

Basic Statistics

Introduction

Statistics and its Types

- Descriptive Statistics

- Inferential Statistics

Population And Sample

Sampling Techniques

Descriptive Statistics

- Frequency Distribution

- Bar Graph, Histograms

- Measures of Central Tendency : Mean, Median, Mode

- Why Measures of Central Tendency is not enough?

- Measure of Dispersion/Spread

 - Variance, Standard Deviation

 - Percentiles and Quartiles

 - Range, Interquartile Range

- Box and Whisker plot for Outlier Removal

 - Five number summary

What is Statistics?

Statistics is the science of collecting, organizing and analysing data. It involves methods for summarizing data, drawing conclusions from data, and making decisions in the presence of uncertainty.

Descriptive statistics provide a snapshot of the data, while inferential statistics extend the insights to a larger population.

Aspect	Descriptive Statistics	Inferential Statistics
Objective	Organize and Summarize and describe the main features of data. Summarize and present data in an understandable way. No broader conclusions or predictions, just focuses on Data characteristics within the sample.	Make inferences or predictions or conclusions about a larger population based on a sample. Make informed decisions, test hypotheses, draw broader conclusions. Broader conclusions, focuses on population characteristics based on sample information.
Methods	Measures of central tendency, dispersion, and graphical representations.	Hypothesis testing, confidence intervals, regression analysis, etc.
Example	Calculate mean income, create histograms, bar charts.	Test if a new drug is effective, estimate population parameters, predict future outcomes.
Real Life	When we go and buy a house, we describe it and make decision on house itself.	When we buy Rice, we check only a handful of rice and make decision on the overall lot.

Suppose you have a sample of heights from a school, denoted as:

$$X=\{x_1, x_2, \dots, x_{50}\}$$

Descriptive Statistics: Descriptive statistics provide summary measures of the sample. It summarizes the heights of the 50 students in your sample.

Mean (Average): You calculate the mean height of the 50 students to determine the typical height in the sample.

Histogram: You create a histogram to visualize the distribution of heights in the sample, showing the frequency of students in different height ranges.

Inferential Statistics: Inferential statistics are used when you want to make inferences about a larger population based on a sample data. In this case, you might want to estimate the average height of all students in the school.

Confidence Interval: You calculate a confidence interval for the mean height, which gives you a range of heights within which you can be confident the true population mean falls.

Hypothesis Testing: You perform a hypothesis test to determine whether the average height of students in your school is different from the national average.

Sample Size: Inferential statistics can also help you decide how large your sample should be to make reliable inferences about the entire student population.

Answering questions like “Are marks of students of this classroom similar to marks of Maths classroom in the college?”

Population (N) And Sample (n)

Population and Sample are two fundamental concepts in statistics that help researchers study and draw conclusions about a group of interest.

Population:

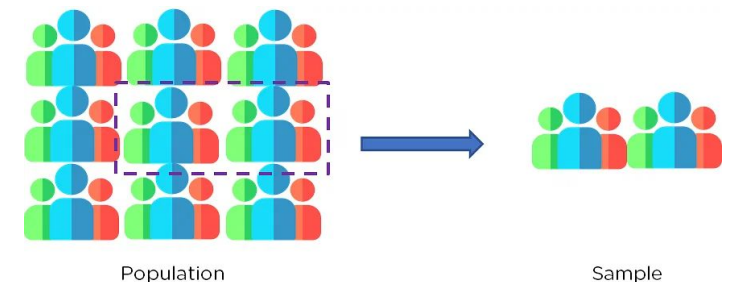
Definition: The population refers to the entire group or collection of individuals, objects, or data points that you are interested in studying. It represents the complete set of items that fall under the scope of your research.

Example: Consider you want to study the average income of all working professionals in a specific city. The population, in this case, is the entire group of working professionals in that city, which could number in the thousands or more.

Sample:

Definition: A sample is a subset or a smaller, representative group of individuals or data points selected from the population to represent the population. It is chosen in a way that allows you to make inferences and draw conclusions about the entire population without having to study every single member of the population as that might be impractical or too costly.

Example: To study the average income of all working professionals in the city, it may not be feasible to collect data from everyone. Instead, you randomly select 500 working professionals from different neighborhoods in the city. This group of 500 individuals is your sample. By analyzing the income data from this sample, you can make estimates and draw conclusions about the average income of the entire population.



Sampling Techniques

Each of these sampling techniques has its own advantages, limitations, and appropriate use cases. The choice of sampling method depends on the research objectives, available resources, and the desired level of representation within the sample. The key to effective sampling is ensuring that the sample is representative and unbiased, allowing for valid generalizations to be made about the population.

Sampling Technique	Description	Example
Simple Random Sampling (SRS)	Each individual in the population has an equal chance of being selected.	Randomly selecting students from a class list to conduct a survey.
Stratified Sampling	The population is divided into subgroups (strata), and random samples are drawn from each stratum.	Dividing a population of employees into different job roles (strata) and selecting random samples from each stratum to ensure representation of gender in all job roles.
Systematic Sampling	Sampling is performed by selecting every kth element from a list.	Selecting every 10th customer from a list of customers in a database for a customer satisfaction survey.
Convenience Sampling	Sampling based on convenience and accessibility.	Conducting surveys at a shopping mall by approaching individuals who happen to be present at a given time.
Snowball Sampling	Initially, one or a few participants are selected, and they refer or recruit additional participants.	Studying a rare disease by identifying one individual who has the disease and then asking them to refer other affected individuals.
Quota Sampling	The researcher divides the population into subgroups (quotas) and selects individuals non-randomly until each quota is filled.	Surveying a fixed number of males and females in a specific age group to ensure gender and age representation in the sample.

Frequency Distribution

Consider a dataset of exam scores for a class of 30 students. The exam scores range from 60 to 100.

