t + \sigma B_t,\ t \geq 0,$ with B = SBM.

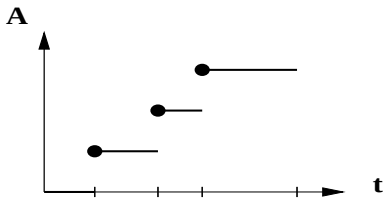
Dynamic Randomness: The **Poisson Process**

Hall, Chapter 3: The *Arrival Process*

Counting Process $A = \{A_t,\ t \geq 0\}$, where A_t = cumulative number of arrivals during $[0, t]$.

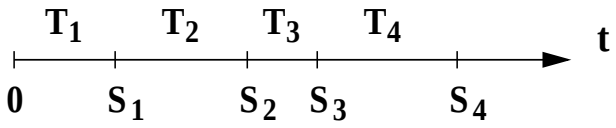
Assume: $A_0 = 0$; a single arrival at a time.

Characterization via sample paths of A:



or via times of *events* = jumps $S_1, S_2, S_3, \ldots$

or via *inter-arrival times* $T_1, T_2, \ldots$: $S_n = T_1 + \cdots + T_n,\ n \geq 1$.



- Completely *deterministic* arrivals at a constant rate λ : $T_n \equiv \frac{1}{\lambda}$.
- Completely *random* arrivals at a constant rate λ : ?

Today: a mathematical *model* for completely random arrivals at a *constant* rate.
(Later: *varying rates*.)

Contents

- Mathematical Framework: Levy Processes;
- Constructions:

  Intuitive (via Bernoulli ⇒ Poisson);

  Explanatory (via "must" properties: order-statistics);

  Axiomatic (Levy + counting);

  Practical (exponential interarrivals).

- Properties; PASTA; Biased-sampling & paradoxes.
- Inference & simulation.

Hall, Chapter 3: **The Arrival Process** $N = \{N(t),\ t \geq 0\}$

§3.1 *Definition* 3.2 requires too much. As discussed, Levy + counting $\Rightarrow$ $\exists \lambda > 0 \ni N(t) - N(s) \sim$ Poisson $[\lambda(t-s)]$.

In particular,

$$\begin{array}{rrl} P\{ & N(t+dt) - N(t) = 1\} & = \lambda dt + o(t) \\ \{ & = 0\} & = 1 - \lambda dt + o(t). \\ \{ & > 1\} & = o(t) \end{array}$$

§3.2 *Derivation* of the Poisson distribution from Bernoulli.

§3.3 *Properties* of the Poisson Process.

1. Poisson *marginals*; number of events in any interval is Poisson;

$$\begin{aligned} EN_t &= \lambda t\,, \text{Var } N_t = \lambda t \\ &\Rightarrow C = \frac{\sigma}{E} = \frac{\sqrt{\lambda t}}{\lambda t} = \frac{1}{\sqrt{\lambda t}} \text{ small for } t \text{ large.} \end{aligned}$$

2. *Interarrival times* which are iid exp (λ).
   Beginning of proof: $P(T_1 \geq t) = P(N_t = 0) = e^{-\lambda t},\ t \geq 0$.
   This is a characterizing property that is practical for simulation.
   Extensions to $T_2, T_3, \ldots,$ and their independence, if rigorous, requires more than the "it should be apparent" in Hall, pg. 58.
3. *Memoryless* property: time till next event does not depend on the elapsed time since the last event.
4. $S_n = T_1 + \cdots + T_n \sim$ Gamma (n, λ) = Erlang.
5. *Order-statistics* property: Given $N(t) = n$, the unordered event times are distributed as n iid r.v., uniformly distributed on $[0, t]$.

   $\Rightarrow$ simulation over $[0,t] : N(t) \sim$ Poisson $(\lambda t);\ U_1, U_2, \ldots, U_{N(t)}$ iid $U[0,t]$.

§3.4 *Goodness* of *Fit*

How well does a Poisson model fit our arrival process?

Qualitative assessments:

Airplanes landing times at a single runway, during an hour:	no
Airplanes landing times at a large airport, during an hour:	plausible
Job candidates that arrive at their appointments during an hour:	no
Visits to a zoo, most of which arrive in groups, during an hour:	no
Arrival times at a bank ATM = **A**utomatic **T**eller **M**achine, during an hour:	plausible

§3.5 *Quantitative Tests*

Graphical Tests:

cumulative arrivals vs. a straight line (Fig. 3.2)

paired successive interarrivals (Fig. 3.4)

exponential interarrivals
(How do you identify exp $(\cdot)$ when you see one? Use Histograms!)

§3.6 *Parameter Estimation*

Estimate $\lambda =$ arrival rate.

MLE (Max. Likelihood Estimator), given $A(t),\ t \leq T : \hat{\lambda} = \frac{A(T)}{T}$.

Confidence intervals for $\frac{1}{\lambda} : \frac{T}{A(T)} \pm z_\alpha \frac{T}{A(T)^{3/2}}$ (3.34)

Sample-size: for $(1-\alpha)$-confidence interval of width w, $N \geq [\frac{2z_\alpha}{w\lambda}]^2$.

Thus, for $w = \epsilon \cdot \frac{1}{\lambda}$, we need $N \geq [\frac{2z_\alpha}{\epsilon}]^2$.
(Eg.: 95%-confidence interval of width = 10% of mean, requires $N \geq [\frac{2\times 1.96}{0.1}]^2 \approx$ 1500!)

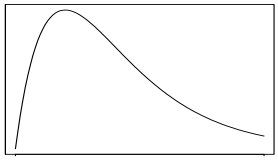
Modelling and Design of Service-Time; **Phase-Type** Distributions.

References.

- Issaev, Eva,: "Fitting phase-type distributions to data from a call center", M.Sc. thesis, Technion IE&M, 2003. (With ample references and a literature review.) http://iew3.technion.ac.il/serveng/References/Thesis.pdf
- Neuts, M.F., *Matrix Geometric Solutions in Stochastic Models*, John Hopkins University Press, 1981.
- Mandelbaum, A. and Reiman, M., "On pooling in queueing networks", Management Science, 44, 971-981, 1996. http://iew3.technion.ac.il/serveng/References/pooling.pdf
- Buzacott, J.A. and Shanthikumar, J., *Stochastic Models of Manufacturing Systems*, pages 63–64, pages 540–541; pages 64–67.

Buzacott and Shanthikumar, on pages 154–155, provide an IE-discussion and references on human task-time, stochastic variability and working rates. Their sources are likely to be manufacturing-based. These are their **key points:**

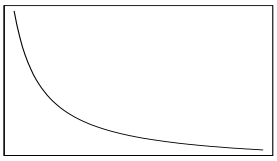
- Much of the variability is beyond control of the operator.
- There exists minimum time of task duration.
- Paced vs. *Unpaced* work. (In Services, it is typically unpaced: no upper bound is imposed on task duration.)
- Typically, for experienced operators service-time distribution is skewed with $P\{T \leq \text{ mean }\} \approx 0.65 \approx 1 - \frac{1}{e}$, which is consistent with an exponential distribution. (For many practical purposes, a distribution with CV ≈ 1 "is exponential".)

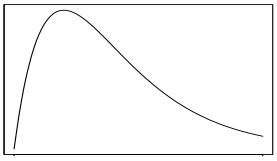


- For inexperienced operators: greater mean and less skewness. (My understanding of this: greater mean and more symmetry.)

- It is not clear from the literature whether there is serial correlation in performing successive tasks.
- Variation of working rate is rarely due to physical fatigue or exhaustion, but rather to inserted-idleness, while the distribution of task-time is stationary over the day.

Personal Experience. Two patterns of service-duration density are prevalent:





One pattern "is exponential" and the other "is lognormal".
It is interesting to understand what does the shape depend on: is it experience-related? job-related? Research is needed to answer such questions.

Examples: Service (Process) Design

- **Pooling Resources** without changing the service process. (At the City Hall of Haifa, moving the Treasury Department to a new location gave rise to phase-type service durations, and motivated M. and Reiman.)
- **IVR/VRU:** Design of search protocols (Comverse).
- **Phone Scripts:** Design of phone/chat services at call/contact centers. (Electric Company).

Contents

- Design of a service system (service time): Pooling.
- What is Service Time?
  - Single- vs. multiple-visits.
  - Time- and State-dependency.
  - Sample Size
  - Estimation and Prediction, Workload.
  - Production of Health: Hernia, Even-Doctors-Can-Manage.
  - After-service work: managing accessibility.
  - Averages do not tell the whole story: the need for the distribution.
  - Service Time is a Statistical Distribution: for example, Log-Normal, Exponential, Mixture.
  - Heterogeneity of Servers.
- Stochastic Ordering (of distributions).
- Exponential Service Times in Human Services:
  - How does one recognize an exponential distribution in a histogram.
  - mean = standard deviation ($CV = 1$) sometimes suffices, but *not* always. (For example, in the QED regime things are yet unclear).
  - Meta Theorem: Durations of human *homogeneous* services are either exponential or lognormal.
- Service Time is a Process: Phase-type models natural and useful.
- Beyond CV's: Some subtle effects of the service-time distribution in the QED regime.

On Pooling in Queueing Networks

Avishai Mandelbaum
Faculty of Industrial Engineering and Management
Technion
Haifa, Israel

Martin I. Reiman
Bell Labs, Lucent Technologies
Murray Hill, New Jersey 07974

February 18, 1996
Revised: October 24, 1996; May 12, 1997; May 4, 1998

Abstract

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.

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

Two Local Municipalities

"Theorem" : **Durations of human homogeneous services "are" exponential**
"Proof": **Empirical (see below); Theoretical (phase-type dense); Scientific?**

Department	Station No.	Total Customers	Avg. Service Time (Mins)	STD (Mins)	Utilization	Maximal Service Time (Mins)
Collection - Reception	1	370	7.55 ± 0.68	7.96	37%	79.32
	2	951	5.42 ± 0.33	6.27	68%	105.20
	3	510	6.51 ± 0.50	6.94	44%	63.33
	4	377	8.41 ± 0.75	8.90	42%	58.15
Collection - Immigrants	5	493	11.59 ± 0.80	10.88	76%	74.60
	6	569	10.38 ± 0.62	8.98	78%	50.87
Collection - Back office	7	114	10.80 ± 1.98	12.82	16%	93.73
	8	28	9.07 ± 3.56	11.50	3%	52.07
	9	47	18.32 ± 4.90	20.34	10%	113.57
	10	28	23.39 ± 5.52	17.75	9%	63.77
	11	59	11.99 ± 3.16	14.75	9%	70.30
	12	128	16.73 ± 2.34	16.08	28%	88.68
Cashier	13	1460	2.51 ± 0.21	4.92	48%	52.18
	14	1416	3.86 ± 0.18	4.16	72%	46.92
Billing - Reception	15	340	13.74 ± 1.07	12.02	62%	63.68
	16	363	10.88 ± 0.92	10.60	52%	87.92
	17	473	6.66 ± 0.50	6.68	42%	49.93
	18	302	11.22 ± 1.30	13.81	45%	100.60
Billing - Back office	19	34	19.29 ± 5.64	19.99	8%	78.27
	20	13	12.20 ± 3.86	8.47	3%	29.28
Total (1 month)		8075				

Water	1	57	7.80 ± 1.70	7.61	6.5%	31.28
	2	130	9.34 ± 1.20	8.37	19.3%	54.68
Tellers	3	336	9.04 ± 0.80	8.93	48.2%	49.05
	4	208	9.93 ± 1.00	8.82	33.0%	49.12
	5	417	8.97 ± 0.70	8.55	59.4%	49.37
	6	144	9.53 ± 1.20	8.75	21.8%	41.70
	7	156	8.03 ± 1.10	7.96	19.8%	35.27
	8	67	3.74 ± 0.70	3.58	4.0%	21.03
Cashier	9	757	6.64 ± 0.40	6.94	79.7%	29.95
Manager	10	190	1.99 ± 1.00	8.44	24.1%	38.97
Discounts	11	317	4.59 ± 0.40	4.54	23.1%	36.72
Total (1 month)		2779				

* Service time ranges given with 90% confidence

"STD = Mean" is what often (but not always) "counts" towards Exponentiality.

Local Municipalities

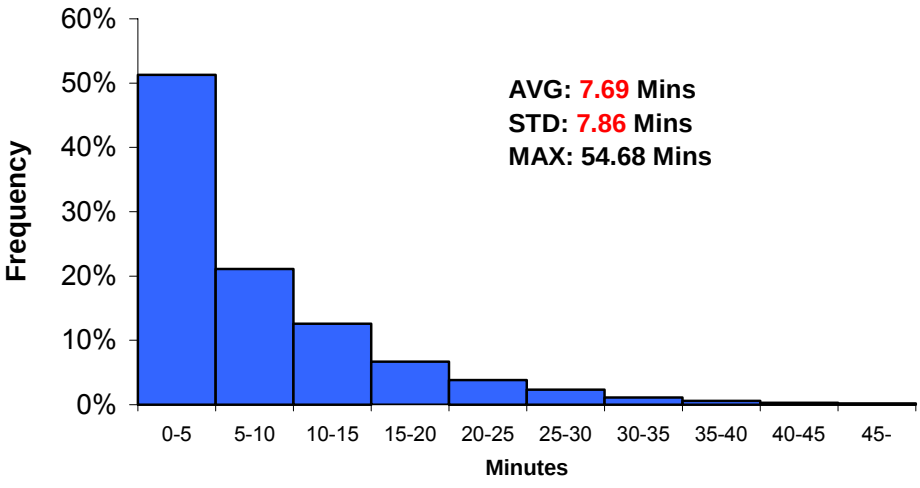
Department	Station No.	Total Customers	Avg. Arrival Rate (1/Hr)	Avg. Service Time (Mins)	STD (Mins)	Maximal Service Time (Mins)	Utilization	Avg. Waiting Time (Mins)
Water	N/A	187	1.8 ± 0.2	**8.87** ± 1.0	**8.15**	54.68	13.3%	4.76
Tellers	N/A	1328	12.6 ± 0.5	**8.82** ± 0.4	**8.55**	49.37	30.8%	7.73
Cashier	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



Government Office

Hour	Frequency	Service Time		Waiting Time		Waiting / Service Ratio	Avg. Time In System
		Avg.	STD	AVG.	STD.		
8	421	**2.7**	**2.7**	17.1	14.1	6.5	19.8
9	404	**2.8**	**2.5**	30.7	25.9	11.1	33.4
10	349	**2.9**	**2.7**	29.4	30.7	10.1	32.3
11	218	**2.9**	**2.7**	26.6	32.8	9.1	29.5
12	317	**2.2**	**2.5**	23.3	20.3	10.6	25.5
13	255	**2.6**	**2.6**	38.7	31.8	14.9	41.3
14	244	**2.3**	**2.7**	34.1	38.6	14.5	36.4
15	53	**2.1**	**2.6**	22.2	18.8	10.8	24.3
Total	**2261**	**2.6**	**2.6**	**27.6**	**28.1**	**10.5**	**30.2**

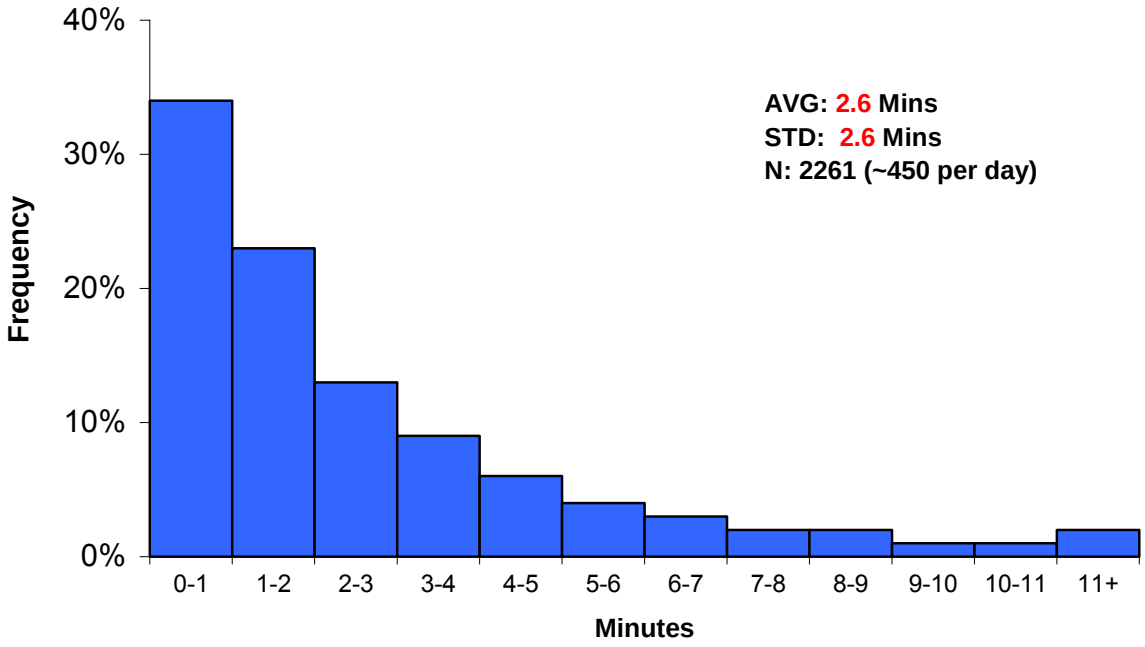
Date	Frequency	Service Time		Waiting Time		Waiting / Service Ratio	Avg. Time In System
		Avg.	STD	AVG.	STD.		
23/10	411	**3.0**	**2.7**	26.4	25.9	8.8	29.4
24/10	470	**2.7**	**2.7**	23.6	23.4	8.9	26.2
25/10	515	**2.3**	**2.5**	29.3	30.3	12.6	31.6
26/10	447	**2.4**	**2.5**	31.9	30.5	13.2	34.3
27/10	418	**2.8**	**2.9**	26.6	29.0	9.4	29.4

Department	Frequency	Service Time		Waiting Time		Waiting / Service Ratio
		Avg.	STD	AVG.	STD.	
Women	444	**2.4**	**2.1**	43.2	30.5	17.7
Income	412	**1.8**	**2.1**	23.7	19.1	13.5
Free Professions	332	**2.9**	**3.0**	7.6	12.4	2.6
Men	279	**2.3**	**2.7**	29.1	34.0	12.7
Released Soldier	239	**2.2**	**2.3**	26.4	21.8	11.9
Registration	201	**3.8**	**2.3**	37.5	26.6	9.8
Disabled People	181	**3.5**	**3.4**	18.4	30.2	5.3
New Immigrants	173	**3.3**	**3.0**	32.6	29.0	9.8

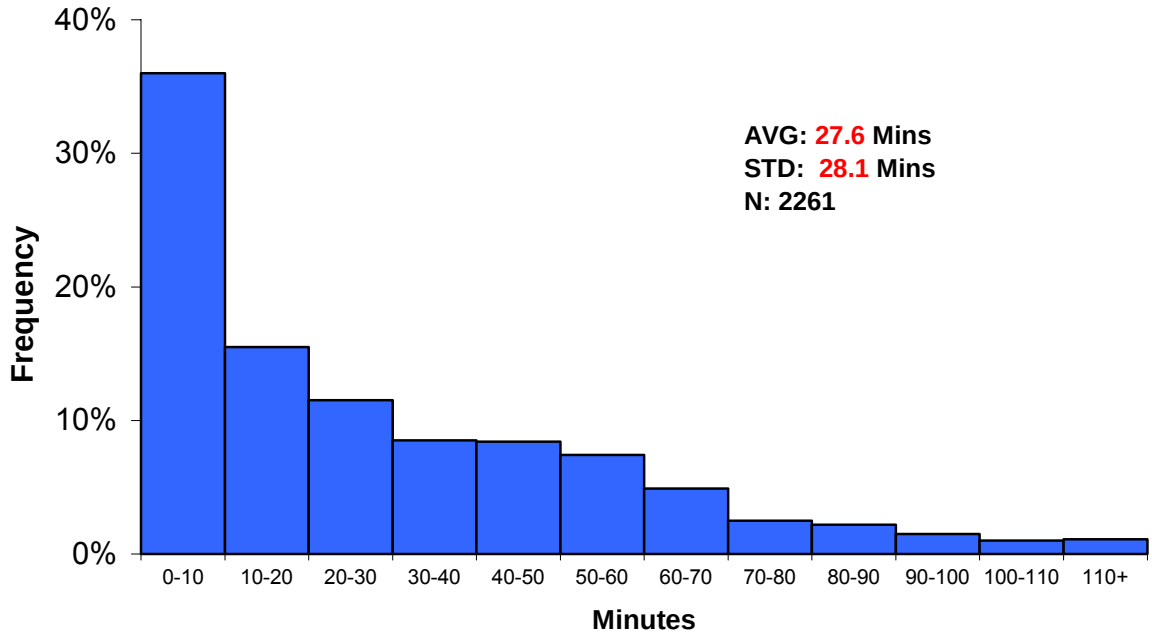
Service Type	Frequency	Service Time		Waiting Time	
		Avg.	STD	AVG.	STD.
Reporting	1746	**2.1**	**2.4**	27.4	27.8
Forwarding	239	**3.3**	**2.8**	14.9	25.6
Registration	223	**3.8**	**2.3**	35.4	27.2
Health Committee	85	**3.2**	**2.8**	22.0	37.8
Endorsements	52	**3.1**	**3.2**	6.8	11.8
Course Offering	28	**2.9**	**3.9**	16.0	18.4
Counseling Offering	26	**3.8**	**3.7**	7.5	13.7
Income Completion	6	**5.8**	**4.9**	20.6	22.1

Government Office - Cont'd

Service Times Histogram:



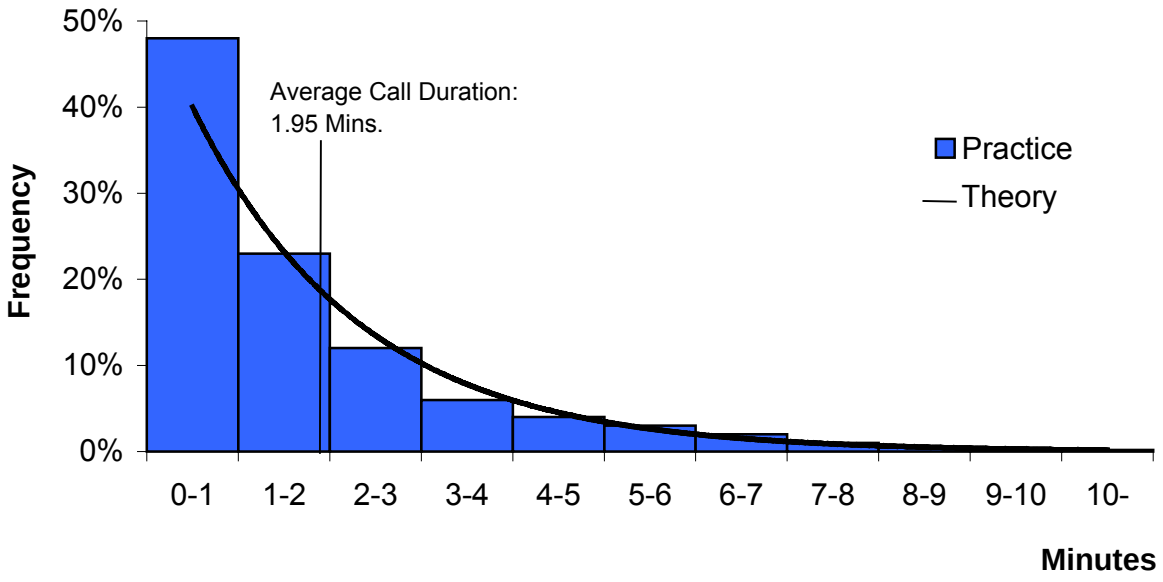
Waiting Times Histogram:



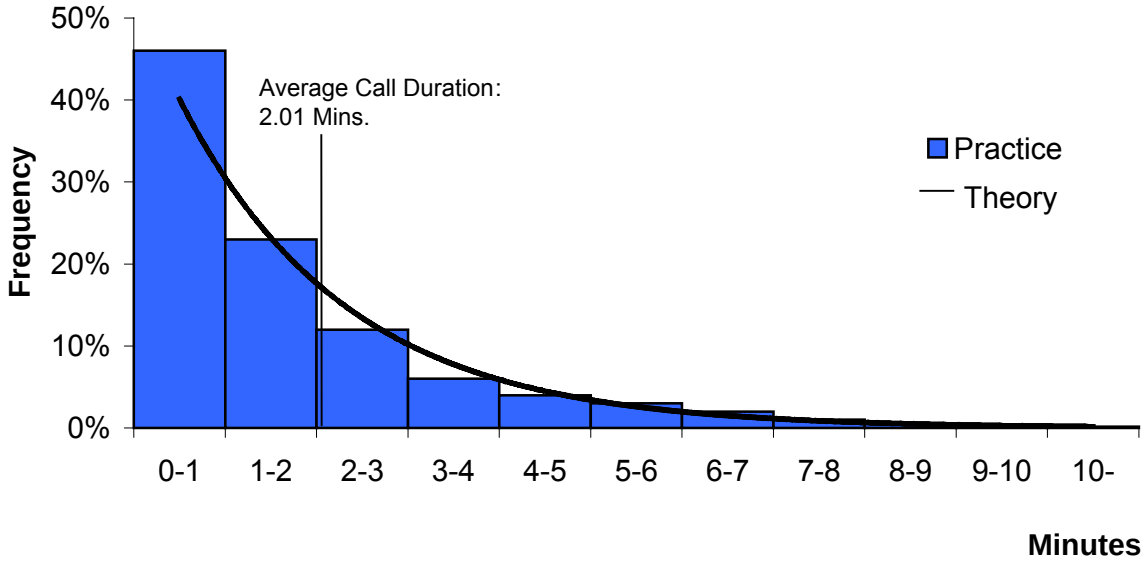
Note: Average sojourn time is 30.2 mins. Hence Service Index = 0.086. Too Low!

