Basic Queueing Theory

CS 450 : Operating Systems

Michael Saelee <lee@iit.edu>

Computer Science

Science

IIT College of Science

ILLINOIS INSTITUTE OF TECHNOLOGY

Agenda

- Queueing theory? Huh?
- Probability refresher / Crash course
- Queueing theory & Kendall's notation
  - Mean value analysis of basic queues

- Queueing Theory? Huh?

Remember, we started our discussion of scheduling at a high level — "policy"

- mostly described *heuristics*-based (i.e., hand-wavy) approaches
  - makes it very important to measure and *evaluate* scheduling systems!

assignment 2 treads middle ground —
evaluation based on *simulation*:

- some basis in reality, but hard to predict real workloads
- no mathematical/computational rigor

to obtain *empirical* data, should examine a "live" operating system's scheduler

— low-level (coming later!)

to exercise *rigor*, should leverage some branch of mathematics well-suited to analyzing scheduling systems

… queueing theory!

queueing theory models *wait queues*
using (mostly) probabilistic techniques

- e.g., arrival/service rate *distributions*

- supports mathematical analysis & rigor wide application:

- checkout lines
- telecom switch
- traffic light system
- network quality of service

we'll barely scratch the surface — queueing theory is an area ripe for research — but you'll see some basic applications

- will also help explain underpinnings of simulators used for assignment 2!

§Probability refresher / Crash course

Probability theory = quantitative analysis of *random* phenomena

- assign a weighted *probability* to every *event* in a *sample space* (Ω)
- use these probability *distributions* to better understand the behavior of the phenomena

Core abstraction: *random variable*

- a R.V. is a function that maps the sample space onto numeric values (e.g., $X:\Omega\to\mathbb{R}$)
  - *discrete* R.V.s map to a *countable* set
  - *continuous* R.V.s map events onto an *uncountable* set (e.g., real-valued)

E.g., double coin toss (discrete event space)

$$\Omega = \{TT, TH, HT, HH\}$$

$$X(\omega) = \begin{cases} 0, & \text{if } \omega = TT \\ 1, & \text{if } \omega = TH \\ 2, & \text{if } \omega = HT \\ 3, & \text{if } \omega = HH \end{cases}$$

Typically interested in a variety of *statistics* of random variables (and corresponding events):

- probability of event n: $P(X=n)$ (or $p(n)$)
- expected value (mean): $E(X)$
- variance: $\sigma^2(X)$; standard deviation: $\sigma(X)$
- coefficient of variance: $C_X = \sigma(X) / E(X)$ (unitless measure)

e.g., (6-sided) dice roll

$$P(X = n) = \frac{1}{6},\ n = 1, 2, 3, 4, 5, 6$$

— *probability mass function*

Note: $\sum_n P(X = n) = 1$

$$E(X) = \sum_n n \cdot p(n) = 3.5$$

e.g., (6-sided) dice roll

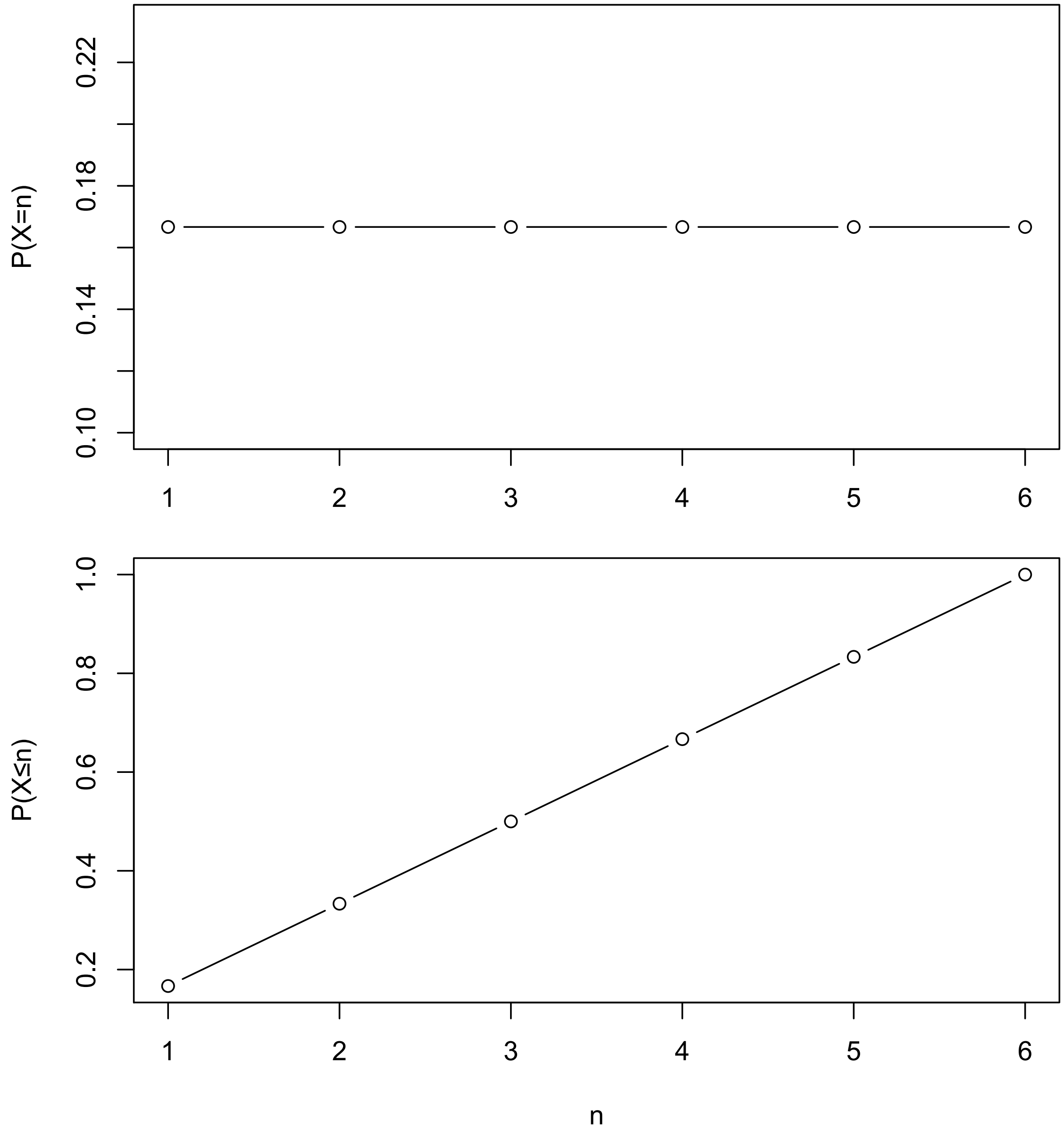
cumulative distribution function (CDF):

