



EN3150 Pattern Recognition

Motivation and Introduction to the Course

M. T. U. Sampath K. Perera,
Department of Electronic and Telecommunication Engineering,
University of Moratuwa.
(sampathk@uom.lk).
Semester 5 – Batch 20.

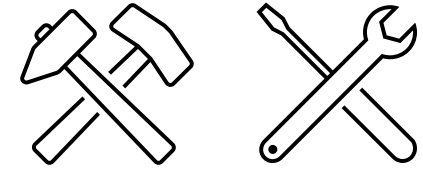
Administrative Information

- The course web page on LearnOrg will contain all relevant information about the course and serve as the primary source for essential announcements related to this class.
- Lecturer - Dr. Sampath K. Perera
- Email- sampathk@uom.lk
- Room No- EB 111, Department of Electronic & Telecommunication Engineering, Faculty of Engineering, University of Moratuwa.
- Office Hours- Tuesdays 13:15-15:15 (contact to schedule an appointment)
- No. of Credits : 3

(one credit = 50 notional hours of study (lecture hours + Labs and Projects + self-learning))

- Lecture Time:
 - Tuesday – 15:15-17:15
 - Thursdays – 15:15-17:15 Labs and self study

Textbooks and software tools



■ Textbooks:

1. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer NY, 2006.
2. Kevin P. Murphy, *Probabilistic machine learning: an introduction*. MIT press, 2022.

■ Software and toolboxes:

1. Python programming language (<https://www.python.org/>)
2. Scikit-learn Machine Learning in Python (<https://scikit-learn.org/stable/index.html>)
3. Pytorch open-source machine learning library (developed by Facebook) (<https://pytorch.org/get-started/locally/>)
4. Tensorflow Open-source machine learning library (developed by the Google) (<https://www.tensorflow.org/>)
5. Seaborn: statistical data visualization (<https://seaborn.pydata.org/>)

Textbooks and software tools

How to run the code?

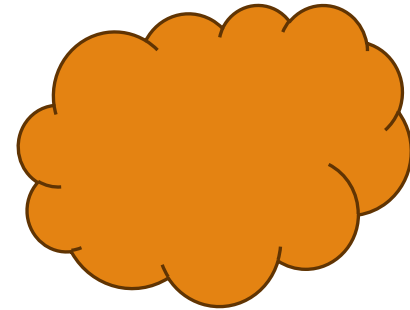
Local Machine

- Install Python and required libraries (e.g., NumPy, pandas, scikit-learn) on your local machine.
- Write your Python machine learning code using a code editor or Integrated Development Environment (IDE). E.g. [PyCharm](#), [Visual Studio Code](#), or [Spyder](#)
- Run the code directly on your local machine.



Online Code Editors

- Cloud-based platforms allow you to write and run Python machine learning code directly in your web browser
- [Google Colab](#) and [Kaggle Kernels](#).



Learning Outcomes

LO1	Explain the process of learning from data and related challenges.
LO2	Characterize a wide class of pattern recognition/machine learning (ML) algorithms by the underlying mathematical structures and limitations.
LO3	Demonstrate the utility of pattern recognition/ML algorithms with the help of publicly available software libraries and data sets.
LO4	Implement different pattern recognition/ML algorithms in a range of practical applications.
LO5	Build a simple convolutional neural network to perform classification.

Module Outline

	Topic	Learning Outcomes
1.	Introduction [2 hours] Learning from data and related challenges, supervised vs unsupervised learning, model selection and bias-variance trade-off.	LO1, LO2
2.	Linear Models for Regression [6 hours] Linear regression models and least squares, subset selection, regularized linear models (e.g., Ridge, LASSO), prediction and related confidence intervals.	LO2 to LO4
3.	Classification [6 hours] Linear models of classification, discriminant functions, generative models, probabilistic discriminative models, optimal separating hyperplanes and SVM.	LO2 to LO4
4.	Kernel Methods [4 hours] Feature maps, representer theorem, kernels and kernel trick, kernel density estimation.	LO2 to LO4
5.	Additive Models and Mixtures [4 hours] Tree based methods, boosting, ensemble methods, mixture of Gaussians, EM algorithm.	LO2 to LO4
6.	Unsupervised Learning Techniques [2 hours] Cluster analysis, principal components analysis, independent component analysis, multidimensional scaling.	LO2 to LO4
7.	Deep Neural Networks [4 hours] Introduction to neural networks (NN) and backpropagation, architecture of convolutional neural networks, implementing NN using frameworks, training neural networks and performance analysis.	LO2, LO3, LO5

Evaluation Criterion

Continuous Assessment – **70%**

End of Semester Examination – 30%

Continuous Assessment: *Individual and group-based activities.*

- *Schedule will be provided.*

Evaluation Criterion

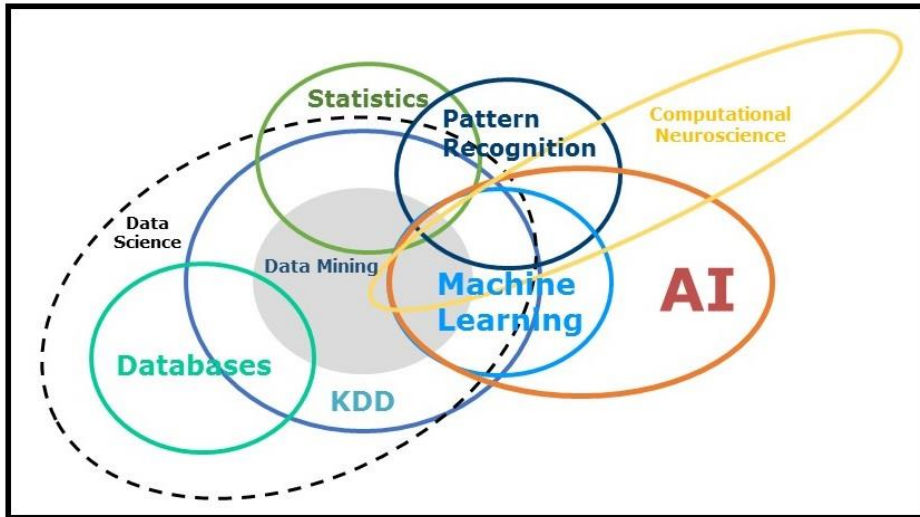
- Assignment 01: Regression*15%
- Assignment 02: Classification* 10%
- Assignment 03: Quiz (Closed book, written exam) 15%
- Assignment 04: Kernel Methods, Additive Models and Mixtures, Unsupervised Learning Techniques* 10%
- Assignment 05 : Convolutional neural network to perform classification* (Group) 20%
- Final exam: Closed book, written exam 30%

➤ Homework exercise after each section/lecture (will not be graded)

*Manual calculation and derivations + computer simulation + report

Motivation and Introduction

- Pattern recognition



- Searching for patterns in data is a fundamental task in data analysis, machine learning, and artificial intelligence.
- The goal is to identify regularities, trends, correlations, or any **meaningful structure** within the data that can provide **valuable insights** or be used for predictive modeling.
- Overall, pattern recognition plays a crucial role in enabling machines to learn from data and make informed decisions, making it a core component of modern AI systems

Artificial intelligence refers to the simulation of human intelligence in machines that are programmed to think and learn like humans. The primary goal of AI is to create intelligent systems that can perform tasks that typically require human intelligence, such as problem-solving, reasoning, learning, perception, speech recognition, and decision-making.

Machine learning is a type of artificial intelligence (AI) that allows software applications to become more accurate in predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

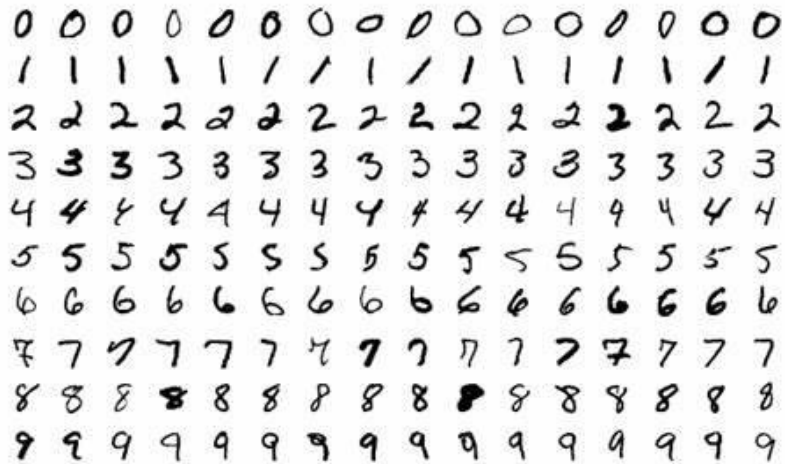
[AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the difference? \(ibm.com\)](https://ibm.com)

Image is adapted from <https://blogs.sas.com/content/subconsciousmusings/2014/08/22/looking-backwards-looking-forwards-sas-data-mining-and-machine-learning/>

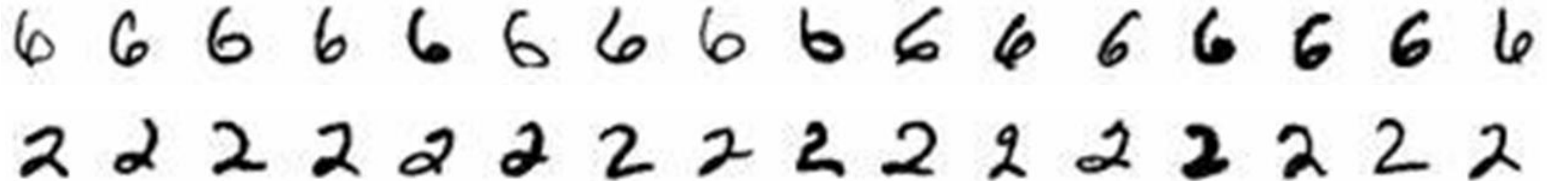
Motivation and Introduction

Example: Handwritten Digit Recognition.

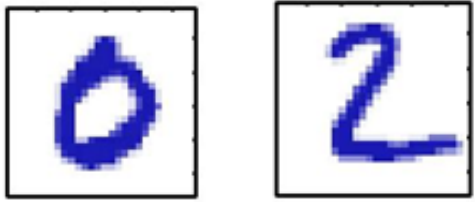
Can we use a predefined program to detect digits based on shapes of the strokes?



Gives **poor results** due to the **wide variability** of handwriting.



Sample of the MNIST dataset of handwritten digits
(https://en.wikipedia.org/wiki/MNIST_database)



Motivation and Introduction

Example: Handwritten Digit Recognition.

If you are given a task to distinguishing between the digits 0 and 2, which features would you choose to focus on?



Closed loop: The digit 0 is a closed-loop shape. Detecting a closed loop with continuous edges can be a strong indicator of the digit 0.

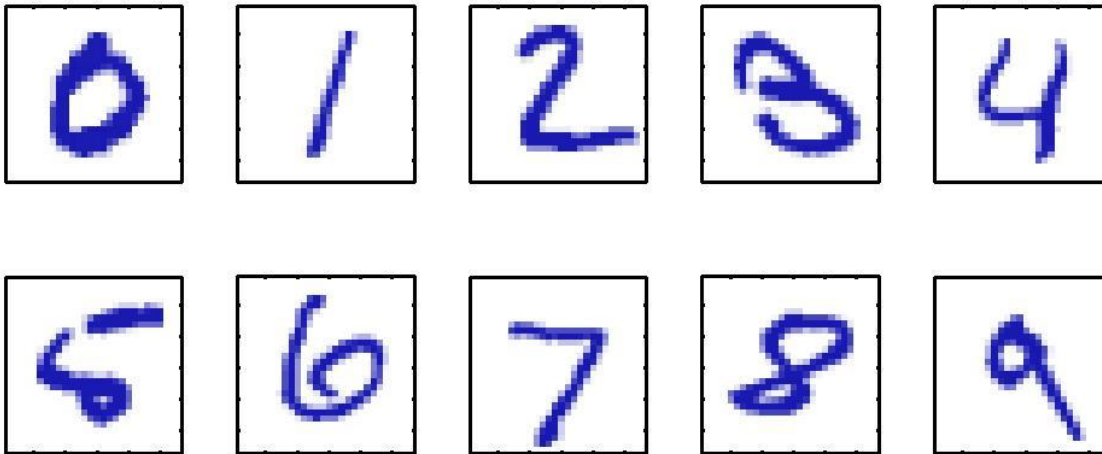
Look for a curved edge: The top part of the digit 2 usually has a curved shape, resembling a semi-circle. Detecting this curved edge can be helpful in distinguishing it from other digits.



No horizontal line: Unlike digit 2, digit 0 does not have a horizontal line at the bottom. Instead, the loop is typically the only edge feature at the bottom part of the digit.

