Case study 3: Waiting in line

Stephen Childress

April 27, 2005

1 Background

We will look at the problem of modeling queues. A typical problem is illustrated by a police roadblock on a one-lane highway. Cars arriuve at infrequent intervals and are stopped, inspected or drivers questioned, then waived on. The question is, how long a delay can be expected when you arrive at the roadblock. Obviously the answer depends upon the rate of arrive of cars and the time the police spend on inspecting each car. Depending on the circumstances, both the arrival rate and the inspection times can fluctuate wildly. It is not at all clear based say on the average times, just how long one is likely to have to wait. Presumably the waiting time can also fluctuate considerably.

Another example is teller service at a bank. Most banks now have a single line for multiple tellers, which must therefore be a better solution than letting a queue form at each window. We should be able to model these cases and understand the difference. The processing of transmissions along a communication network presents similar problems.

In these notes we will look at the simplest example of a queuing model, comprising a single server and a single line waiting to be served.

2 A deterministic flow model

Suppose that a line is served at a rate of μ customers per unit time, and customers arrive at a rate of λ customers per unit time. We will allow these numbers to be arbitrary real numbers (as in our population models), so we are thinking of the "unit" of customers as being a large number. Suppose that initially no one is in line, and customers begin arriving. At time t the number in the line is therefore

$$n = \begin{cases} t(\lambda - \mu), & \text{if } \lambda > \mu, \\ 0, & \text{if } \mu > \lambda. \end{cases} \tag{1}$$

Assume that if this number exceeds N, no one will join the line. The *waiting time* of a customer in the queue when $\lambda > \mu$ is

$$T = \begin{cases} t(\lambda - \mu)/\mu, & \text{if } t(\lambda - \mu) < N, \\ K/\mu, & \text{if } t(\lambda - \mu) \geq N. \end{cases} \tag{2}$$

If there are M servers with a line for each, we may assume, since the number of customers are large, that they will at any time make the lines of equal length at each server, so that λ/M will be the arrival rate for each server. Assuming that $\lambda/M > \mu$ we get a waiting time of

$$T = \begin{cases} t(\lambda/M - \mu)/\mu, & \text{if } t(\lambda/M - \mu) < N, \\ N/\mu, & \text{if } t(\lambda/M - \mu) \geq N. \end{cases} \tag{3}$$

If we have a single line for M servers, then the effective rate of service of the line is $M\mu$. Also the customer will be willing to tolerate a longer line, of MN customers. Thus in the two cases, above, we get $t(\lambda - \mu M)/(\mu M)$ and $MN/\mu M$ respectively, so the same waiting times result.

If $\mu > \lambda$ the single server flow model predicts zero waiting time. Also for multiple lines or servers we have just seen that the model finds no distinction between the two. Neither of these properties correctly reflects the true situation, due to the fact that the numbers involved in realistic queues are not that large. Waiting for 6 people to be served at a bank can be time consuming and frustrating. The flow model essentially says that either a line grows indefinitely (assuming incoming customers always join it), or else it has zero length. In practice we are willing to tolerate a short wait and the servers we a queuing for would presumably be set up so that *in a steady state* the incoming customers are served reasonably promptly but there can still be a short line. In other words, in real life the situation is not one of queues of zero or infinite length. It is the need to handle the "in-between" cases that forces us to abandon the simple flow model.

Suppose that we try to track the individuals in this flow model. Let a customer arrive at a server every $\frac{1}{\lambda}$ units of time. If $\mu > \lambda$ then the person gets served in a time $1/\mu < 1/\lambda$ and so the line is emptied before the next customer arrives. If we count the customer being served as part of the line (we always do this), then we can say that either 1 person or no persons are in the line. This is clearly not what actually happens, even when the service is rapid. The *fluctuations* in arrival times means that in general when you arrive you can expect to find one or more customers already in line. There will also be stretches of time when no customers have arrived, and so the servers are doing nothing. In a sense the waiting times we experience even with adequate servers are paying the price of the time when the servers were idle.

