# Day-6

11 August 2024 17:38

Delivered content

| # | Topic | Number of Use cases/ Examples discussed | Explained Y/N, comments if any |
|---|---|---|---|
| 1 | Poisson dist | - 7 excel based exercises | |
| 2 | Binomial | - 1 exercise set | |
| 3 | Normal distribution | - 2 python code files, 7 general purpose examples<br>- PDF/CDF on ca_housing dataset<br>- Example code on why pdf can be more than 1 | |
| 4 | Exp vs Poisson - context | - Example of call center - modeling | |
| 5 | Z-score | - Examples on where to use this | |
| 6 | CLT (central limit theorem) | - Example of inferring on population | CLT to be covered in Inferential stat module |
| 7 | | | |
| 8 | | | |
| 9 | | | |
| 10 | | | |
| | | | |

Topics = { Concept, Code, examples)

**Exercise on Binomial distribution**

- Hepatitis C ("Hep C") is a virus affecting the liver, whose symptoms include inflammation, cirrhosis, and all sorts of other nastiness.
- According to the World Health Organization (WHO), 4.1% of people have Hep C. In humans, it is discovered via a routine blood test.
- In screening for Hep C, some health care providers, to save time and money, combine blood samples from 5 patients to test, and
  - if the combined sample comes back negative, it means that all 5 folks are not infected with Hep C.
  - if the combined sample comes back positive, it means that at least one of the people in the combined sample has Hep C.
    - Then, each of the 5 must be individually tested to see who is infected.
- Suppose 5 randomly selected folks have their blood taken, and their blood is placed into a combined

sample. This combined sample is then tested for Hep C.

p(individual person H-C) = 0.041 → 4.1%

p(no H-C) = 1 − 0.041 → 96%

Qs 1 : Find the chance the combined sample comes back negative, to the nearest percent.

p(none of the 5 have H-C) = 18

$0.959 \times \cdots \times 5 = 82\%$

Qs 2 : Finding the chance the combined sample comes back positive

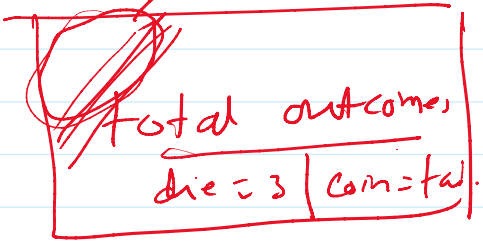
1 − 82 = 18%

p(die = 3) = $\frac{1}{6}$ | p(coin = tails) = $\frac{1}{2}$

p(die = 3 & coin = tail)

p1 Joint prob p2

multiply p1 * p2




Qs 3 : Suppose a Hep C test (whether done on a single blood sample, or 5) costs around $100.

If you sample 100 random people, approximately how much is saved by using combined samples instead of individual testing (assuming you want to individually ID all those infected with Hep C)?

According to a 2010 CDC report, approximately 85% of Americans have health insurance. Suppose we randomly select 10 Americans.

ind

10,000

grp-

20000

↓

individual if grp test

individual if grp test +ve

prob of how many grps may fail/+ve.

$p(H \sim c) = 18\%$

20 grps $\Rightarrow$ 18% of 20

$= 4$

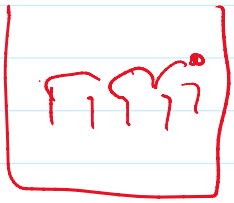
20 additional test

$= 2000$

$\Rightarrow$ 5000 savings

Qs 4 : Can you think of why it might not make financial sense to combine test?

Qs 5 : Does this sampling represent independent or dependent sampling?



Qs 6 : Find P(all of them have health insurance)

$S = 10$

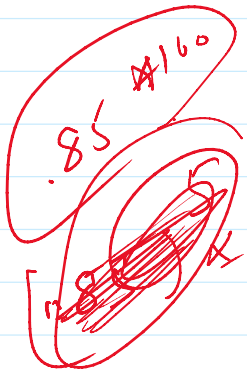
$p$(all 10 have ins) $= .85 \times 10$ 10

$= .10$

- Qs 7 : Find P(at least one of them does not have health

- Qs 7 : Find P(at least one of them does not have health insurance).

$1 - .10 = .90\%$



Qs 8 : Find P(half of them have health insurance).

$p(\text{insured}) = .85$
$p(\text{not-ins}) = .15$

$n = 10$
$k = 5$

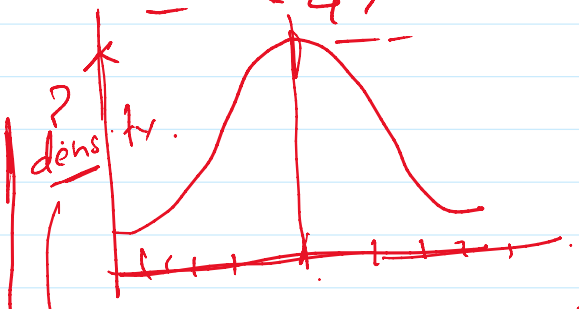
$pmf(X = 5) = \binom{n}{k} p^{k} (1-p)^{n-k}$

$= \binom{10}{5} p^{5} (1-p)^{5}$

$= .247$

**Continuous prob distribution**

- Normal dist
- Standard dist
- Uniform dist



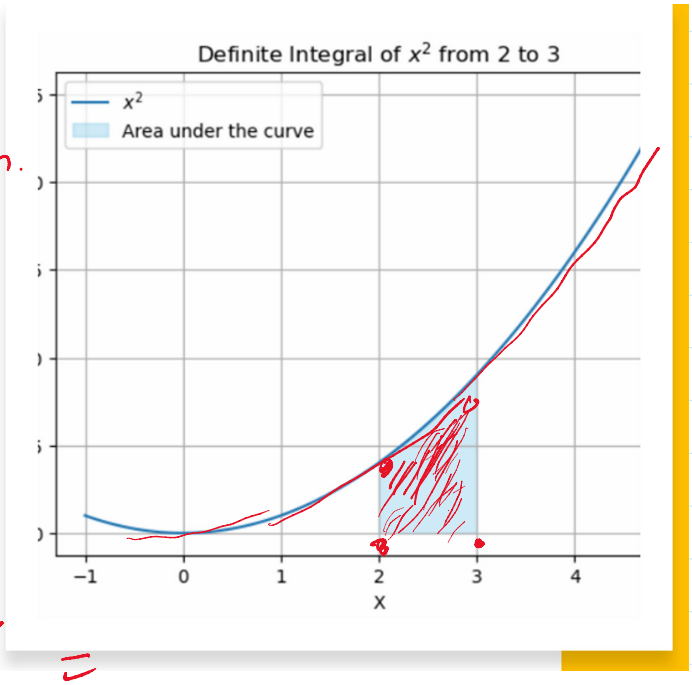

range of values & density ?

density > 1 ?

TINKU's pov : it can't be greater than 1 as the highest freq is .4

# Some maths - Integrals

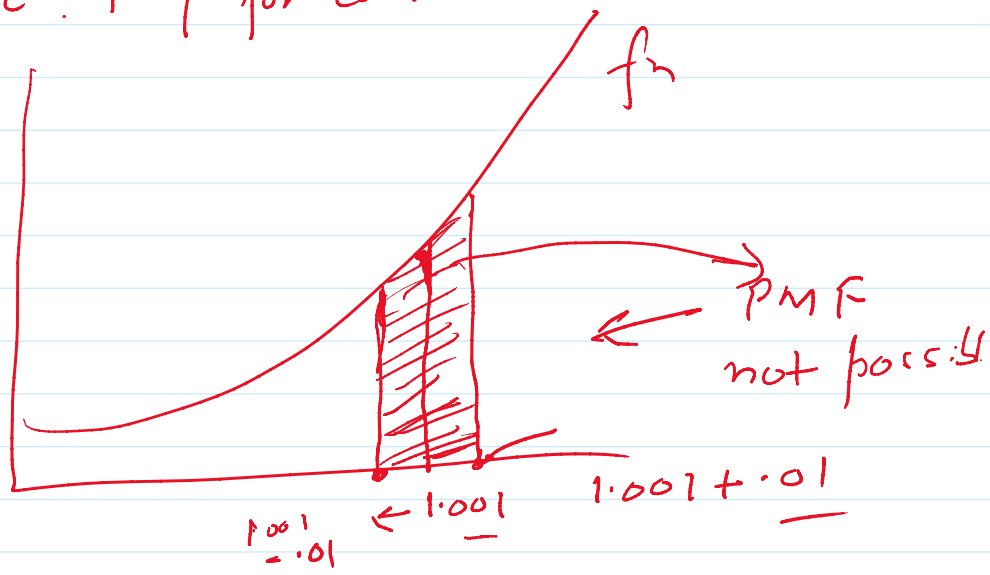
- The function $f(x)=x^2$ is plotted over the interval [-1, 5]
- compute the definite integral of $x^2$ from 2 to 3 → Area under the curve shaded
- Definite Integral - represents the signed area under the curve, and in this case, it corresponds to the area of the shaded region in the plot




Sum.

decimal/numeric $\int x^2$ =

1.001001009 ↖ specific? PMF for continuous dist. X



# What is probability density function?

Probability Density Function (PDF) represents the probability distribution of a continuous random variable.

describes the likelihood of the random variable taking on specific values within a given range.

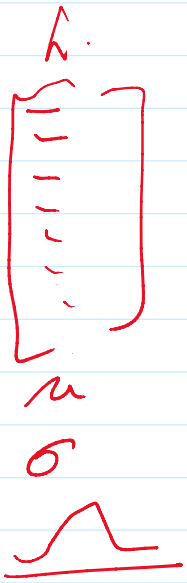
## Example 1 - Height of Adult Population

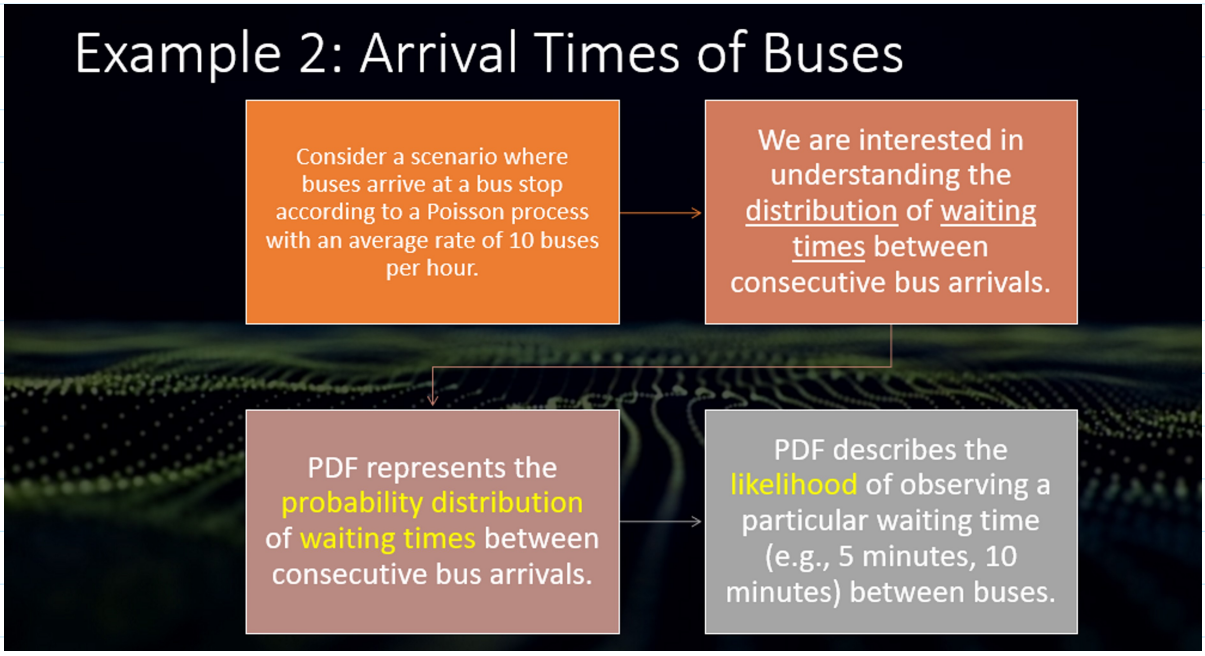
Suppose we are interested in understanding the distribution of heights among adults in a certain country.

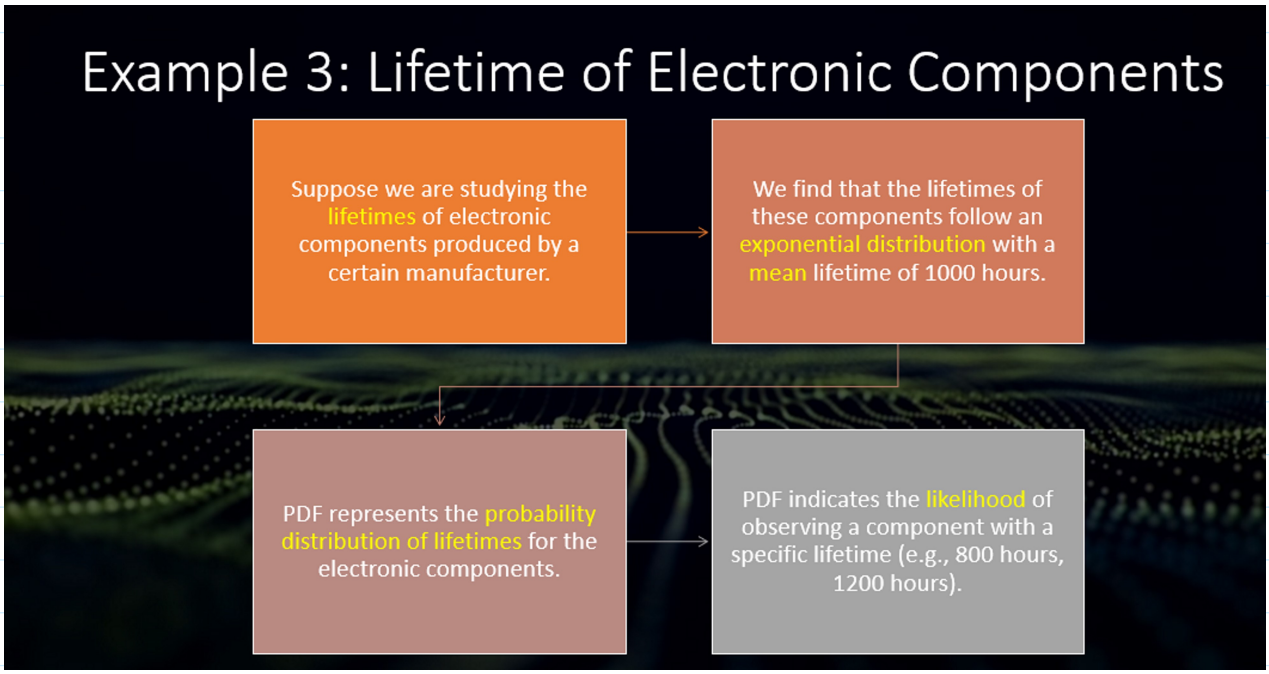
We collect data from a large sample of adults and find that their heights follow a normal distribution with a mean of 170 cm and a standard deviation of 10 cm.

PDF represents the probability distribution of heights among adults in the population.

For any given height value within a certain range (e.g., between 160 cm and 180 cm), the PDF provides a measure of the likelihood that a randomly selected adult will have a height falling within that range.




Example 2: Arrival Times of Buses
Consider a scenario where buses arrive at a bus stop according to a Poisson process with an average rate of 10 buses per hour.
We are interested in understanding the distribution of waiting times between consecutive bus arrivals.
PDF represents the probability distribution of waiting times between consecutive bus arrivals.
PDF describes the likelihood of observing a particular waiting time (e.g., 5 minutes, 10 minutes) between buses.

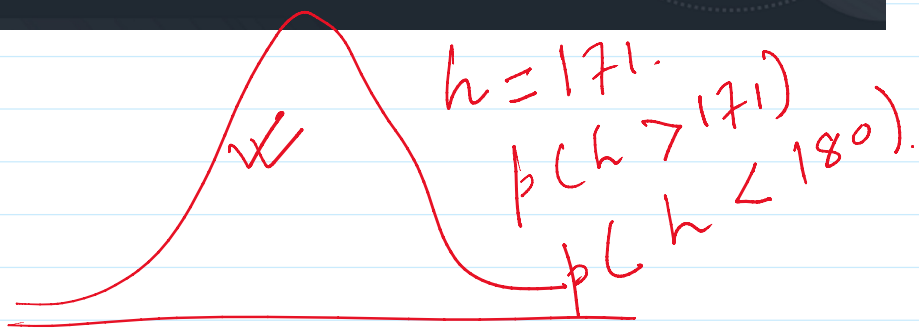

Example 3: Lifetime of Electronic Components
Suppose we are studying the lifetimes of electronic components produced by a certain manufacturer.
We find that the lifetimes of these components follow an exponential distribution with a mean lifetime of 1000 hours.
PDF represents the probability distribution of lifetimes for the electronic components.
PDF indicates the likelihood of observing a component with a specific lifetime (e.g., 800 hours, 1200 hours).

# correlation between definite integral and probability OR probability density

correlation between definite integrals and probability is observed in continuous probability distributions.

↓

area under the probability density function (PDF) curve over a specified interval represents the likelihood of a random variable falling within that interval.




## Example – for normal distribution

- **Distribution Parameters:** Mean ($\mu$), Standard Deviation ($\sigma$)
- **Objective:** Calculate the probability that a random variable $X$ lies between two values $a$ and $b$.
- **Probability Density Function (PDF) for Normal Distribution**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$ ← pdf.

- **Probability Calculation**

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

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

# Example – for a standard normal distribution

- **Distribution Parameters:** Mean ($\mu$) = 0, Standard Deviation ($\sigma$) = 1
- **Objective:** Calculate the probability that a random variable *X* lies between two values *a* and *b*.
- **Probability Density Function (PDF) for Normal Distribution**

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

- **Probability Calculation -** Let's calculate the probability of the random variable *X* lying within the interval $-1 \leq X \leq 1$ using the PDF:

$$P(-1 \leq X \leq 1) = \int_{-1}^{1} \frac{1}{\sqrt{2\pi}} e^{\frac{x^2}{2}} dx$$

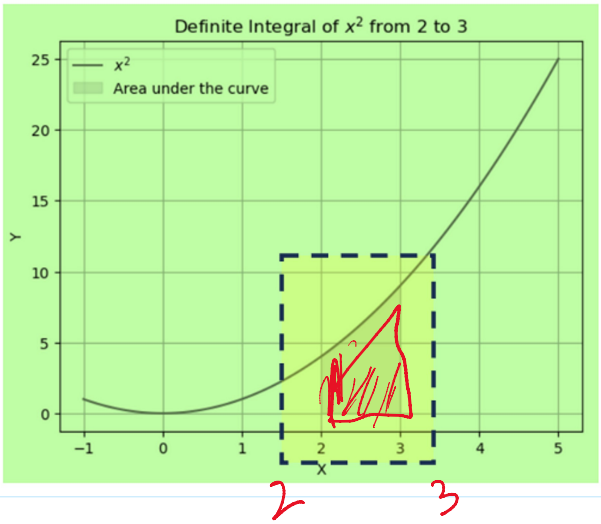
## Note

| Probability Density Function (PDF) | The PDF is denoted as f(x) and satisfies the properties |
|---|---|
| • represents the probability distribution of a continuous random variable. | • f(x)≥0 for all x within the range.<br>• total area under the PDF curve is equal to 1. |

# Example 1: Definite Integral of a Simple Function

Area under the curve between the points [2,3] gives the probability density (likelihood) that a value will fall between [2, 3]

```
# Define a simple function to integrate
def f(x):
    return x**2
```

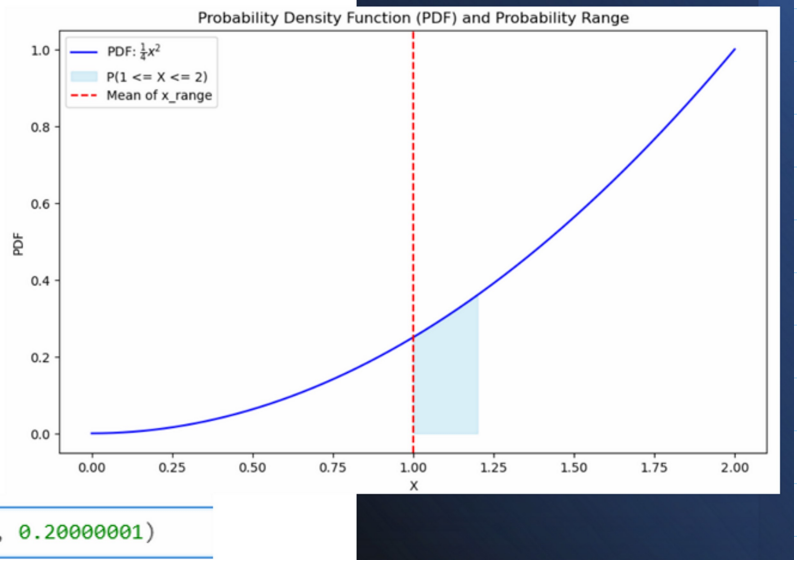



# Exercise 1: PDF Calculation

- Given the probability density function (PDF) for a continuous random variable X:

$$f(x) = \frac{1}{4}x^2, \quad \text{for } 0 \le x \le 2$$

- Calculate the probability that $X$ lies between 1 and 2 .
- Use SciPy library to compute the area under the curve




```
spi.quad(pdf_function, 0.19999, 0.20000001)
```

↳ area to you
↳ likelihood .
↳ density ≤ 4 ??

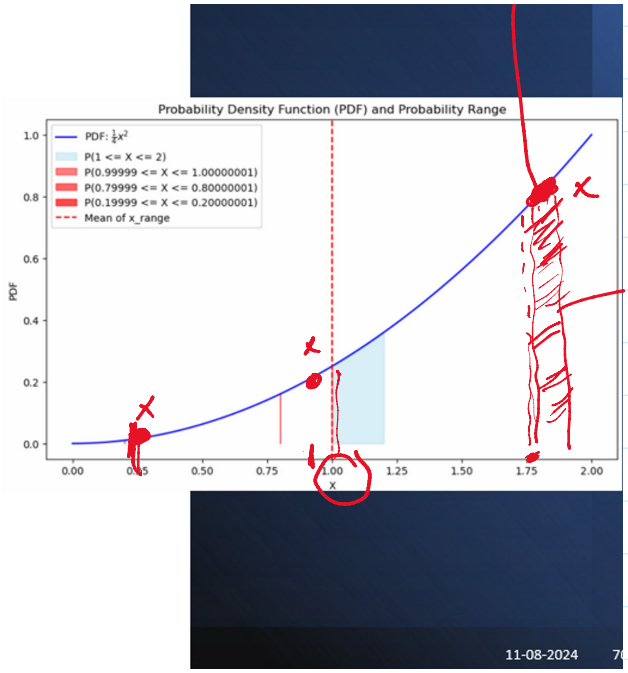
# PDFs at various points

- **Around 1 (the mean):**

point on curve

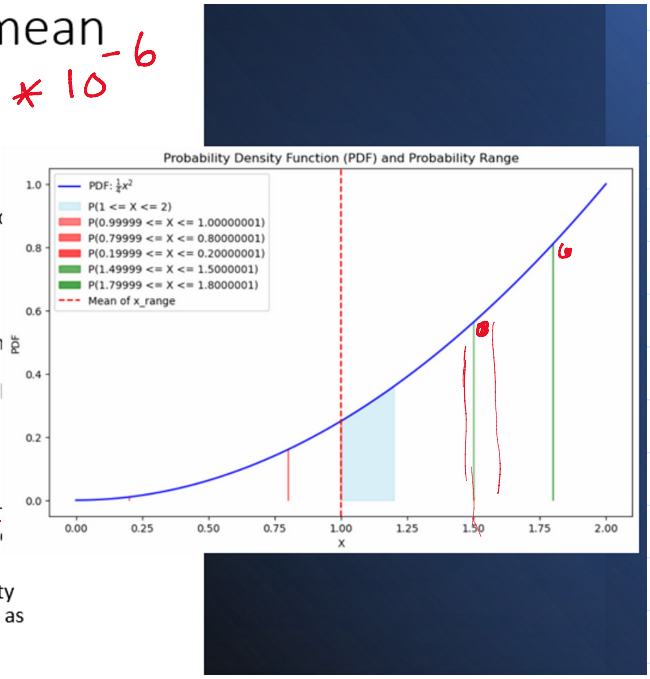
# PDFs at various points

- **Around 1 (the mean):**
  - Result: 2.502475×10−6
  - Interpretation: probability density around the mean of 1 is low within the extremely narrow interval [0.99999, 1.00000001]. While the density is small, it's not zero, indicating that there is some probability mass within this tiny range.
- **Around 0.8:**
  - Result: 1.60158×10−6
  - Interpretation: Similarly, the probability density around 0.8 is low within the narrow interval [0.79999, 0.80000001]. The small value suggests a concentration of probability density around this point.
- **Around 0.2:**
  - Result: 1.00095×10−7
  - Interpretation: The probability density around 0.2 is also low within the narrow interval [0.19999, 0.20000001]. The very small value indicates a concentration of probability density around this point.




# PDFs at points after the mean

- Around 1.5:
  - Result: 5.6812125039030244×10−06
  - Interpretation: The probability density around 1.5 is relatively higher within the narrow interval [1.49999, 1.5000001] compared to the previous examples. This suggests a concentration of probability density aroun 1.5.
- Around 1.8:
  - Result: 8.180955004683693×10−06
  - Interpretation: Similarly, the probability density aroun 1.8 is relatively higher within the narrow interval [1.79999, 1.8000001]. The larger value indicates a hig concentration of probability density around 1.8 compared to the previous examples.
- In summary
  - the probability density is still small, but it's larger thar the values we obtained for intervals around 1, 0.8, an 0.2.
  - aligns with the general expectation that the probability density is highest around the mean (1) and decreases as you move away from it.



## Cumulative density function (CDF)

The Cumulative Distribution Function (CDF) provides the cumulative probability of a random variable being less than or equal to a specific value.

CDF is denoted as $F(x)$ and is defined as the integral of the PDF up to that value.

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

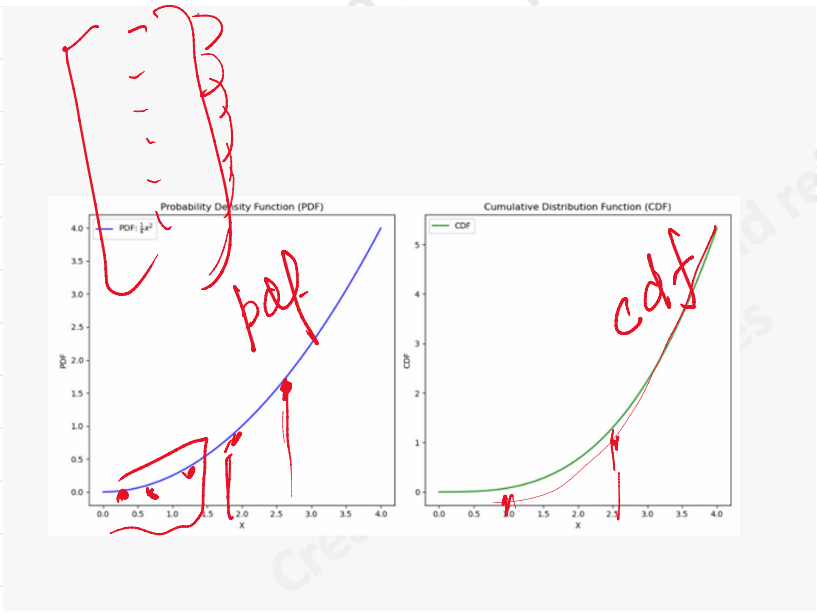
# Exercise 1: CDF Calculation

Given the probability density function (PDF) for a continuous random variable X:

$$f(x) = \frac{1}{4}x^2, \quad \text{for } 0 \le x \le 2$$

Using the same PDF as in Exercise 1, calculate the Cumulative Distribution Function (CDF) for

P(1≤X≤2).




## Results

- The blue curve represents the PDF function
- CDF plot (green) illustrates how the cumulative probability evolves across the range of x.
- It starts from 0 and approaches 1 as x increases.
- steepness of the CDF curve indicates the rate at which probability accumulates.

# Interpretation of CDF Values

- Point A (x=1):
  - CDF Value: P(X≤1)
  - Interpretation: The probability that X is less than or equal to 1.
- Point B (x=2):
  - CDF Value: P(X≤2)
  - Interpretation: The probability that X is less than or equal to 2.
- Point C (x=3):
  - CDF Value: P(X≤3)
  - Interpretation: The probability that X is less than or equal to 3.



# Percent Point Function (PPF)

The Percent Point Function (PPF), also known as the inverse cumulative distribution function (CDF),

It maps a probability value to the corresponding value of the random variable.

In other words, given a probability $p$, the PPF returns the value $x$ such that $P(X \leq x) = p$.

Use cases of ND

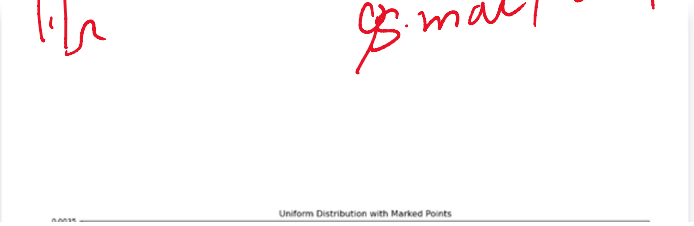
-

**Uniform distribution**

- process many (defects) → empirical
- explore (data) $x_1, x_2, x_3$ ... → flag small differences

# Example

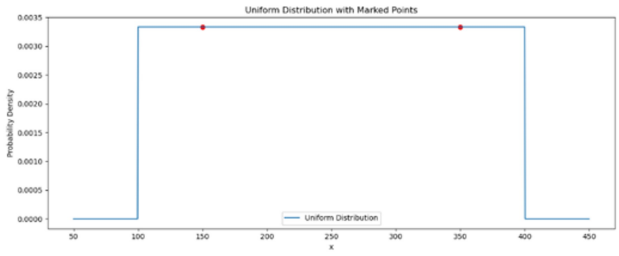
- between a and b is a horizontal line,



# Example

- between a and b is a horizontal line, indicating a constant probability density within the specified interval.
- For illustration, let's take a uniform distribution between *a*=100 and *b*=400
- The probability density is the same for all values within the range [a, b].

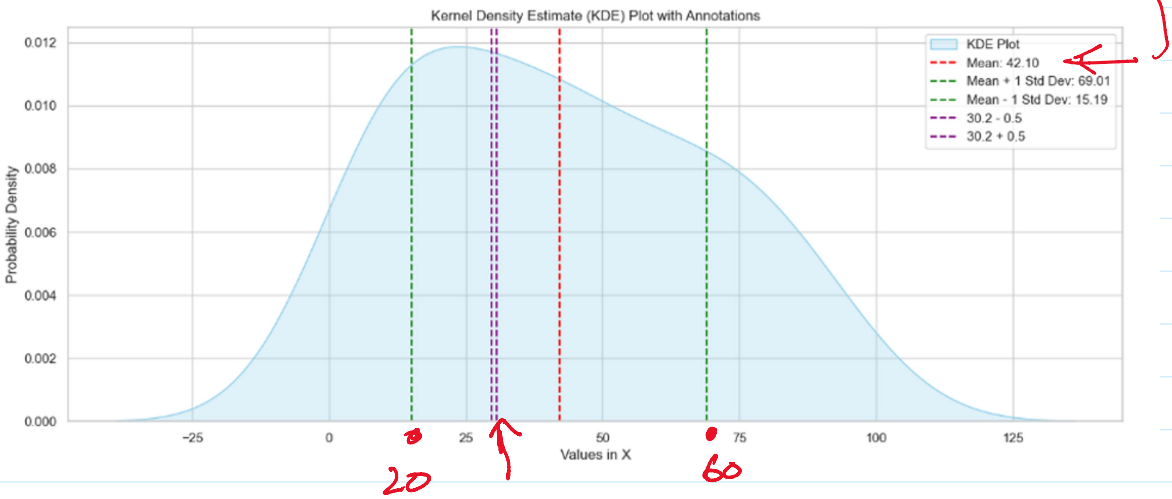
$$f(x) = \frac{1}{b-a}$$




pdf value at x1 (150): 0.0033333333333333335
pdf value at x2 (350): 0.0033333333333333335

# Explanation

- PDF(20.2) = 0.0106
- PDF(30.2) = 0.0134
- PDF(60.2) = 0.0118



# Summary

**01**

the PDF values provide a way to quantify the likelihood of the variable taking specific values in a continuous distribution.

**02**

A higher PDF value implies a higher concentration of values around that point, making it more likely, while a lower PDF value suggests a lower concentration and lower likelihood.

**03**

The interpretation is relative, comparing the likelihood of different values within the distribution.

??

$P(X = 30.2) = 13'$.

relative ??

**Poisson dist - exponential dist**

discrete continuous.

# 68-95-99.7 Rule

**Understand the empirical rule for normal distributions:**

- Approximately 68% of the data falls within one standard deviation of the mean.
- Approximately 95% falls within two standard deviations.
- Approximately 99.7% falls within three standard deviations.

**Example:**

For a standard normal distribution ($\mu$=0, $\sigma$=1),

- about 68% of the data falls between -1 and 1,
- 95% falls between -2 and 2,
- and 99.7% falls between -3 and 3.

# basis of the 68-95-99.7 Rule

also known as the Empirical Rule or the Three Sigma Rule, is a statistical guideline that describes the approximate percentage of data within certain intervals from the mean in a normal distribution.

rule is based on the properties of the normal distribution curve, which is symmetric and bell-shaped.

# Summary of 3-sigma rule

**Approximately 68% within (±1$\sigma$):**

- bulk of the data is concentrated in this central region, and the curve starts to slope down gradually as you move away from the mean.

**Approximately 95% within (±2$\sigma$):**

- wider range captures a larger proportion of the data, encompassing both the central region and the tails of the distribution.

**Approximately 99.7% within (±3$\sigma$):**

- vast majority of the data is now included, covering the central region and extending into the tails of the distribution.


- not all data in the real world perfectly follows a normal distribution.
- However, the central limit theorem states that the sampling distribution of the mean of a sufficiently large random sample from any population will tend to approximate a normal distribution, which is why the normal distribution is widely used in statistical analysis.


- **Income Distribution**: distribution of income in most countries is highly skewed, with a large portion of the population earning relatively low incomes and a smaller percentage earning very high incomes.
- **Aging Population**: age distribution of a population is typically not normally distributed. It often follows a right-skewed distribution, with more people in younger age groups and a decreasing number of individuals as age increases.
- **Product Defects**: The number of defects found in manufactured products often follows a Poisson distribution, which is not normal. The Poisson distribution is discrete and describes rare events that occur at random intervals.


## EXAMPLES OF DATA THAT ARE NOT NORMALLY DISTRIBUTED

- **Web Page Hits**: The number of hits on a website per day can often follow a skewed distribution, where a few pages receive a large number of hits, while the majority of pages receive very few.
- **Customer Transaction Amounts**: In retail or e-commerce, the distribution of transaction amounts can be right-skewed, with many small transactions and a few large ones.
- **Stock Price Returns**: While stock prices themselves may not follow a normal distribution, the returns (percentage changes) on stock prices are often not normally distributed. They can exhibit fat tails and volatility clustering.

## EXAMPLES OF DATA THAT ARE NOT NORMALLY DISTRIBUTED

- **Social Media Engagement**: The number of likes, shares, or comments on social media posts often follows a highly skewed distribution, with a few viral posts receiving a disproportionate amount of engagement.
- **Hospital Stay Duration**: The duration of hospital stays for patients can be right-skewed, with many patients having relatively short stays and a smaller number of patients requiring long hospitalizations.
- **Wait Times**: Wait times in queues, such as lines at a grocery store or traffic congestion, are often right-skewed, with most people experiencing short waits and a few people experiencing very long waits.
- **Natural Disasters**: The occurrence of natural disasters, such as earthquakes or hurricanes, follows a distribution known as the power-law distribution, which is characterized by infrequent but highly impactful events.

## should we worry too much about non normally distributed data?

The normal distribution assumption is a common assumption in many statistical methods

it's important to note that not all machine learning algorithms require normally distributed data.

## Considerations … details in the ML course

Many machine learning algorithms, such as decision trees, random forests, and k-nearest neighbors, are non-parametric and don't assume a specific distribution of the data.

Linear regression and some other parametric methods may assume normality of errors, but they can still perform reasonably well with non-normally distributed predictors if other assumptions (like homoscedasticity) are met.

Consider using robust models that are less sensitive to outliers and distributional assumptions.

Support Vector Machines and Neural Networks, for example, can often handle non-normally distributed data.

Normalizing data

$$Z = \frac{X - \mu}{\sigma}$$

$$= \frac{10 - 22}{2} = \frac{-12}{2} = -6\sigma$$

away from $\mu$

10
20
30
40
115

$\mu = 22$
$\sigma = 2$

$Z(20) = \frac{20 - 22}{2} = (-1)\sigma$ away from $\mu$

11-08-2024

# Example

- let's take a set of 10 odd data points:
  - {1,3,5,7,9,11,13,15,17,19}

- They indicate how many standard deviations each data point is from the mean of the distribution.
- Positive Z-scores indicate values above the mean, while negative Z-scores indicate values below the mean.

1. For $X = 1$:
   $Z = \frac{1-10}{5} = -1.8$
2. For $X = 3$:
   $Z = \frac{3-10}{5} = -1.4$
3. For $X = 5$:
   $Z = \frac{5-10}{5} = -0.8$
4. For $X = 7$:
   $Z = \frac{7-10}{5} = -0.6$
5. For $X = 9$:
   $Z = \frac{9-10}{5} = -0.2$
6. For $X = 11$:
   $Z = \frac{11-10}{5} = 0.2$
7. For $X = 13$:
   $Z = \frac{13-10}{5} = 0.6$
8. For $X = 15$:
   $Z = \frac{15-10}{5} = 1$
9. For $X = 17$:
   $Z = \frac{17-10}{5} = 1.4$
10. For $X = 19$:
    $Z = \frac{19-10}{5} = 1.8$

# Usage of z-scores

important

DP

**Standardization**: Z-scores are used to standardize data from different distributions into a common scale, facilitating comparison and analysis.

**Outlier Detection**: Z-scores help identify outliers, which are data points that deviate significantly from the mean of the distribution.

**Probability Calculation**: In a standard normal distribution (with a mean of 0 and standard deviation of 1), the Z-score directly corresponds to the probability of obtaining a value equal to or less than the given Z-score.