Phone Calls: Information

Call-Duration Frequency - North:



Call-Duration Frequency – Central:



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

A. Geometric Approximation.

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

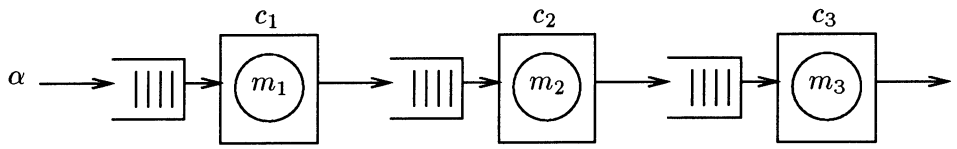


Figure 1: A specialized model with tasks attended by specialized servers.

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 $\sum_k c_k$.

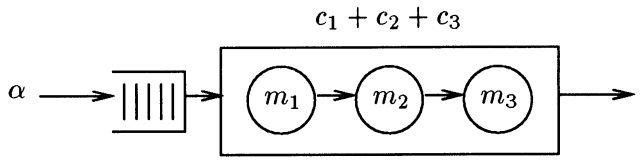


Figure 2: A flexible model with complete pooling into a single queue and a single flexible server.

Customer arrivals are assumed Poisson and task durations exponential. (We comment on these distributional assumptions in the Addendum.) As articulated in Section 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 [19], as in Figure 3, and the flexible (pooled) architecture is modeled by an M/PH/1 system [26], as in Figure 4.

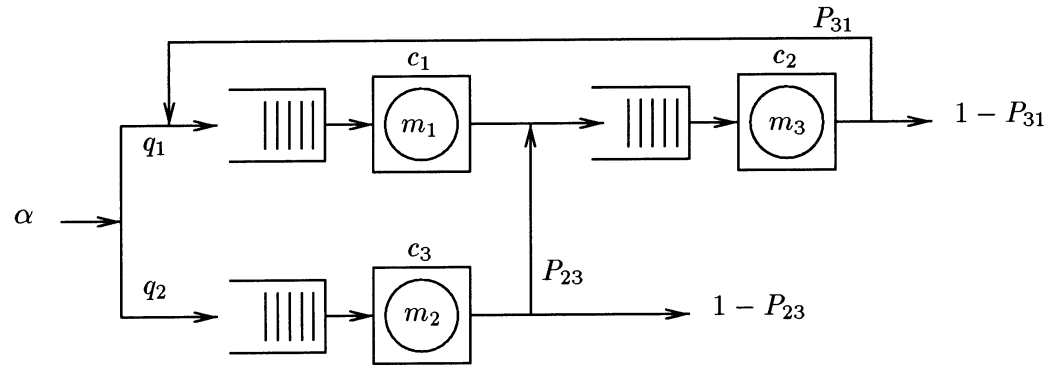


Figure 3: A specialized model with task repetition and feedback.

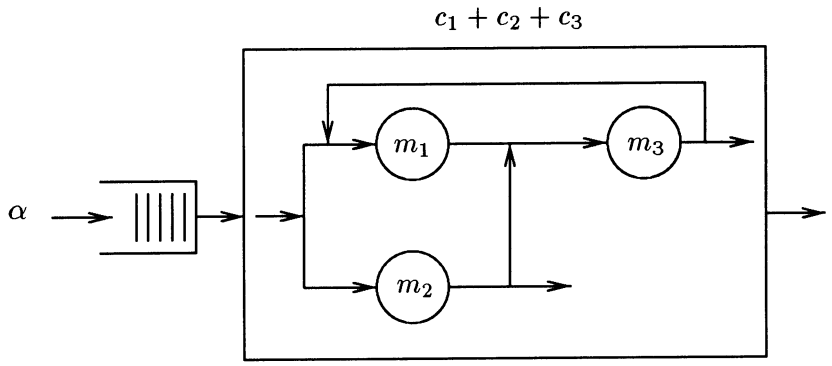


Figure 4: The flexible model, under complete pooling, that corresponds to Figure 3.

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 Section 5.3). Figure 6 depicts partial pooling of only queues and servers 1

The Model

Customers/Tasks:

- Customers arrive in a Poisson process of rate α
- A service is made up of a random sequence of tasks
  - K types of tasks, $k = 1, \ldots, K$
  - Work content in task k exponentially distributed with mean m_k
  - q_k = probability that task k is first
  - P_{jk} = probability that task k immediately follows task j (P is transient)

The Specialized Model

Task k has a server with service capacity c_k (units of work per unit of time) dedicated to it

$$\text{Processing times} \ : \ \text{mean} \ \frac{m_k}{c_k}$$

This yields a *Jackson network*:

K single server stations

Arrival rates $(\alpha q_1, \alpha q_2, \ldots, \alpha q_K)$

Service rates $(c_1/m_1, c_2/m_2, \ldots, c_K/m_K)$

Routing matrix P

The Flexible Model

There is 1 *flexible* server with service capacity

$$\boldsymbol{c}^T\boldsymbol{e} = \sum_{k=1}^{K} c_k$$

This yields an M/PH/1 queue:

K phases

Arrival rate α

Mean phase 'duration' $(m_1/\boldsymbol{c}^T\boldsymbol{e},\ m_2/\boldsymbol{c}^T\boldsymbol{e},\ \ldots,\ m_K/\boldsymbol{c}^T\boldsymbol{e})$

Initial phase probabilities $(q_1,\ q_2,\ \ldots,\ q_K)$

Routing matrix P

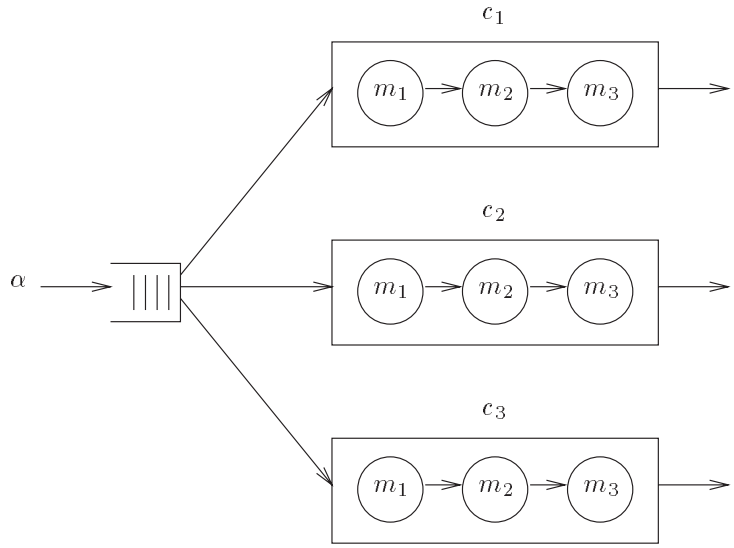


Figure 5: Complete pooling of queues only; servers are made flexible but maintain individual identities.

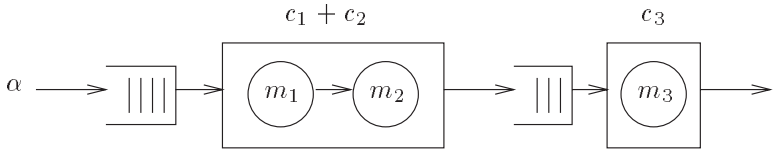


Figure 6: Partial pooling.

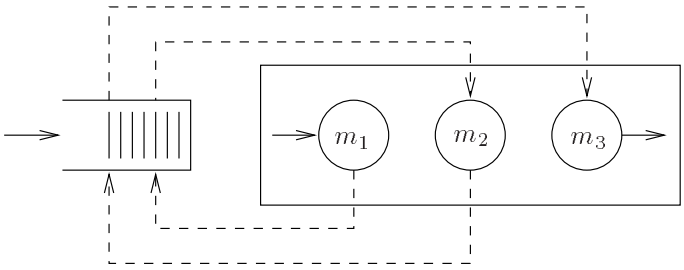


Figure 7: Splitting services. Each task returns to the end of the queue.

and 2 (see Section 5.4). Figure 7 depicts a split of the service so that a customer, upon completion of a task, rejoins the queue (see Section 5.5), and additional designs are possible

Carefull: Recall the Appendix in HW2

Bramson [6] chose first $\frac{399}{400} \leq d < 1$, then K large enough for $d^{K-2} < 1/50$, and finally δ small enough so that $0 \leq \delta < (1-d)/50(K-2)^2$. The specialized network is, therefore, stable ($\rho_k^s < 1,\ 1 \leq k \leq K$) and its (complete) pooling, as in Subsection 5.1, is advantageous.

We consider now two (related) poolings. In the first, depicted in Figure 8, the K servers are pooled into 3 servers as follows: server 1 attends to tasks 1 and K; server 2 serves

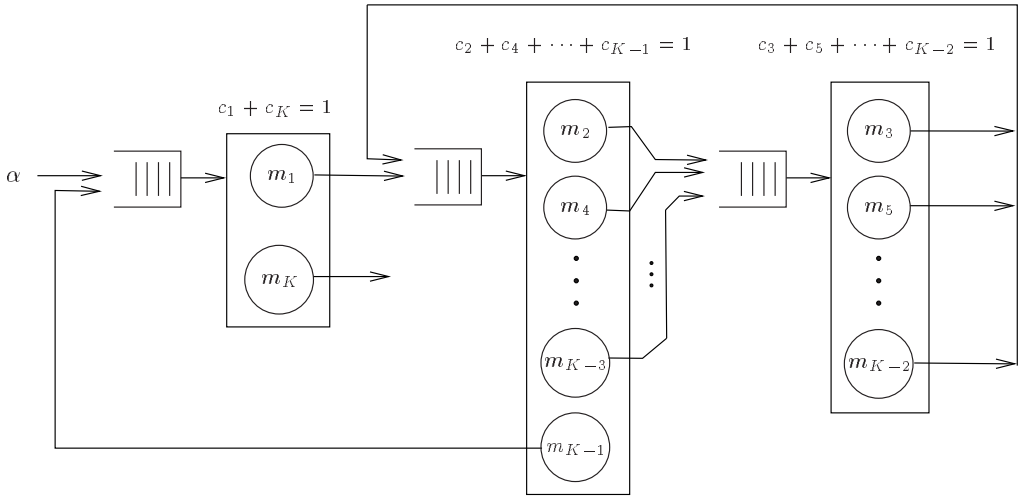
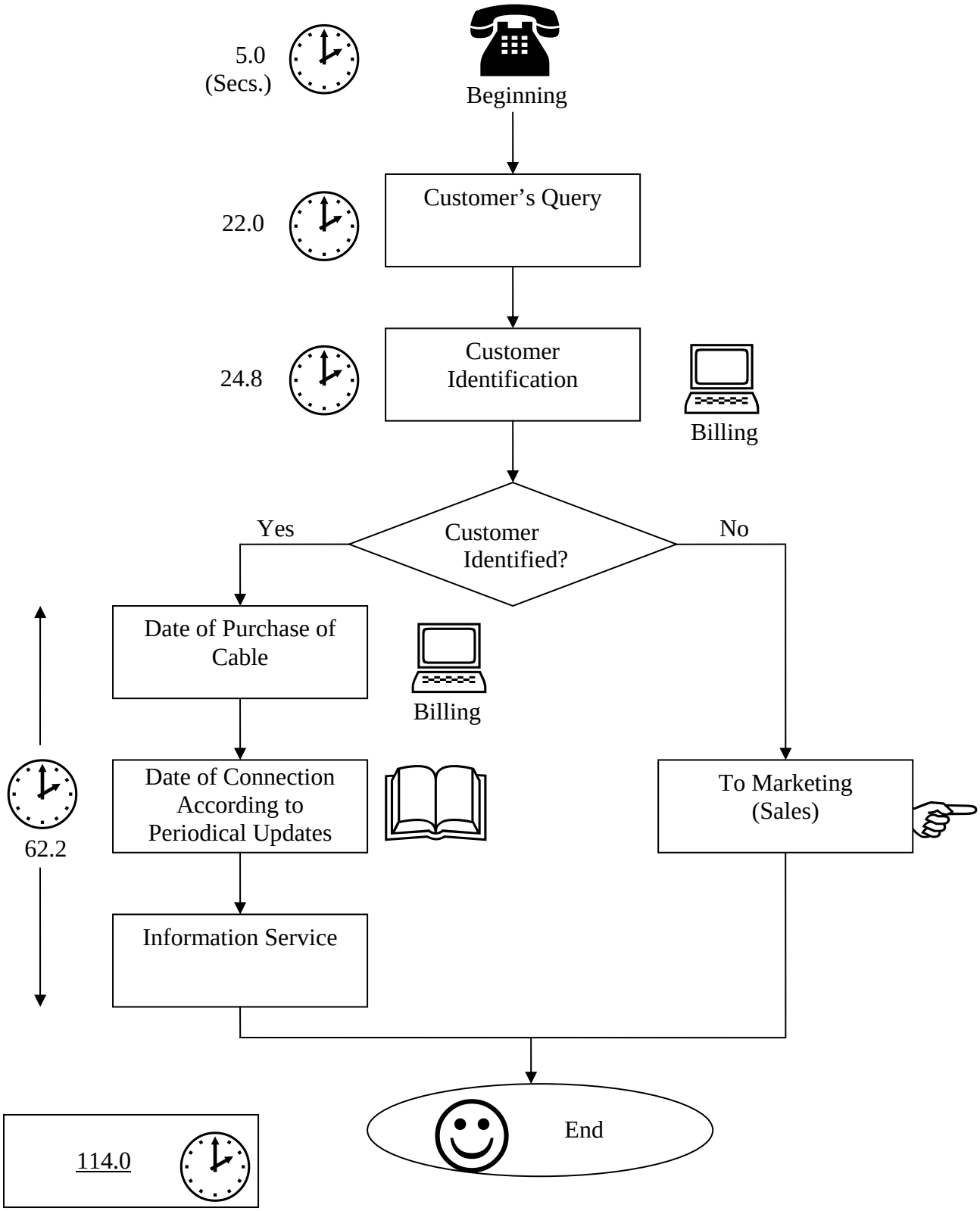


Figure 8: Bramson's unstable network obtained by partial pooling.

tasks $2, 4, \ldots, K-1$; server 3 cares for tasks $3, 5, \ldots, K-2$. Thus, a customer starts with server 1, moves on to 2, then 3, back to 2, and so on, until service $K-1$ at server 2, then the last service back at 1 and finally out. Each server uses the FIFO discipline, under which Bramson [6] proved that the network is unstable. (See his comment, immediately following the statement of Theorem 1.) In particular, with probability 1, the sojourn time of customers increases to infinity, as $t \uparrow \infty$. Instability arises because the system roughly alternates between busy periods of server 2, attending mainly to incoming tasks 2 while starving server 1, and busy periods of server 1, attending to tasks K while starving server 2. The starvation of both servers is a consequence of FIFO, under which ample δ-tasks are forced into queueing behind few d-tasks. (A more refined and quantitative intuition is provided in [6].)

Service (Process) Design; Phase-Type Service

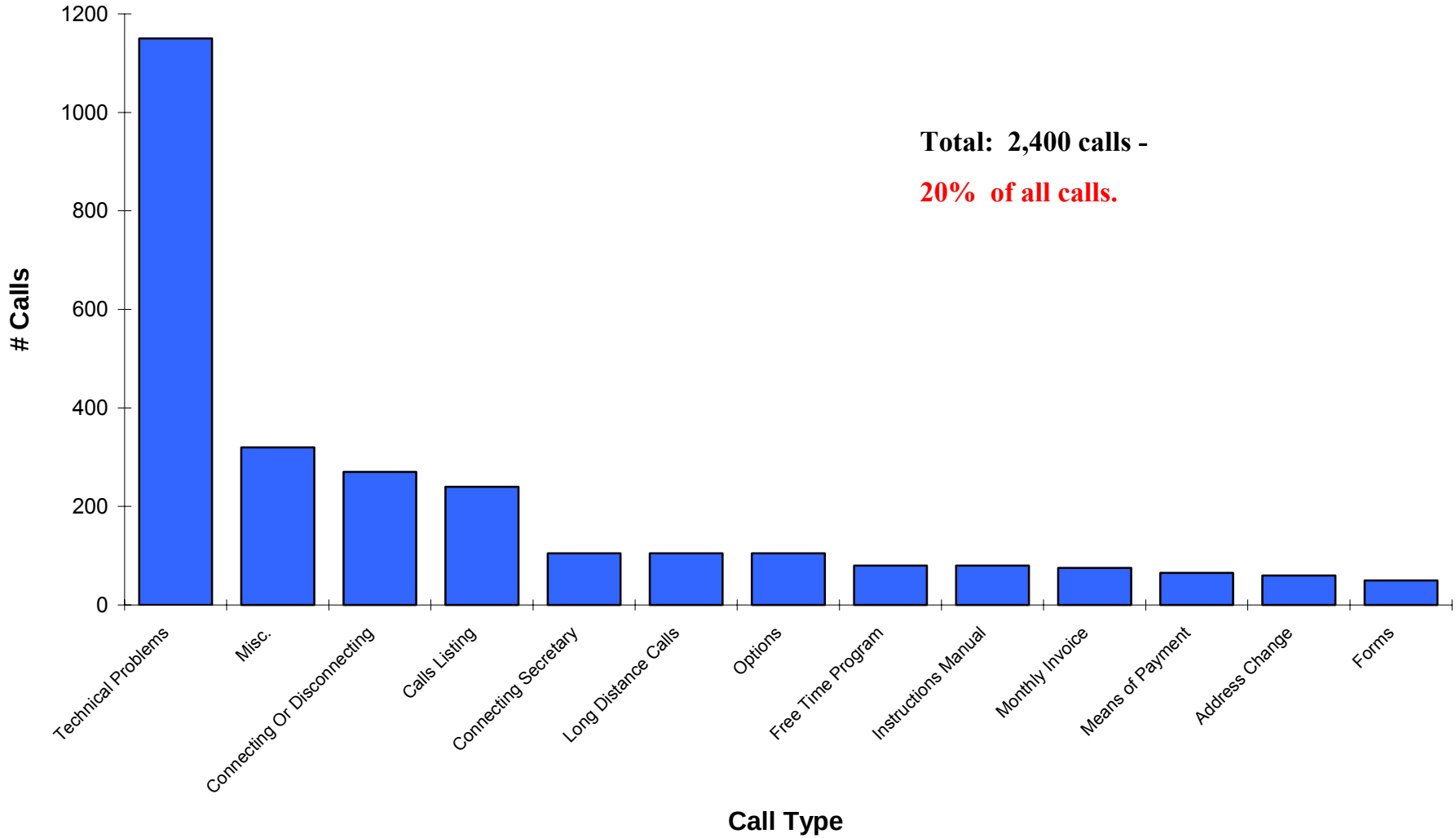
Late Connections



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

What is “Service Time”?

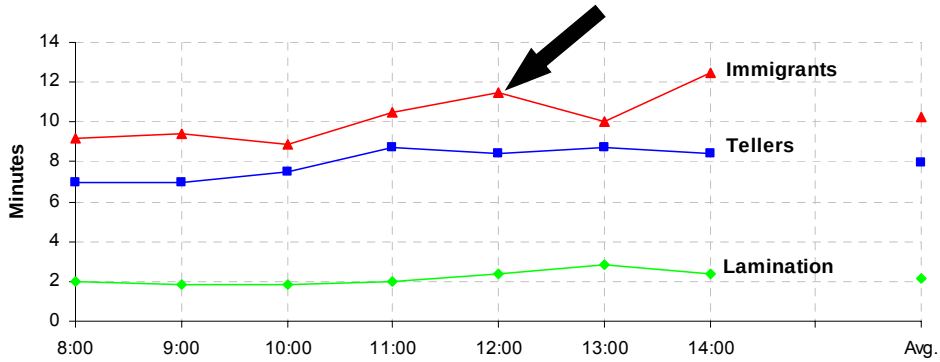
Bank Classification of “Continued – Calls”



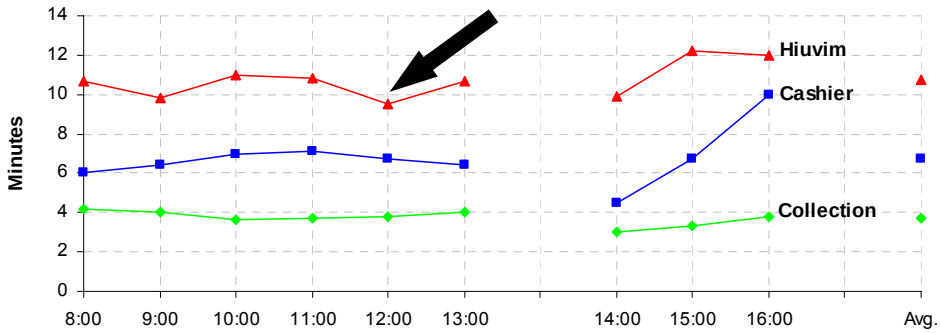
Total: 2,400 calls -

20% of all calls.

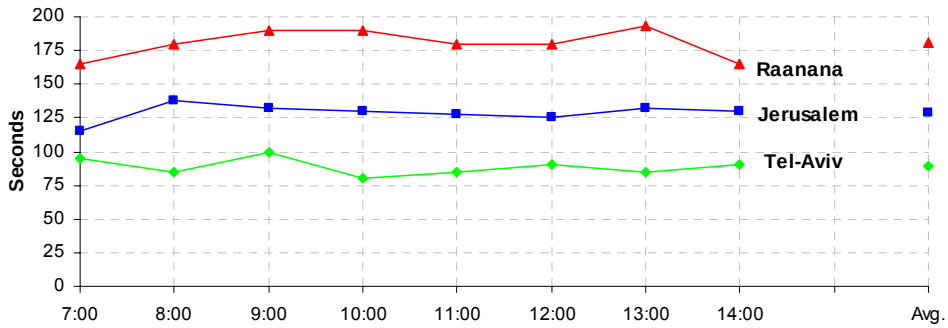
Average Service Durations Over The Day



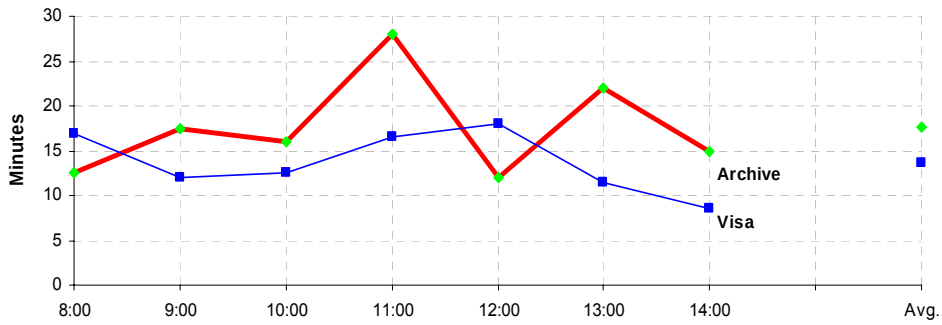
Time or State Dependent ?



