Statistics

Statistics is like the strong foundation of data analysis. Knowing how to gather, understand, and find essential insights from this information is incredibly important in a world flooded with data. Whether you're a seasoned data scientist or just beginning your journey in data analysis, understanding statistics is essential. This chapter introduces you to the fundamental concepts and techniques of statistics and how they play a crucial role in making sense of data.

In statistics, we have two main branches: descriptive statistics, which helps summarize and present data, and inferential statistics, which aids in drawing conclusions and making predictions based on data.

Descriptive statistics is like the detective that helps us unravel the story hidden in the data. It's all about organizing, summarizing, and presenting data in a way anyone can understand. This branch of statistics gives us tools to analyze data, find patterns, and describe what's happening without making predictions or drawing significant conclusions.

On the other hand, inferential statistics goes beyond describing data. It involves making educated guesses or conclusions about a larger population based on a sample of data. This is where we get into hypothesis testing, confidence intervals, and more advanced statistical methods to make predictions and draw inferences.

In this chapter, we'll look at descriptive statistics and the tools it offers to make data more accessible and insightful.

To dive deeper into statistics, let's begin by exploring a fundamental concept that forms the bedrock of statistical analysis: the distinction between 'population' and 'sample.'

Let's use an election example to understand the concepts of population and sample:

Population: The population refers to the entire group of eligible voters in a specific geographical area, such as a country or a state. This population includes every individual who has the right to vote in that election. For example, in a national election in the United States, the population would consist of all eligible American citizens, which could number in the hundreds of millions.

Sample: A sample is a subset of the population that is selected to conduct surveys, studies, or polls. It is a smaller, manageable group that represents the larger population. In the election example, if you wanted to gauge voters' preferences before the election, you might select a sample of, say, 1,000 registered voters from the entire population. This sample should be chosen in a way that it accurately reflects the diversity and characteristics of the entire population to make meaningful inferences.

We commonly use "N" to represent the entire population and "n" to stand for a sample from that population.

So

Population: The population is the entire group or collection of individuals, objects, or elements that are the subject of study, observation, or analysis. It encompasses all the entities that share a common characteristic, and it is often too large to study or analyze in its entirety.

Sample: A sample is a smaller, carefully chosen subset of a population. It is used to gather data, conduct research, or make inferences about the population as a whole. The process of selecting a sample should be done in a way that minimizes bias and ensures that it accurately represents the characteristics of the larger population. The key to the validity of such inferences is ensuring that the sample is representative of the population so that the findings from the sample can be generalized to the larger group. That’s where the sampling technique comes into the picture.

But first, let’s discuss what is sampling.

Sampling is a fundamental concept in statistics that allows researchers to study a portion of a larger group or population, to draw conclusions about the whole without examining every single

member. It's like taking a bite-sized piece of cake to understand the entire flavor.

Sampling Technique:

A sampling technique is a method used to select a representative subset, or sample, from a larger population to draw valid conclusions about that population. This process involves carefully choosing a subset of individuals, items, or elements from the larger group, to ensure that the sample accurately reflects the characteristics, trends, or properties of the entire population. Sampling techniques are essential in research, as they make data collection more manageable, cost-effective, and efficient while maintaining the integrity and reliability of the findings. By using various sampling methods, researchers can extrapolate meaningful insights and make informed decisions based on a smaller, manageable portion of the whole, without having to examine every single element.

Here are some common sampling techniques

- **Random Sampling**: Imagine you have a big bowl of colorful candies. To do a random sample, you close your eyes, stir the candies, and pick a few without looking. This way, every candy has an equal chance of being chosen. Random sampling helps get a fair and unbiased selection from a group.

- **Stratified Sampling**: In stratified sampling, the population is divided into subgroups or strata based on certain characteristics (e.g., age, gender, income). Then, random samples are drawn from each stratum proportionate to its size. This technique ensures that each subgroup is represented in the sample, making it useful to study specific subpopulations within a larger population. For example, suppose you have a box of crayons in different colors. In stratified sampling, you decide how many crayons you want from each color and randomly choose that number from each group. This method ensures you get a good mix of all colors in your sample.

- **Systematic Sampling**: Here, you select every nth member from a list of the population. For example, if you have a list of 1,000 individuals and want a sample of 100, you could select every 10th person. The key is to choose a random starting point to avoid periodic patterns that could introduce bias. For example, think of a long line of people waiting to get ice cream. Instead of picking randomly, you decide to choose every 5th person in line. This is systematic sampling, which is an organized way to select samples at regular intervals.

- **Cluster Sampling**: Cluster sampling involves dividing the population into clusters or groups, often based on geographic location. Then, you randomly select some clusters and sample all the individuals within those clusters. This approach is useful when the population is dispersed and hard to reach individually. For example in a big city, researchers want to study traffic patterns. They divide the city into 10 neighborhoods (clusters). They randomly select 3 neighborhoods (clusters) and observe traffic within those chosen clusters. The data from these 3 clusters represent their sample, simplifying the data collection process compared to studying all neighborhoods individually.

In cluster sampling, the clusters are typically chosen randomly, ensuring that every cluster has an equal chance of being included.

- **Convenience Sampling**: Convenience sampling involves selecting individuals who are easy to access or readily available. While this method is quick and cost-effective, it can introduce bias because it doesn't ensure that the sample represents the population accurately. For example, sometimes, you might pick the easiest or most convenient subjects to study. For example, if you ask your friends about their favorite movies, that's convenience sampling. It's quick and easy but may not represent the whole population.

- **Snowball Sampling**: Snowball sampling is commonly used in studies where the population is hard to identify or reach. It starts with one or a few initial participants who are asked to refer other potential participants. For example, imagine you want to find people with a rare hobby. You ask one person you know who has that hobby, and then you ask them to introduce you to others with the same interest. It's like a snowball rolling and getting bigger as it goes.

These sampling techniques help researchers collect information from a group of items or people in different ways, depending on their research goals and available resources. It's essential to select the most appropriate technique to ensure that the sample accurately represents the larger population, enabling meaningful and valid inferences to be drawn from the collected data.

These sampling techniques also play a crucial role in the process of collecting information during the train-test split phase in research. Train-test split involves dividing our dataset into two parts: the training set, used to train and build our model, and the testing set, used to evaluate its performance.

Now, imagine you have a large group of items or people and want to ensure that the data you use to train your model accurately

represents the entire population. This is where different sampling techniques come in like Random Split, Stratified Split, and others.

By employing these sampling techniques during the train-test split, researchers can create balanced and representative datasets, enhancing the reliability of their models and the validity of their findings.

Variables and their types:

A variable is like a container that holds information or data. It can represent different things, such as numbers, names, or categories. In the world of statistics and data analysis, variables are the building blocks used to understand and describe the characteristics of things we're studying. They help us organize, compare, and make sense of information. For example, a variable could be someone's age, a product's price, or a city's name. Variables are the essential tools that help us answer questions and draw conclusions from data.

Imagine you want to study the ages of students in a classroom. Each student's age is a variable because it can take on different values. For instance, you might have students with ages 7, 8, 9, 10, and so on. In this case, "age" is the variable, and the values it can take are the specific ages of the students in the classroom.

Variables can be broadly categorized into two main types: qualitative and quantitative. Now, let's delve deeper into the two main types of variables: qualitative and quantitative.

Qualitative variablesare also known as**categorical variables**. They represent categories or labels that don't have a numerical value.

Consider a survey about car colors. The variable "Car Color" is qualitative because it classifies cars into categories like "Red," "Blue," "Green," and so on. These categories are labels, not numerical values.

Qualitative variables can be further divided into two types: nominal and ordinal.

- **Nominal Variables**: Nominal variables represent categories with no inherent order or ranking. For example, "Eye Color" (Blue, Brown, Green) is a nominal variable because eye colors have no natural order.
- **Ordinal Variables**: Ordinal variables represent categories with a specific order or ranking. For instance, "Educational Level" (High School, College, Graduate) is an ordinal variable because there's a clear hierarchy in educational attainment.

Quantitative variables, on the other hand, are numerical in nature. These variables represent measurable quantities.

Let's go back to our example of "Age." Age is a quantitative variable measured in years and can take on a wide range of numerical values.

Quantitative variables can be further divided into two types: Discrete and Continuous

- **Discrete Variables**: Discrete variables can only take specific, distinct values, usually in whole numbers. For example, "Number of Children in a Family" is a discrete variable because you can have 0, 1, 2, 3, etc., but you can't have 2.5 children.
- **Continuous Variables**: Continuous variables can take any value within a range and can have decimal points. "Height" is an example of a continuous variable because it can be measured with great precision and can take values like 160.5 cm or 170.2 cm.

Understanding the type of variable you're dealing with is crucial because it determines the appropriate statistical methods and tools you should use for analysis.

Independent and dependent features:

Let’s understand the Concept of**independent and dependent features**with an example of predicting house prices.

