Queuing Theory I
© 2006 Samuel L. Baker

Queuing theory is the analysis of waiting lines, or "queues".
The goal of this unit of the course is to acquaint you with the existence of queuing theory, and to show what kinds of assumptions underlie its results.

Queue = waiting line

A queue is a waiting line. The elements of a queue are
1.   Arrivals that need service of some kind,
2.   Service facilities that take care of the arrivals,
3.   The queue, where the arrivals wait until they can be serviced.

Here are some examples of queues:

Arrivals	Servicing Facilities, or "Channels"	The queue
Shoppers	Clerks	Checkout line
Patients	Doctor	In waiting room
Patients	Operating teams	Waiting list
Customers	Stock	Back orders
Machine breakdowns	Repair persons	Broken machines
Finished goods	Dealer	Inventory
Pharmaceuticals	Pharmacy in hospital	Inventory
Airplanes	Runways	Stack in air
Telephone calls	Circuits	Uncompleted calls
Fires	Fire fighters	Burning buildings
Automobiles	Intersection	Traffic jam
Streams, rain	Water users	Reservoir

Queues can vary according to:

Arrival factors:
1.   arrival distribution -- when new customers arrive
2.   size of population from which customers are drawn
                Is the population effectively infinite or
                is it small enough that each arrival means one less new customer in the future?

Service factors:
1.   service time distribution -- how long it takes to serve an arrival
2.   number of channels, such as how many checkout counters at the grocery store
3.   how many stages of service there are

Queue factors:

1.   how many queues
              one or more than one
2.   service priority among customers
              first come first served is most common
              other possibilities: random, customers with shorter service time go first,
              triage and priority assignment, with or without preemption of customers already being
                   served
3.   customer behavior in queue
              balking -- customers don't join line that's too long
              reneging -- customers quit line if wait gets too long
              bounded queue -- waiting list can only be so long

Some flow charts of queues:

Single-Server Single-Stage Queue
```
      Arrivals ---> Queue ---> Service ---> Done
```

Several Single Server Single Stage Queues in Parallel
```
      Arrivals ---> Queue ---> Service ---> Done
      Arrivals ---> Queue ---> Service ---> Done
      Arrivals ---> Queue ---> Service ---> Done
```

Multiple Server Single Stage Queue
```
                              ┌-> Service ---> Done
      Arrivals ---> Queue ----├-> Service ---> Done
                              └-> Service ---> Done
```

Single Server Multiple Stage Queue
```
      Arrivals ---> Queue ---> Service ---> Service ---> Service ... --
-> Done
```
       The Multiple Stage model is like a cafeteria, in which customers waiting to get their main dishes
       may prevent customers behind them from getting their salads.

Single Server Queues in Series
```
      Arrivals ---> Queue ---> Service ---> Queue ---> Service ... --->
Done
```

       The Queues-in-Series model has queues between the stages of service, which the Multiple Stage
       model does not. This might be appropriate for a doctor's office. Patient wait in the waiting room
       until seen by a nurse. Then they wait again in examining rooms until the doctor arrives.

Poisson Queues

A Poisson queue is a queuing model in which the number of arrivals per unit of time and the number of
completions of service per unit of time, when there are customers waiting, both have the Poisson
distribution.

The Poisson distribution is good to use if the arrivals are all random and independent of each other.

For the **Poisson distribution**, the probability that there are exactly x arrivals during t amount of time is:

The Poisson
Distribution

$$Prob\,(Arrivals{=}x) \;=\; \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

t is a duration of time. Its units are, e.g., hours or days.
λ (Greek letter lambda) is the expected (average) number of arrivals per hour, or day, or whatever units t is measured in.
λt is therefore the expected number of arrivals during t amount of time.
x is a possible number of arriving customers.
x! ("x factorial") means $1\times2\times...\times(x-1)\times x$. For example, $5! = 1\times2\times3\times4\times5$. We define $0! = 1$.

If arrivals are distributed according to the above formula, then we say "the arrivals are Poisson" or "have the Poisson distribution."

Spreadsheet Implementation

Here is a spreadsheet layout for calculating values for the Poisson distribution. Detailed explanations of the formulas will follow.

