The Offered-Load Process: Modeling, Inference and Applications

Michael Reich

Technion - Israel Institute of Technology

November 10, 2011

Advisors: Prof. Avishai Mandelbaum, Prof. Ya'acov Ritov

Outline

Motivation

Standard assumption in modeling service systems: The service-time and the patience of customers are independent

Is this assumption valid ?

Motivation

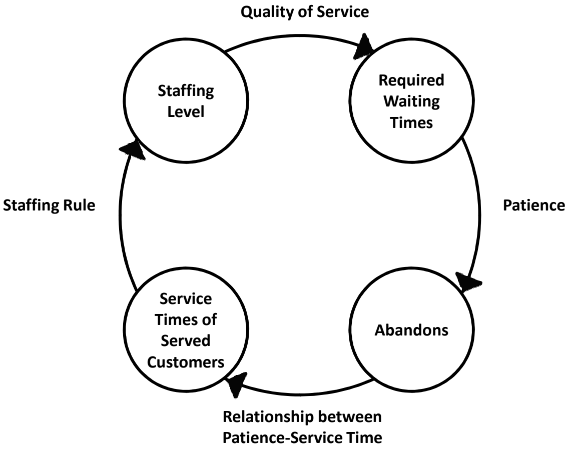
Standard assumption in modeling service systems: The service-time and the patience of customers are independent

Is this assumption valid ?

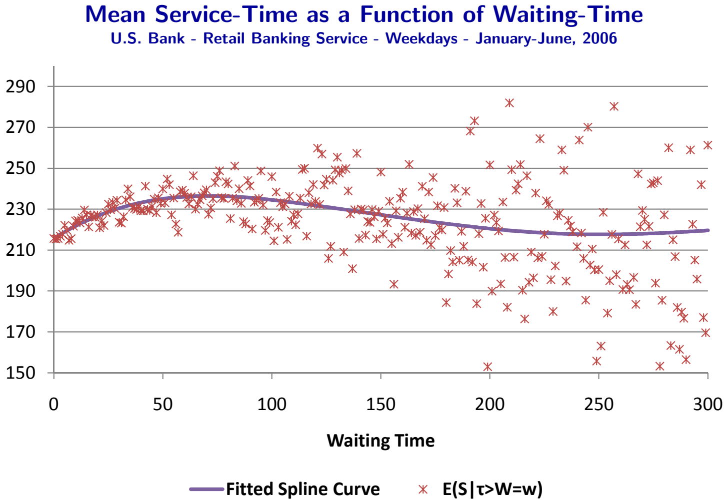
- I can pay my bills on another occasion
- I must immediately consult my broker in order to protect my equities profile

Motivation

Various performance measures are influenced by the presence of dependency between patience and service-time



Motivation


Mean Service-Time as a Function of Waiting-Time
U.S. Bank - Retail Banking Service - Weekdays - January-June, 2006
290
270
250
230
210
190
170
150
0
50
100
150
200
250
300
Waiting Time
Fitted Spline Curve
E(S|τ>W=w)

Research Outline

- A model for the relationship between patience and service-time
- Definitions of the offered-load process and function
- Estimation and prediction of the offered-load
- Performance analysis of both real and simulated data

Definitions

The $M_t/GI^*/n_t + GI$ Queue:

- M_t - Nonhomogeneous (time-dependent) Poisson arrivals
- GI^* - General service-time distribution; May depend on patience
- n_t - The number of telephone agents over time
- GI - General patience distribution

The parameters are assumed to be uninfluenced by system's state

Definitions

Notations:

- S - Service-time
- τ - (Im)patience
- V - Virtual waiting-time, namely the time a customer is **required** to wait before entering service
- W - Waiting-time of a customer, calculated over both served and abandoning customers

Note: $W = min\{V, \tau\}$

Definitions

Assumptions:

- The pair (τ, S) is independent of V
- For those who abandon, we observe V **censored** by their (im)patience
- For those who are served, we observe V
- W is observable for all customers
- S is observable only for customers who have $\tau > V$

Definitions

Biased Sampling:

- In common applications, service-time is measured over all served customers
- The prevalent estimator for mean service-time is actually $E(S|\tau > W)$
- **One tends to observes more customers with longer patience**

Definitions

Biased Sampling:

- In common applications, service-time is measured over all served customers
- The prevalent estimator for mean service-time is actually $E(S|\tau > W)$
- **One tends to observes more customers with longer patience**

What if the patience and service-times (associated with the same customer) are positively correlated ?

