Probability & Distribution Stats

Topics:

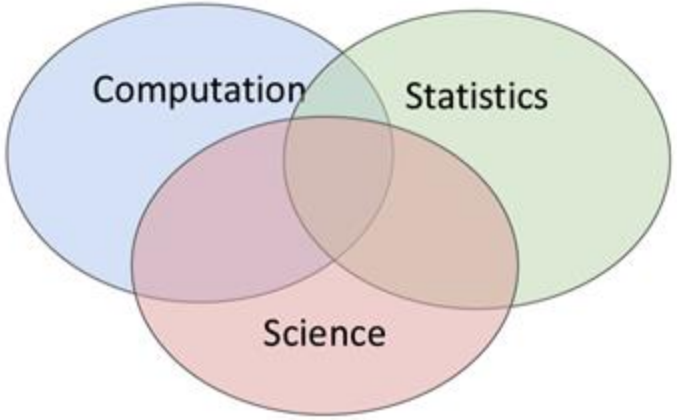
1. Probability Theory:
    a. Basic concepts (events, sample space, probability axioms)
    b. Conditional probability
    c. Bayes' theorem
    d. Probability distributions (discrete and continuous)
    e. Common distributions (e.g., normal, binomial, Poisson)

Statistics for Data Science

Statistics - methods for evaluating hypotheses in the light of empirical facts.

(Stanford Encyclopedia of Philosophy, 2014)

Data Science - a **field** focused on using statistical, scientific, and computational techniques to gai



Approximately equal:

Data Science ≈ Data Mining ≈ Analytics ≈ Quantitative Science

Highly Related

Data Science , Big Data , Machine Learning , Artificial Intelligence

By combining statistical techniques with computational and scientific approaches, **data science expands the scope of traditional statistics to address more complex and diverse data-driven challenges.**

Statistics for Data Science

Statistics - Statistics is the science and art of prediction and explanation.

- Involves statistical methods for **gaining knowledge** and insights from data.

A pathway to knowledge about...

... what was, (past)

... what is, (present)

... what is likely (future, the full population)

Probability Theory

Statistics relies on **probability theory** to understand and interpret data.

Probability Theory and Data Science

Foundation of Data Science:

- **Probability Theory**: Essential for **understanding and quantifying uncertainty** in data analysis.
  a. The mathematical framework for quantifying uncertainty.
  b. Applied to problems of making inferences and decisions under uncertainty.

"Mathematical language for quantifying uncertainty" - Wasserman

In data science, we often deal with incomplete or noisy data, and probability provides a framework to assess the likelihood of different outcomes. By assigning probabilities to events, we can make informed decisions and evaluate the associated risks.

Probability Theory and Data Science

Applications of Probability in Data Science

- **Predictive Modeling**: Uses probability to forecast future events (e.g., predicting customer churn).
- **Machine Learning**: Many algorithms rely on probabilistic models (e.g., Naive Bayes, Bayesian networks).
- **Statistical Inference**: Helps make data-driven decisions by estimating parameters and testing hypotheses (e.g., A/B testing).
- **Practical Examples**:
  - **Dice Rolls**: Modeling the probability of different outcomes (e.g., rolling a die).
  - **Coin Flips**: Understanding the chances of heads or tails (e.g., flipping a coin).

What is Probability?

People talk loosely about probability all the time:

- “What are the chances our Cricket team will win this weekend?”
- “What’s the chance of rain tomorrow?”

What Is Probability?

Probability quantifies uncertainty.

- Measures the likelihood of events.
- Values range from 0 (impossible) to 1 (certain).

Classic Example: What is the chances of rolling a one (Number 1) on the dice?

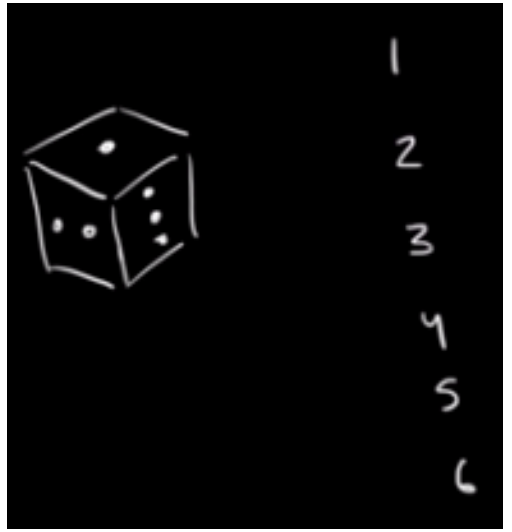
Basic Probability Formula

To find the probability of an event happening we use the formula:

$$P(A) = \frac{\text{Number of times A occurs}}{\text{Total number of possible outcomes}}$$

Example: The probability of getting number 1 is: $P(1) = 1/6$

can also be written as a percentage
$P(1) = 1/6 \sim 16.67\%$



There are six different outcomes. (**sample space**)

Try Yourself: What is probability of **getting an even number** when a die is rolled?

Some More Examples

- Stock Market → The chance of stock market's rising above some point, or falling below some point is ‘x’%
- Weather Forecast → The chance for rain is 45% tonight.
  - The likelihood of rainfall in terms of probability is 0.45 that the event, rainfall, might occur

So Essentially probability is a measure between equate to 1.

Key Concepts in Probability

Help measure and express uncertainty:

Events: Outcomes or results of experiments or observations (e.g., whether a customer will buy a product).

Random Variables: Variables that can take different values due to chance (e.g., the number of purchases a customer makes).

Probability Distributions: Functions that describe the likelihood of different outcomes (e.g., using a normal distribution to model customer spending patterns). Helps in understanding the spread of data.

$X =$ Number of heads in 10 coin tosses

$X = 10$

$X = 9$

$X = 9$

Example: Random Variable and Probability Distributions

A random variable represents the outcomes of a random event.

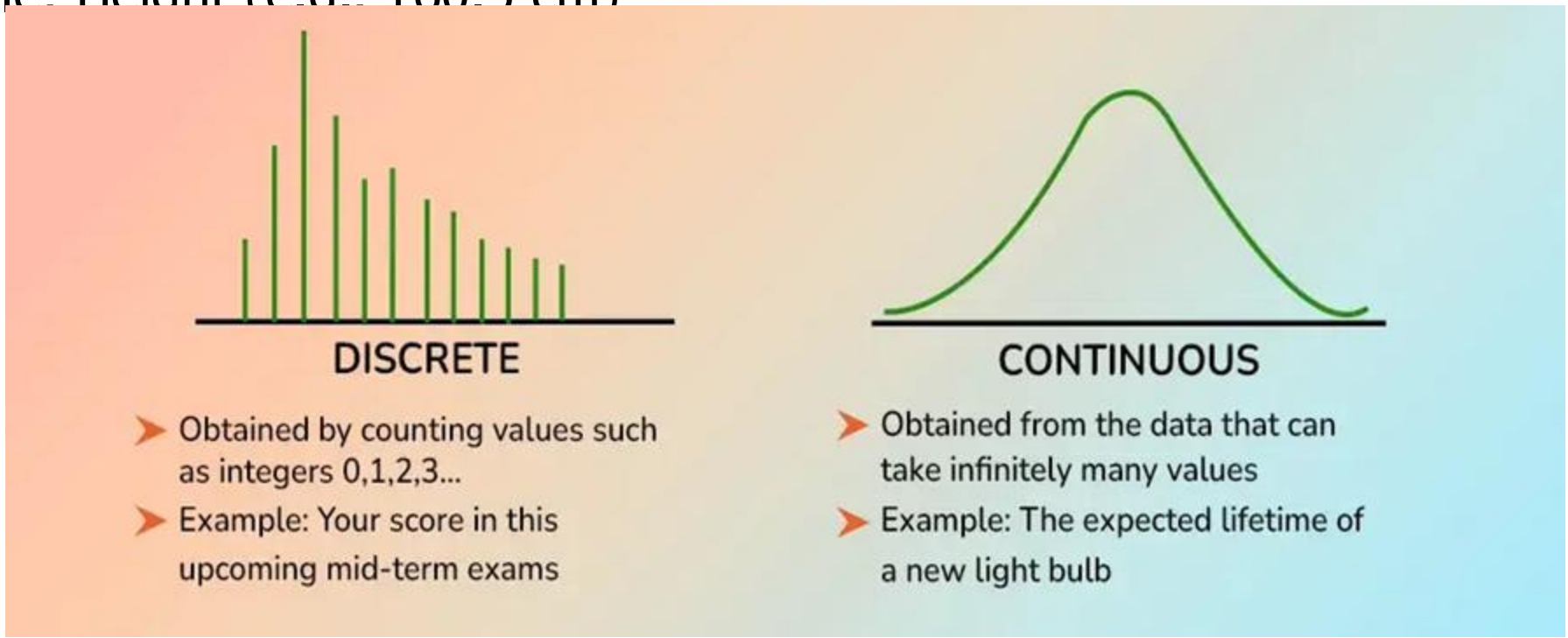
- **E.g: Imagine rolling two dice.**
- The outcome (the combination of numbers we get) is uncertain.
- So, we use a random variable, X to represent the sum of the numbers on the two dice.
- X could take on values from 2 to 12, depending on the sum of the two dice.

P(X=2)= 1/36
P(X=3)= 2/36 (rolling a 1 & a 2 or rolling a 2 & a 1)
P(X=4)= 3/36
P(X=5)= 4/36
P(X=6)= 5/36
P(X=7)= 6/36
P(X=8)= 5/36
P(X=9)= 4/36
P(X=10)= 3/36
P(X=11)= 2/36
P(X=12)= 1/36

Probability Distributions: Describe the likelihood of different outcomes occurring.

Random Variable Types:

- **Discrete Random Variable:** Takes countable values. Example: Rolling a die (1-6).
- **Continuous Random Variable:** Takes infinite possible values within a range. Example: Height (e.g., 160.5 cm)



More Probability Formulas

Conditional Probability

Sometimes we want to figure out the chance of something happening when we already know else has happened/occurred?

"Conditional Probability"

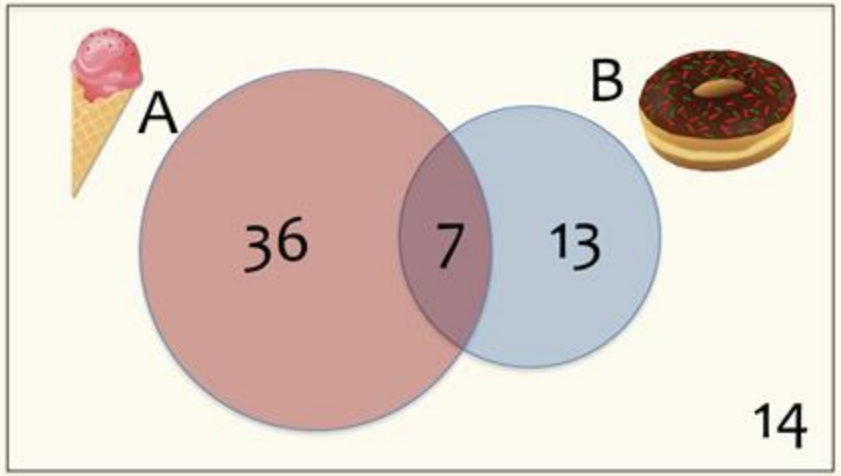
(Understanding Likelihood Given Information)

The probability of an event A happening given that another event B has already occurred/happened.

Conditional Probability

P(A | B) → read as the probability of "A given B."

Examples: P(A | B) = P(Passing the class | Not sleeping the night before the final)



What's the probability that someone likes ice cream **given** they like donuts?

Conditional Probability Formula

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

Probability of A given B

Probability of A and B

Probability of B

Conditional Probabilities are Fundamental to Data Science

Machine Learning: Most modern ML learning techniques try to estimate

P (outcome | data)

Example(s):

- Probability a customer buys a product after viewing it online: P(customer buys | viewed product).
- Algorithms like Naive Bayes use probabilities based on **known features to predict outcomes**, calculating conditional probabilities.

Causal inference: identify and understand cause-and-effect relationships in data.

Does a treatment (i.e new drug, policy) cause a specific outcome (i.e recovery, sales increase)?

- Compare P(outcome | treatment) =/= P(outcome| no treatment) *
- Causal Effect: P(outcome | treatment) – P(outcome | no treatment).

*also requires random sampling of treatment conditions

Formula: Conditional Probability

Conditional Probability Formula

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

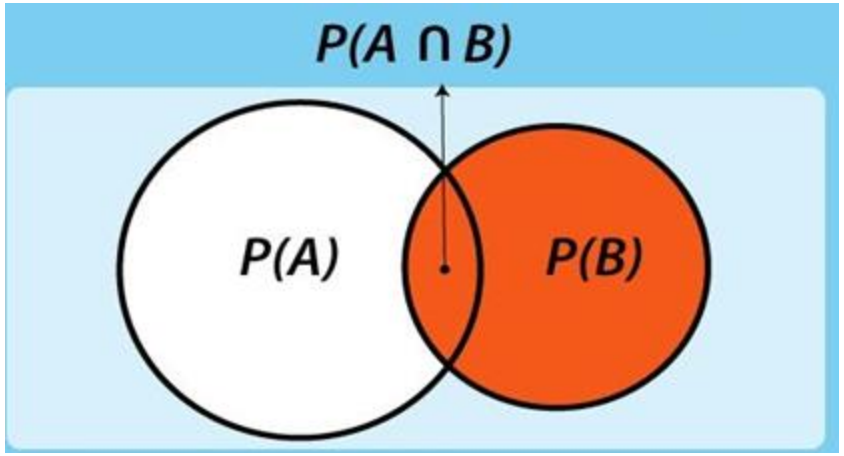
Probability of A given B

Probability of A and B

Probability of B

Assuming P(B) != 0

We take the chance of both things happening together, then divide it by the chance of the thing we know.



More Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a **shiny** marble from the bag if I know it's red?"

A represents the event of picking a shiny marble

B represents the event the event of picking a red marble.

Example:

Imagine you have a bag of colored marbles - 5 red and 5 blue. If you know that 3 out of the 5 red marbles are also **shiny**, you might wonder:

"What's the chance of picking a **shiny** marble from the bag if I know it's red?"

