

# CPE345: Modeling and Simulation

## Lecture 5

# Today's topics

- More on random variables and distributions of random variables
- Introduction to queueing models (chapter 6)
- Announcement:
  - Midterm – March 14

# More details about the exponential distribution

$$\text{pdf: } f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{ow} \end{cases} \quad \begin{aligned} E(X) &= \frac{1}{\lambda} \\ \text{var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

$\lambda$  is a rate: e.g. arrival rate, service rate, failure rate, etc...

- Some important properties:

- **Memory-less property:**  $P(X > s + t \mid X > s) = P(X > t)$

conditional probability: for two events A, B:

$$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

We can then show the memory-less property of the exponential r.v.

$$P(X > s + t \mid X > s) = \frac{P(X > s + t, X > s)}{P(X > s)} = \frac{P(X > s + t)}{P(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t}$$

3

# Example for exponential distribution

- Suppose a bus arrives at a bus station, such that the inter-arrival time between buses is exponential distributed with mean  $\mu = 10$  minutes.
- Suppose that you already have waited for the bus for 10 minutes. Questions:

- What is the probability that you will still have to wait for at least another 15 minutes?

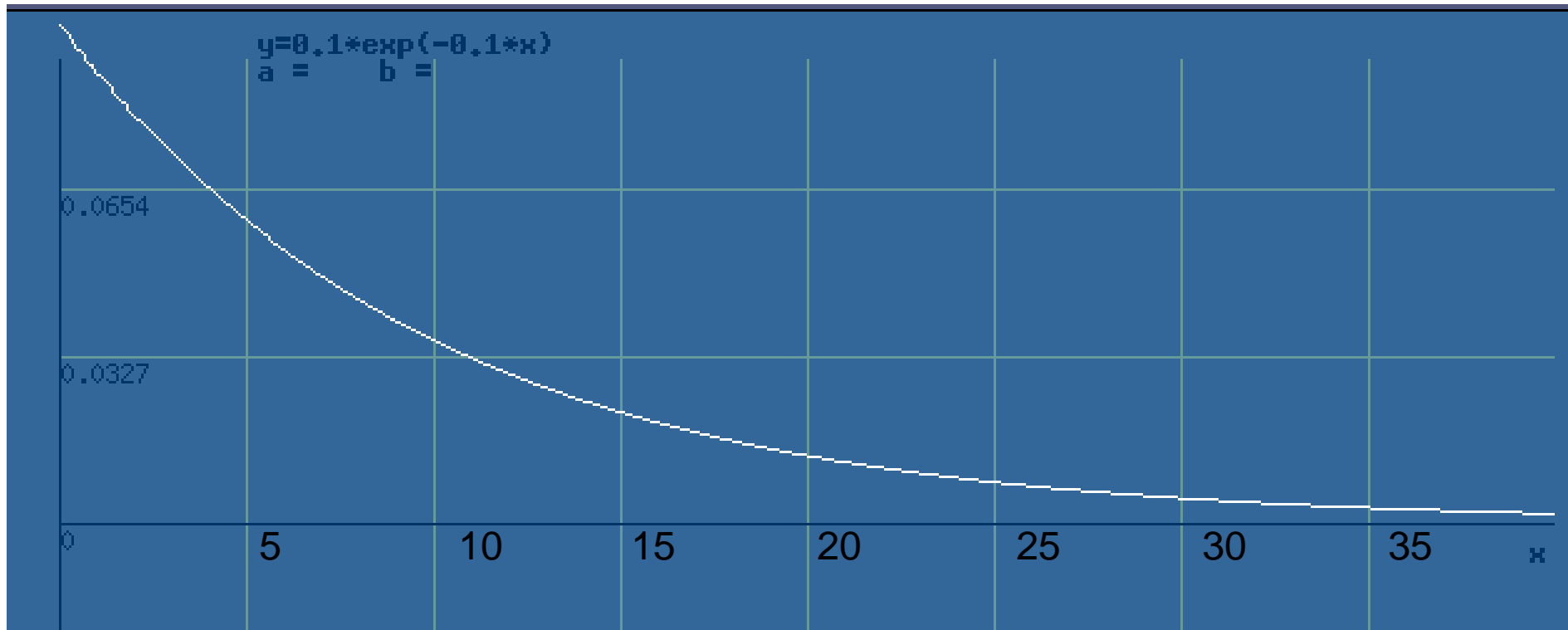
$$P(X > 10 + 15 | X > 10) = P(X > 15) = \int_{15}^{\infty} 0.1e^{-0.1x} dx = -e^{-0.1x} \Big|_{15}^{\infty} \\ = e^{-0.1 \cdot 15} = 0.22$$

