Statistics for Computing MA4413

Lecture 9

The Exponential Distribution and Queueing Theory

Kevin Burke

kevin.burke@ul.ie

No Poisson Events in t Intervals

We saw in Lecture8 that

$$X \sim \text{Poisson}(\lambda) \text{ in 1 interval.}$$

$$\Rightarrow X \sim \text{Poisson}(\lambda t) \text{ in } t \text{ intervals.}$$

The probability that there are *no events* in t intervals (of time) is:

$$\Pr(X = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = \frac{1}{1} e^{-\lambda t} = e^{-\lambda t}.$$

No event in t intervals of time $\Rightarrow$ it happens sometime after this.

Thus, $\Pr(X = 0)$ is the probability that the event *time* is greater than t.

Exponential Distribution

X is the number of events in an interval of time.

Now let T be the random variable representing the *waiting time* between Poisson events. Note that $T \sim$ **exponential distribution**.

From the previous slide we have that

$$\Pr(T > t) = \Pr(X = 0) = e^{-\lambda t}.$$

This is the probability function for the exponential distribution.

(recall that the Poisson distribution also applies to intervals other than time, e.g., distance / area etc. In such cases T is the distance / space between Poisson events)

Exponential Distribution

The **exponential distribution** is used for calculating the probability of waiting more than t units of time / distance / space for a Poisson event:

$$\boxed{T \sim \text{Exponential}(\lambda)}$$

$$\boxed{\Pr(T > t) = e^{-\lambda t}}$$

where $\boxed{t \in [0, \infty)}$ is *continuous*

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

$$\boxed{Var(T) = \frac{1}{\lambda^2}}$$

Continuous Vs Discrete Distributions

The exponential distribution is **continuous** (i.e., the random variable T is continuous) unlike the Binomial and Poisson distributions which are *discrete*.

We *never* calculate "equal to" probabilities for continuous distributions!

(recall from Lecture8 that the probability of a system crashing at any precise moment in time is zero, i.e., $\Pr(T = t) = 0$)

Instead, we calculate the probability that T is greater than or less than some value or the probability that T lies in some range of values.

$\Rightarrow$ The probability function used in the exponential case is $\Pr(T > t)$.

Continuous Vs Discrete Distributions

For discrete variables we had $\Pr(X > 1) = \Pr(X \geq 2)$, i.e., the next *discrete* number after 1 is 2.