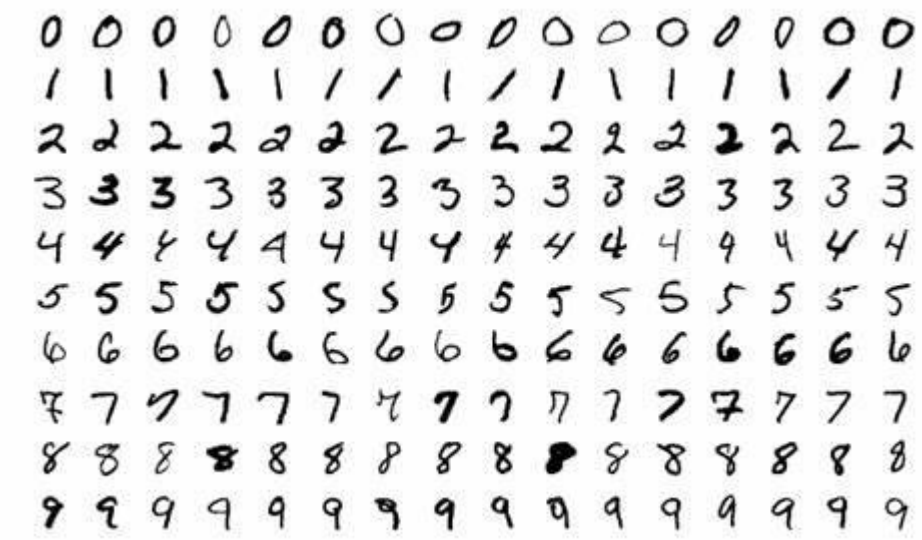


Horizontal line: Digit 2 often has a horizontal line at the bottom, which is usually straight and doesn't have abrupt changes in intensity. This line can be used as an additional distinguishing feature.



Motivation and Introduction

Example: Handwritten Digit Recognition
 Task: recognize handwritten digits it has never seen before.



Pattern recognition techniques are employed to extract relevant features from the raw pixel data of the images. These features could include the presence of **edges**, **corners**, **loops**, or other distinctive characteristics that help differentiate between different digits.

Sample of the MNIST dataset of handwritten digits
 (https://en.wikipedia.org/wiki/MNIST_database)

Motivation and Introduction

Example: Face recognition

Task: Automatic identification and verification of individuals based on their facial features which never seen before.

In the context of face recognition, pattern recognition techniques are applied to identify and interpret patterns within facial images. These patterns could be specific facial features like the arrangement of eyes, nose, and mouth, or more abstract features extracted using advanced algorithms such as deep learning models.



Image credit <https://media.istockphoto.com/>

Motivation and Introduction

Applications

- Machine Vision
- Speech Recognition
- Character Recognition
- Medical Imaging
- Robotics
- Speech and Natural Language Processing



Siri and Google Assistant

Netflix Recommendations

Grammarly

Google photos

Image credit <https://www.zwillgen.com/>

AI & ML Black box nature



- The "black box" nature of ML/AI refers to the lack of interpretability or transparency in understanding how these complex models arrive at their decisions or predictions.
- Possible reasons
 - Complexity
 - Non-linearity
 - High-dimensional data

[Unpacking black-box models | MIT News | Massachusetts Institute of Technology](#)

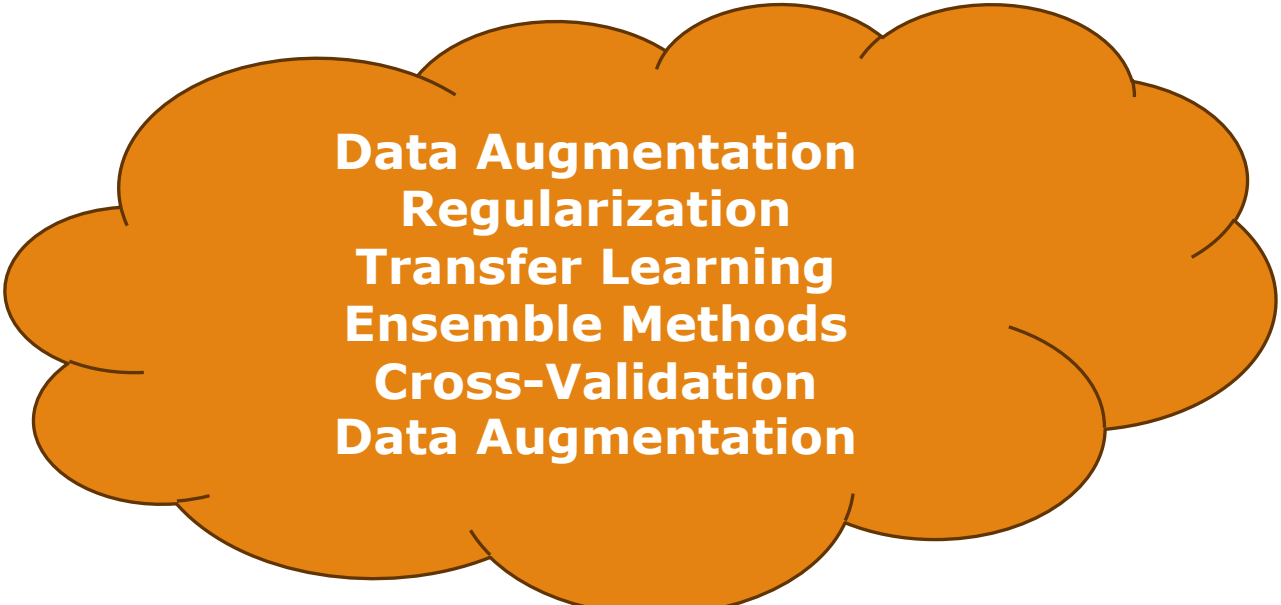
[How well do explanation methods for machine-learning models work? \(mit.edu\)](#)

AI & MI

- Generalizability
- Interpretability
- Explainability and Causality
- Scalability
- Learn effectively from limited data
- Robustness
- Multi-modal Learning
- Domain-Specific Adaptations
- Privacy and Security

AI & ML Black Challenges

- Improving the **Generalizability** of machine learning models
 - Generalizability refers to the ability of a trained model to perform well on unseen data, especially when the new data is significantly different from the data on which the model was trained. The lack of generalizability can lead to severe performance degradation and limit the practical applicability of ML models.
- Possible causes
 - Data Distribution Discrepancy
 - Overfitting
 - Insufficient Data

An orange cloud-like shape with a black outline, containing a list of techniques to improve generalizability.

Data Augmentation
Regularization
Transfer Learning
Ensemble Methods
Cross-Validation
Data Augmentation

AI & MI Black Challenges

➤ Interpretability

➤ The lack of interpretability is indeed a critical limitation of many machine learning (ML) models, especially as they become more complex and sophisticated. Interpretability refers to the ability to understand and explain how a model arrives at its predictions or decisions.

➤ Why?

➤ Complexity

➤ Non-linearity



The lack of interpretability in ML has several implications

Trust and Reliability (e.g., applications like healthcare)
Debugging and Improvement

Probability Theory

Slides are based on C. M. Bishop, Pattern Recognition and Machine Learning, Springer NY, 2006.

Probability Theory

Key Concept: Uncertainty in Pattern Recognition.



- Uncertainty is a fundamental concept in pattern recognition.
- It emerges from noise in measurements and the limited size of data sets.
- In pattern recognition, we must account for uncertainty to make reliable decisions and predictions.
- Probability theory serves as a **consistent framework** for quantifying and manipulation uncertainty, making it a **central foundations** in the field of pattern recognition.

Probability Space



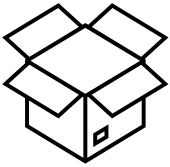
- Probability is associated with uncertain events.
- For example, we can talk about the probability of rain on a specific date.
- To discuss probability theory formally, we need to define the possible events to which we want to assign probabilities.
- Probability space (Ω, \mathcal{F}, P)
 - Outcome space or sample space (Ω)
 - E. g, rolling a fair six-sided die, the sample space would be $\{1, 2, 3, 4, 5, 6\}$
 - Event space $\mathcal{F} \subseteq 2^\Omega$ (power set of Ω -which is the set of all possible subsets of Ω , including the empty set and Ω itself)
 - Events are subsets of the sample space ($E \in \mathcal{F}$), representing specific outcomes or combinations of outcomes
 - E.g., Getting number 2: $\{2\}$, Getting an even number: $\{2, 4, 6\}$, Getting an odd number: $\{1, 3, 5\}$
 - Probability measure P
 - Maps an event ($E \in \mathcal{F}$) to a real value between 0 and 1
 - E.g, for a fair six-sided die
 - $P(\text{Getting an even number}) = 3/6 = 1/2$

Probability

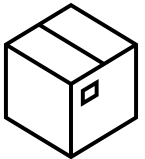
- Non-negativity $P(a) \geq 0$ for $\forall a \in F$
- Range: Probability values always lie between 0 and 1,
- Additivity (two mutually exclusive events) $a, b \in F$
 - $P(a \cup b) = P(a) + P(b)$.
- $P(\Omega) = 1$ (probability of sample space)
- Independence: Two events are independent if the occurrence of one does not affect the probability of the other occurring.
- Conditional Probability: It refers to the probability of an event happening given that another event has already occurred.
- Joint Probability: For two or more events, it represents the probability of their simultaneous occurrence.
- Marginal Probability: The probability of a single event without considering other events.

Probability Theory

Classical probability



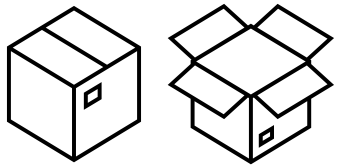
- The probability of an event is calculated by comparing the number of favorable outcomes to the total number of equally likely outcomes.
- We have information about all the possible options.
- Example: Drawing a Card from a Standard Deck.



Frequentist probability

- It is based on the idea of long-run frequencies or repeated trials of an experiment.
- Example: Drawing a Card from a Deck in which we do not know the card distribution.

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{\# occurrences of the event}}{\text{\# Total occurrences}}$$

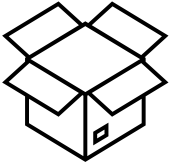


Bayesian probabilistic view

- Probability as a measure of subjective belief or confidence.
- In the Bayesian framework, probabilities are assigned to events based on both prior knowledge and new evidence.
- As more evidence becomes available, the initial beliefs (prior probabilities) are updated to incorporate the new information, resulting in revised probabilities known as posterior probabilities.

Probability Theory

Classical probability

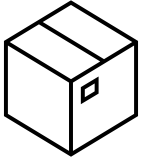


Example: Drawing a Card from a Standard Deck

Sample Space = {hearts, diamonds, clubs, and spades} and 13 cards from each.

Event: Getting diamond

Probability of the event (P) = Number of Favorable Outcomes / Total Number of Outcomes = $13/52 = 0.25$



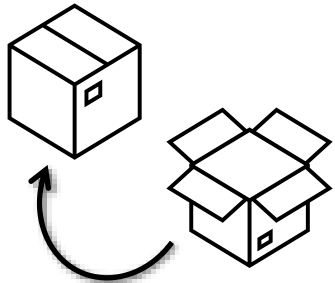
Frequentist probability

Example: Drawing a Card from a Deck in which we do not know the card distribution.

After 10,000 trials, we got 5000 diamonds.

$$P(A) = \lim_{n \rightarrow \infty} \frac{\text{\# occurrences of the event}}{\text{\# Total Outcomes}}$$

Probability of the event (P) = Number of occurrences of the event / Total trials = $5000/10,000 = 0.5$



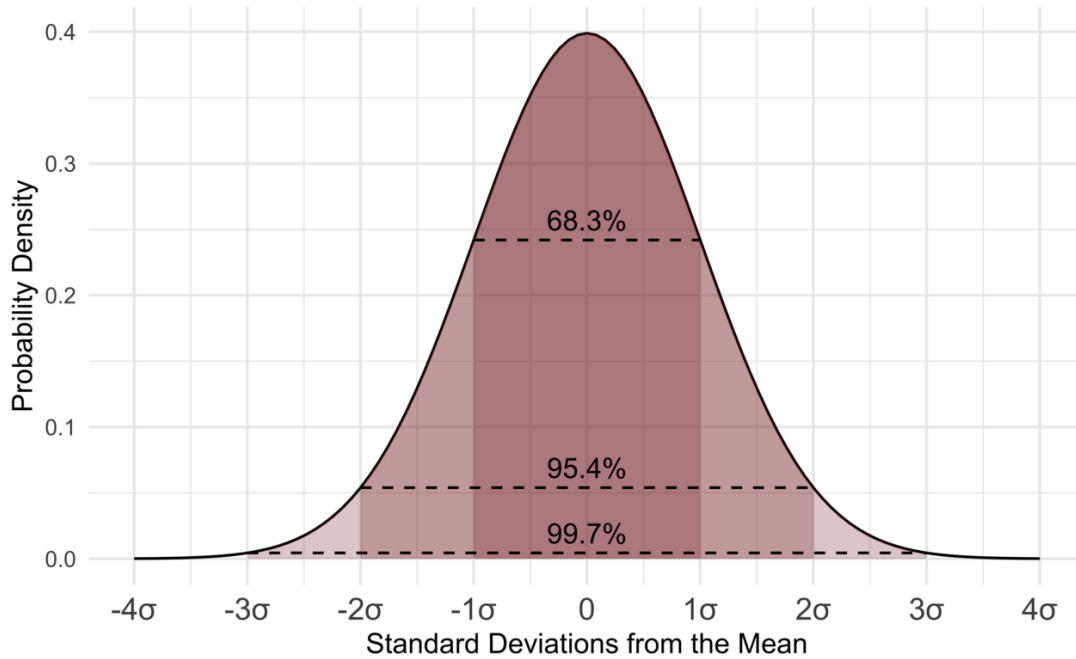
Bayesian probabilistic view

Example: Drawing a Card from a Deck in which we do not know the card distribution.

