QUEUEING THEORY WITH APPLICATIONS AND SPECIAL CONSIDERATION TO EMERGENCY CARE

JAMES KEESLING

1. Introduction

Much that is essential in modern life would not be possible without queueing theory. All communication systems depend on the theory including the Internet. In fact, the theory was developed at the time that telephone systems were growing and requiring more and more sophistication to manage their complexity. Much of the theory was developed by Agner Krarup Erlang (1878-1929). He worked for *Copenhagen Telephone Company*. His contributions are widely seen today as fundamental for how the theory is understood and applied. Those responsible for the early development of what has become the Internet relied on the work of Erlang and others to guide them in designing this new system. Leonard Kleinrock was awarded the National Medal of Honor for his pioneering work leading to the Internet. His book [7] reworked Queueing Theory to apply to this new developing technology.

These notes are being compiled from a seminar in the Department of Mathematics at the University of Florida on *Queueing Theory Applied to Emergency Care*. The seminar meets weekly and the material is updated based on the presentations and discussion in the seminar. In the notes an attempt is made to introduce the theory starting from first principles. We assume some degree of familiarity with probability and density functions. However, the main distributions and concepts are identified and discussed in the notes along with some derivations. We will try to develop intuition as well. Often the intuition is gained by reworking a formula in a way that the new version brings insight into how the system is affected by the various parameters.

If one makes complete rigor the principal goal in Mathematics, then the subject can become extremely dry. We will be content to be sure that the results are correct and any arguments elucidating. We will focus on the insights that might be relevant to the various applications that we have in mind.

We will first introduce *Poisson processes*. This forms the basic underpinning of elementary queueing theory. Next we will introduce the *simple queue*. This is a queueing system with a single server with Poisson arrivals and exponential service times. We then discuss more complex queueing systems. As we introduce new ideas we will try to give applications and hint how the ideas will apply to emergency care. The general applications will range from telephone communications to stochastic modeling of population dynamics and other biological systems. The most complex queueing systems are frequently beyond mathematical analysis. This is likely the case for a realistic model of emergency care. Such cases will be studied by simulation. The tools of simulation will be gradually developed through the notes. One of the major accomplishments of the seminar is a realistic model of the flow of patients in the emergency room.

There are several texts that we recommend on the subject of queueing theory. The book by Donald Gross, John Shortle, James Thompson, and Carl Harris, *Fundamentals of Queueing Theory* [5] is recommended for those involved in this project. Some others that may be consulted are *Probability, Markov Chains, Queues, and Simulation: the Mathematical Basis of Performance Modeling* by William Stewart [9], *An Introduction to Queueing Theory and Matrix-Analytic Methods* by Lothar Breuer and Dieter Baum [2], *Queueing Theory and Telecommunications: Networks and Applications* by Giovanni Giambene [4], *Optimal Design of Queueing Systems* by Shaler Sticham, Jr. [6], and *Elements of Queueing Theory, Palm Martingale Calculus and Stochastic Recurrences* by François Baccelli and Pierre Brémaud [1]. Efficiency and reduced waiting times are only part of the management concerns of emergency care. A valuable general guide to operational improvement for emergency departments is given in [3].

As mentioned, this material is compiled from a seminar that meets regularly to study the subject of *Queueing Theory Applied to Emergency Care*. Here is a picture of the participants at our meeting on October 25, 2012.

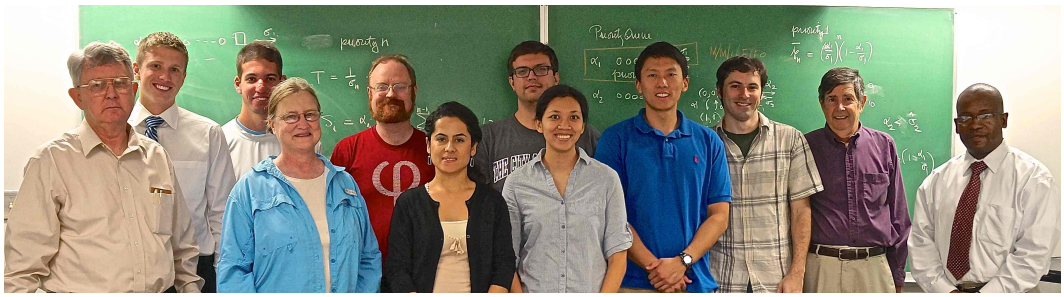

FIGURE 1. Emergency Care/Queueing Seminar: (Left to Right) Jed Keesling, Trent Register, Joshua Hurwitz, Jean Larson, James Maissen, Hayriye Gulbudak, Evan Milliken, Jo Ann Lee, David Zhou, Scott McKinley, Lou Block, Adrian Tyndall (Head of Emergency Services at Shands)

2. Poisson Processes

Randomness is hard to recognize. If points are thought to be random on a line or in the plane, then we would be suspicious if the points were regular distances apart. On the other hand, if the points are truly random and independent, then there will appear to be clustering. We will see why this is so when we cover exponential waiting times in the next section. Arrivals in queueing theory are assumed to be random and independent, but at some given rate. This is a Poisson process. We first give the axioms for a Poisson process which intuitively describe a process in which the events are random and independent. In the following let $\alpha > 0$ be a real constant.

Definition 2.1. *A* **Poisson process** *is a random sequence of events such that the following three axioms hold.*

(1) *The probability of an event occurring in a small interval of time* Δt *is given by*

$$Pr = \alpha \cdot \Delta t + o(\Delta t).$$

(2) *If I and J are disjoint intervals, then the events occurring in them are independent.*

(3) *The probability of more than one event occurring in an interval Δt is $o\left(\Delta t^2\right)$.*

From these axioms one can derive properties of the distribution of events. The first formula we will derive is the probability of exactly k events occurring in an interval of length t. To derive the formula we divide the interval into n equal subintervals each of length $\frac{t}{n}$. For large enough n, the probability of an event in each of these intervals is approximately $\frac{\alpha t}{n}$. The probability that an event will not occur in a particular subinterval is $1-\frac{\alpha t}{n}$. Thus, the binomial probability for k occurrences is given by

$$B(n,k) = \left(\begin{array}{c} n \\ k \end{array} \right) \left(\frac{\alpha t}{n}\right)^k \cdot \left(1-\frac{\alpha t}{n}\right)^{n-k}.$$

Using the fact that $\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a$ one can easily calculate the limit of the above.

$$\lim_{n\to\infty} B(n,k) = \frac{(\alpha t)^k}{k!}\exp(-\alpha t)$$

This calculation tells us the probability of k occurrences in an interval of length t. The probabilities for $k = 0, 1, 2, \ldots$ give a Poisson distribution. The average number in the interval is $\overline{k} = \alpha t$ and the variance is $\sigma^2 = \alpha t$. For large αt this is approximately a normal distribution. Below is a graph of the Poisson distribution for $\alpha t = 5, 10,$ and 20.

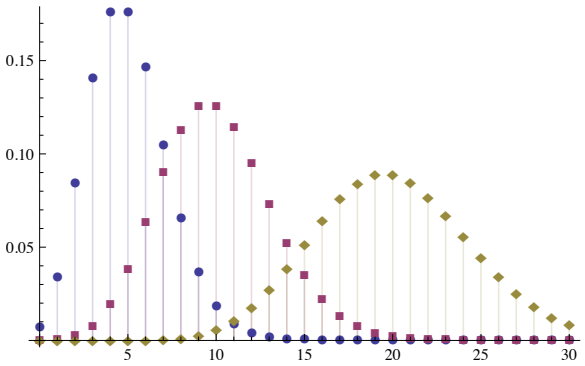

FIGURE 2. Poisson Distributions

3. Exponential Waiting Times

A Poisson process is equivalent to points being placed on a line by a stochastic process such that the distribution of distances between the points are independent from the density function $\alpha\exp(-\alpha t)$. If we think of the line as being time and the events as occurring at certain times, the density function is called the *exponential waiting time* with rate α or average waiting time $\frac{1}{\alpha}$. The average waiting time for this density function is $\frac{1}{\alpha}$ and the variance is $\frac{1}{\alpha^2}$.

Suppose that we have a Poisson process as described in §2. Suppose that we want to know the density function for the time to the next event. What is the probability that the next event will occur between t and $t+\Delta t$? For the time T to the next event to lie between t and $t+\Delta t$, there

must not be an event in the interval $[0, t]$. This is the probability of 0 events in the interval of length t which is $\frac{(\alpha t)^0}{0!}\exp(-\alpha t) = \exp(-\alpha t)$. In addition, there must be an event in the interval $[t, t+\Delta t]$. The probability of this is approximately $\alpha\Delta t$. So, by the independence of events in these two intervals, the probability that both would occur is approximately $\alpha\Delta t \cdot \exp(-\alpha t)$. Dividing by Δt and taking the limit leads us to the density function, $\alpha\exp(-\alpha t)$. So, a Poisson process has exponential waiting times. It is easy to show that these are independent since events in disjoint intervals are independent.

On the other hand, it can also be shown that if we have independent exponential waiting times between events, it will be a Poisson process. It is a good test of your understanding to prove this.

Exercise 3.1. *Suppose that points are distributed on a line with the intervals between being independent exponential waiting times with parameter α. Show that the points come from a Poisson process with rate α.*

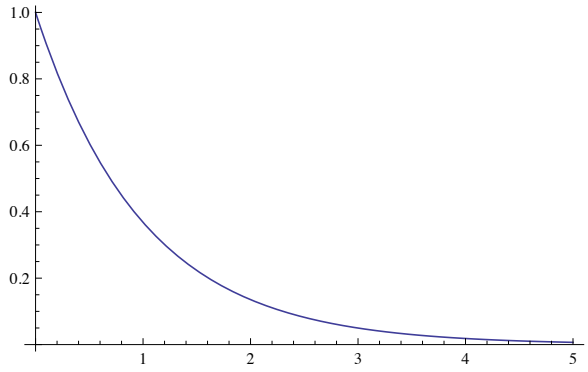

FIGURE 3. Exponential Distribution with $\alpha = 1$

Note that the distribution implies that short times are more frequent than long times since the probability density function is highest at zero. A longer time is less frequent than a shorter time since the function is decreasing. It is also helpful to have the cumulative distribution function $F(t)$. This is the function such that the probability of the time T to the next event being less than t is given by $F(t)$. Clearly, $F(t) = \int_0^t \alpha\exp(-\alpha\tau)d\tau = 1 - \exp(-\alpha t)$. The cumulative distribution function is useful for simulating waiting times. First we give the graph of the cumulative distribution function with $\alpha = 1$.