The Model

Consider the mean service-time of customers who were served after waiting exactly w time units:

$$g(w) = E(S|\tau > W, W = w) = E(S|\tau > w, V = w) = E(S|\tau > w)$$

Calculations show that

$$g(w) = \frac{\int_{u=w}^{\infty} f_\tau(u) \cdot E(S|\tau = u)\, du}{\int_{u=w}^{\infty} f_\tau(u)\, du}$$

or alternatively,

$$E(S|\tau = w) = g(w) - \frac{g'(w)}{h_\tau(w)} \, ,$$

where $f_\tau(w)$ and $h_\tau(w)$ are the cdf and the hazard-rate function of the patience, τ

Testing the Relationship

We present a statistical test for the relationship between patience and service-time:

- $H_0 : E(S|\tau = w) = E(S), \forall w \geq 0$
- H_1 : Otherwise

Since $E(S|\tau = w)$ is not observable, the test is based on $g(w) = E(S|\tau > W, W = w)$

Under the null hypothesis:

- $g(w) = E(S|\tau = w)$
- $E(S|\tau = w)$ is constant if and only if $g(w)$ is constant

Testing the Relationship

Let $g(W)$ be a random variable which takes the value $g(w)$ according to the density function $f_{W|\tau>W}(w)$

We test if the variance of $g(W)$ can be assumed to be zero:

$$Var(g(W)) = E(g^2(W)) - E^2(g(W))$$

Consequently, we choose the statistic for our test to be

$$T = \int_{u=0}^{\infty} g^2(u) \cdot f_{W|\tau>W}(u)\,du - \left[\int_{u=0}^{\infty} g(u) \cdot f_{W|\tau>W}(u)\,du\right]^2$$

Testing the Relationship

Construct a permutation distribution for the test statistic:

1. **Data**
   - Take all the observations of served customers with strictly positive waiting-time
2. **Discretization**
   - Divide the observations into several groups of a similar size, according to the ranking of their waiting-times
   - Calculate the probability of an observation to belong to each group
3. **Permutations**
   - Generate a large number (say 4,000) of permutations by randomly pairing between the waiting-times and the service-times
   - For each permutation, calculate the statistic's value

Testing the Relationship

Denote:

- $t_1, t_2, \ldots, t_K$ - The values of the test statistic in any of the random pairing permutations
- t^* - The original sample's statistic

The p-value is then approximated by the proportion of samples with the value of this statistic larger than the original sample's statistic. Explicitly:

$$p - value \approx \frac{\sum_{i=1}^{K} \mathbb{1}_{\{t_i > t^*\}}}{K}$$

Definitions

Introduce the following definitions:

1. **Resource-k Offered-Load Process** - A **stochastic process**, representing the amount of work being processed by resource k at time t, under the assumptions of *infinitely* many resources of type k, and that a task that reaches resource k enters service immediately upon arrival

2. **Resource-k Offered-Load Function** - A **function** of time $t \geq 0$, representing the average of Resource-k Offered-Load Process at time t

The Offered-Load of $M_t/GI/N_t + GI$

Theorem:

For any time $t > 0$, $L(t)$ has a Poisson distribution with mean

$$R(t) = E[L(t)] = E[\lambda(t - S_e)] \cdot E[S] = E\left[\int_{t-S}^{t} \lambda(u)du\right] =$$

$$= \int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du,$$

where

S is a generic service-time

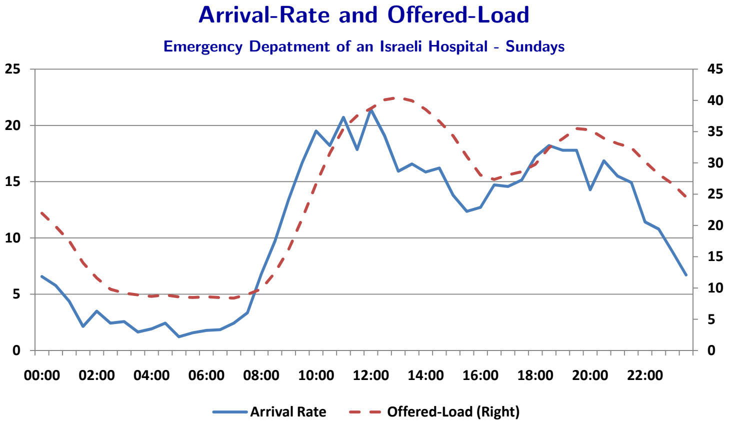
S_e is a generic excess service-time

$R(t)$ is by definition **the Offered-Load** function

The Offered-Load of $M_t/GI/N_t + GI$

Insights:

The offered-load function **lags** after the arrival rate -
$R(t) = E[L(t)] = E[\lambda(t - S_e)] \cdot E[S]$

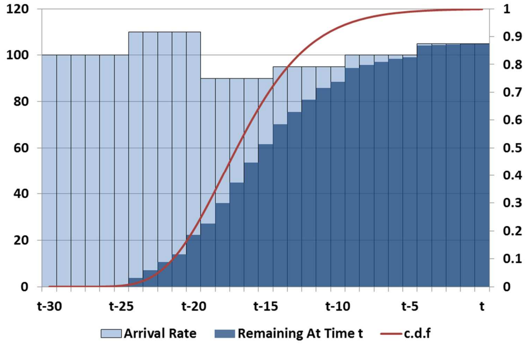


The Offered-Load of $M_t/GI/N_t + GI$

Insights:

The expected number of customers at time t, who arrived during the time interval $[s_1, s_2]$, is $\int_{s_1}^{s_2} \lambda(u) \cdot [1 - G(t-u)]du$

Alternative Arrival Process View



Estimation of the Offered-Load

- **$\mathbf{M_t/GI/\infty}$ queue:**
  1. The offered-load process, $L = \{L(t),\ t \geq 0\}$, equals the number of customers in service ($L(t)$ at time t)
  2. The offered-load function, R, is estimated as the average number of customers in service (over all available periods)
- **$\mathbf{M_t/GI/n_t}$ queue:**
  1. Eliminate the customers' waiting-times and shift their service period to start right upon arrival
  2. Then, follow the procedure of the $M_t/GI/\infty$ queue
- **$\mathbf{M_t/GI/n_t + GI}$ queue:**
  1. For the offered-load process, impute the service-times of abandoning customers and follow the $M_t/GI/n_t$ queue
  2. We propose a method to estimate the offered-load function, based on the expression $R(t) = \int_{-\infty}^{t} \lambda(u) \cdot [1 - G(t-u)]du$

Estimation of the Offered-Load

- $\mathbf{M_t/GI^*/n_t + GI}$ **queue:**

1. For the offered-load process
   - Estimate the distribution of the conditional service-time, $S|\tau = t$
   - Impute the service-times of abandoning customers
   - Follow the $M_t/GI/n_t$ procedure
2. For the offered-load function
   - Estimate the marginal service-time distribution from non-waiting customers
   - Apply the $M_t/GI^*/n_t + GI$ estimation procedure with this service-time distribution

Staffing and Performance Measures

Iterative Staffing Algorithm (ISA), a simulation code developed by Feldman et al. ['07]

- Determines time-dependence staffing levels aiming to achieve time-stable delay probability (hence time-stable performance)

- In our implementation, we added the feature of defining the relationship between patience and service-time in the time-varying $M_t/GI^*/n_t + GI$ queue

Remark: In this part of the thesis, we analyze only queues with homogenous Poisson arrival rate

Staffing and Performance Measures

ISA was applied to three types of $M/M^*/n+M$ queueing systems:

- Arrivals are according to a homogeneous Poisson process with arrival rate of 100 customers per time unit
- Patience is exponentially distributed with mean 1 time unit
- Mean service-time (unconditional) is equal to 1 time unit
- Service time, conditional on the patience of a customer, is exponentially distributed
- Relationship between patience and service-time differs accross models

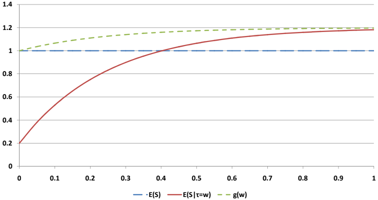
All performance measures are calculated as an average of 5000 replications

Staffing and Performance Measures

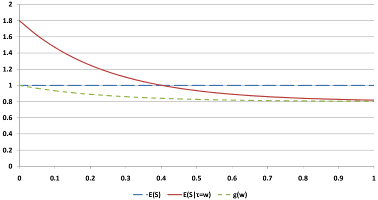
Description of the Relationship:

Increasing Monotone Function	$E(S\|\tau = w)$	$g(w)$
No Relation	1	1
Increasing Monotone Function	$1.2 - e^{-4 \cdot w}$	$1.2 - 0.2 \cdot e^{-4 \cdot w}$
Decreasing Monotone Function	$0.8 + e^{-4 \cdot w}$	$0.8 + 0.2 \cdot e^{-4 \cdot w}$

(a) Increasing Monotone Function

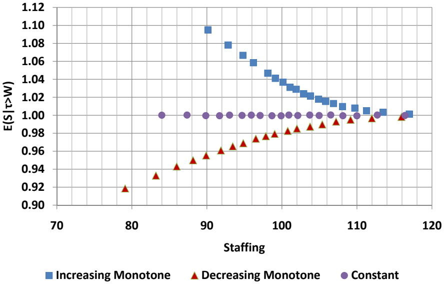


(b) Decreasing Monotone Function

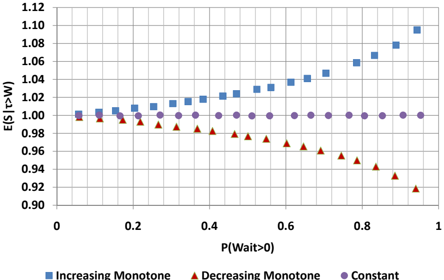


Staffing and Performance Measures

Comparison between the mean service-times of served customers as a function of the staffing level



Comparison between the mean service-times of served customers as a function of the probability of waiting



Staffing and Performance Measures

Analysis of the Square-Root Staffing Rule:

The rule for M/M/n+M queue is given by

$$n = R + \beta\sqrt{R},$$

where,

- R is the offered load function
- β is the Quality of Service (QoS) parameter, determined by the **Garnett function**:

$$\alpha = \left[1 + \sqrt{\frac{\theta}{\mu}} \cdot \frac{h(\beta\sqrt{\mu/\theta})}{h(-\beta)}\right]^{-1}, \quad -\infty < \beta < \infty,$$

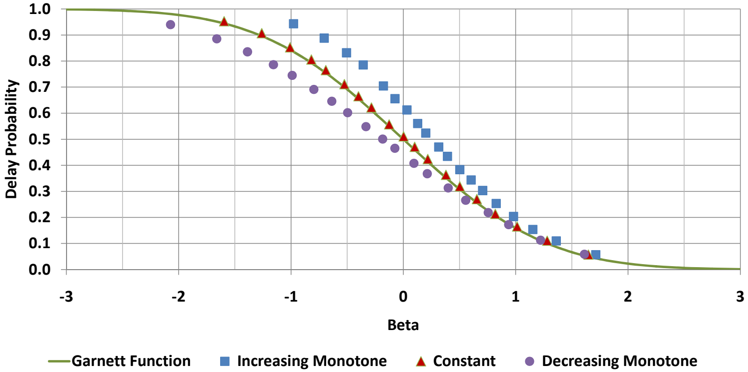
where α is the required probability of delay