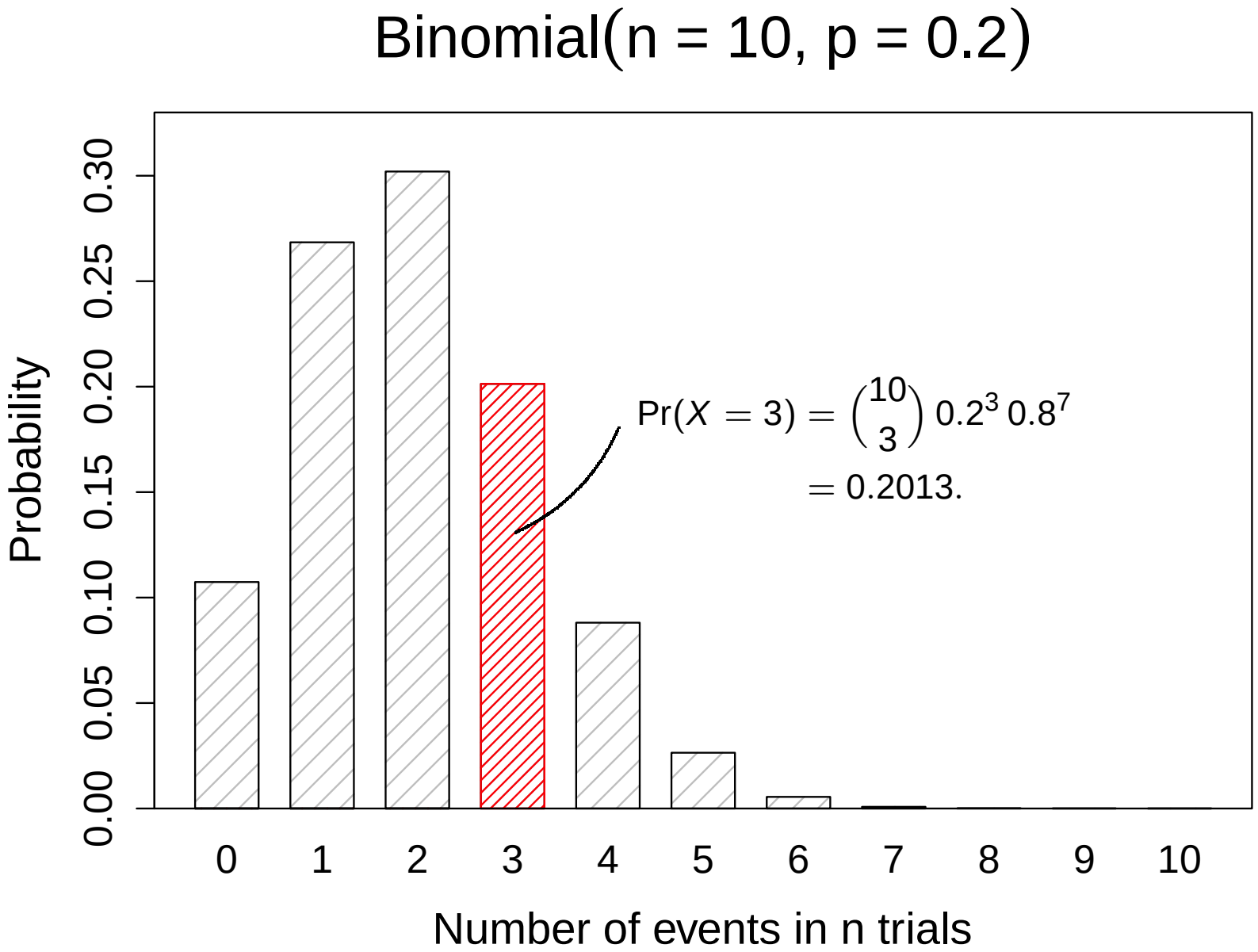
For *continuous* variables there is really no "next number". Think about values such as $1 + 0.01$, $1 + 0.00001$, $1 + 10^{-10}$, $1 + 10^{-1000}$...

