Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 949726, 5 pages
http://dx.doi.org/10.1155/2014/949726

Research Article

Computing the Expected Cost of an Appointment Schedule for Statistically Identical Customers with Probabilistic Service Times

Dennis C. Dietz

CenturyLink, Inc., Boulder, CO 80301, USA

Correspondence should be addressed to Dennis C. Dietz; dennis.dietz@centurylink.com

Received 29 August 2013; Accepted 13 November 2013; Published 30 January 2014

Academic Editors: E. K. Aydogan, D. Oron, and M. D. Toksari

A cogent method is presented for computing the expected cost of an appointment schedule where customers are statistically identical, the service time distribution has known mean and variance, and customer no-shows occur with time-dependent probability. The approach is computationally efficient and can be easily implemented to evaluate candidate schedules within a schedule optimization algorithm.

1. Introduction

Consider a medical service system where a fixed number of patients are to be scheduled for appointments with a single provider over a finite time horizon [1]. The common service time distribution for every patient is known, as is the time-dependent probability that any patient will fail to show up for the appointment. During the time each arriving patient waits for service, an institutional waiting cost of c_w per hour is incurred. Expected waiting times can be kept small by planning large gaps between appointments, but the gap sizes are constrained by a finite provider availability period. To ensure that all patients are served, the availability period can be extended at an institutional overtime cost of c_o per hour. The objective is to determine a schedule of patient arrival times that will minimize the expected total cost (waiting and overtime) of operating the appointment system.

Similar problems could arise in many other operational contexts. For example, a maritime shipping entity may need to optimally schedule vessel dockings within a fixed window of rented dock time (overtime penalties for docks can be quite severe). A surgical suite manager may need to determine a schedule for procedures that minimizes expected waiting times for surgical teams while avoiding intrusion on a subsequent high-priority commitment. Analogous to medical environments, providers in legal, financial, or other personal service professions that operate on an appointment basis are usually concerned with both server efficiency and client waiting times.

Appointment optimization has received substantial attention in operations research literature, originating with a brief reference by Herne in 1951 in the discussion following Kendall's important article on queueing theory [2]. Numerous attempts followed to obtain reasonable scheduling rules through simulation modeling, beginning with Bailey [3]; relevant surveys are presented in Magerlein and Martin [4], Przasnyski [5], and Ho et al. [6]. Analytic modeling efforts began with steady-state approximation of appointment systems, where statistically identical customers could be scheduled to arrive on a continuous time horizon [7–9]. Transient models and solution methods were later addressed [10–14]. The mathematical complexity of optimization in continuous time led other researchers to constrain arrivals to discrete time points (e.g., a 15-minute lattice) [15, 16]; operationally, this approach was more realistic for most applications. Service time distributions were assumed to be identical and follow exponential, Erlang, or Coxian probability distributions. Vanden Bosch and Dietz [17] generalized the service time model and proposed a heuristic sequencing algorithm for statistically nonidentical customers. Kaandorp

and Koole [18] subsequently derived an exact sequencing and scheduling algorithm for customers having exponentially distributed service times with distinct mean values.

Consistent with the standard practice of scheduling appointments on a lattice of time slots with fixed width Δ, assume that each of N customers is appointed to enter the service queue at one of K discrete times $0, \Delta, 2\Delta, \ldots, (K-1)\Delta$. All customers arrive exactly on time, and the server works under a first-come-first-served (FCFS) discipline whenever one or more customers are in the system. Let a_k be the number of appointments scheduled at time $(k-1)\Delta$, and define a *schedule* as a vector $S = (a_1, a_2, \ldots, a_K)$ such that $\sum_{k=1}^{K} a_k = N$. Any single component of S has a feasible range of 0 to N. For example, a trivial problem where $c_w = 0$ and $c_o > 0$ clearly yields an optimal schedule of $S = (N, 0, 0, \ldots, 0)$.

