Day · 76, Feb 14, 2025 (Falgun 2, 2081 B.S.)

Probability Distributions:

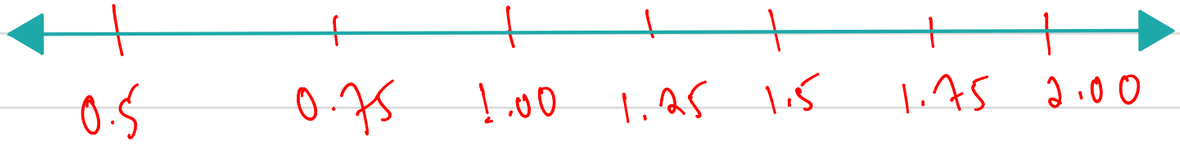
→ Describe the likelihood of the possible outcomes of a random event. Eg: probability of succesful clinical trials.

Random Variables. Represents the values for the possible outcomes of a random event.

Discrete Random Variable → has a countable possible values of P

Continuous Random variable → Takes all the possible values in some range of numbers. It measure the outcome (continuous).

Continuous values.



↳ Discrete Distributions represent discrete random variables.

↳ Continuous Distributions represent Continuous Random variables.

Sample Space:

- The set of all possible values for a random variable.
  - Sample space for single coin toss = {Heads, Tails}.
  - Sample space for single Die Roll = {1, 2, 3, 4, 5, 6}

Single Die Roll

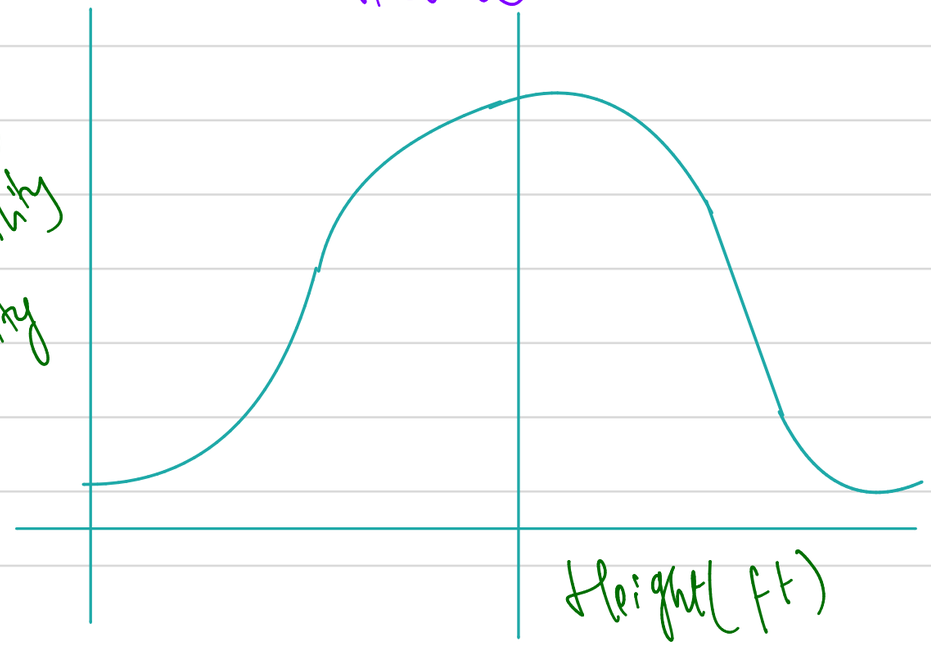
- Sample space = {1, 2, 3, 4, 5, 6}

- probability of each outcome = 16.7%

Infinite number of values = Continuous Random Variable.

eg. 5.67 + 5.66 + 5.15 + ... + ∞

So, we need Continuous Probability Distribution using Probability Density Function —



Binomial Distribution

A discrete distribution that models the probability of events with only two possible outcomes, success or failure.

- Each event is independent
- the probability of success is the same for each event.
- Eg: tossing the coin for 10 times where each toss is 'F or 'S'.

Mutually Exclusive

Two outcomes are mutually exclusive if they cannot occur at the same time.

↳ Binomial Distributions used in ML, Banking, medicine and investing.

Random Experiment

→ A process whose outcome cannot be predicted with certainty.

→ More than one possible outcome and each outcome of the experiment depends on chance.

Binomial Experiment

↳ Consists of a number of repeated trials

↳ Each trial has only two possible outcomes and the 'p' is Success is the same for each trial.

