Random variables

Dr Ahmad S. Tarawneh

Definition

- **A random variable,** denoted by *X,* is a function that maps each outcome in the sample space of a random experiment to a real number.
- This mapping allows us to quantify outcomes in a way that facilitates mathematical analysis and probability calculations
- There are two types of random variables
  - Discrete random variable: countable and the values can be listed $x_1, x_2, \; x_3..$
  - Continuous random variable: not countable and cannot be listed, presented as interval

Example

- In the experiment of flipping a coin two times, and X is a random variable represents the number of heads
  - what is the possible values of X

- **Solution**

$$S = \{HH, HT, TH, TT\}$$

2 1 1 0

- $x = 0, 1, 2$
- we can ask what is the probability of x $P(X = x)$, such that x is an outcome

$$P(X = 0) = \frac{1}{4}, P(X = 1) = \frac{2}{4}$$

Other examples

- X is a random number representing the number of customers who enter a store each hour.
  - X is considered a random variable because this number can vary and is subject to change
- Y is a random variable representing the service time in minutes of a customer by a specific server
- Z is a random variable representing the money you hold in your wallet in JOD
- W is a discrete random variable representing the number of cars expected to enter a gas station on a particular day of the month.

Probability mass function

- For a discrete random variable X with possible values $x_1, x_2,\ x_3..$ The probability mass function (PMF) is a function such that:
  1. $f(x_i) \geq 0$
  2. $\sum_{i=1}^{n} f(x_i) = 1$
- $f(x_i)$ is equal to $P(X = x_i)$,

- Think of the probability of the color value in an image

Color V	0	1	2	3	...	255
Probability	0.01	0.09	0.02	0.05		0.001

  - The some of these probabilities must equal 1, if not then it is not PMF

Example

- given the following mass function:

x	-2	-1	0	1	2
$f(x) = P(X = x)$	1/8	2/8	2/8	2/8	1/8

- Find
  1. $P(X \leq 2)$
  2. $P(X \geq 0)$
  3. $P(-2 < X < 2)$
  4. $P(X < 0 \text{ or } X = 1)$

Cumulative Distribution Function (CDF)

- For a random variable $\boldsymbol{X}$, the Cumulative Distribution Function, $\boldsymbol{F}(\boldsymbol{x})$, is defined for every number $\boldsymbol{x}$ as:

$$F(x) = P(X \leq x)$$

- This means ***F(x)*** represents the probability that the random variable $\boldsymbol{X}$ will have a value less than or equal to $\boldsymbol{x}$

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

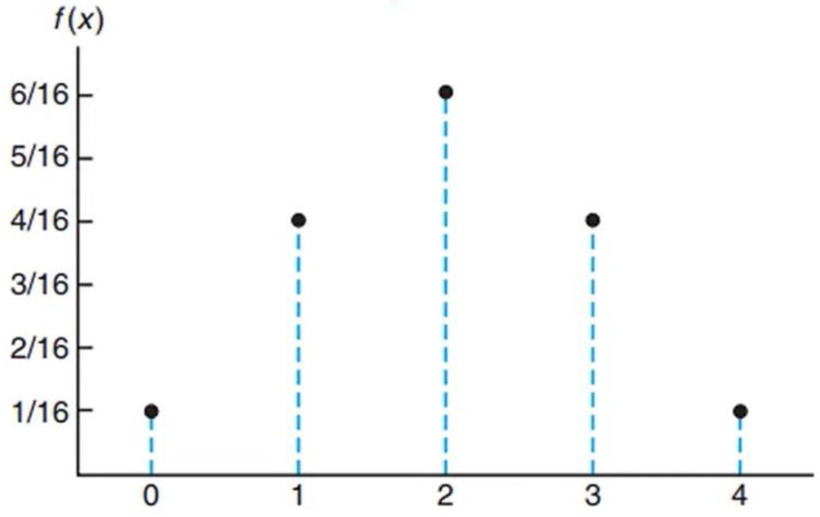
Cont.

