A Quick Introduction to Queuing Theory

Examples of Queuing Systems

System	Servers	Customers
Bank	Tellers	Customers
Hospital	Doctors, nurses, beds	Patients
Computer System	CPU, I/O devices	Jobs
Manufacturing System	Machines, workers	Parts
Airport	Runways, gates, security check-in stations	Airplanes, travelers
Communications network	Nodes, links	Messages, packets

A Quick Introduction to Queuing Theory

Components of a Queuing System.

1) **Arrival Process**. This specifies how customers arrive to the system.

The ***inter-arrival time***- A_i denotes the time interval between the arrival of the $(i - 1)$st and ith customer.

The successive times $A_1, A_2, \ldots, A_n, \ldots$ are Independent Identically Distributed (IID) random variables with mean of $E(A)$.

The *arrival rate* of the customers- $\lambda = 1/E(A)$**.**

the physical units: if A_i is in seconds, then λ is in **reciprocal seconds**.

The probability distribution of an arrival rate: unless specified, it usually means an exponential distribution with given mean.

A Quick Introduction to Queuing Theory

2) Service Mechanism.

number of servers, usually denoted by the variable s.

the service times *S1, S2, …, Sn, …* are IID random variables giving the service times of a sequence of customers .

the *service rate* of a server $\omega = 1/E(S)$.

3) **Queue Discipline**.

This is the rule by which we choose the next customer to be served. Common ones are

FIFO: First In First Out (a standard queue)

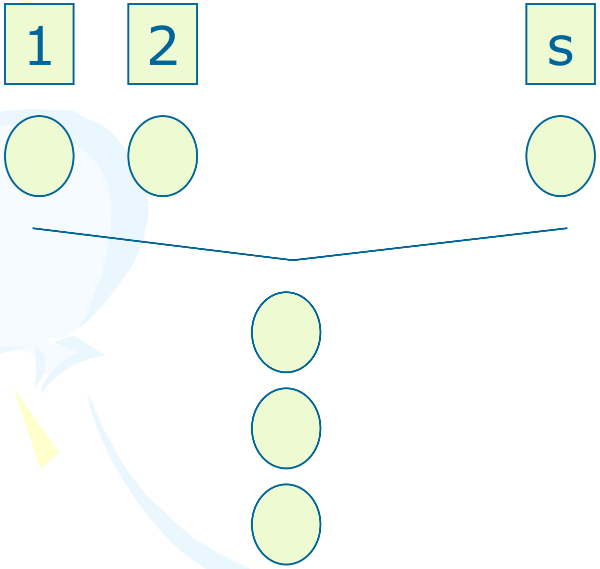
LIFO: Last In First Out (a stack)

Priority: some way is defined to determine the priority of a customer (a priority queue)

A Quick Introduction to Queuing Theory

Notation for Queuing Systems.

A notational system originally introduced by D. G. Kendall has become standard.



1. s servers in parallel and one FIFO queue feeding servers.

2. Queue capacity is infinite.

3. $A_1, A_2, \ldots$ are IID random variables.

4. $S_1, S_2, \ldots$ are IID random variables.

5. The A_i's and S_i's are independent.

Such a system is denoted by ***GI/G/s***, for a *General Independent* distribution of the A_i's, a General distribution of the S_i's and the presence of s servers.

A Quick Introduction to Queuing Theory

- $M/M/1$: A system with one server, exponential arrivals and exponential service; M is used for Markovian, or memory-less.
- E_k is used for *k-Erlang distribution*, obtained as a sum of k exponential random variables.
- D is used to denote deterministic (or constant) arrivals or service.
- $\rho = \lambda/(s\omega)$ is called the utilization factor. ω is the service rate of a single server, and there are s servers.

ρ is a measure of how heavily the resources of the queuing systems are used.

Note that, under normal circumstances, $\rho \leq 1$.

