Service Engineering

Class 8

Customers' (Im)Patience & Abandonment; Hazard Rates

- 14-Years Modeling Gallery.
- Customers' (Im)Patience: Introduction.
- Understanding (Im)Patience:
  Observing, Describing, Managing, Estimating, Modeling.
- Examples.
- Abandonment and (Im)Patience: Theoretical and Practical Significance.
- Modeling (Im)Patience: Patience-Time and Offered-Wait (or Time-Willing and Time-Required to Wait).
- Patience Distribution: Survival Function and Hazard Rate.
- Palm's Law of Irritation.
- Paying an Old Debt: Longest Service Times at Peak Congestion.
- Estimating Exponential Patience.
- A Patience Index.
- Probability to Abandon and Average Wait, or the
  "Law: $\mathrm{P}\{\mathrm{Ab}\} = \theta \cdot \mathrm{E}[W_q]$," and relatives.
- Estimating General (Im)Patience (Kaplan-Meier).
- Some Human (Psychological) Aspects of (Im)Patience.
- Adaptivity and Learning.
- Next: Queues – Integrating the Building Blocks.

Call Centers = Q's w/ Impatient Customers 14 Years History, or "A Modelling Gallery"

1. Kella, Meilijson: Practice ⇒ Abandonment important
2. Shimkin, Zohar: No data ⇒ Rational patience in Equilibrium
3. Carmon, Zakay: Cost of waiting ⇒ Psychological models
4. Garnett, Reiman; Zeltyn: Palm/Erlang-A to replace Erlang-C/B as the standard Steady-state model
5. Massey, Reiman, Rider, Stolyar: Predictable variability ⇒ Fluid models, Diffusion refinements
6. Ritov; Sakov, Zeltyn: Finally Data ⇒ Empirical models
7. Brown, Gans, Haipeng, Zhao: Statistics ⇒ Queueing Science
8. Atar, Reiman, Shaikhet: Skills-based routing ⇒ Control models
9. Nakibly, Meilijson, Pollatchek: Prediction of waiting ⇒ Online Models and Real-Time Simulation
10. Garnett: Practice ⇒ 4CallCenters.com
11. Zeltyn: Queueing Science ⇒ Empirically-Based Theory
12. Borst, Reiman; Zeltyn: Dimensioning M/M/N+G
13. Kaspi, Ramanan: Measure-Valued models and approximations
14. Jennings; Feldman, Massey, Whitt: Time-stable performance (ISA)

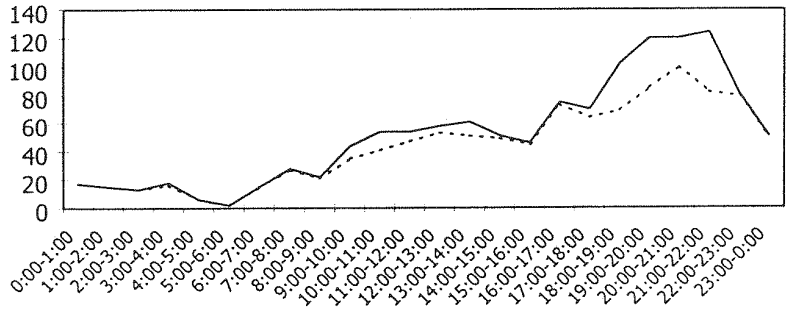
Understanding (Im)Patience

- **Observing** (Im)Patiecne – Heterogeneity:
  Under a single roof, the fraction abandoning varies
  from 6% to 40%, depending on the type of service/customer.
- **Describing** (Im)Patience Dynamically:
  Irritation proportional to Hazard Rate (Palm's Law).
- **Managing** (Im)Patience:
  - VIP vs. Regulars: who is more "Patient"?
  - What are we actually measuring?
  - (Im)Patience Index:
    "How long Expect to wait" relative to
    "How long Willing to wait".
- **Estimating** (Im)Patience: Censored Sampling.
- **Modeling** (Im)Patience:
  - The "Wait" Cycle:
    Expecting, Willing, Required, Actual, Perceived, etc.
    The case of the Experienced & Rational customer.
  - (Nash) Equilibrium Models.

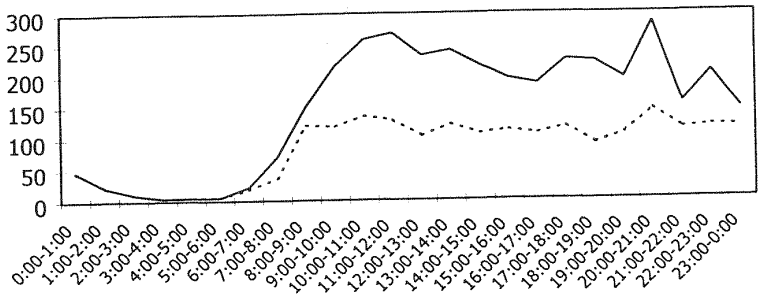
Example: "A Catastrophic situation"

Marketing Campaign at a Call Center

Average wait 72 sec, 81% calls answered (Saturday)

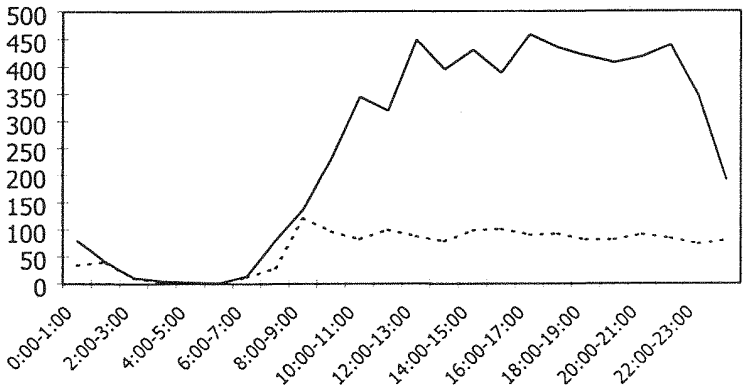


Average wait 217 sec, 53% calls answered (Thursday)



Avg. wait **376** sec, Max wait **1214** sec, **24%** calls answered (Sunday)

Note: Systems's capacity about 100 customers per hour.



% answered

מוקד שירות לקוחות

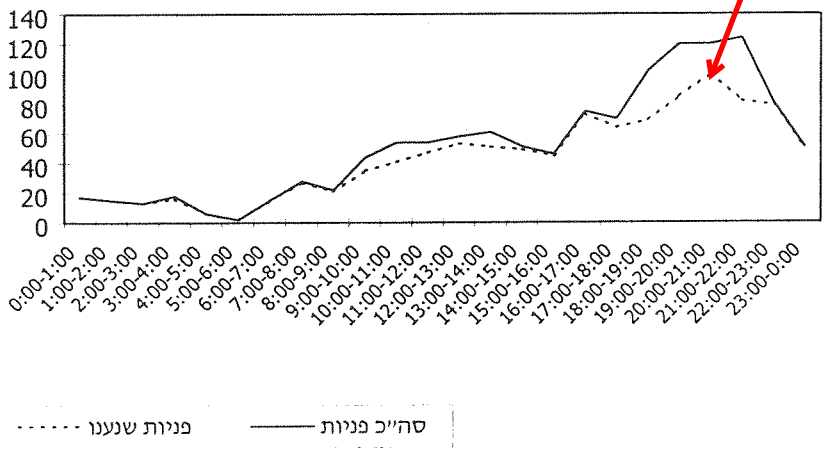
(1013) יום שבת 06/11/99

מופע הפניות והתפלגות על פני היום

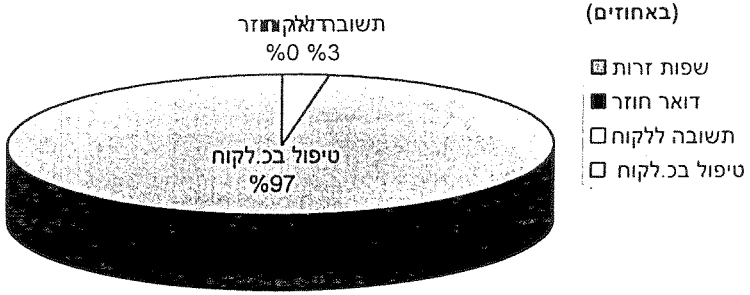
שעה	פניות שנענו	פניות שננטשו	סה"כ פניות	אחוז פניות	אחוז מענה	זמן המתנה מקסימלי (שניות)	זמן המתנה ממוצע למענה (שניות)
0:00-1:00	17	0	17	1%	100%	2	0
1:00-2:00	15	0	15	1%	100%	2	1
2:00-3:00	13	0	13	1%	100%	2	1
3:00-4:00	16	2	18	1%	89%	74	6
4:00-5:00	6	0	6	0%	100%	2	0
5:00-6:00	2	0	2	0%	100%	2	1
6:00-7:00	15	0	15	1%	100%	2	0
7:00-8:00	27	1	28	2%	96%	66	9
8:00-9:00	21	1	22	2%	95%	2	0
9:00-10:00	35	9	44	4%	80%	292	46
10:00-11:00	41	13	54	4%	76%	260	75
11:00-12:00	47	7	54	4%	87%	340	35
12:00-13:00	53	5	58	5%	91%	222	40
13:00-14:00	51	10	61	5%	84%	396	73
14:00-15:00	49	2	51	4%	96%	192	17
15:00-16:00	45	1	46	4%	98%	120	10
16:00-17:00	73	2	75	6%	97%	118	18
17:00-18:00	64	6	70	6%	91%	132	23
18:00-19:00	69	33	102	8%	68%	532	130
19:00-20:00	85	35	120	10%	71%	370	131
20:00-21:00	100	20	120	10%	83%	214	58
21:00-22:00	82	42	124	10%	66%	340	106
22:00-23:00	79	2	81	7%	98%	94	9
23:00-0:00	49	1	50	4%	98%	42	5
	1054	192	1246	100%	90%	532	33

% answered

התפלגות הפניות ע"פ היום
(סה"כ פניות מול פניות שנענו)



התפלגות חזרה ללקוח
(באחוזים)



התפלגות חזרה ללקוח

שפות זרות	0
דואר חוזר	0
תשובה ללקוח	15
טיפול בכ.לקוח	515
סה"כ	530

שעות 08:00-22:00

זמן המתנה ממוצע (שניות)	72
זמן המתנה מקס'	532
אחוז שיחות נענות	81%

סה"כ נציגים	22
זמן עבודה נטו (דקות)	7260

% answered

אחוז עבודה במיוחדים מתוך סה"כ זמן עבודה כולל	13%
אחוז האיפיונים ביחס למספר השיחות שנענו	113%

% answered

מוקד שירות לקוחות

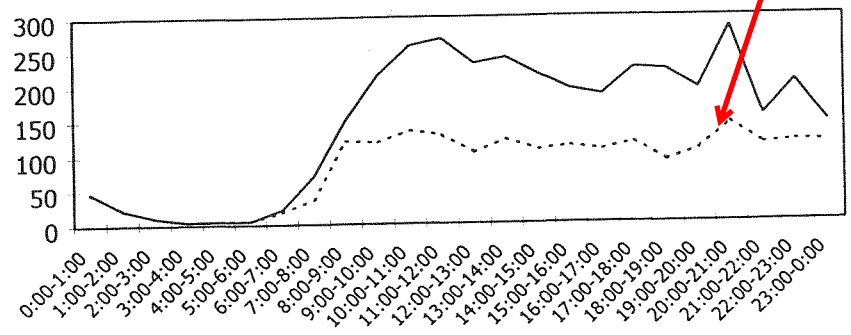
(1013) יום ה' 25/11/99

מופע הפניות והתפלגות על פני היום

שעה	פניות שנענו	פניות שנותרו	סה"כ פניות	אחוז פניות	אחוז מענה	זמן המתנה מקסימלי (שניות)	זמן המתנה ממוצע למענה (שניות)
0:00-1:00	47	0	47	1%	100%	44	2
1:00-2:00	21	0	21	1%	100%	2	0
2:00-3:00	10	0	10	0%	100%	148	15
3:00-4:00	4	0	4	0%	100%	2	1
4:00-5:00	5	0	5	0%	100%	2	0
5:00-6:00	5	0	5	0%	100%	86	18
6:00-7:00	17	4	21	1%	81%	198	50
7:00-8:00	36	34	70	2%	51%	528	268
8:00-9:00	120	31	151	4%	79%	482	100
9:00-10:00	118	98	216	6%	55%	682	255
10:00-11:00	135	123	258	7%	52%	478	283
11:00-12:00	128	140	268	8%	48%	494	307
12:00-13:00	102	130	232	6%	44%	738	350
13:00-14:00	121	119	240	7%	50%	634	292
14:00-15:00	106	110	216	6%	49%	464	250
15:00-16:00	111	83	194	5%	57%	416	188
16:00-17:00	105	81	186	5%	56%	600	253
17:00-18:00	116	108	224	6%	52%	514	226
18:00-19:00	89	133	222	6%	40%	636	461
19:00-20:00	105	89	194	5%	54%	548	254
20:00-21:00	144	140	284	8%	51%	402	197
21:00-22:00	112	42	154	4%	73%	318	109
22:00-23:00	116	88	204	6%	57%	638	191
23:00-0:00	115	29	144	4%	80%	334	46
	1988	1582	3570	100%	68%	738	172

% answered

התפלגות הפניות ע"פ היום
(סה"כ פניות מול פניות שנענו)



שעות 08:00-22:00

זמן המתנה ממוצע (שניות)	217
זמן המתנה מקס'	738
אחוז שיחות נענות	53%

סה"כ נציגים	55
זמן עבודה נטו (דקות)	18150

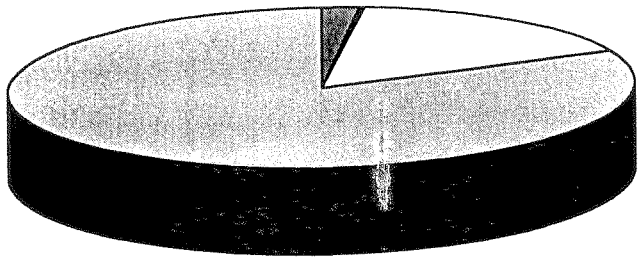
% answered

אחוז עבודה במיוחדים מתוך סה"כ זמן עבודה כולל	19%
אחוז האיפיונים ביחס למספר השיחות שנענו	124%

התפלגות חזרה ללקוח

שפות זרות	50
דואר חוזר	10
תשובה ללקוח	439
טיפול בכ.לקוח	2200
סה"כ	2699

התפלגות חזרה ללקוח
(באחוזים)



מוקד שרות לקוחות

(1013) יום א' 21/11/99

מופע הפניות והתפלגות על פני היום

שעה	פניות שנענו	פניות שנטשו	סה"כ פניות	אחוז פניות	אחוז מענה	זמן המתנה מקסימלי (שניות)	זמן המתנה ממוצע למענה (שניות)
0:00-1:00	34	46	80	1%	43%	656	230
1:00-2:00	40	0	40	1%	100%	2	0
2:00-3:00	10	1	11	0%	91%	2	1
3:00-4:00	3	2	5	0%	60%	110	37
4:00-5:00	3	0	3	0%	100%	0	0
5:00-6:00	1	0	1	0%	100%	0	0
6:00-7:00	12	2	14	0%	86%	104	27
7:00-8:00	29	50	79	1%	37%	564	312
8:00-9:00	121	16	137	2%	88%	338	66
9:00-10:00	96	135	231	4%	42%	702	377
10:00-11:00	82	263	345	6%	24%	866	657
11:00-12:00	99	220	319	5%	31%	736	528
12:00-13:00	87	361	448	7%	19%	1022	681
13:00-14:00	78	317	395	7%	20%	1214	896
14:00-15:00	98	332	430	7%	23%	944	718
15:00-16:00	101	288	389	6%	26%	864	624
16:00-17:00	90	368	458	8%	20%	994	761
17:00-18:00	92	342	434	7%	21%	812	596
18:00-19:00	81	338	419	7%	19%	1100	851
19:00-20:00	81	326	407	7%	20%	860	682
20:00-21:00	91	327	418	7%	22%	834	642
21:00-22:00	84	355	439	7%	19%	922	651
22:00-23:00	73	273	346	6%	21%	882	694
23:00-0:00	80	110	190	3%	42%	678	412
1566	4472	6038	100%	45%	1214	435	

התפלגות הפניות ע"פ היום
(סה"כ פניות מול פניות שנענו)

שעות 08:00-22:00

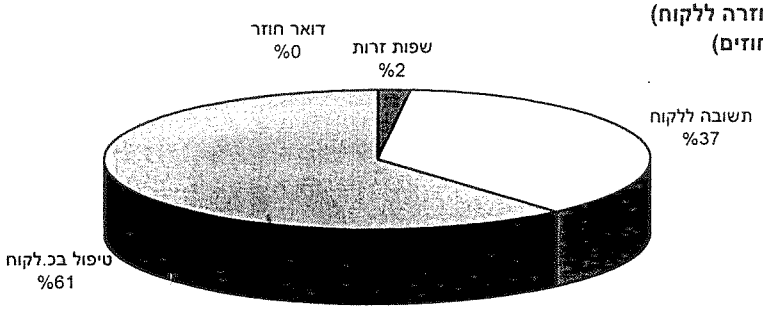
זמן המתנה ממוצע (שניות)	376
זמן המתנה מקס'	1214
אחוז שיחות נענות	24%

סה"כ נציגים	47
זמן עבודה נטו (דקות)	15510

התפלגות חזרה ללקוח

שפות זרות	20
דואר חוזר	0
תשובה ללקוח	395
טיפול בכ.לקוח	662
סה"כ	1077

התפלגות חזרה ללקוח
(באחוזים)



? !

Common Performance

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN
Skill: 37
Skill Name: !BA AUTH1

Date: 7:00 pm WED MAR 10, 1999

Acceptable Service Level: 30

DAY	ACD CALLS	AVG SPEED ANS	ABAND CALLS	AVG ABAND TIME	AVG TALK TIME	TOTAL AFTER CALL	FLOW IN	FLOW OUT	TOTAL AUX/ OTHER	AVG STAFF	% IN SERV LEVL
3/04/99	637	0:19	219	0:26	1:57	92:05	0	0	4310:06	8.7	66
3/05/99	849	0:06	135	0:06	1:35	179:58	0	0	4299:43	11.3	85
3/06/99	1330	0:11	363	0:13	1:42	280:22	0	0	5592:29	13.2	73
3/07/99	1213	0:12	358	0:18	1:46	226:20	0	0	4830:15	11.5	72
3/08/99	631	0:26	382	0:33	1:57	150:50	0	0	3743:04	7.9	49
3/09/99	570	0:40	487	0:43	1:52	148:41	0	0	3979:04	6.7	38
3/10/99	512	0:29	292	0:28	1:41	243:06	0	0	3046:00	7.9	50
SUMMARY	5742	0:18	2236	0:26	1:46	1321:22	0	0	****:**	9.6	63

Arrivals Abandons 40%

Switch Name: FDC/HAMPDEN
Skill: 46
Skill Name: !BA AUTHORIZATION

Date: 7:00 pm WED MAR 10, 1999

Acceptable Service Level: 30

DAY	ACD CALLS	AVG SPEED ANS	ABAND CALLS	AVG ABAND TIME	AVG TALK TIME	TOTAL AFTER CALL	FLOW IN	FLOW OUT	TOTAL AUX/ OTHER	AVG STAFF	% IN SERV LEVL
3/04/99	1185	0:22	479	0:31	2:08	190:16	0	0	4213:22	8.4	61
3/05/99	1805	0:05	308	0:04	1:38	337:20	0	0	4299:43	11.3	84
3/06/99	2437	0:12	642	0:12	1:51	444:03	0	0	5592:29	13.2	73
3/07/99	2260	0:13	558	0:14	1:46	326:33	0	0	4830:14	11.5	74
3/08/99	1260	0:35	676	0:28	2:06	308:19	0	0	3743:04	7.9	48
3/09/99	1126	0:40	653	0:34	2:10	250:40	0	0	3979:04	6.7	44
3/10/99	890	0:30	472	0:32	2:16	162:13	0	0	3046:00	7.9	51
SUMMARY	10963	0:19	3788	0:22	1:55	2019:24	0	0	****:**	9.6	65

30%

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN
Skill: 33
Skill Name: GA Authorization

Date: 7:01 pm WED MAR 10, 1999

Acceptable Service Level: 30

DAY	ACD CALLS	AVG SPEED ANS	ABAND CALLS	AVG ABAND TIME	AVG TALK TIME	TOTAL AFTER CALL	FLOW IN	FLOW OUT	TOTAL AUX/ OTHER	AVG STAFF	% IN SERV LEVL
3/04/99	1248	0:27	61	0:42	1:57	330:04	0	0	4390:04	9.5	72
3/05/99	1521	0:14	37	0:20	1:58	353:48	0	0	6035:35	13.0	85
3/06/99	2388	0:20	130	0:34	2:10	550:16	0	0	6369:58	14.4	76
3/07/99	1748	0:14	66	0:30	2:08	432:16	0	0	4616:11	11.7	82
3/08/99	925	0:18	50	1:00	1:53	191:06	0	0	3835:19	8.4	81
3/09/99	856	0:26	57	0:53	1:54	125:16	0	0	4388:02	8.1	73
3/10/99	959	1:15	125	1:55	1:48	186:44	0	0	4198:39	8.9	53
SUMMARY	9645	0:25	526	0:57	2:02	2169:30	0	0	****:**	10.6	76

6%

BCMS SKILL REPORT
Switch Name: FDC/HAMPDEN

Date: 7:02 pm WED MAR 10, 1999

Example: QED Operation (at most times)

ACD Report: Health Insurance (Charlotte)

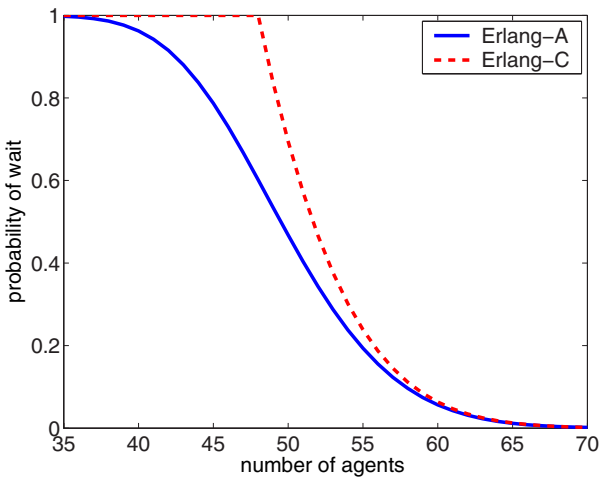
Time	Calls	Answered	Abandoned%	ASA	AHT	Occ%	# of agents
Total	20,577	19,860	**3.5%**	**30**	307	95.1%	
8:00	332	308	7.2%	27	302	87.1%	59.3
8:30	653	615	5.8%	58	293	96.1%	104.1
9:00	866	796	8.1%	63	308	97.1%	140.4
9:30	1,152	1,138	1.2%	28	303	90.8%	211.1
10:00	1,330	1,286	3.3%	22	307	98.4%	223.1
10:30	1,364	1,338	1.9%	33	296	99.0%	222.5
11:00	1,380	1,280	7.2%	34	306	98.2%	222.0
11:30	1,272	1,247	2.0%	44	298	94.6%	218.0
12:00	1,179	1,177	0.2%	1	306	91.6%	218.3
12:30	1,174	1,160	1.2%	10	302	95.5%	203.8
13:00	1,018	999	1.9%	9	314	95.4%	182.9
13:30	**1,061**	**961**	**9.4%**	**67**	**306**	**100.0%**	**163.4**
14:00	1,173	1,082	7.8%	78	313	99.5%	188.9
14:30	**1,212**	**1,179**	**2.7%**	**23**	**304**	**96.6%**	**206.1**
15:00	1,137	1,122	1.3%	15	320	96.9%	205.8
15:30	1,169	1,137	2.7%	17	311	97.1%	202.2
16:00	1,107	1,059	4.3%	46	315	99.2%	187.1
16:30	914	892	2.4%	22	307	95.2%	160.0
17:00	**615**	**615**	**0.0%**	**2**	**328**	**83.0%**	**135.0**
17:30	420	420	0.0%	0	328	73.8%	103.5
18:00	49	49	0.0%	14	180	84.2%	5.8

"The Fittest Survive" and Wait Less - Much Less!