Ex 1 : Here are the exam scores for the students:

80, 85, 72, 90, 92, 78, 85, 88, 70, 78, 88, 92, 95, 80, 75, 85, 92, 88, 80, 78, 75, 90, 92, 80, 85, 70, 78, 88, 85, 75

Ex 2 : List of favourite colors of students in a class:

- Student A: Red
- Student B: Blue
- Student C: Green
- Student D: Red
- Student E: Blue
- Student F: Green
- Student G: Green
- Student H: Yellow
- Student I: Blue
- Student J: Purple

Continuous

Score Range	Frequency
60-70	0
70-80	10
80-90	13
90-100	7

Discrete

Color	Frequency
Red	2
Blue	3
Green	3
Yellow	1
Purple	1

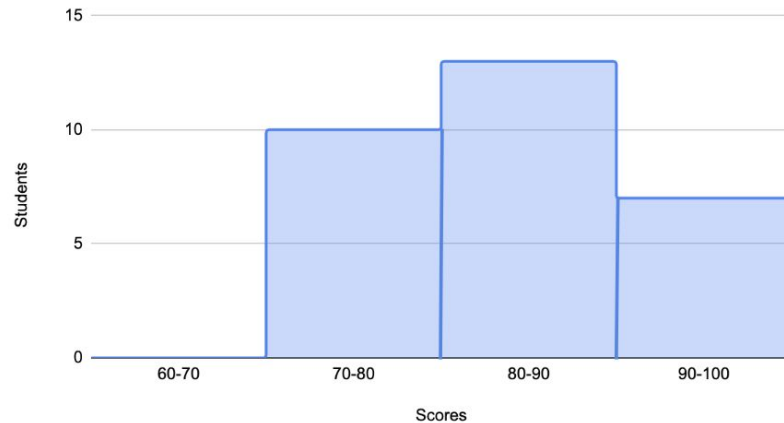
Above are frequency distributions to count how many times each score occurs in the dataset and how many students like what color respectively.

Bar Graphs and Histograms

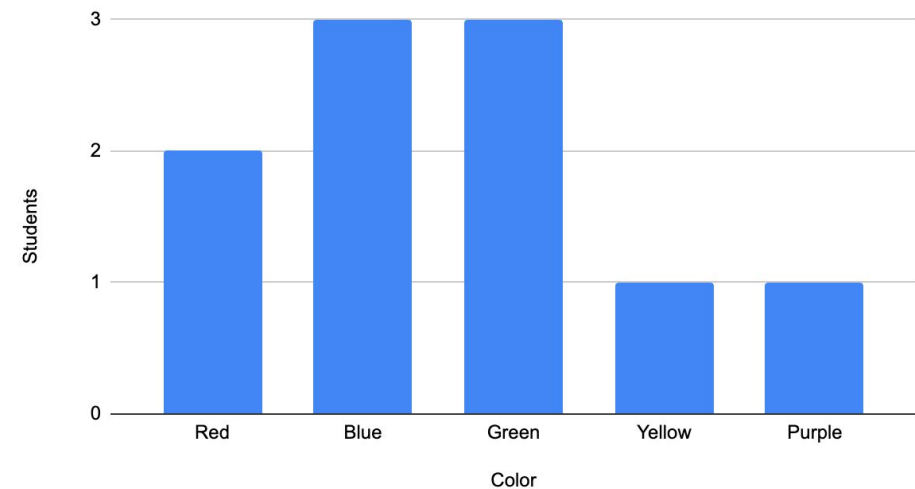
By creating a frequency distribution, bar graph, and histogram, we can easily understand the distribution of exam scores and identify patterns within the data.

While both bar graphs and histograms use bars to represent data, they serve different purposes. Bar graphs are suitable for categorical or discrete data, where each category is distinct and unrelated. Histograms, on the other hand, are used for continuous data, where values are grouped into intervals to visualize the distribution and identify patterns. The key distinction lies in the nature of the data and the use of fixed intervals in histograms.

Students vs Score Histogram



Students vs Color BarGraph



Use Cases of Stats

In a survey, we collected the ages of participants. What is the average age of the participants?

In a marathon race, what is the "typical" finishing time for runners? Should we use the fastest or slowest time?

In a survey, we collected students placement data. Which company is the most popular among the students?

In a college, how do we get to know average placement range of students?

We have test scores from two different schools. How do we compare the spread of scores between the two schools?

In a sales dataset, what is the difference between the highest and lowest sales figures in a given period?

We have employee salaries for a company. How can we understand the distribution of salaries within the company?

In a test score dataset, how does a student's score compare to the rest of the students? Is it in the top 10%?

Measures of Central Tendency

A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. It is used to determine the center of the distribution of the data. They are sometimes called measures of the central location and are also classed as summary statistics.

Measure of Central Tendency	Definition	Calculation	Usefulness	Example
Mean/Average	Sum of all values divided by the number of observations.	$\text{Mean}(\mu) = (x_1 + x_2 + \dots + x_n) / n$	The mean is most appropriate for data that follows a roughly symmetrical distribution, such as the normal distribution.	Mean of height of students 5,5.5,4,4.5,5.2,5.8 = 5 (Used in Normalisation, Standardisation, MSE, etc)
Median	Middle value when dataset is arranged.	n is odd, median = value at $(n+1)/2$; n is even, median = average of values at $n/2$ and $(n/2)+1$	Useful in case of outliers as median is less impacted by outliers. Good measure for skewed data.	Median of annual salaries in college placement where 100L is an outlier: 5L,6L,10L,3L,8L,7L,100L = 7L (Used in Numerical Data Imputation)
Mode	Most Frequent Value	The mode may not be unique or may not exist in some datasets. There can be unimodal, multimodal or no mode models.	Useful when looking for the most common value or generally when the data is categorical or nominal.	Replacing by mode in a Gender data: M, M, M, F, M, F, _, _ Mode = M (Useful in Categorical Data Imputation)

Why Measures of Central Tendency not sufficient Enough ?

