The Wait for an Ambulance

Queuing Theory for Managers

by George K. Barton, MBA, EMT-P

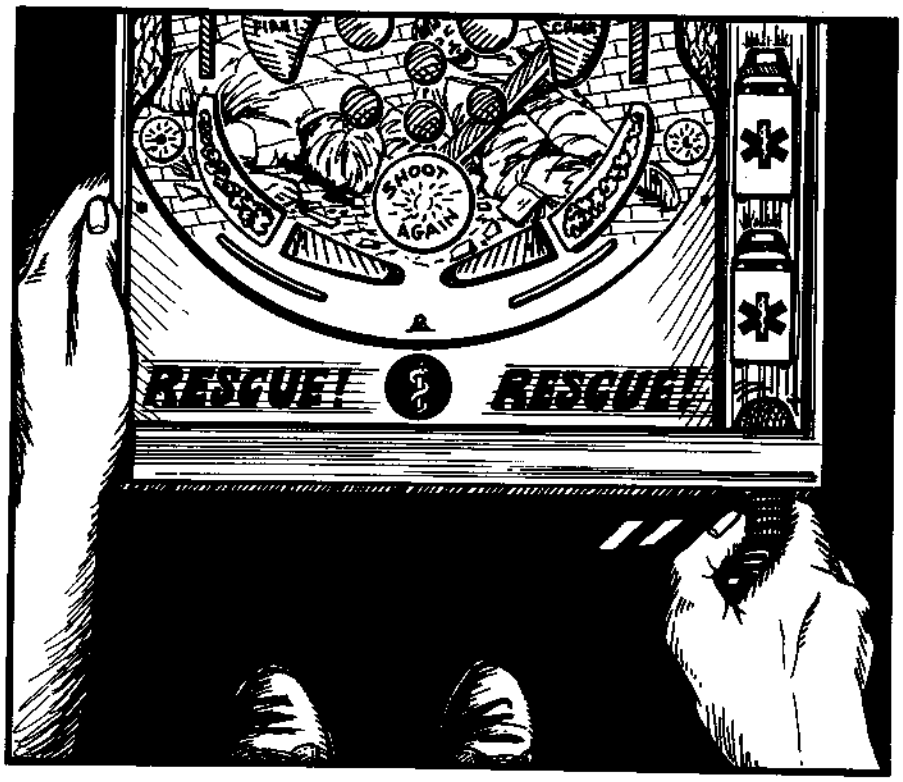
No one likes to wait in a line, and none of the systems with which we work would permit a lengthy wait to occur for a patient while the dispatcher searches for an ambulance to send on the critical call. On the other hand, no system can place an ambulance on every corner and call in a backup crew when each call is dispatched. Somewhere between having too many or too few ambulances lies a viable solution: queuing theory. This simple mathematical model or set of formulas will enable you to determine the number of ambulances needed, by hour of day and day of the week, to meet calls for service in an efficient manner, and will provide objective information to modify previously committed resources.

The skills developed from using the model in this article will be beneficial for all systems managers, even those who dread quantitative analysis. The model will serve as a supplement to the commonly used rules-of-thumb that have provided reliable guidance for experienced managers.

Most managers have found that an EMS system will receive one emergency call a day per 10,000 population. Moreover, transfers can be estimated from the previous year's transfer calls if the current hospital utilization rate and the local population growth rate are known. The average total number per day for all types of calls usually follows the same population growth rate. Finally, we all check the transfer load for the next day and can add an extra day crew if an extra eight calls are scheduled. Those are the basic "seat-of-the-pants" rules, and new quantitative decision aids found in the queuing theory will not replace them. However, just as you have learned to improve your skills in the field of medicine to better serve your patients, you can also advance your planning and strategic skills to better manage your operation through an exposure to some analytically based developments.

"Queuing" is, admittedly, an odd term and may seem obscure to some American readers. The term originates from the British "queue," which means "stand in line." Of course, it is important to avoid having your patients wait in a line for the next available ambulance while also ensuring that you are not wasting valuable staff time and resources that could be scheduled more efficiently.

