# SNS COLLEGE OF ENGINEERING

Kurumbapalayam (Po), Coimbatore – 641 107

**An Autonomous Institution**

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

## DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

## COURSE NAME : 19CS501 Introduction to Machine Learning

III YEAR /V SEMESTER

Unit 1- Introduction

Topic : Probability theory – Probability Distributions – Decision Theory

# Probability

- ❑ When we study real world processes we want to learn about numerous random events that distort our experiments. Uncertainty is everywhere and we must tame it to be used for our needs. That is when probability theory and statistics come into play

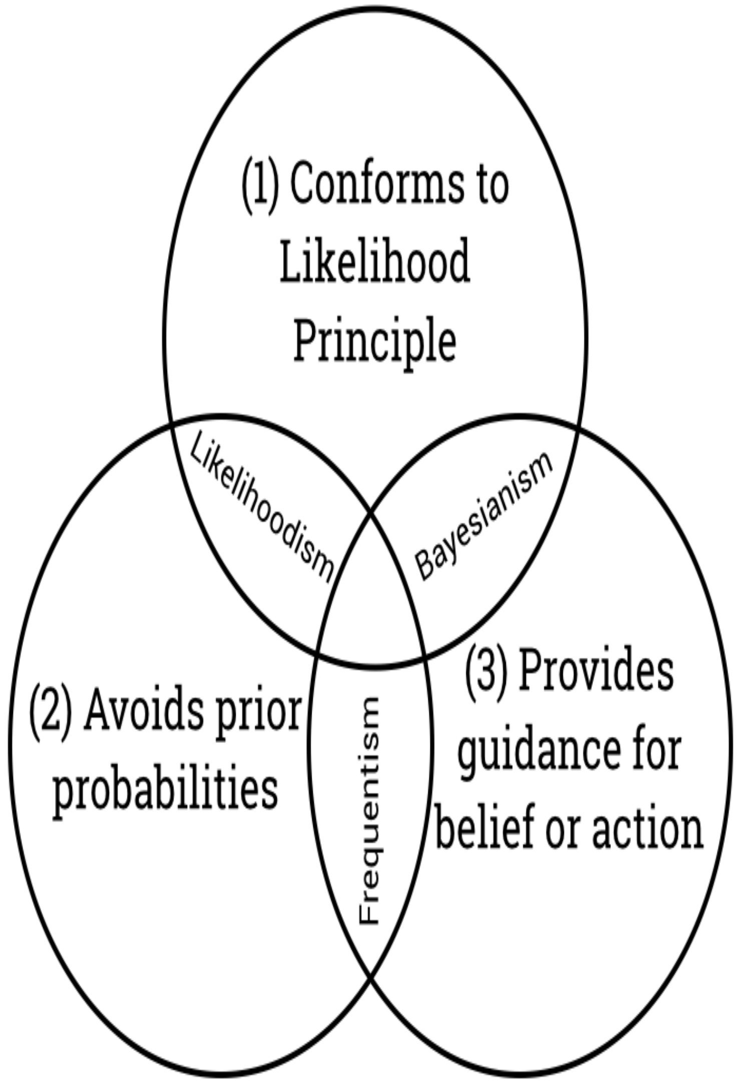
***Now a days those disciplines lie in the center of artificial intelligence, particle physics, social science, bio-informatics and in our everyday lives***

# Frequentist probabilities

- ❑ Imagine we were given a coin and want to check whether it is fair or not. How do we approach this? Let's try to conduct some experiments and record 1 if heads come up and 0 if we see tails.

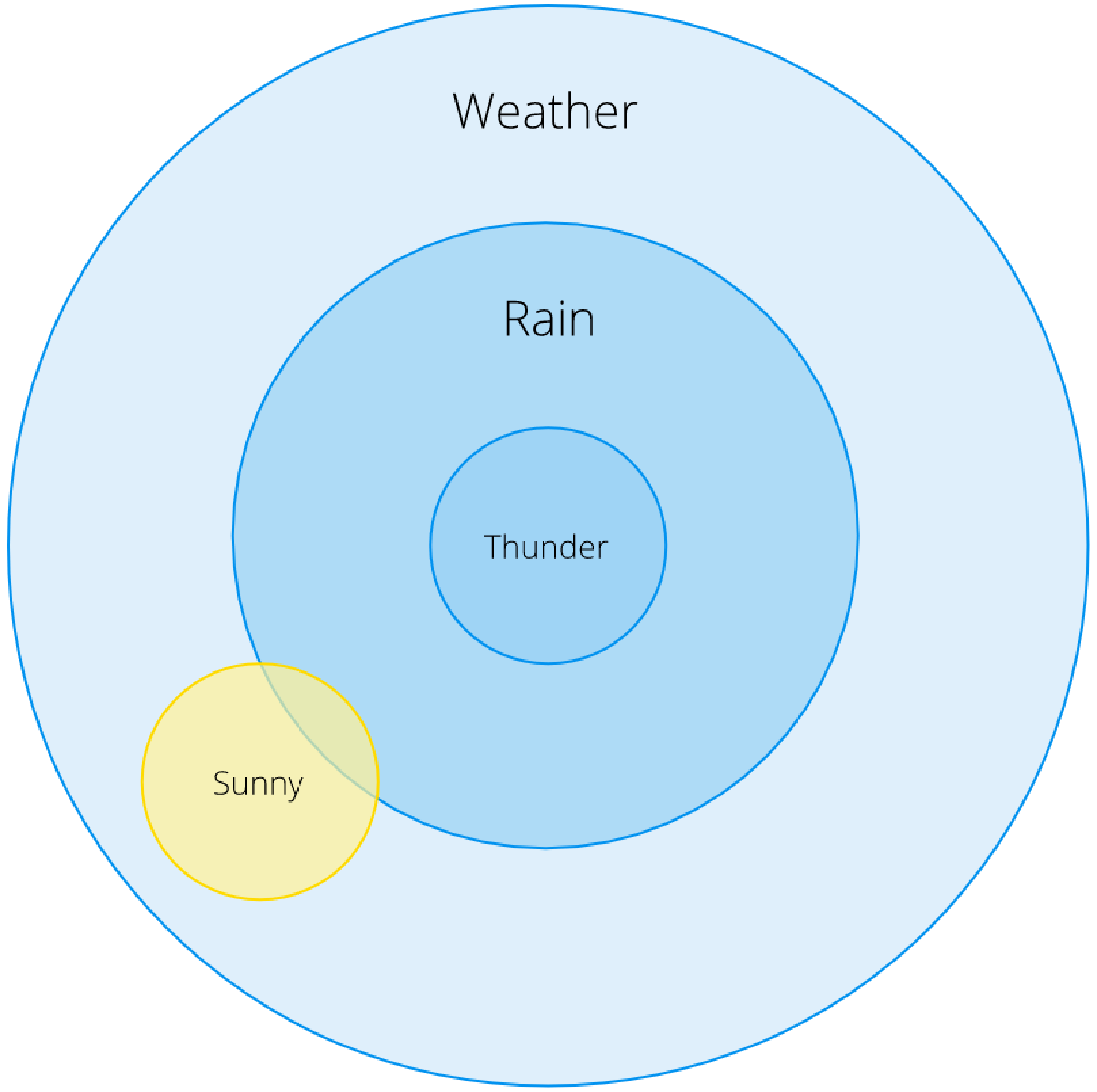
- ❑ Repeat this 1000 tosses and count each 0 and 1. After we had some tedious time experimenting, we got those results: 600 heads (1s) and 400 tails (0s).