Staffing and Performance Measures

For any target α, we ran an ISA simulation for each model
Define the implied quality of service grade

$$\beta^{ISA} \equiv \frac{n^{ISA} - R}{\sqrt{R}},$$

where n^{ISA} denotes the result staffing level of the simulation



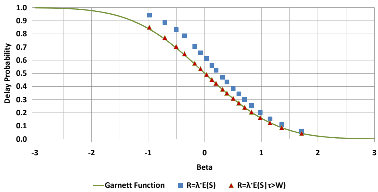
Staffing and Performance Measures

In fact, the mean service-time that the system faces, S^*, (due to served customers) is different from the unconditional mean service-time and is affected by the quality of service (determined by staffing level, N)

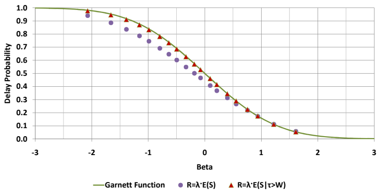
We define a modified offered-load expression, given by:

$$R^* = \lambda \cdot E(S^*(N))$$

(a) Increasing Monotone Function



(b) Decreasing Monotone Function



Notice that the offered-load is not influenced by the relationship

Analysis of the Service-Time vs Patience Relationship

Consider a simulation model:

The service-time of a customer with patience $\tau = t$ is a Log-Normal random variable, $S|\tau = t \sim LogNorm(\mu(t), \sigma^2)$, with *pdf*

$$f_{S|\tau}(s|t) = \frac{1}{s\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln s - \mu(t))^2}{2\sigma^2}}, \; s, t > 0$$

Then,

- $E(S|\tau = t) = e^{\mu(t) + \frac{\sigma^2}{2}}$
- $E(S^2|\tau = t) = e^{\sigma^2} e^{\mu(t) + \frac{\sigma^2}{2}} = e^{\sigma^2} E^2(S|\tau = t)$
- $Var(S|\tau = t) = (e^{\sigma^2} - 1) e^{\mu(t) + \frac{\sigma^2}{2}} = (e^{\sigma^2} - 1) E^2(S|\tau = t)$

Analysis of the Service-Time vs Patience Relationship

Assume that:

- τ is exponentially distributed, with mean $\frac{1}{\theta}$
- $E(S|\tau = t)$ be of the form:

$$E(S|\tau = t) = a \cdot (b - e^{-t\cdot(\beta-\theta)}),\ \beta > \theta,\ a > 0,\ b > 1$$

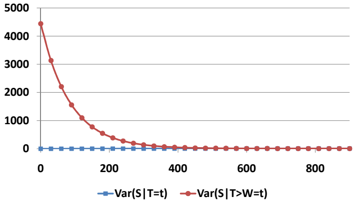
The analysis covers the following:

- $E(S|\tau = w) = 230 \cdot (1.2 - e^{-w\cdot(\frac{29}{3600} - \frac{8}{3600})})$
- $\theta = \frac{8}{3600}$
- Four values for σ are considered

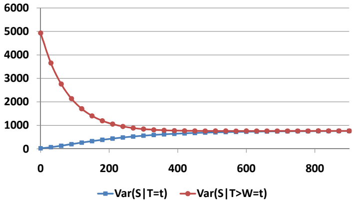
Analysis of the Service-Time vs Patience Relationship

A comparison between $Var(S|\tau = w)$ and $Var(S|\tau > W = w)$ with different values of σ