George Barton, an EMT for 16 years and paramedic for 10, is assistant director for Albuquerque Ambulance and flight coordinator for Presbyterian Air Ambulance. This article completed his MBA degree, and is one of the cornerstones of computerized strategic and financial analysis for the Albuquerque system.



Predicting Service Demands

No one knows when the next call for an ambulance will occur. All you can hope to accomplish will be a prediction based upon the call patterns of yesterday, last week, or last month. The queuing theory approach involves the use of the probability theory to predict the pattern of service call arrivals using the historical frequency of calls. Patient demands for service are rarely distributed in the familiar bell-shaped normal curve. Instead, they often follow a special probability distribution called Poisson (named after a famous French mathematician). Poisson-demand distributions allow for the possibility that an ambulance will *not* be needed during a given time interval such as the next hour.

If the rate of patient calls for ambulance services follows a Poisson pattern, it is possible to use queuing models to help predict average EMS service characteristics. In particular, the use of an average call rate will predict whether the ambulance staffing will be sufficient. The skeptical manager may argue that staffing for an average call frequency will create problems during an abnormally high call period. Certainly *any* method of predicting the needs of a system will not address every possibility. Unusual peak periods of demand occur for both the general approximations now used by most EMS managers and the queuing theory model proposed as a replacement for those approximations. The queuing theory has simply proven to be far more reliable than the rough approximations.

Some Experiential Results

The queuing theory allows us to take into account the markedly different service needs of our patients. This approach considers three categories of calls: emergency calls, non-emergency transfers, and no-service calls. Forty percent of all calls are triaged as true emergencies. They are dispatched immediately and are generally responded to using lights and sirens. The length of time necessary for a crew to treat a patient, transport, and return to service will vary according to crew training, geography, and hospital location. Our statistics show that the average length of an emergency call varies between 40 and 43 minutes, and a crew running back-to-back calls can treat an average of 1.46 patients per hour (60 minutes per hour divided by 43 minutes per call). Although your times may differ, this value of 1.46 patients per hour represents the *service capacity* for one of our EMS vehicles.

The other 60 percent of our work is made up of non-emergency transfers and no-transport emergencies. Their effect on the emergency response capability is markedly different, and they should be analyzed separately.

Non-emergency transfers require a crew to respond without lights and siren and may require the crew to spend additional time moving the patient from a bed in the home or hospital. Transfer calls average between 52 and 56 minutes, and thus a crew can treat an average of slightly more than one patient per hour. However, transfers can normally be prescheduled by the dispatcher and can be delayed in order to accommodate emergency patients. A separate queuing analysis of the average number of transfer calls by hour and day of the week will display expected service demands. If your system uses the same crews for both emergency and non-emergency calls, the average call volume analyzed by this model must reflect both types of service, and you will have to recalculate the average treatment time.

No-transport emergency calls have had a negligible effect on our emergency staffing needs. They usually comprise 30 percent of all lights and siren emergency responses and average only seven minutes per call. Due to the minimal average length of service, no-transport calls rarely become significant. However, they should be analyzed quarterly to confirm that their rate of occurrence and length of service time has not changed.

A Quantitative Analysis of EMS Performance

Because emergencies clearly take priority in staffing any EMS data model, we will use the queuing theory model for a hypothetical ambulance service that responds to emergencies before it responds to transfers. We will examine a single hour of the week as an example of how all hours should be analyzed.

Between six and seven o'clock on a Monday morning, the data indicates that the ambulance company responds to an average of two calls per hour. Due to its historically heavy call volume that occurs later in the morning, this ambulance service has chosen to have five units on duty for the six o'clock hour, "just to be ready." You, the manager, want to determine if the queuing theory will indicate which, if any, changes are necessary.

