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

Lawrence D. Brown [a] [b]

The Wharton School

University of Pennsylvania

November 5, 2002

[a] Joint work with Noah Gans, Avishai Mandelbaum, Anat Sakov, Haipeng Shen, Sergey Zeltyn and Linda Zhao.

[b] Manuscript and talk slides will be available at http://ljsavage.wharton.upenn.edu/∼lbrown.

Outline

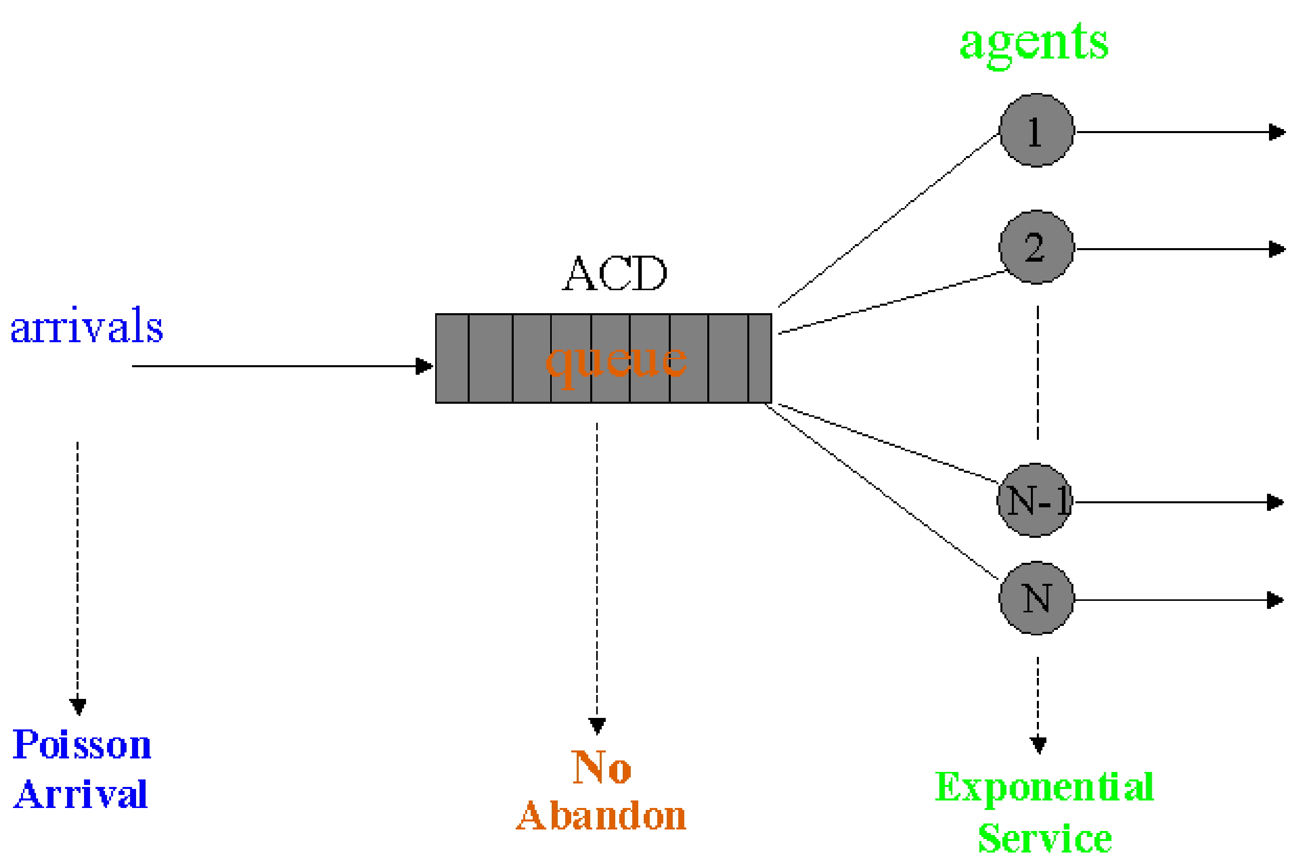
1. Call Centers
2. Bank Call Center Data
3. Queueing Theory
4. Analysis
   (a) Arrivals
   (b) Service Durations
   (c) Queueing Behavior
   (d) Forecasting of Workload
5. Validation of Queueing Model Outputs
6. Summary

The Vast Call Center World

- 70% of all customer-business interactions occur in call centers.
- 3% of the U.S. working population is currently employed in call centers.
  - ▷ 1.55 - 6 million agents
  - ▷ more than in agriculture
- 20% growth rate of the call center industry.
- State-of-art technology, but 70% for human cost.
- Factory floor of modern commerce.

Classical Queueing Theory

Erlang-C: M/M/N



Some Calculations based on Erlang-C Model

Assume Poisson arrival rate λ, service rate μ/server (mean $1/\mu$), λ/μ: *offered load*,

- M/M/1:
  - ▷ system stable iff $\rho = \lambda/\mu < 1$;
  - ▷ **Ave. # of customers in system**: $L = \frac{\rho}{1-\rho}$;
  - ▷ **Ave. # of customers in queue**: $Q = \frac{\rho^2}{1-\rho}$;
  - ▷ **Ave. # of customers in service**: $L_s = \rho = L - Q$;
  - ▷ **Ave. waiting time in system**: $w = L/\rho = 1/(\mu - \lambda)$;
  - ▷ **Ave. delay in queue**: $d = Q/\rho = w - 1/\mu$.

M/M/N and M/G/N Models

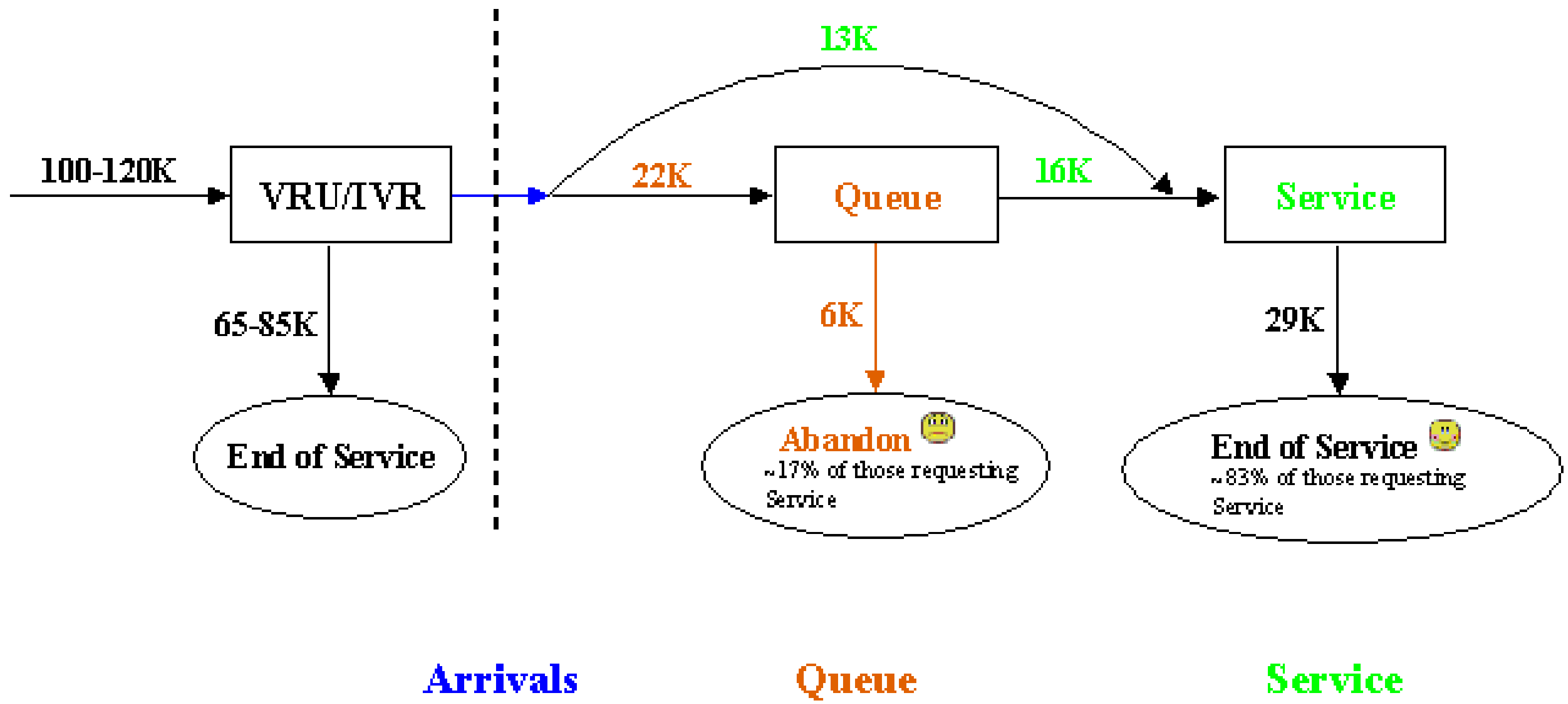
- M/M/N: exact formulae exist but are complicated.
- M/G/N: approximate formulae exist – Khintchine-Pollaczek Formula.
  - ▷ General service time distribution;
  - ▷ Exact for $M/G/1$;
  - ▷ Approximate linear relationship between w and $\frac{\rho}{1-\rho}$;
    - ∗ $\frac{N}{\mathrm{E}(G)} w \approx K_G \frac{\rho}{1-\rho}$

