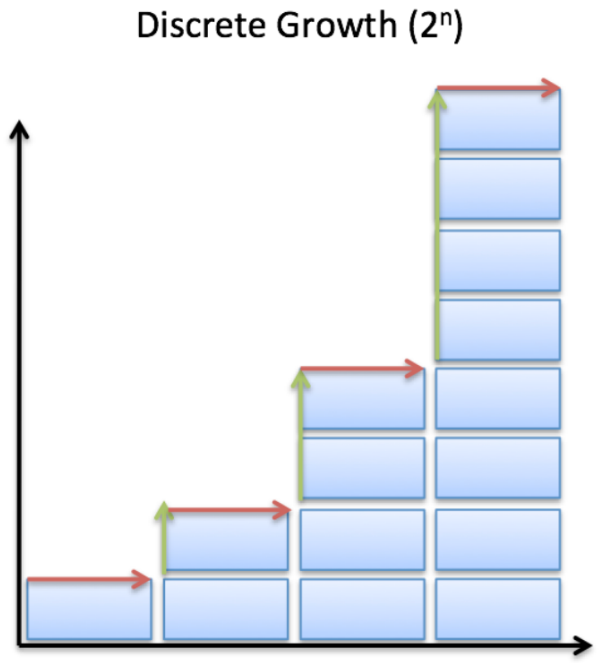
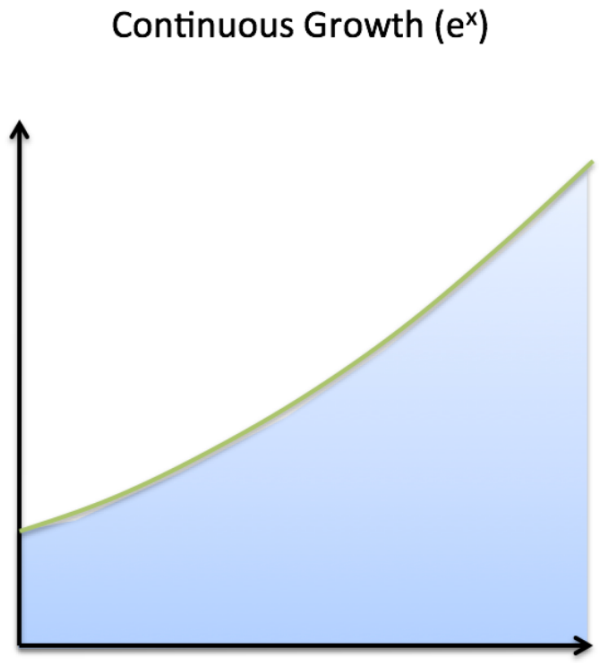
CSCI 3022: Intro to Data Science

Lecture 9: Continuous Variables and their Distributions

Rachel Cox

Department of Computer Science





Announcements & Reminders

- Quizlet 03 Due Wednesday at 11:59pm
- Midterm - next Tuesday

What will we learn today?

- ❑ Probability Density Function
- ❑ Continuous Density Function
- ❑ Continuous Uniform Distribution
- ❑ Normal (Gaussian) Distribution
- ❑ Exponential Distribution

- ❑ *A Modern Introduction to Probability and Statistics, Chapter 5*

Review from Last Time

A **discrete random variable** (r.v.) X is a function that maps the elements of the sample space Ω to a finite number of values $a_1, a_2, \ldots, a_n$ or an infinite number of values $a_1, a_2, \ldots$

A **probability mass function** (pmf) is the map between the random variable's values and the probabilities of those values.

$$f(a) = P(X = a)$$

A **cumulative distribution function** (cdf) is a function whose value at a point a is the cumulative sum of probability masses up until a.

$$F(a) = P(X \leq a) = \sum_{x \leq a} f(x)$$

Continuous Random Variables

Many real-life random processes must be modeled by random variables that can take on continuous (i.e. not discrete) values. Some examples:

People's heights: $X \in$ $(0, 15\text{ ft})$

Final grades in a class: $X \in$ $[0, 100]$

97.13875

Time between people checking out in a line at the store: $X \in$ $(0, \infty)$

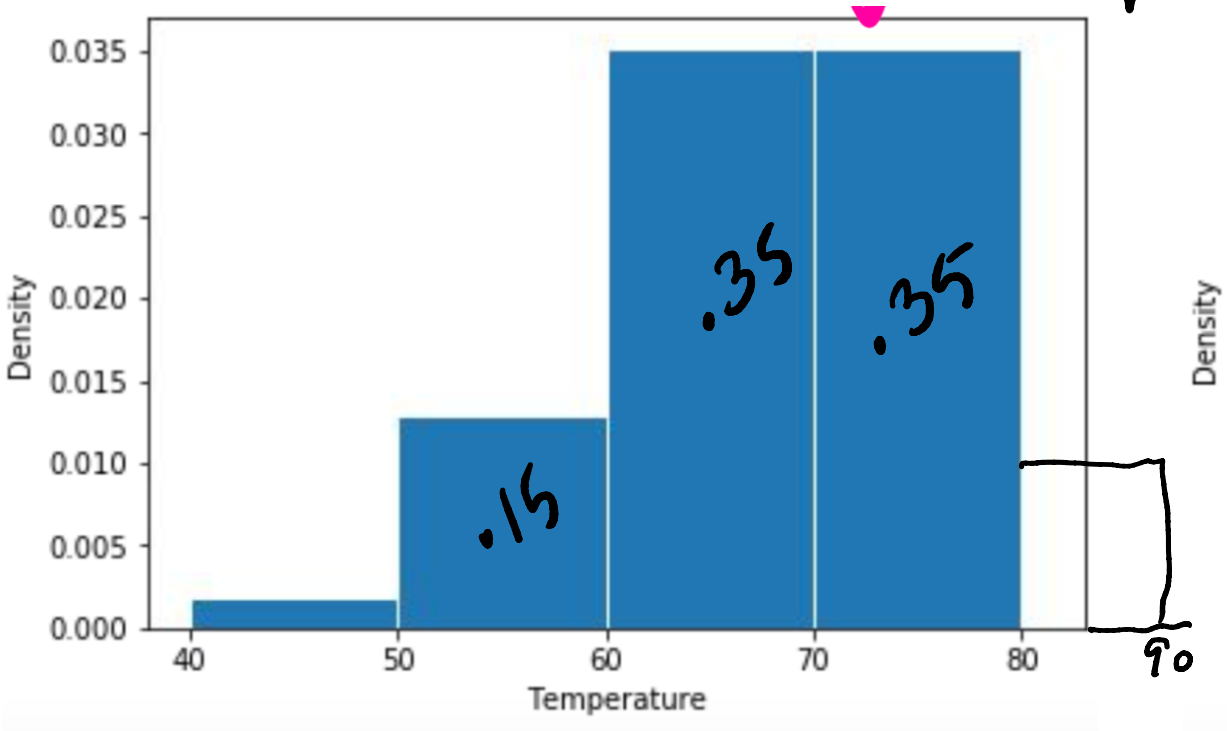
Can you think of other examples?