The first customer should always be scheduled in the first time slot, since a later arrival would waste server time with no improvement in waiting times. The total number of schedules that must be considered is the number of ways in which the remaining $N-1$ appointments can be assigned to K time slots, which can be computed as $\binom{N+K-2}{N-1}$. Hence, a typical problem involving 20 customers and 60 time slots would generate $\binom{78}{19} = 6.71 \times 10^{17}$ candidate schedules. In general, the large number of candidate schedules will prohibit an optimality search by exhaustive enumeration. Assuming that a method exists for determining the total cost $C(S)$ of any schedule, optimality can be efficiently achieved using the algorithm below [1].

(1) Determine a schedule that is assured to have a lower cost than all earlier schedules, S_E, using the following procedure.

   (a) Establish an early incumbent schedule S_E. If no better bound is available, let $S_E = (N, 0, 0, \ldots, 0)$.

   (b) Let m be the largest integer for which the mth arriving customer in S_E is not scheduled at time $(K-1)\Delta$.

   (c) Establish a candidate early schedule S by shifting the arrival of the mth arriving customer in S_E by Δ later, unless this shift causes the order of customer arrivals to change. If all customers but the first are scheduled at $(K-1)\Delta$, stop (recall that the first customer's arrival is fixed at zero).

   (d) If $C(S) < C(S_E)$, let $S_E = S$ and return to step 1(b).

   (e) If $m > 2$, decrement m and return to step 1(c). Otherwise, each customer of the current S_E but the first has shifted without improvement, and S_E is established.

(2) Establish a candidate late schedule S_L by shifting each arrival time by Δ, if feasible. Perform a parallel procedure to step (1) (shifting arrivals by Δ earlier rather than later).

(3) If S_E and S_L differ, the optimal arrival time of each customer is that defined either by S_E or by S_L. Enumerate each of the possible intermediate schedules and evaluate their costs to find the optimum.

Although this algorithm is NP-hard due to the enumeration sometimes required by step (3), S_E and S_L often coincide in practice and seldom differ by more than a few customer arrivals. Computation time is roughly linearly dependent on N and K in the coincidental case, but the algorithm always requires the evaluation of numerous candidate schedules. Computational efficiency is therefore dependent on a cogent and efficient method for computing the expected cost of a specified appointment schedule, which is the contribution offered by this paper.

2. Service Time Model

Suppose that the customer service times are independent, identically distributed random variables with known mean θ and variance σ^2, so the coefficient of variation is $\xi = \sigma/\theta$. For $\xi \leq 1$, the service time distribution can be modeled as a mixture of Erlang$(\mu, r-1)$ and Erlang(μ, r) distributions with density function

$$f(t) = \alpha \frac{\mu^{r-1} t^{r-2}}{(r-2)!} e^{-\mu t} + (1-\alpha) \frac{\mu^r t^{r-1}}{(r-1)!} e^{-\mu t}, \quad t \geq 0, \tag{1}$$

where $0 \leq \alpha \leq 1$. When the parameters r, α, and μ are chosen such that

$$r = \left\lceil \frac{1}{\xi^2} \right\rceil,$$
$$\alpha = \frac{\xi^2 - \left\{ r\left(1+\xi^2\right) - r^2\xi^2 \right\}^{1/2}}{1+\xi^2}, \tag{2}$$
$$\mu = \frac{r-\alpha}{\theta},$$

the distribution will have the required mean and variance (see Tijms [19, page 358]). This model is desirable, since $f(t)$ is always unimodal and is similar in shape to the commonly occurring gamma density function.