↳ Each trial is independant.

Binomial Experiment Examples:

- 10 repeated coin toss
  - 2 possible outcomes: heads or tails
  - 'p' of success for each toss is the same: 50%.
  - the outcome of one coin toss does not affect the outcome of any other coin toss.

- 100 customer visits with two possible outcomes: return or not return
  - p of success for each customer visit is the same = 10%.
  - the outcome of one customer visit does n't affect the outcome of any other customer visit.

Binomial Distribution formula

$$P(X=k) = \frac{n!}{k!\,(n-k)!}\, p^k (1-p)^{n-k}$$

- k refers to no. of Success
- n refers to the number of trials
- p refers to the probability of Success in a given trial.

$$P(X=k) = \frac{n!}{k!\,(n-k)!} * p^k * (1-p)^{n-k}$$

$\frac{n!}{k!\,(n-k)!} \rightarrow {}^{n}C_{k}$ (n-choose k).

$C\binom{n}{k} \rightarrow$

$P(X=0) = 0.729$

$P(X=2) = 0.027$

$P(X=3) = 0.001$.

Normal Distribution:

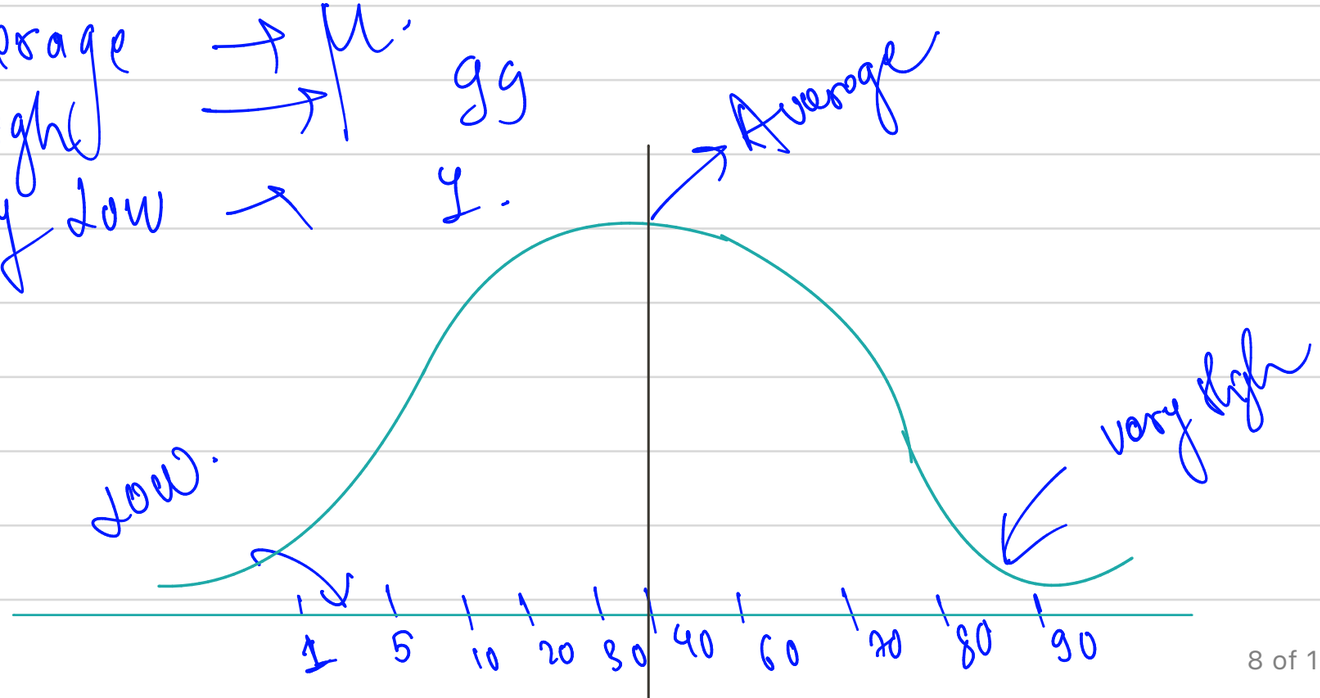
↳ While Discrete probability Distributions like Binomial & poisson Distribution can only take the discrete value into an account.

Normal Distribution:

→ A continuous probability distribution that is symmetrical on both sides of mean and bell-shaped.

Eg: Maximum number score Average → μ.
very less " " high → 99
very " " " very low → 1.

