CS 237: Probability in Computing

Wayne Snyder
Computer Science Department
Boston University

Lecture 11:

- Poisson Process
- Poisson Distribution
- Continuous Distributions
- Exponential Distribution

Poisson Process

An important kind of Random Variable involves counting how many events or objects occur in time or space….

X = "How many people arrive at Starbucks every minute?"

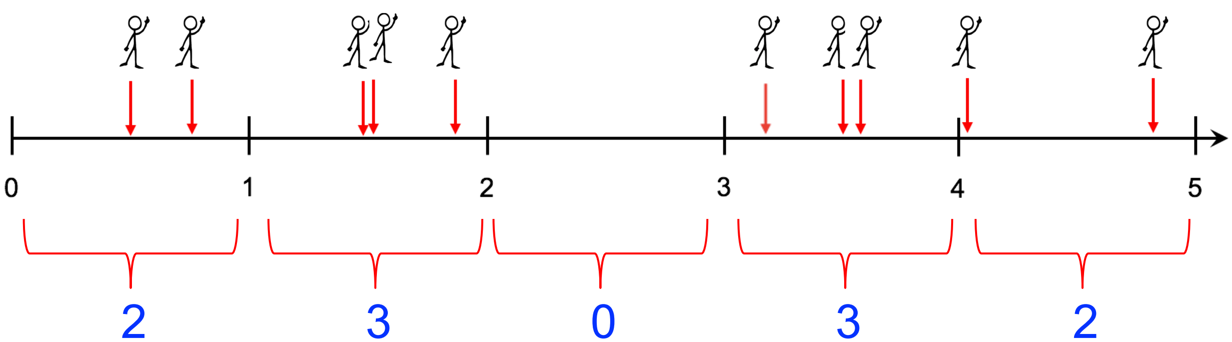
Y = "How many typos per page are in this book?"

Z = "How many raisins occur in a loaf of cinnamon-raisin bread?"

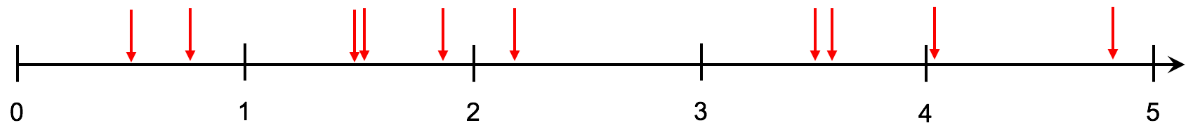
A Counting Random Variable is Discrete ($R_X = [0,1,2,\ldots]$)
but has a continuous component (when or where occurs) that affects the outcomes:

Timeline
per minute:

X produces:

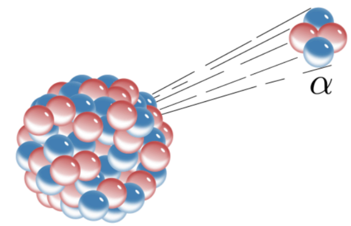


Poisson Process



Examples in the time domain:

- Sneezes in a classroom
- Alpha particles emitted from U 238
- Email arriving in my inbox
- Accidents at an intersection
- Earthquakes, volcanoes, asteroids, …



What if an asteroid hit the Earth?

BY MARSHALL BRAIN

UP NEXT »

An illustration of an asteroid on its way to Earth. See more space dust images.
PHOTOGRAPHER: ANDREUS AGENCY: DREAMSTIME.COM

An asteroid striking our planet -- it's the stuff of science fiction. Many movies and books have portrayed this possibility ("Deep Impact," "Armageddon," "Lucifer's Hammer," and so on).

An asteroid impact is also the stuff of science fact. There are obvious craters on Earth (and the moon) that show us a long history of large objects hitting the planet. The most famous asteroid ever is the one that hit Earth 65 million years ago. It's thought that this asteroid threw so much moisture and dust in to the atmosphere that it cut off sunlight, lowering temperatures worldwide and causing the extinction of the dinosaurs.

Yellowstone volcano eruption: NASA to SAVE the world from supervolcano erupting

NASA scientists are creating an ambitious plan to prevent an explosion of a Yellowstone volcano that could even end human life by drilling a hole.

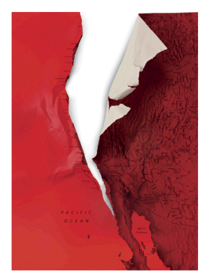
THE REALLY BIG ONE

An earthquake will destroy a sizable portion of the coastal Northwest. The question is when.

By Kathryn Schulz

When the 2011 earthquake and tsunami struck Tohoku, Japan, Chris Goldfinger was two hundred miles away, in the city of Kashiwa, at an international meeting on seismology. As the shaking started, everyone in the room began to laugh. Earthquakes are common in Japan—that one was the third of the week—and the participants were, after all, at a seismology conference. Then everyone in the room checked the time.

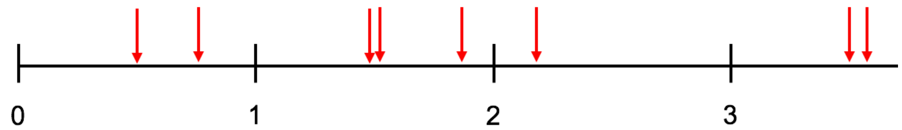
Seismologists know that how long an earthquake lasts is a decent proxy for its



The next full-margin rupture of the Cascadia subduction zone will spell the worst natural disaster in the history of the continent.

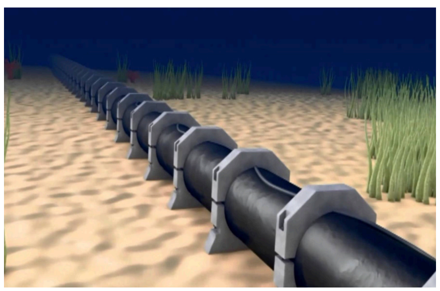
Every year The Federal Highway Administration reports approximately 2.5 Million intersection accidents. Most of these crashes involve left turns.

Poisson Process

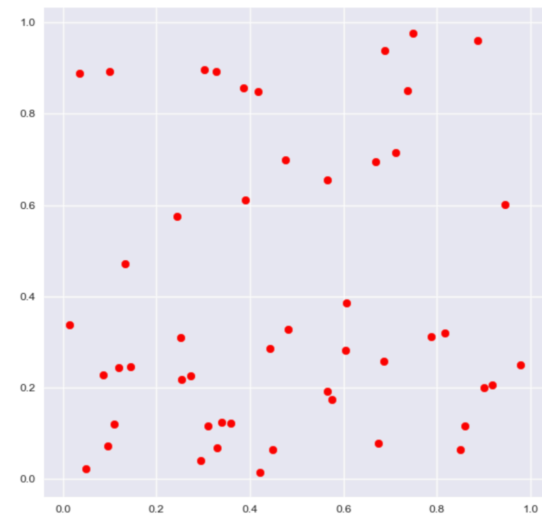


It is also possible that the continuous dimension is distance in space, in 1 dimension or more than 1. Examples include:

