IEEE FCI - Helwan Student Branch

Statistics and visualization

Hossam Ahmed
Ziad Waleed
Mario Mamdouh

Types of Statistics

- **Statistics** deal with the **collection**, **organization**, **analysis**, **interpretation** and **presentation** of data.
- **Statistics** can be categorized into two main types.
- **Descriptive statistics** summarize and describe the dataset quantitively.
  - Mean, median, mode, range,…
- **Inferential statistics** use sample data to make generalizations, predictions, or decisions about a larger population.
  - Estimation, Hypothesis testing, Correlation, Regression.
- Inference is the use of sample of data to predict (estimate) something about the larger population, we call the process of prediction in ML inference for this reason.

Sampling

- Sampling is the process of selecting a subset of data points (called a sample) from a larger group (a population) to study and draw conclusions about the entire population.
- Sampling is used when it is impractical or impossible to analyze the entire population due to constraints like time, cost, or accessibility.
- Sampling methods are divided into two categories: **probability sampling** and **non-probability sampling.**
- **probability sampling** every member of the population has a known, non-zero chance of being selected, this ensure the sample is representative of the population.
- **Non-probability sampling** not all members have a chance of being selected.
  - *Selecting individuals who are easiest to reach or available.*

Probability Sampling

- Random sampling : each member of the population has an equal chance of being selected.
  - *Randomly selecting 50 students from a school.*
- Systematic Sampling : Selecting every ***k-th*** individual from a list after randomly choosing a starting point
  - *Surveying every **10th** customer entering a store.*
- Stratified Sampling : Dividing the population into subgroups (strata) based on specific characteristics (e.g., age, gender) and then sampling proportionally from each subgroup.
  - *Divide students into two subgroups based on age, **sample 20 student**.*
    - *10-12 years : 40 student (0.4 from the total population)*
    - *13-15 years : 60 student (0.6 from the total population)*
    - *Sample 40% from (10-12)* $40 \times \frac{20}{100} = 8\ students$
    - *Sample 60% from (13-15)* $60 \times \frac{20}{100} = 12\ students$

Sampling with replacement

- Random sampling : a sampling method where each selected item is **returned** to the population before the next draw, meaning the same item can be selected multiple times.
- Used to create random samples where each individual in the population has an equal chance of being selected each time.
- Example : population $\{A, B, C\}$ sample size 3
  - **Process:** Randomly select items, allowing duplicates.
  - Draw 1 : A
  - Draw 2 : B
  - Draw 3 : A
  - Sample : $\{A, B, A\}$

Bootstrapping

- Bootstrapping : a statistical technique that uses **sampling with replacement** from an observed dataset (treated as a proxy for the population) to estimate something.
- Commonly used in machine learning, inferential statistics, and situations with small datasets.
- Operates on a sample, treating it as if it were the population.
- Example : original sample or dataset $\{5, 10, 15, 20, 25\}$
  - sample 1 : $\{5, 15, 15, 20, 25\}$
  - sample 2 : $\{10, 10, 25, 5, 25\}$
  - sample 3 : $\{15, 20, 20, 25, 5\}$
  - sample 4 : $\{15, 15, 20, 20, 5\}$

Variables and data types

Variable

a variable is a characteristic or attribute that can take different values or categories

Qualitative

non-numerical and represent different categories or groups

Ordinal

the categories have a ***natural ordering*** or ranking to them

Nomial

where there is ***no inherent order*** or ranking to the categories

Quantitative

are ***numerical*** in nature and can be measured on a numerical scale

Discrete

values are ***countable*** and typically take on whole numbers

Continuous

values are infinite and can take on any ***value within a range***

Variables and data types

- In mathematical modeling the model expect to take numbers as inputs not strings, because you can't do mathematical operations on strings.

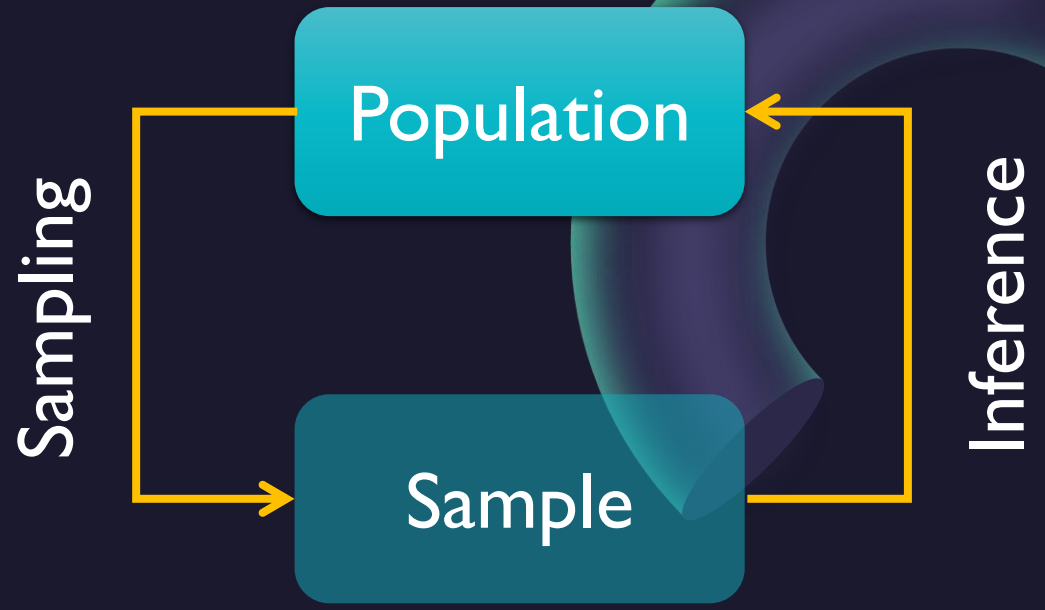
- We need to map the Qualitative (categorical) data to numbers as a preprocessing step before inputting the data to our ML model, this process is called encoding.

```
colors = ['Red', 'Blue', 'Green', 'Red', 'Green', 'Blue', 'Red']
# Create a dictionary for encoding
encoding_dict = {'Red': 1,'Blue': 2,'Green': 3}
# Encode the categorical features
encoded_data = [encoding_dict[item] for item in colors]
# Print the results
print("Original Data:", colors)
print("Encoded Data:", encoded_data)
```

```
Original Data: ['Red', 'Blue', 'Green', 'Red', 'Green', 'Blue', 'Red']
Encoded Data: [1, 2, 3, 1, 3, 2, 1]
```

Populations and Random Samples

- **Statistic :** is a characteristic or measure obtained by using the data values from a sample.
- **Parameter :** is a characteristic or measure obtained by using all the data values from a specific population.



- If you are using a statistic obtained from a sample to infer or conclude something about the whole population this is Estimation.
- Like collecting data from a sample of employees and use their average salary as an estimation of the real average of all the employees working a specific profession.

Measures of location

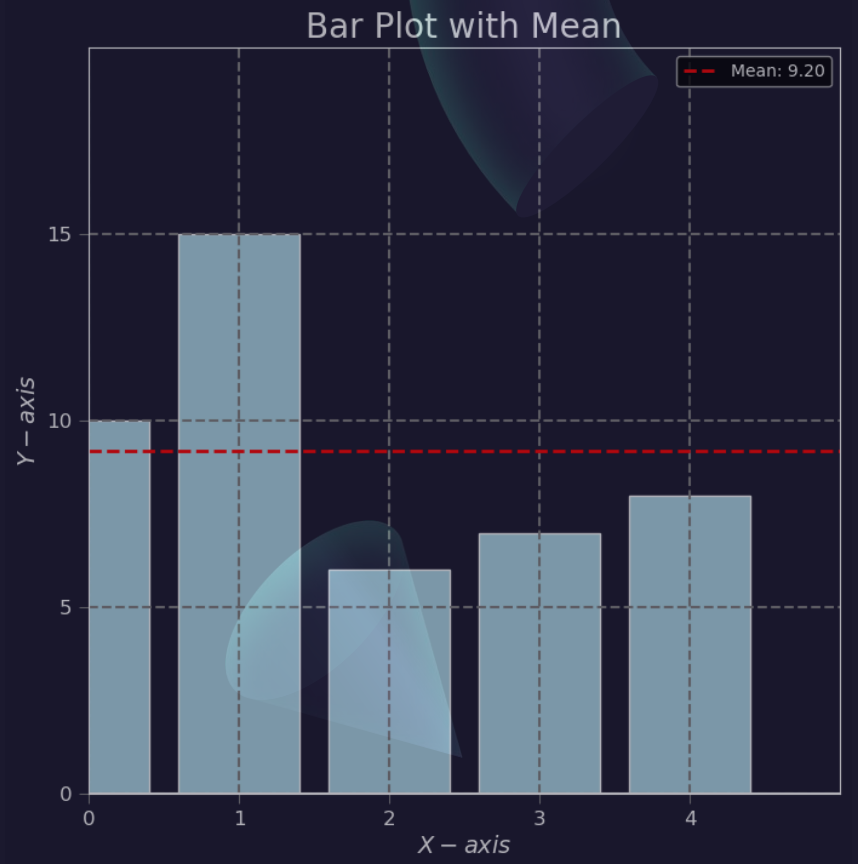
- **Mean** (arithmetic average) is found by adding the values of the data and dividing by the total number of values.
  - Sample mean (statistic) obtained from a sample denoted $\bar{X}$.
  - Population mean (parameter) obtained from the population donated μ.

- Sample mean $\bar{X} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_i^n x_i$

- Population mean $\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{1}{N}\sum_i^N x_i$

```
import numpy as np
np.mean([10, 15, 6, 7, 8])
```

9.2



Measures of location

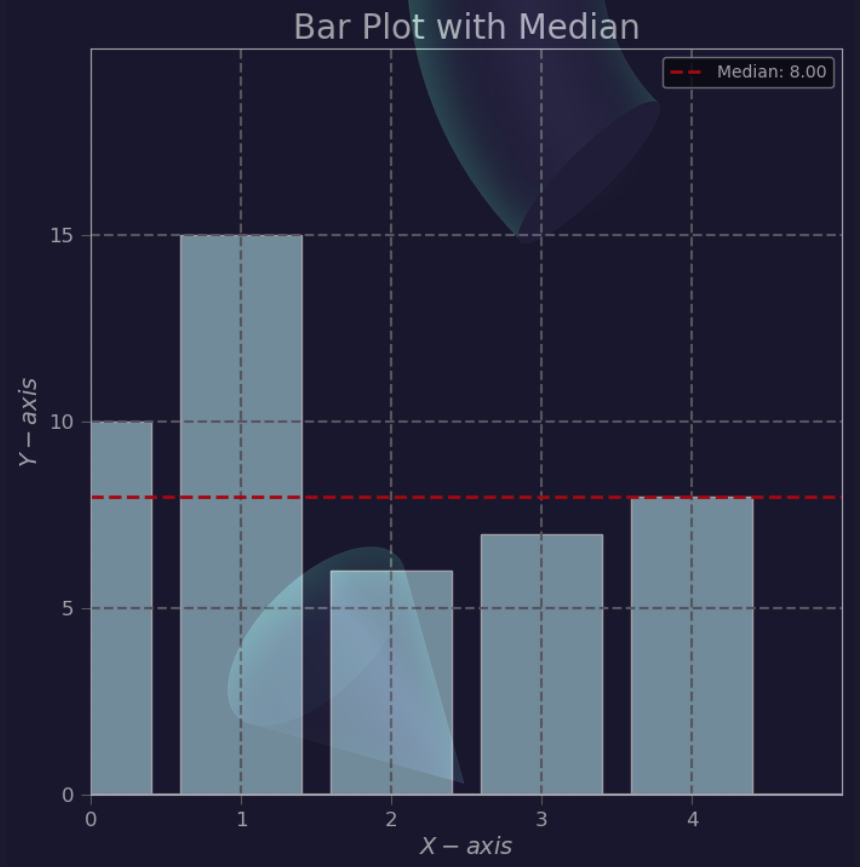
- **Median** is the mid point of the sorted array of the data.
  - Step 1 : arrange the data values.
  - Step 2 : Select the middle point, if even number of points take the average of the two middle points.

- For [10, 15, 6, 7, 8]
  - Sort [6, 7, 8, 10, 16]
  - Mid point 8

```
import numpy as np
np.median([10, 15, 6, 7, 8])
```

8.0

- For [10, 15, 6, 7, 8, 9]
  - Sort [6, 7, 8, 9, 10, 16]
  - Mid point (8+9)/2 = 8.5



Measures of location

- **Mode** the value that occur the most in a dataset.
- For example, [1,1,2, 2, 2, 2] the mode is 2.
- **The weighted mean** of a variable X by multiplying each value by its corresponding weight and dividing the sum of the products by the sum of weights.
- Weighted mean $\bar{X} = \frac{w_1x_1+w_2x_2+\cdots+w_nx_n}{w_1+w_2+\cdots+w_n} = \frac{\sum_i^n w_ix_i}{\sum_i^n w_i}$
- **The weighted mean** is used to calculate the GPA (Grade Point Average), where the hours are the weights and the grade is the value.
  - A+ : 4, A : 3.75, B+ : 3.4, B : 3.1, C+ : 2.8, C : 2.5, D+ : 2.25, D : 2, F: 1
- **Midrange** is defined as the sum of the lowest and highest values in the data set, divided by 2.

Measures of location

- The mean is affected by extremely high or low values, called outliers, and may not be the appropriate average to use in these situations.
- The median is used when it is necessary to find out whether the data values fall into the upper half or lower half of the distribution.
- The median is affected less than the mean by extremely high or extremely low values.
- The mode can be used when the data are nominal or categorical, such as religious preference or gender.
- A data set can have more than one mode, or the mode may not exist for a data set.
- The midrange is affected by extremely high or low values in a data set.

Measures of variation

- **Measures of Variation** describe how spread out or scattered data points are in a dataset.

- **Range** is the difference between the highest and lowest values in the dataset.
  - $Range = Max\ value\ - Min\ Value$

- **Variance** is the average of the squares of the distance each value is from the mean.
  - $\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}$ population variance
  - $s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$ sample variance, using $n - 1$ to reduce the bias resulted from the use of the sample mean $\bar{x}$ which reduce the apparent spread because it's closer to the sample points than the true mean μ (Bessel's correction).

What are Degrees of Freedom (DOF)?

- **Degrees of freedom (DOF)** refer to the number of independent values that **can vary** when calculating a statistic.

- If we know n data points we can freely choose n values.
- But once we computed the sample mean $\bar{x}$, one of the datapoints is no longer free to vary because the sum of all values must match the mean.
  - $\bar{x} = \frac{1}{n}\sum_i^n x_i$
  - Multiply both by n so we can express one data point in terms of the others:
  - $n\bar{x} = n\frac{1}{n}\sum_i^n x_i$
  - $n\bar{x} = \sum_i^n x_i$
  - $x_n = n\bar{x} - \sum_i^n x_i = 0$
  - $x_n = n\bar{x} - (x_1 + x_2 + \cdots + x_n) = 0$
- If we know the first $n - 1$ points the last one is determined automatically. So, instead of having n fully independent numbers, we only have $n - 1$ independent choices.

What are Degrees of Freedom (DOF)?

- Suppose you are asked to pick **5 numbers** randomly. You have full freedom to choose all 5.

- However, if I tell you that these 5 numbers **must have a specific mean** (say, 10), suddenly, you **don't have full freedom** anymore.
  - You can freely choose the **first 4 numbers**
  - But the **5th number is already determined** because the sum must add up to match the given mean.

- When we use $\bar{x}$ sample mean to estimate μ the population mean, we can freely choose $n-1$ points but the last point is just automatically calculated, so that $\sum_{i=1}^{n}(x_i - \bar{x})$ the sum of deviations must equal zero, using $\bar{x}$ in the sample variance tends to underestimate the true variance and to solve this we divide by the DOF.

What are Degrees of Freedom (DOF)?

- Let's generate a large population of 100,000 integers between 0 and 100.

```
import numpy as np
np.random.seed(42)
population = np.random.randint(0, 100, size=100000)
```

- Compute the true population variance μ.

```
true_variance = np.var(population, ddof=0)
```

- Take 1000 random samples of size 30 and compute sample variance

```
sample_variances_n = \
[np.var(np.random.choice(population, 30, replace=False), ddof=0) for _ in range(1000)]
sample_variances_n_minus_1 = \
[np.var(np.random.choice(population, 30, replace=False), ddof=1) for _ in range(1000)]
```

What are Degrees of Freedom (DOF)?

- Let's check the results.

```
# Compare the means
print(f"True Population Variance: {true_variance:.4f}")
print(f"Mean Sample Variance (dividing by n): {np.mean(sample_variances_n):.4f}")
print(f"Mean Sample Variance (dividing by n-1): {np.mean(sample_variances_n_minus_1):.4f}")
```

True Population Variance: 838.1360

Mean Sample Variance (dividing by n): 801.4821 (underestimate)

Mean Sample Variance (dividing by n-1): 840.4902

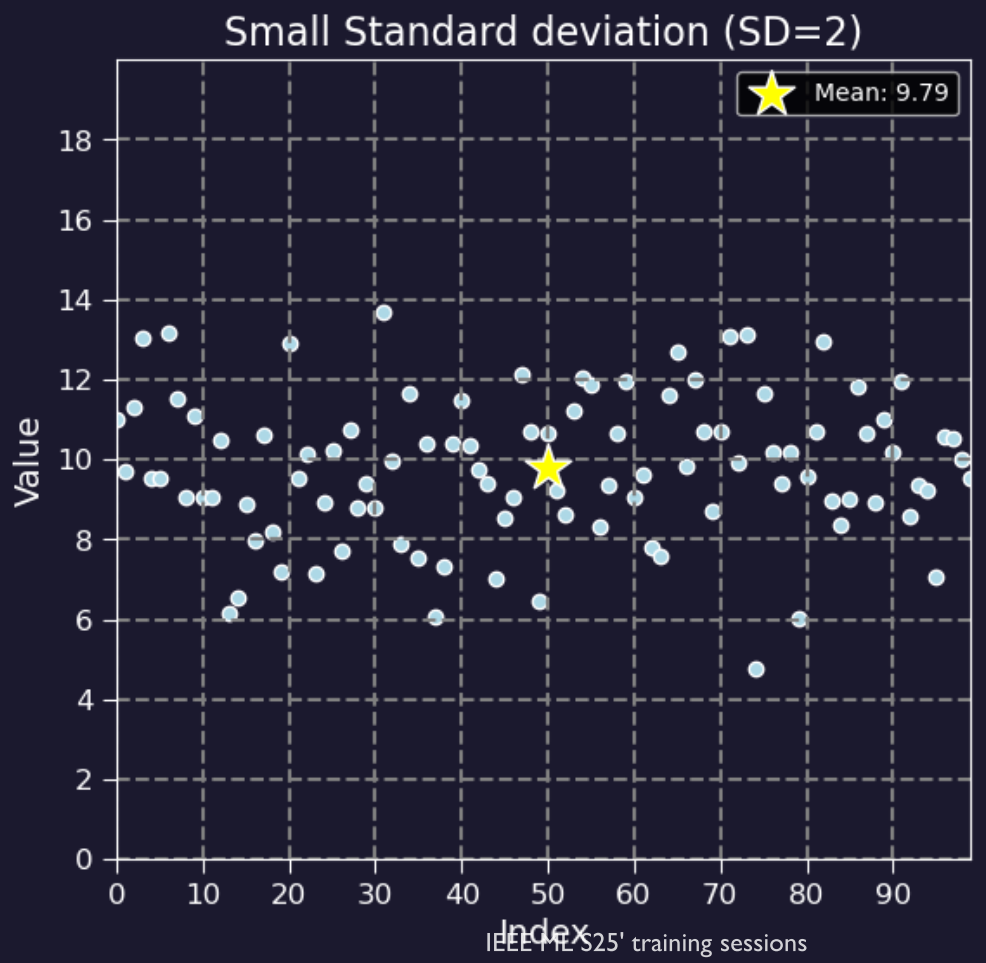
- So, by dividing on $n-1$ in the sample variance equation we reduced the underestimation bias resulted from the use of $\bar{x}$ sample mean that limit our DOF to be $n-1$.

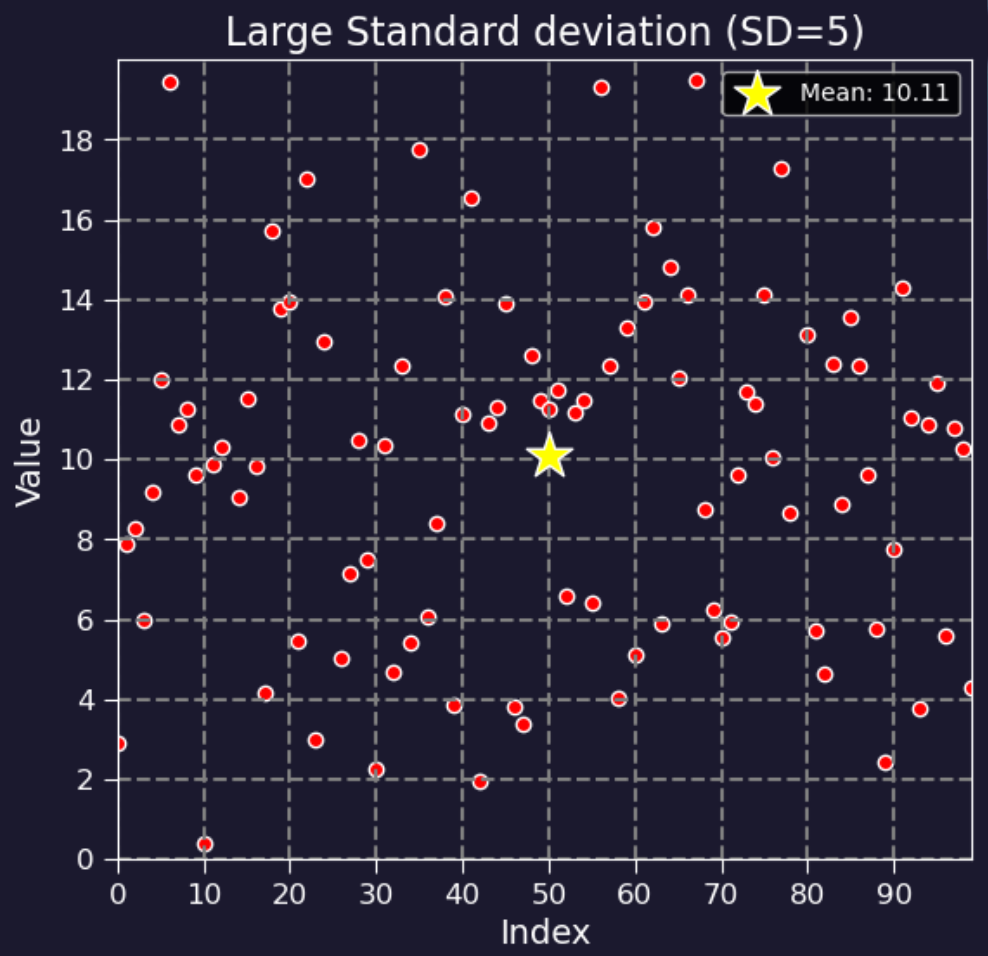
Measures of variation

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

$s = \sqrt{s^2}$ sample

- **Standard deviation** is the square root of variance, representing data spread in the same units as the original data.





Measures of variation

- **Coefficient of Variation** is the standard deviation divided by the mean.
  - The result is expressed as a percentage.
  - $CV = \left(\frac{\sigma}{\mu}\right) \times 100$
  - $CV = \left(\frac{s}{\bar{x}}\right) \times 100$

- If two datasets have different units (e.g., income in dollars vs. weight in kg), CV helps compare their variability in a **scale-independent** way.
  - It tells us how spread out the data is relative to its mean.
  - **Lower CV** means **less variation**, and **higher CV** means **more variation** compared to the mean.

Measures of variation

- The average salary in country X is $50,000 with a SD of $5000, while in country Y, the average salary is $20,000 with a SD of $4000.

- $CV_X = \frac{5000}{50000} \times 100 = 10\%$

- $CV_Y = \frac{4000}{20000} \times 100 = 20\%$

- Even though Country X has a higher absolute salary difference ($5,000 vs. $4,000), the CV shows that salaries in Country Y are relatively more dispersed compared to their mean.

- **Higher CV (20%)** means **greater income inequality** in Country Y than in Country X.

Measures of variation

- The mean of the number of sales of cars over a 3-month period is 87, and the standard deviation is 5. The mean of the commissions is $5225, and the standard deviation is $773. Compare the variations of the two

- $CV_{sales} = \frac{5}{87} \times 100 = 5.7\%$
- $CV_{commissions} = \frac{773}{5225} \times 100 = 14.8\%$

- Since the coefficient of variation is larger for commissions, the commissions are more variable than the sales.

Measures of position

- **Measures of position** are statistical tools that describe where a specific value falls within a dataset. They help compare individual data points to the rest of the distribution.

- **Percentiles** divide a dataset into **100 equal parts**.
  - The $k-th$ percentile P_K is the value below which $k\%$ of the data falls.
  - If your exam score is in the **90th percentile**, it means you scored **higher than 90%** of the students.
  - $P_k = \frac{k}{100} \times (n+1)$, where n the total number of observation.

- Consider these exam scores sorted in ascending order :
  - [50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]
  - let's find the **70th percentile** $P_{70} = \frac{70}{100} * (11+1) = 8.4$
  - This means P_{70} lies between the **8th** and **9th** scores, and **8.4 nearest to the 8th element so it would be 85**

Measures of position

- **Quartiles** divide the dataset into four equal parts:
  - Q1 the 25th percentile (lower Quartile)
  - Q2 the 50th percentile (Median)
  - Q3 the 75th percentile (upper Quartile)

- Using the same points **[50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100]**
  - $Q_1 = P_{25} = \frac{25}{100} * (11 + 1) = 3$, Q1 is 60
  - $Q_2 = P_{50} = \frac{50}{100} * (11 + 1) = 6$, Q1 is 75 (median)
  - $Q_3 = P_{75} = \frac{75}{100} * (11 + 1) = 9$, Q1 is 90

- Another way to calculate them, is to sort the numbers in ascending order, then find the median Q2, and use the median to split the data into two parts, find the median of each part to find Q1 from the first part and Q3 from the second part.

Measures of position

```
q1 = np.percentile(data, 25, method="nearest")
q2 = np.percentile(data, 50,  method="nearest")
q3 = np.percentile(data, 75,  method="nearest")

print(f"First Quartile (Q1): {q1}")
print(f"Second Quartile (Q2): {q2}")
print(f"Third Quartile (Q3): {q3}")
```

First Quartile (Q1): 60

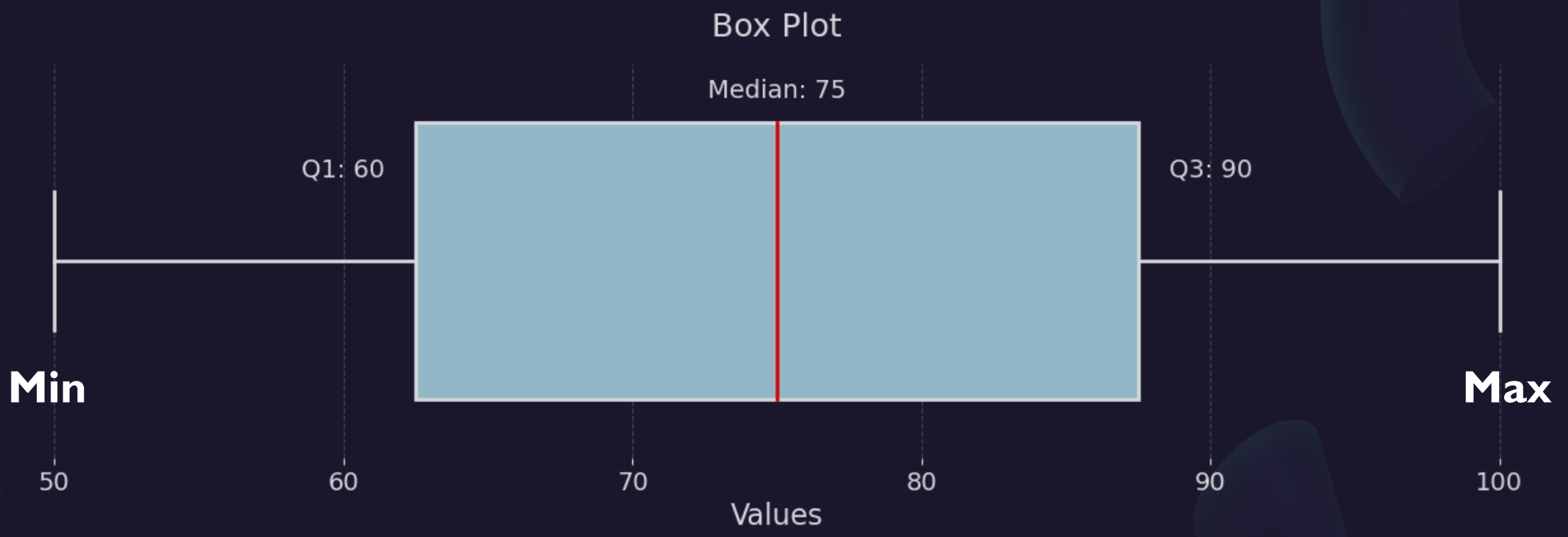
Second Quartile (Q2): 75

Third Quartile (Q3): 90

- There are other ways to calculate the percentiles, you can specify them using method parameter.
- The “linear” is the default for this NumPy function, but we used “nearest” to select an exact value from the points we have that’s nearest to the percentile rank.
- The “linear” method provide a smooth estimation (continuous data).

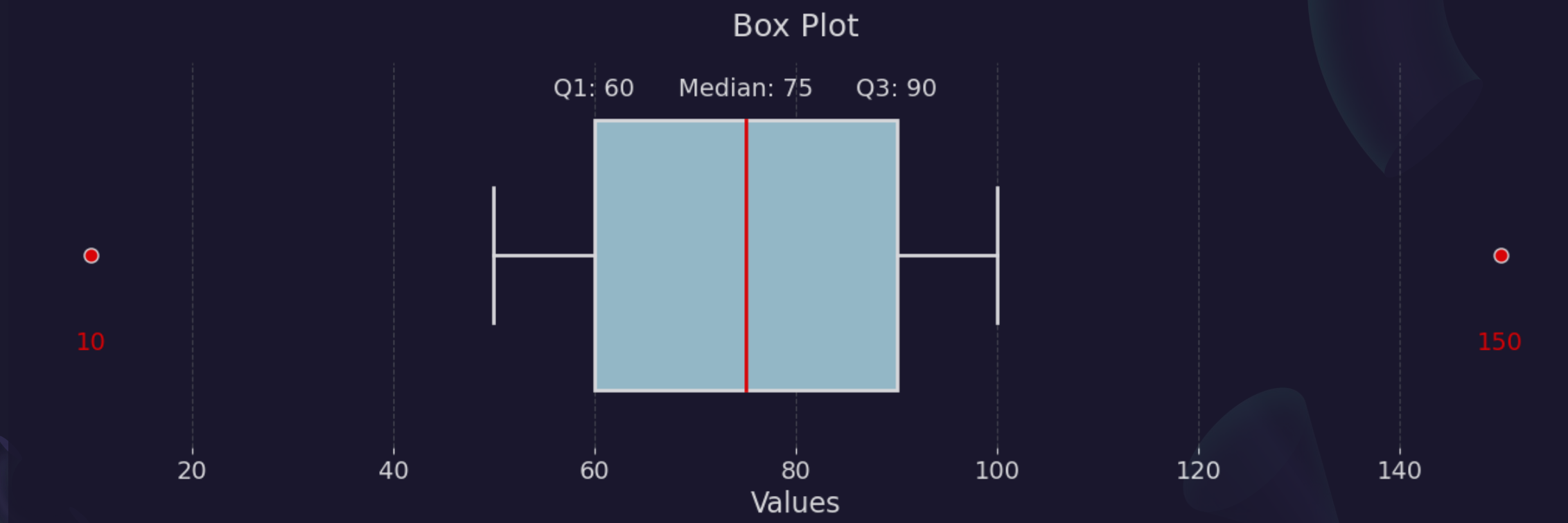
Box plots

- A **box plot** displays the five-number summary of a set of data. The five-number summary is the minimum, first quartile, median, third quartile, and maximum.



Box plots and outliers

- An **outlier** is a data point that significantly deviates from the rest of the dataset. It is an **unusually high** **or** **low value** compared to the general distribution of the data.
  - For these data points [50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 150, 10]

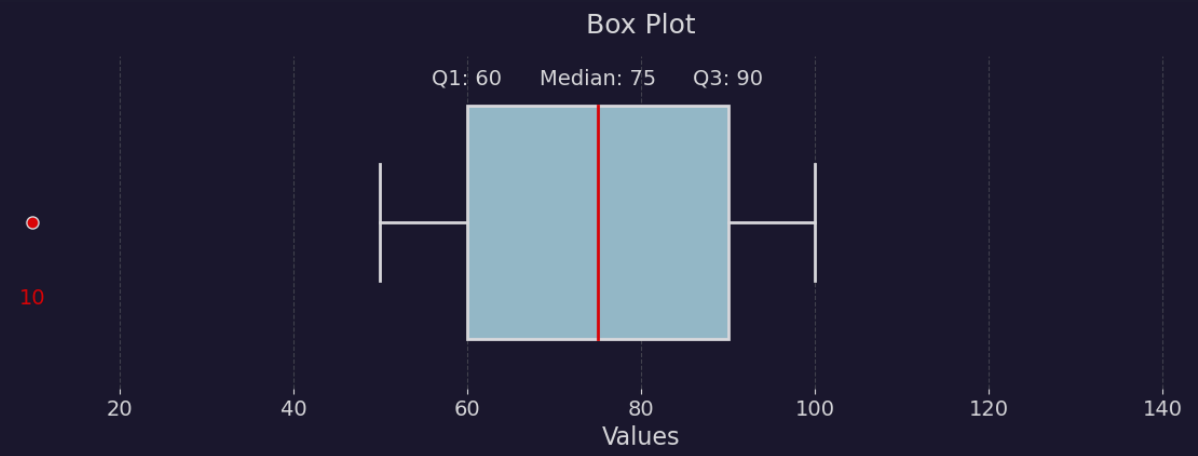


Any point beyond the fences (outer lines) would be considered an outlier.

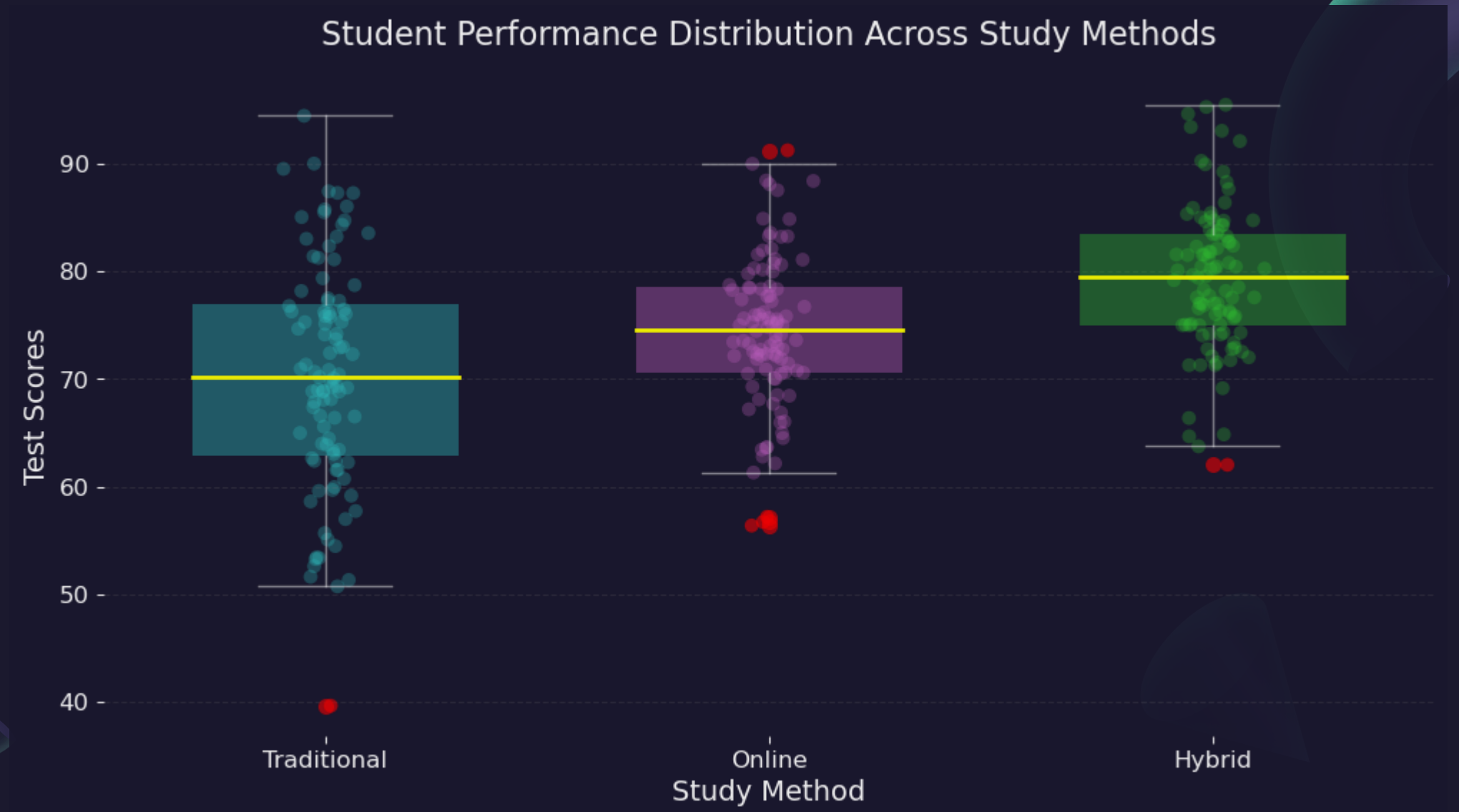
Box plots and outliers

- The box itself represent the range in which 50% of the data lies in, this lies between Q1 and Q3 and often referred to as **Interquartile range (IQR)**
  - ***Notice that the min/max that were previously shown without outliers.***
  - $IQR = Q3 - Q1$

- Outliers are data points that lies far beyond the typical range of dataset (IQR)
  - Upper bound $Q3 + 1.5\,IQR$ (minimum non-outlier value)
  - Lower bound $Q1 - 1.5\,IQR$ (maximum non-outlier value)
  - Any data points outside this range are considered **outliers**.



Box plots and outliers



Measures of position

- Standard scores (z-score) measures how far a data point from the mean in term of standard deviations.
  - $Z = \frac{x-\mu}{\sigma}$
  - $Z = 0$ the value is exactly the mean
  - $Z < 0$ the value is below the mean
  - $Z > 0$ the value is above the mean
  - $|Z| = 2$ potential outlier
- Z-scores also used in standardization, which convert data from different scales into a common scale.
- If a student scores **1400** on the SAT and the mean SAT score is **1100** with a standard deviation of **200**, we calculate:
- $Z = \frac{x-\mu}{\sigma} = \frac{1400-1100}{200} = \frac{300}{200} = 1.5$ the student scored 1.5 standard deviation above the mean.

Measures of position

- Standardization is a technique used in statistics and machine learning to **scale data** so that it has a **mean of 0** and a **standard deviation of 1**. This transformation ensures that data from different sources or distributions can be compared on the same scale.

```python
from sklearn.preprocessing import StandardScaler
import numpy as np

# Sample data (heights in cm)
data = np.array([[150], [160], [170], [180], [190]])

# Standardize data
scaler = StandardScaler()
standardized_data = scaler.fit_transform(data)

print(standardized_data)
```

```
[[-1.41421356]
 [-0.70710678]
 [ 0.        ]
 [ 0.70710678]
 [ 1.41421356]]
```

Probability

- The probability of an event is a number between 0 and 1,the larger the probability, the more likely an event is to occur.
  - $\boldsymbol{P(A) = \frac{Number\ of\ favorable\ outcomes}{Total\ number\ of\ outcomes}}$

- Two events are **mutually exclusive events** if they cannot occur at the same time (i.e., they have no outcomes in common).
  - Single die 🎲 rolled
    - Event A rolling even number, $A = \{2,4,6\}$
    - Event B rolling odd number, $B = \{1,3,5\}$
    - Since rolling a die can't be both Odd and Even in the same time, event A and B are mutually exclusive.
    - That mean $\boldsymbol{P(A \cap B) = 0}$ and $\boldsymbol{P(A \cup B) = P(A) + P(B) = 1}$ this is true for any set of mutually exclusive events.

Probability multiplication rule

- Two events A and B are **independent** if the occurrence of one does **not** affect the other.
  - $\boldsymbol{P(A \cap B) = P(A) \times P(B)}$ works only for independent events.
  - Event A : tossing a coin 🪙 and getting "head", $\mathrm{P}(A) = \frac{1}{2}$
  - Event B : rolling a die 🎲 and getting "4", $P(B) = \frac{1}{6}$
  - Probability of both happening together
    - $\boldsymbol{P(A \cap B) = P(A) \times P(B) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}}$

- Drawing cards ♠️ ♥️ ♦️ ♣️ with replacement
  - Event A : Drawing a king $\mathrm{P}(A) = \frac{4}{52} = \frac{1}{13}$ (return the card to the deck)
  - Event B : Drawing another king $\mathrm{P}(B) = \frac{4}{52} = \frac{1}{13}$
  - $\boldsymbol{P(A \cap B) = P(A) \times P(B) = \frac{1}{13} * \frac{1}{13} = \frac{1}{169} = 0.0059}$

Probability multiplication rule

- If events A and B **dependent,** the probability of B changes if A has already occurred.
  - To attend a session the link of the meeting should happened first else no session.
  - $\boldsymbol{P(A \cap B) = P(A) \times P(B|A)}$ given that A event happened.

- Drawing cards ♠ ♥ ♦ ♣ without replacement
  - Event A : Drawing an Ace first $\mathrm{P}(A) = \frac{4}{52} = \frac{1}{13}$
  - Event B : Drawing a king second $\mathrm{P}(B) = \frac{4}{51}$
  - $\boldsymbol{P(A \cap B) = P(A) \times P(B) = \frac{1}{13} * \frac{4}{51} = \frac{1}{169} = 0.0060}$

- Conditional probability is the probability of an event A occurring given that B has already occurred.
- $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}}$

Conditional probability

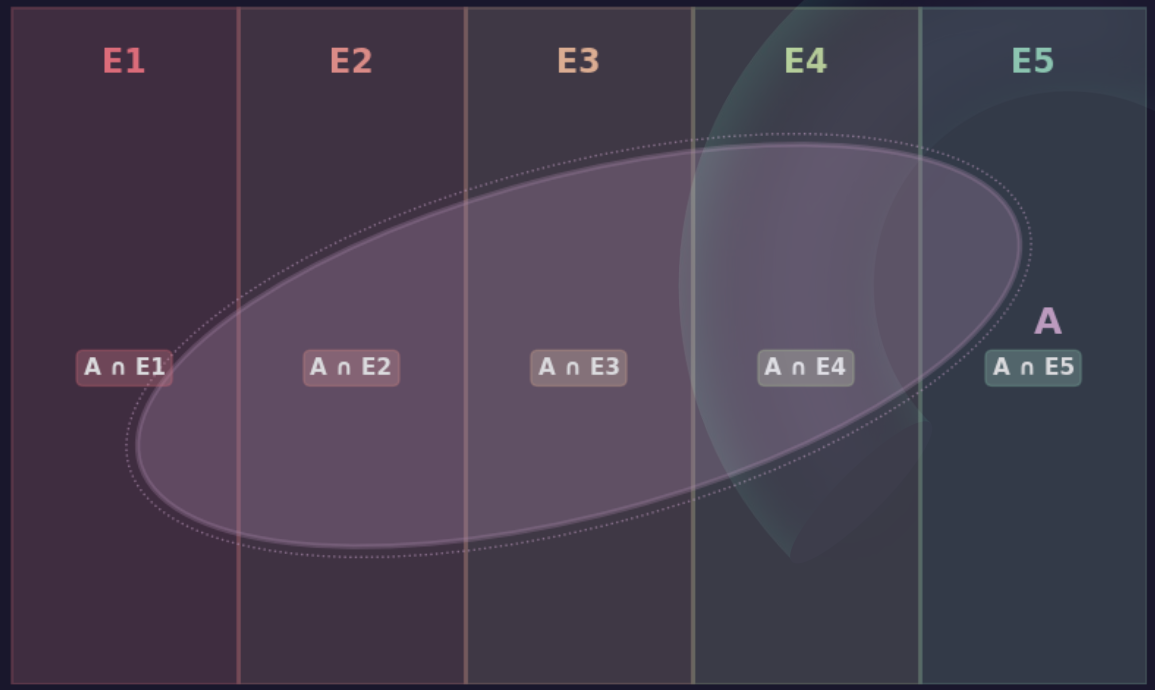
- Conditional probability is the probability of an event A occurring given that B has already occurred $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}}$

- Event A :A student scores above 80%
- Event B :The student solved 3 tasks at least
  - If $P(A \cap B) = 0.4$ (40% of the students scored above 80% and solved at least 3 tasks) and $P(B) = 0.6$ (60% solved at least 3 tasks)
  - $\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.4}{0.6} = 0.67}$ (67% chance to get 80% if you solved at least 3 tasks)

- If $A_1, A_2, \dots, A_k$ are mutually exclusive events then
  - $P(A_1, A_2, \dots, A_k|B) = P(A_1|B) + P(A_2|B) + \cdots + P(A_k|B)$

Bayes' Theorem

- Let the events $E_1, E_2, \ldots, E_k$ constitute a portion of the sample space S, and this set of events are mutually exclusive and exhaustive $S = E_1 \cup E_2 \cup \cdots \cup E_k$
- A is an event that can be written as the union of k mutually exclusive events,
- $\boldsymbol{A = (A \cap E_1) \cup (A \cap E_2) \cup \cdots \cup (A \cap E_k)}$



- Then $P(A) = \boldsymbol{P(A \cap E_1) \cup P(A \cap E_2) \cup \cdots \cup P(A \cap E_k) = \sum_{i=1}^{k} P(A \cap E_i)}$
- $\boldsymbol{\sum_{i=1}^{k} P(A \cap E_i) = \sum_{i=1}^{k} P(E_i) \times P(A|E_i)}$
- $\boldsymbol{P(E_k|A) = \frac{P(E_k \cap A)}{P(A)} = \frac{P(E_k) \times P(A|E_k)}{\sum_{i=1}^{k} P(E_i) \times P(A|E_i)}}$

$$\boldsymbol{P(A \cap B) = P(A) \times P(B|A)}$$

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)}}$$

Bayes' Theorem

- Bayes' theorem gives a mathematical rule for inverting conditional probabilities, allowing one to find probabilities of a cause given its effect.

- Bayes' Theorem is a way to **update** what we believe about something when we get new information.
  - You have an initial belief (called the prior probability) $P(E_k)$.
  - You get **new evidence** (something happens, or new data comes in) $P(A|E_k)$.

- You have a deck of cards with **4 red cards** and **6 black cards** (total **10 cards**).
  - You randomly pick one card, but you don't see it.
  - Your friend looks at it and tells you:
    - "The card is a face card" (King, Queen, or Jack).
  - Now, you wonder: **What's the chance the card is red given that it's a face card?**

Bayes' Theorem

- **R = "Card is red"** , **B="Card is Black"**, F="Card is a face card"
- You want to find $P(R|F)$ the probability that the card is red **given** that it is a face card.
- **Red Cards:** 4 total → 2 are face cards (King ♦, Queen ♦)
- **Black Cards:** 6 total → 3 are face cards (King ♠, Queen ♠, Jack ♠)
- $P(R) = \frac{4}{10}$, $P(B) = \frac{6}{10}$, $P(F|R) = \frac{2}{4}$ (If it's red, 2 out of 4 are face cards), $P(F|B) = \frac{3}{6}$ (If it's black, 3 out of 6 are face cards).

$$P(E_k|A) = \frac{P(E_k \cap A)}{P(A)} = \frac{P(E_k) \times P(A|E_k)}{\sum_{i=1}^{k} P(E_i) \times P(A|E_i)}$$

- $$P(R|F) = \frac{P(R \cap F)}{P(F)} = \frac{P(R) \times P(F|R)}{P(R)P(F|R) + P(B)\,P(F|B)} = \frac{\left(\frac{4}{10} \times \frac{2}{4}\right)}{\left(\frac{4}{10} \times \frac{2}{4}\right) + \left(\frac{6}{10} \times \frac{3}{6}\right)} = 0.4$$

Bayes' Theorem

- Now, suppose your friend **gives you another clue**:
  - "The card is a King." 👑
  - Previous information $P(R|F) = 0.4$ that the card is red

- Now, we update our probability to find $P(R|F,K)$ the probability that the card is **Red**, given that it is a **Face Card and** a **King** 👑**.**

- From the **face cards**, we now focus only on **Kings** 👑:
  - Red Kings: 1 (King ♦️)
  - Black Kings: 1 (King ♠️)
  - The total number of possible cards **shrinks** to only these **2 Kings out of 10 cards**.

- $$P(R|F,K) = \frac{P(R \cap F \cap K)}{P(F \cap K)} = \frac{P(R)P(F,K|R)}{P(F,K)} = \frac{\left(\frac{4}{10} \times \frac{1}{4}\right)}{\left(\frac{2}{10}\right)} = 0.5$$

Discrete distributions

- A **probability distribution** is the mathematical function that gives the probabilities of occurrence of possible outcomes for an experiment.
- A **discrete distribution** is a probability distribution that describes the probability of occurrence of each possible value of a discrete random variable.
- A **discrete random variable** is one that can take on only a finite or countably infinite set of distinct values.
- Examples include the outcome of a dice roll 🎲, the number of heads 🪙 in a series of coin flips, or the number of students 👩‍🎓🏫 in a class.
- The sum of the probabilities of all the events in the sample space must equal 1, and probability of each event in the sample space must be between or equal to 0 and 1.

Discrete distributions and PMF

- Let X be a random variable representing the outcome of rolling a fair 6-sided die. The possible values of X are:
  - $X \in \{1,2,3,4,5,6\}$, $P(X = x) = \frac{1}{6}$ for $x = \{1,2,3,4,5,6\}$
  - This forms a **discrete uniform distribution**, where every outcome has the same probability.

- Probability Mass Function (PMF) gives you the probability of each discrete value.

- PMF $P(X = x) = f(x)$ of a discrete random variable X is a function that satisfy
  - $P(X = x) = f(x) > 0, x \in S$ (non-negative)
  - $\sum_{x \in S} f(x) = 1$ (total probability is 1)
  - $P(X \in A) = \sum_{x \in A} f(x)$
    - The probability of any event A is the sum of the PMF values for all x in A
    - $P(X \in \{1,2,3\}) = P(X = 1) + P(X = 2) + P(X = 3)$

Discrete distributions and PMF

- Determine whether each distribution is a probability distribution

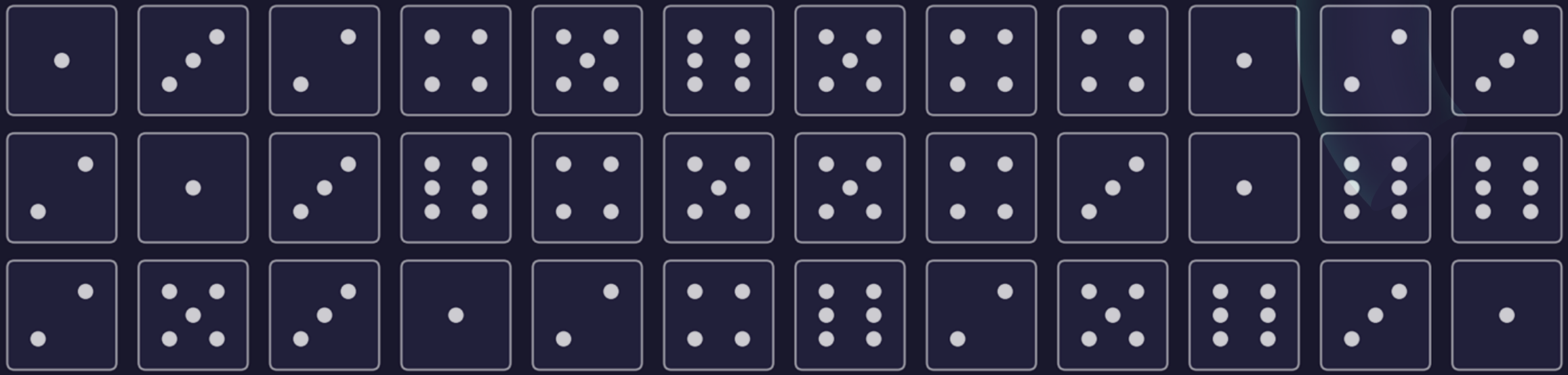
X	4	6	8	10
$P(X)$	-0.6	0.2	0.7	1.5

X	8	9	12
$P(X)$	$\frac{2}{3}$	$\frac{1}{6}$	$\frac{1}{6}$

X	1	2	3	4
$P(X)$	0.25	0.25	0.25	0.25

Mathematical expectation

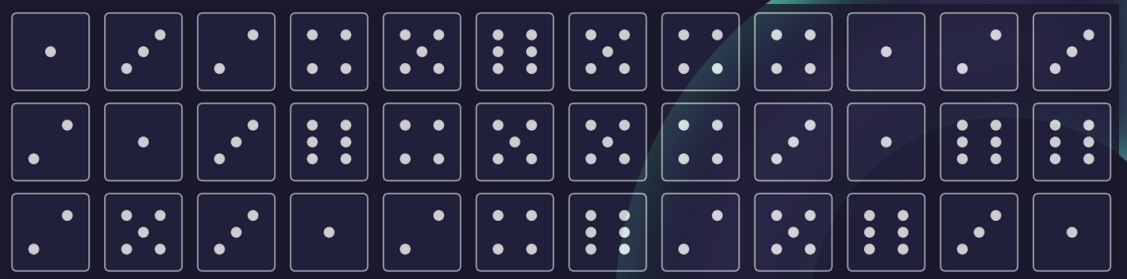
- Toss a die 🎲 many times, in the long run, what would the average (mean) of these tosses be?



- There are 12*3 = 36 tosses, let' calculate the average, by summing all the numbers over the count of tosses.

Mathematical expectation

- $\text{Mean} = \frac{1+3+2+4+\cdots+5+6+3+1}{36}$



$$\text{Mean} = \frac{(1+1+\cdots+1)+(2+2+\cdots 2)+\cdots(6+6+\cdots+6)}{36}$$

$$\text{Mean} = 1\left(\frac{6}{36}\right)+2\left(\frac{6}{36}\right)+\cdots+6\left(\frac{6}{36}\right)$$

$$\text{Mean} = 1\left(\frac{1}{6}\right)+2\left(\frac{1}{6}\right)+\cdots+6\left(\frac{1}{6}\right)$$

$$\text{Mean} = 1.P(X=1)+2.P(X=2)+\cdots+6.P(X=6) = \sum u(x)f(x) = E[u(x)]$$

Mathematical expectation

- The expected value (also called the mathematical expectation or mean) of a discrete random variable X is a measure of the "center" of its probability distribution.
- It represents the average value of X if we were to repeat an experiment infinitely many times.
- The expected value is a **weighted average** of all possible values of X, where each value is weighted by its probability.
- It's a theoretical mean and the value we got doesn't have to be from the data points we have, for example rolling a six-sided die: $E[X] = 3.5$, meaning over thousands of rolls, the average outcome per roll will approach 3.5.

Mathematical expectation

- The concept of expected value is fundamental in AI and machine learning, especially in decision-making, reinforcement learning, probabilistic models, and loss functions.

- AI agent plays a game where it can take **3 actions** with different rewards

Action	Reward R	Probability P
Jump 🦘	10	0.3
Run 🏃	5	0.5
Stop 🛑	2	0.2

The expected reward is $E[R] = \sum P_i R_i = (0.3 \times 10) + (0.5 \times 5) + (0.2 \times 2) = 5.9$

- The AI expects to get **5.9 reward points** on average if it picks actions randomly.

Mathematical expectation

- We can then select the action that maximize the expected reward value.
  - $\boldsymbol{E[jump\ 🦘] = 10 * 0.3 = 3}$ ✅
  - $E[run\ 🏃] = 5 * 0.5 = 2.5$
  - $E[stop\ 🛑] = 2 * 0.2 = 0.4$

- Can be used in risk analysis let's prove gambling is bad financially (and ethically of course) :
  - Lottery ticket costs \$5, and the prizes are
    - 1% chance to win \$100
    - 10% chance to win \$10
    - 89% chance to win nothing
    - $E[X] - 5\ 💸 = (0.01 * 100) + (0.10 * 10) + (0.89 * 0) - 5 = -3$
    - On average, each ticket loses \$3 💸 —so it's not a good investment.

Mathematical expectation

- Properties of Expectation
  - $E[c] = c$ if c is a constant, $E[5] = 5$
  - $E[cu(x)] = cE[u(x)]$, c is a constant and $u(x)$ is a function
  - $E[c_1u_1(x) + c_2u_2(x)] = c_1E[u_1(x)] + c_2E[u_{2(x)}]$

X	0	1	2	3
$P(X)$	0.2	0.1	0.4	0.3

- $E[X] = (0.2 * 0) + (0.1 * 1) + (0.4 * 2) + (3 * 0.3) = 1.8$
- $E[X^2] = \sum x^2 . P(X = x) = (0.2 * 0^2) + (0.1 * 1^2) + (0.4 * 2^2) + (0.3 * 3^2) = 4.4$
- $E[4X^2] = 4 * 4.4 = 17.6$
- $E[3X + 2X^2] = 3E[X] + 2E[X^2] = 3 * 1.8 + 2 * 4.4 = 14.2$

Mathematical expectation

- Properties of Expectation
  - $E[c] = c$ if c is a constant, $E[5] = 5$
  - $E[cu(x)] = cE[u(x)]$, c is a constant and $u(x)$ is a function
  - $E[c_1u_1(x) + c_2u_2(x)] = c_1E[u_1(x)] + c_2E[u_{2(x)}]$

X	0	1	2	3
$P(X)$	0.2	0.1	0.4	0.3

- $E[X] = (0.2 * 0) + (0.1 * 1) + (0.4 * 2) + (3 * 0.3) = 1.8$
- $E[X^2] = \sum x^2 . P(X = x) = (0.2 * 0^2) + (0.1 * 1^2) + (0.4 * 2^2) + (0.3 * 3^2) = 4.4$
- $E[4X^2] = 4 * 4.4 = 17.6$
- $E[3X + 2X^2] = 3E[X] + 2E[X^2] = 3 * 1.8 + 2 * 4.4 = 14.2$

Bernoulli distribution

- A Bernoulli distribution is a discrete probability distribution for a **single** trial that has exactly **two possible outcomes**: success (1) or failure (0).
  - Success p, or failure $q = 1 - p$
- The PMF of this distribution, over possible outcomes k is
  - $f(k;p) = f(x) = \begin{cases} p, & if\ k = 1 \\ q = 1 - p, & if\ k = 0 \end{cases}$
  - can be written as $f(k;p) = p^k \times q^{1-k} = p^k \times (1-p)^{1-k}$ for $k \in \{0,1\}$
  - Or $f(k;p) = pk + (1-p)(1-k)$ for $k \in \{0,1\}$
- Expected value of the Bernoulli distribution of a random variable X
  - $E[X] = 1 * P(X = 1) + 0 * P(X = 0) = 1 * p + 0 * p = p$
- A factory produces **light bulbs**, and each bulb is either **defective (1) or non-defective (0)** If a bulb has a **5% chance** of being defective ($p = 0.05$) and ($q = 0.95$)

Binomial distribution

- A binomial distribution represents the **sum of multiple independent Bernoulli trials**
  - Parameters : n number of trials, p probability of success in each trial
  - The Bernoulli distribution is a special case of the Binomial distribution.

- PMF : $P(X = x) = \binom{n}{k} p^k \times q^{n-k}$ for $k = 0,1,2,3, \dots, n$

- A factory produces **10 light bulbs** 💡 per batch, where each bulb has a **10% chance** of being defective. Let X be the number of defective bulbs in a batch of 10.
  - Each 💡 is independent $X \sim Binomial(n = 10, p = 0.1)$
  - Compute the Probability of Exactly 2 Defective Bulbs?

- Ans: $P(X = 2) = \binom{10}{2}(0.1)^2(0.9)^{10-2} = 0.193$
  - Thus, the probability of exactly 2 defective bulbs in a batch of 10 is 19.3%.

Poisson distribution

- The **Poisson distribution** models the probability of a given number of events occurring in a fixed interval of time or space, assuming that these events happen independently at a constant rate.
- PMF : $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$ for $k = 0,1,2,3, \dots, n$
  - λ = expected number of occurrences in the given interval (mean),
- A call center receives an average of **5 calls per hour**. We want to **predict the probability of getting exactly 8 calls in an hour** using Poisson distribution.
- Ans: $P(X = 2) = \frac{5^8 e^{-5}}{8!} = 0.065$
  - The probability of receiving **exactly 8 calls in an hour** is **6.5%**.

Poisson distribution

- The **Poisson distribution** used in
  - 📊 Queueing: Predicts customer arrivals, traffic flow, and service demand to optimize staffing, scheduling, and efficiency.
  - 🤖 Machine Learning: Used in fraud detection, traffic monitoring, predictive maintenance, and network security.
  - 🏥 Healthcare: Predicts emergency patient arrivals to allocate doctors efficiently.
  - 🚗 Ride-Sharing & Logistics: Helps Uber & Lyft optimize surge pricing and food delivery time predictions.

```python
from scipy import stats
lambda_ = 180  # Average car arrivals per hour
k = 200        # Expected cars

probability = stats.poisson.pmf(k, lambda_)
print(f"Probability of 200 cars passing in an hour: {probability:.4f}")
```

```
Probability of 200 cars passing in an hour: 0.0097
```

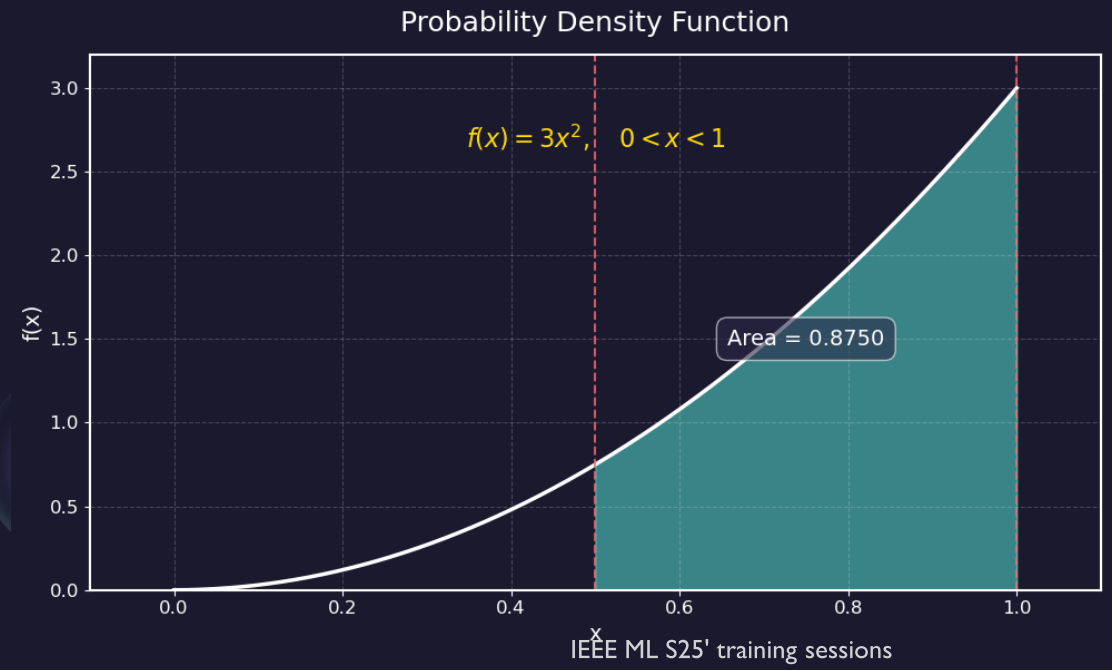
Continuous distributions

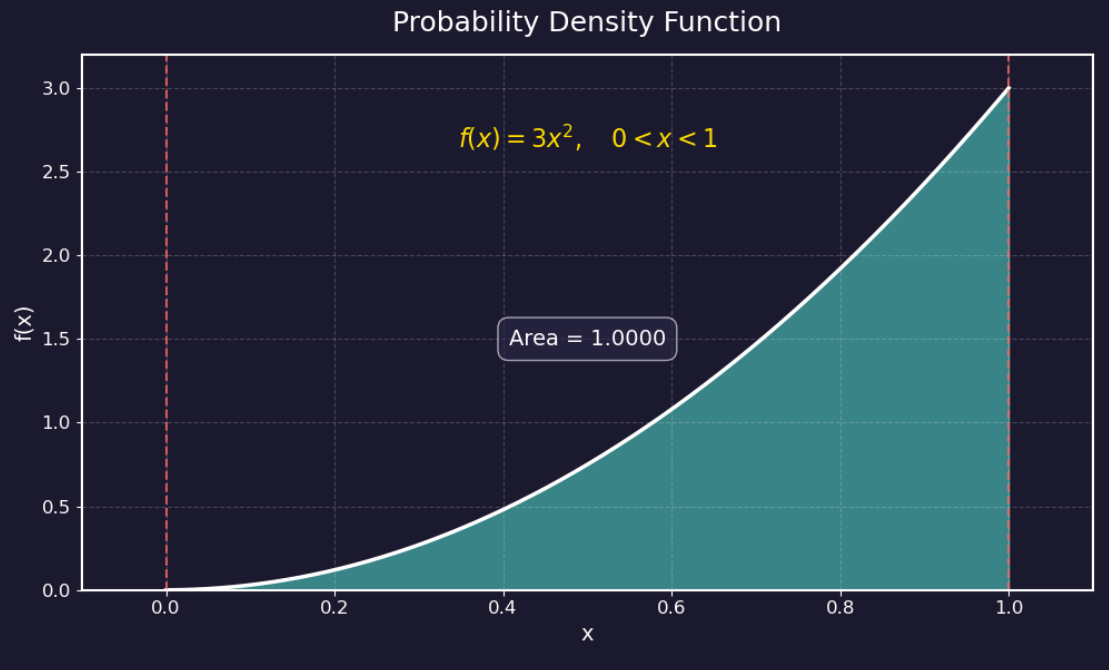
- A **continuous distribution** is a probability distribution that describes continuous random variables, meaning variables that can take an infinite number of values within a given range.
  - Examples on continuous variables (height, weight, temperature, time, and speed)
- **Infinity possible values** → the variable can take any value within a range (e.g., between 0 and 1 or $-\infty$ to ∞).
- Probability Density Function (PDF) of a continuous random variable X is an integrable function $f(x)$ satisfying the following
  - $f(x) > 0$ positive everywhere
  - $\int_S f(x)dx = 1$ the area under the curve equal 1
  - If A is some interval using the PDF over that interval $P(X \in A) = \int_A f(x)dx$
  - We just changed the summation and used integrals; in the end the integration is just is a summation of shapes that keep shrinking for infinity.

Continuous distributions

- Let X be a continuous random variable whose PDF is $f(x) = 3x^2, 0 < x < 1$
  - Note that $f(x) \neq P(X = x), f(0.9) = 3(0.9)^2 = 2.43$ which is not a probability
  - That 's in the continuous case.
  - $f(x)$ is the height of the curve at $X = x$

- What is the probability X falls between $\frac{1}{2}$ and 1, $P(\frac{1}{2} < X < 1)$

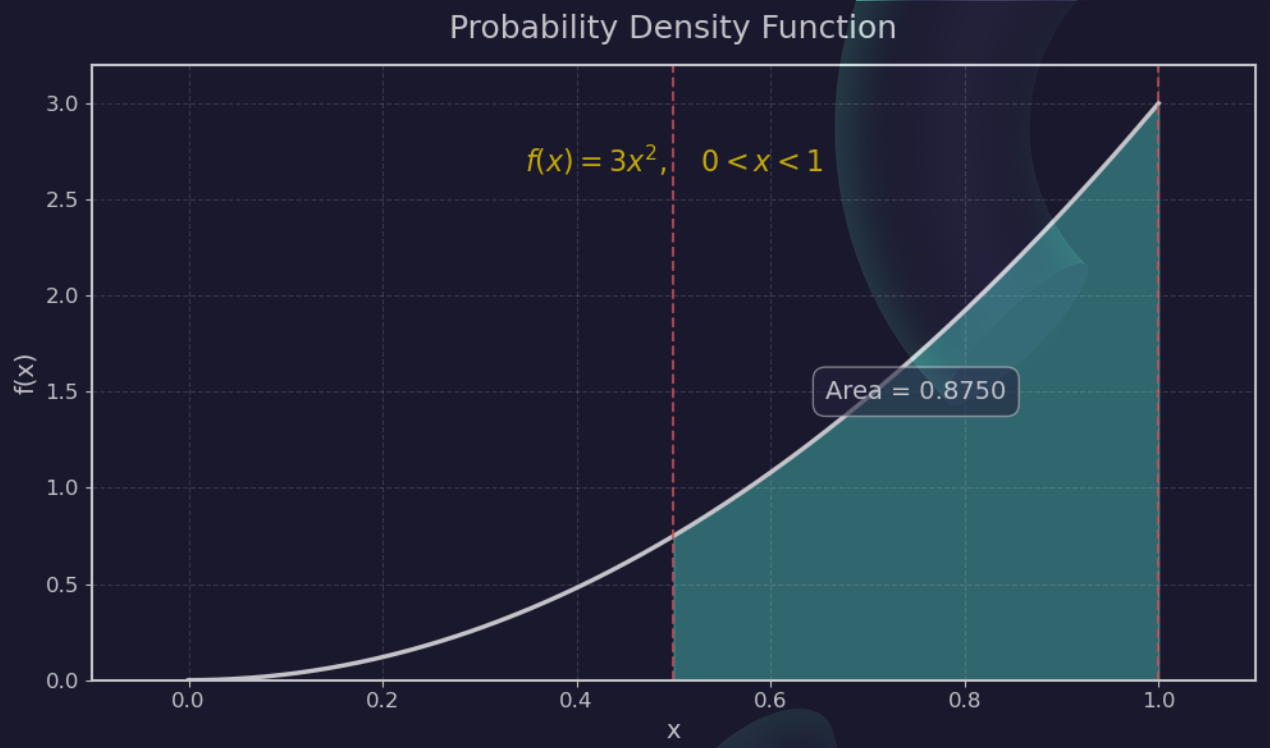




Continuous distributions

- What is the probability X falls between $\frac{1}{2}$ and 1, $P(\frac{1}{2} < X < 1)$

- $P\left(\frac{1}{2} < X < 1\right) = \int_{0.5}^{1} 3x^2 dx$
  $= [x^3]_{0.5}^{1} = [(1)^3 - (0.5)^3]$
  $= 0.875$



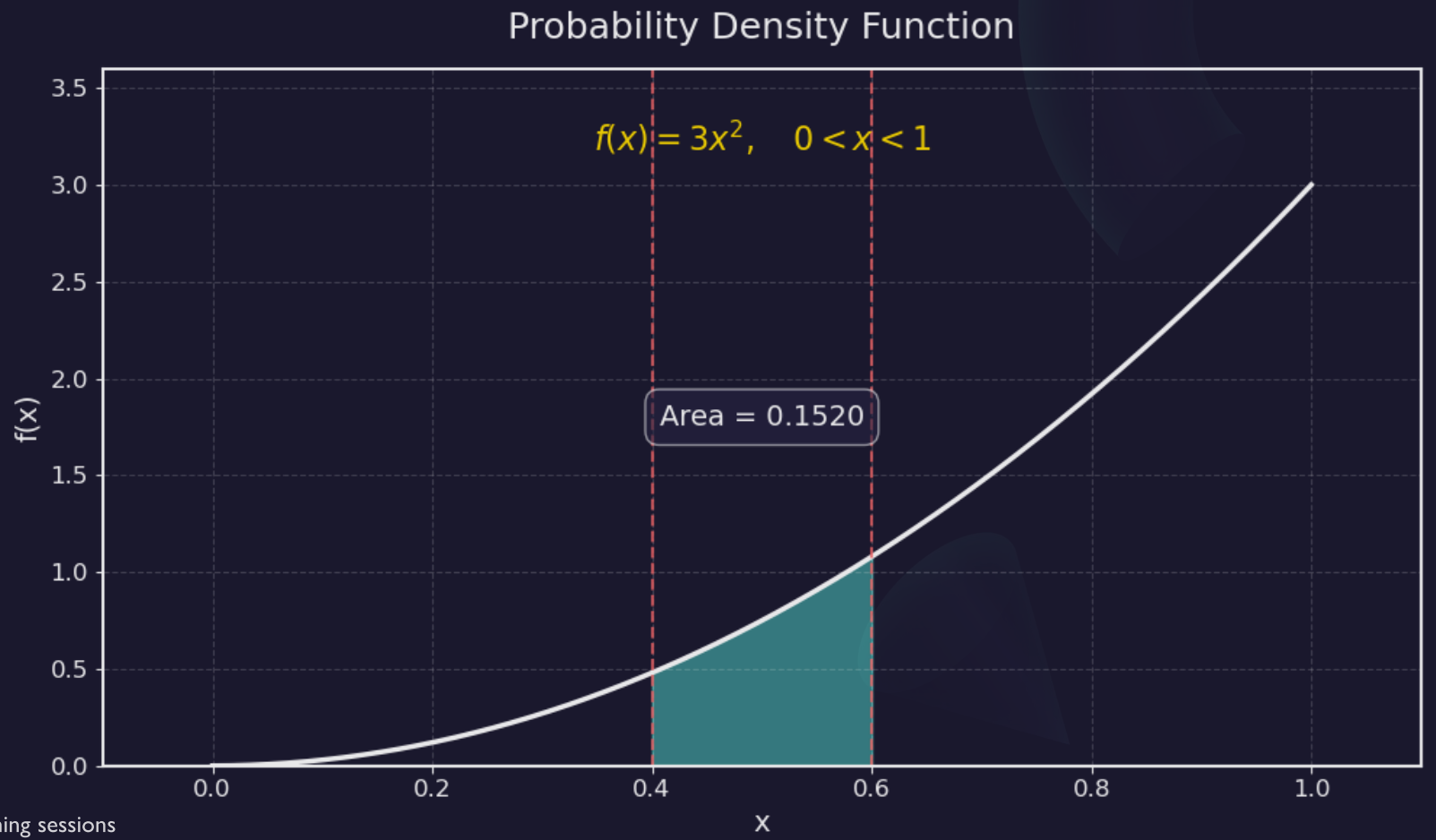
- What is $P\left(X = \frac{1}{2}\right)$?

- $P\left(X = \frac{1}{2}\right) = \int_{0.5}^{0.5} 3x^2 dx = \frac{1}{8} - \frac{1}{8} = 0$ always zero because it's not interval !

Continuous distributions

- An implication of $P(X = x) = 0$ for all x in the X, you can be careless about the endpoint of the intervals
  - $P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b)$

What is the probability X falls between 0.4 and 0.6, $P(0.4 < X < 0.6)$ find algebraically



Cumulative Distribution Functions (CDF)

- The **Cumulative Distribution Function (CDF)** of a random variable X gives the probability that X takes a value less than or equal to some value x.
  - $F(x) = P(X \leq x) = \int_{\infty}^{x} f(t)dt$
  - $f(x)$ is the PDF
  - CDF is a monotonically increasing function (never decreases).

- Properties of CDF
  - $F(x)$ between 0 and 1, $0 \leq F(x) \leq 1$
  - $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
  - The PDF is the derivative of the CDF
    - $f(x) = \frac{d}{dx} F(x)$

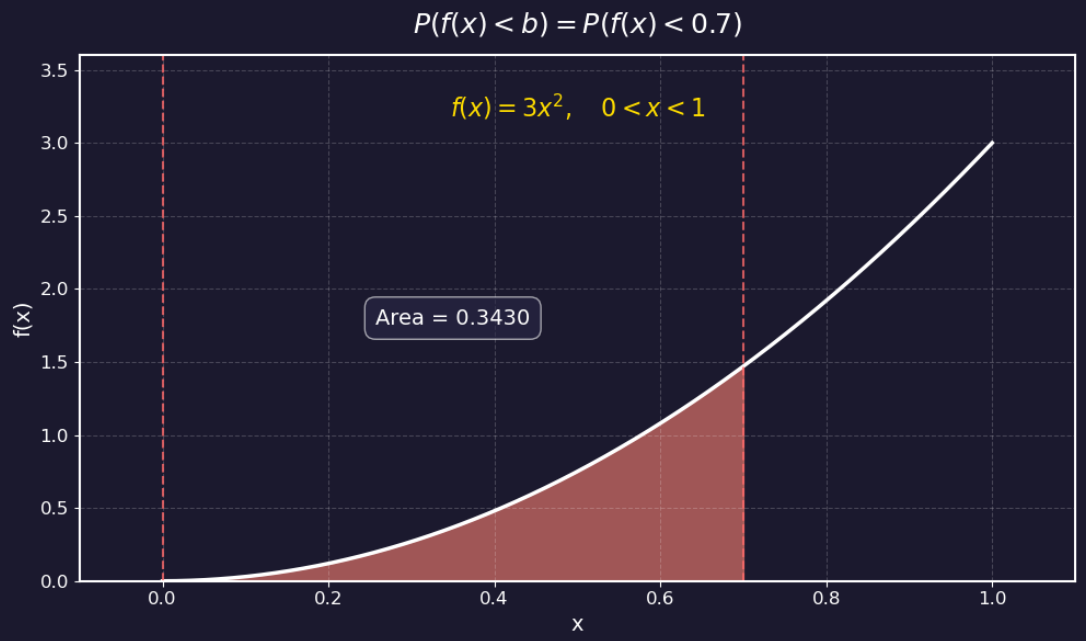
Cumulative Distribution Functions (CDF)

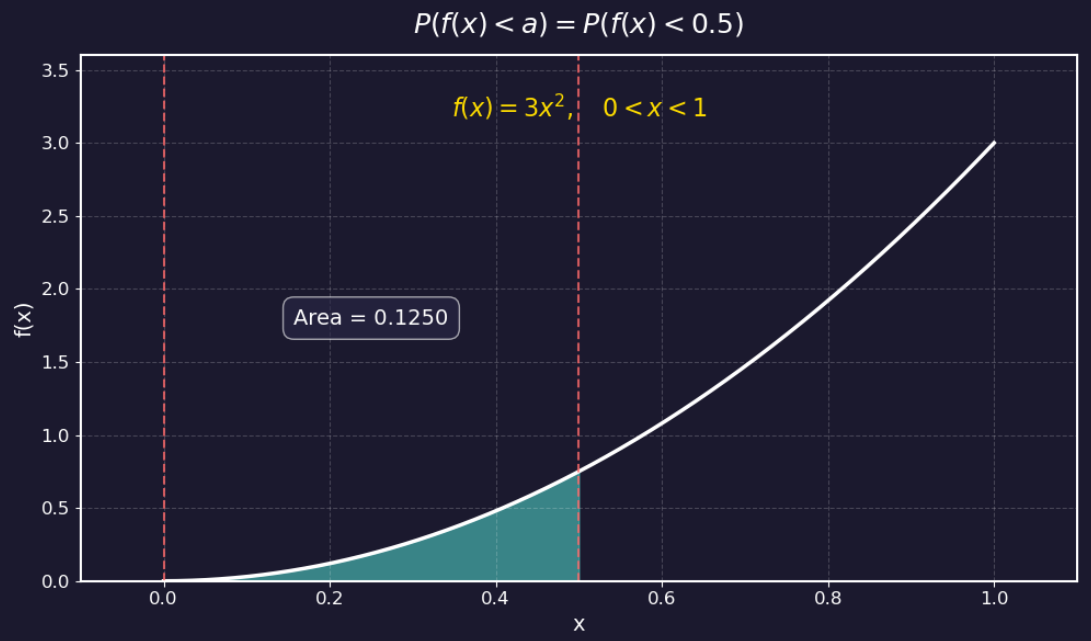
- The **Cumulative Distribution Function (CDF)** of a random variable X gives the probability that X takes a value less than or equal to some value x.
  - $F(x) = P(X \leq x) = \int_{\infty}^{x} f(t)dt$
  - $f(x)$ is the PDF
  - CDF is a monotonically increasing function (never decreases).

- Properties of CDF
  - $F(x)$ between 0 and 1, $0 \leq F(x) \leq 1$
  - $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
  - The PDF is the derivative of the CDF
    - $f(x) = \frac{d}{dx} F(x)$
  - $P(a \leq X \leq b) = F(b) - F(a)$

Cumulative Distribution Functions (CDF)

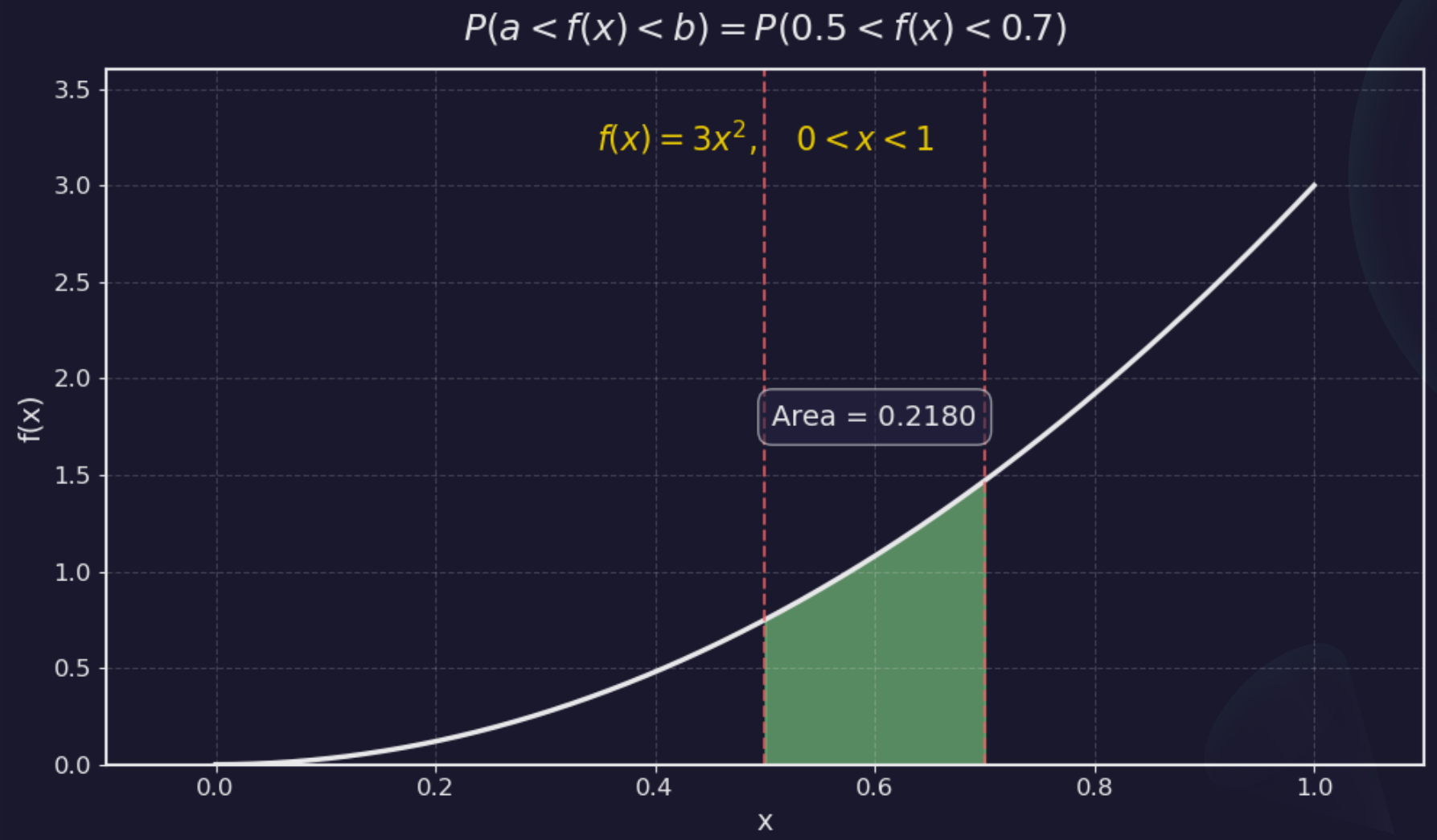
- $P(a \leq X \leq b) = F(b) - F(a)$
  - Let's say $a = 0.5$ and $b = 0.7$



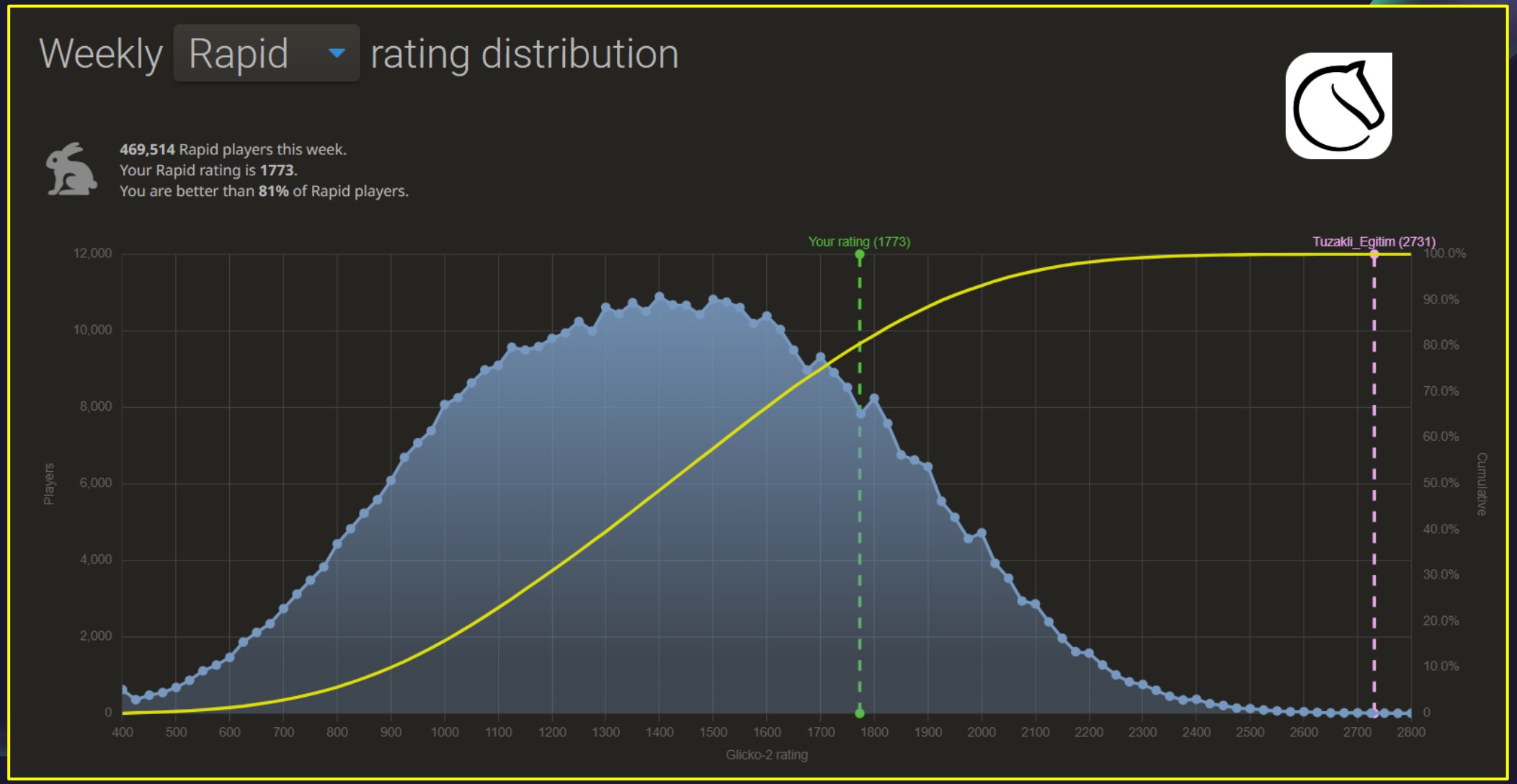


- $P(a \leq X \leq b) = F(b) - F(a) =$
- $P(0 < X < 0.7) - P(0 < X < 0.5) = 0.343 - 0.125 = 0.218$

Cumulative Distribution Functions (CDF)



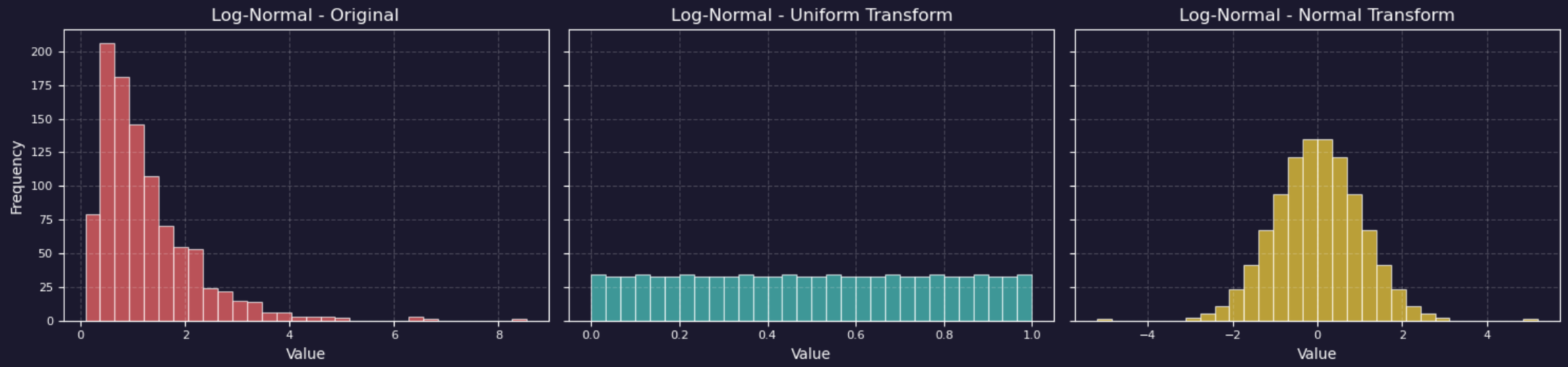
Cumulative Distribution Functions (CDF)



Cumulative Distribution Functions (CDF)

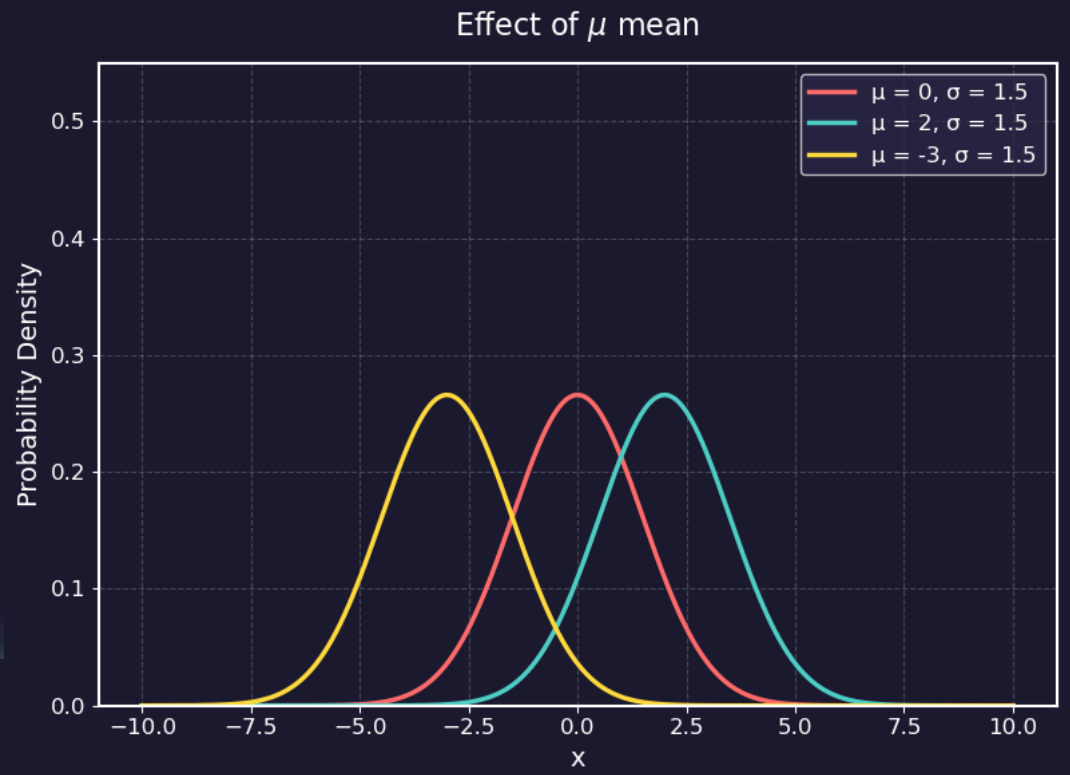
- CDF-based transformations are **data preprocessing techniques** that use the **Cumulative Distribution Function (CDF)** to reshape a dataset.

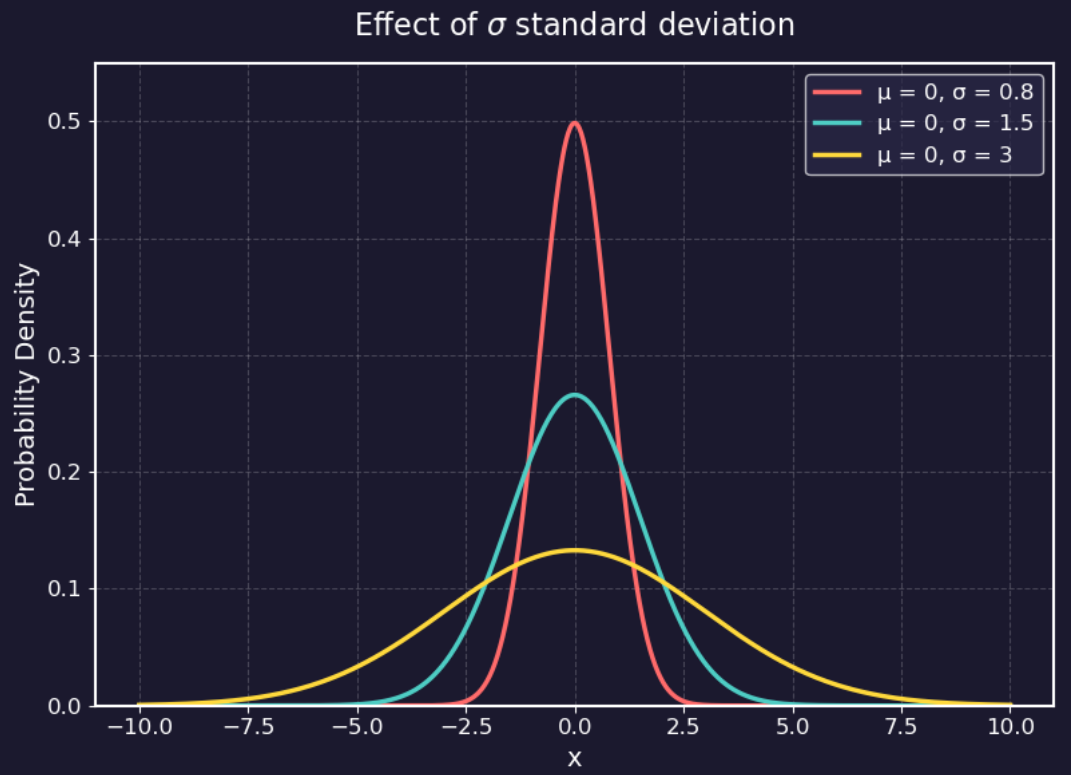
- These transformations are especially useful for normalizing **skewed** data and improving machine learning model performance.
  - QuantileTransformer from sklearn



Normal distribution

- Normal distribution or Gaussian is a continuous distribution take the shape of bell 🔔
  - PDF : $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
  - μ is the mean, the median and the mode; σ^2 is the variance





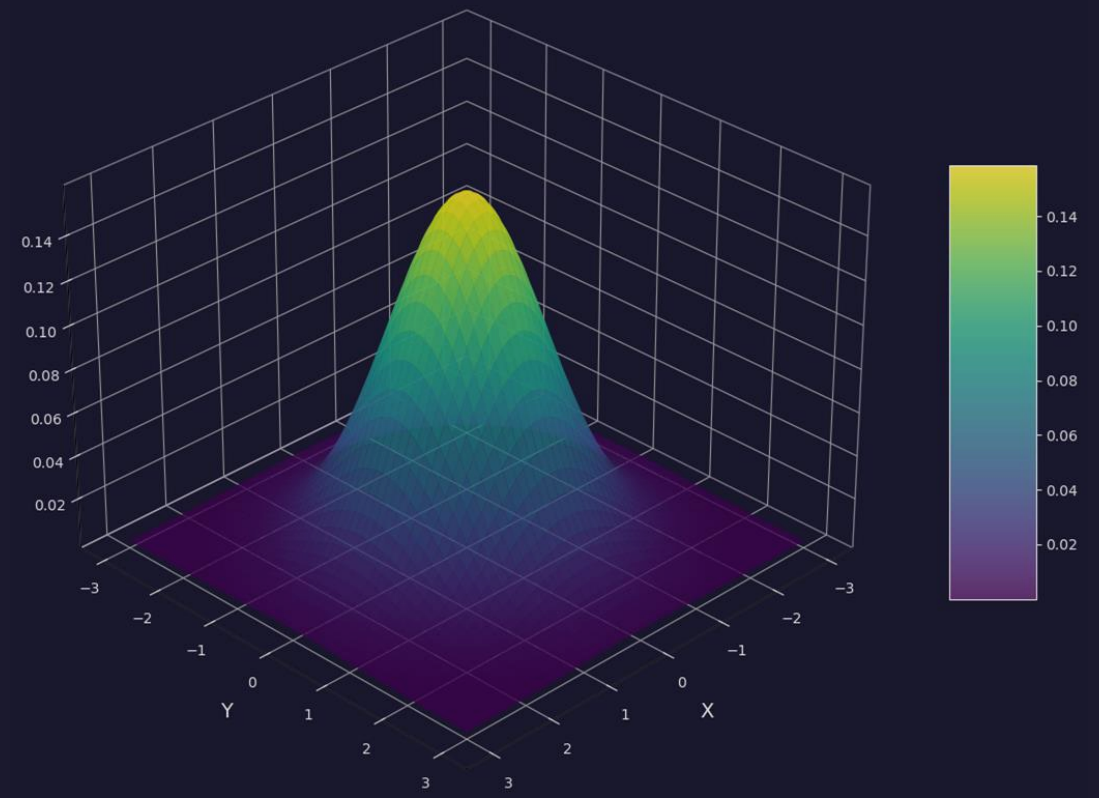
Standard Normal distribution

- If $Z \sim N(\mu = 0, \sigma = 1)$ we say that Z follow a Standard normal distribution
  - PDF : $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(z)^2}{2}}$
- It's a Normal distribution thar has a mean of zero (symmetric around zero) and a standard deviation of 1.
- CDF $\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{(w)^2}{2}} dw$
  - $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$
  - $P(Z > a) = 1 - \mathrm{P}(\mathrm{Z} \leq \mathrm{a}) = 1 - \Phi(a)$
  - $\Phi(-a) = 1 - \Phi(a)$
  - $P\big(Z > (-a)\big) = 1 - P(Z \leq -a) = 1 - \Phi(-a) = 1 - 1 + \Phi(a) = \Phi(a)$

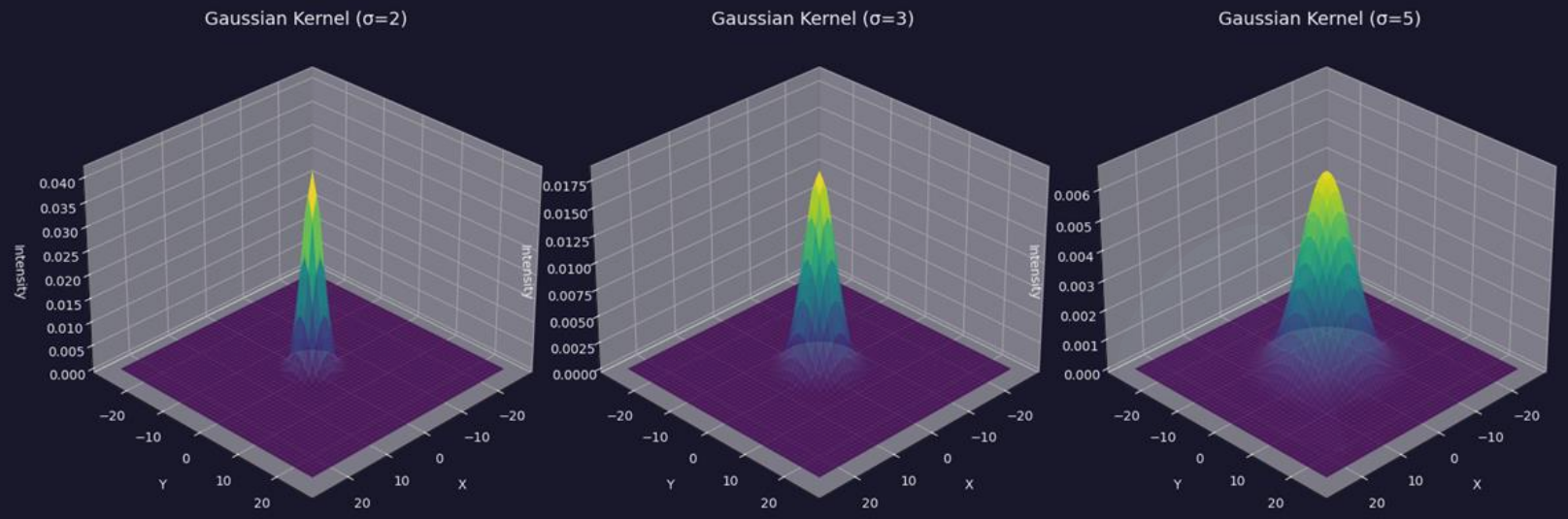
Normal distribution

- A simplified equation for a 2D, normal equation (pdf)

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$



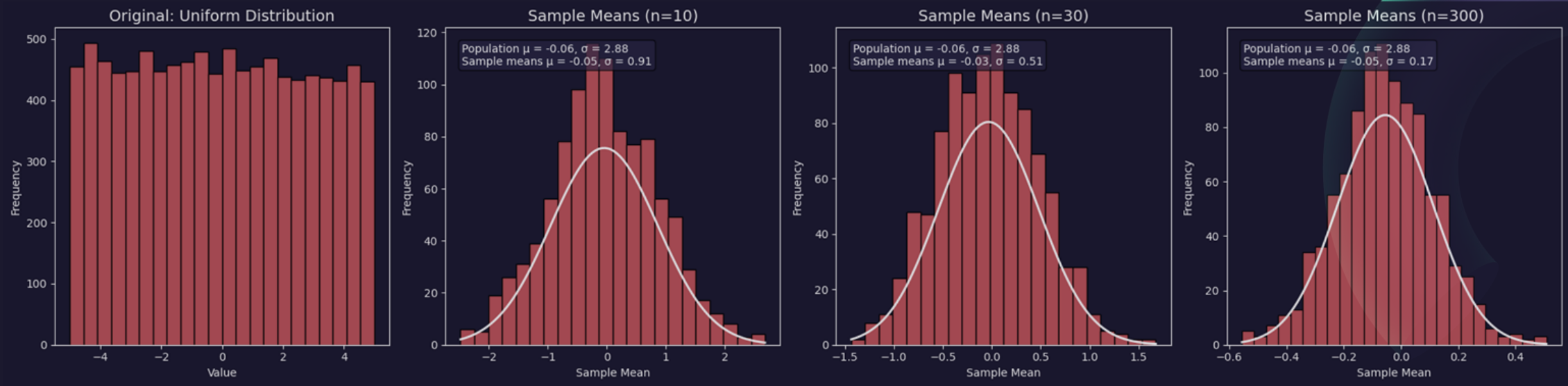
Normal distribution and image blurring

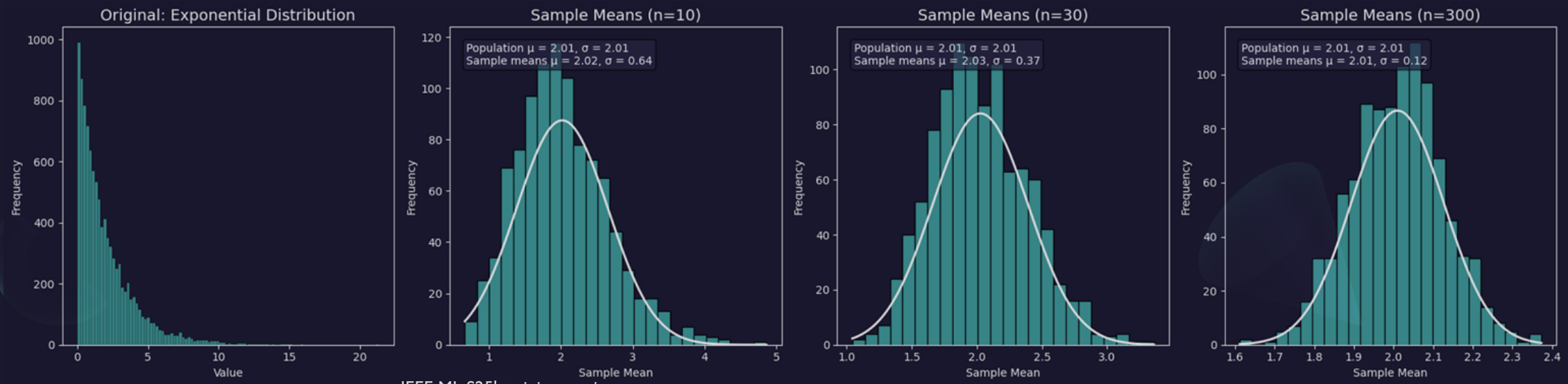


Central Limit Theorem (CLT)

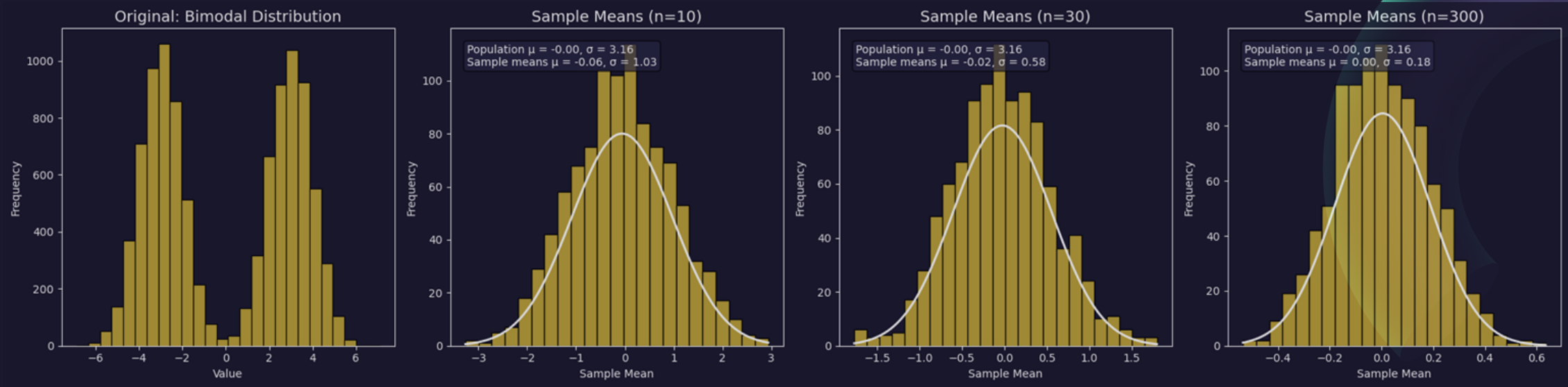
- The distribution of the sample mean approaches a normal distribution as the sample size increases, regardless of the original population distribution
- This means that even if the **original data is not normally distributed**, the **mean of sufficiently large random samples** will be **approximately normal**.
- The **larger the sample size** (typically **n ≥ 30**), the closer the sample mean distribution is to **normal**.
- The samples should be **random** and **independent** for CLT to hold.
- The original population can be **skewed, uniform, exponential, or even multimodal**, but the sample means will still be **normally distributed**.

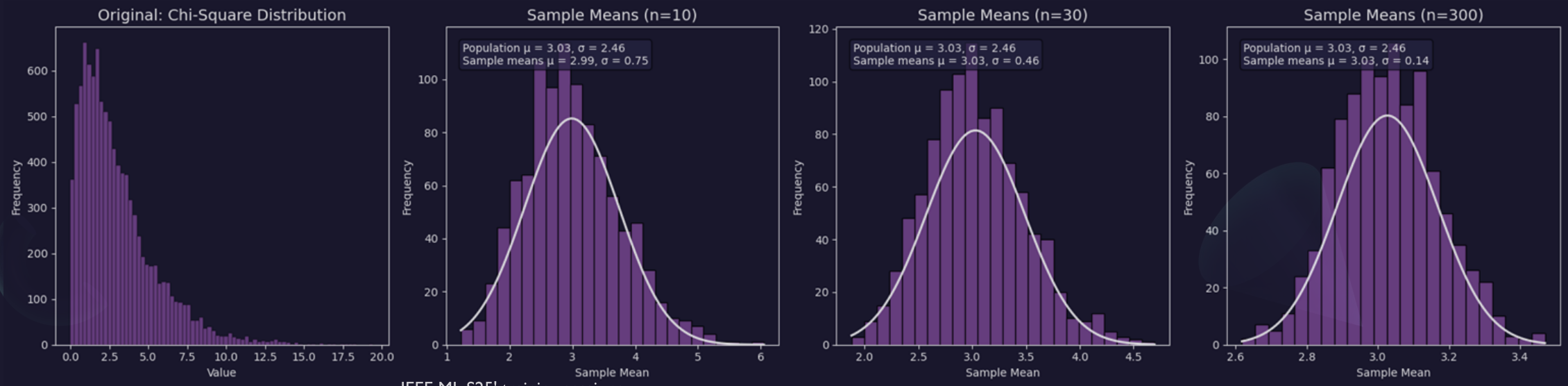
Central Limit Theorem (CLT)





Central Limit Theorem (CLT)





Likelihood

- Probability measures the chance of observing a specific outcome given a known model or distribution.
  - It is a function of the outcome (data), assuming the model parameters are fixed and known.
  - $P(X = x; \theta)$, θ represent the parameter of the model.
  - $P(data|distribution)$

- Likelihood refers to the process of determining the best data distribution given a specific situation in the data.
  - $P(distribution\ |data)$

- Likelihood is a function that measures how plausible a given set of parameters (or model) is, given observed data.
  - $L(\theta|X) = P(X|\theta)$, how likely a parameter θ given the observed data X.

Likelihood

- Likelihood is a function of the parameters given observed data.
- It quantifies how plausible a parameter value is, given the data.
- Unlike probability, likelihood does not sum to 1 over all possible outcomes
- You flip a coin 🪙 10 times and observe 7 heads. You want to estimate the probability p of getting heads.
  - **Probability** (Forward-looking): If we assume $p = 0{,}6$, what is the probability of observing 7 heads in 10 flips?
  - **Likelihood** (Reverse-thinking): Given that we observed 7 heads, what is the most likely value of p?
- Likelihood is finding the best parameters that would fit the data, and that comes a lot in ML.

Maximum Likelihood Estimation

- $X_1, X_2, \dots, X_n$ a random sample, from distribution depends on unknows parameters $f(x; \theta_1, \theta_2, \dots, \theta_m)$ with PMF or PDF.

1 $$L(\theta_1, \theta_2, \dots, \theta_m) = f(x_1; \theta_1, \theta_2, \dots, \theta_m) * f(x_2; \theta_1, \theta_2, \dots, \theta_m) * \cdots * f(x_m; \theta_1, \theta_2, \dots, \theta_m)$$

$$L(\theta_1, \theta_2, \dots, \theta_m) = \prod_{i=0}^{n} f(x_i; \theta_1, \theta_2, \dots, \theta_m)$$

2 $$l(\theta) = ln\, L(\theta)$$

3 $$\frac{\partial l(\theta)}{\partial \theta} = 0$$

See 👀

- https://youtu.be/9ONRMymR2Eg?si=rI27KzWtottYf9Gp 📺 (Sample variance and why n-1?)
- Welcome to STAT 414! | STAT 414 (course from Pennsylvania State university)
- https://www.columbia.edu/~ww2040/8100F16/Riquelme-Johari-Banerjee.pdf (see how Uber 🚗 use poisson distribution)
- https://conference.nber.org/confer/2017/MDf17/Castillo_Knoepfle_Weyl.pdf
- https://hossam-ahmed.notion.site/0fc336c14f994eff8a40b17abba23b5d?v=e4777f06cc9a4d6684227e0eee201bfe
- MLE solved examples
- https://youtu.be/sguol03tfWo?si=S4R8I4j47PgwgXo5 📺 (MLE)