Think of independent features as the things that influence or affect something else. In the case of predicting house prices, independent features are characteristics of the house that might affect its price. These can include the number of bedrooms, the size of the house, the neighborhood it's in, and so on.

On the other hand, the dependent feature is what you're trying to predict or understand based on the independent features. In our house price example, the dependent feature is the actual price of the house.

So, to predict the house price, you use the independent features (like the number of bedrooms, size of the house, and distance to the nearest school) to make an educated guess about the dependent feature (the price of the house).

In a nutshell:

Independent features are the factors you think might influence something.

Dependent features are what you're trying to predict or understand based on those factors.

Central Tendencies:

A critical part of descriptive statistics is central tendencies. Central tendencies tell us where the "center" of our data is. It helps us understand where most of the data points cluster or concentrate.

The three measures of central tendency are mean, median, and mode.

Imagine you have many friends, and you want to know how tall they are. You measure their heights and get the following data (in inches): 60, 62, 64, 66, 68, 70, 72.

Now, let's calculate the three central tendencies for this data:

- Mean: The mean is the average of all the numbers. To find it, you add up all the numbers and then divide by how many numbers there are.

In the above given example Add up all the heights and divide by the number of friends.

(60 + 62 + 64 + 66 + 68 + 70 + 72) / 7 = 462 / 7 ≈ 66.

So, the mean height of your friends is approximately 66.

Now there are two terms known as Population Mean and Sample Mean.

Population Mean (μ): The population mean represents the average of all values in an entire population, and it is denoted as "μ."

Sample Mean ($\bar{x}$): The sample mean represents the average of values in a subset (sample) of a population, and it is denoted as "$\bar{x}$."

- Median: The median is another vital measure of central tendency.

The median is a measure of central tendency that represents the middle value of a dataset when the data is ordered from smallest to largest. If there is an even number of values, it is the average of the two middle values.

Let’s calculate it with this example: data : 60, 62, 64, 66, 68, 70, 72.

Median for Odd Values:

First, you need to order the data from smallest to largest:

60, 62, 64, 66, 68, 70, 72.

Since there is an odd number of values (7 in this case), the median is the middle value. In this dataset, the middle value is 66, as it has three values below it and three values above it.

So, for this odd dataset, the median is 66.

Median for Even Values:

If you have an even number of data points, finding the median is slightly different. Let's use a modified dataset as an example: 60, 62, 64, 66, 68, 70.

First, order the data from smallest to largest:

60, 62, 64, 66, 68, 70.

Calculate the median by taking the average of the two middle values. In this dataset, the two middle values are 64 and 66.

Median = (64 + 66) / 2

Median = 130 / 2

Median = 65

So, for this even dataset, the median is 65. The median is the average of the two middle values when you have an even number of data points.

- Mode: The mode is the value that appears most frequently. Let’s understand this with an example:

First, identify the value(s) with the highest frequency. In this case, the value "70" appears twice, which is more frequent than any other value. So, the mode is 70.

Let's move on to unimodal, bimodal, and multimodal.

When a set of data has only one mode, it is called unimodal. In other words, there is one number that appears more often than any other.

Consider the set of numbers 10, 15, 20, 25, 20, 30, 35. The number 20 appears twice, while the other numbers only appear once. Therefore, this set is unimodal with a mode of 20.

If a set of data has two modes, it is called bimodal. This means there are two numbers that appear with the highest frequency.

Let's take the set of numbers 5, 10, 5, 15, 20, 15, 25. Both 5 and 15 appear twice, more often than any other number. Hence, this set is bimodal with modes 5 and 15.

When a set of data has more than two modes, it is referred to as multimodal. This means there are multiple numbers that are repeated with the highest frequency.

Imagine a set of numbers 5, 10, 5, 15, 20, 15, 25, 20, 25. Here, 5, 15, 20, and 25 all appear twice, making the set multimodal with modes 5, 15, 20, and 25.

So, in summary, mode helps us identify the most frequent value, and we can describe datasets as unimodal, bimodal, or multimodal based on the number of modes they have.

Central tendency measures are like tools in statistics that help us understand the main or usual value in a set of data. Which tool we use depends on the kind of data and what we want to find out.

Now, imagine you have a bunch of numbers, and some are missing or not available (we call them "nan" values). If you want to fill in these gaps, the choice of which central tendency measure to use depends on your specific situation.

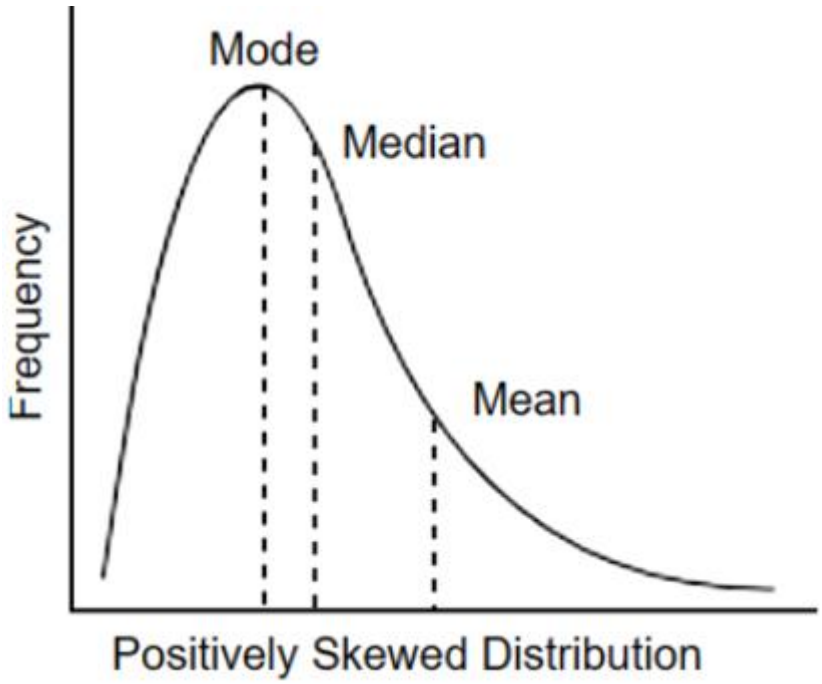
For instance, if you have some missing values in a set of test scores, and you want to fill them in with a representative value, you might choose the mean (average). This is helpful because using the mean

ensures that the missing values won't make the overall result 'lean' too much in one direction or the other. It helps to keep things balanced and fair in understanding the overall performance.

It's worth noting that for numerical data that can be measured or counted, both the mean and median are applicable. However, it's crucial to consider the characteristics of the distribution and whether there are outliers present. This careful consideration ensures a more accurate representation of the data's central tendency.

If your data is roughly symmetrically distributed without extreme values or outliers, the mean is often a good choice as it considers all values in the dataset.

If your data is skewed or contains outliers, the median might be a more robust choice. It is less affected by extreme values and provides a better representation of the central location in such cases.



Suppose we have the following dataset, which is roughly symmetrically distributed without extreme values or outliers:

4,7,8,10,12.

So the Mean will be (4+7+8+10+12)/5=8.2

Since the dataset is sorted, the median is the middle value, which is 8 in this case.

In symmetrically distributed data, the mean (8.2) and median (8) are close, indicating that the mean is a reasonable choice for representing the central location.

Now, let's introduce an outlier to the dataset:

4,7,8,10,12,100

So the updated Mean will be (4+7+8+10+12+100)/5=23.5

Since the dataset is sorted, the median is still the middle value, which is now 9.

With the addition of the extreme value (100), the mean has shifted even more dramatically to 23.5, indicating how sensitive it is to extreme values. On the other hand, the median remains relatively unaffected at 9.

This example highlights that in the presence of outliers, the mean can be heavily influenced, making it less representative of the central location. In such cases, the median is a more robust choice, providing a better representation of the typical value in the dataset.

For categorical data, the mode is typically used. It represents the most frequently occurring category or value.

Let's consider an example to understand this:

Student Name	Favorite color
John	red
ben	blue
jerry	green
peter	
jean	blue
David	yellow

Here, we have a dataset containing data about students' names and their favorite colors. However, there's a missing value for Peter's favorite color. When we encounter missing data like this, it can cause issues when analyzing the dataset. One common approach to handle missing values is to replace missing values with estimated ones based on the rest of the data.

In this case, we can use the mode to fill the missing value for Peter's favorite color. The mode is simply the most frequent value in the dataset, and it's a good choice for categorical data like favorite

colors. Looking at the dataset, Blue appears twice, and Red, Green, and Yellow each appear once. Since Blue is the most frequent color, we'll use it as the mode. So, we'll replace the missing value for Peter's favorite color with Blue. This way, we maintain the integrity of the dataset and ensure that it's ready for analysis without losing valuable information.

But there is no one-size-fits-all answer, and the choice depends on the specific characteristics of your data and the goals of your analysis. It's often helpful to consider multiple measures of central tendency and other descriptive statistics to comprehensively understand the data distribution.

Let's take another example to understand why it is important to not rely only on one measure of central tendency:

Let's consider two schools A and B each with 5 students.