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

Arrivals: Inhomogeneous Poisson

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

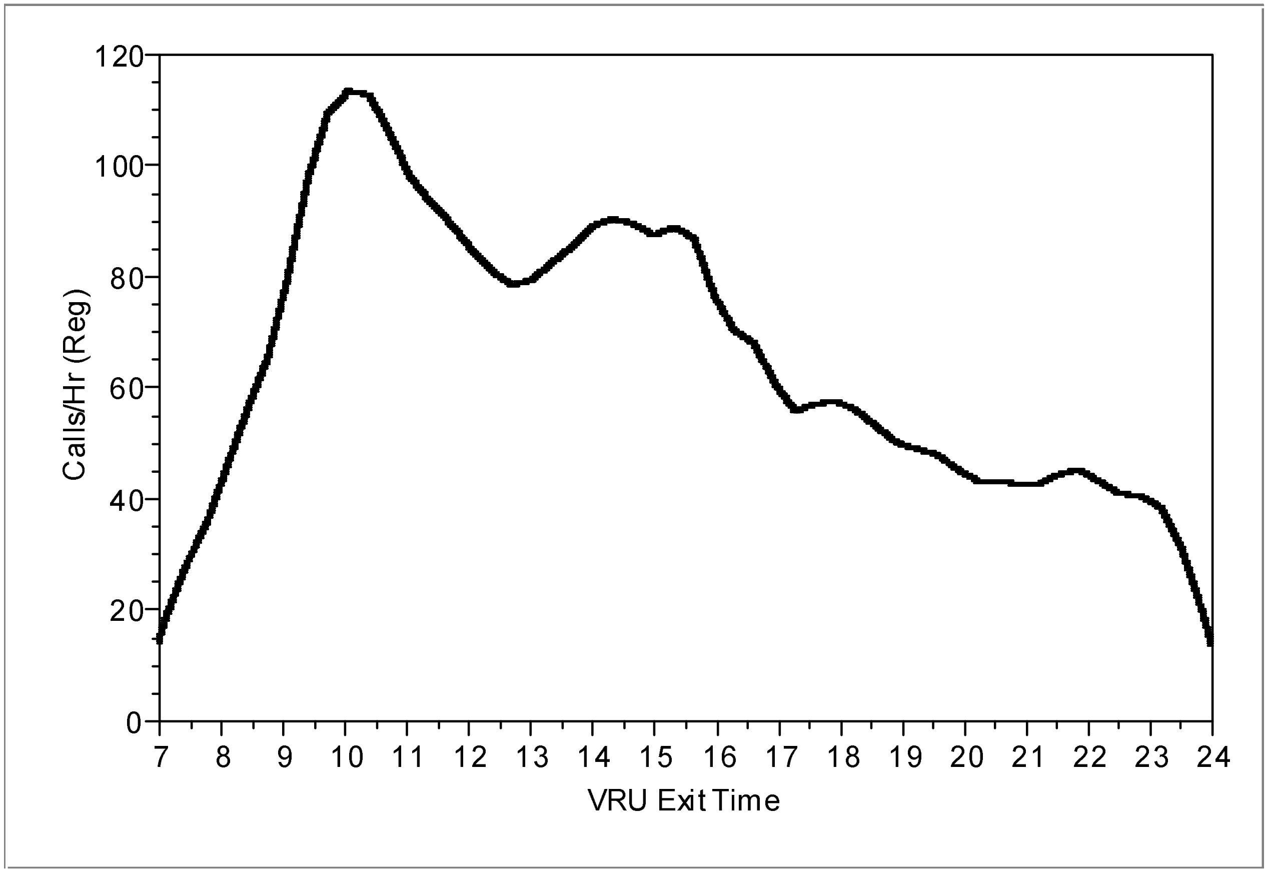
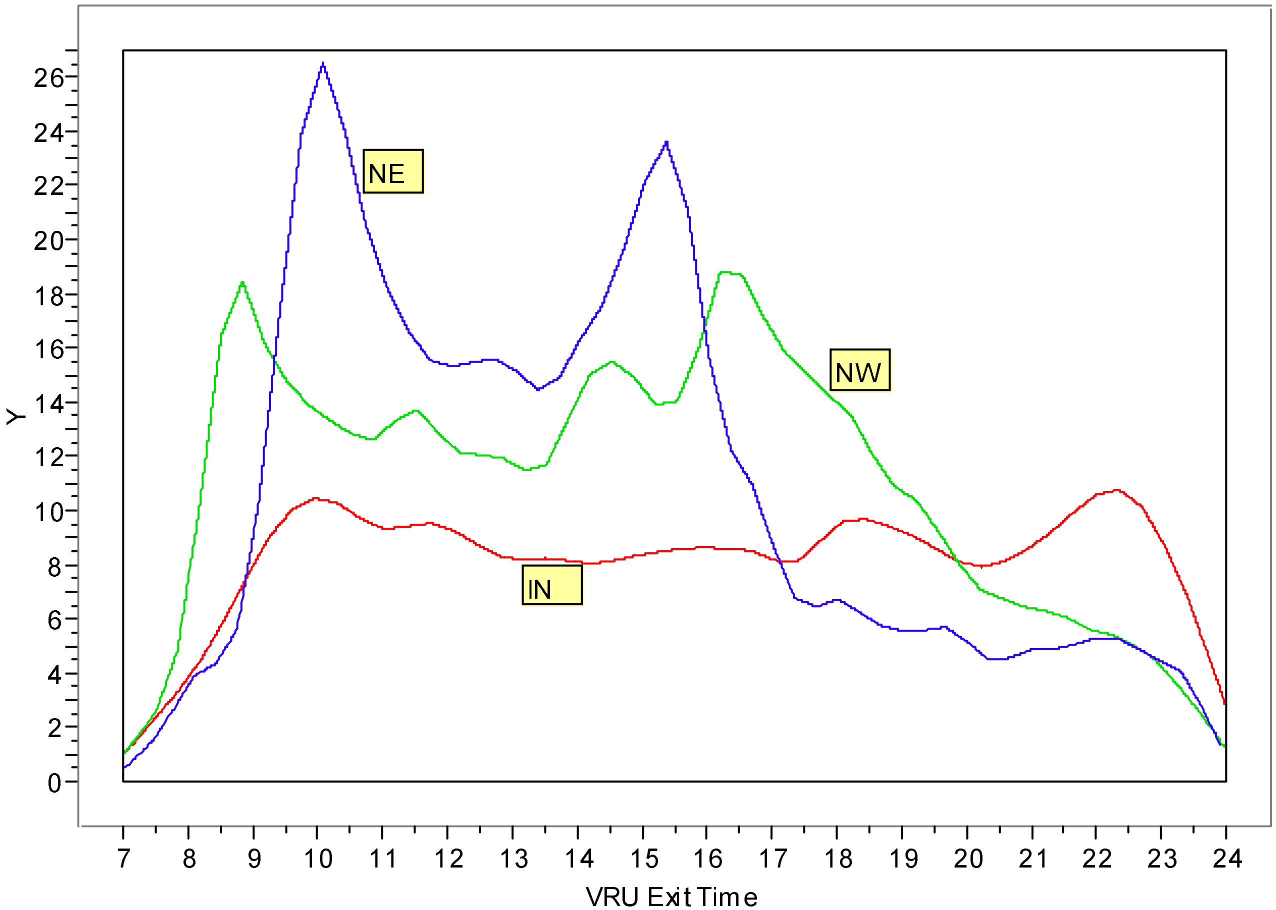


Figure 2: Arrivals (to queue or service) – IN, NW, and NE Calls

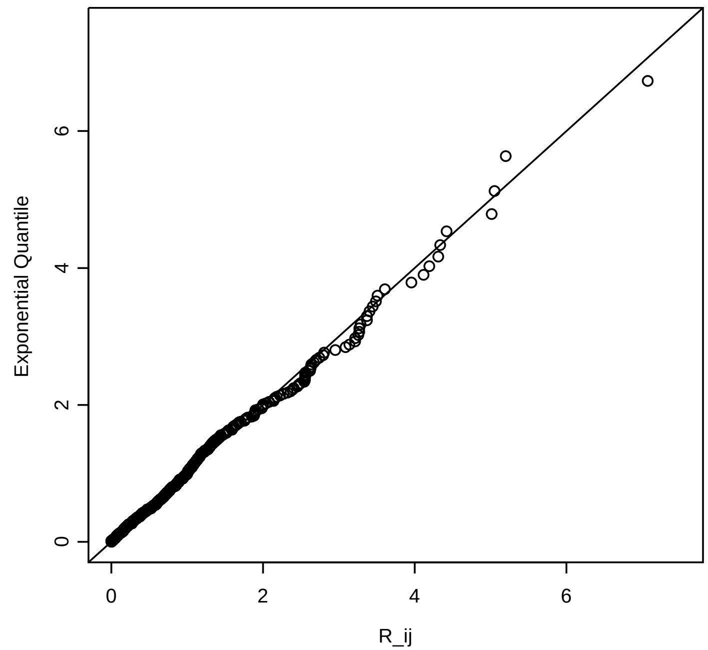


IN = INternet Consulting; NW = New Customer Service; NE = Stock Exchange.

A Test for Inhomogeneous Poisson Process

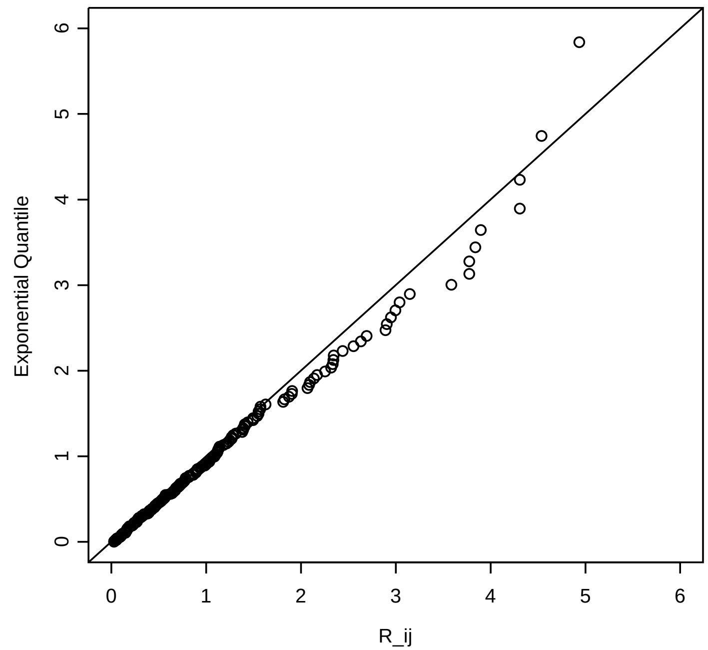
1. Break up the interval of a day into short blocks of time, say I (equal-length) blocks of length L.
2. Let $T_{i0} = 0$ and
   T_{ij}: the j-th ordered arrival time in the i-th block, $i = 1, \ldots, I$ and $j = 1, ..., J(i)$,
   then define
   $$R_{ij} = (J(i) + 1 - j)\left(-\log\left(\frac{L - T_{ij}}{L - T_{i,j-1}}\right)\right).$$
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. Use any customary test for the exponential distribution; for example, Kolmogorov-Smirnov test.

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



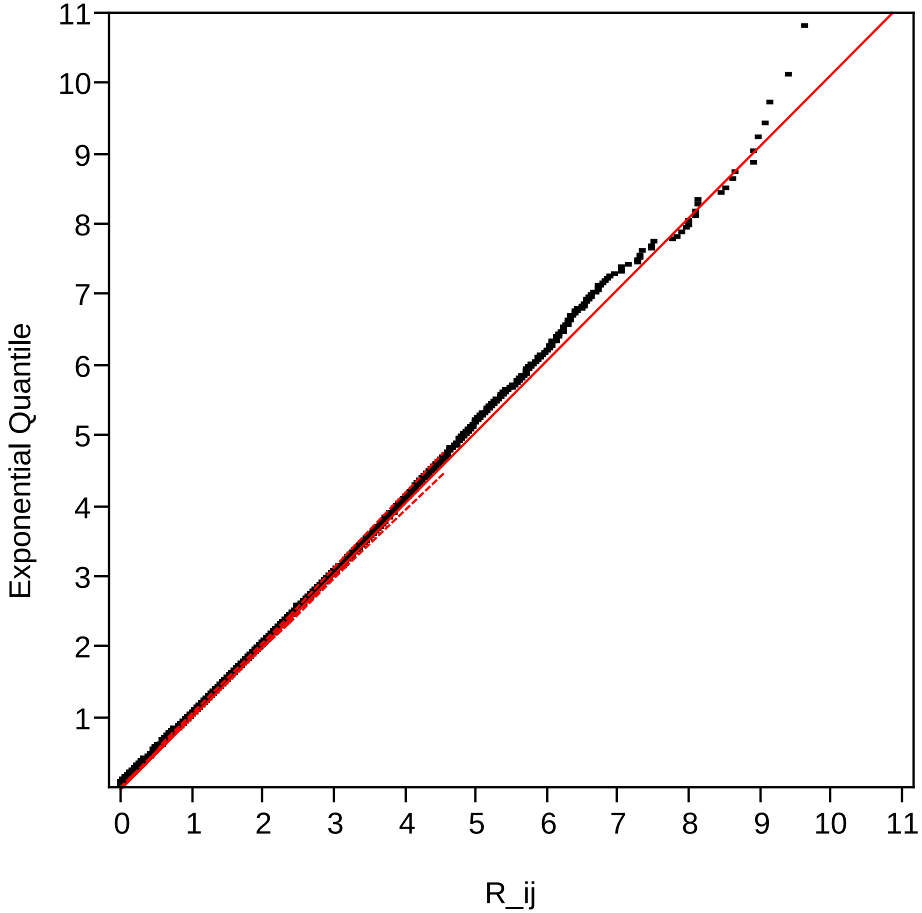
$L = 6$ min, $n = 420$, Kolmogorov-Smirnov statistic $K = 0.0316$ and the P-value is 0.2.