$$F_X(n) = P(X \leq n) = \sum_{x \leq n} p(x)$$

$$\text{e.g., } F_X(3) = P(X \leq 3)$$

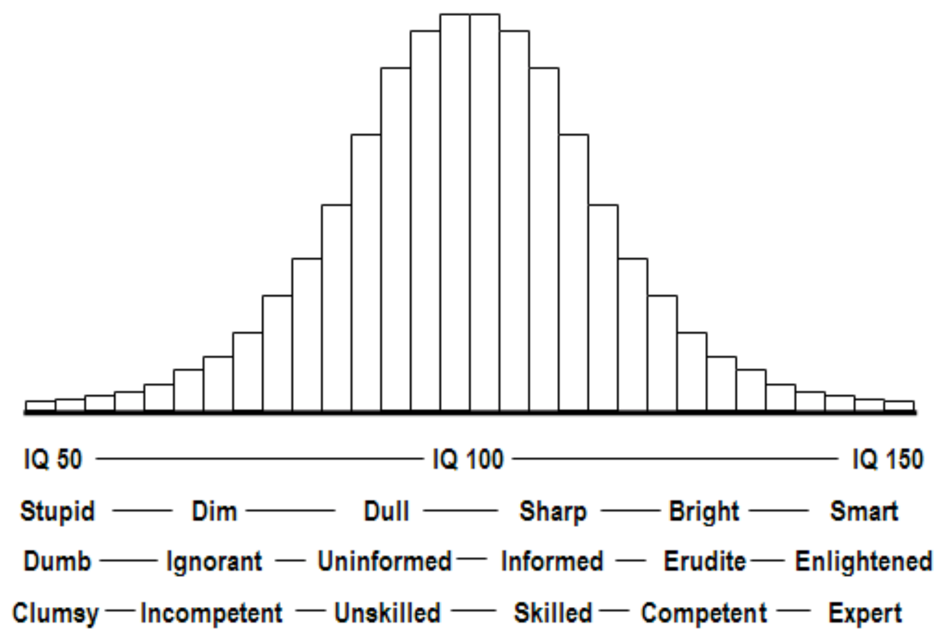
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

$$F_X(6) = 1$$



6-sided dice PMF, CDF

Many *well known probability distributions* are used to *model* real world phenomena



from “Brains and Careers,” by D. Keirsey

¶ Two *discrete* probability distributions

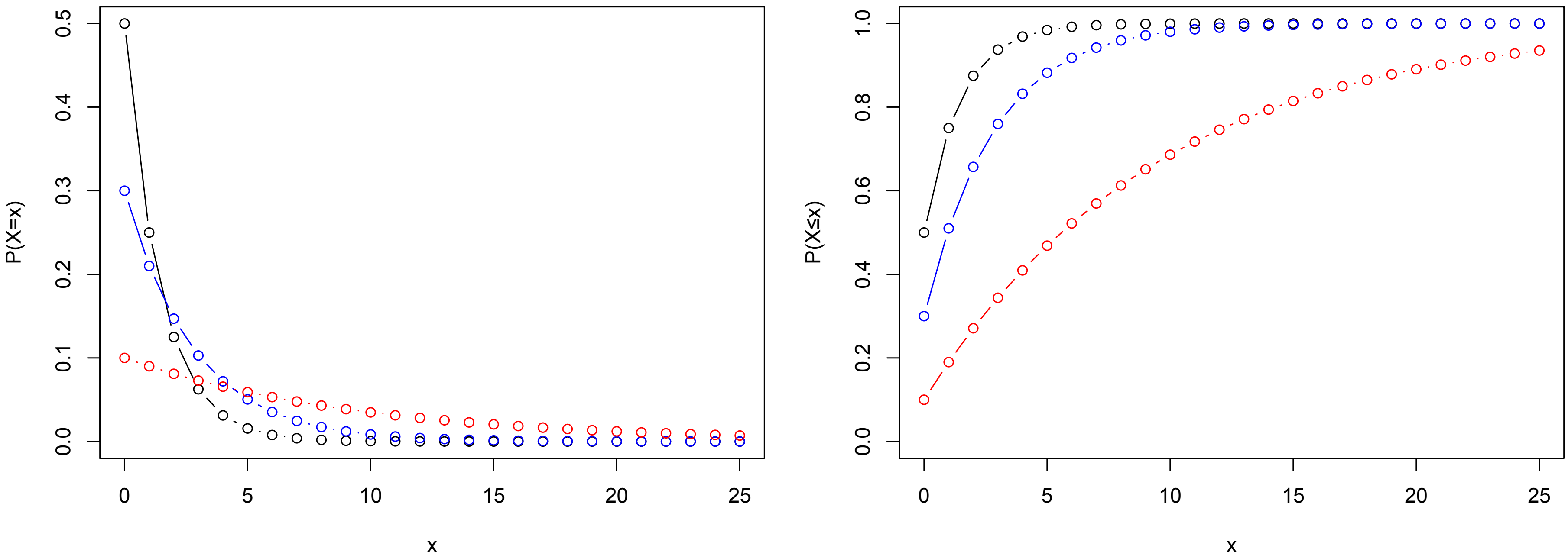
Geometric distribution

$$P(X = n) = (1 - p)^n p, n = 0, 1, 2, \ldots$$

- parameter p = chance of success in a trial
- gives probability of n failures before success