	A	B	C
1	Lambda	4	
2	t	2	
3	Lambda*t	=B1*B2	
4			
5	x	P of x	P of <=x
6	0	=EXP(-B3)*B3^A6/FACT(A6)	=B6
7	=A6+1	=EXP(-B3)*B3^A7/FACT(A7)	=B7+C6
8	=A7+1	=EXP(-B3)*B3^A8/FACT(A8)	=B8+C7
9	=A8+1	=EXP(-B3)*B3^A9/FACT(A9)	=B9+C8
10	=A9+1	=EXP(-B3)*B3^A10/FACT(A10)	=B10+C9
11	=A10+1	=EXP(-B3)*B3^A11/FACT(A11)	=B11+C10
12	=A11+1	=EXP(-B3)*B3^A12/FACT(A12)	=B12+C11

To show the formulas, I set the cells' numeric format to Text. You do not need to do that for your work. Instead, you will want the cells to display the numbers, which they do by default.

For illustration, I've picked a λ of 4 arrivals per hour expected, and a t of 2 hours. This makes P of x the probability that exactly x people will arrive in the next 2 hours, given that the expected number of arrivals is 4.

P of <= x, in column C, is the probability that x or fewer people will arrive in the next two hours.

Column A has the values of x. I use formulas like =A6+1, rather than typing in 1,2,3,..., to make it easier to extend the table down with one copying operation.

Here's an explanation of the formula =EXP(-B3)*B3^A6/FACT(A6) in B6:

EXP is the Excel function for raising e to a power. =EXP(-B3) means $e^{-\$B\$3}$, which is $e^{-\lambda t}$, since

-λt is in B3. The dollar signs keep the cell reference on B3 when the formula is copied and pasted to another cell.

The caret symbol ^ is for raising a number to a power. B3^A6 means (B3)^{A6}, which means $(λt)^x$.

The FACT function (@FACT in Quattro Pro) calculates factorials. Here, it is calculating x!, the factorial of the x value in column A.

The formula in the C column, starting with C7, gives a running total of the P of x's. Each cell is the cell above, which is the previous total, plus the cell to the left, the current P of x.

When you are constructing the table of probabilities in the spreadsheet:
1.   Fill in row 6 as shown.
2.   In A7, type =A6+1 as shown.
3.   Copy cell B6 and paste to B7.
4.   In C7, type =B7+C6 as shown. That should give you row 7 complete as in the diagram above.
5.   Copy A7:C7. Paste to a range starting in A8 and extending down the A column as far down as you want. (For the homework, you'll want to go down about twenty rows.)

	A	B	C
1	Lambda	4	
2	t	2	
3	Lambda*t	8	
4			
5	x	P of x	P of <=x
6	0	0.000335	0.000335
7	1	0.002684	0.003019
8	2	0.010735	0.013754
9	3	0.028626	0.04238
10	4	0.057252	0.099632
11	5	0.091604	0.191236
12	6	0.122138	0.313374
13	7	0.139587	0.452961
14	8	0.139587	0.592547
15	9	0.124077	0.716624
16	10	0.099262	0.815886
17	11	0.07219	0.888076
18	12	0.048127	0.936203
19	13	0.029616	0.965819
20	14	0.016924	0.982743
21	15	0.009026	0.991769
22	16	0.004513	0.996282
23	17	0.002124	0.998406
24	18	0.000944	0.99935
25			

Here are the numbers I see after I copy A7:C7 to A8:A24.

As mentioned, Column B, rows 6 and below, gives us probability of x events happening, for values of x from 0 to 18. For instance, cell B14 tells us that the probability of exactly 8 arrivals in the next two hours is 0.139587.

Column C is a running total of B, the probability that x or fewer events happen, for values of x from 0 to 18. Cell C14 tells us that the probability of 0 through 8 arrivals is 0.592547.

As x increases from 0, the probability of exactly x events rises at first. The probability is highest at x = λt-1 and x = λt, which in this example is x = 7 and x = 8. For higher values of x, the probability of x diminishes, approaching 0 as x gets very large.

As x increases, the probability of x or fewer events rises, approaching 1 as x gets very large.

The Poisson distribution is skewed. It's not symmetrical about its mean. The probability of 7 or less arrivals (.453) is greater than the probability of 9 or more arrivals (.407 = 1-.593). The average of the distribution is 8, because of the possibility of a very large number of arrivals.
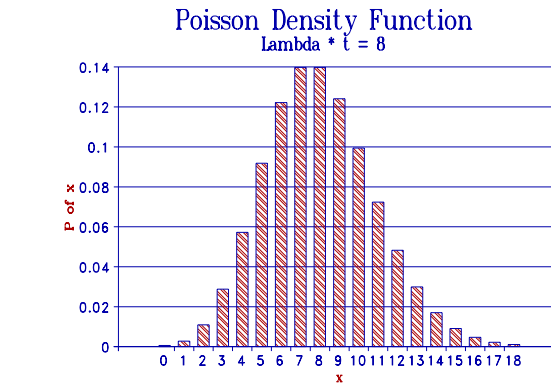
This
is a graph of the P of x numbers in column B.
The shape is a skewed bell. The right tail is larger
than the left tail.