- ❑ If we then count how frequent heads or tails came up in the past, we will get 60% and 40% respectively. Those frequencies can be interpreted as probabilities of a coin coming up heads or tails. This is called a frequentist view on the probabilities.


(1) Conforms to Likelihood Principle
Likelihoodism
Bayesianism
(2) Avoids prior probabilities
Frequentism
(3) Provides guidance for belief or action

# Conditional probabilities

Frequently we want to know the probability of an event given some other event has occurred. We write conditional probability of an event A given event B as P(A | B). Take rains for example:



What is the probability of a rain given we see thunder
What is the probability of a rain given it is sunny?

$$P(Rain|Sunny) = \frac{P(Rain, Sunny)}{P(Sunny)}$$

# Dependent and independent events

Events are called independent if the probability of one event does not influence the other in any way. Take for example the probability of rolling a dice and getting a 2 for the first time and for the second time. Those events are independent.

$$P(roll2) = P(roll2_{\text{1st time}})P(roll2_{\text{2nd time}})$$

First, let’s rename events for 1st and 2nd tosses as A and B to remove notational clutter and then rewrite probability of a r oll explicitly as joint probability of both rolls we had seen so far:

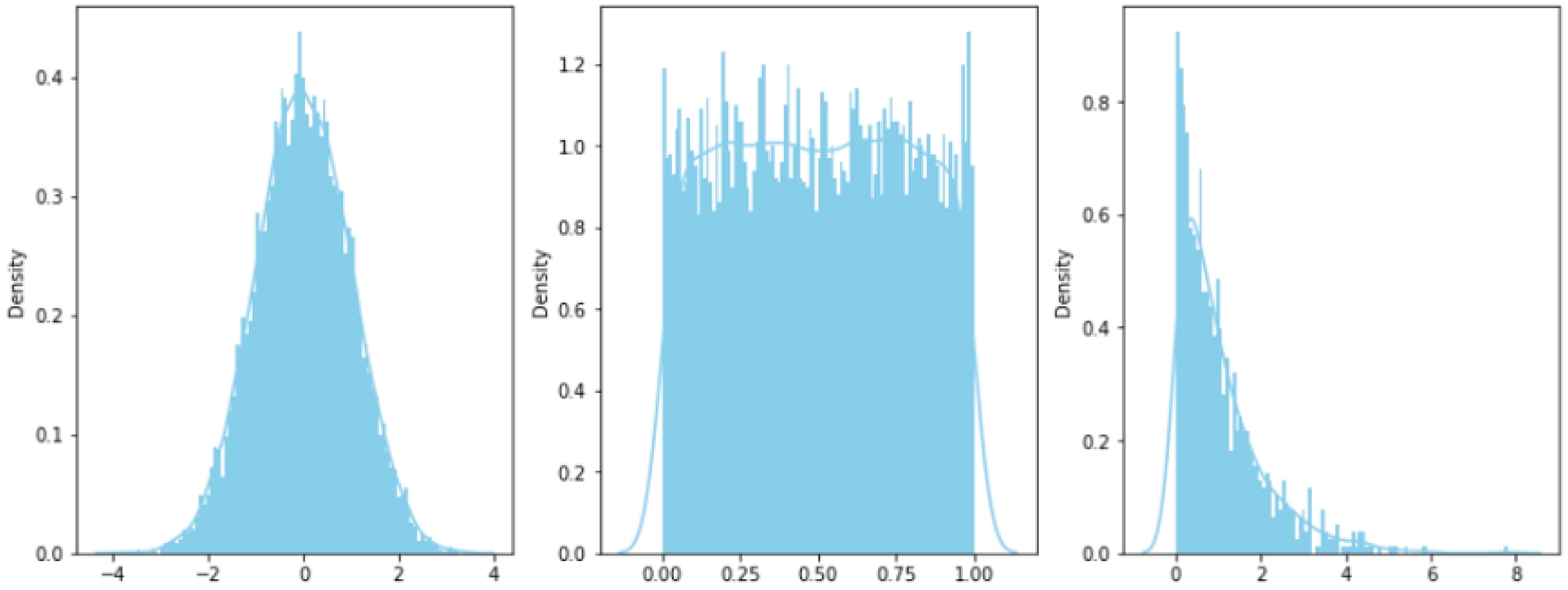
$$P(A, B) = P(A)P(B)$$

And now multiply and divide P(A) by P(B) (nothing changes, it can be cancelled out) and recall the definition of conditional probability:

$$P(A) = \frac{P(A)P(B)}{P(B)} = \frac{P(A, B)}{P(B)} = P(A \mid B)$$

# Probability Distributions

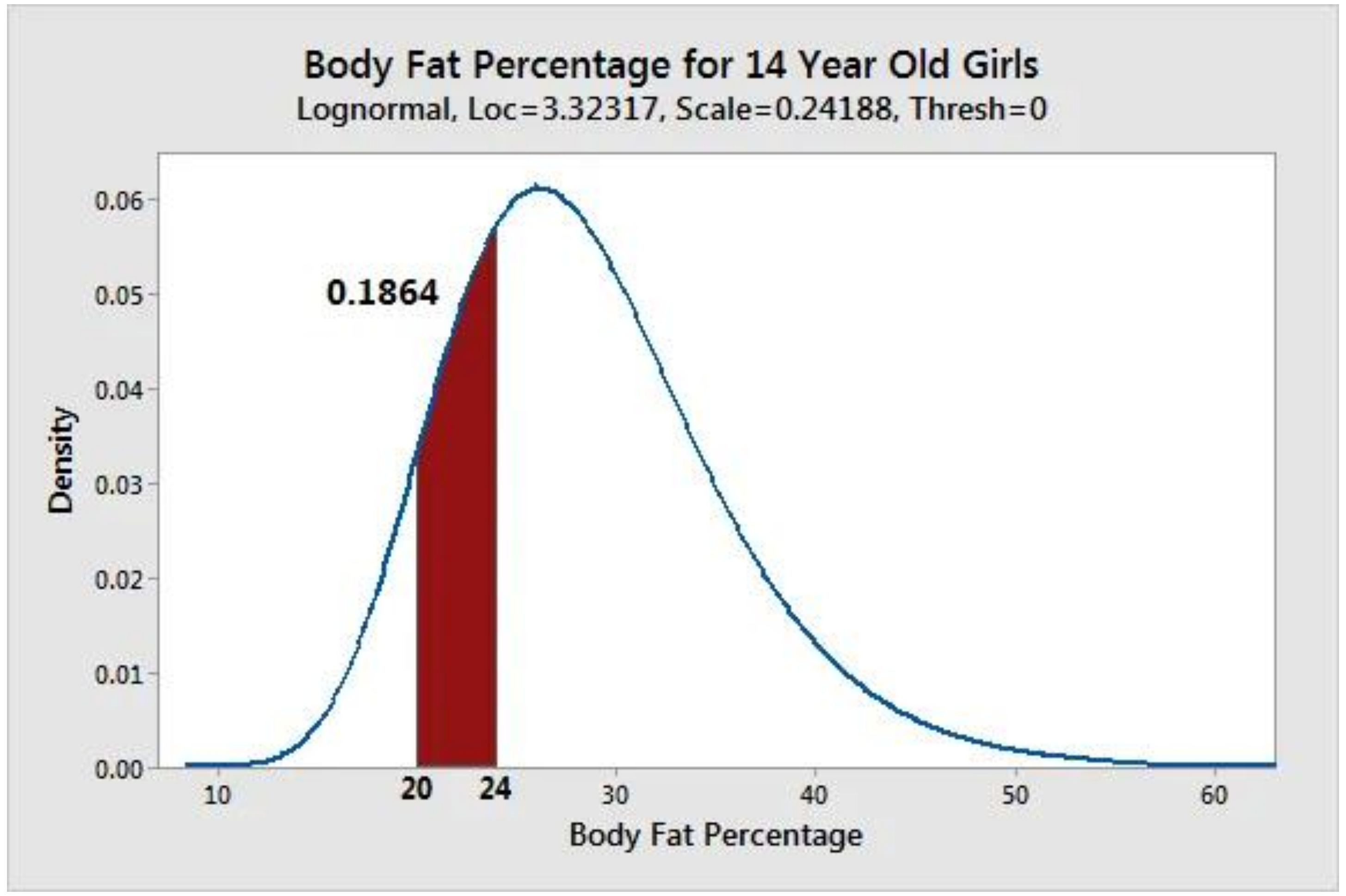
- Probability is the basic building block of Machine Learning and Data Science. In fact, some of the underlying principle of modern machine learning algorithms are partially built on these statistical understanding



- Normal distribution
- Uniform distribution
- Cauchy distribution
- Gamma distribution
- Exponential distribution