Figure 4: Exponential (λ=1) Quantile plot for $\{R_{ij}\}$ of INternet calls (Nov. 23)



$L = 60$ min, $n = 172$, Kolmogorov-Smirnov statistic $K = 0.0423$ and the P-value is 0.2.

Figure 5: Exponential Quantile plot of $\{R_{ij}\}$ for all Regular calls (Two outliers omitted.)



Service Times

- Erlang-C assumes exponential distribution.
  - ▷ simple for calculation
  - ▷ memoryless property
- How about this call center?

Figure 6: *Service Time Distribution from This Call Center*

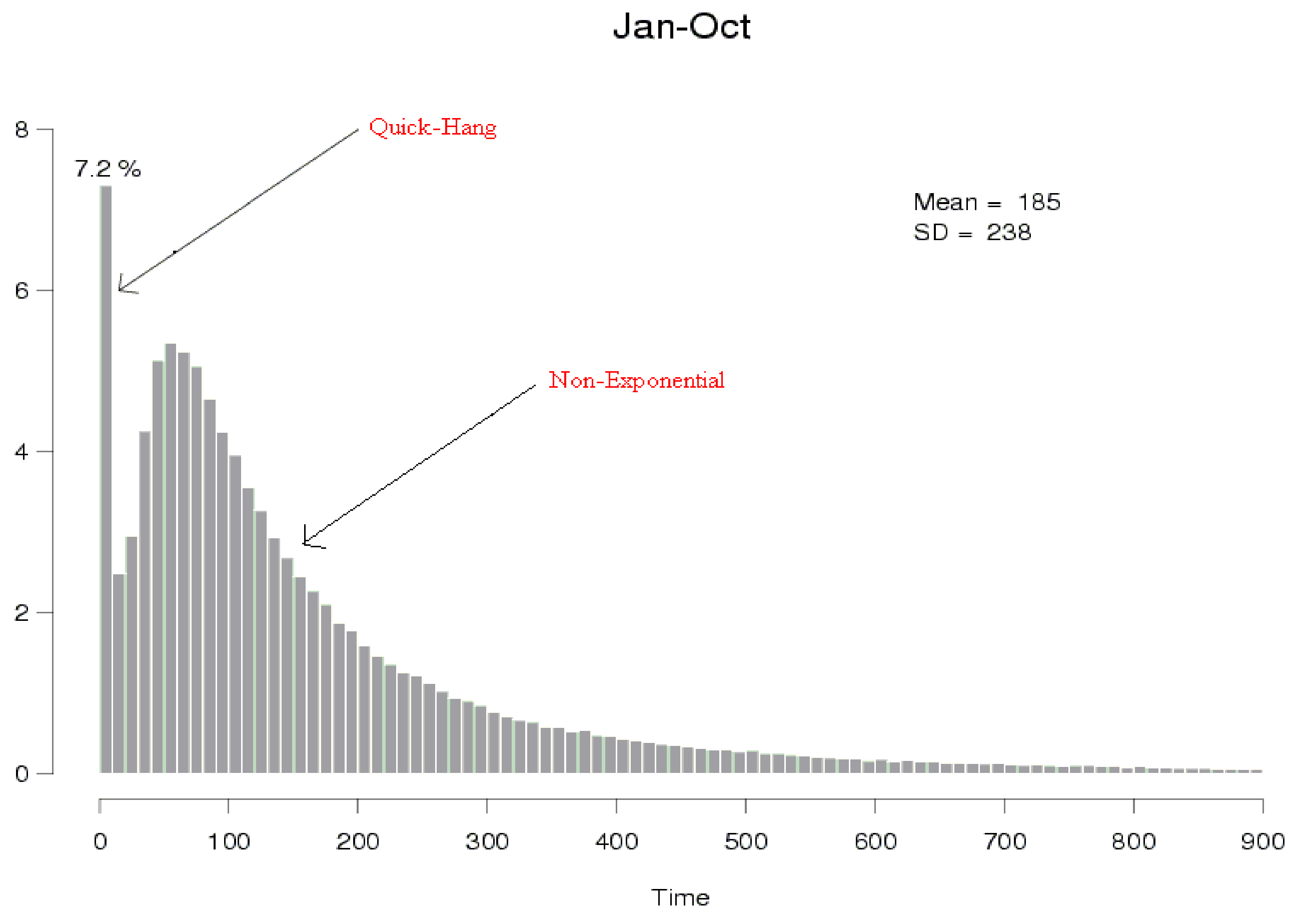
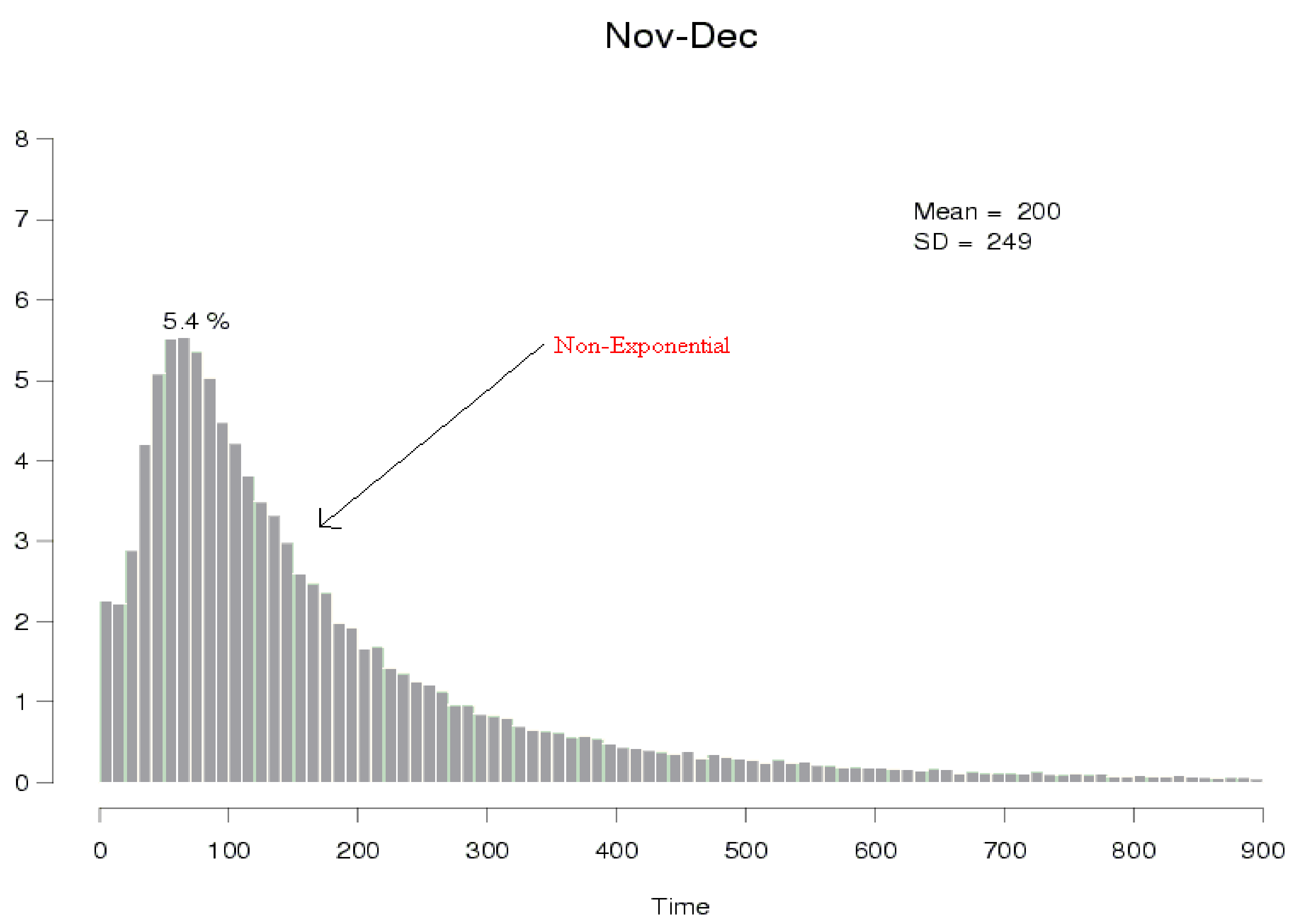


Figure 7: *Service Time Distribution from This Call Center*



Service Times are Lognormal

Figure 8: *Histogram of Log(Service Time) (Nov + Dec)*

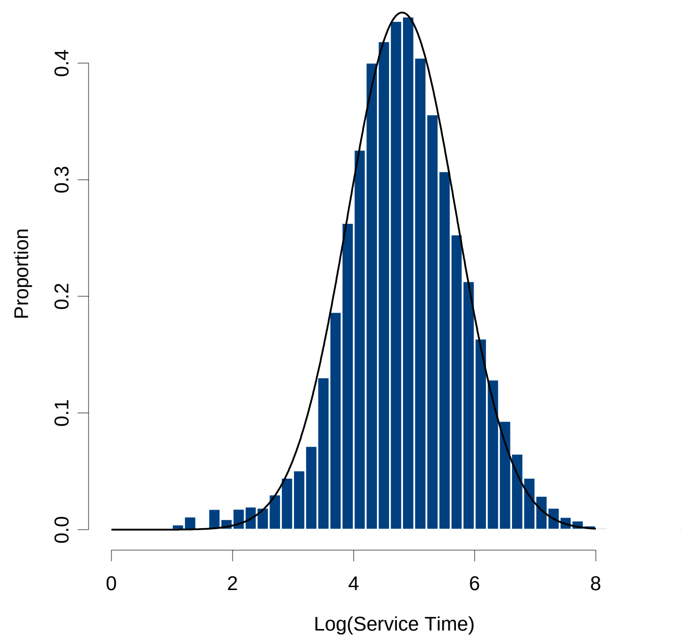
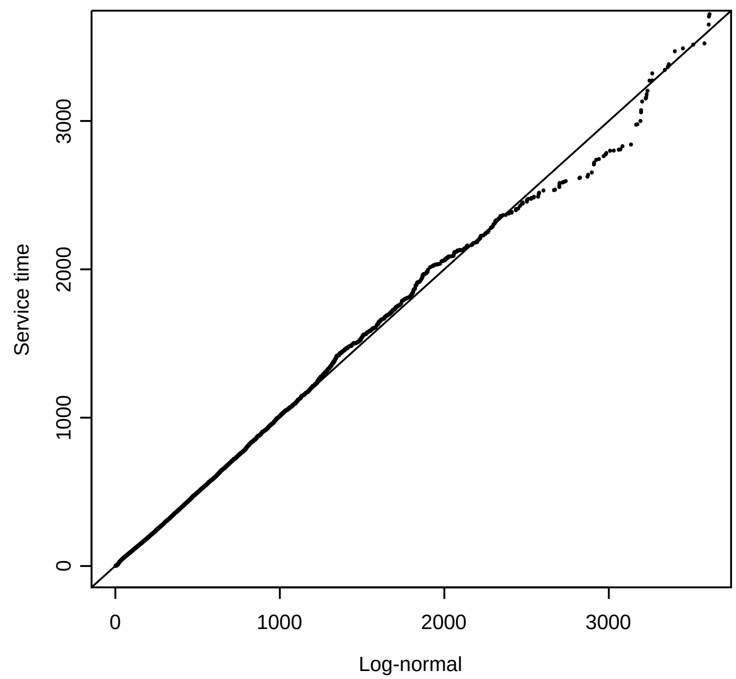