The next step is to figure how to simulate independent exponential waiting times. First we suppose that we have a method of generating a sequence of independent random numbers from the uniform distribution on $[0, 1]$. Such random number generators are usually included with any modern programming language. We will not go into the theory of how to generate such numbers or test that they are independent. We will take on faith that it can be done and that the program available will do that. So, suppose that we have a finite sequence of such numbers $\{u_i\}_{i=1}^n$. With each of these numbers u_i we associate a time from the exponential waiting time $\alpha\exp(-\alpha t)$. The method is straightforward, simply solve $u_i = F(t_i)$. Since $F(t)$ is monotone increasing, there will be a unique t_i for each u_i. In the case at hand, $u_i = 1 - \exp(\alpha t)$. Solving we get

$$t_i = -\frac{\ln(1-u_i)}{\alpha}.$$

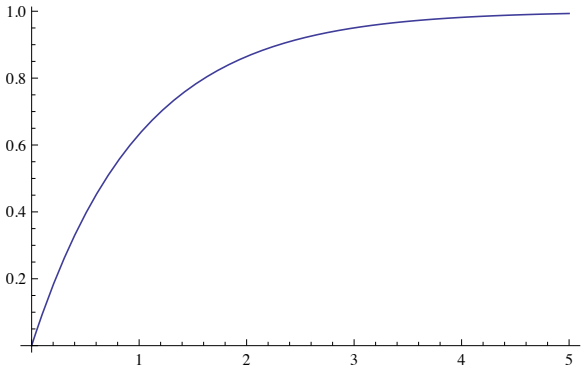

FIGURE 4. Cumulative Distribution of the Exponential Distribution with $\alpha = 1$

Since $1 - u_i$ also comes from a uniform distribution on $[0, 1]$, we could replace $1 - u_i$ by u_i in the formula. Doing so would reduce the computation by one floating point operation: $t_i = -\frac{\ln(u_i)}{\alpha}$.

Figure 5 gives a randomly generated list of twenty random numbers u_i followed by the associated exponential waiting times with $\alpha = 1$.

1	0.288238	1.24397
2	0.939854	0.0620311
3	0.620859	0.476652
4	0.901637	0.103544
5	0.30665	1.18205
6	0.254485	1.36851
7	0.406778	0.899489
8	0.760417	0.273889
9	0.643364	0.441045
10	0.22541	1.48984
11	0.321007	1.13629
12	0.664928	0.408076
13	0.686137	0.376678
14	0.0231776	3.76457
15	0.677097	0.38994
16	0.0368309	3.30142
17	0.914729	0.0891275
18	0.37808	0.97265
19	0.220679	1.51105
20	0.776203	0.253341

FIGURE 5. Twenty Independent Exponential Waiting Times with $\alpha = 1$

1	0.367842	2.26853
2	0.0418932	0.132377
3	0.598283	−3.13514
4	0.0238941	0.0752067
5	0.75493	−0.969491
6	0.925518	−0.238359
7	0.517458	−18.2144
8	0.60534	−2.91062
9	0.973564	−0.0832437
10	0.365848	2.23058
11	0.712502	−1.26848
12	0.606356	−2.88065
13	0.182457	0.6455
14	0.848425	−0.515772
15	0.785122	−0.800524
16	0.750969	−0.993933
17	0.992575	−0.0233298
18	0.708066	−1.3055
19	0.39754	2.99864
20	0.719243	−1.21466

FIGURE 6. Twenty Independent Samples from $\frac{1}{\pi} \cdot \frac{1}{1+x^2}$

Simulation of Poisson Process

In[1]:= $n = 10; \alpha = 1;$

In[2]:= $M = \text{Table}[\text{RandomReal}[], \{i, 1, n\}]$

Out[2]= {0.684511, 0.471398, 0.20682, 0.508947, 0.668194, 0.00381792, 0.315233, 0.241728, 0.0614914, 0.709117}

In[3]:= $T = -\text{Log}[M]/\alpha$

Out[3]= {0.379051, 0.752053, 1.57591, 0.675412, 0.403177, 5.56805, 1.15444, 1.41994, 2.78886, 0.343734}

In[4]:= $\text{T2} = \text{Table}[\{\text{Sum}[T[[i]], \{i, 1, j\}]\}, \{j, 1, n\}];$

In[5]:= $\text{T3} = \text{Table}[\{i - 1, x \le \text{T2}[[i]][[1]]\}, \{i, 1, n\}]$

Out[5]=
$$\begin{pmatrix} 0 & x \le 0.379051 \\ 1 & x \le 1.1311 \\ 2 & x \le 2.70701 \\ 3 & x \le 3.38242 \\ 4 & x \le 3.7856 \\ 5 & x \le 9.35365 \\ 6 & x \le 10.5081 \\ 7 & x \le 11.928 \\ 8 & x \le 14.7169 \\ 9 & x \le 15.0606 \end{pmatrix}$$

In[6]:= $A = \text{Piecewise}[\text{T3}]$

Out[6]=
$$\begin{cases} 0 & x \le 0.379051 \\ 1 & x \le 1.1311 \\ 2 & x \le 2.70701 \\ 3 & x \le 3.38242 \\ 4 & x \le 3.7856 \\ 5 & x \le 9.35365 \\ 6 & x \le 10.5081 \\ 7 & x \le 11.928 \\ 8 & x \le 14.7169 \\ 9 & x \le 15.0606 \end{cases}$$

In[7]:= $\text{Plot}[A, \{x, 0, \text{T2}[[n]][[1]]\}]$

Out[7]=

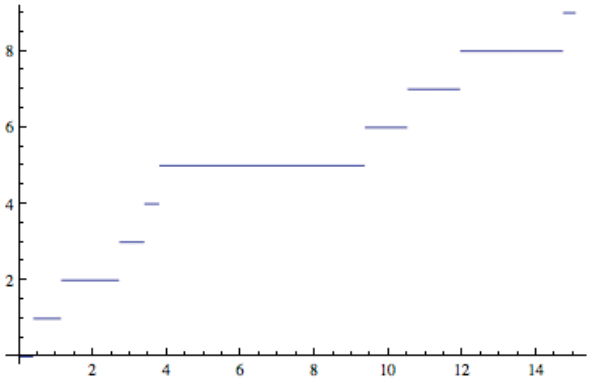

FIGURE 7. *Mathematica* Program for Poisson Simulation

4. Independent Numbers from an Arbitrary Distribution

In the derivation and simulation in the last section, the distribution that we were using was the exponential distribution with probability density function $p(t) = \alpha \exp(\alpha t)$ and cumulative distribution function $F(t) = 1 - \exp(\alpha t)$. However, if we wanted to generate a set of independent random numbers from an arbitrary probability density function, $p(t)$, we would go through the same process using its cumulative distribution function $F(t)$. We would use a set of independent random numbers from the uniform distribution on $[0, 1]$, $\{u_i\}_{i=1}^N$, and then solve for $\{x_i\}_{i=1}^N$ using the equation $u_i = F(x_i)$.

For example, suppose that our probability density function is $p(x) = \frac{1}{\pi} \cdot \frac{1}{1+x^2}$. Then the cumulative distribution is $F(x) = \frac{1}{\pi} \cdot \arctan(x)$. As before we generate twenty independent numbers from this distribution using random numbers from $[0, 1]$. Figure 5 gives a list numbers generated from this distribution.

In the next few sections, we will assume exponential exponential waiting times. The reason is to be able to make certain calculations. When we derive the *Kolmogorov-Chapman differential equations*, we need the processes to only depend on the most recent state. This is called the *Markov* property for a stochastic process. If the waiting time to leave a given state is exponential, then the Markov property will be satisfied. When this assumption is not appropriate, we may have to resort to simulation. So, the ability to generate independent values from a general distribution will be valuable tool.

Figure 6 is a program in *Mathematica* that simulates a Poisson process. The plot is $n(t)$ the number of events that have occurred in the interval $[0, t]$. In the simulation, we generate a table of independent exponential waiting times. We use these times as the times between successive events in the Poisson process. The output of the various computations in the simulation are shown to make the program more transparent. In *Mathematica* one could have hidden this output. Later in this document we will be simulating much more complex examples. In the program one can change the number of events n and the rate α.

What are examples of Poisson processes? Radioactive decay is one example. If you take the transformation of one of the atoms in the radioactive sample as an event, then the sequence of these decays will be a Poisson process. In telephone networks one generally assumes that a customer trying to make a call is an event. It is assumed that this is a Poisson process and this assumption is supported in practice. Of course, the rate α will change with the time of day and day of the week. We will be assuming that α is constant, but it is not hard to analyze the case that α is time dependent. The arrival of information packets at a given node in the Internet is assumed to be Poisson. The distances between successive cars in a lane on a highway are sometimes assumed to be independent and exponential.

We can also derive a theory of Poisson processes for events in the plane or space. How would you modify the axioms in these cases? How would the properties that we derived change in these settings? Can you think of a way of simulating a Poisson process in a region of the plane? In three space? In R^n?

An example of a Poisson process in the plane would be the location of a species of plant whose seeds disperse widely from the parent plant. The location of stars in a star chart down to a given magnitude will be Poisson. The three dimensional location of stars is Poisson. However, the density α varies through space.