3 Patterns



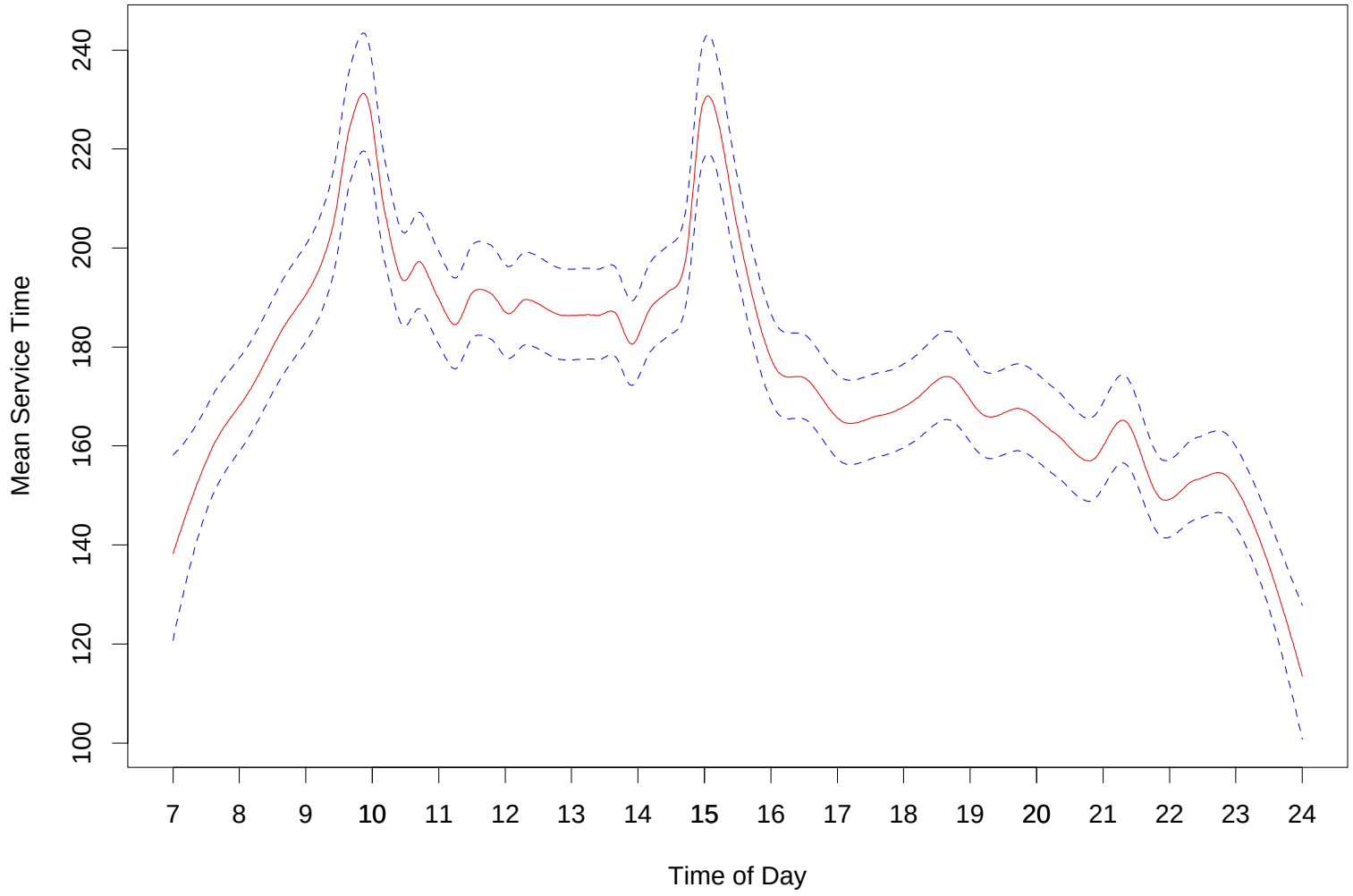
3 Branches Provide the Same Tele-Service



? Sample Size !

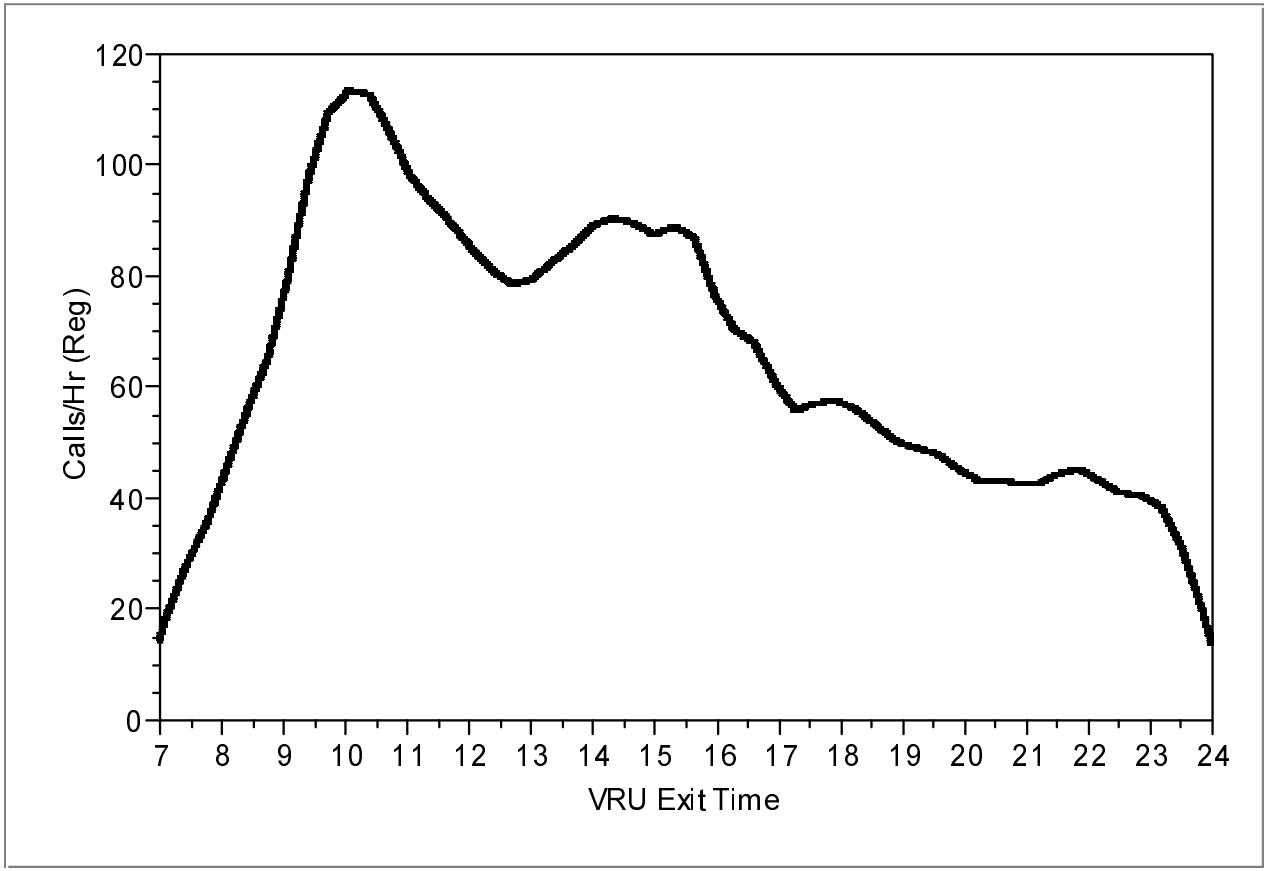
“Fluid” – view, but …

Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) ($n = 42613$)



Arrivals: Inhomogeneous Poisson

Figure 1: Arrivals (to queue or service) – "Regular" Calls

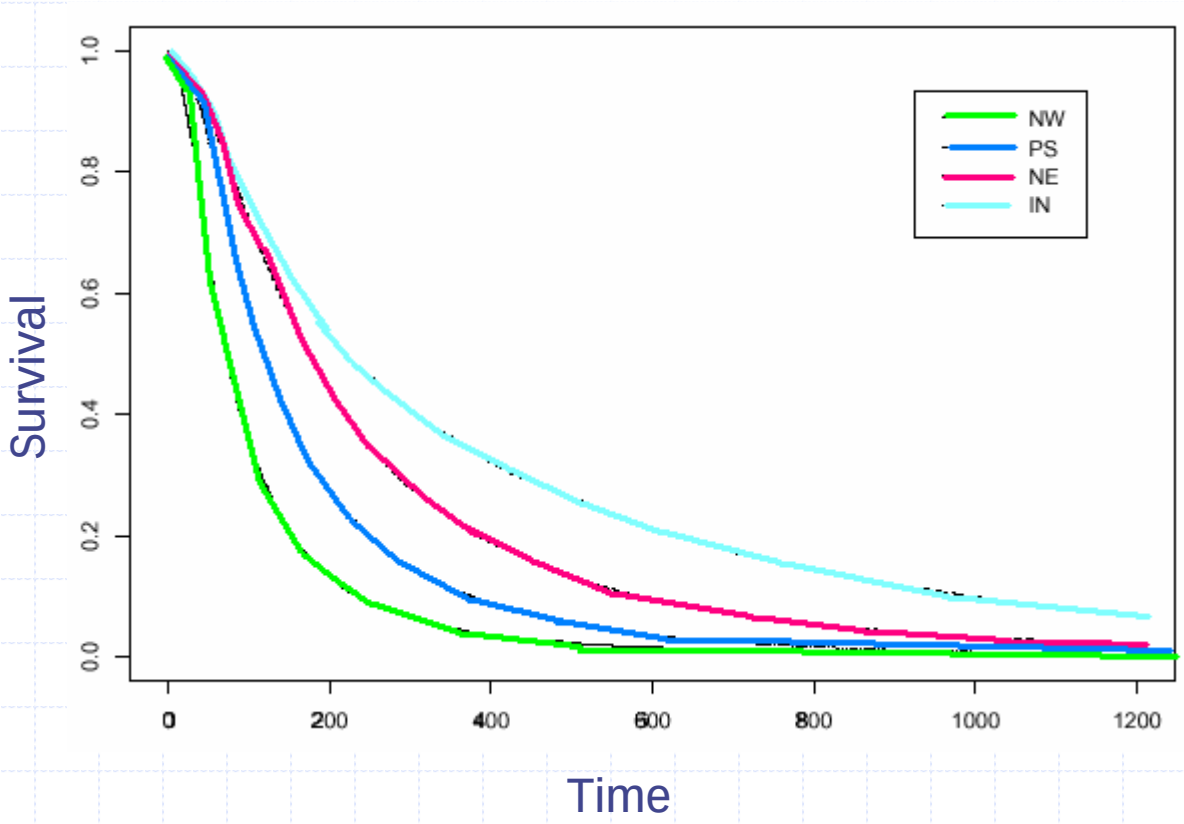


Service Time

	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

Service Time

Survival curve, by Types



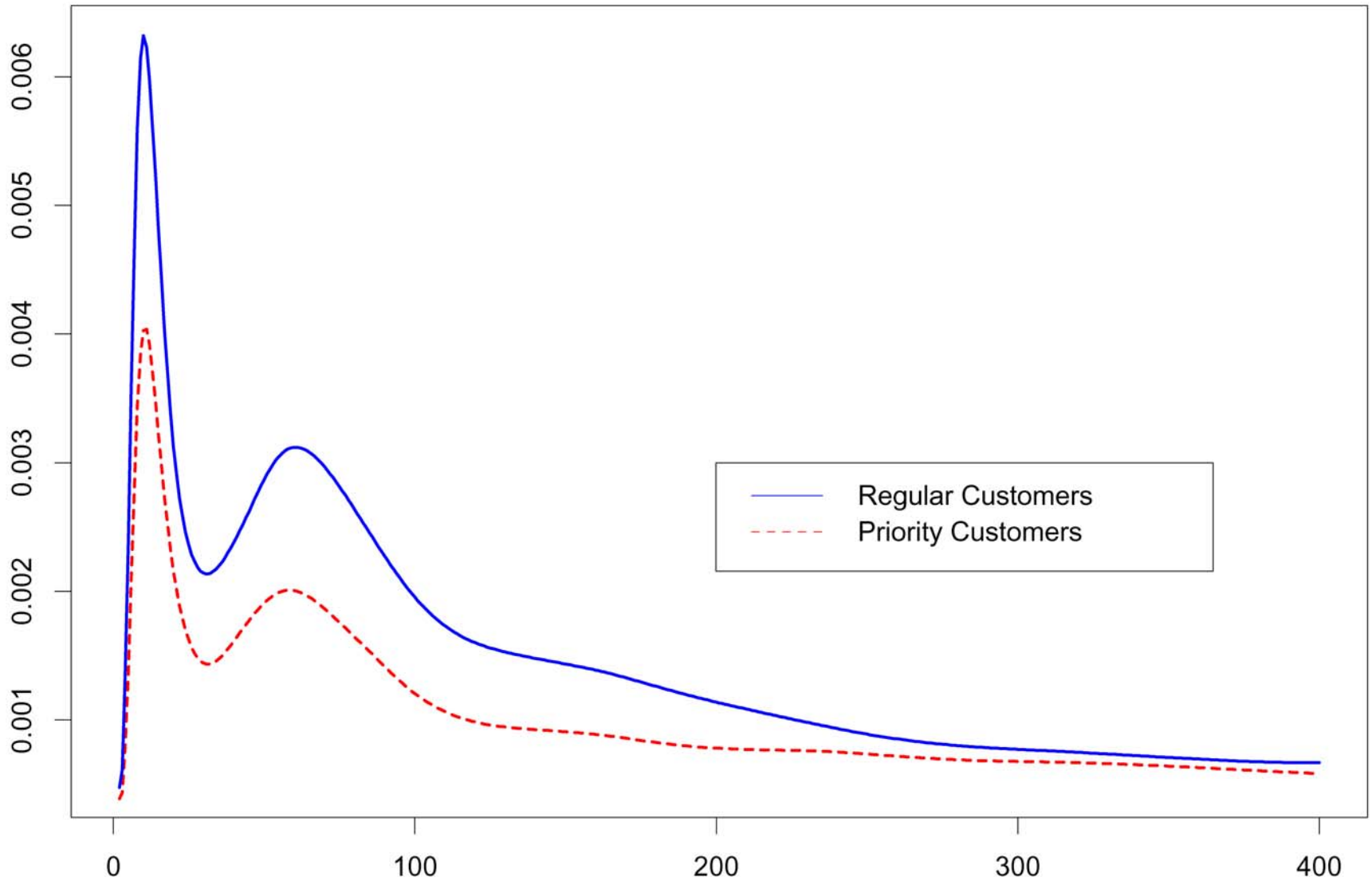
Means (In Seconds)

NW (New) = 111

PS (Regular) = 181

NE (Stocks) = 269

IN (Internet) = 381


Hazard Rate: Empirical (Im)Patience
0.001
0.002
0.003
0.004
0.005
0.006
0
100
200
300
400
Regular Customers
Priority Customers

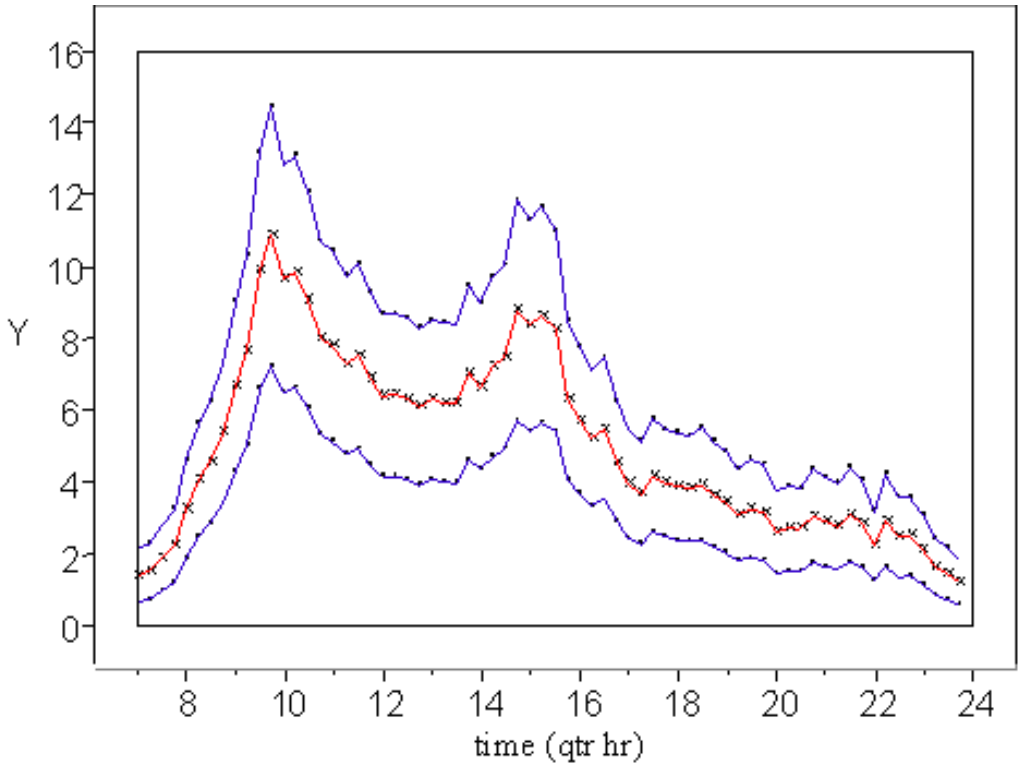
Workload

Suppose at time t, the arrival rate is $\Lambda(t)$ and the mean service time is $\nu(t)$, then the workload at time t is defined as

$$L(t) = \Lambda(t)\nu(t).$$

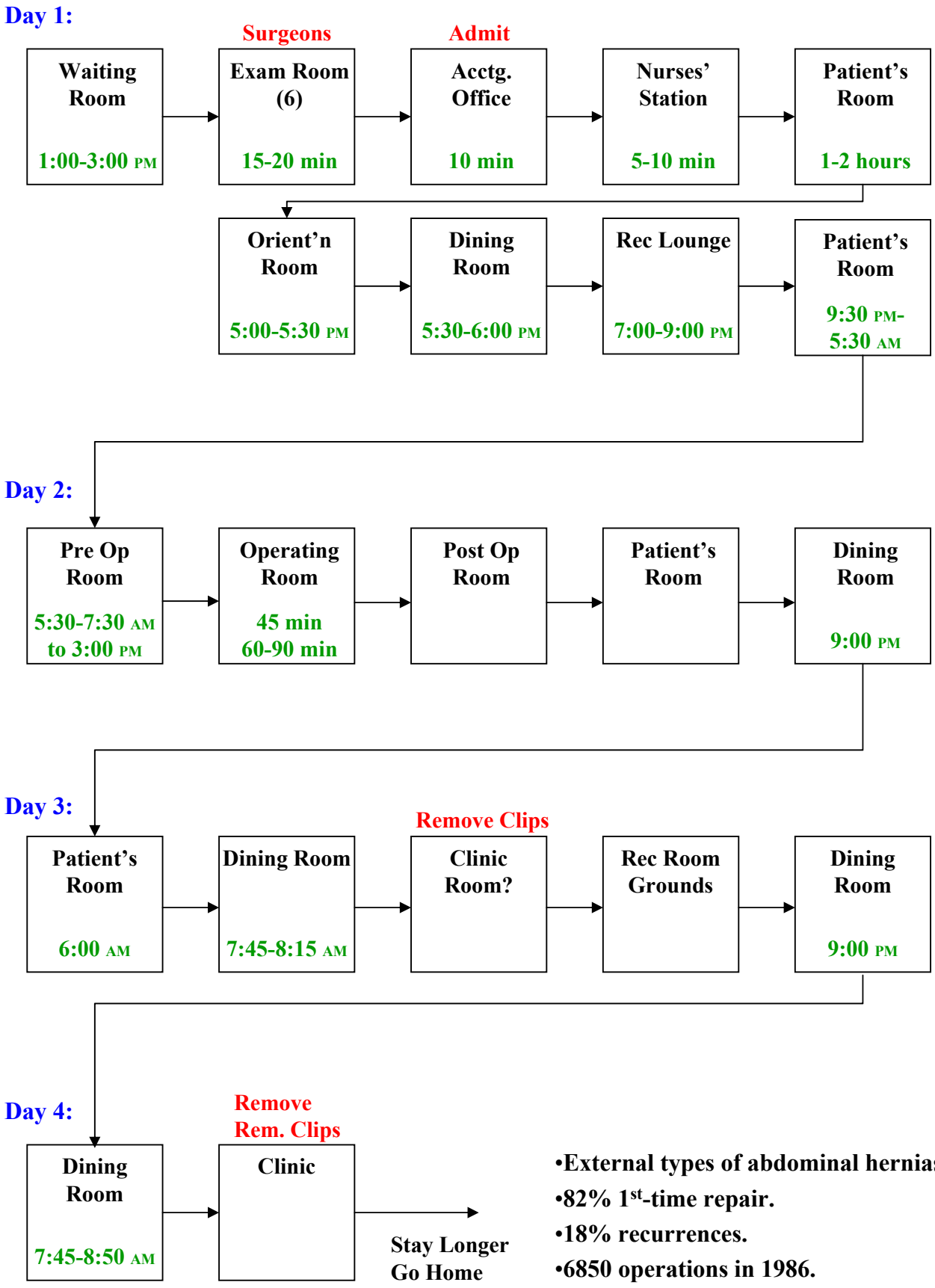
- the expected time units of work arriving per unit of time.
- primitive quantity in building classical queueing models and setting staffing levels.

Figure 18: 95% prediction intervals for the load, L, following a day with $V_+ = 340$.



Units on vertical axis are "required agents".

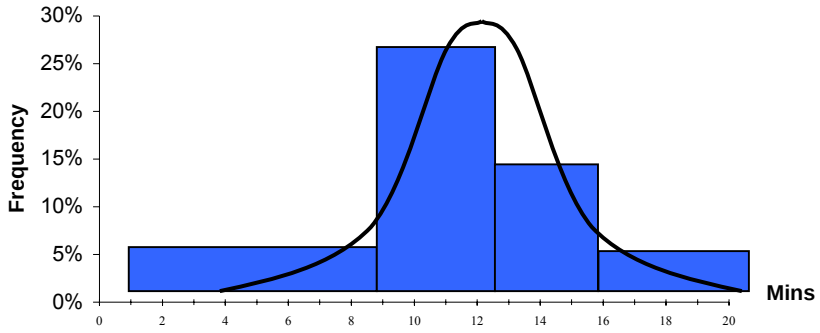
Shouldice Hospital: Flow Chart of Patients' Experience



- External types of abdominal hernias.
- 82% 1st-time repair.
- 18% recurrences.
- 6850 operations in 1986.
- **Recurrence rate: 0.8% vs. 10% Industry Standard.**

Ambulatory Operations Time
Production of Health

Cystoscopy:

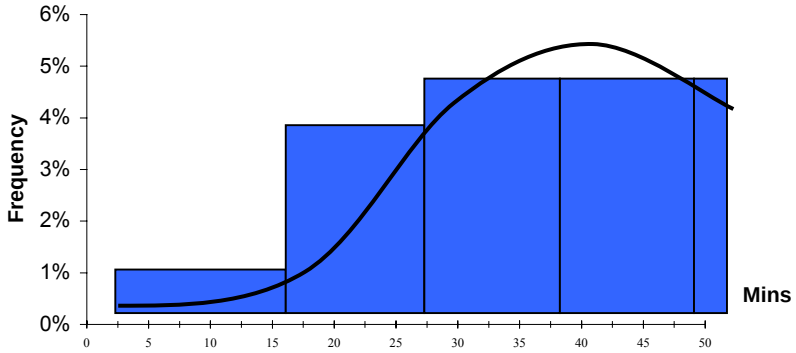


Practice
Theory

AVG: 11.33 Mins.
STD: 2.83 Mins.
N: 48

$4.1057 = \chi^2 < \chi^2_{0.95} = 5.991$ => Do not Reject

TURT / TURP:

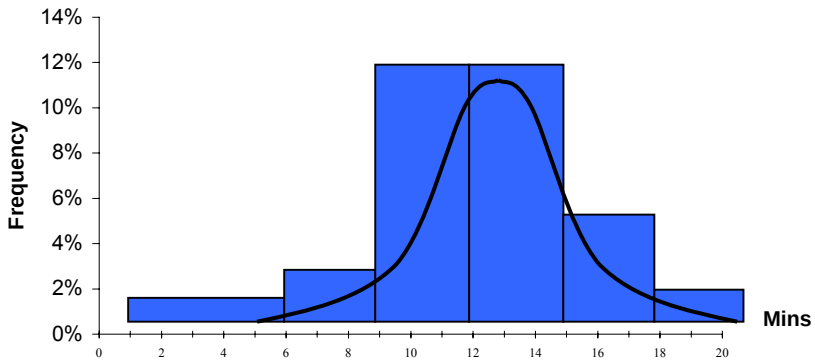


Practice
Theory

AVG: 37.1 Mins.
STD: 14.41 Mins.
N: 20

$0.5113 = \chi^2 < \chi^2_{0.95} = 5.991$ => Do not Reject

Curettage:



Practice
Theory

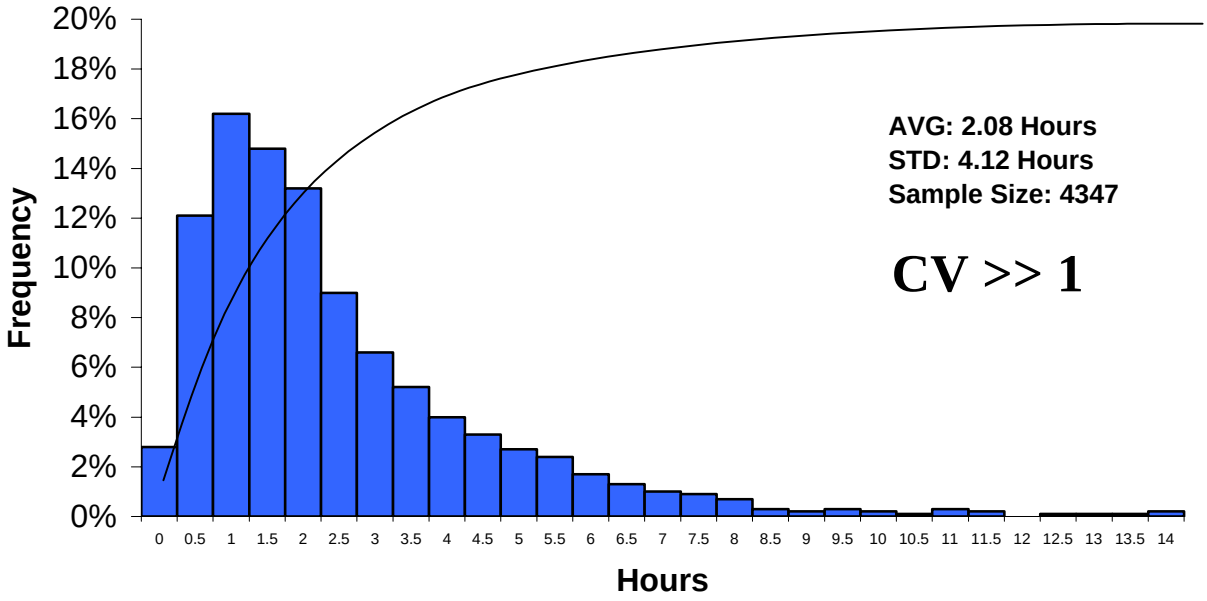
AVG: 12.08 Mins.
STD: 3.08 Mins.
N: 40

$2.4887 = \chi^2 < \chi^2_{0.95(6-3)} = 7.815$ => Do not Reject

CV << 1

Operations Time In a Hospital

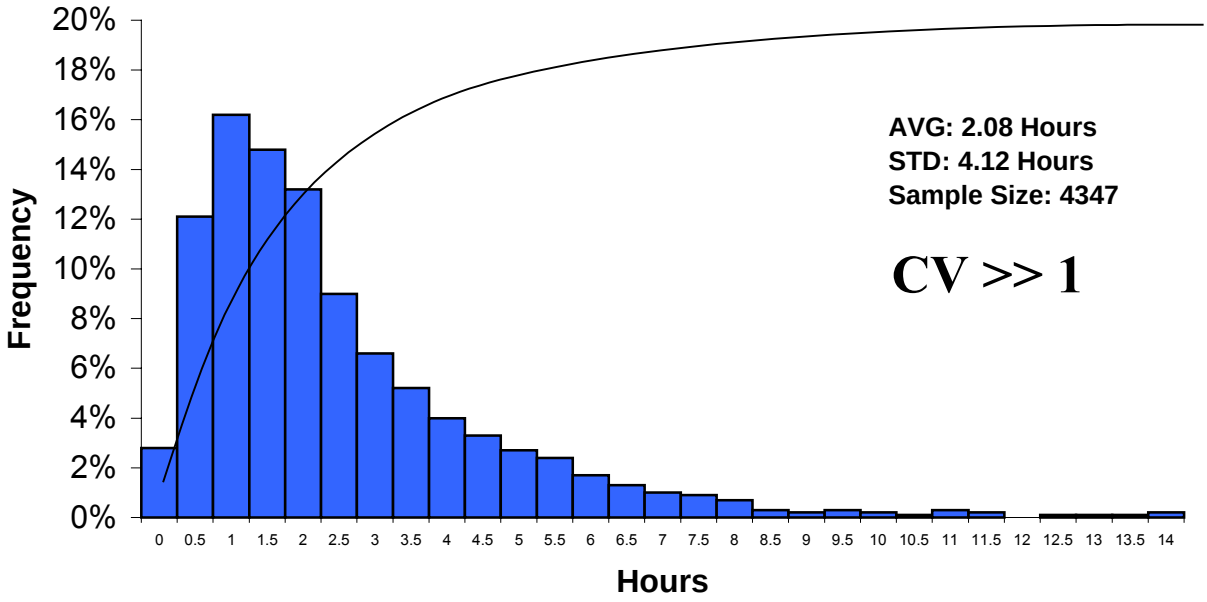
Operations Time Histogram:



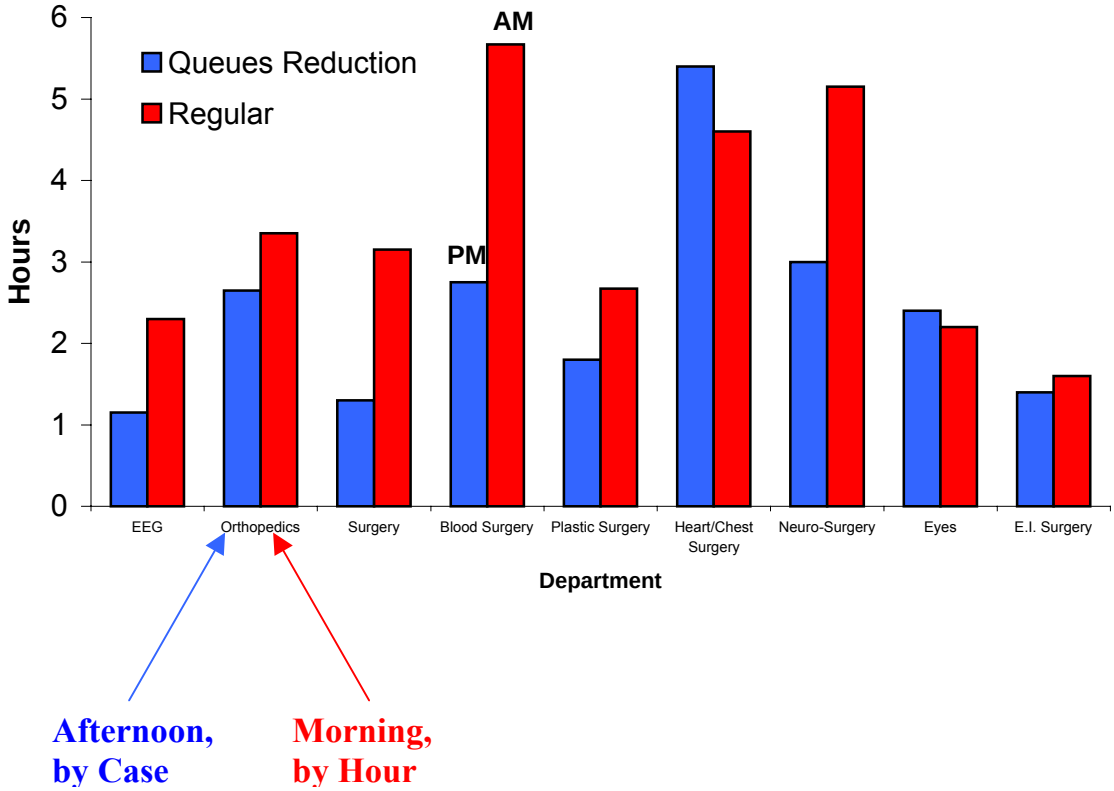
O1 O2 All Clear

Operations Time In a Hospital

Operations Time Histogram:



Operations Time - Morning vs. Afternoon:



Afternoon, by Case

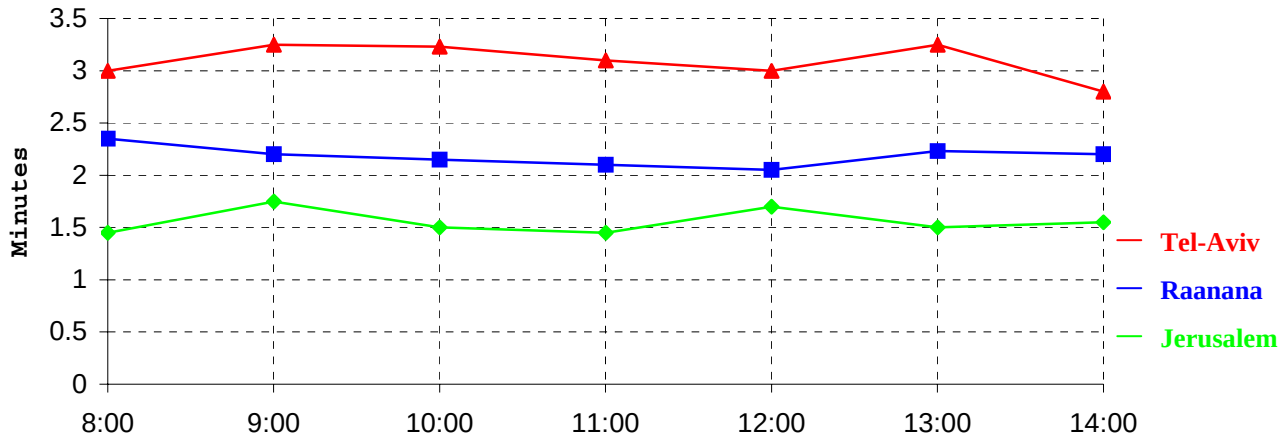
Morning, by Hour

Ethical?

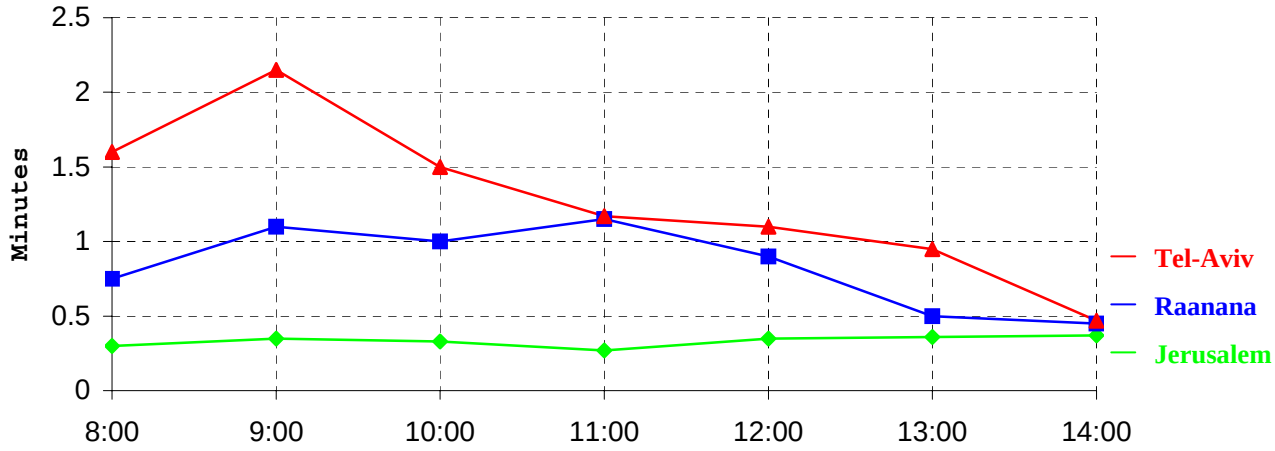
Even Doctors Can Manage!

Service Performance

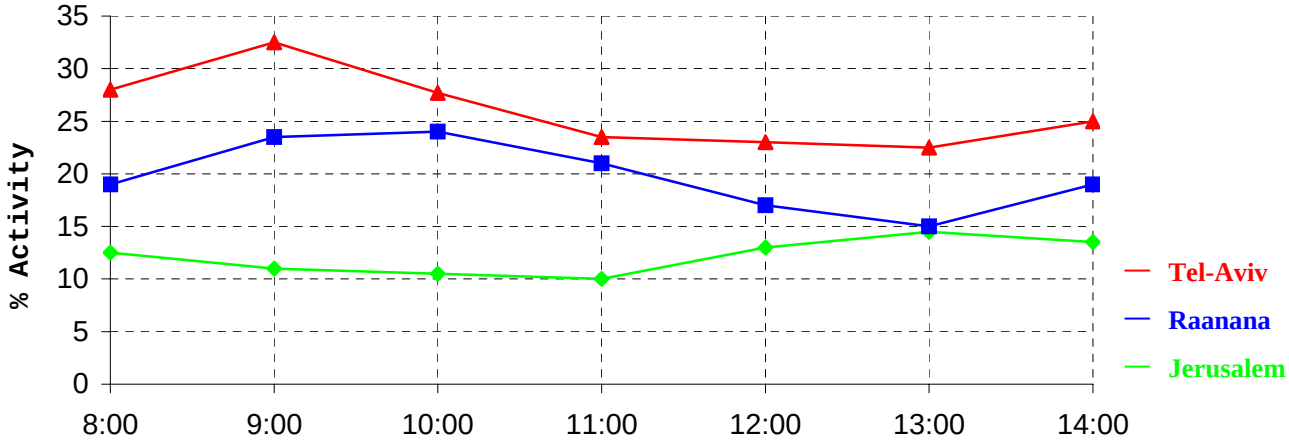
Service Time – Average:



Waiting Time – Average:



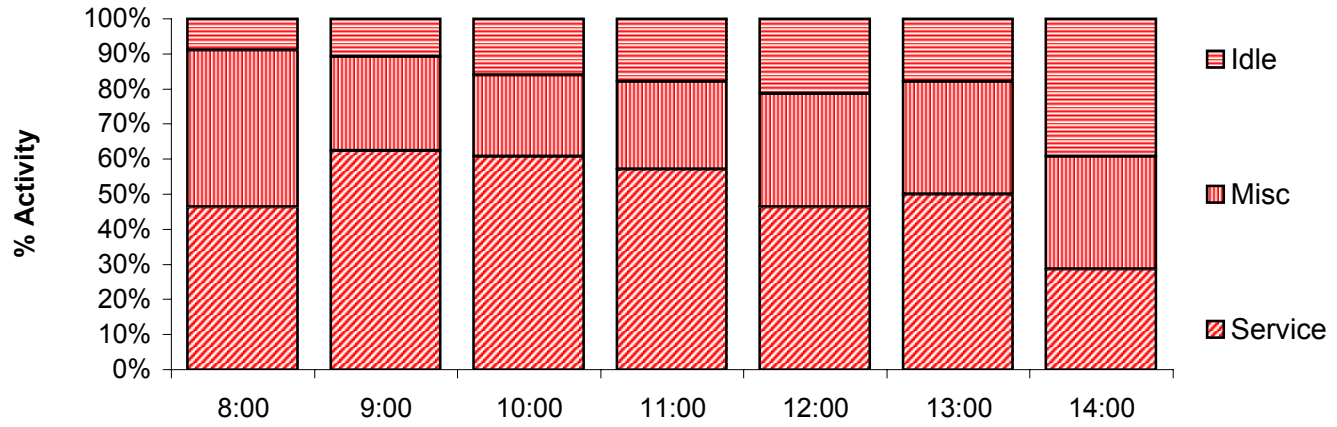
% Abandonment:



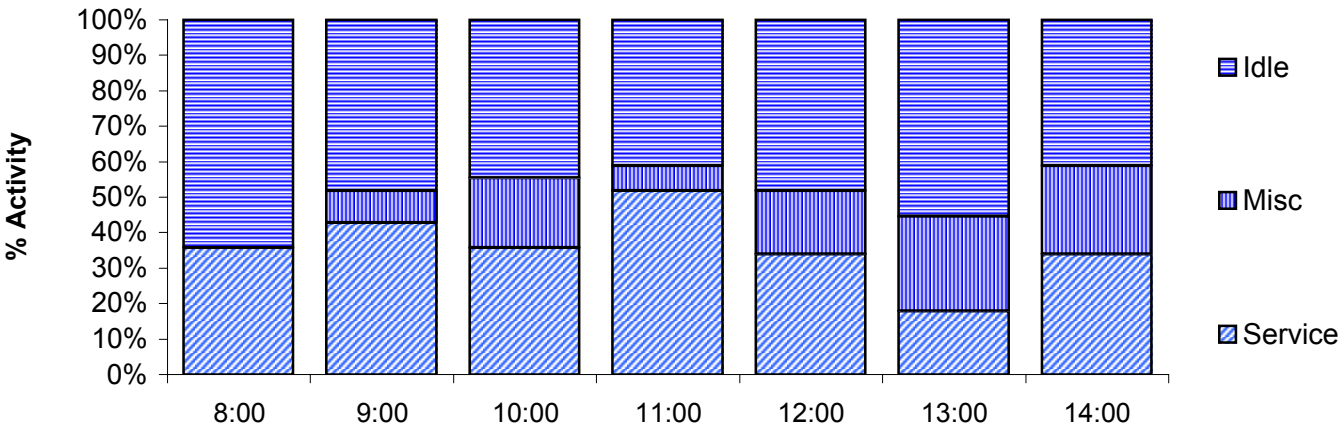
What is “Service Time” ?

Utilization Profile in 3 Call Centers Doing the Same Thing

Tel-Aviv:



Raanana:



Jerusalem:

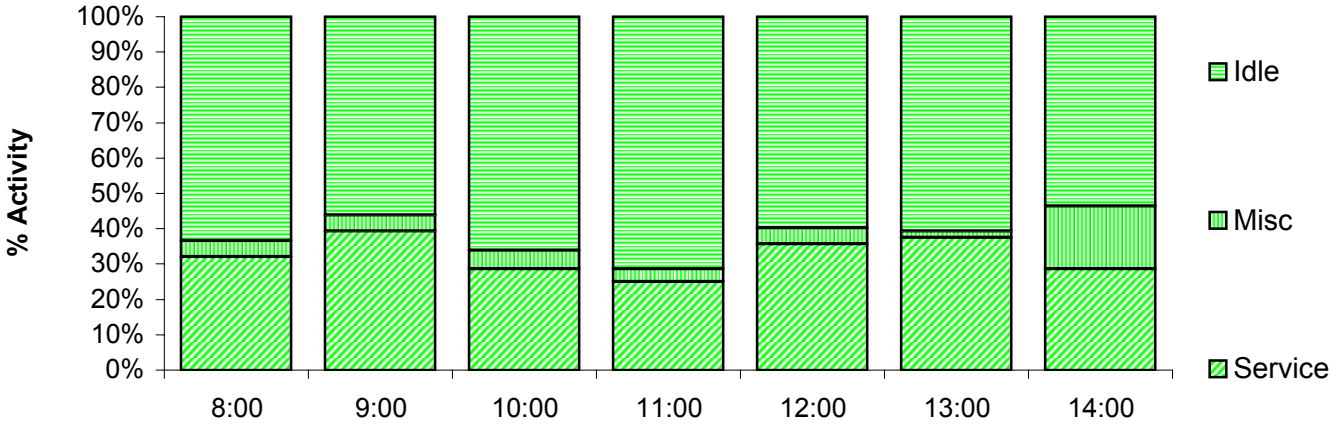


Figure 6: Histogram of Service Times (in seconds)

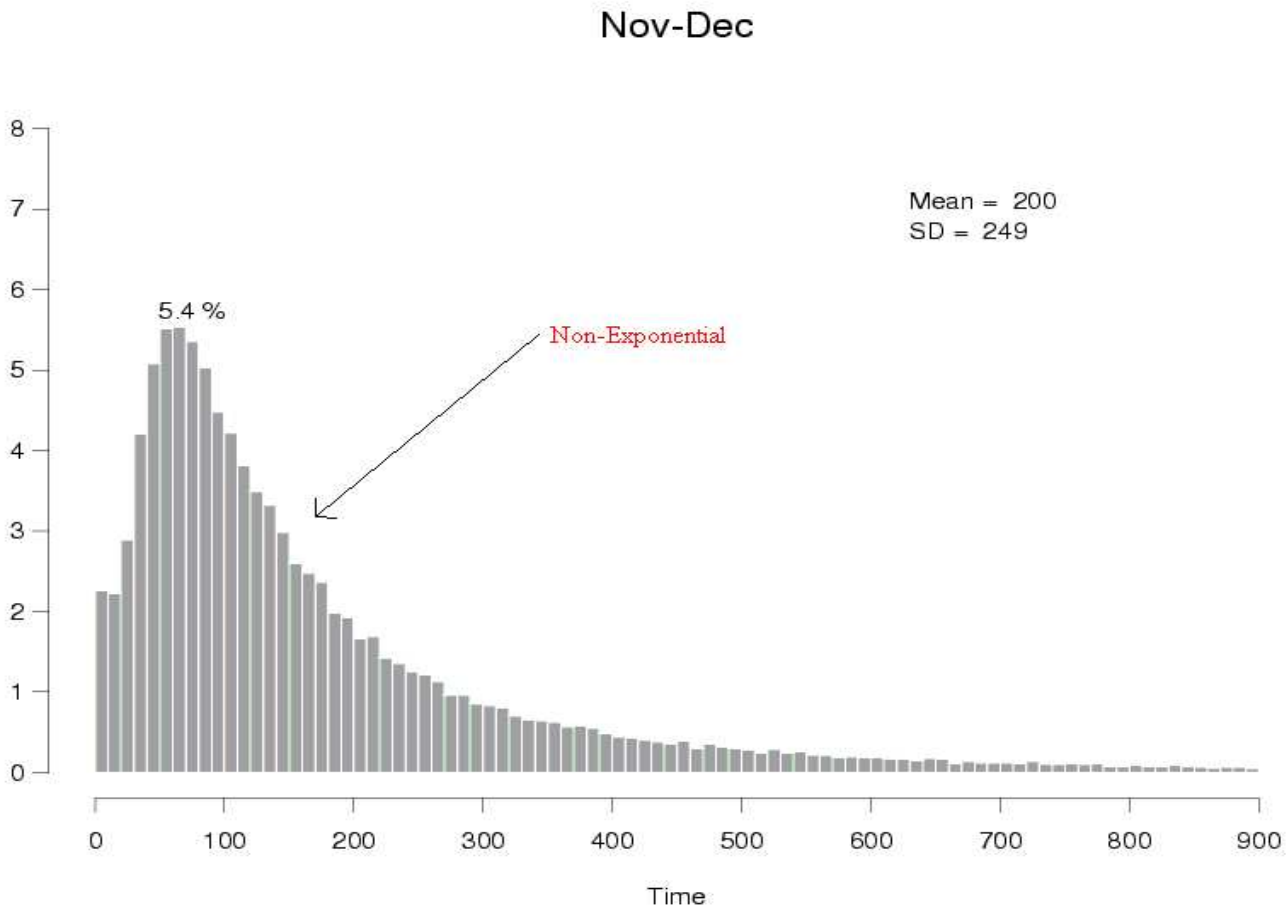
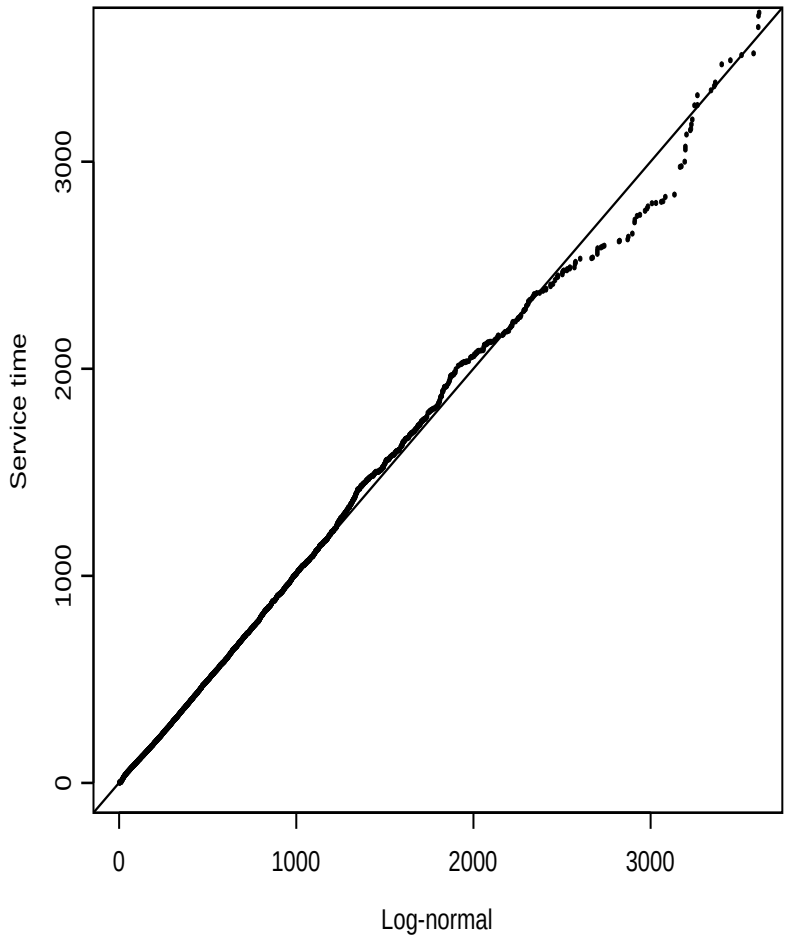
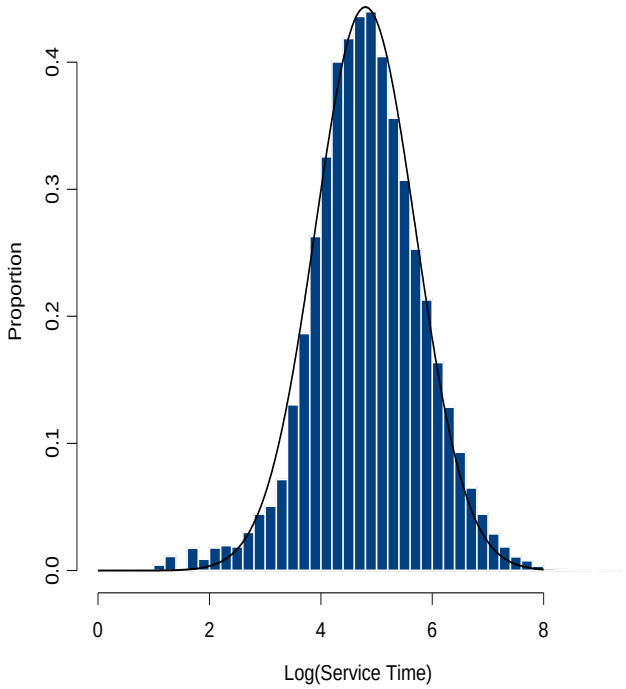