↳ used in Advanced Statistics like Hypothesis testing. (In terms of score.)



Features of Gaussian Distribution:

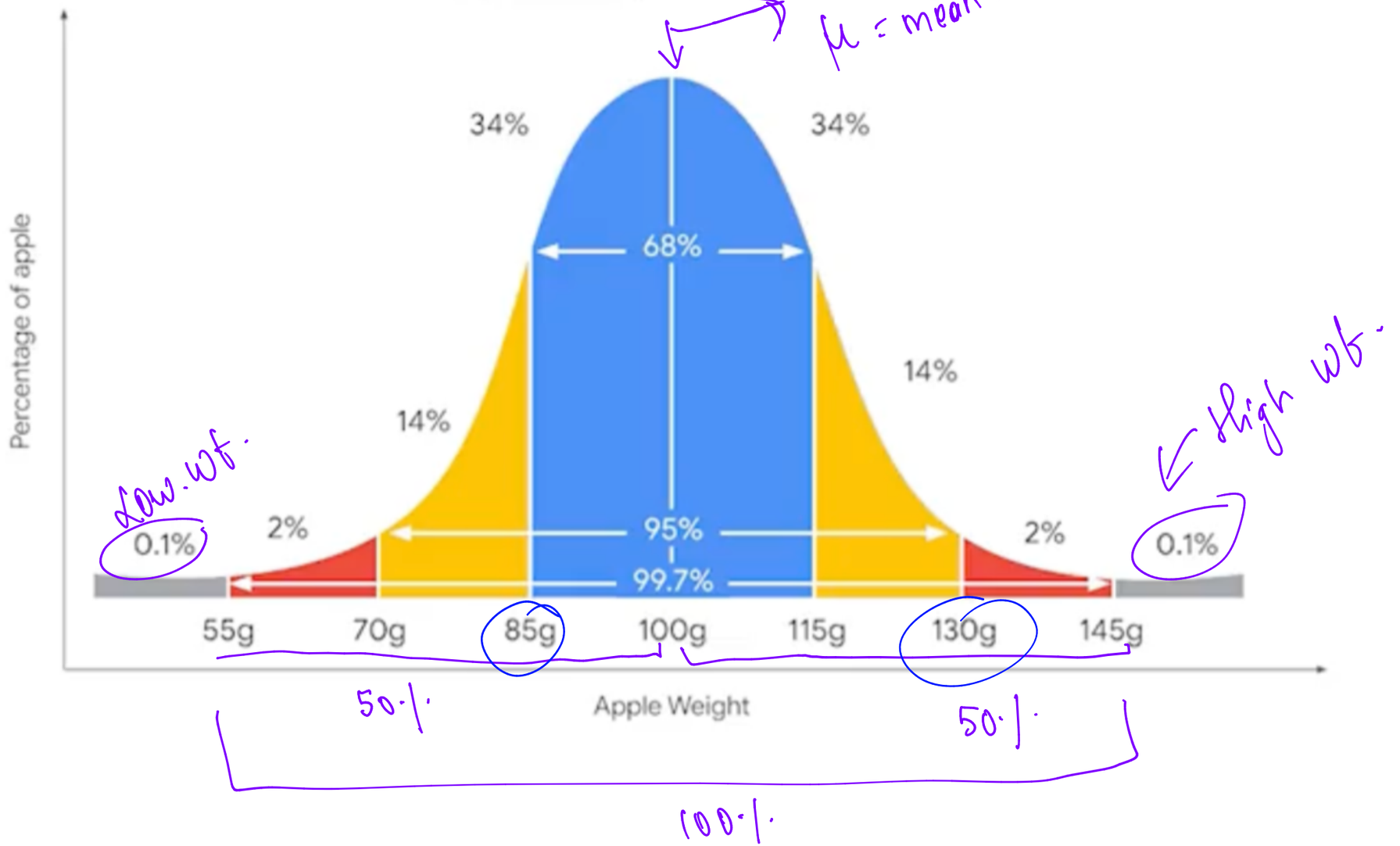
1. Shape is a bell curve.
2. Mean is located at the center of the curve
3. Curve is symmetrical on both sides of the center
4. the total area under the curve equals 1.

Standard Deviation:

Calculates the typical distance of a data point from the mean of your dataset.