Clearly the next continuous number after 1 is indistinguishable from 1

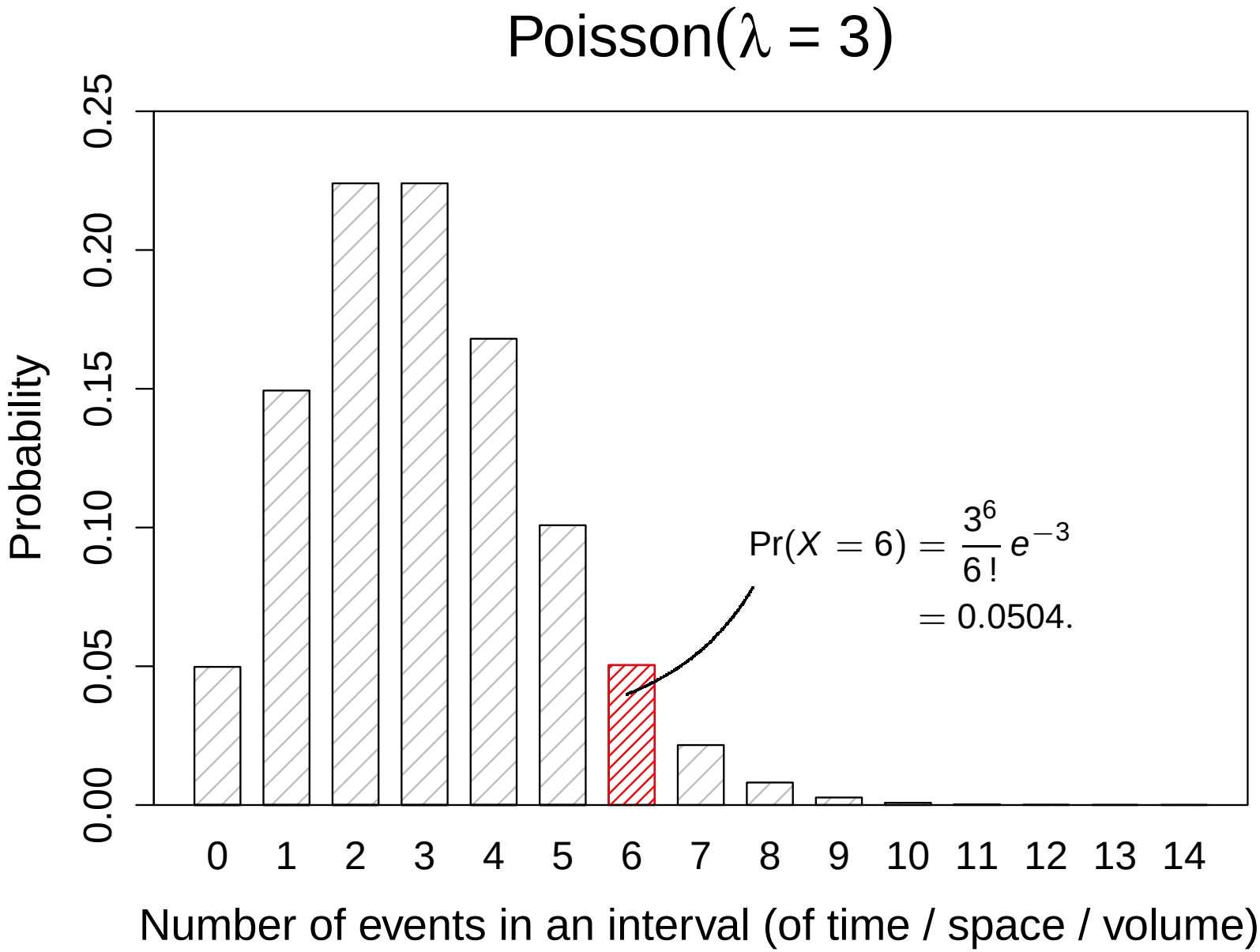
$\Rightarrow \Pr(T > 1) = \Pr(T \geq 1)$.

