Ambulance Staffing Using Queuing Theory

When designing an EMS response system, the primary concern is response time. Response time is determined by 1) the location of the ambulance base or, in the case of SSM, the posting location; and 2) the availability of the ambulance to respond. Ambulance availability, in turn, is determined by call demand, the number of ambulances staffed, and the time required to service a call. Historically, ambulance staffing has been based on relatively unsophisticated statistical demand analysis, supplemented with the intuition of the EMS manager. We will use a more sophisticated means of analyzing call demand called Queuing Theory. Through the use of queuing theory, we will be able to determine the number of ambulances needed in a given response district such that an ambulance is available for immediate dispatch with a predetermined reliability. We will also use this technique to determine necessary staffing levels not only by district, but also by hour of day and day of week. Once this information is available, planning shift rotations becomes more exact and cost-effective.

Review of Mathematical Concepts:

1. Factorial (!) – to calculate the factorial of any number, take the number and multiply it by one less that number, then take that quantity and multiply it by two less than the original number. Repeat this process until you are finally multiplying the quantity by 1.

   Examples:

   $3! = 3x2x1 = 6$
   $4! = 4x3x2x1 = 24$
   $5! = 5x4x3x2x1 = 120$

   Note: $0! = 1$
   $1! = 1$

2. Exponentiation (X^y) – multiply x by itself y times

   Examples:

   $2^2 = 2x2 = 4$
   $2^3 = 2x2x2 = 8$
   $2^4 = 2x2x2x2 = 16$

   Note: $X^0 = 1$ any number raised to the power of zero equals 1 ($3^0 = 1$, $10^0 = 1$)

   $X^1 = 1$ any number raised to the power of 1 equals itself ($3^1 = 3$, $10^1 = 10$)

3. Summation (∑) – asks you to repeat a mathematic operation, replacing certain values each time, then adding together (or summing) the results of each repetition. The summation sign (∑) may also have an index. The index tells you which variable will be substituted in the mathematical operation, and what values to use during each substitution. For example:

$$\sum_{n=0}^{n=k-1} 5\,x\,n$$

Suppose we know that the value of k is 4. The notation under the summation sign indicates that we repeat the function (5*n), beginning with n = 0. Then we increase the value of n by 1 for each repetition until we reach the value above the summation sign. In this case, n = (4-1) = 3. Once we have calculated the value of the operation for each iteration, we SUM all of the values across all iterations.

Example:

For n = 0	5 x 0 = 0
For n = 1	5 x 1 = 5
For n = 2	5 x 2 = 10
For n = 3	5 x 3 = 15

Then we add together the results (0 + 5 + 10 + 15) = 30.

Components of Queuing Theory

Arrival rate - the number of customers (ambulance calls) received during a given period of time (usually calculated on an hourly basis). For most problems, we assume that arrivals follow a particular probability distribution called a **Poisson pattern** or **Poisson probability distribution**. The Poisson probability distribution is described by the following equation:

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where,

x = number of arrivals in a specified period of time
λ = average or expected number of arrivals for the specific period of time
e = 2.71828

Example: During a given hour, past data reveal that the average number of ambulance calls received is 3. With this single parameter, we can calculate the probability of receiving any specified number of calls during this given hour.

$$P(x) = \frac{3^x e^{-3}}{x!}$$

The probabilities of receiving 0 thru 4 calls during this hour are as follows:

$$P(x=0) = \frac{3^0 e^{-3}}{0!} = \frac{1x2.71828^{-3}}{1} = 0.0498$$

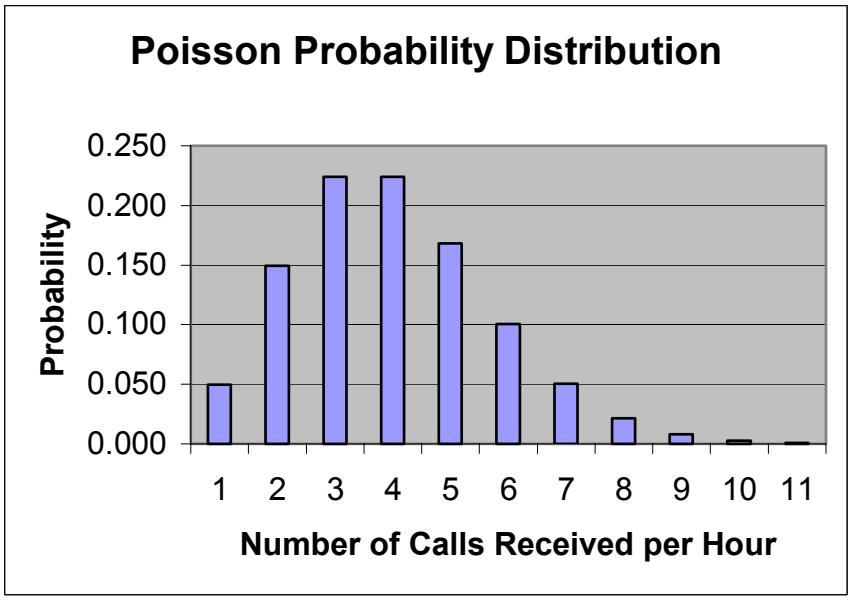
$$P(x=1) = \frac{3^1 e^{-3}}{1!} = \frac{3x2.71828^{-3}}{1x1} = 0.1494$$

$$P(x=2) = \frac{3^2 e^{-3}}{2!} = \frac{9x2.71828^{-3}}{2x1} = 0.2241$$

$$P(x=3) = \frac{3^3 e^{-3}}{3!} = \frac{27x2.71828^{-3}}{3x2x1} = 0.2241$$

$$P(x=4) = \frac{3^4 e^{-3}}{4!} = \frac{81x2.71828^{-3}}{4x3x2x1} = 0.1680$$

This means that we would expect to receive no calls during this hour 4.98% of the time, one call 14.94% of the time, and 2 calls 22.41% of the time, and so on. If we continued on and calculated the probabilities for receiving 5, 6, 7, etc., calls, we would get a probability distribution such as the one below:



Service Time Distribution - Similar to describing the distribution of call arrival rates, we need a mechanism to describe how long it takes to service the call. Because of the variation in response times, the time to treat patients at the scene, and transport times, the total time required to complete or "service" a call varies from call to call. In general, we can describe the variation in service time using the **exponential probability distribution**. The exponential probability distribution is defined as follows:

$$f(x) = \mu^{\,e-\mu x}$$

where,

x = service time

μ = the number of calls that can be handled during a specified period of time, calculated as 60 divided by the average time required to service a call.

Example: An EMS system takes an average of 45 minutes to complete a call.

μ = 60 minutes/45 minutes = 1.33.

Therefore, the service capacity of a single ambulance is 1.33 calls per hour.

With this equation, we can compute the probability of any one call taking *x* minutes to complete, although we really don't need to for our purposes.

Queue Discipline - When the call arrival rate exceeds the service capacity of an EMS system, calls will "back up", "stack", or more accurately, queue. How these calls are handled is referred to as queue discipline. Examples of queue disciplines include:

1. **FIFO** - "first in, first out". This is what we call "first come, first served" or “serviced in the order in which they are received".

2. **LIFO** - "last in first out". Similar to the queue discipline of an elevator.

3. **Truncated** - we assume that calls are not allowed to queue, i.e., they are serviced by someone else.

4. **Infinite** - calls are allowed to pile up until all are serviced.

5. **Prioritized** - service certain calls before others, such as emergency calls before non-emergency

For our purposes, the equations we use will assume FIFO with infinite queuing. Although other methods exist for dealing with the other queue disciplines, they are mathematically complex and some queue disciplines cannot be solved mathematically; they must be modeled using computer simulation.

Channels - Queuing systems are either single or multi-channeled. Single channel systems means that there is a single queue and a single service facility (i.e., ambulance). A multiple channel queuing system means that there is more than one ambulance. Although the equations for analyzing a single channel queuing system are simpler than those for the multiple channel system, they have limited applications in EMS (maybe for calculating service times for a single 911 dispatcher). So, we will focus only on a multiple channel queuing system with Poisson arrivals and exponential service times, where:

- ***k*** = number of channels (ambulances)
- **λ** = mean arrival rate for the system **<u>for a specific hour</u>**
- **μ** = mean service **<u>rate</u>** for each channel (can assume to be identical across all ambulances within a given system or district, but will probably vary across districts due to variation in response times and transport times)

We can learn much about how an EMS system operates using queuing theory and the equations given below. We can use queuing theory to examine issues related to staffing, waiting times, and queue lengths, all of great importance to an EMS administrator.