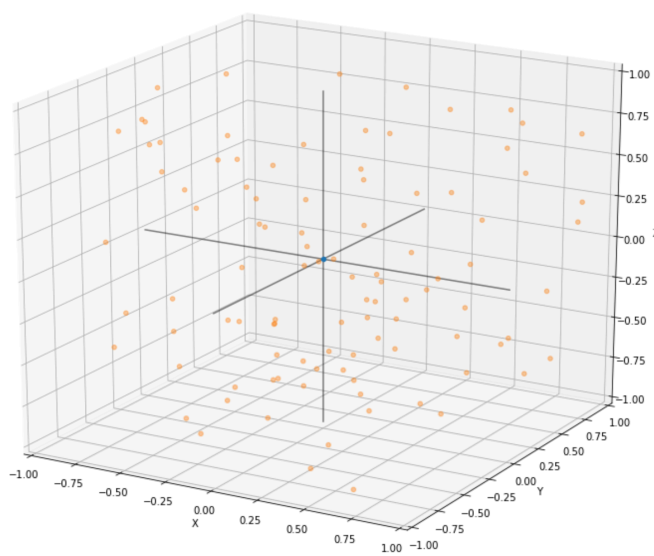
The occurrence of leaks in an undersea pipeline (1D):



Location of trees in a 1 square mile plot of land:



Location of supernovas in a given cubic gigaparsec volume of space in the last billion years:



The important point is that events (discrete) occur along 1 or more (continuous) dimensions.

Poisson Process

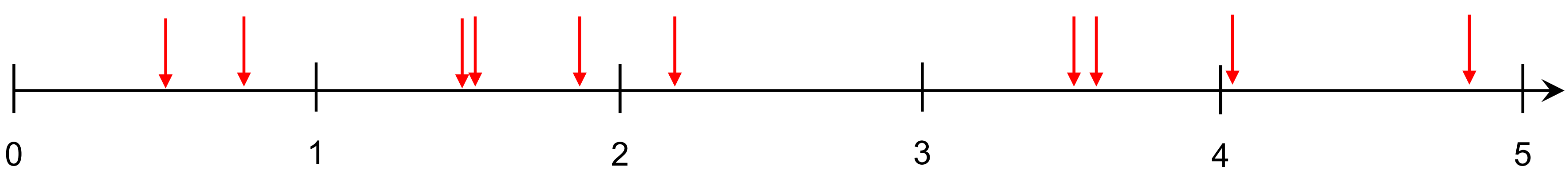
The Poisson Process concept captures an important way of thinking about events randomly occurring through time (or space)...

Two things to remember are

- **Events are discrete** (they happen or they don't – you can think of it as a Bernoulli trial with an outcome of success or failure), but
- **Time and space are continuous**..... the random behavior here is the time of an event.

When an event has happened we say it has arrived. You can think of this as a sequence of real numbers giving the **arrival time** of each successive event:

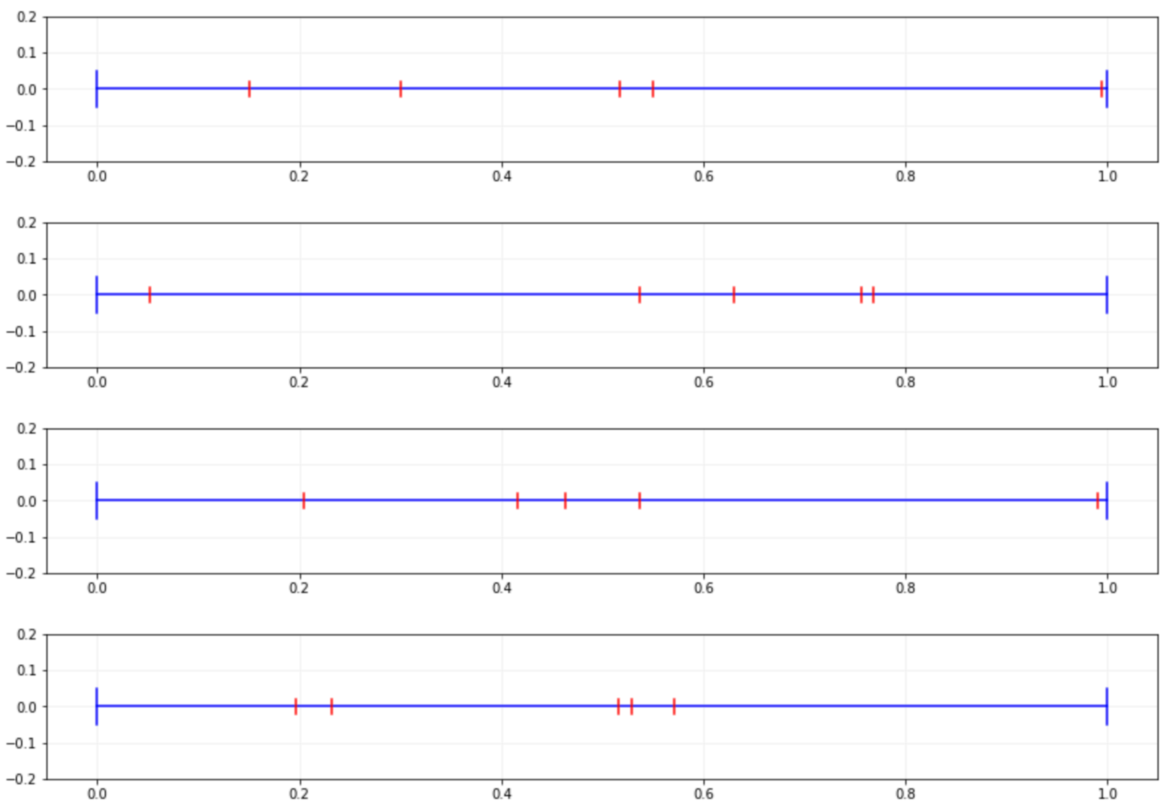
Arrival Times = { 0.4324... , 0.734 , 1.389... , 1.453... , 2.1546... , ... }



Time t in seconds/days/years/whatever

Poisson Process

We can motivate the way a Poisson process is formally defined by considering what happens when we randomly generate arrivals in a unit interval. Suppose each trial of the experiment we generate 5 random numbers in the interval [0..1):



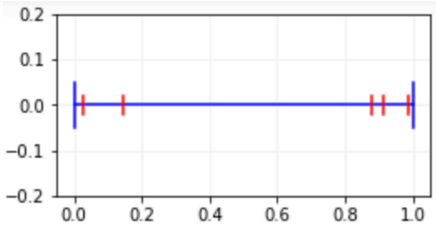
We know that the probability that a particular arrival occurs in the interval [0.0 .. 0.1) is 1/10; for [0.2 .. 0.5) is 0.3, and for any interval [a..b) it is (b-a).