A Quick Introduction to Queuing Theory

Measures of Performance for Queuing Systems.

Given a queuing system, what are "relevant" elements to measure?

1) D_i = delay in queue of ith customer;

2) $W_i = D_i + S_i$ = waiting time in system of ith customer;

3) $Q(t)$ = number of customers in queue at time t;

4) $L(t)$ = number of customers in system at time t

= number of customers in queue + number of customers being served.

A Quick Introduction to Queuing Theory

Steady-State Value.

a) Steady-State Average Delay in queue:

$$d = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} D_i}{n}, \quad w.p.\,1$$

where $w.p.$ 1 stands for *with probability* 1 and means that the limit holds for almost all sequences $D_1, D_2, \ldots$

b) Steady-State Average Waiting Time in system:

$$w = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} W_i}{n}, \quad w.p.\,1$$

A Quick Introduction to Queuing Theory

c) **Steady-State Time-Average Number in Queue**:

$$Q = \lim_{T \to \infty} \frac{\int_0^T Q(t)dt}{T}, \quad w.p.1$$

d) **Steady-State Time-Average Number in System**:

$$L = \lim_{T \to \infty} \frac{\int_0^T L(t)dt}{T}, \quad w.p.1$$

Note that in all cases $\rho < 1$ is a necessary condition for the limits to exist (the average number of arrivals must be less than the average number of *possible* departures).

A Quick Introduction to Queuing Theory

Little's law (two conservation equations, or Little's formula)

$$Q = \lambda d \,, \qquad L = \lambda w$$

They are valid for any queuing systems for which the limits d and w exist.

In words: the *time-average number in queue* is given by the product of the *mean arrival rate* and the *average delay*.

The *time-average number in system* is given by the product of the *mean arrival rate* and the *average waiting time*.

Since time in queue and waiting time are related, we can also obtain a formula relating their averages:

$$w = d + E(S)$$

A Quick Introduction to Queuing Theory

arrival and service distributions

The number of arrival and service distributions for which analytical results can be obtained is rather small.

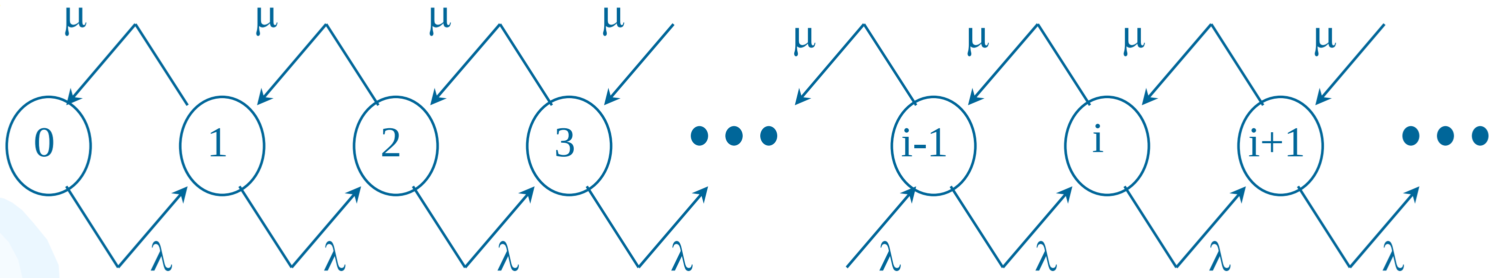
Exponential distributions (and their Erlang variants) are good candidates.

Usually, at least one of arrival and service distributions must be exponential for an analytical attack to be successful.

The simplest case, **the *M/M/1* queue**.

A Quick Introduction to Queuing Theory

Model of *M/M/1* queue



Poisson Process and exponential distribution

An arrival process with exponentially distributed inter-arrival times is also called "Poisson Process".

This refers to arrivals (for inter-arrival times) or to service times (for servers).