We have a prior belief that the deck might be biased, favoring the diamond suit more than others.

Before observing any evidence, we assign a prior probability to our belief that a card drawn will be a diamond with $P(\text{diamond})=0.3$ (Prior Probability)

Now, we draw a card from the deck and based on this evidence, we update the probability of getting a diamond (Posterior Probability).



Probability Theory

Probability theory is a branch of mathematics that deals with the study of uncertainty and randomness.

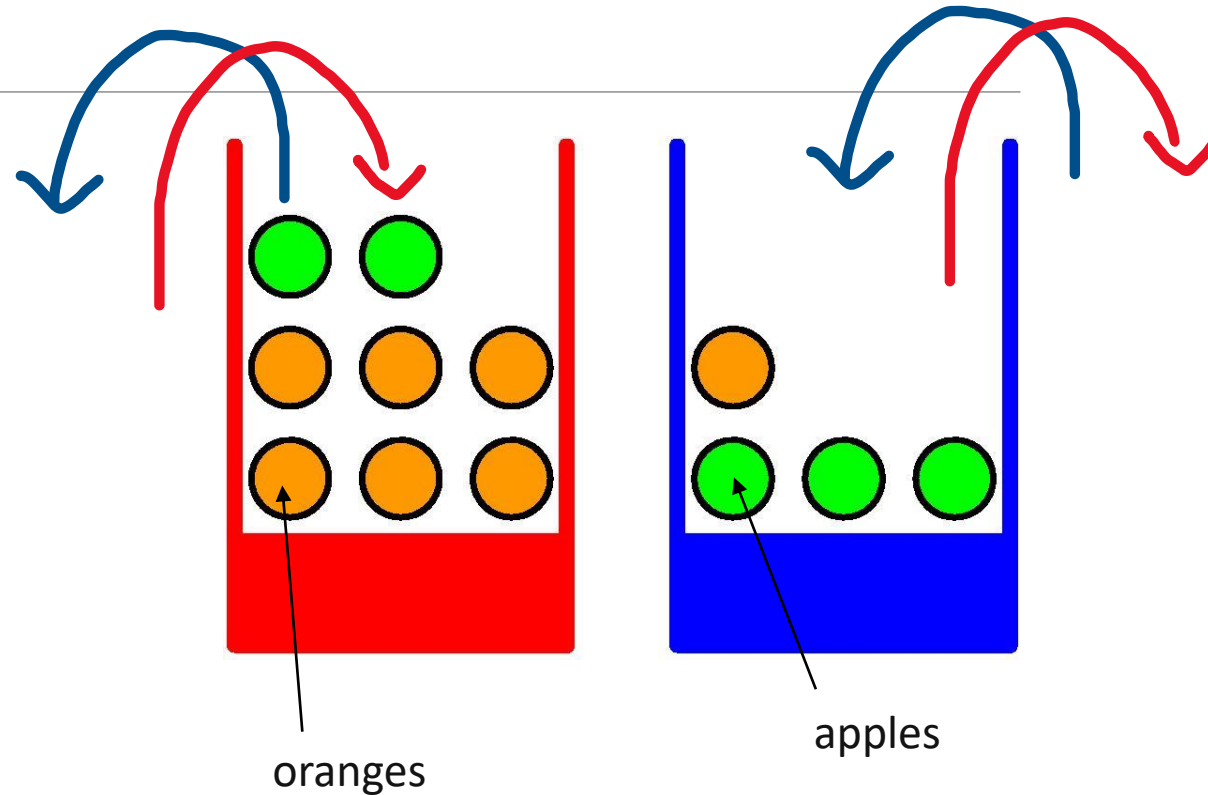
It provides a formal framework for quantifying the likelihood of events or outcomes occurring in various situations.

Fundamental concept in statistics, data analysis, and machine learning.

Provides a formal framework for handling uncertainty in various fields.

Probability Theory

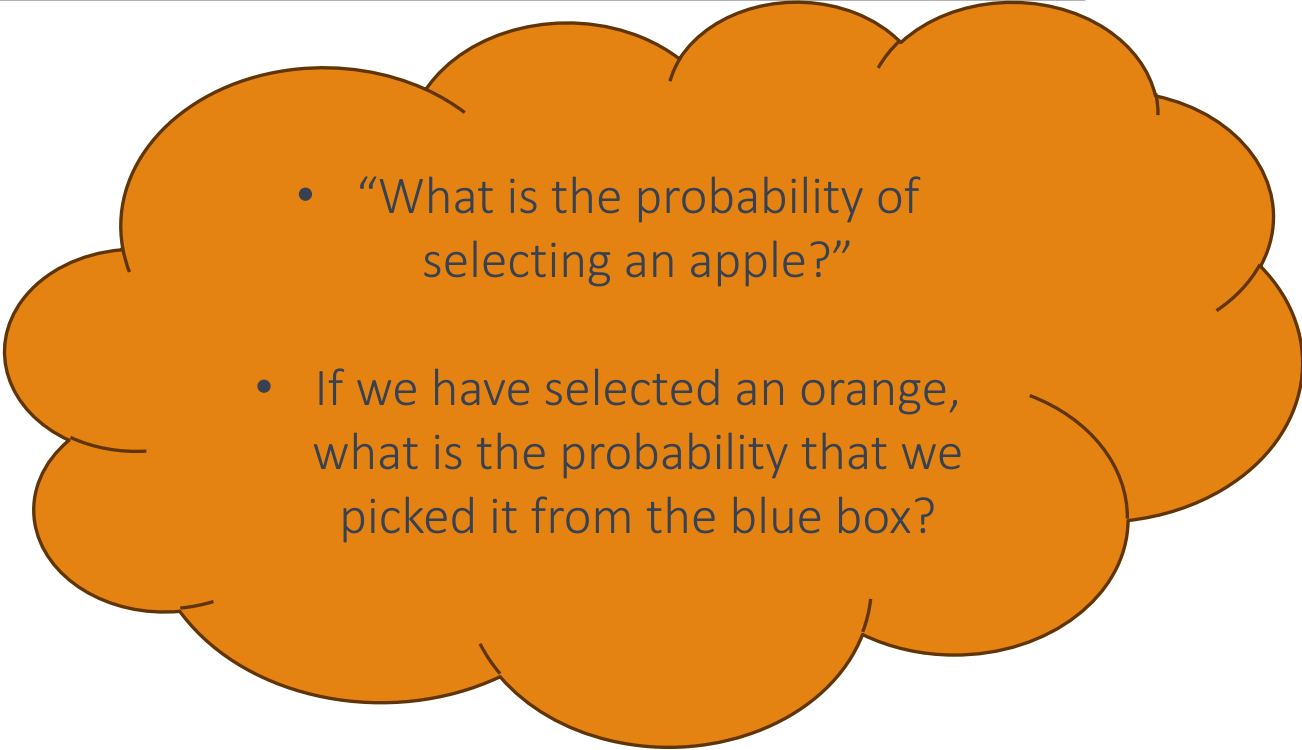
- Randomly pick a box (red or blue) with a 40% chance for the red box and a 60% chance for the blue box.
- From the chosen box, randomly select an item of fruit.
- Observe and note the type of fruit selected.
- Place the selected fruit back into the same box from which it was taken.
- Repeat the process many times.



Probability Theory

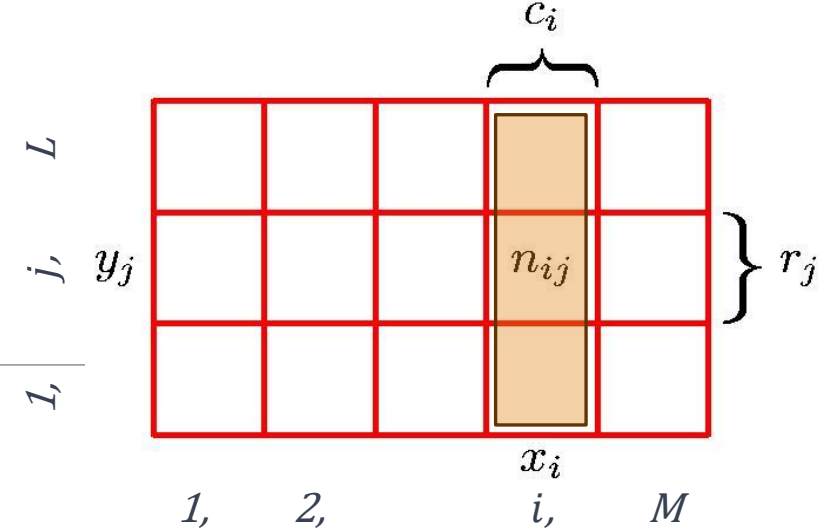
How many random variable*?

- Box that will be chosen (B)
 - Two possible values (red (r) or blue (b)) B
- Fruit will be chosen (F)
 - Two possible values (apple (a) or orange (o))
- $p(B = r) = 40/100$ and $p(B = b) = 60/100$.

- 
- “What is the probability of selecting an apple?”
 - If we have selected an orange, what is the probability that we picked it from the blue box?

*A random variable is a variable in statistics and probability theory that can take on different values based on the outcomes of a random event as opposed to being fixed or deterministic. A random variable is usually denoted by the capital letter.

Probability Theory



- Two random variables X and Y
- Total of N trials
- The count of points in column , where $X = x_i$, is represented as c_i , and the count of points in row j , where, $Y = y_j$ is denoted as r_j .
- Joint Probability (probability of two or more events occurring simultaneously)

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

- Marginal Probability (probability of a single event or outcome, irrespective of the values of other variables)

$$p(X = x_i) = \frac{c_i}{N} = \frac{1}{N} \sum_{j=1}^L n_{ij} = \sum_{j=1}^L p(X = x_i, Y = y_j)$$

$$p(Y = y_j) = \frac{r_j}{N}.$$

sum rule of probability

Probability Theory

➤ Conditional Probability of $Y = y_j$ given $X = x_i$

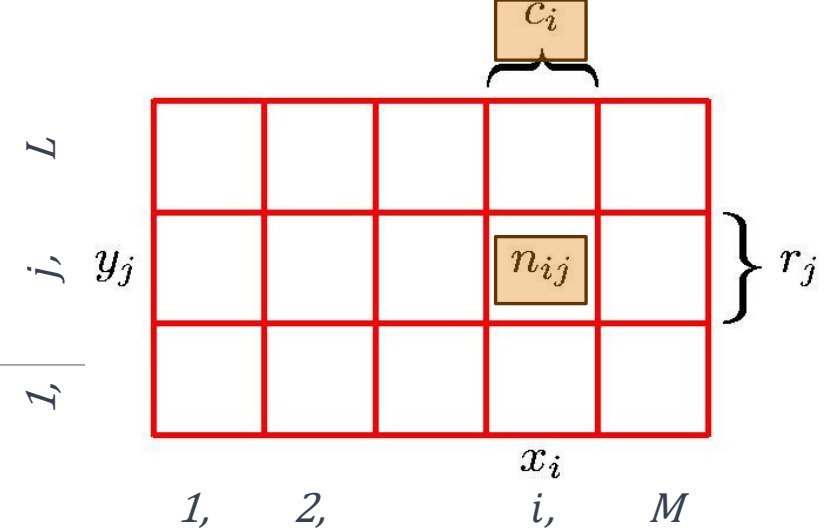
$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

➤ Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= p(Y = y_j | X = x_i) p(X = x_i)$$

Product Rule



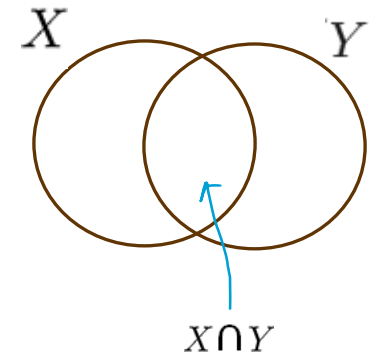
The Rules of Probability

Sum and product rule

- The probability of Y given X $p(Y|X)$
- The probability of X and Y $p(X, Y)$
- Marginal probability of X (the probability of X) $p(X)$