- When X is discrete

x	-2	-1	0	1	2
$f(x) = P(X = x)$	1/8	2/8	2/8	2/8	1/8
$\boldsymbol{F(x) = P(X \leq x)}$	**1/8**	**3/8**	**5/8**	**7/8**	**8/8**

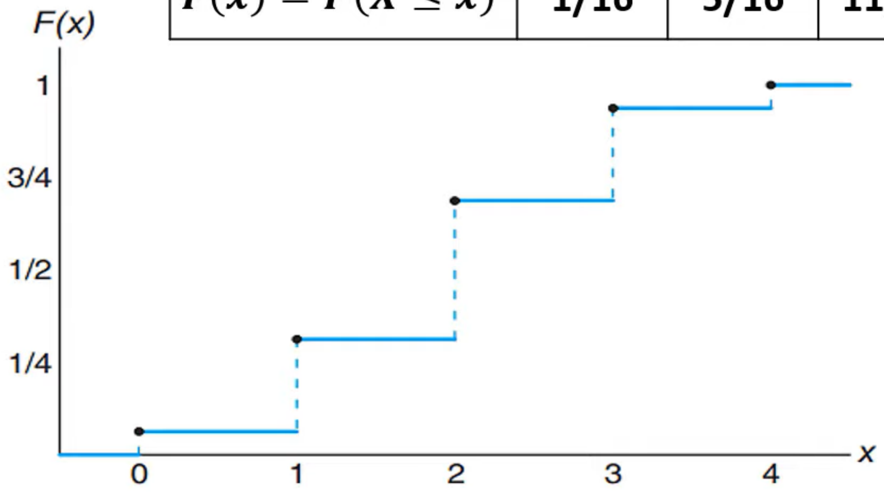
PMF vs CDF plot

x	0	1	2	3	4
$f(x) = P(X = x)$	1/16	4/16	6/16	4/16	1/16



Probability mass function plot.

x	0	1	2	3	4
$f(x) = P(X = x)$	1/16	4/16	6/16	4/16	1/16
$\boldsymbol{F(x) = P(X \leq x)}$	**1/16**	**5/16**	**11/16**	**15/16**	**16/16**



Discrete cumulative distribution function.

Illustration

- let us assume that x represent the number of typos in a given page of a book

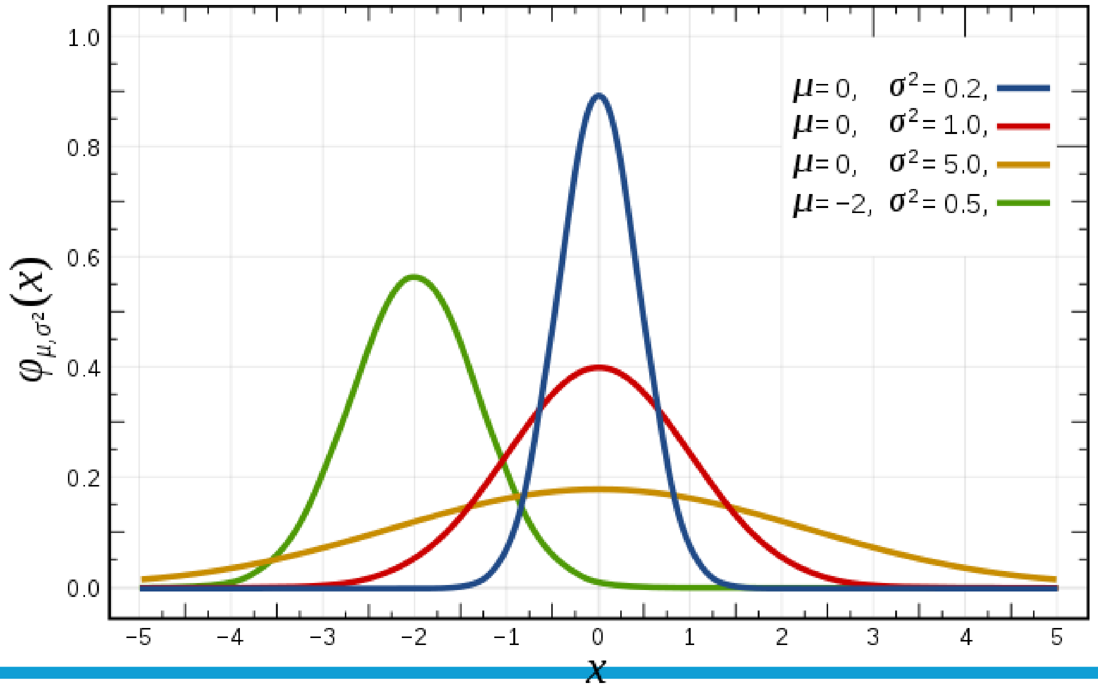
x	0	1	2	3	4
$f(x) = P(X = x)$	1/16	4/16	6/16	4/16	1/16
$\boldsymbol{F(x) = P(X \leq x)}$	**1/16**	**5/16**	**11/16**	**15/16**	**16/16**