A Quick Introduction to Queuing Theory

Poisson Processes. There are a number of different (but equivalent) definitions of a Poisson process. We will introduce one and derive a number of properties. You should be aware that there are multiple way of deriving the properties.

Definition: a function $f(x)$ is a member of the $o(h)$ class, denoted by $f \in o(h)$ if $\lim_{h \to 0} \frac{f(h)}{h} = 0$.

Let $P_n(t)$ denote the probability that exactly n arrivals will occur in a time interval of length $t \geq 0$. Let λ be a positive number. We introduce two postulates:

a) The probability that an arrival occurs during $(t, t + h)$ is $\lambda h + o(h)$.

b) The probability that more than one arrival occurs during $(t, t + h)$ is $o(h)$.

A Quick Introduction to Queuing Theory

The next step is the derivation of a system of differential equations for $P_n(t)$, $n = 0, 1, 2, \ldots$. Examine the interval $(0, t + h) = (0, t) \cup [t, t + h)$. $P_n(t + h), n \geq 1,$ can be computed as:

a) the probability of n arrivals during $(0, t)$ and no arrivals during $[t, t + h)$;

b) the probability of $n - 1$ arrivals during $(0, t)$ and one arrival during $[t, t + h)$;

c) the probability of $x \geq 2$ arrivals during $[t, t + h)$ and $n - x$ arrivals during $(0, t)$.

These are three mutually exclusive and exhaustive possibilities (they cover all cases). They give:

$$P_n(t + h) = P_n(t)(1 - \lambda h - o(h)) + P_{n-1}(t)\,\lambda h + o(h)$$

$$= P_n(t)(1 - \lambda h) + P_{n-1}(t)\,\lambda h + o(h)$$

Rearranging and dividing by h gives:

A Quick Introduction to Queuing Theory

$$\frac{P_n(t+h)-P_n(t)}{h} = -\lambda P_n(t) + \lambda P_{n-1}(t) + \frac{o(h)}{h}, \quad n\geq 1, \quad t\geq 0$$

Taking the limit as $h \to 0^+$, gives the (infinite) system of differential equations (actually claim the two-sided limit is justified):

$$P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t), \quad n\geq 1, \quad t\geq 0.$$

Applying similar reasoning to the case $n = 0$, we have

$$P_0'(t) = -\lambda P_0(t), \quad for\, t \geq 0.$$

A moment of thought allows us to conclude that $P_0(0) = 1$ (there have been no arrivals at all), so the last equation comes to us complete with initial condition. Its unique solution is $P_0(t) = e^{-\lambda t}$. We can also observe that $P_n(0) = 0$ for all $n \geq 1$. Using the just computed expression for $P_0(t)$, we obtain:

$$P_1'(t) = -\lambda P_1(t) + \lambda\, e^{-\lambda t},\ P_1(0) = 0.$$

A Quick Introduction to Queuing Theory

This is a non-homogeneous equation and its solution involves a trick: first fine the general solution of the corresponding homogeneous equation:

$$P_1'(t) = -\lambda P_1(t) \quad -> \quad P_1(t) = C_1 e^{-\lambda t}.$$

Replace the constant by a function of t: $z(t)e^{-\lambda t}$, and attempt to determine $z(t)$ by simply inserting into the **original** equation and seeing what happens:

$$z'(t)\, e^{-\lambda t} - \lambda\, z(t)\, e^{-\lambda t} = - \lambda\, z(t)\, e^{-\lambda t} + \lambda\, e^{-\lambda t}$$

$$z'(t) = \lambda \;\; -> \;\; z(t) = C_1 + \lambda t$$

$$P_1(t) = (C_1 + \lambda t)\, e^{-\lambda t}, \;\; P_1(0) = 0.$$

Finally:

$$P_1(t) = \; \lambda t\, e^{-\lambda t}.$$

We can repeat the construction for $P_2(t)$, $P_3(t)$, … etc., and a final induction proof lets us conclude that:

A Quick Introduction to Queuing Theory

$$P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$$

the probability of exactly n arrivals during $(0, t)$. It is easy to verify that

$$\sum_{n=0}^{\infty} P_n(t) = \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \equiv 1$$

From this, we can also compute the probabilities of at least N arrivals, or at most N arrivals, etc. Something we have not quite addressed, yet, is the meaning of λ. It **seems** to be an expected arrival rate, but we need to show it.