Two or more distributions may have averages which are exactly alike, even though the distributions are dissimilar in other aspects. Central Tendency does not provide any information about the degree to which the data tend to spread or scatter about the average value.

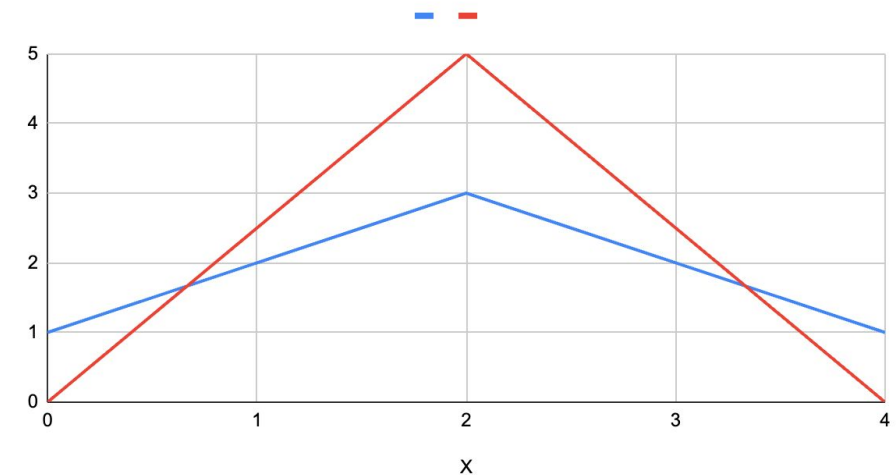
$\{0,2,2,2,4\}$ -> Mean = 2, Median = 2, Mode = 2

$\{2,2,2,2,2\}$ -> Mean = 2, Median = 2, Mode = 2

The mode is primarily used to identify the most frequently occurring value in a dataset, but it may not always provide clear differentiation between distributions, especially if the distributions have different shapes or patterns. While the mode can be a useful measure, especially for identifying common values in a dataset, it may not always be sufficient to differentiate between two distributions if their overall shapes and characteristics are similar.

To differentiate between two distributions, especially when the mean and median are the same, you may need to consider other statistical methods, such as measures of spread (variance or standard deviation), or visual techniques like histograms or probability density functions. These methods can help us gain a better understanding of the underlying characteristics of the distributions and how they differ.

Data1 v/s Data2



Example

Given the initial set of marks: 48, 49, 50, 51, 52, 0, 100, 10, 90, 20, 80

Set 1 (around 50): Marks: 48, 49, 50, 51, 52, 50, 50, 50, 50, 50, 50

Explanation: In this set, most values are centered around 50 to maintain the same mean and median as the original set.

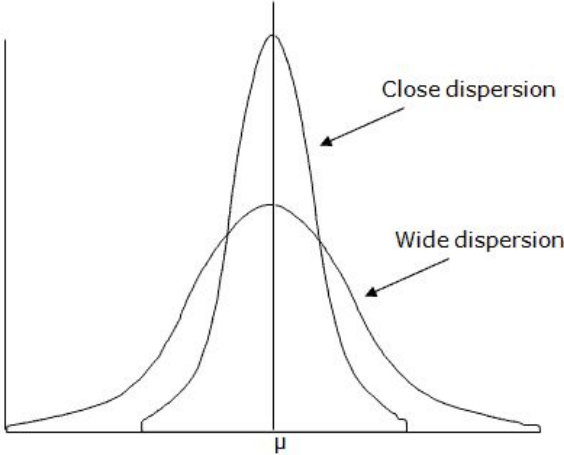
Set 2 (mix of low and high values): To achieve the same mean and median but with a mix of low and high values:

Marks: 0, 0, 0, 0, 0, 100, 100, 100, 100, 100, 50

Explanation: This set is designed to maintain the same mean and median as the first set while incorporating extreme values at the lowest and highest ends (0 and 100) with fewer values around the median (50).

Both sets have the same mean and median (approximately 50), but Set 1 centers around 50, and Set 2 involves extreme values (0 and 100) to achieve the same statistical measures.

Measures of Dispersion (Spread)

Measure	Definition	Formula	Graph
Variance	<ol style="list-style-type: none"> Variance is an essential statistical tool for quantifying and understanding the spread of data. Variance is the expected value of the squared differences from the mean. The use of squared differences has some mathematical advantages, such as simplifying calculations and being friendly to certain statistical tests. <p>Note : Square is done to avoid cancelling of -ve and +ve variances.</p>	<p>Population Variance</p> $\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$ <p>Sample Variance</p> $s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$	 <p>Close dispersion</p> <p>Wide dispersion</p> <p>Same mean and different dispersion</p>
Standard Deviation	<ol style="list-style-type: none"> Standard deviation is an essential statistical tool for quantifying and understanding the spread of data and understanding how far a value is from the mean. Standard deviation is the square root of the variance. Standard deviation is more interpretable and intuitive. For example, if you're measuring the heights of individuals in centimeters, the standard deviation will also be in centimeters, whereas the variance will be in square centimeters. 	<p>Population Standard Deviation</p> $\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$ <p>Sample Standard Deviation</p> $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n - 1)}}$	<p>More Variance -> More spread -> More dispersion</p> <p>Variance and Standard Deviation can be calculated for all kind of data.</p> <p>Is Std robust to outliers?</p> <p>Why Variance when we have standard deviation and vice versa?</p>

Why $n-1$ in Sample Variance and Sample Standard Deviation?

Sample Variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{(n - 1)}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{(n - 1)}}$$

However, when we are working with a sample (a subset of the population), we often use sample statistics to estimate population parameters. Dividing by $n-1$ instead of n in the denominator is known as using Bessel's correction. This correction is applied because the sample data typically underestimates the population variability. Dividing by $n-1$ instead of n helps to adjust for this underestimation and provides a more accurate estimate of the population variability.