Figure 3: *Log-normal QQ Plot of Service Time (Nov + Dec)*



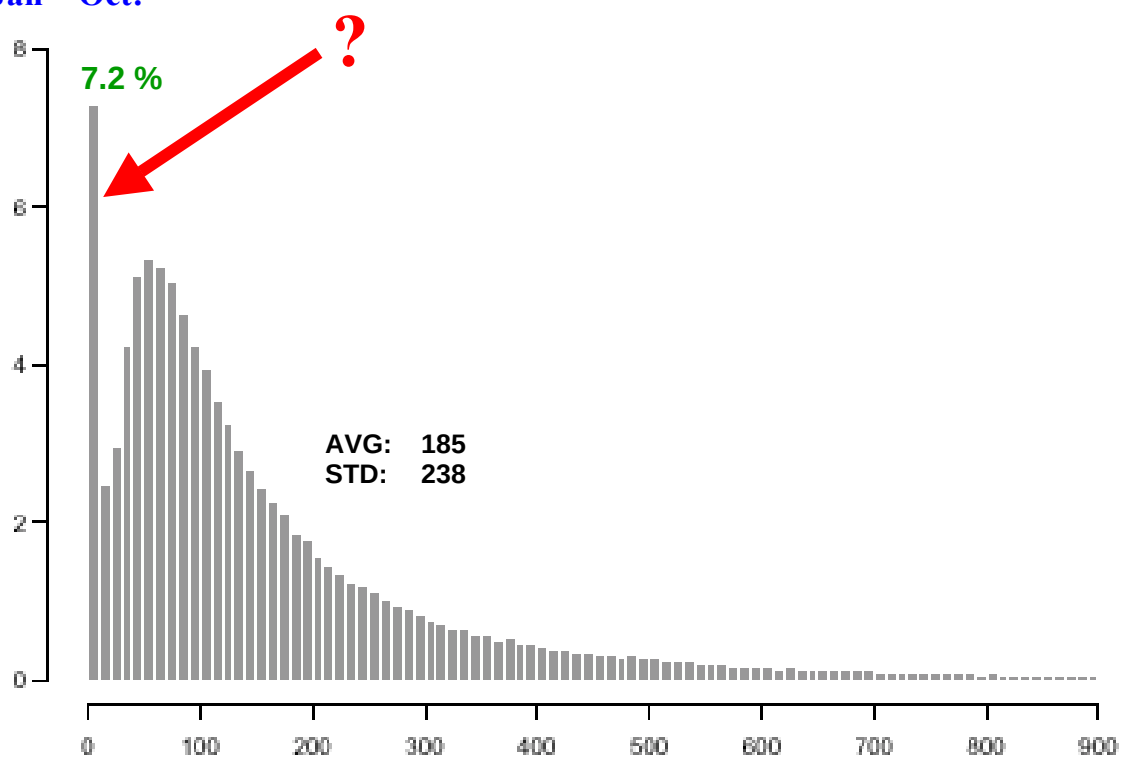
Lognormal Service Time

Figure 2: *Histogram of Log(Service Time) (Nov + Dec)*

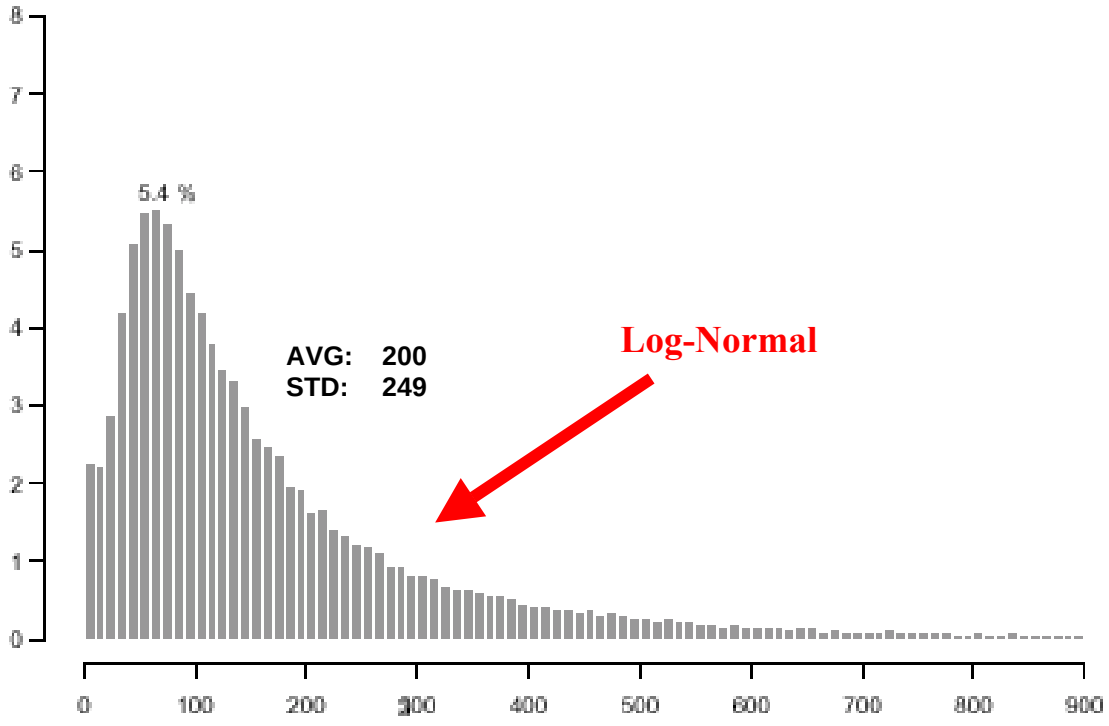


Beyond Data Averages
Short Service Times

Jan – Oct:



Nov – Dec:



Percent Calls w/Service < 10sec

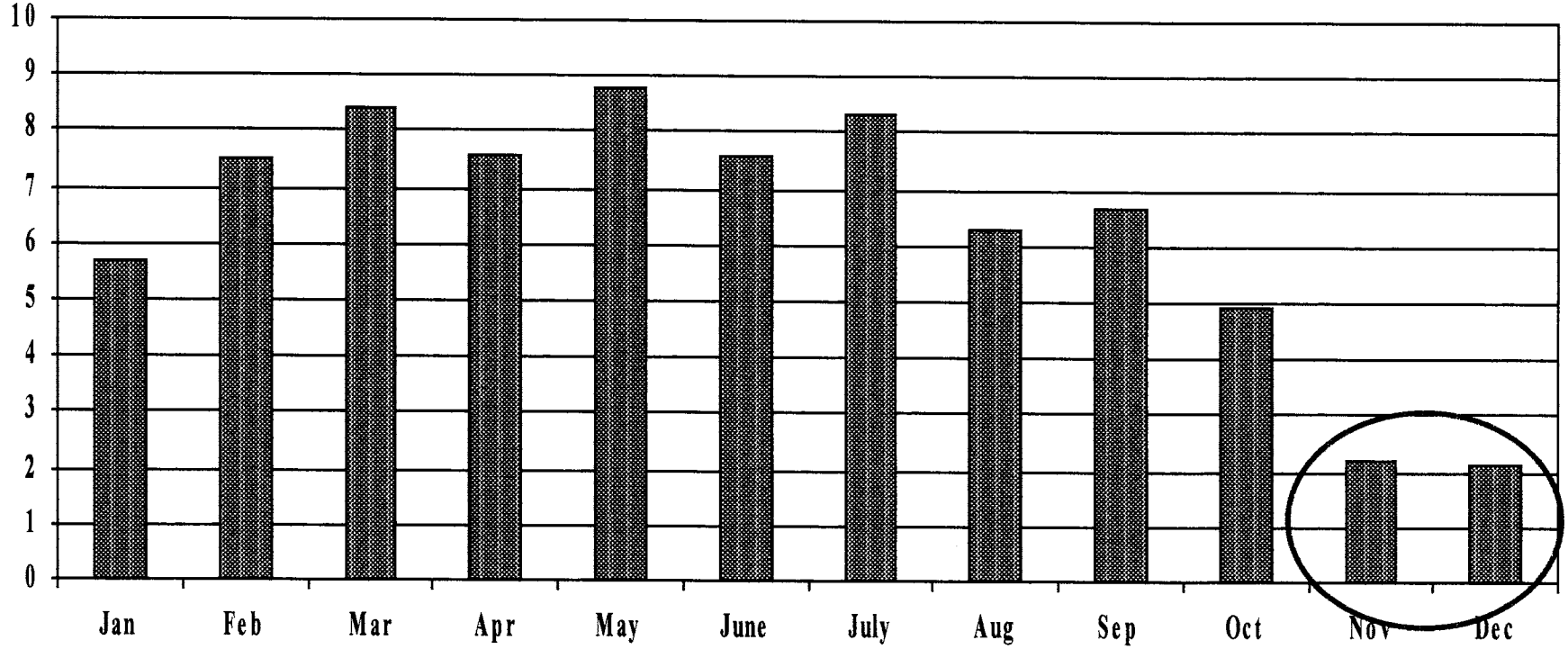


Table 52: Number of calls handled by an agent

	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

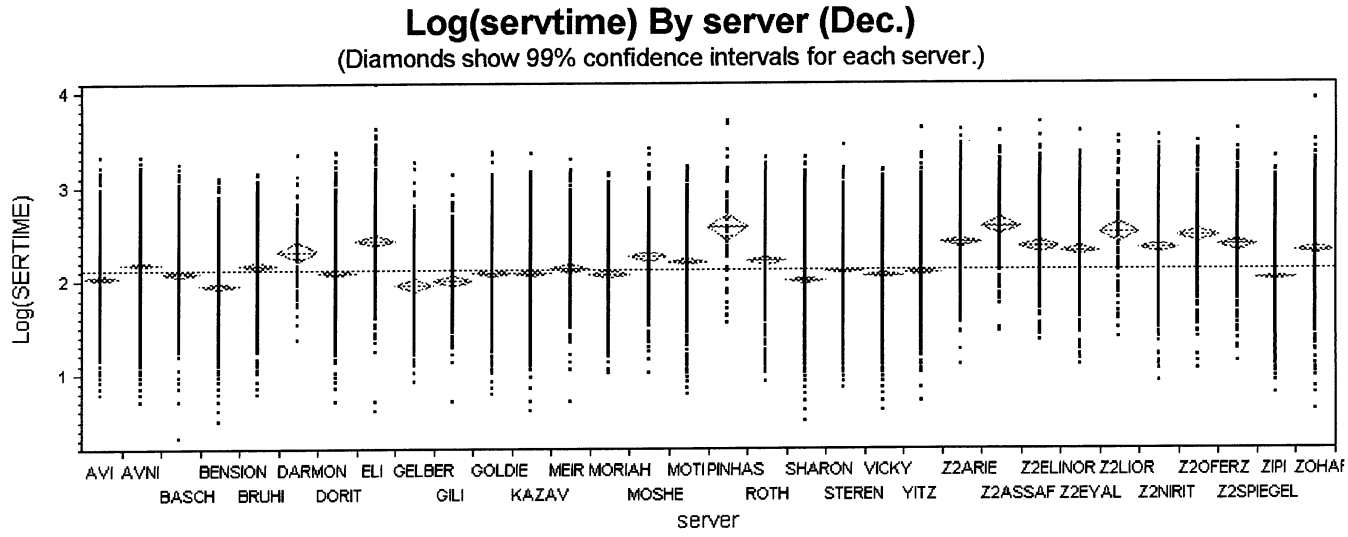
Table 53: Number of calls with short service time

	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

Heterogeneous Servers (Recall the Judges)

Example: Comparison of Service times among various servers.

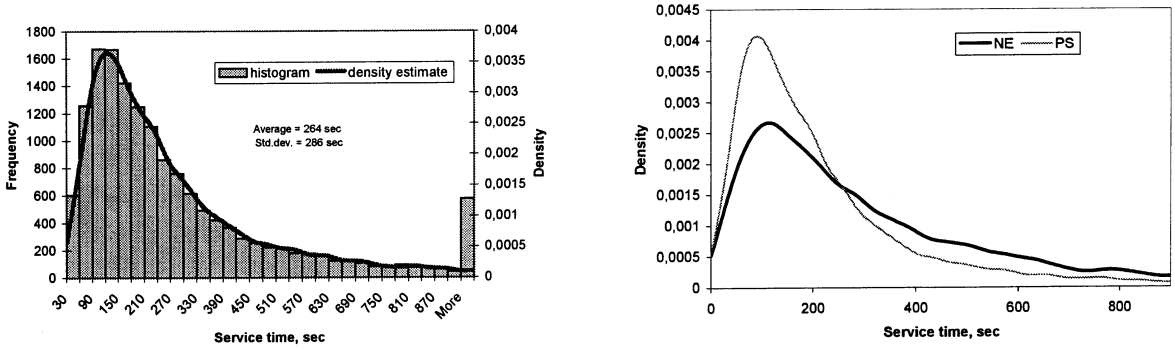
If one breaks the service time down according to the individual servers, the differences are quite noticeable. Here is the situation in December.



[----------]

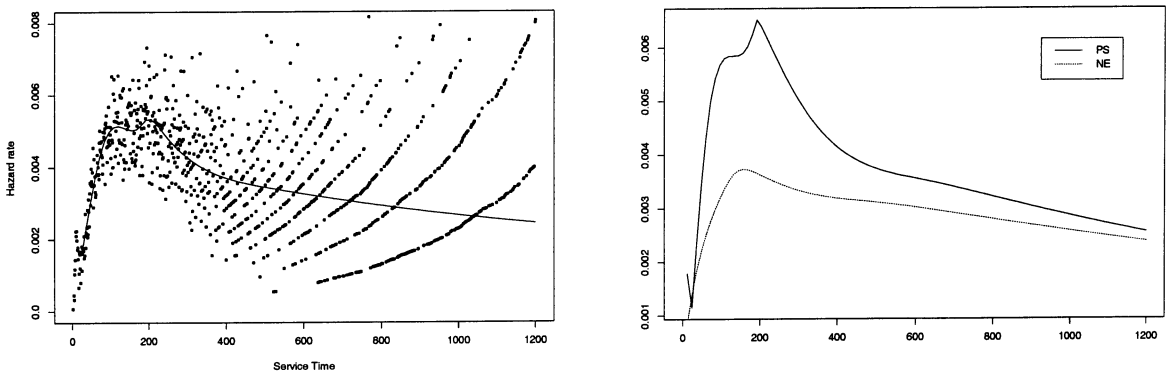
IN Servers

Figure 22: ZOHARI's service time distribution



Next we look at the hazard rate for ZOHARI's service time. The hazard rates were smoothed using HEFT [24]. Figure 23 shows the HEFT estimate, superimposed on the empirical hazard rates, and the HEFT estimate for types PS and NE. The shapes of both the density and the hazard rates are similar to those observed for the overall agent population; see Section 7.

Figure 23: Hazard rate for ZOHARI's service time

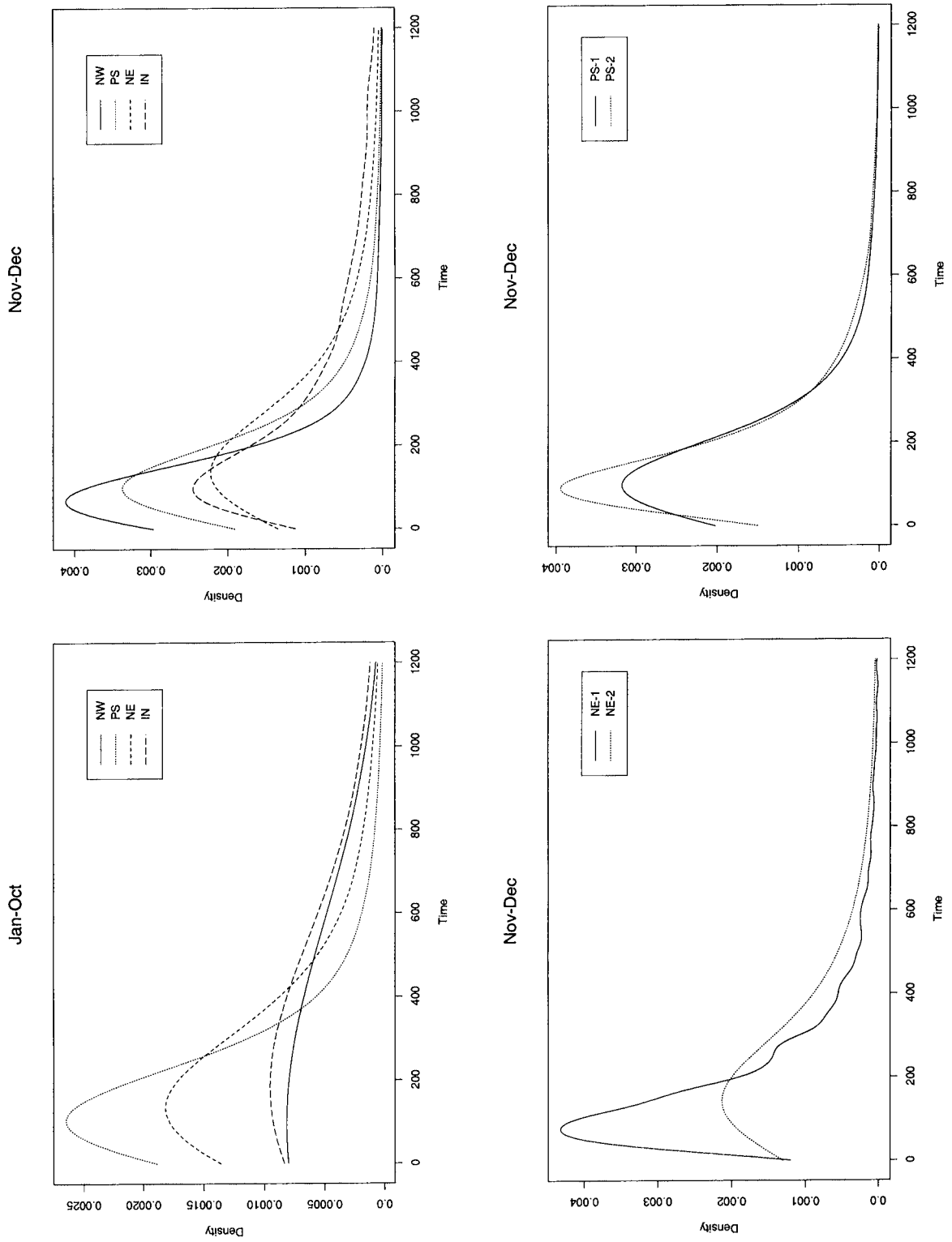


Service Time

	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

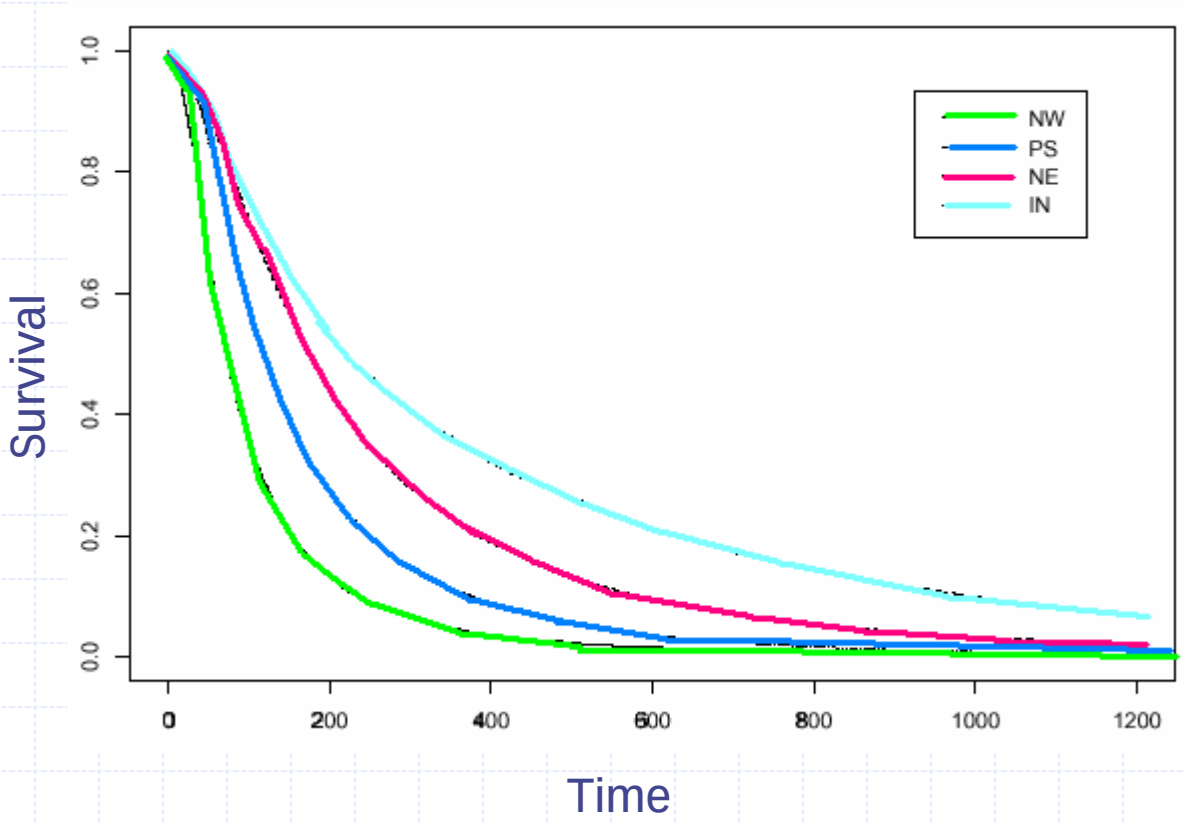
A digression on Stochastic Ordering

Figure 19: Densities of service times, by types and priorities



Service Time

Survival curve, by Types



Means (In Seconds)

NW (New) = 111

PS (Regular) = 181

NE (Stocks) = 269

IN (Internet) = 381

Stochastic ordering:

$X \geq_{st} Y \iff \bar{F}_X(t) \geq \bar{F}_Y(t) \quad \forall t \geq 0.$

(Here $\bar{F} = 1 - F$ survival function)