- What is the probability that you will still have to wait less than 5 minutes?

$$P(X < 10 + 5 | X > 10) = P(X < 5) = 1 - \int_5^{\infty} 0.1e^{-0.1x} dx = 0.39$$

# Exponential distribution pdf

Exponential:  $\lambda = 0.1$ ;  $\mu = 10$



Source for the plot: <http://www.wessa.net/math.wasp>

# Relation with Poisson r.v.

- If the interval between generation of events (e.g. arrival, service) is an exponential r.v. with mean  $\mu = 1/\lambda$ , then the event generation process is a Poisson process, with mean  $\lambda$ .
  - Example: If buses arrive at the station at intervals that are exponentially distributed, the arrival process for the buses is Poisson.
    - Questions: If the mean time between arrivals is  $\mu = 10$  minutes,
      - (1) What is the probability that a traveler has to wait for the bus for more than 15 minutes?
      - (2) What is the probability that at most 2 busses will arrive in the station within the first  $\frac{1}{2}$  hour?

$$(1) \quad P(t > 15) = e^{-0.1 \cdot 15} \approx 0.22$$

$$(2) \quad P(N \leq 2) = \sum_{i=0}^2 \frac{e^{-0.1 \cdot 30} \cdot 3^i}{i!} \approx 0.049 + 0.15 + 0.22 = 0.4195$$

# Poisson process

- A counting process  $\{N(t), t \geq 0\}$  ( $N(t)$  represents the number of events that occurred in the interval  $[0, t)$ ) is a Poisson process if
  - Arrivals occur one at a time
  - $\{N(t), t \geq 0\}$  has stationary increments: the distribution of the number of arrivals for the interval  $t+s$ , depends only on the length of the observation interval  $s$ , and is independent on the initial starting point  $t$
  - $\{N(t), t \geq 0\}$  has independent increments: the number of arrivals for non-overlapping time intervals are independent random variables.
  - The probability of  $n$  arrivals in the interval  $[0, t)$  is given as

$$P(N(t) = n) = \frac{e^{-\lambda t} (\lambda t)^n}{n!}, \quad t > 0, \quad n = 0, 1, 2, \dots$$

# Some useful properties of the Poisson process

- Random splitting
  - If a Poisson arrivals process with rate  $\lambda$  is split using a coin flipping (probability of a head =  $p$ ) into two types of arrivals A and B, the resulting arrival processes are also Poisson with rates  $\lambda_A = \lambda p$ , and  $\lambda_B = \lambda(1 - p)$ , respectively
- Pooling of two or more arrival streams
  - If  $n$  arrival streams are pooled together, the resulting arrival process will be Poisson, with the rate equal to the sum of the rates of the individual processes.

$$\lambda_p = \sum_{i=1}^n \lambda_i$$

# More on random variable distributions

- Poisson and exponential random variables are extensively used for queueing theory analysis and modeling of queueing systems
- If you add  $k$  independent exponential random variables, with rate  $\lambda$ , the resulting random variable has an Erlang distribution of order  $k$ :

$$f(x) = \frac{\lambda^k e^{-\lambda x} x^{k-1}}{(k-1)!}, \quad x \geq 0$$

- *For  $k=1 \rightarrow$  exponential*

- *CDF:* 
$$F(x) = 1 - \sum_{i=0}^{k-1} \frac{(\lambda x)^i e^{-\lambda x}}{i!}$$

- *Mean and variance:* 
$$E(X) = \frac{k}{\lambda}; \quad \text{var}(X) = \frac{k}{\lambda^2}$$

# Gamma distribution

- The gamma distribution generalizes the Erlang distribution

$$f(x) = \frac{\lambda^\alpha e^{-\lambda x} x^{\alpha-1}}{\Gamma(\alpha)}, \quad x \geq 0,$$

$$\text{where, } \Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1} e^{-t} dt, \quad \Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$$