Figure 9: *Log-normal QQ Plot of Service Time (Nov + Dec)*



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 as a function of time-of-day - **Nonparametric Regression**
  - ▷ Mean service time across different categories, like service types, day-of-week - **Anova**

A Property of Lognormal Distributions

If Z is lognormally distributed with mean ν, and $Y = \log(Z) \sim \mathrm{N}(\mu, \sigma^2)$, then

$$\nu = e^{\mu+\sigma^2/2}.$$

Similarly, if $Y|X = x \sim \mathrm{N}(\mu(x), \sigma(x)^2)$, then

$$\nu(\mathbf{x}) = \mathbf{e}^{\mu(\mathbf{x})+\sigma(\mathbf{x})^{\mathbf{2}}/\mathbf{2}} \tag{1}$$

The Problem

The data:

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

where $Z|X = x$ has a lognormal distribution. For example, in a **regression** setup,

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

and

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

We are interested in estimating

$$\nu(x) = \mathbb{E}(Z|X = x)$$

with confidence band attached. *Shen (PhD Thesis 2003)*

The Idea

1. Transform the original problem into a problem with normal errors;
2. Make inferences based on the transformed data;
3. Back transform the inference results to the original scale.

The Procedure

1. $\{X_i, Z_i\} \;\Rightarrow\; \{X_i, Y_i\}$ with $Y_i = \log(Z_i)$;
2. The model is
$$Y_i = \mu(X_i) + \sigma(X_i)\epsilon_i$$
where $\epsilon_i | X_i \overset{i.i.d.}{\sim} \mathrm{N}(0,1)$.
3. Estimate $\mu(x)$ using any good existing nonparametric regression method, e.g. local polynomial method.
4. Estimate $\sigma(x)$ using some good local nonparametric regression method, like difference-based estimate plus local polynomial smoothing. We can get $\mathrm{se}_\mu(x)$ from $\hat{\sigma}(x)$.

The Procedure (cont')

5. We also need to estimate the variance of $\hat{\sigma}^2(x)$ in order to get $\mathrm{se}_{\sigma^2}(x)$.

6. Suppose $\hat{\mu}(x)$ is (approximately) unbiased, then

$$\hat{\mu}(\mathbf{x}) \pm \mathbf{Z_{1-\alpha/2}}\mathrm{se}_{\mu}(\mathbf{x})$$

   will be an approximate $100(1-\alpha)\%$ confidence interval for $\mu(x)$.
   Similarly a $100(1-\alpha)\%$ confidence interval for $\sigma^2(x)$ is approximately

$$\hat{\sigma}^{\mathbf{2}}(\mathbf{x}) \pm \mathbf{Z_{1-\alpha/2}}\mathrm{se}_{\sigma^{\mathbf{2}}}(\mathbf{x}).$$

   Note: *Calculation of* se_{σ^2} *requires care, depending on the method for estimating* $\hat{\sigma}^2$. *(See Levins PhD Thesis (2003).)*

The Procedure (cont')

7. Back-transform to the original scale and obtain the following plug-in estimate for $\nu(x)$,

$$e^{\hat{\mu}(x)+\hat{\sigma}^2(x)/2};$$

The corresponding $100(1-\alpha)\%$ confidence interval for $\nu(x)$ is

$$\mathbf{e}^{(\hat{\mu}(\mathbf{x})+\hat{\sigma}^{\mathbf{2}}(\mathbf{x})/\mathbf{2})\pm\mathbf{Z}_{\mathbf{1}-\alpha/\mathbf{2}}\sqrt{\mathrm{se}_{\mu}(\mathbf{x})^{\mathbf{2}}+\mathrm{se}_{\sigma^{\mathbf{2}}}(\mathbf{x})^{\mathbf{2}}/\mathbf{4}}}.$$

Note:

$\hat{\mu}(x)$ and $\hat{\sigma}^2(x)$ are asymptotically independent and very nearly independent at any sample size, which gives us

$$\mathrm{se}(\hat{\mu}(x)+\hat{\sigma}^2(x)/2) \approx \sqrt{\mathrm{se}_{\mu}(x)^2+\mathrm{se}_{\sigma^2}(x)^2/4}.$$

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

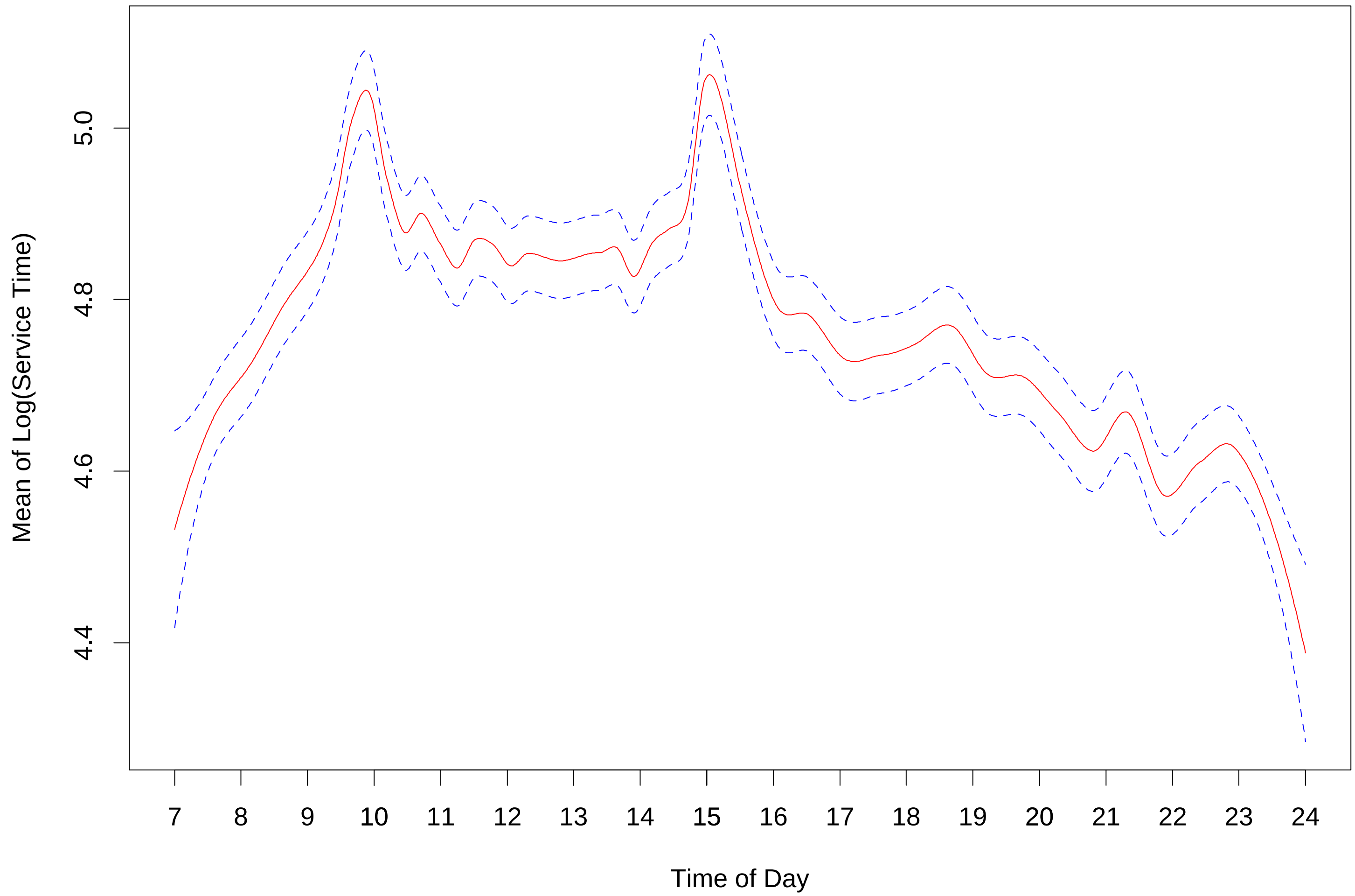


Figure 11: Variance of Log(Service Time) (Regular) vs. Time-of-day (95% CI) ($n = 42613$)