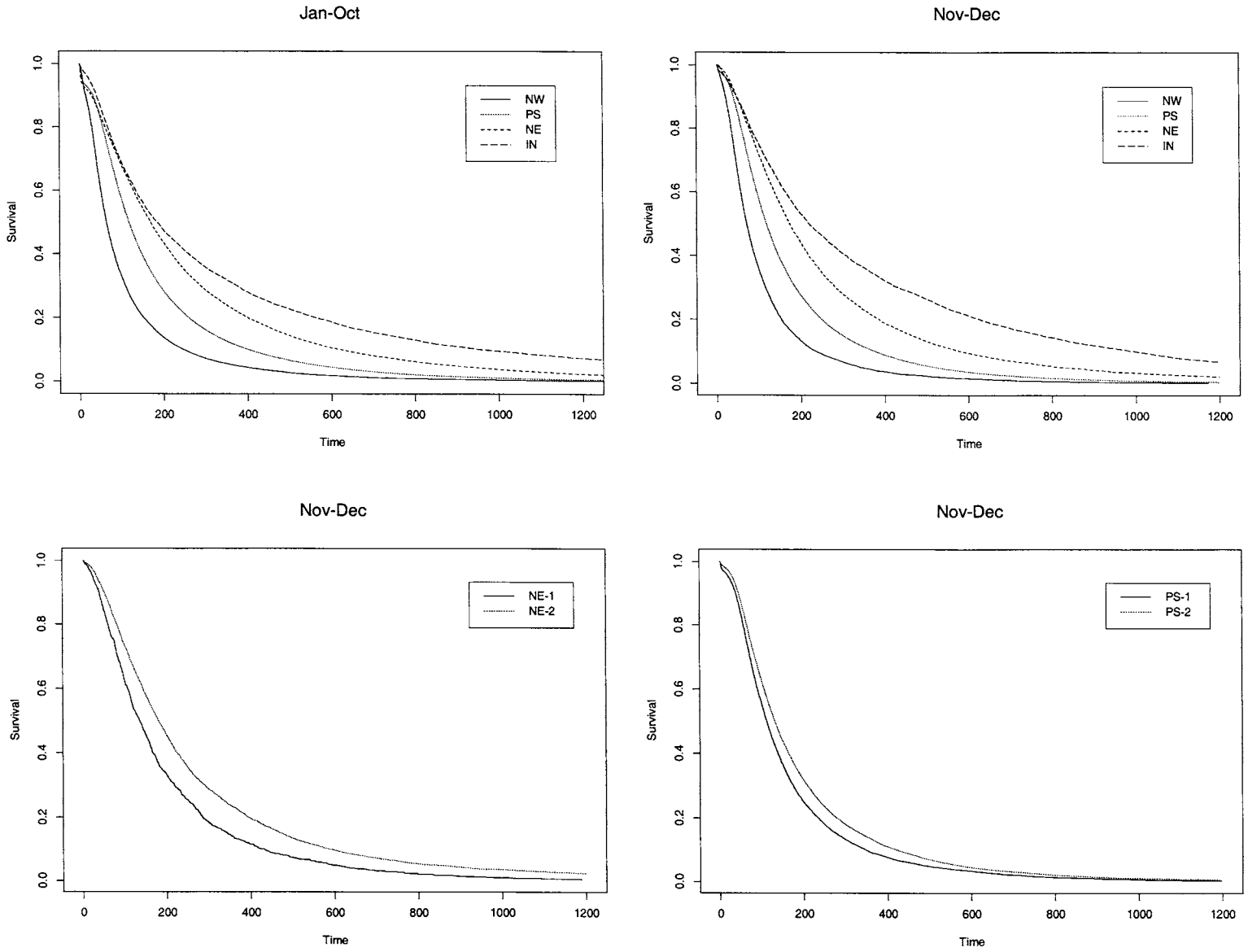
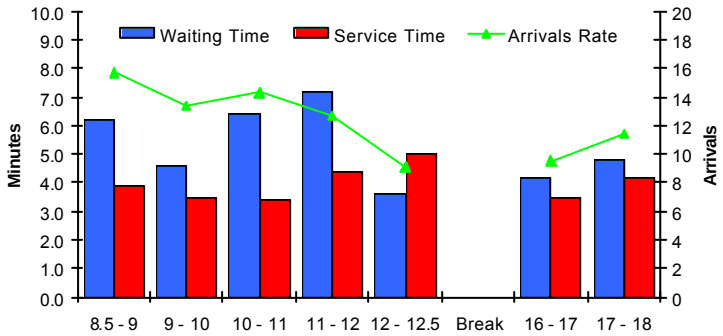


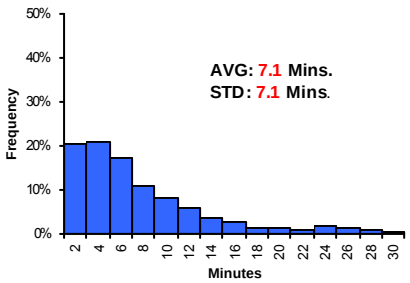
Figure 20: Survival function of service time, by types and priorities

A Bank
Private Banking

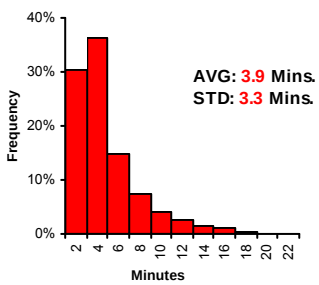
Hour	Arrivals Rate (In Hr)	Waiting Time (Mins)		Service Time (Mins)		Sample Size	Avg. #Tellers
		Avg.	STD	Avg.	STD		
8.5 - 9	15.7	6.2	5.7	3.9	2.7	110	1.00
9 - 10	13.4	4.6	5.4	3.5	2.7	188	1.00
10 - 11	14.3	6.4	6.8	3.4	3.0	200	1.00
11 - 12	12.7	7.2	7.0	4.4	4.4	165	0.94
12 - 12.5	9.1	3.6	4.1	5.0	3.9	41	1.00
Break							
16 - 17	9.5	4.2	4.8	3.5	2.8	63	0.88
17 - 18	11.4	4.8	5.6	4.2	3.3	76	0.95
Average	**12.3**	**5.7**	**6.2**	**3.9**	**3.3**	**843**	**0.97**



Waiting Time Histogram

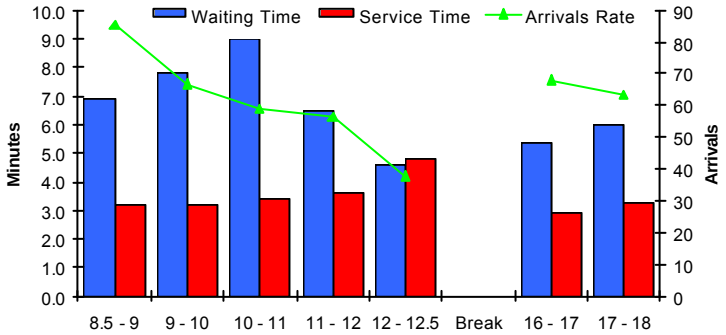


Service Time Histogram

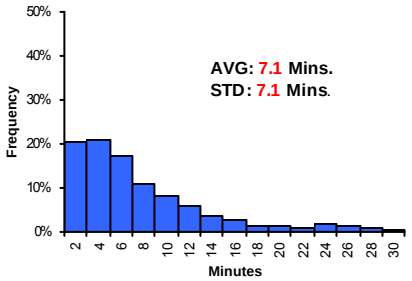


A Bank
General Services

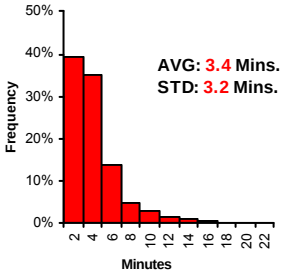
Hour	Arrivals Rate (In Hr)	Waiting Time (Mins)		Service Time (Mins)		Sample Size	Avg. #Tellers
		Avg.	STD	Avg.	STD		
8.5 - 9	85.4	6.9	7.4	3.2	2.7	598	4.5
9 - 10	66.8	7.8	8.7	3.2	3.0	935	4.5
10 - 11	58.9	9.0	8.0	3.4	3.3	825	4.6
11 - 12	56.6	6.5	5.5	3.6	3.4	736	4.2
12 - 12.5	37.8	4.6	4.0	4.8	5.5	227	4.5
Break							
16 - 17	68.1	5.4	4.6	2.9	2.3	465	3.7
17 - 18	63.3	6.0	5.2	3.3	2.7	440	4.2
Average	**62.4**	**7.1**	**7.1**	**3.4**	**3.2**	**4,226**	**4.3**



Waiting Time Histogram

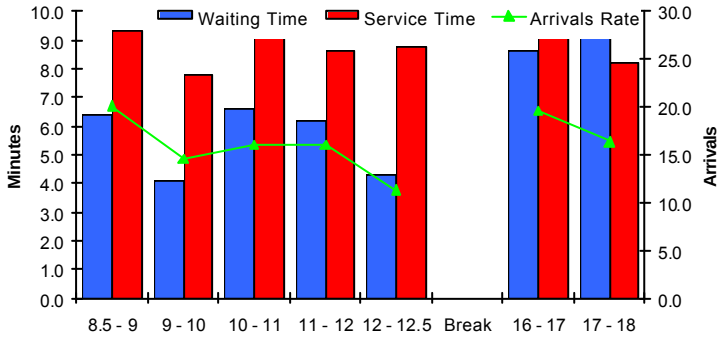


Service Time Histogram

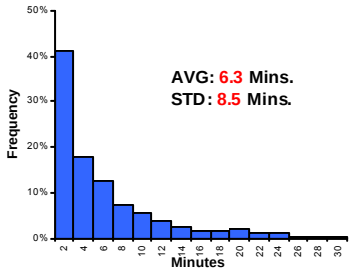


A Bank
Comprehensive Services

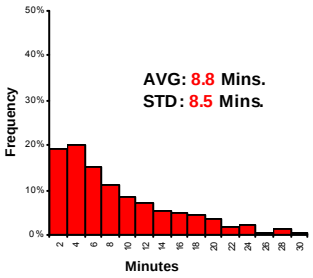
Hour	Arrivals Rate (In Hr)	Waiting Time (Mins)		Service Time (Mins)		Sample Size	Avg. #Tellers
		Avg.	STD	Avg.	STD		
8.5 - 9	20.1	6.4	10.0	9.3	8.8	141	4.1
9 - 10	14.6	4.1	5.7	7.8	8.3	205	4.0
10 - 11	16.0	6.6	7.8	9.2	9.2	224	3.9
11 - 12	16.0	6.2	7.4	8.6	8.7	208	3.7
12 - 12.5	11.3	4.3	6.6	8.8	7.2	68	3.9
Break							
16 - 17	19.6	8.6	11.4	9.8	8.8	135	3.1
17 - 18	16.4	9.4	10.1	8.2	7.2	107	3.6
Average	**16.3**	**6.3**	**8.5**	**8.8**	**8.5**	**1,088**	**3.8**



Waiting Time Histogram

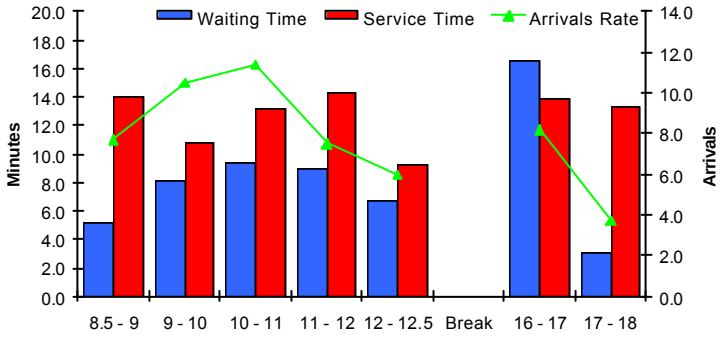


Service Time Histogram

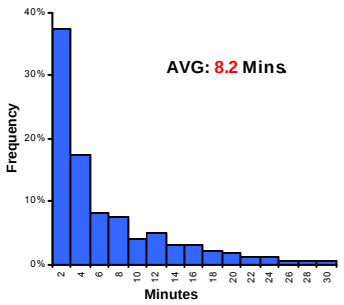


A Bank
Tourists / Business Services

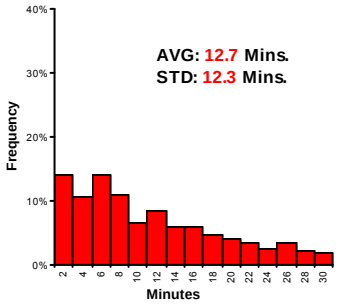
Hour	Arrivals Rate (In Hr)	Waiting Time (Mins)		Service Time (Mins)		Sample Size	Avg. #Tellers
		Avg.	STD	Avg.	STD		
8.5 - 9	7.7	5.1	6.0	14.0	12.3	74	3.44
9 - 10	10.5	8.1	10.7	10.8	11.0	184	3.57
10 - 11	11.4	9.4	12.2	13.2	13.1	201	3.64
11 - 12	7.5	9.0	13.0	14.3	13.2	140	3.64
12 - 12.5	6.0	6.7	12.5	9.3	7.5	46	3.62
Break							
16 - 17	8.2	16.5	12.1	13.9	12.4	61	2.81
17 - 18	3.7	3.0	2.7	13.3	13.7	29	2.71
Average	**7.9**	**8.2**	**11.4**	**12.7**	**12.3**	**735**	**3.34**



Waiting Time Histogram

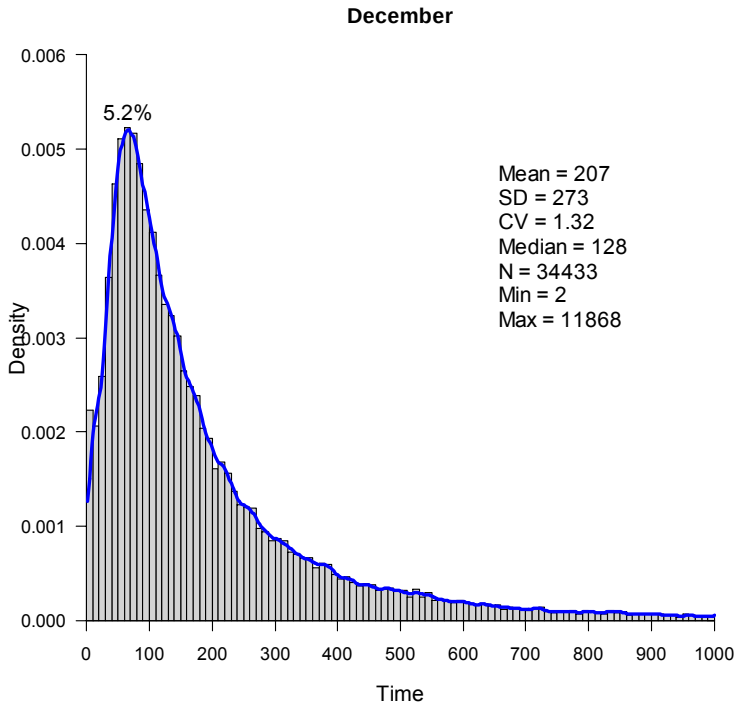


Service Time Histogram



Eva Ishay: M.Sc. Thesis, 2003.

Kernel Density Estimator of Service Time



- **Histogram with h = 10**
  - easy to construct and interpret
  - discontinuous estimator
  - choice of bandwidth (h), tradeoff – bias versus variance
- **Kernel density estimator with a Gaussian kernel of width = 30**
  - continuous and smooth estimator
- Shape is not exponential !
- **Density function**
  - proportion of customers that departure from the service in any time interval
  - the peaks of high frequency of departure from the service

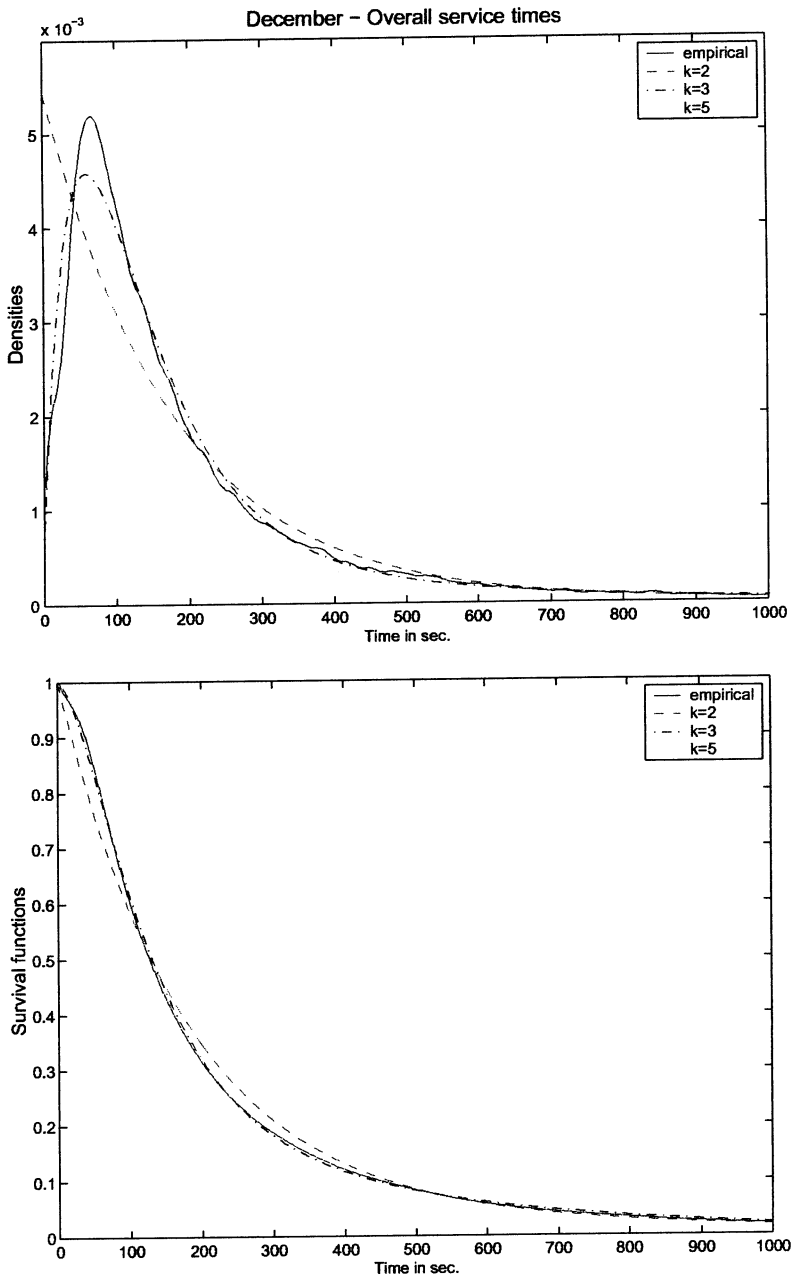
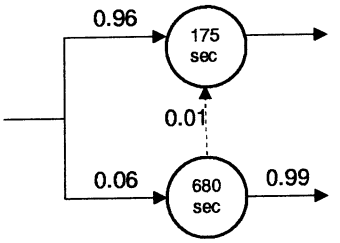
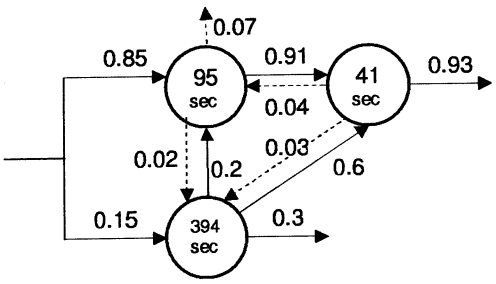


Figure 8.4: Phase-type fits to December service time by a general structure of order $k = 2 - -$, $k = 3 - \cdot -$, $k = 5 \cdots$. In the top plot, the solid line is the kernel density estimator, given as a comparison to the fitted densities. In the bottom plot, the solid line is the empirical survival function.

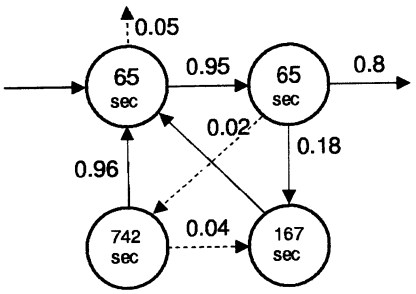
Figure 8.11: Overall service time - December. PH-type structures of order $k = 2, 3, 4, 5, 6$.



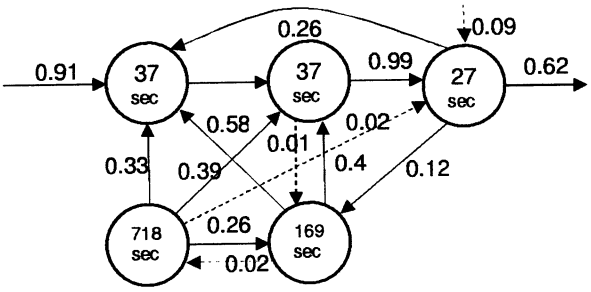
Log-likelihood = -217288



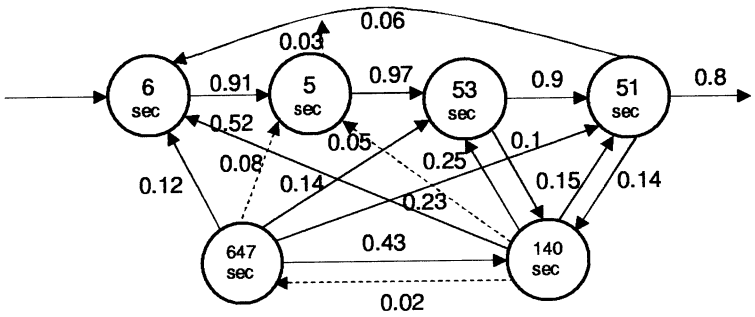
Log-likelihood = -215736



Log-likelihood = -215648



Log-likelihood = -215544

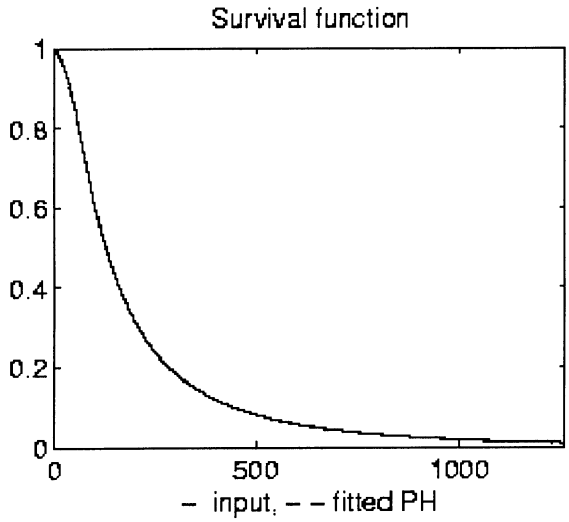


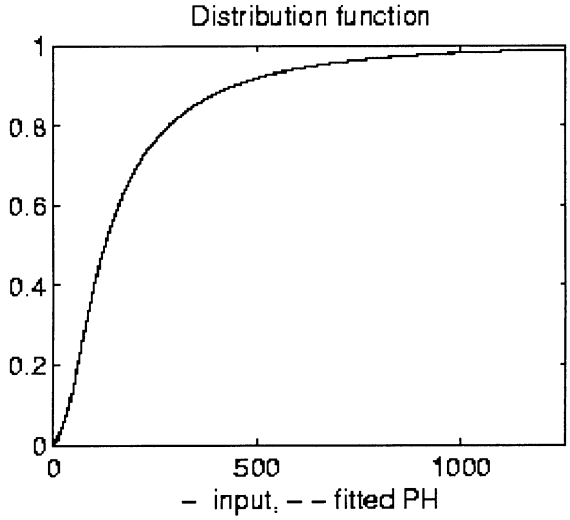
Log-likelihood = -215425

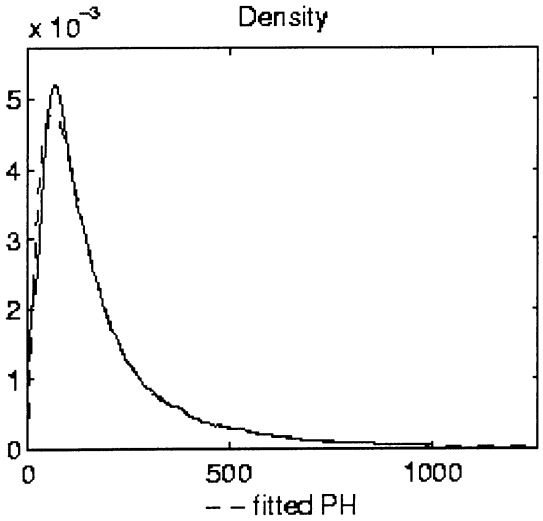
Call Center Data

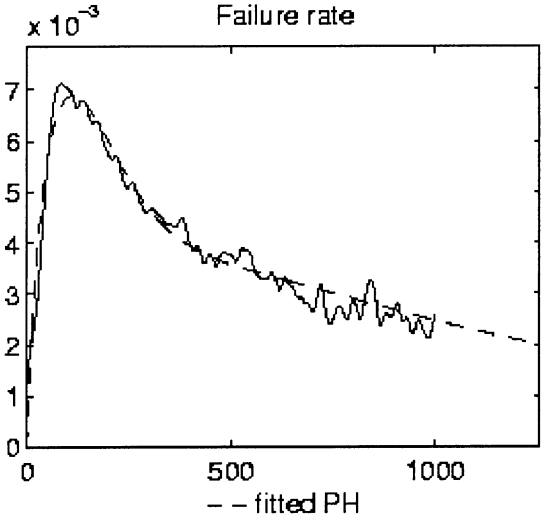
From Eva's M.Sc. Thesis (website)

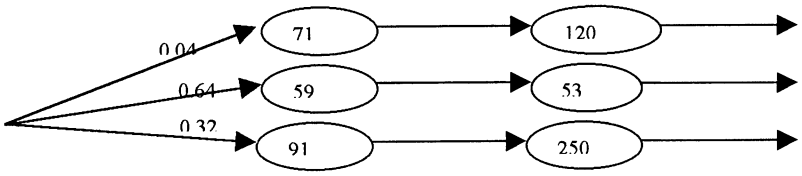
Service times. December. 34433 observations.
Sample mean = 207.0 sec. Sample standard deviation = 272.6 sec.
Type of PH-distributions – Erlang mixtures, p=6.