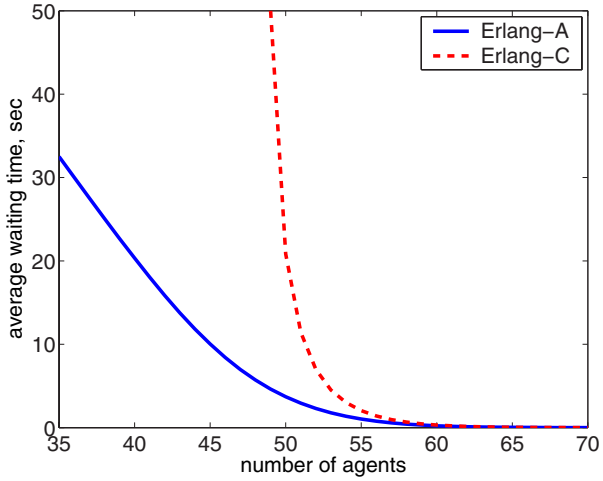
Erlang-A vs. Erlang-C

48 calls per min, 1 min average service time,
2 min average patience

probability of wait
vs. number of agents



average wait
vs. number of agents



If 50 agents:

	M/M/n	M/M/n+M	M/M/n, $\lambda \downarrow$ 3.1%
Fraction abandoning	–	3.1%	-
Average waiting time	20.8 sec	3.7 sec	8.8 sec
Waiting time's 90-th percentile	58.1 sec	12.5 sec	28.2 sec
Average queue length	17	3	7
Agents' utilization	96%	93%	93%

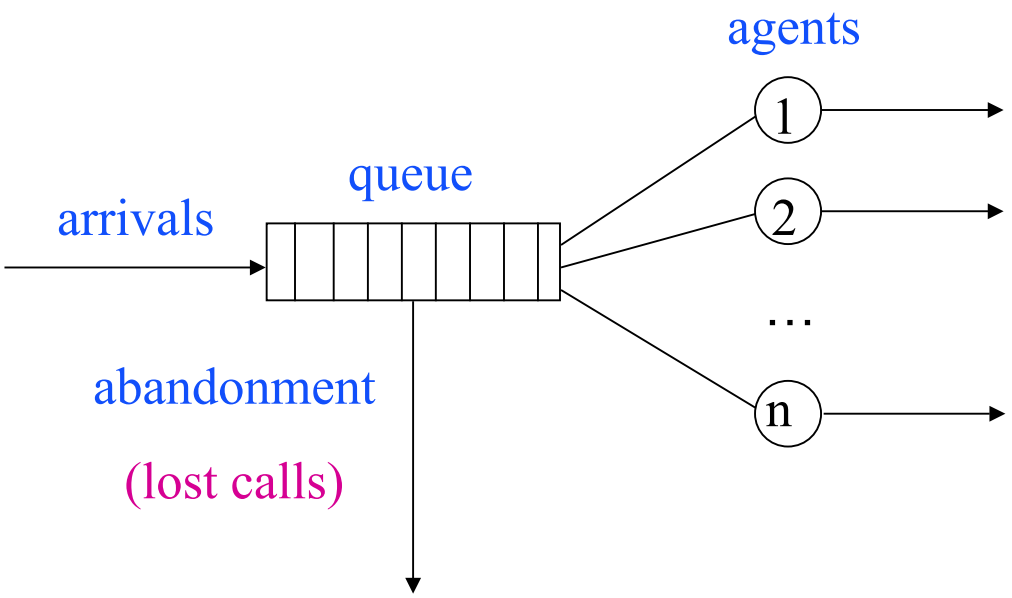
Practical Significance
Abandonment and (Im)Patience

- One of two **customer-subjective** performance measures (2^{nd}=Redials).
- **Lost business** (present losses).
- **Poor service** level (future losses).
- **1-800 costs** (present gains: out-of-pocket vs. alternative).
- **Self-selection:** the "fittest survive" and wait less (possibly much less).
- Must account for (carefully) in models and performance measures. Otherwise, distorted picture of reality, hence misleading goals and staffing levels:
  - Over-Staffing (Efficiency): If one uses models that are (im)patience-ignorant in order to determine staffing levels.
  - Under-Staffing (Quality): If one uses performance measures (eg. average delay) of only those who got served, ignoring those who abandoned. (The latter, in turn, could also lead to unacceptable protocols.)
- Robust models, numerically but, even more importantly, with respect to deviations in underlying model-assumptions (eg. service-time distribution).

Theoretical Significance
Abandonment and (Im)Patience

- **Queueing Theory**: Extend classical queueing models to accommodate call center features, notably Abandonment (and Redials).
- **Queueing Science**: The classical scientific paradigm of Measure, Model, Experiment, Validate, Refine, etc.
- **Multi-Disciplinary** Research, fusing
  Operations Research + Psychology + Marketing, through
  Models: Empirical, Mathematical (Software: 4CC), Simulation, in
  steady-state (Erlang-A), transience (Fluid), (Nash) equilibrium.
- **Applications** beyond Call Centers:
  - **VRU/IVR:** Opt Out Rate (**OOR**) to a live agent;
  - **Internet:** 60% and more abandon in mid-trasaction;
  - **Multi-Media** Contact Centers: eg. Chatting (completely open);
  - **Hospitals:** Left Without Being Seen (**LWBS**); in Emergency Departments (ED) can reach 5-10% (and then?).
  - **Other services:** Abandoning a bus station to take a taxi, ... , more?

(Im)Patience in Models: (Im)Patience-Time & Offered-Wait



- **(Im)Patience Time τ** (random variable/distribuion):
  Time a customer is **willing to wait** for service.
- **Offered Wait V**:
  Time a customer **must wait** for service;
  equivalently, waiting time of a customer with infinite patience.
- **Actual wait W** $= \min\{\tau, V\}$.
- If $\tau < V$, customer **abandons** (after waiting τ);
  otherwise ($\tau \geq V$), **gets service** (after waiting V);

Predicting Performance with Models

Model **Primitives**:

- **Arrivals** to service (stochastic process, eg. Poisson)
- **(Im)Patience** while waiting τ (r.v. $\equiv$ distribution)
- **Service** times (r.v., eg. Exponential, LogNormal)
- **# Servers / Agents** (parameter, sometimes r.v.)

Model **Output**: **Offered-Wait** V (r.v.)

Operational Performance Measure calculable in terms of (τ, V):

- eg. Average Wait $= \mathrm{E}[\min\{\tau, V\}]$
- eg. % Abandonment $= \mathrm{P}\{\tau < V\}$
- eg. Average Wait of Served (ASA) $= \mathrm{E}[V | \tau > V]$

Application: **Staffing – How Many Agents?**
(vs. When? Who?)

- The Mathematical Model (Palm, Erlang-A), based on an Empirical Model
- Base for software implementation (4CallCenters), and Simulation

Designing a Call Center with Impatient Customers

O. Garnett* A. Mandelbaum*† M. Reiman ‡

March **26**, 2002

ABSTRACT. The most common model to support workforce management of telephone call centers is the $M/M/N/B$ model, in particular its special cases $M/M/N$ (Erlang C, which models out busy-signals) and $M/M/N/N$ (Erlang B, disallowing waiting). All of these models lack a central prevalent feature, namely that impatient customers might decide to leave (abandon) before their service begins.

In this paper we analyze the simplest abandonment model, in which customers' patience is exponentially distributed and the system's waiting capacity is unlimited ($M/M/N+M$). Such a model is both rich and analyzable enough to provide information that is practically important for call center managers. We first outline a method for exact analysis of the $M/M/N+M$ model, that while numerically tractable is not very insightful. We then proceed with an asymptotic analysis of the $M/M/N+M$ model, in a regime that is appropriate for large call centers (many agents, high efficiency, high service level). Guided by the asymptotic behavior, we derive approximations for performance measures and propose "rules of thumb" for the design of large call centers. We thus add support to the growing acknowledgment that insights from diffusion approximations are directly applicable to management practice.

*Davidson Faculty of Industrial Engineering and Management, Technion, Haifa 32000, ISRAEL.

†Research supported by the fund for the promotion of research at the Technion, by the Technion V.P.R. funds - Smoler Research Fund, and B. and G. Greenberg Research Fund (Ottawa), and by the Israel Science Foundation (grant no. 388/99).

‡Bell Laboratories, Murray Hill, NJ 07974, USA.

- In website
- Published in MSOM, 2003

4CallCenters™

Personal Optimization Tools for Call Centers

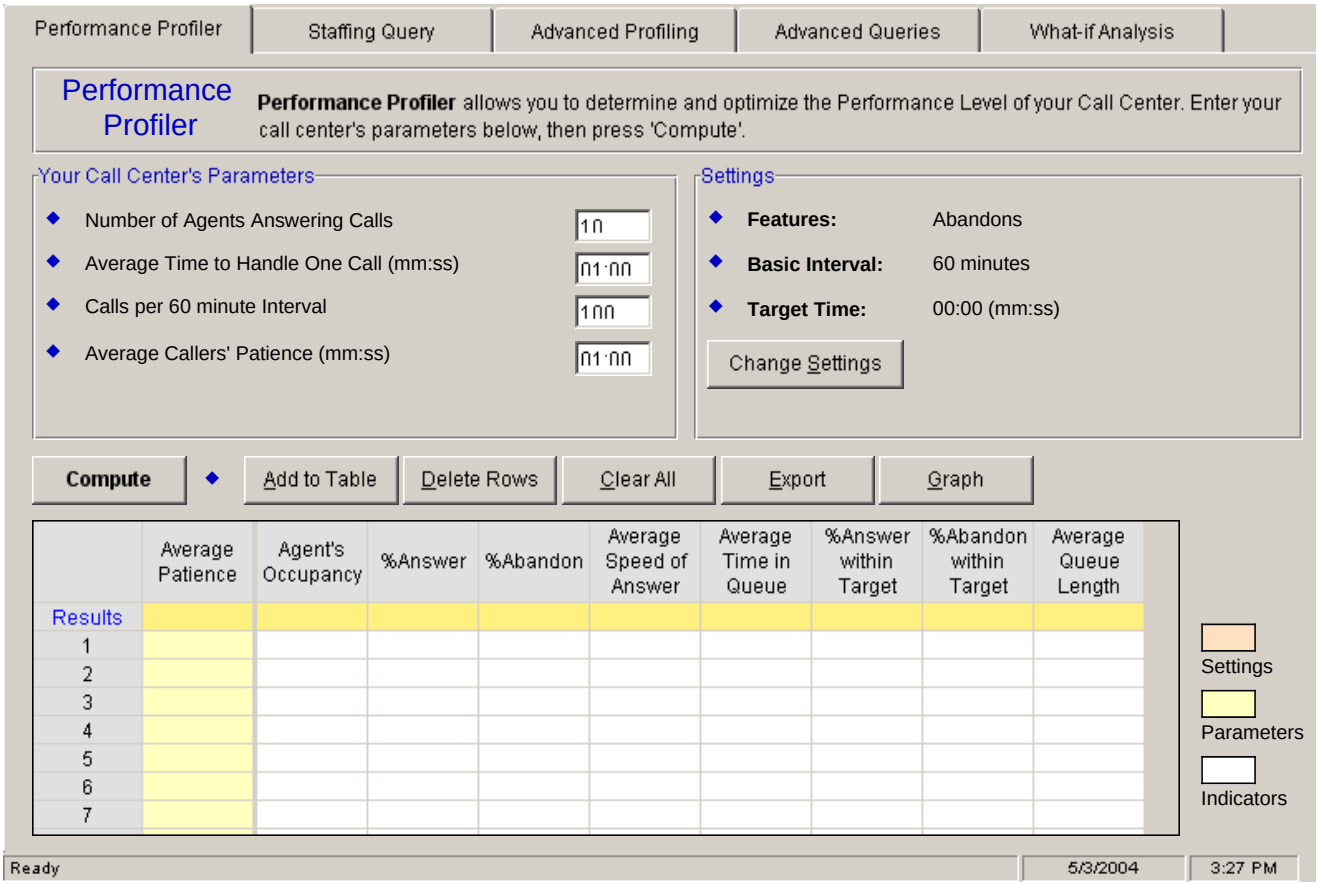
Downloads:

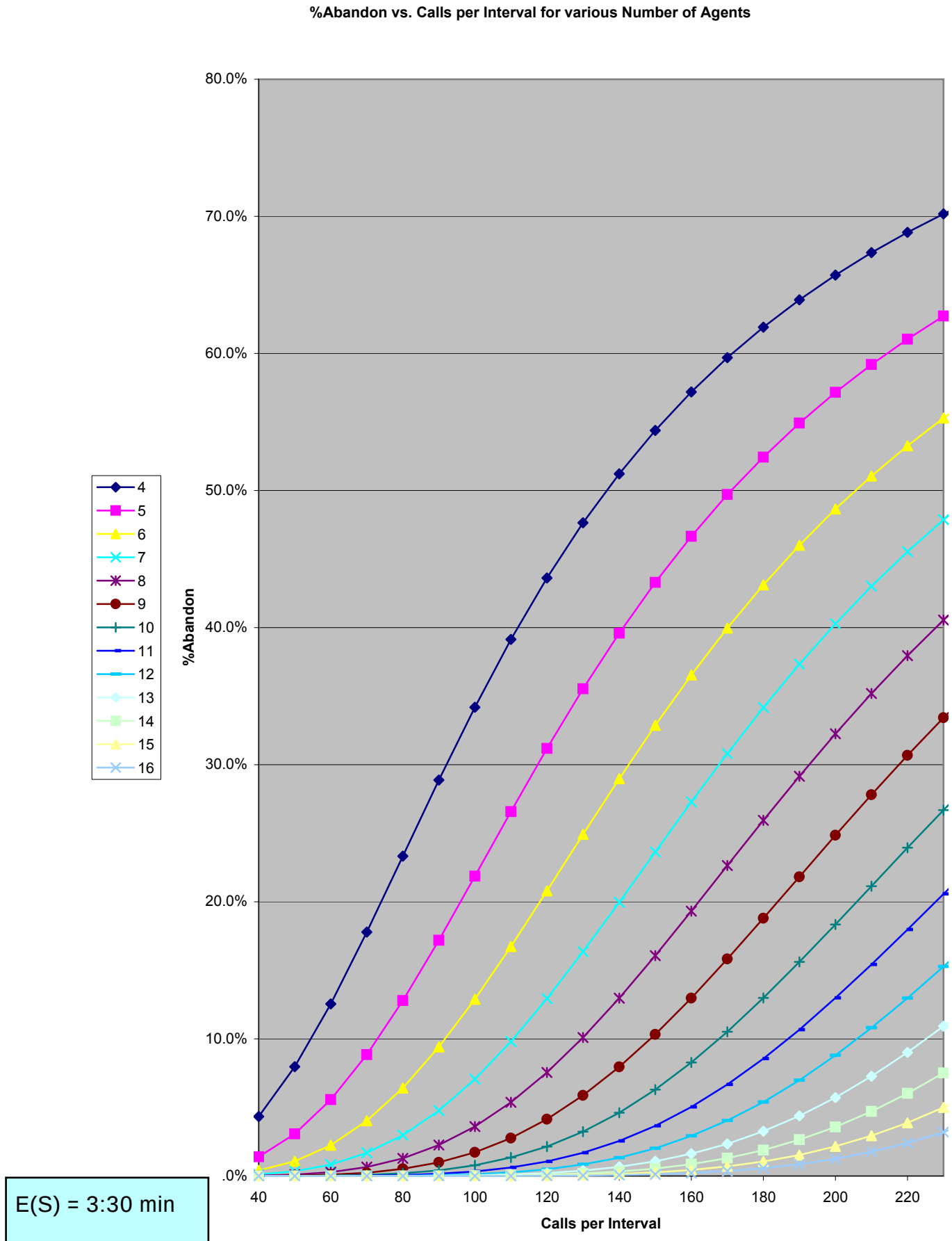
1. 4CallCenters v2.01(zip file- **5.4mb**)
   Desktop application offering personal profiling and optimization tools.

   - **For installation: Download the zip file, open it, activate setup.exe and follow the instructions.**
   - **To uninstall the installed software: Go to Start/Programs/4CallCenters v2.01/Uninstall 4CallCenters v2.01**

2. 4CallCenters v2.01 - Help Document (90**kb**)
   Word document containing the 4CallCenters application's help pages.

QSetup





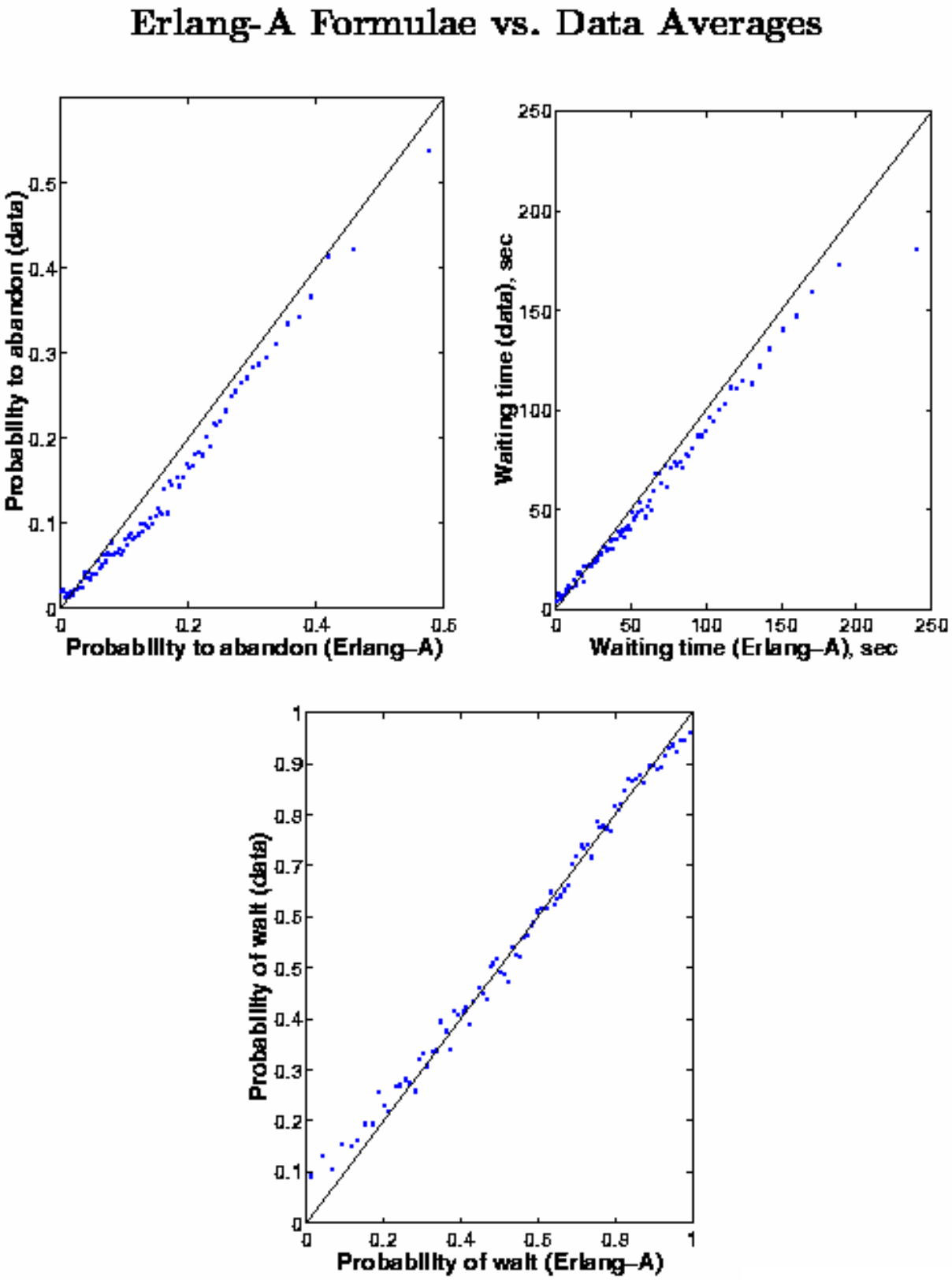
E(S) = 3:30 min

E(R) = 6:00 min

Interval = 1 **hour**

Fitting a Simple Model to a Complex Reality

13.



Measuring Patience: Censored Data

Israeli Bank Data

Statistics	Average wait	Interpretation
360K served (80%)	2 min	? **must** wait
90K abandoned (20%)	1 min	? **willing** to wait

Interpretation is wrong!

Both waiting times are **censored**:

- If customer abandoned, patience is known: $\tau = W$.
- If customer served, only a lower-bound known: $\tau > W$.

To estimate the distribution of τ and V, must **"un-censor"**:
How? Later, via techniques from Statistical **Survival Analysis**.

Censoring prevalent:

- Recall "length of stay of elderly people in institutional long-term care", when we studied phase-type service times;
- Medical Trials (Source of Terminology): duration between successive recurrences of a disease,...
- Insurance: durations between accidents,...
- Social Sciences: duration of marriage, time to find a job,...
- Marketing: duration between successive purchases of a product, ...

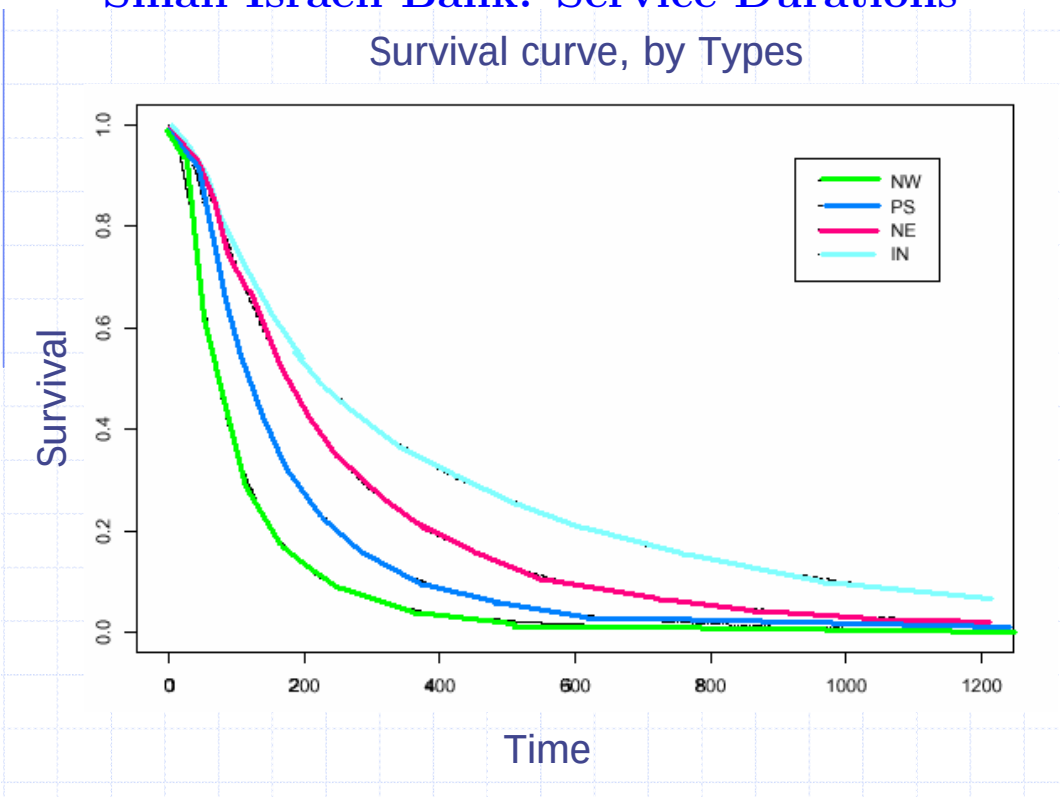
Survival Function & Stochastic Order

Survival Function: $S(t) = \mathrm{P}\{X > t\} = 1 - F(t)$.

Stochastic Order:

$X \overset{\text{st}}{\leq} Y \Leftrightarrow \mathrm{P}\{X > t\} \leq \mathrm{P}\{Y > t\} \Leftrightarrow S_X(t) \leq S_Y(t)$
for all t.