- Find:
  1. The probability of having less than or equal 2 errors
     - this can be found using the CDF $F(2)=11/16$
  2. Errors fewer than or equal to 4 errors
     - $F(4)= 16/16$
  3. Errors in-between 1 and 3
     - $f(1)+f(2)+f(3) =4/16 + 6/16 + 4/16 = 14/16$
     - $OR \quad F(3)-F(0)= 15/16 - 1/16 =14/16$

- **Such data can be collected from real life.**
- **we will come later in this course on how to collect data and make such an analysis**

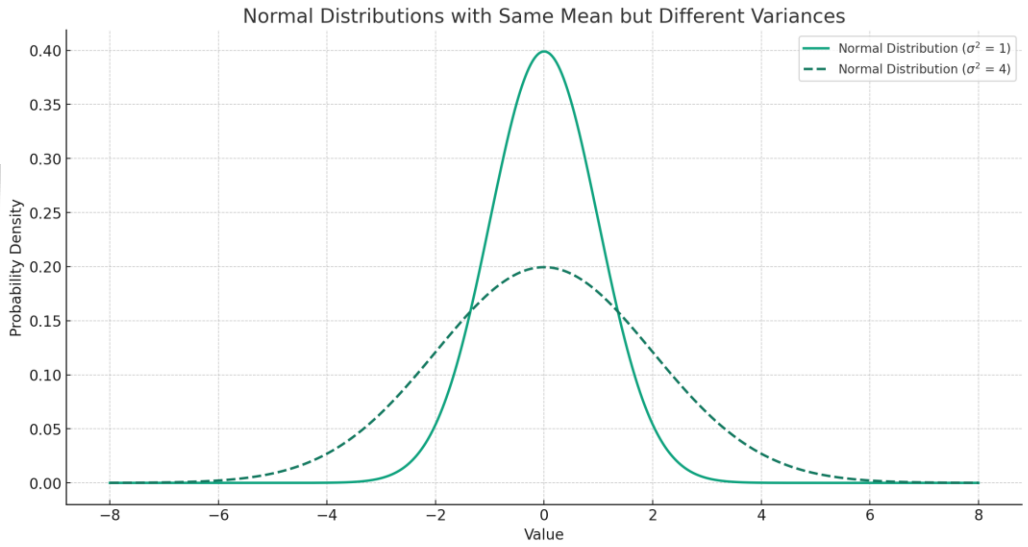
Mean and Variance

- Mean and variance are very important metrics used to summarize a probability distribution for a random variable.
- Mean represents the center of the probability distribution, while variance measures the dispersion of the distribution