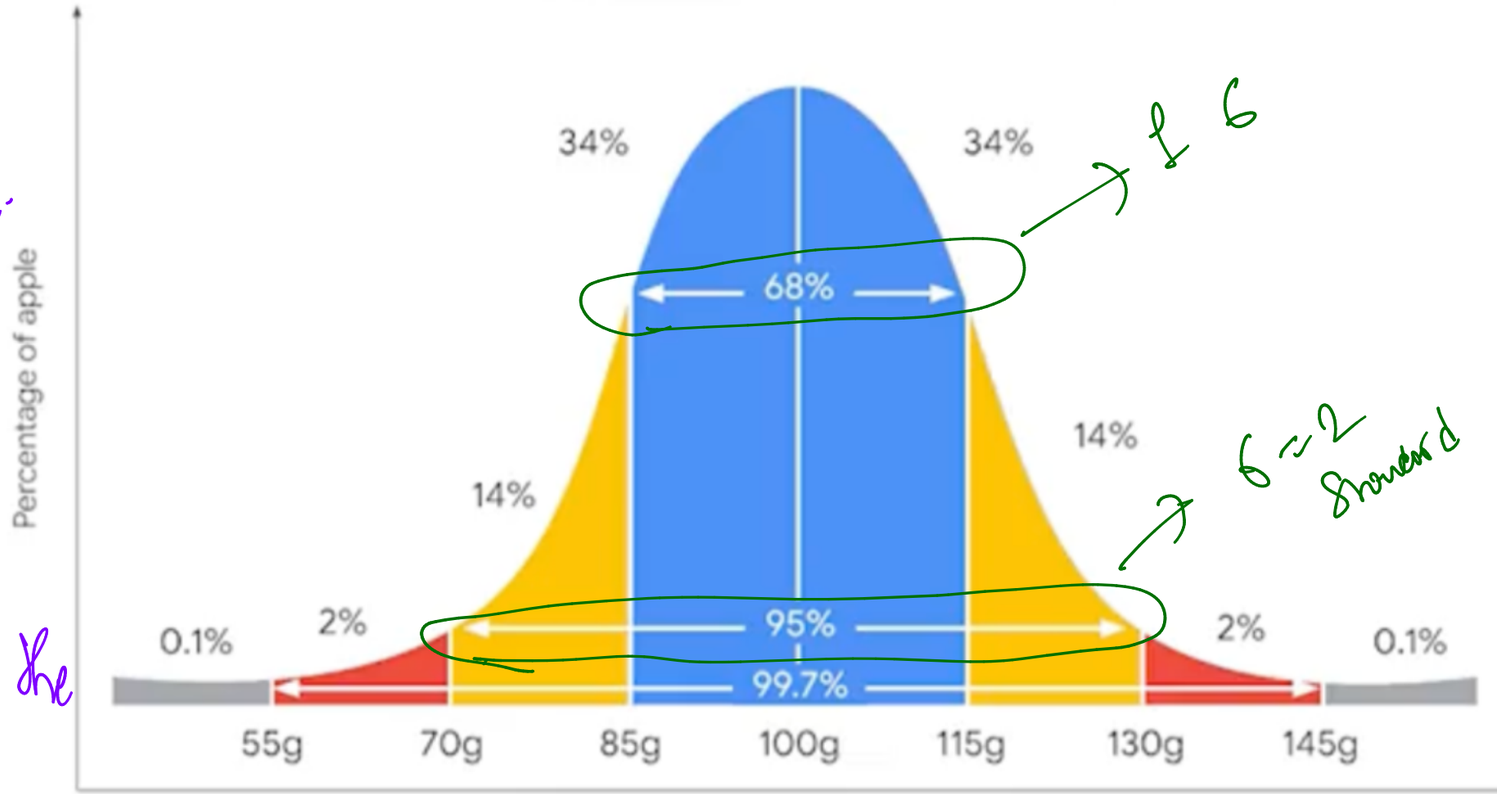
mean → center

σ ⇒ Spread of data from the center or mean.


Apple Weight Distribution
μ = mean = mode = median
34%
34%
68%
14%
14%
95%
2%
2%
99.7%
0.1%
0.1%
low wt.
high wt.
55g
70g
85g
100g
115g
130g
145g
Percentage of apple
Apple Weight
50%
50%
100%

Empirical Rule:

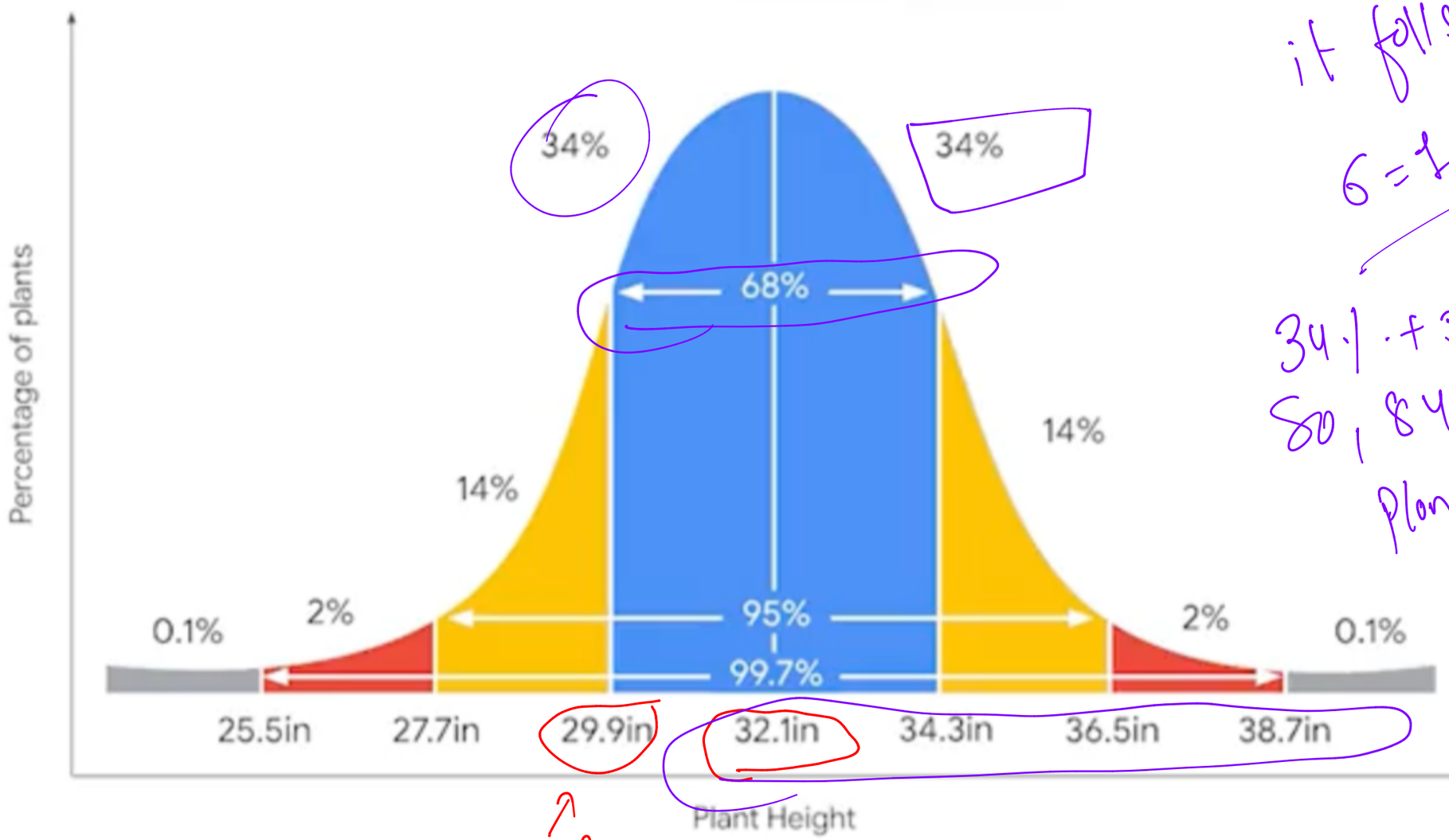
- 68.1% of values fall within 1 standard deviation of the μ.
- 95.1% of values ... 2 standard deviation of the μ.
- 99.7% of values fall under the 3 standard deviation of μ.



→ 6 = 1 means 68% which means 85g and 115 g between the ranges of apple weight ranges.

→ 6 = 2 means of ranges between 70g and 130g.

So, 34% from the centre of mean for plant to be atleast 29.9 inch tall because if falls under the $\sigma = 1$ standard.



$34\% + 50\% \Rightarrow 84\%$

So, 84% of Plants are taller than 29.9 in.

At least 29.9 in tall?

Explanation:

If you want to determine the percentage of plants taller than 29.9 inches for your backyard landscape design, you can use the Empirical Rule of normal distribution.