For the less typical situation where $\xi > 1$, we can resort to modeling the service time distribution as a mixture of Erlang$(\mu, 1)$ and Erlang(μ, r) distributions with density function

$$f(t) = \alpha\mu e^{-\mu t} + (1-\alpha) \frac{\mu^r t^{r-1}}{(r-1)!} e^{-\mu t}, \quad t \geq 0, \tag{3}$$

where $0 \leq \alpha \leq 1$. In this case, the required mean and variance can be realized when

$$r = \max\left\{2, \min\left\lceil k : \frac{k^2+4}{4k} \geq \xi^2 \right\rceil\right\},$$
$$\alpha = \frac{2r\xi^2 + r - 2 - \left(r^2 + 4 - 4r\xi^2\right)^{1/2}}{2(r-1)\left(1+\xi^2\right)}, \tag{4}$$
$$\mu = \frac{\alpha + r(1-\alpha)}{\theta}.$$

This approach may be reasonable and useful, but should be applied with caution if distribution characteristics beyond the mean and variance are known. Simulation studies have demonstrated that appointment schedule performance is generally insensitive to higher order moments of the service time distribution [20], but these studies are limited to cases where $\xi \leq 1$.

3. Cost Computation

With the chosen service time model, we can compute the expected total cost of a schedule by exploiting the memoryless property of the exponential distribution. The service times are conceptually comprised of exponentially distributed service phases, each with mean duration $1/\mu$. Hence, the state of the system at any time can be completely described by the number of unfinished phases remaining in the system. Let $t_k = (k-1)\Delta$, $t_k^- = \lim_{\delta \to 0}(t_k - \delta)$, and $t_k^+ = \lim_{\delta \to 0}(t_k + \delta)$. Since each phase completion is an event within a Poisson processs, the probability that s phases of service remain at t_k^- given v remains at t_{k-1}^+ can be written as

$$p(s \mid v) = \begin{cases} \dfrac{e^{-\mu\Delta}(\mu\Delta)^{v-s}}{(v-s)!}, & 0 < s \leq v, \\ 1 - \displaystyle\sum_{m=0}^{v-1} \frac{e^{-\mu\Delta}(\mu\Delta)^{m}}{m!}, & s = 0, \\ 0, & \text{otherwise}, \end{cases} \tag{5}$$

for $k = 2, \ldots, K+1$. Now, let $A_k = \sum_{j=1}^{k} a_j$ and let $q_k(s)$ be the probability that s phases remain at t_k^-. For notational convenience, define a binomial operator $b(i, a, \pi) = \binom{a}{i}\pi^i(1-\pi)^{a-i}$, $0 \leq i \leq a$, $0 \leq \pi \leq 1$, and let

$$z = \begin{cases} 1, & \xi \leq 1, \\ r - 1, & \xi > 1. \end{cases} \tag{6}$$

By conditioning on the number of scheduled arrivals that enter the system with fewer than r service phases (i.e., with $r-1$ phases when $\xi \leq 1$ or with a single phase when $\xi > 1$), we have

$$q_2(s) = \sum_{i=0}^{a_1} b(i, a_1, \alpha)\, p(s \mid ra_1 - zi), \quad s = 0, \ldots, ra_1, \tag{7}$$

and, by recursion,

$$q_k(s) = \sum_{i=0}^{a_{k-1}} b(i, a_{k-1}, \alpha) \times \sum_{m=\max(0, s - ra_{k-1} + zi)}^{rA_{k-2}} q_{k-1}(m)\, p(s \mid ra_{k-1} - zi + m), \quad k = 3, \ldots, K+1, \quad s = 0, \ldots, rA_{k-1}. \tag{8}$$

Expected total cost is then given by

$$C = c_w \left(\sum_{k=1}^{K} \frac{a_k(a_k - 1)\theta}{2} + \sum_{k=2}^{K} \sum_{s=1}^{rA_{k-1}} q_k(s) \frac{a_k s}{\mu} \right) + c_o \left(\sum_{s=1}^{rA_K} q_{K+1}(s) \frac{s}{\mu} \right). \tag{9}$$

The first summation in (9) is equivalent to $\sum_{k=1}^{K} \sum_{j=1}^{a_k - 1} a_j \theta$ and accounts for waiting due to multiple arrivals scheduled in the same time slot.