Because the average service capacity of a single unit is 1.46 patients per hour, the service capacity of all five units is 7.3 patients (five units x 1.46 single-unit service capacity). The system's utilization factor would be determined by dividing the two-call average for the hour by its 7.3 patient service capacity. The .274 value indicates that the five units were actively responding to patient demands only 27 percent of the time (.274 rounded and converted to percent). Stated another way, 73 percent of the crew time is idle!

See formula 1 for calculating average system utilization.

Although efficient schedules and operations are important, idle crews and vehicles are not our primary concern. Remember that the patient's life and our economic livelihood depends on how long our patients have to wait for our service to dispatch an ambulance. To determine the average time patients must wait, we must first determine the average number of patients waiting for an ambulance.

Before working through the formula for estimating the average number of patients waiting, two mathematical terms, P-zero and exclamation point, need some further explanation. As stated earlier, there is the possibility that no patients will call for an ambulance during the six o'clock hour. The Poisson distribution allows for that fact in the value P-zero, that is, the probability that zero patients will be receiving service at any point during a given hour. Most operations research textbooks will contain a table in the appendix that will allow you to determine the value of this factor (for example, Anderson, 1982, p. A111). In our hypothetical system with an average service demand of two patients per hour, a system utilization factor of .274 with five units on duty would have a tabular P-zero value of .2590. P-zero tells you that there is a 26 percent probability that there will be no patients needing service at the given point in time between six and seven o'clock.

The other symbol is an exclamation point (!). In addition to the ordinary exponents, the formula uses a "factorial," which is a number with an exclamation point (5!). This factorial of five means 5 x 4 x 3 x 2 x 1, which equals 120. (Factorials grow fast. Try 7! on your pocket calculator – it equals 5,040.)

Now, to determine the average number of patients waiting, you will have to use a complex division process (see formula 2).

The P-zero value and utilization value are already known. The last value in the numerator is determined by dividing the average number of calls (2) by a single unit's service capacity (1.46 patients) (2/1.46 = 1.37). The 1.37 value is raised to the fifth power. (The number times itself five times equals 4.82. Verify on your pocket calculator.) The power (exponent) you use is directly equivalent to the number of units on duty. Four units would require the fourth power. The numerator by itself now appears like this:

$$(0.2590)\ (0.274)\ (4.82) = 0.342$$

The denominator first contains the factorial 5!. This value is also directly related to the number of units on duty. The second part of the denominator is simply one, minus the utilization factor of .274 and

[illegible]

System utilization = [illegible] per hour = average number of service calls / service capacity = 2.0 calls / 1.46 calls/hour × five vehicles = .274

[illegible]

Average number of patients waiting = (P zero) [illegible] / [illegible]

equals .726. That value is "squared" or multiplied times itself to equal .527. The denominator by itself appears like this:

(120) (0.527) = 63.24

The entire formula is now .342 divided by 63.24 and equals the seemingly inconsequential number of an average of .0054 patients who are waiting for an ambulance.

$$\frac{0.342}{63.24} = 0.0054$$

A final formula will allow us to determine the average amount of time a patient will have to wait for an ambulance between six and seven a.m.

The formula for calculating the expected waiting time is:

$$\frac{\text{Average number of patients waiting}}{\text{Average number of patient calls}}$$

Take the average number of patients waiting (.0054) and divide by the average number of patients who need an ambulance (2). The result of .0027 is a fraction of an hour, .16 of a minute, and we finally conclude that our patients will wait an average of 9.7 seconds for the dispatcher to find and dispatach an available ambulance.

The queuing theory has demonstrated that the historical staffing pattern that places five units on duty on Monday morning is certainly meeting our patients' needs. However, the thinking manager may have concluded early in this analysis that five units were more than adequate for a two-call average. The question is, how do you determine the degree to which staffing can be reduced without risking the well-being of your patients?

Sensitivity Analysis

Sensitivity analysis is a widely used method for determining the impact of changes in the input data values on the recommended solutions. That is, to what extent will the average patient's waiting time for an ambulance change if the number of units on duty is reduced from five to three or four.