- Some properties:

$$\Gamma(1/2) = \sqrt{\pi}$$

$$\text{if } \alpha \text{ integer, } \Gamma(\alpha) = (\alpha - 1)!$$

# Rayleigh and Lognormal Distributions

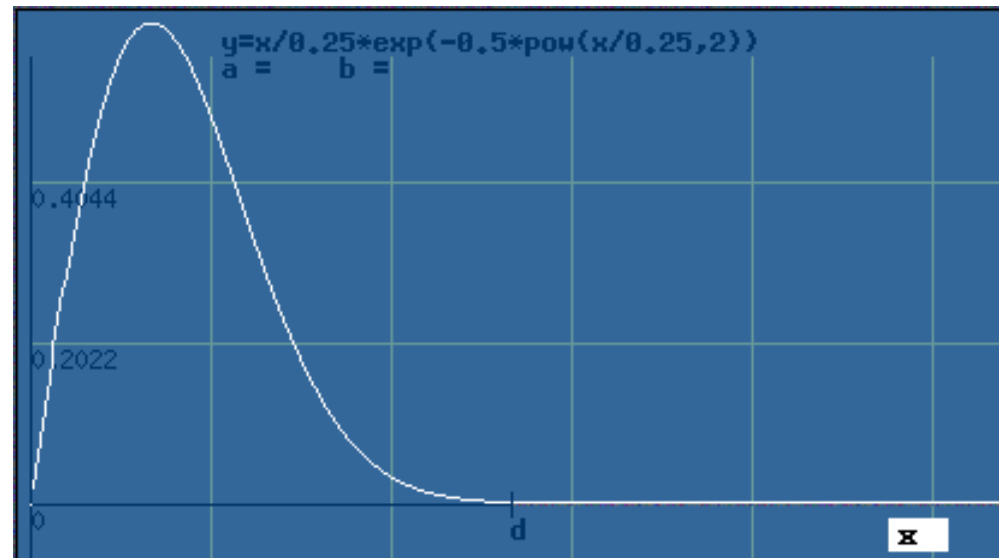
- Both are used in wireless communications for modeling different types of fading experienced by the radio transmission
  - Fast fading: modeled by the Rayleigh distribution (appears as an effect of the motion)
  - Slow fading: modeled by the Lognormal distribution (appears as an effect of the environment)

## Rayleigh distribution

$$f(x) = \frac{x}{p} \exp\left(-\frac{x^2}{2p}\right), x \geq 0$$

$$E(X) = \sqrt{\frac{\pi}{2}} p; \quad \text{var}(X) = \frac{4 - \pi}{2} p$$

Rayleigh:  $p = 0.25$



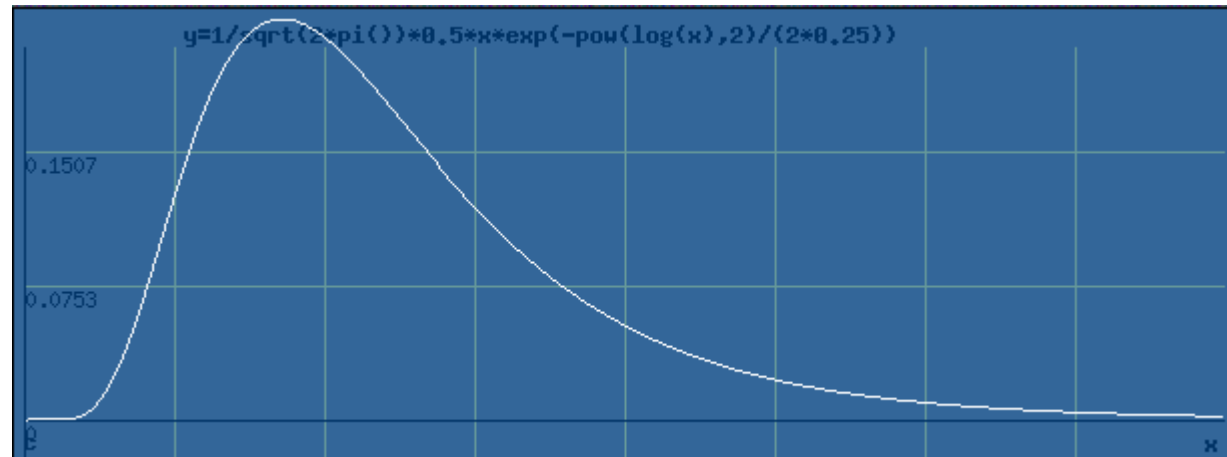
# Lognormal distribution

- pdf:  $f(x) = \frac{1}{\sqrt{2\pi\sigma x}} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right], \quad x > 0$
- If  $X$  is lognormal,  $\ln(X)$  is normal distributed with mean  $\mu$  and variance  $\sigma^2$
- Mean and variance for the lognormal distribution