$$E(X) = \frac{1 - p}{p}, \ \sigma^2(X) = \frac{1 - p}{p^2}$$

Geometric distribution



p=0.5, p=0.3, p=0.1

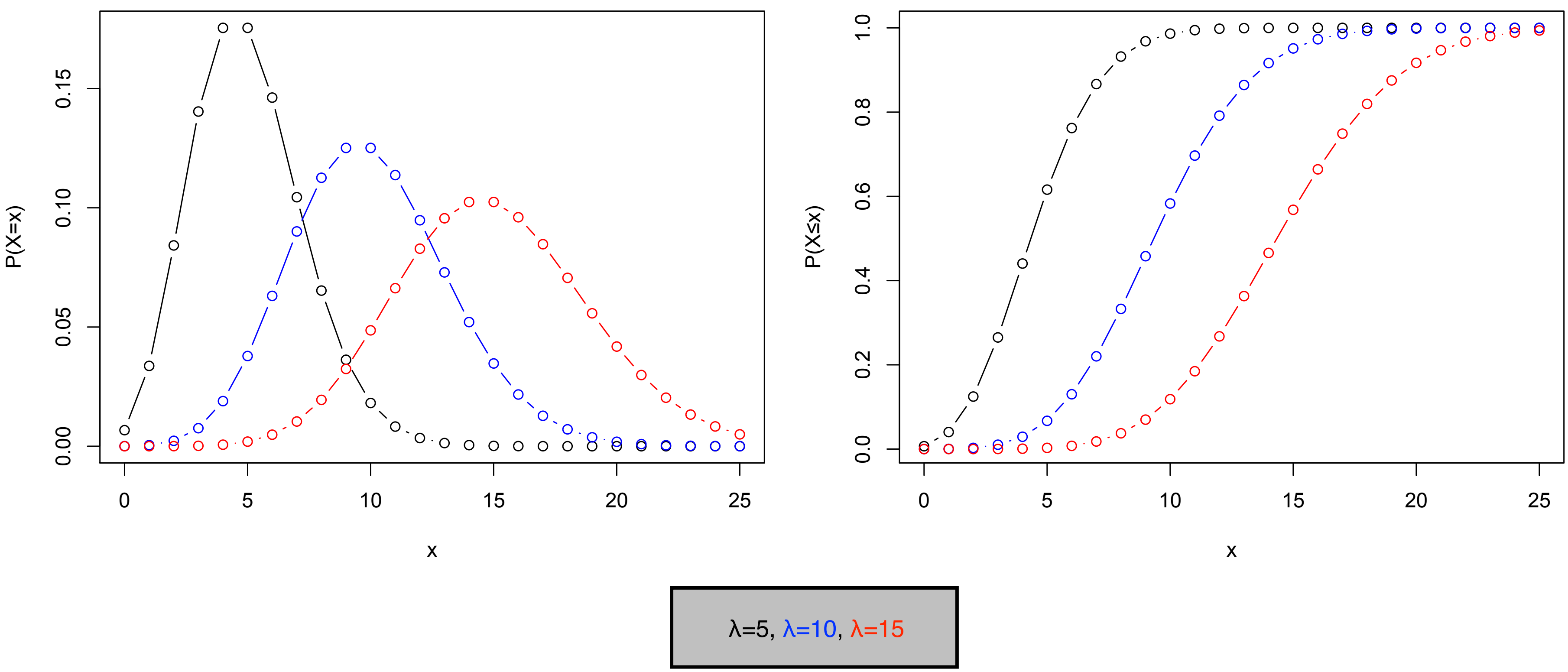
Poisson distribution

$$P(X = n) = \frac{\lambda^n}{n!} e^{-\lambda},\ n = 0, 1, 2, \ldots$$

- gives probability of n events occurring in a fixed time interval, when
  - average rate λ is known, and
  - each event is *independent* of previous ones

$$E(X) = \lambda,\ \sigma^2(X) = \lambda$$

Poisson distribution



¶ Two *continuous* probability distributions

It doesn't make sense to measure probability at a point (sample space is *uncountable*!)

Instead, assign probabilities to *intervals*:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

f is the *probability density function*

$$P(X \leq x) = F(x) = \int_{-\infty}^{x} f(t)dt$$

F is the *cumulative distribution function*

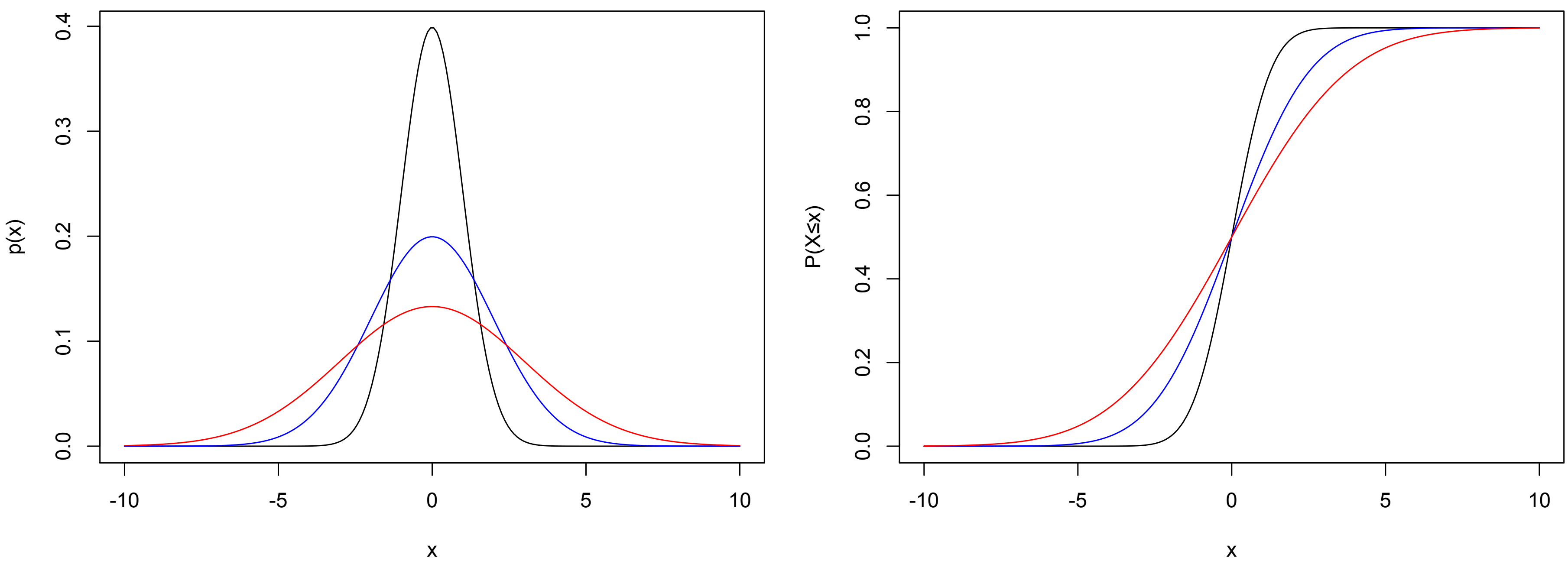
Gaussian (Normal) distribution

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- parameters μ (mean) and σ^2 (variance)
- produces "bell curve" around μ

$$E(X) = \mu, \ \sigma^2(X) = \sigma^2$$

Gaussian (Normal) distribution



$\mu=0$; $\sigma^2=1$, $\sigma^2=2$, $\sigma^2=3$

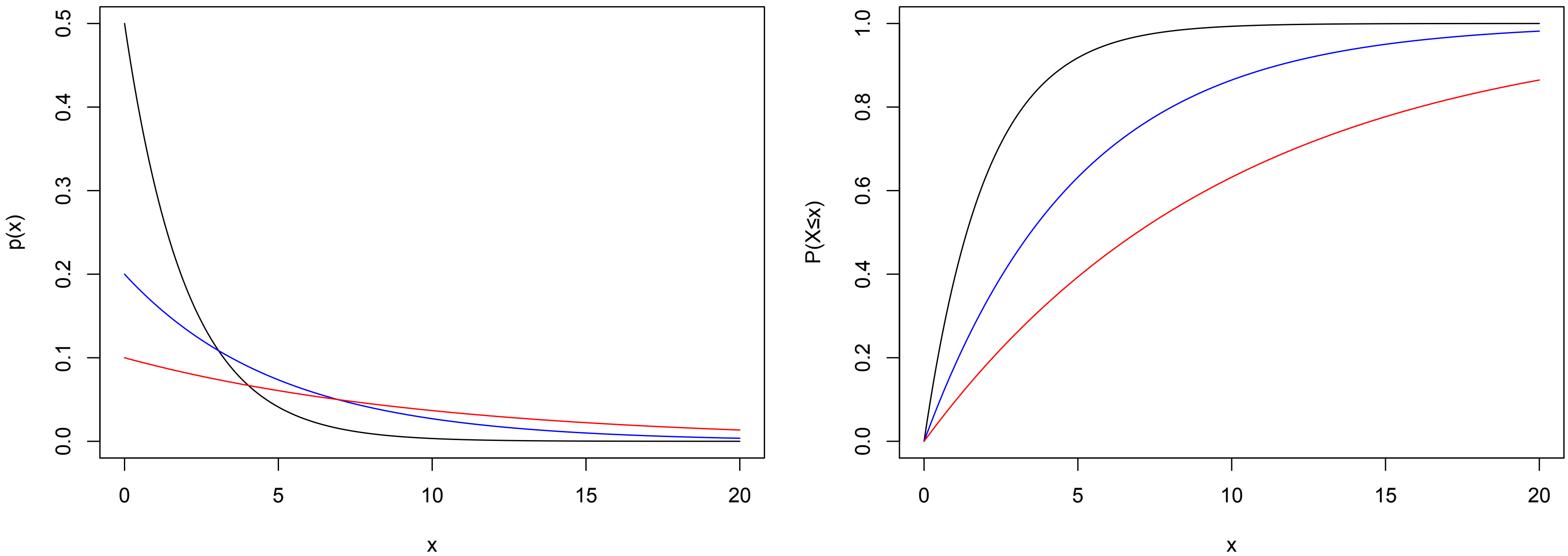
Exponential distribution

$$f(t;\mu) = \mu e^{-\mu t}, \quad t \geq 0$$

- parameter μ = rate
- gives probability of time t elapsing between successive *independent* events

$$E(X) = \frac{1}{\mu},\ \sigma^2(X) = \frac{1}{\mu^2},\ C_X = 1$$

Exponential distribution (continuous)



μ=0.5, μ=0.2, μ=0.1

Important property of the exponential distr.