The marks of the 5 students of school A is: 70, 70, 70, 70, 70

The marks of the 5 students of school B is: 60, 60, 70, 75, 85

The mean of School A is 70 and the mean of School B is also 70. In such a case, if we only rely on the mean of the data we may be misled and have incorrect insights. so we must also look for other central tendencies such as median and mode.

In this case, the median of School A is 70 and the mode of School A is 70 whereas the median of School B is 70 and the mode of School B is also 60.

We should not only be limited to the central tendencies of a dataset because at times that might also be deceiving. So to understand our data more precisely we must also check the measure of dispersion.

Measures of dispersion:

But first, consider having the exam scores of a class, and the goal is to determine the average or most common score. So in such a scenario, one would employ measures like mean, median, or mode to identify this central point.

Now let's say we have two classes with the same average exam score. Now, we want to see how the individual scores vary around this average.

For this, we use measures of dispersion like the standard deviation. If Class A and Class B both have the same mean but different standard deviations, it tells us that even though their averages are similar, the scores in one class might be more tightly clustered around the average, while the other class may have more scattered or spread-out scores.

So, while central tendency helps us find the midpoint or typical value, measures of dispersion give us insight into how much the scores deviate from this midpoint in different classes or groups.

Therefore, in statistics, "Measures of Dispersion" are like detectives that help us understand how spread out or scattered the numbers are in a data group. They show us if the data points are huddled close together or are all over the place.

Imagine you're playing a game where you have to shoot darts at a target. If most of your darts land very close to the center, you have low dispersion. But if your darts are all over

the place, you have high dispersion.

Measures of dispersion help us figure out if our data is tightly clustered or widely spread, and they come in different types like the range, variance, and standard deviation.

They provide valuable insights into the degree to which data points deviate from the central tendency. These measures give us insights into the variability of our data, making it easier to draw conclusions and make decisions.

Let's delve into three key measures of dispersion: variance, standard deviation, and range.

Variance: Variance measures the average squared difference between each data point and the mean (average) of the dataset. It

tells us how much each data point varies from the mean. A high variance indicates that the data points are spread out over a large range.

So the mathematical formula to calculate the variance is:

Formula for Variance:

$$\text{Variance} = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

where:

- N is the number of data points.
- X_i represents each data point.
- $\bar{X}$ is the mean of the dataset.

Example:

Let's say we have a dataset of daily temperatures in Celsius for a week: [25, 28, 26, 22, 30, 27, 24]. First, calculate the mean ($\bar{X}$):

$$\bar{X} = \frac{25+28+26+22+30+27+24}{7} = 26.43$$

Next, calculate the variance:

$$\text{Variance} = \frac{1}{7}\left[(25 - 26.43)^2 + (28 - 26.43)^2 + (26 - 26.43)^2 + (22 - 26.43)^2 + (30 - 26.4\right.$$

$$\text{Variance} \approx 6.18$$

Next, we have is

Standard Deviation: The standard deviation is the square root of the variance. It provides a more interpretable measure of how data points deviate from the mean, it is expressed in the same units as the data, whereas variance is in squared units. This makes standard deviation more interpretable because it aligns with the scale of your data. For example, if you're measuring the heights of trees in meters, the standard deviation will also be in meters, which is easier to relate to.

When you calculate the standard deviation, you don't need to worry about squared values.

The formula for Standard Deviation:

$$\text{Standard Deviation} = \sqrt{\text{Variance}}$$

The notation for standard deviation is "σ" (sigma) for a population and "s" for a sample.

The notation for variance is "σ²" (sigma squared) for a population and "s²" for a sample.

Using the previous example, the standard deviation is the square root of the variance:

$$\text{Standard Deviation} \approx \sqrt{6.18} \approx 2.48$$

So a low standard deviation indicates that the data points are close to the mean or Data points are tightly packed together, as your friends huddled in one corner of the party, while a high standard deviation indicates that the data points are spread out over a larger range or Data points are more scattered or dispersed like your friends spread out in different rooms at the party. And Dispersion is another word for this "spread".

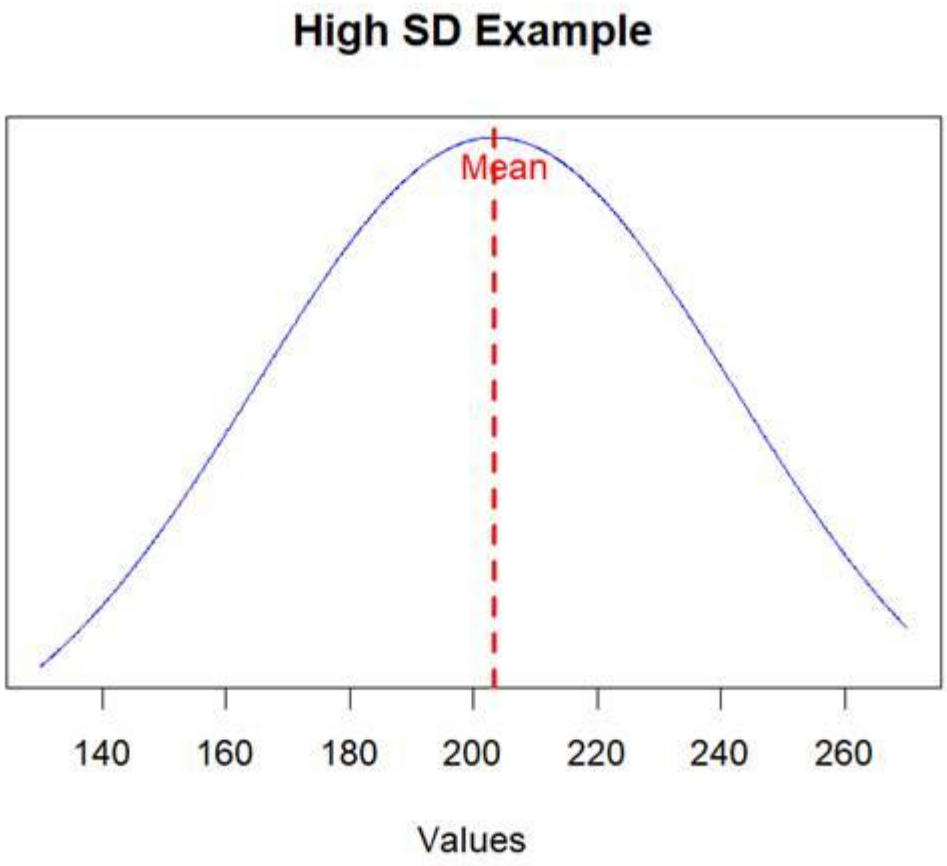
If the value of the standard deviation for a dataset is high, it indicates that the data points are widely spread out from the mean. In other words, the individual data points are highly variable or dispersed around the average.

Whereas if the value of the standard deviation for a dataset is low, it indicates that the data points are closely clustered around the mean, suggesting that there is minimal variability or dispersion in the dataset. The values are relatively consistent and close to each other.

Let's assume a dataset with a high standard deviation value, say set A:

A = 230, 140, 260, 150, 240, 170, 250, 180, 220, 200, 210, 190

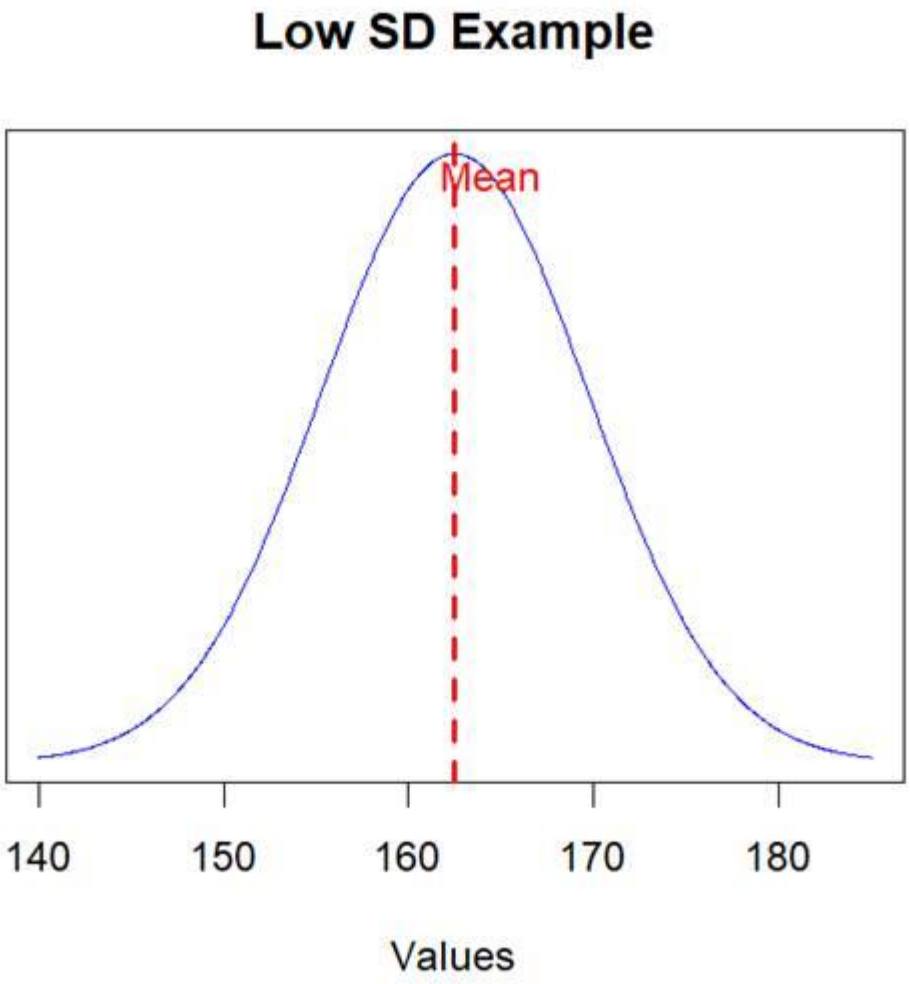
The standard deviation value for the data is: 38.69