Let $E(n, t)$ denote the expected number of arrivals during $(0, t)$. From standard probability theory - and some elementary algebra - we have:

$$E(n,t) = \sum_{n=0}^{\infty} n \cdot P_n(t) = \sum_{n=0}^{\infty} n \frac{(\lambda t)^n}{n!} e^{-\lambda t} = \lambda t$$

so that λ is the expected number of arrivals in unit time, or the **arrival rate**. What is the relationship between λ and inter-arrival times?

A Quick Introduction to Queuing Theory

Let $\Lambda(t)$ denote the number of arrivals during $(0, t)$. The probability that the first arrival time, τ_1, satisfy $\tau_1 > t$ is given by

$$P(\tau_1 > t) = P(\Lambda(t + 0) - \Lambda(0) = 0) = P_0(t) = e^{-\lambda t}.$$

If we assume that the number of arrivals over an interval depends only on the length of the interval, and not on when the interval occurs, we also have (τ_2 is the interval between first and second arrival):

$$P(\tau_2 > t| \ \tau_1 = x) = P(\Lambda(t + x) - \Lambda(x) = 0) = P_0((t +x) - x) = P_0(t)= e^{-\lambda t},$$

and $P\left(\tau_{i+1} > t \mid \sum_{j=1}^{i} \tau_j > x\right) = P\left(\Lambda(t+x) - \Lambda(x) = 0\right) = P_0(t) = e^{-\lambda t}$

We conclude that the probability that **any** inter-arrival time τ is greater than t is given by $G(t) = P(\tau > t) = P_0(t) = e^{-\lambda t}$. Inter-arrival times are thus IID random variables with **cumulative probability distribution function** $F(t) = 1 - G(t) = 1 - e^{-\lambda t}$, and **probability density function** $f(t) = \lambda e^{-\lambda t}$:

$$F(t) = \int_0^t \lambda e^{-\lambda\tau}\, d\tau = -e^{-\lambda\tau}\Big|_0^t = 1 - e^{-\lambda t}$$

A Quick Introduction to Queuing Theory

The **expected inter-arrival** time is thus given by

$$E(\tau)=\int_0^\infty \tau f(\tau)d\tau=\int_0^\infty \tau\lambda e^{-\lambda\tau}d\tau=-\tau e^{-\lambda\tau}\Big|_0^\infty-\int_0^\infty -e^{-\lambda\tau}d\tau$$

$$=0+\int_0^\infty e^{-\lambda\tau}d\tau=-\frac{1}{\lambda}e^{-\lambda\tau}\Big|_0^\infty=\frac{1}{\lambda}$$

Another interesting question - useful to help us choose theoretical distributions matching empirical data sets - is that of the position of the median relative to the mean.

Let τ be the random variable representing inter-arrival time.

$$P\left(\tau\le\frac{1}{\lambda}\right)=F\left(\frac{1}{\lambda}\right)=1-e^{-\lambda\frac{1}{\lambda}}=1-e^{-1}\approx 0.632121;$$

$$P\left(\tau>\frac{1}{\lambda}\right)=1-F\left(\frac{1}{\lambda}\right)=e^{-1}\approx 0.367879.$$

A way of interpreting this result is that, for an exponential distribution, the **median is less than the mean**.

A Quick Introduction to Queuing Theory

Let $p_n(t)$ denotes the probability that an $M/M/1$ system with mean arrival rate λ and mean service rate μ have n customers in the system at time t .