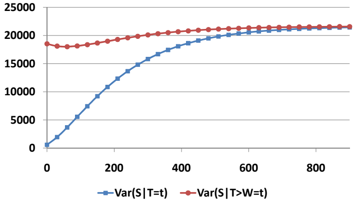
(a) $\sigma = 0.01$



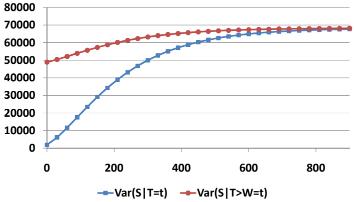
(b) $\sigma = 0.1$



(c) $\sigma = 0.5$



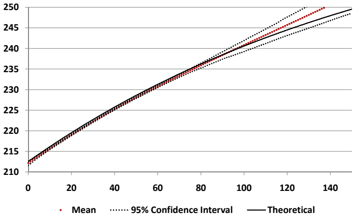
(d) $\sigma = 0.8$



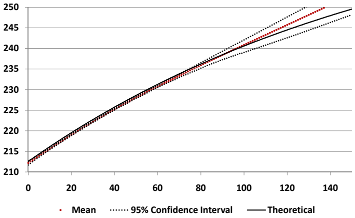
Analysis of the Service-Time vs Patience Relationship

95% percent confidence intervals of the spline estimator for $g(w) = E(S|\tau > W = w)$

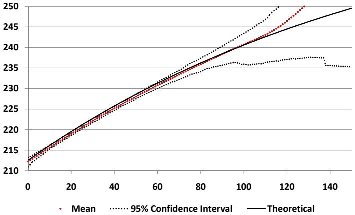
(a) $\sigma = 0.01$



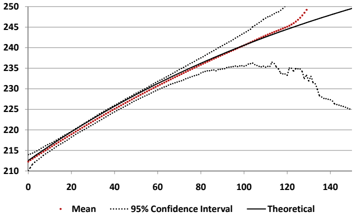
(b) $\sigma = 0.1$



(c) $\sigma = 0.5$



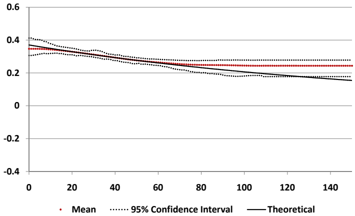
(d) $\sigma = 0.8$



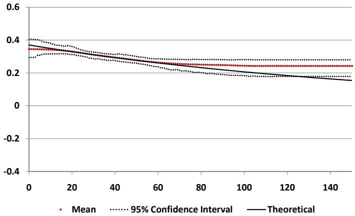
Analysis of the Service-Time vs Patience Relationship

95% percent confidence intervals of the derivative of the spline estimator for $g(w)$

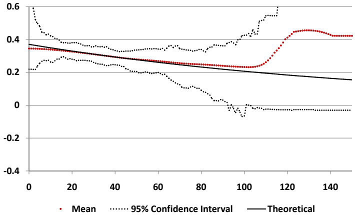
(a) $\sigma = 0.01$



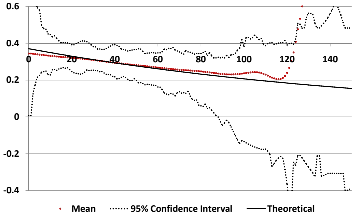
(b) $\sigma = 0.1$



(c) $\sigma = 0.5$



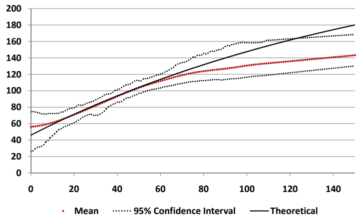
(d) $\sigma = 0.8$



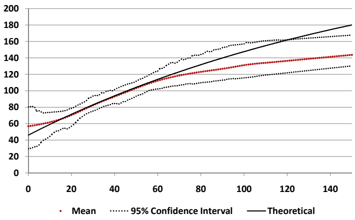
Analysis of the Service-Time vs Patience Relationship

