Simple Queuing Theory Tools You Can Use in Healthcare

Jeff Johnson
Management Engineering Project Director
North Colorado Medical Center

Abstract

Much has been written about queuing theory and its powerful applications. But only recently have healthcare professionals discovered the benefits of applying queuing theory techniques. Many have been discouraged by the mathematical mystery shrouding queuing theory. Surprisingly there are some simple queuing functions. The functions I use in this paper have been written for Excel by the University of Alberta's School of Business. These functions allow anyone to apply the techniques for better process understanding leading to better decision making and optimization of your healthcare budget.

Staffing numbers, patient arrival rates, and patient service times are the key data to collect from processes when applying queuing theory techniques. Even if this data isn't as accurate as you would like, you can still perform "what if" scenarios for any combination of the appropriate data. The simple Excel functions are the basis for creating data tables and graphs which show the effects of changing staffing, arrival rates, and service times on utilization, wait times, and time patients spend in the "system". For example, utilization and wait time can be estimated based on the current staffing, arrival rates and service time. Then you can determine if you can change staffing levels and still have acceptable utilization and wait times or you can determine if you need to focus on improving your patient service times to meet necessary utilization or wait time.

Description of a Queuing Problem

A queuing system can be described as patients arriving for service, waiting for service if it is not immediate, utilizing the service, and leaving the system after being served.

Characteristics of Queuing Process

- Patient arrival distribution or pattern
- Patient service distribution or pattern
- Number of servers
- Capacity of system
- Queue discipline

Patient Arrival Distribution or Pattern

In queuing situations it is necessary to estimate the probability distribution or pattern of the arrival times between successive patient arrivals (inter-arrival times). What is the reaction of the patients when they arrive? Will the patients wait, no matter what the length of the queue is? Will they balk, leave immediately because the line is too long, or will they renege, leave after a period of time? And finally, does this distribution or pattern change over time?

Patient Service Distribution or Pattern

The probability distribution or pattern of service times is also important to understand. The distribution may depend on the number of patients in line

(the server may work faster as the line becomes longer) or the experience of the server. (experienced servers may work faster) The patient arrival pattern and the service pattern are assumed to be independent, that is, one doesn't depend on the other.

Number of Servers

This refers to the number of servers available for patients to use simultaneously. Servers can be fed from a single line or queue or multiple lines or queues. Waiting in line at an airport for an agent is an example of the former situation while waiting in multiple lines at a grocery store is an example of the latter.

Capacity of system

The capacity refers to the physical limitation of the system such as a waiting room. When the waiting room is full the next patient must leave since there is not enough room to wait.

Queue Discipline

The discipline describes the manner in which the patients are served after a queue has formed. The common disciplines are FCFS – First Come First Served, LCFS – Last Come First served and RSS – Random Selection Service.

Model Notation

Queuing processes or models are described by a series of symbols and slashes.

A/B/X/Y/Z

The letters denote the following:

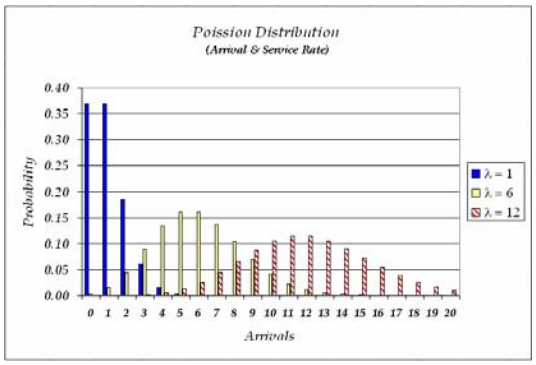
A – Patient Arrival Distribution
B – Patient Service Distribution
X – Number of Servers
Y – System Capacity
Z – Queue Discipline

Based on this notation it is evident that queuing theory has the capability of handling complex and fairly sophisticated models. But for most of the work I have been involved with, simple models are adequate to provide all of the information I need to make proper decisions.

Using the notation described above the most common queuing models are the M/M/c/infinite/FCFS. The M refers to a Markovian process which assumes the arrival or service rate follows a Poisson distribution and the time between arrivals or service time follows an exponential distribution. The c refers to the number of servers which in the simplest case is one. For this system the system capacity is infinite and queue discipline is FCFS (First Come First Served). These last two terms are typically left off the notation so the model is written as M/M/c.

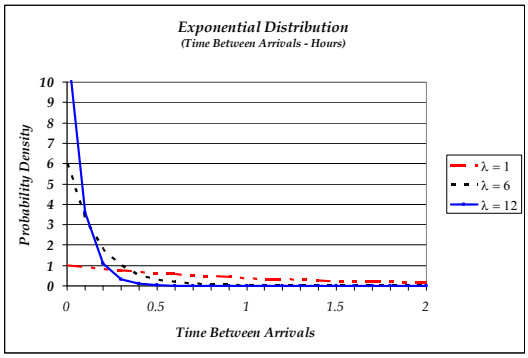
Explanation of Distributions

The Poisson distribution provides a realistic model for many random phenomena. In healthcare this could be arrival rate to the ED or requests for service from the transport department. The Poisson distribution is characterized by the mean and variance being equal. Distributions with three different means are plotted on the next page. (Graph 1)



Graph 1 – Poisson distribution

The Exponential distribution is an appropriate model for time between events when a process exhibits a Poisson distribution for the occurrence of the event. So the time between patient arrivals or transport request would follow an exponential distribution. The exponential distribution is characterized by the mean and standard deviation being equal. See Graph 2.



