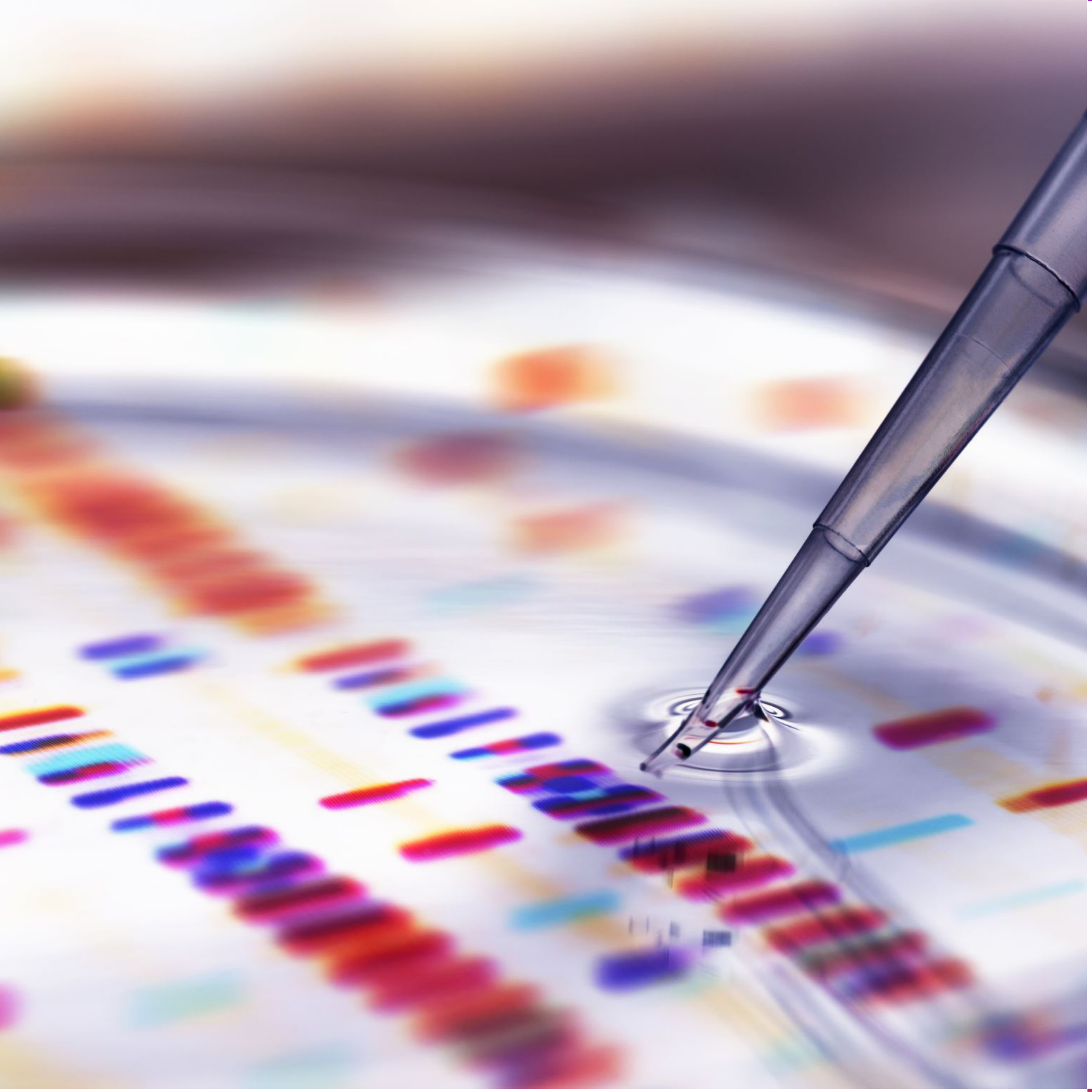


Module 2

2.2 Percentiles. Skewness and kurtosis.
2.3 Statistical analysis: Measures of central tendency (mean, median, mode). Measures of dispersion (range, variance, standard deviation). Quartiles, percentiles. Stem and leaf plots, Box plots

Dr. Bhakti Palkar

Statistical analysis

- Simplest form of data analysis as during this we use just one variable to research the info.
- The standard goal is to know the underlying sample distribution/ data and make observations about the population. Outlier detection is additionally part of the analysis.
- The characteristics of population distribution include:
  1. **Central Tendency**
  2. **Spread**
  3. **Skewness and kurtosis**

Statistical analysis

Central tendency:

The central tendency or location of distribution has got to do with typical or middle values.