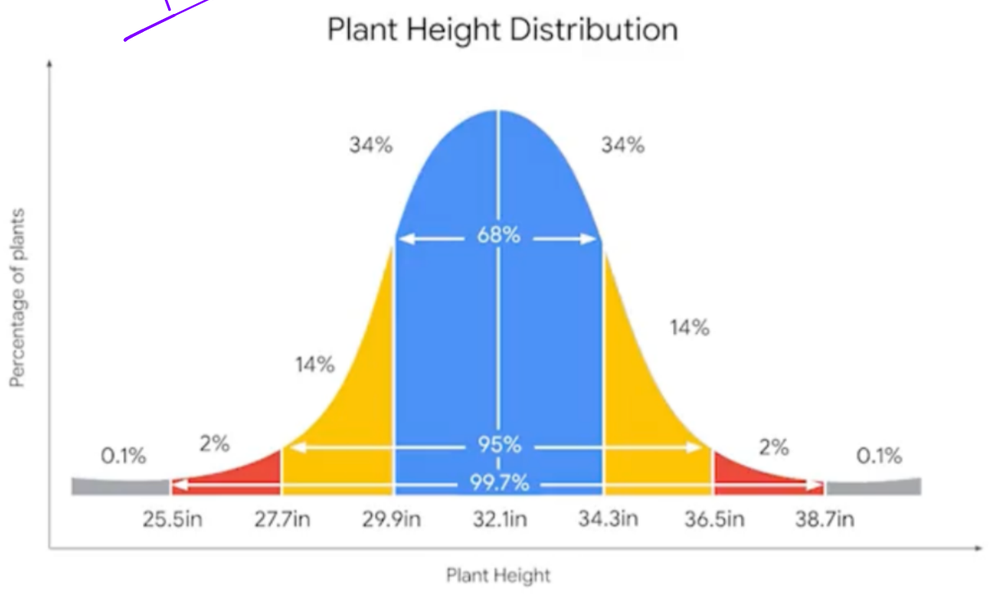
Since 29.9 inches is one standard deviation below the mean, the Empirical Rule states that 68% of values fall within one standard deviation of the mean. This means 34% of values lie between 29.9 and the mean. Additionally, in a normal distribution, 50% of values are above the mean.

To find the percentage of plants taller than 29.9 inches, add these two values:
34% (between 29.9 and the mean) + 50% (above the mean) = 84%.

Thus, 84% of your plants meet the height requirement. This quick estimation method helps analyze data patterns efficiently. As a future data professional, understanding the normal distribution will be key to interpreting and making data-driven decisions.

Source: The power of Statistics Coursera Course provided by Google.

Explanation:



Standardize Data using Z-scores:

→ Z-score is a measure of how many σ below or above the population mean a data point is.

→ Z-score gives idea on how far the data is from the μ.

→ Z-score is 0 if the value is equal to the mean.

→ Z-Score is +ve if the value is greater than the μ.

→ Z-Score is -ve if the value is less than the μ.

Standardization the process of putting different variables in the same scale.

Z-score helps to understand the relationship between different Dataset.

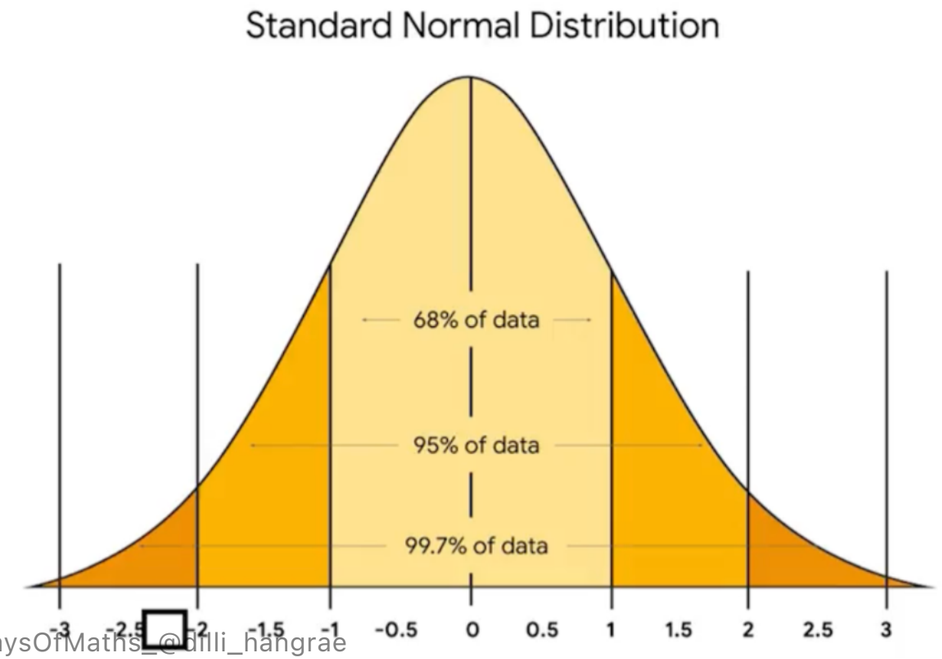
Z-Score Application in Anomaly Detection Application