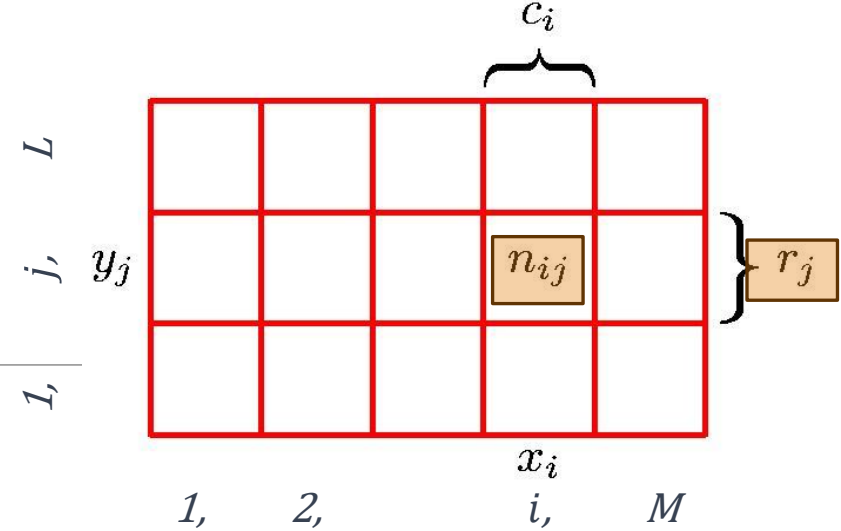
Sum Rule
$$p(X) = \sum_Y p(X, Y)$$

Product Rule
$$\underbrace{p(X, Y)}_{p(X \cap Y)} = p(Y|X)p(X)$$



Exercise

- Derive an expression for $p(Y|X)$
- Now, find the relationship between $p(X|Y)$ and $p(Y|X)$



$$\begin{aligned}
 p(Y = y_j, X = x_i) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{r_j} \cdot \frac{r_j}{N} \\
 &= p(X = x_i | Y = y_j) p(Y = y_j)
 \end{aligned}$$

$$p(Y = y_j, X = x_i) = p(X = x_i, Y = y_j)$$

$$p(X = x_i | Y = y_j) p(Y = y_j) = p(Y = y_j | X = x_i) p(X = x_i)$$

Bayes' Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

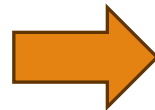
Sum Rule $p(X) = \sum_Y p(X, Y)$



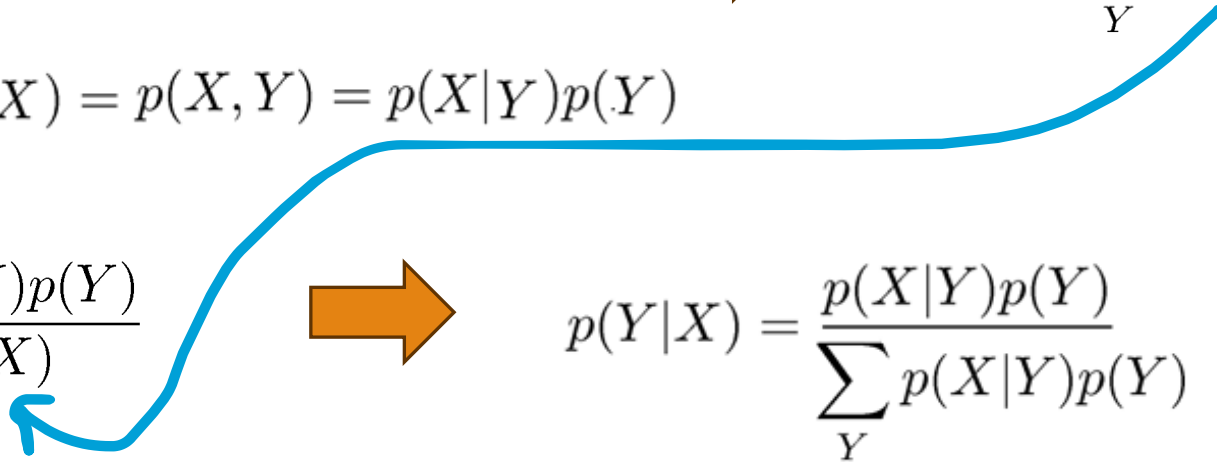
$$p(X) = \sum_Y p(X|Y)p(Y)$$

Product Rule $p(Y, X) = p(X, Y) = p(X|Y)p(Y)$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$



$$p(Y|X) = \frac{p(X|Y)p(Y)}{\sum_Y p(X|Y)p(Y)}$$



Bayes' Theorem

The likelihood of Y under the assumption that event X is true.

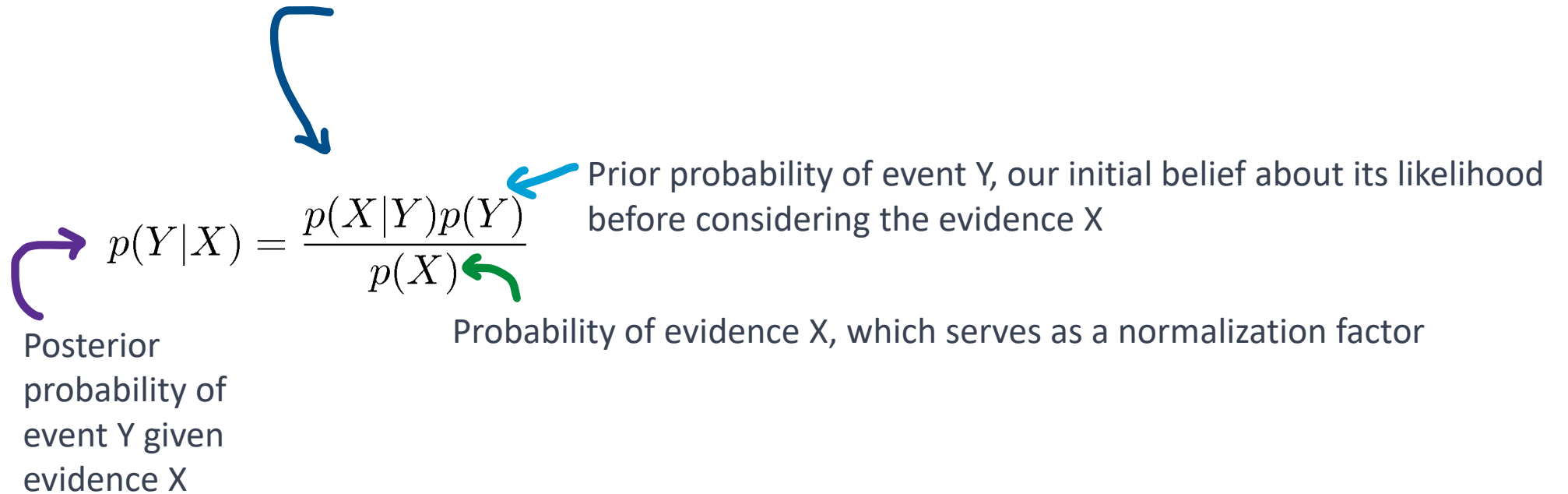
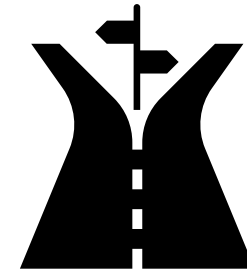


Diagram illustrating Bayes' Theorem with annotations:

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- Posterior probability of event Y given evidence X** (points to $p(Y|X)$)
- The likelihood of Y under the assumption that event X is true.** (points to $p(X|Y)$)
- Prior probability of event Y, our initial belief about its likelihood before considering the evidence X** (points to $p(Y)$)
- Probability of evidence X, which serves as a normalization factor** (points to $p(X)$)

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

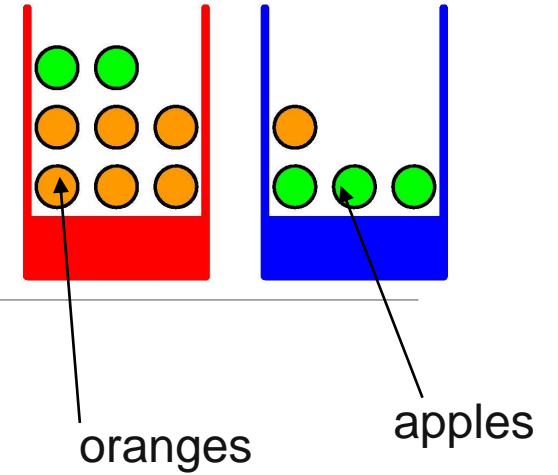


Likelihood

- Likelihood of hypothesis (H) given data (D) is proportional to $P(D|H)$ multiplied by an arbitrary positive constant (K).
- $L(H|D) = K \cdot P(D|H)$.
- Likelihood is not a probability and does not follow the rules of probability.
- Likelihoods do not necessarily sum to 1.
- Example:
 - Data (D): The batsman strike rate in both day matches and night matches.
 - Hypothesis 1 (H 1): The hypothesis is that the batsman's scoring rate is higher during day matches.
 - Hypothesis 2 (H 2): The hypothesis is that the batsman's scoring rate is higher during night matches.
 - Likelihood ($L(H1|D)$) $\propto P(D|H1)$. The likelihood of our hypothesis (batsman scores faster during day matches) given the observed data.
 - Likelihood ($L(H2|D)$) $\propto P(D|H2)$. The likelihood of our hypothesis (batsman scores faster during night matches) given the observed data.

Data: Strike rate for 50 Day Matches: = 1.2 runs per ball, Strike rate for 50 For Night Matches:= 0.90 run per ball

Bayes' Theorem Example



- $p(r) = 0.4$, $p(b) = 0.6$
 - Calculate $p(a|r)$, $p(o|r)$, $p(a|b)$ and $p(o|b)$
 - $p(a|r) = 2/8$, $p(o|r) = 6/8$, $p(a|b) = 3/4$, $p(o|b) = 1/4$
 - Overall probability of choosing an orange ?
 - $p(r) \times p(o|r) + p(b) \times p(o|b) = 0.4 \times 6/8 + 0.6 \times 1/4 = 9/20$
 - We have chosen an orange from one of the boxes, and now we want to determine which box it originated from.
 - $p(r|o)$? $p(b|o)$?
 - $p(r|o) = p(o|r)p(r)/p(o) = 6/8 \times 0.4 / (9/20) = 2/3$ and $p(b|o) = 1 - p(r|o) = 1/3$.
 - Can we interpret this result?
-
- Prior probability $p(r) = 0.4$, $p(b) = 0.6$ (The probability is available before we know which fruit is selected (orange or apple)) → more likely to select blue box.
 - Posterior probability $p(r|o)$ → probability after we know the selected fruit is orange → it is more likely to select red box, once we know the selected fruit is orange.

Bayes' Theorem

We start with a prior probability, which represents our initial belief or knowledge about an event before observing any data. As new information or evidence is collected, we update the probabilities using Bayes' theorem to obtain the posterior probability. The posterior probability represents our updated belief after incorporating the new information.

Bayesian theorem provide a powerful framework to handle uncertainty, incorporate prior knowledge, and make better decisions based on probabilistic reasoning.

Independence

- $P(X) = P(X|Y)$ → Random variable X is independence of Y.

The knowledge of Y does not change the probability distribution of X.

- If $P(X) = P(X|Y)$ → Joint Probability $P(X, Y) = P(X)P(Y)$.

- Conditional independence

- $P(X|Z) = P(X|Y, Z)$ or $P(X, Y | Z) = P(X | Z) P(Y | Z)$.

Two random variables X and Y are conditionally independent given a third random variable Z if, once we know the value of Z, the relationship between X and Y becomes independent.

Probability Density function

- PDF provides the probability distribution for the random continuous variable over its entire range*.

$$p(x) \geq 0 \quad \int_{-\infty}^{\infty} p(x) dx = 1$$

- Probability that x will lie in an interval (a, b)

$$p(x \in (a, b)) = \int_a^b p(x) dx$$

- The Cumulative Distribution Function (CDF)

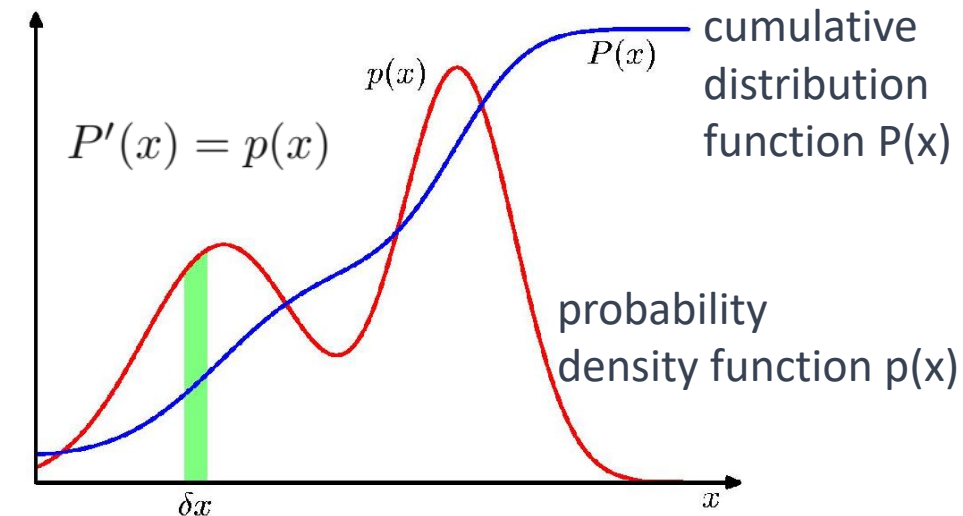
$$P(z) = \int_{-\infty}^z p(x) dx$$

- Cumulative probability that a random variable takes on a value less than or equal to a given value.