We have used several average values in the queuing model up to this point. We must avoid the conclusion that the single answer is correct and therefore need to assess a range of solutions. We will systematically test alternative unit staffing assignments before making changes on the current number of ambulances in service.

An initial test of the model under the sensitivity analysis approach may propose that an average service demand of two calls per hour could be covered by three units. It would seem logical that one unit in reserve would cover unexpected calls. The queuing theory model's formulas will show that three units, which are available to handle an average of two calls per hour, will result in markedly different values for the important average performance characteristics such as system utilization, patients waiting, and time waiting. Expected results from using the model for three units are shown below next to the original results for five units.

5 Ambulances on duty	
Utilization	= 27.4%
Patients waiting	= .0054
Time waiting	= 9.7 sec.
3 Ambulances on duty	
Utilization	= 45.7%
Patients waiting	= .16
Time waiting	= 4.79 minutes

Why is the change so dramatic? Queuing theory incorporates the fact that nonproportional returns to scale occur in waiting-line situations. If you increase available ambulances from five units to 10 units, you do not see the expected drop in waiting time of 50 percent. The decrease in waiting time is 90 percent! The reason is that the mathematics of queuing situations show disproportional effects on performance from a *one* unit change in availability.

What then is an appropriate unit staffing and patient waiting time for six o'clock? Five units give a more adequate waiting time but indicates some wastefulness. Three units obviously produced an unacceptable average waiting time of almost five minutes before the system was able to respond to a call. The final test of the model under sensitivity analysis is with four units on duty, and it results in the following values.

5 Ambulances on duty	
Utilization	= 27.4%
Patients waiting	= .0054
Time waiting	= 9.7 sec.
4 Ambulances on duty	
Utilization	= 34%
Patients waiting	= .03
Time waiting	= 53 seconds

Sensitivity analysis has produced an acceptable decrease in idle staffing while maintaining an average waiting time for an ambulance below one minute.

Why should the one-minute waiting time be considered an acceptable value? A system that waits one minute or less to have a unit available for dispatch meets a clinically sound criterion as well as representing managerial efficiency. This standard has produced successful operational performance over the past two years. Moreover, the "one minute for dispatch" standard certainly has logical appeal to senior managers who have reviewed the model.

The reason this set of procedures is called sensitivity analysis is that a manager needs to know how "sensitive" critical standards of system utilization and average patient waiting time are to changes in controllable factors such as the number of ambulance units on duty. If waiting time remains below the one-minute standard after changes in the number of available units, then a manager need not be concerned with the proposed alteration to the schedule. This sensitivity analysis shows that a single-unit change can have a disproportionate and rapidly increasing impact on ambulance response.

Implications for Managers

At this point it is useful to stop and consider what all of this means for the manager. Why should you consider such a fundamental change in operational decision procedures? How do you proceed with implementation? What is the value of doing this considerable task?

You should attempt to replicate the queuing model for two main reasons. First, it successfuly prescribes staffing needs for our mid-sized system that has an annual call volume of 33,000. Moreover, it has demonstrable face validity. It has not produced results that in any way conflict with the experience the field crews report as heavy or light hours. Second, managers generally seek quantitative supplements for their intuitive decisions. With queuing theory, staffing is now varied using a reliable, quantitative model rather than by subjective, non-verifiable suggestions that are little more than "hunches."

Implementation of the model requires an adequate computer and several decisions relating to the distribution of calls and service time. Data on call volume must be maintained by the hour for each day of the week for emergencies, transfers, and no-service calls. The average length of a call must be calculated monthly for each call category. Data should be kept in monthly aggregates that can be compared for three months to check trends. Monthly and seasonal trending should be reviewed before settling on a weekly trend.

You will find it necessary to test your data to determine if patient call rates and ambulance service times fit a Poisson distribution. This is required to utilize the queuing model that has been presented.

Finally, the model will not yield its most reliable results without the benefit of a systematic sensitivity analysis, which will demonstrate to crew members and upper management alike that the recommended changes are appropriate.