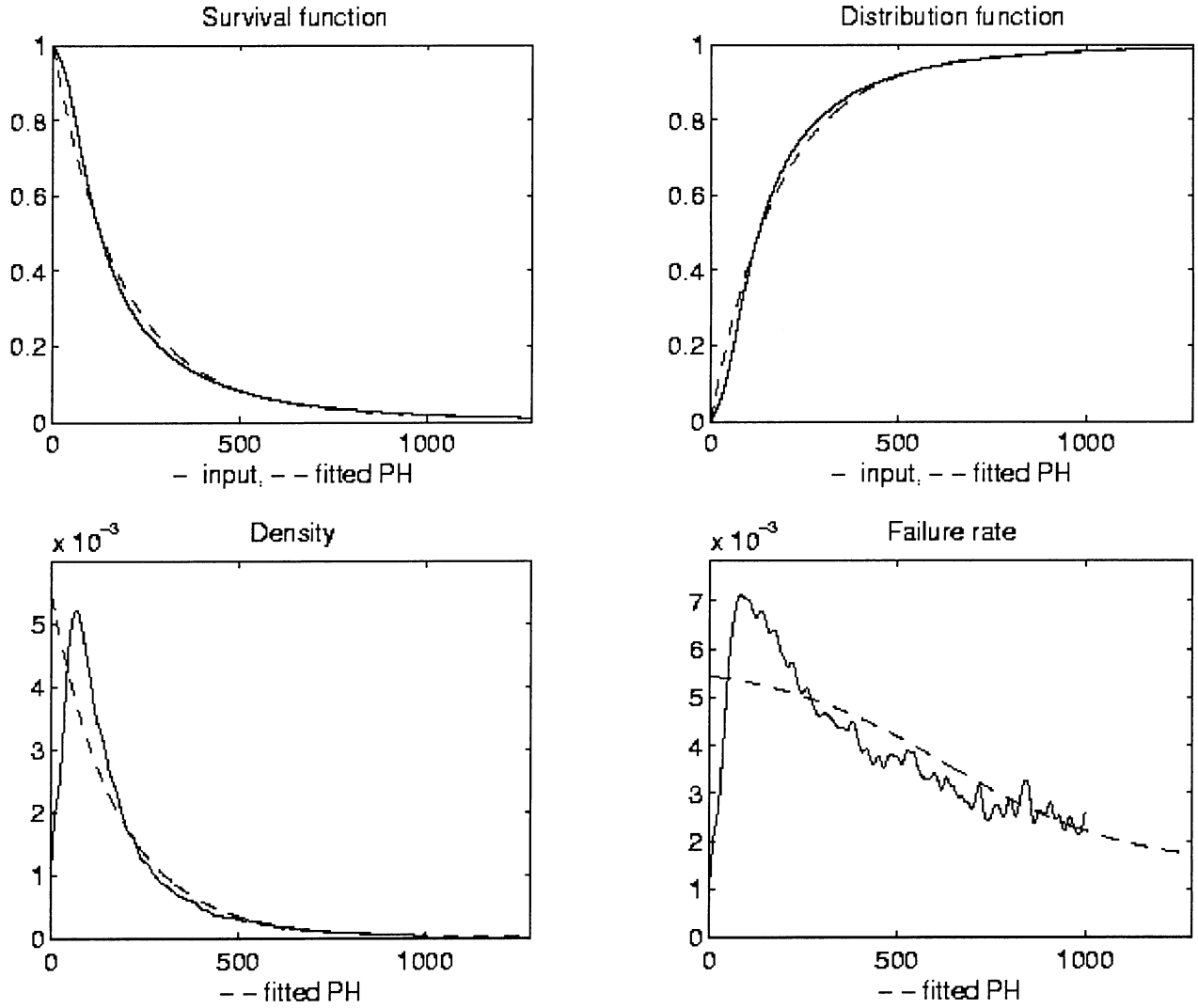




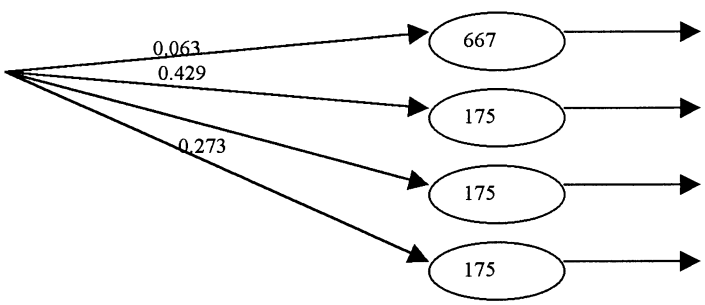




Service time – December.
Type of PH-distributions – Hyperexponential, p=4.

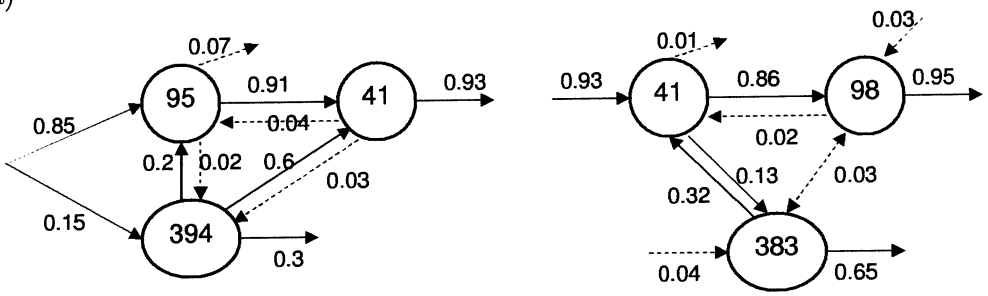


Fitting PH-distribution mean = 207.0286
Fitting PH-distribution standard-deviation = 270.3563



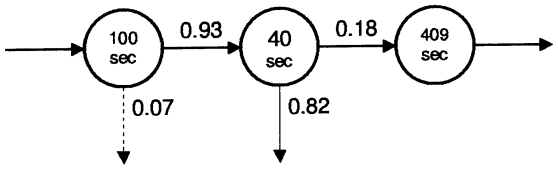
likelihood function is coincide. Figure 8.5 shows three different structures of PH-distribution of order $k = 3$ with the same log likelihood function. The

a)



Log-likelihood = -215735.877179 (final)

b)



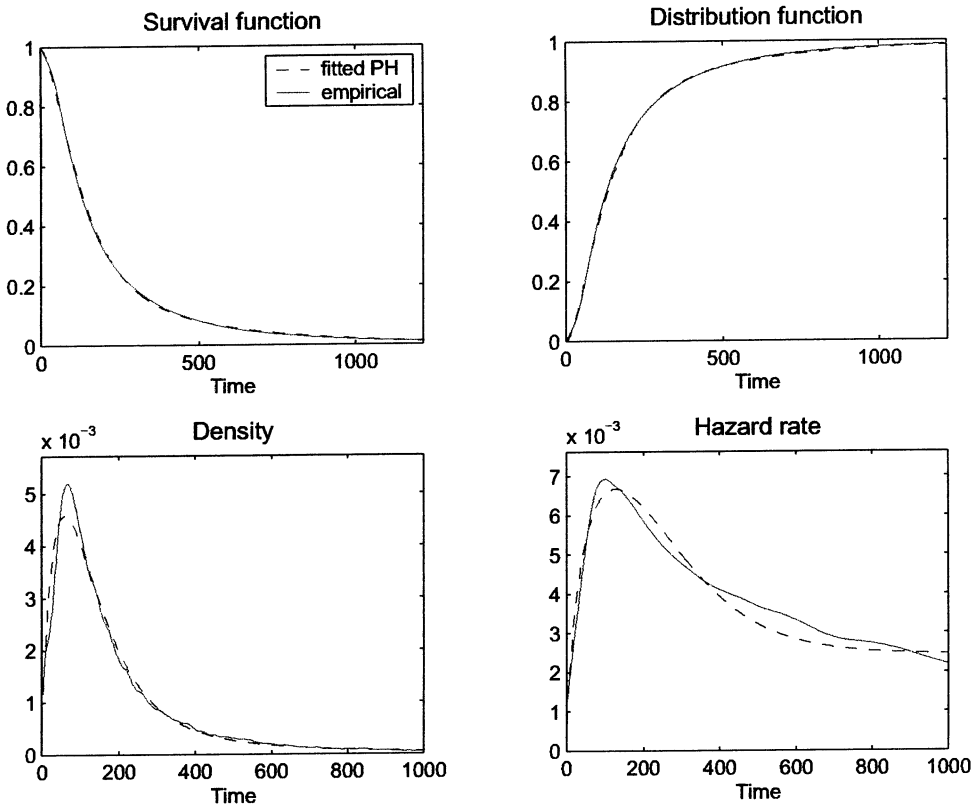
Log-likelihood = -215735.877179 (final)

Figure 8.5: Two different structures of the same order, $k = 3$, of PH-type fit to the service time - December, starting with different initial values, Figure a) above. The fitted Coxian structure of the same order, Figure b) above.

corresponding densities are, when plotted, difficult to distinguish from each other. Then, Figure 8.6 (p. 47) demonstrates the fitted distribution, survival, density and hazard functions together with corresponding empirical functions for the PH-distribution of order $k = 3$ of the structure at left in Figure 8.5. In addition, Figure 8.5 b) demonstrates the fitted Coxian structure of order $k = 3$. It has the same log-likelihood function and, correspondingly, the same fitted mean and standard deviation.

When one looks at the two structures above by ignoring the small probabilities (the dashed arrows in Figure 8.5 a)), it can be seen that despite of different estimated set-up of the parameters at first sight, there are similar length time in the states. Moreover, these two structures can be simplified to the following, showed in Figure 8.7 (p. 48), with corresponding two different setups of parameters $(\mathbf{q}, \mathbf{R})$ and $(\boldsymbol{q}, \boldsymbol{R})'$:

Figure 8.6: PH-type fit of order $k = 3$ of general structure (dashed curve) with empirical functions (solid curve).



The fitted PH-distribution has mean = 207 and standard-deviation = 253, CV = 1.22.

(1)

$$\mathbf{q} = \begin{pmatrix} q \\ 0 \\ 1-q \end{pmatrix} \quad \mathbf{R} = \begin{bmatrix} -\lambda_1 & \lambda_1 & 0 \\ 0 & -\lambda_2 & 0 \\ 0 & \lambda_3 & -\lambda_3 \end{bmatrix}$$

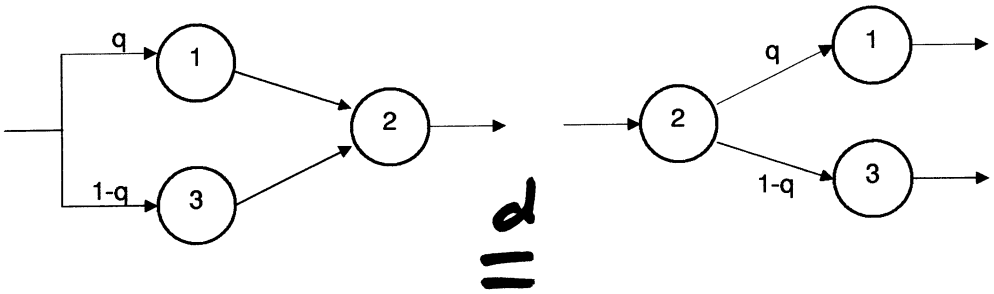
(2)

$$\mathbf{q}' = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \quad \mathbf{R}' = \begin{bmatrix} -\lambda_1 & 0 & 0 \\ q\lambda_2 & -\lambda_2 & (1-q)\lambda_2 \\ 0 & 0 & -\lambda_3 \end{bmatrix}$$

Let X_1, X_2, X_3 - three independent random variables exponentially distributed

Identifiability ?

Figure 8.7: An example of PH-distribution of third order represented by two different setups of parameters.



with parameters $\lambda_i, i = 1, 2, 3$, and an indicator

$$I = \begin{cases} 1 & \text{w.p. } q \\ 0 & \text{w.p. } 1-q \end{cases}$$

independent of $X_i, i = 1, 2, 3$. Then $Y = X_2 + IX_1 + (1 - I)X_3$ is time to absorption, and in both cases:

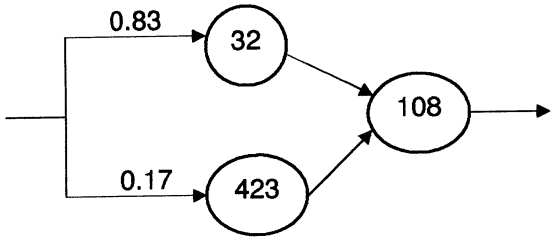
$$f_Y(y) = \frac{\lambda_1\lambda_2 q}{(\lambda_1 - \lambda_2)}\left(e^{-\lambda_2 y} - e^{-\lambda_1 y}\right) + \frac{\lambda_2\lambda_3(1-q)}{(\lambda_3 - \lambda_2)}\left(e^{-\lambda_2 y} - e^{-\lambda_3 y}\right), y > 0.$$

The PH-distribution of order $k = 3$ of the structure at left in Figure 8.7 is fitted to compare its results with the fitted PH-distribution of the same order of the general structure, given in figure 8.5. Figure 8.8 (p. 49) shows the derived specified structure with corresponding log-likelihood function and the graph of fitted PH-density functions of general and specified structures together with kernel density estimator. It is difficult to distinguish between the two structures, according to the graph of their survival functions. It can be seen, according to their fitted density functions that there is a little difference between them at the top of the mode, with preference to specified structure. However, according to the likelihood function, the general structure has larger likelihood (or consequently, the smaller log-likelihood).

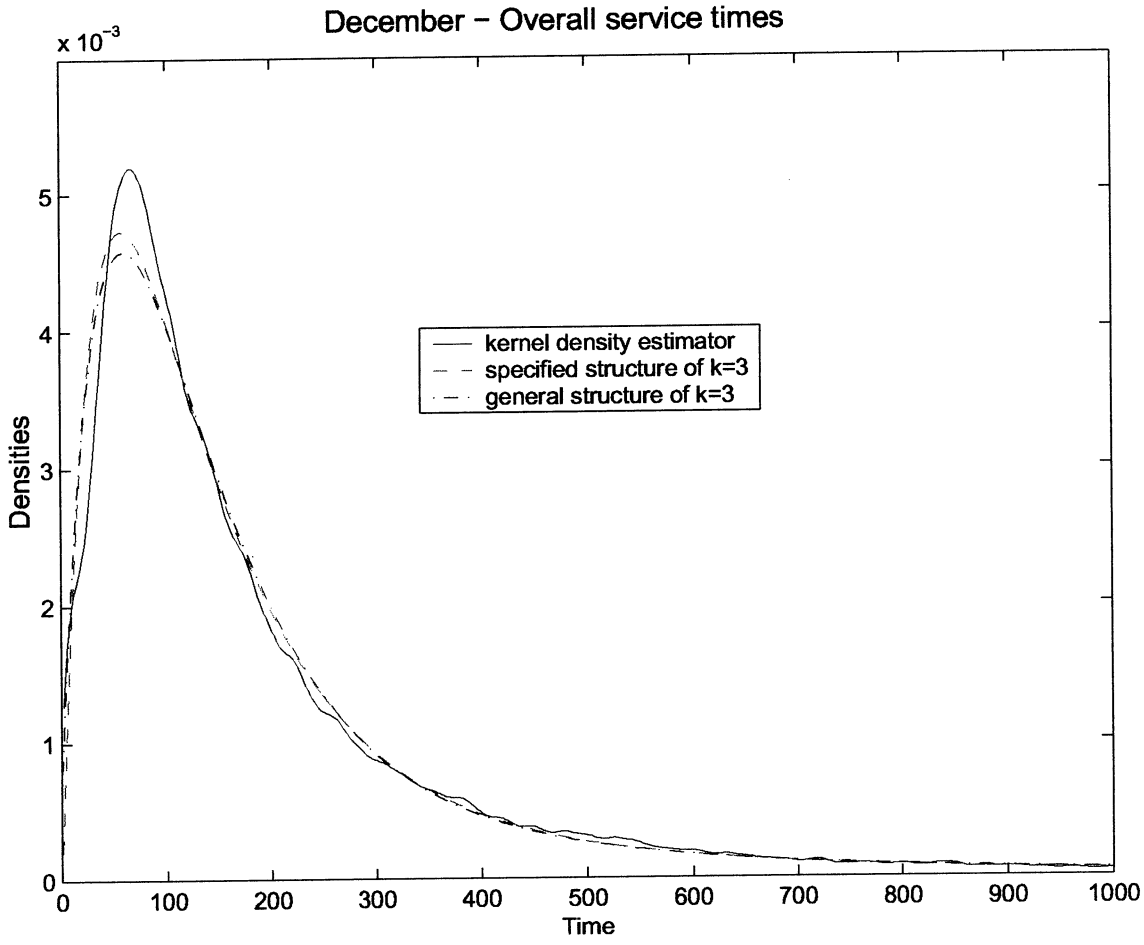
Table 8.1 presents the fitted PH-distribution mean (Mean), standard-deviation (SD), coefficient of variation (CV) and log-likelihood function (Log-L) for the fitted general structure of order $k = 2, 3, 4, 5, 6$ to the service time - December.

Table 8.2 (p. 50) shows the results of applying EDF tests – the D^* and A^2 statistics associated with the K-S and A-D tests, respectively. These statistics heavily depend on size of the sample data. In table at top, the

Figure 8.8: The specified PH-structure of order $k = 3$ fitted to the service time - December (at top). The fitted PH-density functions with empirical one (at bottom).



Log-likelihood = -215792.369456 (final).



by fitting PH-distributions of general structure to service time - December, by priorities.

Figure 8.17: Service time - December, by priorities. PH-type structures of order $k = 3, 4, 5$.

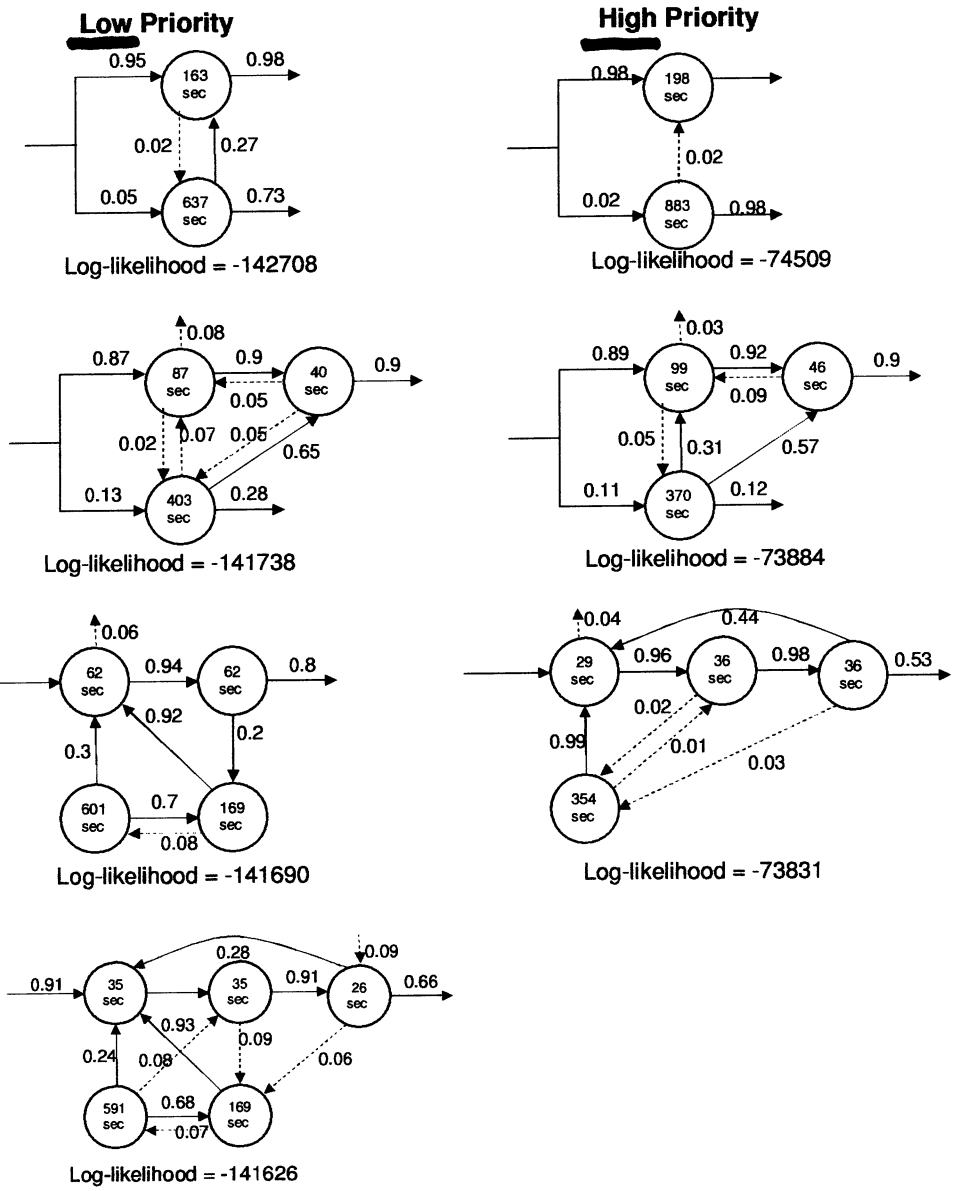
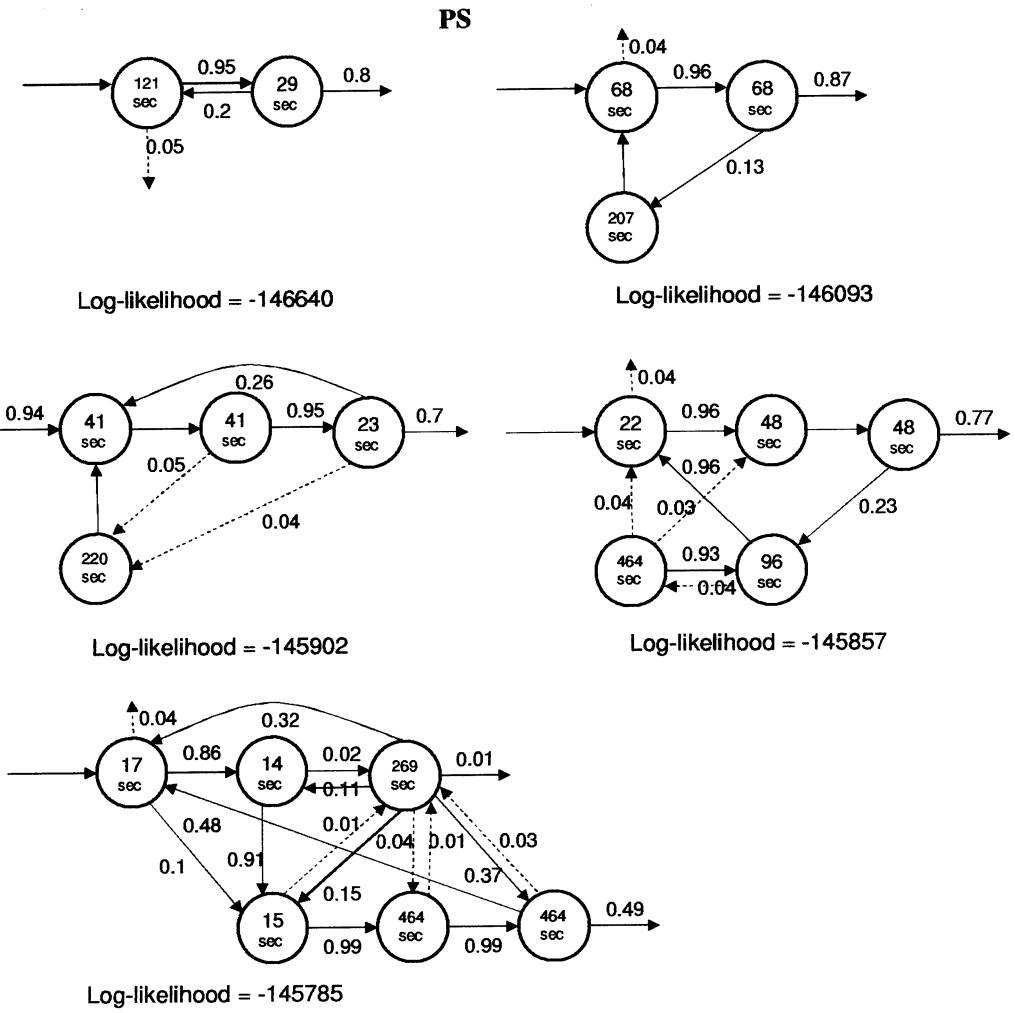
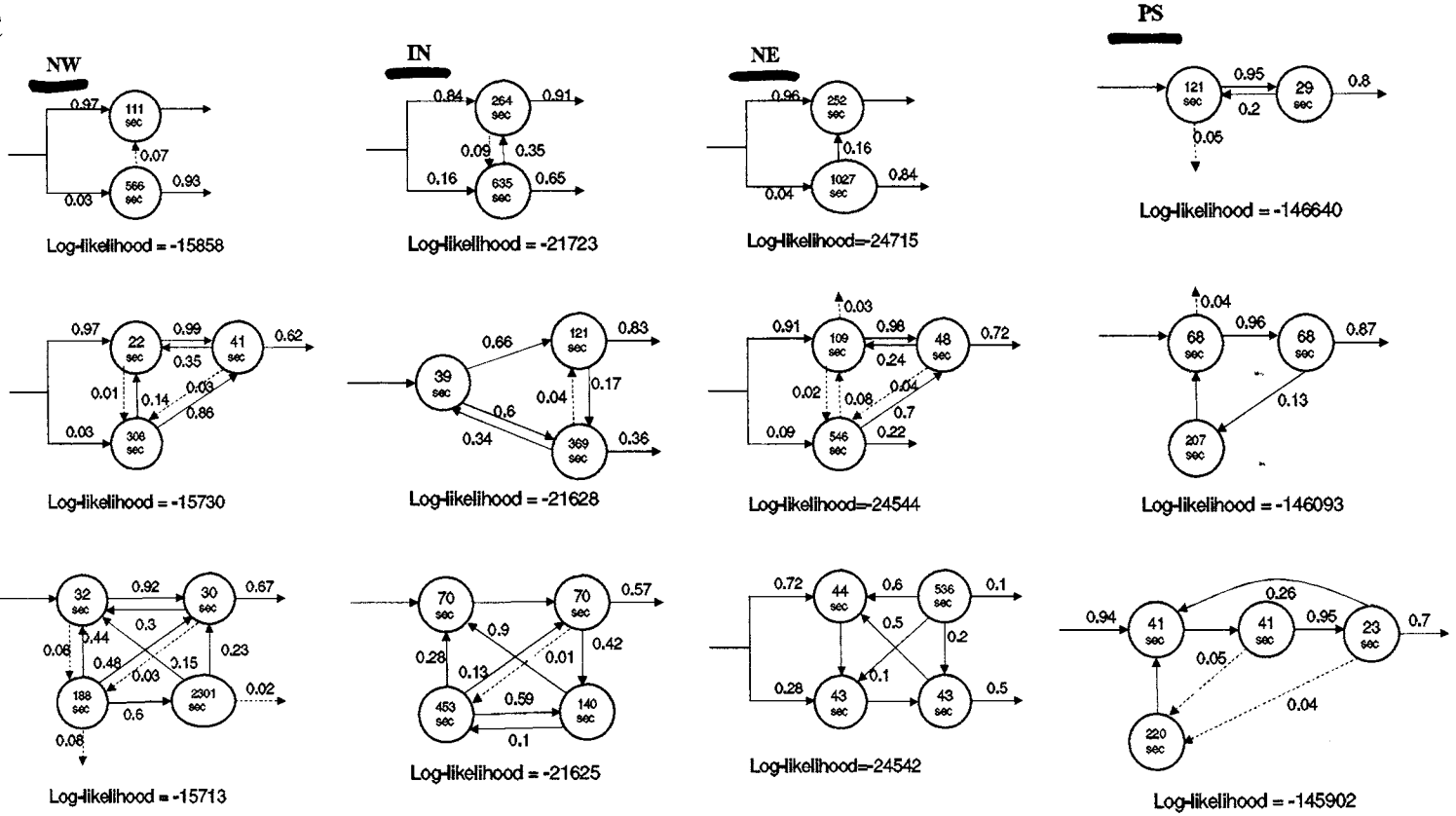