The commonly useful measures of central tendency are statistics called mean, median, and sometimes mode during which the foremost common is mean.

For skewed distribution or when there’s concern about outliers, the median may be preferred.

Statistical analysis

Spread:

The variance and standard deviation are two useful measures of spread.

The variance is the mean of the squares of the individual deviations. The standard deviation is the square root of the variance.

For Normally distributed data, approximately 95% of the values lie within 2 sd of the mean..

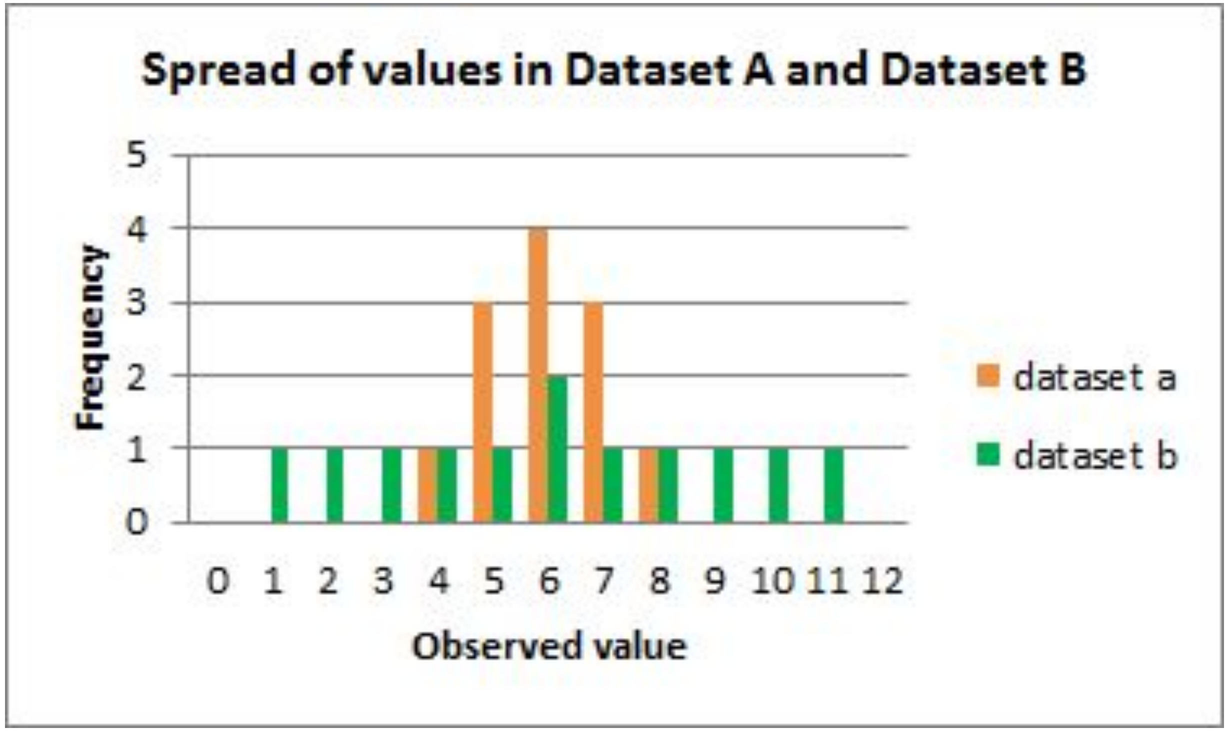
The interquartile range (IQR) is a robust measure of spread.

For example:

If we look at the spread of the values in the graph, we can see that Dataset B is more dispersed than Dataset A. Used together, the measures of central tendency and measures of spread help us to better understand the data.