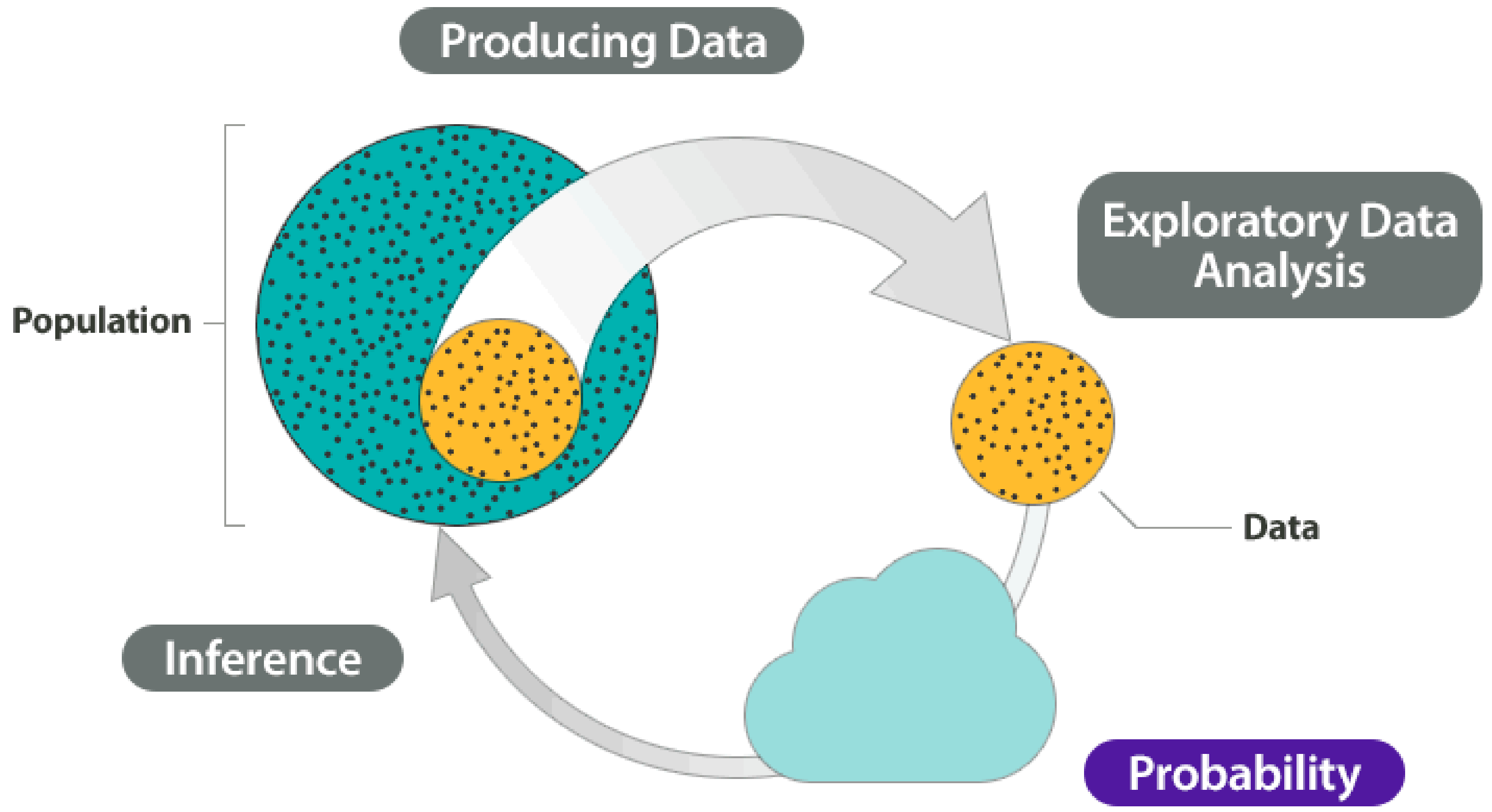
# What is probability distribution

- Probability distribution is a function that gives the probabilities of occurrence of different possible outcomes for an experiment.
- To illustrate, given a 6-sided dice, there are 6 possible outcomes it can be rolled: 1,2,3,4,5,6. If the dice is fair, then the probability across all possible outcomes is identical: 1/6. The probability distribution, therefore, is 1/6 for all possible values of x [1…6].



# Why is probability distribution important?

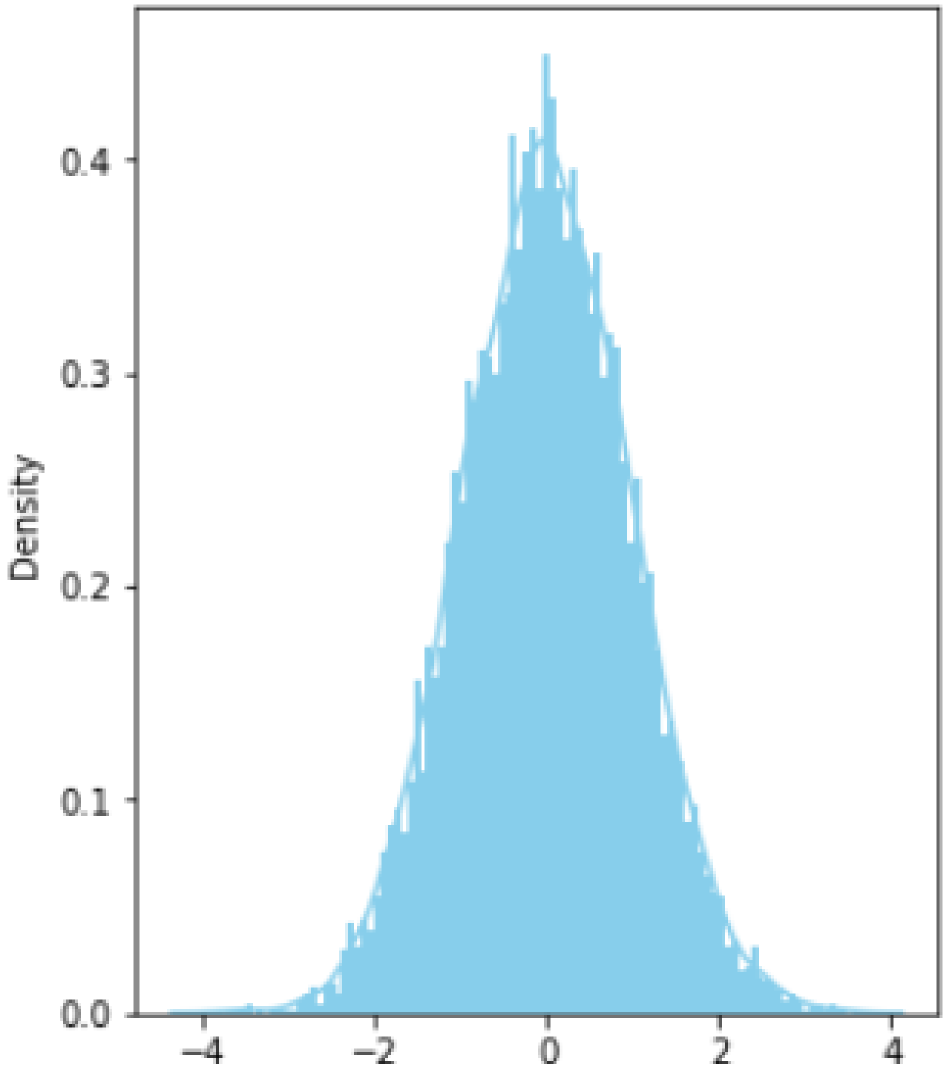
- Helps you to perform sampling given a probability distribution. For instance, your friends ask you to roll the 6-sided dice 100 times in order to evaluate whether the dice is fair or not.
- Allows you to build statistical models by making assumptions of the observations' or parameters' underlying distribution (e.g. assuming that a particular population parameter in a Bayesian model follows an exponential distribution).



# Normal distribution

- Normal distribution, is, well, normal because it describes many of the natural phenomenon out there: blood pressure, measurement error, IQ scores, etc.
- The mathematical formula for normal distribution is as follows, where μ (read: miu) is the mean and σ (read: sigma) is the deviation of the observations.