The degree of skewness depends on λt. The
larger λt is, the more the Poisson distribution
approaches the normal distribution, which has a
symmetrical bell shape.

The Poisson distribution is a limiting case of the
binomial distribution. That means that you are
making two important assumptions when you use
the Poisson:

1. **λ is constant.** The expected arrival rate
doesn't change over the period.

*The Poisson distribution is skewed – not
symmetrical.*

2. **Independence**, which means **No Memory** and **No Groups**. Customers do not tend to arrive in
groups. The fact that somebody just walked in doesn't make it any more or less likely that somebody else
is coming soon. If nobody came or if ten times λ people came during the last time period, the expected
number of people during the next time period is still just λ.

If and only if those two assumptions hold, then the number of events (arrivals) in any specified time
period has the Poisson distribution.

Poisson application -- a model of the need for inventory

Here is an example of how the Poisson distribution can be applied to model how much inventory is
needed. The example explores the effect of a Medicaid policy change on how much inventory
pharmacies need to keep.

Inventory is goods on hand, waiting to be used. Inventory is what retail trade is all about. The store
keeps an inventory of goods so that customers can walk in at times convenient to them and get what they
want.

South Carolina Medicaid has controlled spending on pharmaceuticals with, among other policies, limits
on how many active prescriptions a Medicaid recipient can have. At one time, the limit was three
prescriptions per month. This was problematic for patients with multiple problems, particularly chronic
psychiatric problems, so the policy was changed to allow pharmacists to sell 3-months of doses at once
per prescription, rather than limiting them to selling 1-month of doses at a time. This changed allowed
patients to have up to nine prescriptions active at once, rather than just three, if they could plan well
enough to rotate their purchases. (Since then, there have been more policy changes.)

Pharmacists complained that this new scheme would cost them money. Why? Inventory is the answer.
Let's explore this with a simplified model that shows what's going on.

We use the Poisson distribution to model how often people come to the drug store to buy a prescription.

We can see how much inventory the pharmacists will need under different assumptions about the size of purchases and how often they happen.

Best for the pharmacist, as far as inventory need is concerned, would be if the purchases were on a strict schedule. If the pharmacist knows when every customer is going to appear and what the customer will buy, little or no inventory is required. The pharmacist could arrange to have the drugs delivered just before the patient comes in to buy them.

If the patients' purchases are randomly timed, however, inventory is required. How much inventory? The store does not want to turn away customers, so it wants enough inventory to make running out unlikely. No level of inventory is enough to be 100% sure of never running out, but the more inventory a store has, the less likely it is to run out. There is a cost to inventory, however. Goods on the shelf have to be bought. Inventory is capital. The more inventory a store keeps, the higher is the store's cost of operating.

We will assume for this example that the pharmacist will settle for a 95% probability of having enough stock on hand to satisfy the customers. In other words, the pharmacist will tolerate a 5% chance of running out on any given day.

To simplify the numbers as much as possible, suppose that a pharmacy's customers switch from buying one bottle of pills per month to buying three bottles at once, but only once every three months. Suppose that the pharmacy has enough customers so that it sells an average of three bottles per day, both before and after the policy change.

We also assume that the pharmacy gets one shipment per day. Every night, after closing, the store places an order. The shipment arrives the next morning, before the store opens.

Here is the Poisson model for three 1-bottle sales per day. Lambda is 3, expected bottles sold per day. The time period t is 1, for 1 day.

```
          A          B          C          D
1    Lambda              x P of x   P of <=x
2           3           0 0.049787 0.049787
3    t                  1 0.149361 0.199148
4           1           2 0.224042 0.423190
5    Lambda*t           3 0.224042 0.647232
6           3           4 0.168031 0.815263
7                       5 0.100819 0.916082
8                       6 0.050409 0.966491
```

Look down the running total of the probabilities in column D and find the first number that is bigger than 0.95. That number, 0.966491, is in the row where x is 6, so 0.966491 is the probability that 6 or fewer customers will arrive. This means that the pharmacist needs to start the day with 6 bottles in inventory to make the probability of running out less than 5%.