The method can be easily extended to accommodate probabilistic customer no-shows. Let γ_k be the probability that a customer actually appears for an appointment scheduled in slot k. By conditioning on the number of customers showing up, (7) can be rewritten as

$$q_2(s) = \sum_{j=0}^{a_1} b(j, a_1, \gamma_1) \sum_{i=0}^{j} b(i, j, \alpha)\, p(s \mid rj - zi), \quad s = 0, \ldots, ra_1, \tag{10}$$

and (8) can become

$$q_k(s) = \sum_{j=0}^{a_{k-1}} b(j, a_{k-1}, \gamma_{k-1}) \sum_{i=0}^{j} b(i, j, \alpha) \times \sum_{m=\max(0, s - rj + zi)}^{rA_{k-2}} q_{k-1}(m)\, p(s \mid rj - zi + m), \quad k = 3, \ldots, K+1, \ s = 0, \ldots, rA_{k-1}. \tag{11}$$

Expected total cost is then computed as

$$C = c_w \left(\sum_{k=1}^{K} \sum_{j=2}^{a_k} b(j, a_k, \gamma_k) \frac{j(j-1)\theta}{2} + \sum_{k=2}^{K} \sum_{j=1}^{a_k} b(j, a_k, \gamma_k) \sum_{s=1}^{rA_{k-1}} q_k(s) \frac{js}{\mu} \right) + c_o \left(\sum_{s=1}^{rA_K} q_{K+1}(s) \frac{s}{\mu} \right). \tag{12}$$

The cost computation procedure can be embedded in an appointment optimization algorithm for statistically identical customers. Unlike more sophisticated and generalized approaches [17], the distribution parameters are trivially calculated and cost computation requires no matrix exponentiation. Schedule optimization can thus be easily performed on a personal computer with common software such as a macroenabled spreadsheet. It should be noted that the cost evaluation of an individual candidate schedule may not require complete computation of C; summation and supporting calculations can be terminated as soon as the partial cost of the candidate schedule exceeds the cost of the incumbent.

Table 1: Effect of service time variability.

ξ	r	Optimal schedule	Total cost	Schedules evaluated	Execution time (sec)
0.125	64	(1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0)	1.4072	309	15
0.250	16	(1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0)	2.7861	304	2
0.500	4	(1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0)	6.7935	287	1
1.000	1	(2, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 0)	15.9581	253	1
1.500	9	(2, 1, 1, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 0)	25.2274	217	1
2.000	16	(2, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0)	32.8035	189	2

4. Implementation and Performance

Now consider a specific appointment scheduling problem where the number of statistically identical customers is $N = 10$, the number of time slots is $K = 16$, the time slot width is $\Delta = 0.5$, the mean service time is $\theta = 0.75$, the service time variance is $\sigma^2 = 0.25$, the show probability is $\gamma_k = 0.95, k = 1, \ldots, 16$, and the cost of server overtime is estimated at ten times the cost of customer waiting (we notionally set $c_w = 1$ and $c_o = 10$, since only relative values affect the optimal schedule). Because $\xi = 0.6667 < 1$, the applicable service time model is given by (1) and the associated parameter values are computed as $r = 3$, $\alpha = 0.5234$, and $\mu = 3.3022$. Imbedding the cost computation method within the optimization algorithm described above yields an optimal schedule of

$$S = (1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 0). \quad (13)$$

The associated waiting, overtime, and total costs are 4.8603, 4.9541, and 9.8144, respectively. Optimality is achieved after evaluation of 287 candidate schedules, and complete execution requires less than one second of processing time on a personal computer with a 2.26 GHz processor.

Table 1 illustrates the effect of service time variability by parameterizing on ξ. The table quantifies the intuitive positive relationship between variability in service time and the expected cost of the optimal schedule. The number of schedules evaluated diminishes as ξ increases, although computation times are longer for very high or low values of ξ due to the larger number of exponential phases r in the associated service time models.