95% percent confidence intervals of the estimator for $E(S|\tau = w)$

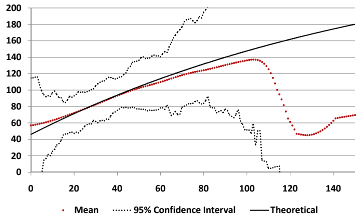
(a) $\sigma = 0.01$



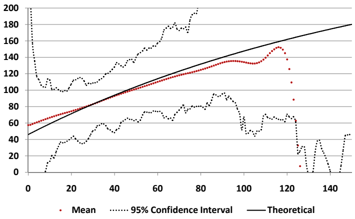
(b) $\sigma = 0.1$



(c) $\sigma = 0.5$



(d) $\sigma = 0.8$

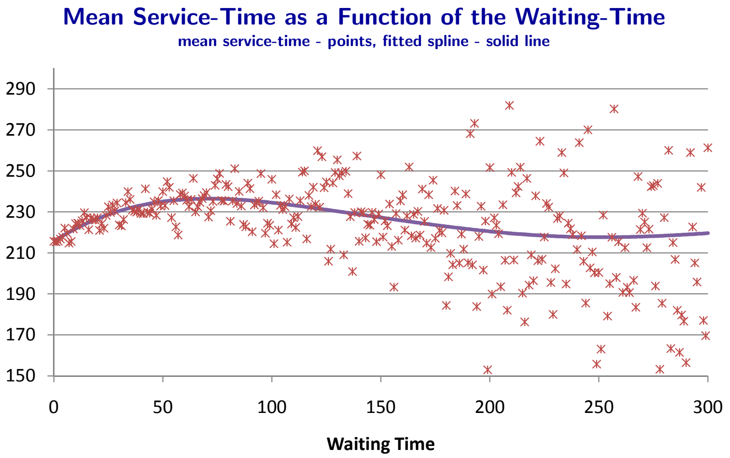


U.S. Bank Case Study

Data Description:

- A large North American commercial bank (U.S. Bank)
- Analysis is of the Retail Banking service
- Period - all weekdays (Monday through Friday) between January-June, 2006
- Observe arrivals between 10:00 and 16:00
- Total number of observations is 2,722,129, out of which 2,683,418 calls where served

U.S. Bank Case Study

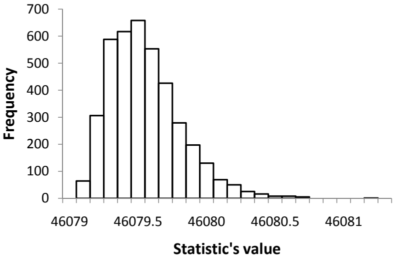


U.S. Bank Case Study

Testing the relationship between patience and service-time

- Consider only served calls
- Omit all non-waiting observations
- Divide the observations into 9 groups, by the ranking their waiting times
- Perform a random pairing permutation test (4000 replications)
- The value of the original permutation statistic is 46,140.62

A histogram for the distribution of the test statistic



U.S. Bank Case Study

In order to estimate $E(S|\tau = w)$ we use the formula

$$E(S|\tau = w) = g(w) - \frac{g'(w)}{h_\tau(w)}$$

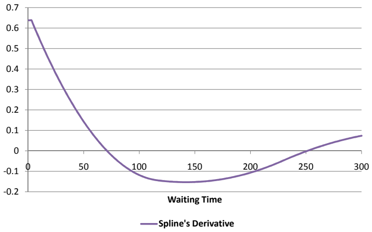
We fit a cubic smooth spline for $g(w)$ - the mean service-time, as a function of the waiting-time

- Designed to handle smooth functions
- Enables to simply extract the derivatives

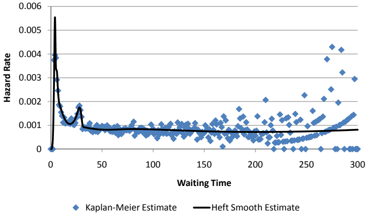
We choose a spline with only 5 knots - smoothness of the derivative

U.S. Bank Case Study

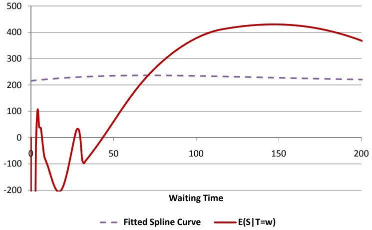
(a) Splines Derivative



(b) Hazard-Rate of the Patience



Estimator for the mean service-time as a function of the patience of a customer



Future Research

- A refinement of the estimation procedure of the mean service-time as a function of the patience
- Further research and modeling of the staffing rules and performance measures in the $M_t/GI^*/n_t + GI$ queue
- Apply the presented model to other databases

Thank You