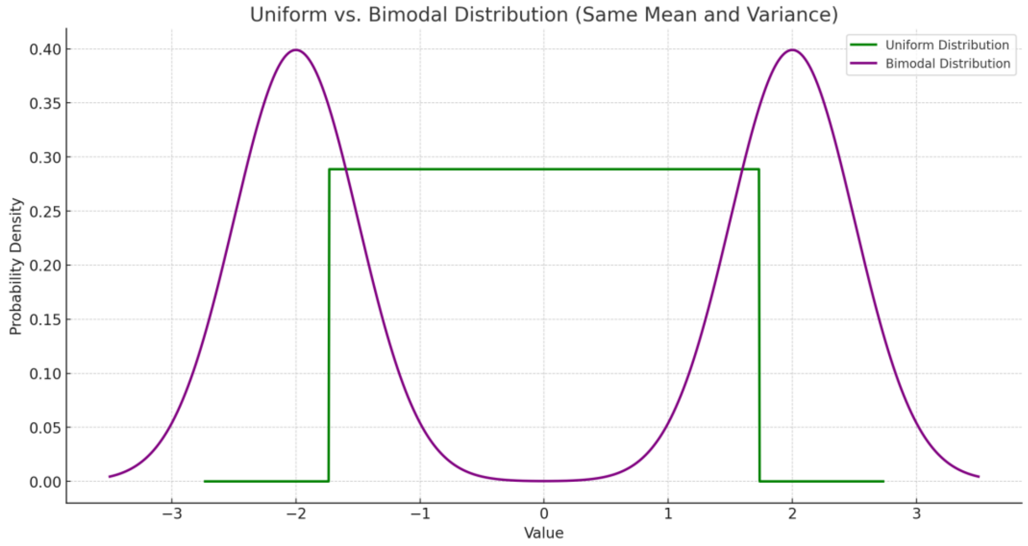


Cont.

- Sometimes, the probability distribution has the same mean but different variance



- Or the same mean and same variance with different distribution



Mean/Variance/Standard deviation

- The mean μ (aka expected value $E(X)$) can be calculated using the following formula

$$\mu = E(X) = \sum_{x} x f(x)$$

- Variance σ^2 or $V(X)$

$$\sigma^2 = E(X^2) - \left(E(X)\right)^2 = \sum_{x} x^2 f(x) - \mu^2$$

- Standard deviation

$$\sigma = \sqrt{\sigma^2}$$

Example

x	1	2	3	4	5	6
$f(x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

- Calculate the expected value and variance of the following distribution
- **Expected value**

$$E(X) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$$

- **Variance**

$$E(X^2) = 1^2\left(\frac{1}{6}\right) + 2^2\left(\frac{1}{6}\right) + 3^2\left(\frac{1}{6}\right) + 4^2\left(\frac{1}{6}\right) + 5^2\left(\frac{1}{6}\right) + 6^2\left(\frac{1}{6}\right) = 15.16$$

$$V(X) = 15.16 - 3.5^2 = 2.92$$

Continuous Random variable

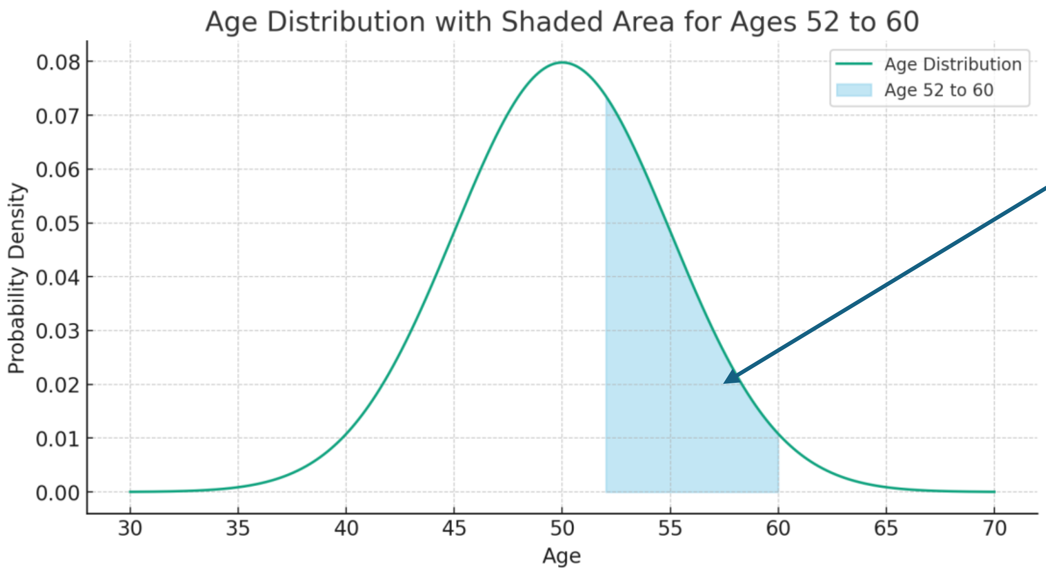
- If the range of the random variable X is an interval e.g. 1-2, then this random variable is continuous, as the numbers in this range are not countable and can not be listed