Graph 2 – Exponential distribution

Measurements of Interest

Generally there are two items of interest.

1) Utilization – How much of the time are the servers busy or working?
2) Wait Time – How long do patients have to wait for service?

These two items generally conflict. As utilization (better use of your resources – staff) increases, wait time increases also (more time waiting for or patient) and alternately as utilization decreases wait time decreases.

The tools described in this paper allow you to easily see the trade-offs between utilization and wait time and allow you to make better decisions related to staffing and potential waiting times for patients.

Queuing Notation

The key equation for queuing is as follows:

$\rho = \lambda / (c * \mu)$

ρ = utilization
λ = arrival rate
c = number of servers
μ = service rate per server

By obtaining estimates of the actual arrival and service rate and knowing the number of servers in the system we can easily calculate the utilization.

This information and addition equations allows us to calculate other useful measures of the system such as:

- W & W_q – Average wait time in system or queue
- L & Lq – Average number of patients in system or queue
- p_0 – Probability of all servers being idle
- p_n – Probability of n patients in the system at any time.

Fortunately it isn't necessary to make all of these calculations (see appendix) because they have been programmed into the Excel functions. We just need to use the appropriate function,

understand the results of the output and present the output in a form that is most useful for those making the decisions.

The formulas can be obtained free of charge at the following website:

http://www.bus.ualberta.ca/aingolfsson/QTP/

These are the functions I used most often, although the appendix lists additional functions for additional measurements and uses the general distributions assumption instead of the exponential distribution.

- QTPMMS _ Util - Average server utilization
- QTPMMS _ L – Average number of patients in the system.
- QTPMMS_Lq – Average number of patients in the queue.
- QTPMMS_W – Average time a patient spends in the system.
- QTPMMS_ Wq – Average time a patient spends in the queue.
- QTPMMS_PrEmpty - Probability that the system is empty of patients.
- QTPMMS_ PrFull – Probability that the system is full.
- QTPMMS_PRWait – Probability that a patient has to wait before receiving service.

All of these Excel function require four input values in the following order:

- Arrival Rate
- Service Rate
- Number of Servers
- Queue Capacity

Note: Be careful that the units for arrival rate and service are the same. (i.e. If arrival rate is in patients per hour the service rate should be in patients per hour also. Often service is expressed in minutes and must be converted to patients per hour.) Queue capacity is assumed infinite for most applications so putting a large number into that function location is sufficient.

Process for Data Sheet

For the best results, estimates of the arrival rate, service rate and current number of servers should be obtained. With that information you can begin to build your data table (Table 1) by varying your current estimates by appropriate amounts to cover the ranges that you could expect to see with your process. I would suggest having 5-7 values for each input in the table. In my example the arrival rate varied from 3 – 15 in increments of 3. The service rate varied from 10 – 30 in increments of 5. The number of servers varied from 1 – 7 in increments of 1.

The 1st three columns of the Excel data table will have the values noted above. Every combination should be listed. The 4th column is the maximum queue size.

Additional columns in the data table calculate the queuing statistics that are of interest of you. They range from utilization to average waiting (queue) time. I've included all of the functions listed in this paper. By copying the queuing formulas in the cells of the data table down the columns the data table is relatively easy to build.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Requests (per hour)	Service Time (minutes)	Servers	Max Queue Size	Effective Utilization	Average number of requests in queue	Average number of requests in system	Average time in queue (minutes)	Average time in system (minutes)	Probability of empty system	Probability of full system	Probability of having to wait	Usually Calculated Utilization
147	3	30	6	10	25%	0.0	1.5	0.0	30.0	22.3%	0.0%	0.5%	25%
148	6	30	6	10	50%	0.1	3.1	1.0	31.0	4.9%	0.0%	9.9%	50%
149	9	30	6	10	75%	1.0	5.5	6.9	36.9	0.9%	0.6%	41.1%	75%
150	12	30	6	10	93%	4.0	9.6	21.5	51.5	ation >=	7.3%	79.9%	100%
151	15	30	6	10	99%	6.8	12.7	34.2	64.2	ation >=	21.0%	96.0%	125%
152	3	10	7	10	7%	0.0	0.5	0.0	10.0	60.7%	0.0%	0.0%	7%
153	6	10	7	10	14%	0.0	1.0	0.0	10.0	36.8%	0.0%	0.0%	14%
154	9	10	7	10	21%	0.0	1.5	0.0	10.0	22.3%	0.0%	0.1%	21%
155	12	10	7	10	29%	0.0	2.0	0.0	10.0	13.5%	0.0%	0.5%	29%
156	15	10	7	10	36%	0.0	2.5	0.0	10.0	8.2%	0.0%	1.5%	36%