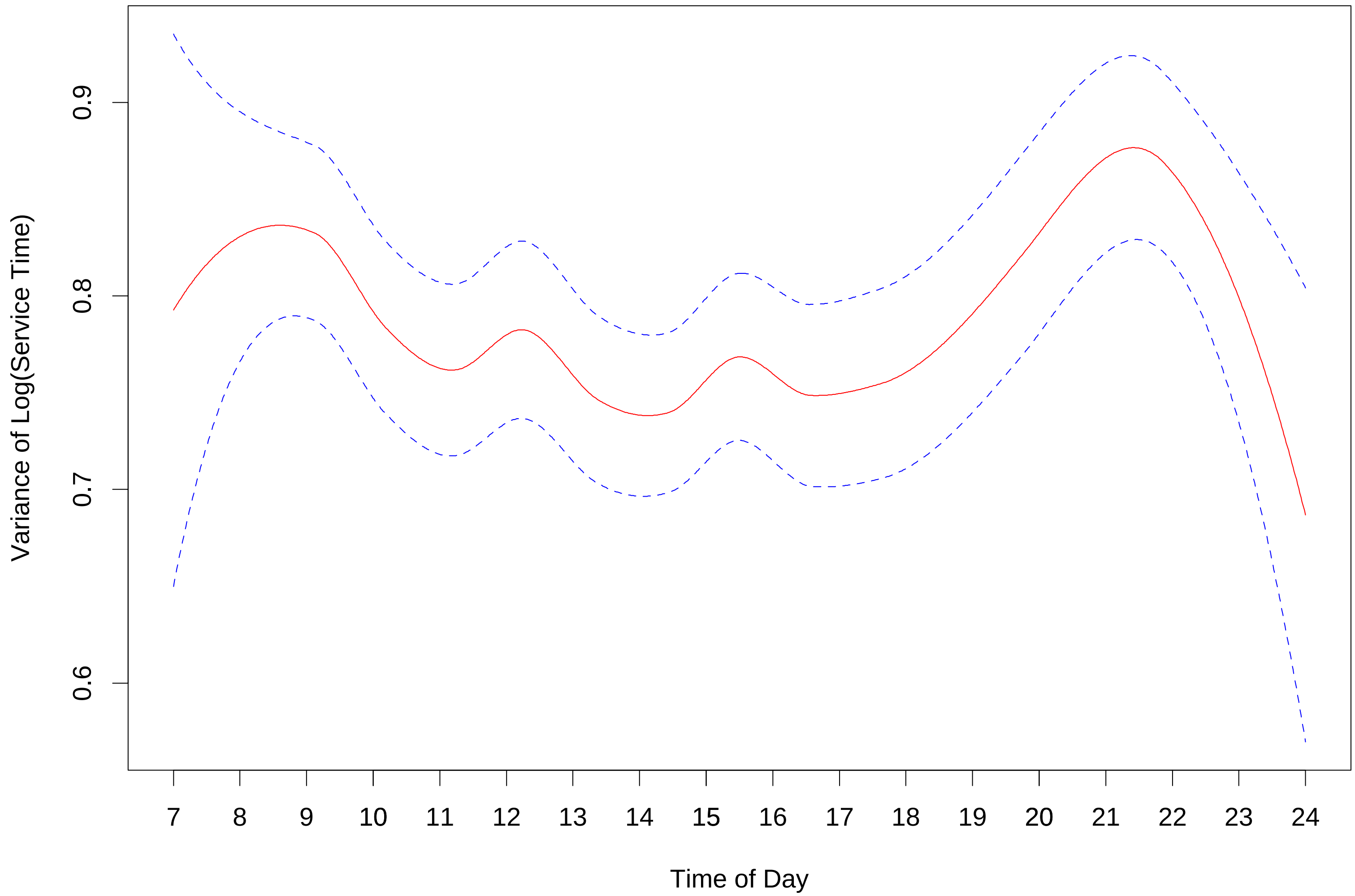


Figure 12: Mean Service Time (Regular) vs. Time-of-day (95% CI) ($n = 42613$)

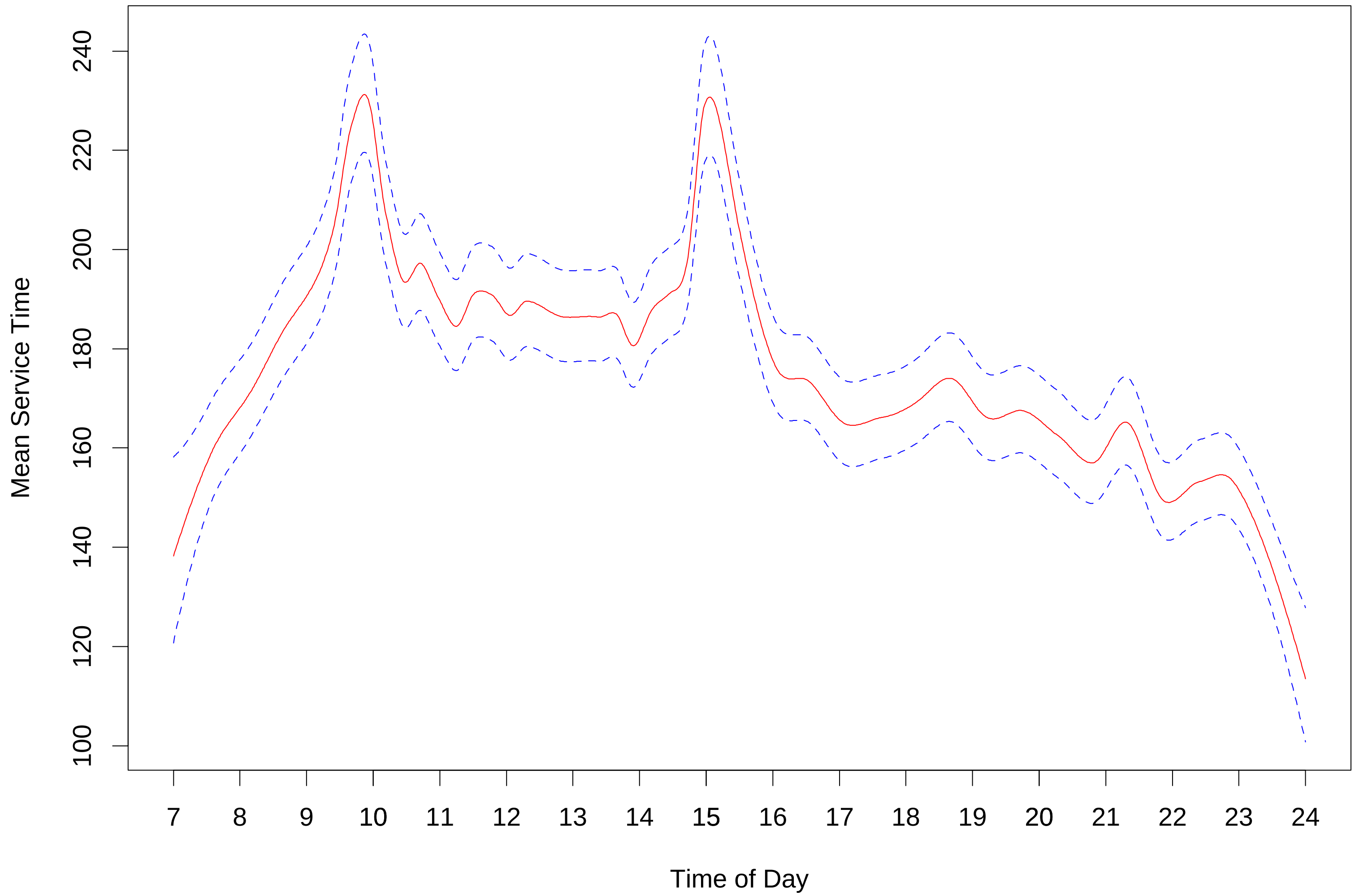
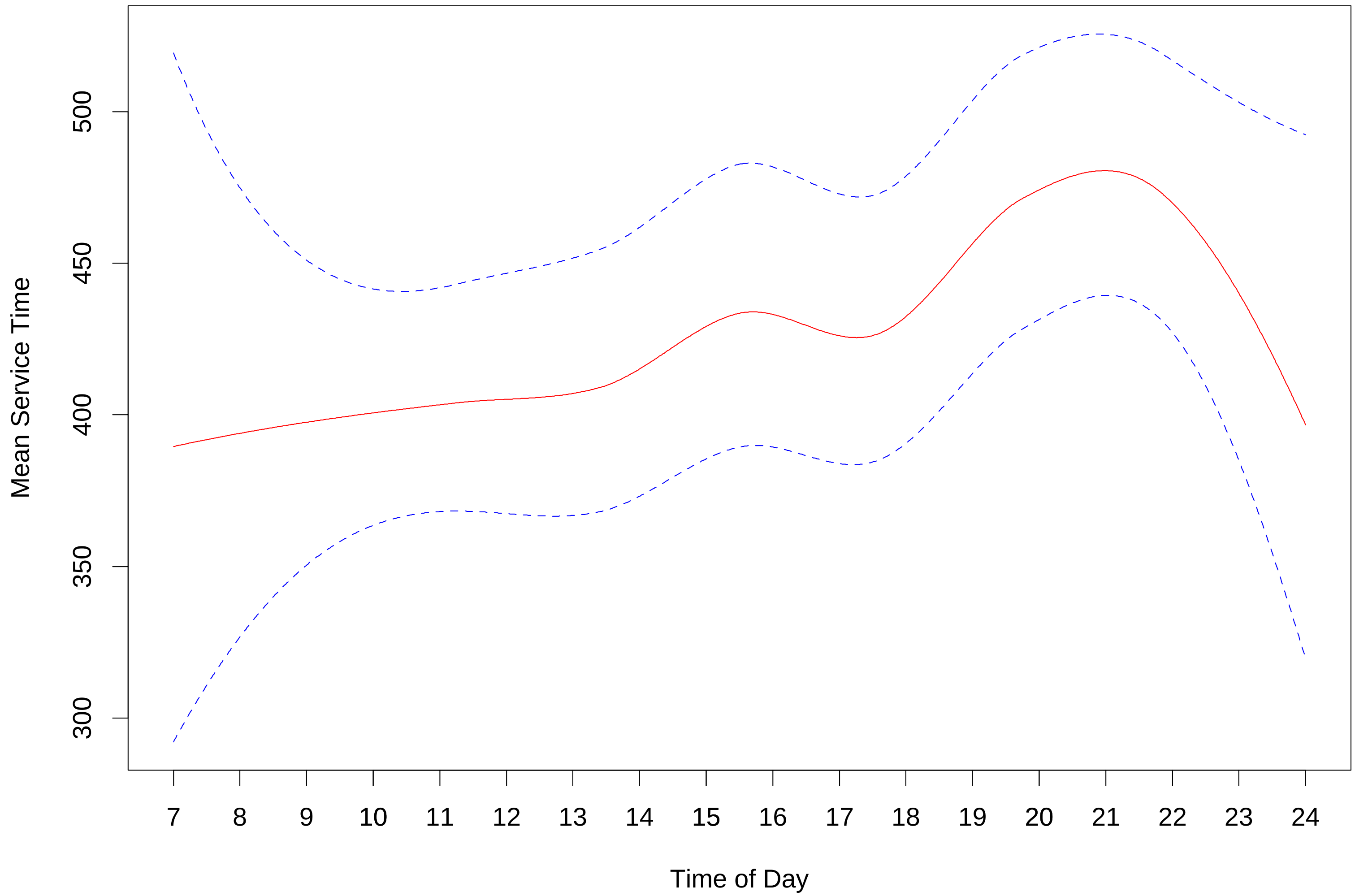


Figure 13: Mean Service Time (INternet) vs. Time-of-day (95% CI)($n = 5066$)

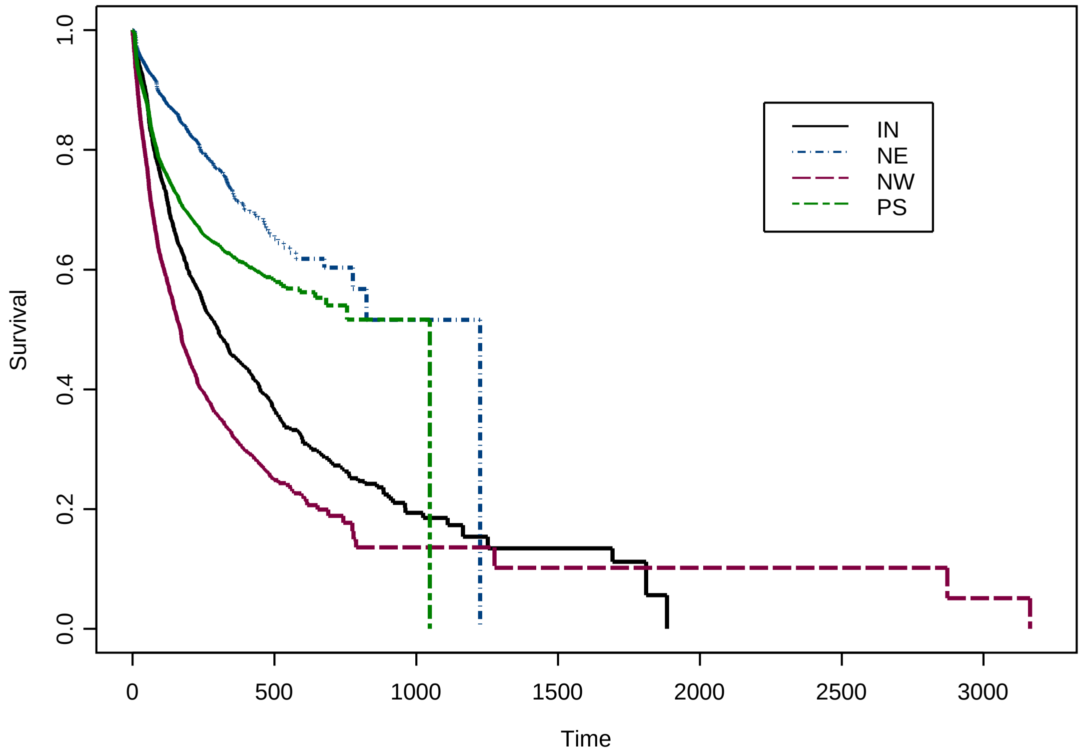


Customer Patience and Abandonment Behavior

- Censored data
- 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.
- Also observe the indicator $I_{R<V}$.
- Thus, V and R are censored.