Dataset A	Dataset B
4, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7, 8	1, 2, 3, 4, 5, 6, 6, 7, 8, 9, 10, 11

Mode (most frequent value):6
Median (middle value*):6
Mean (arithmetic average) :6



Range of values for Dataset B is larger than Dataset A

Statistical analysis

- **Quartiles:**
- Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point *between* the quarters. A dataset may also be divided into quintiles (five equal parts) or deciles (ten equal parts).

Quartiles						
25% of values	Q1	25% of values	Q2	25% of values	Q3	25% of values

The **lower quartile (Q1)** is the point between the lowest 25% of values and the highest 75% of values. It is also called the **25th percentile**.

The **second quartile (Q2)** is the middle of the data set. It is also called the **50th percentile**, or the **median**.

The **upper quartile (Q3)** is the point between the lowest 75% and highest 25% of values. It is also called the **75th percentile**.

Calculating Quartiles and IQR

The interquartile range (IQR) is the difference between the upper (Q3) and lower (Q1) quartiles. The IQR is often seen as a better measure of spread than the range as it is not affected by outliers.

Dataset A														
4	5	5	Q1	5	6	6	Q2	6	6	7	Q3	7	7	8

As the quartile point falls between two values, the mean (average) of those values is the quartile value:

Q1 = (5+5) / 2 = 5

Q2 = (6+6) / 2 = 6

Q3 = (7+7) / 2 = 7

IQR = Q3 – Q1= 7 - 5= 2

Dataset B														
1	2	3	Q1	4	5	6	Q2	6	7	8	Q3	9	10	11

As the quartile point falls between two values, the mean (average) of those values is the quartile value:

Q1 = (3+4) / 2 = 3.5

Q2 = (6+6) / 2 = 6

Q3 = (8+9) / 2 = 8.5

IQR = Q3 - Q1= 8.5 - 3.5= 5

Calculating the Population Variance σ^2 and Standard Deviation σ

Dataset A

Calculate the population mean (μ) of Dataset A.
(4 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 8) / 12
mean (μ) = 6

Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset

$X_i - \mu$ = -2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2

Square each individual deviation value
$(X_i - \mu)^2$ = 4, 1, 1, 1, 0, 0, 0, 0, 1,1,1, 4

Calculate the mean of the squared deviation values =

$$\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

(4 + 1 +1 +1 + 0 + 0 + 0 + 0 +1 +1 +1 + 4) / 12

Variance σ^2= 1.17
Calculate the square root of the variance
Standard deviation σ = 1.08

Dataset B

Calculate the population mean (μ) of Dataset B.
(1 + 2 + 3 + 4 + 5 + 6 + 6 + 7 + 8 + 9 + 10 + 11) / 12
mean (μ) = 6

Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset

$X_i - \mu$ = -5, -4, -3, -2, -1, 0, 0, 1, 2, 3, 4, 5,

Square each individual deviation value
$(X_i - \mu)^2$ = 25, 16, 9, 4, 1, 0, 0, 1, 4, 9, 16, 25

Calculate the mean of the squared deviation values =

$$\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$$

= (25 + 16 + 9 + 4 + 1 + 0 + 0 + 1 + 4 + 9 + 16 + 25) / 12

Variance σ^2 = 9.17
Calculate the square root of the variance
Standard deviation σ = 3.03

The larger Variance and Standard Deviation in Dataset B further demonstrates that Dataset B is more dispersed than Dataset A.