— it is "*memoryless*", i.e.,

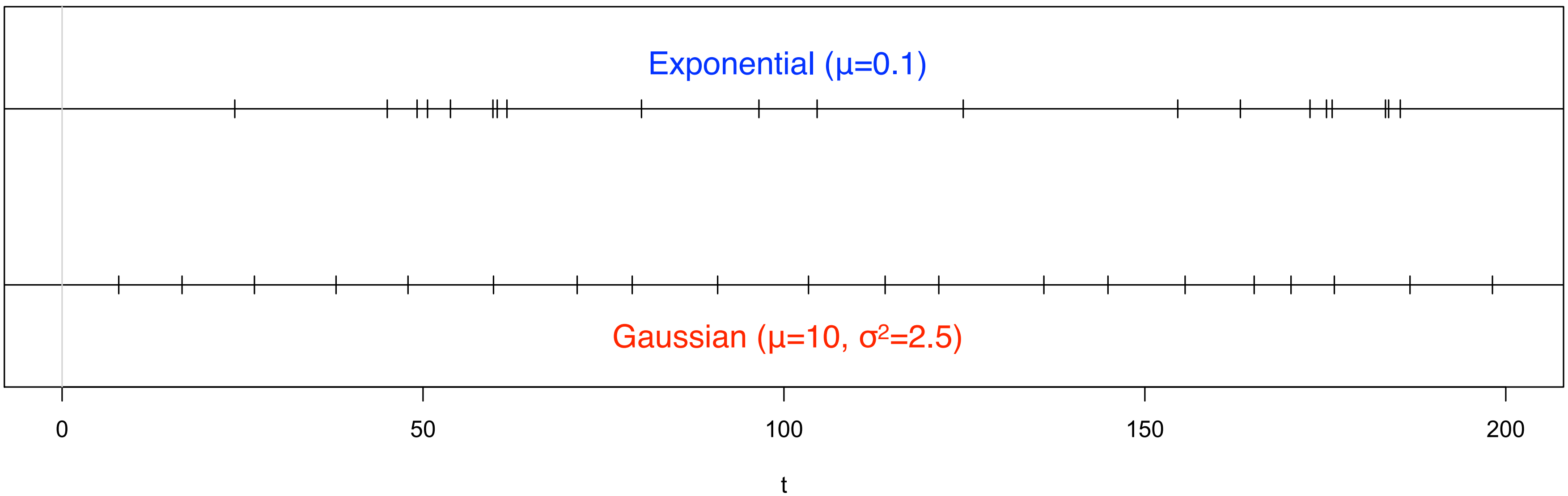
$$P(X > t + \Delta t \mid X > t) = P(X > \Delta t) \text{ for all } t, \Delta t \geq 0$$

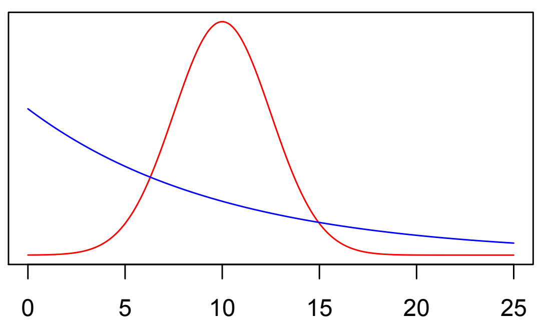
e.g., Given exponential bus arrival times:

If $P(X > 20\textbf{min}) = 0.3$, and you've already waited 15 minutes, how likely is it that the bus won't arrive for another 20 minutes?

$$P(X>35 \mid X>15) = P(X>20) = 0.3$$

Exponential vs. Gaussian arrival times





¶ Stochastic processes

A *stochastic process* is a collection of random variables $\{F_t, t \in T\}$ defined on Ω

- t is typically a time parameter
  - so F_t may describe how some system behaves over time period t

Poisson Process; $\{N_t,\ t \geq 0\}$

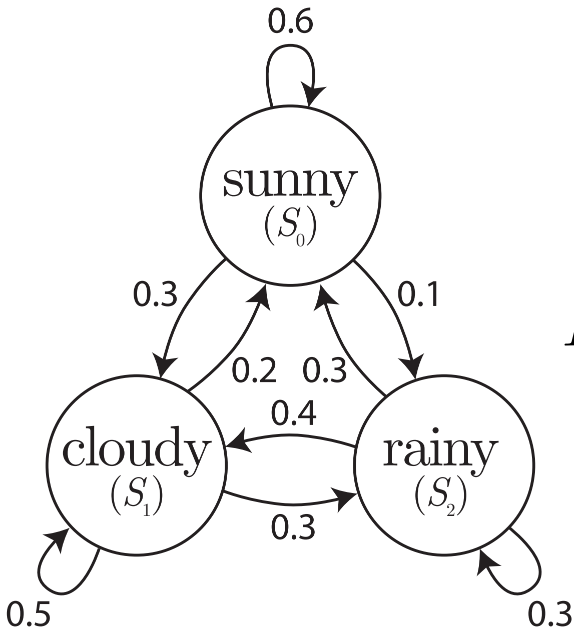
- $N_t =$ number of arrivals in $[0, t]$
- N_t is Poisson distributed with param λt
- time between arrivals is exponentially distributed with rate $1/\lambda$
- inherently memoryless

Markov Chain

- sequence of r.v.s, X_1, X_2, X_3, … such that:

$$P(X_{t+1} = x \mid X_t = x_t, X_{t-1} = x_{t-1}, \ldots, X_2 = x_2, X_1 = x_1)$$
$$= P(X_{t+1} = x \mid X_t = x_t)$$

- next state depends only on the current state (future is *independent* of past)
- range of X_i = *state space* (S) of the chain



"transition matrix"

$$P = \begin{pmatrix} p_{00} & p_{01} & p_{02} \\ p_{10} & p_{11} & p_{12} \\ p_{20} & p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.3 & 0.1 \\ 0.2 & 0.5 & 0.3 \\ 0.3 & 0.4 & 0.3 \end{pmatrix}$$

$$p_{ij} = P(X_{t+1} = j \mid X_t = i)$$

e.g., $P(X_{\mathrm{t+1}}{=}\text{sunny} \mid X_{\mathrm{t}}{=}\text{rainy}) = p_{20} = 0.3$

