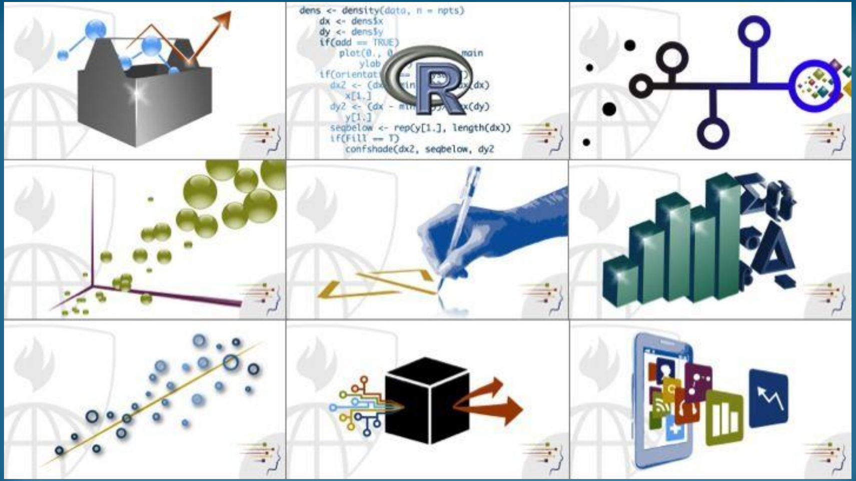
Descriptive Statistics

Understanding Descriptive Statistics

Descriptive statistics, a powerful set of techniques used to represent and understand data. Descriptive statistics help us make sense of data by summarizing, organizing, and presenting it in a meaningful way.

Understanding Descriptive Statistics

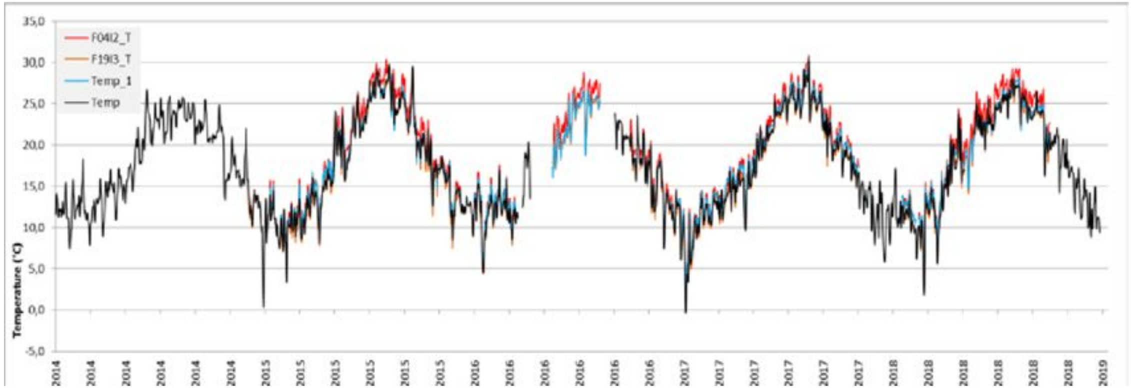
Descriptive statistics are like tools in our data toolbox. They help us explore and describe data without drawing conclusions or making predictions. Instead, we aim to present data in a way that makes it easier to comprehend.



```
dens <- density(data, n = npts)
    dx <- dens$x
    dy <- dens$y
    if(add == TRUE)
      plot(0., 0                main
        ylab
    if(orientat
      dx2 <- (dx    in              x(dx)
        x[1.]
      dy2 <- (dx - min              x(dy)
        y[1.]
      seqbelow <- rep(y[1.], length(dx))
      if(Fill == T)
        confshade(dx2, seqbelow, dy2
```

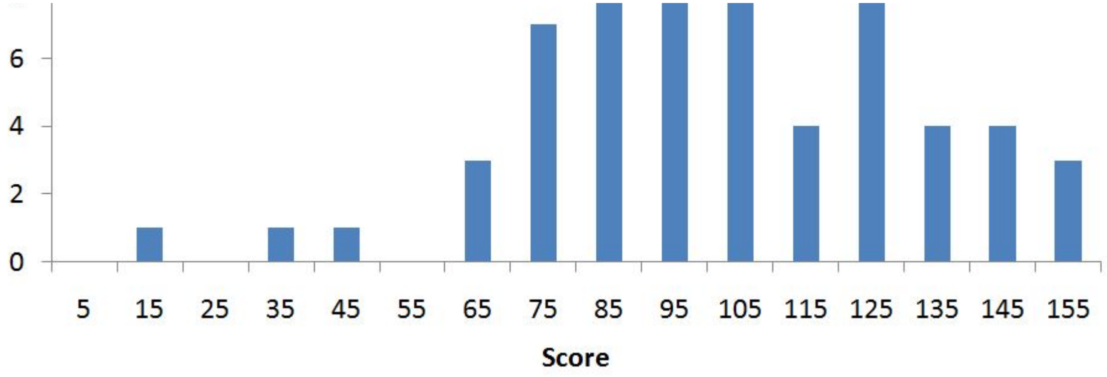
Example 1: Daily Temperature

Imagine you have collected daily temperature data for a month. Descriptive statistics can help you summarize this data by providing insights such as the average temperature, the most common temperature, and how much temperatures vary.