Can we obtain a steady-state probability, i.e., does

$\lim_{t\to\infty} p_n(t)$ exist?

$$p_n(t+\Delta t) = p_n(t)\big[(1-\mu\,\Delta t)(1-\lambda\,\Delta t) + \mu\,\Delta t\,\lambda\,\Delta t\big] + o(\Delta t)$$

$$+\, p_{n+1}(t)\big[\mu\,\Delta t(1-\lambda\,\Delta t)\big] + o(\Delta t)$$

$$+\, p_{n-1}(t)\big[\lambda\,\Delta t(1-\mu\Delta t)\big] + o(\Delta t), \quad n \geq 1.$$

For $n = 0$, the first term on the right hand must reflect a probability 0 for a departure, and the third term must be absent.

$$p_0(t+\Delta t) = p_0(t)(1-\lambda\Delta t) + o(\Delta t)$$

$$+\, p_1(t)\big[\mu\Delta t(1-\lambda\Delta t)\big] + o(\Delta t)$$

A Quick Introduction to Queuing Theory

Assuming smoothness in t for $p_n(t)$, we can use the approximation $p_n(t + \Delta t) = p_n(t) + p_n'(t)\Delta t$.

Replacing into the equation above, and letting $\Delta t \to 0$, we have

$$p_n'(t) = -(\lambda + \mu)p_n(t) + \mu p_{n+1}(t) + \lambda p_{n-1}(t)$$

$$p_0'(t) = -\lambda p_0(t) + \mu p_1(t)$$

A necessary condition for a steady-state solution is that $p_n'(t)$ vanish identically, leaving us with a time-independent system of infinitely many equations:

$$\lambda p_0 = \mu p_1;$$

$$(\lambda + \mu) p_n = \mu p_{n+1} + \lambda p_{n-1}, \quad for\ n \geq 1.$$

A Quick Introduction to Queuing Theory

We recall that $\rho = \lambda/\mu$ is what we called **utilization**. We can now proceed with, essentially, and induction:

$p_1 = \rho\, p_0$;

$p_2 = (\rho + 1)p_1 - \rho p_0 = \rho^2 p_0$;

….

$p_n = \rho^n p_0$, for all $n \geq 0$.

Since $\sum_{n=0}^{\infty} p_n \equiv \sum_{0}^{\infty} \rho^n p_0 = 1$, we must have $\rho < 1$. Summing the geometric series: $p_0 = (1 - \rho),\ \ p_n = (1 - \rho)\, \rho^n$. From $\rho = 1 - p_0$, we can conclude that ρ is the probability the system is not empty.