- Sum and product rule

$$p(x) = \int p(x, y) dy$$

$$p(x, y) = p(y|x)p(x).$$



*For discrete variable, probability mass function (PMF), It gives the probability of a discrete random variable X taking on a specific value.
The PMF is defined for all possible values of the random variable

Expectations

$$\mathbb{E}[f] = \sum_x p(x) f(x)$$

discrete distribution

$$\mathbb{E}[f] = \int p(x) f(x) dx$$

continuous variables

- The expectation of a function $f(x)$ under a probability distribution $p(x)$ is the average value of $f(x)$ when x is drawn from the distribution $p(x)^*$.

- N of points drawn from the probability distribution

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^N f(x_n)$$

Approximate Expectation
(discrete and continuous)

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x)$$

Conditional Expectation (discrete)

Ex:- X be outcomes of a roiling of a die

$E(X) = ?$

$E(X) = (1) \times (1/6) + (2) \times (1/6) + \dots + 6 \times (1/6)$

*Also known as mean, expected value, or first moment.

Expectations

- Linearity of Expectations (For independent or dependent random variables)

$$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$$

- For independent variables $E(XY) = E(X)E(Y)$

Variances and Covariances

- Variance*: It measures how far the data points are from the mean of the data set. In other words, variance indicates the variability or diversity of the data points around the average.

$$\text{var}[f] = \mathbb{E} \left[(f(x) - \mathbb{E}[f(x)])^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- Covariance

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y} [\{x - \mathbb{E}[x]\} \{y - \mathbb{E}[y]\}] & \text{cov}[\mathbf{x}, \mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}} [\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\} \{\mathbf{y}^T - \mathbb{E}[\mathbf{y}^T]\}] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] & &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^T] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^T]. \end{aligned}$$

For two random variables x and y

two vectors of random variables \mathbf{x} and \mathbf{y}

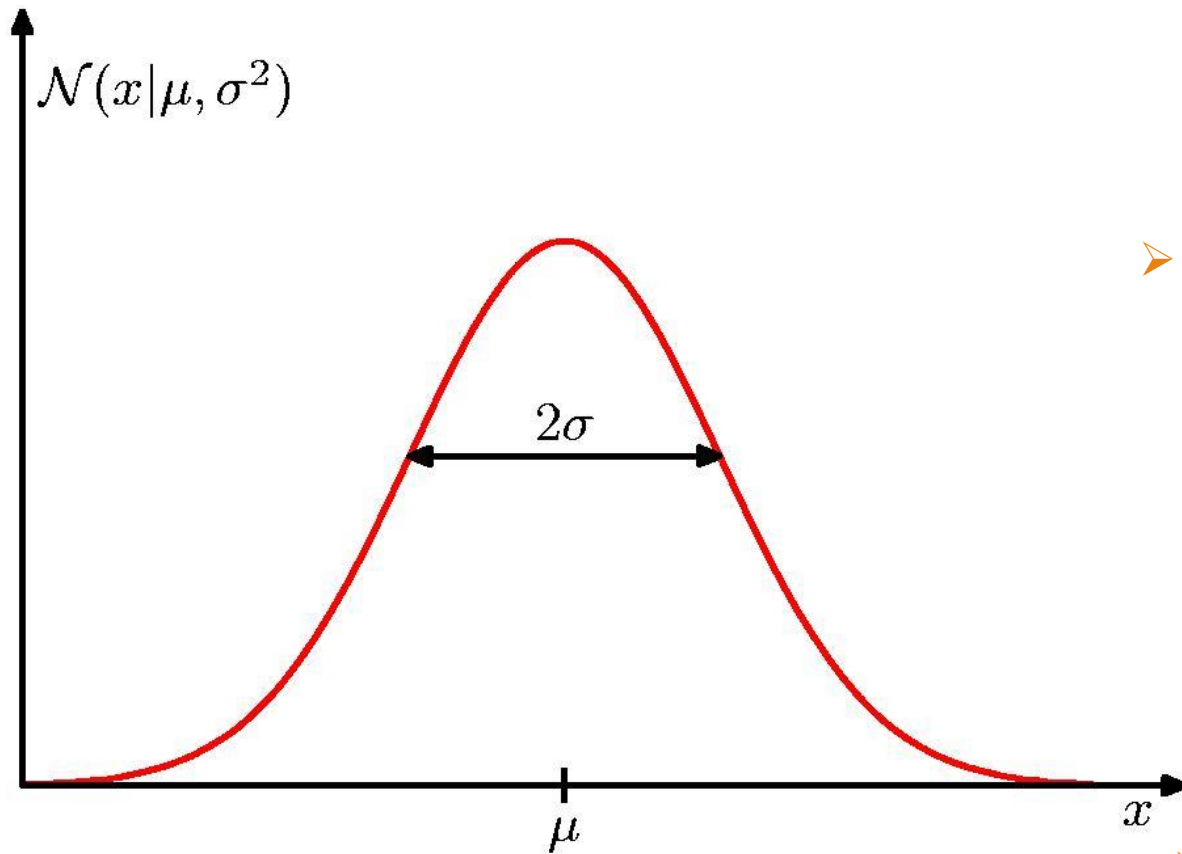
- Covariance measures the degree to which two variables, x and y , vary together. When x and y are independent, their covariance is zero. In other words, if the values of x and y do not show any systematic relationship, the covariance between them is negligible.

Variances and Covariances

- $\text{Var}(X+b)$ for constant b
- $\text{Var}(X+b) = \text{Var}(X)$
- Why?
 - It's important to note that the constant " b " does not contribute to the variance since it is just a shift or translation of the data along the X-axis, which doesn't change the spread or variability of the data.
- $\text{Var}(aX+b) = a^2 \text{Var}(X)$ for constant a and b
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ (if X and Y are independent)

Not linear as expectation

Normal or Gaussian distribution



This distribution is defined by its mean (μ) and standard deviation (σ), and it is symmetric around the mean.

- For a single real-valued variable x , the Gaussian distribution is given by

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\}$$

$$\mathcal{N}(x|\mu, \sigma^2) > 0$$

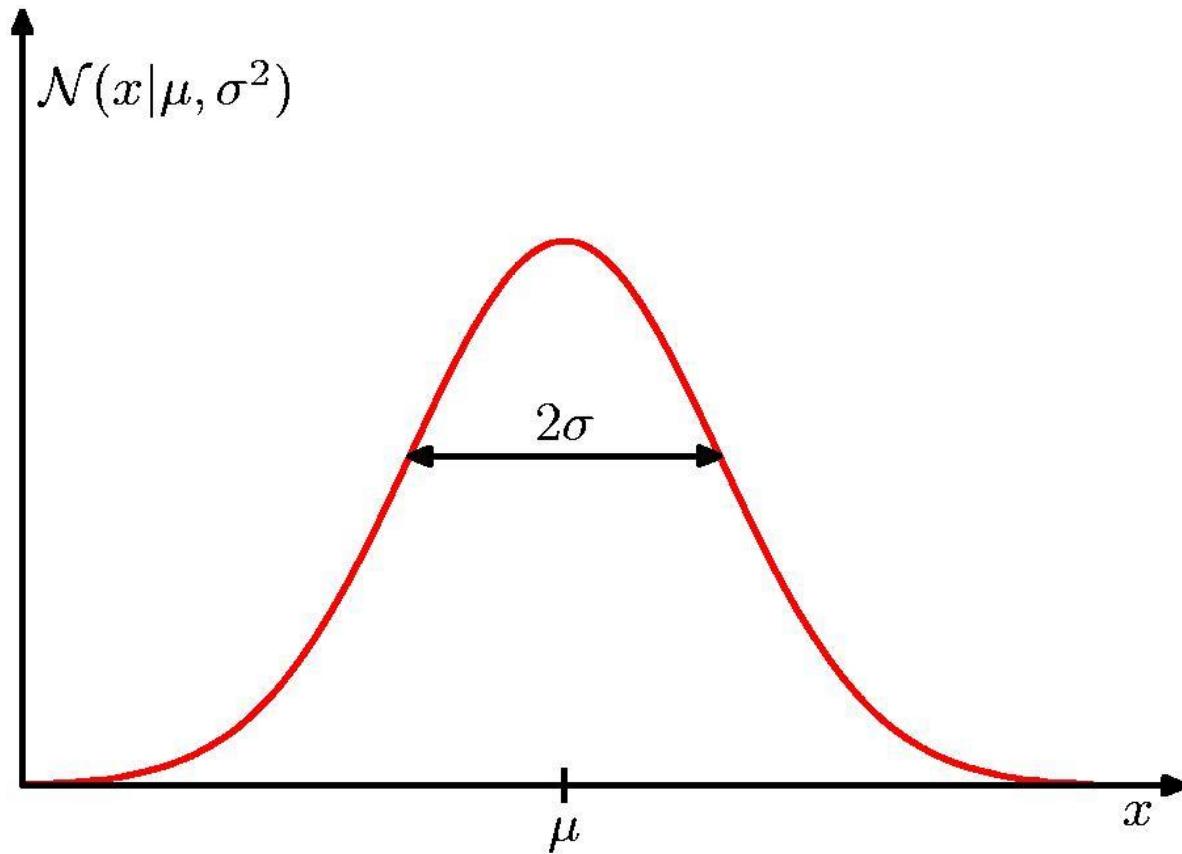
$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) dx = 1$$

- Precision

$$\beta = 1/\sigma^2$$

The Gaussian distribution is a fundamental and widely used probability distribution in statistics and data analysis due to its many important properties and widespread occurrence in natural phenomena