We make *no distinction* between "$>$" and "$\geq$" in the continuous case.

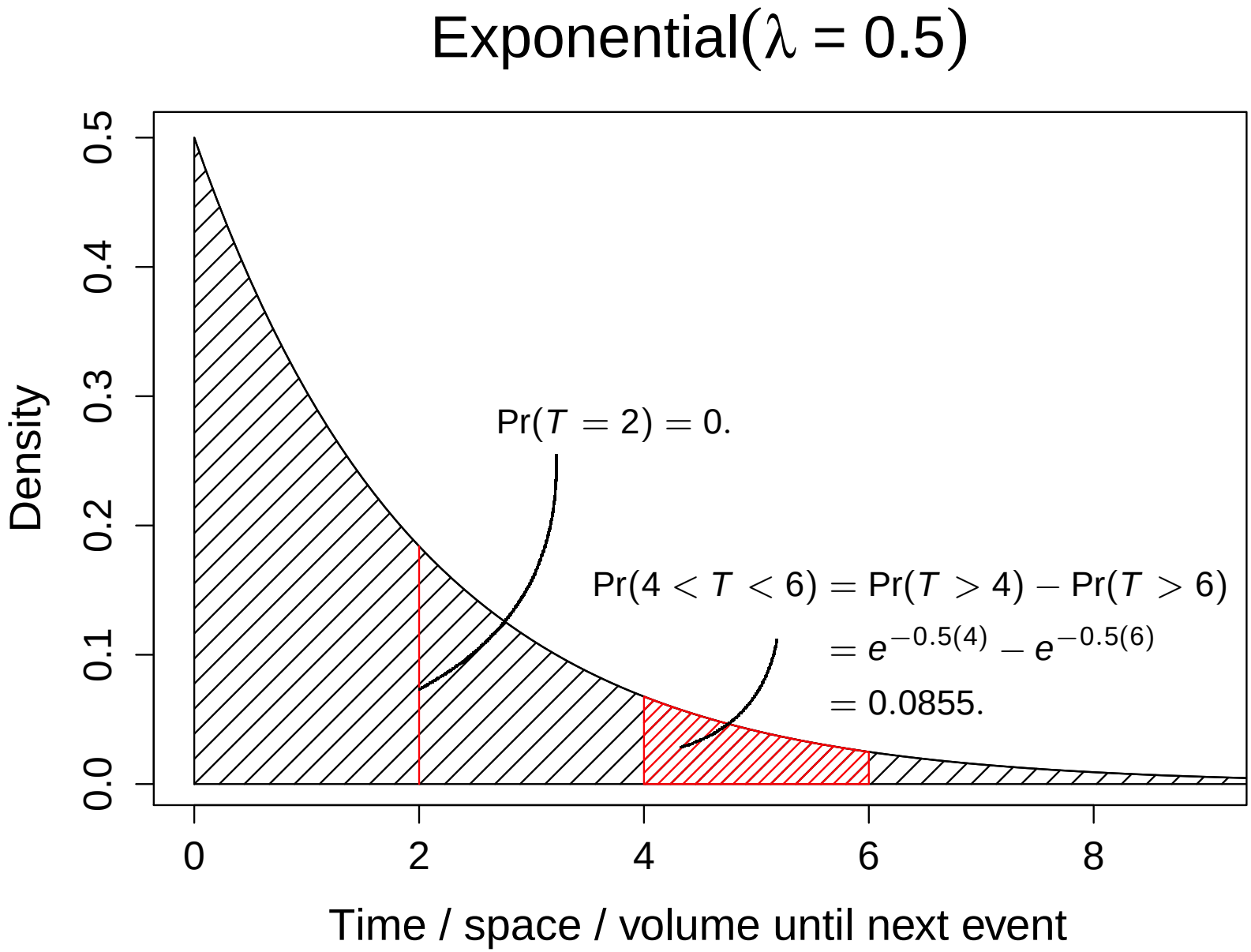
Discrete


Binomial(n = 10, p = 0.2)
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
1
2
3
4
5
6
7
8
9
10
Number of events in n trials
$\Pr(X = 3) = \binom{10}{3} 0.2^3 0.8^7$
$= 0.2013.$

