Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective

Haipeng Shen [a] [b]

Department of Statistics and Operations Research

University of North Carolina at Chapel Hill

July 24, 2004

---

[a] Joint work with Larry Brown, Noah Gans and Linda Zhao at Wharton, Avi Mandelbaum, Anat Sakov and Sergey Zeltyn at Technion.

[b] Available at http://www.unc.edu/∼haipeng.

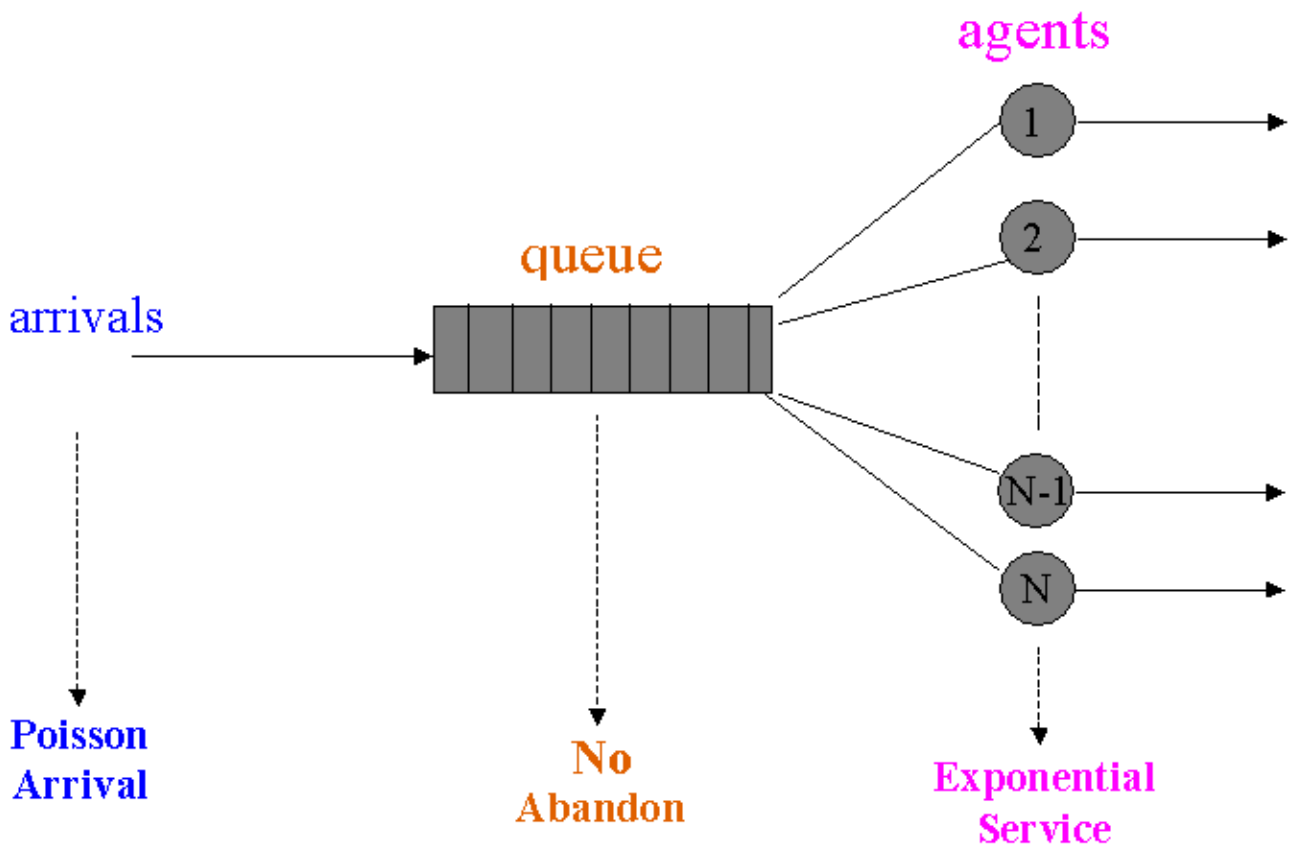
Research Goal

- Analyze data collected at a small Israeli financial call center.
- Model/understand call center operations.
  - ▷ Primitives: arrivals, service durations and patience
  - ▷ Performance measure: waiting time
- Provide forecasting for arrivals and workload.
- Serve as a prototype for further work.
  - ▷ Call-by-call data, instead of aggregated data.
  - ▷ ongoing analysis of a moderate US call center network.

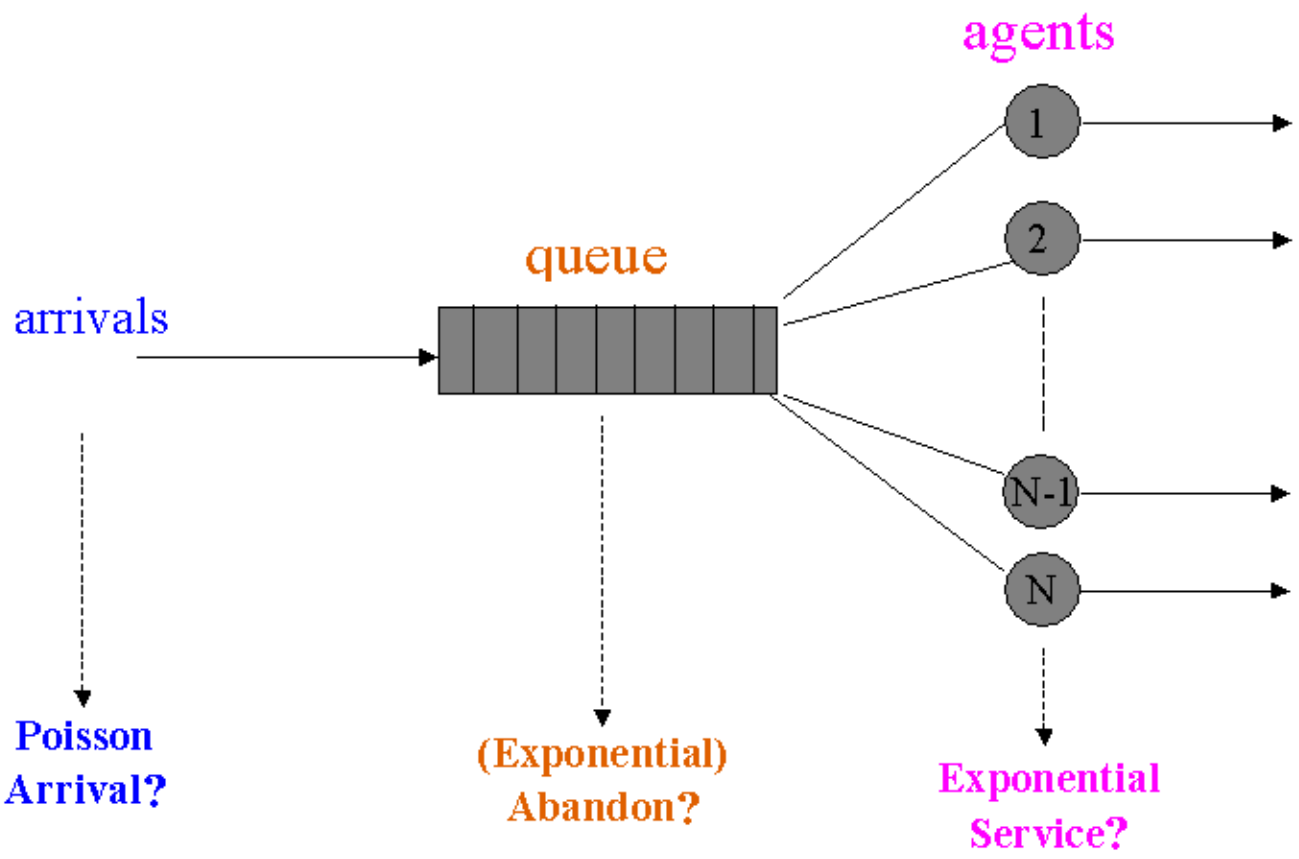
Outline

1. Queueing Theory
2. *Bank Anonymous* Call Center Data
3. Analysis
   (a) Arrivals
   (b) Service Durations
   (c) Queueing Behavior
   (d) Forecasting of Workload
4. Application of Queueing Science
5. Summary

Classical Erlang-C $(M/M/N)$



Reality?

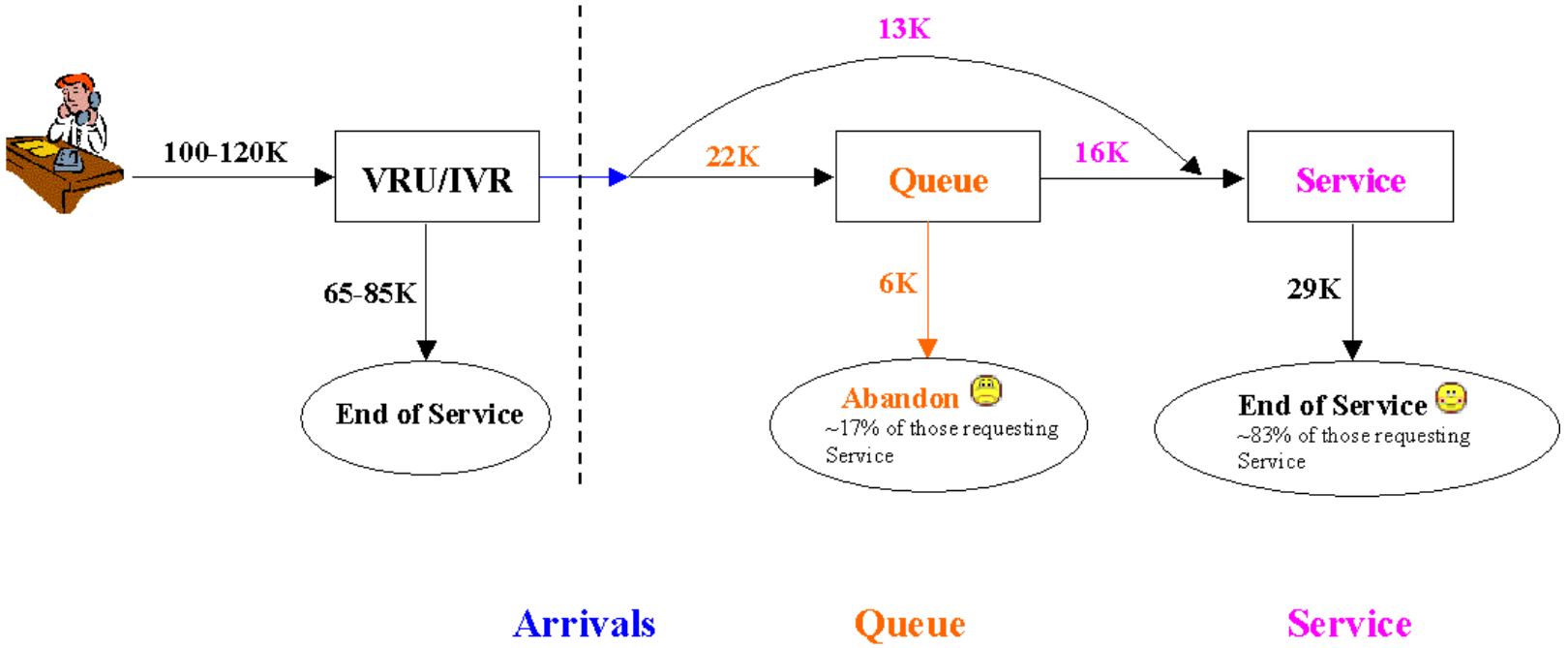

agents
1
2
queue
arrivals
N-1
N
Poisson
Arrival?
(Exponential)
Abandon?
Exponential
Service?