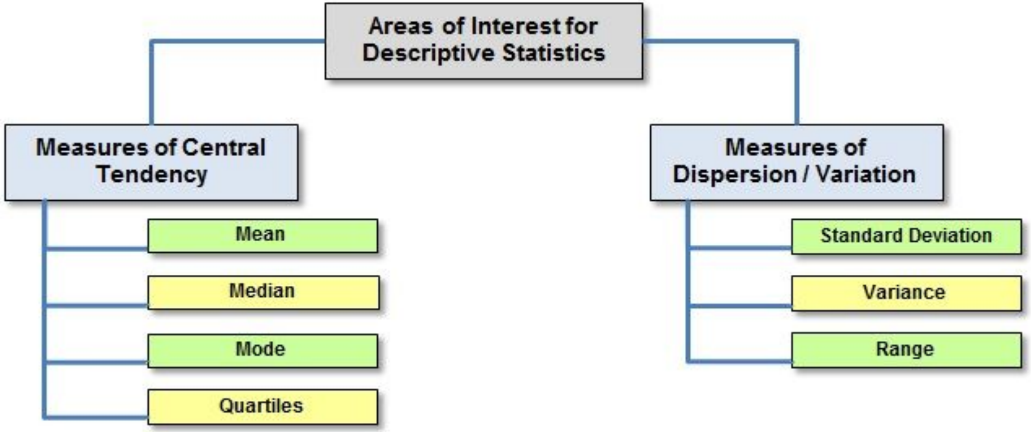
Example 2: Exam Scores

Suppose you have the exam scores of your classmates. Descriptive statistics can reveal the class's average score, the score that occurs most frequently, and how scores are dispersed across the class.



Key Descriptive Techniques

We'll explore various techniques, including **measures of central tendency** (like the mean, mode, and median), **measures of variability** (like the range and standard deviation), and ways to visualize data using graphs and charts.



Areas of Interest for Descriptive Statistics

Measures of Central Tendency

Mean

Median

Mode

Quartiles

Measures of Dispersion / Variation

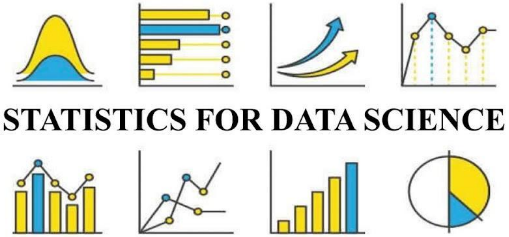
Standard Deviation

Variance

Range

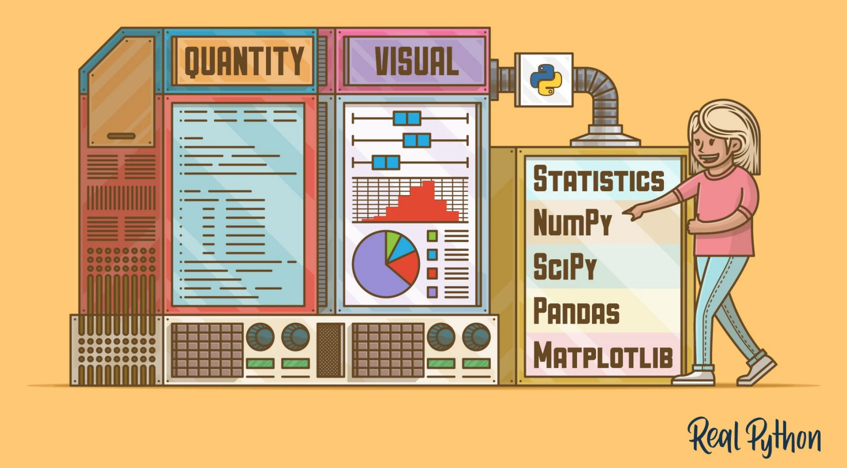
Why Descriptive Statistics Matter

Understanding data is crucial in many fields. For instance, businesses use it to track sales trends, scientists use it to analyze research results, and healthcare professionals use it to monitor patient health.



Conclusion

Descriptive statistics are the foundation of data analysis. They help us turn raw data into valuable insights. As we delve deeper into this topic, you'll discover how these techniques are applied and why they are essential.



THANK YOU

Measures of Central tendency

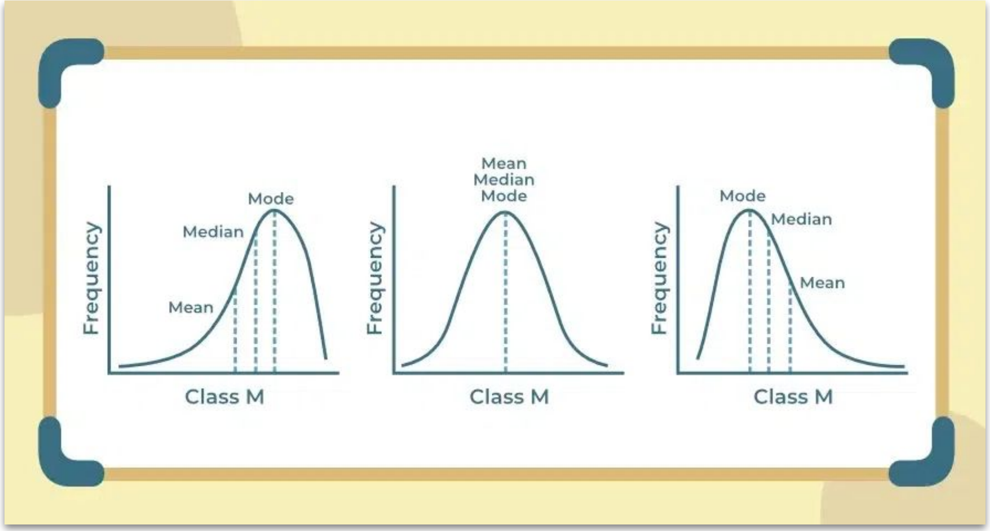
Mean, Median & Mode

Measures of Central Tendency in Statistics

- Central Tendencies in Statistics are the numerical values that are used to represent mid-value or central value a large collection of numerical data. These obtained numerical values are called central or average values in Statistics.
- The representative value of a data set, generally the central value or the most occurring value that gives a general idea of the whole data set is called the Measure of Central Tendency.