Now let's model one sale per day, with three bottles in each sale. We still expect to sell three bottles per day, but now the sales are in three-bottle lumps. Lambda is now 1, because the expected rate of sales is 1 three-bottle order per day.

In the table below, x is the number of sales events. x is not the number of bottles in the table below, as it was in the table above, because each sale is a sale of three bottles.

```
        A          B          C          D
1    Lambda           x P of x    P of <=x
2            1       0 0.367879 0.367879      0 bottles
3    t               1 0.367879 0.735759      3 bottles
4            1       2 0.183940 0.919699      6 bottles
5    Lambda*t        3 0.061313 0.981012      9 bottles
6            1
```

Look down the running total of the probabilities in column D and find the first number that is bigger than 0.95. This time, it is 0.981012, in the row for x=3. The pharmacist has to be prepared for three sales events, if he or she wants at least a 95% chance of not running out. Three sales means nine bottles. The pharmacist needs to start the day with 9 bottles in inventory to make the probability of running out less than 5%.

Lumpy sales increases the inventory required from six bottles to nine. That adds capital cost, cutting into the profit that the pharmacist makes selling to Medicaid patients. That is why the pharmacists were unhappy with the change.

The Exponential Distribution

If the number of events during a specified period of time has the Poisson distribution, then the amount of time between events has what is called the **exponential distribution**.

The Exponential Distribution	$$Prob(next\ arrival\ in\ time\ \leq\ t) = 1 - e^{-\lambda t}$$

λ is the expected number of arrivals per unit of time, as before.

The Poisson and the exponential distributions are mathematically equivalent. They are two ways of looking at the same thing.

Here is an example of an exponential distribution calculation:
If λ is 1 arrival per hour, then the probability that the next arrival will be in less than 1 hour is:
$1 - e^{-1} = 0.632$.
This result may surprise you. Many people expect the answer to be ½. The actual probability is greater than ½ because the exponential distribution is skewed, as the Poisson is.

Another exponential distribution calculation example:

Suppose λ is 4. The probability that the next arrival will come in the next 0.25 and 0.5 hours is calculated by subtracting the probability that the next arrival will come sooner than 0.25 hours from the probability that the next arrival will come sooner than 0.5 hours.

$(1 - e^{-4 \times 0.5}) - (1 - e^{-4 \times 0.25}) =$
$\ \ .8646647 - \ \ .6321206 = \ \ .2325442$

The mean of the exponential distribution is $1/\lambda$. If we expect 4 events per hour, then the expected interval between events is ¼ of an hour. This does make intuitive sense.

Your turn:

Assignment 7

Suppose we expect 2 arrivals per hour and arrivals have the Poisson distribution.

1. What is the probability that no one will come during the next hour?

2. What is the probability that exactly 2 people will come during the next hour?

3. What is the probability that 1, 2, or 3 people will come during the next hour?

4. What is the probability that more than 18 people will come during an 8-hour day?

5. A charitable organization asks you to give a copy of their brochure to everyone who comes in during the next 8-hour day. How many brochures do they need to give you before the start of the day so that the probability is less than 0.01 that you will run out of brochures before the day is over?

Analyzing a Queue

A full queuing situation involves arrivals and service, so we need some more Greek letters:

λ "lambda" is the average customer arrival rate per unit time
μ "mu" is the average customer service rate (when customers are waiting)
         $\mu = 1/$(average service time)
ρ "rho" is λ/μ, also called the **server utilization factor**

We will use the Poisson and exponential distributions to model both arrival times and service times. As mentioned, the Poisson and exponential distributions are mathematically related. If the number of service completions per unit of time, when there is a backlog of customers waiting, has the Poisson distribution, then service time has the exponential distribution. It's conventional to think of service in terms of the length of service time.

The standard simple queuing model assumes that:
1.   arrivals have the Poisson distribution
2.   service times have the exponential distribution
3.   arrivals and service times are all independent.
      (Independence means, for example, that:
           arrivals don't come in groups, and
           the server does not work faster when the line is longer.)

Steady State

When a service first opens, the queue may have 0 length, or it may have some other length if people are lining up before the doors open. Once the doors do open, customers get served and new ones arrive. The queue will grow, if it started from 0, or it might shrink, if it started out long. Eventually, if the average arrival rate and service rate stay steady, the queue will tend to settle down to what is called a steady state. The steady state does not mean that the queue is of constant length. The queue varies as customers arrive or get served. What stays steady is the probability that any particular number of people are waiting. In the steady state, the variation in the queue length can be predicted without regard to how the queue started.