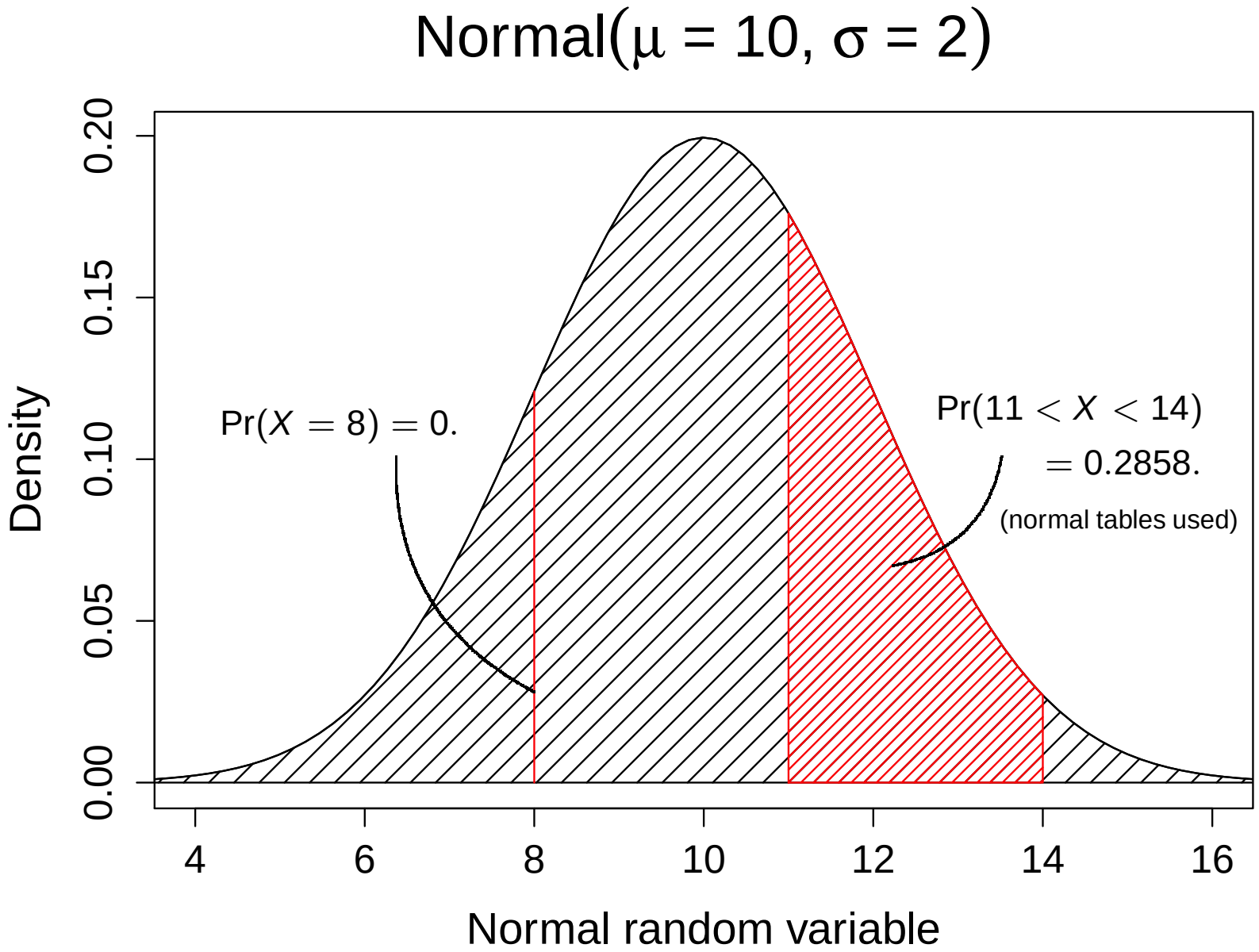
Discrete


Poisson(λ = 3)
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of events in an interval (of time / space / volume)
$\Pr(X = 6) = \frac{3^6}{6!} e^{-3}$
$= 0.0504.$

Continuous


Exponential(λ = 0.5)
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
Pr(T = 2) = 0.
Pr(4 < T < 6) = Pr(T > 4) − Pr(T > 6)
= e^−0.5(4) − e^−0.5(6)
= 0.0855.
Time / space / volume until next event

Continuous


Normal(μ = 10, σ = 2)
Density
0.00
0.05
0.10
0.15
0.20
Pr(X = 8) = 0.
Pr(11 < X < 14)
= 0.2858.
(normal tables used)
4
6
8
10
12
14
16
Normal random variable

Example: System Crash Time

The number of system crashes in a year is $X \sim \text{Poisson}(\lambda = 3)$.

$\Rightarrow$ the *time between* system crashes is $T \sim \text{Exponential}(\lambda = 3)$.

$\Rightarrow \Pr(T > t) = e^{-3t}$.

What is the probability that the system works for more than 6 months (i.e., 0.5 years) without a crash?

In other words, we wait more than 0.5 years for the next crash:

$$\Pr(T > 0.5) = e^{-3(0.5)} = e^{-1.5} = 0.2231.$$

Example: System Crash Time

What is the probability that the next crash happens within 3 months?

In other words, we wait less than $\frac{3}{12} = 0.25$ years for the next crash:

$$\begin{aligned}
\Pr(T < 0.25) &= 1 - \Pr(T \geq 0.25) && (\geq \text{ the same as } >) \\
&= 1 - \Pr(T > 0.25) && (\text{since } T \text{ is continuous}) \\
&= 1 - e^{-3(0.25)} \\
&= 1 - 0.4724 \\
&= 0.5276.
\end{aligned}$$

Example: System Crash Time

What is the probability that the next crash occurs between 6 months and 1 year from now.

$$\begin{aligned}\Pr(0.5 < T < 1) &= \Pr(T > 0.5) - \Pr(T > 1) \\ &= e^{-3(0.5)} - e^{-3(1)} \\ &= e^{-1.5} - e^{-3} \\ &= 0.2231 - 0.0498 \\ &= 0.1733.\end{aligned}$$

Example: System Crash Time

What is the *average* waiting time?

$$E(T) = \frac{1}{\lambda} = \frac{1}{3} \text{ years},$$

i.e., on average we wait $(\frac{1}{3} \times 12 =)$ 4 months for a crash. This should make sense as there are 3 crashes per year.