A Call Center of *Bank Anonymous* of Israel

- Small: 15 seats at most.
- Types of service:
  - ▷ information for current and prospective customers
  - ▷ transactions for bank accounts
  - ▷ stock-trading
  - ▷ IT support for users of the bank's website
- Working hours:
  - ▷ Sundays-Thursdays: 7AM – 12AM
  - ▷ Fridays: 7AM – 2PM
  - ▷ Saturdays: 8PM – 12AM

Event history of an incoming call

(units of rates are calls per month)



The Call Center Data

- Data ⇒ whole history of every agent-seeking call in 1999.
- 450,000 observations.
- Two operational changes:
  - ▷ Separate agent pool for Internet Consulting since Aug;
  - ▷ One aspect of the service-time data changed since Nov.
- Focus on
  - ▷ weekdays of Nov and Dec.
  - ▷ normal business hours – 7AM to midnight.

Arrivals: Inhomogeneous Poisson

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

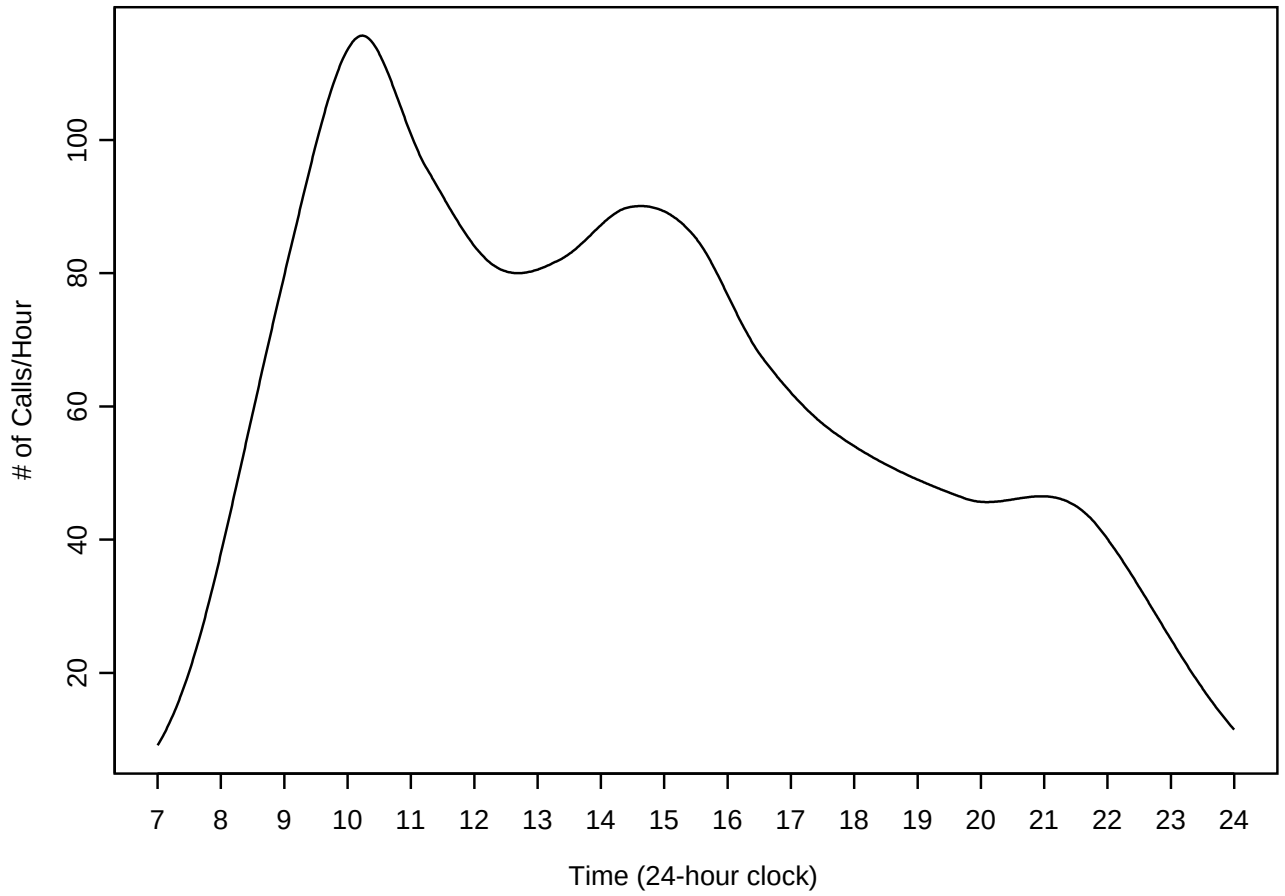


Figure 2: Arrivals (to queue or service) – Other Calls

A Test for Inhomogeneous Poisson Process

1. Break up the interval of a day into short blocks of time, say I (equal-length) blocks of length L.

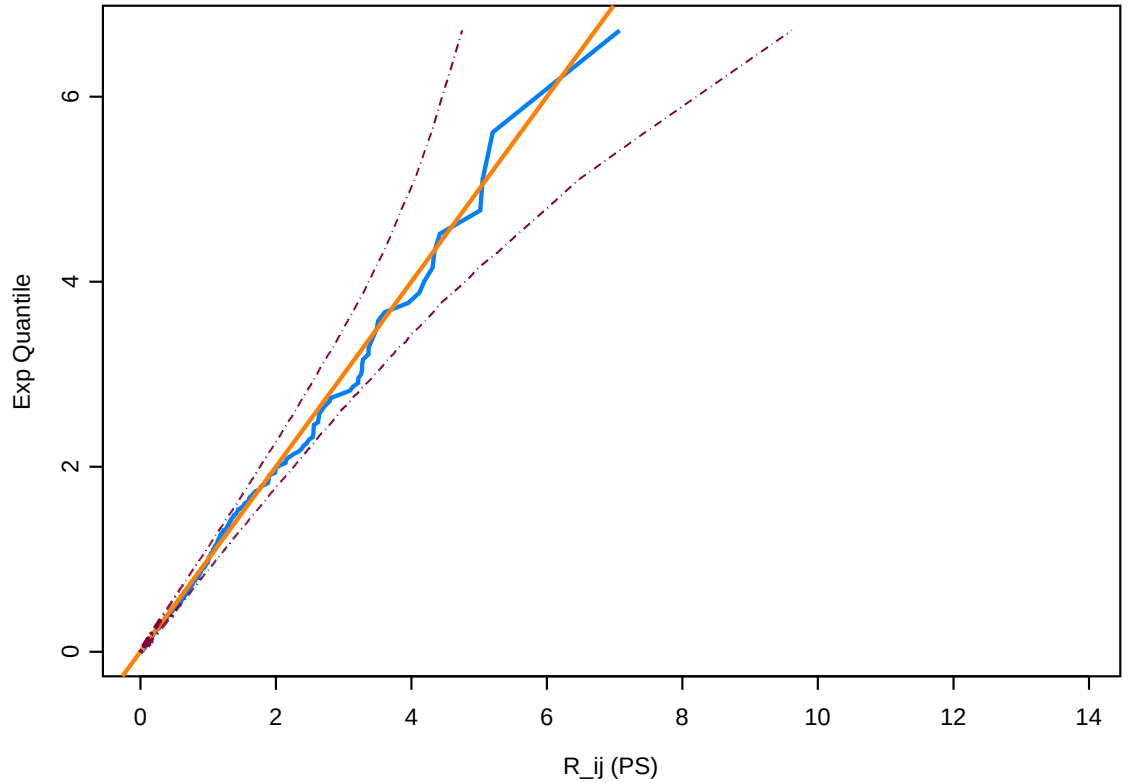
2. Let $T_{i0} = 0$ and
   T_{ij}: the j-th ordered relative arrival time in the i-th block,
   $i = 1, \ldots, I$ and $j = 1, ..., J(i)$,
   then define

$$R_{ij} = (J(i) + 1 - j)\left(-\log\left(\frac{L - T_{ij}}{L - T_{i,j-1}}\right)\right).$$

3. Under the null hypothesis that the arrival rate is constant within each given time interval, the $\{R_{ij}\}$ will be independent standard exponential variables.

4. Test for the exponential distribution; for example, Anderson-Darling (A^2) test (D'Agostino and Stephens 1986).

Figure 3: Exponential (λ=1) Quantile plot for $\{R_{ij}\}$ from Regular calls (11:12am – 11:18am)



$L = 6$ min, $n = 420$, Anderson-Darling statistic $A^2 = 0.6422$ and the P-value is 0.61.

Arrival Rate

- Determines the whole arrival process.
- Not a deterministic function of available covariates like service type, day-of-week and time-of-day.
  - ▷ Brown and Zhao (2002).
- Has to be modelled as a stochastic process.
  - ▷ Doubly-Stochastic Poisson process.
- More on this later.

Service Times

- Service time distribution is another key input for queueing theory. The mean is especially important, also the second moment.
  - ▷ system delay
  - ▷ workload
  - ▷ staffing

Figure 4: Service time cumulative distribution function (by type)