$P(X_{\mathrm{t+2}}{=}\text{sunny} \mid X_{\mathrm{t}}{=}\text{rainy})?$

$= p_{20}p_{00} + p_{21}p_{10} + p_{22}p_{20} = 0.35$

$$p_{20}^{(2)} = p_{20}p_{00} + p_{21}p_{10} + p_{22}p_{20} = 0.35$$

$$p_{ij}^{(2)} = \sum_{k \in S} p_{ik}p_{kj} = (P \times P)[i][j]$$

$$P \times P = P^2 = \begin{pmatrix} 0.45 & 0.37 & 0.18 \\ 0.31 & 0.43 & 0.26 \\ 0.35 & 0.41 & 0.24 \end{pmatrix}$$

$$p_{ij}^{(n)} = P^n[i][j]$$

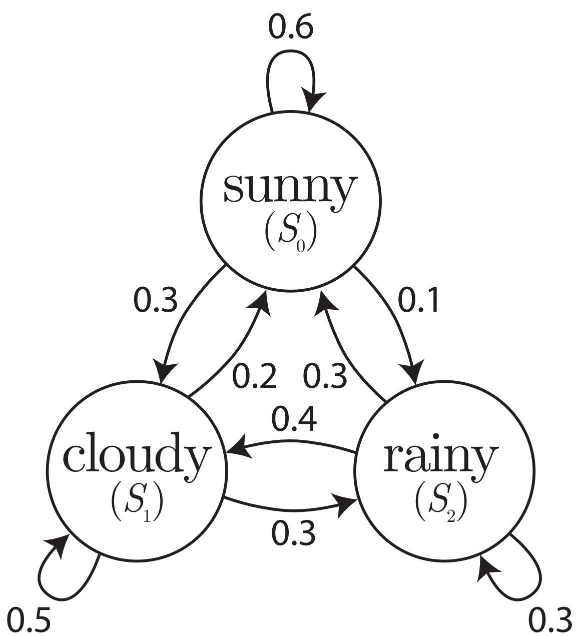
$$P^2 = \begin{pmatrix} 0.45 & 0.37 & 0.18 \\ 0.31 & 0.43 & 0.26 \\ 0.35 & 0.41 & 0.24 \end{pmatrix} \quad P^3 = \begin{pmatrix} 0.398 & 0.392 & 0.210 \\ 0.350 & 0.412 & 0.238 \\ 0.364 & 0.406 & 0.230 \end{pmatrix} \quad P^4 = \begin{pmatrix} 0.380 & 0.399 & 0.220 \\ 0.364 & 0.406 & 0.230 \\ 0.369 & 0.404 & 0.227 \end{pmatrix}$$

$$P^5 = \begin{pmatrix} 0.374 & 0.402 & 0.224 \\ 0.369 & 0.404 & 0.227 \\ 0.370 & 0.404 & 0.226 \end{pmatrix} \quad P^6 = \begin{pmatrix} 0.372 & 0.403 & 0.225 \\ 0.370 & 0.404 & 0.226 \\ 0.371 & 0.403 & 0.226 \end{pmatrix} \quad P^7 = \begin{pmatrix} 0.371 & 0.403 & 0.226 \\ 0.371 & 0.403 & 0.226 \\ 0.371 & 0.403 & 0.226 \end{pmatrix}$$

$\lim_{k \to \infty} P^k$ converges to a *steady-state distribution*

all rows are equal to the same vector π, where

$$\pi = \pi \times P \text{ and } \sum_{i \in S} \pi_i = 1$$

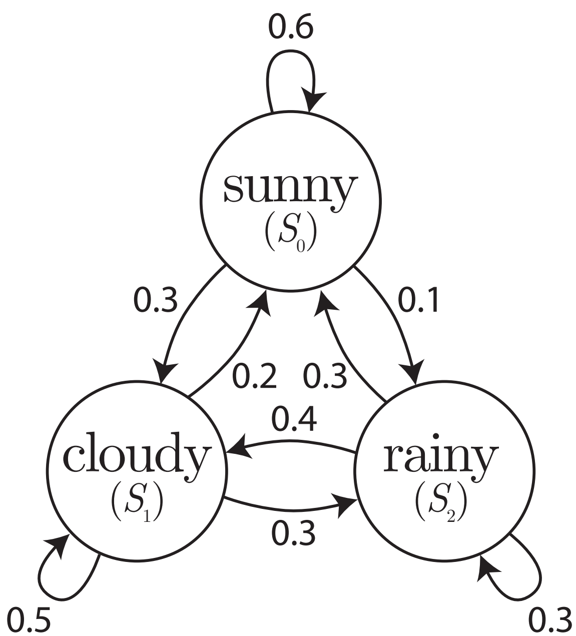


$$\pi = \begin{array}{ccc} \text{sunny} & \text{cloudy} & \text{rainy} \\ [0.371 & 0.403 & 0.226] \end{array}$$

independent of starting state:

$$P(X_t = \text{sunny}) = 0.371$$

i.e., fraction of sunny days $\approx 37\%$



$$\pi = [0.371 \quad 0.403 \quad 0.226]$$

Also note that, for every state,

rate of flow out = *rate of flow in*

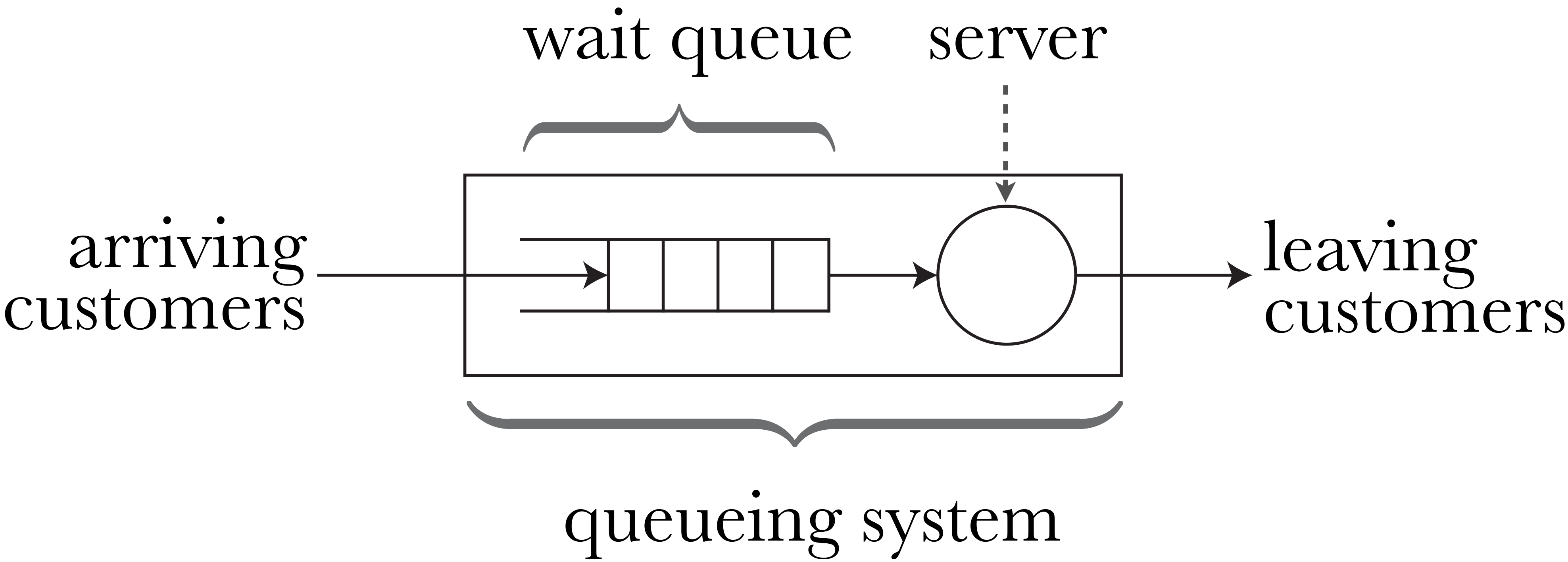
e.g., for S_0

$$\begin{aligned} \text{rate out} &= (0.371)(0.1 + 0.3) \\ &= 0.148 \\ \text{rate in} &= (0.403)(0.2) + (0.226)(0.3) \\ &= 0.148 \end{aligned}$$

i.e., the system is in *equilibrium*

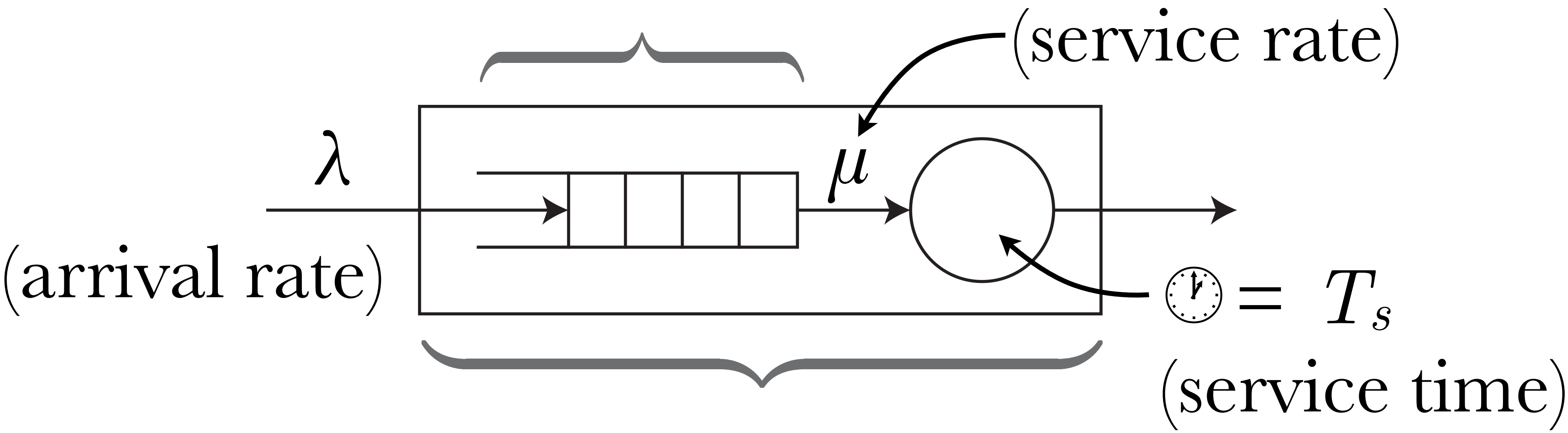
§Queueing theory

Basic model:


wait queue
server
arriving customers
leaving customers
queueing system

= L_q (waiting customers)

🕐 = T_q (wait time)



= L (total customers)

🕐 = T (sojourn / turnaround time)

typically, queue characteristics vary over time…

goal: given distributions for λ, μ, *derive the rest*!

e.g., distribution of T_q, T

distribution of L_q, L

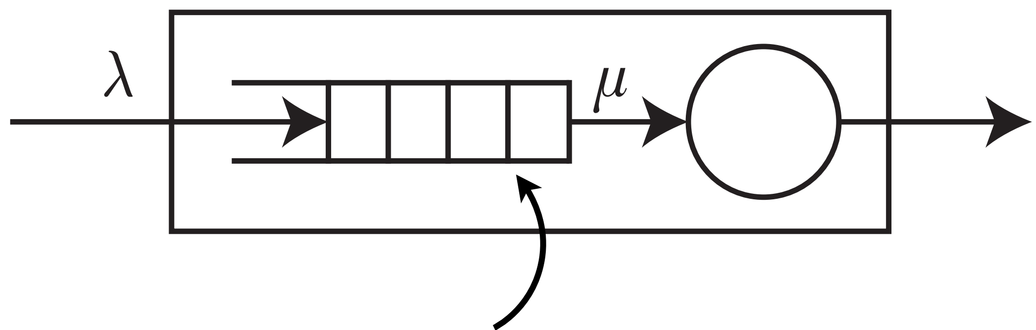
distribution of busy periods (i.e., when server is continuously busy) given distributions, can compute things like:

$P(L_q = 0)$	(queue is empty)
$P(L_q \geq x)$	(x or more in line)
$P(T \leq t)$	(sojourn time threshold)
$E(T_q), E(T)$	(avg. wait, sojourn time)
$E(L_q), E(L)$	(avg. queue/system size)

interested in *limiting distribution;*

i.e., after system reaches equilibrium

— over a long period of time, # customers leaving system = # customers entering



queue cannot grow unboundedly!

require $\rho = \dfrac{\lambda}{\mu}$, "intensity" < 1

ρ is the *utilization* for a *stable system*

¶ Little's Law

Little's Law: $E(L) = \lambda E(T)$

customers in system $=$ arrival rate $\times$ sojourn time

applied to queue: $E(L_q) = \lambda E(T_q)$

applied to server: $\rho = \lambda E(T_s)$

- † Little's Law does *not assume anything* about arrival/service distributions or any other server characteristics!
(but requires that $E(L)$, $E(T)$, λ be *bounded*)

e.g., 35th St. Jimmy John's:

12 customers arrive per hour,
Average time spent in store = 15 minutes.

Average # customers in store?

$$\frac{12}{\text{hour}} \times \frac{1\ \text{hour}}{60\ \text{min}} \times 15\ \text{min} = 3$$

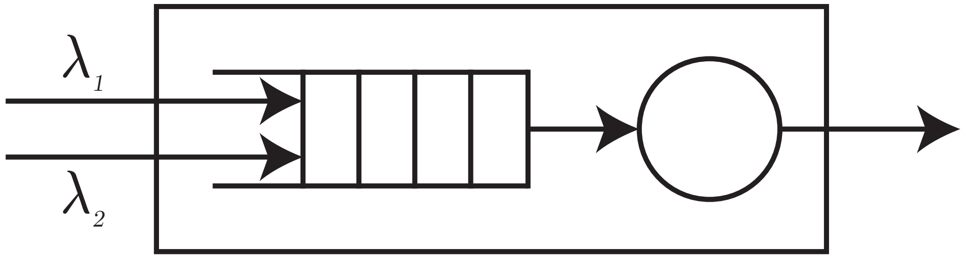
e.g., Customer appreciation day!

100 customers arrive per hour,

Average line length = 15

Average wait time?

$$15 \times \frac{1 \text{ hour}}{100} = 0.4 \text{ hour} = 9 \text{ min}$$



e.g., Packet switching system with 2 inputs:

λ_1=200 packets/s, λ_2=150 packets/s,

On average 2,500 packets in system.