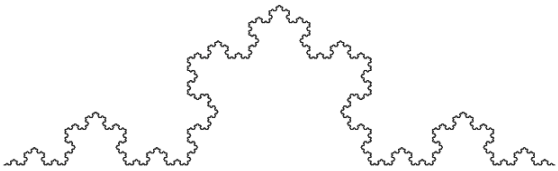

FIGURE 8. The von Koch Curve

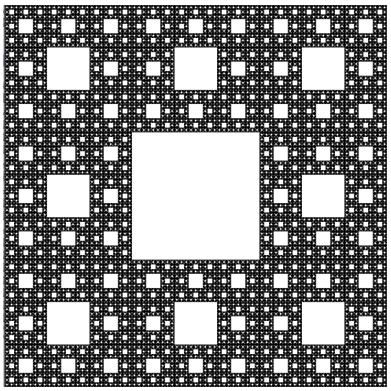

FIGURE 9. The Sierpinski Carpet

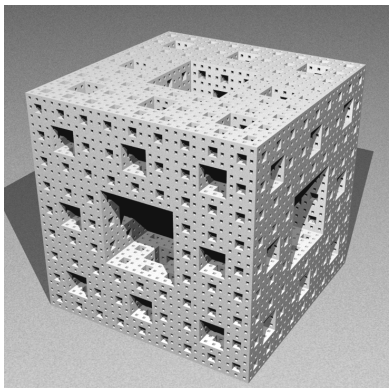

FIGURE 10. The Menger Sponge

One can also define Poisson processes on fractal sets or other spaces having a regular measure. What do you think would be a good definition? How could you simulate such a Poisson process on the von Koch curve, the Sierpiński carpet, or the Menger sponge? Figures 8, 9, and 10 give pictures of these objects if you are not familiar with them.

5. A Simple Queue

In this section we explain a simple queue. This will illustrate the fundamentals of the theory. There are a number of good references for queueing theory. We recommend *Fundamentals of Queueing Theory* by Donald Gross, John Shortle, James Thompson, and Carl Harris [5].

A simple queue has a single server that can serve one customer at a time. The service time is an exponential waiting time with parameter σ. The service times are assumed to be independent and independent of arrivals. The arrival rate is Poisson with rate α. If a customer arrives and the server is occupied, that customer goes to the end of the waiting line. Customers are served in order of arrival.

There is a notation that has become common in the field that communicates the assumptions being made about a given queueing system. The case described above is denoted $M/M/1/\text{FIFO}$. The first M indicates the assumption about arrivals. In this case they are Markovian (Poisson). The second M indicates the assumption about the service times, that they are exponential waiting times. The third indicator gives the number of servers. The last indicates the *queueing discipline*, that is in what order the waiting customers are served. FIFO indicates that they are served in the order "first in first out".

There are also some helpful diagrams that are used to describe this queue. In the one that follows, the clients are denoted by circles and the service facility by a box. The circle in the box denotes the client that is being served by that server.

Figure 11. Diagram of a Simple Queue

We will let the *states* of the system at a given time to be the number of customers that have arrived and not completed service. The state n will be in the set $\{0, 1, 2, \dots\}$. If the state is $n > 0$, then one customer is being served and $n - 1$ are waiting in line. We can also represent the system in the following way.

We can also represent this type of queue with the following diagram.

$$0 \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} 1 \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} 3 \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} \cdots n \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} n+1 \cdots$$

Let $p_n(t)$ denote the probability that the system is in state n at time t. We will be assuming a set of initial probabilities $\{p_n(0)\}_{n=0}^{\infty}$. We will now describe a set of differential equations that govern the system, the *Kolmogorov-Chapman equations*. Suppose that we have the following probabilities at time t, $\{p_n(t)\}_{n=0}^{\infty}$. How can these change in a small interval of time Δt? Consider the case $n = 0$.

$$p_0(t + \Delta t) \approx \sigma \Delta t \cdot p_1(t) + (1 - \alpha \Delta t) \cdot p_0(t)$$

This leads to the differential equation.

$$\frac{dp_0(t)}{dt} = \sigma p_1(t) - \alpha p_0(t)$$

Similarly we can derive a differential equation for each $n > 0$.

$$\frac{dp_n(t)}{dt} = \sigma p_{n+1} + \alpha p_{n-1}(t) - (\alpha + \sigma)p_n(t)$$

The system of differential equations determines the behavior of the probabilities through time after a time t that the values $\{p_n(t)\}_{n=0}^{\infty}$ are known.

If $\sigma > \alpha$, we can expect that these probabilities will have a limiting value. We call the set of these limiting values the *steady-state* for the system. We can solve for the steady-state by setting the derivatives equal to zero in the Kolmogorov-Chapman equations. In the case of a simple queue with $\sigma > \alpha$, this leads to the following probabilities.

$$\overline{p}_n = \left(\frac{\alpha}{\sigma}\right)^n \cdot \left(1 - \frac{\alpha}{\sigma}\right)$$

The average number in the system $E(n) = \overline{n}$ can be calculated.

$$\overline{n} = \frac{\frac{\alpha}{\sigma}}{1 - \frac{\alpha}{\sigma}}$$

Consider a simple example. Suppose that $\alpha = 9$. This could be nine customer arrivals per hour if the unit of time is an hour. Suppose that $\sigma = 10$. At first glance it would seem that this queueing system should run smoothly. In an hour we would expect nine arrivals. The server could handle ten in the hour. So, he should be able to finish the work and have time for a rest before the next hour. However, this naïve analysis does not take proper account of the randomness of the process. The formula for $\overline{n}$ above shows that the average number in the system is

$$\overline{n} = \frac{\frac{9}{10}}{1 - \frac{9}{10}} = 9.$$

This is much more congested than one would have imagined. A customer arriving at a time when the system is in steady-state would fine nine customers already in the system and would have to wait on average ten times as long as one service time $\frac{1}{\sigma} = \frac{1}{10}$, that is, the average wait would be one hour rather than just six minutes.

Let us imagine that this is an approximate of emergency care and a typical time of treatment is one hour. Then $\sigma = 1$ and $\alpha = \frac{9}{10}$. We have changed the time scale, but this does not change the ratio between α and σ which is still $\frac{9}{10}$. So, $\overline{n} = 9$ and the average time that the patient will be in the emergency care facility is a total of ten hours, ten times as long as the treatment period of one hour.

What this demonstrates is that it is important to have an accurate model of how randomness is affecting the system. If we have not properly taken it into account, the congestion that arises due to randomness will be unexpected and may be overwhelming. In what follows we will be showing that the congestion that is due to randomness can be lowered to a manageable level by a modest additional investment of resources.

6. A. K. Erlang

Historically, Agner Krarup Erlang (1878 - 1929) developed queueing theory to analyze telephone systems. He worked for *Copenhagen Telephone Company* during the period that telephone systems were growing in complexity. Obviously, randomness is an integral part such a system. When ARPANET was being considered, the pioneers of this precursor of the Internet used the queueing theory advanced by Erlang and others to show that the system was feasible. One of those involved in those early days was Leonard Kleinrock [7]. Figure 13 shows him receiving the National Medal of Honor for his pioneering work leading to the Internet. Queueing theory continues to play a vital role in analyzing the functioning of the Internet. Figure 12 is a picture of Erlang available from *Wikimedia Commons*.

Figure 12. Agner Krarup Erlang

FIGURE 13. Leonard Kleinrock Receiving the National Medal of Honor

7. Queue with an Infinite Number of Servers

In practice we will have a finite number of servers. However, in order to determine how many servers will be adequate, it makes sense to assume an infinite number of servers and see if this might be helpful in determining what finite number of servers might be adequate. Since every arrival will have a server available, there will be no waiting line for this system. We will continue to assume Poisson arrivals at a rate α and exponential waiting times with parameter σ. The system can be described in our queueing shorthand as $M/M/\infty$. Below is a diagram of this type of queueing system.

The states of the system are $\{n = 0, 1, 2, \dots\}$. However, now the rates are different as represented by the following diagram.

$$0 \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} 1 \underset{2\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} \cdots \underset{(n-1)\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} n-1 \underset{n\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} n \underset{(n+1)\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} n+1 \cdots$$

The steady-state probability of being in each state n is given by the following formula.

$$\overline{p}_n = \frac{\left(\frac{\alpha}{\sigma}\right)^n}{n!} \cdot e^{-\frac{\alpha}{\sigma}}$$

So, we see the Poisson distribution again, this time in the context of a queueing system with an infinite number of servers.

In applications to communication systems and the Internet, $\frac{\alpha}{\sigma}$ will be large. In that case, the Poisson distribution will be approximately normal with mean $\mu = \frac{\alpha}{\sigma}$ and variance $\sigma^2 = \frac{\alpha}{\sigma}$. In

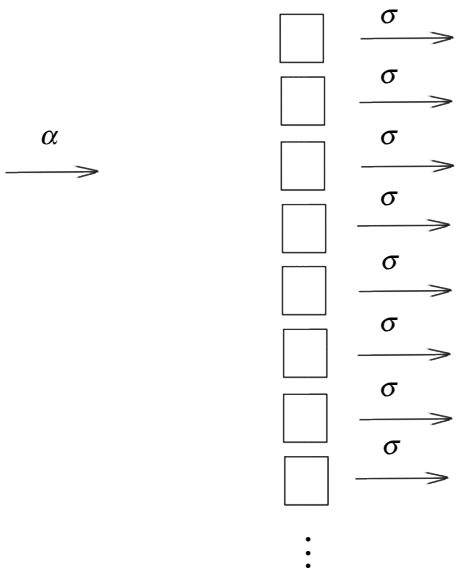

FIGURE 14. Diagram of Queue with an Infinite Number of Servers

practice n will have to be finite. If we choose n to be several, say k, standard deviations beyond the mean $\frac{\alpha}{\sigma}$, then it will be rare that there are more clients than servers.

$$n = \frac{\alpha}{\sigma} + k \cdot \sqrt{\frac{\alpha}{\sigma}}$$

Exercise 7.1. *Suppose that in a certain town there are* $200,000$ *land-line telephones. Suppose that these are used on average seven times per day and each use is equally likely at any time of the day. Assume that the average time of use each time is five minutes and that this is given by an exponential waiting time. Assume that it is unacceptable for there to be greater than* $\frac{1}{1,000,000,000}$ *proportion of the time that a telephone connection cannot be made. If you were designing the system, what should be the capacity for the number of connections in the system?*