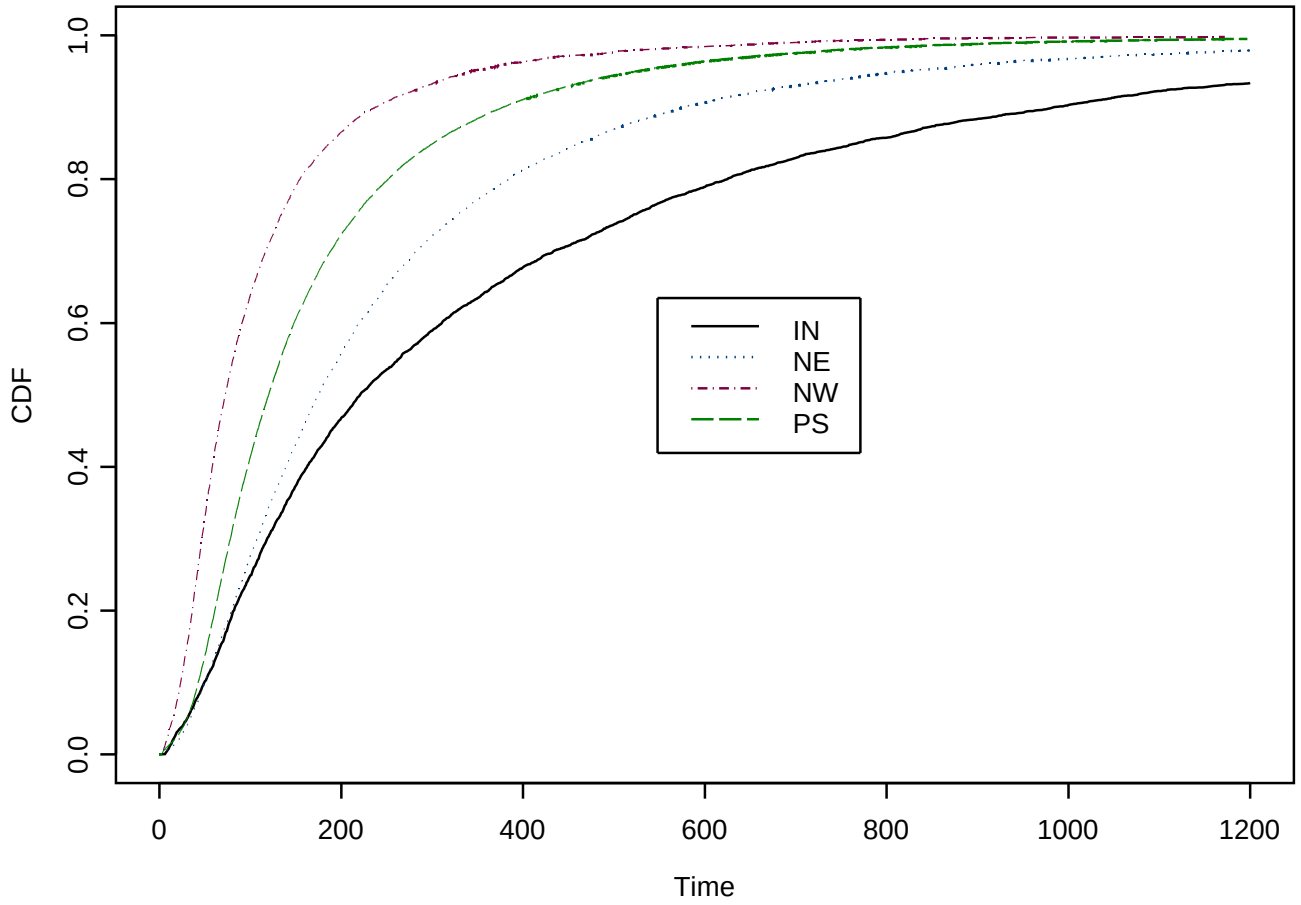


Figure 5: Histogram of Service Times (in seconds)

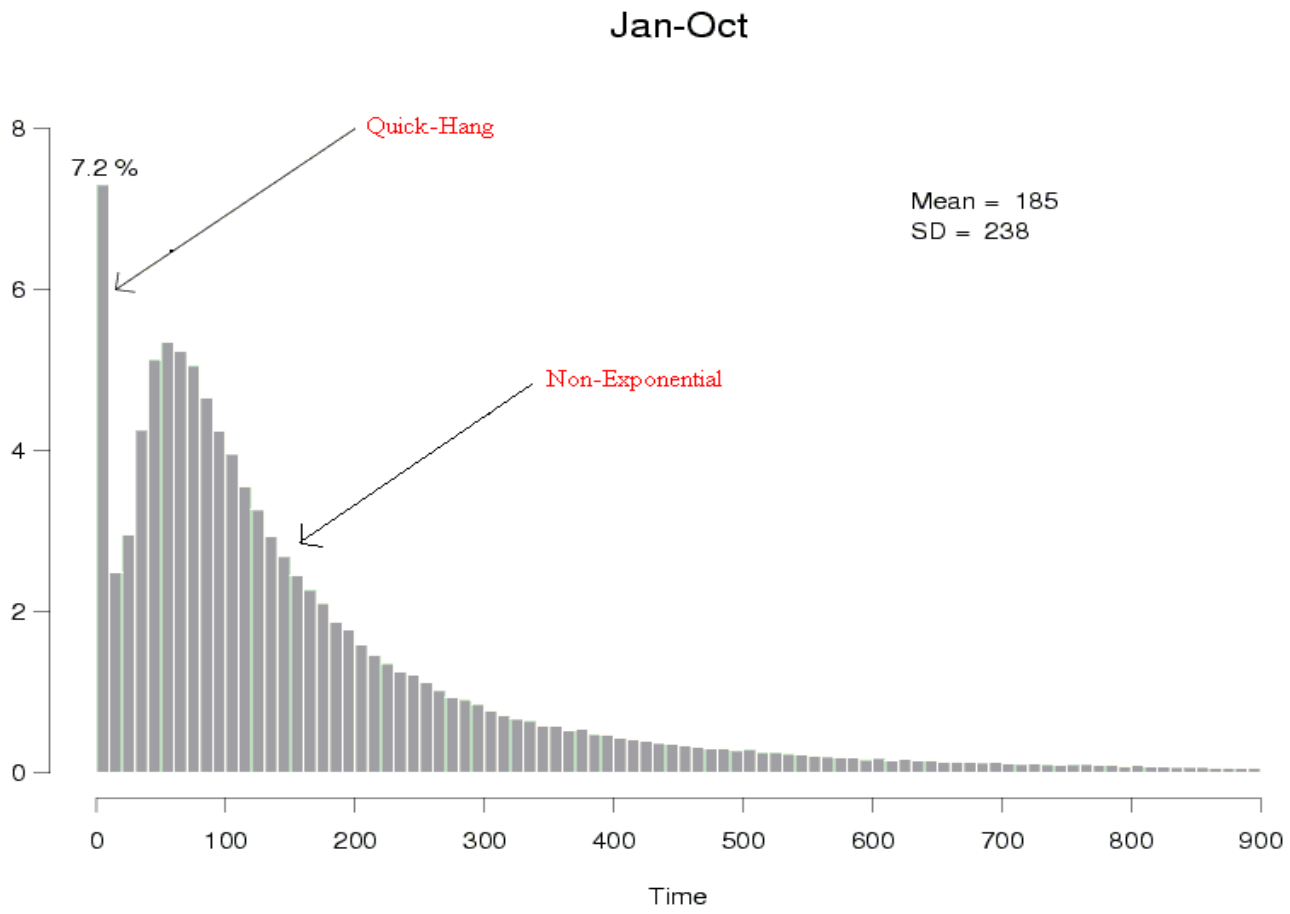
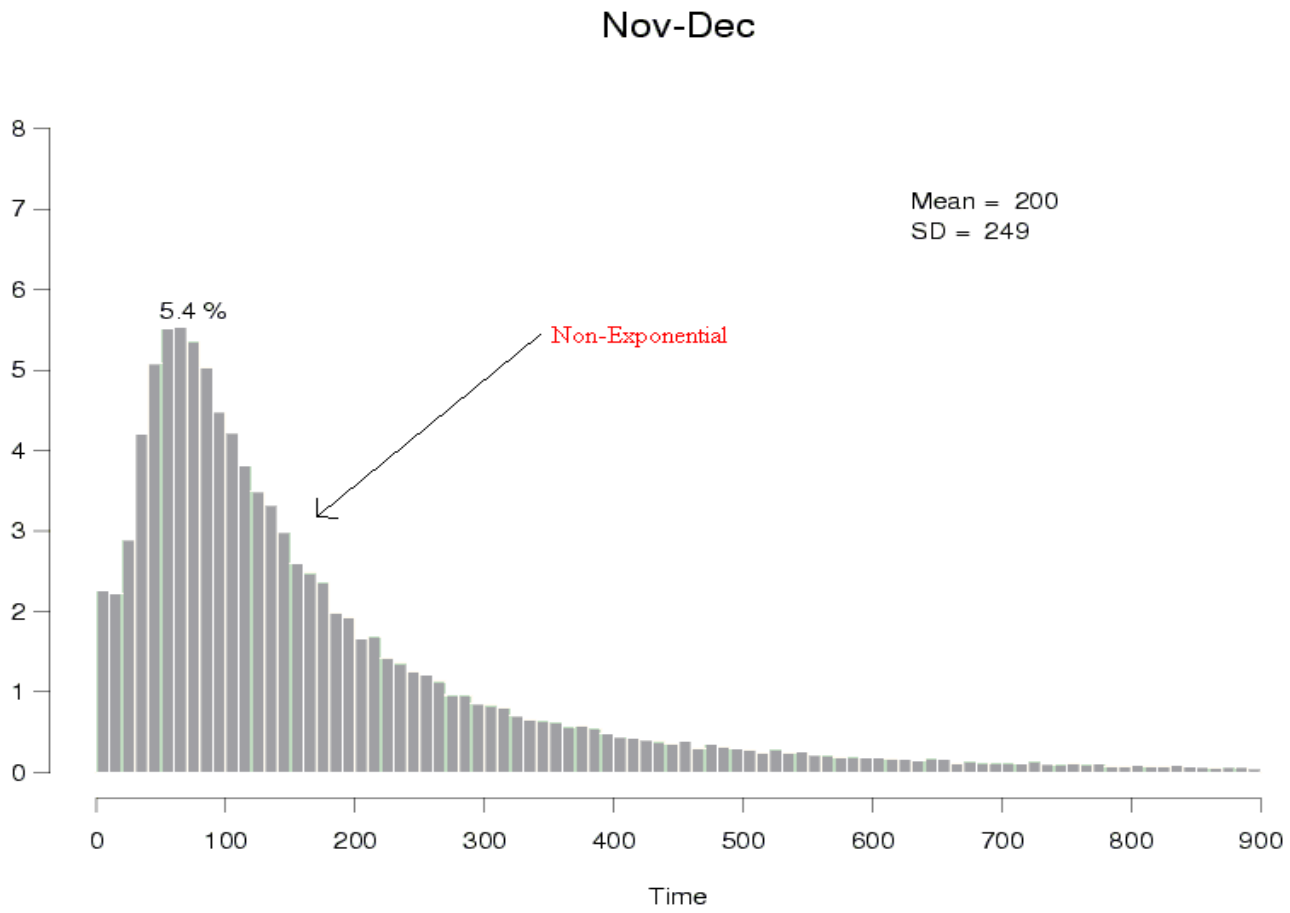


Figure 6: Histogram of Service Times (in seconds)



Service Times are Lognormal

Figure 7: Histogram of Log(Service Time) (Nov–Dec)