What is the standard deviation of T?

$$Var(T) = \frac{1}{\lambda^2} = \frac{1}{9} \text{ years}^2.$$

$$Sd(T) = \sqrt{\frac{1}{9}} = \frac{1}{3} \text{ years}.$$

Example: Laptop Battery Life

Note: *Often you have $E(T)$ and need to calculate λ.*

Let's assume that the battery life of a laptop has an exponential distribution with a mean of 3 hours, i.e., $E(T) = 3$.

We know that $E(T) = \frac{1}{\lambda}$ from which we find that:

$$\boxed{\lambda = \frac{1}{E(T)}}.$$

Laptop battery life is $T \sim$ Exponential($\lambda = \frac{1}{3}$) and, the probability function is $\Pr(T > t) = e^{-\frac{1}{3}t}$.

Question 1

The *average time* between customers arriving to a shop is 5 minutes. We will assume that the time, T, has an exponential distribution. Calculate the following:

a) The average arrival *rate*, i.e., λ customers per minute.

b) The probability that we wait more than 15 minutes for the next customer.

c) The probability that the next customer arrives within 1 minute.

d) The average *number of customers* in a 1 hour period. What is the standard deviation that goes with this average?

e) The probability that *15 or more* customers arrive in a 1 hour period.

R Code

For the exponential distribution we calculate *greater than* probabilities, i.e., $\Pr(T > t)$. Note that `rate` $= \lambda$.

Examples:

`pexp(0.5,rate=3,lower=F)`
gives `0.2231302`.

`pexp(1,rate=3,lower=F)`
gives `0.04978707`.

Compare this with slide 13.

(Warning: although there is also a `dexp` function, it differs from `dbinom` or `dpois` since it does not give $\Pr(T = t)$. The values produced by `dexp` are called *density* points - this fact is alluded to on slides 9 and 10)

R Code

We can *generate* exponential random variables as follows:

Example:

`rexp(100,rate=3)`
generates 100 Exponential($\lambda = 3$) variables.

Queueing Theory

The theory of *Poisson arrivals* and *exponential waiting times* is useful in the study of **queues** (i.e., waiting lines).

Examples:

- **Customers waiting to be served**.
- Jobs waiting to be processed.
- Planes waiting to land.
- Objects manufactured on a factory line.
- Traffic at a busy junction.

Definitions and Notation

- The **system** is the *whole* system, i.e., *queue* + *service*.
- λ_a is the **arrival rate** to the system.
- λ_s is the **service rate** *within* the system.
- N is the **total number** of customers in the system.
- T is the **total time** that a customer spends in the system.

Important note: we require that $\boxed{\lambda_a < \lambda_s}$ to prevent the system from becoming overloaded with customers.

Little's Law (1961)

Little's Law is fundamental to the theory of queueing since it holds for *almost any queueing system* (or component of a system):

$$\boxed{E(N) = \lambda_a E(T)},$$

i.e., the expected number of customers in a system is the rate at which they arrive multiplied by the expected time spent in the system.

Example: people arrive to a museum at a rate of $\lambda_a = 30$ per hour and spend $E(T) = 1.5$ hours there on average $\Rightarrow$ the average number of people in the museum is $E(N) = \lambda_a E(T) = 30(1.5) = 45$.

Utilisation Factor

The **utilisation factor** or *traffic intensity* measures the load on the service component:

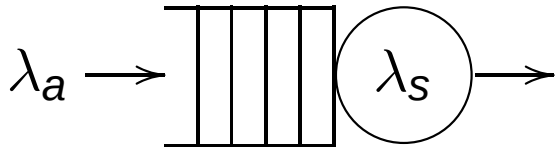
$$\boxed{\rho = \frac{\lambda_a}{\lambda_s}}.$$

(note: ρ is the Greek letter “rho”)

It is the proportion of time that the server is busy. It should be clear that λ_a must be smaller than λ_s or the system becomes overloaded.

Example: customers arrive at a rate of 1 per minute; the server can deal with 4 customers per minute. Thus, $\rho = 1/4 = 0.25$: the server is *busy* 25% of the time and *idle* 75% of the time.

