Analyzing queues with abandon times Kaplan-Meier rivisited

Franca Rinaldi *Paolo Serafini*

Abstract

The analysis of abandons times is a central issue in modeling and simulating call centers due the strong impact of customer abandons on the behavior of a queuing system. In this paper we propose an original approach to estimate the distribution functions of the abandon times and the times-to-service given the observed waiting time of each costumer before either abandon or beginning of service. This approach can be used to derive both parametric and non-parametric estimators. In the non-parametric case it leads to a formula that corresponds to the well-known Kaplan-Meier estimator if applied to untied data but differs when applied to tied data. We also present two simulation examples that validate the proposed estimators.

1 Introduction

This paper has been motivated by the need of analyzing and predicting waiting times in a call center with the goal of improving the performance of the call center. There are many studies in the literature devoted to analyze and possibly improve the performance of a call center. See among others (Avramidis et al., 2010; Brown et al., 2005; Gand et al., 2003).

In a call center customers arrive, possibly wait for service and sometimes they abandon the queue before service. Besides, there are rules assigning calls to particular operators. In order to design effective assignment rules, one has to understand how the queuing system works. Of particular interest is the analysis of abandon times and in this paper we focus on this aspect of the problem.

In principle each customer is not willing to wait for a too long time and sooner or later will give up. In order to model this behavior we should be able to observe each customer, but this is impossible since only a fraction of customers abandons the queue. Similarly we wish to analyze the time-to-service, i.e., the time a customer is supposed to spend in line in case he/she decides not to abandon. Again we observe only partial data since only a fraction of customers is served.

If there are no abandons the time-to-service is just the waiting time, as it is referred to in the queuing literature. The possibility of leaving the queue before service makes the two concepts different.

Analyzing the abandon times is fundamental because a statistical knowledge of the abandon times allows for a more accurate simulation of a queuing system. The question of abandon times has been subject to an extensive treatment in the recent literature (Aksin et al., 2013; Conley, 2013; Ibrahim & Witt, 2011; Mandelbaum & Momčilović, 2012; Mandelbaum & Zeltyn, 2013). Also estimating the time-to-service is important. Although it is not necessary in case we simulate a queuing system, since the time-to-service is indirectly determined by the queue behavior, its knowledge is important if we want to have a direct performance measure of the system without entering into a detailed analysis of all components of the queuing system.

For each customer we observe the first occurrence of one of the two events: either a service beginning or an abandon. The behavior of a queue depends on both abandon times and times-to-service in an intricate way, since abandons speed up the time-to-service of other customers and long times-to-service increase the abandon rate. Furthermore, both events depend on the arrival times of the customers.

However, we may just disregard the complex interactions that take place in a queuing system and consider a generic customer provided with two independent random variables, one related to the abandon time and the other one to the time-to-service. This independence assumption, that can be justified theoretically and is also validated experimentally, is crucial in developing the estimators since it allows to infer the random variables via the theory of the so-called censored data.

The research of parametric and non-parametric estimation techniques in case of censored samples is a central issue of the Survival Analysis (for an extensive overview of the main results of the theory we refer the reader to Aalen et al. (2008); Rinne (2014); Cox & Oakes (1984) and the references therein). With no doubts, the two more popular non-parametric estimators in the case of censored data are the Kaplan-Meier (KM) (Kaplan & Meier, 1958) and the Nelson-Aalen (NA) estimators (Aalen, 1976; Nelson, 1972). These two estimators have been derived by following different points of view and indeed they are based on different formulas.

In particular, the KM-estimator has been proposed in Kaplan & Meier (1958) to estimate the survival function, whereas the NA-estimator has been proposed to estimate the cumulative hazard function. In principle either estimator can be used to estimate the other function due to the strict analytical link between

the survival function and the cumulative hazard function (see (6)). However, this transformation leads to formulas that do not have a direct justification and interpretation, especially in the case of continuous random variables. So the arguments used to derive the two estimators, their statistical properties and the opportunity to adopt either one according to the application context rely heavily on which of the two functions is taken into consideration.

In this paper we approach the estimation problem by first deriving formulas for continuous random variables that can be used both in a parametric and in a non-parametric way. Then we consider a non-parametric estimation based on smoothing the functions derived from the observed data. Our approach leads to a direct estimation of the cumulative hazard function and, via the fundamental relation (6), to an estimation of the survival function and the distribution function. It turns out that the final formula for the survival function corresponds exactly to the KM estimator, although our estimation starts from the cumulative hazard function instead of the survival function as has been done in Kaplan & Meier (1958), and is based on different considerations.

Hence our approach sheds new light on the KM-estimator and seems to further validate this estimator. In addition, our approach allows also to consider in a natural way the case of tied times without resorting to the usual assumption that, in case of ties, the event times occur just before the censored ones. By following simple analytical considerations we arrive to a formula that is the natural extension of the KM-estimator but is different from the one usually suggested.

The paper is structured as follows. In Section 2 we provide a formal definition of the various concepts. In Section 3 we introduce the theoretical formulas to compute the distribution functions of the abandon times and the times-to-service. In Section 4 we apply these formulas to the actual observations and derive the estimators in the case of untied data. We extend the estimator to the case of tied data in Section 5. In Section 6 we test the estimators by simulating a M/M/1 queuing system with exponentially distributed abandon times. While the abandon times are randomly generated, the times-to-service are not directly generated but they are computed according to the simulated interarrival times, service times and abandon times. For this particular queue a theoretical expression for the time-to-service is available (computed in Section 8) and therefore the estimated distribution functions can be directly compared to the theoretical ones. We also provide another simulation in which the abandon times are uniform on a given interval. In this case no theoretical expression is available for the times-to-service and the estimated distribution of the times-to-service is compared to the simulated time-to-service distribution obtained a

posteriori. Finally some conclusions are reported in Section 7.

2 Definitions

For a generic customer of a call center, let A be the real–valued random variable that denotes the maximum time the customer is willing to wait in the queue without being served. When this time is reached and the customer is still in the queue waiting for service, the customer abandons the queue. Let S be the real–valued r.v. that denotes the time-to-service, i.e., the time the customer is supposed to wait before being served, irrespective of the possibility that he/she abandons the queue.

We assume that a generic customer is provided with the two random variables A and S. Clearly S depends on other r.v.'s of the system, i.e., the interarrival times, the service times and the abandon times. The crucial issue regards independence with respect to abandon times. We assume that the r.v. S associated to a specific customer, while clearly dependent on the abandon times of previous customers, is independent of the r.v. A associated to the same customer, because a decision of leaving the queue has no effect on the time-to-service of this particular customer. We also assume that customers have no information when their service is supposed to start. Otherwise it is conceivable that this information strongly conditions the decision of leaving the queue or not. Hence we assume that A and S are independent. This allows for the validity of the relation (2) in the sequel.

Let Q be the r.v. that denotes the actual time when the customer leaves the queue. Since a costumer leaves the queue either because he/she abandons the queue or because he/she starts being served we have

$$Q = \min\{A, S\}. \tag{1}$$

Let the respective distributions functions be

$$F_A(t) = \Pr\{A \le t\}, \quad F_S(t) = \Pr\{S \le t\}, \quad F_Q(t) = \Pr\{Q \le t\},$$

with corresponding density functions

$$f_A(t) = F'_A(t), \quad f_S(t) = F'_S(t), \quad f_Q(t) = F'_Q(t).$$

Note that A and S are not directly observable, because for each customer we can observe either A or S but not both. Our purpose is to infer both A and S from the partial observations. By the independence of A and S and from (1) we have

$$1 - F_Q(t) = (1 - F_A(t))\,(1 - F_S(t)). \tag{2}$$

We assume $F_A(0) = 0$. In other words, no customer leaves the queue in the instant he/she joins the queue. This hypothesis excludes the case of customers that don't want to wait in the queue and therefore give up as soon as they know there is a queue. Although this is not an uncommon behavior we may just disregard these customers as not affecting the process. Indeed, by assuming exponential interarrival times, which is a fairly reasonable assumption, the random exclusion of some arrivals leaves the arrival process still exponential, though with a smaller arrival rate, but it is this rate that we measure and take into account. As a consequence, $F_Q(0) = F_S(0)$.

Differently, we must consider the case $F_S(0) > 0$, that corresponds to the fraction of customers that are served without waiting in line because they find an empty queue. In this case S exhibits a positive mass probability at $t = 0$ and we may express F_S and f_S as

$$F_S(t) = F_S(0)\,\mu(t) + \hat{F}_S(t), \qquad f_S(t) = F_S(0)\,\delta(t) + \hat{f}_S(t), \tag{3}$$

where $\mu(t)$ is the Heaviside function

$$\mu(t) := \begin{cases} 0 & t < 0 \\ 1 & t \geq 0 \end{cases},$$

$\delta(t)$ is the Dirac function for which

$$\mu'(t) = \delta(t), \quad \mu(t) = \int_{-\infty}^{t} \delta(\tau)\,d\tau$$

and $\hat{F}_S(t)$, $\hat{f}_S(t)$ are implicitly defined by (3).

The distributions of the r.v.'s A and S can be equivalently described in terms of the *hazard rate function* and the *cumulative hazard function*. The hazard rate function $h_A(t)$ of the continuous r.v. A has the form

$$h_A(t) = \frac{\Pr\{t \leq A \leq t + dt \,|\, t \leq A\}}{dt}$$

so that $h_A(t)\,dt$ represents the probability that a costumer abandons in $[t, t+dt]$ conditioned to the fact that he/she has not abandoned the queue before t. Therefore

$$h_A(t) = \frac{f_A(t)}{1 - F_A(t)} \qquad \text{for} \quad t : F_A(t) < 1. \tag{4}$$

The associated cumulative hazard function $H_A(t)$ is defined as

$$H_A(t) = \int_0^t h_A(\tau)\,d\tau. \tag{5}$$

Therefore

$$H_A(t) = \int_0^t \frac{f_A(\tau)}{1 - F_A(\tau)} d\tau = -\Big[\ln(1 - F_A(\tau))\Big]_0^t = -\ln(1 - F_A(t))$$

so that

$$F_A(t) = 1 - e^{-H(t)}. \tag{6}$$

Similar results hold for the hazard rate function $h_S(t)$ and the cumulative hazard function $H_S(t)$ of the r.v. S.

3 Theoretical expressions for $F_A(t)$ and $F_S(t)$

Let Q_A be the r.v. that denotes the time when the customer abandons the queue, if this happens, and Q_S the r.v. that denotes the time spent in the queue before being served, if this happens. The r.v.'s Q_A and Q_S are defective because the event 'abandon the queue' may not happen and similarly the event 'beginning of service' may not happen. Let

$$G_A(t) = \Pr\{Q_A \le t\}, \qquad G_S(t) = \Pr\{Q_S \le t\}.$$

Hence $G_A(\infty) < 1$ is the fraction of customers that abandon and $G_S(\infty) < 1$ is the fraction of customers that are served. In any case we have for any t

$$G_A(t) + G_S(t) = F_Q(t). \tag{7}$$

Note that $G_A(0) = 0$, necessarily from $F_A(0) = 0$, so that $G_S(0) = F_Q(0) = F_S(0)$.

Differently from A and S, the r.v.'s Q, Q_A and Q_S are observable. Hence we have to retrieve $F_A(t)$ and $F_S(t)$ from $F_Q(t)$, $G_A(t)$ and $G_S(t)$. The probability that an abandon will be observed in the interval $[t, t+dt]$ is given by

$$G'_A(t)\,dt = f_A(t)\,(1 - F_S(t))\,dt. \tag{8}$$

Similarly, the probability that the beginning of a service will be observed in the interval $[t, t+dt]$ is given by

$$G'_S(t)\,dt = f_S(t)\,(1 - F_A(t))\,dt,$$

which is valid also for $t = 0$.

From (2), (4) and (8) we derive

$$h_A(t) = \frac{f_A(t)}{1 - F_A(t)} = \frac{G'_A(t)}{1 - F_Q(t)}, \tag{9}$$

that allows to express the hazard rate function $h_A(t)$ in terms of the observable functions $G'_A(t)$ and $F_Q(t)$.

If there is a threshold T_A beyond which no abandons will take place and a threshold T_S beyond which no services will start, then $F_Q(t) = 1$ and $G'_A(t) = 0$

for $t \geq \min\{T_A, T_S\}$. Hence the ratio at the right in (9) is indeterminate for $t \geq \min\{T_A, T_S\}$, whereas the hazard rate function is indeterminate for $t \geq T_A$. Thus the identity (9) holds for $t < \min\{T_A, T_S\}$.

By assuming $t < \min\{T_A, T_S\}$ (if such thresholds are defined) we can compute the cumulative hazard function from (5) and (9) as

$$H_A(t) = \int_0^t \frac{G'_A(\tau)}{1 - F_Q(\tau)}\, d\tau \tag{10}$$

and, by applying (6), we can retrieve $F_A(t)$. However, in view of the previous observations, we can retrieve $F_A(t)$ only for $t < \min\{T_A, T_S\}$. If, in particular, $T_S < T_A$, no reconstruction is possible for $T_S \leq t \leq T_A$. This makes sense because all customers wait at most a time T_S to be served and thus no abandons take place after T_S.

We may carry out a similar computation for F_S, but only for $t > 0$, and obtain

$$h_S(t) = \frac{f_S(t)}{1 - F_S(t)} = \frac{G'_S(t)}{1 - F_Q(t)}, \quad t > 0, \tag{11}$$

that allows to express the hazard rate function $h_S(t)$ in terms of the observable functions $G'_S(t)$ and $F_Q(t)$. As before, the ratio at the right in (11) is indeterminate for $t \geq \min\{T_A, T_S\}$ whereas the hazard rate function is indeterminate for $t \geq T_S$. Thus the identity (11) holds for $t < \min\{T_A, T_S\}$.

By integrating both sides of (11) and avoiding the discontinuity at $t = 0$ we have

$$\lim_{t' \to 0^+} \int_{t'}^t \frac{G'_S(\tau)}{1 - F_Q(\tau)}\, d\tau = \lim_{t' \to 0^+} \int_{t'}^t \frac{f_S(\tau)}{1 - F_S(\tau)} d\tau =$$

$$- \lim_{t' \to 0^+} \Big[\ln(1 - F_S(\tau)) \Big]_{t'}^t = -\ln \frac{1 - F_S(t)}{1 - F_S(0)}.$$

If we denote

$$H_S(t) = \lim_{t' \to 0^+} \int_{t'}^t \frac{G'_S(\tau)}{1 - F_Q(\tau)}\, d\tau, \qquad t > 0, \tag{12}$$

we may write

$$F_S(t) = 1 - (1 - F_S(0))\, e^{-H_S(t)}. \tag{13}$$

Again, we can reconstruct $F_S(t)$ from (13) only for $t \leq \min\{T_A, T_S\}$. If, in particular, $T_A < T_S$, no reconstruction is possible for $T_A \leq t \leq T_S$. We see that full reconstruction of both $F_A(t)$ and $F_S(t)$ is possible only if $T_A = T_S$ (or they are both infinite).

Example. We find it useful to show the various relations by referring to a simple example. Assume that A is uniform on $[0, T]$ and that $S = 0$ with probability

1/2 and otherwise uniform on $[0, 2\,T]$ with overall probability 1/2. Hence in this case $T_A = T < T_S = 2\,T$ and

$$F_A(t) = \frac{t}{T}, \quad 0 \le t \le T, \quad F_A(t) = 1, \quad t \ge T, \qquad f_A(t) = \frac{1}{T}, \quad 0 \le t \le T,$$

$$F_S(t) = \frac{1}{2}\,(1 + \frac{t}{2\,T}), \quad 0 \le t \le 2\,T, \quad F_S(t) = 1, \quad t \ge 2\,T,$$

$$f_S(t) = \frac{1}{2}\,(\delta(t) + \frac{1}{2\,T}), \quad 0 \le t \le 2\,T,$$

We remark that these functions are unknown to the observer and his/her task is to retrieve them. We assume that the observer knows the following functions that he/she exactly infers from the observed data (while we derive them from $F_A(t)$ and $F_S(t)$)

$$F_Q(t) = \frac{1}{2} + \frac{3}{4}\,\frac{t}{T} - \frac{1}{4}\,\frac{t^2}{T^2}, \quad 0 \le t \le T, \qquad F_Q(t) = 1, \quad t > T,$$

$$f_Q(t) = \frac{1}{2}\,\delta(t) + \frac{3}{4}\,\frac{1}{T} - \frac{1}{2}\,\frac{t}{T^2}, \qquad 0 \le t \le T,$$

$$G'_A(t) = f_A(t)\,(1 - F_S(t)) = \frac{1}{2\,T}\,(1 - \frac{t}{2\,T}), \qquad 0 \le t \le T,$$

$$G_A(t) := \int_0^t \frac{1}{2\,T}\,(1 - \frac{\tau}{2\,T})\,d\tau = \frac{t}{2\,T}\,(1 - \frac{t}{4\,T}), \qquad 0 \le t \le T,$$

$$G'_S(t) = f_S(t)\,(1 - F_A(t))\,dt = \frac{1}{2}\,(\delta(t) + \frac{1}{2\,T})\,(1 - \frac{t}{T}), \qquad 0 \le t \le T,$$

$$G_S(t) = \frac{1}{2}\int_0^t (\delta(\tau) + \frac{1}{2\,T})\,(1 - \frac{\tau}{T})\,d\tau = \frac{1}{2} + \frac{1}{4}\,(\frac{t}{T} - \frac{t^2}{2\,T^2}).$$

Given these functions the observer may compute the cumulative hazard function according to (10) as

$$H_A(t) = \int_0^t \frac{\frac{1}{2\,T}\,(1 - \frac{\tau}{2\,T})}{\frac{1}{2} - \frac{3}{4}\,\frac{\tau}{T} + \frac{1}{4}\,\frac{\tau^2}{T^2}}\,d\tau = -\ln(1 - \frac{t}{T}), \qquad 0 \le t < T,$$

and retrieve the function $F_A(t)$ as

$$F_A(t) = 1 - e^{-H_A(t)} = 1 - e^{\ln(1 - \frac{t}{T})} = 1 - (1 - \frac{t}{T}) = \frac{t}{T}, \qquad 0 \le t < T.$$

Note that $\lim_{t \to T} F_A(t) = 1$, i.e., full reconstruction of $F_A(t)$ has been possible. As for the time-to-service we have according to (12)

$$H_S(t) = \lim_{t' \to 0^+} \int_{t'}^t \frac{\frac{1}{4\,T}\,(1 - \frac{\tau}{T})}{\frac{1}{2} - \frac{3}{4}\,\frac{\tau}{T} + \frac{1}{4}\,\frac{\tau^2}{T^2}}\,d\tau = -\ln(1 - \frac{t}{2\,T}), \qquad t < T,$$

so that

$$F_S(t) = 1 - (1 - F_S(0))\, e^{-H_S(t)} = 1 - \frac{1}{2}\,(1 - \frac{t}{2\,T}) = \frac{1}{2} + \frac{t}{4\,T}, \qquad t < T.$$

In this case $\lim_{t\to T} F_S(t) = 3/4$, and no reconstruction of $F_S(t)$ is possible for $t > T = T_A$ because there are no data available after T.

4 Empirical computation of $F_A(t)$ and $F_S(t)$

Our goal in this section is to estimate the functions $F_A(t)$ and $F_S(t)$ from the observed data. As already remarked we may compute $F_A(t)$ from the $H_A(t)$ which in turn can be derived from the knowledge of $G_A(t)$ and $F_Q(t)$.

We have chosen a non parametric method of inferring $G_A(t)$ and $F_Q(t)$ by approximating these functions with continuous quasi-stepwise functions, that are obtained by smoothing the usual stepwise estimators in small intervals around the observed time instants. The final estimator of $H_A(t)$ and $F_A(t)$ will be obtained by shrinking these intervals to the time instants.

The smoothing of discrete estimators to have continuous probability distribution functions is a well known and largely adopted technique (Rinne, 2014). We initially smooth the functions in order to carry out some analytical computations that are otherwise not well defined. However, after having carried out these computations we get back to the original sharp functions.

We consider a family of continuous quasi-stepwise functions $\delta_\varepsilon(t)$, parametrized by $\varepsilon > 0$, continuous on $-\varepsilon < t < \varepsilon$, such that

$$\delta_\varepsilon(t) = 0 \quad \text{if } |t| \geq \varepsilon, \quad \delta_\varepsilon(t) = \delta_\varepsilon(-t), \quad \int_{-\infty}^{+\infty} \delta_\varepsilon(t)\, dt = \int_{-\varepsilon}^{+\varepsilon} \delta_\varepsilon(t)\, dt = 1.$$

Let $\mu_\varepsilon(t)$ be the family of continuous functions such that

$$\mu_\varepsilon(t) = \int_{-\infty}^{t} \delta_\varepsilon(\tau)\, d\tau = \int_{-\varepsilon}^{t} \delta_\varepsilon(\tau)\, d\tau.$$

Hence

$$\mu_\varepsilon(t) = 0, \text{ for } t \leq -\varepsilon, \qquad \mu_\varepsilon(t) = 1, \text{ for } t \geq \varepsilon, \qquad \mu_\varepsilon(0) = \frac{1}{2},$$

$$\frac{d\mu_\varepsilon(t)}{dt} = \delta_\varepsilon(t), \quad \text{for } -\varepsilon < t < \varepsilon.$$

For $\varepsilon \to 0$, $\delta_\varepsilon(t)$ tends to the Dirac function $\delta(t)$ and $\mu_\varepsilon(t)$ tends to the Heaviside function $\mu(t)$ with the convention $\mu(0) = 1/2$. For instance we may take

$$\delta_\varepsilon(t) = \frac{1}{2\,\varepsilon}, \quad \mu_\varepsilon(t) = \frac{t + \varepsilon}{2\,\varepsilon}, \qquad -\varepsilon < t < \varepsilon.$$

The rationale behind the use of this family of functions is that instead of considering random variables that occur exactly at the observed time instants we blur the picture and consider random variables that are equally likely in a small neighborhood of the same time instants.

The observed data are the realizations of the random variable $Q = \min\{A, S\}$, i.e., the observed values that we sort as $q_1, q_2, \ldots, q_n$. Moreover we know which q_i's are realizations of Q_A and which q_i's are realizations of Q_S. Let $a_1, \ldots, a_p$ be the sorted realizations of Q_A and $s_1, \ldots, s_r$ be the sorted realizations of Q_S. Clearly both $a_p \leq \min\{T_A, T_S\}$ and $s_r \leq \min\{T_A, T_S\}$ must hold.

We approximate $G_A(t)$ by the quasi-stepwise function $\bar{G}_A(t)$ defined as

$$\bar{G}_A(t) = \frac{1}{n} \sum_{k=1}^{p} \mu_\varepsilon(t - a_k),$$

from which

$$\bar{G}'_A(t) = \frac{1}{n} \sum_{k=1}^{p} \delta_\varepsilon(t - a_k).$$

Similarly, we approximate $F_Q(t)$ by the quasi-stepwise function

$$\bar{F}_Q(t) = \frac{1}{n} \sum_{k=1}^{n} \mu_\varepsilon(t - q_k), \qquad 1 - \bar{F}_Q(t) = \frac{1}{n} \left(n - \sum_{k=1}^{n} \mu_\varepsilon(t - q_k)\right).$$

Therefore we approximate the cumulative hazard function $H_A(t)$ in (10) with the function

$$\bar{H}_A(t) = \int_0^t \frac{\sum_{k=1}^{p} \delta_\varepsilon(\tau - a_k)}{n - \sum_{k=1}^{n} \mu_\varepsilon(\tau - q_k)}\, d\tau = \sum_{k=1}^{p} \int_0^t \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau =$$

$$\sum_{k=1}^{p} \int_{\min\{t, a_k - \varepsilon\}}^{\min\{t, a_k + \varepsilon\}} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau. \tag{14}$$

Let us first assume that all q_i's are different, i.e., we have untied data. We analyze the case of equal data in Section 5.

In case of non equal data, there exists $\varepsilon > 0$ such that for each $k = 1, \ldots, p$, no value q_h, $h \neq k$, falls in the interval $[a_k - 2\,\varepsilon, a_k + 2\,\varepsilon]$. Consequently, for $\tau \in [a_k - \varepsilon, a_k + \varepsilon]$, $\mu_\varepsilon(\tau - q_h) = 1$ if $q_h < a_k$ and $\mu_\varepsilon(\tau - q_h) = 0$ if $q_h > a_k$. This implies that for $\tau \in [a_k - \varepsilon, a_k + \varepsilon]$ we have

$$n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h) = n - |\{h : a_k > q_h\}| - \mu_\varepsilon(\tau - a_k) = |\{h : a_k \leq q_h\}| - \mu_\varepsilon(\tau - a_k).$$

Since

$$\frac{d}{d\tau}(|\{h : a_k \leq q_h\}| - \mu_\varepsilon(\tau - a_k)) = -\delta_\varepsilon(\tau - a_k), \quad \text{for } a_k - \varepsilon < \tau < a_k + \varepsilon$$

we obtain

$$\int_{a_k-\varepsilon}^{a_k+\varepsilon} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau = -\Big[\ln(|\,\{h : a_k \le q_h\}\,| - \mu_\varepsilon(\tau - a_k)) \Big]_{a_k-\varepsilon}^{a_k+\varepsilon} =$$

$$\ln\left(\frac{|\,\{h : a_k \le q_h\}\,|}{|\,\{h : a_k < q_h\}\,|}\right). \tag{15}$$

Note that this expression is unbounded if k corresponds to the last observed event and this is an abandon. This means that the function $\bar{H}_A(t)$ tends to infinity as t tends to $a_k + \varepsilon$ and a_k is the last observed event. By denoting as

$$\gamma_k := \frac{1}{|\,\{h : a_k < q_h\}\,|} \tag{16}$$

the inverse of the number of customers that are still in the queue just after a_k, we may rewrite (15) as

$$\int_{a_k-\varepsilon}^{a_k+\varepsilon} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau = \ln(1 + \gamma_k).$$

Note that this expression is independent of ε and therefore it remains invariant when ε tends to zero. Furthermore, for $a_k - \varepsilon < t < a_k + \varepsilon$, we have

$$\int_{a_k-\varepsilon}^{t} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau = \ln\left(\frac{1 + \gamma_k}{1 + \gamma_k - \gamma_k\, \mu_\varepsilon(t - a_k)}\right).$$

In particular, if $t = a_k = a_p = q_n$ is the last observed event and thus $\gamma_p = +\infty$, the last expression is equal to $\ln 2$ (due to our assumption $\mu_\varepsilon(0) = 1/2$).

Therefore the function $\bar{H}_A(t)$ has a value independent of ε on all intervals $[a_{k-1} + \varepsilon, a_k - \varepsilon]$. If we let ε tend to zero, the function $\bar{H}_A(t)$ becomes

$$\bar{H}_A(t) = \begin{cases} 0 & t < a_1 \\ \displaystyle\sum_{h=1}^{k} \ln(1 + \gamma_h) - \ln(1 + \frac{\gamma_k}{2}) & t = a_k,\ k = 1, \ldots, p \\ \displaystyle\sum_{h=1}^{k} \ln(1 + \gamma_h) & a_k < t < a_{k+1},\ k = 1, \ldots, p-1 \\ \displaystyle\sum_{h=1}^{p} \ln(1 + \gamma_h) & t > a_p \end{cases} \tag{17}$$

and therefore, the approximated distribution function $\bar{F}_A(t) = 1 - e^{-\bar{H}_A(t)}$ has

the form

$$\bar{F}_A(t) = \begin{cases} 0 & t < a_1 \\ 1 - \prod_{h=1}^{k} (1+\gamma_h)^{-1} (1 + \frac{\gamma_k}{2}) & t = a_k,\ k = 1, \ldots, p \\ 1 - \prod_{h=1}^{k} (1+\gamma_h)^{-1} & a_k < t < a_{k+1},\ k = 1, \ldots, p-1, \\ 1 - \prod_{h=1}^{p} (1+\gamma_h)^{-1} & t > a_p. \end{cases} \tag{18}$$

If $a_p < q_n = s_r$ is the last abandon but not the last observed event, we have that $\bar{F}_A(t)$ is constant and strictly less than 1, for $t \geq a_p$. Therefore full reconstruction of $F_A(t)$ is impossible beyond the last abandon in this case. If, on the contrary, $a_p = q_n$ is the last observed event and thus $\gamma_p = +\infty$, we have $\bar{F}_A(t) = 1$ for $t > a_p$ and full reconstruction is possible.

These considerations are consistent with the previous observations concerning the time horizons T_A and T_S. Suppose that $T_A < T_S$. In this case events close to T_A will be abandons with larger probability than times-to-service. This means that very likely we observe $s_r < a_p = q_n$, i.e., full reconstruction (although clearly approximate) of F_A is possible and this is consistent with the previous observation in Section 3 that in theory full reconstruction of F_A is possible only if $T_A < T_S$.

If on the contrary $T_S < T_A$, we very likely observe $a_p < s_r = q_n$, i.e., full reconstruction of F_A is not possible. However, although no abandon data are available beyond a_p, we know that $a_p < s_r \leq T_A$ and therefore there is a positive probability for abandons up to s_r at least. In other words, if a customer has experienced a time-to-service equal to s_r, this means that his abandon time was at least as large as s_r. Hence $\bar{F}_A(a_p) < \bar{F}_A(t) < 1$ for $a_p < t < q_n$.

As a remarkable fact we observe that the estimator (18) coincides, apart on the points a_k, with the KM estimator proposed by Kaplan & Meier (1958). However, it was derived with a totally different approach that directly estimate the cumulative hazard function $H_A(t)$.

The function $\bar{F}_A(t)$ obtained this way is a step-wise function. Its shape depends on two independent factors. One is the order in which the events of different type happen. This affects the height of the steps as apparent from (18). In particular on passing from $t < a_k - \varepsilon$ to $t > a_k + \varepsilon$ the step height is

$$\prod_{h=1}^{k-1} (1+\gamma_h)^{-1} - \prod_{h=1}^{k} (1+\gamma_h)^{-1} = \prod_{h=1}^{k-1} (1+\gamma_h)^{-1} \frac{\gamma_k}{1+\gamma_k}$$

and depends on how abandons and times-to-service are intertwined but does not depend on when they happen. The other factor is the time interval between

two abandons which is responsible for the length of the steps. The dependence on the order of the events is crucial and heavily influences the estimation in case of equal times as we shall see in Section 5.

For the computation of $\bar{F}_S(t)$ let us assume that $s_k = 0$ for $1 \le k \le r'$ and $|s_{k+1} - s_k| > 2\,\varepsilon$ for $r' < k \le r$. In this case we have to take into account that $F_S(0)$ may be positive. Hence, by following the same approach as for the abandon times, we may write

$$\bar{H}_S(t) = \lim_{t' \to 0^+} \int_{t'}^{t} \frac{\sum_{k=1}^{r} \delta_\varepsilon(\tau - s_k)}{n - \sum_{k=1}^{n} \mu_\varepsilon(\tau - q_k)}\, d\tau = \sum_{k=1}^{r} \lim_{t' \to 0^+} \int_{t'}^{t} \frac{\delta_\varepsilon(\tau - s_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau.$$

The limit in the integral implies that only the points for which $s_k > 0$ must be evaluated. Hence

$$\bar{H}_S(t) = \sum_{k>r'}^{r} \int_{\min\{t, s_k - \varepsilon\}}^{\min\{t, s_k + \varepsilon\}} \frac{\delta_\varepsilon(\tau - s_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau.$$

Let

$$\eta_k := \frac{1}{n - \sum_{h=1}^{n} \mu(s_k - q_h)} = \frac{1}{|\,\{h : s_k < q_h\}\,|}, \qquad r' < k \le r. \tag{19}$$

Then, by carrying out the same steps as for the abandon times, and taking into account that

$$\bar{F}_S(0) = \bar{F}_Q(0) = \frac{|s_k : s_k = 0|}{n} = \frac{r'}{n},$$

and, by (13), we end up with, for $s_k < t < s_{k+1}$,

$$\bar{F}_S(t) = 1 - (1 - \bar{F}_S(0))\, e^{-\bar{H}_S(t)} = 1 - (1 - \frac{r'}{n}) \prod_{h>r'}^{k} (1+\eta_h)^{-1}, \quad k = r'+1, \ldots, r. \tag{20}$$

5 Computing $\bar{F}_A(t)$ with equal data

We now address the issue of equal data, i.e., how the previous analysis changes when the experimental data present a mix of equal abandon times and/or times-to service. Although in our model the time is continuous and the probability of equal data is zero, nonetheless it is useful to consider the possibility of equal data for two reasons. For practical reasons, it is quite normal that time measurements in data coming from real observations are discretized and it may happen that some data are equal. This is a common situation when, as it happens for call-centers, the samples are large and the time is measured in a rough unit, usually seconds. For theoretical purposes, we would like an estimate to show some continuity properties. In other words, if two events get closer we would like the estimate to converge to the same estimate obtained by considering equal data.

We limit ourselves to consider the estimator $\bar{F}_A(t)$ for the abandon times. In this case the occurrence of equal data that are all times-to-service does not introduce any change in the analysis reported in Section 4. So we consider the occurrence of equal data of both types, i.e., a mix of abandons and times-to-service. The results for equal data that are all abandons will be deduced as a particular case.

Let us first consider the general case and assume that there are r_A abandons and r_S times-to-service with the same value $q_i = \cdots = q_{i+r-1}$ where $r = r_A + r_S$. For the sake of notational simplicity we assume that equal data occur only at q_i, but the analysis can be easily extended assuming there are multiple events in other time instants. Let $a_k = q_i$ and $a_{k-1} < q_i$, so that all the abandon times indexed from k up to $k + r_A - 1$ are equal to q_i.

We mimic the previous analysis with the only difference given by the r equal data. As before, we assume that there exists $\varepsilon > 0$ such that no value q_h, $h < i$ and $h > i + r - 1$, falls in the interval $[a_k - 2\,\varepsilon, a_k + 2\,\varepsilon]$. As a consequence, for $\tau \in [a_k - \varepsilon, a_k + \varepsilon]$, it holds $\mu_\varepsilon(\tau - q_h) = 1$ if $q_h < a_k$ and $\mu_\varepsilon(\tau - q_h) = 0$ if $q_h > a_k$, so that

$$n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h) = n - |\,\{h : a_k > q_h\}\,| - r\,\mu_\varepsilon(\tau - a_k) = |\,\{h : a_k \le q_h\}\,| - r\,\mu_\varepsilon(\tau - a_k).$$

Since

$$\frac{d}{d\tau}(|\,\{h : a_k \le q_h\}\,| - r\,\mu_\varepsilon(\tau - a_k)) = -r\,\delta_\varepsilon(\tau - a_k),$$

we obtain

$$\int_{a_k-\varepsilon}^{a_k+\varepsilon} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau = -\frac{1}{r}\left[\ln(|\,\{h : a_k \le q_h\}\,| - r\,\mu_\varepsilon(\tau - a_k))\right]_{a_k-\varepsilon}^{a_k+\varepsilon} =$$

$$\frac{1}{r}\,\ln\left(\frac{|\,\{h : a_k \le q_h\}\,|}{|\,\{h : a_k \le q_h\}\,| - r}\right) = \frac{1}{r}\,\ln\left(\frac{|\,\{h : a_k \le q_h\}\,|}{|\,\{h : a_k < q_h\}\,|}\right).$$

By defining as before

$$\gamma_k := \frac{1}{|\,\{h : a_k < q_h\}\,|}$$

we may rewrite the previous equality as

$$\int_{a_k-\varepsilon}^{a_k+\varepsilon} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau = \frac{1}{r}\,\ln(1 + r\,\gamma_k). \tag{21}$$

Since the expression (14) of $\bar{H}_A(t)$ contains r_A summands equal to (21), we obtain for ε tending to 0

$$\bar{F}_A(t) = 1 - \prod_{h=1}^{k-1} (1 + \gamma_h)^{-1}\,(1 + r\,\gamma_k)^{-r_A/r} \qquad a_k < t < a_{k+r}. \tag{22}$$

The occurrence of equal data is thus captured by the factor

$$(1 + r\,\gamma_k)^{-r_A/r}. \tag{23}$$

Furthermore, for times t with $a_k - \varepsilon < t < a_k + \varepsilon$ we have

$$\int_{a_k-\varepsilon}^{t} \frac{\delta_\varepsilon(\tau - a_k)}{n - \sum_{h=1}^{n} \mu_\varepsilon(\tau - q_h)}\, d\tau = \frac{1}{r}\, \ln\left(\frac{1 + r\,\gamma_k}{1 + r\,\gamma_k - r\,\gamma_k\,\mu_\varepsilon(t - a_k)}\right)$$

and for ε tending to 0

$$\bar{F}_A(a_k) = 1 - \prod_{h=1}^{k-1}(1 + \gamma_h)^{-1} \left(\frac{1 + r\,\gamma_k}{1 + \frac{r}{2}\,\gamma_k}\right)^{-r_A/r}.$$

In the particular case when the equal data are all abandons, i.e., $r_S = 0$, the factor (23) and the expression (22) become, respectively,

$$(1 + r_A\,\gamma_k)^{-1} \tag{24}$$

$$\bar{F}_A(t) = 1 - \prod_{h=1}^{k-1}(1 + \gamma_h)^{-1}\,(1 + r_A\,\gamma_k)^{-1} \qquad a_k < t < a_{k+r}. \tag{25}$$

It is easy to see that formula (25) is the same one obtains in the hypothesis that r_A abandons occur on different (very close) times $a_k < \cdots < a_{k+i} < \cdots < a_{k+r_A-1}$ without any time-to-service in between. Indeed, by directly applying the expression (18) of $\bar{F}_A(t)$, valid in the case of non equal data, for the values

$$\gamma'_i = \frac{1}{\frac{1}{\gamma_k} + r_A - i - 1}, \qquad i = 0, \cdots, r_A - 1,$$

one obtains, for $a_k < t < a_{k+r}$,

$$\bar{F}_A(t) = 1 - \prod_{h=1}^{k-1}(1+\gamma_h)^{-1} \prod_{i=0}^{r_A-1} \frac{1 + (r_A - i - 1)\,\gamma_k}{1 + (r_A - i)\,\gamma_k} = 1 - \prod_{h=1}^{k-1}(1+\gamma_h)^{-1}\,(1+r_A\,\gamma_k)^{-1}.$$

Therefore the same formula holds both for very close data and for equal data. In fact, it is not necessary that the events are close. The only thing that matters is that there is a sequence of abandons without any time-to-service in between. If the r_A abandons are equal, the function $\bar{F}_A(t)$ exhibits a step of height

$$\prod_{h=1}^{k-1}(1 + \gamma_h)^{-1} \frac{r_A\,\gamma_k}{1 + r_A\,\gamma_k}$$

whereas if they are not equal, but there are no times-to-service in between, the function exhibits r_A steps of total height the same quantity.

Unfortunately, the same continuity property does not hold in the case of events of different types. Indeed, the formula one obtains by applying (18) in the hypothesis that the times q_i, $t \leq i \leq t+r-1$, are different (even if arbitrarily close), strongly depends on the order in which the two types of events occur. For instance, consider the two limit situations in which all abandons occur first and then the times-to-service and the symmetric one in which all times-to-service come first and then the abandons. In both cases we can apply the results obtained for equal data of abandon type only. With regard to the first situation, this requires to compute expression (24) for the value $\gamma'_k := 1/(1/\gamma_k + r - r_A)$. This leads to the factor

$$\frac{1 + (r - r_A)\,\gamma_k}{1 + r\,\gamma_k}. \tag{26}$$

Clearly, (23) and (26) are different unless $r = r_A$. The power series expansions of (23) and (26) are

$$(1 + r\,\gamma_k)^{-r_A/r} \approx 1 - r_A\,\gamma_k + \frac{1}{2}\,r_A\,(r + r_A)\,\gamma^2 - \frac{1}{6}\,r_A\,(2\,r^2 + 3\,r\,r_A + r_A^2)\,\gamma^3$$

$$\frac{1 + (r - r_A)\,\gamma_k}{1 + r\,\gamma_k} \approx 1 - r_A\,\gamma_k + r\,r_A\,\gamma_k^2 - r^2\,r_A\,\gamma_k^3$$

from which we may see that (23) and (26) are equal only at the first order.

With regards to the second situation when all times-to-service precede the abandons, we have simply to apply the factor (24)

$$(1 + r_A\gamma_k)^{-1}.$$

This expression is also different from (23). Its power series expansion is

$$(1 + r_A\,\gamma_k)^{-1} \approx 1 - r_A\,\gamma_k + r_A^2\,\gamma_k^2 - r_A^3\,\gamma_k^3.$$

It can be shown that

$$(1 + r_A\,\gamma_k)^{-1} < (1 + r\,\gamma_k)^{-r_A/r} < \frac{1 + (r - r_A)\,\gamma_k}{1 + r\,\gamma_k}. \tag{27}$$

If we mix together abandons and services in an arbitrary way we obtain an expression in between the two extremes.

As observed before the order of the events is crucial. A different function $\bar{F}_A(t)$ is obtained for a different ordering of the events. Hence compressing all events in a single time instant has the effect of loosing any ordering and the expression we obtain is like an average. In the special case $r_A = r_S$ the intermediate term in (27) is the geometric average of the two extremes.

It is interesting to remark that Kaplan & Meier (1958) (see also Aalen et al. (2008)) propose to deal with equal data of the two types by using the expression

(26). Indeed the authors suggest to practically face such a situation as if all the r_A abandons would occur just before q_i and all the r_S services would occur just after q_i. However, this assumption is not justified theoretically. In fact they also report some alternative formulas already proposed in the literature to deal with the uncertainty regarding the ordering of the events occurring on a given interval of time. One of these formulas, mentioned in Kaplan & Meier (1958) as *joint risk estimate* is just the expression (23). Therefore, the approach proposed in this paper shows that a natural generalization of the Kaplan-Meier estimator to the case of equal data is provided by the expression (23) rather than (26).

The difference among the various formulas can be striking if the originally different data are aggregated into large bunches. This will be shown on simulated data in the next section.

With obvious modifications, the above analysis works also for the estimator $\bar{F}_S(t)$.

6 Two examples on simulated data

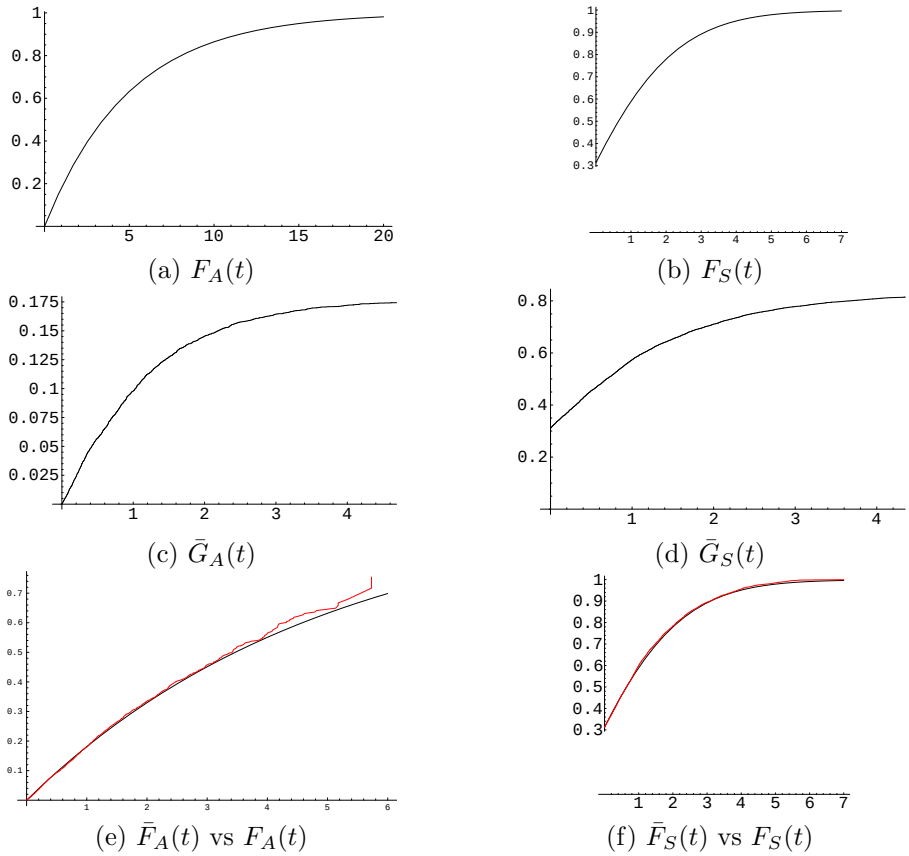
We now simulate a queue behavior so that $F_A(t)$ and $F_S(t)$ are known in advance and the validity of the above procedure can be checked by comparing $F_A(t)$ with $\bar{F}_A(t)$ and $F_S(t)$ with $\bar{F}_S(t)$. Note that S in the simulation is not directly generated but it is computed according to the arrivals, service times and abandon times.

We simulate a queue $M/M/1$ with $\lambda = 1$, $\mu = 1.2$ and $n = 10,000$ customers. The abandon times are exponential with rate $\alpha = 0.2$, i.e., $F_A(t) = 1 - e^{-0.2\,t}$ (see Fig. 1-(a)). This is equivalent to a $M/M/1$ queue modeled as a Markov chain with transition rates $\mu + k\,\alpha$ from state $k+1$ to state k (and clearly λ from k to $k+1$). It is possible to prove that

$$F_S(t) = F_S(0)\left(1 + e^{\lambda/\alpha} \sum_{h \geq 0} \frac{\lambda^{h+1}}{(\mu + h\,\alpha)} \frac{(-1)^h}{h!\,\alpha^h} \left(1 - e^{-(\mu + h\,\alpha)\,t}\right)\right)$$

where $F_S(0)$, the probability that a generic customer finds an empty queue, can be numerically computed from the Markov chain model. For the given data we have $F_S(0) = 0.313611$ and the distribution function $F_S(t)$ is shown in Fig. 1-(b). A derivation of the above formula is provided in the Appendix. The formula can be also derived from the results in Tagaki (2014).

We observe the variables Q, Q_A and Q_S. In this simulation we have counted 1,752 abandons and 8,248 times-to-service. See in Fig. 1-(c,d) the functions $\bar{G}_A(t)$ and $\bar{G}_S(t)$. Note that $\bar{G}_S(0) > 0$ because many customers in this simulation find an empty queue.



(a) $F_A(t)$ (b) $F_S(t)$

(c) $\bar{G}_A(t)$ (d) $\bar{G}_S(t)$

(e) $\bar{F}_A(t)$ vs $F_A(t)$ (f) $\bar{F}_S(t)$ vs $F_S(t)$

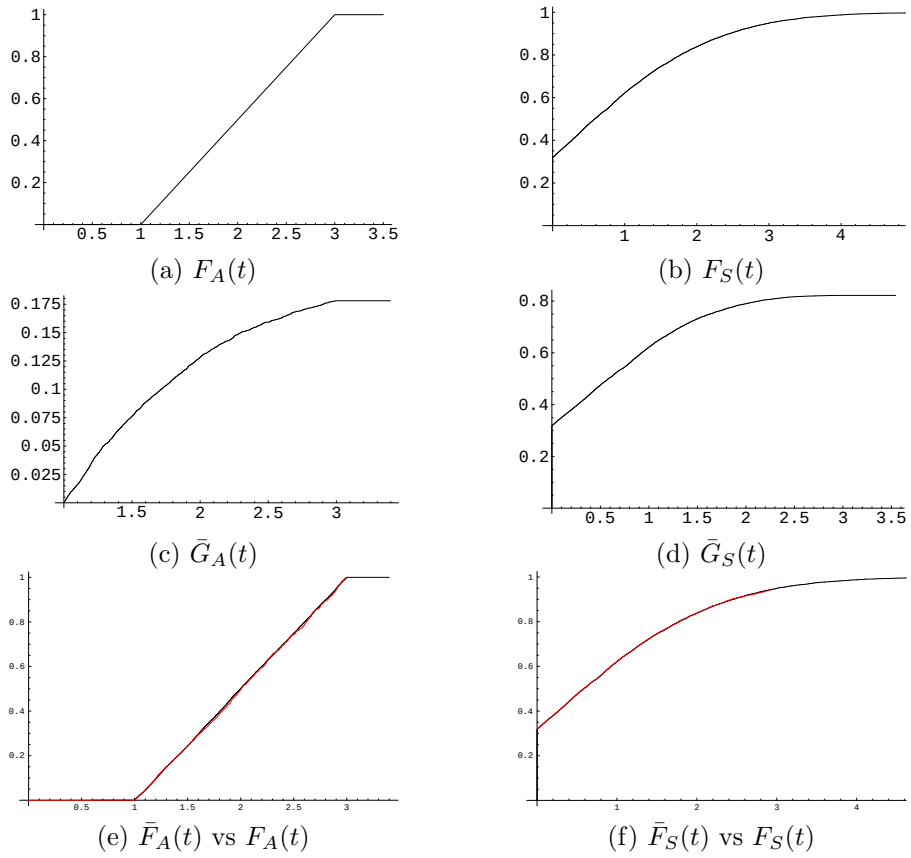
Figure 1: Exponential abandon times

The functions $\bar{F}_A(t)$ and $\bar{F}_S(t)$ we compute by the described procedure are shown in Fig. 1-(e,f) in red plotted against the known values in black (shown also on Fig. 1-(a,b) on a different scale).

We note a striking accuracy for the computed function $\bar{F}_S(t)$, while the accuracy for $\bar{F}_A(t)$ is good for low values of time but is decreasing with time. The reason is clear. For larger times we have less data available for a faithful reconstruction of $F_A(t)$.

Furthermore, the accuracy in the estimation empirically shows the validity of the assumption of stochastic independence between the abandon time and the times-to-service of a particular costumer.

The second example is again a queue $M/M/1$ with $\lambda = 1$, $\mu = 1.2$ and $n = 10,000$ customers. The abandon times are uniformly distributed between a minimum value equal to 1 and a maximum value equal to 3, so that $T_A = 3$.



(a) $F_A(t)$ (b) $F_S(t)$

(c) $\bar{G}_A(t)$ (d) $\bar{G}_S(t)$

(e) $\bar{F}_A(t)$ vs $F_A(t)$ (f) $\bar{F}_S(t)$ vs $F_S(t)$

Figure 2: Uniform abandon times

(see Fig. 2-(a)). In this case an analytic expression for $F_S(t)$ is not available and we use the empirical distribution function that can be inferred from the simulation. This empirical $F_S(t)$ is shown in Fig. 2-(b). Note that $T_S = +\infty$.

In this simulation we have counted 1,781 abandons and 8,219 times-to-service. See in Fig. 2-(c,d) the functions $\bar{G}_A(t)$ and $\bar{G}_S(t)$. The functions $\bar{F}_A(t)$ and $\bar{F}_S(t)$ we compute by the described procedure are shown in Fig. 2-(e,f) in red plotted against the known values in black (shown also on Fig. 2-(a,b) on a different scale)

In this case both $F_A(t)$ and $F_S(t)$ are accurately reconstructed up to $T_A = 3$. Clearly no reconstruction is possible for $F_S(t)$ for $t \geq 3$ because there are no data for $t \geq 3$. The good approximation of $F_A(t)$ on the whole range $[0, T_A]$, in contrast to the previous example, is due to the availability of many data up to T_A.

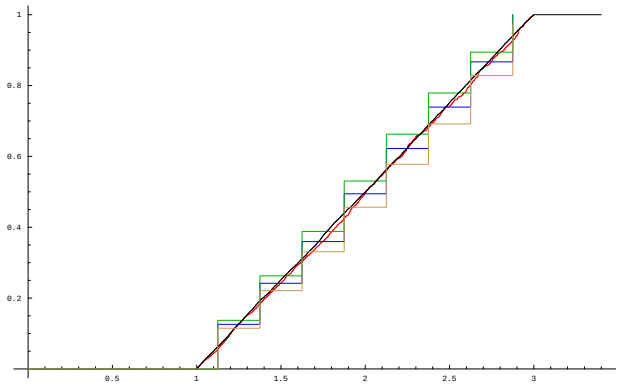


Figure 3: Different reconstructions of $\bar{F}_A(t)$

Now we use this second example to compare the three different factors appearing in (27) for the case of tied data. To this aim, we divide the time into consecutive intervals of length 0.25 and round all the time values that fall in a same interval to the same value, so that we loose the original ordering of these events. We report in Fig. 3 the three step-wise functions $\bar{F}_A(t)$ computed as explained in Section 5 according to the three factors. In particular the top function (in green) corresponds to the case of considering first the times-to-service and then the abandons. The bottom function (brown) corresponds to the opposite case, and the middle function (blue) is obtained by our formula. The almost straight function in black is the 'true' abandon distribution function and the one in red is the abandon distribution function reconstructed by using the tools of Section 4. The difference can be easily perceived. We have also computed the mean squared errors of the three aggregated distribution functions with respect to the reconstructed function (the red line). These are the mean squared errors: $2.1434 \cdot 10^{-3}$ for our formula, $2.49861 \cdot 10^{-3}$ for the formula obtained by putting first the abandons and then the times-to-service, $2.32769 \cdot 10^{-3}$ for the formula obtained by putting first the times-to-service and then the abandons. So on this example our formula provides a better estimate.

7 Conclusions

We have proposed an approach to compute the distribution functions of the abandon times and the times-to-service for a call center queueing system. To the best of our knowledge, the approch is new and leads to an estimator that coincides with the Kaplan-Meier estimator in the case of observed untied data but is different in the case of tied data. Experimental simulations seem to

validate the proposed formulas.

8 Appendix

Consider a queue $M/M/1$ with fixed values λ and μ and abandon times that are exponential with rate α, i.e., $F_A(t) = 1 - e^{-\alpha\, t}$. We want to find an expression $F_S(t)$. This particular queue with abandons is equivalent to a $M/M/1$ queue modeled as a Markov chain with transition rates $\mu + k\,\alpha$ from state $k+1$ to state k and λ from k to $k+1$. The stationary probabilities π_k of the states $k = 0, 1, \ldots$, can therefore be computed numerically with negligible error from

$$\pi_k\, \lambda = \pi_{k+1}\, (\mu + k\, \alpha) \tag{28}$$

In particular we can compute $\pi_0 = F_S(0)$.

If a customer enters the system and finds k customers, we have to compute its time-to-service by allowing only the other waiting customers to abandon the queue. Hence the time-to-service is the sum of k exponential random variables, with rates μ, $\mu + \alpha$, $\mu + 2\,\alpha$, $\ldots$, $\mu + (k-1)\,\alpha$. Let $F_S^k(t)$ be the distribution function of this sum. Note that $F_S^0(t) = 1$ for $t \geq 0$. Then

$$F_S(t) = \sum_{k \geq 0} \pi_k\, F_S^k(t).$$

Let $\tilde{F}_S(t) = \sum_{k \geq 1} \pi_k\, F_S^k(t)$. Since the Laplace transform of the density of an exponential r.v. with rate μ is $\mu/(\mu + s)$, the Laplace transform of $dF_S^k(t)/dt$ is

$$\prod_{h=0}^{k-1} \frac{\mu + h\,\alpha}{\mu + h\,\alpha + s},$$

and the Laplace transform of $d\tilde{F}_S(t)/dt$ is

$$\sum_{k \geq 1} \pi_k \prod_{h=0}^{k-1} \frac{\mu + h\,\alpha}{\mu + h\,\alpha + s}.$$

We want to compute coefficients $A_{h,k}$ such that

$$\prod_{h=0}^{k-1} \frac{\mu + h\,\alpha}{\mu + h\,\alpha + s} = \sum_{h=0}^{k-1} A_{h,k}\, \frac{\mu + h\,\alpha}{\mu + h\,\alpha + s} \tag{29}$$

from which we have

$$F_S^k(t) = \sum_{h=0}^{k-1} A_{h,k}\, (1 - e^{-(\mu + h\,\alpha)\, t})$$

and therefore

$$F_S(t) = \pi_0 + \sum_{k\geq 1} \pi_k \sum_{h=0}^{k-1} A_{h,k}\,(1-e^{-(\mu+h\,\alpha)\,t}) = \pi_0 + \sum_{h\geq 0}(1-e^{-(\mu+h\,\alpha)\,t}) \sum_{k\geq h+1} \pi_k\, A_{h,k}.$$

Let

$$B_h := \sum_{k\geq h+1} \pi_k\, A_{h,k}$$

So we may write

$$F_S(t) = \pi_0 + \sum_{h\geq 0} B_h\,(1 - e^{-(\mu+h\,\alpha)\,t})$$

To compute the coefficients $A_{h,k}$ we use the standard tool of multiplying both sides of (29) by $\mu + j\,\alpha + s$ and substituting $s = -\mu - j\,\alpha$. We get

$$A_{j,k} = \prod_{h=0,h\neq j}^{k-1} \frac{\mu + h\,\alpha}{(h-j)\,\alpha}.$$

However, the coefficients $A_{j,k}$ grow quickly with k and their direct computation becomes numerically unstable. It is therefore convenient to work directly on the terms $\pi_k\, A_{h,k}$ that vanish as k grows. From

$$A_{j,k+1} = A_{j,k}\,\frac{\mu + k\,\alpha}{(k-j)\,\alpha}, \qquad k > j, \qquad \pi_{k+1} = \pi_k\,\frac{\lambda}{\mu + k\,\alpha},$$

we have

$$\pi_{k+1}\, A_{j,k+1} = \pi_k\, A_{j,k}\,\frac{\lambda}{(k-j)\,\alpha}, \qquad k > j,$$

from which we have

$$\pi_{h+r}\, A_{h,h+r} = \pi_{h+1}\, A_{h,h+1}\,\frac{\lambda^{r-1}}{(r-1)!\,\alpha^{r-1}}, \qquad r \geq 1.$$

This leads to

$$B_h = \pi_{h+1}\, A_{h,h+1} \sum_{r\geq 1} \frac{\lambda^{r-1}}{(r-1)!\,\alpha^{r-1}} = \pi_{h+1}\, A_{h,h+1}\, e^{\lambda/\alpha}.$$

Hence the only $A_{h,k}$ coefficients that need to be computed are

$$A_{h,h+1} = \prod_{j=0}^{h-1} \frac{\mu + j\,\alpha}{(j-h)\,\alpha} = \frac{(-1)^h}{h!\,\alpha^h} \prod_{j=0}^{h-1} (\mu + j\,\alpha).$$

From (28) we have

$$\pi_{h+1} = \frac{\lambda^{h+1}}{\prod_{j=0}^{h}(\mu + j\,\alpha)}\,\pi_0$$

and therefore

$$B_h = \frac{\lambda^{h+1}}{(\mu + h\,\alpha)}\,\frac{(-1)^h}{h!\,\alpha^h}\,e^{\lambda/\alpha}\,\pi_0$$

and finally

$$F_S(t) = \pi_0\,\big(1 + e^{\lambda/\alpha}\sum_{h\geq 0}\frac{\lambda^{h+1}}{(\mu + h\,\alpha)}\,\frac{(-1)^h}{h!\,\alpha^h}\,(1 - e^{-(\mu+h\,\alpha)\,t})\big).$$

As a check, we have for $\alpha \to 0$:

$$f_S(t) = \pi_0\,\delta(t) + \pi_0\,e^{\lambda/\alpha}\sum_{h\geq 0}\lambda^{h+1}\,\frac{(-1)^h}{h!\,\alpha^h}\,e^{-(\mu+h\,\alpha)\,t}) =$$

$$\pi_0\,\delta(t) + \pi_0\,\lambda\,e^{-\mu\,t}\,e^{\lambda/\alpha}\sum_{h\geq 0}\lambda^h\,\frac{(-1)^h}{h!\,\alpha^h}\,(e^{-\alpha\,t})^h) =$$

$$\pi_0\,\delta(t) + \pi_0\,\lambda\,e^{-\mu\,t}\,e^{\lambda/\alpha}e^{-\lambda\,e^{-\alpha\,t}/\alpha}$$

By using the asymptotically valid approximation $e^{-\alpha\,t} = 1 - \alpha\,t$ and the known value for the M/M/1 queue with constant λ and μ, $\pi_0 = 1 - \lambda/\mu$, we get the known formula

$$f_S(t) = (1 - \frac{\lambda}{\mu})\,\delta(t) + \frac{\lambda}{\mu}(\mu - \lambda)\,e^{-(\mu-\lambda)\,t}.$$

The only term that has to be numerically computed in the expression for $F_S(t)$ is π_0. It has to be stressed that the computation is numerically unstable for low values of α because the B_h coefficients grow as α tends to zero and alternate in sign. For the data of the first example in Section 6 we obtain

$B_0 = 38.7866$, $B_1 = -166.228$, $B_2 = 363.625$, $B_3 = -538.703$,
$B_4 = 606.041$, $B_5 = -550.947$, $B_6 = 420.862$, $B_7 = -277.491$,
$B_8 = 161.044$, $B_9 = -83.5044$, $B_{10} = 39.1427$, $B_{11} = -16.7455$,
$B_{12} = 6.58968$, $B_{13} = -2.4011$, $B_{14} = 0.814658$, $B_{15} = -0.258622$,
$B_{16} = 0.0771456$, $B_{17} = -0.0217034$, $B_{18} = 0.00577752$, $B_{19} = -0.00145958$

References

Aalen, O.: Nonparametric inference in connection with multiple decrement models. Scandinavian Journal of Statistics 3, 15–27 (1976).

Aalen, O., Borgan, O. & Gjessing H.K.: Survival and event history analysis: a process point of view, Springer - New York (2008).

Aksin, Z., Ata, B., Emadi, S.M., & Su, C-L: Structural estimation of callers' delay sensitivity in call centers. Management Science 59, 2727–2746 (2013).

Avramidis, A. N., Chan, W., Gendreau, M., L'Ecuyer, P., & Pisacane, O.: Optimizing daily agent scheduling in a multiskill call center. European Journal of Operational Research 200, 822–832 (2010).

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., & Zhao, L.: Statistical analysis of a telephone call center: a queueing-science perspective. J. Amer. Statist. Assoc. 100, 36–50 (2005).

Conley, Q. D.: Simulating abandonment using Kaplan-Meir survival analysis in a shared billing and claims center. in: R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl (eds.) Proceedings of the 2013 Winter Simulation Conference, 1805-1817 (2013).

Cox, D.R., & Oakes, D.: Analysis of Survival Data, Chapman and Hall (1984).

Gans N., Koole G., & Mandelbaum, A.: Telephone call centers: Tutorial, review, and research prospects. Manufacturing & Service Operations Management 5, 79–141 (2003).

Ibrahim, R., & Whitt, W.: Wait-time predictors for customer service systems with time–varying demand and capacity. Operations Research 59, 1106–1118 (2011).

Kaplan, E.L., & Meier,P.: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assoc. 53, 457–481 (1958).

Mandelbaum, A., & Momčilović, P.: Queues with many servers and impatient customers. Mathematics of Operations Research 37, 41–65 (2012).

Mandelbaum, A., & Zeltyn, S.: Data-stories about (im)patient customers in tele-queues. Queueing Syst 75, 115-146 (2013).

Nelson, W.: Theory and applications of hazard plotting for censored failure data. Technometrics 14, 945–966 (1972).

Rinne, H., The rate: Theory and inference (with supplementary MATLAB-Programs), visited 2016, `http://geb.uni-giessen.de/geb/volltexte/2014/10793/` (2014).

Tagaki, H.: Waiting time in the $M/M/m/(m+c)$ queue with impatient customers. International Journal of Pure and Applied Mathematics 90, 519–559 (2014).