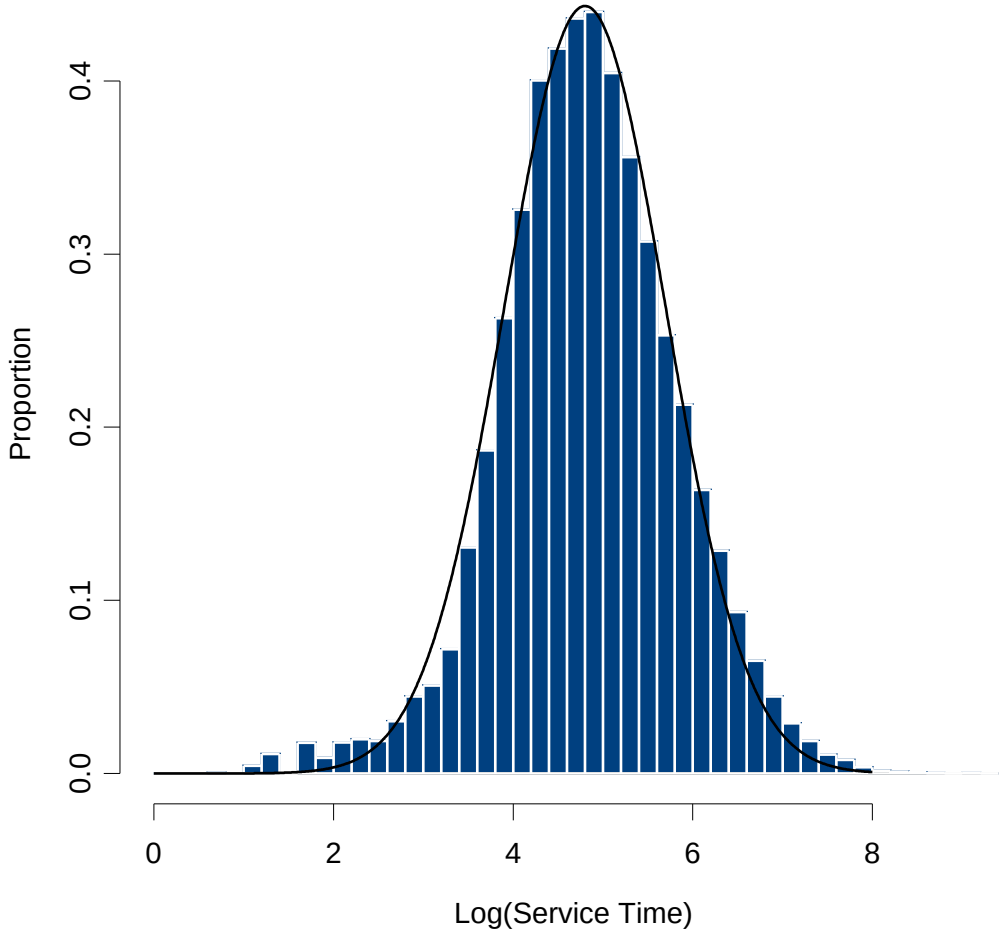
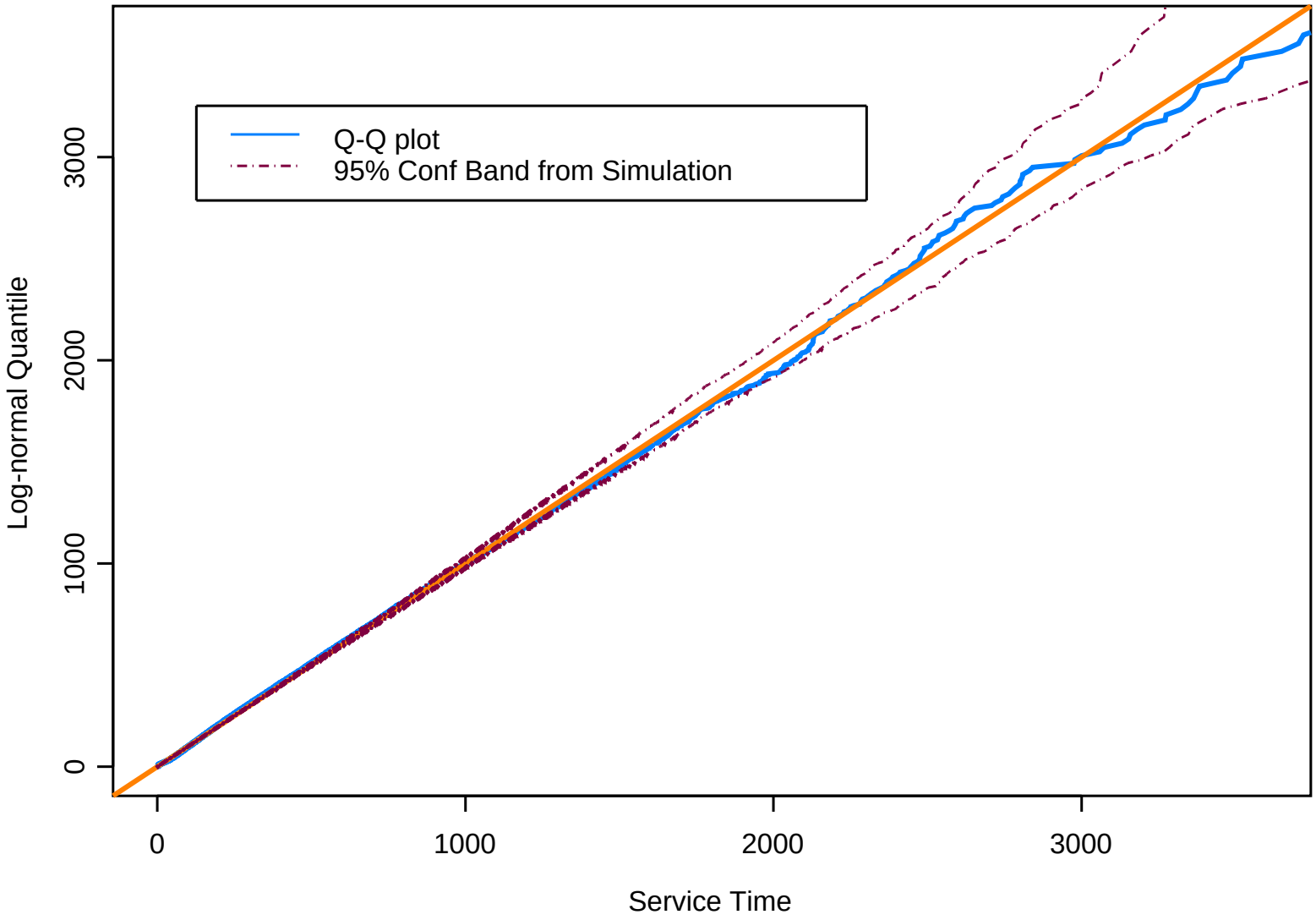


Figure 8: Log-normal QQ Plot of Service Time (Nov–Dec)



Lognormality

- Queueing data:
  - ▷ telecommunication - line usage (Bolotin 1994).
  - ▷ psychology - parallel information process time (Ulrich and Miller 1993, Breukelen 1995).
  - ▷ data from a large US financial call center ...
- So What???
  - ▷ Distribution assumption; parameter estimation.
  - ▷ System performance: Mandelbaum and Schwartz (2002).

Analysis of Service Times

- Lognormality holds
  - ▷ Overall, and at
  - ▷ Lower levels:
    - ∗ when conditioning on time-of-day;
    - ∗ for types of service, priorities of customers, individual servers and days of the week.
- Analysis: Data with lognormal errors
  - ▷ Mean service time across different categories, like service types, day-of-week
  - ▷ Mean service time as a function of time-of-day

Inference of A Lognormal Mean

Suppose $Y_i = \log(Z_i) \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. Want to estimate the lognormal mean

$$\nu = e^{\mu + \frac{1}{2}\sigma^2}$$

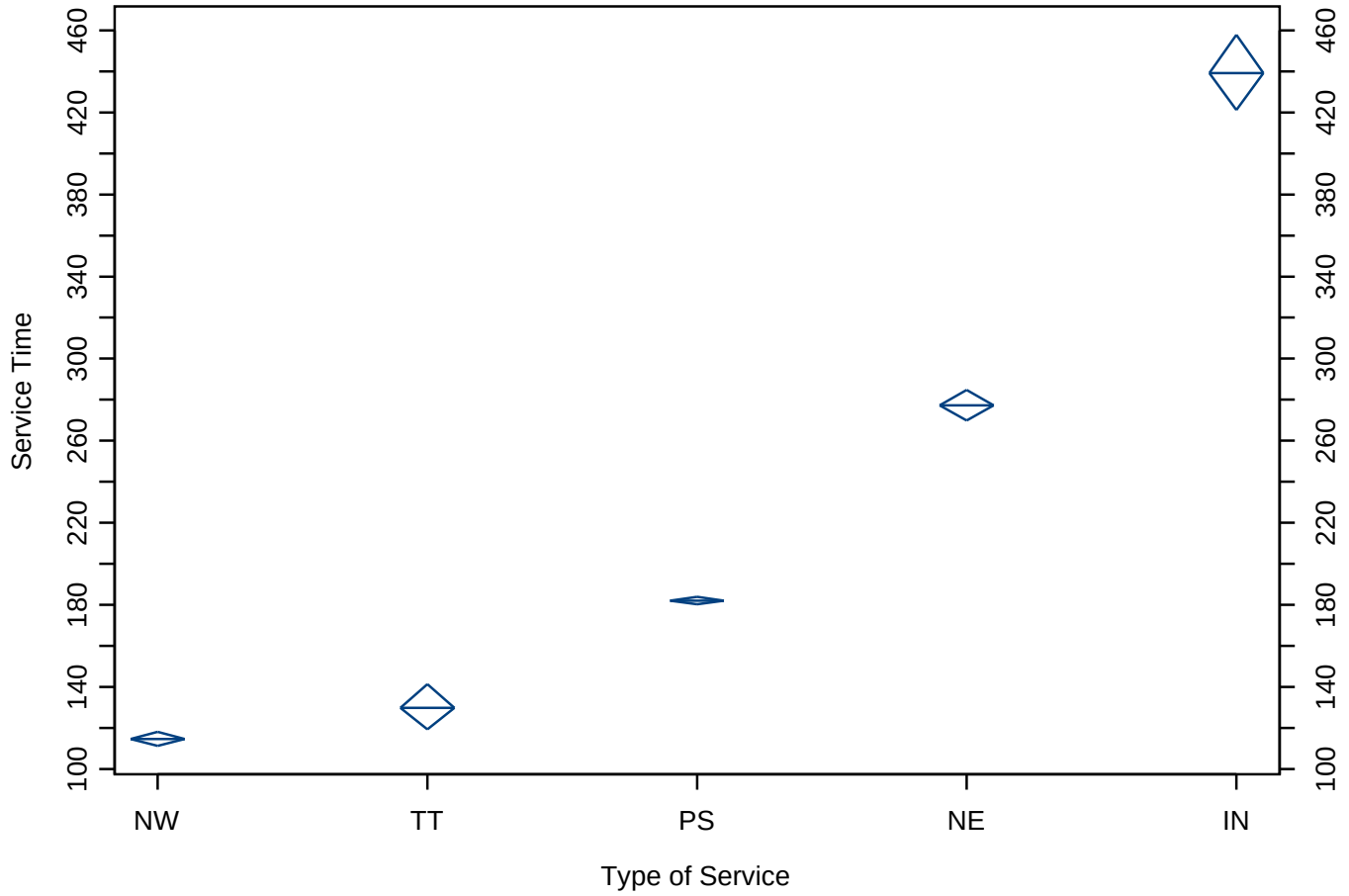
with confidence interval.

Define $S^2 = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$,

- Estimator for ν: $e^{\overline{Y} + \frac{S^2}{2(n+4)}}$
  - ▷ Shen, Brown and Zhi (2004).
  - ▷ smaller squared error risk than $\overline{Z}$, $e^{\overline{Y}}$, $e^{\overline{Y} + \frac{S^2}{2n}}$.
- Cox's confidence interval (Land 1972):

$$e^{\overline{Y} + \frac{1}{2(n-1)} S^2 \pm Z_{1-\alpha/2} \sqrt{\frac{S^2}{n(n-1)} + \frac{S^4}{2(n-1)^3}}}.$$

Figure 9: Mean Service Time vs. Service Time (95% CI)



Nonparametric Regression with Lognormal Errors

The data:

$$\{X_i, Z_i\}_{i=1}^n \overset{i.i.d.}{\sim} \{X, Z\}$$

where $Z|X = x$ has a lognormal distribution with

$$\log(Z)|X = x \sim N(\mu(x), \sigma(x)^2).$$

For example, in our call center setup,

$$X : \text{the time-of-day of a call}$$

and

$$Z : \text{the corresponding service time.}$$

We are interested in estimating

$$\nu(x) = \mathbb{E}(Z|X = x) = e^{\mu(x)+\sigma(x)^2/2}$$

with confidence band attached.