Normal or Gaussian distribution

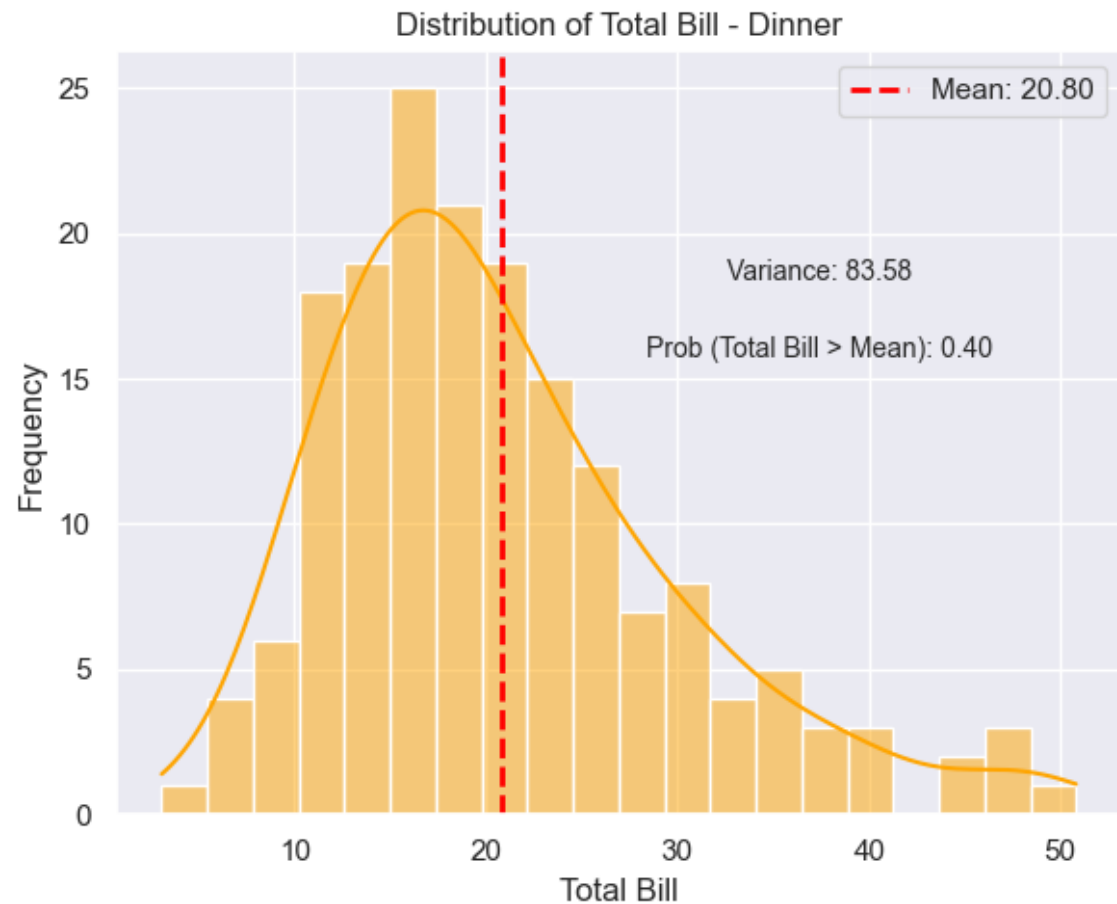
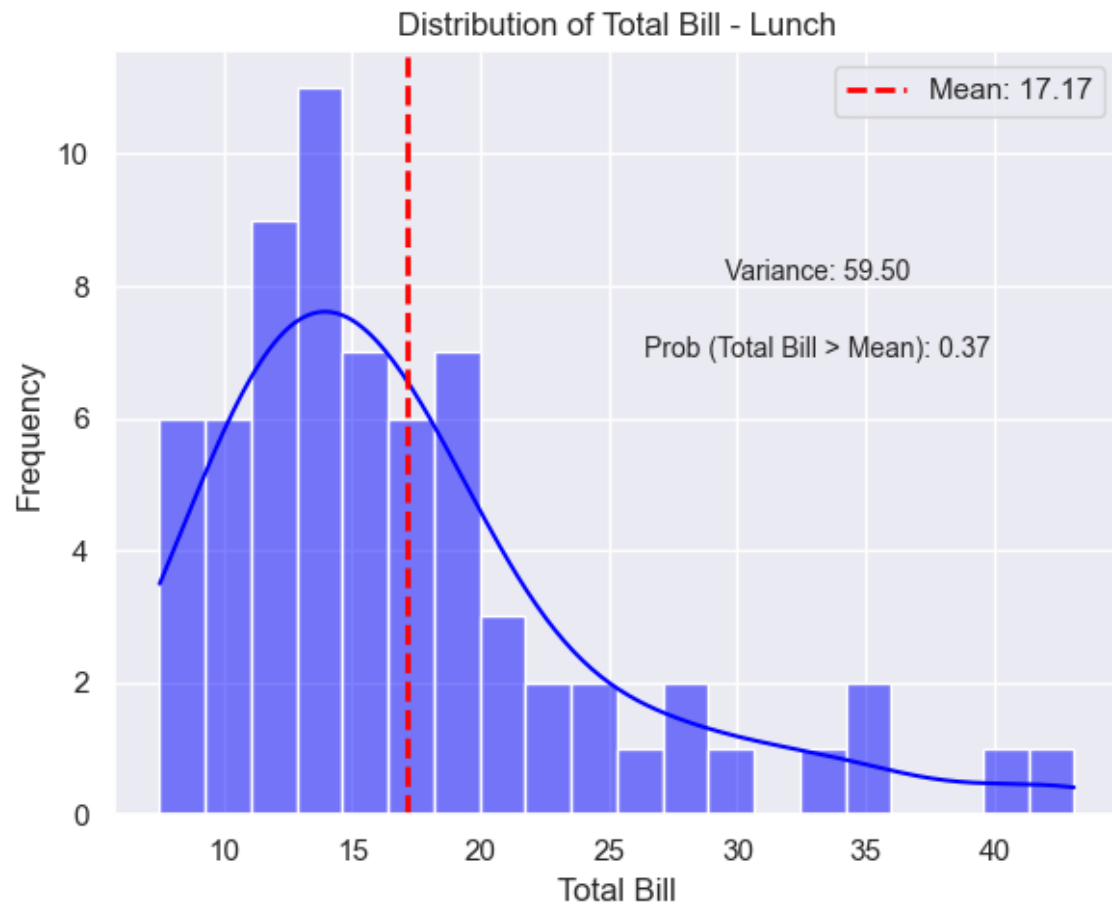


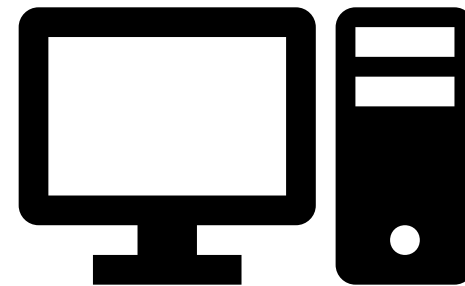
$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x \, dx = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}(x|\mu, \sigma^2) x^2 \, dx = \mu^2 + \sigma^2$$

$$\text{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Variances and Covariances





Code

```
import seaborn as sns

# Load the tips dataset
tips = sns.load_dataset("tips")

# Display the first few rows of the
dataset
print(tips.head())
```

	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3

Variables:

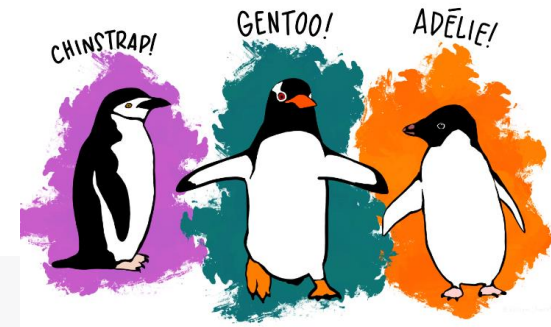
- 1.total_bill: The total bill amount, including the cost of food and drinks, for each customer (numeric).
- 2.tip: The tip amount given by each customer (numeric).
- 3.sex: The gender of the customer (categorical: "Male" or "Female").
- 4.smoker: Whether the customer is a smoker or not (categorical: "Yes" or "No").
- 5.day: The day of the week when the meal was taken (categorical: "Thur", "Fri", "Sat", "Sun").
- 6.time: The time of the day when the meal was taken (categorical: "Lunch" or "Dinner").
- 7.size: The number of people in the dining party (numeric).



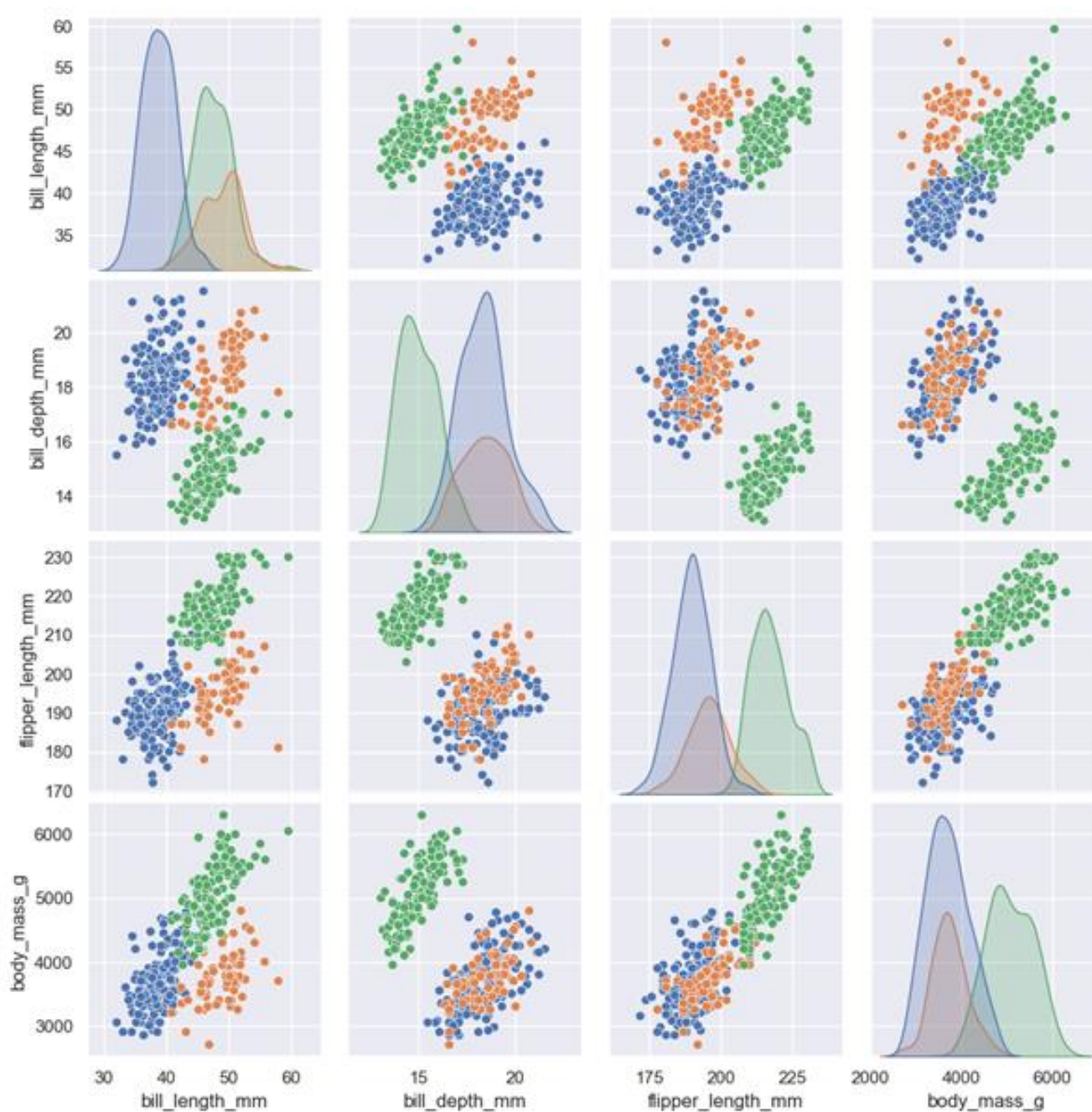
Code

```
import seaborn as sns
import matplotlib.pyplot as plt
# Load the tips dataset
tips = sns.load_dataset("tips")
# Create separate dataframes for lunch and dinner
lunch_data = tips[tips["time"] == "Lunch"]
dinner_data = tips[tips["time"] == "Dinner"]
# Set up the figure with two subplots side by side
fig, axes = plt.subplots(1, 2, figsize=(12, 5))
# Plot histogram for lunch data
sns.histplot(lunch_data["total_bill"], kde=True, color='blue', bins=20, ax=axes[0])
mean_lunch = lunch_data["total_bill"].mean()
axes[0].axvline(mean_lunch, color='red', linestyle='dashed', linewidth=2, label=f"Mean: {mean_lunch:.2f}")
axes[0].set_xlabel("Total Bill")
axes[0].set_ylabel("Frequency")
axes[0].set_title("Distribution of Total Bill - Lunch")
axes[0].legend()
# Plot histogram for dinner data
sns.histplot(dinner_data["total_bill"], kde=True, color='orange', bins=20, ax=axes[1])
mean_dinner = dinner_data["total_bill"].mean()
axes[1].axvline(mean_dinner, color='red', linestyle='dashed', linewidth=2, label=f"Mean: {mean_dinner:.2f}")
axes[1].set_xlabel("Total Bill")
axes[1].set_ylabel("Frequency")
axes[1].set_title("Distribution of Total Bill - Dinner")
axes[1].legend()
#Show the plot
plt.tight_layout()
plt.show()
```

Variances and Covariances



species: The species of the penguin
 bill_length_mm: The length of the penguin's bill in millimeters
 bill_depth_mm: The depth of the penguin's bill in millimeters
 flipper_length_mm: The length of the penguin's flipper in millimeters
 body_mass_g: The body mass of the penguin in grams

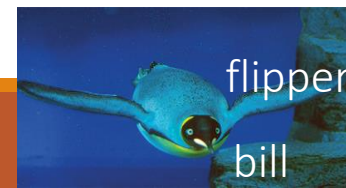


Cov(flipper_length_mm, bill_length_mm) 5.67
 Cov(bill_depth_mm, bill_length_mm) 1.27
 Cov(bill_depth_mm, flipper_length_mm) 2.44
 Cov(flipper_length_mm, body_mass_g) 1404.03
 Cov(body_mass_g, bill_length_mm) 670.35
 Cov(body_mass_g, bill_depth_mm) 321.43

"Adelie"

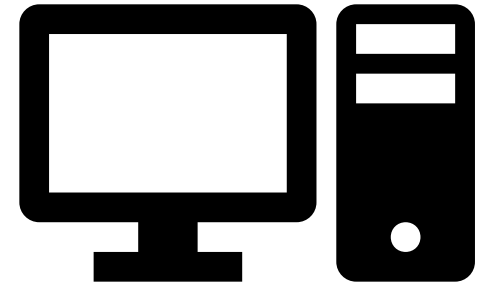
Cov(flipper_length_mm, bill_length_mm) 13.21
 Cov(bill_depth_mm, bill_length_mm) 1.94
 Cov(bill_depth_mm, flipper_length_mm) 4.49
 Cov(body_mass_g, flipper_length_mm) 2297.14
 Cov(body_mass_g, bill_length_mm) 1039.62
 Cov(body_mass_g, bill_depth_mm) 355.69