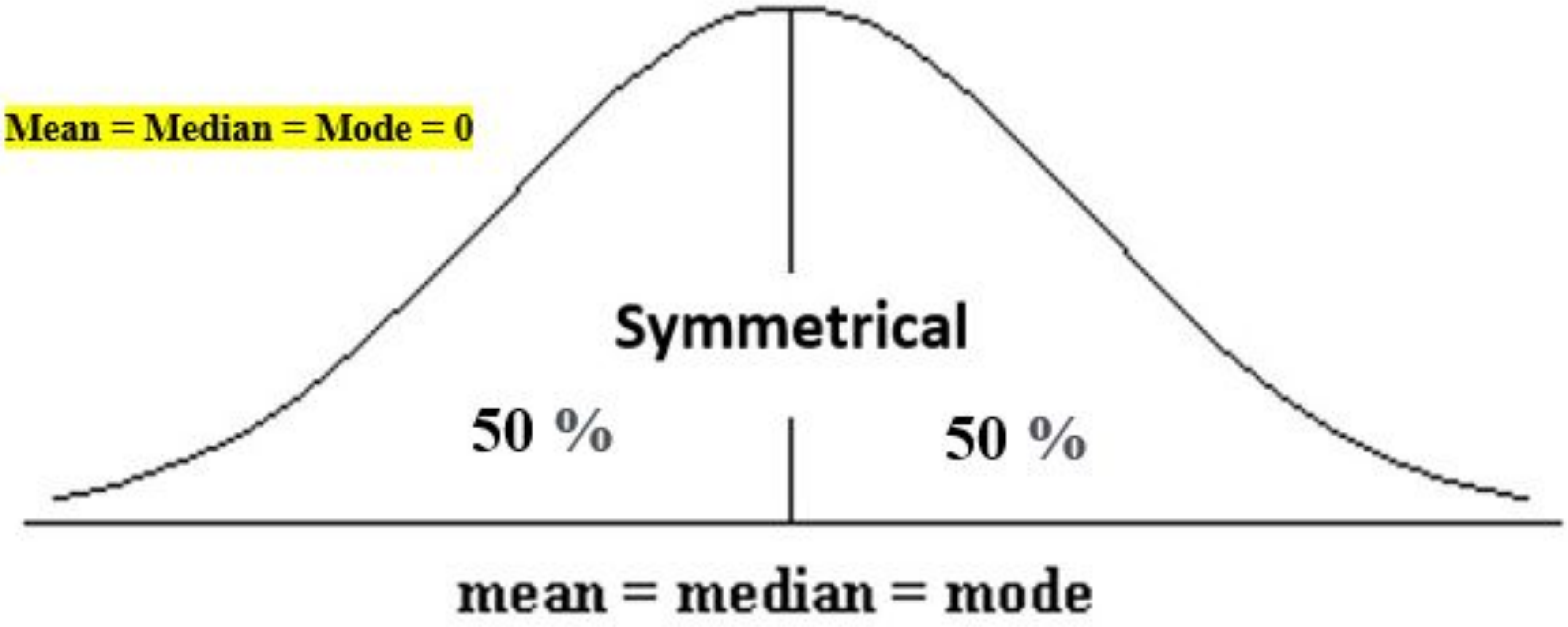
Statistical analysis.

Skewness and kurtosis:

“Skewness essentially measures the symmetry of the distribution, while kurtosis determines the heaviness of the distribution tails.”