Application

- Problem:

  Model the changing pattern of the mean service time across a day.

- Data:
  - ▷ The weekdays of November and December in 1999.
  - ▷ The normal business hours – 7AM to midnight.
  - ▷ Service types:
    - ∗ Regular Services (PS)
    - ∗ Internet Consulting (IN)

Figure 10: Mean Service Time (Regular) vs. Time-of-day (95% CI) ($n = 42613$)

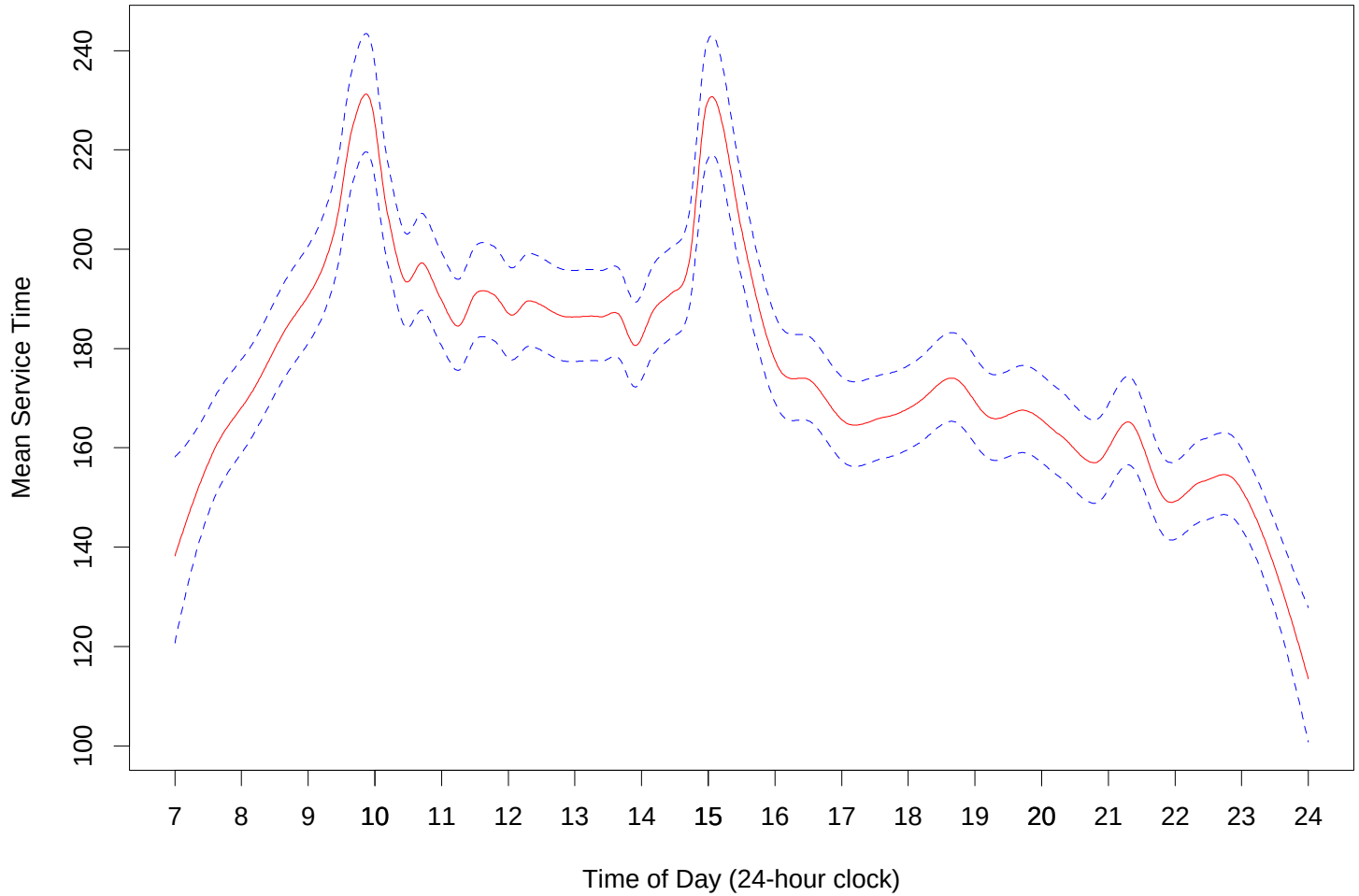
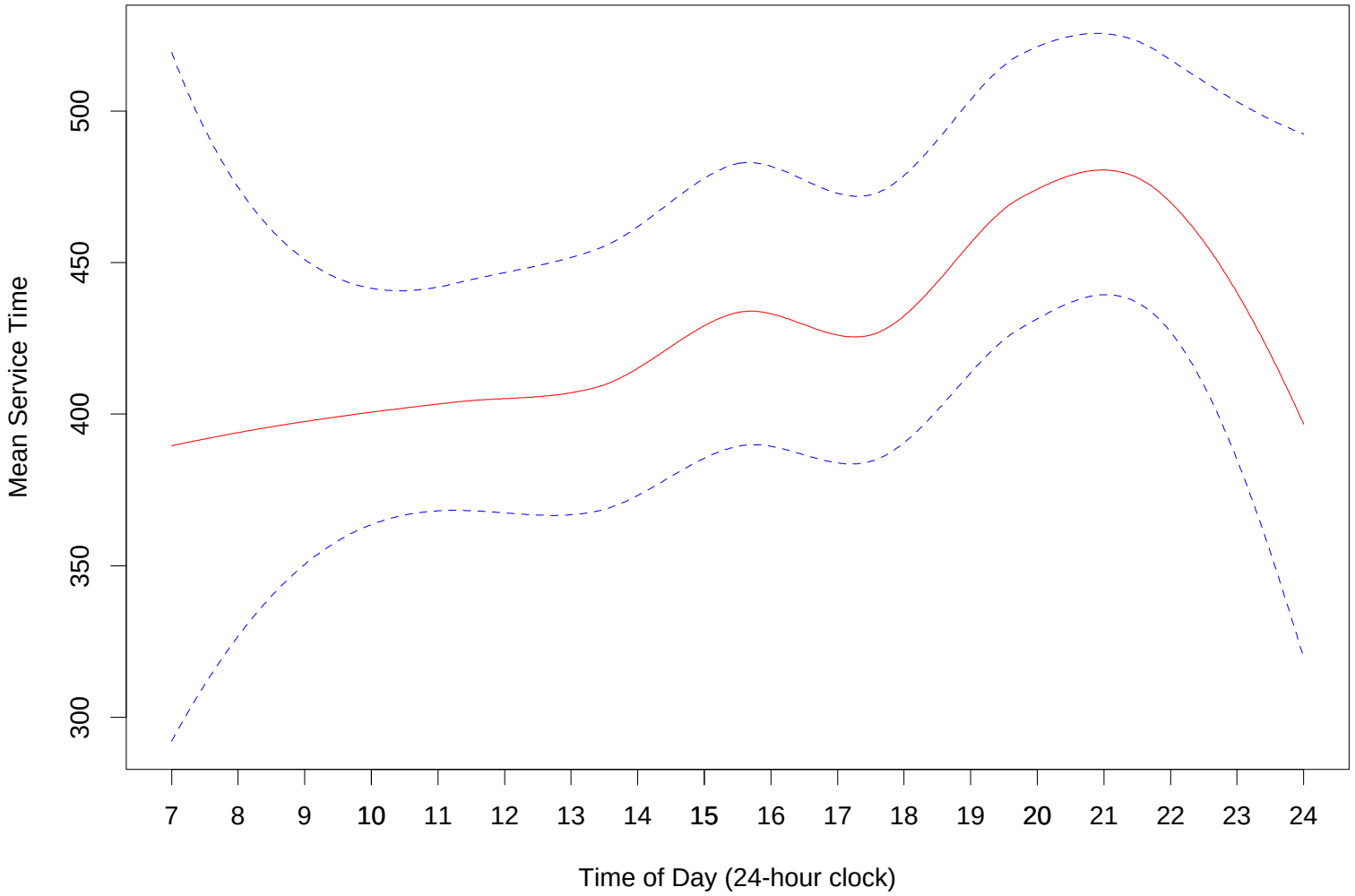


Figure 11: Mean Service Time (Internet Consulting) vs. Time-of-day (95% CI)($n = 5066$)



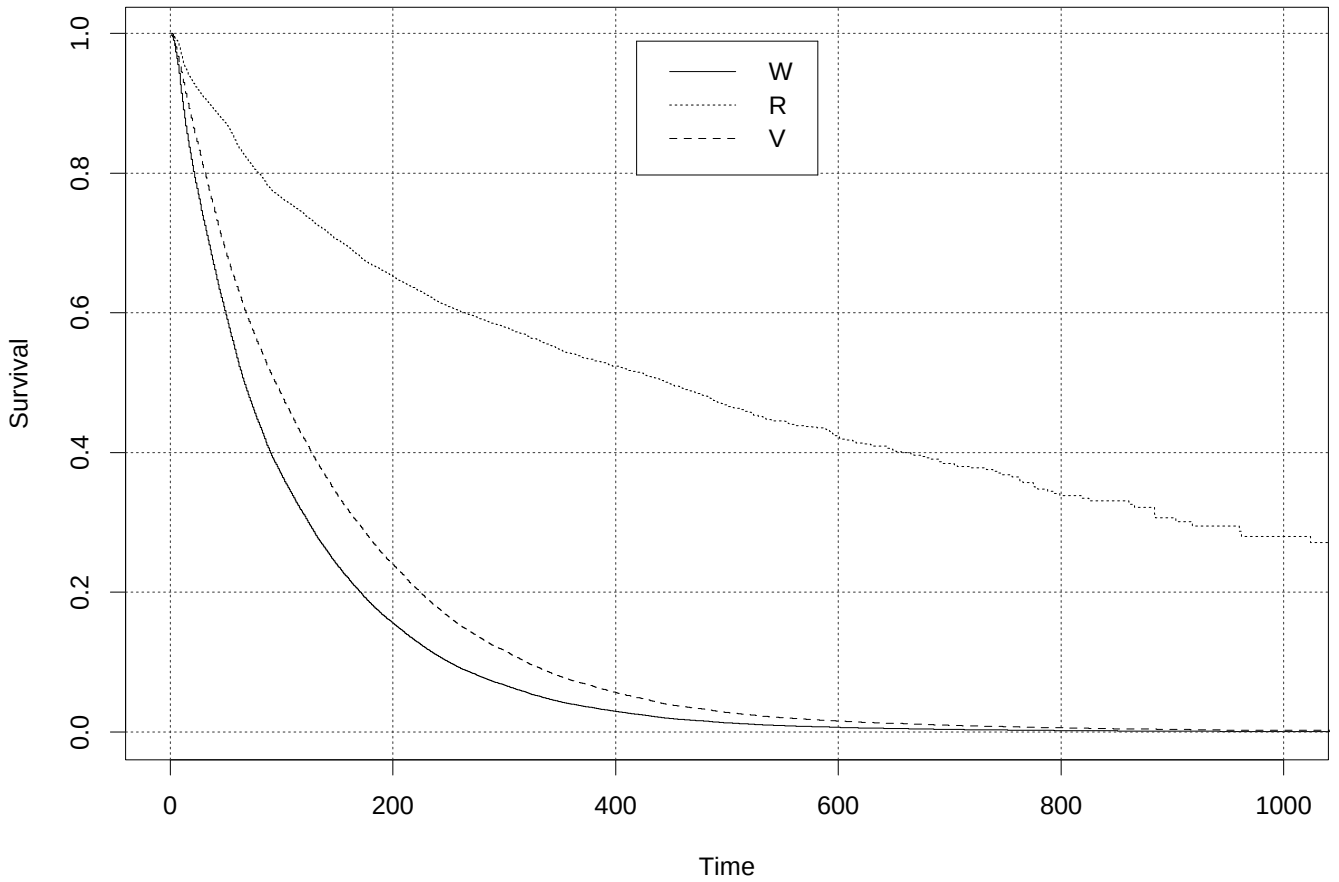
Similarity between Arrival and Service Time Patterns

- Bi-modal with similar mode locations.
- Three hypotheses:
  1. Different mix of customers; more customers with lengthier service.
  2. Agents slow down during peak periods.
  3. More abandonments; customers with relatively short service times.
- Negative correlation between the number of calls and average service time (of each quarter hour).
- Independent conditional on time-of-day.