$M/M/1$ System

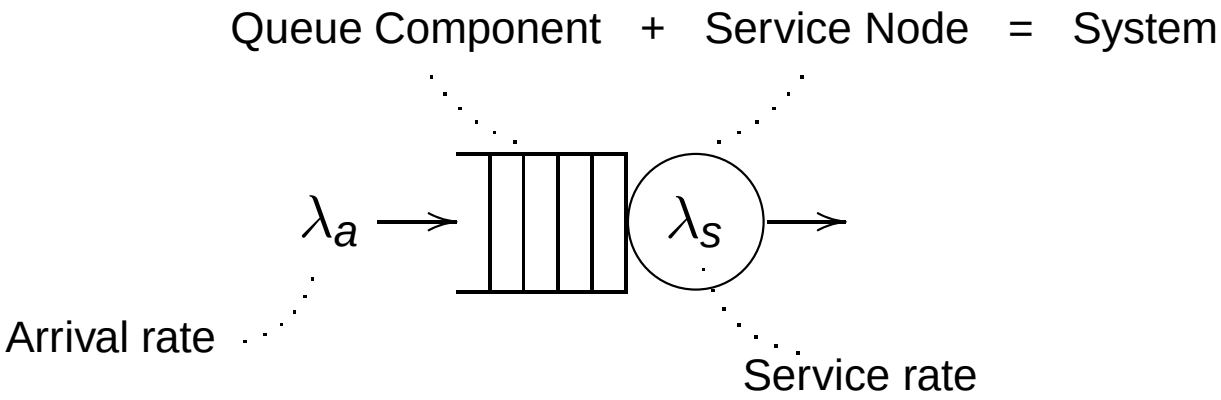


The $\boldsymbol{M/M/1}$ is the simplest queue structure. It is a *first in first out* queue which consists of:

- ***M*** The arrival process is *memoryless*: customer arrivals are *independent* according to a Poisson(λ_a) distribution.
- ***M*** The service process is *memoryless*: service times are *independent* according to an Exponential(λ_s) distribution.
- **1** There is one service node, i.e., only one customer can be served at any given time.

Of course we must have $\lambda_a < \lambda_s$.

$M/M/1$ System



- **Number of arrivals: $X_a \sim$ Poisson(λ_a)** $\Leftrightarrow T_a \sim$ Exponential(λ_a) (time between arrivals)
- **Service times: $T_s \sim$ Exponential(λ_s)** $\Leftrightarrow X_s \sim$ Poisson(λ_s) (service "potential" - *not important!*)
- Note that $\lambda_a < \lambda_s$

$M/M/1$ Results

Based on theory of *birth-death processes* (i.e., arrival-departure) one can derive that the time a customer spends in the system is:

$$\boxed{T \sim \text{Exponential}(\lambda_s - \lambda_a)}.$$

Thus, we know that $\Pr(T > t) = e^{-(\lambda_s - \lambda_a)t}$ and also

$$\boxed{E(T) = \frac{1}{\lambda_s - \lambda_a}}.$$

Little's Law gives us the expected number of customers in the system:

$$\boxed{E(N) = \lambda_a E(T) = \frac{\lambda_a}{\lambda_s - \lambda_a}}.$$

$M/M/1$ Results

We had that service time is $T_s \sim \text{Exponential}(\lambda_s) \Rightarrow E(T_s) = \frac{1}{\lambda_s}$.

Now let T_q represent the time spent in the *queue component only*.

It should be clear that $E(T) = E(T_q) + E(T_s)$. Thus, the expected time waiting in the queue is

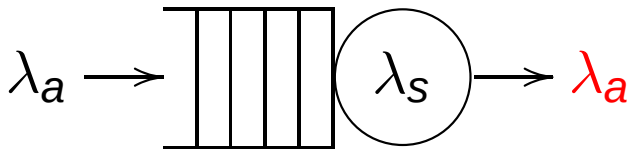
$$\boxed{E(T_q) = E(T) - E(T_s)},$$

and from Little's Law the expected number of customers in the queue is $E(N_q) = \lambda_a E(T_q)$.