Why use Standard Deviation when we have Variance?

Variance is often used in mathematical/statistical calculations, especially when we need to combine variances from different sources or perform certain statistical tests where the squared values are beneficial. While variance has its uses, it is often less intuitive and harder to interpret compared to standard deviation because it is expressed in squared units.

In most cases where we want to interpret variability in the same units as the original data, or where interpretability is crucial, standard deviation is preferred over variance.

Example

Consider the following dataset representing the test scores of a group of students:
78, 85, 92, 88, 75, 80, 92, 80, 85, 70

$$\text{Mean} = (78 + 85 + 92 + 88 + 75 + 80 + 92 + 80 + 85 + 70) / 10 = 845 / 10 = 82.5$$

$$\text{Variance} = [((78 - 82.5)^2 + (85 - 82.5)^2 + (92 - 82.5)^2 + (88 - 82.5)^2 + (75 - 82.5)^2 + (80 - 82.5)^2 + (92 - 82.5)^2 + (80 - 82.5)^2 + (85 - 82.5)^2 + (70 - 82.5)^2) / 10] \approx 46.85$$

$$\text{Standard Deviation} \approx \sqrt{46.85} \approx 6.84$$

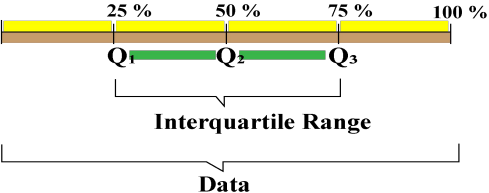
The "1st std" represents the range of values within one standard deviation of the mean. In our example, it would include values within 1 standard deviation of 82.5, so the range is approximately $[82.5 + 6.84, 82.5 - 6.84] = [89.34, 75.66]$

The "2nd std" represents the range of values within two standard deviations of the mean. In our example, it would include values within 2 standard deviations of 82.5, so the range is approximately $[82.5 + 2(6.84), 82.5 - 2(6.84)] = [96.18, 68.82]$

The "3rd std" represents the range of values within three standard deviations of the mean, and so on.

Measures of Dispersion (Spread)

Measure	Definition	Formula	Example
Percentiles	<p>A percentile is a statistical measure that indicates how a particular data point compares to the rest of the data in terms of its position or rank.</p> <p>Percentile is a value below which certain percentage of observation lie.</p> <p>The "kth percentile" is defined as the value below which k% of the data falls. In other words, it divides the data into two parts: k% of the data falls below the kth percentile, and (100 - k)% of the data falls above the kth percentile.</p>	<p>Order the dataset in ascending order.</p> <p>Calculate the rank (position) of the desired percentile using the formula: $\text{Rank} = (P / 100) * (N + 1)$, where P is the desired percentile (e.g., 25 for the 25th percentile), and N is the total number of data points.</p> <p>Percentile of x = $(\text{Rank}/n)*100$</p>	<p>Marks of students : 45 55 60 68 70 75 78 82 90 92</p> <p>Find percentile of students with marks = 60</p> <p>Percentile = $(3/10)*100 = 30^{\text{th}}$ i.e. 30%ile.</p> <p>What value exists at percentile ranking of 25th?</p> <p>index = $(25/100)*11 = 2.75$ value at index = $(55+60)/2 = 57.5$</p>

Measure	Definition	Formula	Comment
Quartiles	Quantiles are values that divide a dataset into equal parts. Quartiles divide data into four parts, quintiles into five parts, and percentiles into 100 parts. Quartiles divide data into four equal parts.	Q1: Rank = (25% of N+1), where N is the total number of data points. Q2 (the median): Rank = (50% of N+1). Q3: Rank = (75% of N+1)	Q1 (lower quartile) is the value below which 25% of the data falls. Q2 (median) divides the data into two equal halves, with 50% of the data below it and 50% above it. Q3 is the value below which 75% of the data falls.
Range	The range is the simplest measure of variability and represents the difference between the maximum and minimum values in a dataset. It provides a quick overview of the spread of the entire dataset but doesn't describe the central spread well.	16, 24, 22, 25, 26, 27, 28, 23 Range = max - min Range = 28 - 16 = 12	Sensitivity to Outliers: The range is sensitive to extreme values (outliers). A single outlier can significantly affect the range, making it less robust in the presence of extreme values.
Interquartile Range (IQR)	The IQR is a measure of statistical dispersion that quantifies the spread of the middle 50% of data. 	IQR = Q3 - Q1	Sensitivity to Outliers : The IQR is used to identify the central spread of data while being less sensitive to extreme values or outliers. It focuses on the middle portion of the data and is not influenced as much by extreme values. Why?