- Speed of a car during a race
- temperature

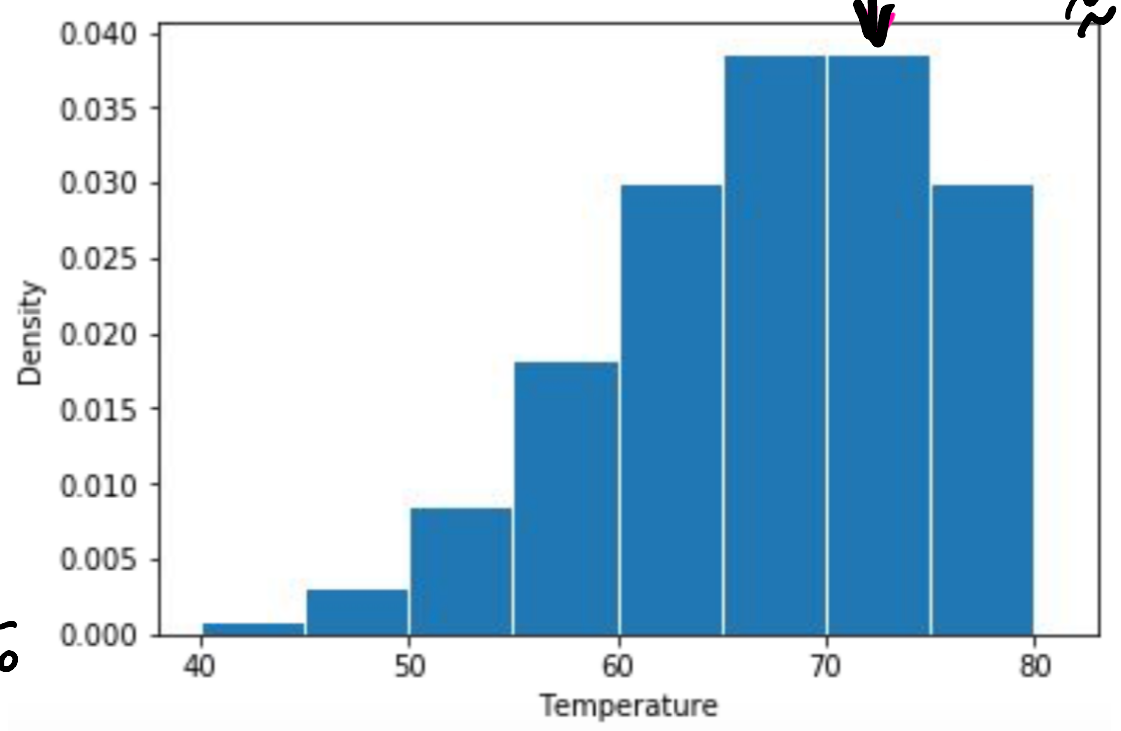
Continuous Random Variables

Example: Suppose your friend asks you what the temperature will be like today. Specifically, they want to know what the probability is that the temperature will be between 70 and 80°F. so that they can decide whether or not to wear shorts.

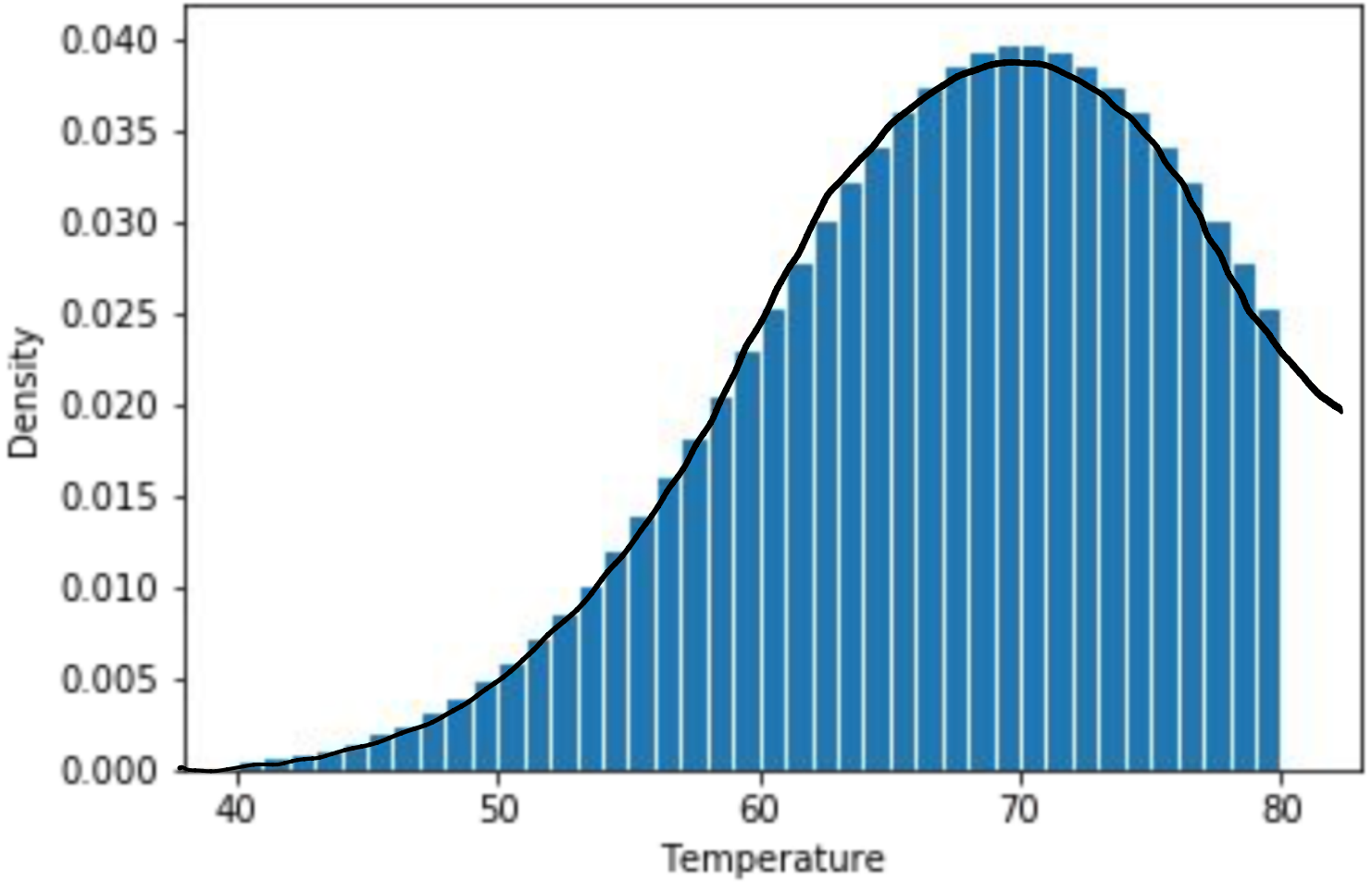
How would you calculate your response?