Small Israeli Bank: Service Durations

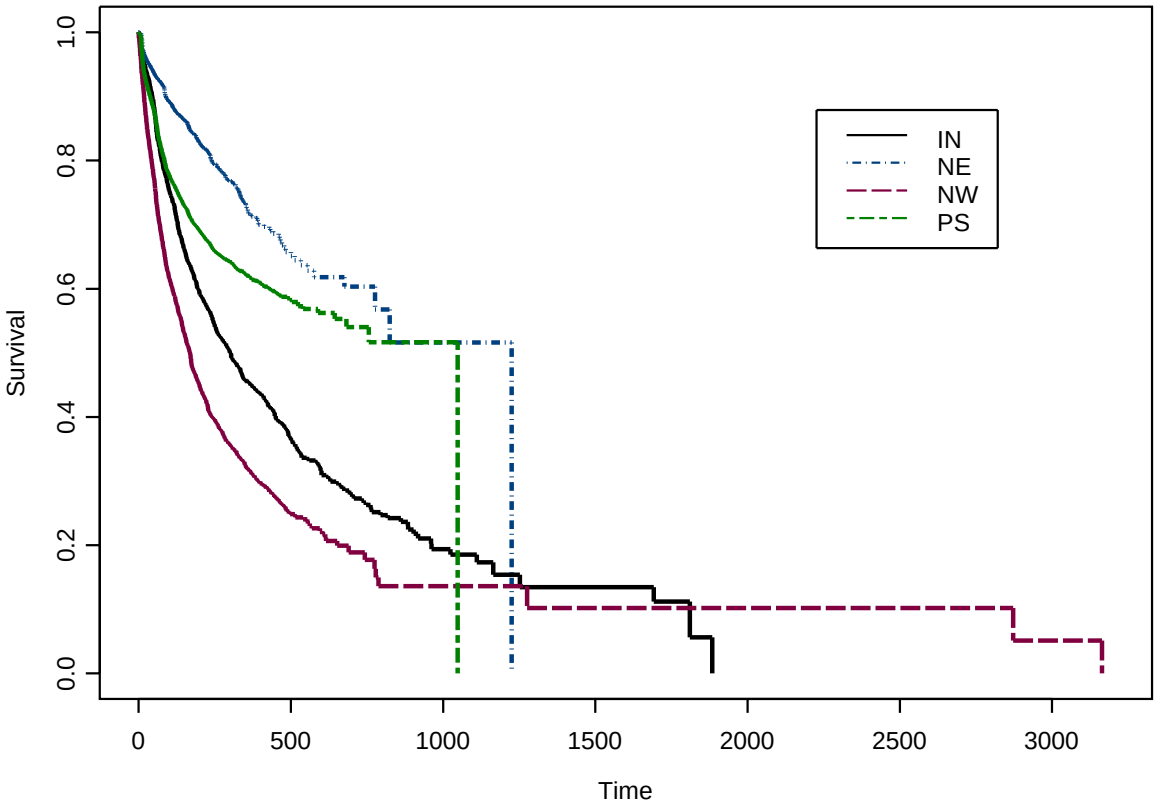


Claim: $X \overset{\text{st}}{\leq} Y \Rightarrow \mathrm{E}[X] \leq \mathrm{E}[Y]$.

Fact: Shorter ($\overset{\text{st}}{\leq}$) service times $\Rightarrow$
less abandonment and shorter waits.

(Im)Patience: Examples of Survival Functions

Small Israeli Bank: (Im)Patience Times



Fact: Shorter ($\overset{\text{st}}{\leq}$) patience times $\Rightarrow$ more abandonment and shorter waits.

Modelling (Im)Patience: Hazard Rates

For $X \geq 0$, an absolutely-continuous r.v., define its
Hazard Rate function to be $h \triangleq f/S$, namely

$$h(t) \triangleq \frac{f(t)}{1-F(t)}, \quad t \geq 0;$$

f = Density function of X,
S = Survival function of X $(S = 1 - F)$,
F = Distribution function of X.

Intuition: $\mathrm{P}\{X \leq t + \Delta | X > t\} \approx h(t) \times \Delta$.

In Discrete-Time: $h(t) = \mathrm{P}\{X = t | X \geq t\}$, $t = 0, 1, \ldots$

Characterizes the distribution:

- Continuous time: $S(t) = e^{-\int_0^t h(u)du}$, $t \geq 0$.
- Discrete time: $S(t) \triangleq \mathrm{P}\{X > t\} = \Pi_{i=0}^{t}[1-h(i)]$, $t = 0, 1, ...$
- Constant Hazard iff Memoryless (Exponential / Geometric)

Estimation: Natural in discrete-time.
In continuous-time, via discrete approximation:

1. Partition time into $0 = t_0 < t_1 < t_2 < \ldots$ (dense "enough");
2. Estimate $\hat{h}(t_i) = \frac{\#\text{ "Events" during } [t_i, t_{i+1})}{\#\text{ "At-Risk" at } t_i}$, $i = 0, 1, \ldots$;
3. Interpolate $\hat{h}(0), \hat{h}(t_1), \hat{h}(t_2) \ldots$.

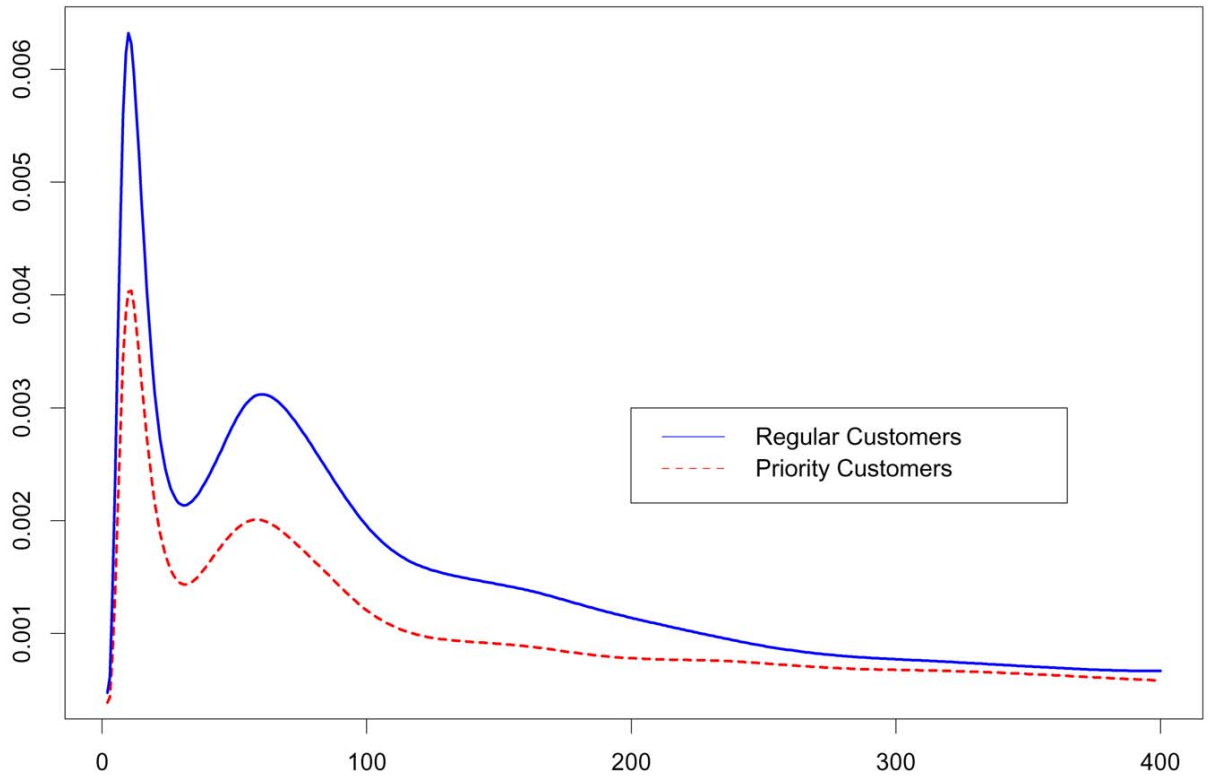
Ordering: Hazard-rate order $(\overset{\text{hr}}{\geq})$ implies Stochastic order $(\overset{\text{st}}{\leq})$.

Hazard Rate: Natural Dynamic Model of (Im)Patience

- Palm's Axiom (1940's): Hazard Rate(t) $\propto$ Irritation(t); Estimated (Im)Patience based on a sample of unlucky customers who called a broken communication-switch and got stuck, till abandoning (hence no censoring).
- Constant hazard rate (Exponential (im)patience): benchmark;
- Increasing hazard rate (IFR): Impatience $\uparrow$ while waiting;
- Decreasing hazard rate (DFR): Patience $\uparrow$ while waiting;
- Other shapes: Bathtab (decreasing, then increasing), or vice versa: both occur for (im)patience.
- More precise tail-description (vs. cdf, density).

Palm's Law of Irritation (1943-53): $\propto$ Hazard-Rate of (Im)Patience Distribution

Small Israeli Bank (1999):
Regular vs. Priority (VIP) Customers

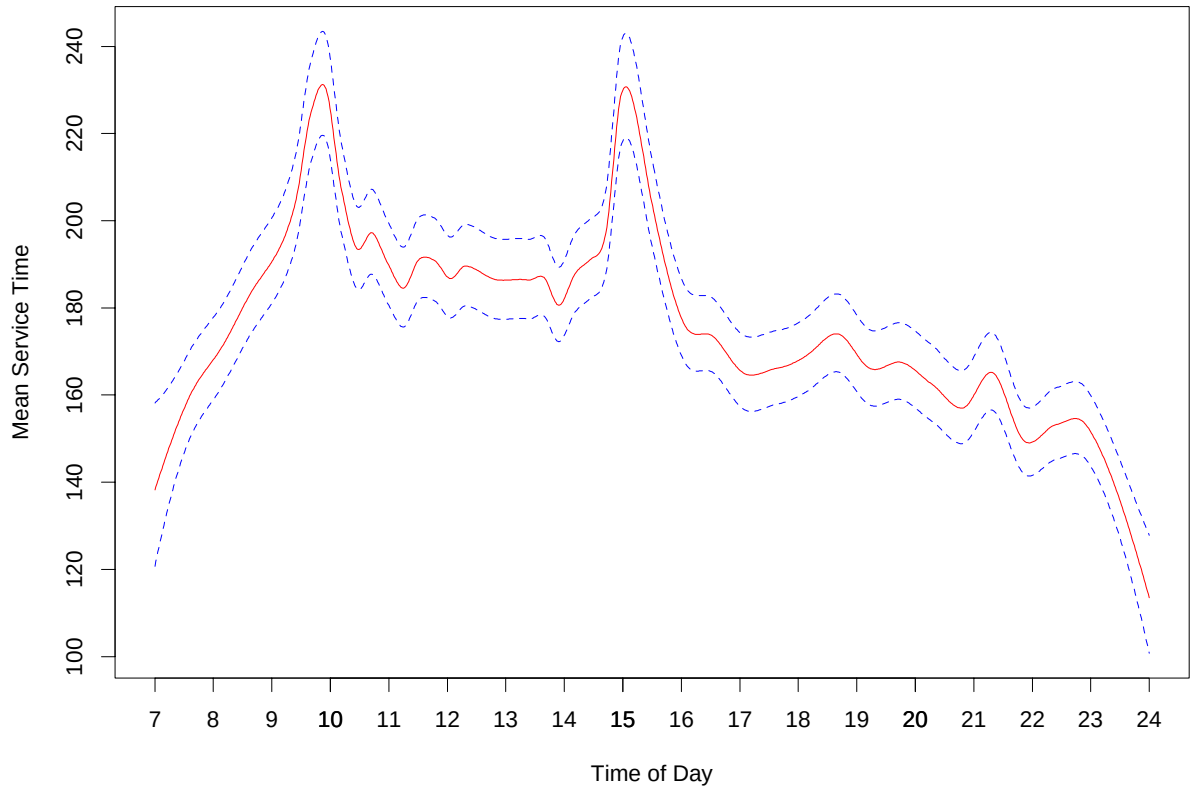


Observations:

- Who is **more patient** – **Regular** or **VIP** ? (stochastically);
- Why the **two peaks** of abandonment (at outset, 60 seconds)?
  - Possibly three **phases of (im)patience**;
  - Possibly three **types of customers**;
  - Actually **human psychology**.

Old Debt: Longest Services at Peak Times ?

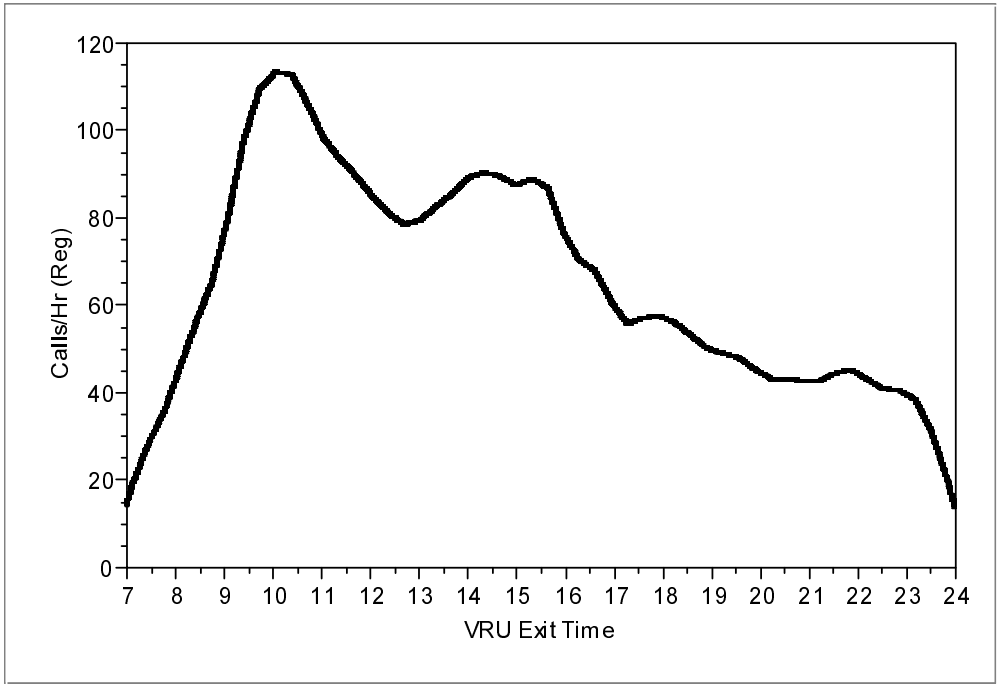
Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) ($n = 42613$)



Peak Loads at 10:00 and 15:00

Arrivals: Inhomogeneous Poisson

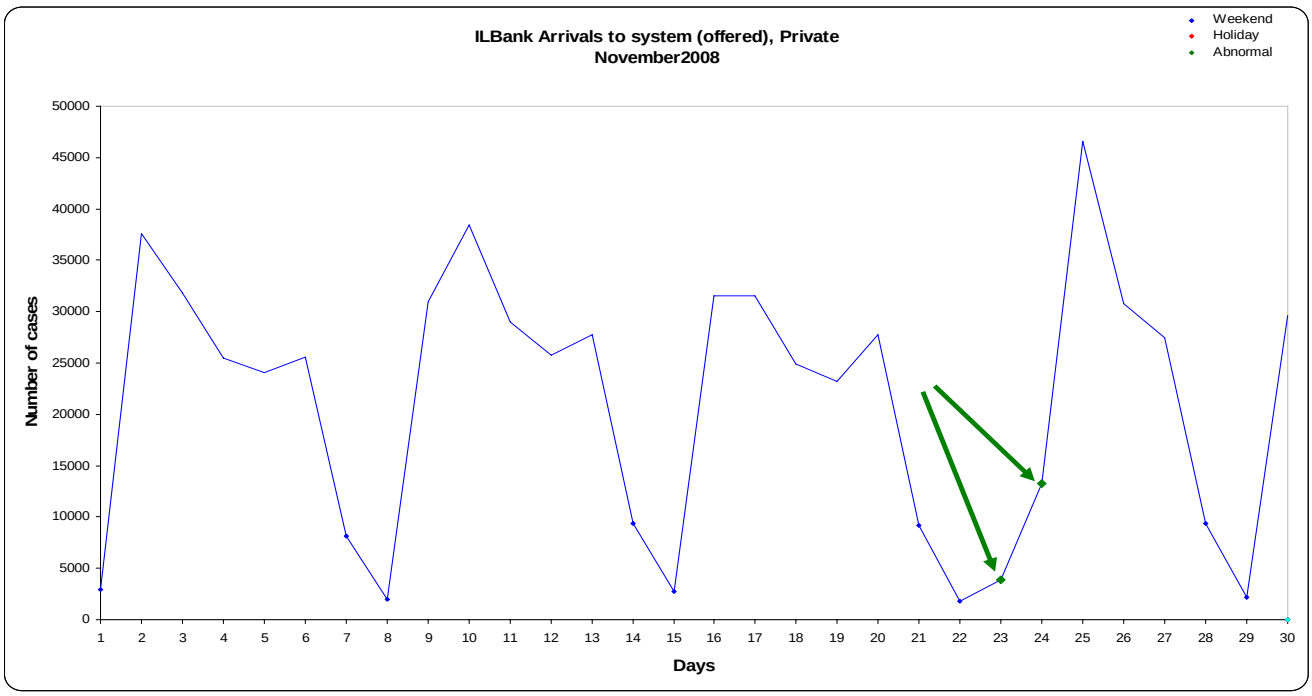
Figure 1: Arrivals (to queue or service) – "Regular" Calls

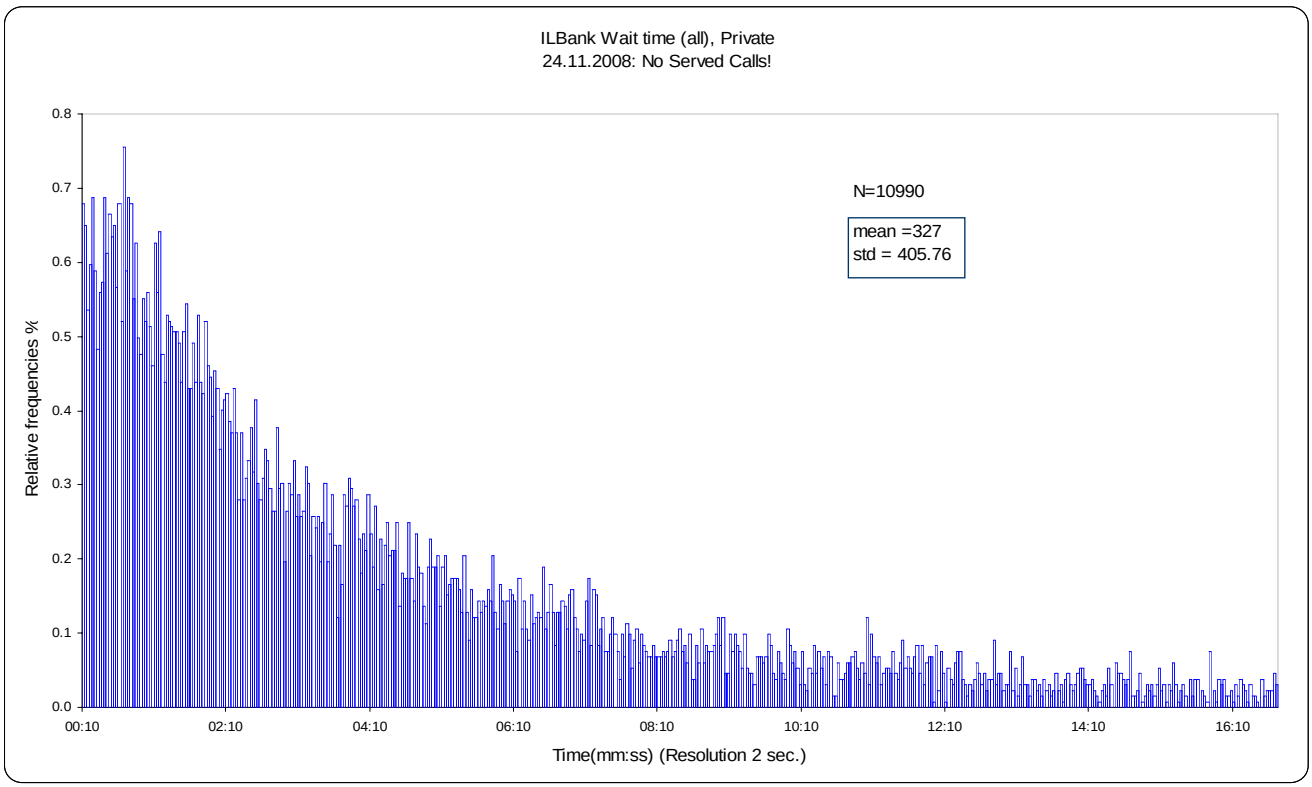


Empirical Adventures: (Im) Patience

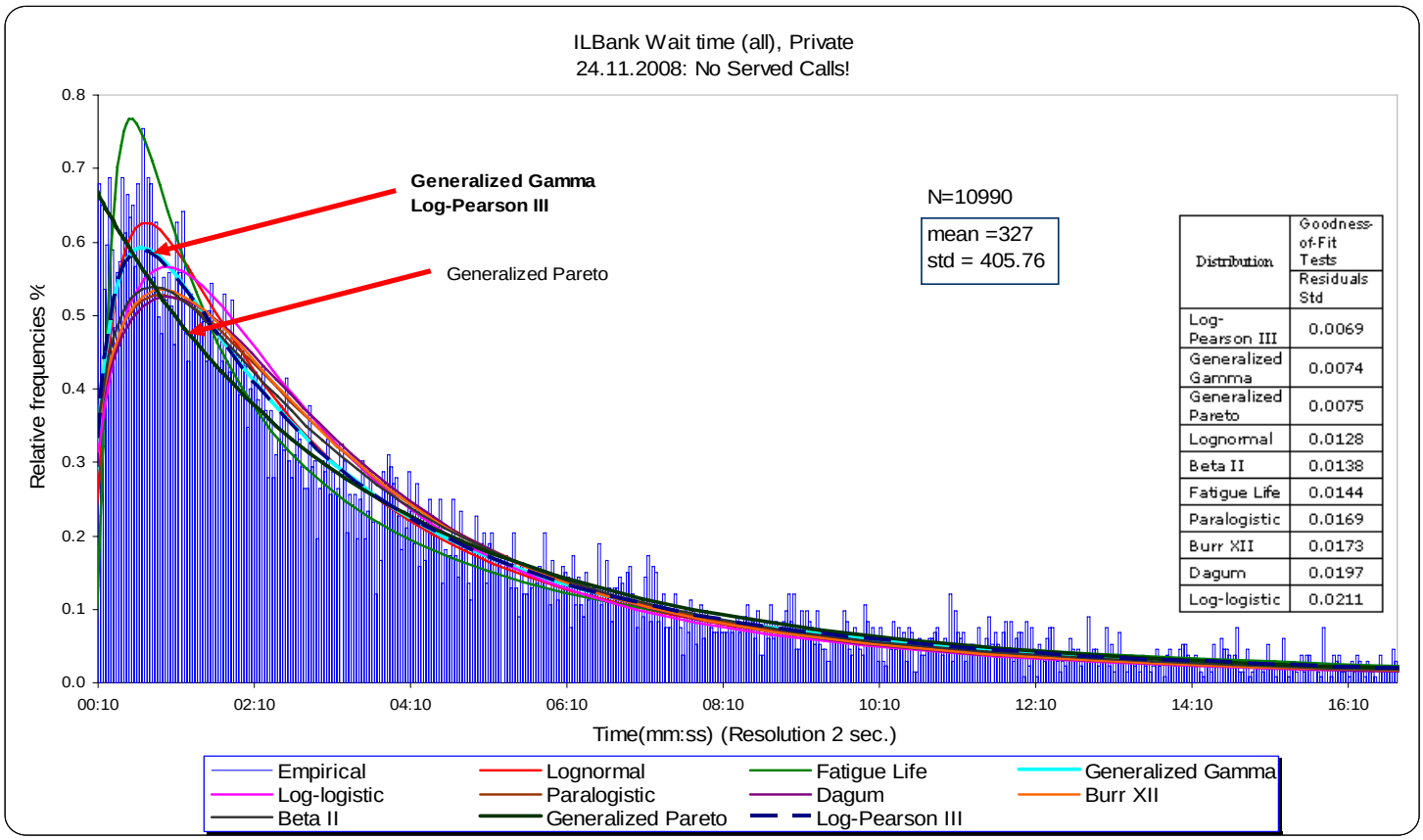
Service Engineering
May 2011

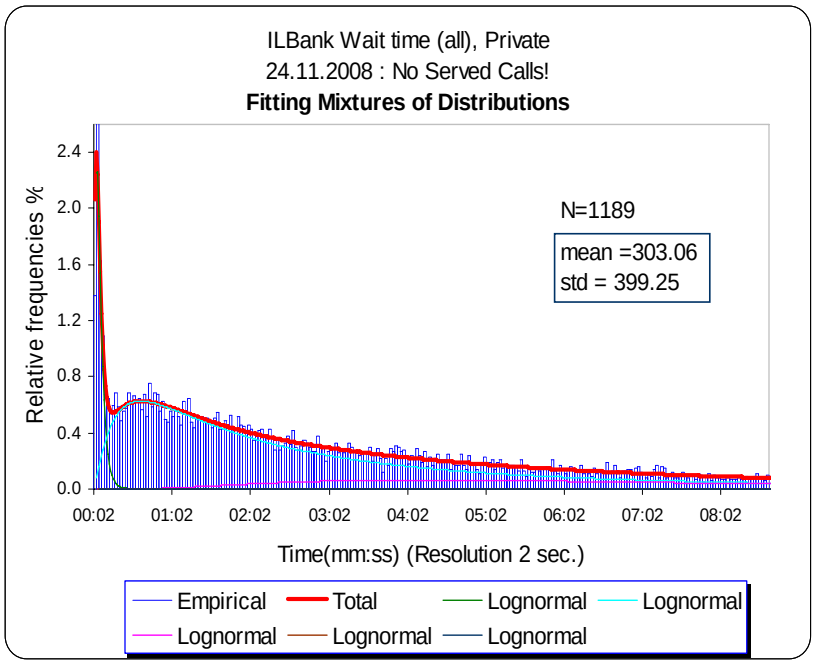
(Im)Patience (Raw)