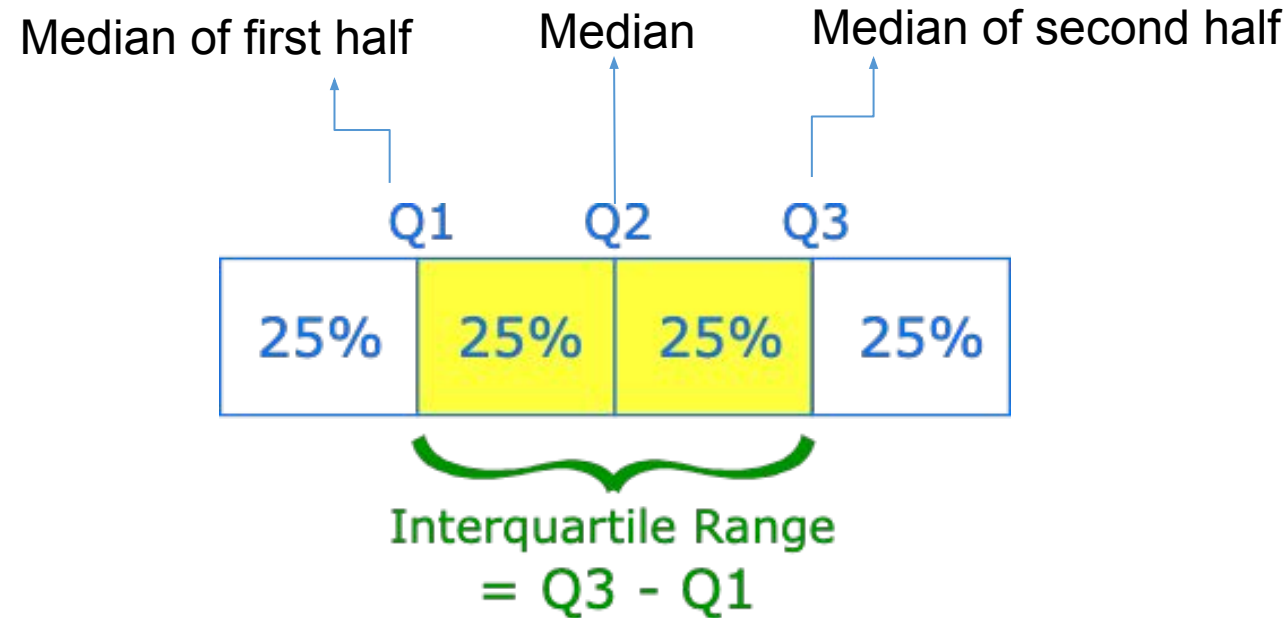
Different Methods of Quartile Calculation

Scores : 72,85,68,92,78,60,88,75,96,45

Step 1: Sort the dataset in ascending order : 45,60,68,72,75,78,85,88,92,96

Method 1: Calculate Quartiles using median of medians.

Method 2: Calculate Quartiles using percentile/quantile method.



Method 1: Calculate Quartiles using median of medians.

45, 60, 68, 72, 75, 78, 85, 88, 92, 96

Step 1: Calculate **Quartiles**

Calculate the first quartile (Q1):

- $Q1 = (25\text{th percentile}) = \text{median of 1st half of the dataset} = 68$

Calculate the second quartile (Q2):

- $Q2 = (50\text{th percentile}) = \text{median} = (75+78)/2 = 76.5$

Calculate the third quartile (Q3):

- $Q3 = (75\text{th percentile}) = \text{median of 2nd half of the dataset} = 88$

Step 2: Calculate the **IQR** = $Q3 - Q1 = 88 - 68 = 20$

The IQR represents the spread of the middle 50% of data and is useful for identifying the variability within the central portion of the dataset. In this example, it helps us understand how scores are distributed around the median score (Q2) while being less influenced by extreme scores.

Step 3: Calculate **Range** = $96 - 45 = 51$

Method 2: Calculate Quartiles using percentile/quantile method.

Scores : 72,85,68,92,78,60,88,75,96,45

Sort the dataset in ascending order : 45,60,68,72,75,78,85,88,92,96

Step 1: Calculate **Quartiles**

Calculate the first quartile (Q1):

- $Q1 = (25\text{th percentile}) = \text{median of 1st half of the dataset}$
- $Q1 = (25/100) \cdot (10+1) = 2.75\text{th Rank}$
 $Q1 \text{ is between the 2nd and 3rd data points} = 60 + 0.75 \cdot (68-60) = 66$

Calculate the second quartile (Q2):

- $Q2 = (50\text{th percentile}) = \text{median}$
- $Q2 = (50/100) \cdot (10+1) = 5.5\text{th Rank}$
 $Q2 \text{ is between 5th and 6th data point} = (75+78) / 2 = 76.5$

Calculate the third quartile (Q3):

- $Q3 = (75\text{th percentile}) = \text{median of 2nd half of the dataset}$
- $Q3 = (75/100) \cdot (10+1) = 8.25\text{th Rank}$
 $Q3 \text{ is between the 8th and 9th data points} = 88 + 0.25 \cdot (92-88) = 89$

Step 2: Calculate the **IQR** = $Q3 - Q1 = 89 - 66 = 23$

The IQR represents the spread of the middle 50% of data and is useful for identifying the variability within the central portion of the dataset. In this example, it helps us understand how scores are distributed around the median score (Q2) while being less influenced by extreme scores.

Step 3: Calculate **Range** = $96 - 45 = 51$

Methods	Median-of-Medians	numpy.percentile()	series.quantile()
Description	Efficiently estimates median for large datasets, used to approximate quartiles.	Built-in NumPy function for calculating percentiles. numpy.percentile offers different interpolation methods (linear by default).	Built-in Pandas function for calculating percentiles. Pandas .quantile method by default uses linear interpolation similar to numpy.percentile.
Advantages	Efficient for large datasets.	Accurate for all dataset sizes. Linear interpolation: It calculates the percentile value by linearly interpolating between the surrounding data points if the desired percentile doesn't fall exactly on an integer data point.	However, pandas offers additional options for interpolation methods through the interpolation argument. The default is 'linear', but you can specify other options like 'nearest' or 'higher' depending on your needs.
Disadvantages	Less accurate for smaller datasets (especially Q1 & Q3)	Less efficient for very large datasets	

- For most data analysis tasks, the slight difference between different methods is due to the underlying interpolation method (how each method handles non-integer percentiles) used but is often negligible.
- If you require very precise control over quartile calculation and need identical results across methods, you might need to explore alternative methods or adjust the interpolation settings.

Practice

Mean : In a survey, we collected the ages of participants. What is the average age of the participants?

Median : In a marathon race, what is the "typical" finishing time for runners? Should we use the fastest or slowest time?

Mode : In a survey, we collected students placement data. Which company is the most popular among the students?