1. The probability that all ***k*** service channels (ambulances) are idle (i.e., the probability of zero calls in the system):

$$P_0 = \frac{1}{\left[\sum_{n=0}^{n=k-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \frac{k\mu}{k\mu-\lambda}} \qquad \text{for } k\mu > \lambda$$

2. The probability that the next call has to wait for service:

$$P_w = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \frac{k\mu}{k\mu-\lambda} P_0$$

There are other statistics that can be calculated using queuing theory, including the average number of calls pending; the average amount of time a call will spend waiting for an ambulance to become free for dispatch; and the average number of calls being serviced and awaiting dispatch. Even though all of these statistics may be of interest to the EMS administrator, the two equations above are of primary interest because they can be used to determine the number of ambulances needed to provide adequate service. For example, an accepted standard for the industry is to staff enough ambulances such that there is at least one ambulance available for dispatch 90% of the time. In other words, the probability that an arriving call has to wait for service (P_W) is 10% or less. Or put another way, the ambulance will spend only 10% of the time busy servicing a call.

Realize, however, that there is an inverse relationship between availability and utilization. As you increase your availability, you decrease your utilization, and low utilization rates mean higher costs.

Sample Problem:

An EMS district has two ambulances staffed 24 hours per day. This particular district has suffered from poor response times and the crews are complaining about being overworked. You are considering increasing staffing of this district from three to four ambulances. You collect data for the busiest hour during the day (i.e., where performance is likely to be the poorest), and determine that the average call arrival rate is 3 calls per hour and that the average time required to service a single call is 35 minutes. Analyze the **current** performance of the system.

We know that:

μ = 60/35 = 1.7 calls per hour (service rate)
λ = 3 calls per hour
k = 3 ambulances

1. Calculate the probability of an idle system.

$$P_0 = \frac{1}{\left[\sum_{n=0}^{n=k-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right] + \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \frac{k\mu}{k\mu-\lambda}}$$

$$\left[\sum_{n=0}^{n=k-1} \frac{1}{n!}\left(\frac{\lambda}{\mu}\right)^n\right]$$

$$n=0: \quad \frac{1}{0!}\left(\frac{3}{1.7}\right)^0 = 1$$

$$n=1: \quad \frac{1}{1!}\left(\frac{3}{1.7}\right)^1 = 1.76$$

$$n=2: \quad \frac{1}{2!}\left(\frac{3}{1.7}\right)^2 = (0.5)(3.114) = 1.557$$

$$1+1.76+1.557 = 4.317$$

$$P_0 = \frac{1}{4.317 + \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \frac{k\mu}{k\mu-\lambda}}$$

$$P_0 = \frac{1}{4.317 + \frac{1}{3!}\left(\frac{3}{1.7}\right)^3 \frac{(3)(1.7)}{(3)(1.7)-3}}$$

$$P_0 = \frac{1}{4.317 + (0.167)(5.495)\left(\frac{5.1}{2.1}\right)}$$

$$P_0 = \frac{1}{4.317 + 2.228} = \frac{1}{6.545} = \boxed{0.152}$$

2. Calculate the probability that an incoming call will enter a queue.

$$P_w = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \frac{k\mu}{k\mu-\lambda} P_0$$

$$P_w = \frac{1}{3!}\left(\frac{3}{1.7}\right)^3 \frac{(3)(1.7)}{(3)(1.7)-3} (.152)$$

$$P_w = \frac{1}{6}(1.764)^3\left(\frac{5.1}{2.1}\right)(.152)$$

$$P_w = (.1667)(5.489)(2.43)(.152)$$

$$P_w = 0.337$$

What this means, then, is that with our current staffing plan of 3 ambulances, for this given hour, an incoming call has a 33.7% chance of finding all ambulances busy and having to wait for the next ambulance to become available. If your policy is to have an ambulance immediately available for dispatch 90% of the time, you would have to repeat these calculations with 4 ambulances, then 5 ambulances, until you find the least number of ambulances that will satisfy the inequality $P_w \leq 0.10$

Use of Queuing Theory in Developing Staffing Plans

The calculations used above would then have to be repeated across each district, each hour of the day, and each day of the week. Obviously, a computer (or a consultant) would be needed to perform the time-consuming calculations. Ultimately, you should have a chart (such as the one below) indicating the number of ambulances to meet your stated unit availability goal (95%, 90%, 85%, etc.). You should have a chart for each response district. You will likely want to look at varying the availability goal to see how it affects the minimum number of ambulances necessary to meet the goal. (Hint: This is NOT a linear function; that is, doubling the number of ambulances doesn't double availability. It may only take a single ambulance to move from 80% to 85% availability, but it may take 5 ambulances to move from 90% to 95% availability.)