Exercise 7.2. *The number of erythrocytes in the human body is approximately* $2 - 3 \times 10^{13}$*. The erythrocytes do not reproduce in the human body and have a life-span of approximately 100-120 days. These cells are produced in bone marrow through a process known as erythropoiesis and then released into the bloodstream. How many erythrocytes are made each day? How long could a person survive if the erythropoietic system were to shut down completely. Assume that a person could barely survive with* 10% *of the normal level of erythrocytes.*

8. Queues with a Finite Number of Servers

Now suppose that there are a finite number of servers. Perhaps that finite number was chosen by the method in the last paragraph of the last section, but we can derive the Kolmogorov-Chapman equations and the steady-state probabilities without knowing just how the number was determined. Let us suppose the there are m servers and that any time the system has more than $n > m$ clients,

the excess $n-m$ clients wait in line for the next available server. This will be an $M/M/m$/FIFO queue. Below is a diagram representing the flow between the states of the system.

$$0 \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} 1 \underset{2\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} \cdots \underset{(m-1)\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} m-1 \underset{m\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} m \underset{m\cdot\sigma}{\overset{\alpha}{\rightleftarrows}} m+1\cdots$$

The steady-state probabilities can be easily computed. We let

$$S = 1 + \frac{\alpha}{\sigma} + \frac{\alpha^2}{2\sigma^2} + \cdots + \frac{\alpha^m}{m!\sigma^m} \cdot \frac{1}{1-\frac{\alpha}{m\sigma}}.$$

If we suppose that $m\sigma > \alpha$, then the steady-state probabilities will exist. In that case the steady-state probabilities will be given by the following.

$$\overline{p}_n = \begin{cases} \frac{\alpha^n}{n!\sigma^n S} & n < m \\ \frac{\alpha^n}{m^{n-m}\cdot m!\sigma^m S} & n \geq m \end{cases}$$

9. Reduction in Waiting Time by Adding a Small Number of Servers

In this section we are interested in determining how much we can reduce the waiting time in the system $M/M/n$/FIFO. Suppose that $m_0 \cdot \sigma \geq \alpha$. If for $m_0\sigma > \alpha$, the system would have an equilibrium value. If $m_0 \cdot \sigma = \alpha$, then the expected waiting time for a client would be infinite. The closer $m_0 \cdot \sigma$ is to α, the longer the waiting time would be. So, there could be enormous congestion for such a queueing system even when m_0 leads to an equilibrium value.

How would this system improve if we added just a few more servers? The formulas for the equilibrium probabilities in §8 do not lend themselves to an easy analysis of this question. However, we are able to give a valuable estimate of the waiting time for $m = m_0 + k$ for $k \geq 1$. In this case the system $M/M/m_0 + k$/FIFO will have an equilibrium value since $(m_0 + k) \cdot \sigma > \alpha$. Using the equilibrium probabilities that were derived in the previous section, the expected time waiting in line for a newly arriving client will be less than $\frac{1}{k\sigma}$. This is independent of the value of m_0. In some applications of this theory m_0 may be enormous and time waiting in line could also be very large. So, just a modest investment – adding a few more servers – could bring the waiting time down to a very tractable level. Even when m_0 is fairly small, this theorem is valuable especially when the servers are highly paid personnel.

In the section to follow, we will provide a proof of this theorem.

In the section after that we show how queueing theory apples to models of population dynamics.

9.1. **Introduction.** In this paper we consider the queueing system known as $M/M/m/FIFO$. It has been studied in standard books of queueing theory. The system consists of a line of customers waiting for being served by a server and m servers. We assume the customer arrivals are random and Poisson with certain rate α, which denotes the average number of customers that come to the system in 1 time unit. The serving times we assume to be independent and exponentially distributed, where σ denotes the average number of customers being served by 1 server in 1 time unit. That makes $\frac{1}{\sigma}$ the average time server needs to serve 1 customer. We should also mention that if the server with exponential service rate σ is kept "busy" all the time, that the output of the server is also a Poisson process of rate σ.

Furthermore, we assume that the customers queue in order of arrival and that the next customer in line goes to the first available server. By the waiting time, t_ω, we mean the time spent in line waiting for an available server. We can control this waiting time by changing the number of servers.

9.2. **Estimated waiting time formula.** The estimated waiting time formula is reached by using Poisson probabilities and can be found in the standard books of queueing theory, as [**?**], but we will derive it here again, to make everything clear and the future proofs more understandable.

Let us assume there are n customers in the system. If $n < m$, where m is the number of servers, all of the customers are being served and the rest of the servers are idle. If $n \geq m$, then m customers are being served by m servers and the rest $(n - m)$ of the customers are waiting in the line. What is the rate of changing the state of the system from 0 customers to 1 customer? It can only happen if someone comes into the system. The rate of this event is α. The same rate α we get for the event changing the state of the system from 1 customer to 2 customers and also from n customers to $n + 1$ customers for all n.

On the other hand, what is the rate of changing the state of the system from 1 customer to 0 customers? It can only happen if there is 1 customer in the system (being served by 1 server), the server finishes serving and customer leaves the system. The rate of this event is σ, which is the average service rate for 1 server. But if we want the rate of changing the state from 2 customers to 1 customer, the rate is now 2σ, because there were 2 customers being served by 2 servers and the service rate for each server is σ. So for $n \leq m$, the rate of changing the state from n to $n - 1$ customers is $n\sigma$. For more than m customers in the system, the number stops growing, because the rate of changing the state from $n + 1$ to n customers (for $n > m$) is $m\sigma$. This is because no matter how many customers are in the system, there are only m working servers and the rate of someone leaving is $m\sigma$.

The situation should be clear from the following diagram:

$$0 \underset{\sigma}{\overset{\alpha}{\rightleftarrows}} 1 \underset{2\sigma}{\overset{\alpha}{\rightleftarrows}} 2 \underset{3\sigma}{\overset{\alpha}{\rightleftarrows}} \cdots \underset{m\sigma}{\overset{\alpha}{\rightleftarrows}} m \underset{m\sigma}{\overset{\alpha}{\rightleftarrows}} m+1 \underset{m\sigma}{\overset{\alpha}{\rightleftarrows}} \cdots$$

Let $\bar{p}_n$ be the steady-state Poisson probability that there are n customers in the system. By steady-state probability we mean the probability, while the system is in equilibrium, which means $\frac{\alpha}{\sigma} < m$. In the opposite case the waiting time is infinite, even for $\frac{\alpha}{\sigma} = m$.

By the diagram, $\alpha\bar{p}_0$ is the probability that the number of customers in the system changes from 0 to 1. Similarly, $\sigma\bar{p}_1$ is the probability for the number of customers to change from 1 to 0. This gives us $\alpha\bar{p}_0 = \sigma\bar{p}_1$, or $\bar{p}_1 = \frac{\alpha}{\sigma}\bar{p}_0$. In the same fashion we get

$$\bar{p}_2 = \frac{\alpha}{2\sigma}\bar{p}_1 = \frac{\alpha^2}{2\sigma^2}\bar{p}_0, \quad \bar{p}_3 = \frac{\alpha^3}{3!\sigma^3}\bar{p}_0, \quad \ldots. \ , \ \bar{p}_n = \frac{\alpha^n}{n!\sigma^n}\bar{p}_0 \quad \forall n \leq m,$$

followed by

$$\bar{p}_{m+k} = \frac{\left(\frac{\alpha}{\sigma}\right)^m}{m!}\bar{p}_0 \left(\frac{\alpha}{m\sigma}\right)^k \tag{1}$$

for $k = 0, 1, 2, \ldots..$

The sum of all possible probabilities must be equal to 1, which gives us equality

$$1 = \sum_{k=0}^{\infty} \bar{p}_k = \bar{p}_0 \left(1 + \frac{\alpha}{\sigma} + \frac{\alpha^2}{2\sigma^2} + ... + \frac{\left(\frac{\alpha}{\sigma}\right)^m}{m!} \left(1 + \frac{\alpha}{m\sigma} + \left(\frac{\alpha}{m\sigma}\right)^2 + ... \right) \right).$$

The expression in the last bracket is a geometrical series, so we can write

$$\bar{p}_0 = \frac{1}{S}, \quad \text{where} \quad S = 1 + \frac{\alpha}{\sigma} + \frac{\alpha^2}{2\sigma^2} + ... + \frac{\left(\frac{\alpha}{\sigma}\right)^m}{m!} \left(\frac{1}{1 - \frac{\alpha}{m\sigma}} \right). \tag{2}$$

Until all the servers are taken, customers coming to the system do not wait at all. It means that for 1 to $m-1$ customers in the system, the waiting time for the upcoming customer is 0. For m customers in the system, the waiting time for the next customer is $\frac{1}{m\sigma}$, since it is the rate that someone of the m customers is served and 1 server becomes available. In general, the estimated waiting time for $m+k$ customers in the system is $\frac{k+1}{m\sigma}$. This makes the average waiting time

$$E(t_w) = (m-1) \cdot 0 + \frac{1}{m\sigma} \sum_{k=0}^{\infty} (k+1)\, \bar{p}_{m+k}.$$

By (1) and (2),

$$E(t_w) = \frac{1}{Sm\sigma} \frac{\left(\frac{\alpha}{\sigma}\right)^m}{m!} \sum_{k=0}^{\infty} (k+1) \left(\frac{\alpha}{m\sigma} \right)^k.$$

The last sum can be converted to $\left(1 - \frac{\alpha}{m\sigma}\right)^{-2}$, which gives us the average waiting time

$$E(t_\omega) = \frac{\frac{1}{m\sigma} \frac{1}{m!} \left(\frac{\alpha}{\sigma}\right)^m}{S \left(1 - \frac{\alpha}{m\sigma}\right)^2}, \tag{3}$$

where

$$S = \sum_{n=0}^{m-1} \frac{1}{n!} \left(\frac{\alpha}{\sigma}\right)^n + \frac{1}{m!} \left(\frac{\alpha}{\sigma}\right)^m \left(\frac{1}{1 - \frac{\alpha}{m\sigma}} \right). \tag{4}$$

We recall that this formula only holds for $E(t_\omega)$ provided $\frac{\alpha}{\sigma} < m$, otherwise the waiting time is infinite.