Median : In a college, how do we get to know average placement range of students?

IQR (Interquartile Range) : We have test scores from two different schools. How do we compare the spread of scores between the two schools?

Range : In a sales dataset, what is the difference between the highest and lowest sales figures in a given period?

Quartiles : We have employee salaries for a company. How can we understand the distribution of salaries within the company?

Percentiles : In a test score dataset, how does a student's score compare to the rest of the students? Is it in the top 10%?

Question : Does mode exist in this data? 1,1,2,2,3,3,4,4 -> Multiple modes exist (All are modes) (As per both Python and Excel)

Box and Whisker Plot (Five number summary)

A box and whisker plot—also called a box plot—displays the five-number summary of a set of data.

minimum, first quartile, median, third quartile, and maximum.

Inner Fence = $Q_1 - 1.5 \times IQR$

Outer Fence = $Q_3 + 1.5 \times IQR$

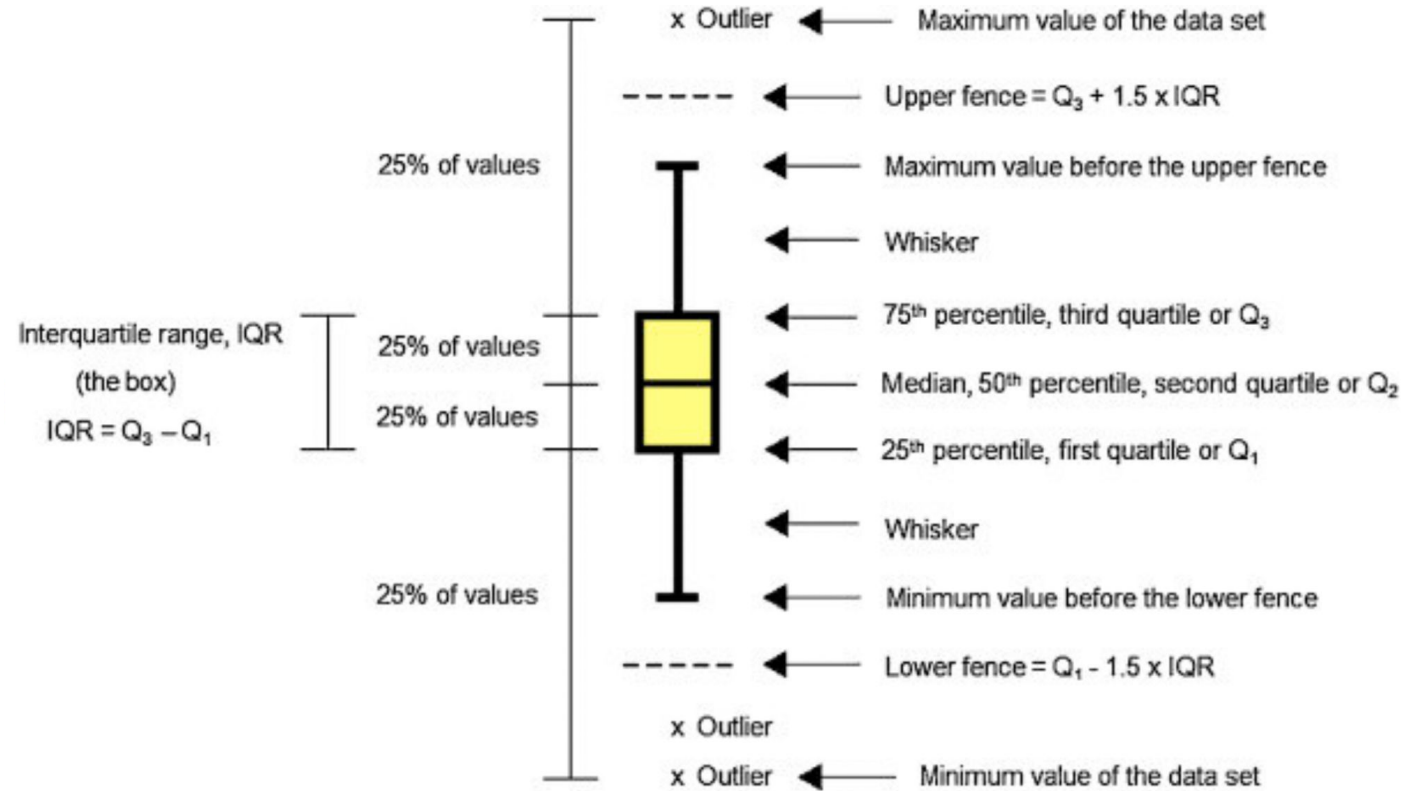
Outliers = Numbers lying beyond the fence

Minimum = smallest value in the dataset

Maximum = largest value in the dataset

In summary, boxplots are versatile tools for visualizing and summarizing data distributions, and they are effective for both symmetrical and asymmetrical distributions.

They provide valuable insights into central tendency, variability, and the presence of outliers, making them widely used in exploratory data analysis and statistical reporting.



Detecting Outliers and Data Visualisation using BoxPlot

1,2,2,2,3 |,3,4,5,5,5 |,6,6,6,6,7,| 8,8,9,11,**27**

Q1 = median of 1st half of dataset = $(3+3)/2 = 3$

Q2 = median = $(5+6)/2 = 5.5$

Q3 = median of 2nd half of dataset = $(7+8)/2 = 7.5$

IQR = $Q3 - Q1 = 4.5$

Inner Fence = $3 - 1.5 \times (4.5) = -3.75$

Outer Fence = $7.5 + 1.5 \times (4.5) = 14.25$

Hence anything outside $[-3.75, 14.25]$ will be considered an outlier.