Now, Let's take an example of a dataset with a low SD value, say set B:

B = 150, 155, 160, 165, 165, 170, 170, 175, 165, 160, 160, 155

The standard deviation value for the dataset is 7.22



On comparing both graphs we can see that the dataset having a lower standard deviation value has more data points clustered near the mean hence the bell curve is steep(pointy). Whereas for the dataset with a high standard deviation value, the data points are more distant from the mean hence the bell curve is wider.

So, Variance Tells us the overall spread of the data. Standard deviation tells us the average distance of each element from the mean.

Range: The range simply measures the difference between the maximum and minimum values in the dataset. It provides a quick way to understand how spread out the data is.

Formula for Range:

Range = Maximum Value − Minimum Value

In summary, these measures of dispersion - variance, standard deviation, and range - help us understand how much the data points vary from the central tendency (mean) in a dataset. Variance and standard deviation provide more detailed information about the spread, while the range quickly tells you how much the data varies overall.

Now let's understand a little about Normal distribution and empirical rule.

Normal Distribution and Empirical Rule

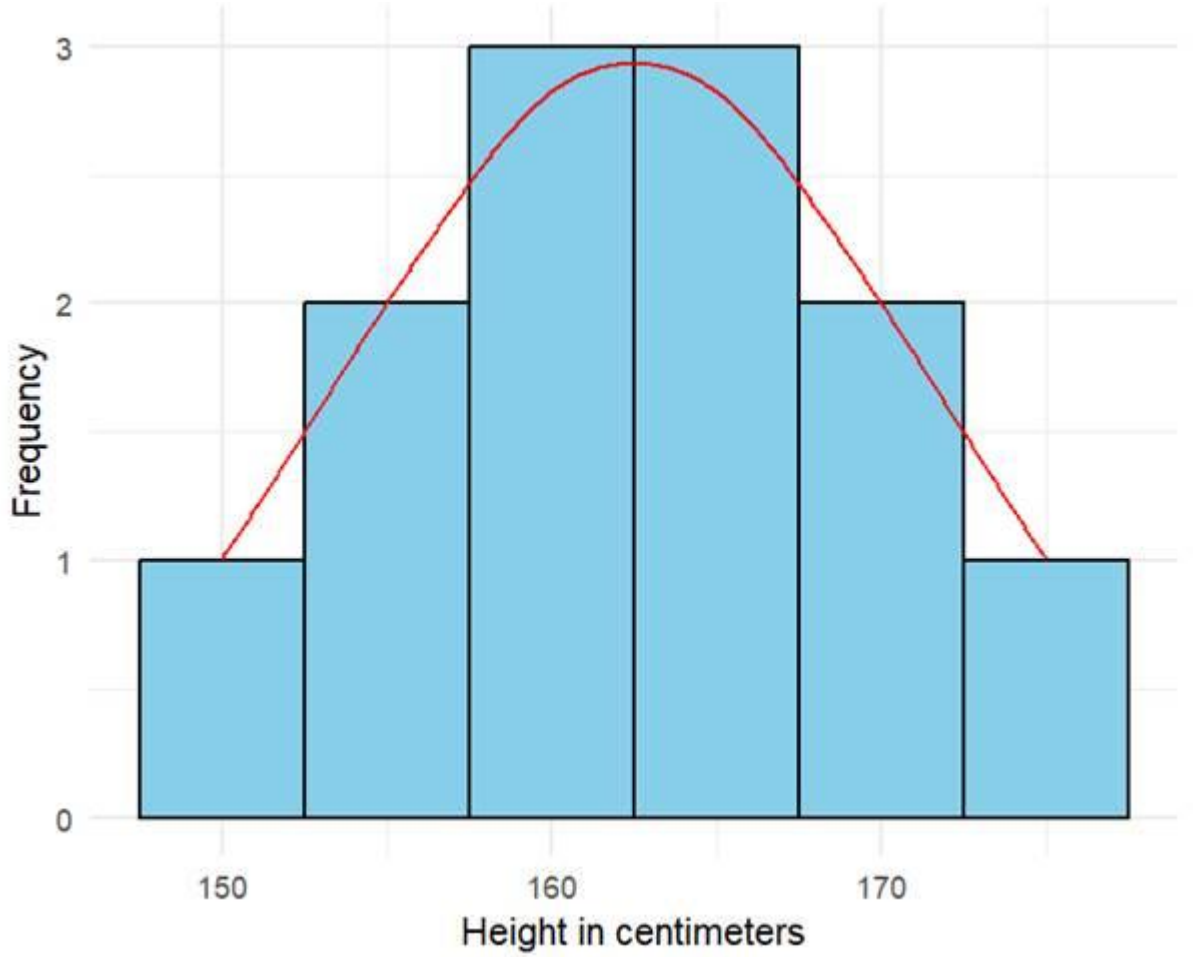
The normal distribution, also known as the bell curve or Gaussian distribution, is a way to describe how data is spread out in many natural and real-world situations. It looks like a bell, and this shape is why people call it the bell curve. The normal distribution is handy when we're dealing with lots of random factors that come together to affect an outcome.

Let's look at an example to understand it better:

Consider a set of 12 observations containing the heights of 12 students (in centimeters)

150,155,160,165,165,170,170,175,165,160,160,155

On plotting the frequency distribution of the above dataset we get a distribution looking like this:



This is called a normal distribution or the bell curve. At the center of the bell curve lies the highest point, called the peak, where the most common or average value occurs. As you move away from the center in either direction, the frequency of values gradually decreases, creating a symmetrical pattern. This means that values near the peak are more common, while those further away occur less frequently.

Now, let's talk about the empirical rule, which is closely related to the normal distribution. This rule tells us that in a normal distribution:

About 68% of the data points fall within one standard deviation range from the mean (the average). The formula for one standard deviation is:

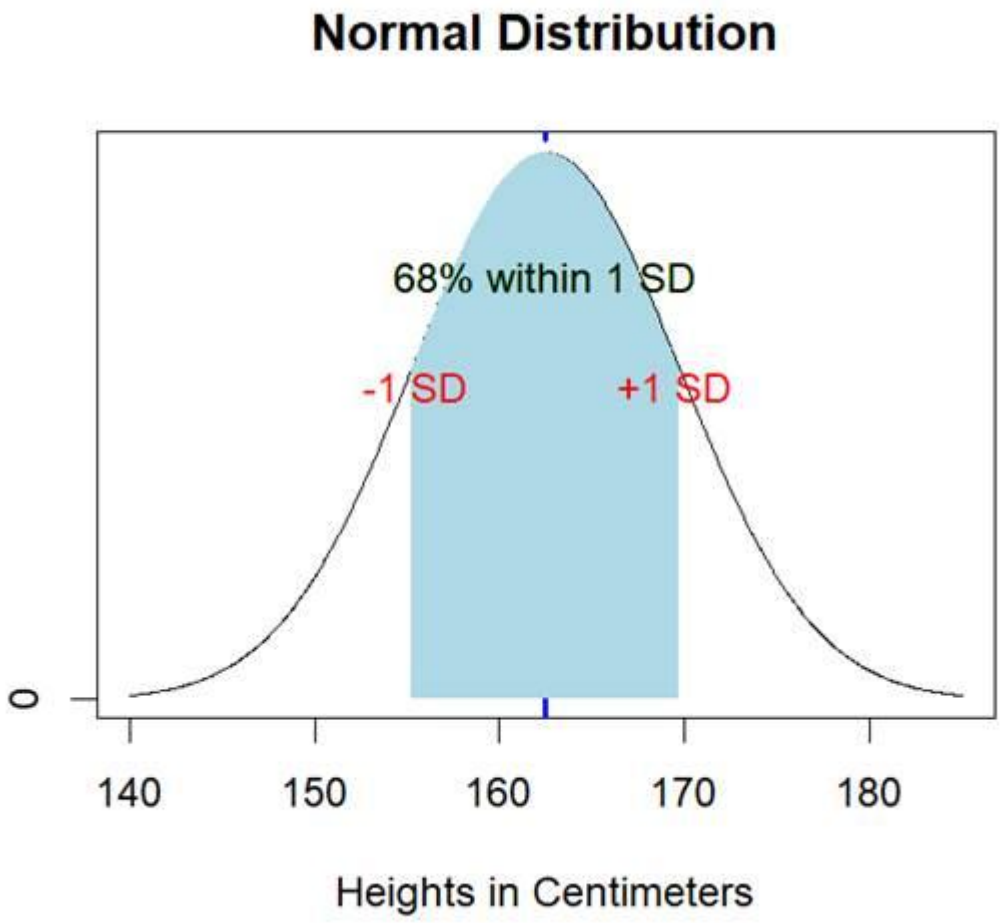
Lower Limit = Mean - Standard Deviation

Upper Limit = Mean + Standard Deviation

The 1 standard deviation for our dataset can be calculated by first calculating the standard deviation value, which is 7.22. So the 1 standard deviation range for our dataset is:

Lower limit = mean - 7.22 = 162.5 - 7.22 = 155.28

upper limit = mean + 7.22 = 162.5 + 7.22 = 169.72



And 68% of the data points fall under this range.