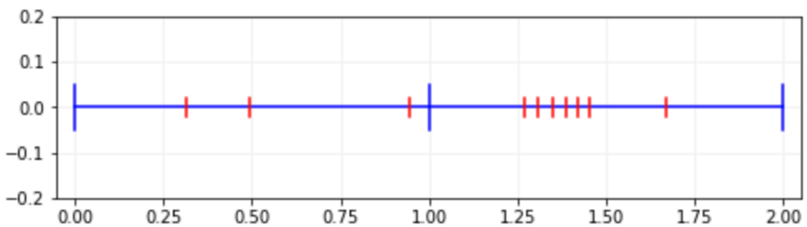
The probability for any one arrival is equal to the length of the interval.

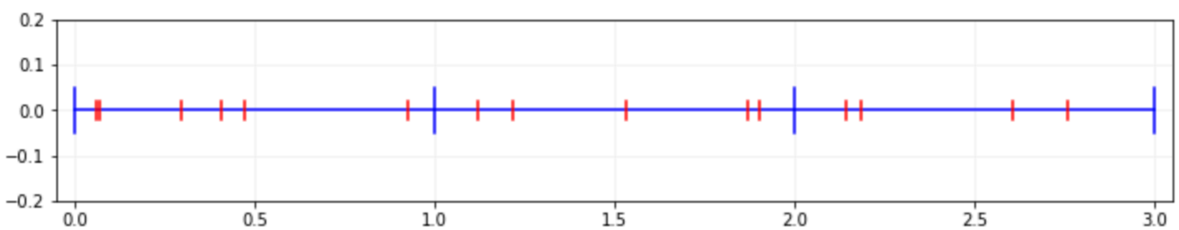
This is because the arrivals are randomly and uniformly distributed in the interval [0..1).

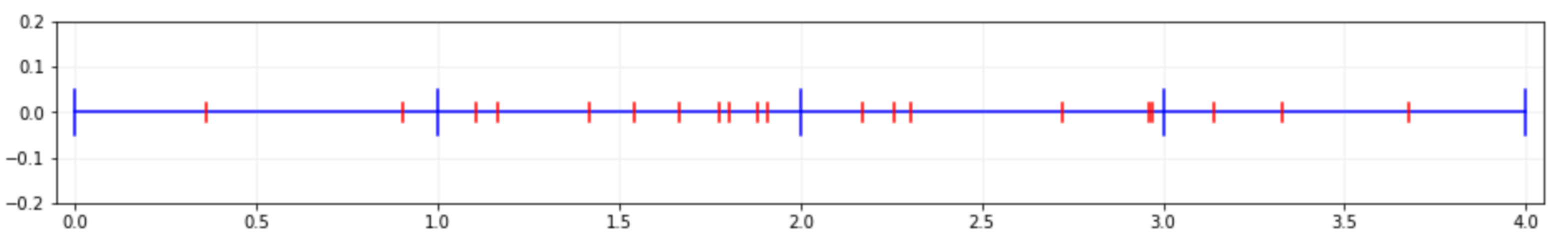
Poisson Process

Now suppose we generate 5 random arrivals in [0 .. 1), 10 random arrivals in [0 .. 2), 15 arrivals in [0 .. 3), 20 in [0 .. 4) and so on, to infinity....









Since the arrivals are independent and distributed uniformly, the mean number of arrivals in each unit interval [0 .. 1), [1 .. 2), [2 .. 3), etc. is still 5.

Also, as the sequence gets longer, the relationship between each interval becomes less and less dependent... in the limit, each intervals' results are independent of every other.

Poisson Process

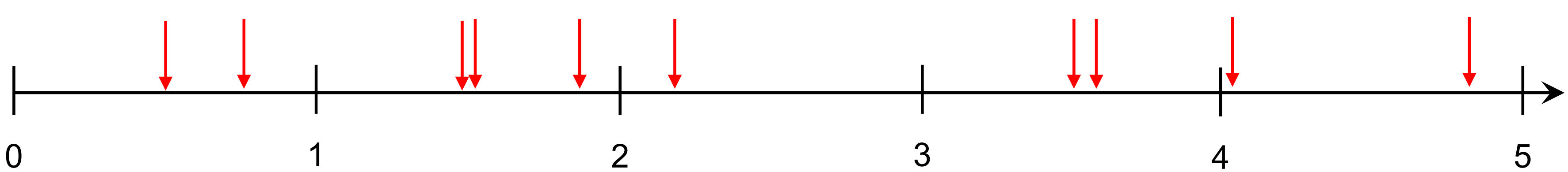
Formally, we have the following definition: suppose we have discrete events occurring through time as just described, and let

$$N[s..t] \;=\; \text{the number of events arriving in the time interval } [s..t]$$

such that

1) The expected value of N[s..t]
   a) Is a fixed constant λ over any unit interval anywhere in the sequence, and
   b) Is proportional to the length $(t - s)$ of the interval; in particular, for any two non-overlapping intervals of the same length, the expected number of occurrences in each is the same;
2) The number of arrivals in two non-overlapping intervals is independent; and
3) The probability of two events occurring at the same time is 0.

Then this random process is said to be a Poisson Process.



Poisson Random Variables

Suppose we have a Poisson Process and we fix the unit time interval we consider (say, 1 second or 1 year, etc.), where the mean number of arrivals in a unit interval is λ, and then each time we "poke" the random variable X we return N[0..1), N[1..2), N[2..3), etc.

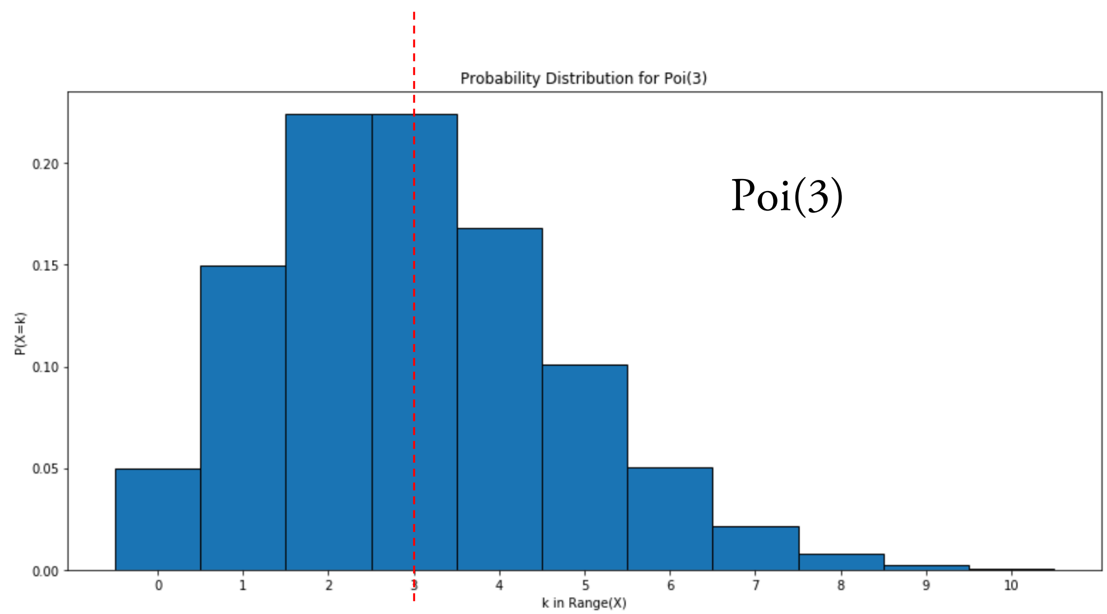
Then we call X a Poisson Random Variable with rate parameter λ, denoted