"Gentoo"



species ● Adélie ● Chinstrap ● Gentoo

<https://allisonhorst.github.io/palmerpenguins/>



Code

```
import seaborn as sns
import matplotlib.pyplot as plt

# Apply the default theme
sns.set_theme()

penguins = sns.load_dataset("penguins")

# Calculate covariance matrix and variance for each variable
cov_matrix = penguins.cov()
variance_values = penguins.var()

# Create the pairplot with covariance plots
g = sns.pairplot(data=penguins, hue="species", diag_kind="kde")

plt.show()
```

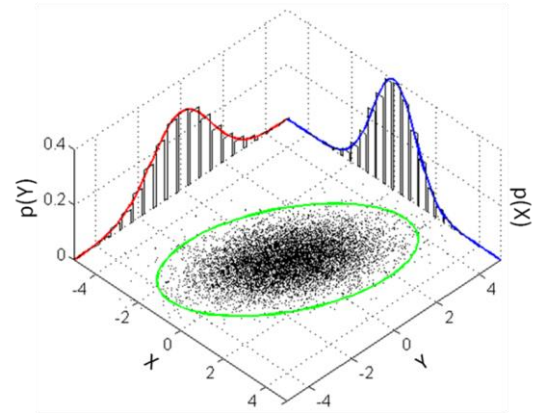
The Multivariate Gaussian

- Gaussian distribution defined over a D -dimensional vector \mathbf{x} of continuous variable

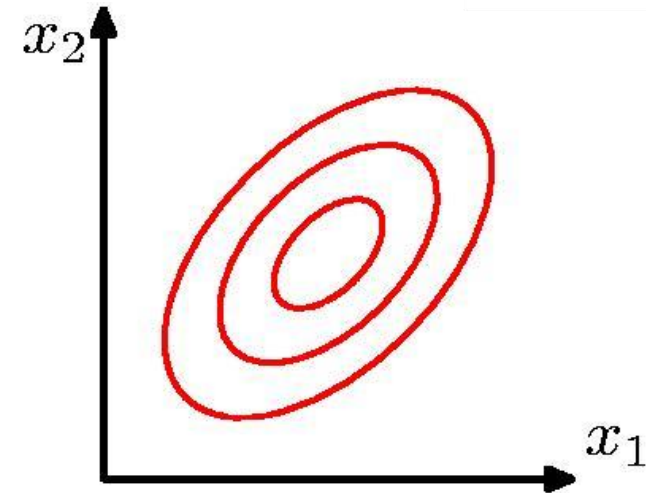
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\}$$

determinant of $\boldsymbol{\Sigma}$

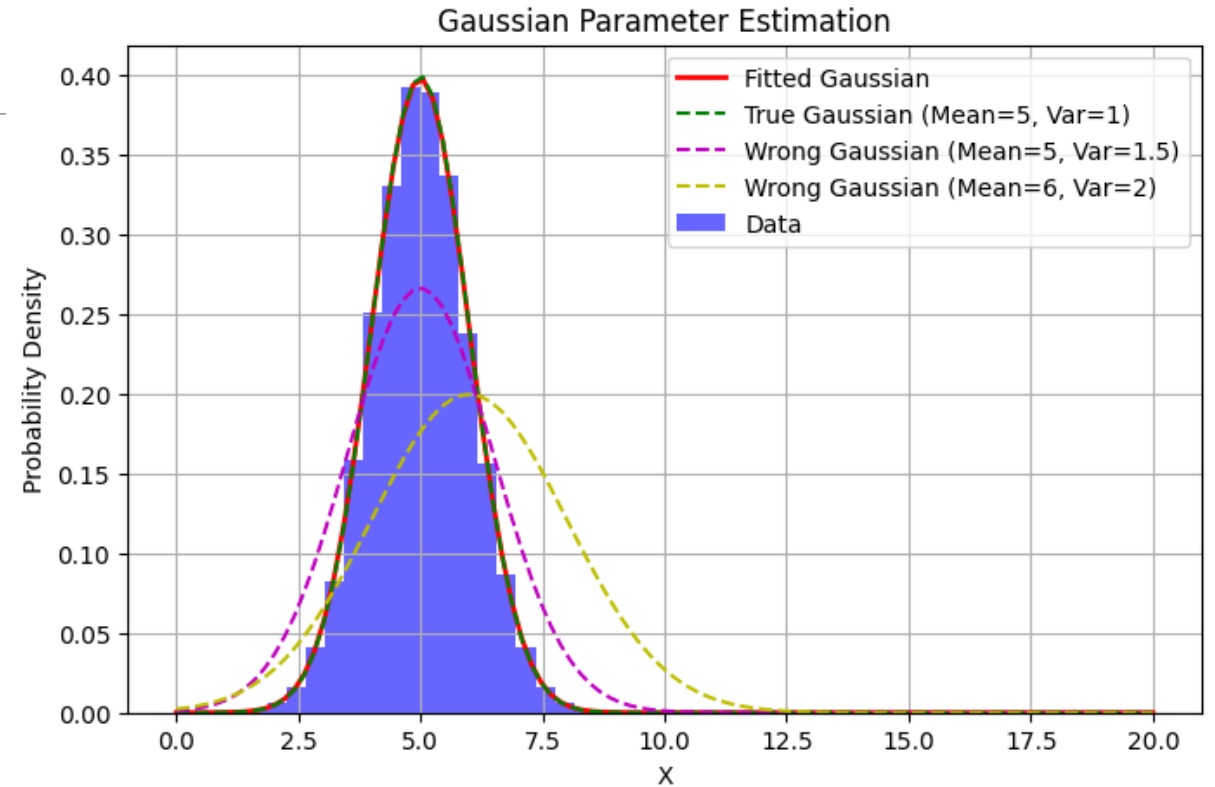
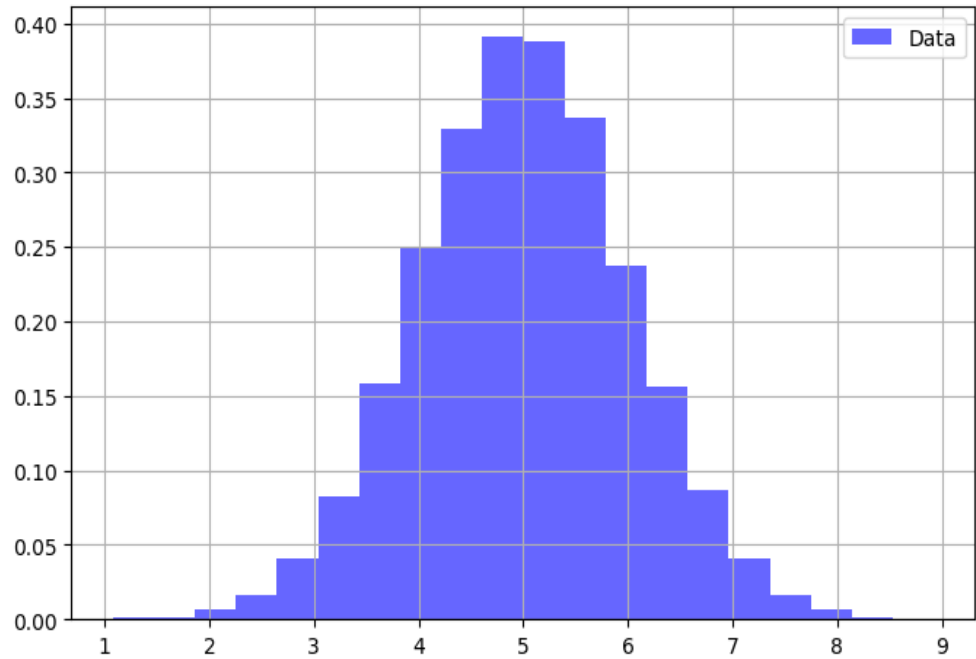
- $\boldsymbol{\Sigma}$ Covariance matrix of size $D \times D$.



- The graph of the multivariate Gaussian distribution is a hyperellipsoid in the multidimensional space.
- In a 2-dimensional case (two variables), the graph of the multivariate Gaussian distribution is an ellipse. As you move to higher dimensions, the graph becomes a hyperellipsoid. The shape of the hyperellipsoid is determined by the covariance matrix.



Gaussian Parameter Estimation



Suppose the observations are independently drawn from a Gaussian distribution with unknown mean μ and variance σ^2 , and we aim to estimate these parameters from the given dataset.

Gaussian Parameter Estimation

- Suppose the observations are independently drawn from a Gaussian distribution with unknown mean μ and variance σ^2 , and we aim to estimate these parameters from the given dataset.

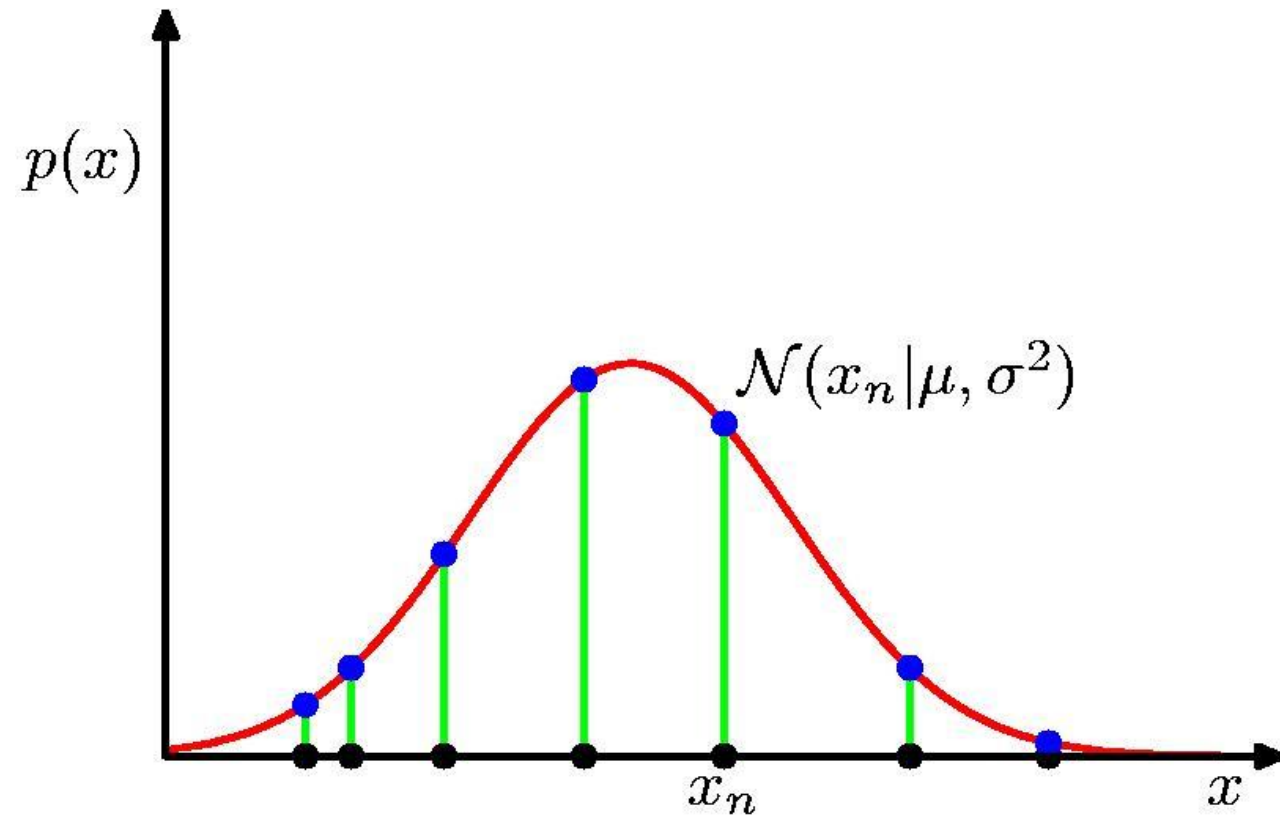
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x - \mu)^2 \right\}$$

- The data points are considered independent and identically distributed (i.i.d.). PDF of the data set is given by

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

- When considering it as a function of μ and σ^2 , this is the likelihood function for the Gaussian distribution.

Gaussian Parameter Estimation



- The likelihood function corresponds to the product of the observed blue values

$$p(\mathbf{x} | \mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n | \mu, \sigma^2)$$

The common approach for estimating the parameters in a probability distribution from an observed dataset is to identify the parameter values that maximize the likelihood function.

Here, the black points denote a data set of values $\{x_n\}$

Gaussian Parameter Estimation

Maximum (Log) Likelihood

$$\ln p(\mathbf{x}|\mu, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi)$$

$$p(\mathbf{x}|\mu, \sigma^2) = \prod_{n=1}^N \mathcal{N}(x_n|\mu, \sigma^2)$$

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\}$$

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad \sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2$$

$$\mathbb{E}[\mu_{\text{ML}}] = \mu$$

- ML estimates correct mean but not the correct variance.
- As $N \rightarrow \infty$ Estimated variance \rightarrow true variance

$$\mathbb{E}[\sigma_{\text{ML}}^2] = \left(\frac{N-1}{N}\right) \sigma^2$$

*In practice, maximizing the log of the likelihood function is preferred for convenience. Since the logarithm is a monotonically increasing function, maximizing the log of a function is equivalent to maximizing the original function.