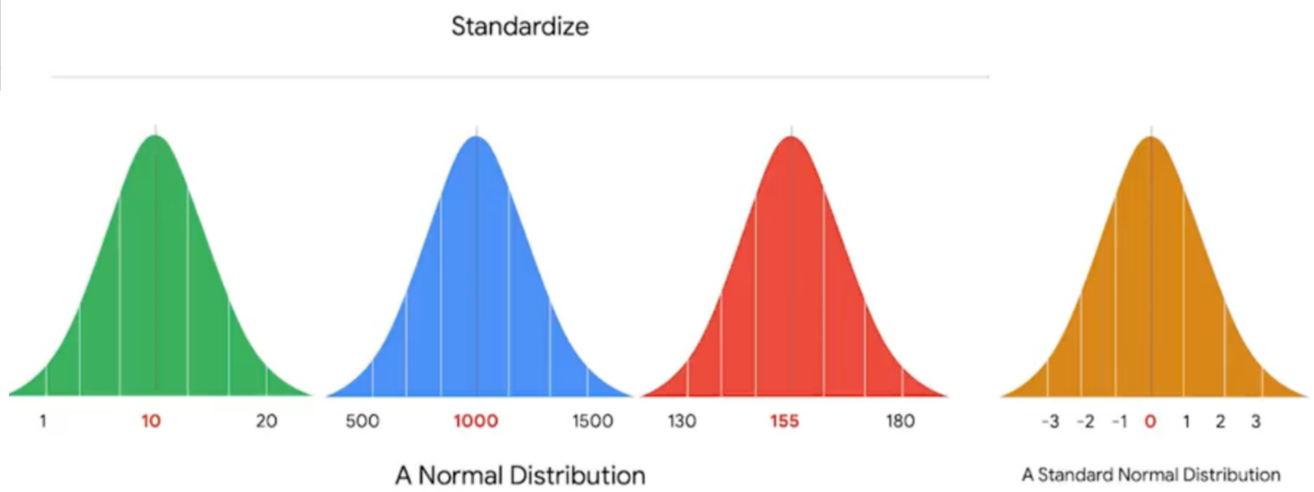
(1) Fraud in financial Transactions

(2) flaws in manufacturing products

(3) Intrusions in Computer Networks.

Standardize





↳ Standardization helps to perform and analyze the distribution statistically efficient & smooth.