9.3. **The Main Theorems.** The main estimate that we give for the waiting time is given by the first theorem. It is not a precise estimate, although in a certain sense it is best possible.

Theorem 9.1. *Let $m_0 = \frac{\alpha}{\sigma}$ be the least integer not less than $\frac{\alpha}{\sigma}$. Let $m = m_0 + k$. Then $E(t_\omega) < \frac{1}{k\sigma}$. Furthermore, this estimate is best possible in the sense that for a fixed k and σ and any $\epsilon > 0$, there is an α, such that $\frac{\alpha}{\sigma}$ is an integer and for that value of α and $m = \frac{\alpha}{\sigma} + k = m_0 + k$, $E(t_\omega) > \frac{1}{k\sigma} - \epsilon$.*

Corollary 9.2. *Let $m_0 = \frac{\alpha}{\sigma}$. Then for $m = m_0 + 1$, the average waiting time is less than the average serving time.*

Proof of Theorem 9.1:

By the formula for the estimated waiting time we have

$$E(t_w) = \frac{\frac{1}{m\sigma}\frac{1}{m!}\left(\frac{\alpha}{\sigma}\right)^m}{\left(1-\frac{\alpha}{m\sigma}\right)^2 \cdot S}\ , \tag{5}$$

where

$$S = \sum_{n=0}^{m-1} \frac{1}{n!}\left(\frac{\alpha}{\sigma}\right)^n + \frac{1}{m!}\left(\frac{\alpha}{\sigma}\right)^m \left(\frac{1}{1-\frac{\alpha}{m\sigma}}\right). \tag{6}$$

Without loss of generality, we can rescale the time axis by putting $\sigma = 1$. Thus, $\alpha = m_0, m = m_0 + k$, where m_0 is the minimal number of servers to ensure the equilibrium of the system.

The formula (5, 6) will change into

$$E(t_w) = \frac{\frac{1}{m_0+k}\ \frac{1}{(m_0+k)!}\ m_0^{m_0+k}}{\left(1-\frac{m_0}{m_0+k}\right)^2 \cdot \left(\sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n + \frac{1}{(m_0+k)!}\ m_0^{m_0+k}\left(\frac{1}{1-\frac{m_0}{m_0+k}}\right)\right)} \tag{7}$$

and we want to show that $\mathbb{E}(t) < \frac{1}{k}$. The form can be simplified as follows

$$E(t_w) = \frac{\frac{1}{m_0+k}\ \frac{1}{(m_0+k)!}\ m_0^{m_0+k}}{\frac{k^2}{(m_0+k)^2} \cdot \left(\sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n + \frac{1}{(m_0+k)!}\ m_0^{m_0+k}\ \frac{m_0+k}{k}\right)} =$$

$$= \frac{\frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}}{k^2 \cdot \left(\sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n + \frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}\ \frac{1}{k}\right)}. \tag{8}$$

By multiplying $E(t_w)$ with k and inversing it, the inequality we want to prove changes to

$$k\ E(t_w) = \frac{\frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}}{k\sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n + \frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}} < 1,$$

followed by

$$\frac{1}{k\ E(t_w)} = \frac{k\sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n}{\frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}} + \underbrace{\frac{\frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}}{\frac{1}{(m_0+k-1)!}\ m_0^{m_0+k}}}_{=1} > 1. \tag{9}$$

The first part of the theorem is proved by (9). To show that the given estimate is best possible it remains to prove

$$\lim_{m_0\to\infty} \frac{k\ (m_0+k-1)!\ \sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n}{m_0^{m_0+k}} = 0. \tag{10}$$

By restriction

$$\sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n \le e^{m_0} \quad \forall m_0,$$

and by using Stirling formula

$$(m_0+k-1)! \sim \sqrt{2\pi}\ (m_0+k-1)^{m_0+k-\frac{1}{2}}\ e^{-m_0-k+1},$$

we can restrict the limit as follows

$$\lim_{m_0\to\infty} \frac{k\ (m_0+k-1)!\ \sum_{n=0}^{m_0+k-1} \frac{1}{n!}\ m_0^n}{m_0^{m_0+k}} \ \leq \lim_{m_0\to\infty} \frac{k\ \sqrt{2\pi}\ (m_0+k-1)^{m_0+k-\frac{1}{2}}\ e^{-m_0-k+1}\ e^{m_0}}{m_0^{m_0+k}}.$$

Finally, by processing the limit we get

$$k\ \sqrt{2\pi}\ e^{1-k} \lim_{m_0\to\infty} \left(\frac{m_0+k-1}{m_0}\right)^{m_0+k} (m_0+k-1)^{-\frac{1}{2}} = k\ \sqrt{2\pi}\cdot 0 = 0$$

as desired.

□

10. Birth-Death Models

Queueing theory is a natural tool for producing models for population dynamics or epidemiology. These models are useful when the randomness of system is an important consideration. First let us consider a model where the system is a simple birth-death process. Think of $n \in \{0,1,2,\dots\}$ as the size of the population being considered. Think of λ_n as being the rate at which the population goes from n to $n+1$ for $n \geq 0$. Think of μ_n as being the rate that the population goes from state n to $n-1$ for $n \geq 1$. The following diagram represents the system so described.

$$0 \underset{\mu_1}{\overset{\lambda_0}{\rightleftarrows}} 1 \underset{\mu_2}{\overset{\lambda_1}{\rightleftarrows}} \cdots \underset{\mu_n}{\overset{\lambda_{n-1}}{\rightleftarrows}} n \underset{\mu_{n+1}}{\overset{\lambda_n}{\rightleftarrows}} n+1 \underset{\mu_{n+2}}{\overset{\lambda_{n+1}}{\rightleftarrows}} n+2 \cdots$$

The steady-state probabilities of this system can be computed easily provided the conditions are met for them to exist. The first condition is that the following sum be finite.

$$S = 1 + \sum_{n=1}^{\infty} \prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}$$

If $S < \infty$, then the steady-state probabilities are given by the following.

$$\overline{p}_0 = \frac{1}{S}$$

$$\overline{p}_n = \frac{\prod_{k=0}^{n-1} \frac{\lambda_k}{\mu_{k+1}}}{S} \quad \text{for} \quad n \geq 1$$

There is another condition required for the steady-state probabilities to exist in addition to $S < \infty$. The probability of an infinite number of transitions to occur in a finite period of time must be zero. We will not go into this requirement in detail. It will not occur for a realistic model of the type considered in these notes.

For many populations $\lambda_n = n \cdot \lambda$ where λ is the birth rate for an individual in the population. Similarly, we often have $\mu_n = n \cdot \mu$ where μ is the death rate for an individual. On the other hand, the individual birth and death rates could vary with population size. So, there is good reason to present the model in this generality. The queueing systems we have already discussed are

birth-death processes. So, the applications of birth-death processes are much more general than to populations of organisms.

In the next section we discussion applications to *Island Biogeography* as developed by MacArthur and Wilson in 1967 [8].

11. MacArthur-Wilson Model for Island Biogeography

The simplest model for a population on an island is called *immigration with death*. This would be a setting where immigration to the island occurs at a rate α. Once individuals of the population have immigrated to the island, they do not reproduce but die at an individual rate μ. This could happen if the conditions on the island were not conducive for the species to reproduce. If the species were a plant that required some special condition to produce seeds and this condition was not satisfied on the island, then we would have this setting.

Let us calculate the steady-state probabilities for this model using the approach of the last section. The diagram for this system is given below.

$$0 \underset{\mu}{\overset{\alpha}{\rightleftarrows}} 1 \underset{2\cdot\mu}{\overset{\alpha}{\rightleftarrows}} \cdots \underset{n\cdot\mu}{\overset{\alpha}{\rightleftarrows}} n \underset{(n+1)\cdot\mu}{\overset{\alpha}{\rightleftarrows}} n+1 \underset{(n+2)\cdot\mu}{\overset{\alpha}{\rightleftarrows}} n+2\cdots$$

It is simple enough to make the calculations required in this setting.

$$S = \sum_{n=0}^{\infty} \frac{\alpha^n}{n!\cdot\mu^n} = \exp\left(\frac{\alpha}{\mu}\right)$$

$$\overline{p}_n = \frac{1}{n!}\cdot\left(\frac{\alpha}{\mu}\right)^n \cdot \exp\left(-\frac{\alpha}{\mu}\right) \quad \text{for} \quad n \geq 0$$

You may recognize this from §7. The calculations are the same as $M/M/\infty$ and in fact we can think of the server as death in this case. Of course, in this case there are enough servers to serve each client no matter how many there may be.

A quantitative theory of island biogeography was developed by Robert MacArthur and Edward Wilson [8] to try to explain the quantity and diversity of species found on an island. They did not consider the simple immigration with death model. However, they did consider the following simple model that we now describe.

Suppose that an invasive species arrives at an island at a rate α. Suppose that once the species has migrated to the island it has an individual birth rate of λ and individual death rate of μ. Suppose that the carrying capacity for the species on the island is K. The *carrying capacity* is the maximum level that the population can attain on the island. Then we can model the population on the island in the following way. Let S be the following sum.

$$\begin{aligned} S &= 1 + \frac{\alpha}{\mu} + \frac{\alpha\cdot\lambda}{2\cdot\mu^2} + \frac{\alpha\cdot 2\cdot\lambda^2}{3\cdot 2\cdot\mu^3}\cdots\frac{\alpha\cdot(K-1)!\cdot\lambda^{K-1}}{K!\cdot\mu^K} \\ &= 1 + \alpha\cdot\sum_{n=0}^{K-1}\frac{\lambda^n}{(n+1)\mu^{n+1}} \end{aligned}$$

Then the steady-state probabilities for the population being in state n is given by the following.

$$\overline{p}_0 = \frac{1}{S}$$

$$\overline{p}_n = \frac{\alpha \cdot \lambda^n}{(n+1)S\mu^{n+1}} \quad \text{for} \quad 1 \leq n \leq K$$

The formulas assume that the immigration rate α is negligible once the population is present on the island.