$$X \sim Poi(\lambda)$$

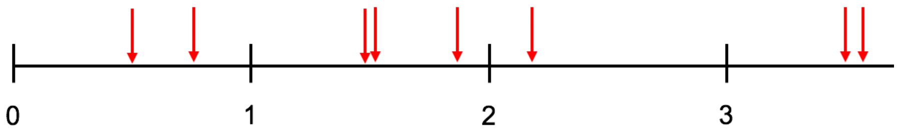
where

$$R_X = \{0, 1, 2, 3, \dots\}$$

$$f_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$



Poisson Random Variables

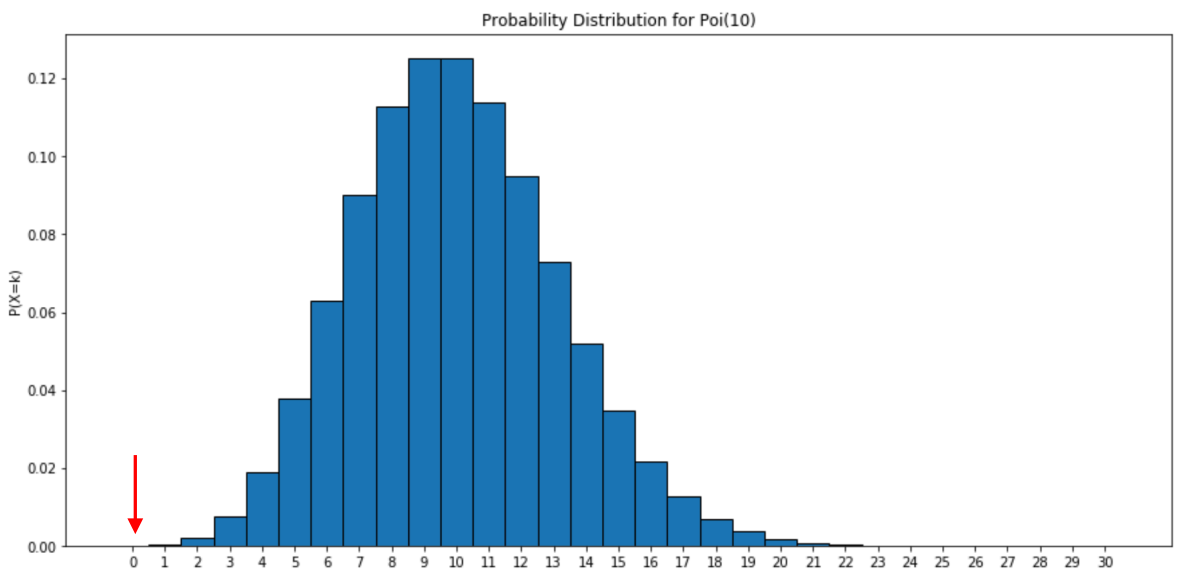


Examples

Assume that arrivals of email in my Inbox are a Poisson Process with rate λ = 10 messages per hour. Then X ~ Poi(10) returns the random number of emails which arrive within any particular hour.

What is the probability that I get no emails in the next hour?

$$P(X = 0) = f_X(0) = \frac{e^{-10}\lambda^0}{0!} = e^{-10} = 4.54 * 10^{-5}$$

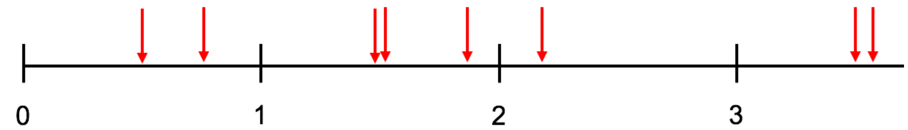


$$X \sim Poi(10)$$

$$R_X = \{0, 1, 2, 3, \dots\}$$

$$P(X = k) = f_X(k) = \frac{e^{-10}10^k}{k!}$$

Poisson Random Variables

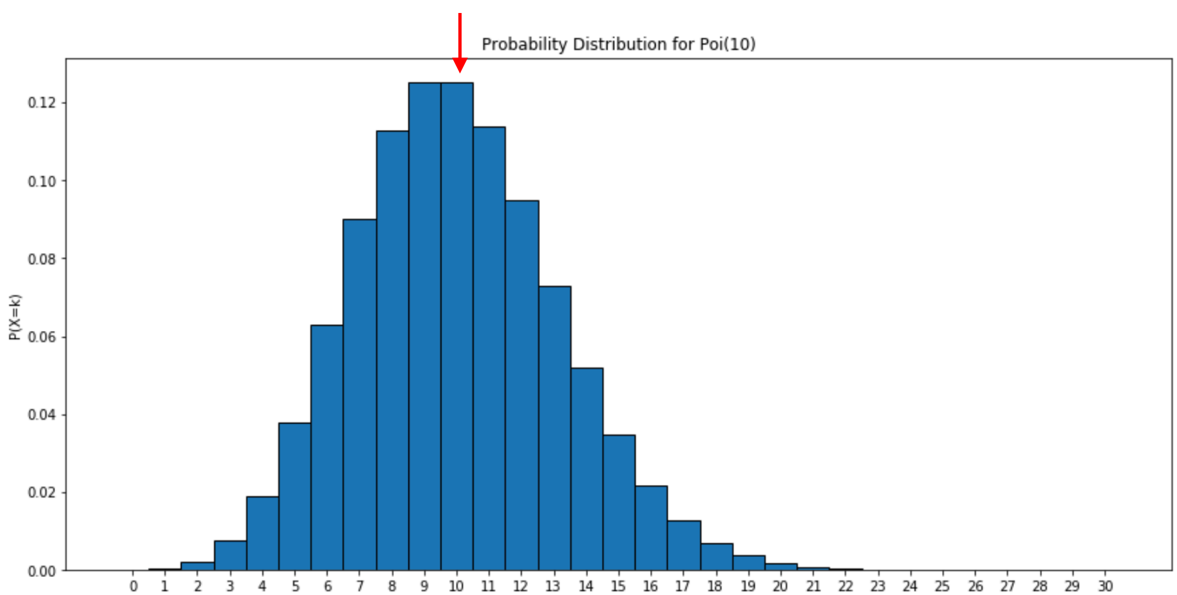


Examples

Assume that arrivals of email in my Inbox are a Poisson Process with rate λ = 10 messages per hour. Then X ~ Poi(10) returns the random number of emails which arrive within any particular hour.