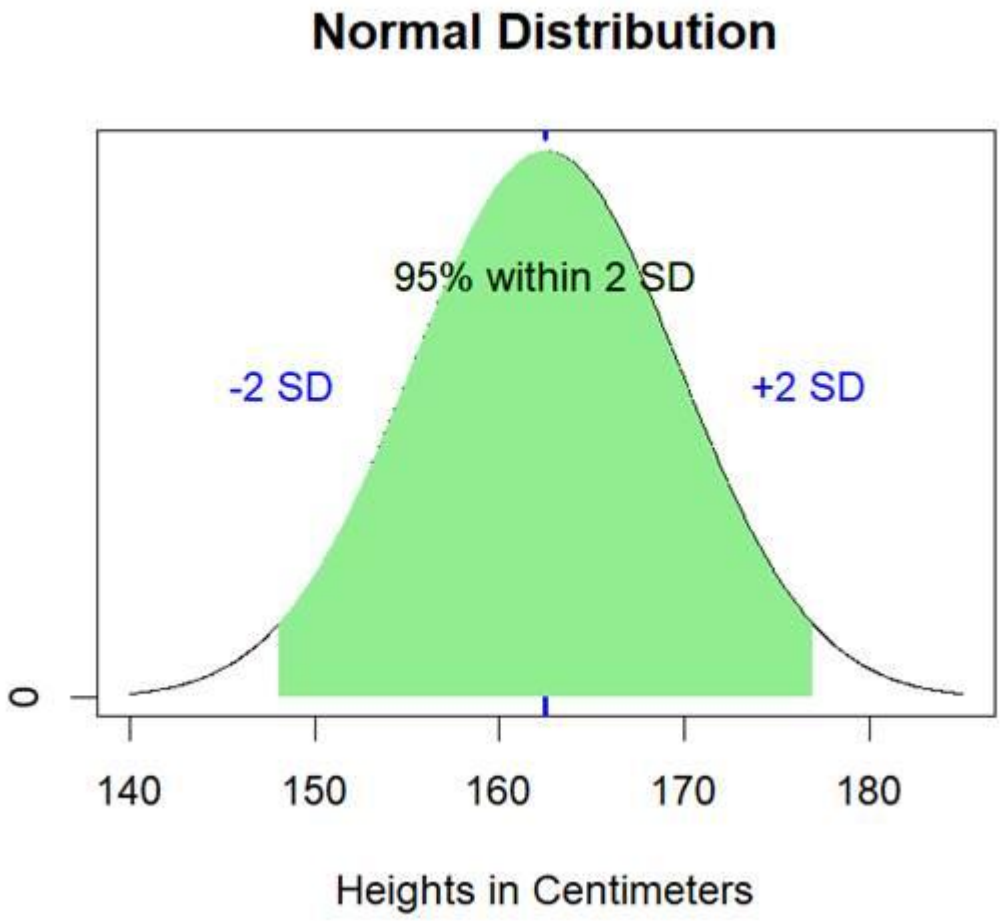
Around 95% falls within two standard deviations range. The formula for the two standard deviations range is:

Lower Limit = Mean - 2 times Standard Deviation

Upper Limit = Mean + 2 times Standard Deviation

Lower limit = mean - 7.22 = 162.5 - (2* 7.22) = 148.06

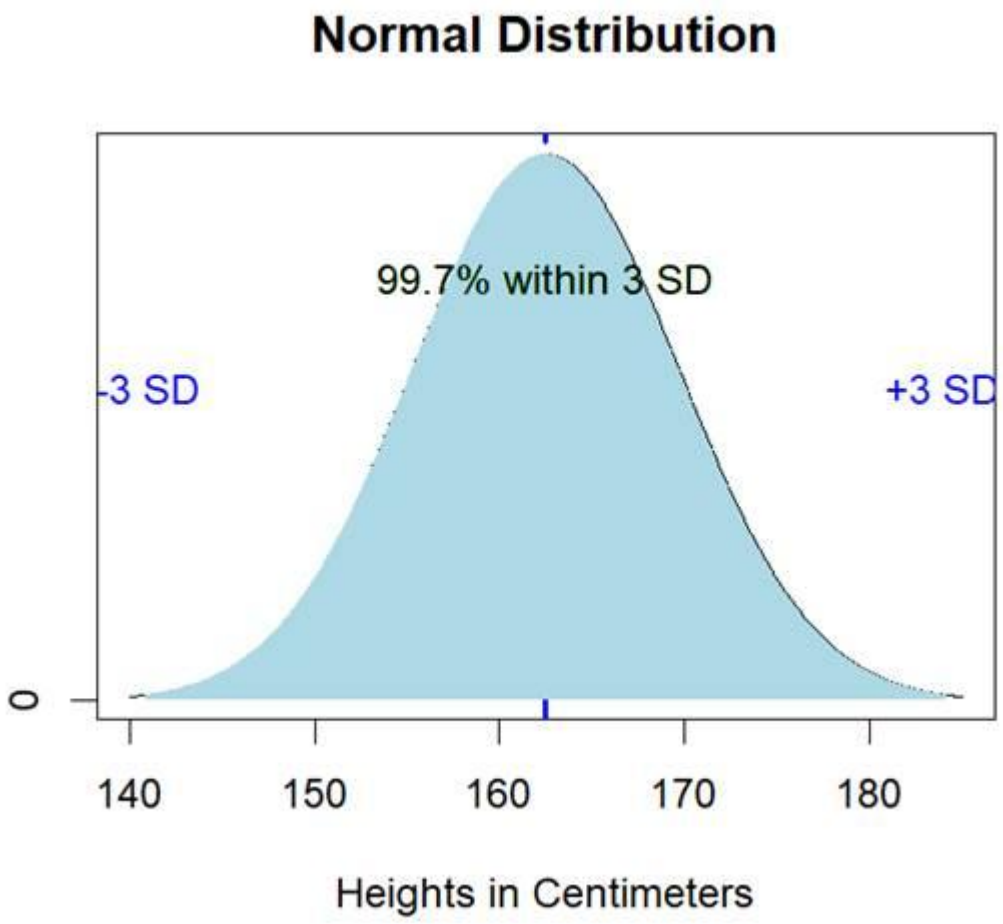
upper limit = mean + 7.22 = 162.5 + (2* 7.22) = 176.94



Almost all, about 99.7%, fall within three standard deviations range. The formula for three standard deviations is:

Lower limit = mean - 3* 7.22 = 162.5 - (3* 7.22) = 140.84

upper limit = mean + 3 *7.22 = 162.5 + (37.22) = 184.16



Z-score:

Alright, now that we've covered central tendencies and measures of dispersion, let's jump into another important concept the Z-score. But first, let’s see an example.

Suppose you have two exam scores from different classes: Student A scored 75 on a math test in Class X, and Student B scored 85 on a math test in Class Y. At first glance, it seems like Student B performed better because they scored higher. But that wouldn't be a fair comparison without considering the context as they come from different classes with potentially different difficulty levels and score distributions. This is where z-scores come in handy!

By converting each score to a z-score, we can standardize the scores and compare them on a common scale, regardless of the original distribution of the scores.

It involves transforming a dataset so that it has a mean of 0 and a standard deviation of 1. The formula for calculating the z-score for a data point X in a distribution with mean μ and standard deviation σ is:

$Z=(X- \mu)/ \sigma$

And here's a general guide to interpreting z-scores:

Z-Score = 0: This means the data point's value is exactly equal to the mean of the dataset. The data point is at the center of the distribution.

Z-Score > 0: The data point is above the mean. The larger the z-score, the farther above the mean the data point is. This indicates an above-average performance or value.