Also it must be admitted that it is wrong to assume that every customer can be served in the same amount of time. There are fluctuations of this time, which can either mean that the service is faster than $1/\mu$, so that the server waits longer for the next customer, or else the service is longer and a line begins to form.

Clearly what is missing in the simple flow model are the fluctuations, both in the arrival times and the service times. We cannot predicts these in a deterministic fashion. Thus the only possible model which allows us to predict the properties of a queue is one which is *stochastic*. In a stochastic model the formulation explicitly introduces the randomness of the phenomenon, and the only predictions that are attempted are based upon *probabilities*.

3 The Poisson process

Before attempting to model a single server we consider the theory necessary to talk about the fluctuations in times of arrival or service. Focusing on arrival times, we can think of a server whose station is closed and suppose that the line is simply building up. We are interested in computing $p_n(t)$ = the probability of n customers arriving within the time interval $0, t$. Of course there is no one answer to this. It could well be that customers arrive at exactly one minute intervals. What we are trying to model is the innumerable small causes that make the arrival of customers completely unpredictable from one instant to the next. The model we adopt here, the *Poisson process*, is based upon two essential ideas. (1) If we divide time up into little intervals of length Δt, then the arrival times of customers in two distinct intervals are completely independent. There is absolutely no influence of the events of one interval on the events within another. (2) Within any such small interval, it is very unlikely that *two* customers arrive. Many such intervals will have zero arrivals, but occasional we will have one arrival. It then makes sense to speak of the probability of one arrival occuring within any such interval.

Note that (1) does not have to be true. For example potential customers may look at the line before deciding if they want to stop. Thus their arrival time is affected by what has happened in a previous interval. Also (2) may not be realistic in some situations. The server may be accessed by an elevator which releases potential customers in bunches of ten or more, all of which arrive at the same time.

But you must agree that the Poisson process is reasonable for a great many situations where customers intent on being served are arriving at times determined only by their own actions, times which have nothing to do with the server or line already present.

We formalize these ideas in the following assumptions:

(i) There is a positive real number λ such that the probability that an arrival occurs between times t and $t + \Delta t$ is

$$\lambda \Delta t + o(\Delta t). \tag{4}$$

(Here $o(\Delta t$ denotes a quantity which, when divided by Δt, forms a quotient which tends to zero as Δt tends to zero, i.e. it "vanishes faster than Δt as $\Delta t \to 0$.)

(ii)The probability that more than one customer arrives in time Δt is $o(\Delta t)$.

(iii) The number of arrivals in non-overlapping time intervals are statistically independent, i.e. the probabilities associated with one are not dependent upon the probabilities associated with the other. According to the theory of probability, (iii) will allow us to compute the probability of two events occuring together on two separate time intervals by multiplying the two separate probabilities, and the probability of either occurring by summing the two probabilities.

Using these assumptions, we now want to show that

$$p_n(t) = \frac{1}{n!}(\lambda t)^n e^{-\lambda t}. \tag{5}$$

To prove (5), first note that the probability that *no* arrivals occured within any time interval of width Δt is $1 - \lambda\Delta t + o(\Delta t)$. We now want to use this fact to compute the probability that no arrivals occur within the interval $t, t + \Delta t$. We consider the non-overlapping time intervals $0, t$ and $t, t + \Delta t$. The events of having no arrivals in these two intervals are independent by (iii) above, so the probability of both happening is obtained by multiplying the two probabilities. The probability that no arrivals occurred up to t is $p_0(t)$, while that of no one arriving in the short interval is, as we have seen, $1-\lambda\Delta t$ up to a quantity $o(\Delta t)$. Thus

$$p_0(t + \Delta t) = p_0(t)(1 - \lambda\Delta t) + o(\Delta t) \tag{6}$$

Rearranging and dividing by Δt we have

$$\frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \frac{o(\Delta t)}{\Delta t}. \tag{7}$$

Taking the limit $\Delta t \to 0$ we obtain

$$\frac{dp_0}{dt} = -\lambda p_0(t). \tag{8}$$

Now clearly $p_0(0) = 1$, so the solution is

$$p_0(t) = e^{-\lambda t}. \tag{9}$$

We have thus established (5) for $n = 0$.