$$\mu_L = e^{\mu + \sigma^2/2}$$

$$\sigma_L^2 = e^{\sigma^2 + 2\mu} \left( e^{\sigma^2} - 1 \right)$$

Lognormal:  $\mu=0, \sigma = 0.5$



# Empirical Distributions

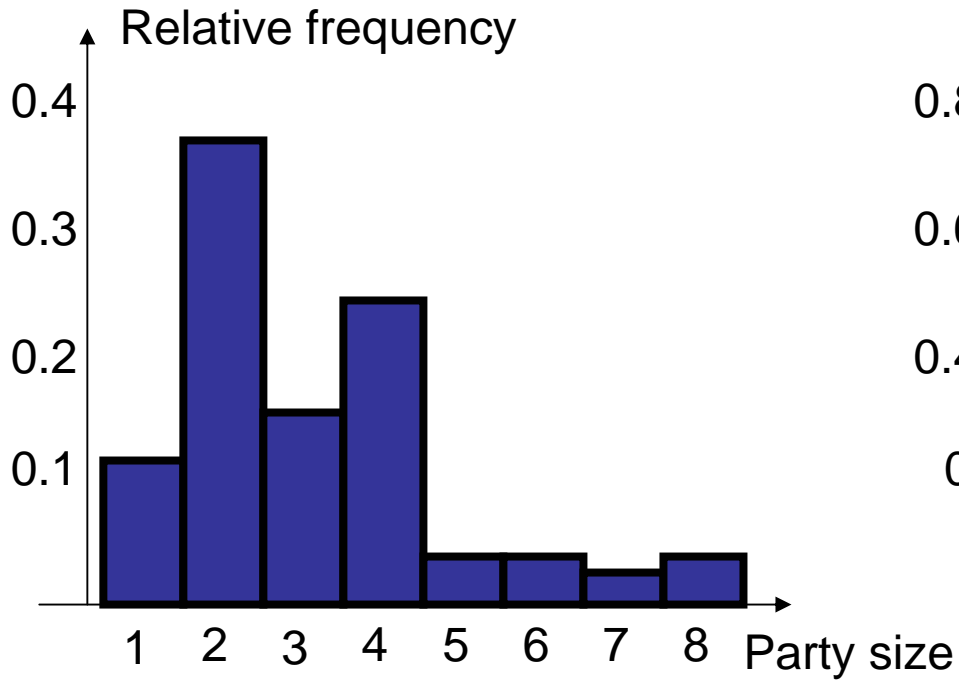
- Sometimes, impossible (or unnecessary) to establish that a random variable has any particular known distribution
- Solution: determine an empirical distribution
  - discrete
  - continuous
- Discrete example
  - Customers at a local restaurant arrive in parties of 1-8 persons. The last 300 arrivals were monitored, and the frequency of different group sizes has been determined

## Discrete example - cont

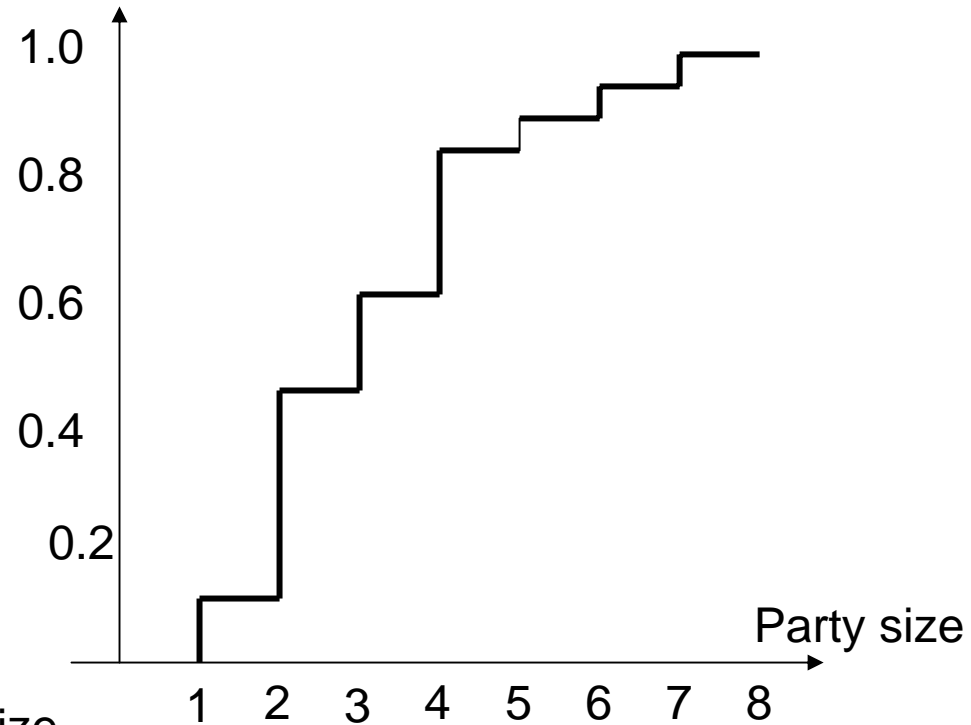
Size of party	Frequency	Relative frequency	Cumulative relative frequency
1	30	0.10	0.10
2	110	0.37	0.47
3	45	0.15	0.62
4	71	0.24	0.86
5	12	0.04	0.90
6	13	0.04	0.94
7	7	0.02	0.96
8	12	0.04	1.00