EMS SYSTEM STAFFING PATTERN ANALYSIS
M/M/N QUEUEING SYSTEM MIN. UNIT AVAILABILITY = .90

DISTRICT 1

DAY/HR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	2	0	1	1	0	0	1	1	1	2	2	2	1	1	0	1	1	0	0	1	1	0	1
2	1	0	0	0	0	1	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
3	0	0	1	0	0	0	0	1	2	2	0	1	0	0	2	0	1	0	1	0	1	0	0	1
4	0	1	0	1	1	1	0	1	0	2	1	2	1	1	2	0	1	0	1	1	0	0	0	1
5	0	0	0	0	0	0	0	1	0	1	2	0	1	1	0	1	0	0	0	1	2	0	0	1
6	1	1	1	1	1	1	0	2	0	1	1	0	0	1	1	0	1	0	0	2	0	1	0	1
7	1	0	0	0	0	0	0	1	1	0	2	0	0	0	0	0	1	1	1	0	0	0	0	0

REGIONAL EMS SYSTEM STAFFING PATTERN ANALYSIS
M/M/N QUEUEING SYSTEM MIN. UNIT AVAILABILITY = .90

DISTRICT 2

DAY/HR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	3	2	3	3	3	2	3	3	2	3	4	3	3	4	4	3	2	3	3	3	3	3	3	2
2	2	3	2	2	2	2	3	3	[illegible]								3	4	3	3	3	2	3	2
3	3	2	3	2	2	2	2	2	[illegible]								3	3	4	3	2	2	2	2
4	2	2	3	2	3	2	2	3	[illegible]								4	3	3	3	3	3	3	2
5	2	2	2	2	2	0	2	2	[illegible]								4	3	2	2	4	3	3	2
6	2	3	3	3	3	2	3	2	[illegible]								4	3	3	4	3	3	4	2
7	2	3	2	2	2	0	2	2	2	3	2	4	3	2	5	4	2	3	3	3	3	3	3	3

Armed with the queuing analysis, the next step will be to develop a staffing plan. This is where art, science, and practicality meet. Rarely will there be demand patterns that will conveniently fit into standardized work schedules. What you will have to do is identify the areas of peak demand, determine if they warrant an extra ambulance, and then design a work shift that will cover those peak demands while not making a shift schedule that no paramedic would be willing to work.

In looking at district 1 above, there is considerable time variation in demand, between 0 and 2 units. We have zero ambulance requirements during some hours because there were no calls during the period in which these data were obtained. So now the tough

decision is deciding whether this district warrants even a single ambulance. If there is another base nearby that can cover this district and provide reasonable response times, then probably not. However, don't negate the political aspects of ambulance deployment. It's relatively easy to add an ambulance base in a community, but closing one down will likely result in a public outcry. This is where you have to earn the big bucks they pay you to make these tough decisions. If I were going to keep the base open, I would likely staff it 24-7 using 1 ambulance. I would do this recognizing that this will be an ambulance with lots of down time ("unprofitable time"), and that occasionally a call in the district will find the ambulance busy and an ambulance from a neighboring district will be required to provide backup.

District 2 shows peak demand during daytime hours, a perfect opportunity for a primetime unit. Overall, it looks like three 24-hour ambulances are needed. Although we could get by with 2 during the early morning hours, it would be difficult (although not impossible) to devise a practical work schedule to cut back to two ambulances during those hours. So I will opt for three 24-hour units. Then we have some options for the primetime units. Although there are some hours that need 5 ambulances, they are few and very scattered, so I will opt to add only 1 primetime unit. Now we have to decide what staffing plan to use. A couple of options exist: one additional ambulance 08:00-20:00 seven days per week (12 hour shifts) (yellow area). This will require 2 shifts of 2 paramedics. Alternatively, we could staff the primetime unit M-F 09:00-17:00 (8 hour shifts) (blue area). This will require only 2 additional paramedics. The final decision must balance adequate coverage with personnel costs, practicality of shift rotations, and of course, politics.