Let's consider the analogous computation of $p_1(t)$, the probability that 1 customer arrives in the interval $0, t$. We want to compute $p_1(t + \Delta t)$. This quantity has an "either-or else" part, namely either one arrives in $0, t$ and none in $t, t + \Delta t$, or else none arrives in $0, t$ and one in $t, t + \Delta t$. In this case we *add* the probabilities of each event. Now examine the first of these, one in the long interval, none in the short. The probability of both of these independent events is the product $p_1(t)(1 - \lambda\Delta t)$ plus $o(\Delta t)$. Similarly for the second we get the product $p_0(t)\lambda\Delta t$. Thus

$$p_1(t + \Delta t) = p_1(t)(1 - \lambda\Delta t) + p_0(t)\lambda\Delta t. \tag{10}$$

Again arranging , dividing by Δt, and taking the limit, we arrive at

$$\frac{dp_1}{dt} = -\lambda p_1 + \lambda p_0 = -\lambda p_1 + \lambda e^{-\lambda t}. \tag{11}$$

Clearly $p_1(0) = 0$. The solution may be obtained using an integrating factor:

$$\frac{de^{\lambda t}p_1}{dt} = -\lambda, \tag{12}$$

so $p_1 = \lambda t e^{-\lambda t} + Ce^{-\lambda t}$. The initial condition implies $C = 0$ and so

$$p_1(t) = \lambda t e^{-\lambda t}. \tag{13}$$

This establishes (5) for $n = 1$.

Now let's proceed by induction. Assume $p_n = \frac{1}{n!}(\lambda t)^n e^{-\lambda t}$, we want to show that $p_{n+1} = \frac{1}{(n+1)!}(\lambda t)^{n+1} e^{-\lambda t}$. Proceding as we did for $n = 1$, we see that

$$\frac{dp_{n+1}}{dt} = -\lambda p_{n+1} + \lambda p_n. \tag{14}$$

We easily get, using $p_{n+1}(0) = 0$, $p_{n+1} = \frac{1}{(n+1)!}(\lambda t)^{n+1} e^{-\lambda t}$ as required.

4 Some properties of the Poisson process

We show several of the $p_n(t)$ for the Poisson process in figure 1.

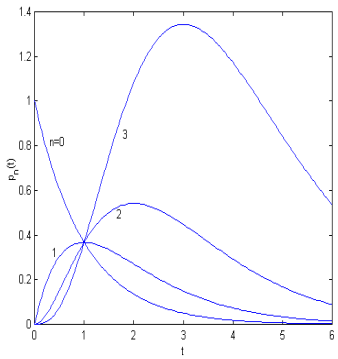

Figure 1. $p_n(t), n = 0, 1, 2, 3$, for the Poisson process with $\lambda = 1$

How can we use these functions to describe the process of arrival of customers? We first ask, what is the probability of one person being in the line at time t, assuming that no one was in the line at time 0? We must add up some possibilities. First, in the interval $0, \Delta t$ the probability of one arrival was $\lambda \Delta t$. So the first possibility is that the new arrival occurred in the first time interval. The next possibility is that the first person arrived in the *second* time interval. The probability that no one arrived in the first time interval is $e^{-\lambda \Delta t}$ and the probability that one arrival occured in the second time interval is again $\lambda \Delta t$. the product of these give the second possibility. Continuing in this way, we are in effect computing the integral

$$\int_0^t \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t}. \tag{15}$$

Let the waiting time for the first customer be T. We have just seen that the probability of one person being in the line at time t is given by (15). This is the same thing as the probability that one person arrived sometime in $0, t$, or that the waiting time for the first person did not exceed t.