Customer Patience and Abandonment Behavior

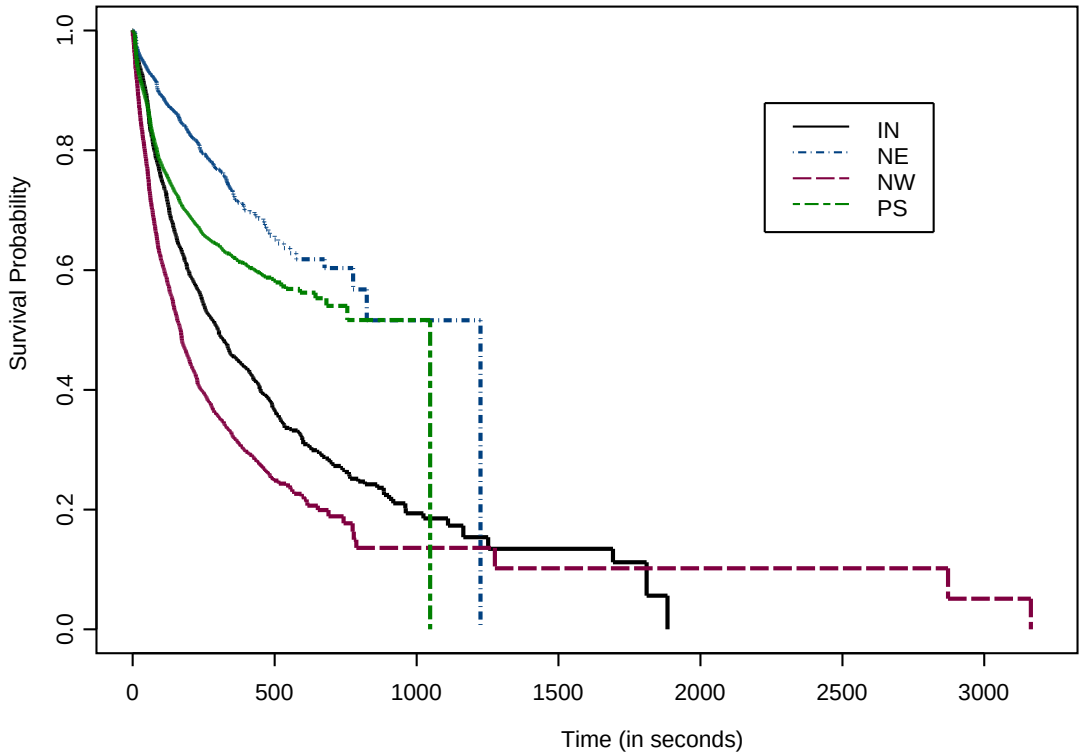
- Need to distinguish 3 Times:
  - ▷ *Virtual waiting time* V: the time a customer *needs* to wait before reaching an agent;
  - ▷ *Time willing to wait* R: the time a customer is *willing* to wait before abandoning the system;
  - ▷ *Waiting time* $W = V \wedge R$: actual observed time a customer waits.
    - ∗ approximately exponential.
    - ∗ Heavily loaded system without abandonment. (Whitt 2002)
- Also observe the indicator $I_{R<V}$.
- Thus, V and R are censored.

Stochastic Order between R, V and W



Customer Patience: Time willing to wait R

Figure 12: Survival curves for time willing to wait (Nov–Dec)



IN = INternet Consulting; NE = Stock Exchange; NW = New Customer Service; PS = Regular Service.

Figure 13: Hazard rate for the time willing to wait for Regular calls (Nov–Dec)

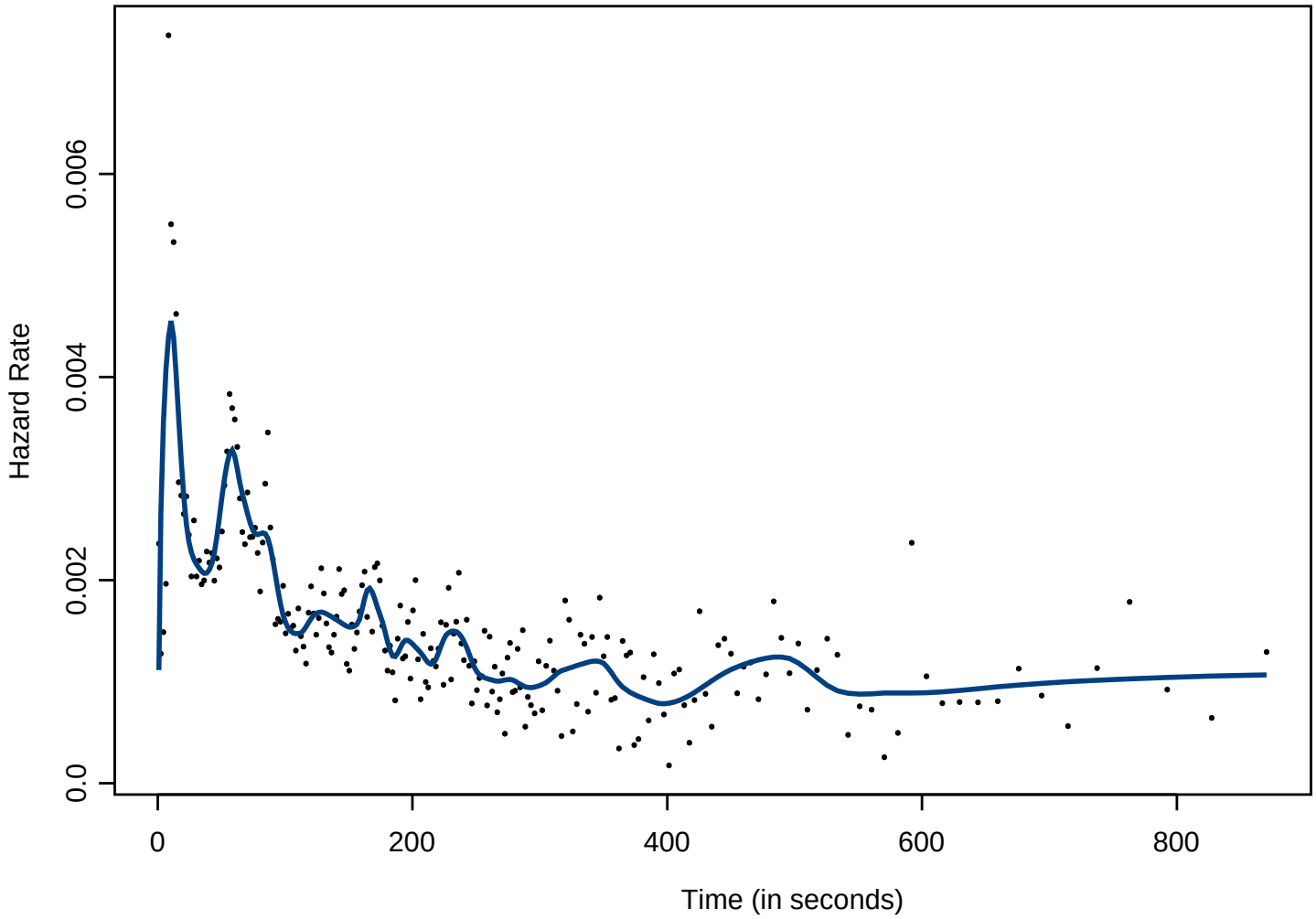
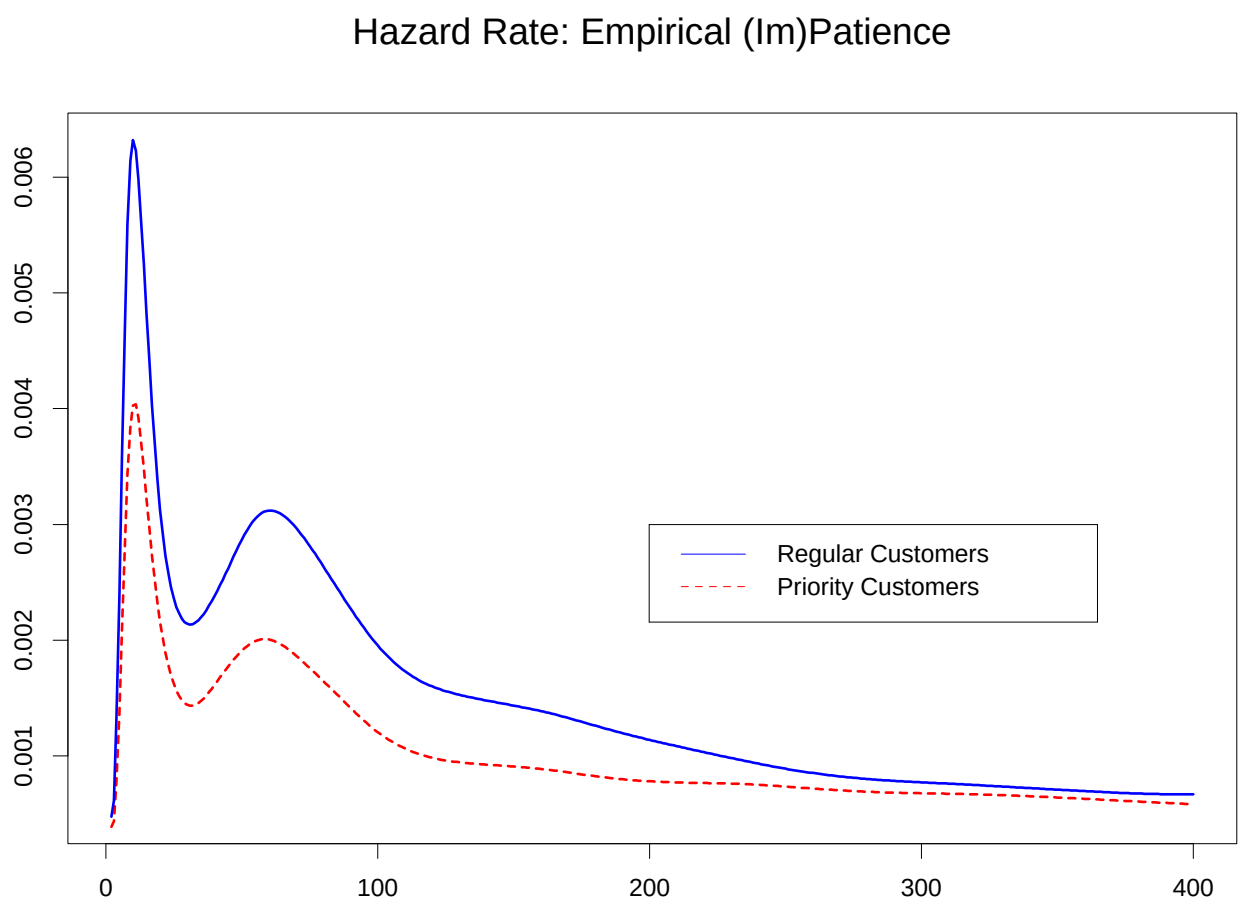