bad histograms



Continuous Random Variables



How would you calculate your response?

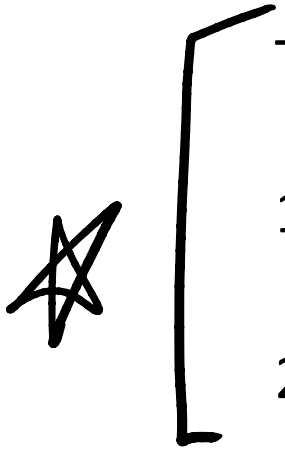
You could use a probability **density** function.

$$P(70 \leq X \leq 80) = \int_{70}^{80} f(x)\, dx$$

Continuous Random Variables

A random variable X is **continuous** if for some function $f: \mathbb{R} \to \mathbb{R}$ and for any numbers a and b with $a \leq b$,

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



The function f must satisfy:

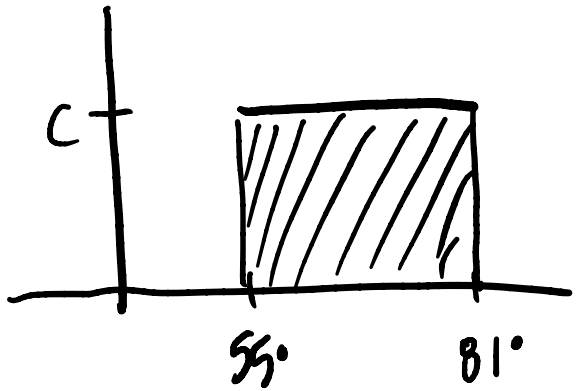
1) $f(x) \geq 0$ for all x

2) $\int_{-\infty}^{\infty} f(x)\, dx = 1$

❖ We call f the probability density function (pdf) of X.

Continuous Random Variables

Example: Suppose you have some reason to believe that the temperature is equally likely to be anywhere between 55° and 81°F. Find the probability density function (pdf) for X.



Need $f(x) \geq 0$ for all x in the support $[55, 81]$

$$f(x) = \begin{cases} c & 55 \leq x \leq 81 \\ 0 & x < 55 \text{ or } x > 81 \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{26} & 55 \leq x \leq 81 \\ 0 & \text{otherwise} \end{cases}$$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

Need this =

$$= \int_{-\infty}^{55} f(x)\,dx + \int_{55}^{81} f(x)\,dx + \int_{81}^{\infty} f(x)\,dx = 1$$

($f(x) = 0$ on $(-\infty, 55)$ and $(81, \infty)$)

$$= \int_{55}^{81} c\,dx = cx\Big|_{55}^{81} = c(81-55) = 1$$

$c = \frac{1}{26}$

Uniform Distribution

A continuous random variable has a uniform distribution on the interval $[\alpha,\ \beta\ ]$ if its probability density function f is given by $f(x) = 0$ if x is not in $[\alpha,\ \beta]$ and

$$f(x) = \frac{1}{\beta\ - \alpha} \quad \text{for } \alpha \le x \le \beta$$

We say $X \sim U(\ \alpha, \beta\ )$

What distribution does temperature follow in the example?