What is the probability that I get exactly 10 emails in the next hour?

$$P(X = 10) = f_X(10) = \frac{e^{-10}\lambda^{10}}{10!} = 0.1251$$

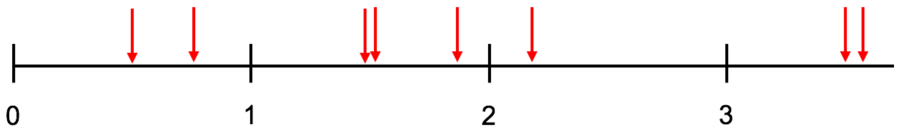


$$X \sim Poi(10)$$

$$R_X = \{0, 1, 2, 3, \ldots\}$$

$$P(X = k) = f_X(k) = \frac{e^{-10}10^k}{k!}$$

Poisson Random Variables

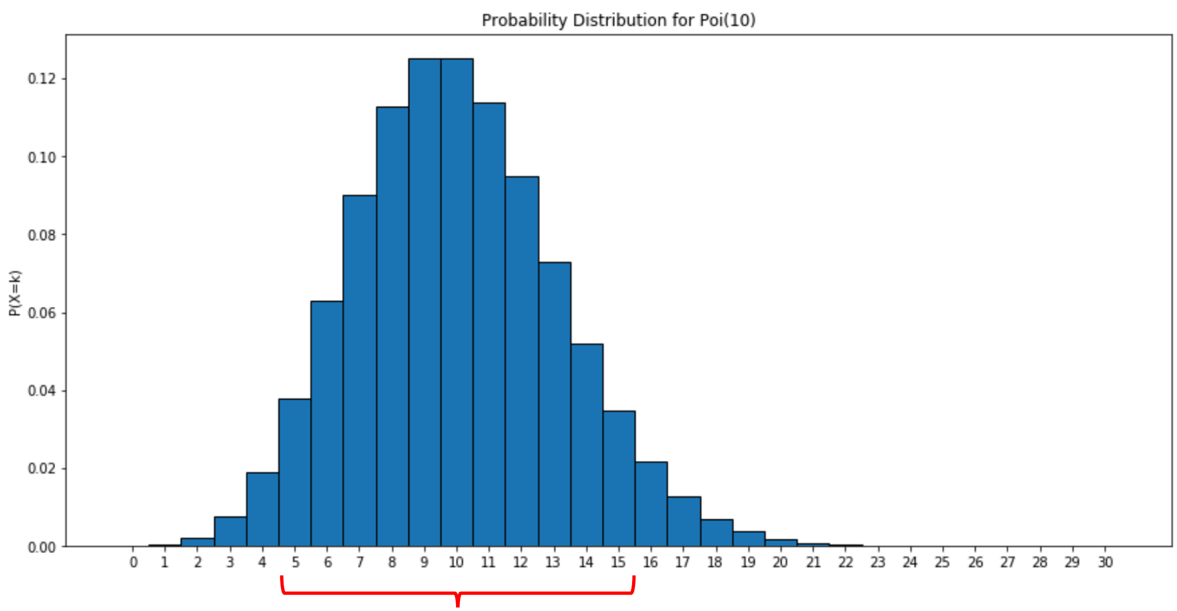


Examples

Unfortunately there is no way to compute the CDF or ranges except by simply adding together all the individual values.

What is the probability that I get between 5 and 15 emails (inclusive) emails in the next hour?

$$P(5 \le X \le 15) = \sum_{k=5}^{15} \frac{e^{-10}10^k}{k!} = 0.922$$



$P(5 \le X \le 15)$

$$X \sim Poi(10)$$

$$R_X = \{0, 1, 2, 3, \dots\}$$

$$P(X = k) = f_X(k) = \frac{e^{-10}10^k}{k!}$$

Recall: Poisson Random Variables

Suppose we have a Poisson Process and we fix the unit time interval we consider (say, 1 second or 1 year, etc.), where the mean number of arrivals in a unit interval is λ, and then each time we "poke" the random variable X we return N[0..1], N[1..2], N[2..3], etc.

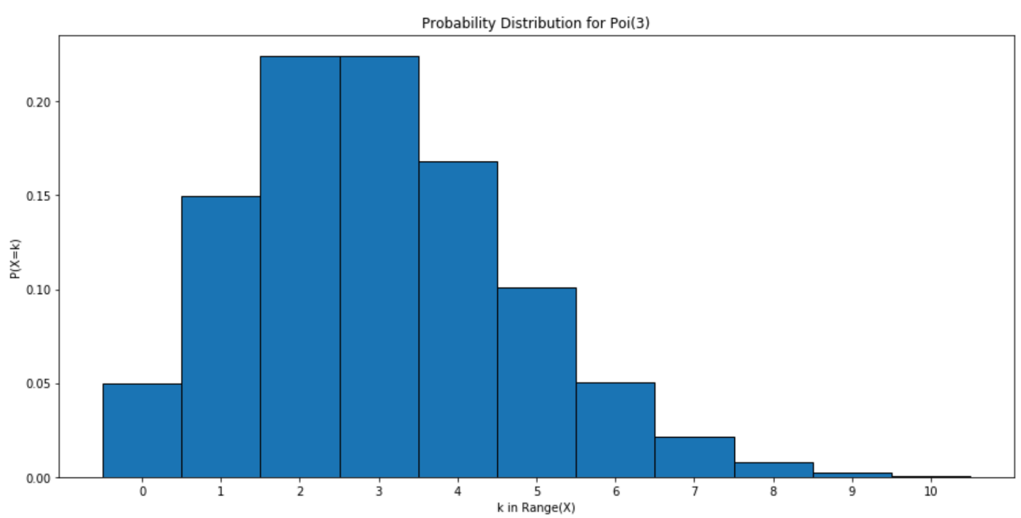
Then we call X a Poisson Random Variable with rate parameter λ, denoted

$$X \sim Poi(\lambda)$$

where

$$R_X = \{0, 1, 2, 3, \dots\}$$

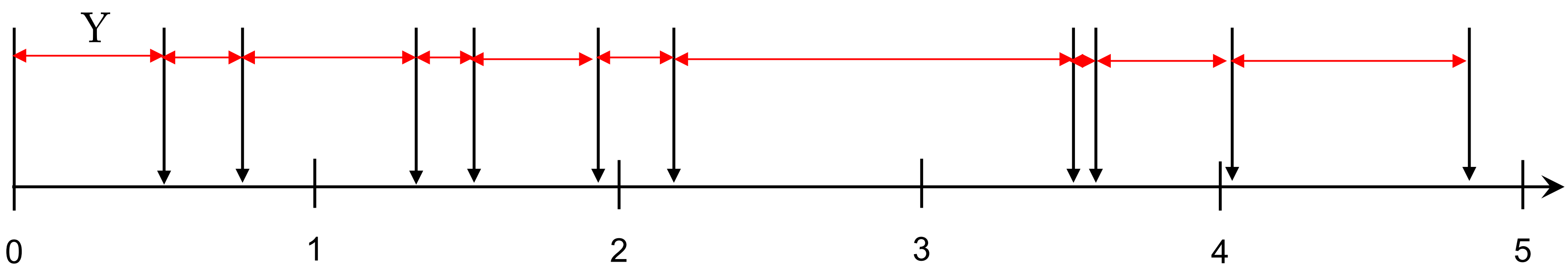
$$f_X(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$



Continuous Distributions: Interarrival Times of a Poisson Process

Suppose we have a Poisson Process, and instead of counting the number of arrivals in each unit interval, we look at the interarrival times, i.e., the amount of time between each arrival.