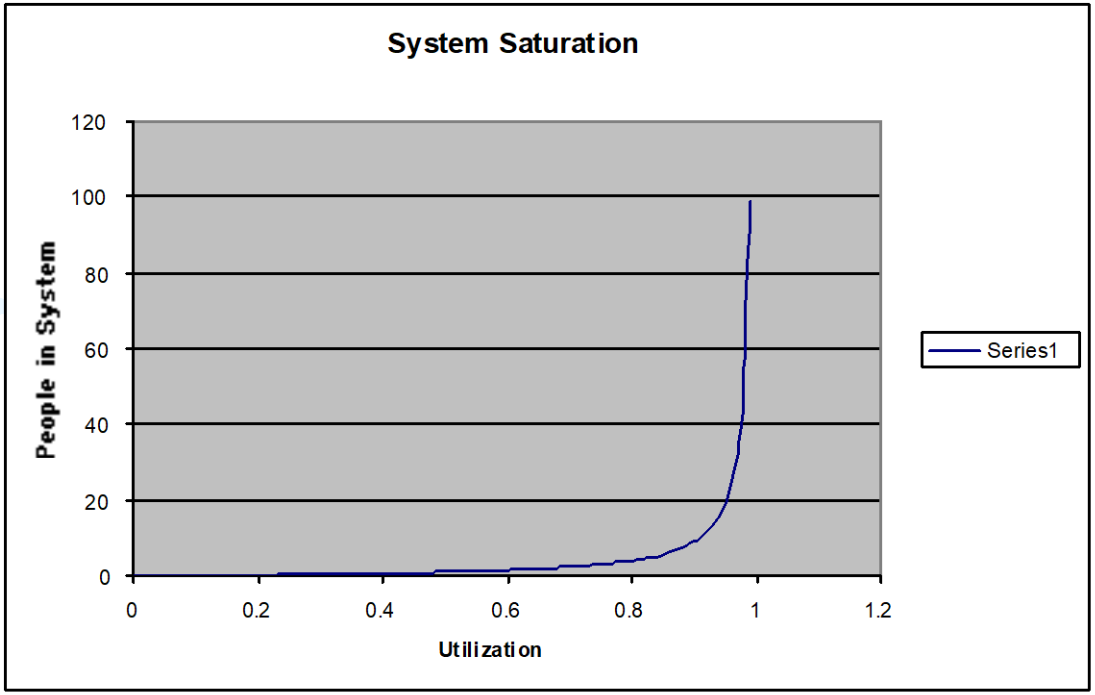
A Quick Introduction to Queuing Theory

The **expected number of customers in the system** is given by

$$E(n)=\sum_0^\infty n\,p_n=\sum_0^\infty n(1-\rho)\rho^n=\frac{\rho}{1-\rho}$$

An easy derivation of the closed form of the sum follows from observing that

$$\frac{1}{(1-\rho)^2}=\frac{d}{d\rho}\left(\frac{1}{1-\rho}\right)=\frac{d}{d\rho}\left(\sum_{n=0}^\infty \rho^n\right)=\sum_{n=0}^\infty n\,\rho^{n-1}=\frac{1}{\rho}\sum_{n=0}^\infty n\,\rho^n$$



A Quick Introduction to Queuing Theory

Note that with a utilization of $\rho = 0.5$, the expected number of customers in the system is 1.

One can also show that, for a general service distribution with exponential arrivals, the steady-state average delay in queue is given by the formula

$$d = \frac{\lambda \left\{ Var(S) + [E(S)]^2 \right\}}{2[1 - \lambda E(S)]}$$

where S is the service distribution, and $Var(S)$ is the variance of the service distribution.

We can go a little further in this direction by computing variance of the number of customers in the system, and the variance of an exponential distribution with mean 1/λ. - the service distributions we have considered up to this point are all exponential anyway (although they don't have to be: for ATM 53Byte packets, the service distribution is uniform).

A Quick Introduction to Queuing Theory

Variance of the number of customers in the system.

$$\sigma^2 = E\left(n^2\right) - E(n)^2 = \sum_{n=0}^{\infty} n^2 p_n - E(n)^2 = (1-\rho)\sum_{n=0}^{\infty} n^2 \rho^n - \frac{\rho^2}{(1-\rho)^2}$$

We use the same differentiation trick as before, but now differentiate twice:

$$\frac{2}{(1-\rho)^3} = \frac{d^2}{d\rho^2}\left(\frac{1}{1-\rho}\right) = \frac{d^2}{d\rho^2}\left(\sum_{n=0}^{\infty} \rho^n\right) = \sum_{n=0}^{\infty} n(n-1)\rho^{n-2}$$

$$= \frac{1}{\rho^2}\sum_{n=0}^{\infty} n^2 \rho^n - \frac{1}{\rho^2}\sum_{n=0}^{\infty} n\,\rho^n = \frac{1}{\rho^2}\sum_{n=0}^{\infty} n^2 \rho^n - \frac{1}{\rho(1-\rho)^2}$$

where we used the result obtained two slides ago. Some simple algebra gives:

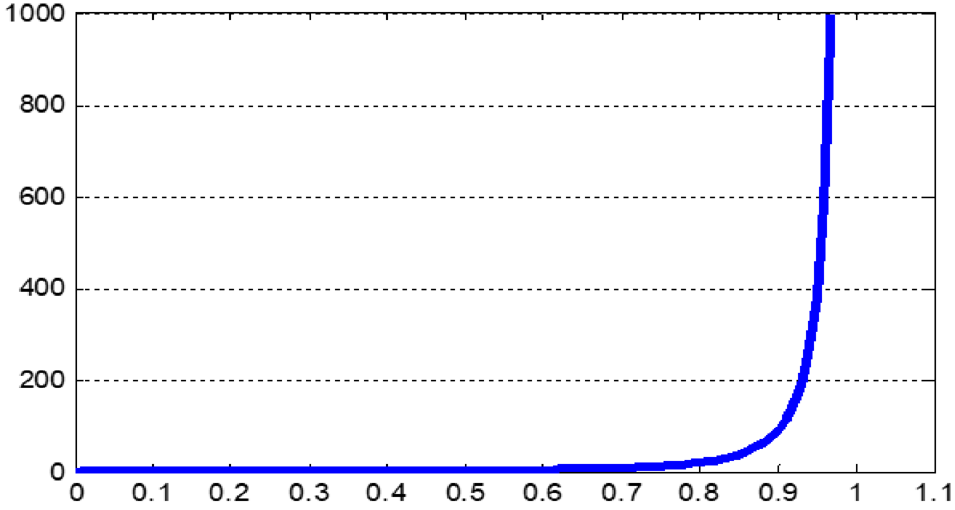
$$\sum_{n=0}^{\infty} n^2 \rho^n = \frac{\rho(1+\rho)}{(1-\rho)^3}$$

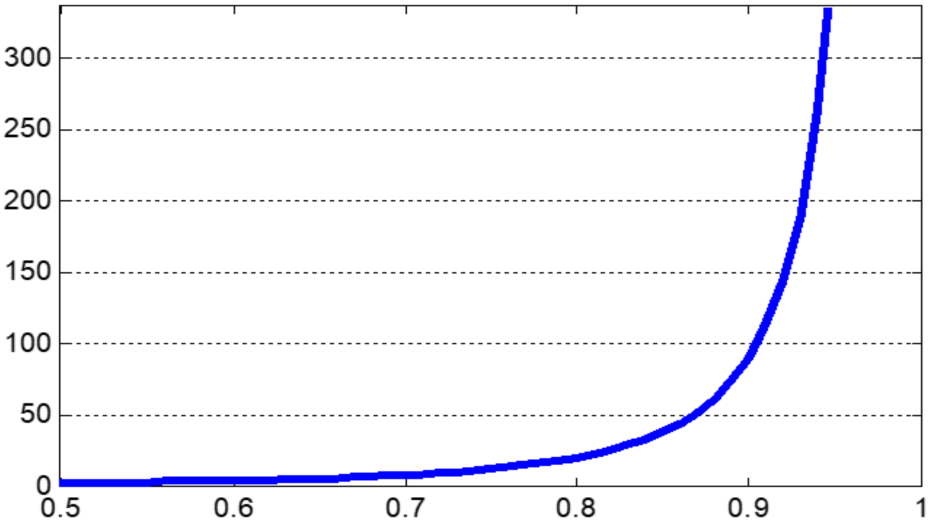
finally leading to:

$$\sigma^2 = \frac{\rho}{(1-\rho)^2}$$

A Quick Introduction to Queuing Theory

$$\sigma^2 = \frac{\rho}{(1-\rho)^2}$$





A Quick Introduction to Queuing Theory

Variance of an exponential distribution with mean $1/\lambda$**.**

Recall that the exponential distribution with mean $1/\lambda$ has probability density function $f(t) = \lambda e^{-\lambda t}$. To compute the variance:

$$Var(\tau) = E\left(\tau^2\right) - E(\tau)^2 = \int_0^\infty \tau^2 \lambda e^{-\lambda\tau}\, d\tau - \left(\frac{1}{\lambda}\right)^2$$

$$= -\tau^2 e^{-\lambda\tau} \Big|_0^\infty - \int_0^\infty 2\tau\left(-e^{-\lambda\tau}\right) d\tau - \left(\frac{1}{\lambda}\right)^2 = \int_0^\infty 2\tau\left(e^{-\lambda\tau}\right) d\tau - \left(\frac{1}{\lambda}\right)^2$$

$$= -2\tau\frac{1}{\lambda} e^{-\lambda\tau} \Big|_0^\infty - \int_0^\infty -2\frac{1}{\lambda} e^{-\lambda\tau}\, d\tau - \left(\frac{1}{\lambda}\right)^2 = \frac{2}{\lambda}\int_0^\infty e^{-\lambda\tau}\, d\tau - \left(\frac{1}{\lambda}\right)^2$$

$$= -\frac{2}{\lambda^2} e^{-\lambda\tau} \Big|_0^\infty - \left(\frac{1}{\lambda}\right)^2 = \left(\frac{1}{\lambda}\right)^2$$

The **variance is the square of the mean**: another point to help us decide whether an empirical distribution is exponential or not.

A Quick Introduction to Queuing Theory

Expected value and variance of the number of customers in queue.

We observe that when the number of customers in the system is 0, we will have 0 customers in the queue, when the number of customers in the system is 1, the number of customers in queue is still 0, while when the number of customers in the system is $n > 1$, the number of customers in queue is $n - 1$. This observation leads to the formula for the expected number of customers in queue:

$$E(n_Q) = \sum_{n=1}^{\infty} (n-1) p_n$$

where $p_n = (1 - \rho)\rho^n$, the probability of having n customers in the system. Making use of the summation formulae just derived, we can compute

$$E(n_Q) = \frac{\rho^2}{1-\rho} = \rho E(n)$$

as well as $E(n) = E(n_Q) + \rho$.: the **expected** number of customers in system is exactly ρ (the utilization) larger than the expected number in queue.

A Quick Introduction to Queuing Theory

We can compute variance:

$$\sigma_Q^2 = \sum_{n=1}^{\infty}(n-1)^2 p_n - E(n)^2 = \frac{\rho^2\left(1+\rho-\rho^2\right)}{(1-\rho)^2}$$

making use of the summation formulae previously derived.

Mean Waiting and Mean Delay Time.

In some texts you will find "waiting time" to be defined as "response time" and "delay time" as "waiting time" (the rationale being that the delay time is the time you are waiting in queue - which is actual waiting - while the total time you spent is the time it takes the system to respond to your request, fulfilling it).   Using our definitions, we have the relation:

$$W_i = D_i + S_i$$

Let $\overline{W}$ denote the mean waiting time, let $\overline{D}$ denote the mean delay time, let $\overline{L}$ denote the mean number of customers in system, and let $\overline{Q}$ denote the mean number of customers in queue.

A Quick Introduction to Queuing Theory

If $1/\lambda$ is the mean interarrival interval, and $1/\mu$ is the mean service time, we have, by definition, $\rho = \lambda/\mu$. Little's Law (see slide 9) gives:

$$\overline{D} = \frac{\overline{Q}}{\lambda} = \frac{\rho^2}{\lambda(1-\rho)} = \frac{\rho}{\mu(1-\rho)}$$

$$\overline{W} = \frac{\overline{L}}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu(1-\rho)}$$

What would this be good for? In the various simulations, you computed time-averaged delays and time-averaged waiting times. This provides you with a theoretical check: if your results don't match (within some statistical confidence level) the theoretically predicted ones, you are in some kind of trouble…

They also allow you to avoid - in some simple cases - going through the whole simulation process: a "back-of-the-napkin" computation can give you the results.