The steady state idea lets us calculate the probability that 0, 1, 2, etc., customers are in line at any given instant without worrying about how long the service has been in operation or how many people were there when the doors opened. We can also develop mathematical formulas for the average queue length and the average time spent waiting. These are presented on the following pages.

ρ must be less than 1 for there to be a steady state. Otherwise, customers are tending to arrive faster than they can be served. The queue will grow and grow. The formulas on the following pages give impossible results if ρ is 1 or bigger.

Steady state formulas for a <u>single-server single-stage queue</u> with
  Poisson distribution arrivals,
  exponential distribution service times,
  first-come first-served queue discipline,
  drawing on an infinite population (so the future arrival rate does not depend on the past arrival
    rate).

The probability of any given number of customers being in the system (waiting or being served):

$$Prob\ (0\ in\ system) = 1 - \rho$$

$$Prob\ (n\ in\ system) = (1 - \rho)\rho^n$$

L is the average number of customers in the system, waiting in queue or being served:

$$L = \frac{\rho}{1 - \rho} = \frac{\lambda}{\mu - \lambda}$$

L_q is the average number of customers waiting in the queue:

$$L_q = \frac{\rho^2}{1 - \rho} = L\rho = L - \rho$$

W is the average time customers spend in system, waiting plus being served:

$$W = \frac{1}{\mu - \lambda} = \frac{L}{\lambda}$$

W_q is the average time a customer spends in waiting in queue before service starts:

$$W_q = \frac{\rho}{\mu - \lambda} = \rho W$$

Some interpretation:

These formulas imply that you can tell how busy the server is just by watching the length of the line.
That's because L_q can be calculated from ρ, the server utilization factor, the proportion of the time that the
server is busy. For example, if you see that there two people in line, on average, then $L_q = 2 = \rho^2/(1-\rho)$.
Solve that for ρ , and you get $\rho = 0.732...$, so the server is busy almost three-fourths of the time. Be
careful about trying this is practice, though, especially if your impression of the length of the line is from
your casual observation. When you look, the line is more likely to be shorter than L_q than it is to be
longer than L_q. That's because the distribution is skewed. Every so often, the line will get very long, and
that's what brings the average length up to L_q.

Notice that L_q is L-ρ, not L-1. When there is someone being served, the number of people in the system is one plus the number in the queue, but the average number of people in line is L-ρ, which is more than L-1. This allows for those times when no one is being served or waiting. If that's hard to grasp, think of it the other way around: L is L_q + ρ, not L_q + 1, because when no one is in the system, the number of people in the system and the number of people waiting are both 0. The probability of no one in the system is 1 - ρ. So, 1 - ρ of the time, the system and the queue are the same (zero) and ρ of the time, the system has one more person it in (the person being served) than the number of people waiting. On average the number of people in the system is ρ more than the number of people waiting.

Spreadsheet Implementation

For how to implement these formulas in a spreadsheet, see the Queuing Theory Cookbook, which is a downloadable file.

Example of Queue Analysis

Suppose the data are:

One pharmacist works filling orders for medications.
10 orders come in per hour.
Filling the orders takes an average of 4 minutes each.
Assume that orders and fills have the Poisson distribution (so service times have the exponential distribution).

The first step is to calculate λ and μ.
λ is easy, because it's given. λ = 10 orders/hour.
μ has to be calculated. Filling an order takes an average of 4 minutes. There are 60 minutes in an hour, so the pharmacist can handle 15 orders per hour on average when busy. μ = 15 orders/hour. Notice how we want λ and μ to have the same units, orders/hour in this case. If one is in terms of hours, the other has to be in hours, too.

ρ = 10/15 = 0.6667. The server is busy two-thirds of the time.

Suppose two chairs are provided at the pharmacy. How much of the time will there by somebody standing?

Answer: Someone must stand if there are more than two people in the system. In this example, the person being served and the people waiting to be served are all in the same area, we presume.

Probability (0 in system) = 1 - ρ = 0.3333
Probability (1 in system) = (1 - ρ)ρ = 0.2222
Probability (2 in system) = (1 - ρ)ρ² = 0.1481
Probability (2 or fewer in system) = P(0) + P(1) + P(2) = 0.7037
Probability (3 or more in system) = 1 - (Probability(2 or fewer in the system) = 1 - 0.7037 = 0.2963

If you have just two chairs, then almost 30% of the time there will be someone standing.