Z-Score < 0: The data point is below the mean. The smaller the z-score, the farther below the mean the data point is. This suggests a below-average performance or value.

Magnitude of Z-Score: The absolute value of the z-score represents how many standard deviations away from the mean a data point is. For example, a z-score of +2 means the data point is 2 standard deviations above the mean, and a z-score of -3 means the data point is 3 standard deviations below the mean.

In short, the sign of the z-score (+ or -) indicates the direction relative to the mean (above or below), and the magnitude of the z-score indicates how many standard deviations away from the mean the data point is.

Now, let's calculate the z-scores for Student A and Student B:

Assuming the mean score in Class X is 70 and the standard deviation is 5: So Z score for A will be:

ZA=75-70/5=1

$$Z_A = \frac{75 - 70}{5} = 1$$

Now Assuming the mean score in Class Y is 80 and the standard deviation is 8: So Z score for B will be:

ZA=85-80/8=0.625

$$Z_A = \frac{85 - 80}{8} = 0.625$$

Now, you can see that Student A's z-score is 1, indicating that their score is 1 standard deviation above the mean in Class X. Student B's z-score is 0.625, indicating that their the score is 0.625 standard deviations above the mean in Class Y.

Since both z-scores are positive, it means both students performed above the mean in their respective classes. However, comparing the magnitudes of the z-scores:

We can see that A's score is 1 standard deviation above the average in their class, and B's is 0.625 standard deviations above their class average.

So, even though B got a higher score, the z-score helps us see that A's performance is relatively better compared to their classmates.

The z-score provides a common scale for comparing data points, which is particularly useful when dealing with diverse datasets with different means and standard deviations. It facilitates a more meaningful interpretation of data by expressing the relative position of a data point within its distribution.

Standardization helps us bring everything to a level playing field for a fair comparison. It is a statistical technique used to transform data in a way that the data has a standard or common scale. This process is essential when working with different units of measurement or when comparing variables with different ranges. The purpose of standardization is to make the data more interpretable and comparable.

Next, we have**Percentiles and Quartiles**

Percentiles and Quartiles:

In the exciting world of statistics, you'll often come across various terms and concepts that might sound similar but have distinct meanings. One such pair of terms that can sometimes cause confusion is "Percentage" and "Percentile."Let’s first understand what Percentage is.

Percentage is a simple way to express a part of something as a fraction of 100. It's a handy tool for understanding proportions and

making comparisons. When you hear someone say, "60% of the students like pizza," they are telling you that out of every 100 students, 60 prefer pizza.

The formula to calculate a percentage is:

- Percentage (%) = (Part / Whole) * 100

Where Part: is the portion of the whole that you want to express as a percentage.

And "Total" is the entire amount or quantity that the "Part" is a part of. It's the whole thing you're considering.

Let’s understand this with an easy example:

Picture a jar with 20 marbles inside. You're curious about how many of those marbles are blue. You count 5 blue marbles.

To find the percentage of blue marbles in the jar, we'll use the formula:

Percentage (%) = (Number of Blue Marbles / Total Number of Marbles) * 100

Now, let's calculate it:

Percentage of blue marbles = (5 / 20) * 100

Percentage of blue marbles = (1/4) * 100

Percentage of blue marbles = 25%

So, 25% of the marbles in the jar are blue.

Now let’s understand what percentile is:

Percentile:

You've probably heard the term "percentile" when discussing standardized tests like the GATE, SAT, or CAT exams.

A percentile is a way to figure out what percentage of people scored less than you on the same test. It helps you see where you stand among the group and gives you a broader understanding of how well you did. In essence, a percentile is a statistical measure that indicates the percentage of people who scored lower than you on a

particular exam or assessment. It's a useful way to gauge your relative standing and comprehend your performance in a broader context.

A percentile is a statistical measure that helps you understand the relative position of a specific data point within a dataset, so your test score represents the data point, while the scores of other students make up the dataset. This concept allows you to assess how your performance compares to that of your peers.

It tells you what percentage of the data falls below or above that data point. We often use percentiles to get a sense of the distribution of data and identify values that are higher or lower than most of the others.

Percentiles are often used in various fields, including education, healthcare, and finance, to compare and analyze data.

Percentiles are often expressed as a percentage.

In simpler terms, a percentile answers the question: "What percentage of the data falls below this particular value?"

They are often expressed as a percentage.

For example, the 95th percentile indicates that:

If someone's performance is at the 95th percentile, it means they have scored better than 95% of the students or individuals in the group. So, 5% of the students or individuals have scored higher than the person at the 95th percentile. In other words, the 95th percentile represents the point below which 95% of the data falls, and there is 5% of the data that is higher than the 95th percentile.

So a percentile is a value below which a certain percentage of observation lies. To illustrate this:

Imagine we have a dataset of 8 test scores: 60, 70, 75, 80, 85, 90, 95, and 100.

Now, let's say you got a score of 85 on the test. What is the percentile rank of your score?

Percentile Rank:

Before we find the percentile rank, let's understand what it means. The percentile rank tells you how your score compares to the scores of others in the dataset. In simple terms, it answers the question: "What percentages of scores in the dataset are equal to or lower than your score?"

Now, let's get into the formula for finding percentile rank. To calculate the percentile rank of a particular score in a dataset, you can use this formula:

Percentile Rank = (Number of Values Less Than or Equal to Given Value/ Tot. Number of Values) x 100

To find the percentile rank of your score (85) in the given dataset of 8 test scores [60, 70, 75, 80, 85, 90, 95, 100], you can follow these steps:

- First, sort the dataset in ascending order.
- Count the number of values in the dataset that are less than or equal to your score (85).
- Calculate the percentile rank using the formula:

Percentile Rank = (Number of Values Less Than or Equal to Given Value/ Tot. Number of Values) x 100

Let's calculate it step by step:

- Sorting the dataset:

Sorted Dataset: [60, 70, 75, 80, 85, 90, 95, 100]

- Count the number of values less than or equal to 85:

There are 5 values (60, 70, 75, 80, 85) that are less than or equal to 85.

Calculate the percentile rank:

Percentile Rank=(5/8)*100= 62.5%

The percentile rank of your score (85) being 62.5% means that you scored as well as or better than approximately 62.5% of the other people who took the same test. In simpler terms, if you imagine all the test scores lined up in order from lowest to highest, your score falls at a point where you've done better than about 62.5% of the

people. So, it's a way to understand how your score compares to everyone else's.

Now you might wonder what to do when the question is reversed, and you're interested in discovering the specific value that corresponds to a percentile within your dataset.

The formula to calculate the specific value that corresponds to a percentile within a the dataset is as follows:

- Position = (Percentile / 100) * (N + 1)

Position: This is the specific data value you want to find, which corresponds to the desired percentile within the dataset.

Percentile: This represents the desired percentile expressed as a percentage.

For example, if you want to find the 25th percentile, you would use Percentile = 25. It signifies the location within the dataset for which you want to find the value. N: N is the total number of data points in the dataset. It represents the size or count of the dataset.

Let's explain what happens when the calculated Value of Position above is a whole number or a decimal:

Whole Number Value:

When the calculated value is a whole number, it means that the desired percentile corresponds to a specific data point within your dataset.

Imagine if the calculated value is, for instance, 7 for the 25th percentile, it means the 25th percentile falls precisely at the 7th data point. In this case, you can directly pick the data point at the 7th position as your 25th percentile value. So, if your data is sorted, the 7th data point is your 25th percentile value.

Decimal Value:

When the calculated value is decimal, it means the desired percentile doesn't point to an exact data point but falls between two data points.

Think of it like a position between two data points. For example, if you calculate a value of 7.5 for the 25th percentile, it means the 25th percentile is halfway between the 7th and 8th data points.

Now let’s understand with an example:

Suppose when the previous question is reversed, and you're interested in discovering the specific value that corresponds to the 25th percentile within your dataset.

Dataset: 60, 70, 75, 80, 85, 90, 95, 100

First, arrange the data in ascending order: 60, 70, 75, 80, 85, 90, 95, 100.

To find the 25th percentile, we start by calculating its position in the dataset. We use the formula: Position = (Percentile / 100) * (Total number of data points + 1).

For the 25th percentile, it's (25 / 100) * (8 + 1) = 0.25 * 9 = 2.25.

Since the position is not a whole number, we need to interpolate. This means the 25th percentile falls between the 2nd and 3rd data points.

To interpolate, we take the average of the values at positions 2 and 3. In this case, it's the average of the 2nd and 3rd data points, which are 70 and 75.

Calculate the average: (70 + 75) / 2 = 145 / 2 = 72.5.

So, the 25th percentile of your dataset is 72.5.

Quartiles:

Now, let's understand another important concept called quartiles. Quartiles divide a dataset into four equal parts , each containing 25% of the data. It's like slicing a cake into four equal pieces so that each piece represents a different part of the whole cake. The three quartiles are often denoted as Q1, Q2, and Q3. Q2 is the same as the median (50th percentile), while Q1 and Q3 represent the 25th and 75th percentiles, respectively. These quartiles help you understand the spread and distribution of the data, and to identify unusual values, called outliers.