Measures of Central Tendency in Statistics

- Some of the most commonly used measures of central tendency are:
  - Mean
  - Median
  - Mode



Mean

- Mean in general terms is used for the arithmetic mean of the data, but other than the arithmetic mean there are geometric mean and harmonic mean as well that are calculated using different formulas.
- Mean for Ungrouped Data
  - Arithmetic mean (x̄) is defined as the sum of the individual observations (xi) divided by the total number of observations N. In other words, the mean is given by the sum of all observations divided by the total number of observations.

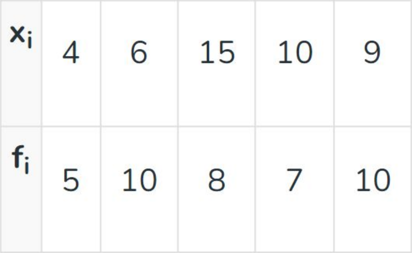
$$\bar{x} = \frac{\sum x_i}{N}$$

  - Mean = Sum of all Observations ÷ Total number of Observations

Mean

- Mean for Grouped Data
  - Mean (x̄) is defined for the grouped data as the sum of the product of observations (xi) and their corresponding frequencies (fi) divided by the sum of all the frequencies (fi).

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$



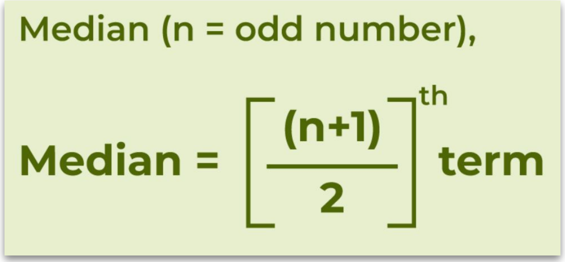
x_i	4	6	15	10	9
f_i	5	10	8	7	10

Disadvantage of Mean

- Although Mean is the most general way to calculate the central tendency of a dataset however it can not give the correct idea always, especially when there is a large gap between the datasets.

Median

- It is the middle value of the data set. It splits the data into two halves. If the number of elements in the data set is odd then the center element is the median and if it is even then the median would be the average of two central elements.
- Median Formula (When n is Odd) - If the number of values (n value) in the data set is odd then the formula to calculate the median is,

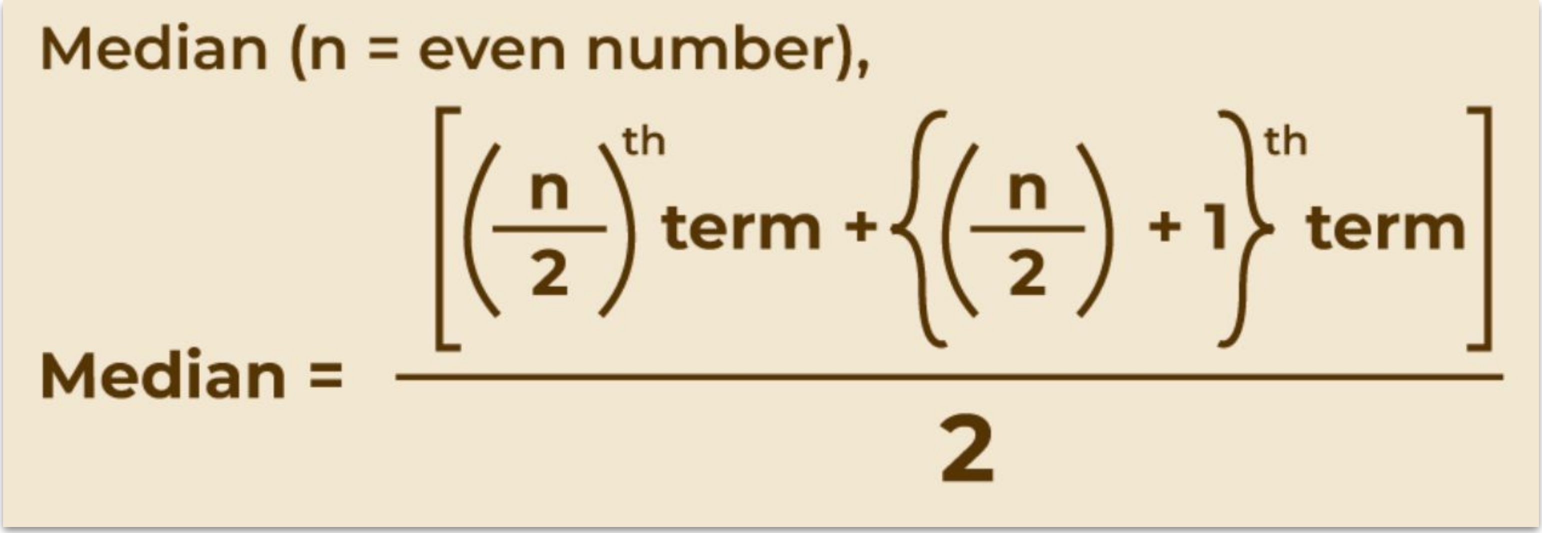


Median (n = odd number),

$$\text{Median} = \left[\frac{(n+1)}{2}\right]^{th} \text{term}$$

Median

- Median Formula (When n is Even) - If the number of values (n value) in the data set is even then the formula to calculate the median is:

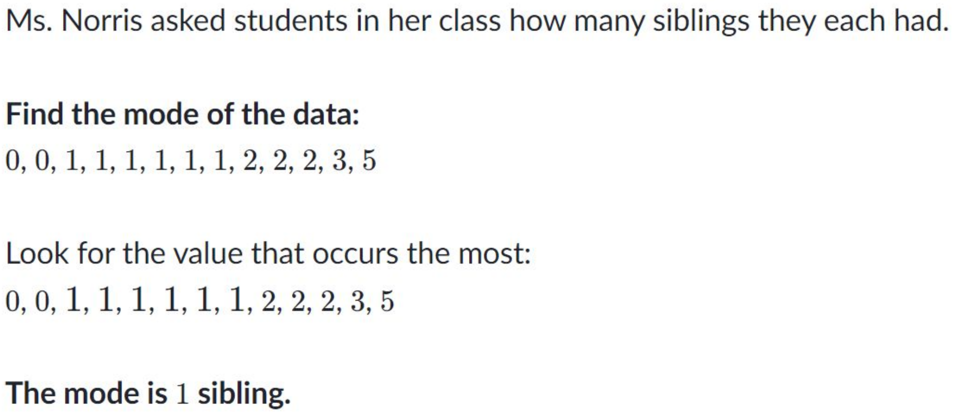


Median (n = even number),

$$\text{Median} = \frac{\left[\left(\frac{n}{2}\right)^{th}\text{term} + \left\{\left(\frac{n}{2}\right) + 1\right\}^{th}\text{term}\right]}{2}$$

Mode

- The mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.



Ms. Norris asked students in her class how many siblings they each had.

Find the mode of the data:
0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5

Look for the value that occurs the most:
0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5

The mode is 1 sibling.

THANK YOU

Measures of variability

Range, Standard Deviation

Range

The range of the data is given as the difference between the maximum and the minimum values of the observations in the data.

For example, let’s say we have data on the number of customers walking in the store in a week.

10, 14, 8, 10, 15, 4, 7

Minimum value in data = 4, Maximum Value in the data = 15

Range = Maximum Value in the data – Minimum value in the data = 15 – 4 = 11

Variance

Variance is a statistical measure that quantifies the spread or dispersion of data points in a dataset. It indicates how much individual data points deviate from the mean (average) of the dataset.

A high variance suggests that the data points are spread out widely, while a low variance indicates that the data points are closely clustered around the mean.

$$\text{Variance } (\sigma^2) = \Sigma(xi - \mu)^2 / N$$

Σ indicates summation (adding up all the values). xi represents each individual data point. μ represents the mean (average) of the dataset. N is the total number of data points.

Variance - Example 1: Exam Scores

Suppose you have the following exam scores for a class of students: 85, 90, 88, 92, 87

Step 1: Calculate the mean (μ): $(85 + 90 + 88 + 92 + 87) / 5 = 88.4$

Step 2: Calculate the variance (σ^2): $(85 - 88.4)^2 = 12.96$

$(90 - 88.4)^2 = 2.56$, $(88 - 88.4)^2 = 0.16$, $(92 - 88.4)^2 = 12.96$, $(87 - 88.4)^2 = 1.96$

Now, sum up these squared differences and divide by the total number of data points:

Variance = $(12.96 + 2.56 + 0.16 + 12.96 + 1.96) / 5 = 30.6 / 5 = 6.12$

So, the variance of the exam scores is 6.12.

Standard Deviation

Standard deviation is a statistical measure of the amount of variation or dispersion in a set of data points. It measures how individual data points differ from the mean (average) of the dataset.

A higher standard deviation indicates greater variability, while a lower standard deviation suggests that data points are closer to the mean.

Standard Deviation (σ) = $\sqrt{[\Sigma(xi - \mu)^2 / N]}$

- σ represents the standard deviation.
- $\sqrt{}$ denotes the square root.
- Σ indicates summation (adding up all the values).
- xi represents each individual data point.
- μ represents the mean (average) of the dataset.
- N is the total number of data points.

Standard Deviation - Example: Daily Stock Returns

Let's say you have collected daily returns for a particular stock over the last five days: -2%, 3%, -1%, 2%, -2%

Step 1: Calculate the mean (μ): (-2% + 3% - 1% + 2% - 2%) / 5 = 0%

Step 2: Calculate the squared differences from the mean:

- $(-2\% - 0\%)^2 = 4\%$
- $(3\% - 0\%)^2 = 9\%$
- $(-1\% - 0\%)^2 = 1\%$
- $(2\% - 0\%)^2 = 4\%$
- $(-2\% - 0\%)^2 = 4\%$

Step 3: Calculate the variance: Variance = $[\Sigma(xi - \mu)^2 / N]$ = (4% + 9% + 1% + 4% + 4%) / 5 = 22% / 5 = 4.4%

Step 4: Calculate the standard deviation: Standard Deviation (σ) = √(Variance) = √(4.4%) ≈ √0.044 ≈ 0.21 (rounded to two decimal places)

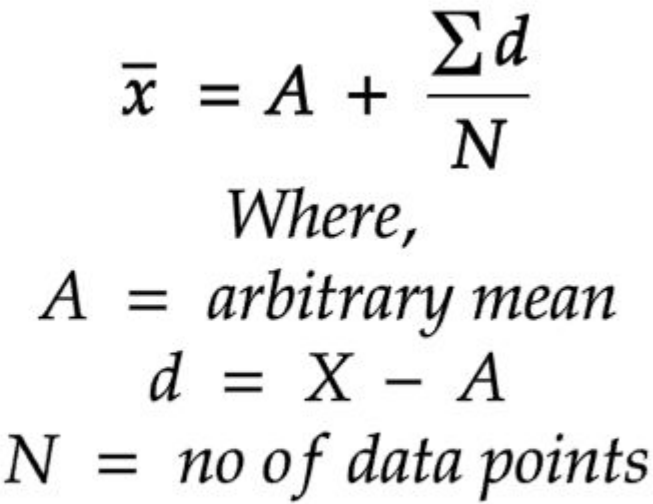
So, the standard deviation of the daily stock returns is approximately 0.21 (or 21%).

Mean

Shortcut method, Deviation

Shortcut method for mean

Ungrouped Data



$$\bar{x} = A + \frac{\sum d}{N}$$

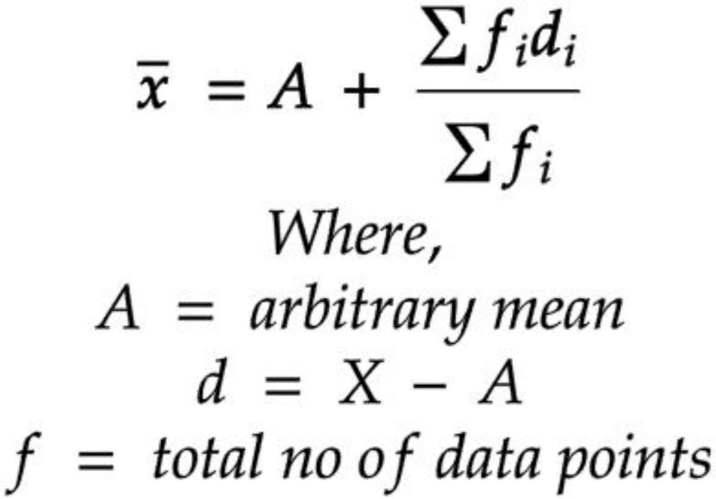
Where,

$A = \text{arbitrary mean}$

$d = X - A$

$N = \text{no of data points}$

Grouped Data



$$\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$$

Where,

$A = \text{arbitrary mean}$

$d = X - A$

$f = \text{total no of data points}$

Mean deviation

Mean deviation is also known as the mean absolute deviation (MAD).

It provides insight into the spread or dispersion of data values. To calculate the mean deviation, follow these steps:

1. Calculate the mean (average) of the data set.
2. Find the absolute difference between each data point and the mean.
3. Take the average of these absolute differences.

Mean Deviation About Median (MDAM)

MDAM measures the average absolute difference between each data point in a dataset and the median of that dataset.

It provides insight into how data points are scattered around the median. Here's how to calculate MDAM:

1. Calculate the median of the dataset.
2. Find the absolute difference between each data point and the median.
3. Take the average of these absolute differences.

Mean Deviation About Mode (MDaM)

MDaM, on the other hand, measures the average absolute difference between each data point in a dataset and the mode (the most frequently occurring value) of that dataset.

It provides insight into how data points are dispersed around the mode. Here's how to calculate MDaM:

1. Calculate the mode of the dataset.
2. Find the absolute difference between each data point and the mode.
3. Take the average of these absolute differences.

Example 1

Calculate the measures of central tendency (mean, median, mode) and variability (range, variance, and standard deviation) for that dataset.

Here's the dataset: [7, 12, 15, 18, 22, 30]

Measures of Central Tendency:

1. Mean (Average):
   Mean = (7 + 12 + 15 + 18 + 22 + 30) / 6 = 104 / 6 = 17.33 (rounded to two decimal places)

2. Median (Middle Value):
   Since the dataset is already in ascending order, the median is the middle value, which is 15.

3. Mode (Most Frequent Value):
   There is no mode in this dataset as all values occur only once.

Example 1 - Continued

Measures of Variability:

4. Range:

Range = Maximum Value - Minimum Value

Range = 30 - 7 = 23

5. Variance:

Variance measures the average squared difference of each data point from the mean.

Variance = [(7 - 17.33)^2 + (12 - 17.33)^2 + (15 - 17.33)^2 + (18 - 17.33)^2 + (22 - 17.33)^2 + (30 - 17.33)^2] / 6

Variance = (114.56 + 28.21 + 5.44 + 0.44 + 21.83 + 160.19) / 6

Variance = 330.67 / 6 = 55.11 (rounded to two decimal places)

6. Standard Deviation:

Standard Deviation is the square root of the variance.

Standard Deviation = √55.11 ≈ 7.42 (rounded to two decimal places)

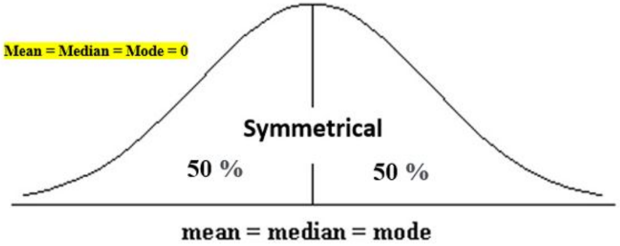
Skewness & Kurtosis

Skewness

Skewness essentially is a commonly used measure in descriptive statistics that characterizes the asymmetry of a data distribution.

It quantifies the extent to which the data is skewed or shifted to one side.

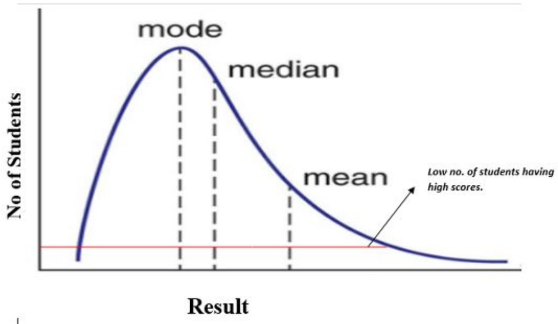
Positive skewness indicates a longer tail on the right side of the distribution, while negative skewness indicates a longer tail on the left side.



Types of skewness

Positive Skewed or Right-Skewed (Positive Skewness)

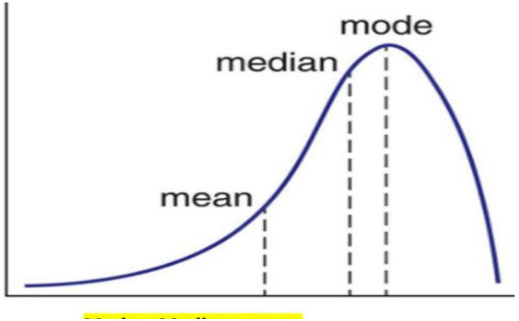
- In positively skewed, the mean of the data is greater than the median.
- The positively skewed or right-skewed distribution has a long right tail.
- This means that the majority of the data points are concentrated on the left side, with a few very large values on the right side.



Mean > Median > Mode

Negative Skewed or Left-Skewed (Negative Skewness)

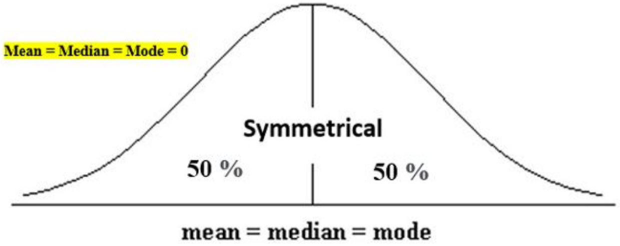
- In negatively skewed, the mean of the data is less than the median.
- The negatively skewed or left-skewed distribution has a long left tail.
- This indicates that the majority of the data points are clustered on the right side, with a few very small values on the left side.



Zero Skewness (Symmetric)

Zero skewness indicates that the data is symmetrically distributed

Both tails having the same length and balance around the mean.



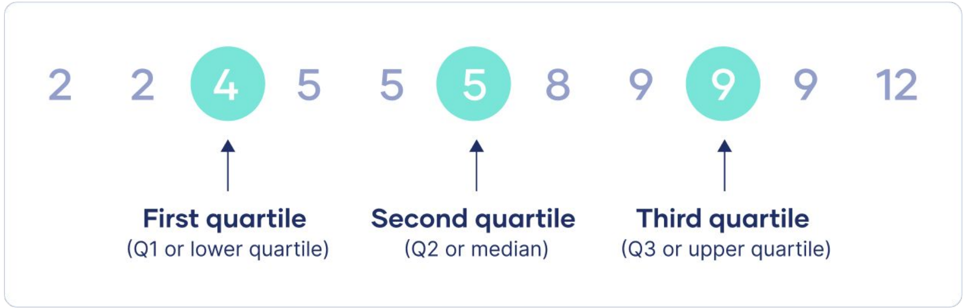
Karl pearson's coefficient of skewness

$$Skewness\ (S_k) = \frac{3 \times (\,Mean - Median\,)}{Standard\ Deviation}$$

Bowley's coefficient of skewness

$$Skewness\ (Sk_B) = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$$

Q_1 = First Quartile = Median of first n/2 data points
Q_2 = Second Quartile = Median of data points
Q_3 = Third Quartile = Median of last n/2 data points



2 2 4 5 5 5 8 9 9 9 12

First quartile
(Q1 or lower quartile)

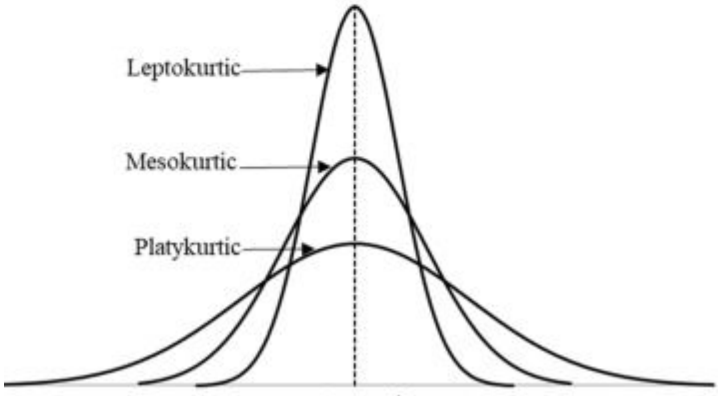
Second quartile
(Q2 or median)

Third quartile
(Q3 or upper quartile)

Kurtosis

Kurtosis

- It provides insight into whether the data is more or less peaked (leptokurtic) or flatter (platykurtic) than a normal distribution.
- Kurtosis can help identify the presence of outliers or extreme values in a dataset.



Types of Kurtosis

Leptokurtic

- A distribution with positive kurtosis (kurtosis > 0) is leptokurtic.
- Leptokurtic distributions have a more peaked central peak and heavier tails compared to a normal distribution.

Mesokurtic

- A distribution with mesokurtic kurtosis has a kurtosis value of 0.
- This indicates that the distribution has a similar shape to a standard normal distribution (bell-shaped curve) with neither a very peaked nor a very flat shape.

Platykurtic

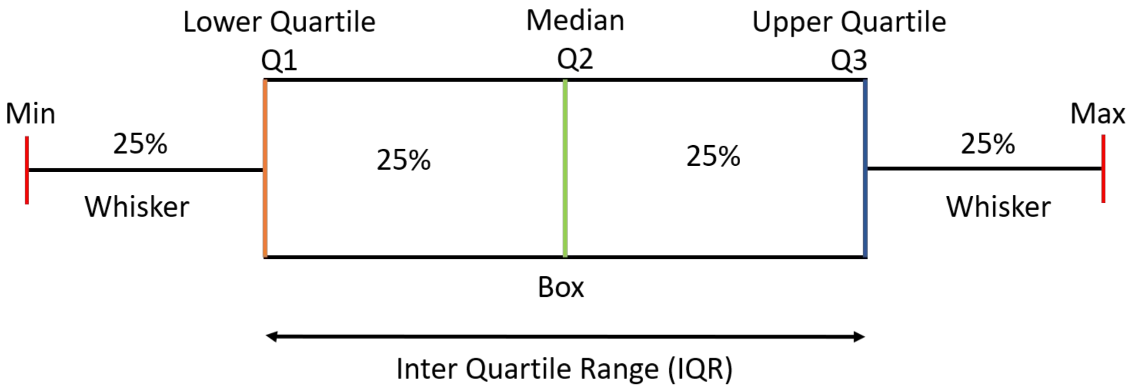
- A distribution with negative kurtosis (kurtosis < 0) is platykurtic.
- Platykurtic distributions have a flatter central peak and lighter tails compared to a normal distribution.

Box Plot

Box Plot

It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

It provides a visual way to display the central tendency, spread, and skewness of the data, as well as the presence of outliers. Box plots are particularly useful for comparing multiple datasets or visualizing the characteristics of a single dataset.



Box Plot

Median: The median is the middle value of the dataset, which divides the given dataset into two equal parts. The median is considered as the second quartile.Third Quartile (Q3): The third quartile is the median of the upper half of the data.

Maximum: The maximum value in the given dataset.

Interquartile Range (IQR): The difference between the third quartile and first quartile is known as the interquartile range. (i.e.) IQR = Q3-Q1

Outlier: The data that falls on the far left or right side of the ordered data is tested to be the outliers. Generally, the outliers fall more than the specified distance from the first and third quartile. (i.e.) Outliers are greater than Q3+(1.5 . IQR) or less than Q1-(1.5 . IQR).

Positively Skewed: If the distance from the median to the maximum is greater than the distance from the median to the minimum, then the box plot is positively skewed.

Negatively Skewed: If the distance from the median to minimum is greater than the distance from the median to the maximum, then the box plot is negatively skewed.

Symmetric: The box plot is said to be symmetric if the median is equidistant from the maximum and minimum values.

Pivot Tables

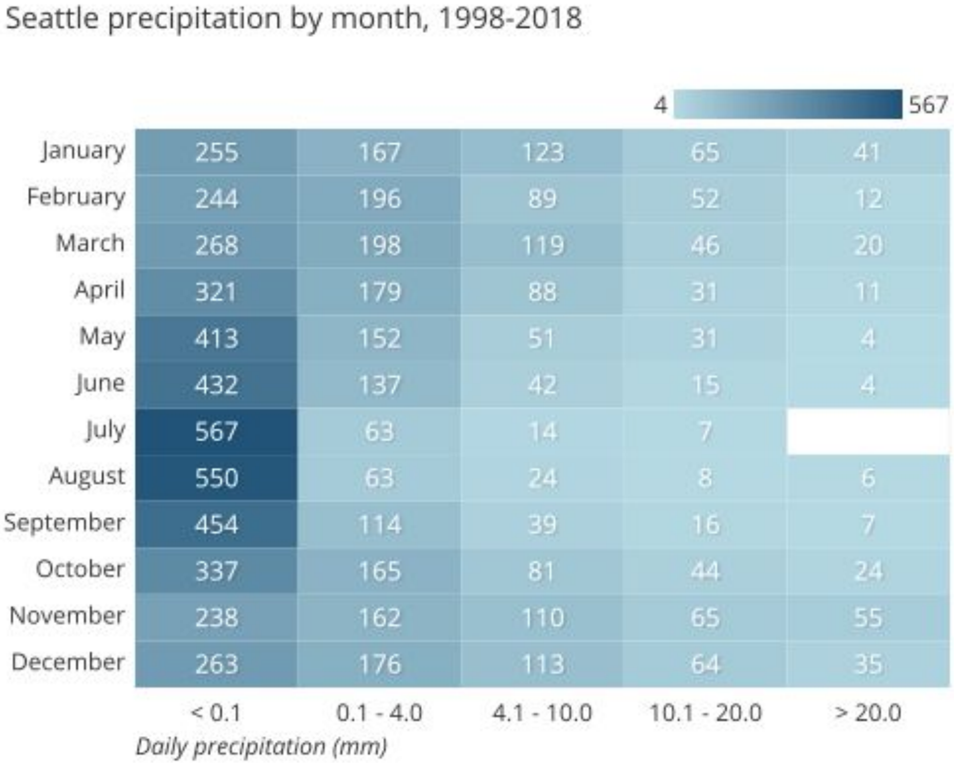
Pivot Tables

- A pivot table is a powerful data summarization tool that can automatically sort, count, and sum up data stored in tables and display the summarized data. Pivot tables are useful to quickly create crosstabs (a process or function that combines and/or summarizes data from one or more sources into a concise format for analysis or reporting) to display the joint distribution of two or more variables.
- Three key reasons for organizing data into a pivot table are:
  - To summarize the data contained in a lengthy list into a compact format.
  - To find relationships within the data those are otherwise hard to see because of the amount of detail.
  - To organize the data into a format that‘s easy to read.

Heatmaps

Heatmap

- A heatmap in descriptive statistics is a graphical representation of a dataset that uses color-coding to display the values of a matrix or a table. It is particularly useful for visualizing relationships, patterns, or trends in data by assigning colors to different data points or cells, making it easier to identify variations and correlations in the data.



	< 0.1	0.1 - 4.0	4.1 - 10.0	10.1 - 20.0	> 20.0
January	255	167	123	65	41
February	244	196	89	52	12
March	268	198	119	46	20
April	321	179	88	31	11
May	413	152	51	31	4
June	432	137	42	15	4
July	567	63	14	7	
August	550	63	24	8	6
September	454	114	39	16	7
October	337	165	81	44	24
November	238	162	110	65	55
December	263	176	113	64	35

Heatmap example

ANOVA

Analysis of Variance

ANOVA

- Analysis of variance, or ANOVA, is a statistical method that separates observed variance data into different components to use for additional tests.
- A one-way ANOVA is used for three or more groups of data, to gain information about the relationship between the dependent and independent variables.

Step 1: Set Hypotheses

1. Null Hypothesis (H0): There are no significant differences in test scores among the three courses (μA = μB = μC).
2. Alternative Hypothesis (Ha): There are significant differences in test scores among the three courses (at least one mean is different).

Step 2: Calculate the Means

- Mean score for Course A:
  - (85 + 90 + 88 + 92 + 86) / 5 = (441) / 5 = 88.2
- Mean score for Course B:
  - (78 + 82 + 80 + 79 + 81) / 5 = (400) / 5 = 80
- Mean score for Course C:
  - (91 + 94 + 90 + 92 + 93) / 5 = (460) / 5 = 92

Step 3: Calculate the Sum of Squares

- Calculate the sum of squares within groups (SSW), which measures the variability within each courses.
    - SSW for Course A = Σ(xi - μA)^2 = (85-88.2)^2 + (90-88.2)^2 + (88-88.2)^2 + (92-88.2)^2 + (86-88.2)^2 = 12.96 + 3.24 + 0.04 + 14.44 + 5.76 = 36.44
    - SSW for Course B = Σ(xi - μB)^2 = (78-80)^2 + (82-80)^2 + (80-80)^2 + (79-80)^2 + (81-80)^2 = 10
    - SSW for Course C = Σ(xi - μC)^2 = (91-92)^2 + (94-92)^2 + (90-92)^2 + (92-92)^2 + (93-92)^2 = 10
    - Total SSW = 36.44 + 10 + 10 = 56.44
- Calculate the sum of squares between groups (SSB), which measures the variability between the course means.
    - SSB = (nA * (μA - μ)^2) + (nB * (μB - μ)^2) + (nC * (μC - μ)^2)
    - nA, nB, and nC are the sample sizes for courses A, B, and C (all 5 in this case).
    - μ is the overall mean of all data points.
    - SSB = (5 * (88.2 - 88)^2) + (5 * (80 - 88)^2) + (5 * (92 - 88)^2) = 400.2

Step 4: Calculate the Degrees of Freedom:

- Degrees of Freedom (DF) within groups: (n - 1) * k = (5 - 1) * 3 = 4 * 3 = 12
- Degrees of Freedom between groups: k - 1 = 3 - 1 = 2

Step 5: Calculate the F-statistic:

- F-statistic = (SSB / (k - 1)) / (SSW / ((n - 1) * k))
- F-statistic = (400.2 / 2) / (56.44 / 12)
- F-statistic = 200.1 / 4.70333333
- F-statistic ≈ 42.57 (rounded to two decimal places)

Step 6: Determine the Critical Value

- For a significance level of α = 0.05 and degrees of freedom (2, 12), consult an F-table or use statistical software to find the critical value. The critical value is approximately 3.89.

Step 7: Compare the Critical Value and F-statistic

- The calculated F-statistic (42.57) is much larger than the critical value (3.89).
- Therefore, we reject the null hypothesis (H0). There are statistically significant differences in test scores among the three courses.

- **Note** :- If F-statistic is less than critical value then null hypothesis is accepted.