Minimum : 1

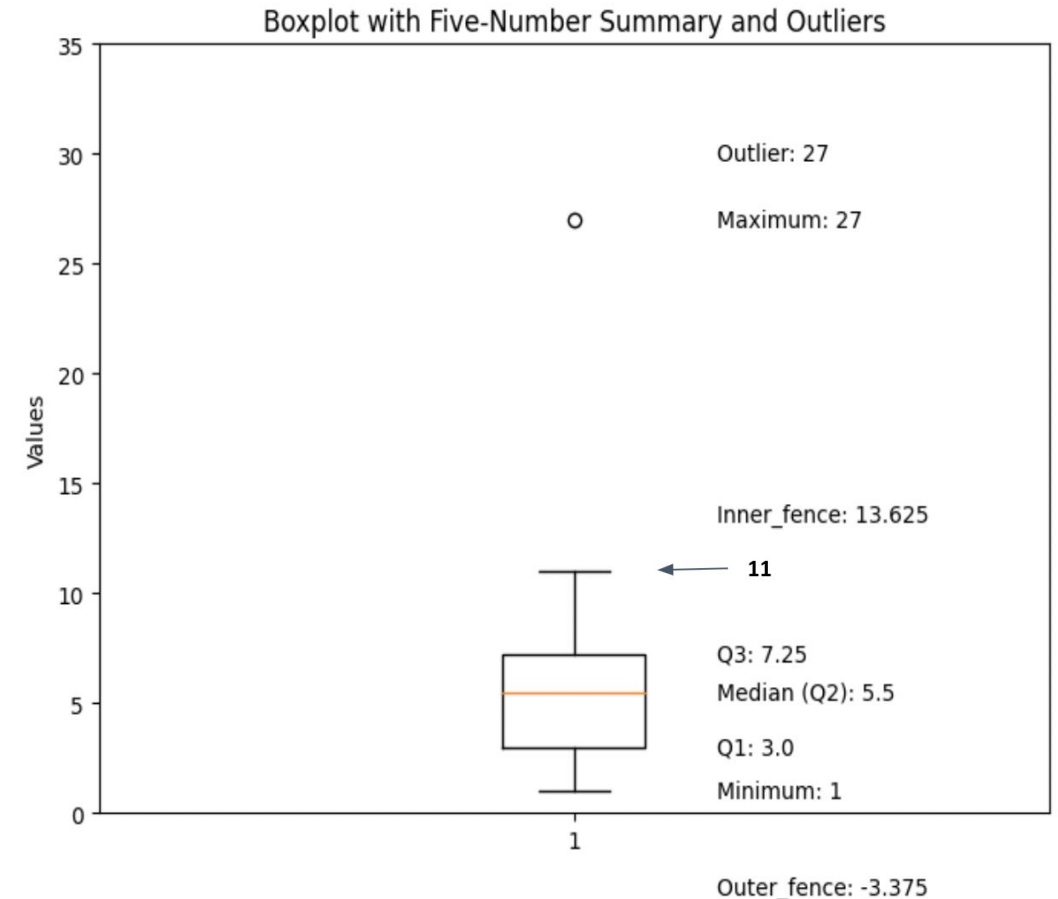
Q1 : 3.0

Median : 5.5

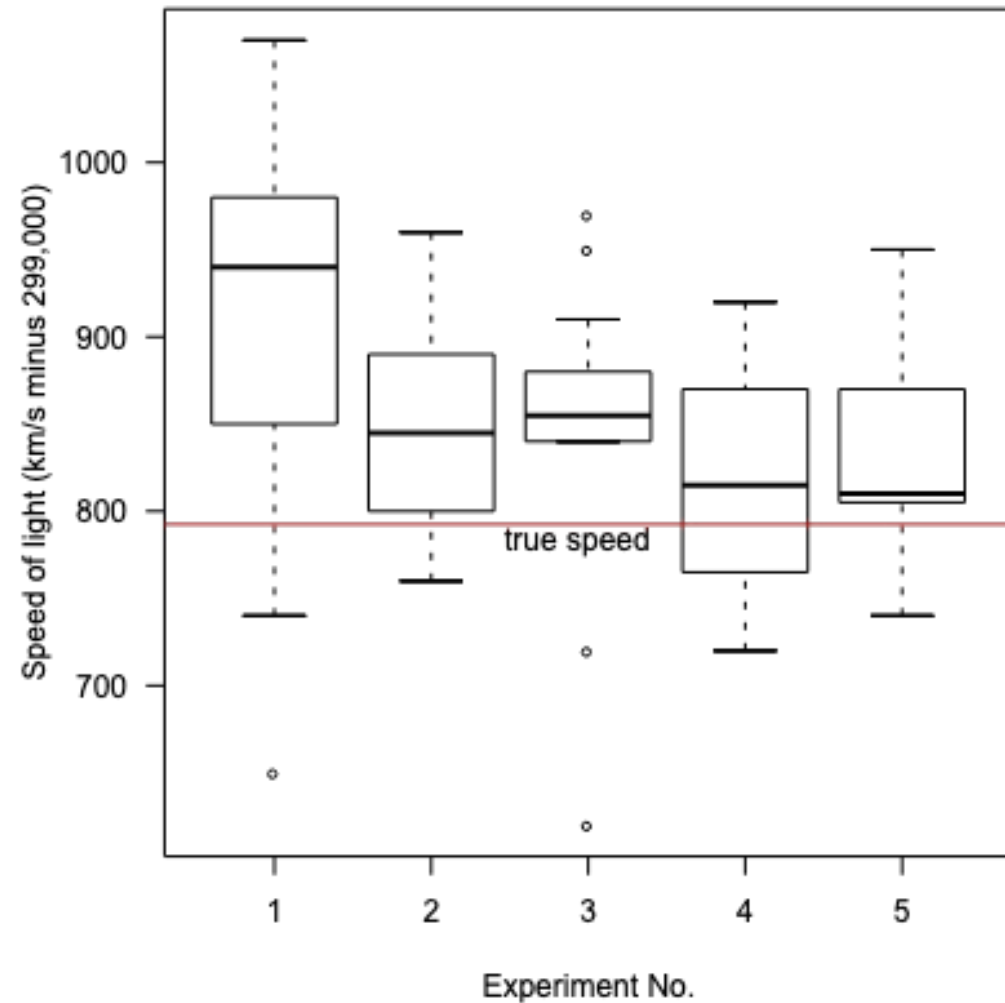
Q3 : 7.5

Maximum : 27

One can also use percentile method to find Q1 and Q3 i.e. 25thile and 75thile points.