$T \sim U(55, 81)$

What is the probability that temperature is between 75° and 81° F?

$$P(75 \le x \le 81) = \int_{75}^{81} \frac{1}{26}\, dx = \frac{1}{26} x \Big|_{75}^{81}$$

$$= \frac{1}{26}(81 - 75)$$

$$= \frac{6}{26} = \boxed{\frac{3}{13}}$$

What is the probability that the temperature is exactly 75°F?

$$P(x = 75) = \int_{75}^{75} \frac{1}{26}\, dx$$

Cumulative Distribution Function

What if we want to compute things like $P(X \leq a)$?

Is there an analog for the cumulative distribution function from the discrete case?

$$F(x) = P(X \leq x) = \sum_{k \leq x} f(k)$$

continuous analog: $F(x) = \int_{-\infty}^{x} f(x)\, dx$ *

pdf

The cdf is the integral of the pdf.

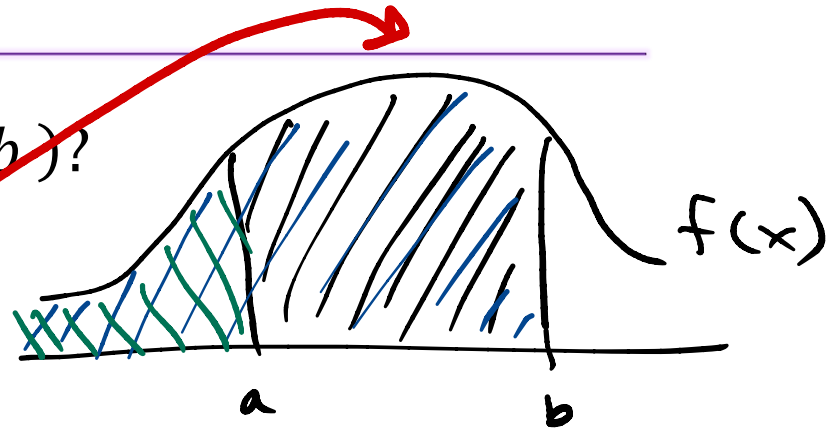
Cumulative Distribution Function

Can we use the cdf to compute things like $P(a \leq X \leq b)$?

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

$$= F(b) - F(a)$$

$$= \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)\,dx$$



Fundamental Theorem of Calculus!

Cumulative Distribution Function

Example: What is we viewed the cdf as a function of x?

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

$$\frac{d}{dx} F(x) = f(x)$$

$$\frac{d}{dx} F(x) = \frac{d}{dx}\left(\int_{-\infty}^{x} f(t)\,dt\right)$$

$$= \frac{d}{dx}\left(\lim_{a \to -\infty} \int_{a}^{x} f(t)\,dt\right)$$

$$= \lim_{a \to -\infty} \frac{d}{dx} \int_{a}^{x} f(t)\,dt$$

$$= f(x)$$

} Because analysis

Normal Distribution

A continuous random variable X has a **normal (or Gaussian) distribution** with parameters μ and σ^2 if its probability density function is given by

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

We say $X \sim N(\mu, \sigma^2)$

μ - mean

$x - \mu$ horizontal shift

σ - standard deviation

function height

Let's play around with this distribution: https://academo.org/demos/gaussian-distribution/

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

$$f(x) \geq 0$$

Exponential Distribution

Sometimes it's easier to first find the cdf and then derive the pdf by taking a derivative.

Example: Recall that the Poisson distribution describes the number of arrivals (or hits), assuming some constant average rate of arrivals per time period.

A Poisson random variable is discrete because we're counting things (e.g. number of cars entering a parking garage). But suppose each "arrival" is someone finishing checking out in a grocery store line.

Exponential Distribution

- Discrete Poisson – how many "events"
- Continuous Exponential – how long between "events"

➢ Now we are interested in the amount of time between each arrival.

- Suppose the # of arrivals follows a Poisson distribution (process) with rate λ arrivals/minute.

- Start from $t = 0$ and let T be the random variable describing the first arrival.

- What is the probability that the first arrival does not occur in the first t minutes?

Exponential Distribution

- Suppose the # of arrivals follows a Poisson distribution (process) with rate λ arrivals/minute.

$$\left(\lambda \frac{\text{arrivals}}{\text{minute}}\right)(t \text{ minutes}) = \mu$$

Recall Poisson pmf:

$$P(X=k) = \frac{\mu^k e^{-\mu}}{k!}$$

- Start from $t = 0$ and let T be the random variable describing the first arrival.