Conditional Probability Formula

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

Probability of A given B

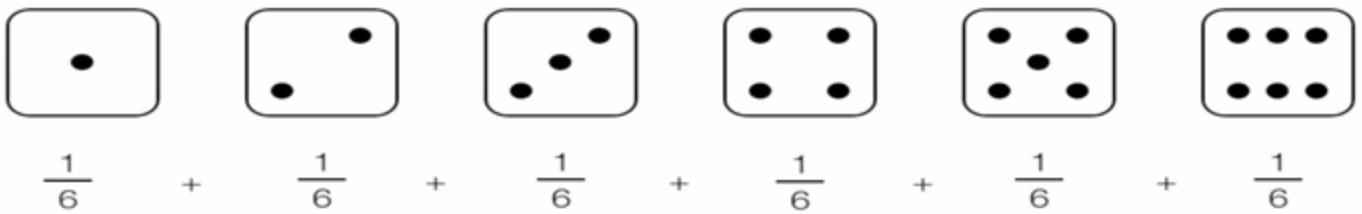
Probability of A and B

Probability of B

- P (A ∩ B) = What is the probability of picking a shiny red marble ? = 3/10
- P (B) = What is the probability of picking a red marble ? = 5/10

P (A|B)? = (3/10) / (5/10) = ⅗ = 0.6 = 60%

Exercise: Conditional Probability



$$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

$$P(B \mid A) = \text{What is the Probability of } \left(\begin{array}{c} \text{rolling a dice and it's} \\ \text{value is less than 4} \end{array} \;\middle|\; \begin{array}{c} \text{knowing that the value is} \\ \text{an odd number} \end{array} \right)$$

Exercise: Conditional Probability

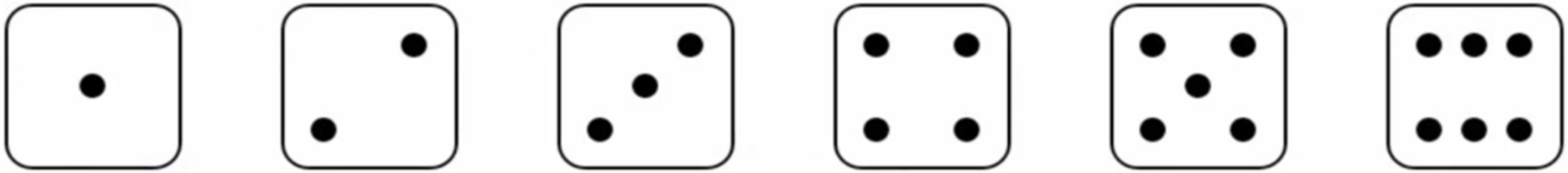
What is the Probability of

rolling a dice and it's value is 1

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

knowing that the value is an odd number

rolling a dice and it's value is 1



Next: Independence

Independence

Conditional independence: the occurrence of one event doesn't affect the probability of another event happening

- Ex: Outcomes of two tosses of a coin are independent.
  - Events A and B are independent if:

$$P(A \cap B) = P(A) \cdot P(B).$$

    - Equivalent to $P(A \mid B) = P(A)$.

- What if P(Passing the class | I drank Coke before final exam)= P(Passing the class)?
- What is P(Drawing a red card | Not sleeping well) =P(Drawing a red card)?
- Think of P(My exercise working out | I'm a hard working person)= P(My exercise working out)?

Independence

Conditional independence:

Two events A and B are independent if and only if P(A∩B)=P(A)P(B).

- Ex: Outcomes of two tosses of a coin are independent.
  - Events A and B are independent if:

$$P(A \cap B) = P(A) \cdot P(B).$$

    - Equivalent to $P(A \mid B) = P(A)$.

$$P(A \cap B) = P(A)\ P(B)$$

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

Differences between disjointness and independence

Concept	Meaning	Formulas
Disjoint	A and B cannot occur at the same time	$A \cap B = \emptyset$,
		$P(A \cup B) = P(A) + P(B)$
Independent	A does not give any information about B	$P(A \mid B) = P(A)$, $P(B \mid A) = P(B)$
		$P(A \cap B) = P(A)\,P(B)$

Conditional Independence in Data Science

Conditional independence helps focus on important data, making models simpler and more effective.

- **Simplifying models:** Reduces unnecessary relationships, improving speed and accuracy.
  - *Example*: In medical diagnosis, cough and fever are independent given a cold, so the model focuses on how each symptom relates to the cold.

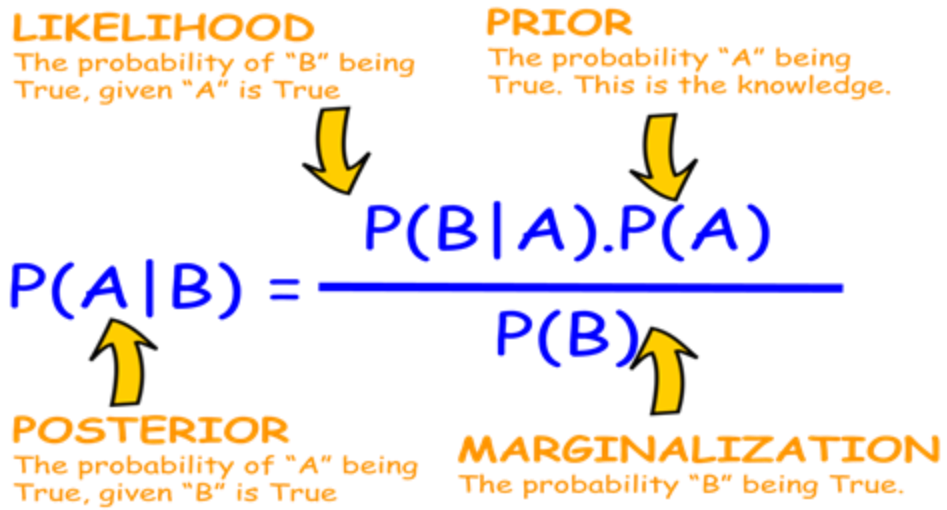
- **Naive Bayes** assumes features are conditionally independent given the target, reducing complexity.
  - *Example*: In spam email detection, words like "buy" and "discount" are independent when predicting spam.

Next: Bayes Theorem

One of the most important formulas in statistics (for our purposes) as well as the most important rule in data science!

Bayes Rule

Bayes' Rule calculates the probability of a hypothesis (event) based on **prior belief** (prior probability) and new evidence



Example:

In spam detection:

- **A** = Email is spam
- **B** = Email contains the word "discount"

Bayes' Rule updates our belief on whether the email is spam based on the word "discount."

- **P(A | B)**: Probability of event A given event B (**posterior probability**).
- **P(B | A)**: Probability of event B given event A (**likelihood**).
- **P(A)**: Prior probability of event A (**prior belief**).
- **P(B)**: Probability of event B (**evidence** or **normalizing constant**).

$P(A)$ is called the **prior** (our belief without knowing anything)
$P(A|B)$ is called the **posterior** (our belief after learning B)

Bayes' Rule is particularly valuable in fields like statistics, machine learning, and Bayesian inference, where we continuously update our beliefs as new data or information becomes available.

Bayes Theorem Proof

It provides a way to calculate conditional probabilities when we have **prior beliefs** and **new evidence**.

By definition Cond. Prob:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \qquad \text{(Equation 1)}$$

Swapping A, B gives:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} \qquad \text{Here: } P(A \cap B) = P(B \cap A),$$

➡ $P(A \cap B) = P(B \mid A) * P(A)$

Now from Equation 1, substitut $P(A \cap B)$:

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

Bayes Theorem Formula

Example: Picnic Day

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

What is the chance of rain during the day?

Scenario: You want to find the chance of rain during the day given that the morning is cloudy. You **have** the following probabilities:

1. Probability of Rain (**P(Rain)**) = 10% (because only 3 out of 30 days are rainy).

1. Probability of Cloud, given that Rain happens (**P(Cloud|Rain)**) = 50% (because 50% of rainy days start off cloudy).
2. Probability of Cloudy morning (**P(Cloud)**) = 40% (because 40% of all days start off cloudy).

Now, use Bayes' Theorem:

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

$$P(Rain|Cloud) = \frac{0.1 \times 0.5}{0.4} = .125$$

Or a 12.5% chance of rain. Not too bad, let's have a picnic!

Example: Fire and Smoke

- dangerous fires are rare (1%)
- but smoke is fairly common (10%) due to barbecues,
- and 90% of dangerous fires make smoke

We can then discover the **probability of dangerous Fire when there is Smoke**:

$$P(\text{Fire}|\text{Smoke}) = \frac{P(\text{Fire})\,P(\text{Smoke}|\text{Fire})}{P(\text{Smoke})}$$

$$= \frac{1\% \times 90\%}{10\%}$$

$$= 9\%$$

So it is still worth checking out any smoke to be sure.

Try: Conditional Prob. Practice

- In a factory there are 100 units of a certain product, 5 of which are defective. We pick three units from the 100 units at random. What is the probability that none of them are defective?

Let us define A_i as the event that the ith chosen unit is not defective, for $i = 1, 2, 3$. We are interested in $P(A_1 \cap A_2 \cap A_3)$. Note that

$$P(A_1) = \frac{95}{100}.$$

Given that the first chosen item was good, the second item will be chosen from 94 good units and 5 defective units, thus

$$P(A_2|A_1) = \frac{94}{99}.$$

Given that the first and second chosen items were okay, the third item will be chosen from 93 good units and 5 defective units, thus

$$P(A_3|A_2, A_1) = \frac{93}{98}.$$

Thus, we have

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_2, A_1) \\ &= \tfrac{95}{100}\tfrac{94}{99}\tfrac{93}{98} \\ &= 0.8560 \end{aligned}$$

Next: Law of Total Probability

Example: Equal Probability of Selecting Each Bag

Example 01: Suppose I have **two bags of marbles**.

- The first bag contains **6** white marbles and **4** black marbles.
- The second bag contains **3** white marbles and **7** black marbles.

Now suppose I put the two bags in a box. If I close my eyes, grab a bag from the box, and then grab a marble from the bag, **what is the probability that it is black?**

Total Probability of Drawing a Black Marble:

$$\frac{11\ Black}{20\ Total} = \frac{11}{20} = 0.55$$

$$\frac{1}{2} * \frac{4}{10} + \frac{1}{2} * \frac{7}{10} = \frac{4}{20} + \frac{7}{20} = \frac{11}{20}$$

How About Unequal Probability of Selecting Each Bag?

Example: Same scenario as before, but now suppose that the first bag is much larger than the second bag.

So that when I reach into the box I'm **twice as likely to grab the first bag as the second**. What is the probability of grabbing a black marble?

Probability of selecting Bag 1 $P(B1) = \frac{2}{3}$

Probability of selecting Bag 2 $P(B2) = \frac{1}{3}$

How About Unequal Probability of Selecting Each Bag?

Example: Same scenario as before, but now suppose that the first bag is much larger than the second bag.

So that when I reach into the box I'm **twice as likely to grab the first bag as the second**. What is the probability of grabbing a black marble?

Probability of selecting Bag 1 $P(B1) = \frac{2}{3}$

Probability of selecting Bag 2 $P(B2) = \frac{1}{3}$

$$\frac{2}{3} * \frac{4}{10} + \frac{1}{3} * \frac{7}{10} = \frac{8}{30} + \frac{7}{30} = \frac{15}{30} = \frac{1}{2}$$

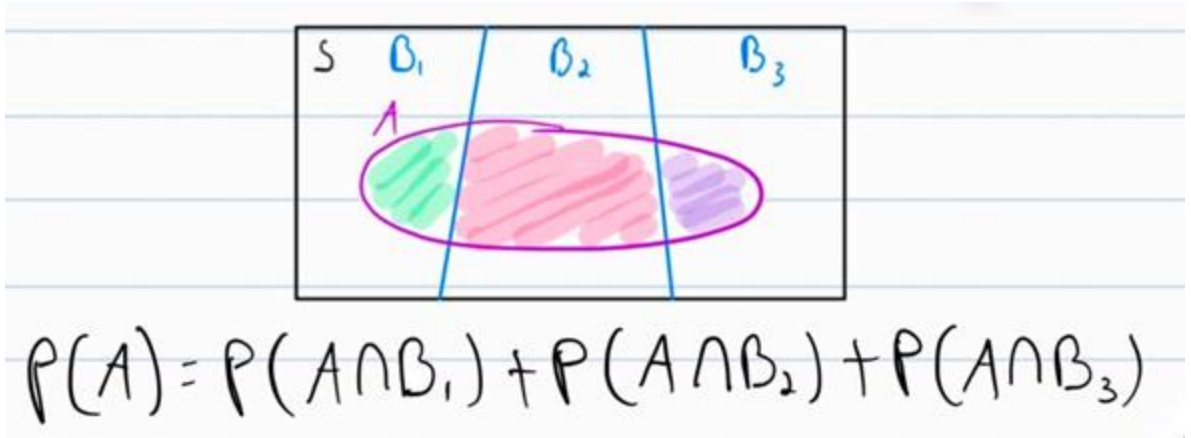
We just applied “Law of Total Probability”

If events B1,B2,...,Bn form a **partition** of the sample space (i.e., they are **mutually exclusive** and **collectively exhaustive**), then

$$P(A) = \sum_{i=1}^{n} P(B_i) \cdot P(A|B_i)$$

Where:

- $P(A \mid B_i)$ = Probability of A given B_i
- $P(B_i)$ = Probability of event B_i



Example:

Let a person come from group A, B, or C with known group probabilities. The total chance they like coffee:

$$P(\text{likes coffee}) = P(C|A)P(A) + P(C|B)P(B) + P(C|C)P(C)$$

Law of Total Probability: back to the example

Sample Space S: The scenario of choosing a bag and drawing a marble.

1. **Mutually Exclusive Events B1 and B2**:
    - B1: Selecting Bag 1.
    - B2: Selecting Bag 2.

    These events are mutually exclusive because you can only choose one bag at a time.

1. **Event A**: Drawing a black marble.

According to the law of total probability, the probability of drawing a black marble (event A) is:

$$P(A)=P(B_1)\cdot P(A|B_1) + P(B_2)\cdot P(A|B_2)P(A)$$

Try: Bayes Rule and Conditional Prob. Practice

Let’s work on a simple NLP problem with Bayes Theorem. By using NLP, I can detect spam e-mails in my inbox. Assume that the word ‘offer’ occurs in 80% of the spam messages in my account. Also, let’s assume ‘offer’ occurs in 10% of my desired e-mails. If 30% of the received e-mails are considered as a scam, and I will receive a new message which contains ‘offer’, what is the probability that it is spam?

$$P(S|O) = \frac{P(O|S) \cdot P(S)}{P(O)}$$

$$P(O) = P(O|S) \cdot P(S) + P(O|D) \cdot P(D)$$

$$P(O) = (0.80 \cdot 0.30) + (0.10 \cdot 0.70)$$

$$P(O) = 0.24 + 0.07$$

$$P(O) = 0.31$$

$$P(S|O) = \frac{0.80 \cdot 0.30}{0.31}$$

$$P(S|O) = \frac{0.24}{0.31}$$

$$P(S|O) \approx 0.774$$

Try Yourself

Scenario: Imagine you are organizing a charity event, and there are three possible venues (A, B, and C) where you can hold the event. The probability of each venue being available on a given day is as follows:

Venue A: 40% chance of being available.

Venue B: 30% chance of being available.

Venue C: 30% chance of being available.

You also know that if **Venue A is available, there's a 70% chance of raising a large amount of money for the charity**, while if **Venue B or C is available, there's a 50% chance of raising a large amount of money**.

What is the overall probability L of raising a large amount of money at the charity event?

Try Yourself

Venue	Availability	Probability of large donation
A	0.40	0.70
B	0.30	0.50
C	0.30	0.50

$$L = P(A) \times P(L|A) + P(B) \times P(L|B) + P(C) \times P(L|C)$$

$$L = (0.40)(0.70) + (0.30)(0.50) + (0.30)(0.50)$$

$$(0.40)(0.70) = 0.28$$
$$(0.30)(0.50) = 0.15$$
$$(0.30)(0.50) = 0.15$$

$$0.28 + 0.15 + 0.15 = 0.58$$

$$L = 0.58$$

Expected Value

Expected Value

The **Expected Value** of a random variable is its **long-run average outcome** over many trials.

- It represents the **center** of a probability distribution.
- The average outcome if an experiment is repeated many times.
- A measure of the center of a probability distribution

$$E[X] = \sum x_i p(x_i)$$

x_i = The values that X takes

$p(x_i)$ = The probability that X takes the value x_i

- X_i: Possible outcomes.
- $P(xi)$" Probability of each outcome.

Expected Value: Example

Someone offers for you to go on a game show. On this gameshow, there is:

- A **5%** chance you will be given a million dollars
- A **95%** chance they will hit you with sticks, causing **$10,000** worth of medical bills (the hits are not particularly bad, your insurance just doesn't cover check-ups)

Your feelings about being hit with sticks aside, should you go on the game show?

Expected Value: Example

- There's a 5% chance of winning $1,000,000, which is 0.05× $1,000,000=$50,000
- There's a 95% chance of incurring $10,000 in medical bills, which is 0.95×−$10,000=−$9,500.
- Minus sign: negative financial outcome

Expected value of the game show:

EV=(Probability of Winning×Prize for Winning)+(Probability of Medical Bills×Cost of Medical Bills)

=(1,000,000 * .05) + (-10,000 * .95) = 40500

Net positive!

Conclusion: on average, you can expect to gain $40,500 by participating in the game show.

Still totally worth it!

Example #2

You are playing a game where you spin a wheel. The wheel has the following sets of rewards:

20% chance you win nothing

30% chance you win a ten dollar gift card

40% chance you win a twenty dollar gift card

10% chance you win a thirty dollar gift card

Example #2

You are playing a game where you spin a wheel. The wheel has the following sets of rewards:

20% chance you win nothing

30% chance you win a ten dollar gift card

40% chance you win a twenty dollar gift card

10% chance you win a thirty dollar gift card

$EV=(.2 * 0) + (.3 * 10) + (.4 * 20) + (.1 * 30) = 0 + 3 + 8 + 3 = 14$

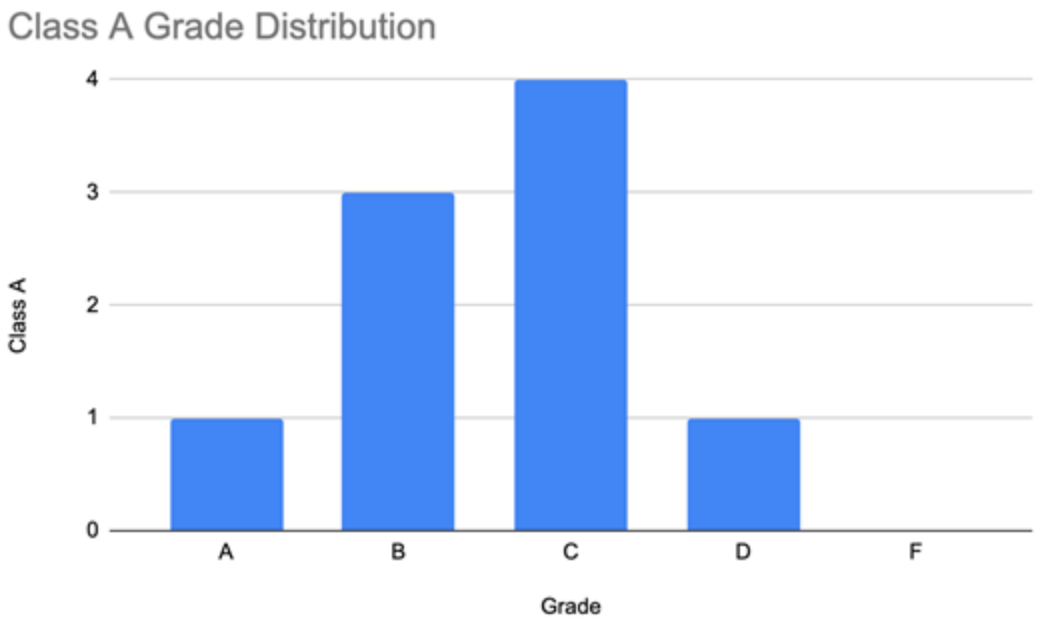
On average, you can expect to win $14 per spin over a large number of spins.

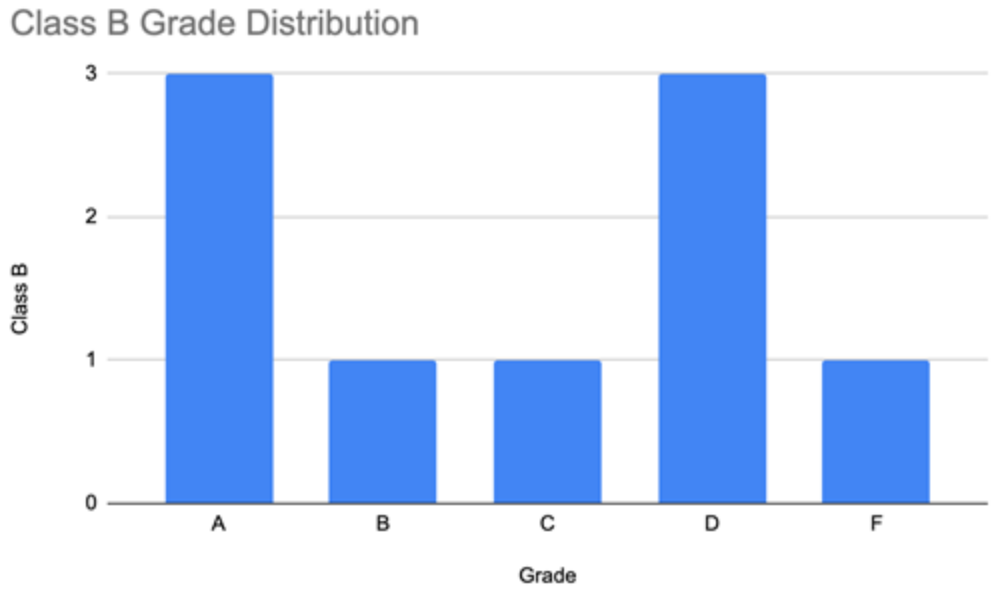
Distributions

What is a distribution?

Class A	**Class B**
C	A
B	A
C	A
A	B
C	C
B	D
D	D
C	D
B	F

Grade Distributions





These distributions show how the grades are distributed among the students in each class, giving an idea of the proportion of students who received each grade.

- It helps understand the performance of the students in terms of their grades.

Distribution Types

Discrete Distributions: describe random variables that can only take on a countable number of distinct values. These values are typically integers.

Examples: Bernoulli distribution, Binomial distribution, Poisson distribution.

Continuous Distributions: describe the probabilities associated with continuous random variables. Unlike discrete distributions, which assign probabilities to specific outcomes, continuous distributions assign probabilities to intervals or ranges of outcomes. This is because continuous random variables can take on any value within a certain range, typically real numbers.

Examples: Normal (Gaussian) distribution, Exponential distribution etc.

Next: We will discuss different distribution types

1. Uniform Distribution
2. Normal (Gaussian) Distribution
3. Not Normal:
    a. Poisson Distribution
    b. Zero-Inflated Poisson Distribution
    c. Binomial Distribution

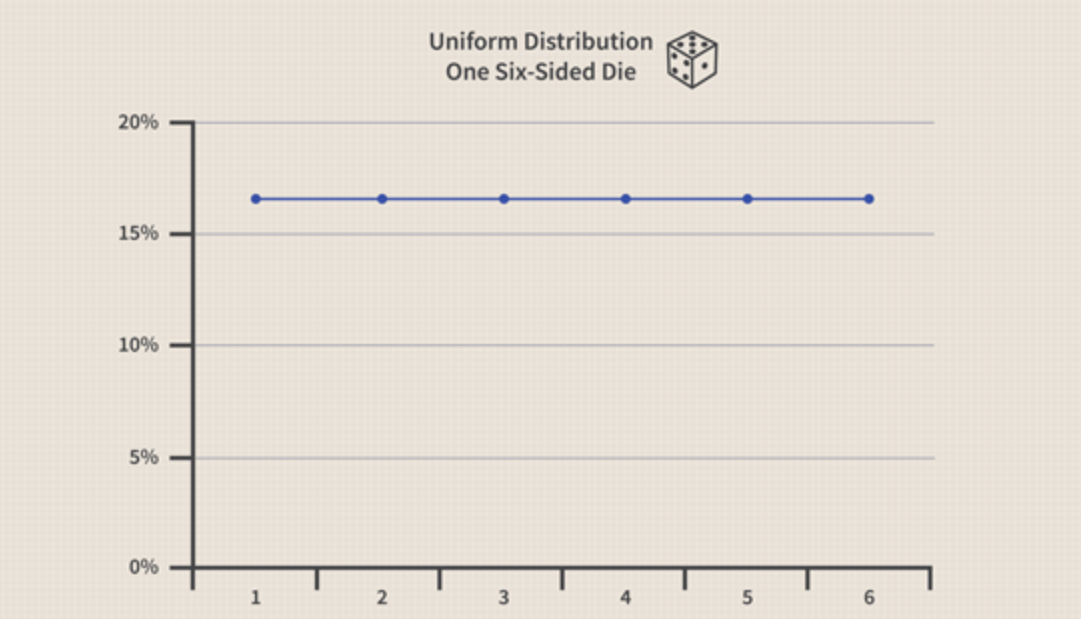
Many more …..

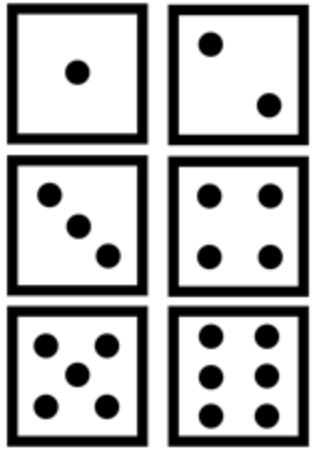
The Uniform Distribution

The Uniform Distribution

Def: describes a set of outcomes where each outcome has an equal probability of occurring within a specified interval.

- All outcomes are equally likely.
- It's like rolling a fair six-sided die, where each number has the same chance of appearing.



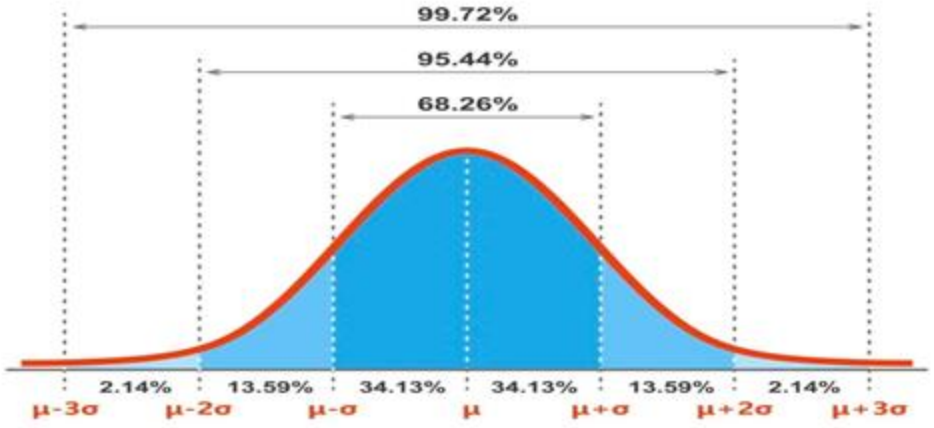


The Normal Distribution (Gaussian)

2. Normal (Gaussian) Distribution

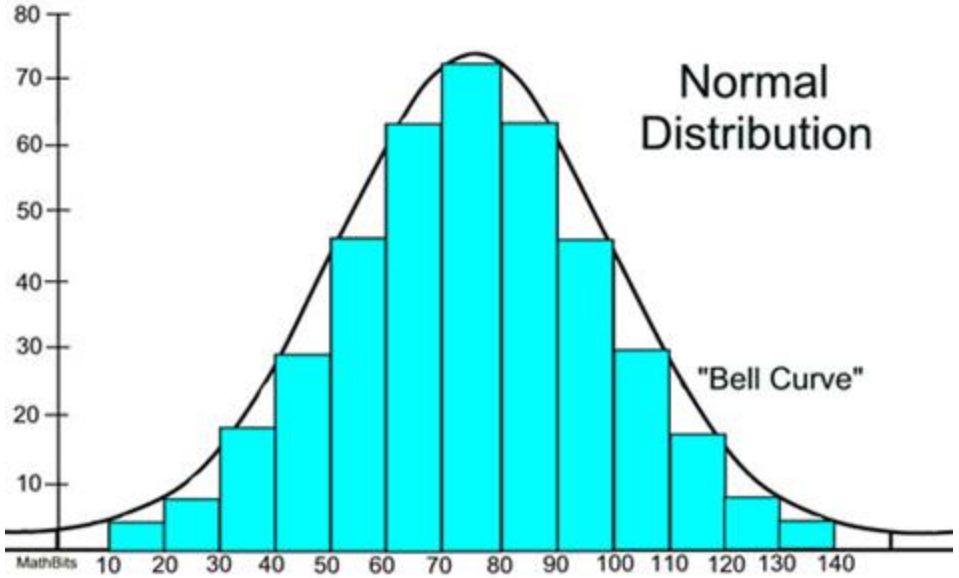
People tend to assume data is **normally distributed** unless there is a reason to think otherwise.

- Bell-shaped, symmetric around the mean
- Defined by: Mean μ: center, Std. Dev: σ: spread
- 68-95-99.7 Rule:
  - ~68% within ±1σ, 95% within ±2σ
- Mean = Median = Mode = μ



Example: Heights of people, measurement errors.

Data is symmetrically distributed with no skew



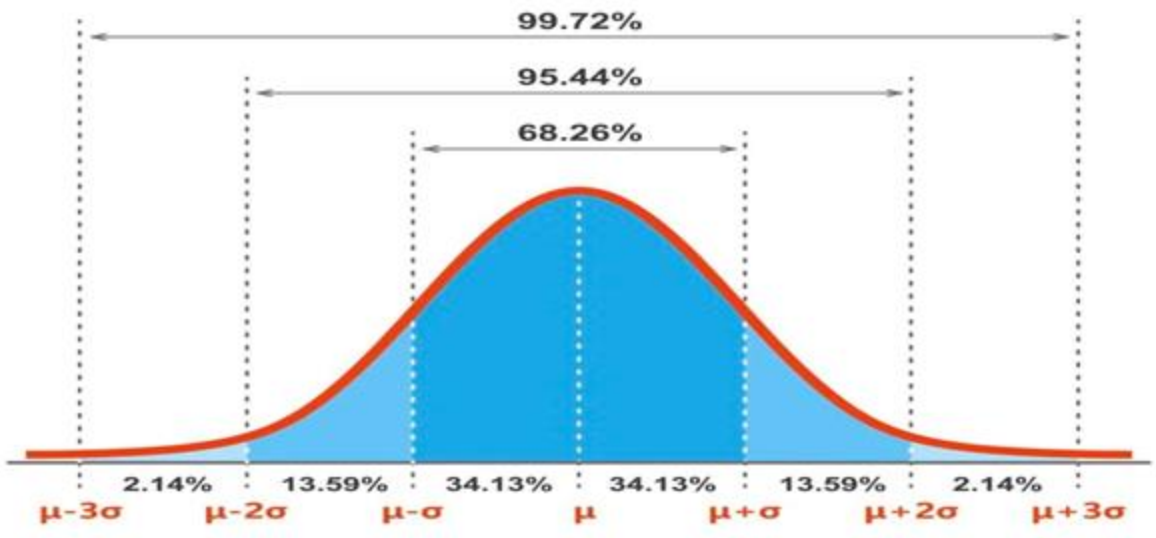
It is symmetric around its mean and is defined by two parameters: **the mean** (μ) and **the standard deviation (σ).**

The Normal Distribution (Gaussian) - A continuous PD

A continuous probability distribution characterized by its bell-shaped curve, symmetric about the mean.

Bell-shaped curve with most values clustering around the mean. **Key Parameters:**

- **Mean (μ):** Central tendency; determines the location of the peak. Change in μ shifts the curve horizontally along the x-axis.
- **Standard Deviation (σ):** Measures variability; controls the width of the curve. *Larger σ = wider spread; smaller σ = narrower spread.*



Properties: The Empirical Rule (68, 95, 99)

- 68% of data within μ±σ, 95% within μ±2σ, 99.7% within μ±3σ.

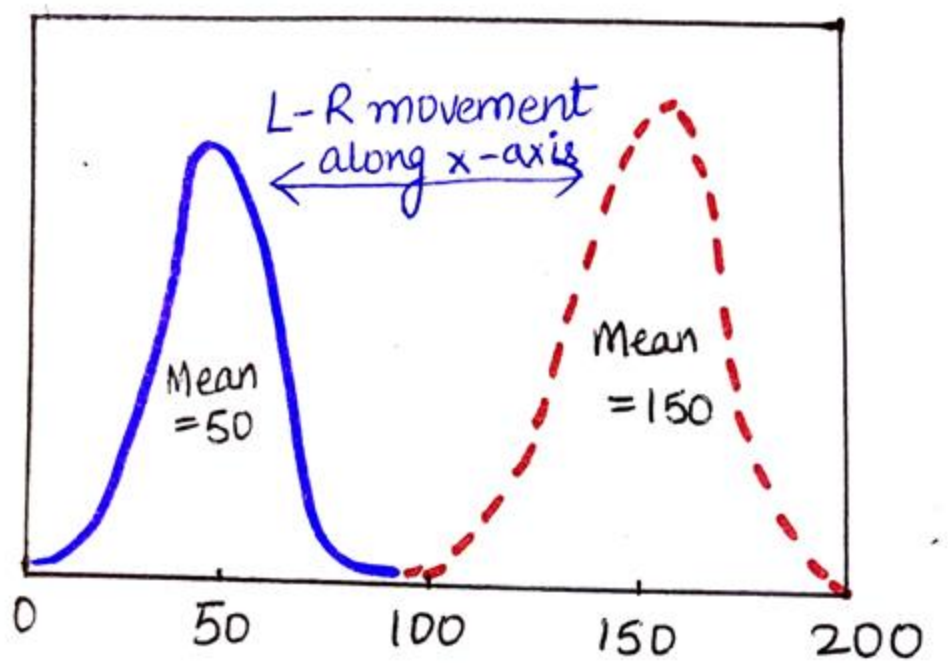
Area Under Curve: Equals 1, as it is a probability distribution.

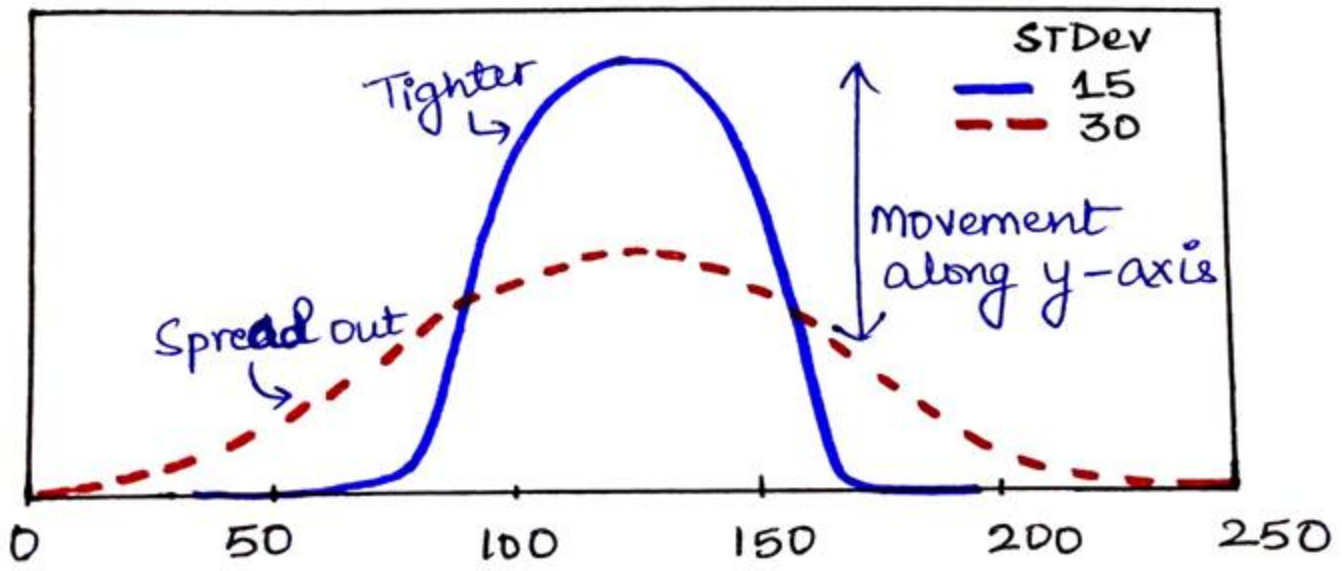
Parameters of the Normal Distribution: μ & σ

Mean (μ): Change in μ shifts the curve horizontally along the x-axis.

Standard Deviation (σ): controls the width of the curve.

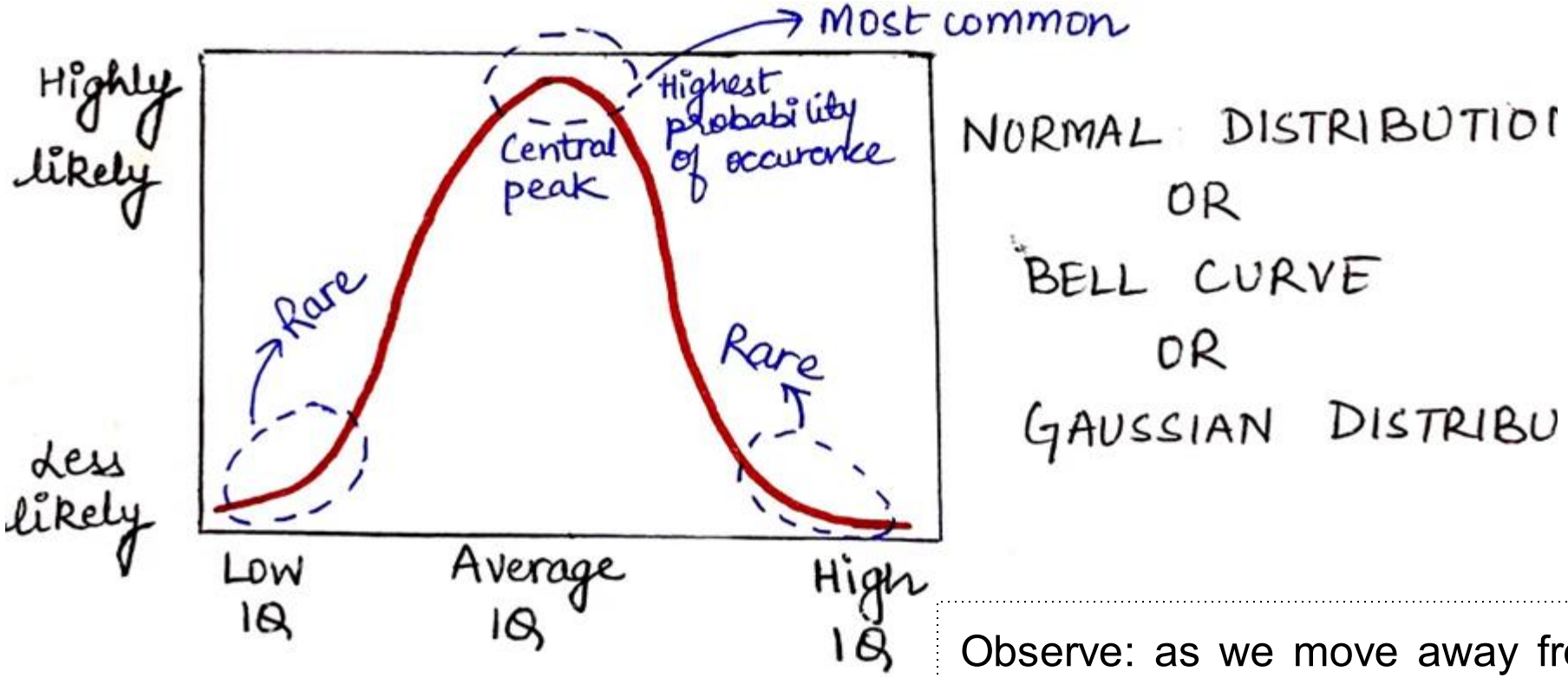
- Tighter Curve: Smaller σ → narrower width, taller peak.
- Spread-Out Curve: Larger σ → wider width, shorter peak





EXAMPLE: The Normal Distribution (Gaussian)

Example: the **IQ levels of a population** follow a normal distribution. It is understandable that people falling in very low IQ levels or very high IQ levels are rare occurrences and the majority of the population lies in the range of Average IQ scores.



Observe: as we move away from the central peak in both the directions, we see that the probability of occurrence of values at the tails of the curve becomes less and less likely.

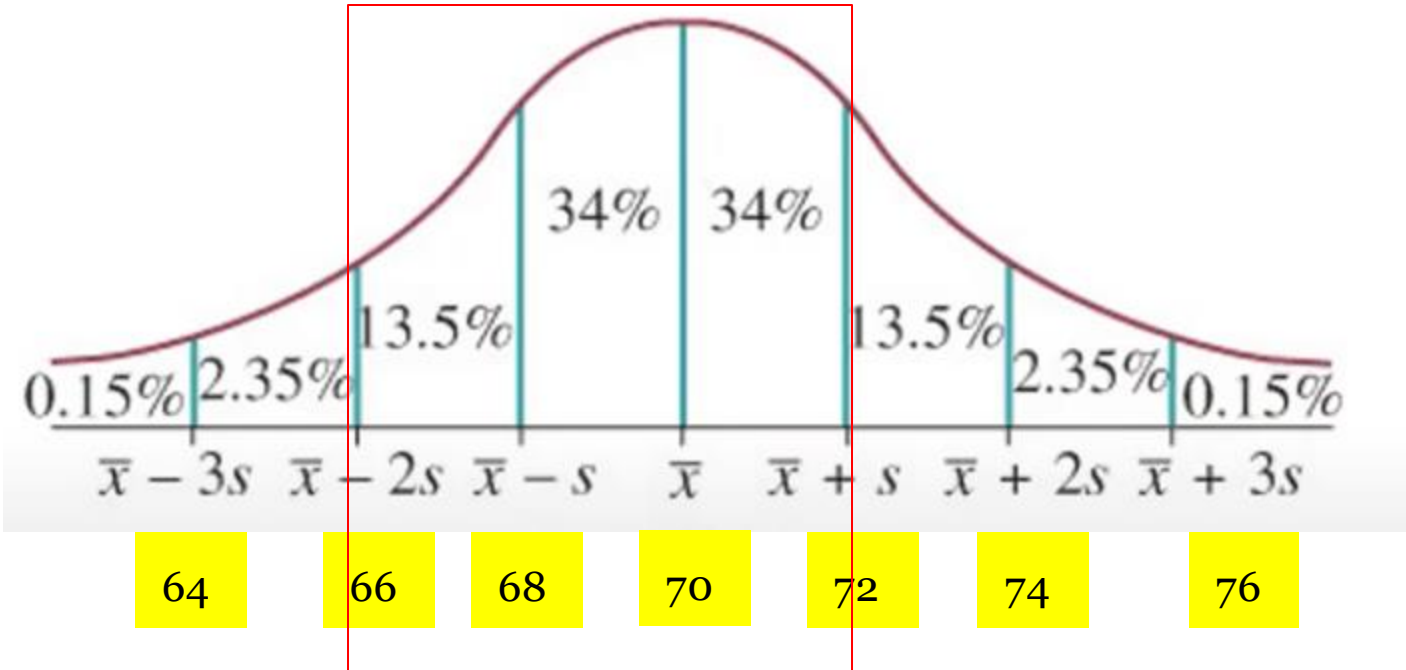
Problem Solving: Finding percentages

Example 1: The time taken to travel between two regional cities is approximately normally distributed with a mean of 70 minutes and a standard deviations of two minutes

Q: What is the percentage of travel times that are between 66 minutes and 72 minutes?

- Mean = 70 min.
- S = 2 min.
- % ?

Percentage = 13.5+34+34= 81.5 %

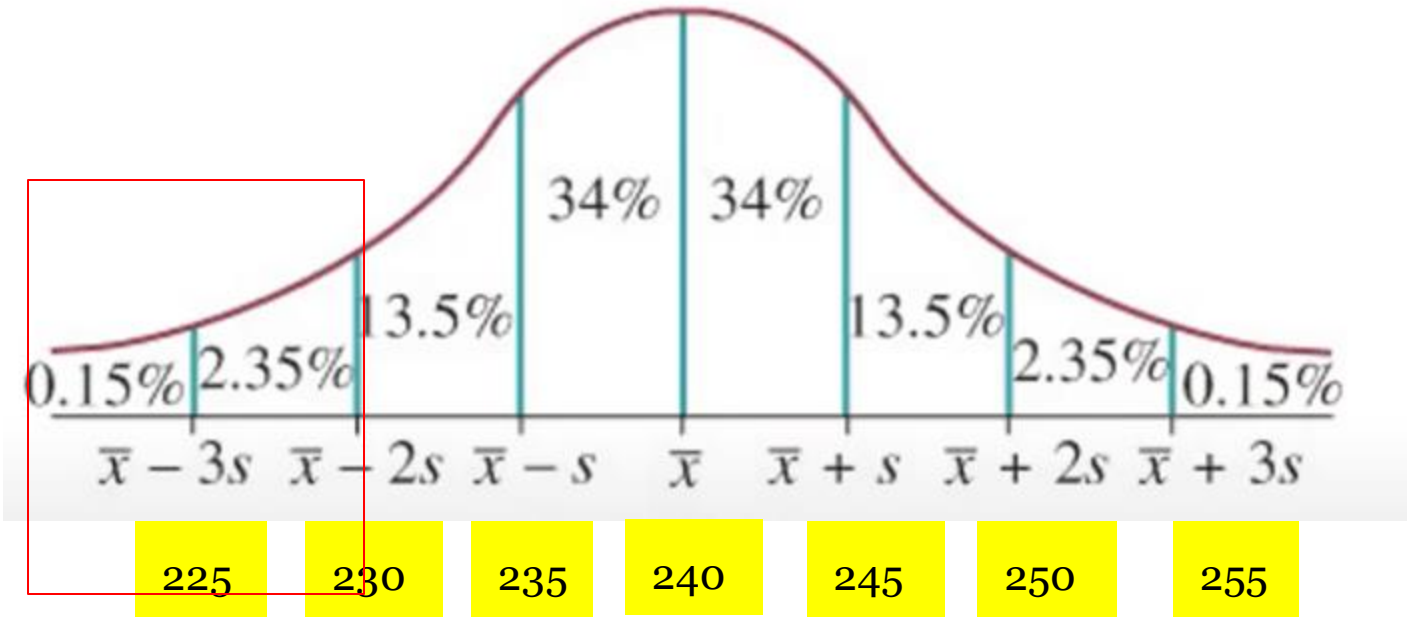


Try Yourself

The volume of a cup of soup serve via machine is normally distributed with a mean of 240 mL and a standard deviation of 5 ml. A fast store used this machine to serve 160 cups of soup

Ques: What number of these cups of soup are expected to contain less than 230 ml?

- Mean = 240ml
- S = 5ml
- % = ?
- cups = ?



- % = 0.15% + 2.35% = 2.5% = 0.025
- cups = 160 x 0.025 = 4 cups

Standard Normal Distribution: Special Case of Normal Distribution

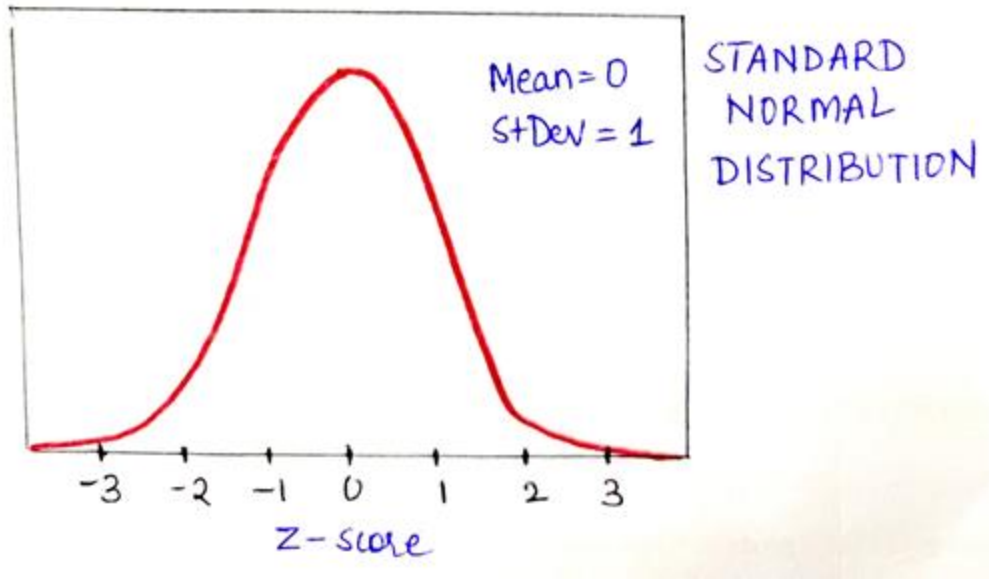
Standard Normal (or z) Distribution: Special Case of Normal Distribution with the mean μ is 0 and the standard deviation, σ is 1.

Allows for **comparison and analysis of data from different distributions** by standardizing them into z-scores, making it a fundamental tool in statistical analysis.

Z-Scores: Any normal distribution can be transformed into a standard normal distribution using:

$$z = \frac{x - \mu}{\sigma}$$

- x = individual value
- μ = mean
- σ = standard deviation



Interpretation of Z-Score: Measures how many standard deviations a data point is from the mean.

- Z = 0: At the mean.
- Z > 0: Above the mean.
- Z < 0: Below the mean.

Symmetric around 0

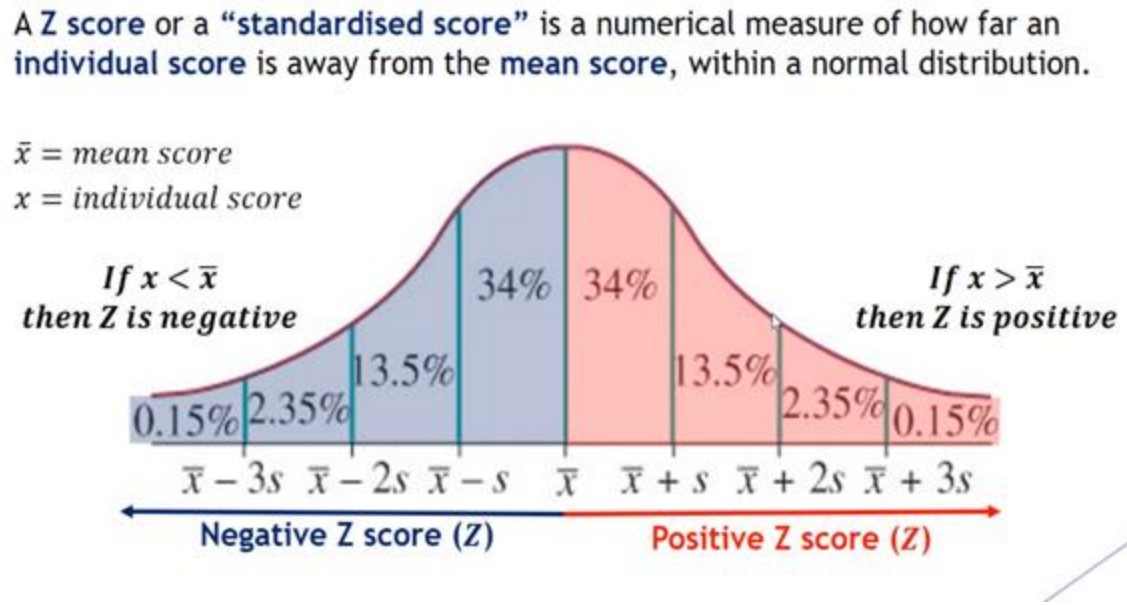
Standardization: Interpreting Z-Scores example

Any normal distribution can be standardized by converting its values into z scores.

Calculating Z-Scores: given μ = 100, σ = 15
If a student scored 115 on the test:

- Z-Score =(115 − 100) / 15 = 1

Interpretation: A Z-score of 1 means the score is 1 standard deviation above the mean, indicating better-than-average performance relative to the spread.



Benefit: When you standardize a normal distribution, the mean becomes 0 and the standard deviation becomes 1. This allows you to **easily calculate the probability** of certain values occurring in your distribution, or to compare data sets with different means and standard deviations.

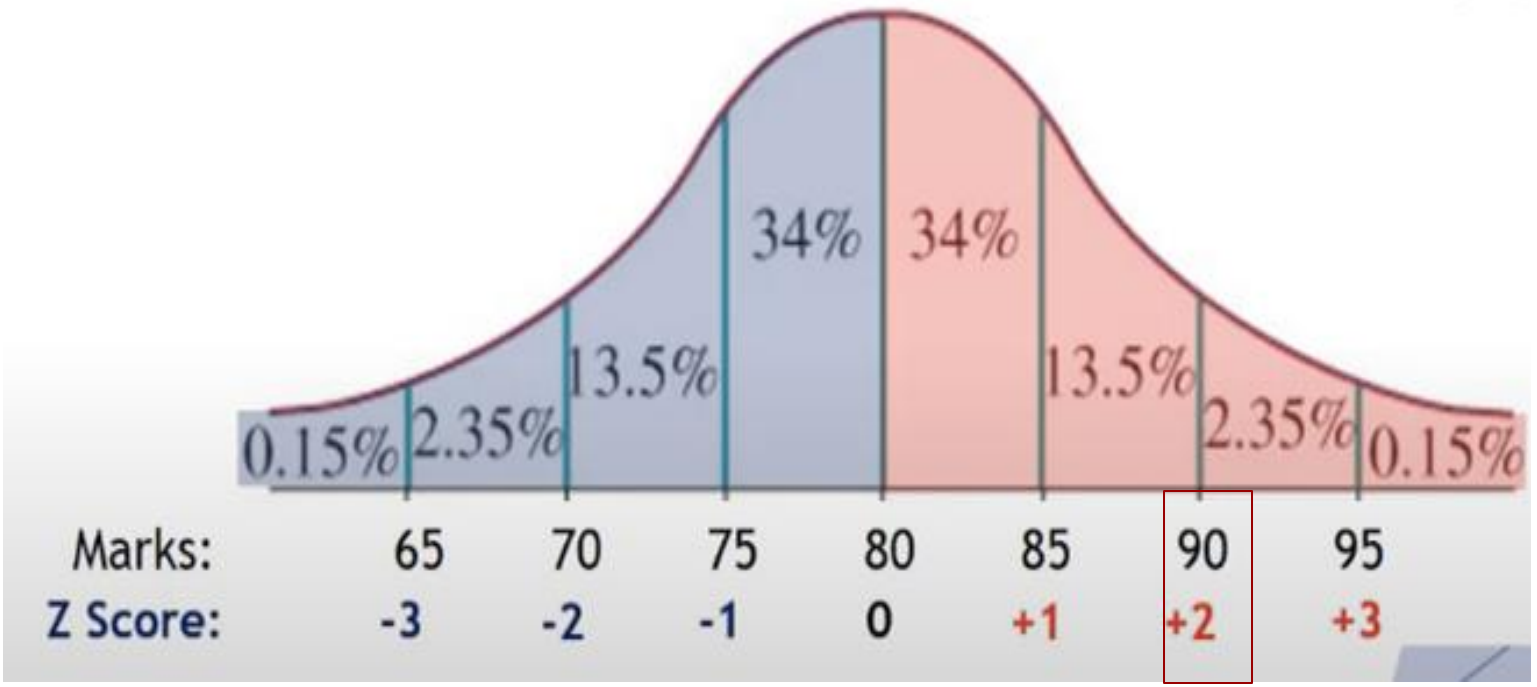
Z-Score Example:

Example: A math class sits a test with a mean score of 80 marks and a standard deviation of 5 marks. The distribution is approximately normally distributed. James achieves score of 90 marks. what is his Z-Score?

Individual score, x = 90
Mean = 80
Std dev = 5
Z = ?

Use equation:

$Z=(90-80)/5=2$



Other Distributions:

Not everything is normal

Poisson Distribution: Definition

A discrete probability distribution that models the **certain number of events (k)** occurring in a **fixed interval of time or space**.

Key Assumptions:

- Independent Events: Occur randomly; one event doesn't affect another.
- **Constant Average Rate (λ):** The average rate (λ) of occurrence is known and constant.

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

We can say, X follows a Poisson distribution with parameter λ

Where,

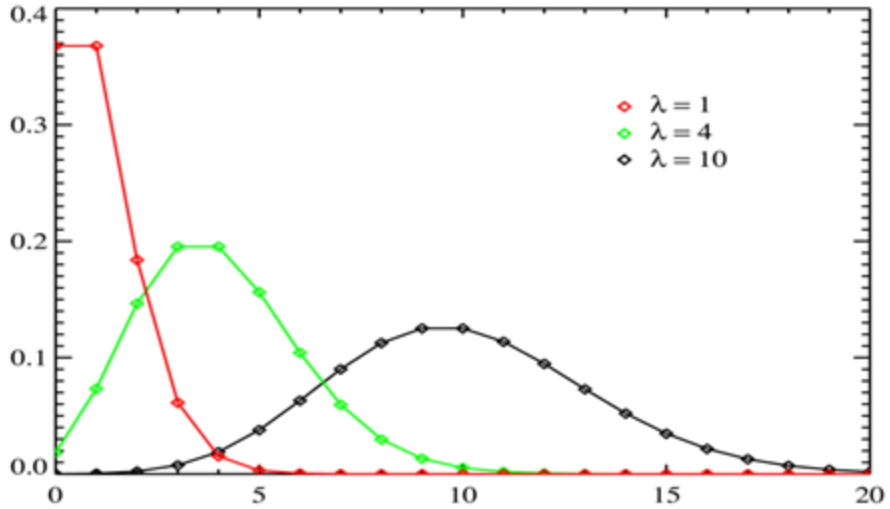
- P(X=k) = the probability of observing exactly k (0,1,2,3,.....) events.
- λ= average number of events per interval (the rate parameter)
- E = Euler's constant ≈ 2.71828

When to Use Poisson Distribution

More Precisely, It gives us the probability of a given specific number of events happening in a fixed interval of time.

Predicts the number of events (k) in a fixed interval (time/space).

- **Events:** Disease cases, customer purchases, meteor strikes, etc.
- **Interval:** 10 days, 5 square inches, etc.



Shape: Skewed right for small λ, becomes more symmetric as λ increases.

Example: A call center receives an average of 5 calls per hour. Here:

- λ=5: This is the **average** number of calls (events) in a **fixed interval of 1 hour**.

Note: When events follow a Poisson distribution, **λ is the only thing you need to know to calculate** the probability of an event occurring a certain number of times.

Example 1: Poisson Distribution

Example: If $\lambda = 3$ (e.g., 3 emails/hour), the probability of receiving exactly 2 emails is:

$$P(X = 2) = \frac{e^{-3} \cdot 3^2}{2!} \approx 0.224$$

TRY: Scenario: <u>how many</u> phone calls you receive during your 8 hours shift?

- **Average Rate λ :** 10 calls per entire 8-hour shift.
- **Assumptions:** Phone calls arrive randomly and independently of each other, with a constant average rate.

Use: Apply the Poisson distribution to model the number of calls.

Example: Calculate the probability **P(X=7)** of r the probability of receiving exactly 7 calls in the 8-hour shift.

Example 2: Poisson Distribution

Example: Births in a hospital occur randomly at an **average rate of 1.8** births per hour. What is the probability of observing **4 births** in a given hour at the hospital?

Sol: Let X =No. of births in a given hour

(i) Events occur randomly

(ii) Mean rate $\lambda = 1.8$

The probability of observing exactly 4 births in a given hour:

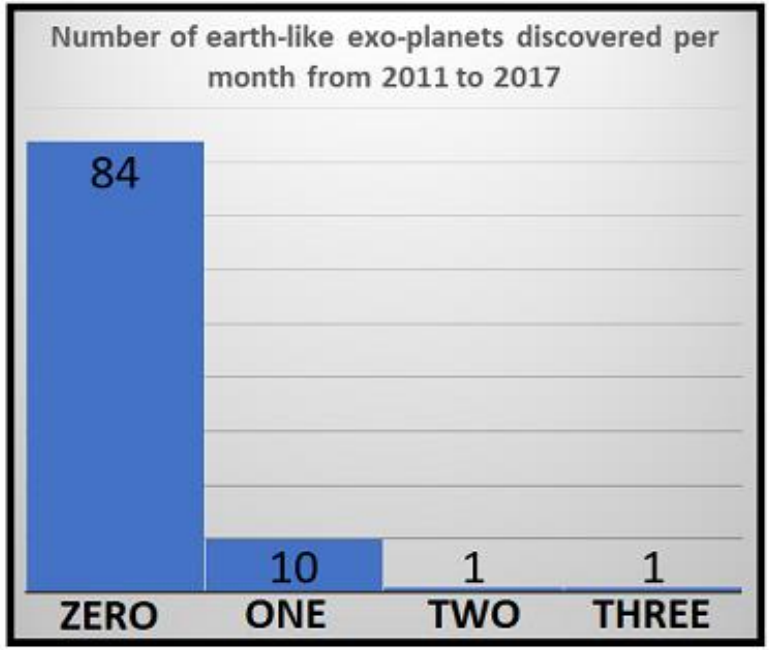
$$P(X = 4) = \mathbf{e}^{-1.8}\left(\frac{1.8^4}{4!}\right) = 0.0723$$

Can you still use Poisson Distribution?

Many real world phenomena produce counts that are **almost always zero**.

For example:

- Number of times a machine fails each month: *Many months, the machine may not fail at all (excess zeros).*
- Number of exoplanets discovered each year
- The number of billionaires living in every single city in the world: *Most cities have zero billionaires (excess zeros).*



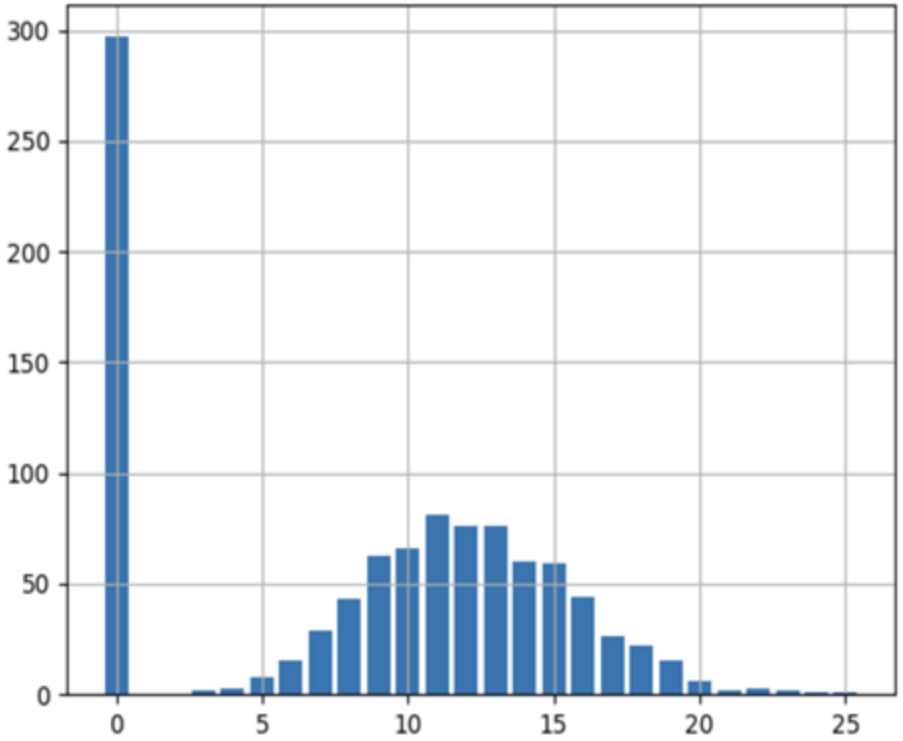
Zero-Inflated Poisson (ZIP) Distribution is a variation of the standard Poisson distribution that accounts for an excess of zero counts in the data.

Can you still use Poisson Distribution?

Oftentimes, Poisson distributions can have a spike at zero.

- Accounts for situations where there's a **higher chance of observing zero occurrences**.

We will use “Zero-inflated distribution” in these cases



Suppose you have data on the number of patients arriving per hour in an emergency room over a period of time. It is observed that sometimes there are hours with zero patients, even though the average number of patients per hour is around 5.

Remember, Any distribution can be zero inflated → You will need to deal with these differently

Zero-Inflated Poisson (ZIP) Distribution

A variation of the Poisson distribution used for count data with too many zeros.

- Model the probability of excess zeros separately from the rest of the count data by combining two components:
  1. **Poisson Component:** Models the count data using a Poisson distribution.
  2. **Inflation Component:** Uses a separate model (e.g., Bernoulli) to handle the extra zeros (represents probability of observing zero counts).

- A **zero-inflated Bernoulli (ZIB)** model handles **excess zeros** in binary data by combining **two Bernoulli distributions**. It assumes the data comes from a mixture of two groups: a "**non-susceptible**" group that is **always zero**, and a "**susceptible**" group that can be **zero or one** according to a standard Bernoulli process.
- This is often used in **regression** where the number of zeros is higher than a regular logistic regression would predict.

Example 2: ZIP Distribution

Scenario: Suppose a local clinic wants to model the number of annual visits by patients. Data shows that **40%** of patients make **zero visits** in a year, while the remaining **60%** have an **average of 2 visits** per year.

Next: The Bernoulli trial and Binomial Distribution

Bernoulli Distribution

Models: A single trial with two possible outcomes: success (1) or failure (0).

Parameters: p: Probability of success.

Probability Mass Function (PMF):

$$P(X = x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

Example: Tossing a coin once:

- p=0.5 (probability of heads).
- P(X=1)=0.5, P(X=0)=0.5.

Remember: Each trial is independent (the outcome of one trial does not affect the outcome of another trial.)

Binomial Distribution

Models: The number of successes (k) in **n** **independent** Bernoulli trials.

Parameters:

n: No. of trials;

p: Probability of success in each trial

Probability Mass Function (PMF):

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

- $\binom{n}{k}$: Number of ways to choose k successes out of n trials.

Example: Tossing a coin 10 times:

- N =10, p=0.5

Probability of exactly 4 heads:

$$P(X = 4) = \binom{10}{4} (0.5)^4 (0.5)^6 \approx 0.205$$

Example: Binomial vs Bernoulli distribution

When to Use:

- **Bernoulli:** For single events (e.g., one machine failure, one customer purchase).
- **Binomial:** For repeated events (e.g., number of failures in 10 machines, number of purchases in 100 customers).

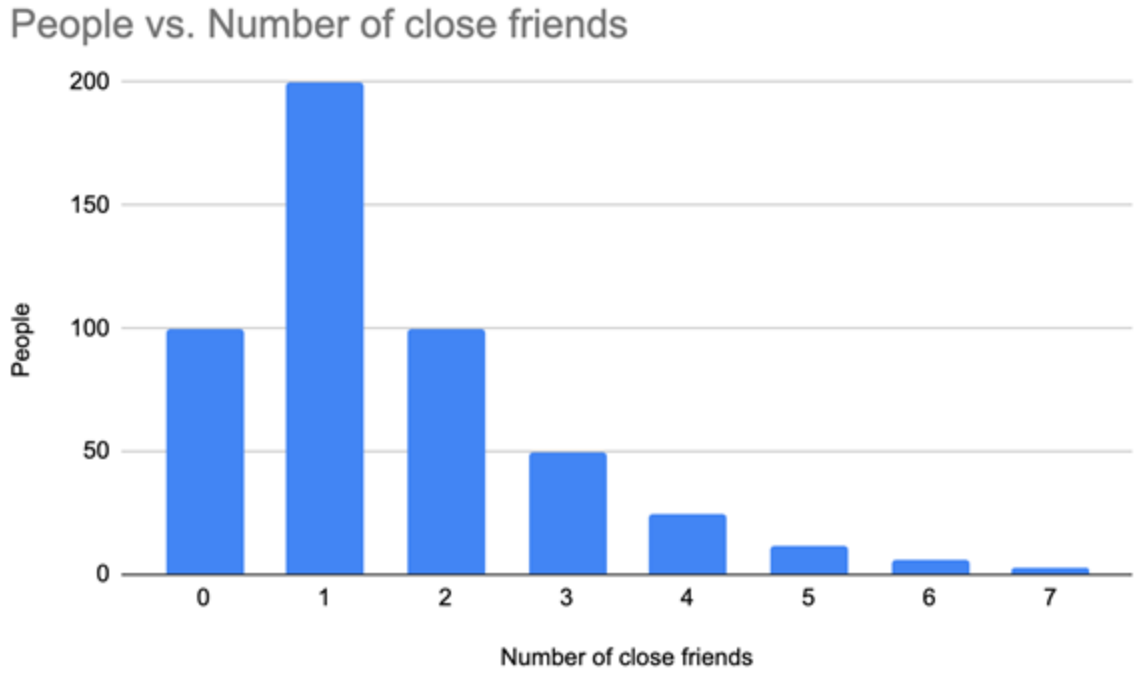
Consider :Attempting a quiz that contains 10 True/False questions.

Bernoulli trial → Answering a single T/F question

Binomial trial → Completing the entire quiz of 10 T/F questions.

Practice Problems

(a) What does the following discrete distribution tell us? (This is called a Power Law distribution)



(b) You are given 1000 datapoints representing how many dollars students have in their bank account. What would you expect the distribution to look like?

Probability distributions in Python

The `SciPy` library contains a `stats` module which has various functions for probability distributions. Functions for the probability distributions discussed in this section are described in the table below. Additional functions and further information about parameters can be found in the SciPy stats documentation.

Table 3.4.1: SciPy functions for probability distributions.

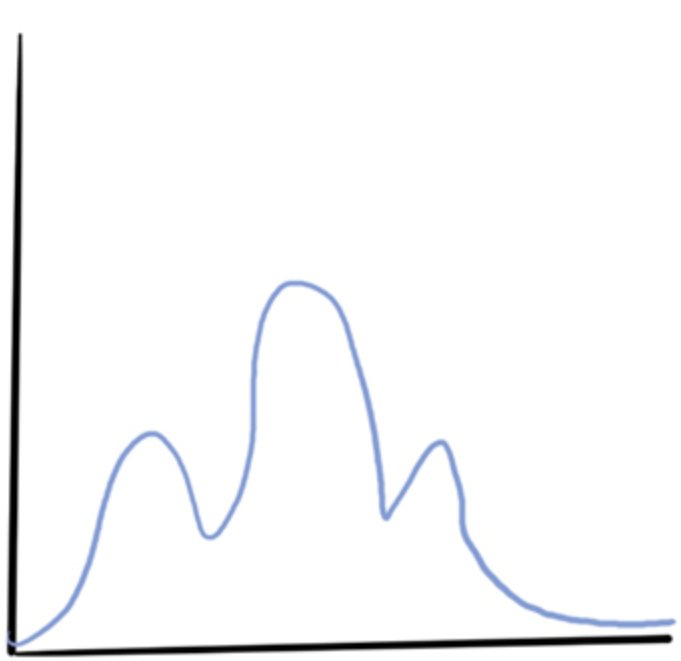
Distribution	Functions	Parameters	Description
Bernoulli	`bernoulli.pmf(k, p)` `bernoulli.cdf(k, p)`	`p`$= \pi$ sets the probability of a "success".	`bernoulli.pmf()` returns the probability $P(X =$ `k`$)$, and the `bernoulli.cdf()` returns the probability $P(X \leq$ `k`$)$.
Binomial	`binom.pmf(k, n, p)` `binom.cdf(k, n, p)`	`n`$= n$ sets the number of observations. `p`$= \pi$ sets the probability of a "success".	`binomial.pmf()` returns the probability $P(X =$ `k`$)$, and the `binomial.cdf()` returns the probability $P(X \leq$ `k`$)$.
Normal	`norm.pdf(x, loc, scale)` `norm.cdf(x, loc, scale)`	`loc`$= \mu$ sets the mean and `scale`$= \sigma$ sets the standard deviation.	`norm.pdf()` returns the density curve's value at `x`, and `norm.cdf()` returns the probability $P(X \leq$ `x`$)$.

The Central Limit Theorem (CLT):

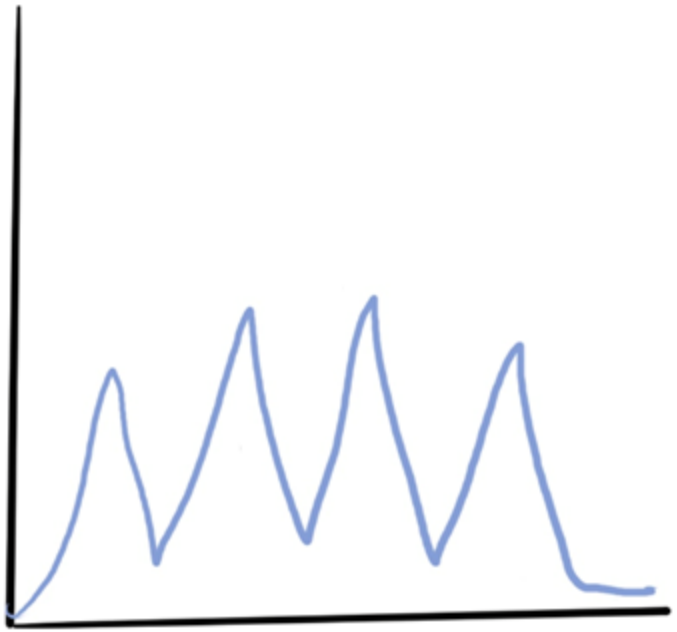
The sampling distribution of the sample mean (or sum) will be approximately **normally distributed** regardless of the original population distribution, **as the sample size increases**, provided the samples are independent and identically distributed.

Before Started

Comparing two non-normal distributions is hard.



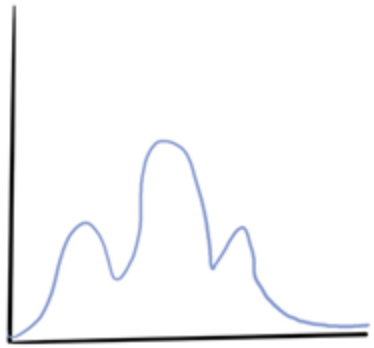
Some annoying distribution



Some other, even more annoying distribution

Before Started

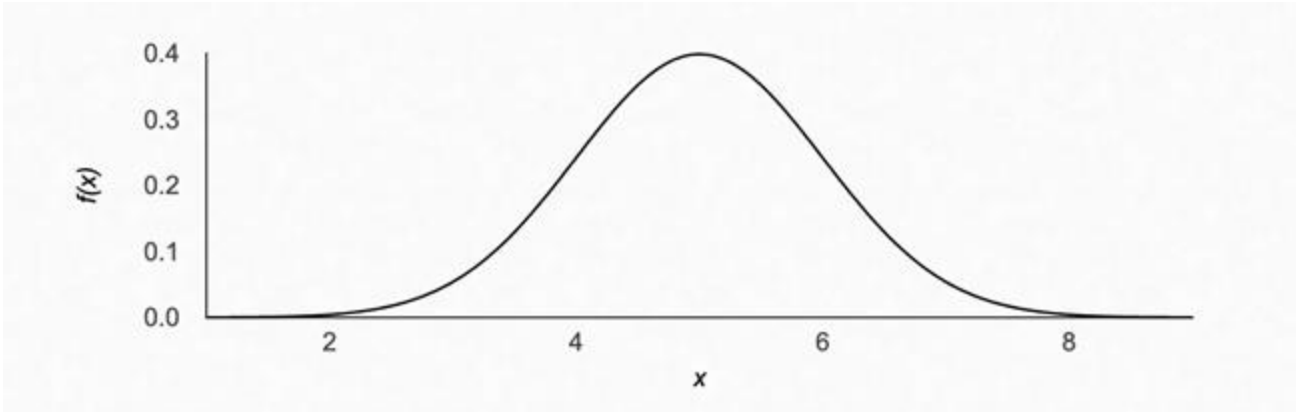
THE OBJECT LEVEL DISTRIBUTION

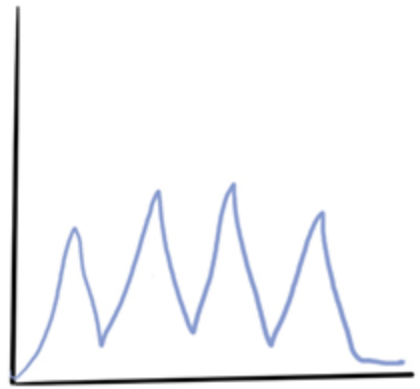


Some annoying distribution

Take a bunch of samples

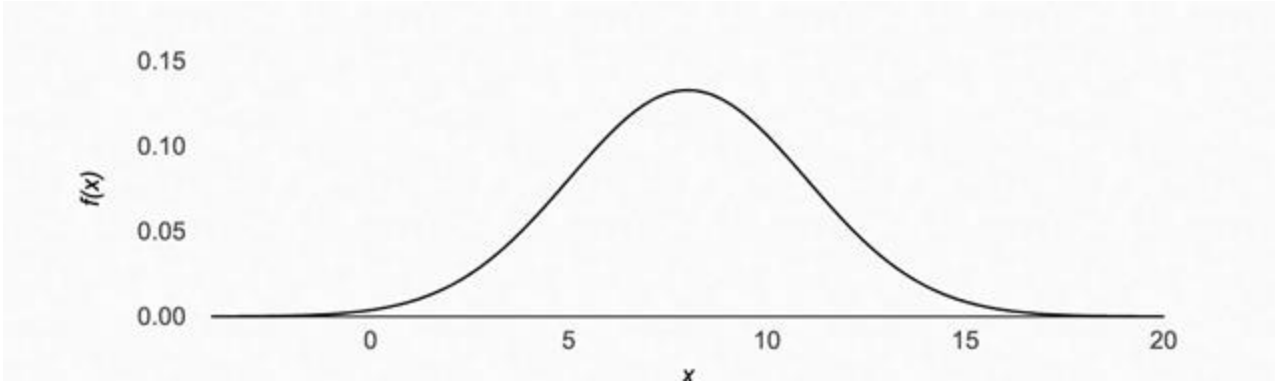
WE TOOK A BUNCH OF SAMPLES AND HERE ARE THEIR MEANS





Some other, even more annoying distribution

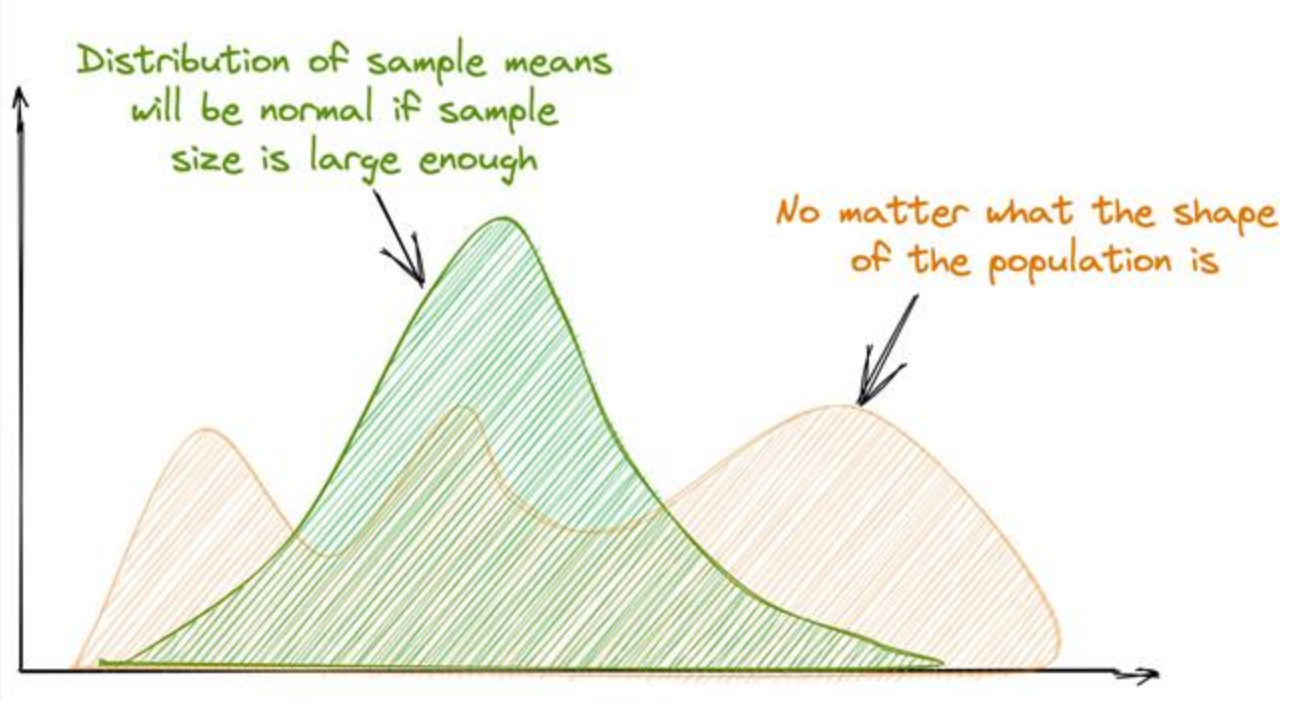
Take a bunch of samples



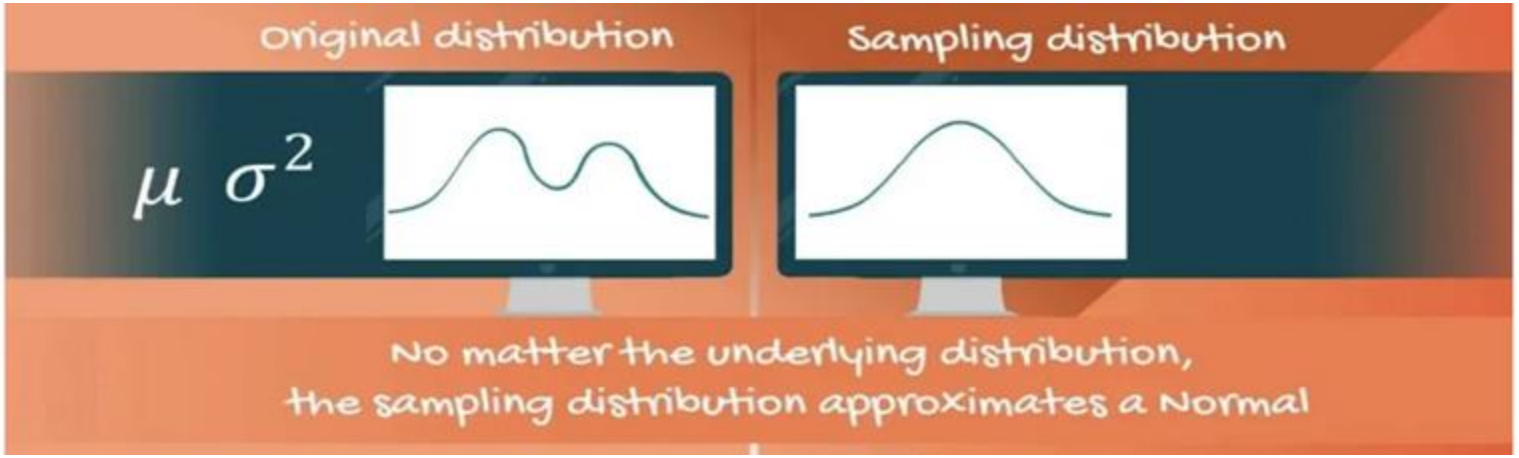
The Central Limit Theorem (CLT):

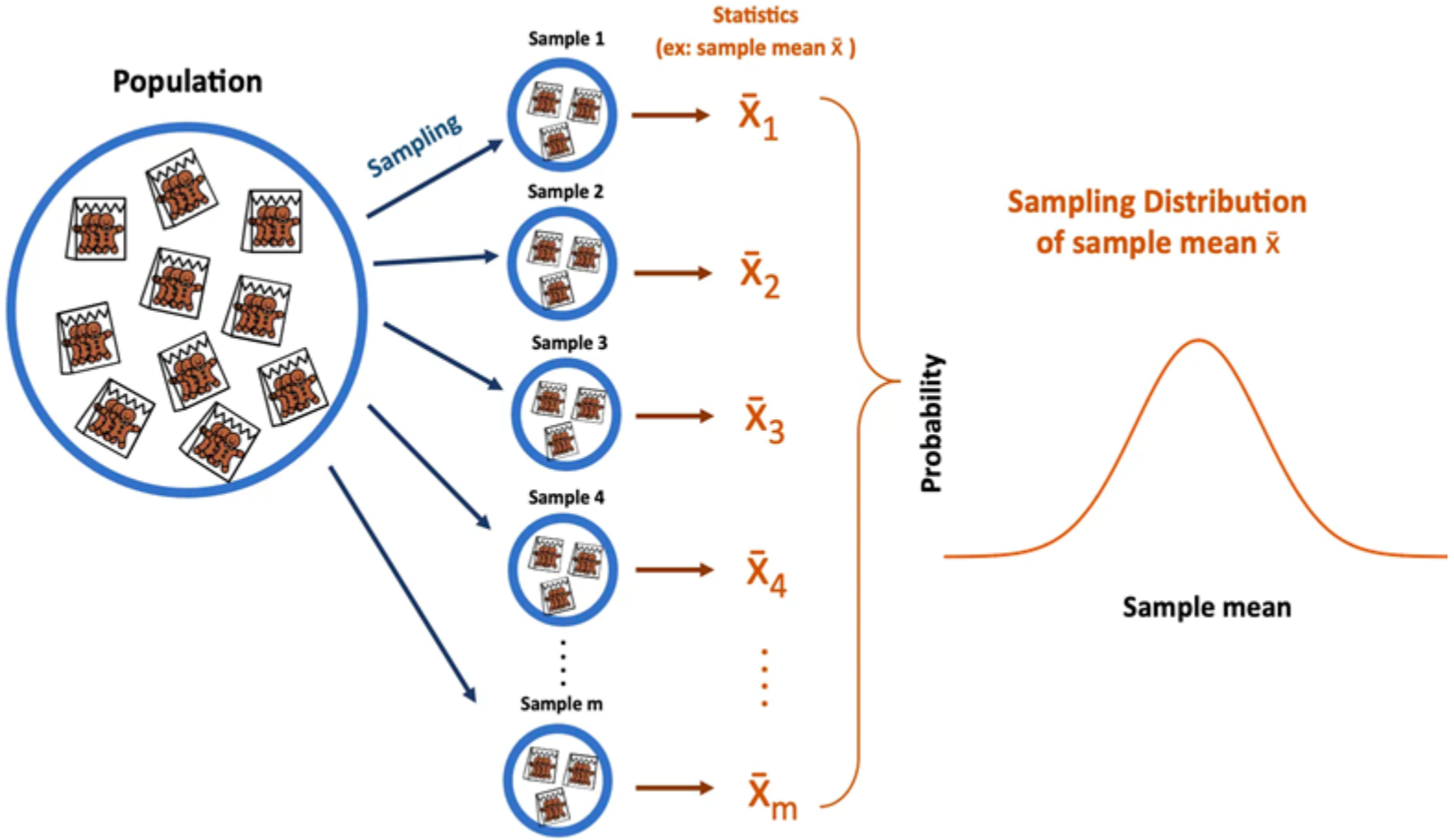
When we take **many random samples** of size n from any population with **finite mean** μ and **finite variance** σ^2, the **sampling distribution of the sample mean** $\bar{X}$ will **approximate a normal distribution** as n becomes large.

"Given a **sufficiently large sample** size, the sampling distribution of the **mean** will approximate a normal distribution, regardless of the population's original distribution."



The Central Limit Theorem





If we sample a distribution a bunch of times, the set of **sample means** is normally distributed.

Let’s see some examples

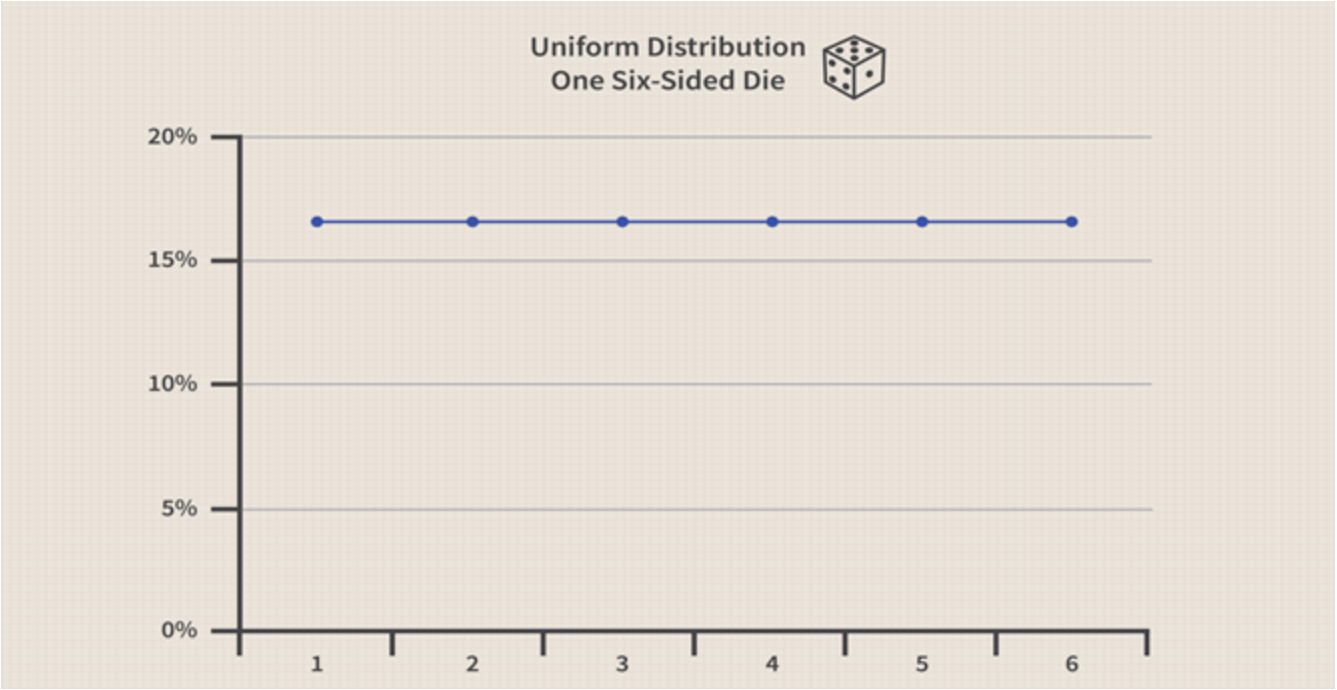
Example 1: Sampling from a Uniform Distribution

Say I take 5 samples from a uniform distribution

Most of the time, the mean will be 3.5

Sometimes, it won't be!

What is the rarest mean?



Example 1: Sampling from a Uniform Distribution

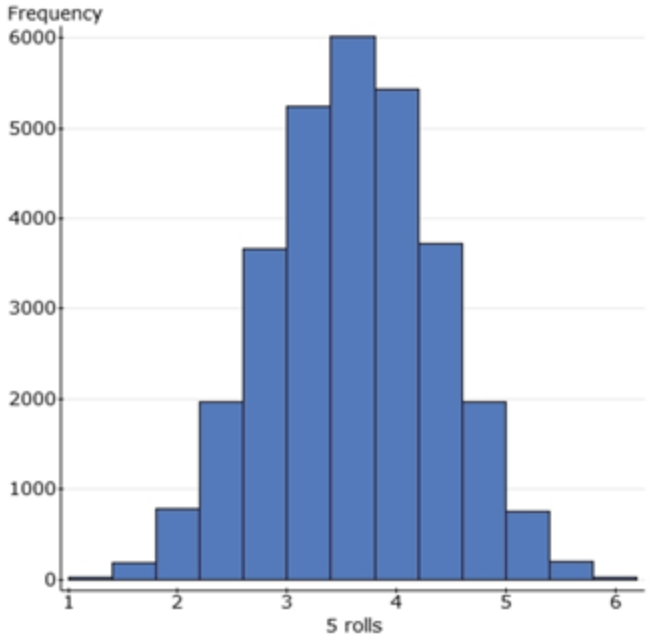
To be clear:

If we roll a dice **five times** we might get:

1. 1 2 3 4 5 6 → Average of 3.5
2. 1 1 6 6 3 4 → Average of 3.5
3. 2 2 3 3 4 4 → Average of 3.5
4. 1 1 1 1 1 1 → Average of 1
5. 6 6 6 6 6 6 → Average of 6

We know that the mean of 3.5 is the most likely outcome (reflects the expected value of a fair six-sided die.).
How can we discuss the likelihood of obtaining other sample means?

Yes. A distribution.



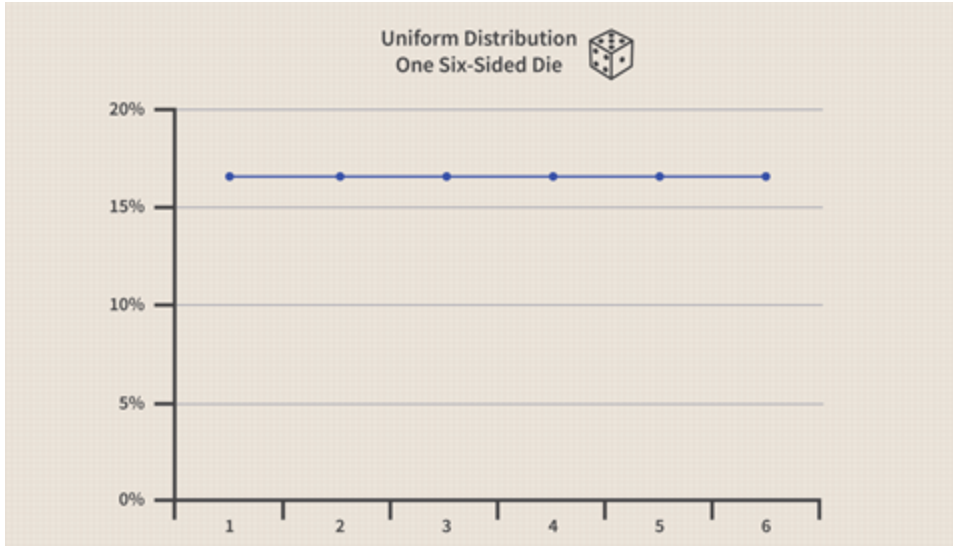
If you roll a dice five times, here is the frequency with which you will get each of the possible means.

Fig: Histogram showing the **frequency of each possible mean when rolling a die five times.**

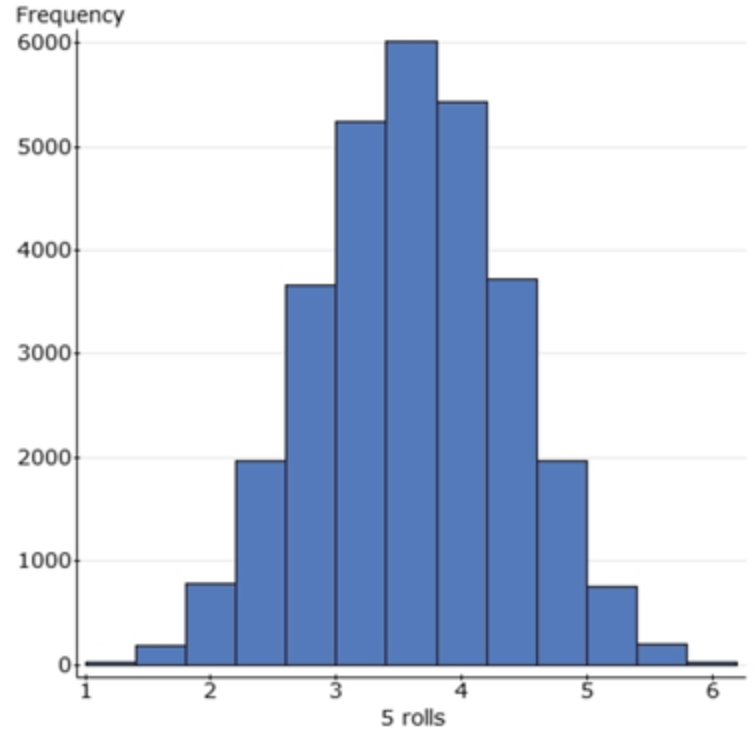
Sampling

According to the Central Limit Theorem, if you roll a die many times and calculate the mean each time, the sample means will follow a normal distribution.

We now have two separate things:



The original distribution



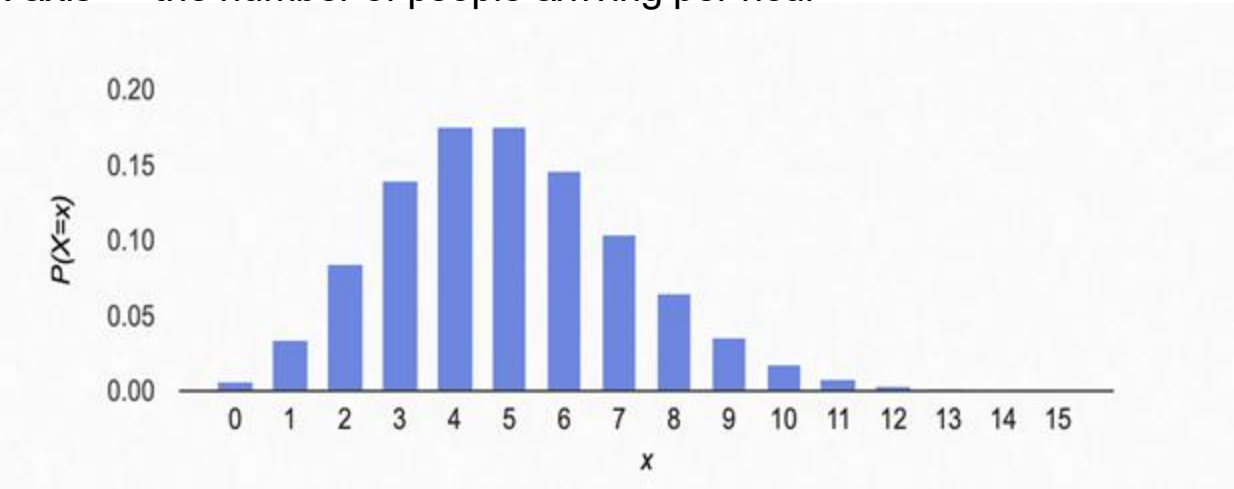
The potential distribution of means when we sample it

Inference: We started with the uniform data distribution but **the means of samples** drawn from it resulted in a normal distribution.

The distribution of the number of people arriving per hour will follow a Poisson distribution

y-axis →the probability of observing each number of arrivals

x-axis → the number of people arriving per hour



Arrivals are discrete (whole numbers, no fractions) and independent; one arrival does not affect another.

Characteristics:

Discrete Counts: Counts exact arrivals (1, 2, 3... people).

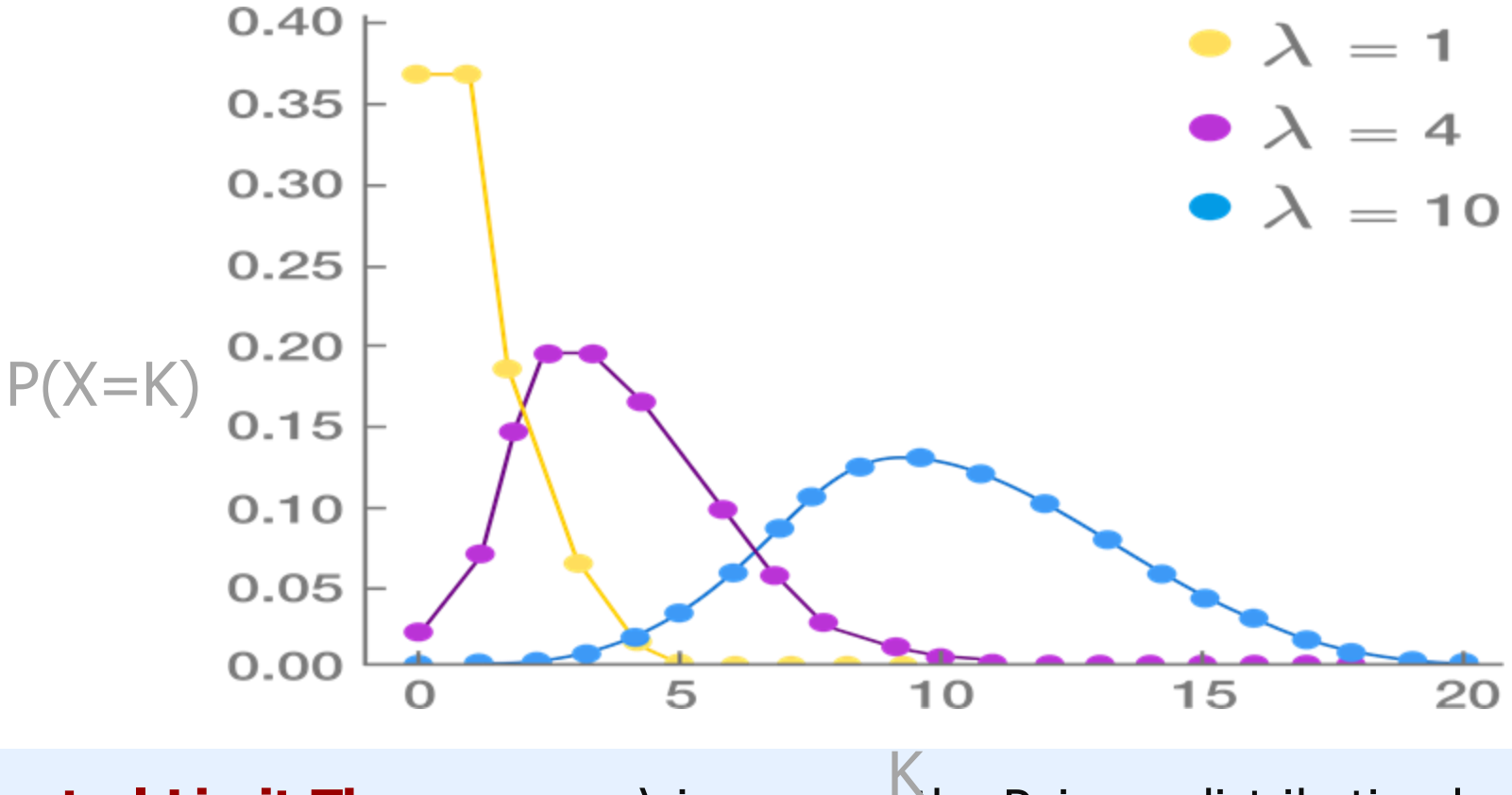
Mean (λ = 5): Highest density around 5 arrivals per hour.

Right-Skewed: Long tail on the right; most likely around 5 people/hour.

Variability: Most hours have around 5 arrivals; fewer chances of very high or low numbers.

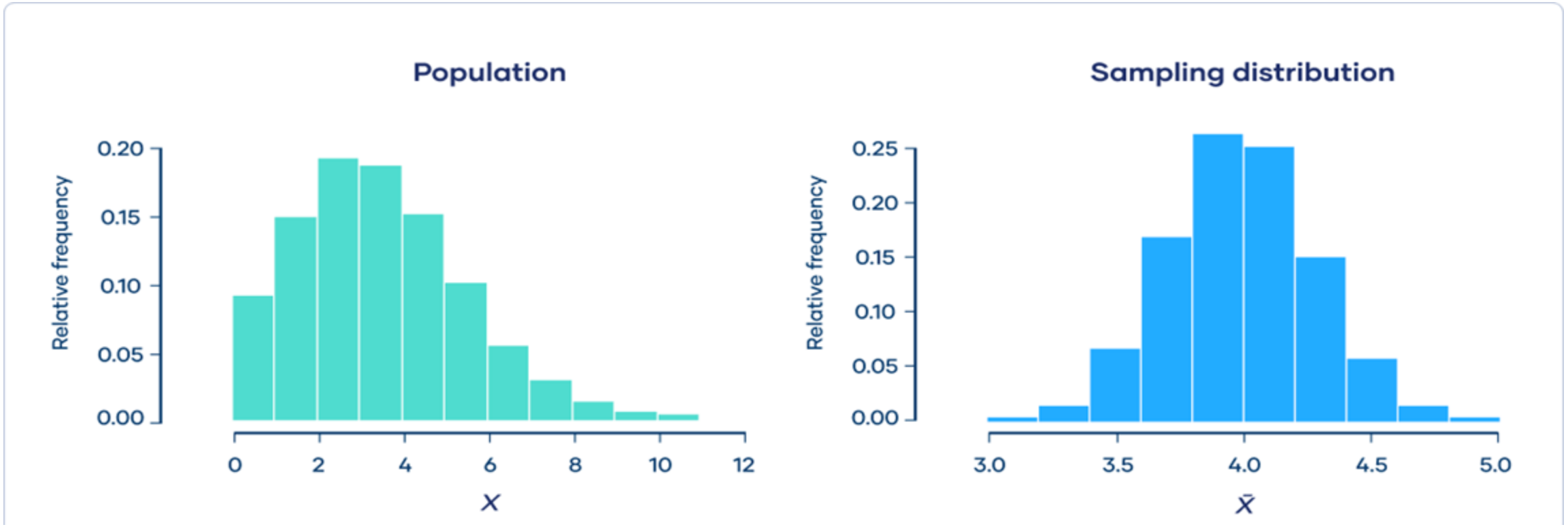
RECAP: Effects of Increasing λ in the Poisson Distribution

as λ increases (i.e 10 or greater), the Poisson distribution becomes more symmetric and bell-shaped, aligning with the characteristics of a normal distribution.



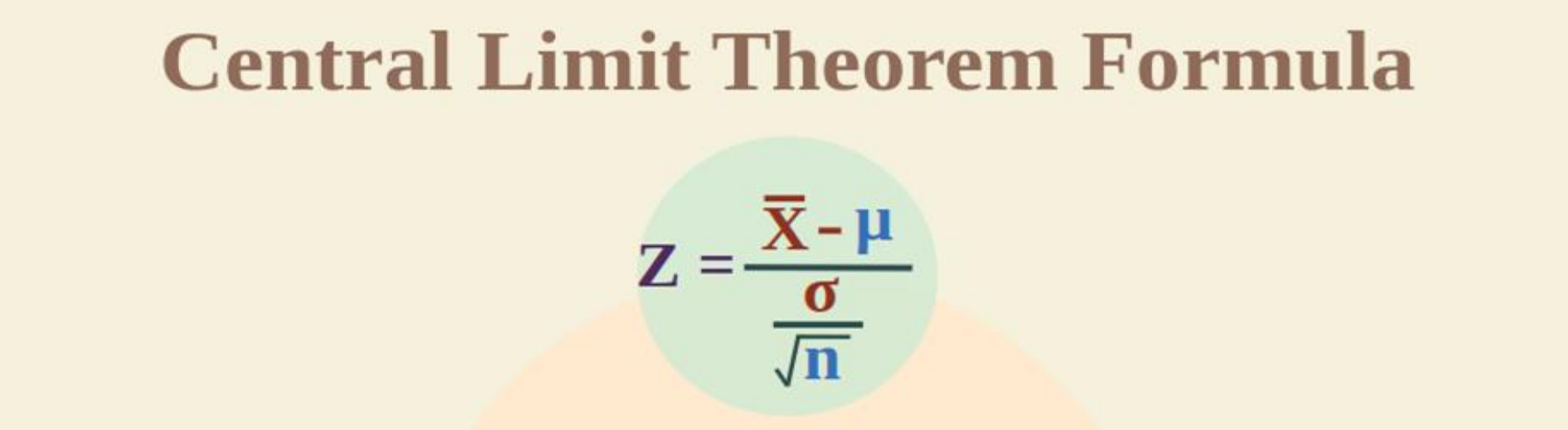
Due to the **Central Limit Theorem**, as λ increases, the Poisson distribution becomes more like a normal distribution, showing a bell-shaped curve

Example 2: Another Example: The Central Limit Theorem



A **population** follows a **Poisson distribution (left image).**
If we take 10,000 **samples** from the population, each with a sample size of 50, the sample means follow a normal distribution, as predicted by the **central limit theorem (right image).**

Formula: The Central Limit Theorem



Sample mean = Population mean = μ

Sample standard deviation = $\frac{\text{(Standard deviation)}}{\sqrt{n}}$

$= \frac{\sigma}{\sqrt{n}}$

NOTES: As n increases, the sample mean distribution approaches a normal distribution, regardless of the population's original distribution

Central Limit Theorem Application

The Central Limit Theorem (CLT) is useful in various scenarios, particularly when analyzing an entire population is difficult. Here are some key applications:

- **Data Science:** The CLT helps make accurate assumptions about a population to build robust statistical models.
- **Applied Machine Learning:** The CLT aids in making inferences about model performance.
- **Statistical Hypothesis Testing:** The CLT is used to determine if a given sample belongs to a specific population.