(note: it turns out that, unlike T and T_s, the time waiting in the queue component, T_q, does *not* have an exponential distribution)

Burke's Theorem (1956)

Burke's Theorem states that for any $M/M/k$ system, the **number of departures** has the same **Poisson(λ_a)** distribution as the arrivals:



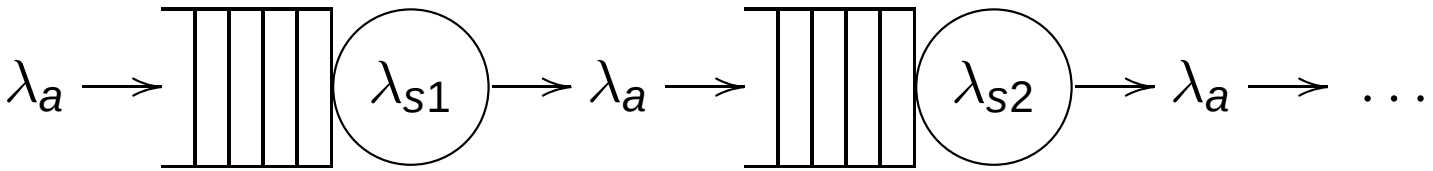
Note: departures are customers served or *jobs completed*.

The above result is intuitive since:

- If *less* customers depart the system than enter: the system is overloaded and the queue is growing infinitely.
- If *more* customers depart the system than enter: where did these extra customers come from??

Tandem Queues

A tandem queue is a finite chain of queues where each customer must visit each one in order:



Burke's theorem tells us the departure process from sub-system one is Poisson(λ_a). This is then the arrival process for sub-system two etc.

Sub-systems are considered *separately* $\Rightarrow E(T) = E(T_1) + E(T_2) + \ldots$ is the average time spent in the entire system. The average number of customers in the system is again found using Little's Law.

Example: Call Centre

On average a call centre worker receives 20 calls per hour and can deal with calls in 2 minutes.

First note that the time-frames are not compatible. We will work in terms of hours $\Rightarrow$ 2 minutes $= \frac{2}{60} = \frac{1}{30}$ hours.

The second thing to notice is that we have the arrival rate $\lambda_a = 20$ but not the service rate. What we have is the average service *time* $E(T_s) = \frac{1}{30} = \frac{1}{\lambda_s}$. Therefore $\lambda_s = \frac{1}{E(T_s)} = 30$ calls per hour.

We can work out everything we need based on the fact that total time in the system is $T \sim \text{Exponential}(\lambda_s - \lambda_a) = \text{Exponential}(10)$.

Example: Call Centre

- $E(T) = \frac{1}{10} = 0.1$ hours $= 6$ minutes is the time that a caller can expect to be in the system, i.e., queue + service.
- $E(N) = \lambda_a E(T) = 20(\frac{1}{10}) = 2$ callers in the system on average.
- $E(T_q) = E(T) - E(T_s) = \frac{1}{10} - \frac{1}{30} = \frac{1}{15} = 0.067$ hours $= 4$ minutes spent in the queue on average.
- $E(N_q) = \lambda_a E(T_q) = 20(\frac{1}{15}) = 1.33$ callers in the queue on average.
- We can also calculate the utilisation factor $\rho = \frac{\lambda_a}{\lambda_s} = \frac{20}{30} = 0.667$. Thus, the call centre worker is busy 66.7% of the time and idle 33.3% of the time.

Example: Call Centre

We can answer questions such as:

What is the probability that a caller spends more than 15 minutes (0.25 hours) in the system? Since $T \sim$ Exponential(10 callers/hour) this is

$$\Pr(T > 0.25) = e^{-10(0.25)} = e^{-2.5} = 0.0821.$$

What is the probability that 25 or more callers are dealt with (i.e., departures) in an hour? Burke's Theorem tells us that departures have the same distribution as arrivals: $X_d \sim$ Poisson(20).

$$\Pr(X_d \geq 25) = 0.1568. \qquad \text{(stats tables used)}$$

Example: Call Centre

What is the probability that the service time is less than 1 minute? (i.e., $\frac{1}{60}$ of an hour). We know that $T_s \sim \text{Exponential}(\lambda_s = 30)$

$$\begin{aligned}\Pr(T_s < \tfrac{1}{60}) &= 1 - \Pr(T_s > \tfrac{1}{60}) \\ &= 1 - e^{-30(1/60)} \\ &= 1 - e^{-2} \\ &= 0.8647.\end{aligned}$$

We can also answer questions about the number of arrivals in some time period since $X_a \sim \text{Poisson}(\lambda_a = 20)$ or time between arrivals since $T_a \sim \text{Exponential}(\lambda_a = 20)$.