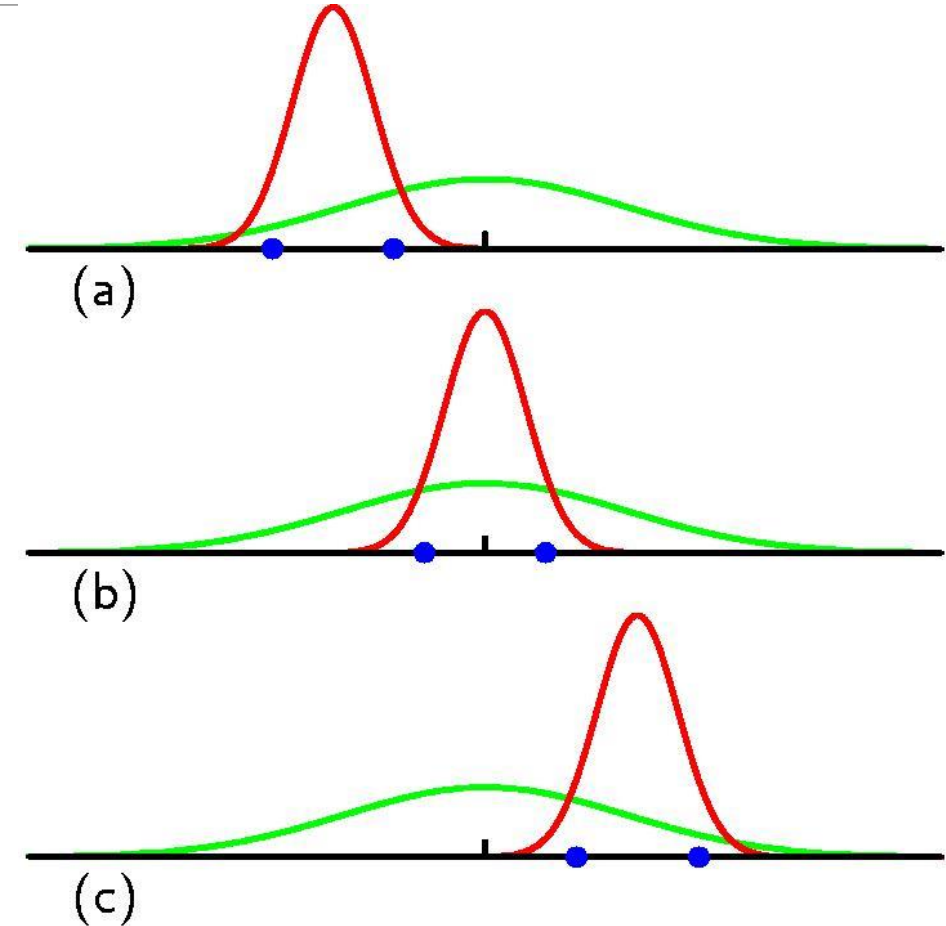
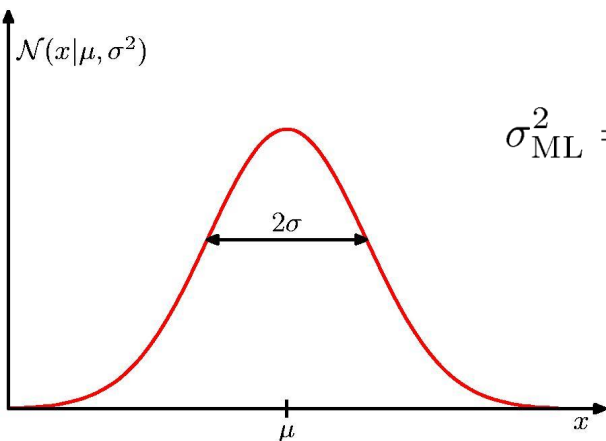
Gaussian Parameter Estimation

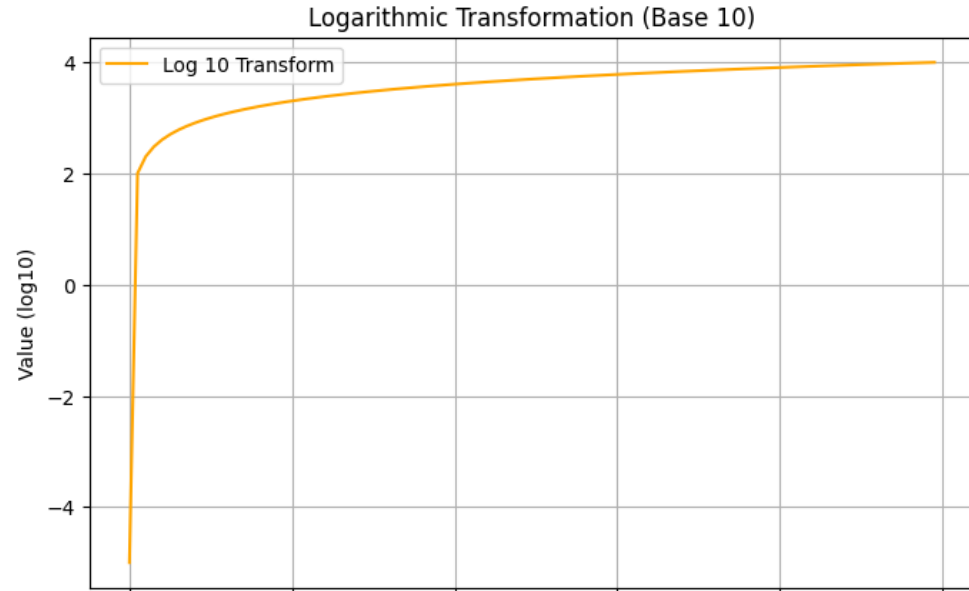
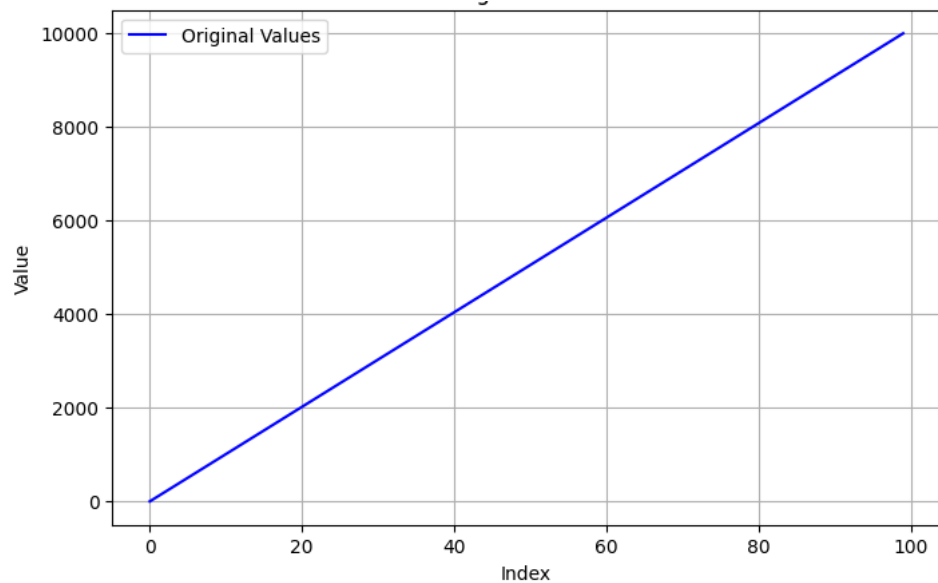
Maximum (Log) Likelihood

- Green curve: Represents the true Gaussian distribution.
- Three red curves: Show estimated distributions from three data sets.
- Each data set has two data points (shown in blue).
- Mean estimates are correct (**averaged across data sets**).
- Variance is systematically under-estimated due to using the sample mean instead of the true mean.

$$\mu_{\text{ML}} = \frac{1}{N} \sum_{n=1}^N x_n \quad \mathbb{E}[\mu_{\text{ML}}] = \mu$$

$$\sigma_{\text{ML}}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{\text{ML}})^2 \quad \mathbb{E}[\sigma_{\text{ML}}^2] = \left(\frac{N-1}{N} \right) \sigma^2$$





Log-transform

The logarithmic transformation is particularly valuable when dealing with likelihood functions, as it converts products of probabilities into sums of logarithmic probabilities.

- In machine learning, we assume independence of different samples.
 - This leads to dealing with products of a large number of distributions.
 - To optimize functions of these products, working with logarithms is easier.
 - The logarithmic function is strictly increasing and preserves the maximum location.
- Useful for dealing with very small or very large values.

Binary Variables

➤ Binary variables are categorical variables that can take on one of two distinct values, typically represented as 0 and 1. These variables are commonly used to represent yes/no or true/false type of information.

1. Married: 1 (married) or 0 (not married)
2. Vaccinated: 1 (vaccinated) or 0 (not vaccinated)
3. Coin flipping: 1(heads) or 0 (tails)

$$p(x = 1|\mu) = \mu$$

The probability of a binary variable x taking the value 1 will be represented by the parameter μ .

$$p(x = 0|\mu) = 1 - \mu$$

Binary Variables

Bernoulli Distribution

- The Bernoulli distribution is a fundamental discrete probability distribution that models a random variable that can take on one of two possible outcomes, typically represented as 0 and 1. It is named after Swiss mathematician Jacob Bernoulli, who introduced the concept in the early 18th century.

$$x \in \{0, 1\}$$

$$\text{Bern}(x|\mu) = \mu^x (1 - \mu)^{1-x}$$

$$\mathbb{E}[x] = \mu$$

$$\text{var}[x] = \mu(1 - \mu)$$

probability of getting a head = 0.6,
probability of getting a tail $1 - p = 0.4$.

Probability of getting a head ($X = 1$):
 $P(X = 1) = 0.6^1 * (1 - 0.6)^{(1 - 1)} = 0.6$

Probability of getting a tail ($X = 0$):
 $P(X = 0) = 0.6^0 (1 - 0.6)^{(1 - 0)} = 0.4$

Binomial Distribution

➤ The binomial distribution is a discrete probability distribution that models the number of successes (say $x=1$) in a **fixed number of independent Bernoulli trials (N)**. It is named "binomial" because it involves two parameters: the number of trials (N) and the probability of success in each trial (p).

➤ N coin flips: $p(m \text{ heads} | N, \mu)$

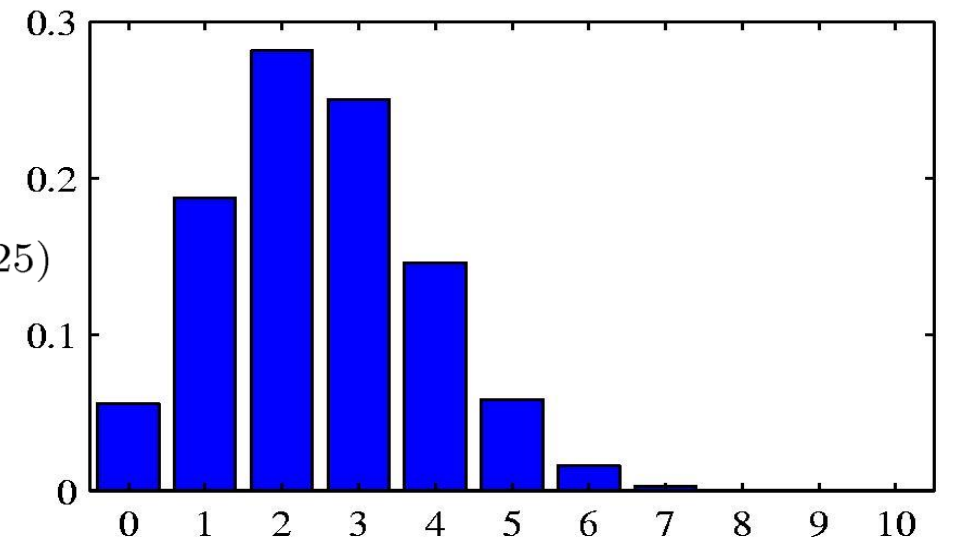
$$\text{Bin}(m | N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^N m \text{Bin}(m | N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^N (m - \mathbb{E}[m])^2 \text{Bin}(m | N, \mu) = N\mu(1 - \mu)$$

$\text{Bin}(m | 10, 0.25)$



binomial distribution for $N = 10$ and $\mu = 0.25$

Binomial Distribution

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m}$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!}$$

$$\mathbb{E}[m] \equiv \sum_{m=0}^N m \text{Bin}(m|N, \mu) = N\mu$$

$$\text{var}[m] \equiv \sum_{m=0}^N (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N\mu(1 - \mu)$$

Probability of getting "Heads" (success) is 0.6. For N=10 calculate probability of getting head for 5 times (m=5).

$$\text{Bin}(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m} = 0.2007$$



Thank You

Q & A