Mean packet delay?

$$E(T) = \frac{E(L)}{\lambda_1 + \lambda_2} = \frac{2,500}{200 + 150} \approx 7.1\text{s}$$

¶ Kendall's Notation

$$A/S/c/k/n/d$$

A : interarrival time distribution

S : service time distribution

c : number of servers

k : buffer size (default=∞)

n : customer population size (default=∞)

d : queueing discipline (default=FCFS)

Arrival/Service distributions:

$$
\begin{aligned}
D &: \text{Deterministic} \\
M &: \text{Markovian (Memoryless)} \\
Geom &: \text{Geometric} \\
G &: \text{General (unknown/arbitrary)}
\end{aligned}
$$

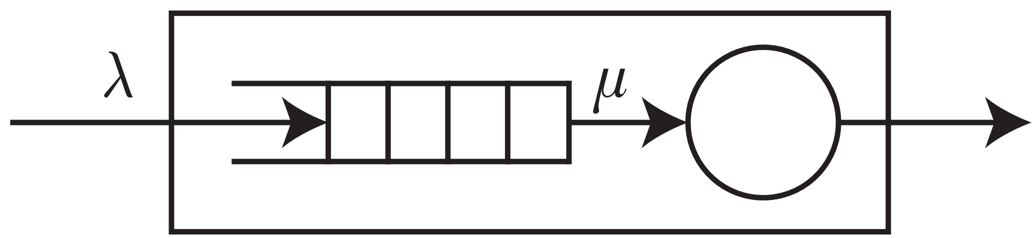
Favorite: Markovian (Exponential)

- combination of multiple independent distributions ⇒ exponential distribution
- when distribution is unknown, exp is a fair compromise: medium variability ($C_X = 1$)
- it really simplifies the math!

¶ M/M/1 queueing system

M/M/1 =

- Poisson arrival process
- Exponential service times
- 1 server
- ∞ buffer length
- ∞ population
- FCFS queue discipline



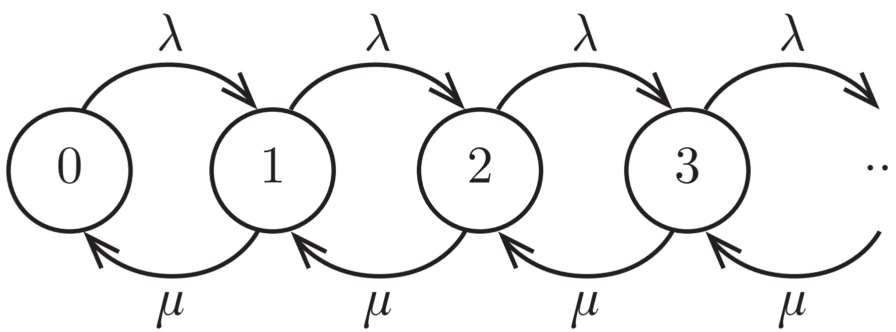
i.e., queueing system above with

exponential arrival rate λ,

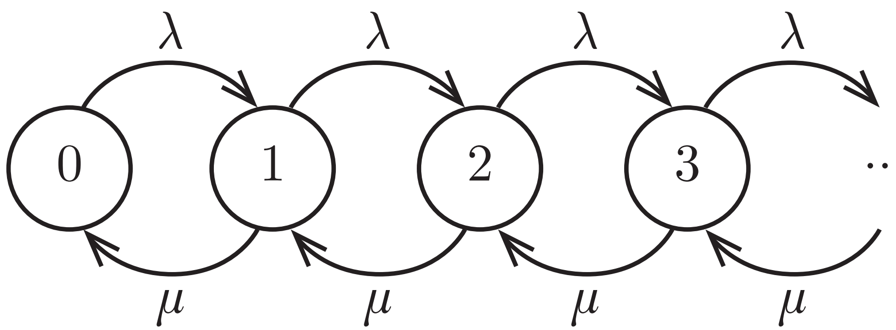
exponential service rate μ

L (# of customers in system) can be used to describe the *state* of the system.

model as a Markov chain:



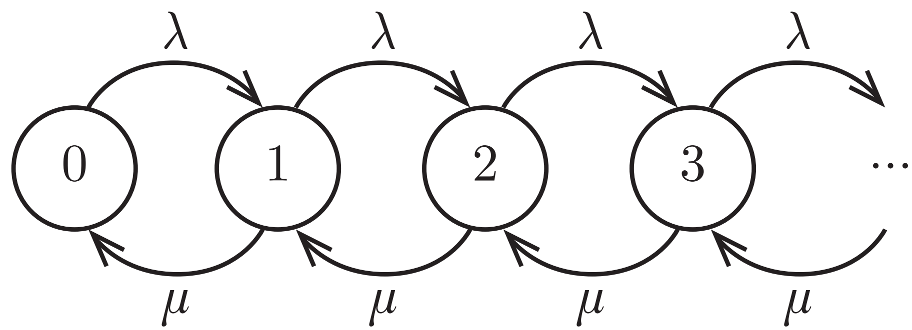
λ and μ are *rates of flow* between each state



$P(L_t = n)$ is probability of $L\!=\!n$ at time t

want the *limiting distribution* (i.e., at *equilibrium*):

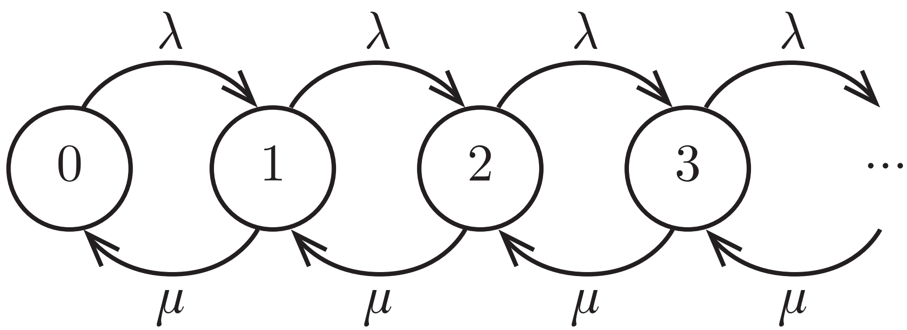
$$p_n = \lim_{t\to\infty} P(L_t = n \mid L_0 = i), i = 0, 1, 2, \ldots$$



“balance” equations (apply at equilibrium):

$$\lambda p_0 = \mu p_1$$

$$(\lambda + \mu) p_n = \lambda p_{n-1} + \mu p_{n+1}$$



$$\lambda p_0 = \mu p_1$$

$$(\lambda + \mu) p_n = \lambda p_{n-1} + \mu p_{n+1}$$

$$\text{together with } \sum_{n=0}^{\infty} p_n = 1,$$

can derive distribution of L

but we will limit ourselves to
mean value analysis

$$P(L_q = 0)$$

$$P(L_q \geq x)$$

$$P(T \leq t)$$

$$E(T_q), E(T)$$

$$E(L_q), E(L)$$

don't really need the distributions to compute these mean values …

Big help: PASTA property

"Poisson Arrivals See Time Averages"

i.e., *arriving customers* in a Poisson process see, *on average*, the same number of customers in the system as predicted by the *steady-state average*

$E(L) = 5$ people in store

- i.e., to the outside observer, there are 5 people in the store on average
- given Poisson arrivals, new customers on average also see 5 people in the store

not true in general!

consider deterministic system:

- arrival times = 1, 3, 5, 7, …
- service time = 1 (constant)
- $E(L) = 1/2$

but arriving customers always see 0 in store!