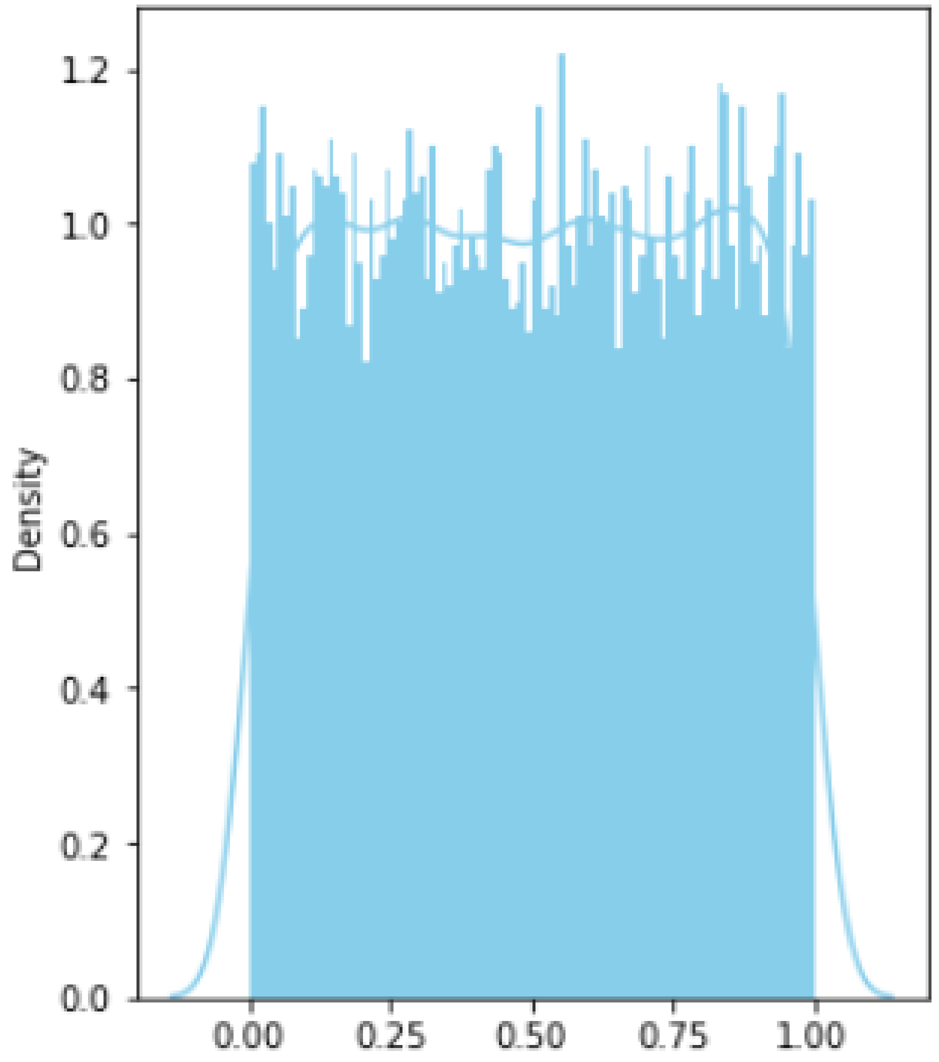
$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



# Uniform distribution

- Uniform distribution describes phenomenon that happen uniformly across possible outcomes. For instance, our earlier 6-sided fair dice or the 52-card playing suite.
- The mathematical formula for uniform distribution will define a limit of a and b (ie. [a,b]). Any value of x below or above b will be assigned a probability of zero, while the rest of the valid observations will be assigned a uniform probability given the number of discrete interval between a and b.

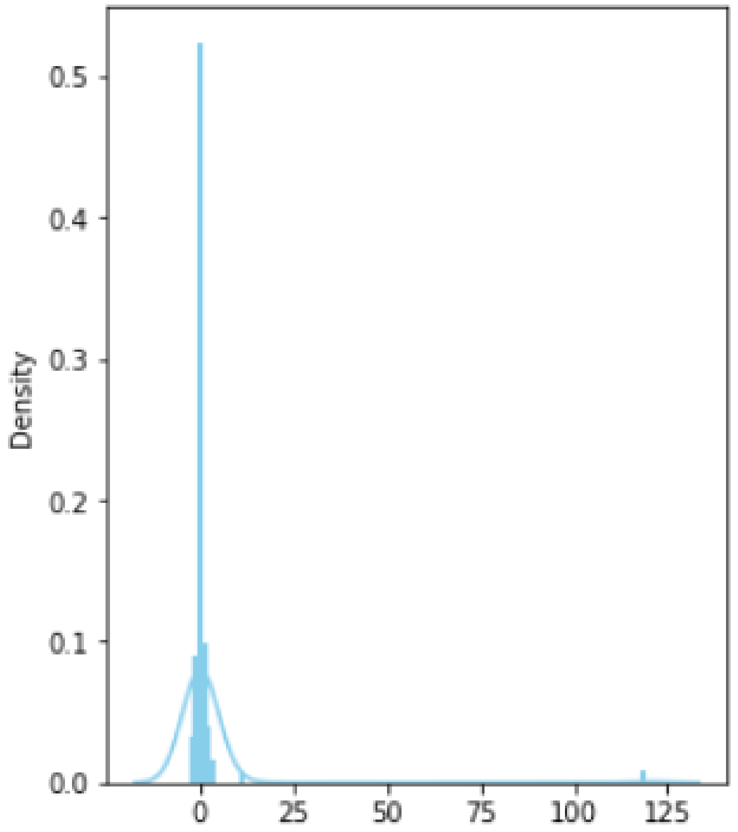
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{for } x < a \text{ or } x > b \end{cases}$$



# Cauchy distribution

- Cauchy distribution is similar in shape to a normal distribution but has differences that are noteworthy. For instance, it has a taller peak than a normal distribution would have. Cauchy’s distribution also has its fat tails to decay much more slowly. This distribution is used a lot in Physics, especially in the field of spectroscopy (ie. study of electromagnetic radiation).
- The mathematical definition is as follows where γ (read: gamma) is the scaling factor to decide how wide or narrow the distribution is going to be.

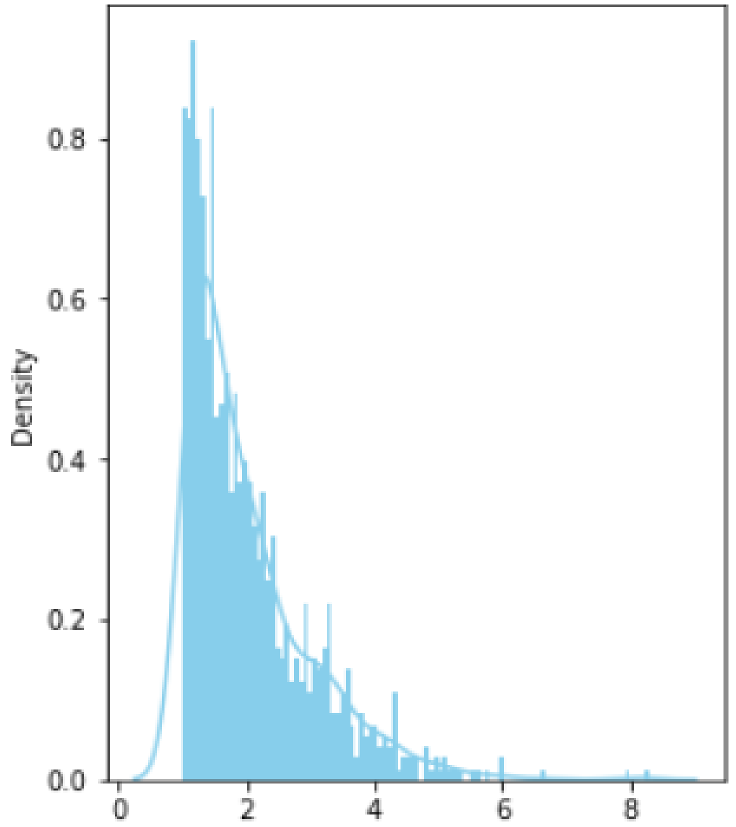
$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x-x_0}{\gamma}\right)^2\right]}$$