# Discrete example - cont

Histogram of party size



Cumulative relative frequency



# Continuous empirical distribution example

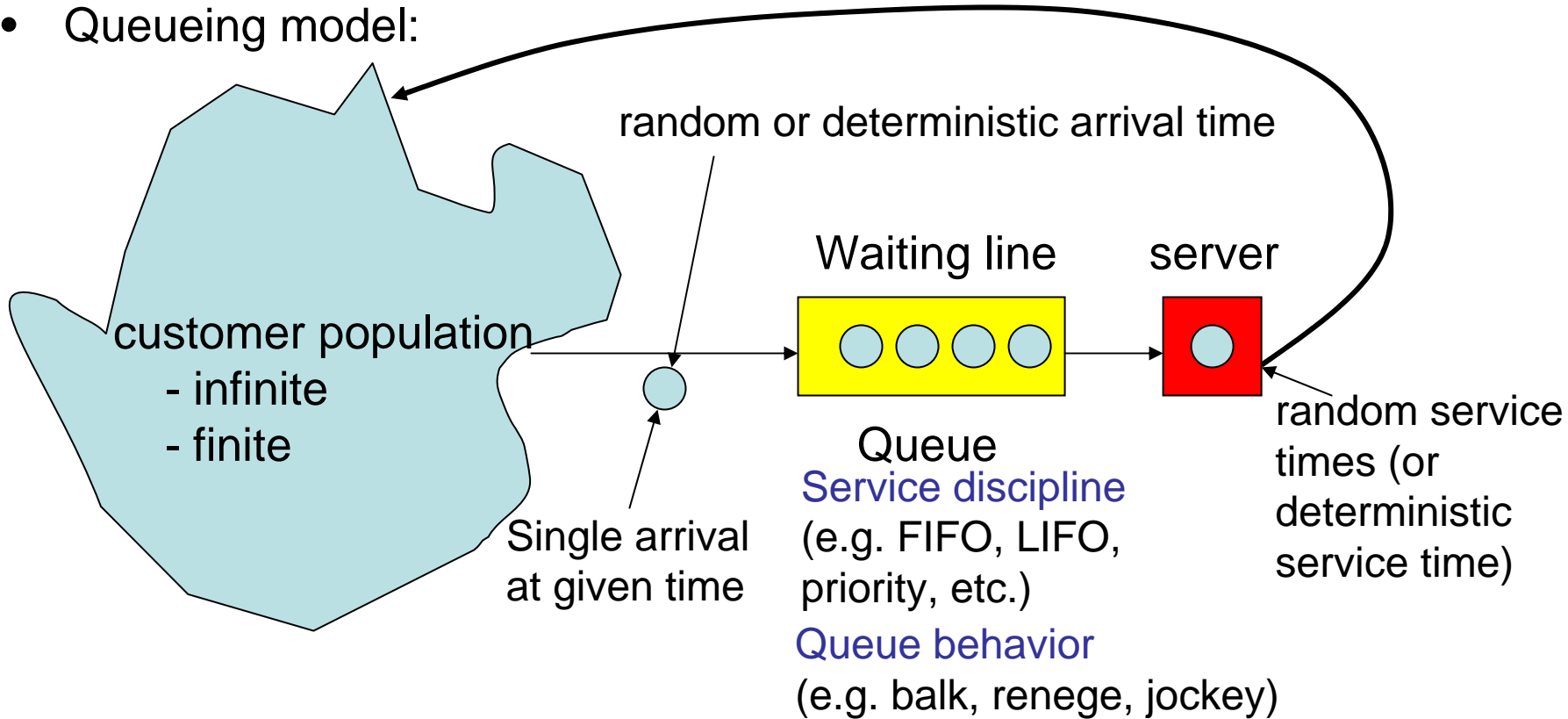
- Determine an empirical cdf and pdf for the driving time from Hoboken to Philadelphia. Consider that you are a commuter that goes 5 days a week, 4 weeks a month from Hoboken to Philadelphia and observes the driving time for the past year:  $12 \cdot 4 \cdot 5 = 240$  days