Figure 8.25: Service time - December, by types. PH-type structures of order $k = 2, 3, 4, 5, 6$.



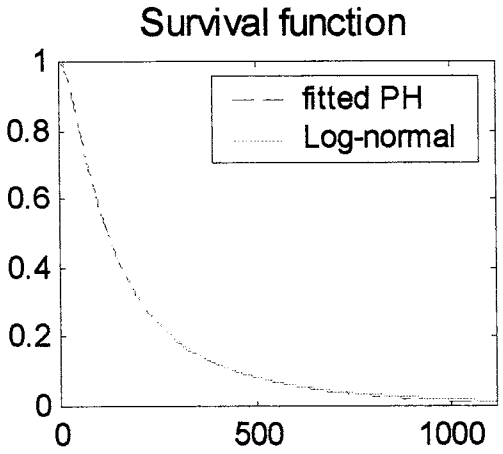
a)

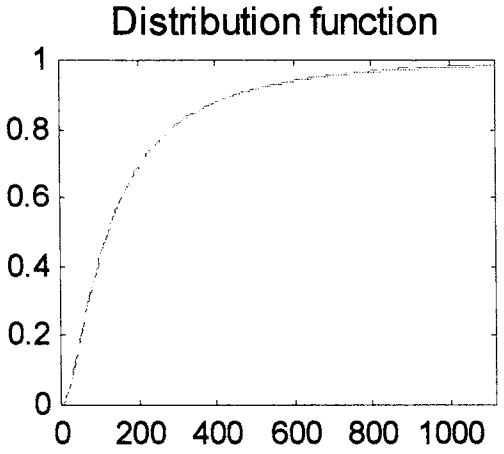
b)

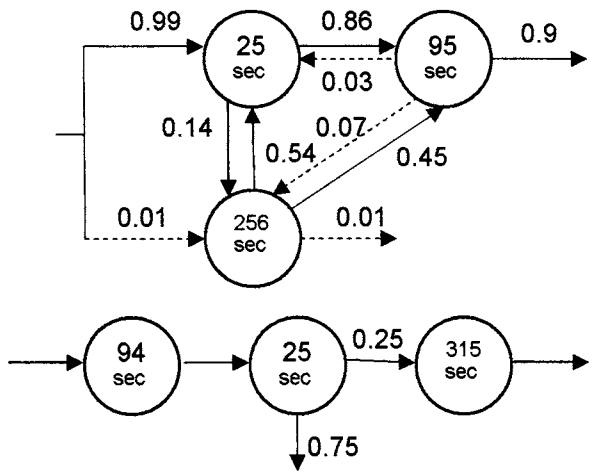

NW
0.97
111 sec
0.07
566 sec
0.93
0.03
Log-likelihood = -15858
IN
0.84
264 sec
0.91
0.09
0.35
0.16
635 sec
0.65
Log-likelihood = -21723
NE
0.96
252 sec
0.16
1027 sec
0.84
0.04
Log-likelihood=-24715
PS
121 sec
0.95
0.2
29 sec
0.8
0.05
Log-likelihood = -146640
0.97
22 sec
0.99
0.35
41 sec
0.62
0.01
0.14
0.03
0.86
306 sec
0.03
Log-likelihood = -15730
0.66
121 sec
0.83
39 sec
0.04
0.17
0.6
0.34
369 sec
0.36
Log-likelihood = -21628
0.03
0.91
109 sec
0.98
48 sec
0.72
0.24
0.02
0.08
0.04
0.7
0.09
546 sec
0.22
Log-likelihood=-24544
0.04
68 sec
0.96
68 sec
0.87
0.13
207 sec
Log-likelihood = -146093
32 sec
0.92
30 sec
0.67
0.3
0.44
0.06
0.48
0.03
0.15
0.23
188 sec
2301 sec
0.02
0.6
0.08
Log-likelihood = -15713
70 sec
70 sec
0.57
0.9
0.28
0.01
0.42
0.13
453 sec
0.59
140 sec
0.1
Log-likelihood = -21625
0.72
44 sec
0.6
536 sec
0.1
0.5
0.2
0.28
0.1
43 sec
43 sec
0.5
Log-likelihood=-24542
0.94
41 sec
0.26
41 sec
0.95
23 sec
0.7
0.05
0.04
220 sec
Log-likelihood = -145902

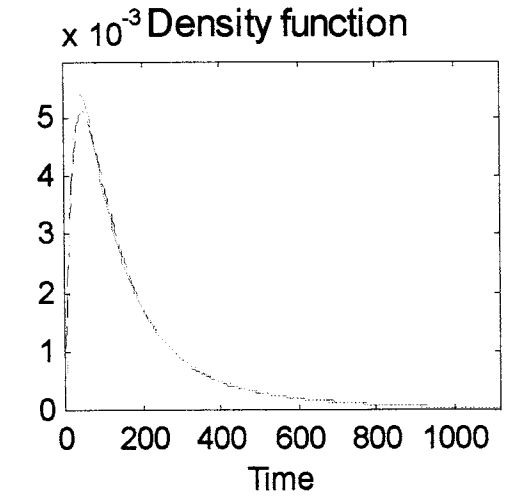
77

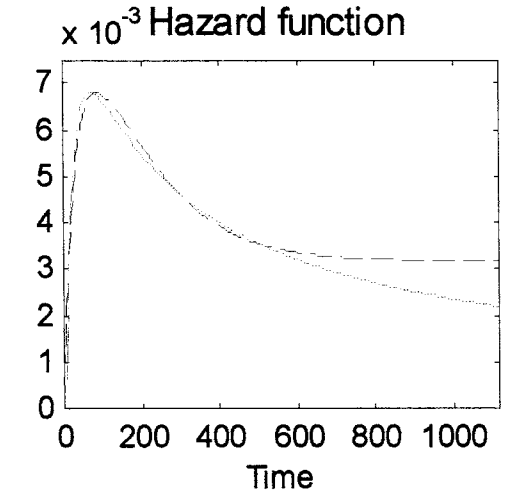
Service Times:
Approximation of Lognormal(μ=4.8, σ=1.03) by PH of order 3











- Fitted mean = 198
- Fitted SD = 230
- CV = 1.16

- E(LN) = 207
- SD(LN) = 284
- CV(LN) = 1.37

Simulation Experiments with $M/G/100$ Queues in the *Halfin-Whitt* (Q.E.D) Regime

Supervised By: Avishai Mandelbaum* Project By: Roy Schwartz*

‡ avim@ie.technion.ac.il
‡ schwartz@ie.technion.ac.il

Draft of July 9, 2002

Industrial Engineering and Management, Technion, Haifa 32000, Israel

Dependence of performance on the service time distribution in the QED regime: M, D, LN

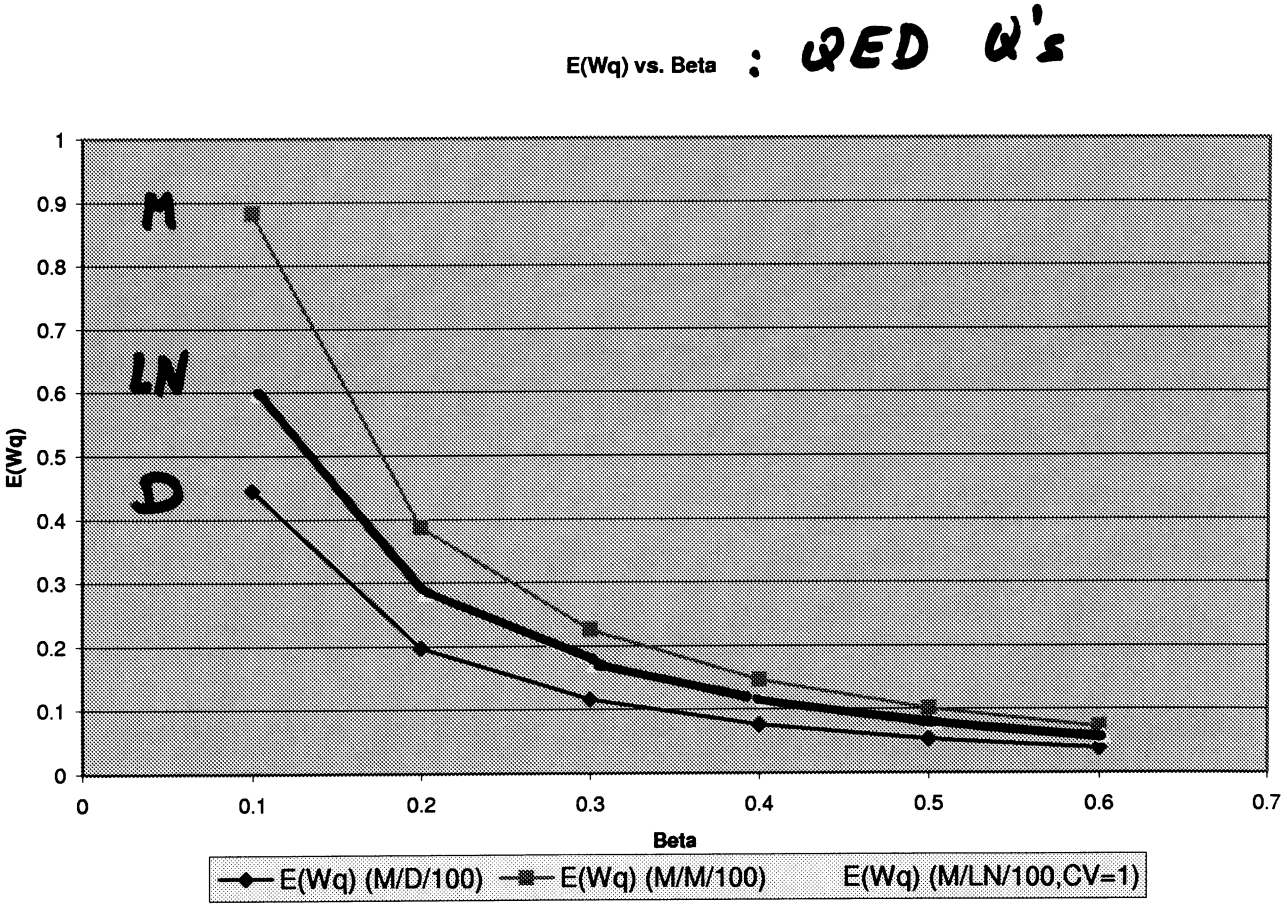
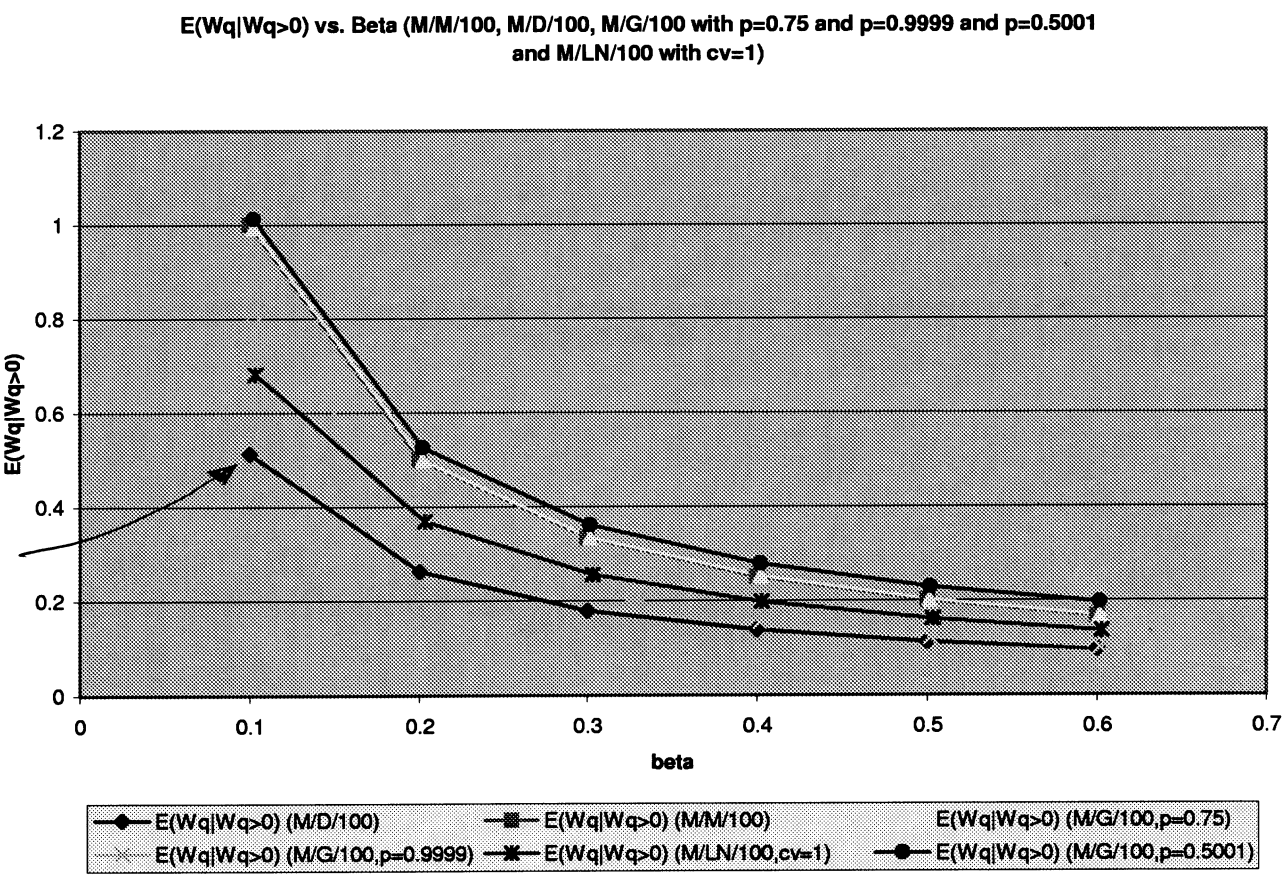


Figure 6: $E(W_q)$ vs. β ($M/M/100$, $M/D/100$ and $M/LN/100$ with $CV = 1$)



M/D/1 and 2-valued w/ CV=1

Figure 19: $E(W_q|W_q > 0)$ vs. β for special service time distribution and *regular* distributions

the case $p = 0.5001$ has the highest results in the $E(W_q|W_q > 0)$ case.

Notice that in all three cases of the special distribution, the slope of the corresponding line of each case becomes steeper as ρ gets closer to 1 (equivalently as β gets closer to 0) than the slope of the lines that correspond to the *regular* distributions. It might be worth mentioning that $E(W_q|W_q > 0)$ is the only statistic for which we have obtained lines that intersect with each other, when making graphs for a statistic vs. β (i.e. the only instance that line intersect is in Figure 19). Again, we have no explanation for this phenomenon (it might be some error in the simulation or a numerical inaccuracy , but we have not been successful in finding one). We would like to point out, that this intersection of lines is most significant in the three following systems: $M/LN/100$ (with coefficient-variance of 1), $M/D/100$ and $M/G/100$ (where $p = 0.9999$). Note that the line that represents $E(W_q|W_q > 0)$ in the case of $M/G/100$ (where $p = 0.9999$), for small ρ's is very close to the line that represents the case of $M/D/100$. However for high ρ's, this line is higher than that of the $M/LN/100$ system.

As in $P(Wait > 0)$, notice that the case $p = 0.75$ is very close to the $M/M/100$ system. Notice also that in the case $p = 0.5001$, $E(W_q|W_q > 0)$ is slightly higher, for each β value, than the $M/M/100$ case (as it was mentioned before, the order between the cases has changed).

The graph of $E(W_q)$ vs. β appears in Figure 20. Notice that the order of the lines in this

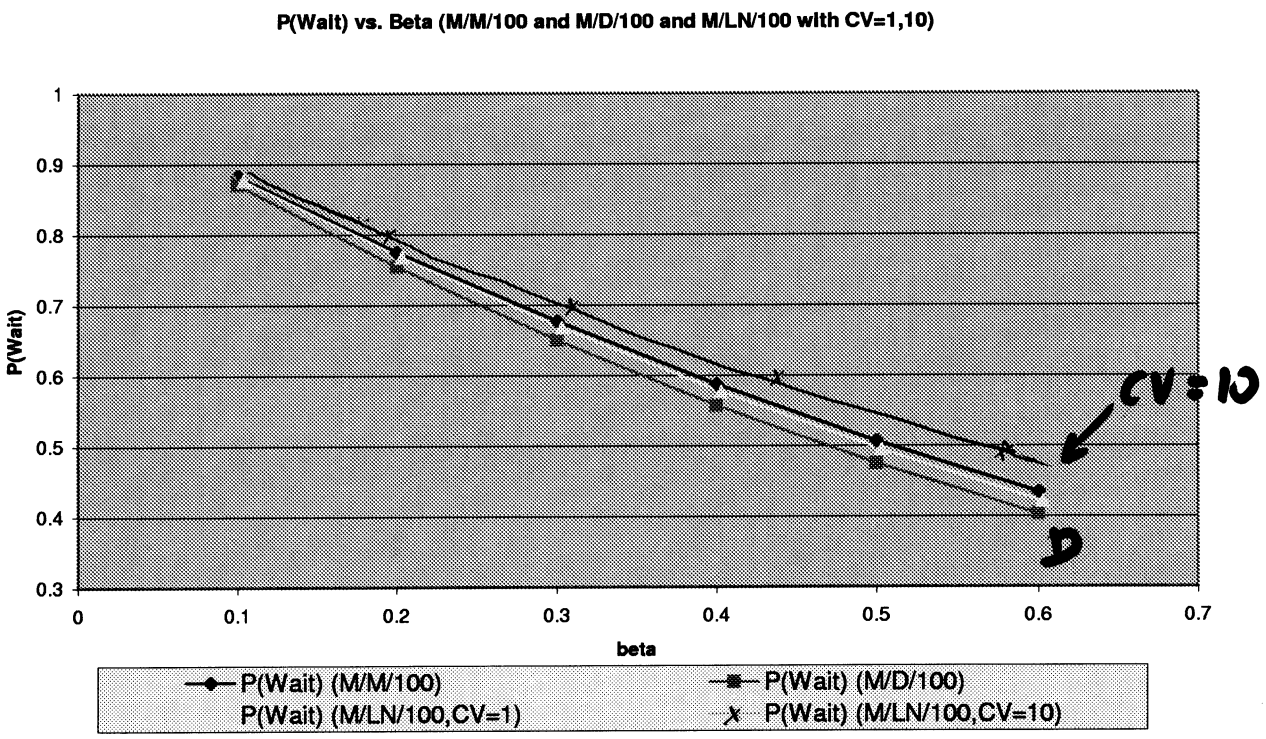


Figure 3: $P(Wait > 0)$ vs. β ($M/M/100$, $M/D/100$ and $M/LN/100$ with $CV = 1$ and $CV = 10$)

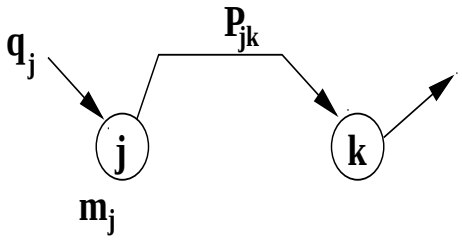
are not correct in the *Halfin-Whitt* regime. In chapter 6 we present a special service time distribution that achieved these results in a much clearer fashion.

Phase-Type Service Times (Durations).

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

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

$1 - \sum_{\ell=1}^{K} P_{k\ell}$ = 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 $\{1, 2, \ldots, K, \Delta\}$, with infinitesimal generator

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

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

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

Recall:

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

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

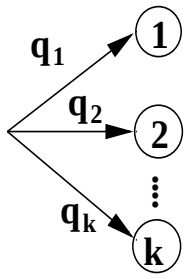
Proof. $P(T > t) = P\{X_t \neq \Delta\} = \sum_k q(e^{tR})_k = qe^{tR}1$.

Parameters:

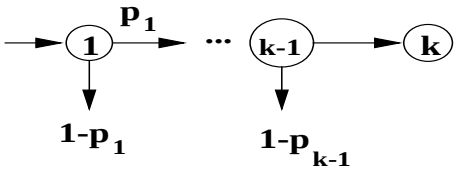
density	$f_T(t) = qe^{Rt}r$
Laplace transform	$\int_0^\infty e^{-xt}F_T(dt) = q[xI - R]^{-1}r$
nth moment	$\int_0^\infty t^n F_T(dt) = (-1)^n n!\, qR^{-n}1$
(mean $= -qR^{-1}1$)	

Special Cases:

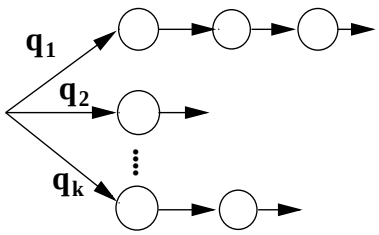
- Exponential (μ) : $R = [-\mu]$ and $q = 1$.
- Erlang: $\rightarrow \boxed{1} \rightarrow \boxed{2} \rightarrow \boxed{\text{K}}$ iid tasks / phases $\left(C^2(T) = \frac{1}{K}\right)$.
- Generalized Erlang: exponential phases in series (tandem) ($C^2 < 1$).
- Hyperexponential: K tasks in parallel (mixture)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



- 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:



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) (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 $E(T) = qRm$ $(= \sum_j q_j R_{jk} m_k)$.

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

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