Thus we can write

$$Prob[T \leq t] = 1 - e^{-\lambda t} = \int_0^t \lambda e^{-\lambda t} dt, \tag{16}$$

Notice that in the last integral we have the function $f = \lambda e^{-\lambda t}$. This function has the property that $\int_0^\infty f(t)dt = 1$.

Definition: The function $f(t) = \lambda e^{-\lambda t}$ is called the *distribution function of waiting times*.

We also remark that

$$Prob[T > t] = e^{-\lambda t}. \tag{17}$$

Do you see why this is true?

We now ask, what is the *expected* waiting time for the first customer? Suppose the first customer arrives in the interval $t, t + \Delta t$. The probability of this is $p_0(t)\lambda\Delta t$. The contribution of this to the expected waiting time is t times this probability, or $tp_0(t)\lambda\Delta t$. Adding up all such intervals, we get the expected waiting time as

$$\int_0^\infty t\lambda e^{-\lambda t} dt = \frac{1}{\lambda}. \tag{18}$$

This gives us a satisfying and useful interpretation of the parameter λ. it has the dimensions of an inverse time, and we have just found that this time is the expected waiting time for the first customer in a line.

5 Simulation of a Poisson process

Consider the buildup of a line. At $t = 0$ no one is in line. At $t = T_1$ a person arrives. At $t = T_1 + T_2$ another person arrives, and so one. The $T_i's$ are thus waiting times from the arrival of one customer to the arrival of the next. The quantities $y_i = 1 - e^{-\lambda T_i}$ are the probabilities of these waiting times, with $0 \leq y_i \leq 1$. Suppose we look at the y_i lying in some interval $0, a \leq 1$. Since a is a probability of the waiting time being $\leq a$, the fraction of y_i in $0, a$ must be proportional to a. In other words to simulate a Poisson process on a computer all we need to do is choose random numbers which are *uniformly distributed* in the interval $0, 1$, From these points y_i we compute the waiting times T_i from $y_i = 1 - e^{-\lambda T_i}$

Here is the result of 50 tries with $\lambda = 1/30$:

• 7.8853
28.0064
19.9649
66.5746
43.0767

18.2900
• 0.5603
51.6794
17.6476
28.6691
47.0974
76.4595
40.2061
•5.8172
15.6114
82.2186
74.6329
15.8427
67.2305
• 1.7890
13.0562
50.3261
• 0.2973
• 4.4860
• 6.7982
• 6.6464
27.7745
• 9.5314
• 6.6499
• 0.4618
41.2056
17.6688
80.5657
18.8205
16.2720
56.1672
22.3428
• 6.7937
33.4548
54.6267
• 0.5950
34.3030
14.3160
53.4773
20.9637
37.0816
16.8053
10.8988
• 6.3088

I have put a bullet next to all the waiting times ≤ 10. There are 14 occurences in the 50 tries, giving the numerical estimate of $Prob[T \leq 10]$ to

be 14/50=.28. The exact value for this probability in a Poisson process is $1-e^{-1/3} = .2835$, so we can see that it doesn't take many tries to se the theory borne out.

6 The single server as a Poisson process

We now suppose that the service times also constitute a Poisson process. The the probability that one customer is fully served in the time interval $t, t+\Delta t$ is $\mu\Delta t$. Let $P_k(t)$ = probability that there are k people in the line at time t. The equation satisfied by P_k is a bit complicated but its pieces are easy to understand. Here it is:

$$P_k(t+\Delta t) = P_k(t)(1-\lambda\Delta t)(1-\mu\Delta t) + P_k(t)(\lambda\Delta t)(\mu\Delta t)$$
$$+P_{k+1}(t)(\mu\Delta t)(1-\lambda\Delta t) + P_{k-1}(1-\mu\Delta t)(\lambda\Delta t). \tag{19}$$

The first term on the right gives the probability that k customers were in the line at time T and that no one either arrived or was served in the interval Δt. The next term is the probability that there were k in the line at time t and that one person arrived and one was served in the time interval Δt. The remaining terms can be similarly described. These are "either-or" probabilities so they must be added.