Intuitively, this is a natural thing to think about: **How long before the next event?**

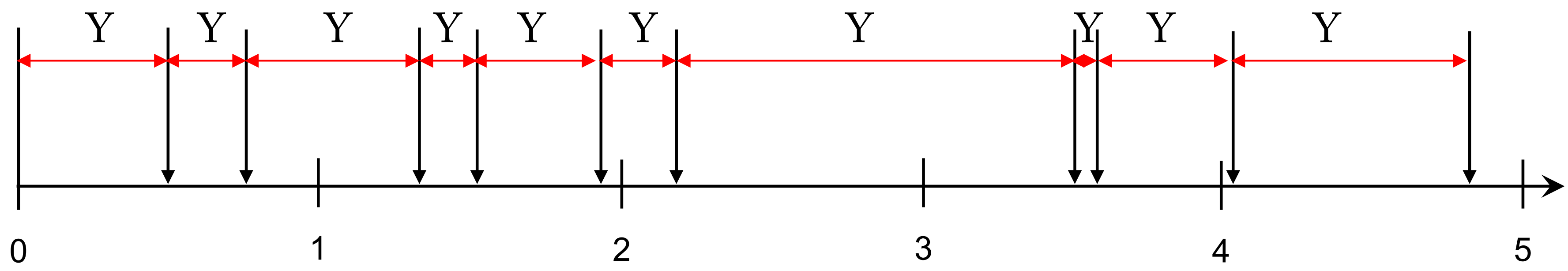


Let's define the random variable Y = "the arrival time of the first event."

Interarrival Times of a Poisson Process

Suppose we have a Poisson Process, and instead of counting the number of arrivals in each unit interval, we look at the interarrival times, i.e., the amount of time between each arrival.

Intuitively, this is a natural thing to think about: **How long before the next event?**



Let's define the random variable Y = "the arrival time of the next event."

In fact, because the arrivals are independent, at any time t, probabilistically the Poisson process starts all over again (the events don't remember the past!), so in fact:

Y = "the interarrival time between any two events"

Now the question is: **What is the distribution of Y? Clearly it is NOT discrete……**

Discrete vs Continuous Distributions

Recall: A Random Variable X is a function from a sample space S into the reals:

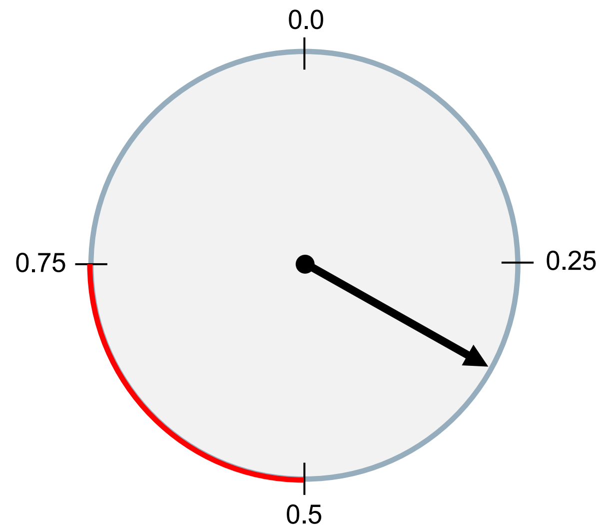
$$X : S \to \mathcal{R}$$

A random variable is called continuous if Rx is uncountable.

What needs to change when working with continuous as opposed to discrete distributions?

Recall: The probability of a random experiment such as a spinner outputting any particular, exact real number is 0:

$$f_X(a) \;=\; P(X = a) \;=\; 0$$



This result extends to any countable collection of real numbers!

So we can **only** think about intervals:

$$P(0.5 < X < 0.75) \;=\; 0.25$$

Discrete vs Continuous Distributions

Because of the anomolies having to do with continuous probability, we need to keep the following important points in mind:

(A) The probability function f_X does NOT represent the probability of a point in the domain, since as just quoted:

$$f_X(a) = P(X = a) = 0$$

therefore we can ONLY work with intervals $P(X \leq a)$, $P(X > a)$, $P(a \leq X \leq b)$, etc. and f_X is not as important as the CDF F_X .

(B) In calculating F_X and working with intervals, we can not use discrete sums $\sum_{x=a}^{b}$ as we did in the discrete case, but will have to use integrals: $\int_a^b$

(C) The range R_X will be all the reals $(-\infty..\infty)$ and so we don't specify it each time.

Discrete vs Continuous Distributions

Discrete Random Variables

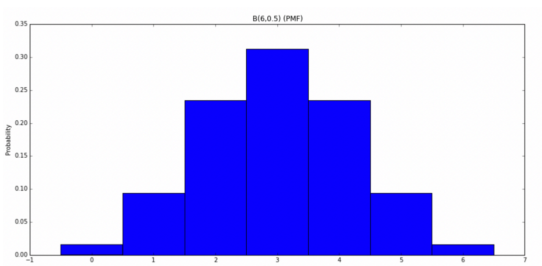
The Probability Density Function of a discrete random variable X is a function from the range of X into $\mathcal{R}$:

$$f : R_x \rightarrow \mathcal{R}$$

such that

(i) $\forall a \in R_x \quad f_X(a) \geq 0$

(ii) $\sum_{x \in R_X} f(x) = 1.0$



Continuous Random Variables

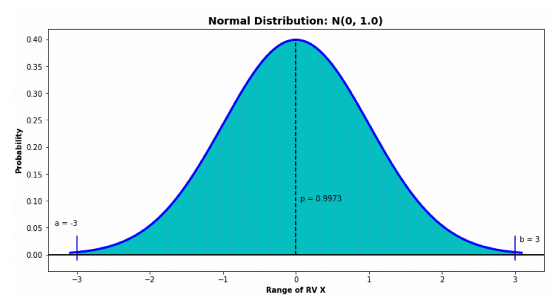
The Probability Density Function of a continuous random variable X is a function from the $\mathcal{R}$ to $\mathcal{R}$:

$$f : \mathcal{R} \rightarrow \mathcal{R}$$

such that

(i) $\forall a.\ f(a) \geq 0$

(ii) $\int_{-\infty}^{\infty} f(x)\, dx = 1.0$

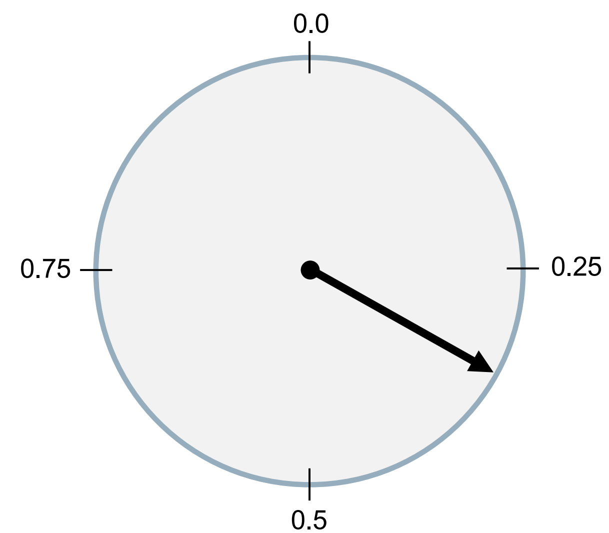


Continuous Distributions

Let's clarify these ideas with an example....