score

mean

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

SD

$$z = \frac{(133 - 100)}{15}$$

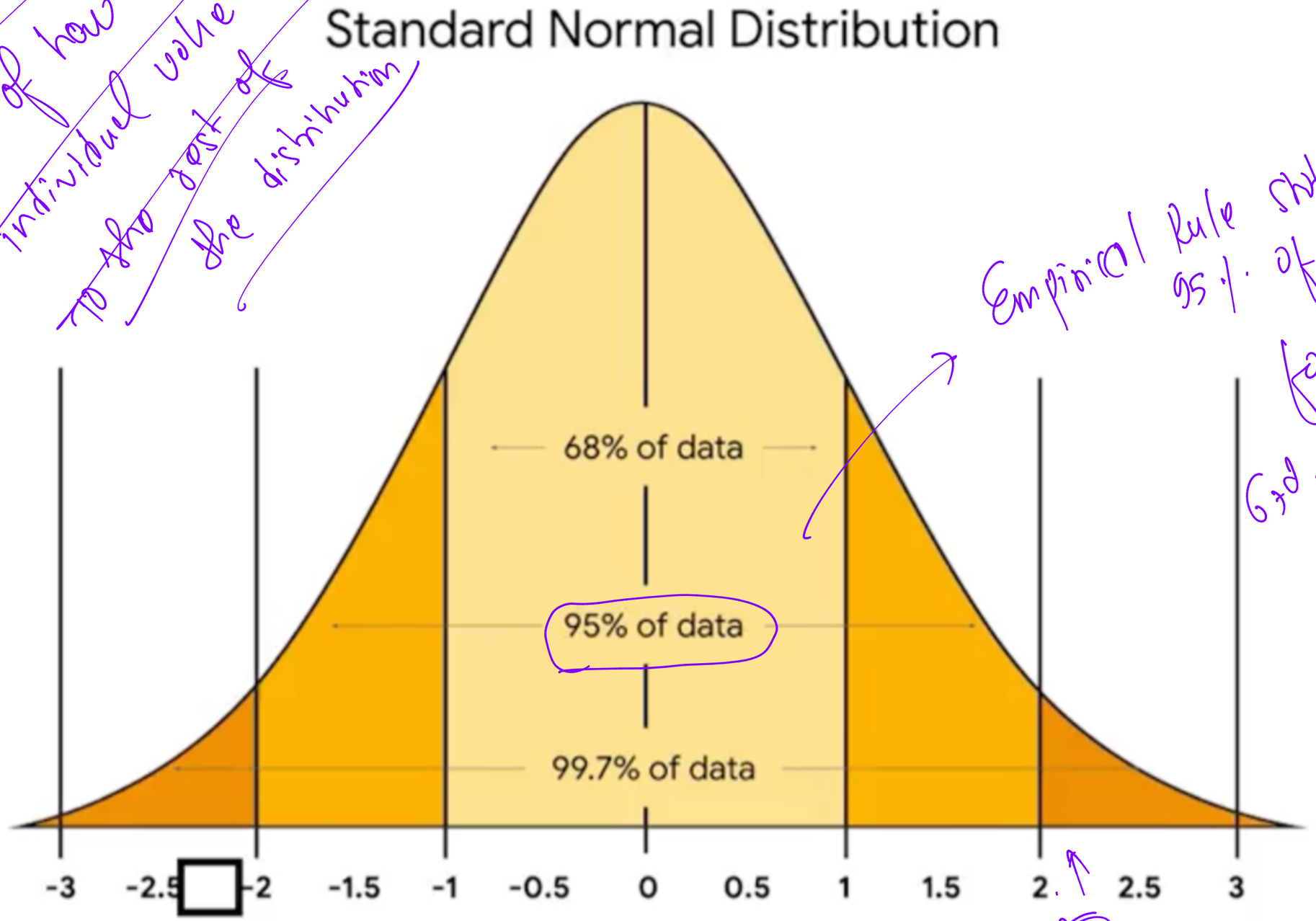
$$z \Rightarrow 2.2$$

2.2 tells us that your test score is 2.2 standard deviations above the mean or average score.

So, the z-score of data value equal to the μ. when z-score $= \mu$.

A z-score is a measure of how many standard deviations below or above the population mean or data point is-

→ Z-score gives idea of how an individual value to the rest of the compares the distribution



→ Empirical Rule states that 95% of the values falls under the 2nd standard deviation

2.2

Example:

Say you score in 85, you want to find out if that's a good score relative to the rest of the class.

Whether or not it's a good score depends on the mean and standard deviation of all exam scores.
Suppose the exam scores are normally distributed with a mean score of 90 and a standard deviation of 4, you can use the formula to calculate the z-score of a raw score of 85.

Your z-score is yours raw score, 85 minus the mean score 90, divided by the standard deviation 4.
This is 85 minus 90 divided by 4 equals -5, divided by 4 equals 1.25. Your z-score of -1.25 tells you that your exam score of 85 is 1.25 standard deviations below the mean or average exam score. Z-scores give you an idea of how individual values compared to the mean.

As a data professional, you'll use z-scores to help you better understand the relationship between specific values in your data set.

Introduction to Sampling:

Types of statistics → Descriptive statistics summarize the main features of a dataset.
→ Inferential statistics use sample data to draw conclusions about a larger population.

#Sampling - process of drawing a subset of data from a population.

Questions Answered by Sampling:

A representative sample accurately reflects the characteristics of a population. If a sample doesn't accurately reflect the characteristics of a population, then the inferences'll be likely be unreliable and inacurate.

Questions answered by sampling

- How many products in an app store do we need to test to feel confident that all the products are secure from malware?
- How do we select a sample of users to run an effective A/B test for an online retail store?
- How do we select a sample of customers of a video streaming service to get reliable feedback on the shows they watch?

This subset is your sample, then you can

↳ Sampling should be drawn randomly or unbiased.

↳ finding the Average Height of People in Nepal we use Sampling?