Figure 14: Comparison Between Different Priority Customers



Patience Index

Let the means of V and R be m_V and m_R, and define

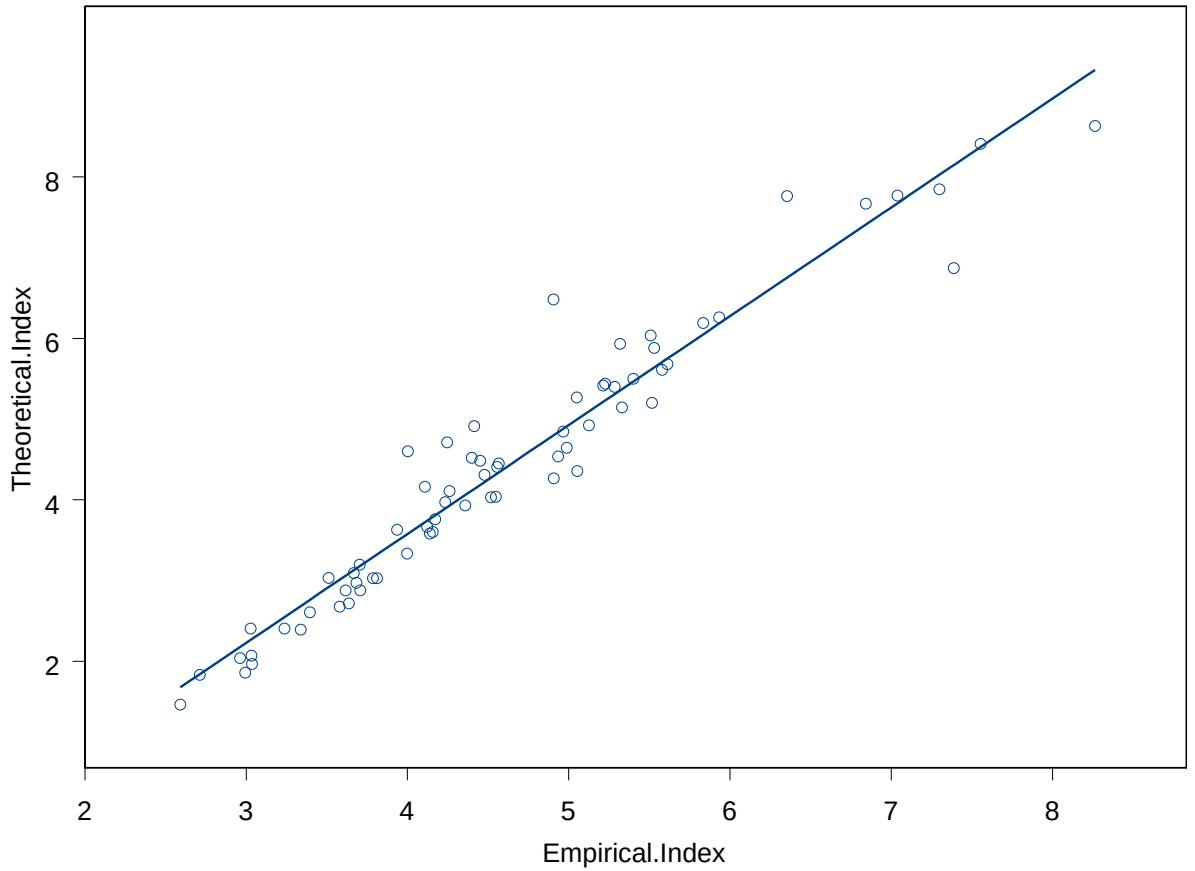
$$\text{Patience Index} \stackrel{\triangle}{=} \frac{m_R}{m_V}.$$

- new customer: 2.36; stock trading: 5.6.
- Problem: Call-by-call data; High-censoring.
- Ancillary measure:

$$\text{Empirical Index} \stackrel{\triangle}{=} \frac{\#\text{ served}}{\#\text{ abandoned}}.$$

  - ▷ The usual plug-in MLE for Patience Index if V and R are independent exponential.
  - ▷ Works (surprisingly) well empirically.

Figure 15: Patience Indices: empirical vs. theoretical ($R^2 = 0.94$)



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 various queueing models and setting staffing levels.
  - ▷ $\sqrt{\cdot}$ Safety-Staffing (Whitt 1992; Garnett, Mandelbaum, Reiman (2002); Borst, Mandelbaum, Reiman (2002)):

$$N = \left\lceil L(t) + \beta(t)\sqrt{L(t)} \right\rceil.$$

- need an accurate forecast for $L(t)$ (maybe) with prediction confidence bounds.

Prediction of Arrival Rate $\Lambda(t)$

- Important for workload forecasting, agent staffing and capacity planning.
- Regularity:
  - ▷ Inter-day dependence: today/tomorrow, weekly, monthly, seasonal, yearly, ...
  - ▷ Intra-day dependence: morning/afternoon/night, ... Avramidis, Deslauriers and L'Ecuyer (2004)
- Anomaly:
  - ▷ Holiday, Promotion, System (hardware/software) failure, ...
  - ▷ Shen and Huang (2004): anomaly detection, feature extraction

Prediction of $\Lambda(t)$

- $\Lambda(t)$ is not a deterministic function of time of day, day of week and type of customer.
  - ▷ Verified by a formal test in Brown and Zhao (2002).
  - ▷ Jongbloed and Koole (2001): Gamma-Poisson model
- Random-effects model.
  - ▷ Regular (non-holiday) weekdays from Aug. to Dec. indexed by j;
  - ▷ Divide the regular working hours from 7AM through midnight into 68 quarter hours indexed by k;

Prediction of $\Lambda(t)$

▷ N_{jk}: number of arrivals within the k-th quarter hour of the j-th day.

▷

$$N_{jk} = \text{Poiss}(\Lambda_{jk}), \;\; \Lambda_{jk} = R_j \tau_k + \varepsilon_{jk},$$

where

* τ_k: fixed deterministic quarter-hourly effects with $\sum \tau_k = 1$;
* R_j: suitable random daily effects;
* ε_{jk}: (Gamma) random errors.

▷ Similar to Jongbloed and Koole (2001).

* one-way vs. two-way
* Estimation model, not a prediction model.
* Correlation structure should be added.

A Property of Poisson Variables

Suppose $X \sim \text{Poiss}(\lambda)$, then Brown, Zhang and Zhao (2002) showed that, asymptotically,

$$V = \sqrt{X + 1/4} \overset{app.}{\sim} N(\sqrt{\lambda}, \frac{1}{4})$$

with good accuracy even for small λ.

An Equivalent Gaussian Model

- Let $V_{jk} = \sqrt{N_{jk} + \frac{1}{4}}$;
- Gaussian model:

$$V_{jk} = \theta_{jk} + \varepsilon^*_{jk} \quad \text{with} \quad \varepsilon^*_{jk} \overset{iid}{\sim} N\left(0, \tfrac{1}{4}\right),$$
$$\theta_{jk} = \alpha_j \beta_k + \varepsilon'_{jk},$$
$$\alpha_j = \mu + \gamma V_{j-1,+} + \varepsilon^{**}_j,$$

  where $\varepsilon^{**}_j \sim N(0, \sigma^{**2})$, $\varepsilon'_{jk} \sim N(0, \sigma^2_\varepsilon)$, $V_{j,+} = \sum_k V_{jk}$, and ε^{**}_j and ε'_{jk} are independent of each other and of values of $V_{j',k}$ for $j' < j$.

- α_j: random effect with an AR(1) type structure.
- $\sum \beta_k^2 = 1$.

Prediction of Tomorrow's Λ_k

- Following today's value of V_+, tomorrow's θ_k is predicted to be

$$\hat{\theta}_k = \hat{\beta}_k \left(\hat{\mu} + \hat{\gamma} V_+\right)$$

  as an estimate of

$$\theta_k = \beta_k \left(\mu + \gamma V_+ + \varepsilon^{**}\right) + \varepsilon \tag{1}$$

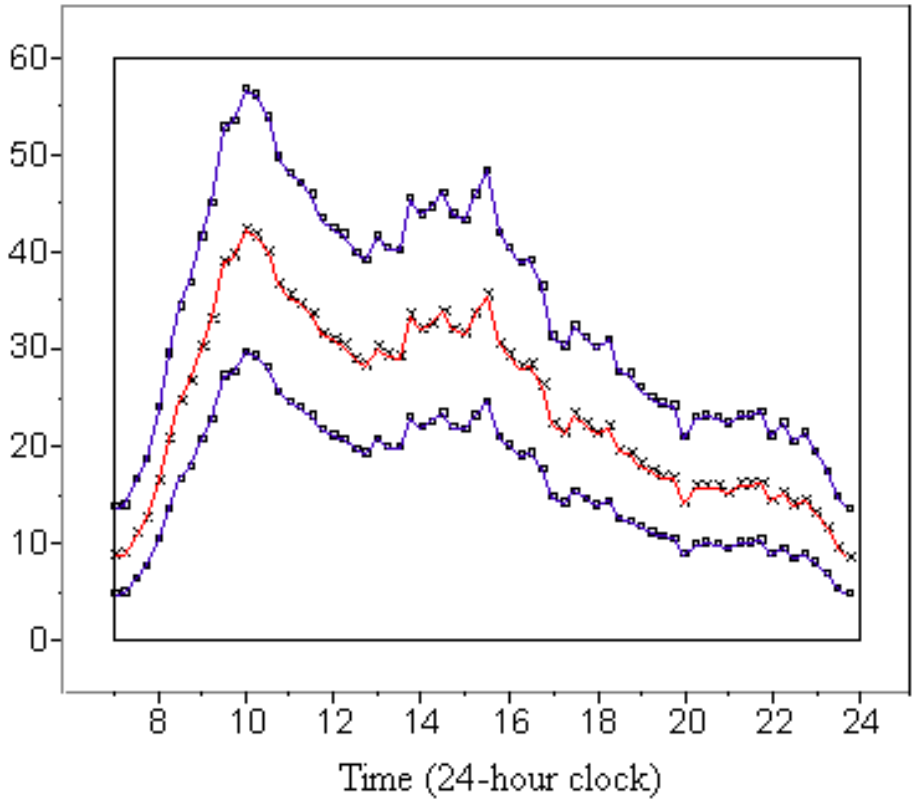
  where $\varepsilon^{**} \sim N(0,\, \sigma^{**2})$ and $\varepsilon \sim N(0,\, \sigma_\varepsilon^2)$ are independent.