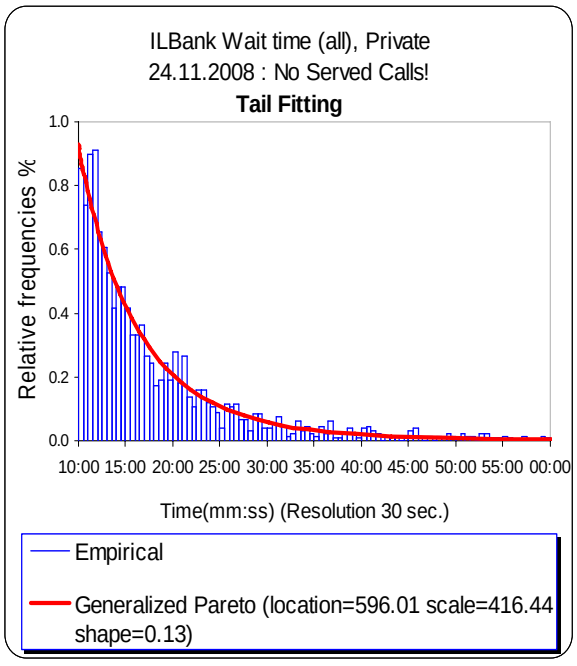




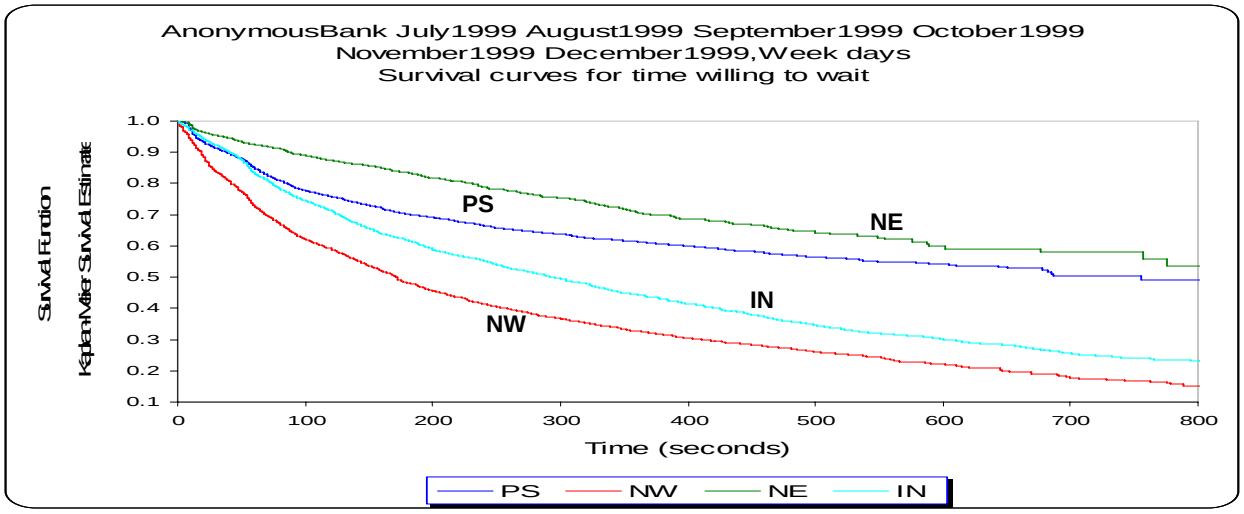
Distribution Fitting

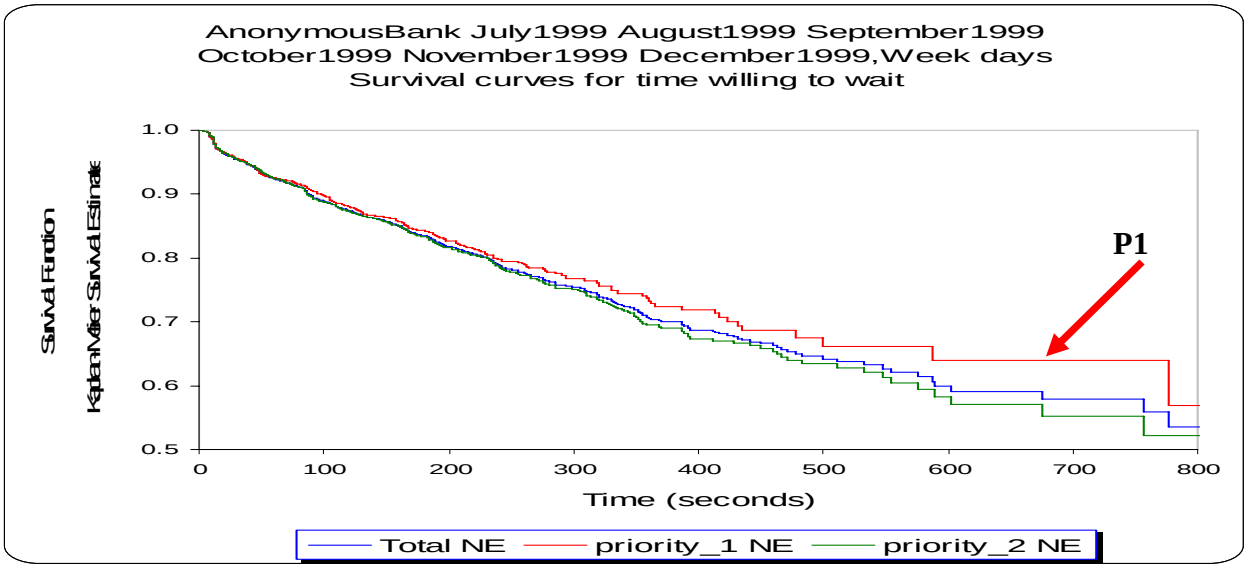


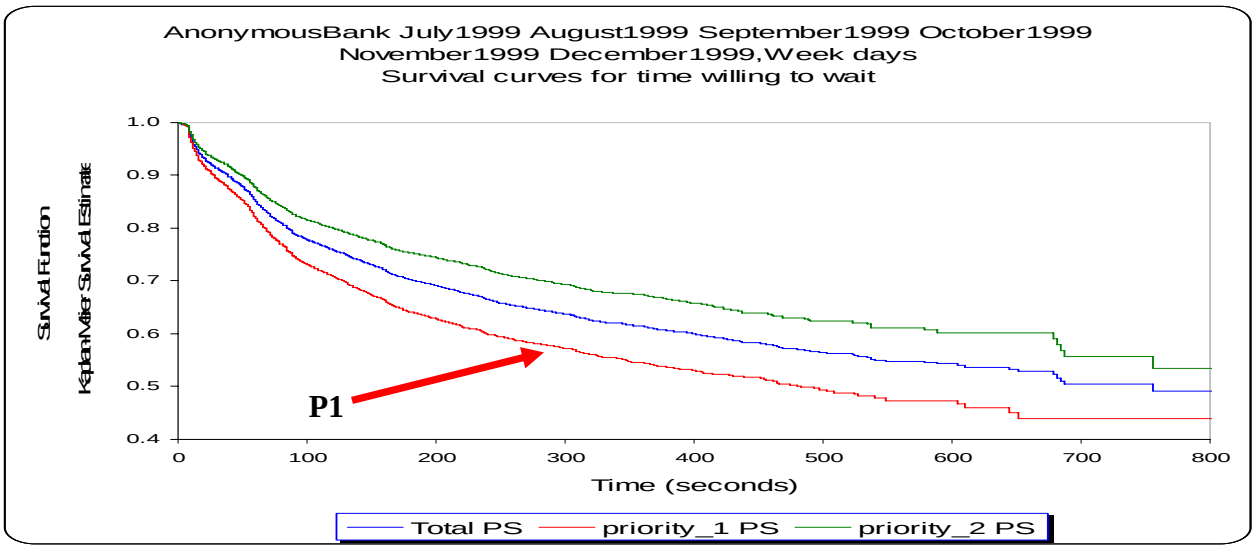




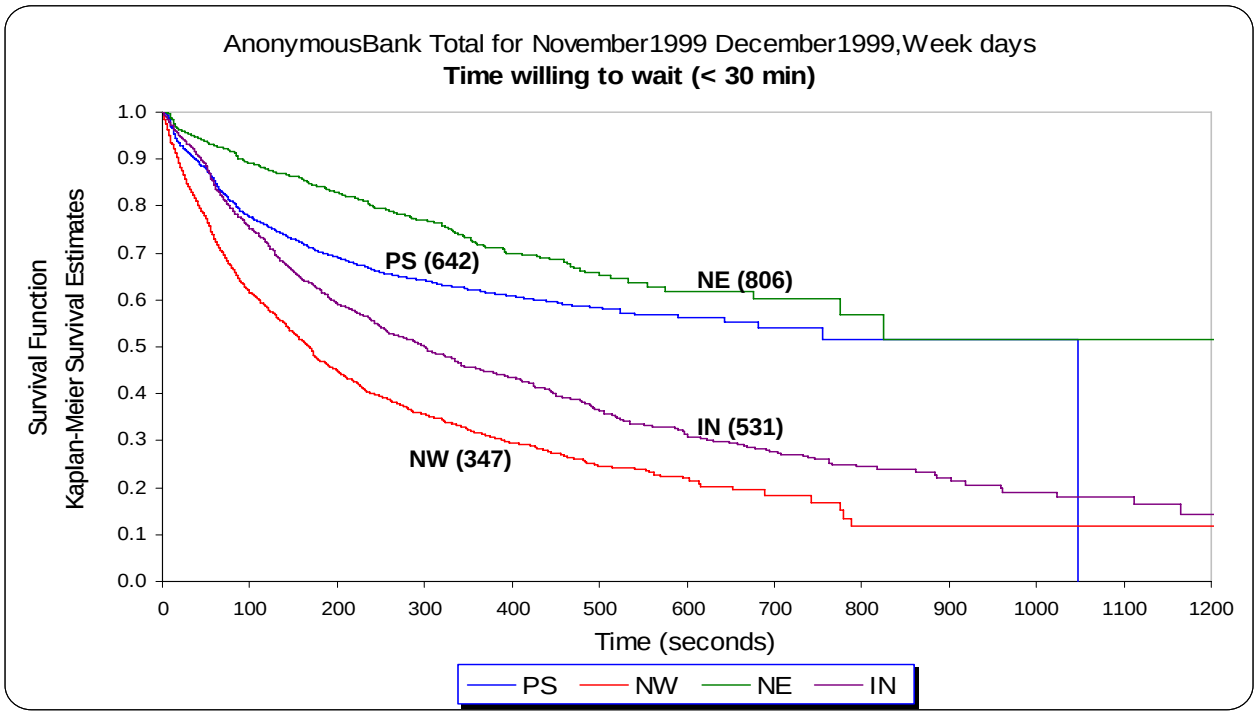
Survival Functions of (Im)Patience

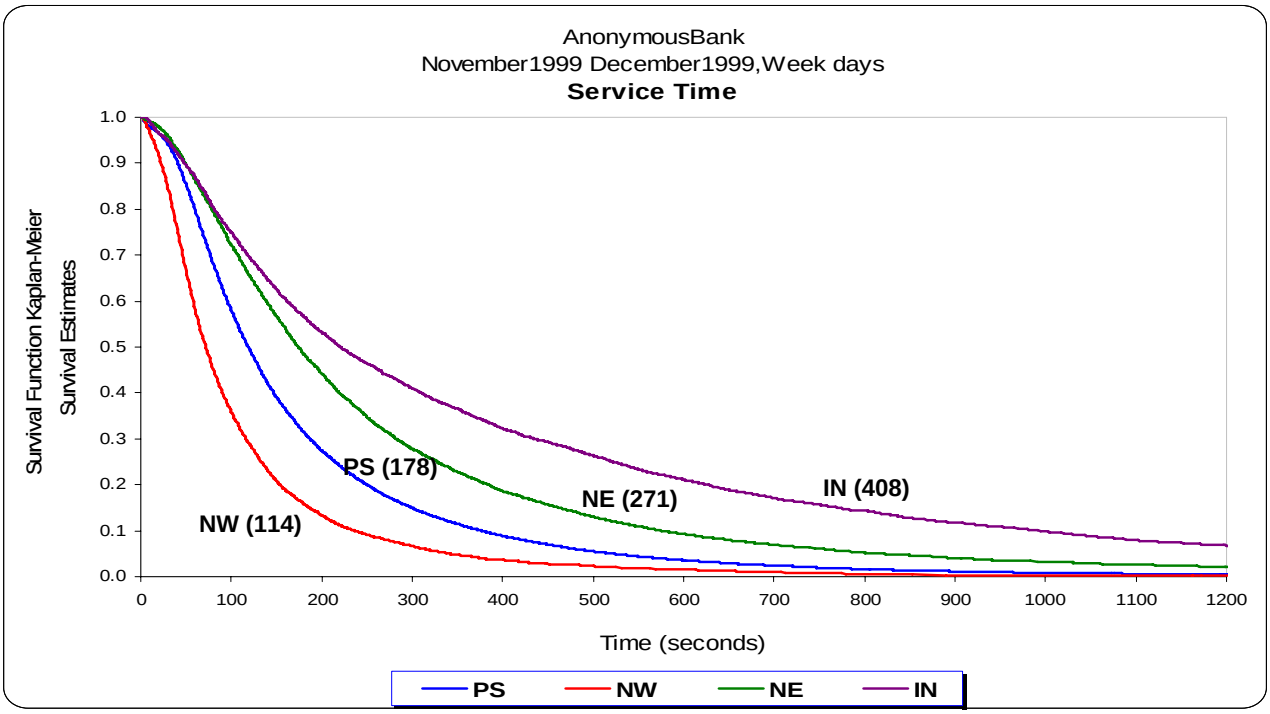




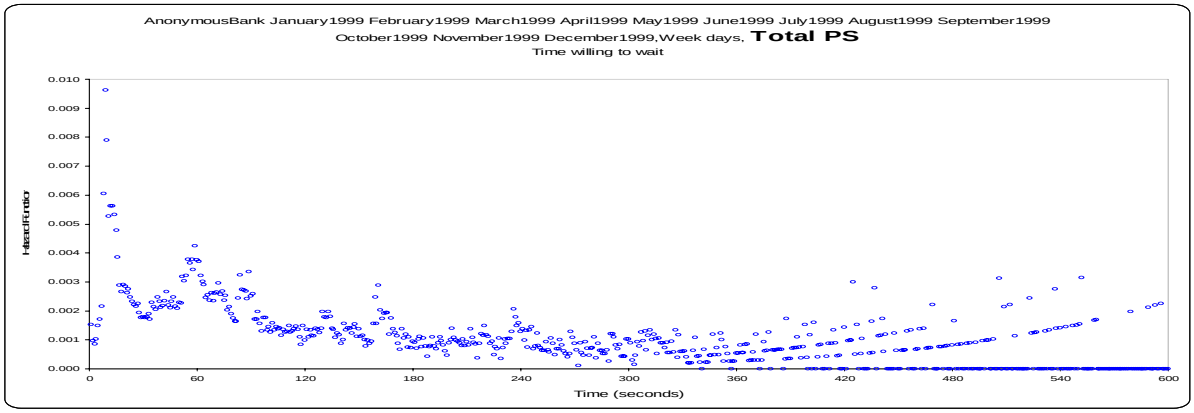


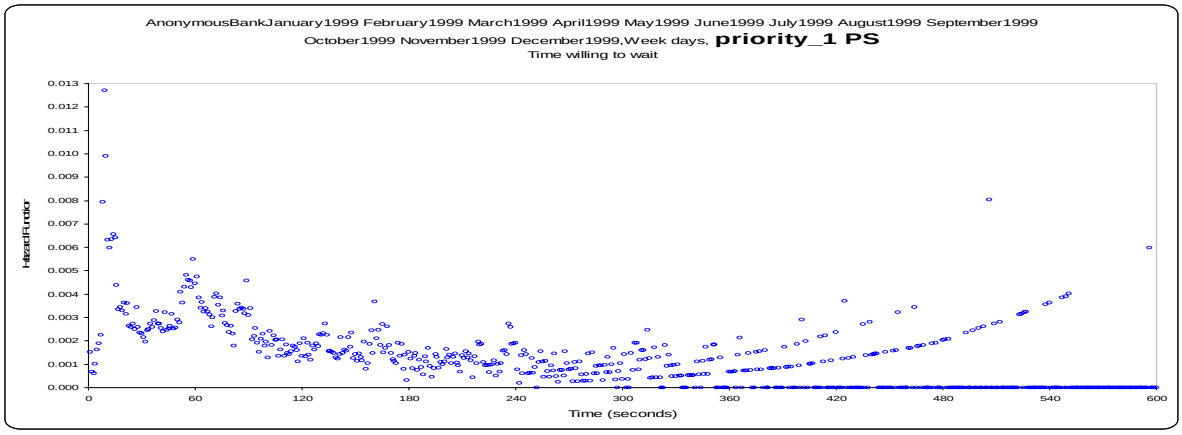
Patience vs. Service Durations (Stochastic Order)

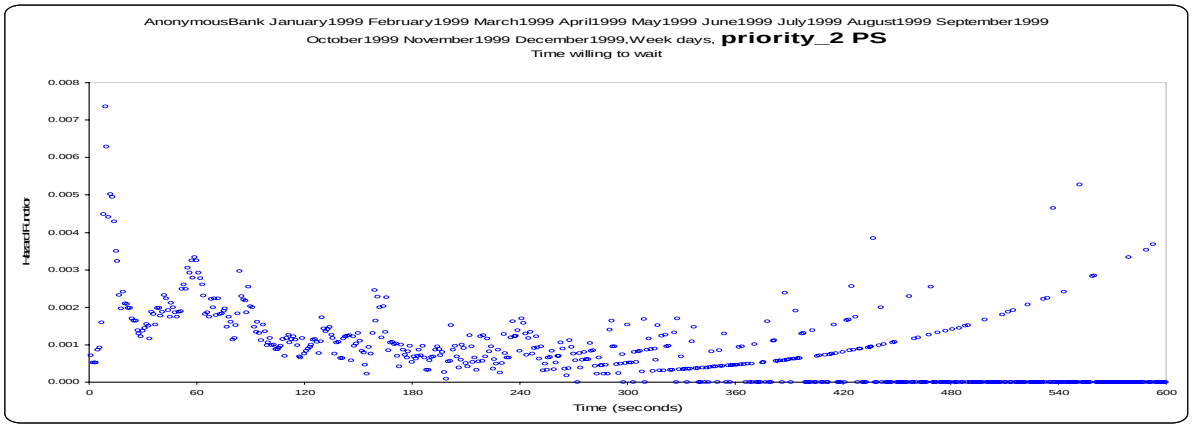




Empirical Hazard Rates

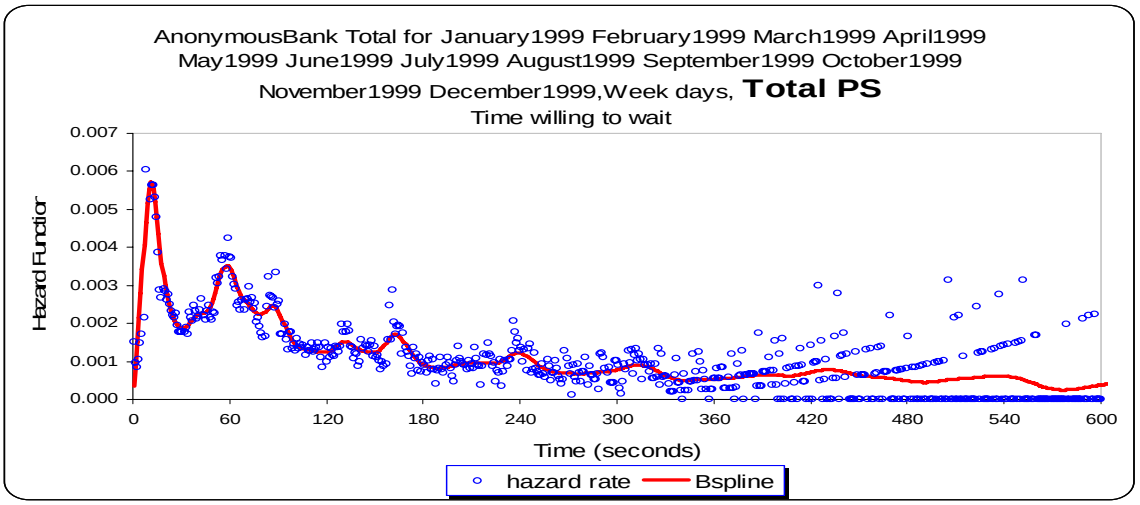


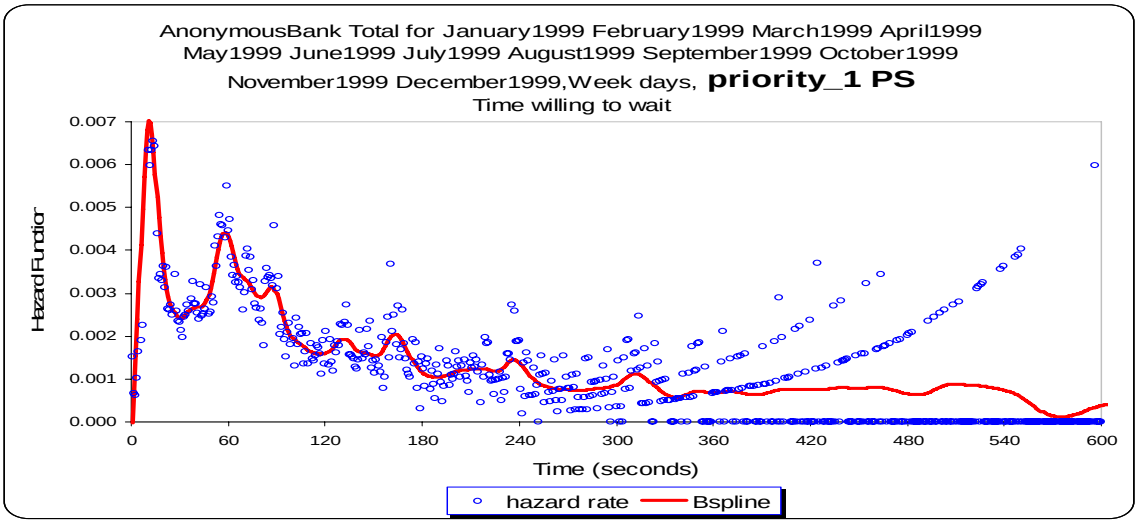


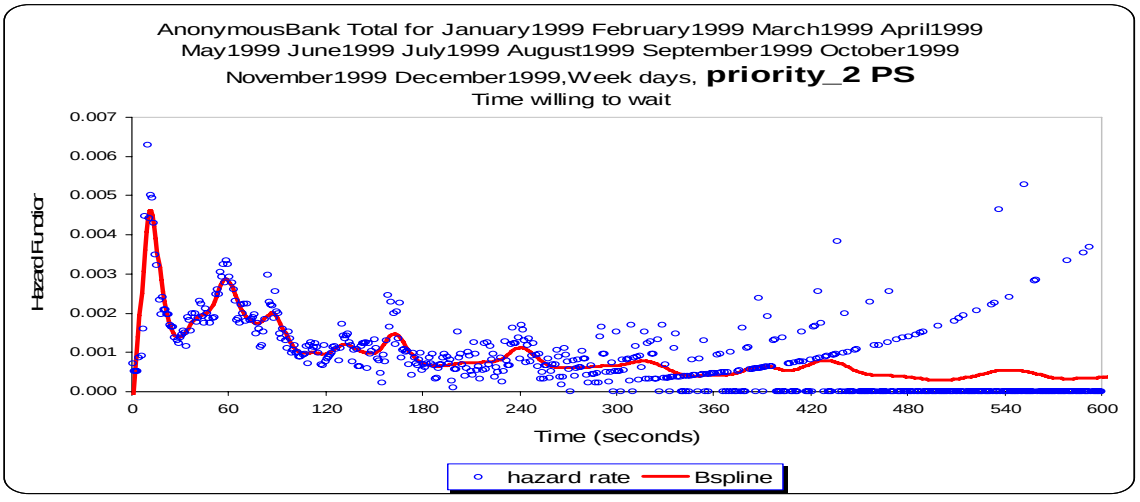


Summary of the Number of Censored and Uncensored Values				
Failure Time: Unhandled Wait Time; Censored Time: Handled Wait Time				
Class	*Number of Cases*	*Failed (Abandoned)*	*Censored (Served)*	*Percent Censored*
Total PS	164817	33006	131811	79.97
priority_1 PS	57007	15206	41801	73.33
priority_2 PS	104762	16042	88720	84.69