- In continuous we cannot find the probability of a single value due to the use of the integration and area-under-curve calculation
  - There is no area of a single line, its width is 0

$$\int_{a}^{a} f(x)\, dx = 0$$

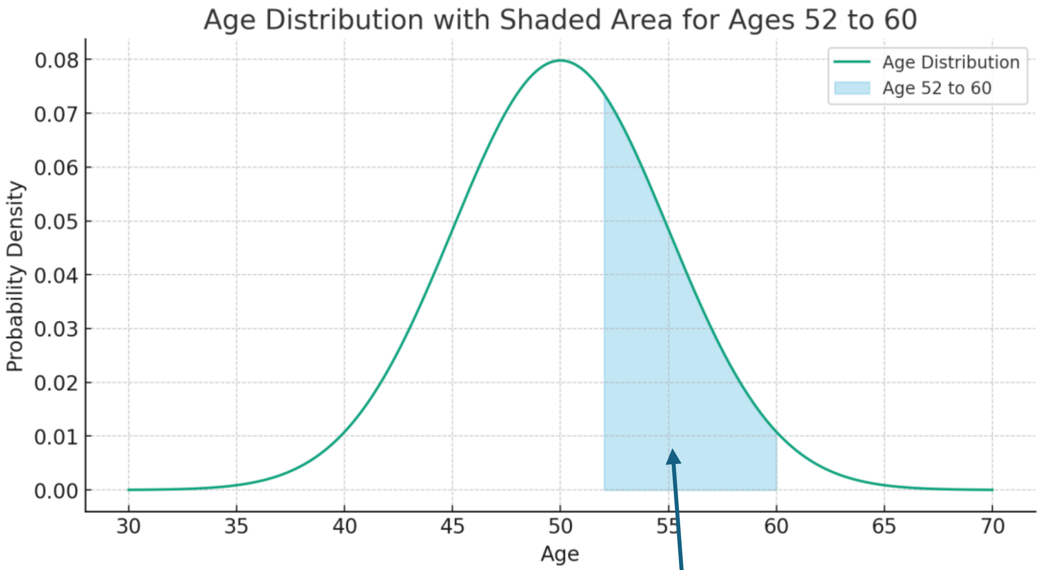
Example

- let's assume that in a certain population, the mean age is 50 years, and the standard deviation is 5 years.
- The shaded area represents the probability of selecting a person from this population whose age is at least 52 but not more than 60.