Consider the spinner example from way back when:

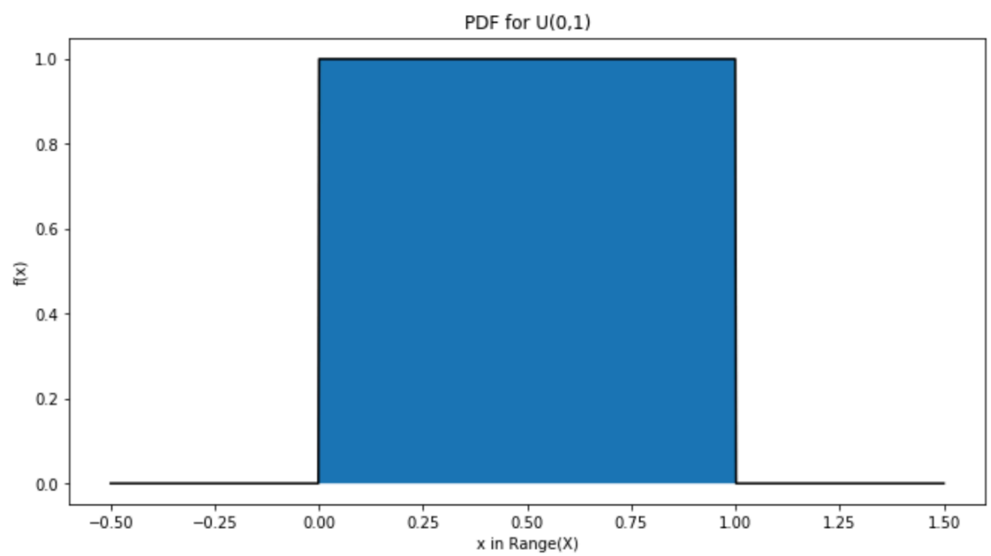
X = "the real number in [0..1) that the spinner lands on"



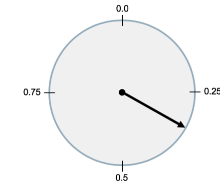
The probability density function is:

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that the area is 1.0 and for any $0 \leq a \leq 1$, we have $f(a) = 1.0$, so it is uniform across [0..1). But clearly $P(X = a) = 0.0$.



Continuous Distributions



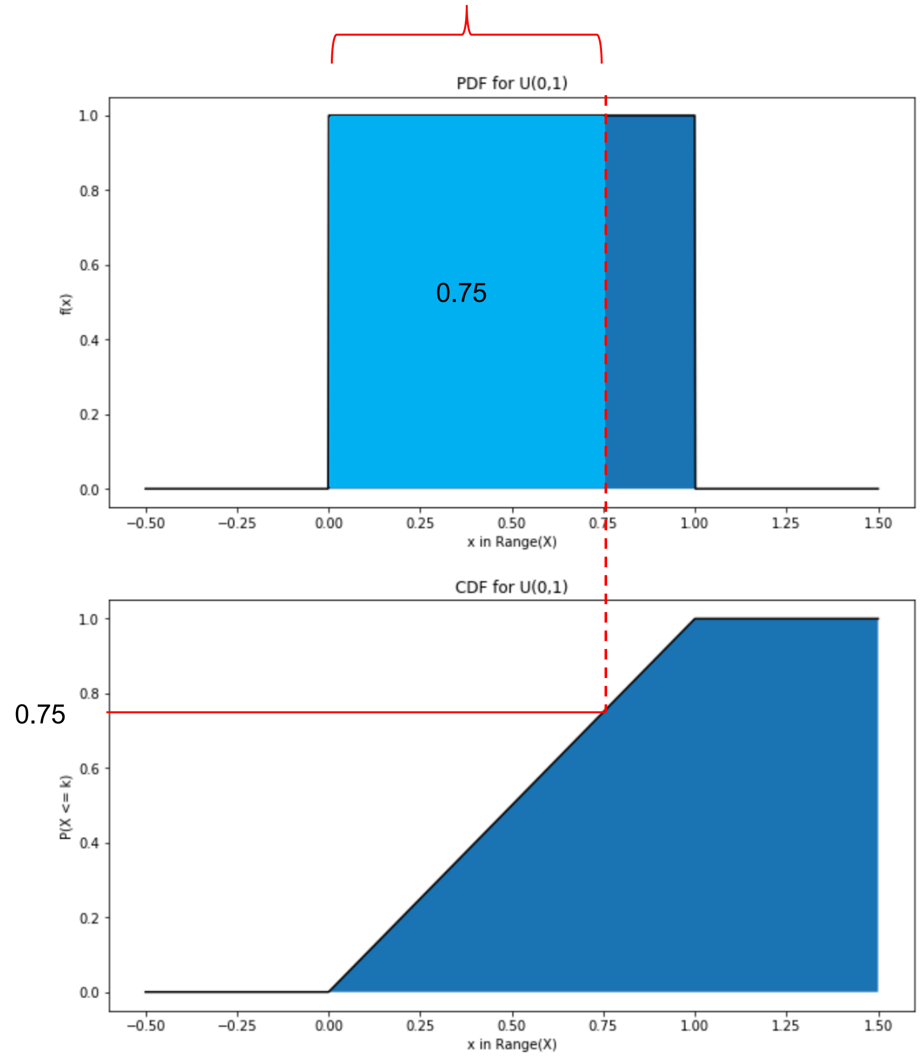
Now recall that the ONLY way to deal with continuous probability is to use intervals and to use area (or extent) for the probability. Hence we will calculate probabilities of intervals using the CDF:

$$P(X < 0.75) = F(0.75) = 0.75$$

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$F(a) = \int_0^a 1\,dx = x\Big|_0^a = a$$

$$F(a) = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } a > 1 \end{cases}$$