Hazard Rate Function (PS)

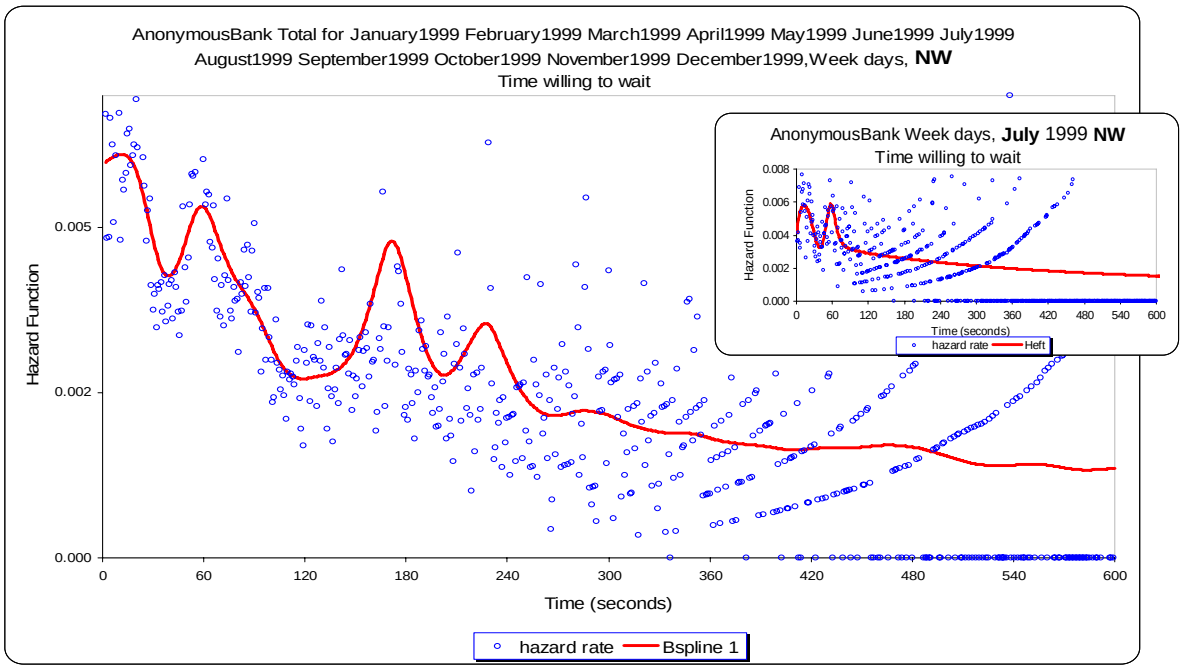


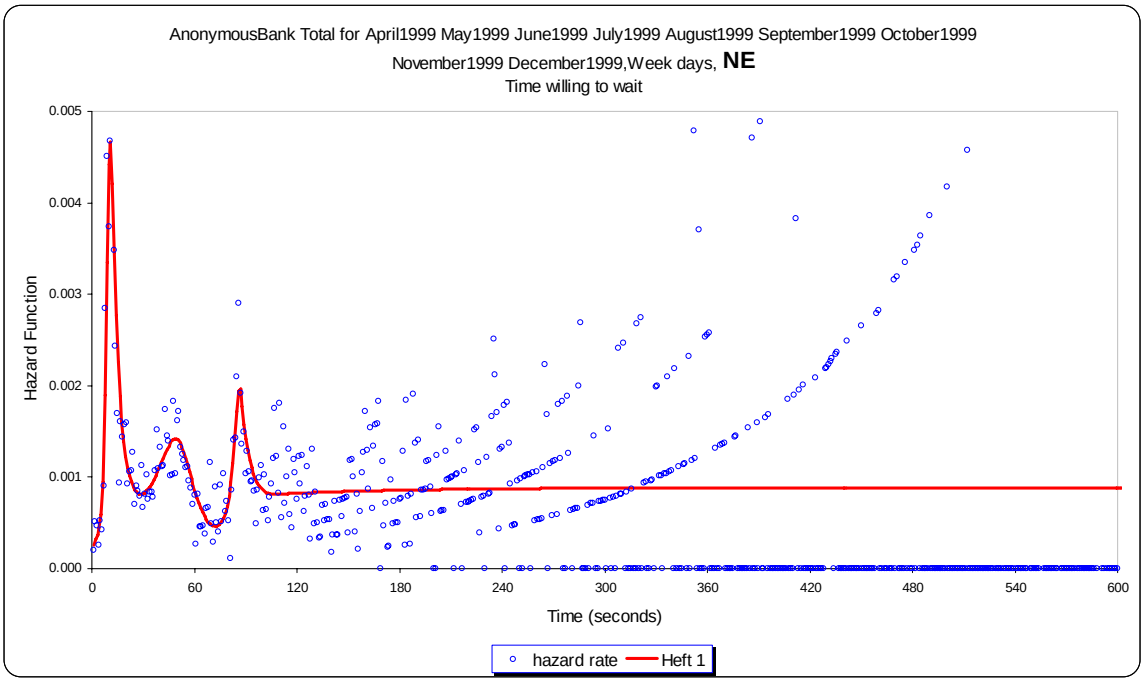




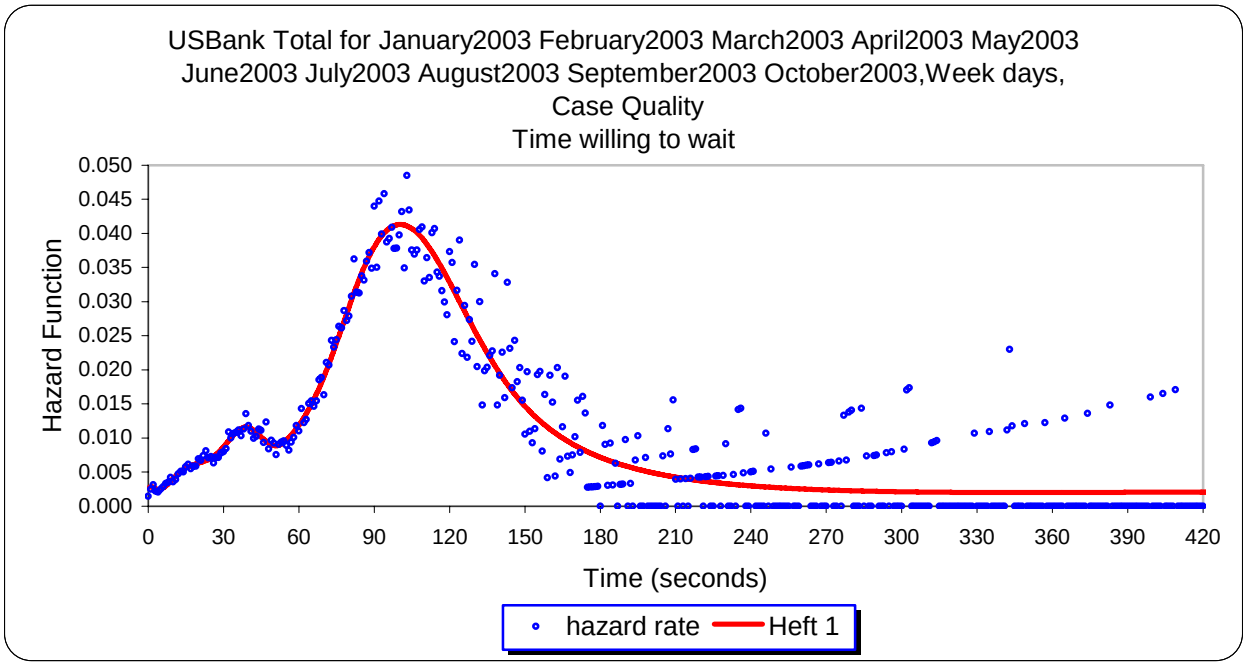
Hazard Rate Function (NW, NE)

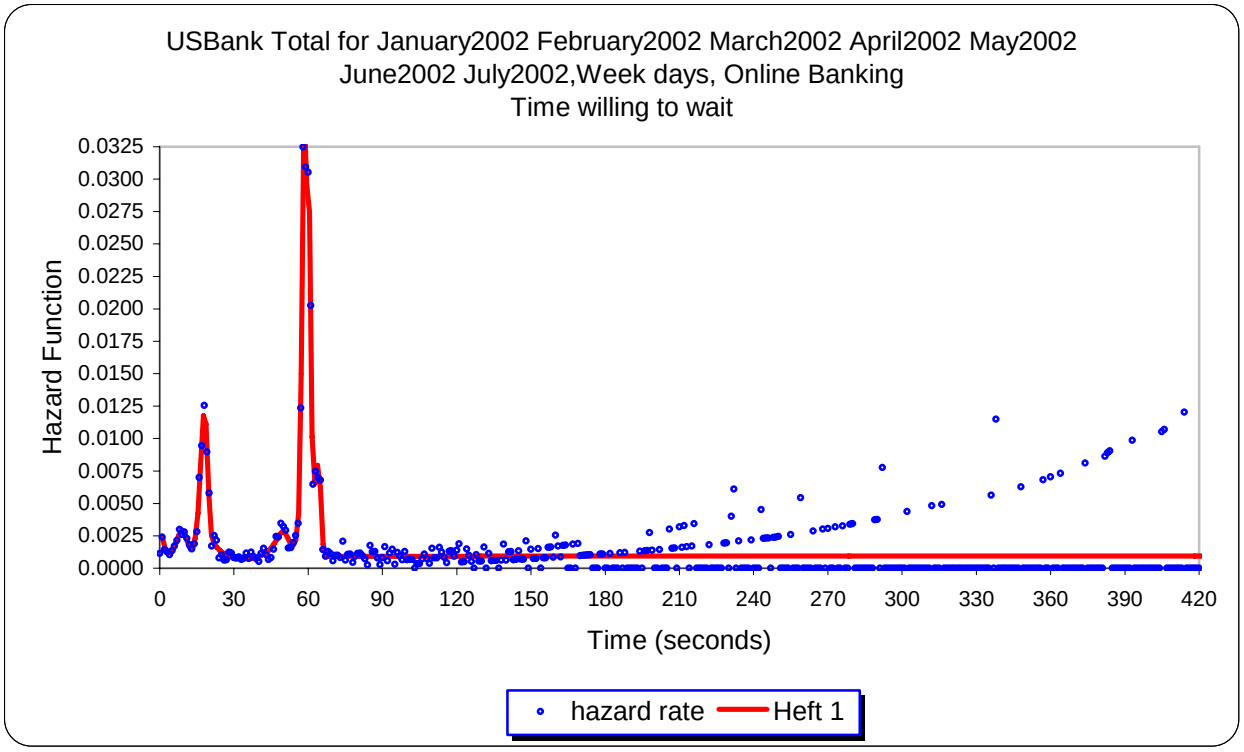
Summary of the Number of Censored and Uncensored Values				
Failure Time: Unhandled Wait Time; Censored Time: Handled Wait Time				
group	*Number of cases*	*Failed*	*Censored*	*Percent Censored*
NW	14709	6886	7823	53.19
NE	19483	2397	17086	87.70





Hazard Rate Function (Case Quality, Online Banking)

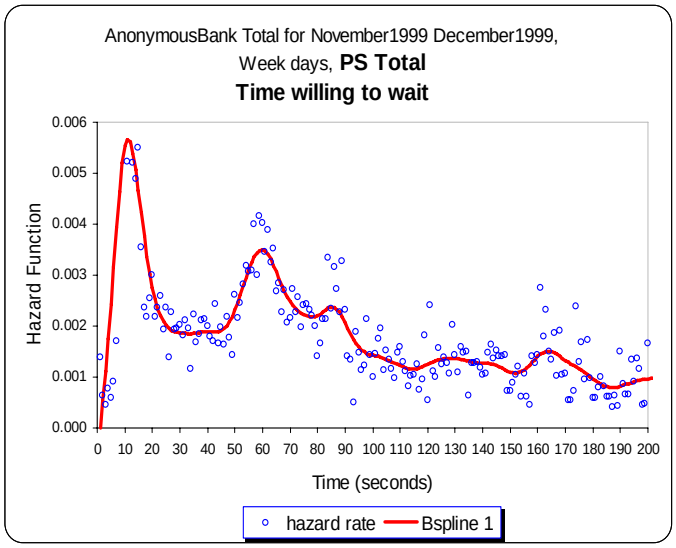


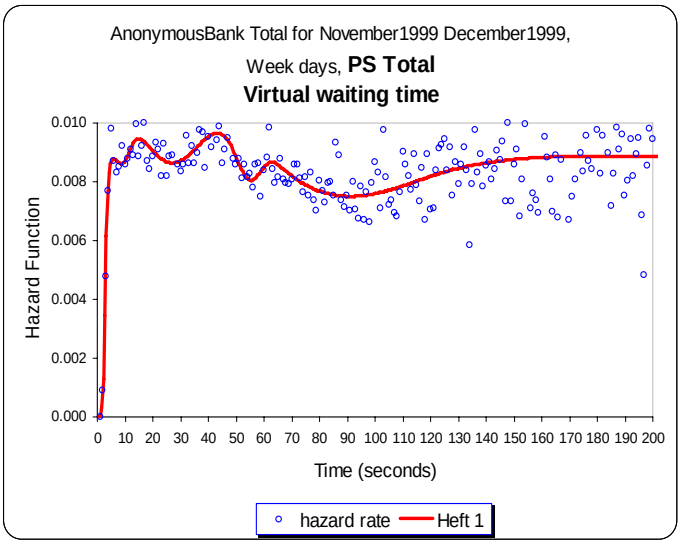


τ = (Im) Patience

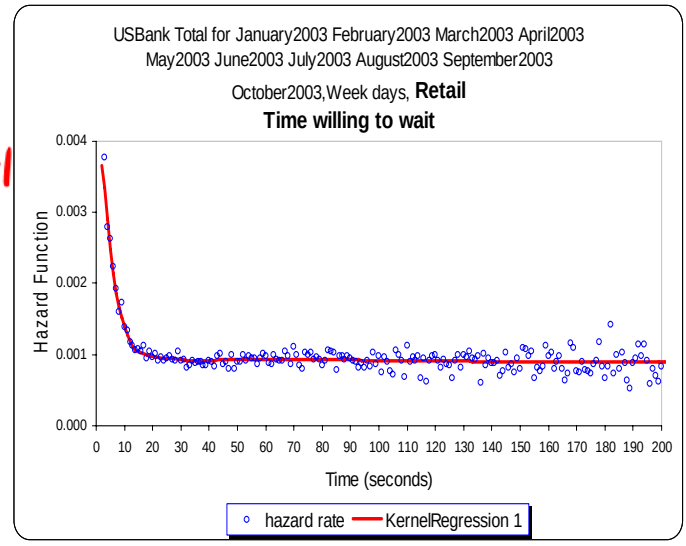
V = Offered (Required) Wait

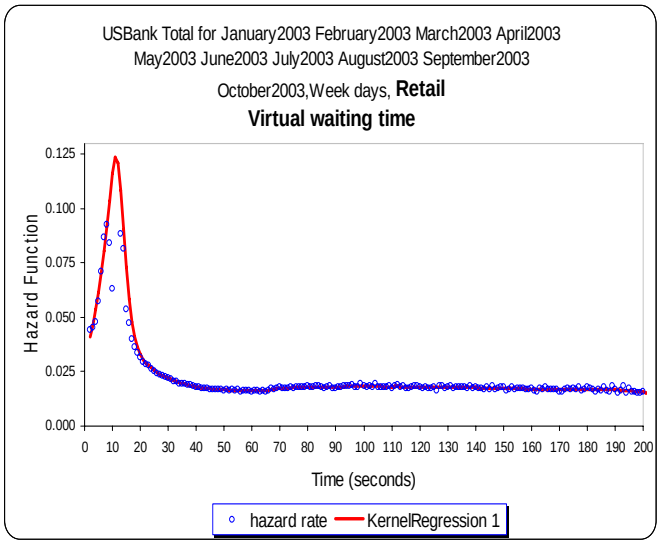
Israel



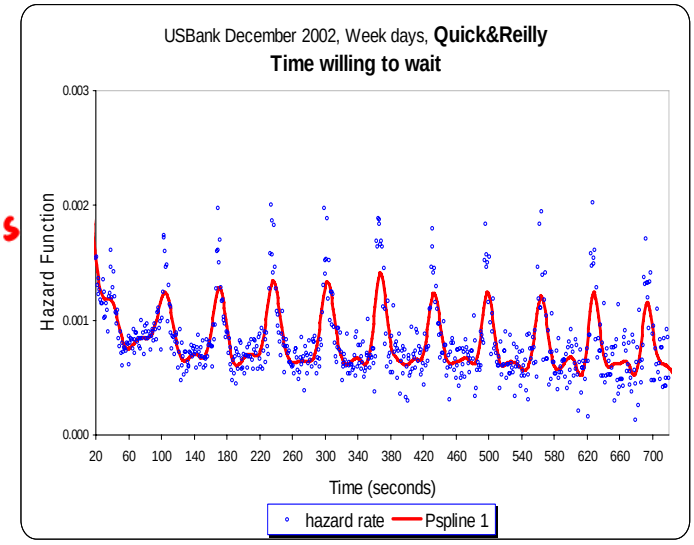


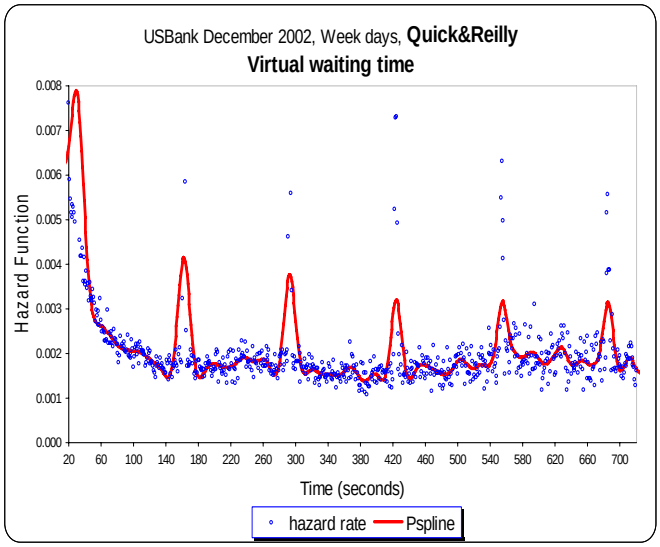
US Retail

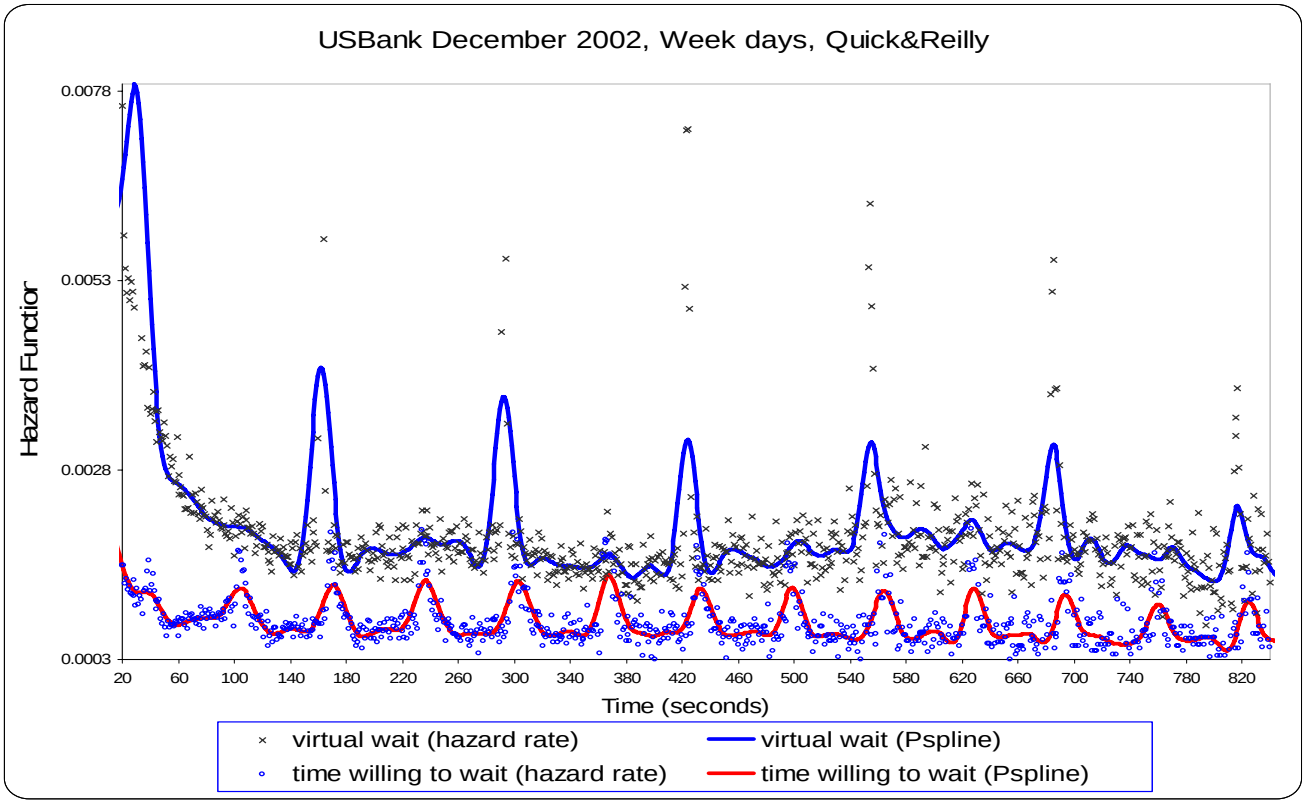




US Stocks






USBank December 2002, Week days, Quick&Reilly
Hazard Function
0.0078
0.0053
0.0028
0.0003
20
60
100
140
180
220
260
300
340
380
420
460
500
540
580
620
660
700
740
780
820
Time (seconds)
virtual wait (hazard rate)
time willing to wait (hazard rate)
virtual wait (Pspline)
time willing to wait (Pspline)

Estimating Average Patience: Warmup

Model: (Im)Patience τ equals

- 2 minutes, with probability p;
- 10 minutes, with probability $1-p$.

What is E$[\tau]$? (equivalently p ?)

Data: n_a abandoned after 2 minutes.
n_s got served (censored) after 3,4,...,9.

- Naive estimator: Average Patience = 2 minutes, which ignores those with the longer patience (who hence got served).