Which of the following Experiment is giving desired results and what do each boxplot convey?



Example

Example: Finding the five-number summary

[-10], [2], [4], [3], [8], [7], [9], [15], [11], [-5]

Make a box plot of the data.

Step 1: Order the data from smallest to largest.

-10, -5, 2, 3, 4, 7, 8, 9, 11, 15

Step 2: Find the median.

The median is $(4+7)/2 = 5.5$

Step 3: Find the quartiles.

The first quartile is the median of the data points to the *left* of the median. [Q1=2]

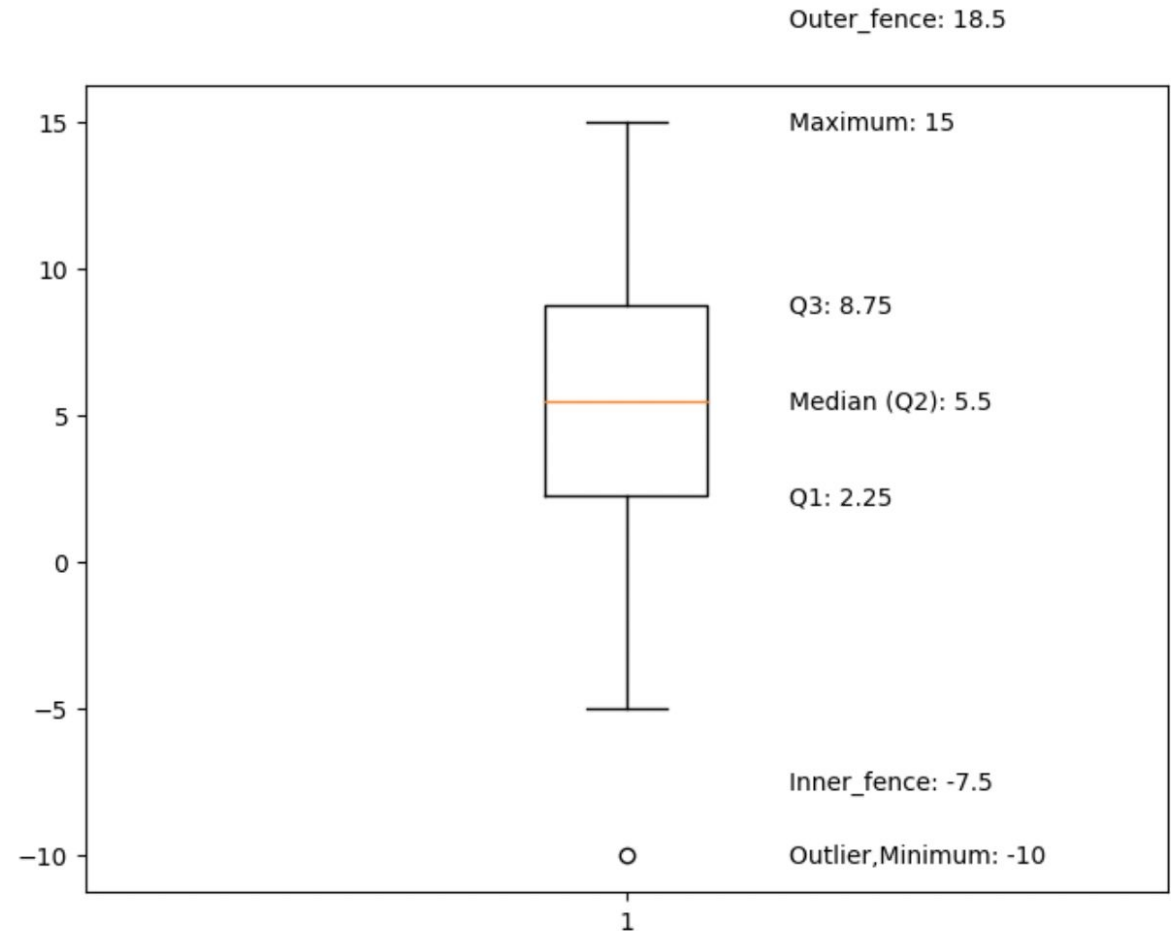
The third quartile is the median of the data points to the *right* of the median. [Q3=9]

Step 4: Complete the five-number summary by finding the min and the max.

The min is the smallest data point, which is [-10].

The max is the largest data point, which is [15].

The five-number summary is [-10], [2], [5.5], [9], [15].



Why 1.5 IQR?

In summary, $Q1 - 1.5IQR$ and $Q3 + 1.5IQR$ is a common criterion which is simple and intuitive for identifying outliers based on the assumption of a symmetric distribution. This approach balances **sensitivity to outliers with robustness against skewed data distributions.**

1.5 IQR rule is a practical and effective method for identifying outliers in both symmetrical and asymmetrical distributions. While the distribution may not be perfectly symmetric, the IQR still represents a robust measure of spread that accounts for the central bulk of the data.

- In asymmetrical distributions, the boxplot may appear elongated on one side (where the data is more spread out) and compressed on the other side (where the data is less spread out).
- Despite the asymmetry, boxplots still effectively convey key summary statistics such as the median, quartiles, and range, making them useful for understanding the shape and variability of the data.
- Boxplots are particularly useful for identifying outliers, regardless of the distribution's symmetry.