$f(52 \leq x \leq 62)$

Cont.



$$f(52 \le x \le 62) = f(52 < x \le 62) = f(52 \le x < 62) = f(52 < x < 62)$$

Probability density function

- For a continuous random variable X, the probability density function (PDF) is a function such that:
  1. $f(x_i) \geq 0$
  2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
  3. $P(a \leq x \leq b) = \int_a^b f(x)\,dx =$ the area under the curve within the defined interval

Example

- let us assume that we have the following equation:

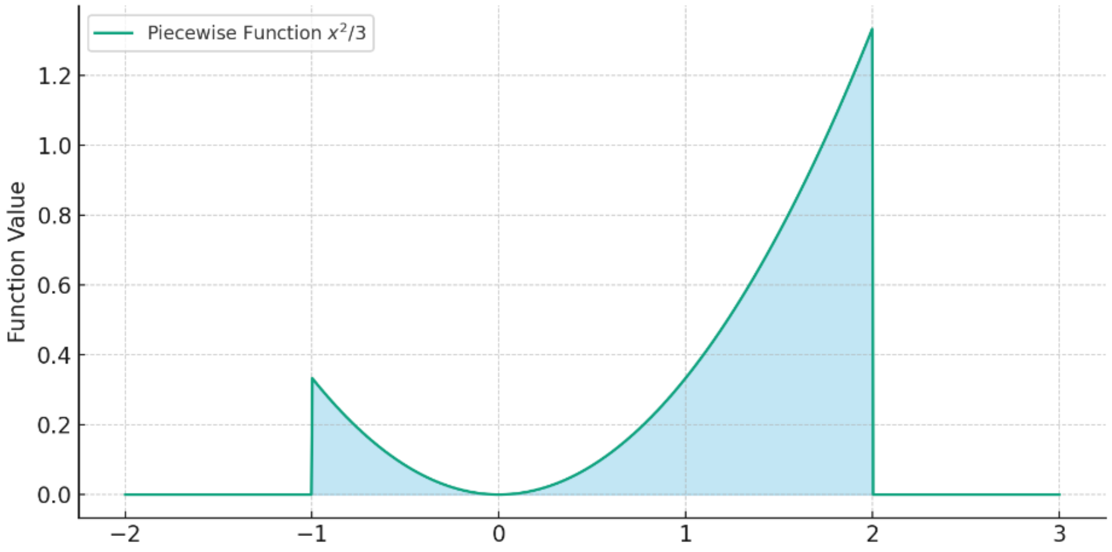
$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ otherwise, & 0 \end{cases}$$

- First, we make sure that this is an actual PDF, by calculating the area under this function within the given interval
  - This can be done by calculating the integral of this function for the defined interval

Cont.

$$area = \int_{-1}^{2} \frac{x^2}{3}\, dx = \frac{x^3}{9}\Bigg|_{-1}^{2} = \frac{8}{9} - \frac{-1}{9} = 1$$

- Therefore, this can be considered as a pdf
- based on this you can calculate the probability of any interval within the function range using the same procedure
  - e.g., $P(0.5 < x < 1)$



Cumulative distribution function

- For the PDF you can also calculate the CDF $F(X)$.

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

- This can help in calculating probabilities without performing integration

Cont.

- given the CDF $F(X)$, one can calculate the probability $P(a < X \leq b)$ using:

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

- Let us take an example of calculating the CDF for the previous function
  - and use it to find the probability $P(0 < X \leq 1)$

$$f(x) = \begin{cases} \frac{x^2}{3}, & -1 < x < 2, \\ otherwise, & 0 \end{cases}$$

Cont.

- We made sure that this is a PDF as its area is exactly 1
- Secondly, we find the integral for this function

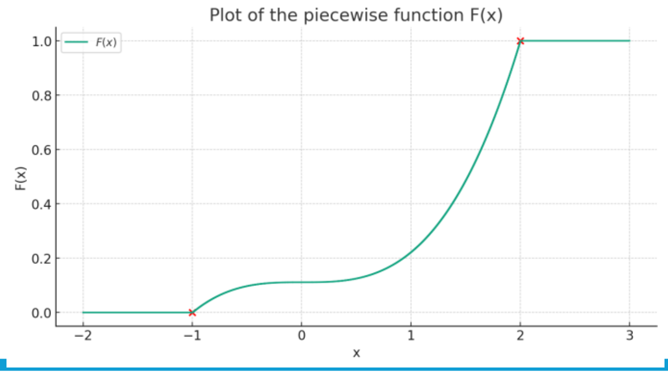
$$\int_{-1}^{+x} \frac{t^2}{3}\, dt$$

  - note x becomes t, as x represents the limit of the integration (interval)
- we calculate the integral of this function to get