- Common-sense estimator: $\hat{p} = \frac{n_s}{n_a+n_s}$

$$\Rightarrow \quad E[\tau] = 2\hat{p} + 10(1-\hat{p}) = 2\frac{n_a}{n_a+n_s} + 10\frac{n_s}{n_a+n_s} = 2 + 8\frac{n_s}{n_a+n_s}.$$

Note:
$E[\tau] \to 10, \quad$ as $n_a/n_s \to 0$;
$E[\tau] \to 2, \quad$ as $n_a/n_s \to \infty$.

General Data: Data could conceivably consist of the times $\{0, 1, 2, \ldots, 9, 10\}$. Then, the 10's are easy to accommodate, and the $\{0, 1\}$'s are simply ignored (as it turns out - see the Kaplan-Meier estimator later, if interested) .

Estimating Average Patience: Practice

(Im)Patience τ is $\exp(\theta)$.
Assume customers' (im)patience times to be i.i.d.

Estimate $\mathrm{E}[\tau]$ (equivalently θ)?

Data: $W_1^a, W_2^a, \ldots, W_{n_a}^a$: n_a times to abandon;
$W_1^s, W_2^s, \ldots, W_{n_s}^s$: n_s times till served (censored).

Geometric Approximation (Intuition):
(Im)Patience Times: Geom(p) (seconds).
(Estimate $1/p$ and deduce an estimator for $1/\theta$.)

Every second flip a coin:
wp p Abandon (Success),
wp $(1-p)$ Wait one more second (Failure).

Coin Flips (in total):

$$= W_1^a + \ldots + W_{n_a}^a + W_1^s + \ldots + W_{n_s}^s \stackrel{\Delta}{=} W_{total}$$

= **Total Waiting Time** (Served + Abandoned).

Successes = # Abandon = n_a.

$$\Rightarrow \quad \hat{p} = \frac{n_a}{W_{total}} = \frac{\#\text{ Abandon}}{\text{Total Waiting Time}},$$

$$\Rightarrow \quad \text{Estimator of Average Patience} = \widehat{1/p} = \frac{\text{Total Waiting Time}}{\#\text{ Abandon}}.$$

Estimating Exponential Patience: Maximum Likelihood Estimator (MLE)

Patience Times: $\exp(\theta)$ i.i.d.

Likelihood:

$$L(\theta) \;=\; \left(\prod_{i=1}^{n_a} \theta \exp\left\{-\theta W_i^a\right\}\right) \cdot \left(\prod_{i=1}^{n_s} \exp\left\{-\theta W_i^s\right\}\right) .$$

Log-likelihood:

$$\begin{aligned} l(\theta) \;&=\; \log(L(\theta)) \\ &=\; n_a \log\theta - \theta \cdot (W_1^a + \ldots + W_{n_a}^a + W_1^s + \ldots + W_{n_s}^s) \\ &=\; n_a \log\theta - \theta \cdot W_{total} \,. \end{aligned}$$

MLE $\hat{\theta}$ attains the maximum in $l(\theta)$:

$$\begin{aligned} l'(\theta) \;&=\; n_a/\theta - W_{total} \;=\; 0\,, \\ \hat{\theta} \;&=\; n_a/W_{total}\,, \\ \widehat{1/\theta} \;&=\; W_{total}/n_a\,. \end{aligned}$$

Note: $\hat{\theta} \;=\; \dfrac{\mathrm{P}\{\widehat{\mathrm{Ab}}\}}{\mathrm{E}[\widehat{W}]}$.

Estimating Patience: Small Israeli Bank

Statistics	Average wait	Interpretation
360K served (80%)	2 min	? **Required** to Wait
90K abandoned (20%)	1 min	? **Willing** to Wait

Both waiting times are **censored**.
If customer abandoned, patience is known: $\boldsymbol{\tau = W}$.
If customer served, a lower bound is known: $\boldsymbol{\tau > W}$.

Total Wait = 90K × 1 min + 360K × 2 min.

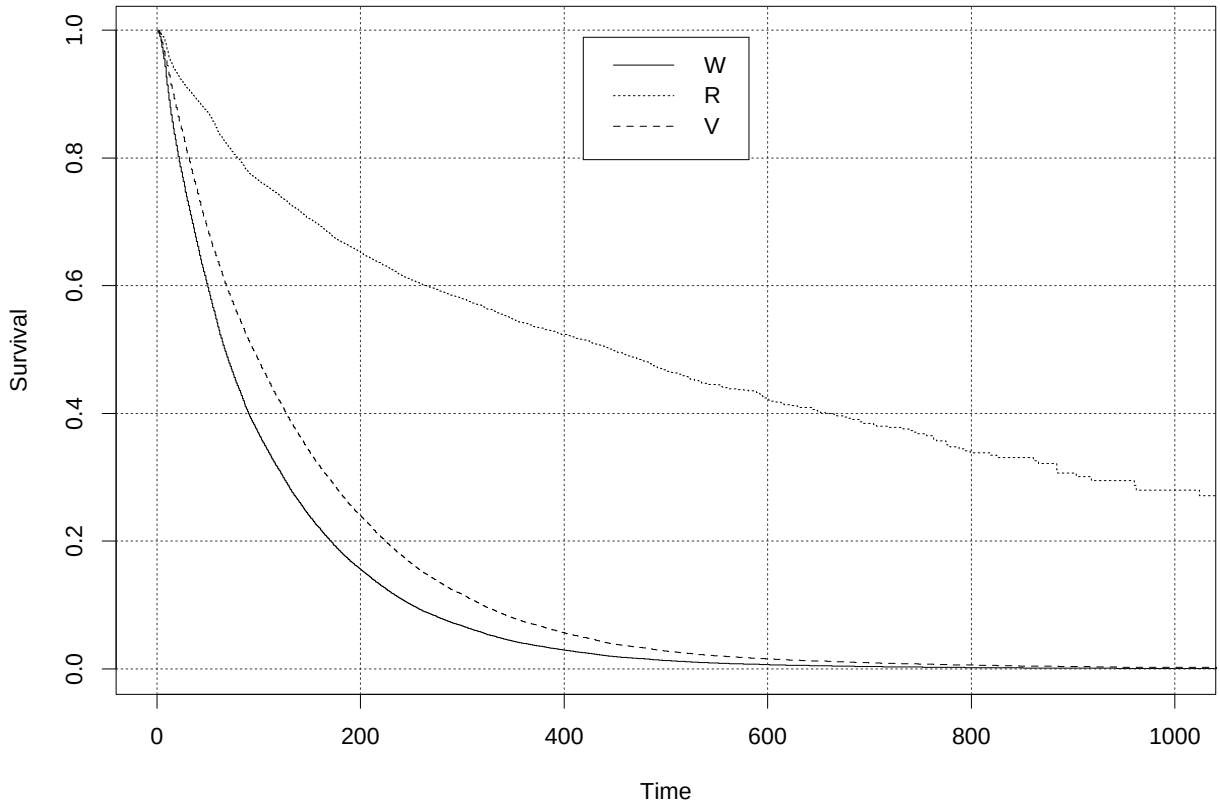
$$\textbf{Willing to Wait} = \frac{90K \times 1 + 360K \times 2}{90K} = 1 + 4 \times 2 = 9 \text{ min!}$$

$$\textbf{Required to Wait} = \frac{90K \times 1 + 360K \times 2}{360K} = 2.25 \text{ min.}$$

Note:
Willing-to-Wait / Required-to-Wait =
9 / 2.25 = 360K / 90K = 4 =
% Served / % Abandoned

Survival Functions: Patience vs. Offered Wait

Small Israeli Bank



E[W] = 98 sec, Med[W] = 62 sec;

E[τ] = 803 sec, Med[τ] = 457 sec; (R in Figure is τ)

E[V] = 142 sec, Med[V] = 96.

Are these customers "Patient"?

What if "E[V] = 1,600 sec" (twice E[τ]) ?

A Patience Index

How to **quantify (im)patience?**

$$\textbf{Theoretical Patience Index} \triangleq \frac{\text{Willing to Wait}}{\text{Expected to Wait}} = \frac{\mathrm{E}[\tau]}{\mathrm{E}[V]},$$

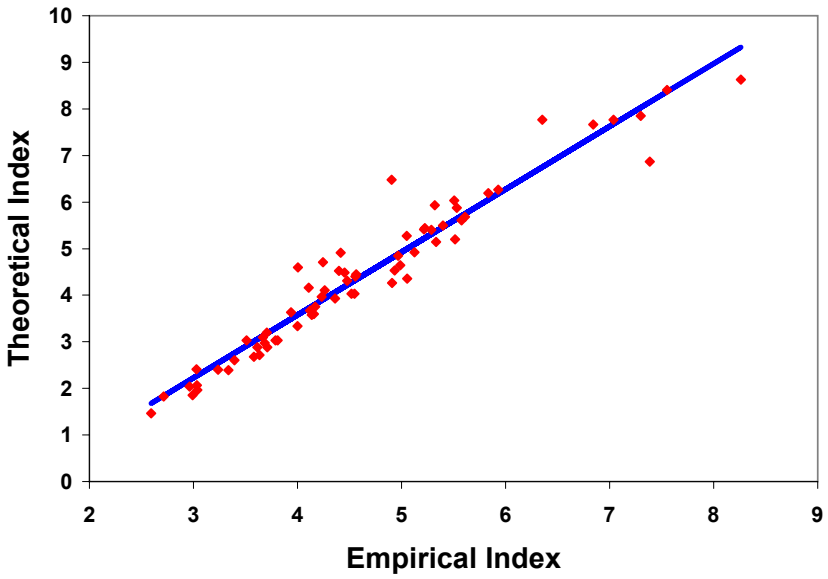
where the last equality (Expected-to-Wait = Required-to-Wait) is plausible for **Experienced Customers**.

We get a calculable quantity, but it still requires "un-censoring". To this end, "pretend" that both τ and V are exponential. Then, the MLE of the "Theoretical Patience Index" is:

$$\textbf{Empirical Patience Index} \triangleq \frac{\%\text{ served}}{\%\text{ abandoned}},$$

which is easily calculable from ACD data.

Patience index – Theoretical vs. Empirical



Patience Index: Willing to wait 10 min (patient ? / impatient ?)

$$\text{Theoretical index} = \frac{\text{Time willing to wait}}{\text{Time required to wait}}$$

$$= \frac{\text{Time willing to wait}}{\text{Time expect to wait}} \qquad \text{(if experienced)}$$

$$\begin{array}{lccl} \text{Index} & \text{large} & \Rightarrow & \text{patient population} \\ & \text{small} & \Rightarrow & \text{impatient} \end{array}$$

$$= \frac{E(R)}{E(V)}. \qquad \text{"Pretend"} \quad exp$$

$$= \frac{\text{Time in test } /\# \text{ abandon}}{\text{Time in test } /\# \text{ served}} \qquad censored.$$

$$\text{Empirical index} = \frac{\# \text{ served}}{\# \text{ abandon}} = \frac{\% \text{ served}}{\% \text{ abandon}}$$

$$= \frac{\% \text{ served / wait} > 0}{\% \text{ abandon / wait} > 0} \qquad \text{(easy to measure)}$$

Summary:

$$\text{Mean Patience} = \text{Mean Wait of Abandoning customers} + \text{Mean Wait of Served customers} \times \text{Patience Index}$$

Law: $\mathbf{P\{Ab\} \propto E[W_q]}$ (Often Enough)

Here we prove for **Exponential** (Im)Patience.

Can be justified theoretically, and validated empirically, **much more generally**.

Claim. Assume a queueing model with exp(θ) (im)patience. Then,

$$\mathbf{P\{Ab\} \;=\; \theta \cdot E[W_q]\,.}$$

Proof. Flow-conservation for **abandoning** customers, namely arrival-rate into queue = departure-rate out of queue,, implies:

$$\lambda \cdot \mathrm{P}\{\mathrm{Ab}\} \;=\; \theta \cdot \mathrm{E}[L_q]\,. \tag{1}$$

By Little's formula:

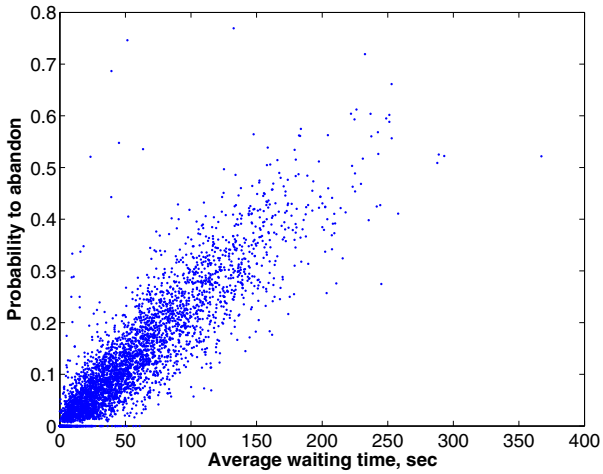
$$\mathrm{E}[L_q] \;=\; \lambda \cdot \mathrm{E}[W_q]\,. \tag{2}$$

Finally, substitute (2) into (1) and cancel λ. ■

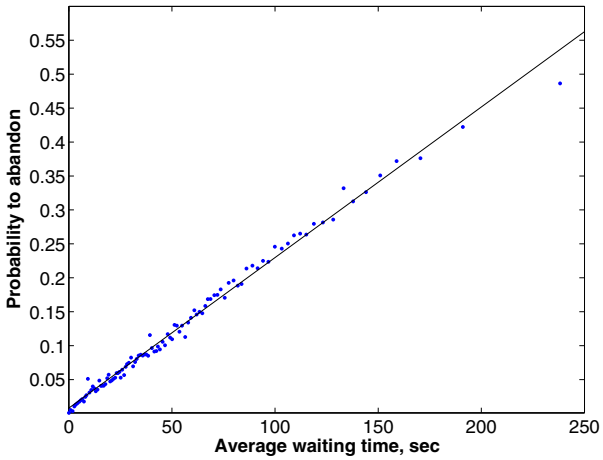
P{Ab} $\propto$ E[W_q]: Empirical Validation

Small Israeli Bank: Yearly Data (4158 hours)

Hourly Data (4158 points)



Aggregated

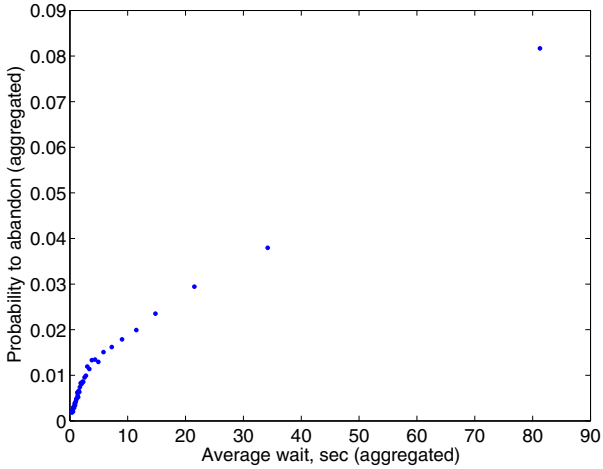


Estimating Average-(Im)Patience via Regression:

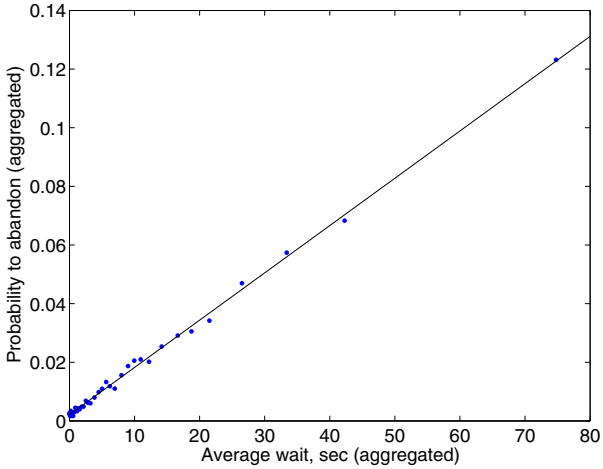
$1/\theta \approx \frac{250}{0.56} \approx 446$ sec.

Large U.S. Bank

Retail



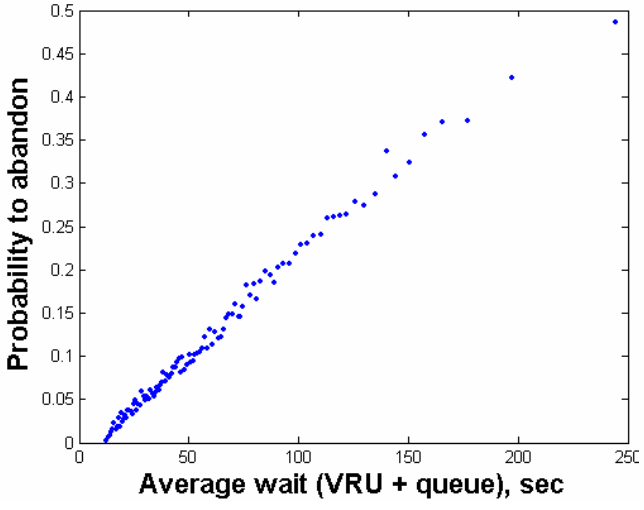
Telesales



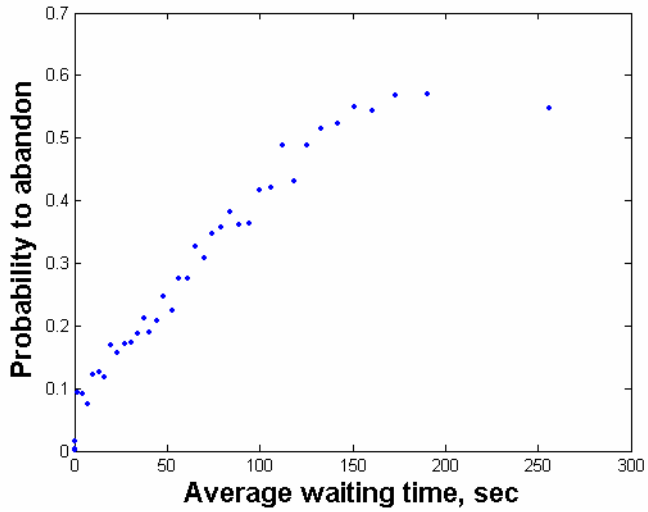
Note: in Retail – many abandon during first seconds of wait.

Queueing Science: Human Behavior

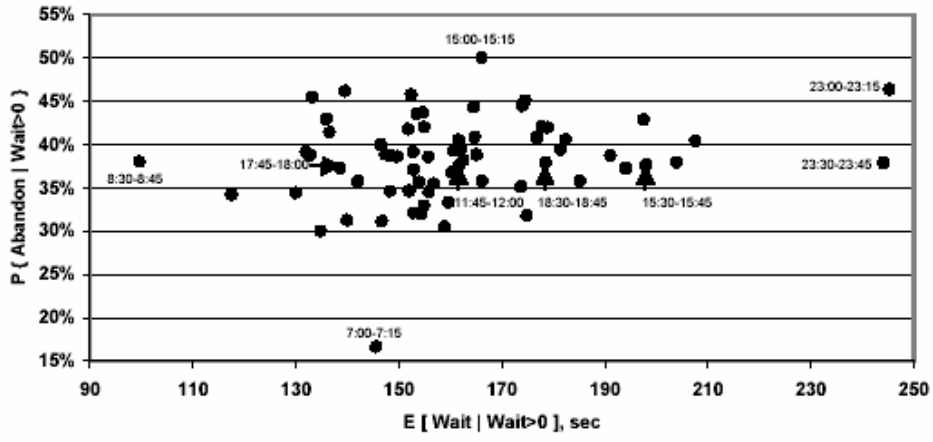
Delayed Abandons (IVR)



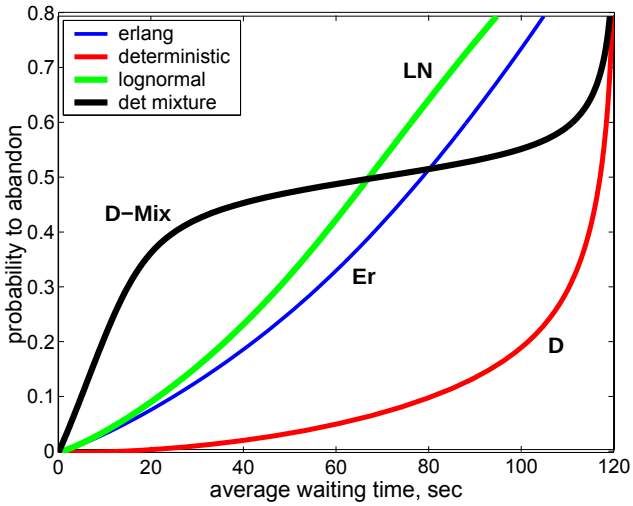
Balking (New Customers)



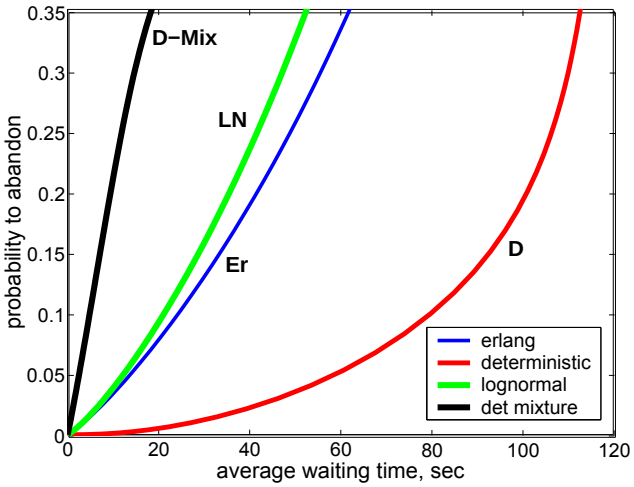
Learning (Internet Customers)



Examples of non-linear relations



moderate loads



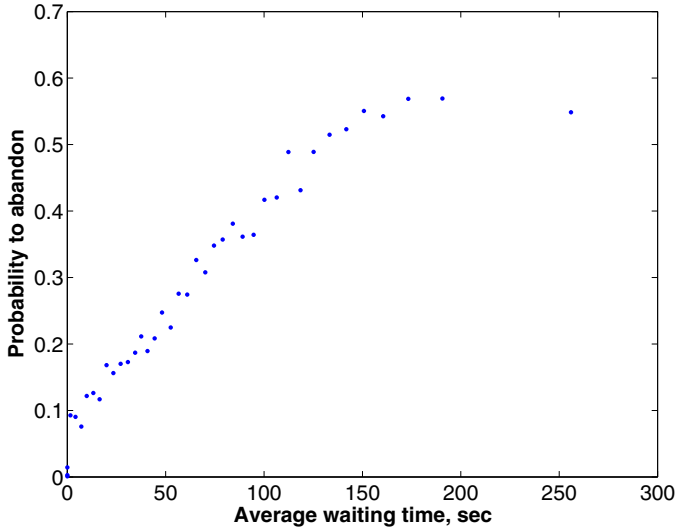
Patience distributions:

- **D**: Deterministic: 2 minutes exactly;
- **Er**: Erlang with two exp(mean=1) phases;
- **LN**: Lognormal, both average and standard deviation equal to 2;
- **D-Mix**: 50-50% mixture of two constants: 0.2 and 3.8.

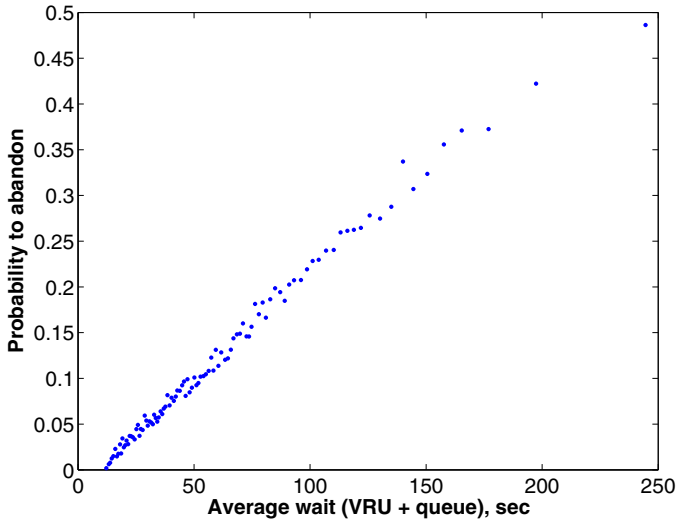
Human Behavior: Mathematical Models

Linear patterns with non-zero intercepts

Israeli Bank: New Customers



U.S. Bank: VRU part of Wait



Left-hand plot $\approx$ exp patience with **Balking**:
0 with probability p, $\exp(\theta)$ with probability $(1-p)$.