Outliers are like the oddballs or the strange ducks in the data. They are values that don't fit with the rest of the numbers. For example, in your class's test scores, most students might score between 60 and 90, but there might be one student who got an incredibly high score of 150. This high score is an outlier because it's way above the usual range. Outliers can also be extremely low values that don't fit with the majority of data. Finding and understanding outliers is important because they can tell us if something unusual or unexpected is happening in our data.

Now that you've got the hang of quartiles and how they help us divide data into four equal parts, let's take our understanding a step further with something called the '5-Number Summary.' Think of it as a way to capture more essential information about your dataset. This summary consists of five key values that provide a quick and meaningful snapshot of your data's distribution**.**

What is the 5-Number Summary?

The five-number summary is a simple way to describe the main characteristics of a set of numbers or data. It consists of five key values that help you understand the data's central tendency and spread.

It consists of five essential values that help us describe a dataset. The first value is the Minimum.

Minimum: This is the smallest number in your data. It represents the lowest point or value in your dataset.

First Quartile: Next up is Q1, which is the first quartile. It's essentially the point below which 25% of the data falls.

Second Quartile: In the middle, we have the Median, also known as Q2 or the second quartile. This is the middlemost value in the dataset. It's essentially the point below which 50% of the data falls, and above which 50% of the data falls.

Third Quartile: After that, we have Q3, the third quartile. This is the point below which 75% of the data falls.

Maximum: And finally, the last value is the Maximum, which is, well, the highest value in the dataset.

These five values are often represented using a graphical tool called a Box-and-Whisker Plot. It's a simple yet powerful way to visualize the 5-Number Summary.

Let's consider a dataset to understand the five-number summary.

Data points: 10, 15, 20, 25, 30, 35, 40, 45, 50, 55

Minimum: It is the smallest number in the dataset. In this case, the minimum is 10.

First Quartile (Q1): It is the point below which 25% of the data falls i.e it is the 25th percentile.

To find Q1, we first need to calculate its position:

Position = (25 / 100) * (10 + 1) = 0.25 * 11 = 2.75

Since the position is not a whole number, we need to interpolate. This means the 25th percentile falls between the 2nd and 3rd data points.

Q1 = Average of the 2nd and 3rd data points: (15 + 20) / 2 = 35 / 2 = 17.5

Second Quartile (Q2): The middlemost value in the dataset also known as Median, below which 50% of the data falls i.e the 50th percentile.

To find Q2, we simply take the median of the dataset. Since there are 10 data points, it's the average of the 5th and 6th values:

Q2 = (30 + 35) / 2 = 65 / 2 = 32.5

Third Quartile (Q3): It is the point below which 75% of the data falls. To find Q3, we calculate its position:

Position = (75 / 100) * (10 + 1) = 0.75 * 11 = 8.25

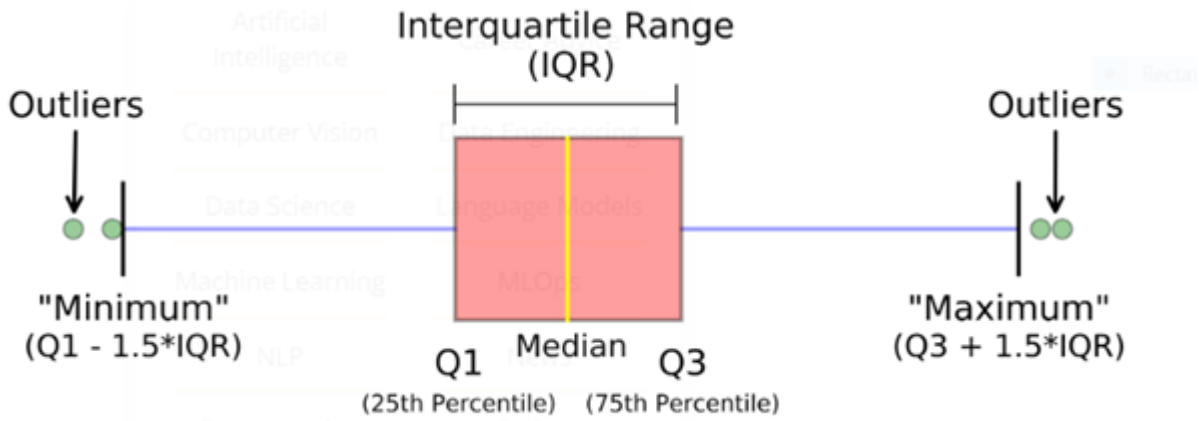
Again, since the position is not a whole number, we interpolate. This means the 75th percentile falls between the 8th and 9th data points.

Q3 = Average of the 8th and 9th data points: (45 + 50) / 2 = 95 / 2 = 47.5

Maximum: The highest value in the dataset. In this case, the maximum is 55.

The 5-number summary is often used to create a box plot that visually represents the spread of data. The "box" in the box plot covers the range from Q1 to Q3, and the "whiskers" extend to the minimum and maximum values.

This boxplot visually displays the central tendency, spread, and the presence of outliers in a dataset.



Now, imagine you have a dataset of ages, just like your group of friends. Most ages are between 25 and 30, but there's that one age of 60. This seems like an outlier, right?

To confirm if a data point is an outlier, we can define something called the lower fence and the upper fence. These fences help us set boundaries for what we consider typical data points and what we consider outliers.

Lower Fence: Usually calculated as Q1 - 1.5 * IQR

Upper Fence: Usually calculated as Q3 + 1.5 * IQR

Where IQR=Q3 - Q1

Any data point that falls below the lower fence or above the upper fence is considered an outlier. In our case, an age below the lower fence or above the upper fence would be an outlier.

Let's consider a numerical example to illustrate the concept of identifying outliers using the lower and upper fences with a dataset of ages.

Suppose we have a dataset of ages from a group of friends, and it looks like this:

Ages: 25, 26, 27, 28, 29, 30, 60

Step 1: Calculate the Five-Number Summary

First, let's calculate the five-number summary:

Minimum: 25

First Quartile (Q1): 26

Median (Q2): 28

Third Quartile (Q3): 30

Maximum: 60

Step 2: Calculate the Interquartile Range (IQR)

Now, we'll calculate the Interquartile Range (IQR) which is used to calculate upper and lower fences, which is the range from Q1 to Q3:

Calculate the IQR:

IQR = Q3 - Q1

IQR = 30 - 26

IQR = 4

Step 3: Calculate the Lower and Upper Fences

Next, calculate the lower fence and upper fence:

Lower Fence = Q1 - 1.5 * IQR

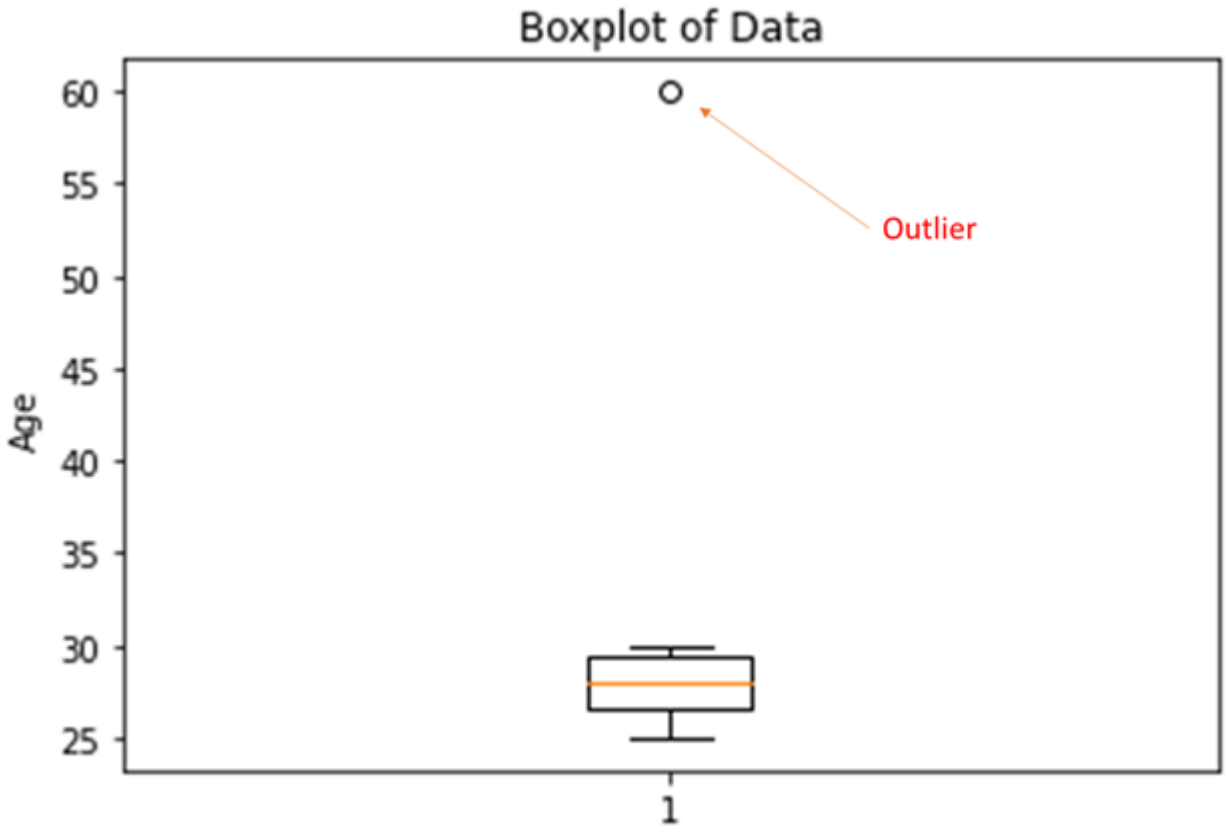
Upper Fence = Q3 + 1.5 * IQR

Lower Fence = 26 - 1.5 * 4= 26 - 6= 20

Upper Fence = 30 + 1.5 * 4= 30 + 6= 36

So, the lower fence is 20, and the upper fence is 36. Any data points outside this range can be considered outliers.