Human Patience: Time willing to wait

Figure 14: Survival curves for time willing to wait (Nov.–Dec.)



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

Hazard Rates Estimation Procedure

- For a time t, choose an interval length δ_t, and estimate the hazard rate at time $t + \delta_t/2$ using

$$\frac{[\,\#\text{ of events during } (t, t+\delta_t]\,]}{[\,\#\text{ at risk at } t]\ \times\ \delta_t}\ .$$

- Usually $\delta_t = 1$ when $t \le t_0$, a pre-specified constant.
- δ_t should be larger when $t \ge t_0$. For example, δ_t can be chosen so that

$$\hat{N}_{(t,t+\delta_t]} \ge n_0$$

where $\hat{N}_{(t,t+\delta_t]}$ is an estimate of the expected # of events during $(t, t+\delta_t]$.

- *Optional* smoothing of the raw hazard rates estimates.

Figure 15: Hazard rate for the time willing to wait for Regular calls (Nov.–Dec.) ($n_0 = 4$)

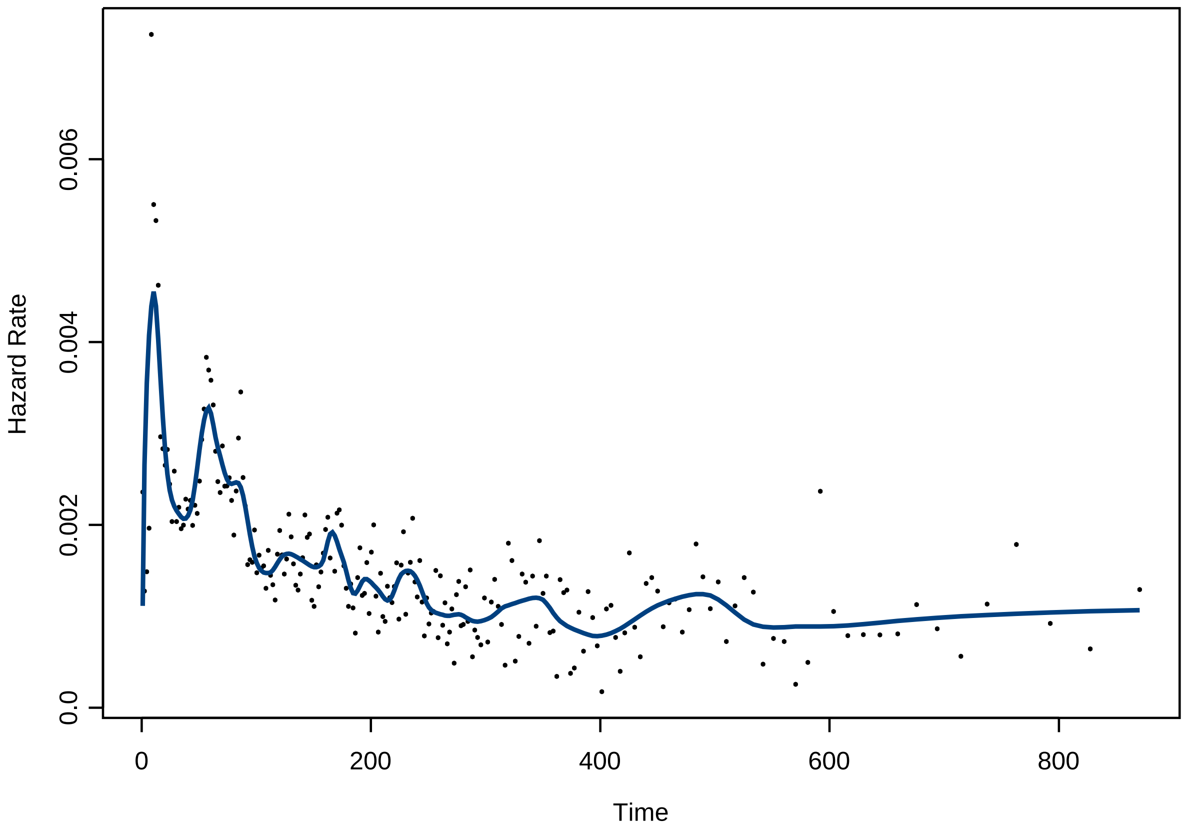
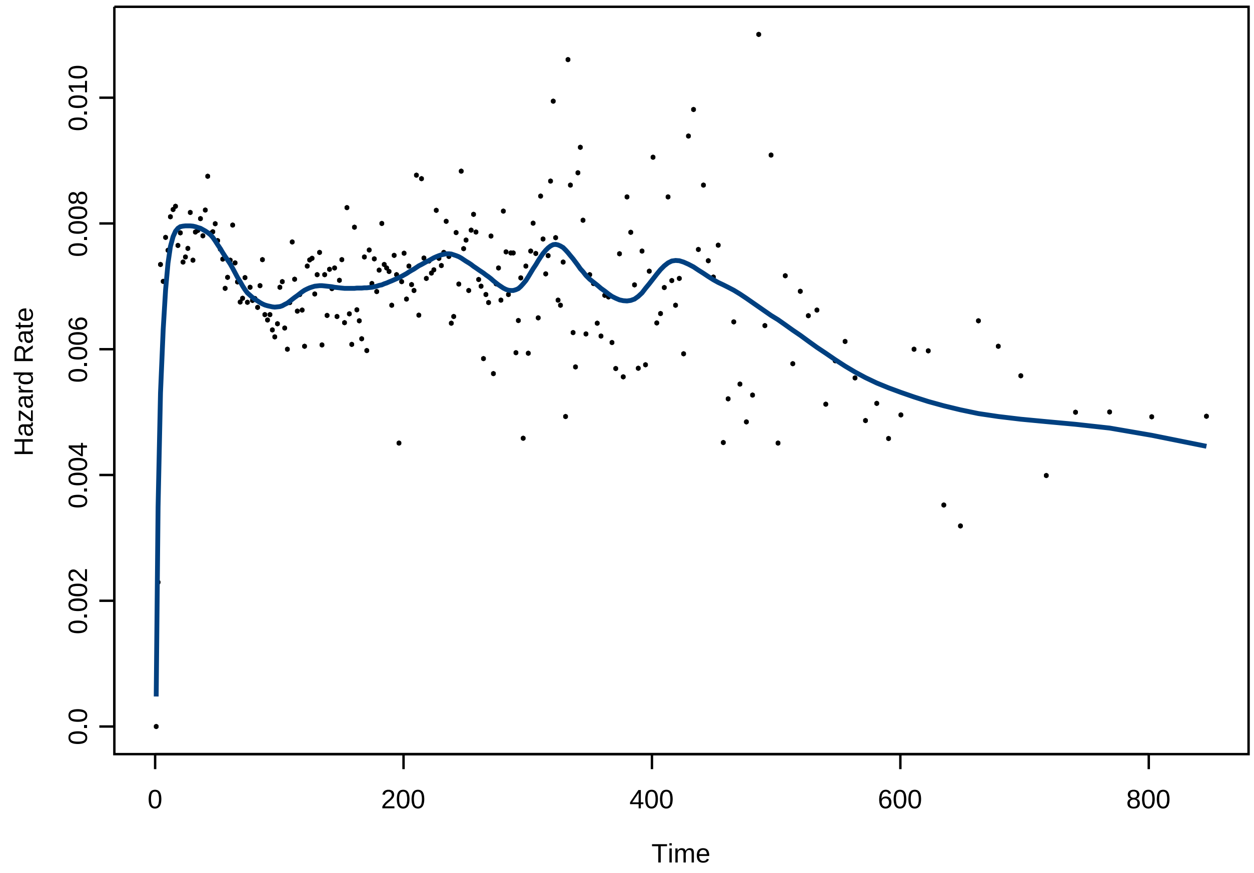


Figure 16: Hazard rate for virtual waiting time (Nov.–Dec.) $(n_0 = 4)$



Workload

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

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

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

Estimation 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).)
- Random-effect model.
  - ▷ Regular (non-holiday) weekdays from Aug. to Dec. indexed by j;
  - ▷ Divide the regular workhours from 7AM through midnight into 68 quarter hours indexed by k;

Estimation of $\Lambda(t)$

- N_{jk}: number of arrivals within the k-th quarter hour of the j-th day.
- $$N_{jk} = \text{Poiss}(\Lambda_{jk}), \;\; \Lambda_{jk} = R_j \tau_k + \varepsilon'_{jk}, \tag{2}$$

  where

  * τ_k: fixed deterministic quarter-hourly effects with $\sum \tau_k = 1$;
  * R_j: suitable random daily effects;
  * ε'_{jk}: random errors.

A Property of Poisson Variables

Suppose $X \sim \text{Poiss}(\lambda)$, then Brown, Zhang and Zhao (2001) showed that, asymptotically,

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

with good accuracy even for small λ.

An Equivalent Gaussian Model

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

$$\begin{aligned} V_{jk} &= \theta_{jk} + \varepsilon^*_{jk} \quad \text{with} \quad \varepsilon^*_{jk} \overset{iid}{\sim} N\left(0, \tfrac{1}{4}\right), \\ \theta_{jk} &= \alpha_j \beta_k + \varepsilon_{jk}, \\ \alpha_j &= \mu + \gamma V_{j-1,+} + \varepsilon^{**}_j, \end{aligned} \tag{3}$$

  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.
- $\sum \beta^2_k = 1$.

Estimation

Here μ, γ, σ^{**2}, σ_ε^2 and β_k can be estimated by a combination of least-squares and method of moments.

Denote the corresponding estimates as $\hat{\mu}$, $\hat{\gamma}$, $\hat{\sigma}^{**2}$, $\hat{\sigma}_\varepsilon^2$ and $\hat{\beta}_k$.