Driving time interval (minutes)	Frequency	Relative frequency	Cumulative relative frequency
(100, 110]	50	0.21	0.21
(110, 120]	110	0.46	0.67
(120, 130]	40	0.17	0.84
(130, 140]	20	0.08	0.92
(140, 150]	18	0.072	0.992
(150,160]	2	0.008	1.00

**Homework:** Plot the empirical pdf and cdf. Hint: determine the histogram using the intervals for the horizontal axis.

# Introduction to queueing models

- Queueing model:



**Requirement:** service time < arrival time

**Performance measures:** delay, queue length, server utilization

**Queueing performance tradeoffs:** server utilization ↗ → Delay and queue length ↗

# To simulate or to analyze?

- Simple queueing models (some of the examples we have already discussed) can be analyzed. Simulation for these cases has only didactic value.
- Complex queueing systems (e.g. non-Poisson arrivals and general distributions for service, tandem queues, complex network of queues) are hard to analyze, need to be simulated

# Characteristics of queueing systems

- Calling population

- Finite: Customers in queue have reduced the available size of population, causing a reduction in the arrival rate
- Infinite: customers already in the queue do not influence the arrival rate process.

- System capacity

- There may be a limit on the queue size
- Customers which arrive and find the queue full, will return to the calling population
- Effective arrival rate: number of customers who arrive and enter the system (are served or are waiting in queue to be served) per unit time

# Characteristics of queueing systems: cont.

- **The arrival process**
  - Characterized in terms of interarrival times between successive customers
  - Arrivals may occur at scheduled times (deterministic, or constant  $\pm$  a small random number) or at random times (characterized by a pdf)
  - Customers may arrive one at a time or in batches, that can be constant size or random size
  - Most important model for arrivals: Poisson arrival process
- **Queue behavior** – customer actions while in queue
  - Balk: incoming customers may leave when they see the queue is too long
  - Renege: leave after being in the line, when they see that the line is moving too slowly
  - Jockey: move from one line to another, if they think they have chosen a slow line

# Characteristics of queueing systems: cont.

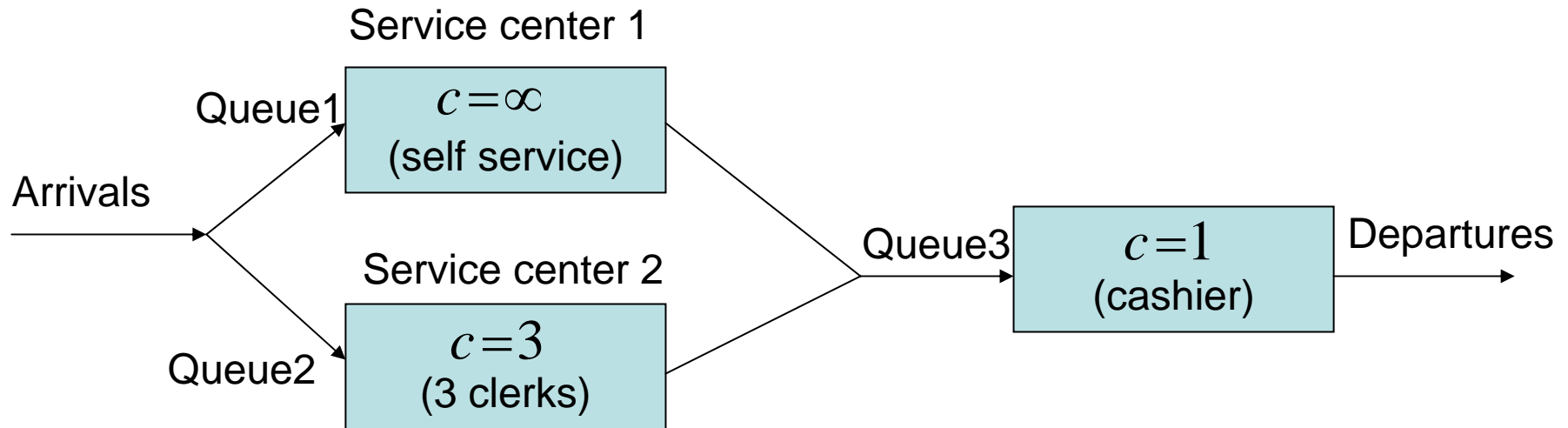
- Queue disciplines
  - FIFO: first-in-first-out
  - LIFO: last-in-first-out
  - SIRO: service in random order
  - SPT: shortest processing time first
  - PR: service according to priority