The graph in this section gives the population model for $\alpha = .01$, $\lambda = 2$, $\mu = 1$, and $K = 1000$. For these parameters, the

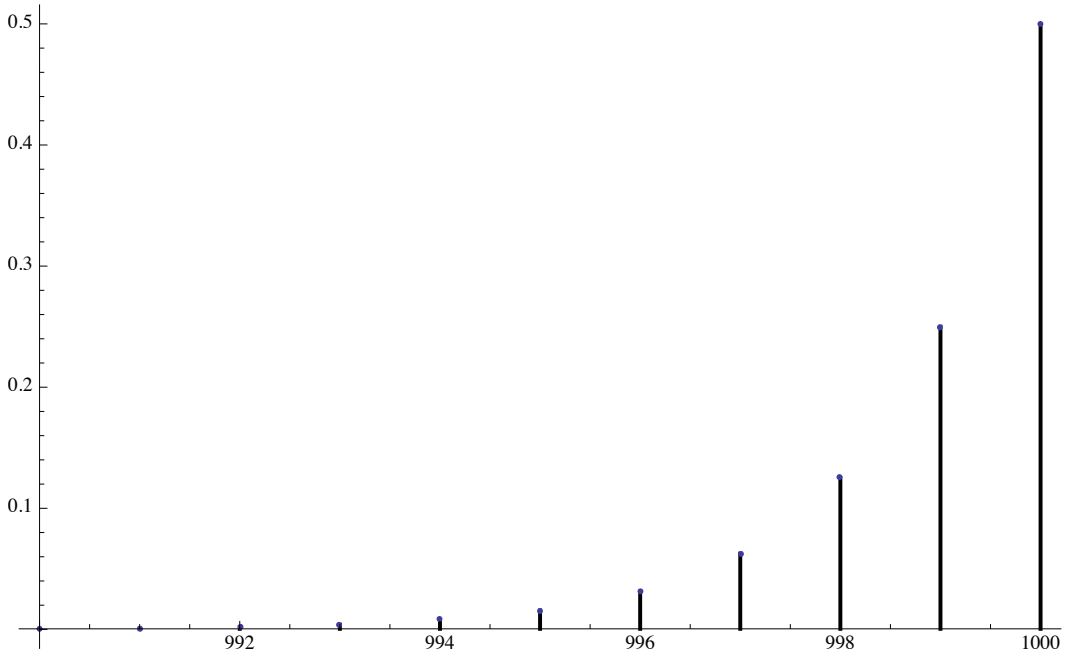

FIGURE 15. Graph of the MacArthur-Wilson Island Population Model
$\alpha = .01$, λ=2, $\mu = 1$, K=1000

For these parameters, $\overline{p}_0 \approx 9.32 \cdot 10^{-297} \approx 0$. Note from the graph that $\overline{p}_K \approx .5$. Note also that the graph is only for 990 to 1000. The other values of $\overline{p}_n$ are so close to zero as to be negligible. This seems very realistic and probably shows that the model is lacking some essential feature. It is probably not the case that the individual birth and death rates are constant for all $1 \leq n \leq K$. Likely the birth rate diminishes with the size of the population and the death rate increases with the size of the population. Note also that $\overline{p}_{K-1} \approx \frac{\mu}{\lambda}$. In fact, a good approximation for $\overline{p}_n$ is given by

$$\frac{\overline{p}_{K-n}}{1-\overline{p}_0} \approx \frac{\left(\frac{\mu}{\lambda}\right)^n}{\frac{1}{1-\frac{\mu}{\lambda}}}$$

In the above case this would give $\overline{p}_{K=1000} \approx \frac{1}{2}$ and $\overline{p}_{K-1=999} \approx \frac{1}{4}$.

There are other questions that might be asked about this model to elucidate its behavior. How long will the population be expected to persist on the island? That is, how long until it dies out?

After extinction, there is an average waiting time of $\frac{1}{\alpha}$ before the next arrival. Let us denote T as the average time to extinction. From the Ergodic Theorem we have that

$$\overline{p}_0 = \frac{\frac{1}{\alpha}}{\frac{1}{\alpha} + T}$$

However, we have a value for $\overline{p}_0$, namely, $\overline{p}_0 = \frac{1}{S}$. From that we can calculate T.

$$T = \frac{S-1}{\alpha} = \sum_{n=0}^{K-1} \frac{\lambda^n}{(n+1)\mu^{n+1}}$$

In this form, the result may not be very useful. However, here is an approximation that gives valuable intuition to the time to extinction.

$$T \approx \frac{\left(\frac{\lambda}{\mu}\right)^{K+1}}{K \cdot (\lambda - \mu)}$$

This estimate is valid when K is large and $\frac{\mu}{\lambda}$ is small. Note that if the carrying capacity K is increased by one, then the average time to extinction is increased approximately by a factor of $\frac{\lambda}{\mu}$. From the analysis in the next section, the probability of a short extinction time is approximately $\frac{\mu}{\lambda}$. This is when the population stays small and never reaches the carrying capacity. The probability of a long extinction time is approximately $1 - \frac{\mu}{\lambda}$. In this case the population reaches the carrying capacity. Let us denote the long extinction times by T'. The long times are approximately given by:

$$T' \approx \frac{\lambda \cdot \left(\frac{\lambda}{\mu}\right)^{K+1}}{K \cdot (\lambda - \mu)^2}$$

Exercise 11.1. *Suppose that* $\lambda = 1.2$ birth per year *and* $\mu = 1$ death per year *for a species. Suppose that the carrying capacity for the species on a given island is* $K = 1,000$. *What is the average time to extinction? What is the average long extinction time? What are these values if* $K = 1,000,000$*?*

12. Logistic Modification of the MacArthur-Wilson Model

The MacArthur-Wilson model of the previous section seems rather naïve. It hardly seems likely that the birth (λ) and death (μ) rates would stay constant until the carrying capacity (K) is reached and then a sudden precipitous change takes place. It is more likely that there is a steady decline of the birth rate or increase in the death rate up to the carrying capacity. We describe an example one such model. In the model we let the carrying capacity of the island be K and let $\lambda_n = \lambda \cdot \frac{K-n}{K}$. We let μ be constant. We think of λ as the *intrinsic* birth rate. As the population increases and competition for resources becomes more intense, the birth rate declines.

Here are the limiting probabilities for this model.

$$S = 1 + \frac{\alpha}{\mu} + \frac{\alpha \cdot \lambda \cdot (K-1)}{2 \cdot \mu^2} + \frac{\alpha \cdot \lambda^2 \cdot (K-1) \cdot (K-2)}{3!\mu^3} + \cdots$$
$$= 1 + \frac{\alpha}{\mu} + \sum_{n=2}^{K} \frac{\alpha \cdot \lambda^{n-1} \cdot \prod_{i=2}^{n}(K-i+1))}{n! \cdot \mu^n}$$

$$\overline{p}_0 = \frac{1}{S}$$

$$\overline{p}_1 = \frac{\left(\frac{\alpha}{\mu}\right)}{S}$$

$$\overline{p}_n = \frac{\left(\frac{\alpha \cdot \lambda^{(n-1)} \cdot \prod_{i=2}^{n}(K-i+1))}{n! \cdot \mu^n}\right)}{S} \quad n \geq 2$$

Below is a plot of $\overline{p}_n$ for $2 \leq n \leq K$. It appears from the graph that the population is approximately normally distributed as we might expect in a realistic setting.

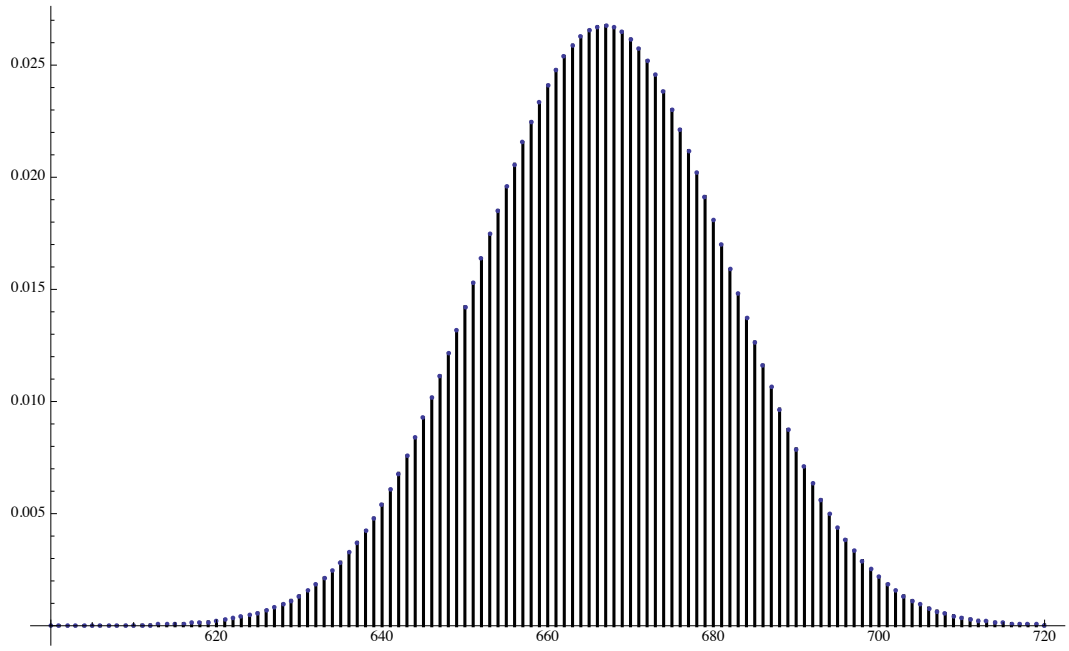

FIGURE 16. Graph of Logistic MacArthur-Wilson Population
$\alpha = .01$, λ=2, $\mu = 1$, K=1000

In fact we can show that if $\overline{p}_0 \approx 0$, then $\overline{p}_n$ is approximately a binomial distribution with K events and probability of success equal $\frac{\lambda}{\lambda+\mu}$. So, the mean population (if present) would be $K \cdot \frac{\lambda}{\lambda+\mu}$. The variance would be $K \cdot \frac{\lambda \cdot \mu}{(\lambda+\mu)^2}$. With the above parameters $\overline{p}_0 \approx 1.5 \cdot 10^{-472} \approx 0$. So, using this

binomial approximation just mentioned, we have a mean population of approximately 667 with a standard deviation of 14.9. This seems a good fit for the graph.