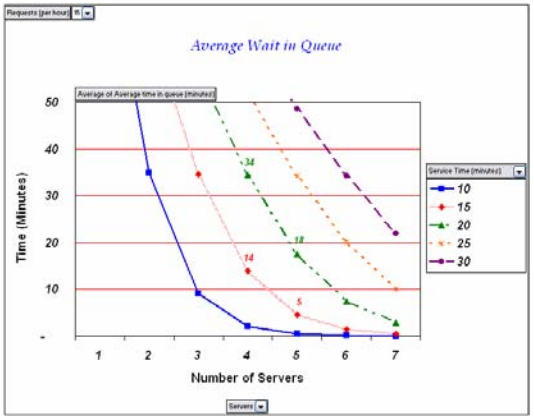
Table 1 – Part of the Data Table

Using the data table shown on the previous page pivot tables (Table 2) can be created that allow you to easily create graphs for analysis.

Requests (per hour)	15					
Average of Average time in queue (minutes)	Service Time					
Servers	10	15	20	25	30	Grand Total
1	93	145	195	245	295	195
2	35	66	93	119	145	92
3	9	35	57	76	93	54
4	2	14	34	52	66	34
5	1	5	18	34	49	21
6	0	1	7	20	34	13
7	0	0	3	10	22	7
Grand Total	20	38	58	79	101	59

Table 2 - Pivot Table (Wait Time)

In this example, multiple pivot tables were generated which allowed me to create multiple graphs. I kept the format the same for ease of interpretation (Graph 3). The number of servers is always the x axis and the statistical value of interest is always the y axis. The multiple lines represent the different service times and the field button (top left) is always the number of requests.



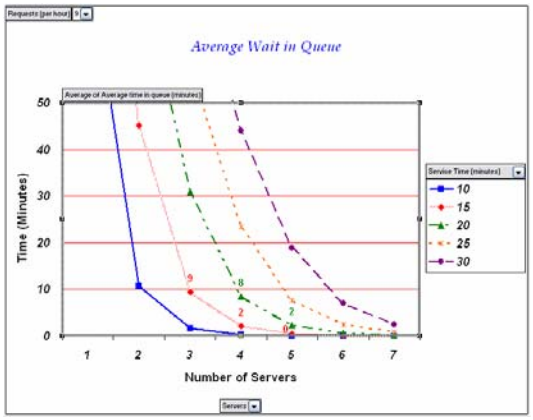
Graph 3 - Graph for Average Wait Time (15 requests per hour)

The graphs are interactive and allow you to perform "what if" scenarios so you can easily see the effect of changing service time, servers or incoming requests on the queuing parameter of interest.

In the previous graph the effect of reducing servers from 5 to 4 when the service time averages 20 minutes increases the wait time from 18 to 34 minutes (green dash & dotted line). When the service time averages 15 minutes the wait time increases from 5 to 14 minutes (red dotted line). This is when the average number of requests per hour equals 15.

So if you have a 20 minute service time and are interested in reducing the servers from 5 to 4 the patients' wait time will nearly double (18 to 34 minutes). On the other hand if you reduce your service time to 15 minutes you can reduce servers to four and the patients actually wait for a shorter amount of time (14 minutes).

If the average number of requests changes from 15 to 9 the wait time changes drastically. (Graph 4) If the server time is decreased to 15 minutes the wait time is 9 minutes with only 3 servers (red dotted line).



Graph 4 – Graph of Average Wait Time (9 requests per hour)

Similar graphs show utilization, the probability of waiting, and the probability of empty or full systems.

As the requests vary throughout the day a different model may be more appropriate. This requires continual monitoring if the optimal trade off between patient service (wait time) and cost (number of servers) is desired.

Conclusion:

Queuing theory is not new but only recently has healthcare begun to use it effectively. The queuing theory graphs are simple yet powerful tools created with the help of simple Excel functions that allow the user to more easily interpret data by looking at different scenarios quickly, accurately, and easily.

Acknowledgements

- Twila Burdick – Banner Heath
- David Datz – Banner Health
- Jimmy Broyles – Arizona State University

References

- Fundamentals of Queueing Theory by Donald Gross & Carl M. Harris
- University of Alberta School of Business
- Introduction to the Theory of Statistics by Alexander M. Mood, Franklin A. Graybill & Duane C. Boes
- Quantitative Analysis for Management by Barry Fender & Ralph M. Stair, Jr.

Biographical Sketch

B.A. Mathematics – University of Colorado
M.S. Statistics – Colorado State University

After graduating from college I worked as a mathematics and statistics instructor at Northern Arizona University in Flagstaff, Arizona for two years. I then moved to Fort Worth, Texas and worked as a test engineer for two years at the General Dynamics Corporation. Moving to Colorado I worked as a statistician and product engineer for the Eastman Kodak Company for 21 years. A year and a half ago I left the manufacturing world and moved into the healthcare field, where I currently work as a Management Engineer for North Colorado Medical Center in Greeley, Colorado.