**Note:** FIFO implies that service begins in the same order as the arrivals, but the customers may leave the system in different order, due to different service times

# Service time and the service mechanism

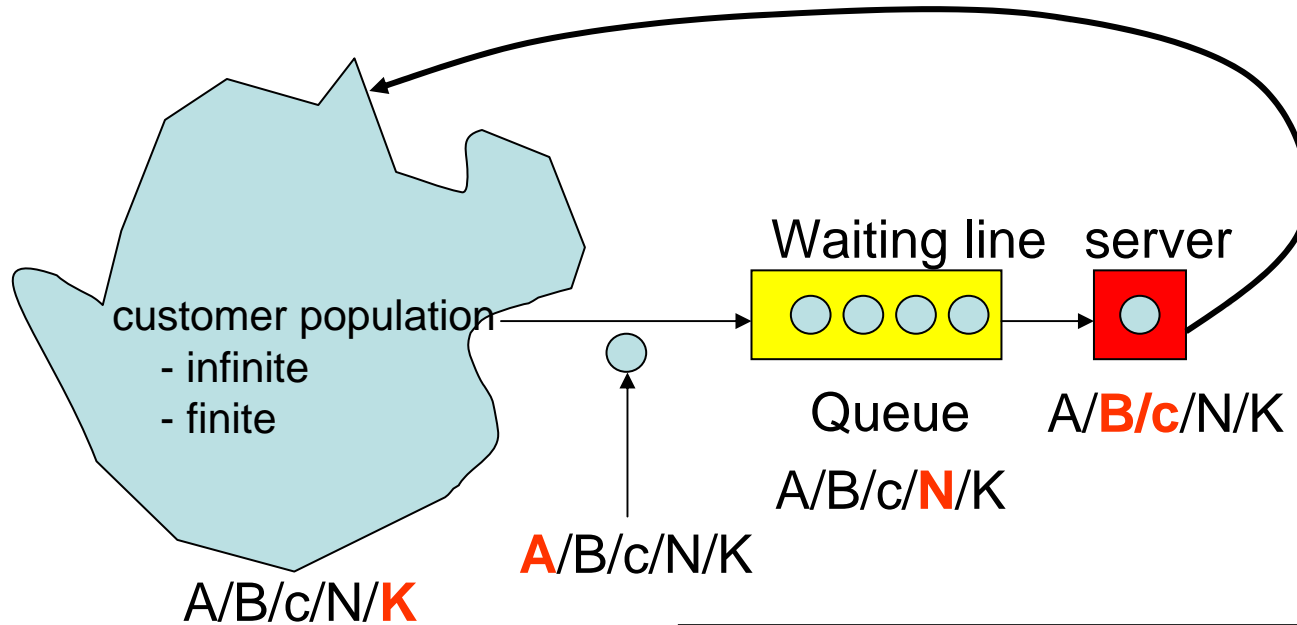
- **Service times** of successive arrivals:  $S_1, S_2, S_3, \dots$ 
  - Deterministic (constant) or random
  - The sequence of random variables  $\{S_1, S_2, S_3, \dots\}$  is i.i.d. (independent and identically distributed)
  - Most commonly used service time: exponential distributed
  - Service times may be identically distributed for customers in the same class, but customers in different classes may have different service times
- **Queueing system** – number of service centers and interconnecting queues
  - Service center – some number of servers,  $c$ , working in parallel

# Queueing system example



Discount warehouse: customers may either serve themselves or wait for one of the 3 clerks, then leave after paying to a single cashier

# Queueing notation: A/B/c/N/K



- A**: inter-arrival time distribution
- B**: service time distribution
- c**: number of parallel servers
- N**: system capacity
- K**: the size of calling population

Some notations for A and B:

- M – exponential distributed (inter-arrivals or service times) – Markov
- D – constant/deterministic
- G – arbitrary/general

Note: if N, K, infinite, they may be dropped from the notation

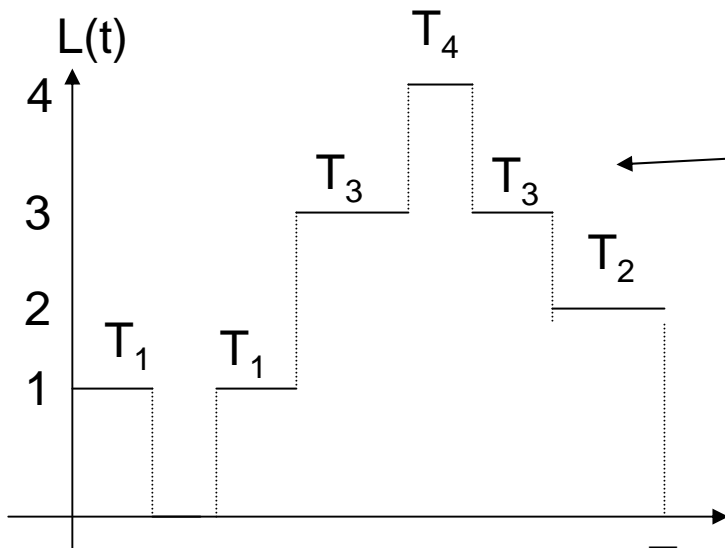
# Queueing notation for parallel server systems

$P_n$	Steady-state probability of having $n$ customers in the system
$P_n(t)$	Probability of $n$ customers in system at time $t$
$\lambda$	Arrival rate
$\lambda_e$	Effective arrival rate
$\mu$	Service rate of one server
$\rho$	Server utilization
$A_n$	Inter-arrival time between customer $n-1$ and $n$
$S_n$	Service time of the $n^{th}$ arriving customer
$W_n$	Total time spent in the system by the $n^{th}$ arriving customer
$W_n^Q$	Total time spent in the waiting line by the $n^{th}$ arriving customer
$L(t)$	Number of customers in system at time $t$
$L_Q(t)$	The number of customers in queue at time $t$
$L$	Long-run time-average number of customers in the system
$L_Q$	Long-run time-average number of customers in queue
$w$	Long-run average time spent in system per customer
$w_Q$	Long-run average time spent in queue per customer

# Long-run measures of performance

- Time average number in system:  $L$ 
  - Observe the system for period  $T$ ,  $L(t) = \text{no. customers at time } t$
  - $T_i = \text{total time during } [0, T], \text{ in which the system contained exactly } i \text{ customers}$

$$\hat{L} = \frac{1}{T} \sum_{i=0}^{\infty} iT_i = \sum_{i=0}^{\infty} i \left( \frac{T_i}{T} \right)$$



$$\sum_{i=0}^{\infty} iT_i = \int_0^T L(t) dt$$

$$\hat{L} = \frac{1}{T} \int_0^T L(t) dt \rightarrow L, \text{ as } T \rightarrow \infty$$

Long run stability in terms of the average performance

# Long-run measures of performance

- Time average number in queue:  $L_Q$

- Same reasoning as before, leads to

$$\hat{L}_Q = \frac{1}{T} \int_0^T L_Q(t) dt \rightarrow L_Q, \text{ as } T \rightarrow \infty$$

- Similarly, we can define

- Average time spent in system by customer

- $N$  = number of arrivals during  $[0, T]$

$$\hat{w} = \frac{1}{N} \sum_{i=0}^{\infty} W_i \rightarrow w, \text{ as } T \rightarrow \infty, N \rightarrow \infty$$

- Average time spent in queue by customer

$$\hat{w}_Q = \frac{1}{N} \sum_{i=0}^{\infty} W_i^Q \rightarrow w_Q, \text{ as } T \rightarrow \infty, N \rightarrow \infty$$

# Homework

- From slide 16:

**Homework:** Plot the empirical pdf and cdf. Hint: determine the histogram using the intervals for the horizontal axis.

- Problem 28, page 199 (chapter 5)