Even if $\overline{p}_0 \not\approx 0$, we can still approximate the conditional probabilities $\frac{\overline{p}_n}{1-\overline{p}_0}$ for $n \geq 1$.

$$\frac{\overline{p}_n}{1-\overline{p}_0} \approx \binom{K}{n} \cdot \left(\frac{\lambda}{\lambda+\mu}\right)^n \cdot \left(\frac{\mu}{\lambda+\mu}\right)^{K-n}$$

13. Probability of Extinction

An island population has to contend with limited resources. This is due to the limiting size of the island. This is what produces the limited carrying capacity. Due to this limited carrying capacity, the population will become extinct with probability one. However, we can study the process by imagining that there are other immigrations after the extinction. For a population with large resources available, there is little loss in assuming that the population could grow without bound. In this case there is a positive probability that the population will never become extinct. In this section we give an indication why this is the case.

Suppose that we have a population that could grow without bound. The states would then be $\mathscr{S} = \{0, 1, 2, \dots\}$. Suppose that the individual birth and death rates are λ and μ, respectively. Then this system can be represented by the following diagram.

$$0 \underset{\mu}{\longleftarrow} 1 \underset{2\cdot\mu}{\overset{\lambda}{\rightleftarrows}} \cdots \underset{n\cdot\mu}{\overset{(n-1)\cdot\lambda}{\rightleftarrows}} n \underset{(n+1)\cdot\mu}{\overset{n\cdot\lambda}{\rightleftarrows}} n+1 \cdots$$

In this case there will be long-term probabilities associated with the system. However, the sum of the $\overline{p}_n$ will not be one. Assume that $\lambda > \mu$. Let us follow the system until either the population becomes extinct, that is, $n = 0$, or until the population reaches some prescribed level, say $n = N$. This system can be modeled as a discrete Markov chain with states $\mathscr{S} = \{0, 1, 2, \dots, N\}$ and with the following transition probabilities. Assuming $0 < n < N$, the probability of going from n to $n+1$ is

$$p_{n,n+1} = \frac{n\lambda}{n\lambda + n\ mu} = \frac{\lambda}{\lambda+\mu}.$$

The probability of going from n to $n-1$ is similarly given by

$$p_{n,n-1} = \frac{\mu}{\lambda+\mu}.$$

The states $n = 0$ and $n = N$ are absorbing states in this analysis. From finite Markov chain theory we can determine the probability of ending in state 0 or N starting in state n. Let us use u_n to denote the probability of ending in state 0 starting in state n. The $1 - u_n$ will denote the probability of ending in state N starting in state n. These probabilities are given below.

$$u_n = \frac{\left(\frac{\mu}{\lambda}\right)^n}{1 - \left(\frac{\mu}{\lambda}\right)^N}$$

$$1 - u_n = \frac{1 - \left(\frac{\mu}{\lambda}\right)^n}{1 - \left(\frac{\mu}{\lambda}\right)^N}$$

We have approached the problem this way so that we can use the theory of finite Markov chains to see what the probabilities are for reaching either of these two absorbing states. However, in the

case we started with $N = \infty$. Taking in the limit in the formulas above we get that the probability of a population at level n going extinct is given below.

$$u_n = \left(\frac{\mu}{\lambda}\right)^n$$

The probability of never going extinct is given by the following.

$$1 - u_n = 1 - \left(\frac{\mu}{\lambda}\right)^n$$

So, for a population not limited by resources and starting in state n, we get that the steady-state probabilities are given by the following.

$$\overline{p}_0 = \left(\frac{\mu}{\lambda}\right)^n$$

$$\overline{p}_n = 0 \quad \text{for} \quad n > 0$$

So, the sum of these steady-state probabilities is not one. It is also the case that if the population does not become extinct, then for any N, the probability is one that the population at some time t will exceed N and never be N or less for all time greater than t.

14. Acuity Analysis for Emergency Care

In this section we analyze a simple queueing system that illustrates a serious problem encountered in emergency care. The problem is that those patients who arrive and are assigned a high acuity level on arrival are given first priority in being treated. The lower acuity level patients have low priority and must wait until all of the high acuity patients have been treated before their treatment begins.

There are give acuity levels in emergency care. These are standardized by the *Emergency Severity Index* or ESI number. The details of the ESI classification system can be found at the following link maintained by the *Agency for Healthcare Research and Quality*.

Our mathematical model is a simplified version of this system. We assume just two levels of acuity. The first level has priority. Even if a patient from the second level of acuity is being treated and a first level patient arrives, the treatment is interrupted and the first priority patient immediately begins treatment. We assume that there is just one server to begin with. Later we will develop a more refined model. The purpose of the model is to show how congestion can occur by the introduction of a priority treatment system.

The parameters for this system is the following $\alpha_1 = \alpha \cdot p_1$, the arrival rate of high priority patients. We denote the arrival rate of the low priority patients by $\alpha_2 = \alpha \cdot p_2$. We assume that these are independent Poisson systems. The service rate for a high priority patient is σ_1. The service rate for a low priority patient is σ_2. There are two lines, one for the priority one patients and one for the priority two patients. The only time that priority two patients are treated is when there are no priority one patients in the system. We visualize the system in the following way.

The patients in the first queue can be treated as an $M/M/1/FIFO$ system. To those patients and the facility, the other patients are virtually invisible. It is when only the system consisting of the first priority patients is in state zero that the patients in the second level of priority can be treated.

We have already analyzed the $M/M/1/FIFO$ system. With the parameters that we have for that system we have the following results for the first priority patients.

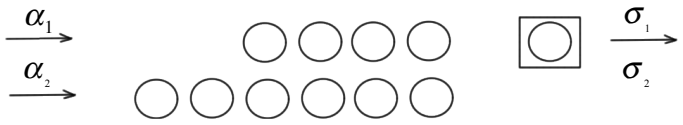

FIGURE 17. Priority Queue with Two Levels of Priority

$$\overline{p}_n = \left(\frac{\alpha_1}{\sigma_1}\right)^n \cdot \left(1 - \frac{\alpha_1}{\sigma_1}\right)$$

$$\overline{n} = \mathbb{E}(n) = \frac{\frac{\alpha_1}{\sigma_1}}{\left(1 - \frac{\alpha_1}{\sigma_1}\right)}$$

From this information we can determine the average time that it takes from the instant that the priority one system goes from empty to non-empty to when it returns to empty once again. Let us call this time T. The *Ergodic Theorem* gives us the formula.

$$\overline{p}_0 = \left(1 - \frac{\alpha_1}{\sigma_1}\right) = \frac{\frac{1}{\alpha_1}}{\frac{1}{\alpha_1} + T}$$

This gives us the following value for T.

$$T = \frac{1}{\sigma_1 - \alpha_1}$$

We can now determine the average time for the priority patient to complete his/her treatment. Let us label this time τ.

$$\tau = \frac{\sigma_2}{\sigma_2 + \alpha_1}\left(\sum_{n=0}^{\infty}(n+1)\cdot\frac{1}{\sigma_2+\alpha_1}\left(\frac{\alpha_1}{\sigma_2+\alpha_1}\right)^n + T\sum_{n=1}^{\infty} n\cdot\left(\frac{\alpha_1}{\sigma_2+\alpha_1}\right)^n\right)$$

This simplifies to the following.

$$\tau = \frac{\sigma_1}{(\sigma_1 - \alpha_1)\cdot\sigma_2} \tag{11}$$

We are now in a position to at least give the average times that patients will spend in the system given the number of patients already present. The priority one patients will spend an average of $(\overline{n}+1)\cdot\frac{1}{\sigma_1}$ time in the system. From the above formula for $\overline{n}$ for the priority one patients, we get the following average time.

$$(\overline{n}+1)\cdot\frac{1}{\sigma_1} = \frac{1}{\sigma_1 - \alpha_1}$$

Now suppose that a patient comes into the system with n_1 priority one patients in line and n_2 priority two patients in line. Let us call the time to complete service for an arriving priority one patient $T_1(n_1, n_2)$ and the time to complete service for the second priority patient $T_2(n_1, n_2)$. If the patient is priority one, then the time waiting is the following.

$$T_1(n_1, n_2) = (n_1 + 1) \cdot \frac{1}{\sigma_1}$$

The n_2 priority two patients play no role in how long it will take this patient to finish treatment.

How long will it take for a priority two patient to finish treatment coming to a system with n_1 priority one patients and n_2 priority two patients? First, the system must be cleared of all priority one patients. This will include the n_1 patients already there plus any others that might arrive during the time that these n_1 are being treated. The time for the priority one system to get to zero from n_1 is just $\frac{n_1}{\sigma_1}$. On the other hand, once the priority one system is in state zero, then the time for the n_2 priority two patients to complete treatment is just $n_2 \cdot \tau$ where τ was computed above. Thus, the total time for the priority two patient to complete treatment will be the following.

$$T_2(n_1, n_2) = \frac{n_1}{\sigma_1} + (n_2 + 1) \cdot \tau = \frac{n_1}{\sigma_1} + \frac{(n_2 + 1) \cdot \sigma_1}{(\sigma_1 - \alpha_1) \cdot \sigma_2}$$

It would be good to know the average waiting time for patients who are assigned priority two. To do this we would need to compute the long-term probabilities $\overline{p}_{(n_1,n_2)}$ of being in each state (n_1, n_2). We do not have simple formulas for these at the present time. However, we still have discovered formulas for the total time waiting and being served.