# Gamma distribution

- The gamma distribution is a two-parameter probability distributions that are always positive and have skewed distributions. It occurs naturally in the processes where waiting times between events are important (i.e. lag time), such as, the effects of newly developed medicine to treat viral infection, etc.
- The gamma distribution can be parameterized by α and β that determines how skewed the distribution is going to be, like the following:

$$f(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)} \quad \text{for } x > 0 \text{ and } \alpha, \beta > 0,$$

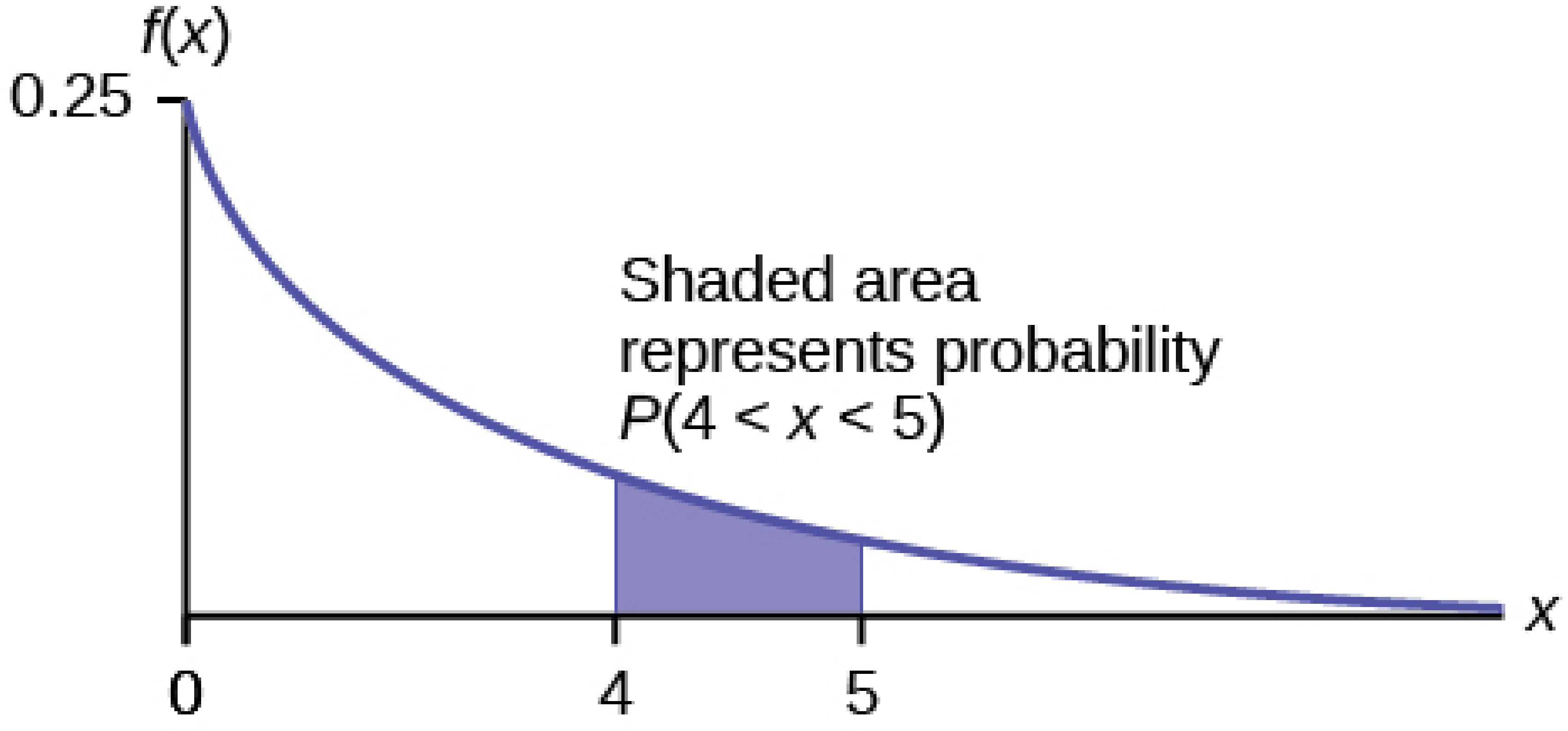


# Exponential distribution

- Exponential distribution is commonly used to model events that are exponentially increasing (or decreasing) such as the trend of population growth (or decline) as time progresses.

- The distribution is parameterized by λ (read: lambda) that regulates how fast the exponential drops or increase as follows:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \ge 0, \\ 0 & x < 0. \end{cases}$$