$$\frac{x^3}{9} - \frac{-1}{9} = \frac{x^3+1}{9}$$

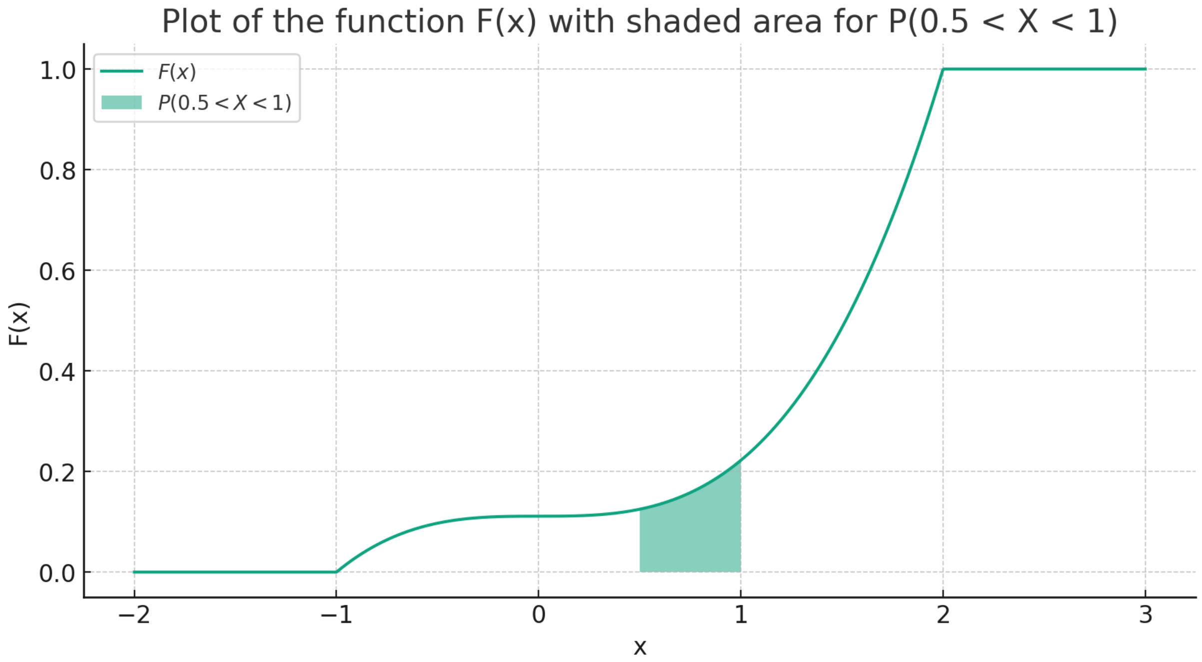
- To make sure that x will not be out of its range, we rewrite it to be:

$$F(x) = \begin{cases} 0, & x < -1 \\ \dfrac{x^3+1}{9}, & -1 \le x < 2 \\ 1, & x \ge 2 \end{cases}$$



Cont.

- now, to calculate $P(0.5 < x < 1)$ we can use this function:
- $F(X) = F(b) - F(a) = F(1) - F(0.5) = \frac{2}{9} - \frac{0.125}{9} = 0.0972$



- Outside the range -1 and 2 the probability is 0 as we are subtracting the number from itself.
  - *P(2.5<X<3) is 0*

Exercise

- Find whether the following functions are a PDFs or not
- if yes, find its cumulative function

$$f(x) = \begin{cases} x^3, & 2 < x < 4, \\ otherwise, & 0 \end{cases}$$

$$g(x) = \begin{cases} 4x^3, & 0 < x < 1, \\ otherwise, & 0 \end{cases}$$