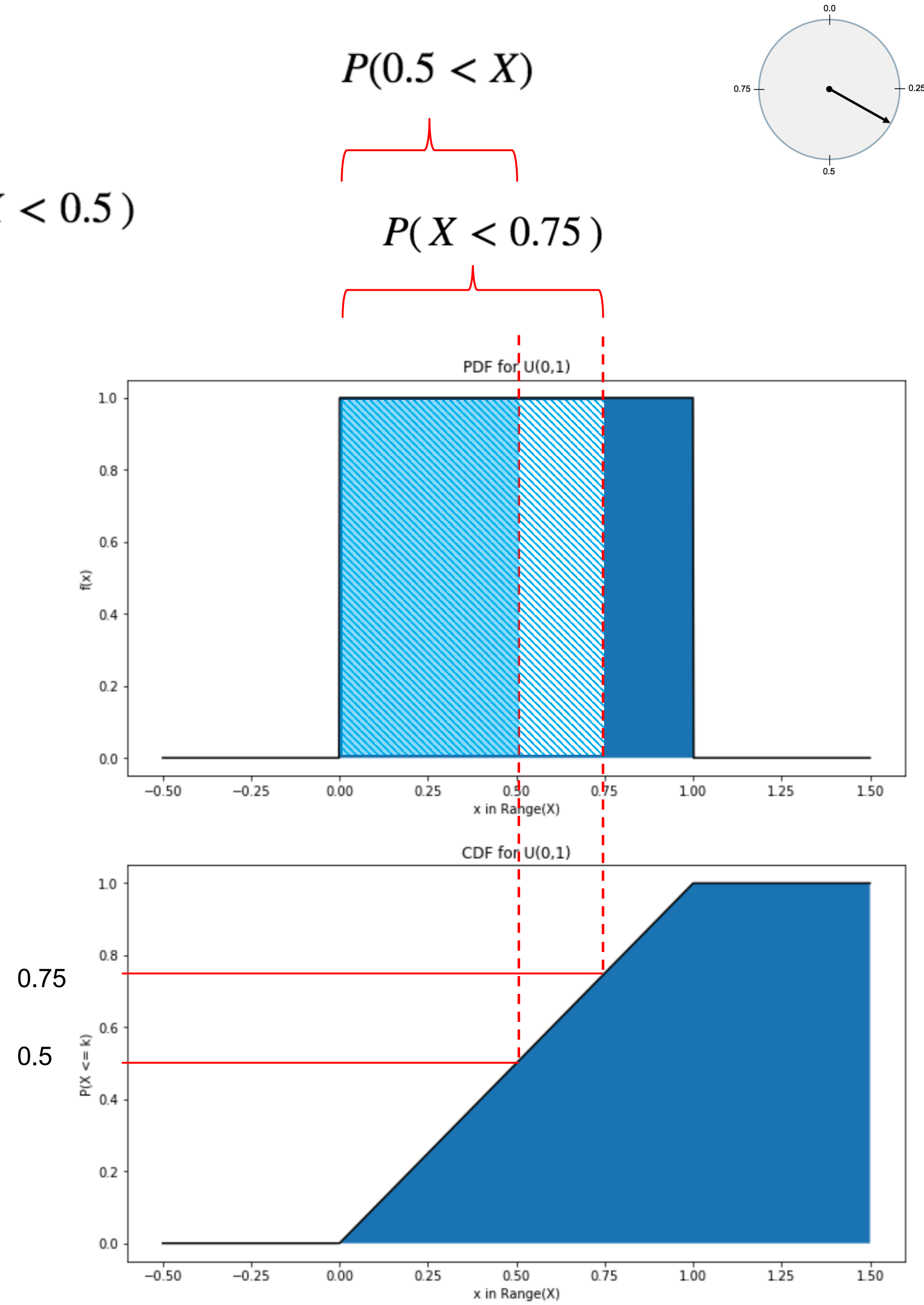
Continuous Distributions

$$P(0.5 < X < 0.75) = P(X < 0.75) - P(X < 0.5)$$
$$= F(0.75) - F(0.5)$$
$$= 0.75 - 0.5$$
$$= 0.25$$

$$f(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

$$F(a) = \int_0^a 1\, dx = x \Big|_0^a = a$$

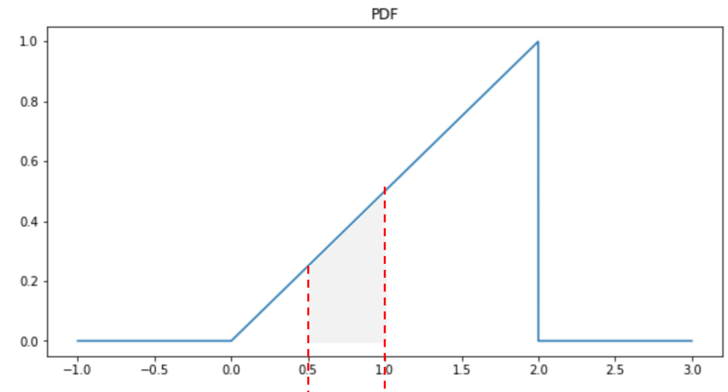
$$F(a) = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } 0 \leq a \leq 1 \\ 1 & \text{if } a > 1 \end{cases}$$



Continuous Distributions

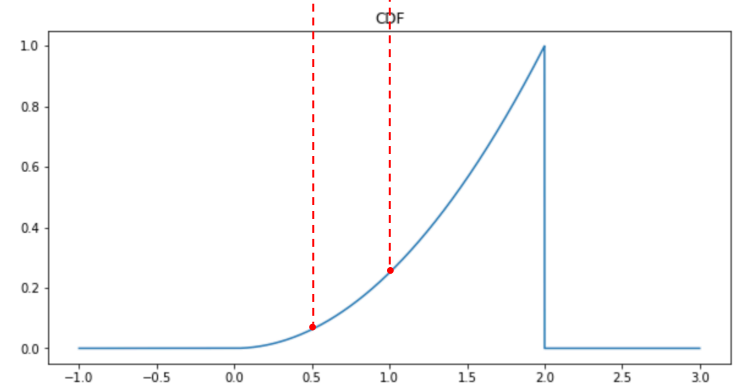
Bottom Line: In order to deal with continuous distributions, you have to do integrals....

Example: Suppose our PDF looked like this:



$$f(x) = \begin{cases} \dfrac{x}{2} & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

To calculate the probability of intervals, we need to determine the CDF, which means doing the following integral:



$$F(a) = \int_{-\infty}^{a} f(x)\, dx = \int_{0}^{a} \frac{x}{2}\, dx = \frac{x^2}{4}\Big|_0^a = \frac{a^2}{4}$$

So for example,

$$P(0.5 < X < 1.0) = \frac{1^2}{4} - \frac{0.5^2}{4} = \frac{4-1}{16} = \frac{3}{16} = 0.1875$$

Continuous Distributions

Discrete Random Variables

$$F_X(b) \;=\; P(\,X \le b\,) \;=_{\text{def}} \sum_{x \le b} f(x)$$

$$P(\,a \le X \le b\,) \;=_{\text{def}} \sum_{a \le x \le b} f(x)$$

$$E(\,X\,) \;=\; \sum_{x \in R_X} x \cdot f(x)$$

Continuous Random Variables

$$F_X(b) \;=\; P(\,X < b\,) \;=_{\text{def}} \int_{-\infty}^{b} f(x)\,dx$$

$$P(\,a < X < b\,) \;=_{\text{def}} \int_{a}^{b} f(x)\,dx$$

$$E(\,X\,) \;=\; \int_{-\infty}^{\infty} x \cdot f(x)\;dx$$

Same for both Discrete and Continuous Random Variables

$$Var(\,X\,) \;=_{def}\; E[\,(\,X - \mu_X\,)^2\,]$$

$$Var(\,X\,) \;=\; E(\,X^2\,) - (\,\mu_X\,)^2$$

$$\sigma_X \;=_{def}\; \sqrt{Var(\,X\,)}$$

$$X^* \;=_{def}\; \frac{X - \mu_X}{\sigma_X}$$