- $$\hat{\Lambda}_k = \hat{\theta}_k^2 = \hat{\beta}_k^2 \left(\hat{\mu} + \hat{\gamma} V_+\right)^2 .$$

- $\mathrm{Var}(\hat{\theta}_k)$ can be derived from (1), which can be used to calculate prediction interval for $\hat{\theta}_k$.

- The above interval can be squared to get prediction interval for $\hat{\Lambda}_k$.

Figure 16: 95% prediction intervals for, Λ, following a day with $V_+ = 340$. ("$V_+ = 340$" $\Rightarrow$ "$N_+ = 1800$" ($> \overline{N}_+ = 1570$))



Vertical axis is prediction of # of arrivals/qtr. hr..

Forecasting of the Load

- Point estimate: $\hat{L}(t) = \hat{\Lambda}(t)\hat{\nu}(t)$.
- Approx. 95% Prediction Interval:

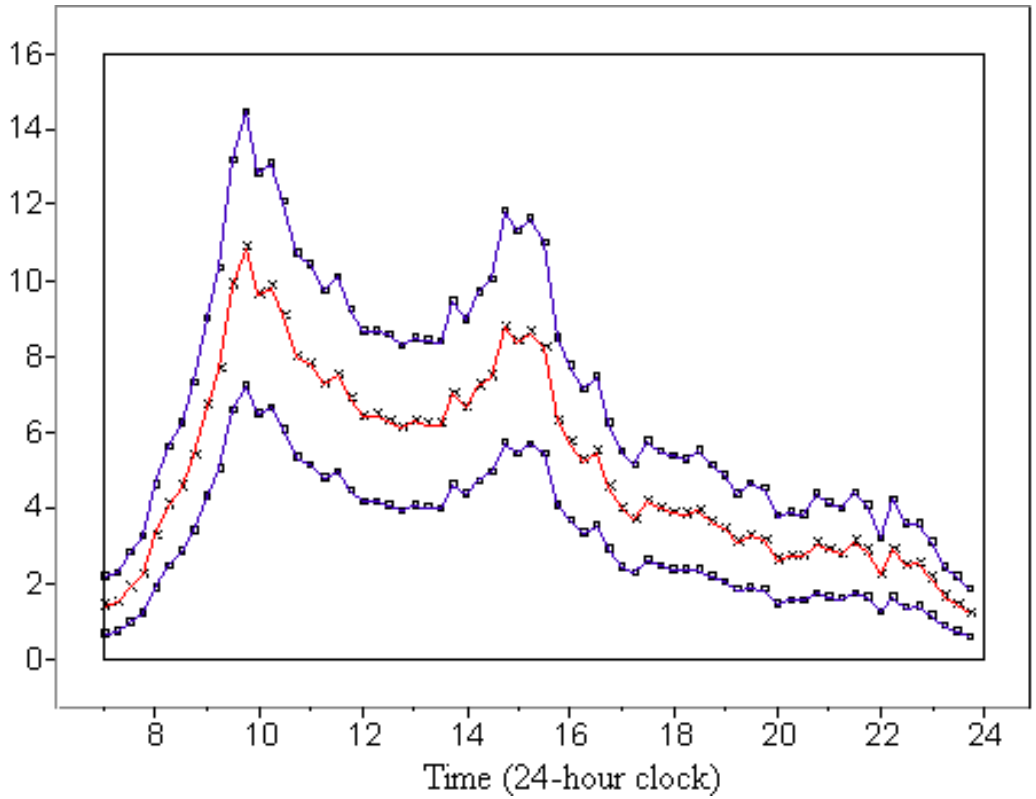
$$\hat{L}(t) \pm 2\hat{L}(t)\widehat{PCV}(\hat{L})(t)$$

where $PCV =$ "$Prediction\ CV$" $= \frac{Prediction\ S.E.}{Mean}$
and

$$\begin{aligned} & \widehat{PCV}(\hat{L})(t) \\ = \quad & \sqrt{\widehat{PCV}^2(\hat{\Lambda})(t) + \widehat{PCV}^2(\hat{\nu})(t) + \widehat{PCV}^2(\hat{\Lambda})(t) \cdot \widehat{PCV}^2(\hat{\nu})(t)} \\ \approx \quad & \sqrt{\widehat{PCV}^2(\hat{\Lambda})(t) + \widehat{PCV}^2(\hat{\nu})(t)} \end{aligned}$$

given the conditional independence of $\hat{\Lambda}(t)$ and $\hat{\nu}(t)$.

Figure 17: 95% prediction intervals for the load, L, following a day with $V_+ = 340$.



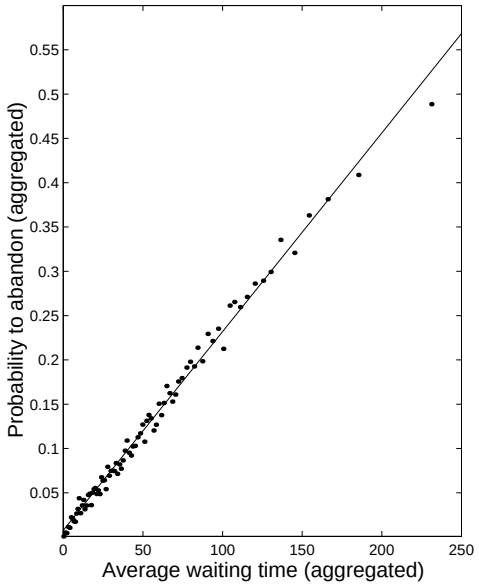
Vertical axis is workload/qtr. hr..

Applications of Queueing Science

- On Patience and Waiting
- On Efficiency and Service Levels
- Fitting the $M/M/N+M$ model (Erlang-A)

% Abandoment=$\frac{E(W)}{E(R)}$

- Exponential patience. (Zohar et al. 2002)



- Mandelbaum and Zeltyn (2004):
  - ▷ Impact of patience dist. on $M/M/N+G$ performance.

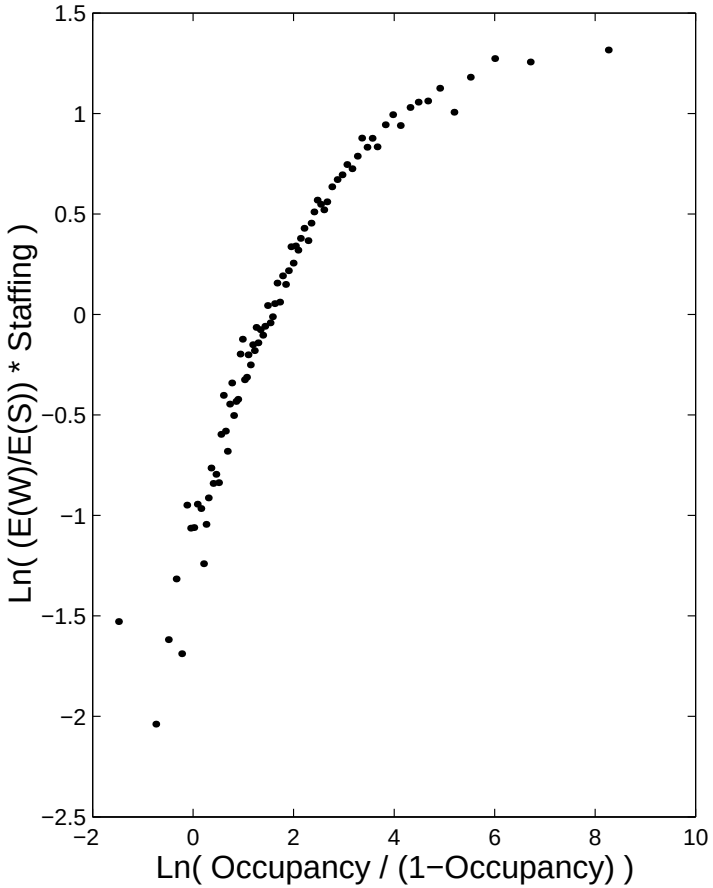
On Efficiency and Service Levels

- $M/G/N$ model
- Heavy traffic analogy of the Khintchine-Pollaczek Formula (Whitt 1993):

$$\boxed{\frac{N}{\mathrm{E}(G)} w \approx K_G \frac{\rho}{1-\rho}}.$$

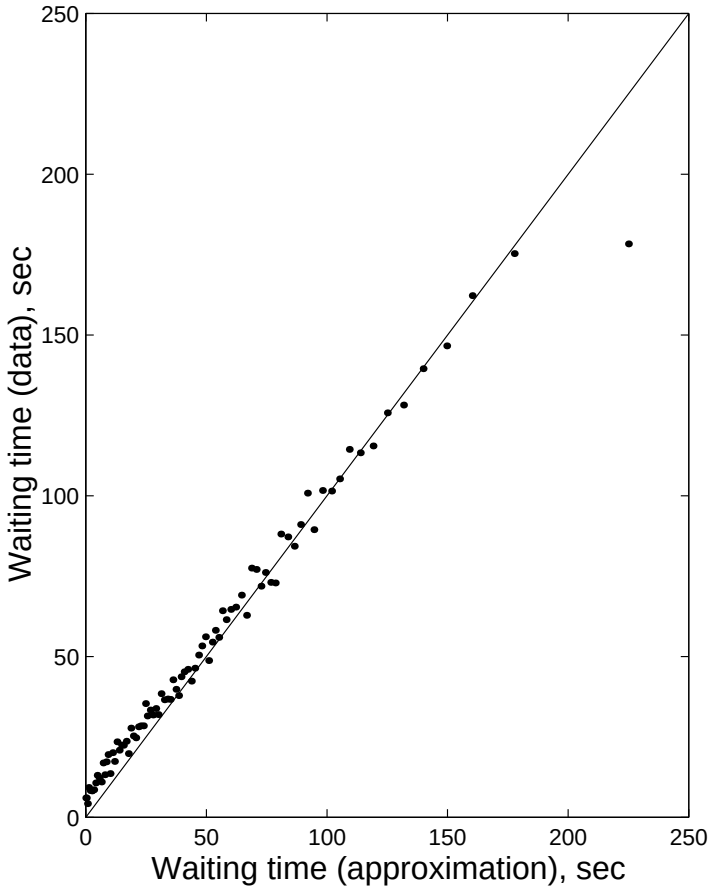
  - ▷ w: average waiting time in system;
  - ▷ ρ: agent occupancy;
  - ▷ Approximate linear relationship between w and $\frac{\rho}{1-\rho}$.

Failure!


1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
0
2
4
6
8
10
Ln((E(W)/E(S)) * Staffing)
Ln(Occupancy / (1−Occupancy))

Robustness of Erlang-A

Figure 18: Waiting Time: Data Ave. vs. Erlang-A Prediction



Summary

- Arrivals
  - ▷ Testing inhomogeneous Poisson process
  - ▷ Test for applicability of fixed effects model
  - ▷ Forecasting Poisson arrival rate
    - ∗ Sqrt-Gaussian Model with an AR structure
- Service Times
  - ▷ Lognormal
  - ▷ Stochastic order among service types
  - ▷ Model daily average service time
- Abandonment Behavior, Customer Patience
  - ▷ stochastic order among patience of different types/priorities of customers
  - ▷ patience index
- Workload Forecasting
- Applications of Queueing Science

Future Research

- Analysis of a much larger US bank call center
- Gamma-Poisson model of call center arrivals
- Mixture modelling of service times
- Survival analysis under high-censoring with large sample size
- Learning curve modelling