- What is the probability that the first arrival does not occur in the first t minutes?

$$P(T > t) = P(\text{no arrivals in the first } t \text{ minutes})$$

This can be described with the Poisson distribution.

$$= \frac{\mu^0 e^{-\mu}}{0!}$$

$$= e^{-\mu}$$

$$= e^{-\lambda t}$$

$$P(T > t) = 1 - P(T \leq t)$$

Exponential Distribution

- Suppose the # of arrivals follows a Poisson distribution (process) with rate λ arrivals/minute.
- Start from $t = 0$ and let T be the random variable describing the first arrival.
- What is the probability that the first arrival does not occur in the first t minutes?

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

$$P(T > t) = 1 - P(T \leq t)$$

$$P(T \leq t) = 1 - P(T > t)$$

$$P(T \leq t) = 1 - e^{-\lambda t}$$

} cdf for exponential distribution

pdf $f(t) = \frac{d}{dt}(1 - e^{-\lambda t})$

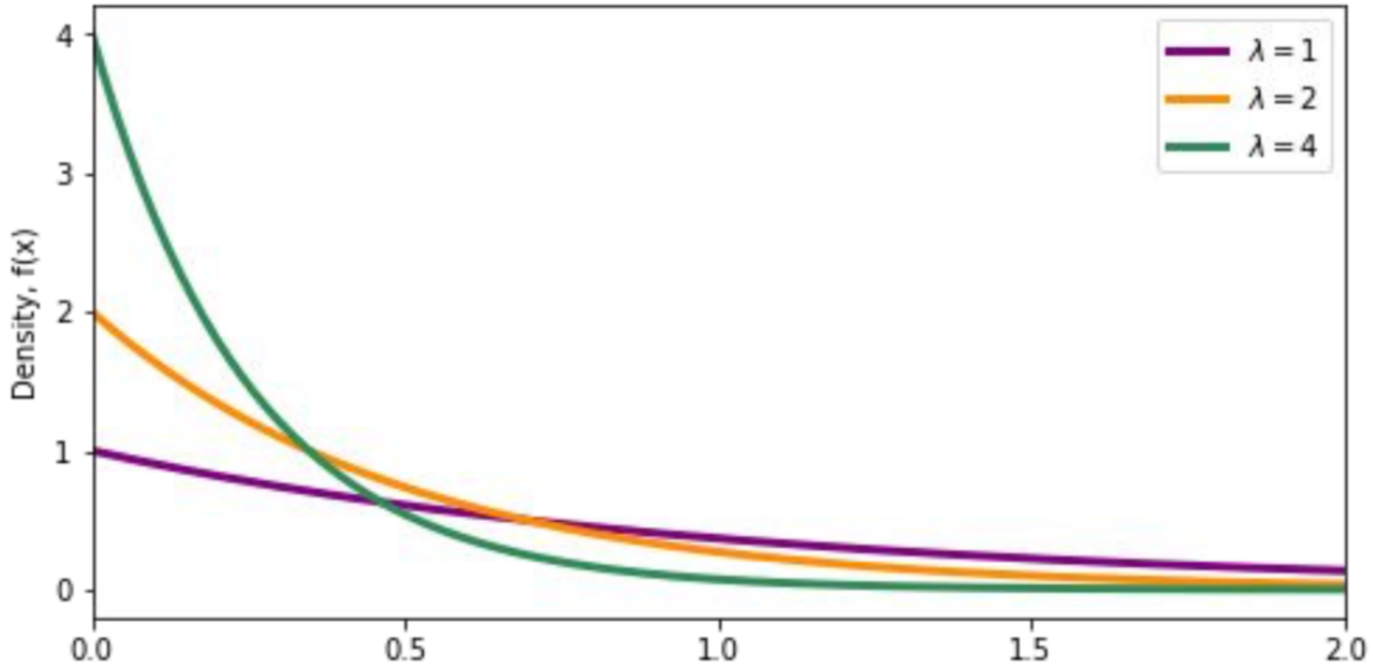
$= -(-\lambda)e^{-\lambda t} = \lambda e^{-\lambda t}$

Exponential Distribution

A continuous random variable X has an **exponential distribution** with rate parameter $\lambda > 0$ if its probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

We say $X \sim Exp(\lambda)$



- That works for the first arrival. But what about the second? Or third?
- We want the distribution of time between all arrivals.

Exponential Distribution

Theorem: (memoryless property)

If $T \sim Exp(\lambda)$, then $P(T > t + t_0 \mid T > t_0) = P(T > t)$

Next Time:

- Expectation