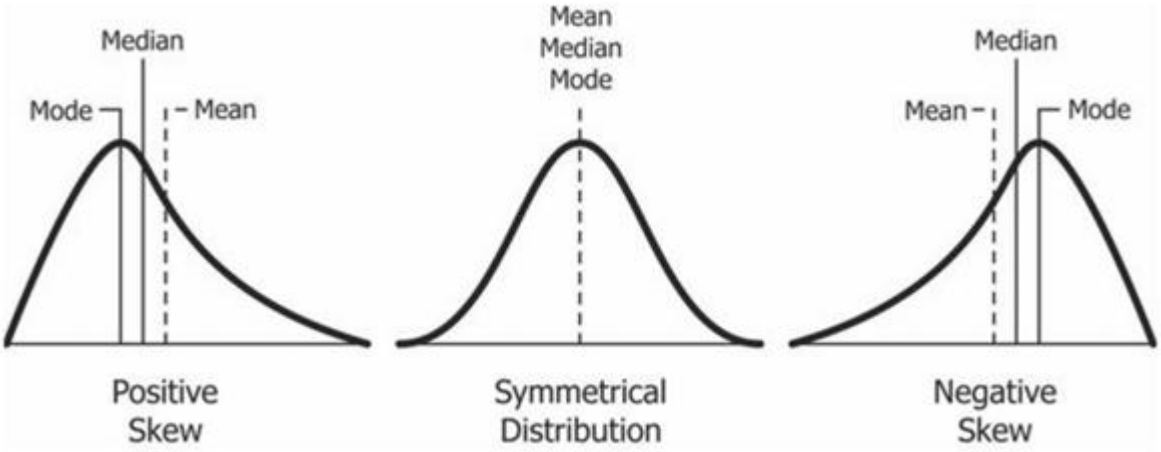
So, based on the lower and upper fences, the age of 60 is considered an outlier in this dataset.



Next, we have Skewness and Kurtosis

Skewness

Skewness is a measure that tells us how much a distribution of data deviates from being symmetrical. In simpler terms, it helps us understand whether the data is leaning more to the left or the right.



A distribution that extends with a longer tail towards the right side is termed a right-skewed or positively skewed distribution. In such distributions, there is a greater concentration of data points towards the right tail compared to the left.

In right-skewed data: mean > median>mode

A symmetrical distribution is a type of distribution where the values are evenly distributed around the center or mean of the distribution. In a symmetrical distribution, the mean, median, and mode are all equal and located at the center of the distribution. Many symmetrical distributions exhibit a bell-shaped curve when plotted.

mean = median =mode

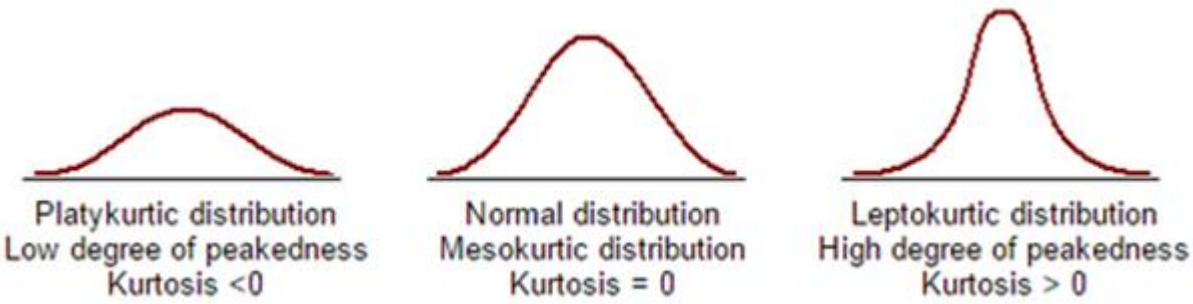
When a distribution displays a longer tail extending towards the left side, it is referred to as a left-skewed or negatively skewed distribution. In left-skewed distributions, there is a higher concentration of data points towards the left tail compared to the right.

In left-skewed data: mean<median<mode

Skewness helps us see if there's a tilt in our data, and it's essential because it can affect the conclusions we draw from statistical analyses. For example, a skewed distribution might indicate that the average (mean) isn't the best measure of central tendency, and we might need to use the median instead.

Kurtosis

Now, let's talk about kurtosis. Kurtosis measures the shape of the distribution, specifically how much data is in the tails and how sharp or flat the peak is compared to a normal distribution (bell curve).



If a distribution has high kurtosis, it means that most of the data points are close to the mean, and there are more extreme values in the tails of the distribution. This creates a peak that is taller and sharper, with longer tails on the sides.

On the other hand, if a distribution has low kurtosis, it means that the data points are more spread out, and there are fewer extreme values

in the tails. This creates a peak that is shorter and flatter, with shorter tails on the sides.

kurtosis helps us understand how much data is concentrated around the mean and how spread out the tails of the distribution are. It gives us insights into the overall shape of the distribution and how it compares to a standard or "normal" distribution.

Why does this matter?

Well, it gives us insights into the nature of extreme values or outliers in our data. High kurtosis suggests there are more outliers, and the data has heavy tails, while low kurtosis implies fewer outliers and lighter tails.

Understanding skewness and kurtosis helps us grasp the shape and behavior of our data, allowing for better-informed statistical analyses and interpretations.

The next topic that we'll discuss is Covariance

Covariance

Covariance is a statistical measure that tells us how two variables change together. It indicates whether an increase in one variable corresponds to an increase or decrease in another variable.

Covariance can be divided into two types: positive covariance and negative covariance. Positive covariance occurs when an increase in one variable corresponds to an increase in another variable.

Imagine you are tracking the daily number of hours students spend studying (variable X) and their corresponding exam scores (variable Y) in a class. You collect data for several students over a semester.

If, on days when a student studies more hours than usual, their exam score tends to be higher, then there is a positive covariance between study hours and exam scores. But if, on days when a student studies fewer hours, their exam score tends to be higher, then there is a negative covariance between study hours and exam scores.

In both cases, covariance provides insights into the direction of the relationship between the two variables. It helps us understand

whether they tend to move in the same direction (positive covariance) or in opposite directions (negative covariance).

So a positive covariance indicates that both variables tend to move in the same direction.

Negative covariance indicates that both variables tend to move in opposite directions. A covariance of 0 tells us that there is no clear relationship between the variables. But Covariance doesn't tell us the strength of the relationship, that's what correlation measures.

Correlation acts like a measuring stick for the intensity of their relationship. Instead of just indicating direction (same together or opposite), it gives you a numerical score between -1 and 1 that tells you how strong that bond is.

The correlation coefficient is often denoted as "r"

If r =1, it indicates a perfect positive linear relationship, meaning that as one variable increases, the other variable also increases proportionally.

If r=−1, it indicates a perfect negative linear relationship, meaning that as one variable increases, the other variable decreases proportionally.

If r=0, it indicates no linear relationship between the variables.

However, In a real-world scenario, encountering correlation coefficients as extreme as 1 or -1 is rare. While perfect positive and negative linear relationships exist theoretically, they are not commonly observed in practice. Real-world data tends to be more complex and subject to various factors, leading to correlations that are less than perfect.

It is more likely for us to encounter a value of r between 1 and -1. For example, if the value for r is 0.7, it indicates a strong positive linear relationship between two variables. This means that as one variable increases, the other variable tends to increase as well, but it's not a perfect relationship. The relationship is considered strong because the correlation coefficient is close to 1.

However, on the other hand, A correlation coefficient (r) of 0.2 indicates a weak positive linear relationship between the variables.

The relationship is considered weak because the correlation coefficient is relatively close to 0. This suggests that there is little to no clear pattern in the data.

The rule of thumb for interpreting the size of correlation coefficient is as follows:

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation