

Statistics of Data Analytics

Mrinal Das

Population

A population is the collection of all people, plants, animals, or objects of interest about which we wish to make statistical inferences (generalizations).

Example: Students in a school.

Population parameter

A population parameter is a numerical characteristic of a Population.

Example: For students in a school as a population, their average height can be a population parameter.

Sample

A sample is the subset of the population that we actually measure or observe.

It is difficult to measure the parameter for the entire population, so we take a sample of the population to measure the parameter.

Sample Statistics

A sample statistic is a numerical characteristic of a sample. A sample statistic estimates the unknown value of a population parameter.

Often also referred to as Descriptive Statistics.

Statistics, as a subject matter, is the science and art of using sample information to make generalizations about populations.

3 Key Aspects of Data

Central Tendency

What is a typical value for each variable?

Dispersion

How far apart are the individual observations from a central value for a given variable?

Association

When more than one variable are studied together, how does each variable relate to the remaining variables?

Central Tendency

What is a typical value ?

The population mean is the measure of central tendency for the population.

Population mean $\mu = E[X]$.

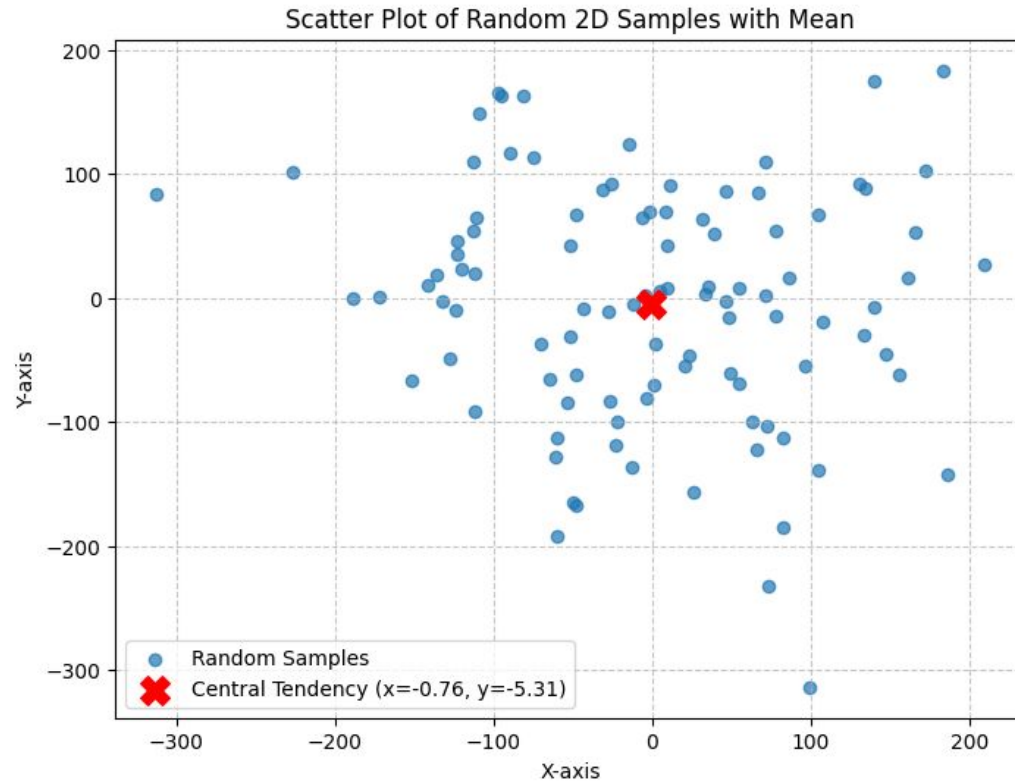
Population mean is estimated by sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Example: what is the typical height of students ?

Answer: sum heights of all students in the sample and divide by the number of students.

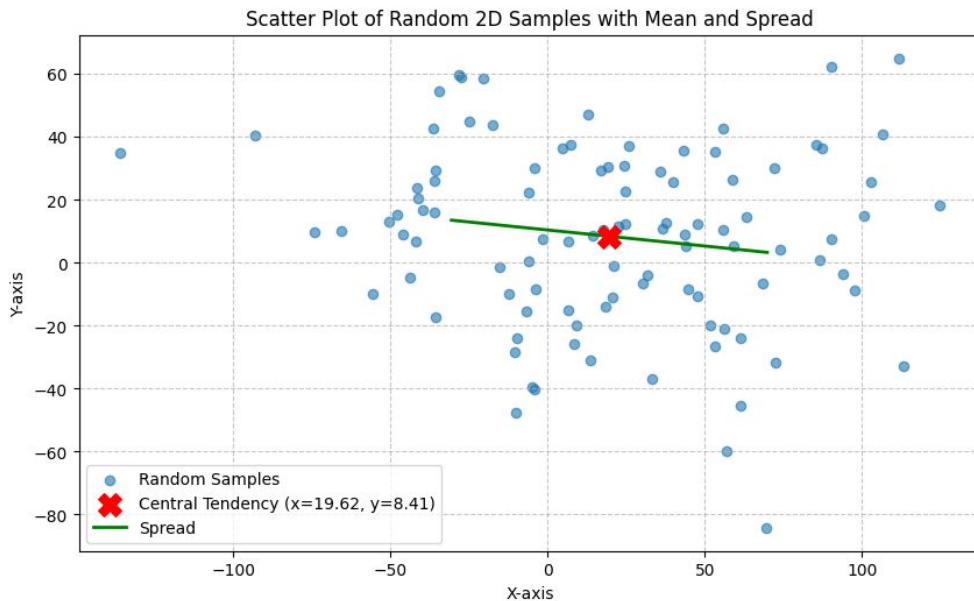
Central Tendency



Dispersion

How far apart are the individual observations from a central value for a given variable?

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



Mean Median Mode

Mean

The mean is the sum of all observations divided by the total number of observations.

Median

The median is the middle value when the data are arranged in ascending order.

Mode

The mode is the value that occurs most frequently.

Mean

Step-by-Step Calculation

Sum of all values:

$$120 + 135 + 125 + 140 + 130 + 125 + 145 = 920$$

Number of observations:

$$n = 7$$

Mean:

$$\frac{920}{7} = 131.43$$

Median

Step 1: Sort the Data

120, 125, 125, 130, 135, 140, 145

Step 2: Find the Middle Value

There are **7 observations**.

The middle (4th) value is:

130

Result

Mode

Example:

120, 125, 125, 130, 135, 140, 145

The value **125** appears **twice**, while every other value appears only once.

Mean Median Mode

Mean: average

Median: middle

Mode: most popular

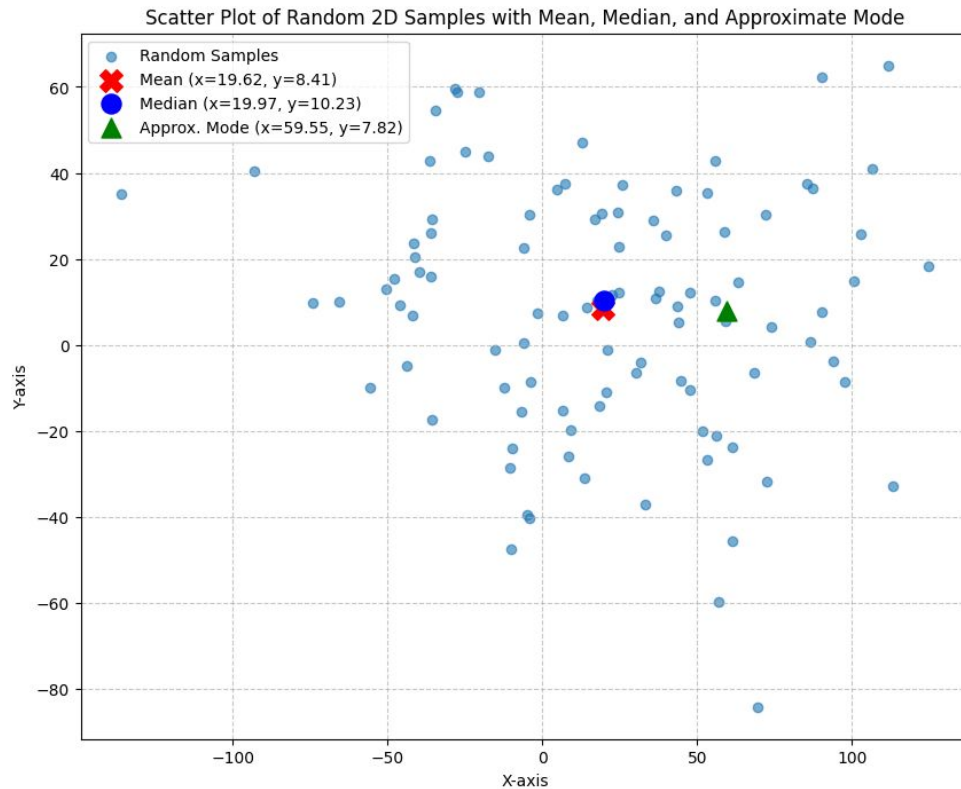
```
120  125  125  130  135  140  145  
|-----|-----|-----|-----|-----|-----|
```

Mode = 125

Median = 130

Mean $\approx$ 131.43

Mean Median Mode



Effect of an Outlier

120, 125, 125, 130, 135, 140, 300

Mean

$$\frac{120 + 125 + 125 + 130 + 135 + 140 + 300}{7} = 153.57$$

Median

120, 125, 125, 130, 135, 140, 300

Median = 130

Mode

Mode = 125

120 125 125 130 135 140 145
|-----|-----|-----|-----|-----|

Mode = 125

Median = 130

Mean $\approx$ 131.43

Importance of Outlier

Often in manufacturing industry, data analytics is used for defect control

When everything is as expected, we see normal outcomes

When something goes wrong, that gets reflected in outcome as abnormality

That abnormality is outlier

So detecting outlier helps in descriptive analytics to identify issues

Box Plot

A **box plot** (or **box-and-whisker plot**) is a graphical tool used to **summarize the distribution of numerical data**. It provides a quick view of:

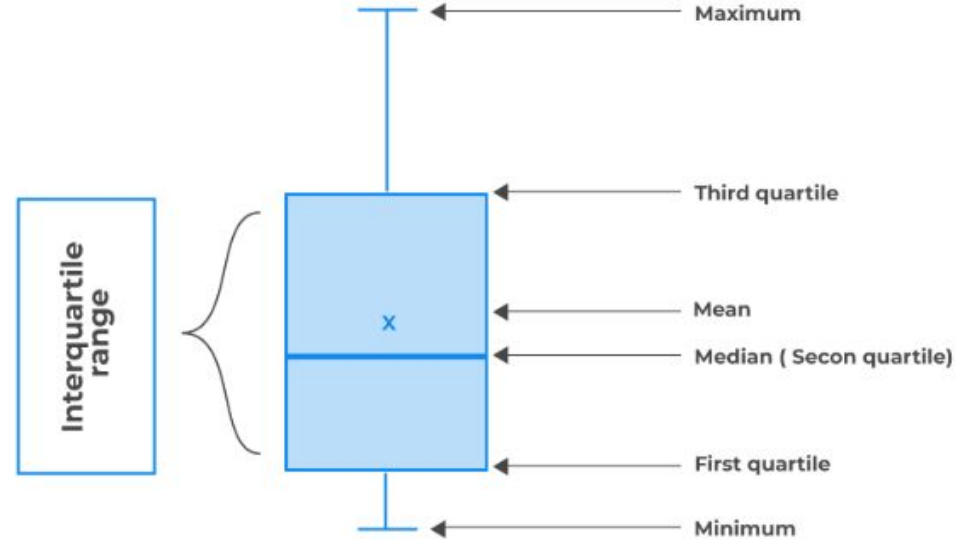
- The **center** of the data (median)
- The **spread** of the data (interquartile range)
- The **variability**
- Any **outliers** (unusually large or small values)

Unlike a histogram, a box plot does not show the exact frequency distribution, but it is excellent for comparing multiple datasets and detecting unusual observations.

Components of a Box Plot

A box plot is based on the **five-number summary**:

1. **Minimum** (smallest non-outlier value)
2. **First Quartile (Q1)** – 25th percentile
3. **Median (Q2)** – 50th percentile
4. **Third Quartile (Q3)** – 75th percentile
5. **Maximum** (largest non-outlier value)



Example

12, 15, 18, 20, 22, 25, 27, 30, 35, 60

Step1: sort the data

Step2: Find the 5 number summary

Step3: Calculate the Interquartile Range (IQR)

Step4: Determine the Outlier Range

Example

12, 15, 18, 20, 22, 25, 27, 30, 35, 60

Step1: sort the data

Step2: Find the 5 number summary

Step3: Calculate the Interquartile Range (IQR)

Step4: Determine the Outlier Range

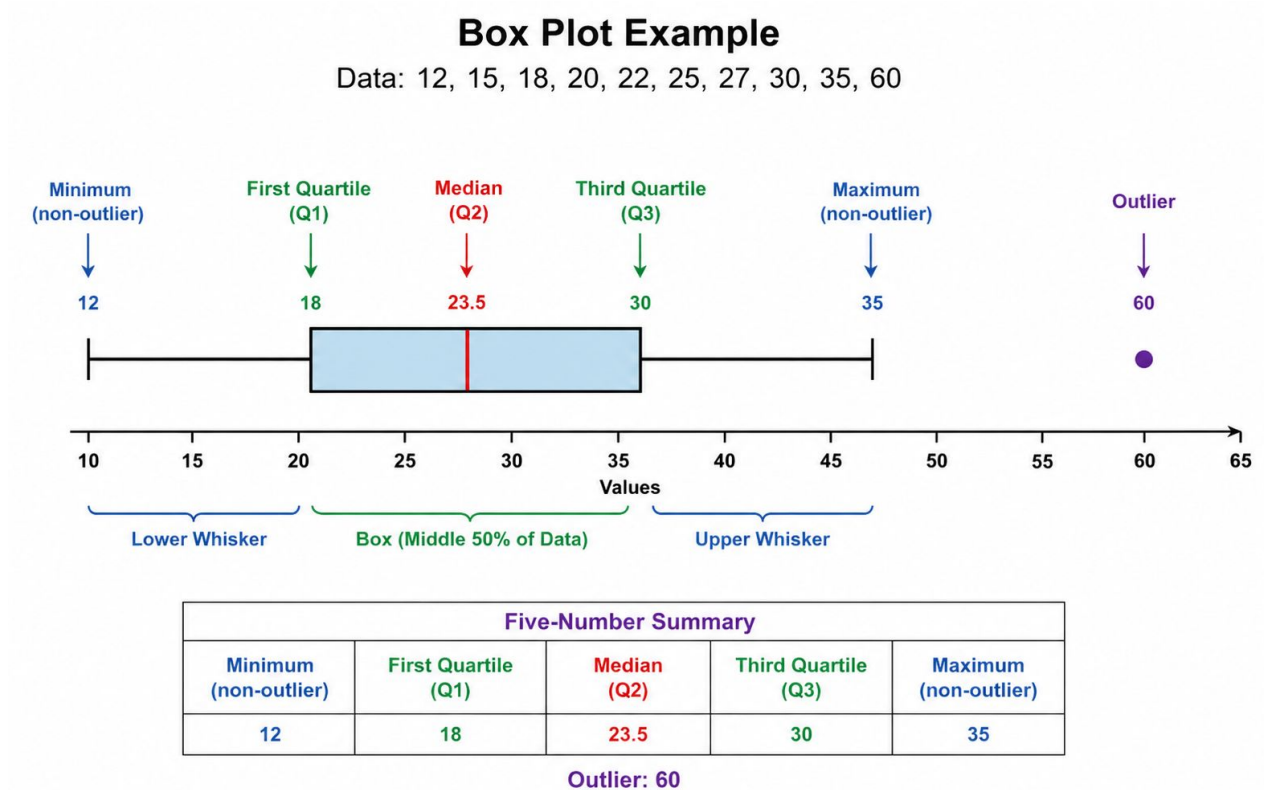
Q1: 18, Median: 23.5, Q3: 30

$$\text{IQR} = Q_3 - Q_1 = 30 - 18 = 12$$

$$Q_1 - 1.5 \times \text{IQR} = 18 - 18 = 0$$

$$Q_3 + 1.5 \times \text{IQR} = 30 + 18 = 48$$

Box Plot Example



Utilities of Box Plot

A box plot is especially valuable because it summarizes an entire dataset in a single figure while clearly showing the **center, spread, skewness, and outliers**.

box plots are widely used in manufacturing, finance, healthcare, and quality control—they summarize an entire distribution in one compact figure while clearly highlighting unusual observations.

A box plot is especially useful when you want to:

- Compare exam scores across different classes.
- Compare salaries across departments.
- Detect defective products in manufacturing.
- Compare machine performance.
- Identify unusually high or low values in a dataset.

Association

Association is a statistical or data mining technique used to discover **relationships or patterns between variables or items** in a dataset.

It answers questions such as:

- Which products are frequently bought together?
- Which symptoms often occur together?
- Which machine conditions commonly lead to a defect?

Quiz-5

Car	Assembly Time
1	8.5
2	8.8
3	8.1
4	8.3
5	8.4
6	7.8
7	8.6
8	8.0
9	8.8
10	11.5

1. Write 2 things that you liked/understood.
2. Write 2 things that you did not like/understand.
3. Show mean median mode in a diagram.
4. Draw the box plot
5. What are your observations ?
6. What are the follow-up questions ?
7. Why is outlier useful ?