Right-hand plot $\approx$ **Delayed Abandonment**:
$c + \exp(\theta)$, $c > 0$.

Formalizing **Learning**:
Experienced customers use **actual** offered-load in order to optimize individual profits, which characterizes (unique) Nash-Equilibrium.

Estimating General Patience: The Kaplan-Meier Estimator

Assume patience and waiting times discreet (seconds).

Hazard rate:
$h(k) = \mathrm{P}\{\tau = k\}/\mathrm{P}\{\tau \geq k\}\,, \qquad k = 0, 1, 2 \ldots$

Survival Function:
$S(k) = S(k-1) \cdot (1 - h(k))\,, \qquad k = 0, 1 \ldots \;\; (S(-1) = 1)$

A_k = number of abandonment exactly at k seconds,

η_k = number of customers that are neither served nor abandoned before k seconds (number-at-risk at time k).

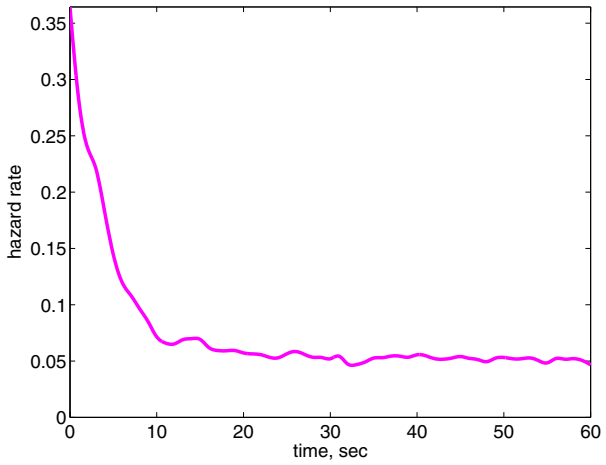
Estimator of Hazard Rate: $\widehat{h(k)} = A_k/\eta_k$.

Estimator of Survival Function (Kaplan-Meier):
$\widehat{S(k)} = \prod_{i=0}^{k} (1 - \widehat{h(i)})$.

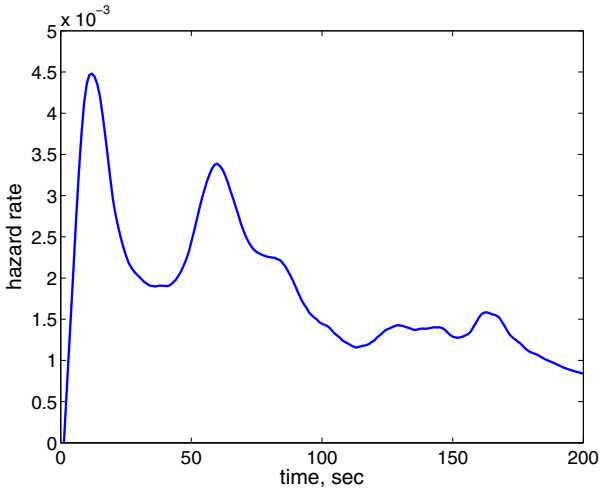
Estimating (Im)Patience Distribution: Real Data

Empirical Hazard Rates of (Im)Patience Times

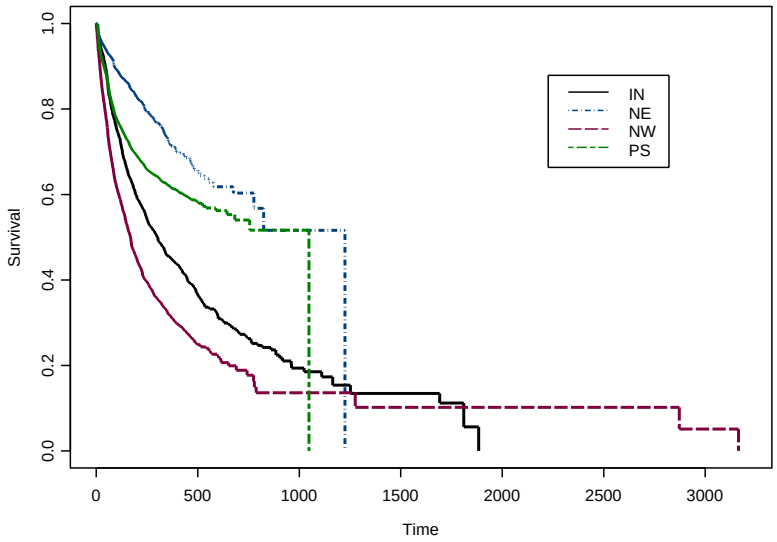
U.S. Bank



Israeli Bank



Israeli Bank: Survival Functions of Service Types



IN – Internet Tech. Support; NE – Stock Transactions;
NW – New Customers; PS – Regular.

The "Phases of Waiting" for Service

TIME IS
Time is Too Slow for those who Wait,
Too Swift for those who Fear,
Too Long for those who Grieve,
Too Short for those who Rejoice;
But for those who Love, Time is not.
(Henry Van Dyke 1852 - 1933)

Common Experience:

- Expected to wait 5 minutes, Required to 10
- Felt like 20, Actually waited 10 (hence Willing $\geq$ 10)

An attempt at "Modeling the Experience":

1. Time that a customer	**expects** to wait	
2.	**willing** to wait	((Im)Patience: $\boldsymbol{\tau}$)
3.	**required** to wait	(Offered Wait:$\boldsymbol{V}$)
4.	**actually** waits	$(\boldsymbol{W_q = \min(\tau, V)})$
5.	**perceives** waiting.	

Experienced customers $\Rightarrow$ Expected = Required
"Rational" customers $\Rightarrow$ Perceived = Actual.

Thus **left with** $\boldsymbol{(\tau, V)}$.

Perceived vs. Actual Waiting: an Example

200 Abandonment in Direct Banking (Students' Project)

Reason to Abandon	**Actual** Abandon Time (sec)	**Perceived** Abandon Time (sec)	Perception Ratio
Fed up waiting (77%)	70	164	2.34
Not urgent (10%)	81	128	1.6
Forced to (4%)	31	35	1.1
Something came up (6%)	56	53	0.95
Expected call-back (3%)	13	25	1.9

Customers' (Im)Patience in Call Centers: Summary

- (Im)Patience time are, in general, **non-exponential**;
- Most tele-customers are **very** (surprisingly) patient;
- Hazard and survival estimators are very informative concerning *qualitative* patterns of (im)patience (abandonment peaks, comparisons, ...);
- Kaplan-Meier can be problematic for estimation of *quantitative* characteristics (eg. mean, variance, median). $\widehat{\mathrm{E}[\tau]} = \int_0^\infty \widehat{S(x)}dx$, where $S(x)$ - survival function of patience. However, $\widehat{S(x)}$ is not reliable for large x.

Practical Question: Can we **apply** models with **exponential** (im)patience as a useful approximation?

Practical Answer: A definite **"YES"**, even in the sense of "Must Apply". In other words, a model that wrongly assumes exponential (im)patience is far better than a model that ignores (im)patience (which, surprisingly, is prevalent in practice).

Estimate Mean Patience that is $\exp(\theta)$.

1. Via $P_{Ab} = \theta E W_q$

$$\widehat{1/\theta} = \frac{EW_q}{P_{Ab}} = \frac{\text{“ total waiting time ” / N}}{\#\text{abandon}/N}$$

$$= \frac{\text{“ total time in test ”}}{\#\text{ uncensored (observed)}}.$$

   Use the above to estimate mean patience , $E(R)$.

2. Note: We get this way the MLE (*maximum likelihood estimation*) of a censored exponential mean.

3. Via Regression of P_{Ab}'s over EW_q 's.

4. Via "Geometric Intuition".

   Suppose measurements are as follows :

   - m abandoned, with time-to-abandon $W_1^a, W_2^a, ..., W_m^a$
   - n served, with time-to-service $W_1^s, W_2^s, ..., W_n^s$ seconds

   Approximate exponential patience with Geometric Patience :
   Every second flip a coin, with

$$\begin{aligned} &\text{probability} \quad p \quad \text{for success} = \textit{abandon}, \\ &\text{probability } 1-p \ \text{ for failure } = \textit{stay one more second}. \end{aligned}$$

Q. What is $1/p$ = mean patience.

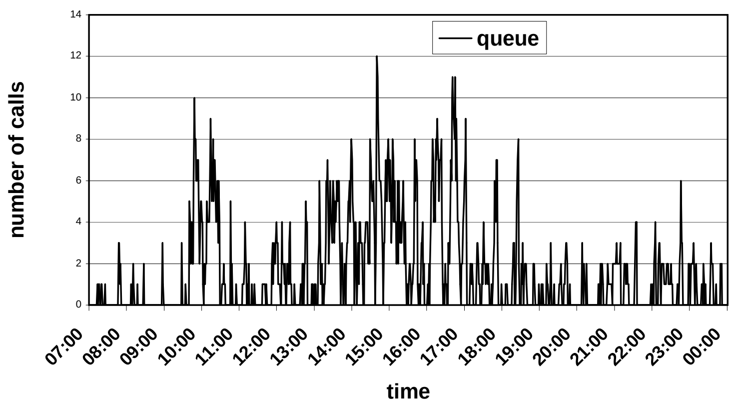
A. Total # of coin flips

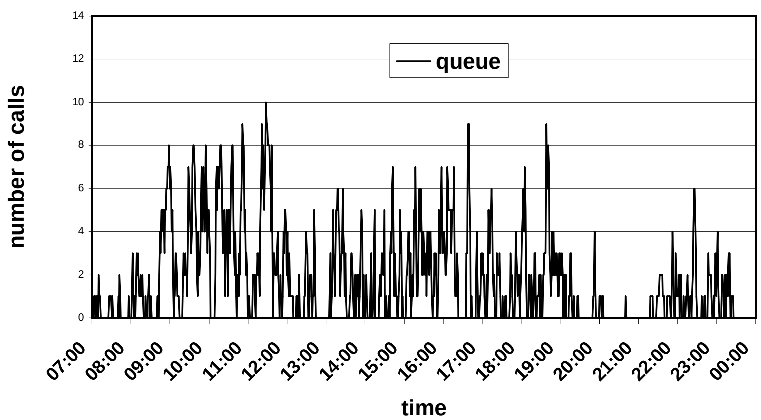
$$\begin{aligned} &= W_1^a + W_2^a + ... + W_m^a + W_1^s + W_2^s + ... + W_n^s \\ &= \text{Total Waiting Time (served + abandoned)}. \end{aligned}$$

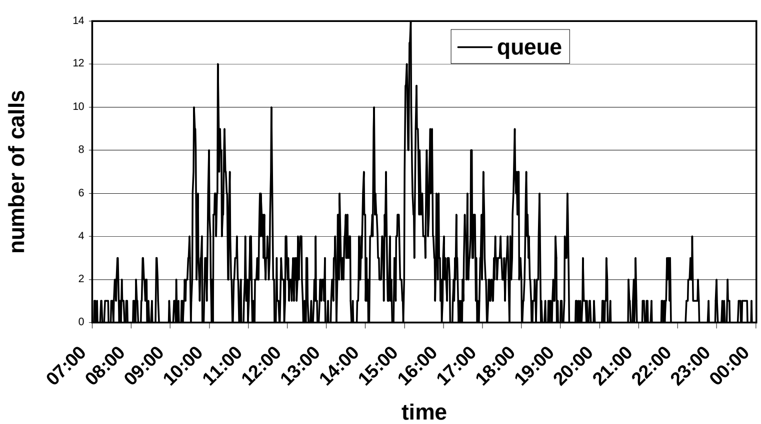
successes = # abandonment = m.

$$\Rightarrow \quad \hat{p} = \frac{\#\text{ abandon}}{\text{Total Waiting Time}} \quad \Rightarrow \quad \hat{1/p} = \frac{\text{Total Waiting Time}}{\#\text{ abandon}}.$$

Queues = Integrating the Building Blocks

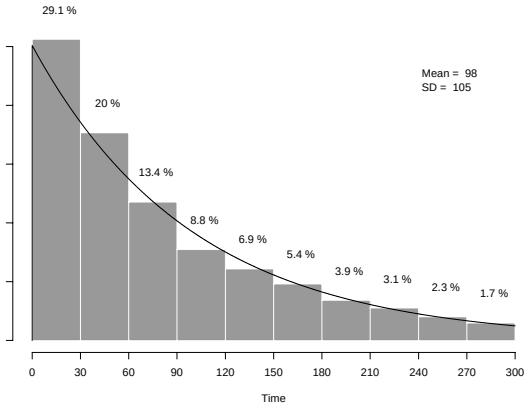


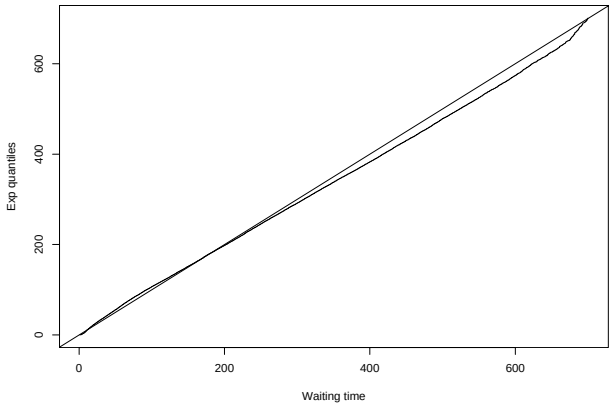




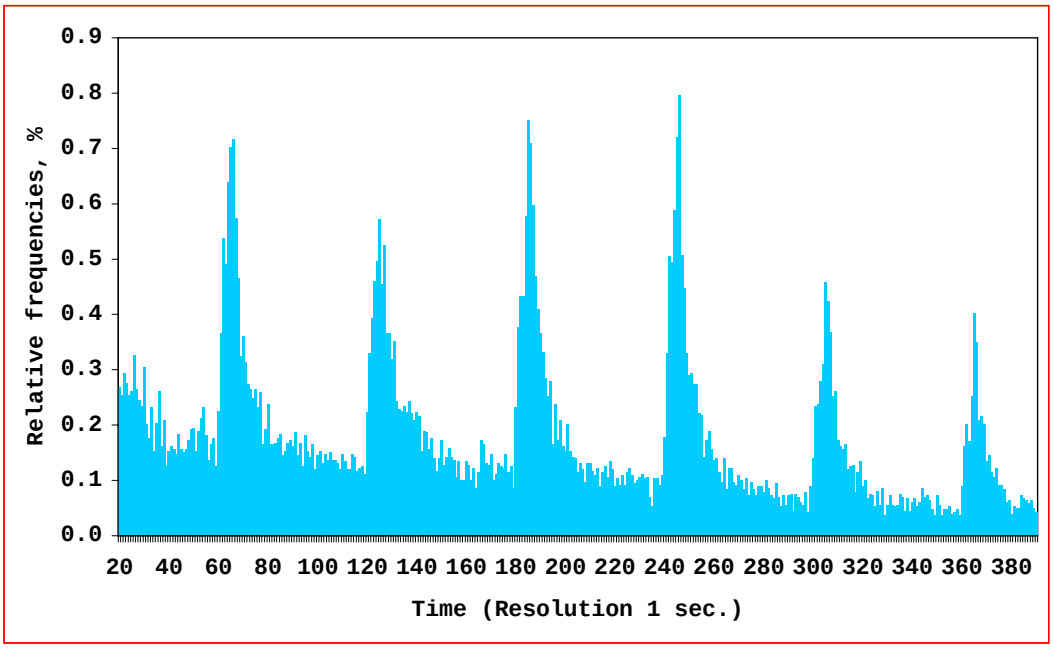
Delays = Integrating the Building Blocks

Exponential Delays:
Small Call Center of an Israeli Bank (1999)





Delays:
Medium-Size Call Center of an Israeli Bank (2006)



Hazard Rate Functions

Examples via Phase-Type Distributions

Definition. If T is an *absolutely continuous* non-negative random variable, its hazard rate function $h(t)$, $t \geq 0$, is defined by

$$h(t) = \frac{f(t)}{S(t)}, \quad t \geq 0,$$

where $f(t)$ is the density of T and $S(t)$ is the survival function:
$S(t) = \int_t^\infty f(u)du = P\{T > t\}$.
Note that $\mathrm{P}\{T \leq t + \Delta \,|\, T > t\} \approx h(t) \cdot \Delta$.

If T is a *discrete* non-negative random variable that takes values $t_1 < t_2 < \ldots$ with corresponding probabilities $\{p_i,\ i \geq 1\}$, then its hazard-sequence $\{h(t_i)\}$ is defined by

$$h(t_i) = \frac{p_i}{\sum_{j \geq i} p_j} = \frac{p_i}{S(t_i-)}, \quad i \geq 1.$$

Note that $\mathrm{P}\{T = t_i \,|\, T > t_{i-1}\} = h(t_i)$.

Why estimate the hazard rates of service times or patience?

- The hazard rate is a *dynamic* characteristic of a distribution.
  (One of the main goals of our note is to demonstrate this statement).
- The hazard rate is a more precise "fingerprint" of a distribution than the cumulative distribution function, the survival function, or density (for example, unlike the density, its tail need not converge to zero; the tail can increase, decrease, converge to some constant etc.)
- The hazard rate provides a tool for comparing the tail of the distribution in question against some "benchmark": the exponential distribution, in our case.
- The hazard rate arises naturally when we discuss "strategies of abandonment", either rational (as in Mandelbaum & Shimkin) or ad-hoc (Palm).

Why do phase-type distributions constitute a convenient class of models for service times ? As discussed in class:

- dense;
- structurally informative;
- meta theorem: homogeneous unpaced human service\task durations are exponential.

Why is it convenient to illustrate the concept of hazard rate via phase-type examples?

- Small number of phases suffices to illustrate the various modes of hazard-rate behavior.
- Simple intuitive explanations of hazard-rate patterns can be demonstrated. (In contrast, try to develop intuition for the hazard rates of normal or lognormal random variables!)

Limitations: *Which patterns of hazard rate cannot be illustrated by phase-type distributions?* **Answer.** We shall see below that the hazard rate of a phase-type distribution has a limit as $t \to \infty$. This limit can be shown to be neither 0 nor ∞. Hence, phase-type distributions can not belong to heavy-tail distributions with hazard rates that converge to zero (recall Pareto) or to distributions with hazard rates that converge to infinity (recall the Normal distribution).

Hazard-rate representation for Phase-Type distributions

Let T be phase-type distributed. Animate T by an absorbing Markov jump-process $X = \{X_t,\ t \geq 0\}$, on a finite state-space S, with an absorbing state Δ. Then the hazard-rate function of T, $h_T(t)$, has the representation:

$$\boxed{h_T(t) \ = \ \sum_{i \in S} q_{i\Delta} \mathrm{P}\{X_t = i | T > t\}, \quad t \geq 0}$$

where $q_{i\Delta}$ is the transition (absorption) rate from state i, that is

$$\mathrm{P}\{X_{t+\epsilon} = \Delta | X_t = i\} \ = \ q_{i\Delta} \cdot \epsilon + o(\epsilon), \ i \in S.$$

The representation above demonstrates the *dynamic approach* to the hazard rate of phase-type distributions: the hazard rate at time t is determined by the conditional distribution of the underlying Markov process X.

For convenience, denote

$$P_i(t) \ = \ \mathrm{P}\{X_t = i | T > t\}, \quad t \geq 0, \quad i \in S.$$

Remark. As $t \uparrow \infty$, the functions $\{P_i(t),\ i \in S\}$ converge to, what is called, the *quasi-stationary* distribution of X. It can be expressed in terms of eigen-values related to the matrix Q (generator of X, restricted to S), and gives rise to a representation for the limit

$$h_T(\infty) = \sum_{i \in S} q_{i\Delta} P_i(\infty) \ .$$

In the examples that follow, $P_i(\infty)$ will be calculated directly.

General description of our (static) simulation.

We consider four examples of phase-type distributions. For each example, 10,000 independent realizations were simulated in Excel. The theoretical hazard rates were plotted and compared against estimates of the hazard rate, based on the simulation data. (The method used for hazard rate estimation is described in the Technical Appendix, at the end of the handout.)

In the examples below, the probabilities $P_i(t)$ for all non-absorbing states $i \in S$ were calculated explicitly. We then tried to illuminate the connection between $P_i(t)$ and the hazard rate, based on the representation above.

(Continued at the END)

Adaptive Behavior of Impatient Customers in Tele-Queues: Theory and Empirical Support [1] [2]

Ety Zohar[3], Avishai Mandelbaum[4] [5] *and* Nahum Shimkin[6]

November 12, 2000

Abstract

We address the modeling and analysis of abandonment from a queue which is invisible to its occupants. Such queues arise in remote service systems, notably the Internet and telephone call centers, hence we refer to them as *tele-queues*. A basic premise of this paper is that customers adapt their patience (modeled by an abandonment-time distribution) to their service expectations, in particular to their anticipated waiting time. We first present empirical support for that hypothesis, and propose an M/M/m-based model which incorporates adaptive customer behavior. In our model, customer patience (and possibly the arrival rate) depend on the *mean* waiting time in the queue. We then characterize the system equilibrium and establish its existence and uniqueness when the growth rate of customer patience is bounded by that of the mean waiting time . The feasibility of multiple system equilibria is illustrated when this condition is violated. We also discuss a decision-theoretic model for customer abandonment, and relate it to our basic model. Finally, a dynamic learning model is proposed where customer expectations regarding their waiting time are formed through accumulated experience. We address certain issues related to censored-sampling that arise in this framework and demonstrate, via simulation, convergence to the theoretically anticipated equilibrium.

Key words: Exponential (Markovian) Queues, Abandonments, Equilibrium Analysis, Invisible Queues, Performance-Dependent Behavior, Tele-services, Tele-queues, Call Centers

[1] Research partially supported by the Israeli Science Foundation, Grant 388/99-2, 1991-2002.

[2] We thank Sergey Zeltyn for his essential proactive contribution to the data analysis. Sergey also read parts of the manuscript and provided helpful feedback that stimulated improvements.

[3] Department of Electrical Engineering, Technion, Haifa 32000, Israel. e-mail: ety@tx.technion.ac.il

[4] Department of Industrial Engineering, Technion, Haifa 32000, Israel. e-mail: avim@tx.technion.ac.il

[5] Research partially supported by Technion V.P.R. fund for the promotion of sponsored research, and by the fund for promotion of research at the Technion.

[6] Department of Electrical Engineering, Technion, Haifa 32000, Israel. e-mail: shimkin@ee.technion.ac.il

[14] for a recent literature review). In particular, patience is unaltered by possible changes in congestion. Such models, however, can *not* accommodate the following scatterplot, that exhibits remarkable patience-adaptivity.

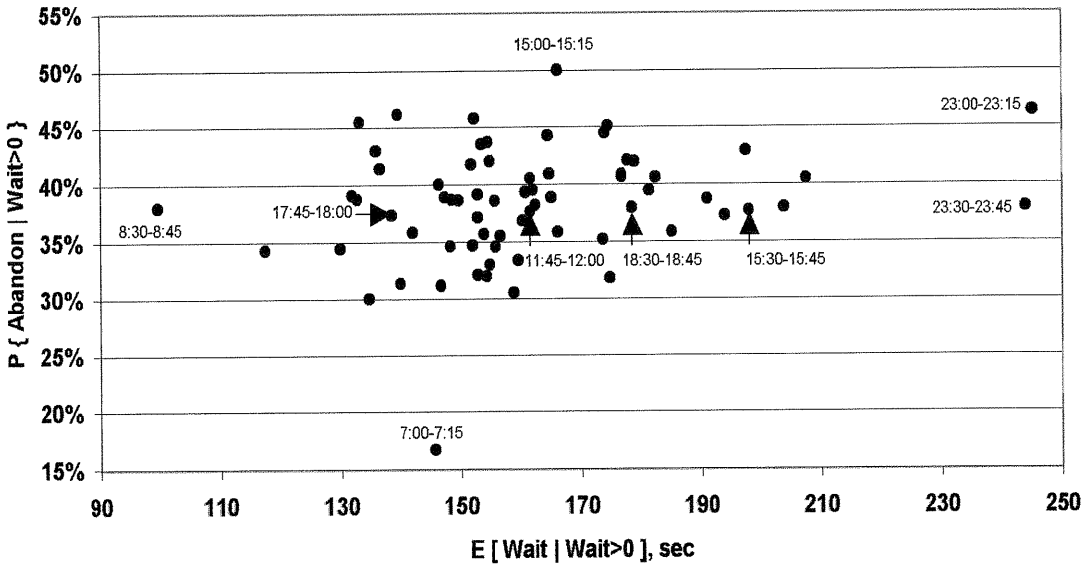


Figure 1: *Adaptive (IN) customers — abandonment probability vs. average offered wait (of customers with positive waits). Each point corresponds to a 15-minute period of a day (Sunday to Thursday), starting at 7:00am, ending at midnight, and averaged over the whole year of 1999.*

The data is from a bank call center [25] (see Section 3 for elaboration and further empirical analysis). We are scatterplotting abandonment fraction against average delay, for delayed customers (positive queueing time) who seek technical Internet-support. It is seen that average delay during 8:30-8:45am, 17:45-18:00, 18:30-18:45 and 23:30-23:45pm is about 100, 140, 180 and 240 seconds respectively. Nonetheless, the fraction of abandoning customers (among those delayed) is remarkably stable at 38%, for *all* periods. This stands in striking contrast to traditional queueing models, where patience is assumed unrelated to system performance: such models would predict a strict increase of the abandonment fraction with the waiting time, as in Figure 2. The behavior indicated in Figure 1 clearly suggests that customers do adapt their patience to system performance.

Appendix – Censored Sampling , via Kaplan-Meier (KM)

The need for accommodating censored data arose first in Section 3. Based on the call center data in [25], we sought to estimate *patience* – the distribution of the time a customer is willing to wait, and relate it to *offered wait* - the time a customer if forced to wait. As explained in Section 3, these two quantities actually censor each other. Then, in Section 6, censored data arose again. Simulated customers sought to estimate the system's offered wait, based on their individual service history where some samples of the offered wait were censored by abandonment. In both Sections 3 and 6, one is required actually to estimate only means, as opposed to the full fledged distribution. (The latter is needed, for example, to support our first observation in Section 3, regarding the non-exponentiality of patience. See [25], Section 6, especially Figures 12 and 14, for interesting hazard-rate estimators of patience and offered wait.)

Techniques for analyzing censored data have been developed within the well-established Statistical branch of Survival Analysis ([27] is an elementary exposition, and [12] is advanced measure-theoretic). As will be explained in the sequel, our needs for such techniques vary from the rudimentary to the unexplored.

In Section 3 we estimated mean patience and mean offered-wait via the means of the corresponding classical Kaplan-Meier (KM) estimator (A.19). KM generalizes the empirical distribution function to accommodate censored samples (see page 46 in [27], or page 4 in [12]). It is a *non-parametric* estimator, proven to have desirable properties, and common enough to be incorporated in essentially all respectable statistical packages. In Section 6 we used again KM, and then continued with a simpler *parametric* estimator, namely the maximum-likelihood estimator (MLE) of the mean of an exponential distribution; it is defined in (A.20) and referred to in our paper as the censored MLE (CMLE). The rest of the Appendix is devoted to a description of KM and CMLE, tailored to the estimation of patience and offered wait.

The KM setup for estimating patience is as follows. We are given a sample $\{W_i\}$ of N waiting times from a call center. Some of the calls end up with abandonment ($W_i = T_i$) and the others with a service ($W_i = V_i$). Denote by $M \leq N$ the number of *distinct* abandonment times in the sample. Let $T^1 < T^2 < ... < T^M$ be the ordered observed abandonment times, and A_k the number of abandonment at T^k, namely those who abandon after exactly T^k units of time. The *Kaplan-Meier estimator* $\hat{S}(t)$, $t \geq 0$, estimates the survival function

$\bar{F}(t) = P(T > t)$, where T is the time to abandon (patience). It is given by

$$\hat{S}(t) = \prod_{k:R_k \leq t} (1 - \frac{A_k}{\bar{B}_k}),$$

where $\bar{B}_k$ denotes the number of customers still present at T^k, that is neither served nor abandoned before T^k. The estimator for mean patience is then based on the tail-formula

$$\widehat{E[T]} = \int_0^\infty \hat{S}(t)dt. \tag{A.19}$$

In the above we estimated patience, which was censored by offered wait. Similarly, KM can be used to estimate the offered wait, by switching the roles of V_i and T_i. This estimate was used both in Section 3 and 6, in the latter by individual customers in order to estimate the system's offered wait that affects their patience.

A simpler alternative for estimating offered wait takes a parametric approach. As above, let $\{W_1, W_2, ..., W_N\}$ denote the collection of all waiting times, both abandoning and served. Assuming that offered wait is exponentially distributed, the standard parametric (maximum likelihood) estimator for its mean is given by ([27], page 22)

$$\widehat{E(T)} = \frac{1}{N_s} \sum_{i=1}^{N} W_i\,, \tag{A.20}$$

where N_s is the number of service experiences that ended up with a service, i.e. were not censored by abandonment. If T is not exponential, the estimator (A.20) is biased enough to be inconsistent.

Remark. *On Independence:* KM assumes independence for the observations whose distribution is to be estimated. Such an independence is plausible for patience (T_k's). It also applies for offered wait (V_i's), if these are sampled during independent sparsely-timed visits to the queue, as in Section 6. Such independence can *not* hold for successive offered loads, that are in fact highly dependent. In this case one is taken out of the KM paradigm. The effect of such dependence has been ignored in Section 3, as well as in [25], and it is the subject of ongoing research.

Remark. *On Robustness:* The KM (Kaplan-Meier) estimator is very sensitive to censored data at the upper tail of the sample. For example, if the longest wait in a customer's history ended up with an abandonment, the KM estimator of the offered wait has a positive mass at infinity, hence its mean is infinity; similarly if one is interested in patience, and the longest

wait ended up with a service. The consequence is that in estimating patience and offered wait, one of the resulting two KM's must be defective, and common practice is to simply truncate it at its last observation. (There are some parametric tail-smoothing techniques, but to the best of our knowledge they are ad-hoc.)

Another alternative is to use medians, rather than means, as more robust estimators of a location-parameter. For example, the analogue of Figure 4 for NW customers, but with medians rather than means, is the following:

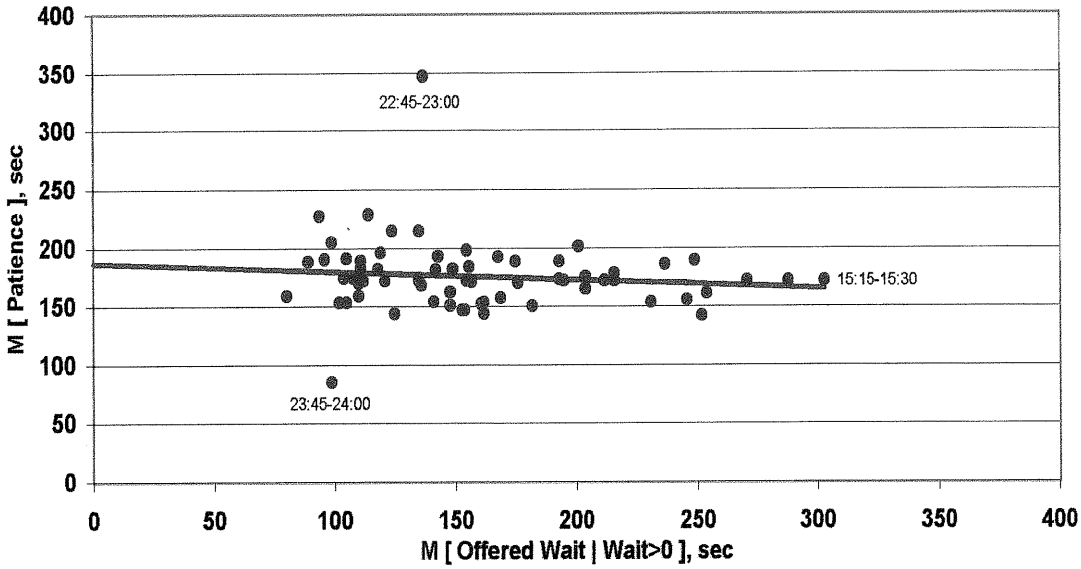


Figure 9: *NW customers.* $M[patience]$ *vs.* $M[offered\ wait | wait > 0]$; $M[\cdot]$ *stands for the* median *of the Kaplan-Meier estimator for the corresponding distribution.*

The flatness, to be compared against the slope in Figure 4, can be attributed to insensitivity of NW patience to congestion, due to their unfamiliarity with the system. As mentioned in Section 3, replacing the medians in Figure 9 with means yields statistically unreliable scatterplots – this is, in fact, the subject of ongoing research.

Two final comments (or reservations) on the use of medians. First, in the context of this paper the the mean seems to be a more natural descriptor of human perception of past performance, and is also more amenable for analysis. Hence the median is not appropriate as a basis for an adaptive theory as developed here. On the technical side, one should note that with ample censoring it is also possible for the KM median to be undefined; this happens, for example, when the whole upper half of the sample consists of customers who were patient

Rational Abandonment from Tele-Queues: Nonlinear Waiting Costs with Heterogeneous Preferences [1]

Nahum Shimkin[2] *and* Avishai Mandelbaum[3]

May 27, 2002

Abstract

We consider the modeling of abandonment from a queueing system by impatient customers. Within the proposed model, customers act rationally to maximize a utility function that weights service utility against expected waiting cost. Customers are heterogeneous, in the sense that their utility function parameters may vary across the customer population. The queue is assumed invisible to waiting customers, who do not obtain any information regarding their standing in the queue during their waiting period. Such circumstances apply, for example, in telephone centers or other remote service facilities, to which we refer as *tele-queues*. We analyze this decision model within a multi-server queue with impatient customers, and seek to characterize the Nash equilibria of this system. These equilibria may be viewed as stable operating points of the system, and determine the customer abandonment profile along with other system-wide performance measures. We provide conditions for the existence and uniqueness of the equilibrium, and suggest procedures for its computation. We also suggest a notion of an equilibrium based on sub-optimal decisions, the *myopic* equilibrium, which enjoys favorable analytical properties. Some concrete examples are provided to illustrate the modeling approach and analysis. The present paper supplements previous ones which were restricted to linear waiting costs or heterogeneous customer population.

Key words: Tele-Queues or Invisible Queues, Abandonment, Impatient Customers, Nash Equilibrium, Telephone Call Centers, Contact Centers, Multi-server Queues

---

[1]This research was partially supported by the Israeli Science Foundation, Grant 388/99-2, by the Technion V.P.R. fund for the promotion of sponsored research, and by the Fund for Promotion of Research at the Technion.

[2]Department of Electrical Engineering, Technion, Haifa 32000, Israel. e-mail: shimkin@ee.technion.ac.il

[3]Department of Industrial Engineering, Technion, Haifa 32000, Israel. e-mail: avim@tx.technion.ac.il

Rational Consistent Equilibrium

Rationality Each customer optimizes own utility

Consistency Perceived virtual waiting time dist
= Actual

Assumptions : multi-types, continuously distributed
linear waiting cost, fixed service reward

Theorem $\exists !$ rational consistent equilibrium
Fixed point of

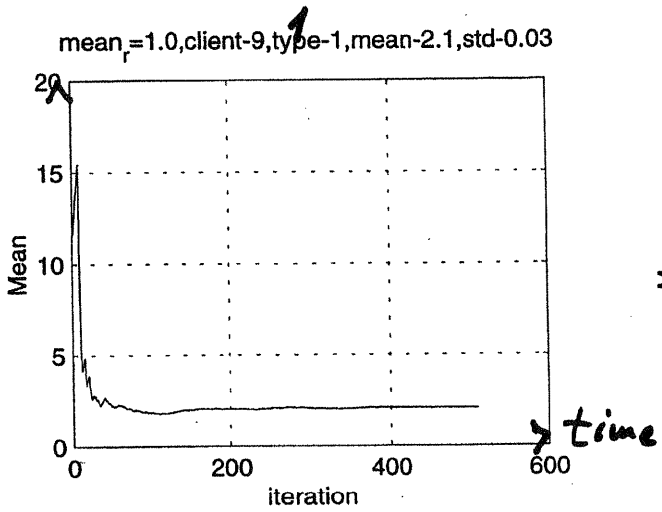
- F perceived dist of virtual wait, by all types
  - abandonment time optimized, per type
  - G patience distribution
- F actual dist of virtual wait in $M/M/N+G$

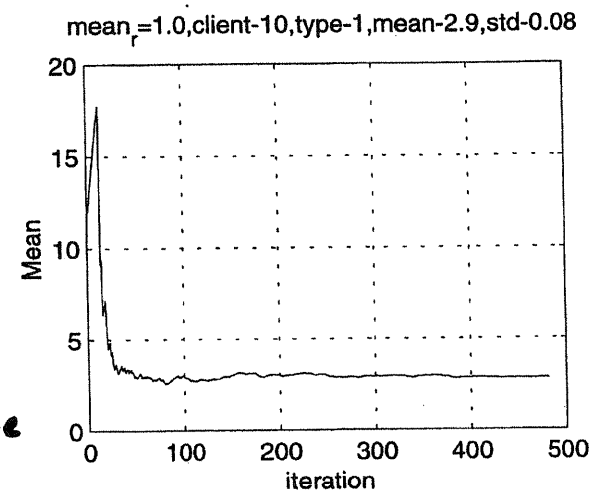
Notes:

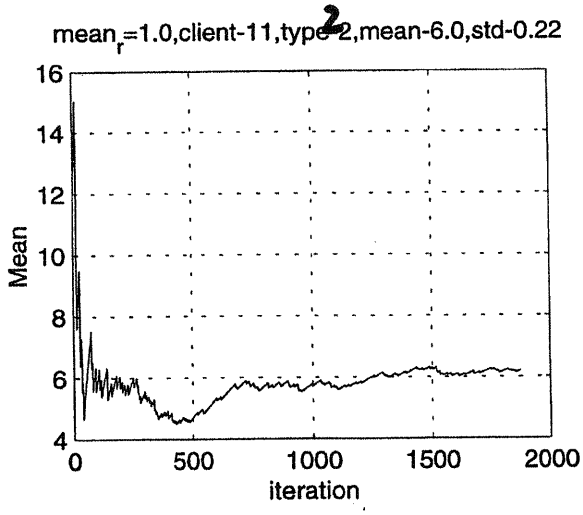
- Equilibrium explicitly computable, up to a scalar fixed-point eq.
- Equilibrium hazard-rate dist is DFR - F

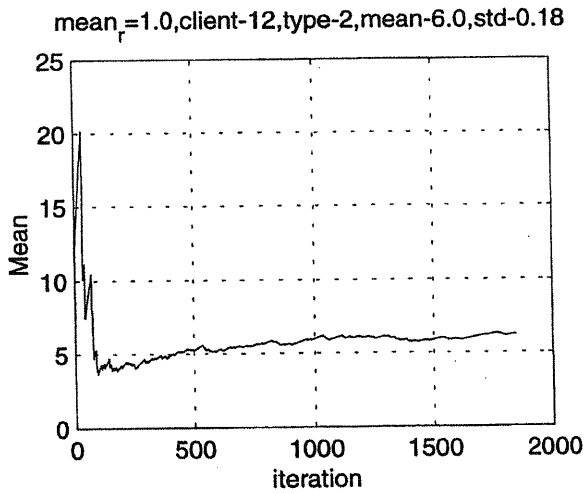
Partial Consistency : "Exponential" Lenses

Mean wait









Type-dependent learning

Censored estimation

Time-Varying Queues (Fluid focus)

Queue Lengths and Waiting Times for Multiserver Queues with Abandonment and Retrials[1]

Avi Mandelbaum
Technion Institute
Haifa, 32000, ISRAEL
avim@tx.technion.ac.il

William A. Massey
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
will@research.bell-labs.com

Martin I. Reiman
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
marty@research.bell-labs.com

Brian Rider
Courant Institute
New York, NY 10012-1185, U.S.A.
riderb@cims.nyu.edu

Alexander Stolyar
Bell Laboratories
Murray Hill, NJ 07974, U.S.A.
stolyar@research.bell-labs.com

April 7, 2000

Abstract

We consider a Markovian multiserver queueing model with time dependent parameters where waiting customers may abandon and subsequently retry. We provide simple fluid and diffusion approximations for both the queue length and virtual waiting time processes arising in this model.

These approximations, which are justified by limit theorems where the arrival rate and number of servers grow large, are compared to simulations, and perform extremely well.

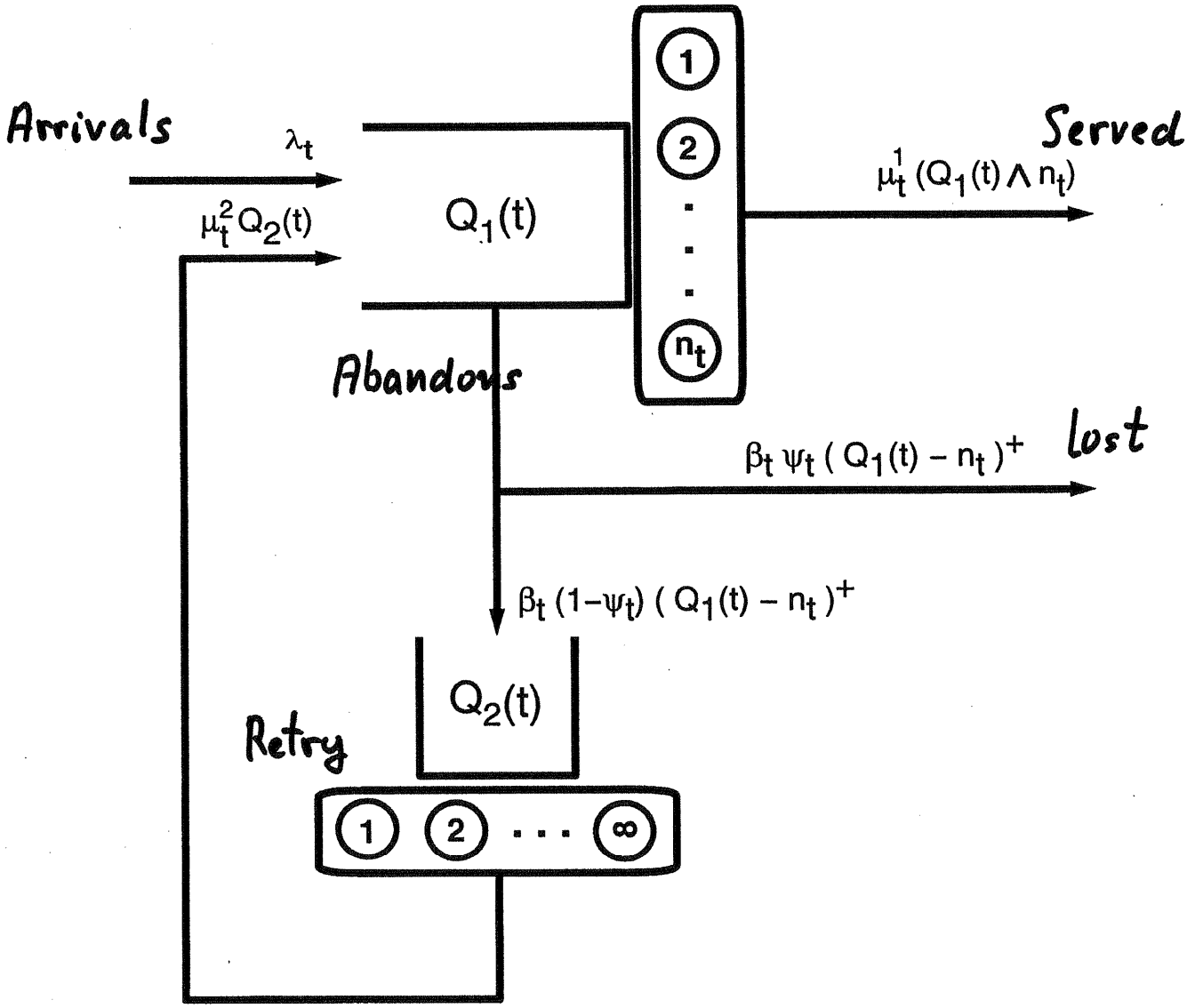
Keywords: Call Centers, Fluid Approximations, Diffusion Approximations, Multiserver Queues, Queues with Abandonment, Virtual Waiting Time, Queues with Retrials, Nonstationary Queues.

[1]Submitted to the Selected Proceedings of the Fifth INFORMS Telecommunications Conference.

Time Varying Multiserver Queues ...

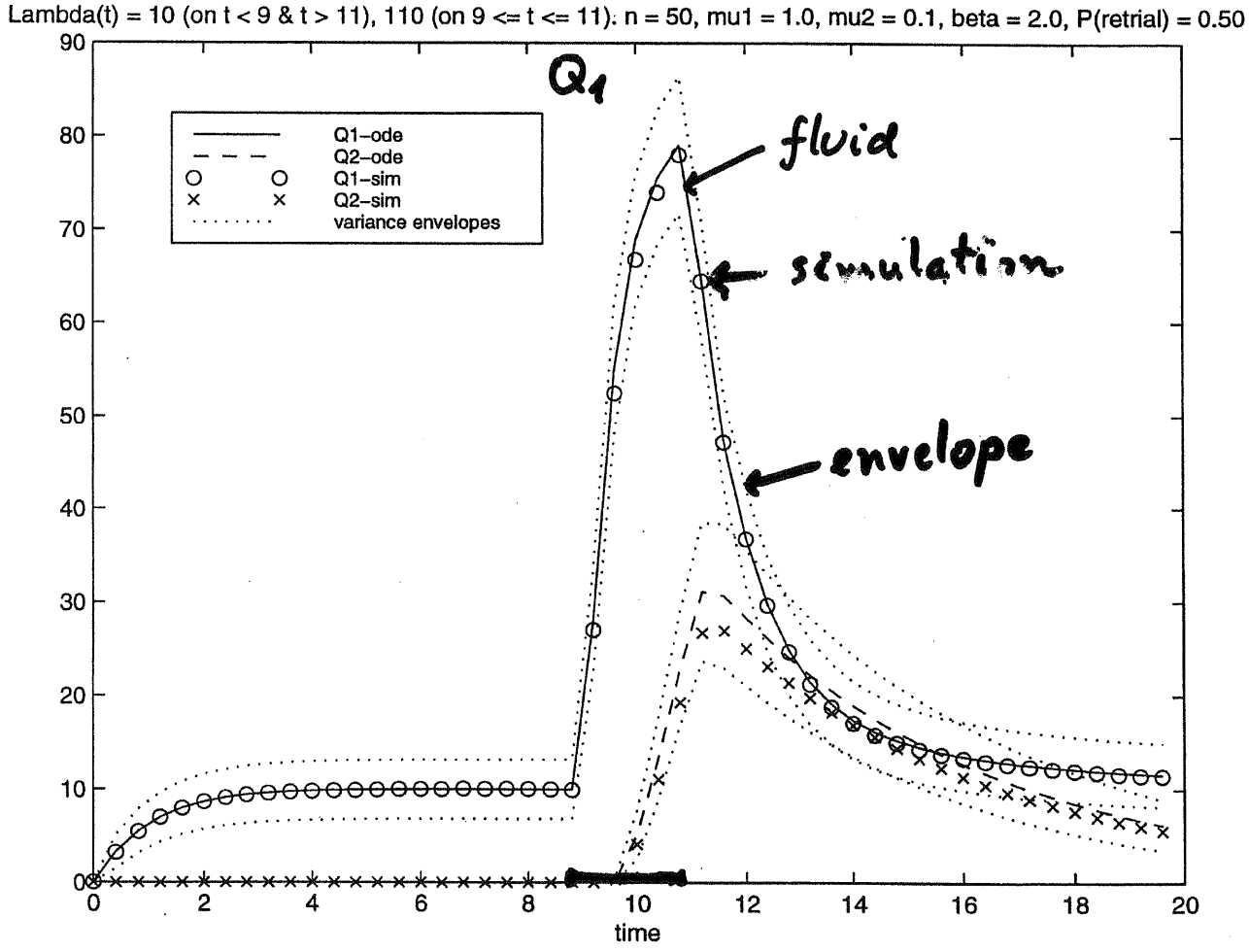
Massey, Reiman, Rider, Stolyar

Call Center: A Multiserver Queue with Abandonment and Retrials



Rush Hour

Deterministic Time-Varying Motion
Predictable Variability



$\lambda_t \equiv 10$

Peak = 110