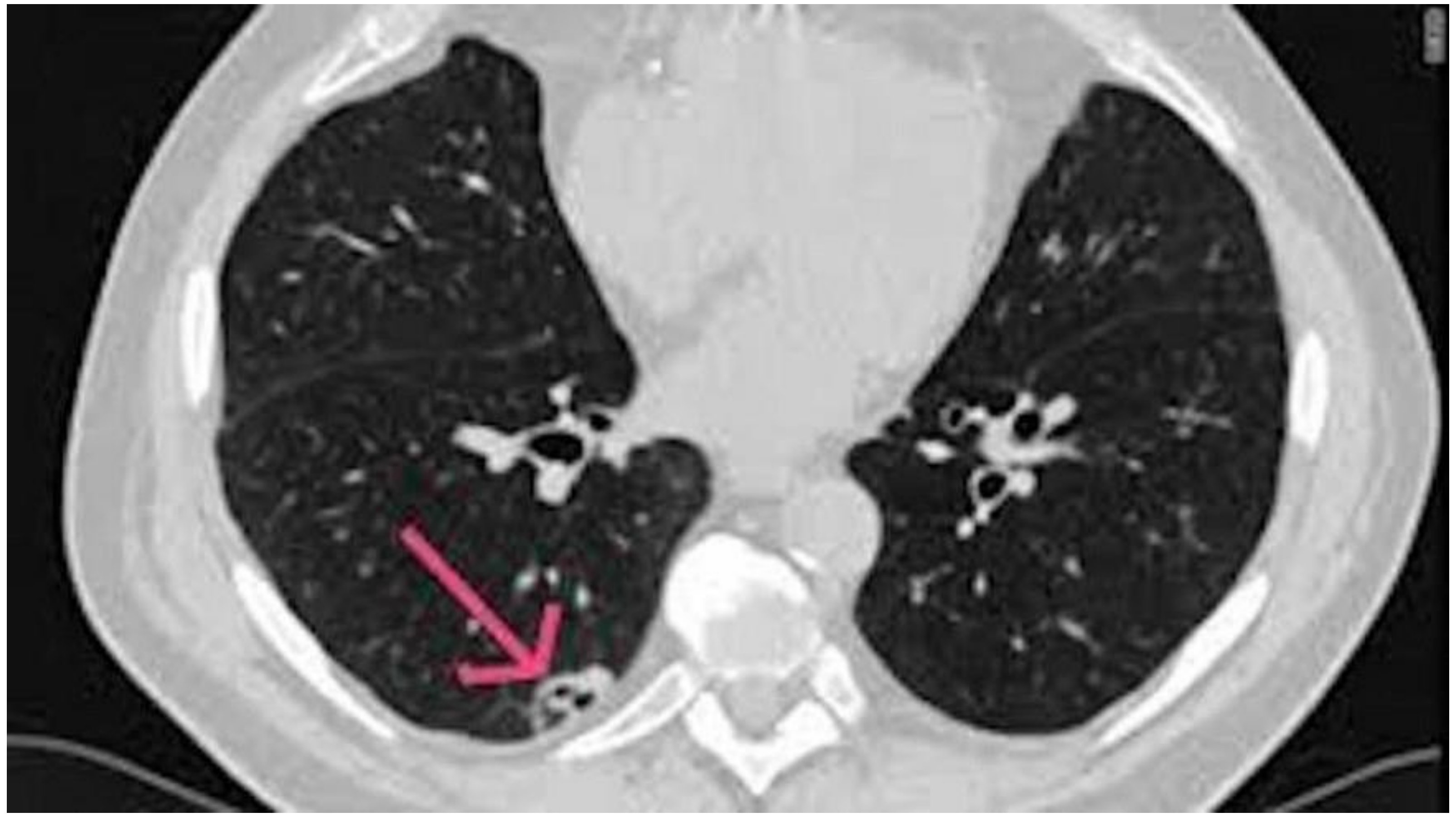
# Decision Theory

- ❑ Whether you are building Machine Learning models or making decisions in everyday life, we always choose the path with the least amount of risk. As humans, we are hardwired to take any action that helps our survival; however, machine learning models are not initially built with that understanding.

- ❑ These algorithms need to be trained and optimized to choose the best option with the least amount of risk. Additionally, it is important to know that some risky decisions can lead to severe consequences if they are not correct

# Decision Theory

- ❑ Consider the problem of cancer detection. Based on a patient's computerized tomography (CT) scan, can a radiologist determine the presence of a tumor?
- ❑ If they believe there is a tumor in the patient, then the physician needs to figure out if the tumor is benign or malignant to determine the proper treatment.
- ❑ Since the purpose of this article is to describe the statistical approach for making these decisions, I will only focus on breaking down the first part of the problem: is there a tumor, yes or no?



# Bayes' Theorem

- One of the most well-known equations in the world of statistics and probability is Bayes' Theorem The basic intuition is that the probability of some class or event occurring, given some feature (i.e. attribute), is calculated based on the likelihood of the feature's value and any prior information about the class or event of interest.
- This seems like a lot to digest, so I will break it down for you.
- First off, the case of cancer detection is a two-class problem. The first class, ω1, represents the event that a tumor is present, and ω2 represents the event that a tumor is not present.

$$P(\omega_j / x) = \frac{p(x / \omega_j) P(\omega_j)}{p(x)} = \frac{\mathit{likelihood} \times \mathit{prior}}{\mathit{evidence}}$$

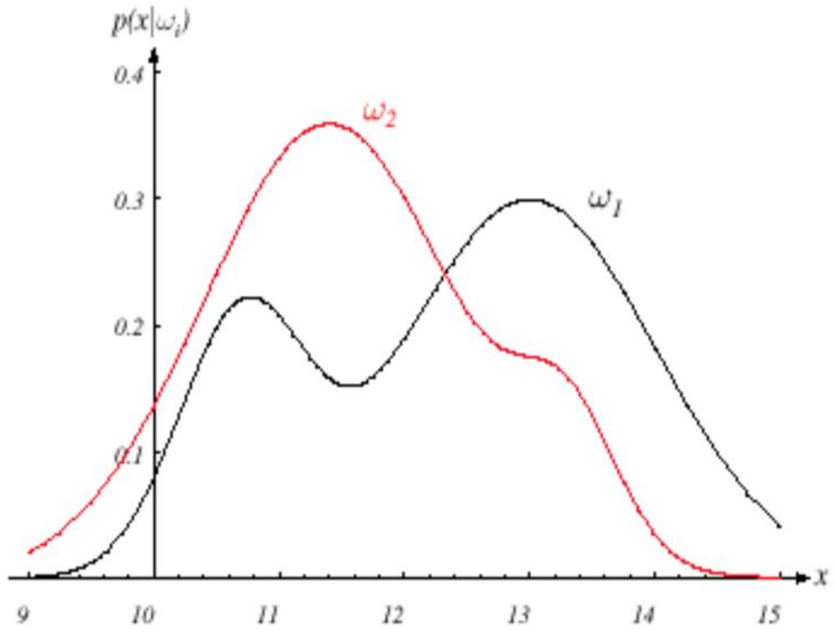
# Bayes’ Theorem

- ❑ Prior
  - ❑There are four parts to Bayes’ Theorem:
    - ❑Prior,
    - ❑Evidence,
    - ❑Likelihood,
    - ❑Posterior.

    - ❑The priors(P(ω1), P(ω2)), define how likely it is for event ω1 or ω2 to occur in nature. It is important to realize the priors vary depending on the situation. Since the objective is to detect cancer, it is safe to say that the probability of a tumor being present is pretty low: P(ω1)<P(ω2). However, no matter the value, all priors must add up to 1.