Now if we neglect terms which are $o(\Delta t)$ in (19) things get simpler:

$$P_k(t+\Delta) = P_k[1-(\lambda+\mu)\Delta t] + P_{k+1}\mu\Delta t + P_{k-1}\lambda\Delta t. \tag{20}$$

Rearranging and taking the limit $\Delta t \to 0$ we have

$$\frac{dP_k}{dt} = -(\lambda+\mu)P_k + \mu P_{k+1} + \lambda P_{k-1}. \tag{21}$$

These eequations hold for $k = 1, 2, \ldots$. When computing the equation for P_0 we have to be careful. nThye only possibilities are that either there was no one in line at time t and no one arrived in the following interval Δt, or else one person was in the line an was served. This leads to

$$\frac{dP_0}{dt} = -\lambda P_0 + P_1\mu. \tag{22}$$

6.1 The steady-state queue

Let us assume that we have a steady state where the probabilities are all independent of time. Then from (22) we have $P_1 = (\lambda/\mu)P_0$. For this steady state (21) yields the second-order difference equation

$$\mu P_{k+1} - (\lambda+\mu)P_k + \lambda P_{k-1} = 0. \tag{23}$$

We see that a solution of (23) is given by

$$P_k = C\rho^k,\ \rho = \lambda/\mu < 1. \tag{24}$$

(There is in fact another solution, namely $P_k = C' =$ another constant, which must be disregarded because all of the $P_k's$ must sum to 1, see below.) Note that we are assuming that expected service time is less that the expected waiting time. This is the only situation where we can expect there to exist a steady state. Otherwise the line just keeps growing.

From $P_1 = (\lambda/\mu)P_0$ we see that $C = P_0$ in (24). Now since the probability of some number ≥ 0 of persons being in the line is 1, we have

$$\sum_0^\infty P_k = 1 = \sum_0^\infty P_0 \rho^k, \tag{25}$$

yielding by summation of the geometric series,

$$P_0 = 1 - \rho. \tag{26}$$

Thus we have our probabilities in terms of λ, μ:

$$P_k = \rho^k (1 - \rho). \tag{27}$$

6.2 Some properties of the queue

What is the expected length of the queue? This is given by

$$L = \sum_0^\infty k P_k = \sum_0^\infty k \rho^k (1 - \rho). \tag{28}$$

To sum this, note that

$$\frac{d}{d\rho} \sum_0^\infty \rho^k = \sum_1^\infty k \rho^{k-1} = \frac{d}{d\rho} \frac{1}{1-\rho} = \frac{1}{(1-\rho)^2}. \tag{29}$$

Thus

$$L = \frac{\rho}{(1-\rho)^2}(1-\rho) = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu - \lambda}. \tag{30}$$

Problem: Suppose we want the expected number of customers in the line actually waiting to be served. Since k in the system means that $k-1$ are actually waiting to be served, show that this is given by

$$\sum_1^\infty (k-1) P_k = \frac{\rho^2}{1-\rho}. \tag{31}$$

Since we are dealing with a steady state, and the expected number of people in the queue is $\frac{\lambda}{\mu - \lambda}$, while waiting in line we will see an average of λT new customers arrive, where T is the average waiting time in the line. Since we are in a steady state we must have $\frac{\lambda}{\mu - \lambda} = \lambda T$, so that

$$T = \frac{1}{\mu - \lambda}. \tag{32}$$

An example: A bank teller finds that the probability of serving a customer within a 1 minute time interval is 1/6. If time is measure in hours, $\Delta t = 1/60$, so $\mu = 10$. The probability of a new customer arriving in the interval $t, t+$ one minute is observed to be 1/12, so $\lambda = 5$. Thus the average waiting time in the line is 1/5 or 12 minutes.