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

  $$\theta_k = \beta_k \left(\mu + \gamma V_+ + \varepsilon^{**}\right) + \varepsilon \tag{4}$$

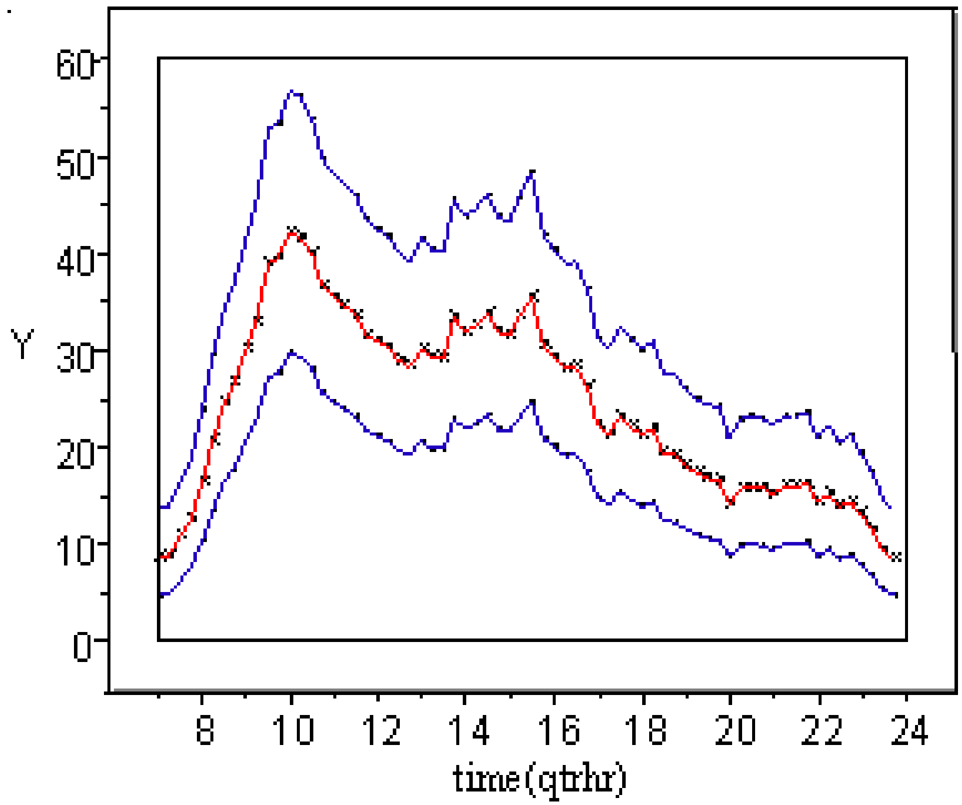
  where $\varepsilon^{**} \sim N(0,\, \sigma^{**2})$ and $\varepsilon \sim N(0,\, \sigma_\varepsilon^2)$ are independent.

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

- $\mathrm{Var}(\hat{\theta}_k)$ can be derived from (4), 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 17: 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:

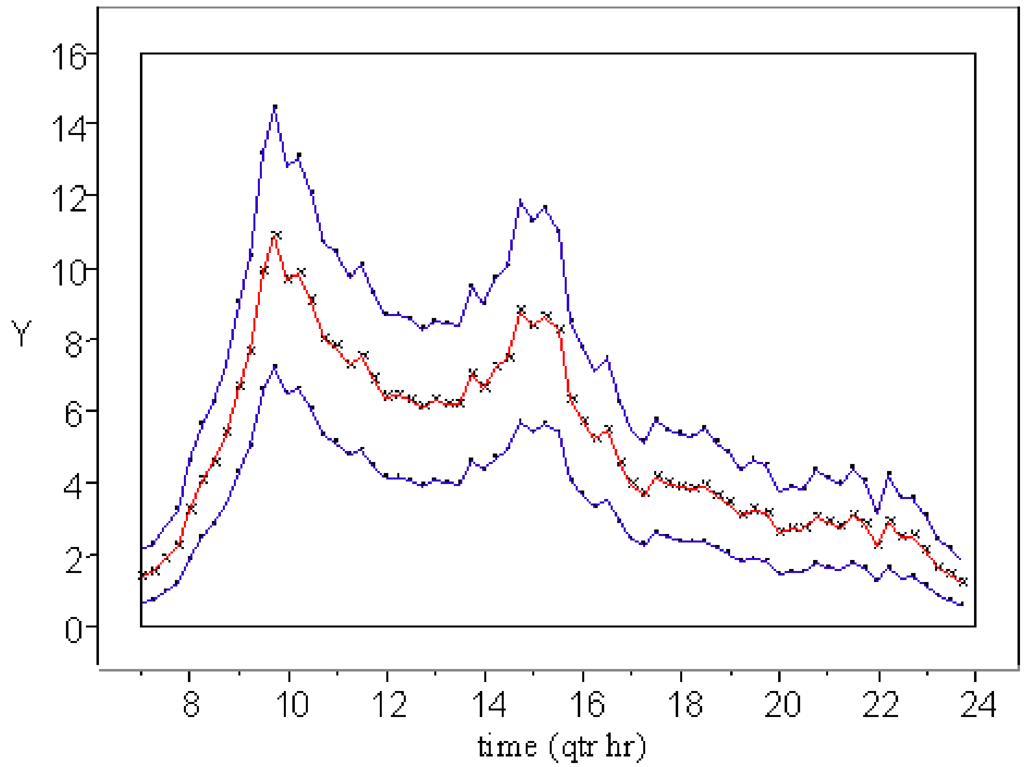
$$\hat{L}(t) \pm 2\hat{L}(t)\widehat{PCV}(\hat{L})(t)$$

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

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

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

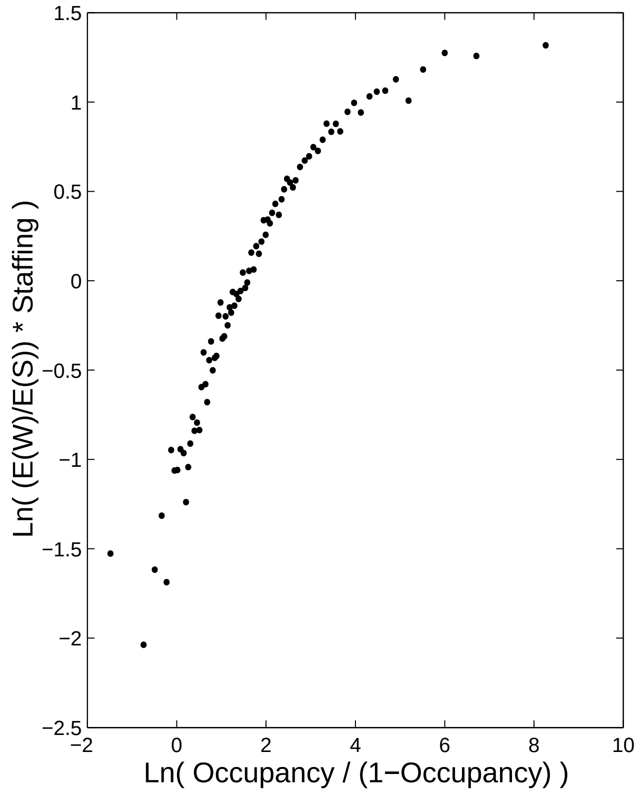
Figure 18: 95% prediction intervals for the load, L, following a day with $V_+ = 340$.



Units on vertical axis are "required agents".

Validation of Queueing Model: Failure of Erlang-C

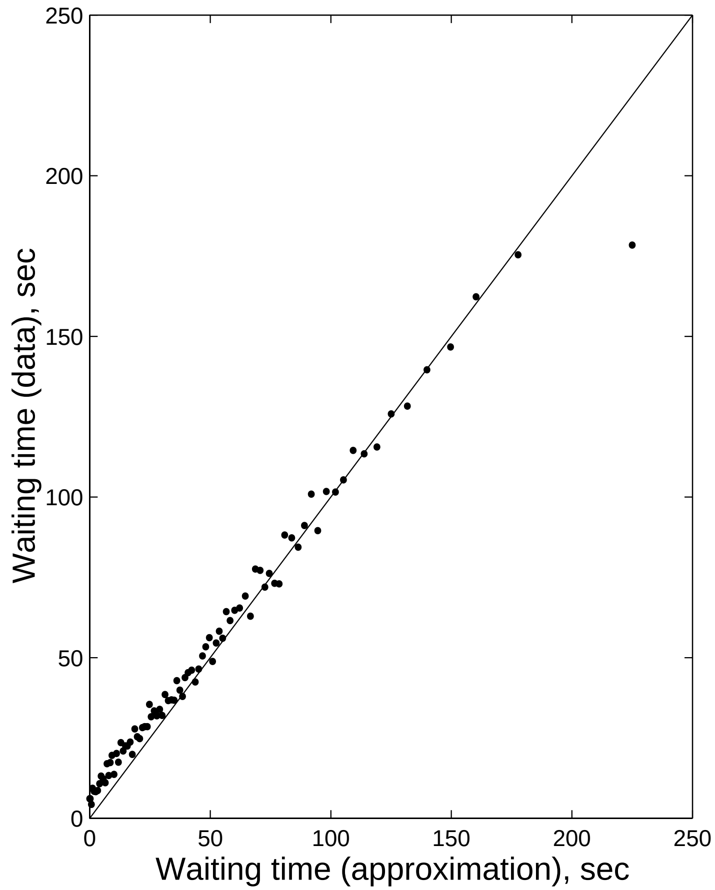
Figure 19: Agents' Occupancy versus Ave. Waiting Time



According to Erlang-C, the plot should be approx. linear with slope=1.

Validation of Queueing Model: Success of Erlang-A

Figure 20: Waiting Time: Data Ave. versus Erlang-A Prediction



Summary

- Arrivals
  - ▷ Testing inhomogeneous Poisson process
  - ▷ Test for applicability of fixed effects model
  - ▷ Forecasting Poisson arrival rate
    - ∗ Sqrt-Gaussian Model
    - ∗ AR structure
- Service Times
  - ▷ Lognormal
  - ▷ Nonparametric regression with lognormal errors
- Abandonment Behavior
  - ▷ Graphical technique for nonparametric hazard rate estimation
  - ▷ Estimation under high-censoring
- Workload
  - ▷ Forecasting with prediction-confidence bands
- Validation of Queueing Models: Erlang-C (No) and Erlang-A(Yes).

Reference

- Brown, L. D., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S. and Zhao, L. H. (2002), "Statistical Analysis of a Telephone Call Center: A Queueing-Science Perspective", submitted to Journal of American Statistics Association.
- Brown, L., Zhang, R., and Zhao, L. (2001), "Root un-root methodology for nonparametric density estimation", *Technical Report*, University of Pennsylvania.
- Brown, L., and Zhao, L. (2002), "A new test for the Poisson distribution", *To appear in Sankhya.*
- Levins, M. (2003), "On The New Local Variance Estimator", *Working PhD Thesis*, University of Pennsylvania.
- Shen, H. (2003), "Estimation, Confidence Intervals and Nonparametric Regression for Problems Involving Lognormal Distribution", *Working Phd Thesis*, University of Pennsylvania.

Model Diagnostics for Service Time Analysis

- Look at the residuals from the regression of Log(Service Time) on Time-of-day for the PS calls.
- Figures 21 and 22 give the histogram and normal quantile plot of the residuals, from which we can see that the residuals are pretty normal.
- Consequently provides additional validation of our assumption of the log-normality of the service times.