Calculating line length and waiting times:

L, the average number of customers waiting or being served = $\rho/(1-\rho)$ = 2 persons.
L_q, the average number of customers waiting in the queue = $L-\rho$ = 1.333 persons.
W, the average time in system, = $1/(\mu-\lambda)$ = 0.2 person-hours/order.
W_q, the average time in queue, = ρW = 0.1333 person-hours/order.

Economic Analysis Example

Typically, managers can make choices that affect how fast the service is. If service is made speedier, that benefits customers two ways: They spend less time being served, and they spend less time waiting to be served. The institution has some savings too, because it needs less space to hold the waiting line. On the other hand, speeding up service usually costs something, in the form of extra resources. To decide whether to speed up service, the manager has to weight the costs against the benefits.

Weighing costs and benefits is easiest if you can express the benefits in dollar terms. The benefit here is waiting time that can potentially be saved. In the following example, we calculate the value of time spent waiting for pharmaceuticals from the aides' pay rate. There are two methods for this.

Method 1: $W \times \lambda \times$ pay rate
Aides who are getting orders filled spend 0.2 hour avg per order (because W=0.2).
There are an average of 10 orders to be filled each hour (because λ=10).
This implies that, on average each hour, aides spend 2 person-hours waiting for orders. (10×0.2=2.0)
Suppose the aides' pay is $8 per hour including fringe, taxes, etc.
Then each hour, on average, your aides spend $16 worth of time waiting for orders. (8×2=$16)

Method 1 Formula

$$W \frac{person-hours}{order} \times \lambda \frac{orders}{hour} \times pay \frac{\$}{person-hour} = \text{Your cost} \frac{\$}{hour}$$

When the formula is written out this way, with the units, you can see how the units multiply out.
Person-hours/order times orders/hour times dollars/person-hour = dollars/hour

Method 2:
(First thought of by a student, not me.)
On average, you have L aides in the pharmacy. Here L = 2 persons.
They cost you their pay rate, $8 per person-hour, while they are there.
Your hourly cost is L $\times$ pay rate = 2 persons $\times$ 8 $/person-hour = 16 $/hour.

Method 2 Formula

$$L \, persons \times pay \frac{\$}{person-hour} = \text{Your cost} \frac{\$}{hour}$$

The units in this formula multiply out properly if you treat the hyphen in person-hour as multiplication (which it is) rather than subtraction.

Now that we have figured out that we are losing an average of $16 per hour thanks to aides waiting, we can evaluate proposals to save some of that time. Suppose automated dispensing equipment could cut service time in half, but it costs $7 per hour to own and operate. Would it be worth buying?

To find the answer double μ, recalculate W (if using Method 1) or L (if using Method 2), and then redo the hourly cost calculation. Then see if the cost savings is bigger than the hourly cost of the machine.

Here is how:

μ was 15; it will be 30 if we buy the new equipment.

Method 1:   W = 1/(μ-λ) = 1/(30-10)=1/20=0.05

                 Wλ8 = 0.05×10×8=4

Method 2   L = λ/(μ-λ) = 10/(30-10)=10/20 =0.5

                 L8 = 0.5×8 = 4

Hourly costs fall to $4 from $16. We save $12, so we come out ahead by $5 per hour if we pay $7 per hour for the equipment.

Your turn again:

Assignment 7A

Patients arrive at your emergency room at a rate of 2 per hour.
The average service time per patient is 20 minutes. In other words, you treat an average of 3 patients per hour if patients come in that fast.

1. To proceed with your analysis, what do you assume about the distributions of arrivals and service times?

2. What is the average number of patients in the ER (waiting or being served?)
         The formulas you need for questions 2, 3, and 4 are on page 10. Try implementing them in a spreadsheet. If you get stuck and need a spreadsheet to copy, see the first model in the Queuing Theory Cookbook.

3. What is the average length of time that patients spend from the time they enter the ER to the time they leave?

4. The waiting area is separate from the examining/treatment room. How many chairs should there be in the waiting area to reduce the probability that someone will have no chair to less than 0.01? Explain how you got your answer.

5. Suppose the hospital has announced, as part of a Continuous Quality Improvement policy, that it will discount each patient's bill by $0.10 per minute ($6 per hour) that the patient waits in the ER waiting area. How much will this discount cost the hospital per hour on the average? Hint: In this example you're paying people only for waiting time. In the pharmacy example we paid people for waiting and service time. This makes the formulas a little different.

6. How much should the hospital be willing to spend to cut the average service time per patient to 15 minutes?