The value of applying the model will be found at several levels in the organization and its environment. Two primary benefits will accrue. Response times will improve as resources are shifted from hours in which they are under utilized to hours in which they will be more frequently utilized. In addition to a decrease in response times, the overall costs for each service response will be lowered due to the fact that idle crews and equipment can now be shifted to potentially higher periods of usage. This will lower the stress on overworked crews and equipment.

These two primary benefits are valued differently by the multiple constituencies to which the manager must respond. The patients will appreciate the faster response time and should be told that the new efficiency in service will not increase their costs. Your employees will note that they have more help when busy hours occur and consequently less idle time. The physicians, hospitals, nursing homes, and government agencies that you serve should find the increased efficiency apparent in response times and will be impressed by the fact that changes are appropriately justified and more thoroughly analyzed. Finally, your success in implementing this program will be advantageous to your professional career and the overall success of the EMS system that employs you!

A Spreadsheet for Queuing Theory

On a monthly basis, 24 hours of data must be analyzed for emergency and non-emergency calls for each day of the week. (Don't do the math—that's 336 series of formulas!)

The first tests of the queuing formulas were worked out by hand on a programmable calculator, and for quick checks on data the calculator remains useful. For 336 formula series per month, the spreadsheet [illegible] invaluable!

Any spreadsheet program with "table" look-up capability [illegible]

Columns [illegible] and rows are numbered [illegible] row given to titles. The [illegible] are listed first with explanations.

Col
A = hour (0100 through 0000)
B = sum (all calls for that hour and day of the week)
C = number of days (number of Sundays, etc., in the month)
D = average call (divides the sum by number of days)
E = number of units (units available for call)
F = S! (refers out to a factorial values table based on the number of units)
G = utilization (picks up an average service rate value and multiples by the number of units, then divides into the average calls to show utilization)
H = P-zero (goes to a second table for the probability of zero calls value)
I = number waiting (calculates the number of patients waiting for an ambulance)
J = seconds waiting (calculates the seconds needed to find an ambulance)

The following list of cell contents was taken from the cells in row two and represent calculations for 0100 hours. The bracketed words [. . .] are commentary only and not part of the cell contents. If the notation is unfamiliar to your spreadsheet format, I suggest working through the commentary before beginning the program.

Col
A = 0100 hours
B = [illegible]
C = [illegible]
D = [illegible]
E = [illegible]
F = [illegible]
G = [illegible] [illegible] placing the number of calls your units can make per hour in cell A2[illegible].]
H = @TBL(G2,E2,$AA2:AN[illegible]) [The probability of zero calls is located at cell AA2 and contains 952 cells. That's why you need a spreadsheet.]
I = (H2 × ((D2/$A16)^E2 × G2)/(F2 × ((G2)^2)) [The number of patients waiting for an ambulance. The extra parentheses keep the calculations orderly.]
J = (I2/D2) × 3600 [Finally we get to the time waiting for an ambulance, The value is in hours, and the 3600 converts it to seconds.]

The number of units column should remain static for each day of the week that you need to calculate. I move the whole column out to an unused section of the spreadsheet in order to save input time each month. I would also advise backing up this program on several disks and keeping them in several safe locations—no one wants to reprogram 1,100 cells. □

Conclusion

Patients are best cared for by EMS systems that take advantage of every opportunity to commit their resources in the most effective and efficient manner. Queuing theory and sensitivity analysis will give EMS managers one set of tools necessary to accomplish this managerial goal. □

References

Anderson, MQ: *Quantitative Management Decision Making.* Monterey: Brooks/Cole, 1982; Atll.

Cooper, RB: *Introduction to Queueing Theory.* New York: North Holland, 1981.

Lee, AM: *Applied Queuing Theory.* London: MacMillan, 1966.

Panico, JA: *Queuing Theory.* Englewood Cliffs: Prentice-Hall, 1969.

Recommended Reading

Gorney, L: *Queuing Theory.* New York: Petrocelli, 1981.