CS 112: Modeling Uncertainty in Information Systems

Prof. Jenn Wortman Vaughan

May 7, 2012

Lecture 10

Reminders & Announcements

- Homework 3 will be posted by the end of the day and is due on Friday, May 18
- This homework assignment will include both a written component and a programming component
- You can start the written component, but some of the problems require material from this Wednesday’s lecture

Today

Continuous random variables

- Probability density functions (the analog to PMFs)
- Cumulative distribution functions

Common continuous random variables

- Uniform continuous random variables
- Exponential random variables
- Normal random variables

Continuous Random Variables

What if a random variable X can take on a continuum of different values?

- Exact execution time of a task
- Exact lifetime of a component
- Time between two requests to the Google server

Continuous Random Variables

What if a random variable X can take on a continuum of different values?

- Exact execution time of a task
- Exact lifetime of a component
- Time between two requests to the Google server

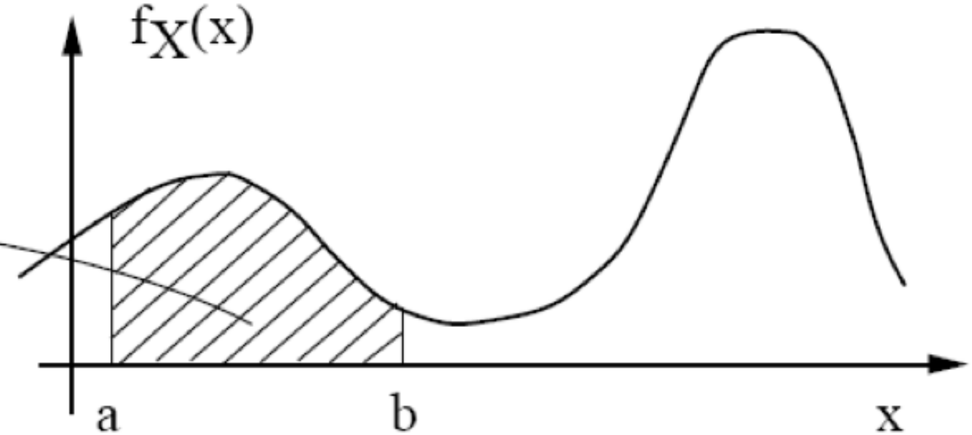
PMFs don't quite make sense anymore…

- Let X be a random variable whose value is drawn uniformly at random from [0,1]. What is $P(X = 0.5)$?

The Probability Density Function

In place of the PMF, we introduce the probability density function (or PDF), denoted f_X. For any a and any $b \geq a$,

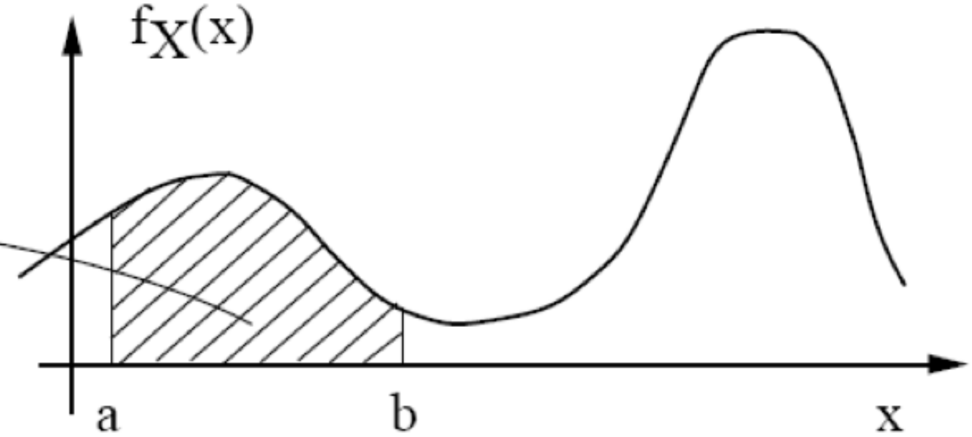
$$P(a \leq X \leq b) = \int_a^b f_X(k)\,dk$$



The Probability Density Function

In place of the PMF, we introduce the probability density function (or PDF), denoted f_X. For any a and any $b \geq a$,

$$P(a \leq X \leq b) = \int_a^b f_X(k)\,dk$$



What do we know about $\int_{-\infty}^{\infty} f_X(k)\,dk$?

The Probability Density Function

For very small values of δ, we can approximate

$$P(a \le X \le a+\delta) = \int_a^{a+\delta} f_X(k)dk \approx f_X(a)\delta$$

This gives us a nice interpretation of $f_X(a)$ as the probability mass per unit length near a

The Probability Density Function

For very small values of δ, we can approximate

$$P(a \le X \le a+\delta) = \int_a^{a+\delta} f_X(k)\,dk \approx f_X(a)\delta$$

This gives us a nice interpretation of $f_X(a)$ as the probability mass per unit length near a

Need to be careful with this interpretation though!!! Note that, for example, $f_X(a)$ can be bigger than 1…

Expectations

Expectation and variance of continuous random variables are similar to the discrete case.

Expectations

Expectation and variance of continuous random variables are similar to the discrete case.

$$E[X] = \int_{-\infty}^{\infty} k f_X(k) dk$$

Expectations

Expectation and variance of continuous random variables are similar to the discrete case.

$$E[X] = \int_{-\infty}^{\infty} k f_X(k) dk$$

$$E[g(X)] = ???$$

Expectations

Expectation and variance of continuous random variables are similar to the discrete case.

$$E[X] = \int_{-\infty}^{\infty} k f_X(k)\,dk$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(k) f_X(k)\,dk$$

Expectations

Expectation and variance of continuous random variables are similar to the discrete case.

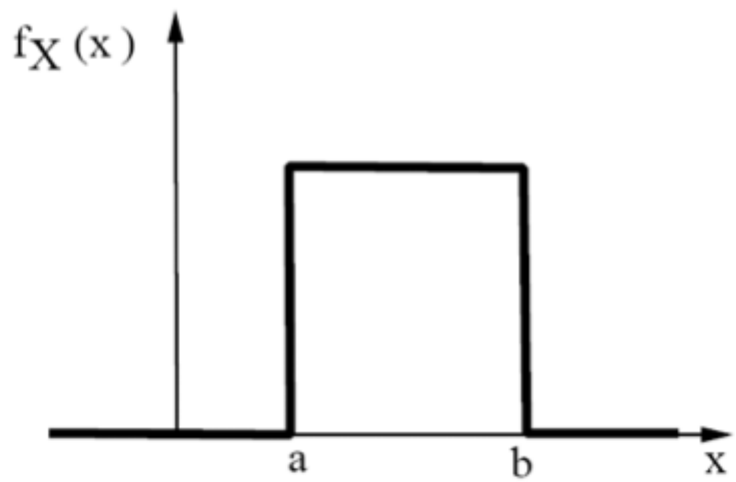
$$E[X] = \int_{-\infty}^{\infty} k f_X(k) dk$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(k) f_X(k) dk$$

$$\mathrm{var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$$

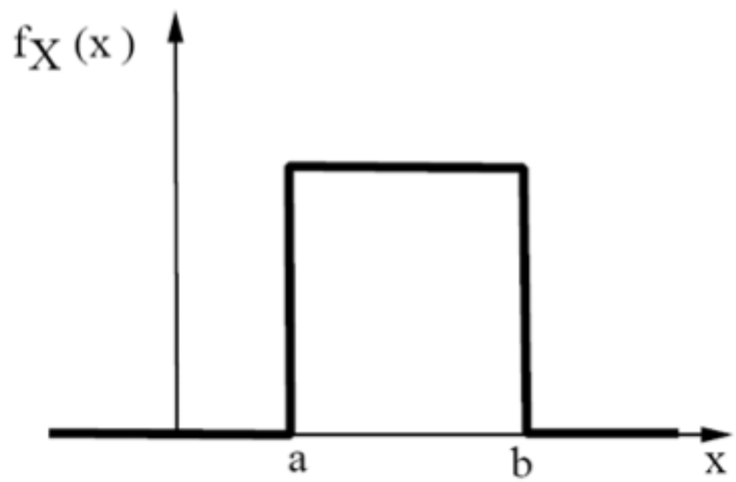
Uniform PDFs

Let X be a random variable whose value is drawn uniformly at random from the range [a, b]



Uniform PDFs

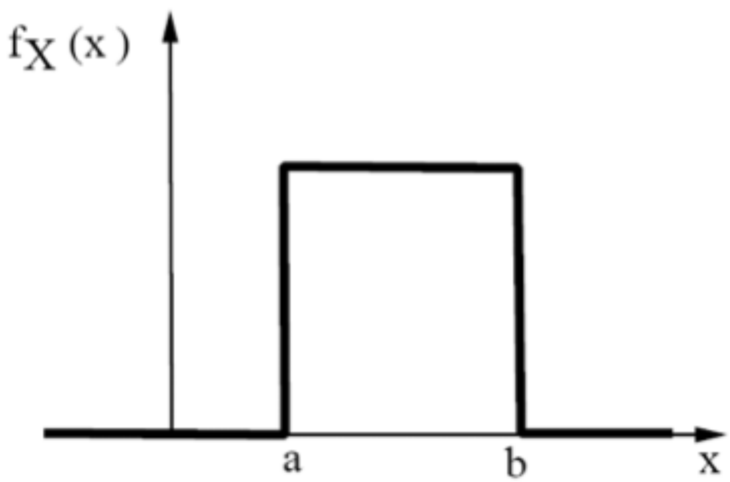
Let X be a random variable whose value is drawn uniformly at random from the range $[a, b]$



- What is f_X? What is E[X]? What is var(X)?

Uniform PDFs

Let X be a random variable whose value is drawn uniformly at random from the range $[a, b]$



- What is f_X? What is E[X]? What is var(X)?
- What if X was chosen from the range (a, b) instead?

Exponential Random Variables

Exponential random variables model the amount of time until an incident of interest takes place

- Length of time before a message arrives at the computer
- Length of time before a light bulb burns out

Exponential Random Variables

Exponential random variables model the amount of time until an incident of interest takes place

- Length of time before a message arrives at the computer
- Length of time before a light bulb burns out

PDF: $f_X(k) = \lambda e^{-\lambda k}$

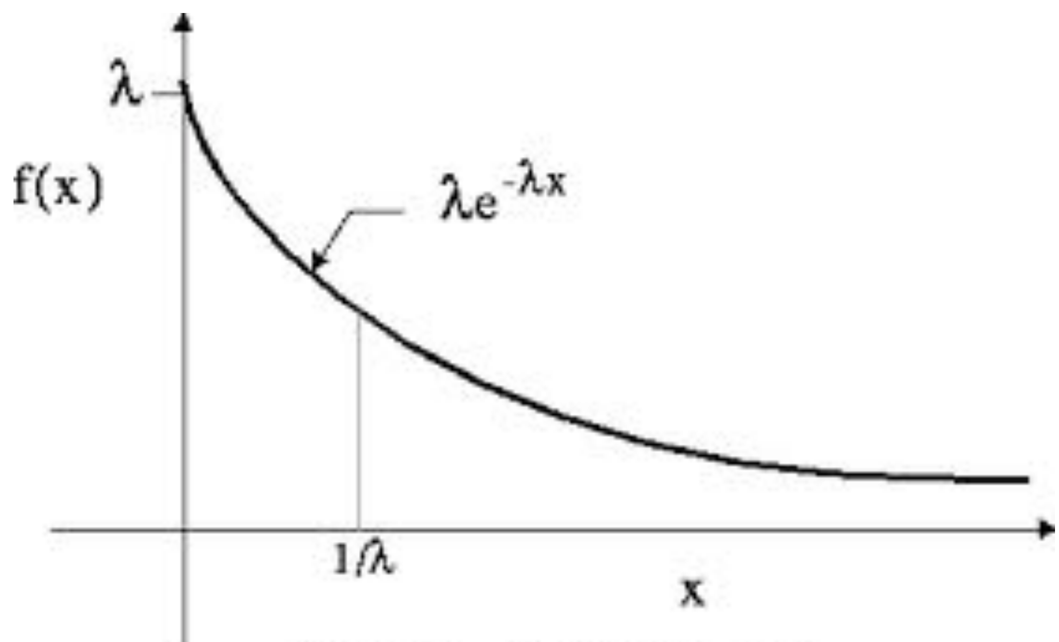


Figure 6. Exponential *pdf*

Exponential Random Variables

Exponential random variables model the amount of time until an incident of interest takes place

- Length of time before a message arrives at the computer
- Length of time before a light bulb burns out

PDF: $f_X(k) = \lambda e^{-\lambda k}$

$E[X] = \lambda^{-1}$ $\quad$ $var(X) = \lambda^{-2}$

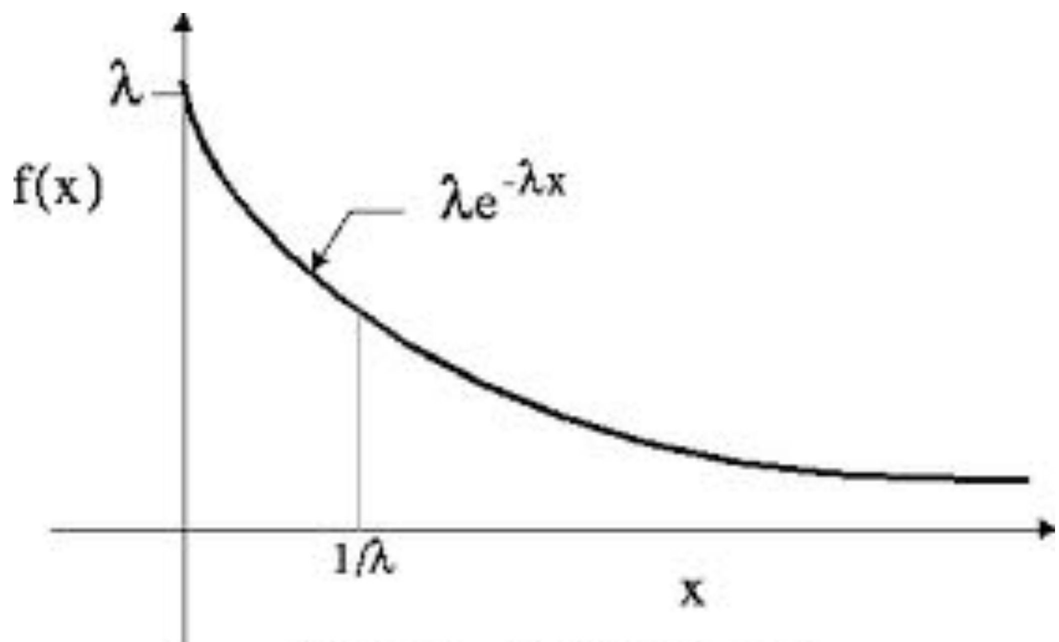


Figure 6. Exponential *pdf*

Cumulative Distribution Functions

A cumulative distribution function (CDF), denoted F_X, "accumulates" probability up to a certain value of X

$$F_X(k) = P(X \leq k)$$

Cumulative Distribution Functions

A cumulative distribution function (CDF), denoted F_X, "accumulates" probability up to a certain value of X

$$F_X(k) = P(X \le k)$$

- For discrete random variables,

$$F_X(k) = ???$$

Cumulative Distribution Functions

A cumulative distribution function (CDF), denoted F_X, "accumulates" probability up to a certain value of X

$$F_X(k) = P(X \leq k)$$

- For discrete random variables,

$$F_X(k) = \sum_{t \leq k} p_X(t)$$

Cumulative Distribution Functions

A cumulative distribution function (CDF), denoted F_X, "accumulates" probability up to a certain value of X

$$F_X(k) = P(X \leq k)$$

- For discrete random variables,

$$F_X(k) = \sum_{t \leq k} p_X(t)$$

- For continuous random variables,

$$F_X(k) = ???$$

Cumulative Distribution Functions

A cumulative distribution function (CDF), denoted F_X, "accumulates" probability up to a certain value of X

$$F_X(k) = P(X \leq k)$$

- For discrete random variables,

$$F_X(k) = \sum_{t \leq k} p_X(t)$$

- For continuous random variables,

$$F_X(k) = \int_{-\infty}^{k} f_X(t)\,dt$$

Cumulative Distribution Functions

A cumulative distribution function (CDF), denoted F_X, "accumulates" probability up to a certain value of X

$$F_X(k) = P(X \leq k)$$

- For discrete random variables,

$$F_X(k) = \sum_{t \leq k} p_X(t)$$

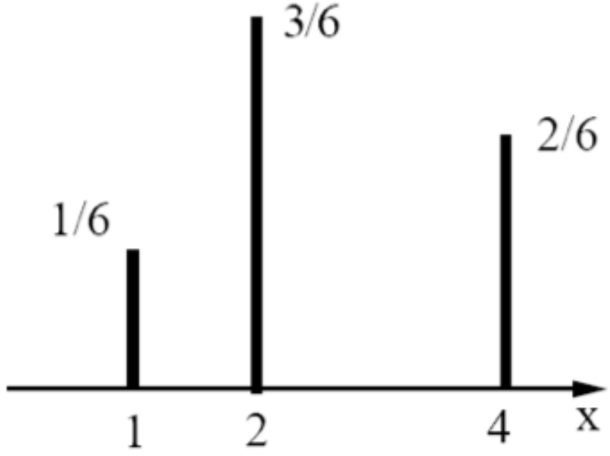
- For continuous random variables,

$$F_X(k) = \int_{-\infty}^{k} f_X(t)\,dt \quad \rightarrow \quad f_X(k) = \frac{dF_X}{dk}(k)$$

Cumulative Distribution Functions

Consider a discrete random variable X

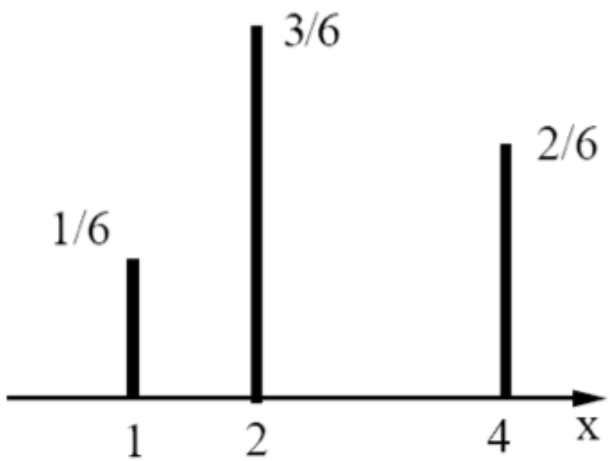
PMF:



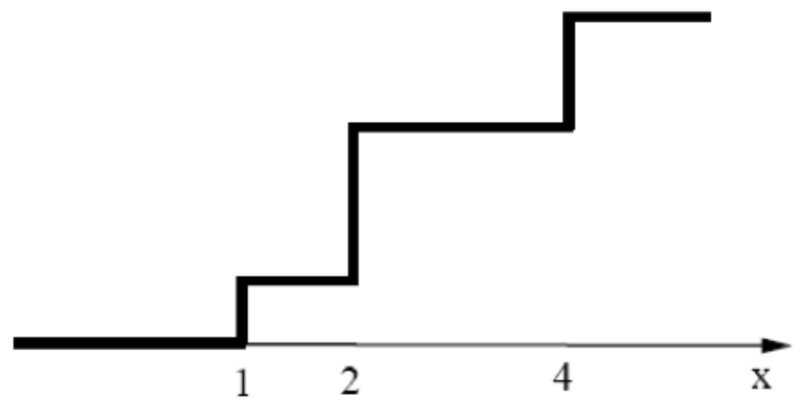
Cumulative Distribution Functions

Consider a discrete random variable X

PMF:

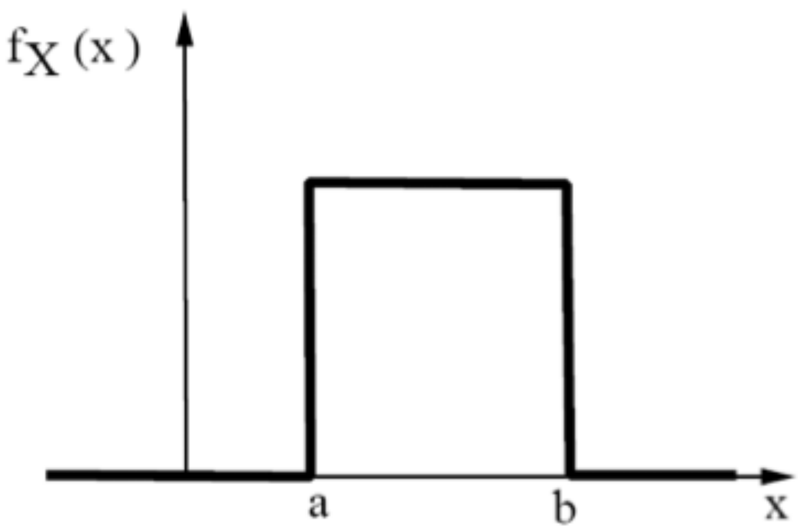


CDF:



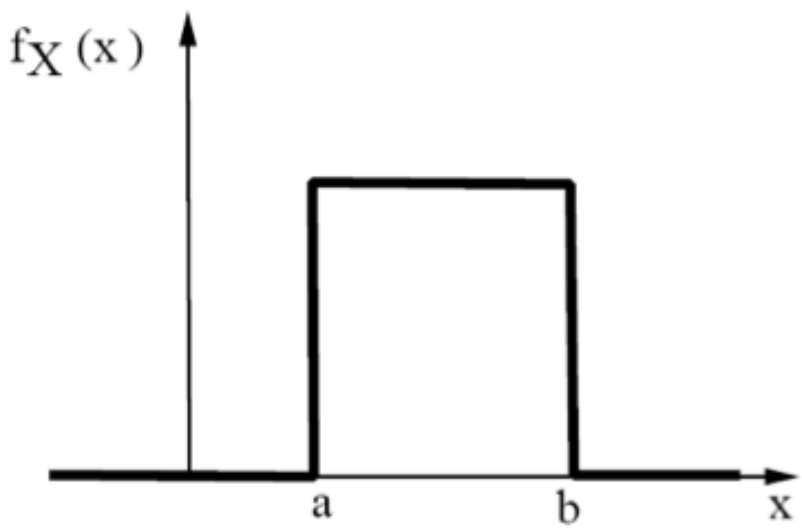
Cumulative Distribution Functions

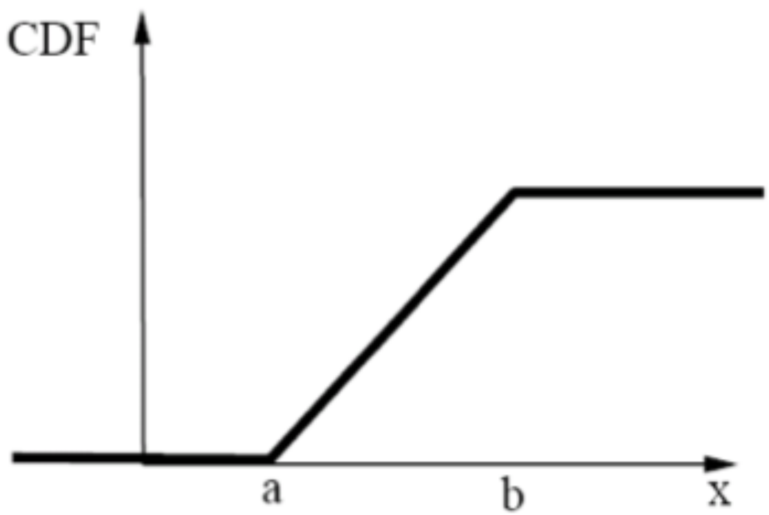
Consider a uniform continuous random variable X



Cumulative Distribution Functions

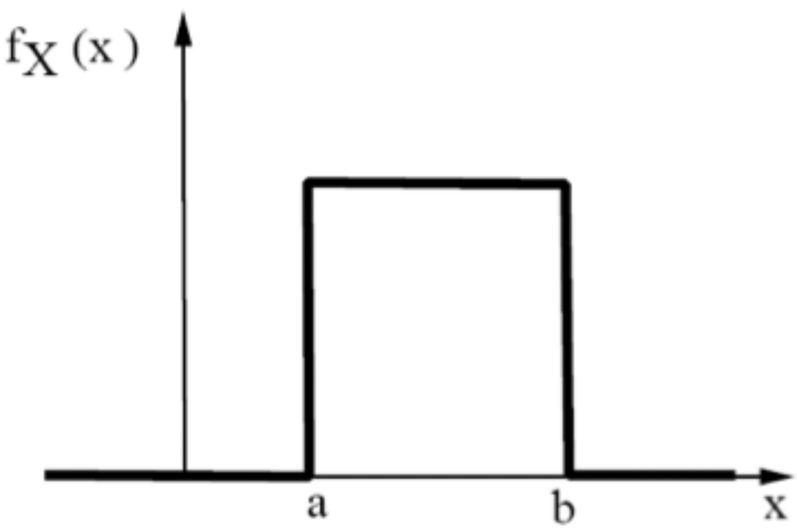
Consider a uniform continuous random variable X

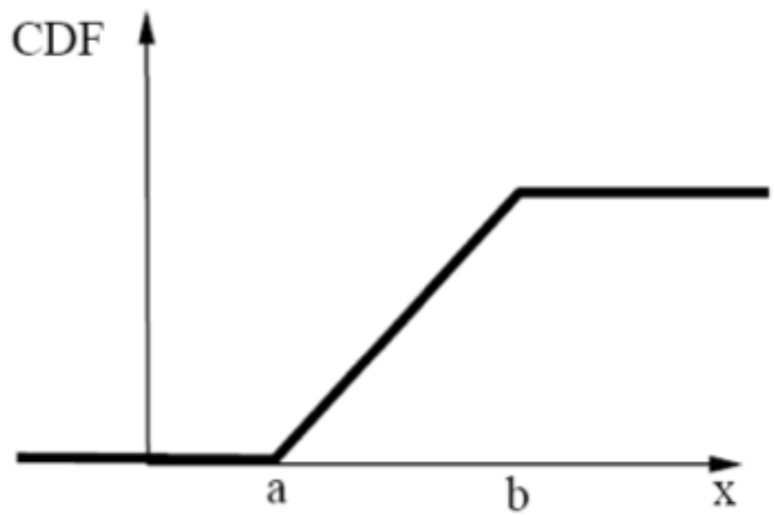




Cumulative Distribution Functions

Consider a uniform continuous random variable X





Note that CDFs are always monotonically non-decreasing

Exponential vs. Geometric

Let X be an exponential random variable with $f_X(k) = \lambda e^{-\lambda k}$.

$F_X(k) = ???$

Exponential vs. Geometric

Let X be an exponential random variable with $f_X(k) = \lambda e^{-\lambda k}$.

$$F_X(k) = 1 - e^{-\lambda k}$$

Exponential vs. Geometric

Let X be an exponential random variable with $f_X(k) = \lambda e^{-\lambda k}$.

$$F_X(k) = 1 - e^{-\lambda k}$$

Let Y be a geometric random variable with $p_Y(k) = p(1-p)^{k-1}$.

$$F_Y(k) = ???$$

Exponential vs. Geometric

Let X be an exponential random variable with $f_X(k) = \lambda e^{-\lambda k}$.

$$F_X(k) = 1 - e^{-\lambda k}$$

Let Y be a geometric random variable with $p_Y(k) = p(1-p)^{k-1}$.

$$F_Y(k) = 1 - (1-p)^{\lfloor k \rfloor}$$

Exponential vs. Geometric

Let X be an exponential random variable with $f_X(k) = \lambda e^{-\lambda k}$.

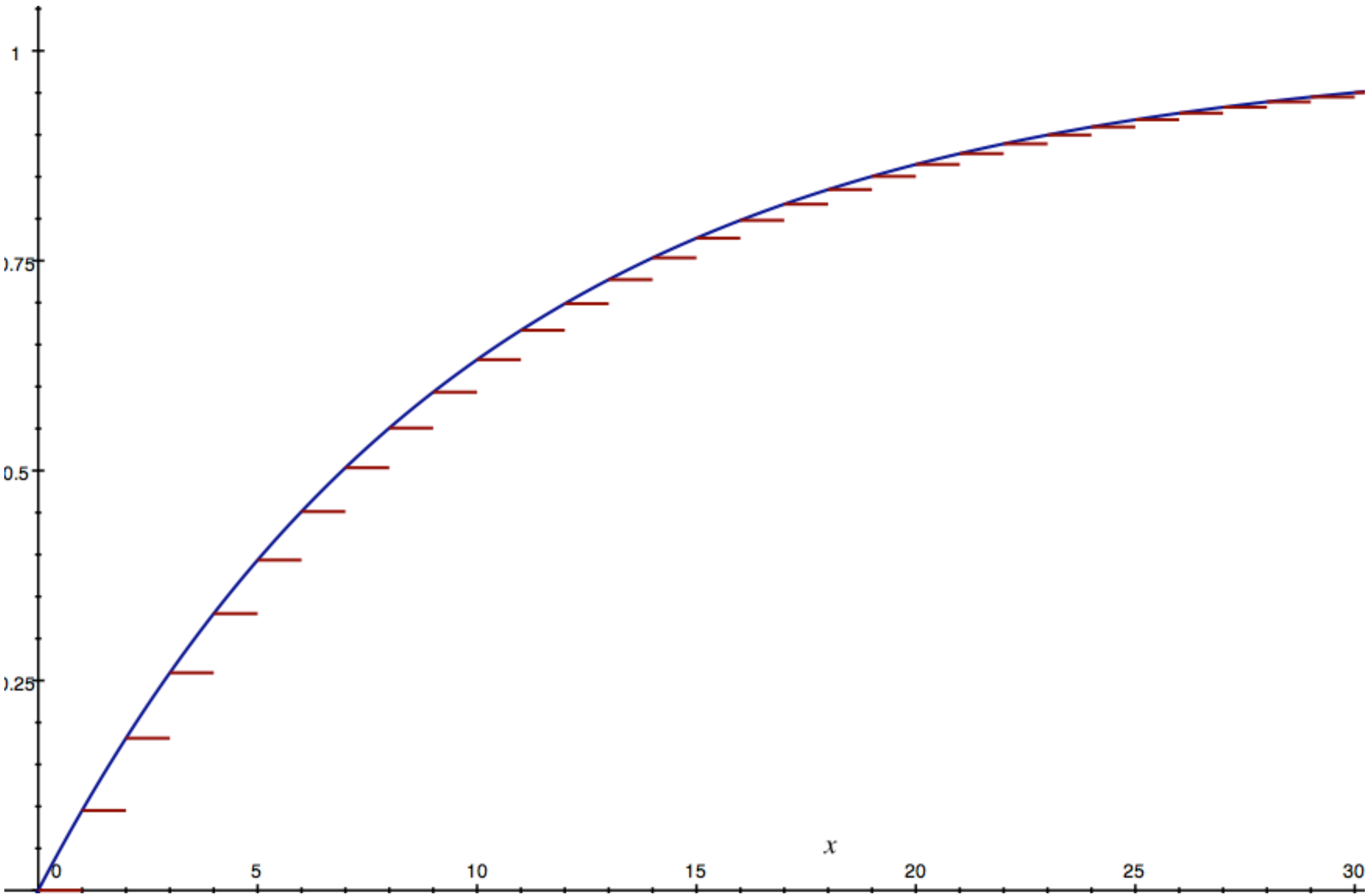
$$F_X(k) = 1 - e^{-\lambda k}$$

Let Y be a geometric random variable with $p_Y(k) = p(1-p)^{k-1}$.

$$F_Y(k) = 1 - (1-p)^{\lfloor k \rfloor}$$

Suppose that we set $p = 1 - e^{-\lambda}$…

Exponential vs. Geometric



Normal Random Variables

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

just for normalization

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

Things to notice:

- Maximized when $k = 0$

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

Things to notice:

- Maximized when $k = 0$
- Symmetric around 0

Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

Things to notice:

- Maximized when $k = 0$
- Symmetric around 0
- Drops exponentially fast

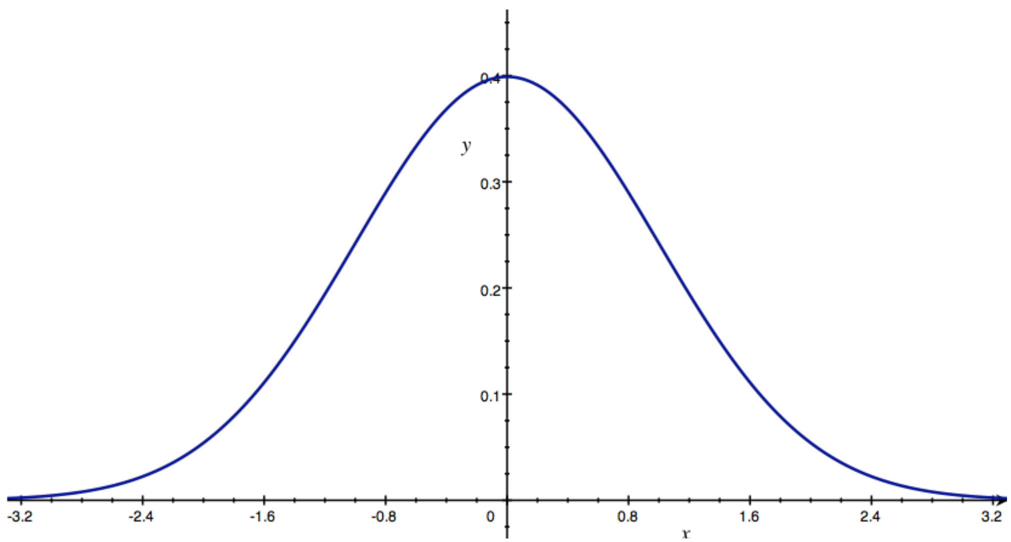
Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

Things to notice:

- Maximized when $x = 0$
- Symmetric around 0
- Drops exponentially fast



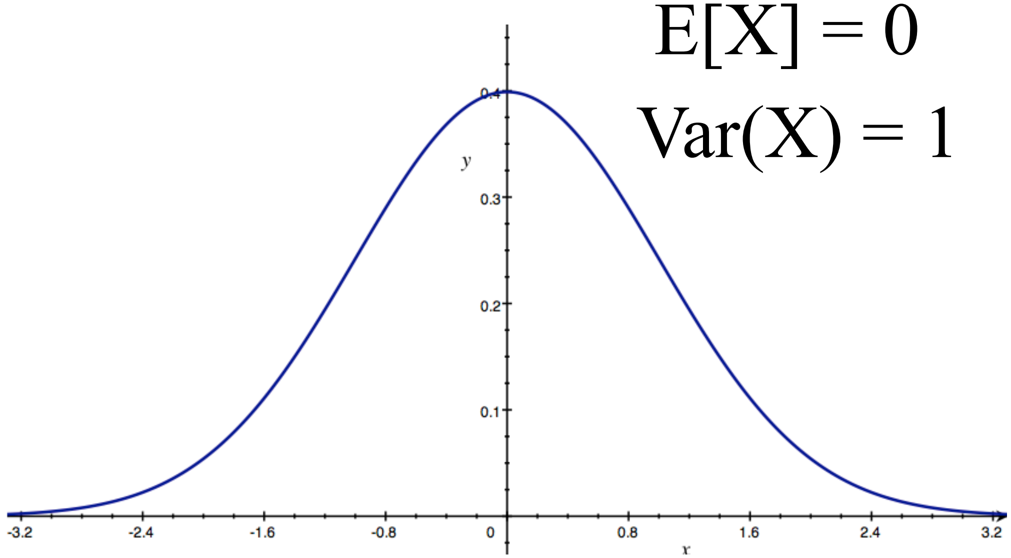
Standard Normal Random Variable

The standard normal random variable is characterized by the following PDF:

$$f_X(k) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}k^2}$$

E[X] = 0

Var(X) = 1



Things to notice:

- Maximized when $x = 0$
- Symmetric around 0
- Drops exponentially fast

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

first term still just normalization

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

Things to notice:

- Maximized when $k = \mu$

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(k) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

Things to notice:

- Maximized when $k = \mu$
- Symmetric around μ

Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(k) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

Things to notice:

- Maximized when $k = \mu$
- Symmetric around μ
- Drops exponentially fast

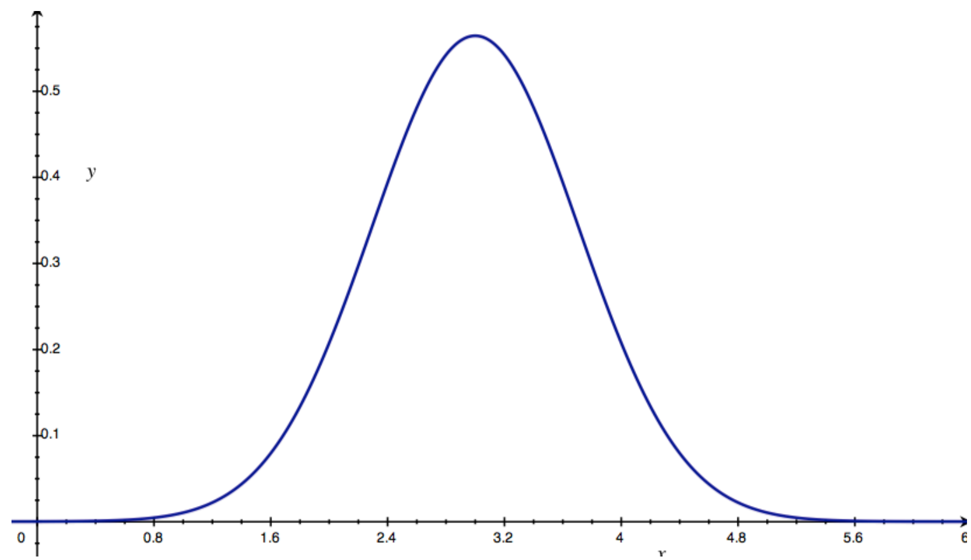
Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

$$f_X(k) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

Things to notice:

- Maximized when $k = \mu$
- Symmetric around μ
- Drops exponentially fast



Normal Random Variable

A continuous random variable X is said to be normal or Gaussian if the PDF has the form:

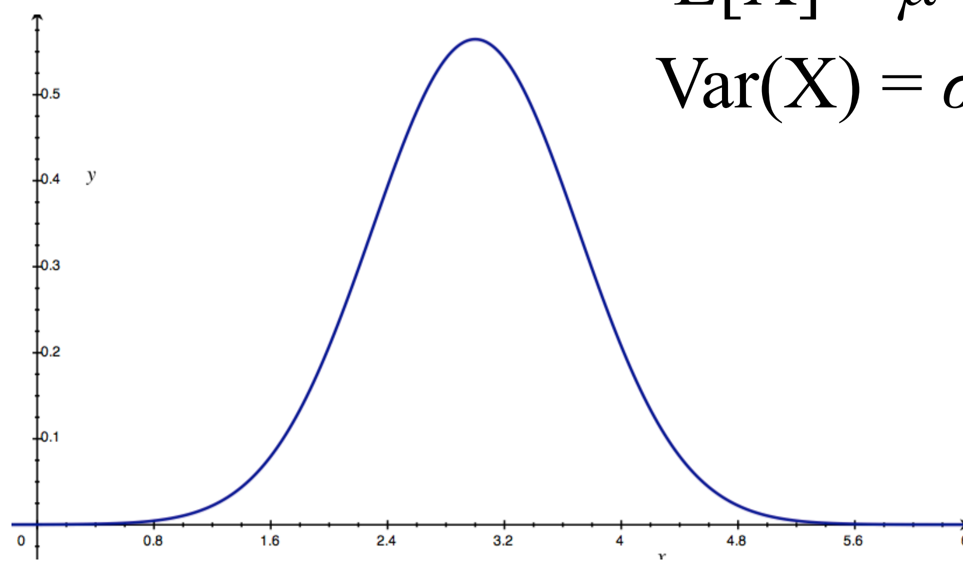
$$f_X(k) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(k-\mu)^2}$$

E[X] = μ

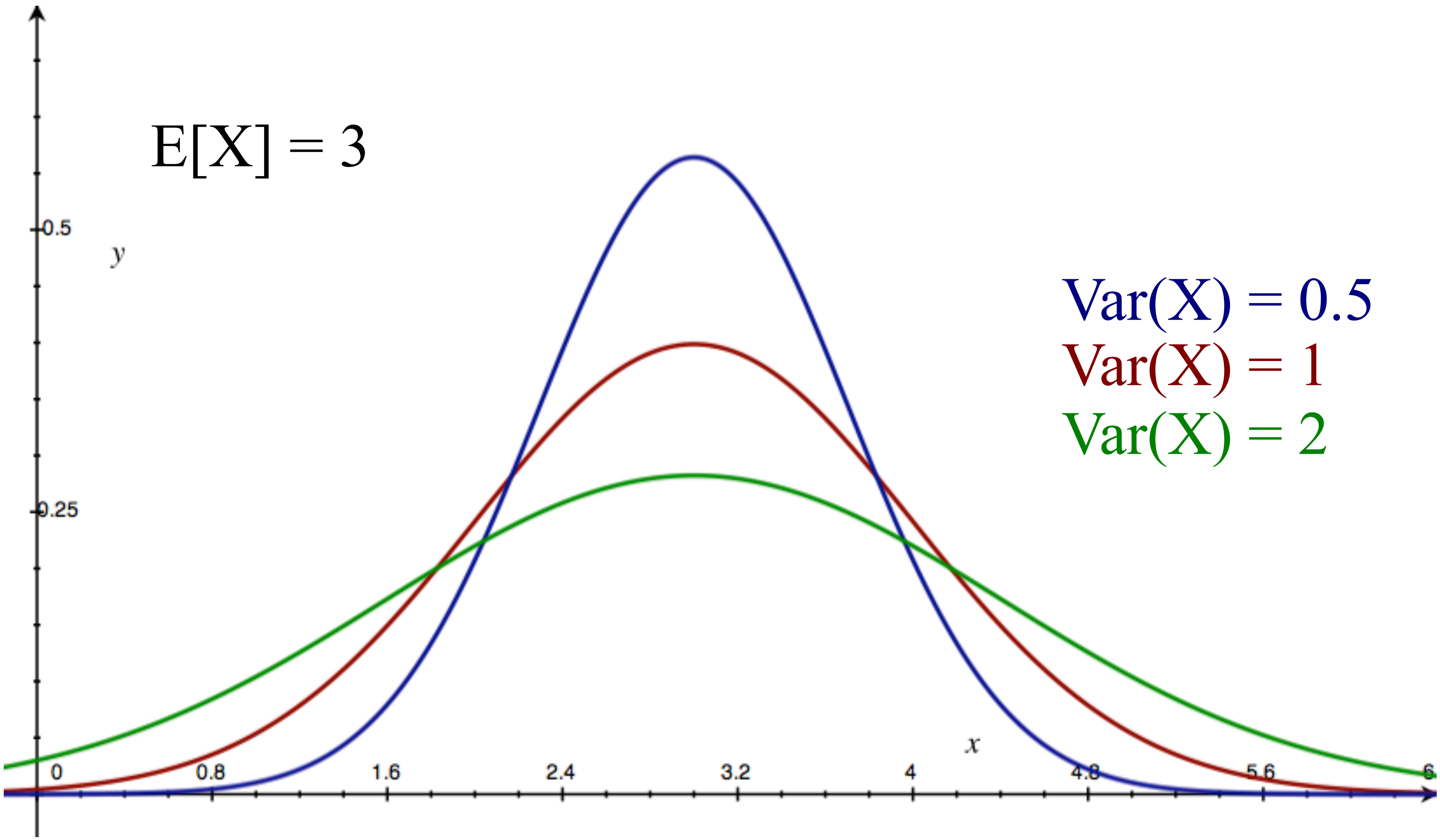
Var(X) = σ^2

Things to notice:

- Maximized when $k = \mu$
- Symmetric around μ
- Drops exponentially fast



Normal Random Variables



Examples

Normal random variables are very tractable analytically, and so are often used to as an approximate model in the natural sciences and the social sciences

- Measurement error or noise in an experiment
- The run time of a program
- A random student’s score on a standardized test (“bell curve grading”)

Examples

Normal random variables are very tractable analytically, and so are often used to as an approximate model in the natural sciences and the social sciences

- Measurement error or noise in an experiment
- The run time of a program
- A random student's score on a standardized test ("bell curve grading")

Could run times or student grades be exactly normal?

Another Example: Sampling

Key fact: The sum or average X of a large number of independent and identically distributed random variables X_1, X_2, ... has a CDF that is approximately normal, regardless of the CDF of the individual random variables.

Another Example: Sampling

Key fact: The sum or average X of a large number of independent and identically distributed random variables X_1, X_2, ... has a CDF that is approximately normal, regardless of the CDF of the individual random variables.

- Example: X is the fraction of surveyed voters who approve of the president (sample mean).

Another Example: Sampling

Key fact: The sum or average X of a large number of independent and identically distributed random variables X_1, X_2, ... has a CDF that is approximately normal, regardless of the CDF of the individual random variables.

- Example: X is the fraction of surveyed voters who approve of the president (sample mean).

We'll come back to this idea again...