# Bayes’ Theorem

- Likelihood
  - From a high level, a CT scan is when x-rays are applied in a circular motion. One of the key metrics that is produced is attenuation — a measurement of x-ray absorption. Objects with a higher density have a higher attenuation and vice-versa. Therefore, a tumor is more likely to have a high attenuation compared to lung tissue.
  - Suppose you only look at attenuation values to help make your decision between ω1 and ω2. Each class has a class-conditional probability density, p(x|ω1) and p(x|ω2), called likelihoods. The figure below shows a hypothetical class-conditional probability density for p(x|ω). These distributions are extracted by analyzing your training data; however, it is always good to have domain expertise to check the validity of the data.



# Bayes’ Theorem

- Evidence
  - The best way to describe the evidence, p(x), is through the law of total probability. This law states that if you have mutually exclusive events (e.g. ω1 and ω2) whose probability of occurrence sum up to 1, then the probability of some feature (e.g. attenuation) is the likelihood times the prior summed across all mutually exclusive events.

$$\sum_i p(\mathbf{x}|\omega_i)P(\omega_i)$$

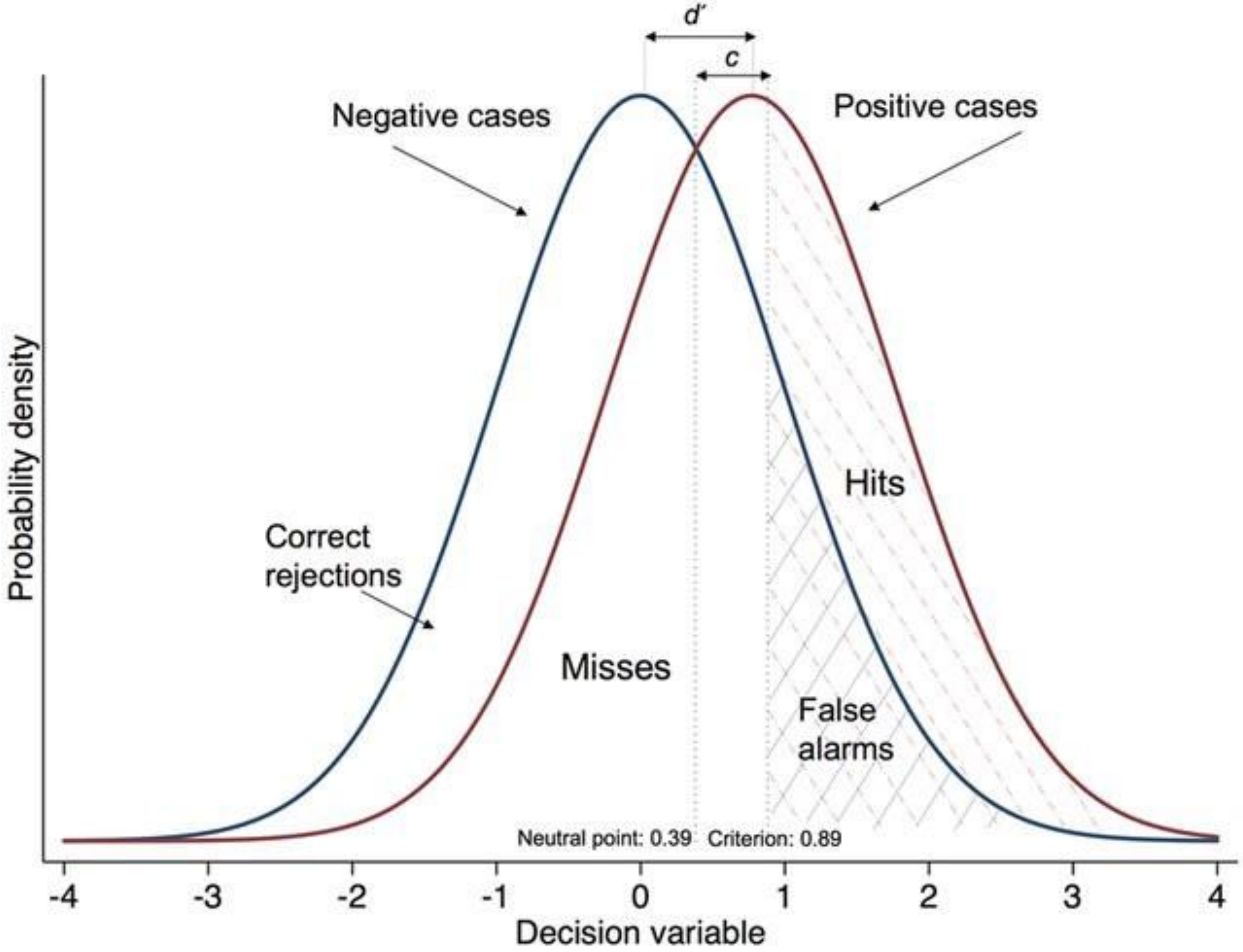
# Bayes' Theorem

- ❑ Posterior
  - ❑The result of using Bayes' Theorem is called the posterior, P(ω1|x) and P(ω2|x).
  - ❑The posterior represents the probability that an observation falls into class ω1 or ω2 (i.e tumor present or not) given the measurement x (e.g. attenuation).
  - ❑Each observation receives a posterior probability for every class, and all the posteriors must add up to 1.
  - ❑ In regards to the cancer detection problem we are trying to solve, there are two posterior probabilities.
  - ❑The image below is a hypothetical scenario of how the posterior values could change with respect to a measurement x. In addition to a connection between the likelihoods and the posteriors, the posterior can be heavily affected by prior P(ω)..

# Decision Rules

- ❑ Now that we have a good understanding of Bayes' theorem, it's time to see how we can use it to make a decision boundary between our two classes.
- ❑ There are two methods for determining whether a patient has a tumor present or not.
- ❑ The first is a basic approach that only uses the prior probability values to make a decision.
- ❑ The second way utilizes the posteriors, which takes advantage of the priors and class-conditional probability distributions.

# Assessment


d′
c
Negative cases
Positive cases
Probability density
Correct rejections
Misses
Hits
False alarms
Neutral point: 0.39
Criterion: 0.89
-4
-3
-2
-1
0
1
2
3
4
Decision variable

# REFERENCES

1. Tom M. Mitchell, “Machine Learning”, McGraw-Hill Education (India) Private Limited, 2013.
2. Trevor Hastie, Robert Tibshirani, Jerome Friedman, “The Elements of Statistical Learning: Data Mining, Inference, and Prediction”, Springer; Second Edition, 2009.

# THANK YOU