Figure 21: Histogram of The Residuals from Modeling Mean Log(Service Time) on Time-of-day (PS)

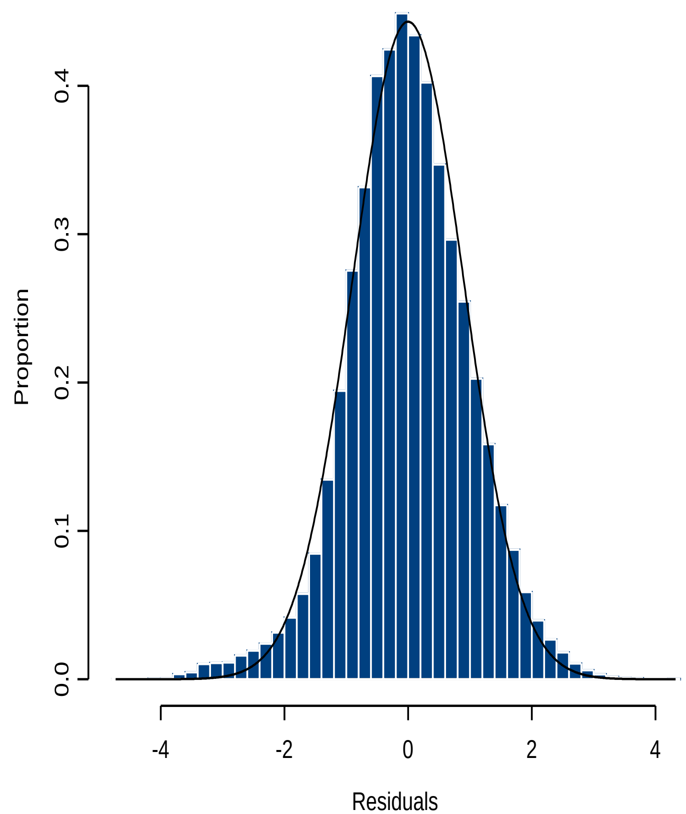
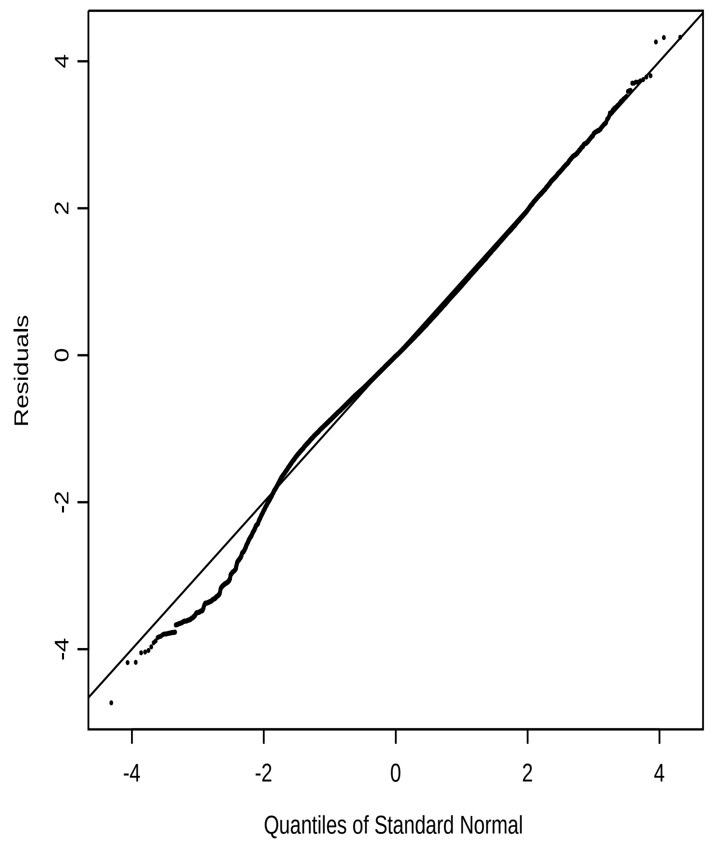


Figure 22: QQ-plot of The Residuals from Modeling Mean Log(Service Time) on Time-of-day (PS)



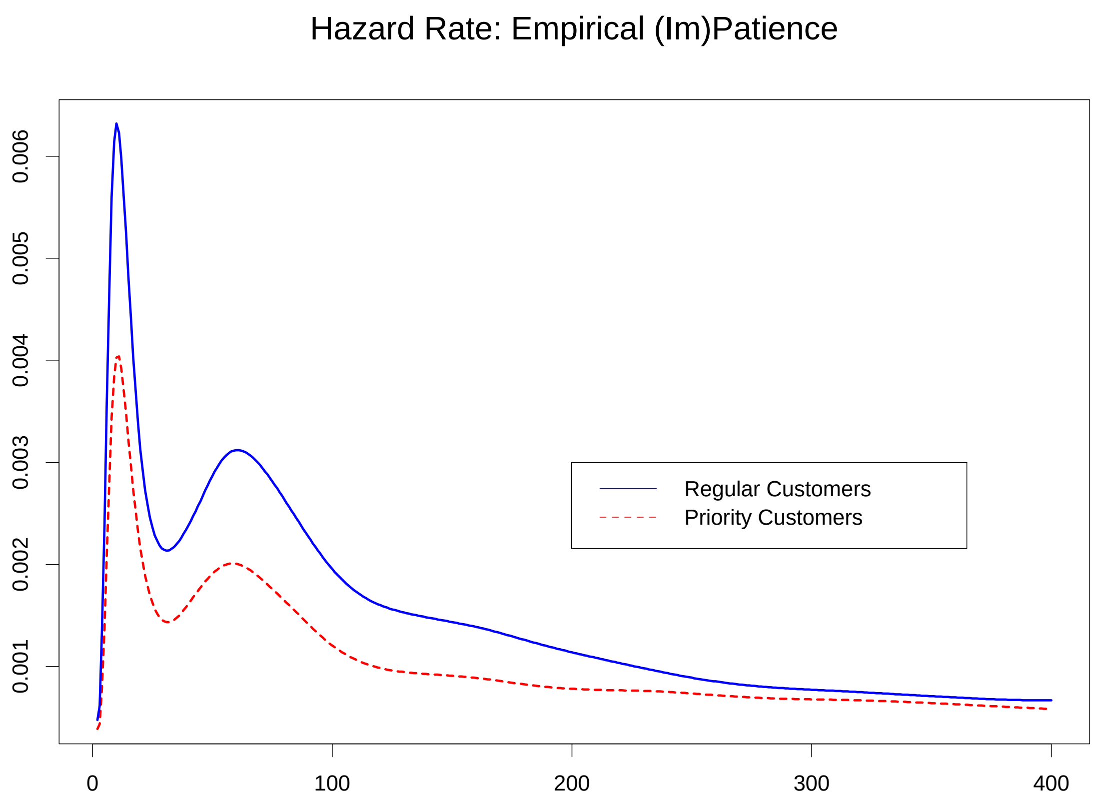
Human Patience

- Clear stochastic order among service types.
- Censoring rate is quite high. Most of the customers are served.
- Anomalistic behavior of K-M estimator, especially for PS and NE calls.

Table 1: Means, SDs and Medians for R (Nov.–Dec.)

	Mean	SD	Median
All Combined	803	905	457
PS	**642**	446	**1048**
NE	**806**	471	**1225**
NW	535	885	169
IN	550	591	302

Figure 23: Comparison Between Different Priority Customers



Patience Index

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

$$\text{Patience Index} \triangleq \frac{m_R}{m_V}.$$

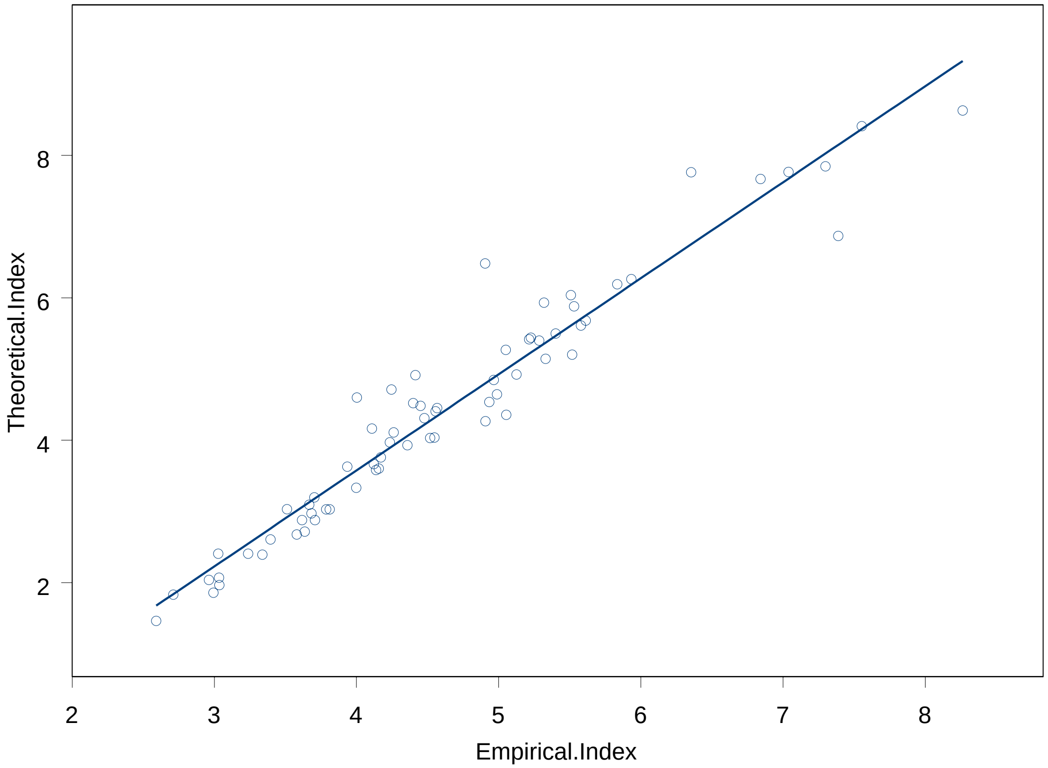
- Call-by-call data
- Survival analysis. High-censoring might be a problem.
- Ancillary measure:

$$\text{Empirical Index} \triangleq \frac{\#\ \text{served}}{\#\ \text{abandoned}}.$$

▷ The usual plug-in MLE for Patience Index if V and R are independent exponential.

▷ Works well empirically .

Figure 24: Patience Indices: empirical vs. theoretical ($R^2 = 0.94$)



Estimation Procedure for the Gaussian Model

μ, γ, σ_α^2, σ_ε^2 and β_k need to be estimated.

- Treat the $\{\alpha_j\}$'s as if they were fixed effects and using least-squares to fit the model

$$V_{jk} = \alpha_j \beta_k + \left(\varepsilon_{jk} + \varepsilon^*_{jk}\right).$$

  This yields estimates $\hat{\alpha}_j$, $\hat{\beta}_k$ and $\hat{\sigma}^2$.

- Estimate σ_ε^2 by method-of-moments as

$$\hat{\sigma}_\varepsilon^2 = \hat{\sigma}^2 - \frac{1}{4}.$$

- Use the "observations" $\hat{\alpha}_j$ to construct the least squares estimates of μ, γ and σ_α^2 by fitting the following model

$$\hat{\alpha}_j = \mu + \gamma V_{j-1,+} + \varepsilon_{\alpha j},$$

  with $R^2 \approx 0.5$.