Start analysis with mean wait time: $E(T_q)$

$$E(T_q) = \begin{matrix}\text{time for waiting}\\ \text{customers ahead}\\ \text{of me to be served}\end{matrix} + \begin{matrix}\text{if the server is busy,}\\ \text{the \textit{residual service time}}\\ \text{of customer in service}\end{matrix}$$

$$\begin{aligned}E(T_q) &= E(L_q)E(T_s) + \rho E(T_r)\\ &= \lambda E(T_q)E(T_s) + \rho E(T_r)\end{aligned}$$

(by Little's Law)

$$E(T_q) - \lambda E(T_q)E(T_s) = \rho E(T_r)$$

$$E(T_q)(1 - \lambda E(T_s)) = \rho E(T_r)$$

$$E(T_q) = \frac{\rho E(T_r)}{1 - \lambda E(T_s)}$$

$$= \frac{\rho E(T_r)}{1 - \rho}$$

Pollaczek-Khinchin formula

$E(T_r)$?

consider deterministic case:

- if mean service time = 1 min, and we arrive to find server occupied, $E(T_r) = ?$
  - Ans: 30 sec $(E(T_s)/2)$

$E(T_r)$?

what about Poisson arrival process?

- by PASTA, new arrivals sees average!
- i.e., $E(T_r) = E(T_s)$

M/M/1 mean value formulae:

$$E(T_r) = E(T_s) = \frac{1}{\mu}$$

$$E(T_q) = \frac{\rho E(T_r)}{1-\rho} = \frac{\rho}{\mu(1-\rho)}$$

$$\begin{aligned} E(T) &= E(T_q) + E(T_s) \\ &= \frac{\rho}{\mu(1-\rho)} + \frac{1}{\mu} = \frac{1}{\mu(1-\rho)} \end{aligned}$$

$$\begin{aligned} E(L) &= \lambda E(T) \\ &= \frac{\lambda}{\mu(1-\rho)} = \frac{\rho}{1-\rho} \end{aligned}$$

e.g., Suppose a network server receives 40 requests per second, and the average service time is 20ms. Assuming requests are exponentially distributed:

1. What is the average server utilization?
2. What is the average time spent in the server's queue?
3. What is the average turnaround time for a request?

Little's Law:

$$E(L) = \lambda E(T)$$

Mean wait time:

$$E(T_q) = \frac{\rho}{\mu(1-\rho)}$$

Given: $\lambda = 40/\text{s}, E(T_s) = 20\text{ms} = 0.02\text{s}$

Find: $\rho, E(T_q), E(T)$

Utilization:

$$\rho = \lambda E(T_s) = 40 \times 0.02 = 0.8$$

Mean wait time:

$$E(T_q) = \frac{0.02 \times 0.8}{1 - 0.8} = 0.08\text{s} = 80\text{ms}$$

Mean sojourn time:

$$E(T) = E(T_q) + E(T_s) = 80\text{ms} + 20\text{ms} = 100\text{ms}$$

e.g., Suppose we upgrade the server and lower mean service time from 20ms to 15ms. By how much does this improve turnaround time? (Arrival rate still = 40/s)

$$\frac{20-15}{20} = 25\%$$ decrease in service time

$$\rho = 40 \times 0.015 = 0.6$$

$$E(T) = \frac{15\text{ms}}{1-0.6} = 37.5\text{ms}$$

$$\frac{100-37.5}{100} = 62.5\%$$ improvement!

e.g., A new cafeteria has just opened on campus, and is set up to service, on average, 2 students/minute. Students are only starting to trickle in, but the manager has already decided that when the average time needed for them to get their food approaches 5 minutes, capacity will be increased. Assuming a M/M/1 system:

1. What would the mean arrival rate need to be for the manager to increase capacity?

2. How many students would be waiting for service at this point?

$$E(L) = \lambda E(T) \qquad E(T) = \frac{1}{\mu(1-\rho)}$$

Given: $\mu = 2/\text{min}, E(T) \to 5\text{min}$

Find: $\lambda, E(L_q)$

$$E(T) = \frac{1}{\mu - \lambda}$$

$$5 = \frac{1}{\mu - \lambda} \Rightarrow \lambda = \mu - \frac{1}{5} = 1.8/\text{min}$$

$$E(T_q) = E(T) - E(T_s) = 5 - \frac{1}{2} = 4.5$$

$$E(L_q) = \lambda E(T_q) = 1.8 \times 4.5 = 8.1$$

e.g., Potbelly's is getting ready to open a new store at the MTCC, and is expecting approximately 8 students to arrive per minute during the lunch rush. If they want to guarantee that no more than 10 students, on average, are waiting in line to get serviced, how quickly must they be able to take and complete orders?

Little's Law:

$$E(L) = \lambda E(T)$$

Mean wait time:

$$E(T_q) = \frac{\rho}{\mu(1-\rho)}$$

Given: $\lambda = 8$, want $E(L_q) \leq 10$

Find: μ

$$E(T_q) = \frac{E(L_q)}{\lambda} = \frac{10}{8} = 1.25 = \frac{\rho}{\mu(1-\rho)}$$

$$1.25 = \frac{\lambda}{\mu^2(1-\rho)} = \frac{\lambda}{\mu^2 - \mu\lambda}$$

$$\mu^2 - 8\mu - 6.4 = 0$$

¶ M/G/1 queueing system

Arbitrary (General) service distribution, but:

- Can assume stable system ($\rho<1$)
- Little's Law still applies
- Still have PASTA!

$$E(T_q) = \frac{\rho E(T_r)}{1-\rho}$$

$E(T_r)$ depends on mean and variance of service times

Intuition: larger variance means that residual time is a bigger fraction of the mean

$$E(T_r) = \frac{\sigma^2(T_s) + E(T_s)^2}{2E(T_s)} = \frac{C_{T_s}^2 + 1}{2} \cdot E(T_s)$$

for exponential, $C_{T_s}^2 = 1$, so $E(T_r) = \frac{1+1}{2} \cdot E(T_s) = E(T_s)$

for deterministic, $C_{T_s}^2 = 0$, so $E(T_r) = \frac{0+1}{2} \cdot E(T_s) = \frac{E(T_s)}{2}$

e.g., A print shop has 4 clients, each of which sends in a job every half hour on average, distributed exponentially. It takes an average of 6 minutes to print each job (there is one printer), and the service distribution can be described with $C^2=1.5$.

1. How many jobs are, on average, waiting?
2. What is the average job turnaround time?

Given:

$$\lambda = 4 \times 2/\text{hour} = 8/\text{hour}$$
$$E(T_s) = 6 \text{ min} = 0.1 \text{ hour}$$
$$C_{T_s}^2 = 1.5$$

Find: $E(L_q), E(T)$

$$\rho = 8 \times 0.1 = 0.8$$

$$E(T_q) = \frac{\rho}{1-\rho} \cdot \frac{C_{T_s}^2 + 1}{2} \cdot E(T_s)$$
$$= \frac{0.8}{0.2} \times \frac{2.5}{2} \times 0.1 = 0.5 \text{ hour}$$

$$E(L_q) = \lambda E(T_q) = 8 \times 0.5 = 4$$

$$E(T) = E(T_q) + E(T_s) = 0.6 \text{ hour} = 36 \text{ min}$$