To further exercise the modeling approach, we can enlarge the baseline problem (with $\xi = 0.6667$) to schedule $N = 50$ customers into $K = 80$ time slots. The cost of the optimal schedule is 51.8026, which is obtained after evaluating 34,162 schedules. Execution time for this very large problem is 483 seconds, which equates to about 0.0141 seconds per schedule evaluated.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] P. M. Vanden Bosch and D. C. Dietz, "Minimizing expected waiting in a medical appointment system," *IIE Transactions*, vol. 32, no. 9, pp. 841–848, 2000.

[2] D. G. Kendall, "Some problems in the theory of queues," *Journal of the Royal Statistical Society B*, vol. 18, no. 2, pp. 151–185, 1951.

[3] N. T. J. Bailey, "A study of queues and appointment systems in hospital outpatient departments, with special reference to waiting times," *Journal of the Royal Statistical Society B*, vol. 14, no. 2, pp. 185–199, 1952.

[4] J. M. Magerlein and J. B. Martin, "Surgical demand scheduling: a review," *Health Services Research*, vol. 13, no. 4, pp. 418–433, 1978.

[5] Z. H. Przasnyski, "Operating Room Scheduling," *AORN Journal*, vol. 44, no. 1, pp. 67–82, 1986.

[6] C.-J. Ho, H.-S. Lau, and J. Li, "Introducing variable-interval appointment scheduling rules in service systems," *International Journal of Operations and Production Management*, vol. 15, no. 6, pp. 59–68, 1995.

[7] W. R. van Voorhis, "Waiting-line theory as a management tool," *Operations Research*, vol. 4, no. 2, pp. 221–228, 1956.

[8] P. M. Morse, *Queues, Inventories, and Maintenance: the Analysis of Operational Systems with Variable Demand and Supply*, Wiley, New York, NY, USA, 1963.

[9] B. Jansson, "Choosing a good appointment system—a study of queues of the type D/M/1," *Operations Research*, vol. 14, no. 2, pp. 292–312, 1966.

[10] G. R. Grape, "Convergence and cost minimization in queuing systems of the type (D,M,1)," *Försvarets Forskningsanstalt*, vol. 2, no. 1, pp. 1–6, 1968.

[11] B. E. Fries and V. P. Marathe, "Determination of optimal variable-sized multiple-block appointment systems," *Operations Research*, vol. 29, no. 2, pp. 324–345, 1981.

[12] C. D. Pegden and M. Rosenshine, "Scheduling arrivals to queues," *Computers and Operations Research*, vol. 17, no. 4, pp. 343–348, 1990.

[13] P. P. Wang, "Static and dynamic scheduling of customer arrivals to a single-server system," *Naval Research Logistics*, vol. 40, no. 3, pp. 345–360, 1993.

[14] P. P. Wang, "Optimally scheduling N customer arrival times for a single-server system," *Computers and Operations Research*, vol. 24, no. 8, pp. 703–716, 1997.

[15] C.-J. Liao, C. D. Pegden, and M. Rosenshine, "Planning timely arrivals to a stochastic production or service system," *IIE Transactions*, vol. 25, no. 5, pp. 62–73, 1993.

[16] P. M. Vanden Bosch, D. C. Dietz, and J. R. Simeoni, "Scheduling customer arrivals to a stochastic service system," *Naval Research Logistics*, vol. 46, no. 5, pp. 549–559, 1999.

[17] P. M. Vanden Bosch and D. C. Dietz, "Scheduling and sequencing arrivals to an appointment system," *Journal of Service Research*, vol. 4, no. 1, pp. 15–25, 2001.

[18] G. C. Kaandorp and G. Koole, "Optimal outpatient appointment scheduling," *Health Care Management Science*, vol. 10, no. 3, pp. 217–229, 2007.

[19] H. C. Tijms, *Stochastic Models: An Algorithmic Approach*, Wiley, West Sussex, UK, 1994.

[20] C. Ho and H. Lau, "Minimizing total costs in scheduling outpatient appointments," *Management Science*, vol. 38, no. 12, pp. 1750–1764, 1992.