15. Acuity Analysis of Emergency Care - Continued

Consider the two level priority queue that was discussed in the last section. Assume that the parameters $\alpha_1, \sigma_1, \alpha_2$ and σ_2 represent that arrival rate and service rate for the priority one and two level patients, respectively. We now derive the service time τ in (11) in a simpler way. The service time, τ, for a priority two patient can be thought of as a sum of the basic service time, $\frac{1}{\sigma_2}$, together with an interruption time, S, due to the arrival of priority one patients..

$$\tau = \frac{1}{\sigma_2} + S$$

The interruption time can be seen to be given by the following equation.

$$S = \alpha_1 \cdot \left(\frac{1}{\sigma_2} + S \right) \cdot \frac{1}{\sigma_1}$$

We can solve for S and substitute into the equation for τ to get the formula that we got by a more complicated process in the previous section.

$$\tau = \frac{\sigma_1}{(\sigma_1 - \alpha_1) \cdot \sigma_2}$$

However, let us write τ in a different way.

$$\tau = \frac{\frac{1}{\sigma_2}}{\left(1 - \left(\frac{\alpha_1}{\sigma_1}\right)\right)} \tag{12}$$

In the form of (12) we can see that τ is the average service time without interruption $\frac{1}{\sigma_2}$ divided by the probability that the priority one system is in the zero state $1 - \frac{\alpha_1}{\sigma_1}$. This tells us precisely how the service time is being increased by the interruptions of priority one patients and how to adjust.

Consider an example. Suppose that the priority one system has $\alpha_1 = 9$ and $\sigma_1 = 10$. Then the traffic intensity is $\frac{\alpha_1}{\sigma_1} = \frac{9}{10}$. In this case, we get a dramatic change in the service time for the priority two patients.

$$\tau = 10 \cdot \frac{1}{\sigma_2}$$

The actual service time is *ten times* what it would normally be without interruptions. This is true for every priority two patient in front of this patient. If there are five patients in front of this patient, then the total time in the system will be *sixty times* the time it would normally take for that patient finally complete treatment. This makes more clear how assigning priority can make the system extremely congested.

Note also that if we had $\alpha_1 = 9$, $\alpha_2 = \frac{1}{2}$, $\sigma_1 = 10$, and $\sigma_2 = 10$, then if there were no distinctions between priority one and two patients, we would have a $M/M/1/FIFO$ queueing system with $\alpha = 9.5$ and $\sigma = 10$. The system would be in equilibrium even though there might be longer lines than would be desirable. The actual average $\overline{n} = \frac{\frac{9.5}{10}}{1-\frac{9.5}{10}} = 19$. Let is compare this with the two priority system. The service rate for priority two patients must be adjusted. The new rate is not σ_2 but

$$\hat{\sigma}_2 = \frac{1}{\tau} = \sigma_2 \cdot \left(1 - \frac{\alpha_1}{\sigma_1}\right)$$

In the case we have been considering, $\hat{\sigma}_2 = \frac{\sigma_2}{10}$. The system is also at equilibrium since $\alpha_2 = \frac{1}{2} < \hat{\sigma}_2 = 1$. However, the average time being served for a priority two patient is ten times the service time. So, if there are three priority patients in line before that patient arrived, then the total time being served would be forty times the service time.

We will have a clearer picture when we can compute the steady-state probabilities. However, the model is giving a sense of the congestion and inconvenience that arises through the priority system.

Of course, in emergency care, there are compelling reasons that priority one patients are treated first. They are facing life-threatening circumstances. However, the overcrowding and gridlock that some emergency rooms are now facing make it clear that something must be done to bring relief to those enduring the long waiting times getting treatment. The average median time for emergency care from registration to discharge is *more than four hours*. This is several times the median time that the care itself would require.

16. Higher Levels of Acuity

Now consider the case that we have several levels of acuity: priority 1, priority 2, ..., priority n. Each level takes priority over those levels below in the same fashion as the case above. Below is a diagram that illustrates this situation.

In this case we can also derive the modified service time for those in priority n.

$$\tau_n = \frac{\frac{1}{\sigma_n}}{1 - \sum_{i=1}^{n-1} \frac{\alpha_i}{\sigma_i}}$$

Of course, for each level of priority we also have a modified service time, namely:

$$\tau_m = \frac{\frac{1}{\sigma_m}}{1 - \sum_{i=1}^{m-1} \frac{\alpha_i}{\sigma_i}}.$$

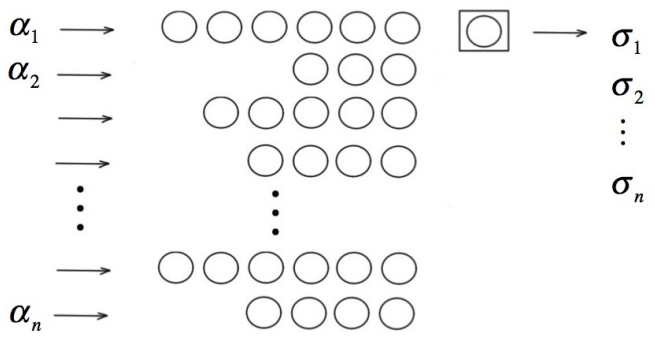

FIGURE 18. Priority Queue with Multiple Levels of Priority

For an equilibrium to exist, it is clear that $\sigma_i > \alpha_i$ must hold for all for all $1 \leq n$. We can now see also that

$$\sum_{i=1}^{n-1} \frac{\alpha_i}{\sigma_i} < 1$$

must hold.

There is also modified rate of service for each i, namely:

$$\hat{\sigma}_i = \sigma_i \cdot \left(1 - \sum_{j=1}^{i-1} \frac{\alpha_j}{\sigma_j}\right).$$

For each of these modified rates of service, $\hat{\sigma}_i > \alpha_i$ must hold for an equilibrium to exist. So, there are more criteria that must hold for equilibrium than first meets the eye.

Note that there is an easy fix for this situation. Instead of one server with all the interruptions caused by multiple priorities, simply have a separate server for each priority or at least for those with the highest $\frac{\alpha}{\sigma}$ ratio. For example, consider a two priority system with $\frac{\alpha_1}{\sigma_1} = \frac{\alpha_2}{\sigma_2} = \frac{1}{2}$. Then the priority system is not stable since $1 - \frac{\alpha_1}{\sigma_1} - \frac{\alpha_2}{\sigma_2} = 0$. However, separating the two priority clients and treating them with separate facilities would yield two $M/M/1/FIFO$ queues each of which has $\frac{\alpha}{\sigma} = \frac{1}{2} < 1$. So, both are stable and will have reasonable waiting times. The cost is the addition of a server. However, the server for priority two patients would likely be less expensive than the server for priority one patients.

In real life many factors must be taken into account. However, the mathematical models considered in the last three sections give insights that are not likely to be seen otherwise. If it is necessary to take other factors into account, more sophisticated models can be constructed and simulations performed to see the effect of those factors.

In the next section we will develop a method for determining the steady-state probabilities for priority queues.

17. Simulation of Priority Queues with Preemption

Here we discuss a realistic model of Emergency Care developed by our research group. It is a simulation program written in *R*. The program was produced by Joshua Hurwitz on collaboration

with Jo Ann Lee. It was a project in a class taught by Professor Scott McKinley who was also involved in the project. The program simulates the number of patients in each of the five ESI levels of priority through a period of several days. For the parameters that fit the arrival rates and staffing at the Shands ED, we get a good fit. The simulation model linked below allows one to adjust the parameters and observe the results of the simulation.

The program assumes that each priority class preempts those of lower priority classes. The preemption includes interruption of treatment of a lower priority patient so that treatment times are longer than would be anticipated. Experimenting with the program shows how easily congestion arises. Examination of the graphs produced by the program shows that one should be able to anticipate the congestion and perhaps take measures to bring on additional personnel to avoid the extreme waiting times. We hope that these simulations will allow hospital management to recognize criteria that indicate when congestion is likely to occur. If we can recognize when congestion is likely to occur, then staffing could be arranged to prevent it.

The paper describing the simulation program was published as, "A comprehensive simulation platform to quantify and manage site-specific emergency department crowding" *Biomed Central Medical Informatics and Decision Making*, `http://www.biomedcentral.com/1472-6947/14/50` (online publication: Joshua Hurwitz, Jo Ann Lee, Kenneth Lopiano, Scott McKinley, James Keesling, Joseph Tyndall). There is a software website demonstrating the simulation platform `http://spark.rstudio.com/klopiano/EDsimulation/`.

References

[1] François Baccelli and Pierre Brémaud. *Elements of Queueing Theory*. Springer-Verlag, Berlin, second edition, 2010.

[2] Lothar Breuer and Dieter Baum. *An Introduction to Queueing Theory and Matrix-Analytic Methods*. Springer, Dordrecht, Netherlands, 2005.

[3] Jody Crane and Chuck Noon. *The Definitive Guide to Emergency Department Operational Improvement*. CRC, Boca Raton, Florida, 2011.

[4] Giovanni Giambene. *Queueing Theory and Telecommunications*. Springer, New York, 2005.

[5] Donald Gross, John F. Shortle, James M. Thompson, and Carl M. Harris. *Fundamentals of Queueing Theory*. John Wiley, Hoboken, New Jersey, 2008.

[6] Shaler Sticham Jr. *Optimal Design of Queueing Sytems*. CRC, Boca Raton, Florida, 2009.

[7] Leonard Kleinrock. *Queueing Systems*. John Wiley, Hoboken, New Jersey, 1975.

[8] Robert H. MacArthur and Edward O. Wilson. *The Theory of Island Biogeography*. Princeton University Press, Princeton, 1967.

[9] William Stewart. *Probability, Markov Chains, Queues, and Simulation: the Mathematical Basis of Performance Modeling*. Princeton Press, Princeton, New Jersey, 2009.