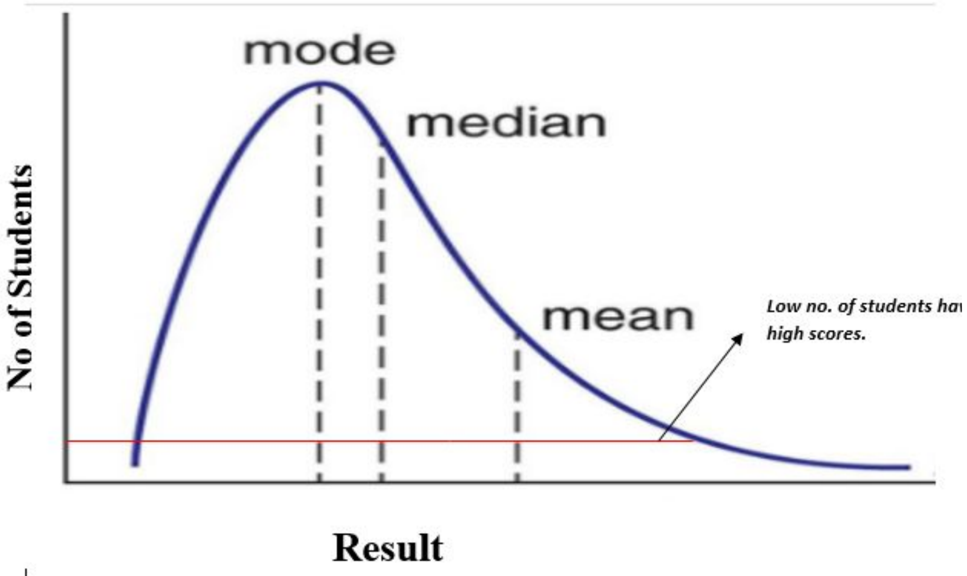
Skewness

- If the values of a specific independent variable (feature) are skewed, depending on the model, skewness may violate model assumptions or may reduce the interpretation of feature importance.
- *In statistics, skewness is a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal distribution (bell curve) in a given set of data.*


Mean = Median = Mode = 0
Symmetrical
50 %
50 %
mean = median = mode

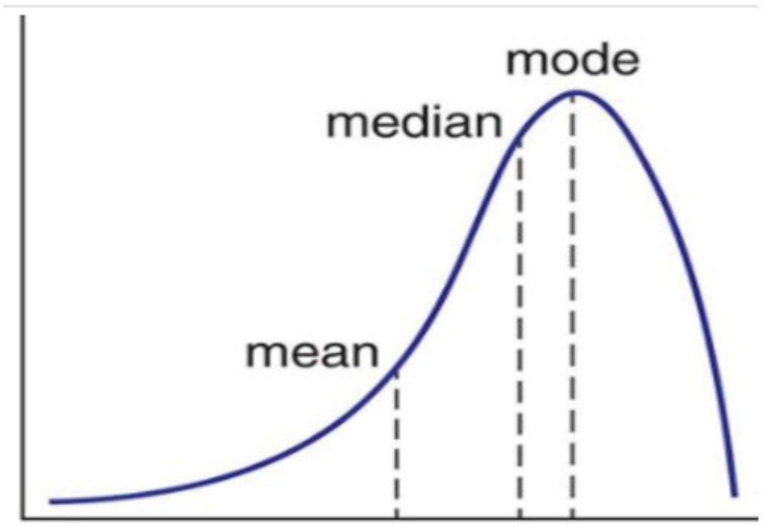
Types of skewness

1. Positive skewed or right-skewed



Mean > Median > Mode

2. Negative skewed or left-skewed



Mode > Median > mean

Skewness coefficient

- ***Pearson’s first coefficient of skewness***

$$\textbf{Pearson's first coefficient} = \frac{\textbf{Mean} - \textbf{Mode}}{\textbf{Standard Deviation}}$$

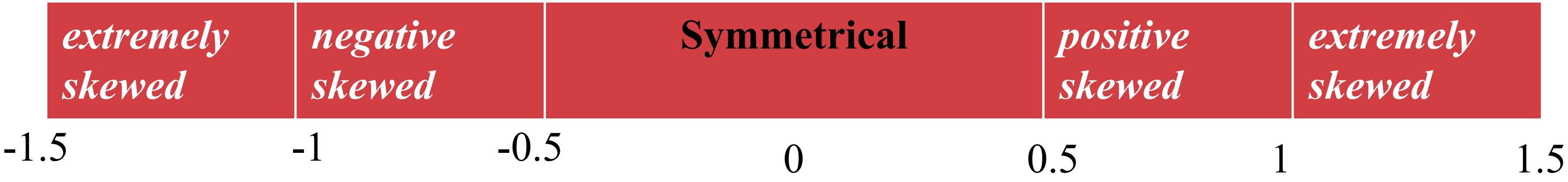
- Range:[-1 to +1]
- 0 indicates no linear relationship
- Preferred if the data has high mode.
- If the data have low mode or various modes, Pearson’s first coefficient is not preferred, and Pearson’s second coefficient may be superior, as it does not rely on the mode.

Skewness coefficient

- *Pearson's second coefficient of skewness*

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

$$\text{Mean} - \text{Mode} \approx 3\,(\text{Mean} - \text{Median})$$



extremely skewed | *negative skewed* | Symmetrical | *positive skewed* | *extremely skewed*

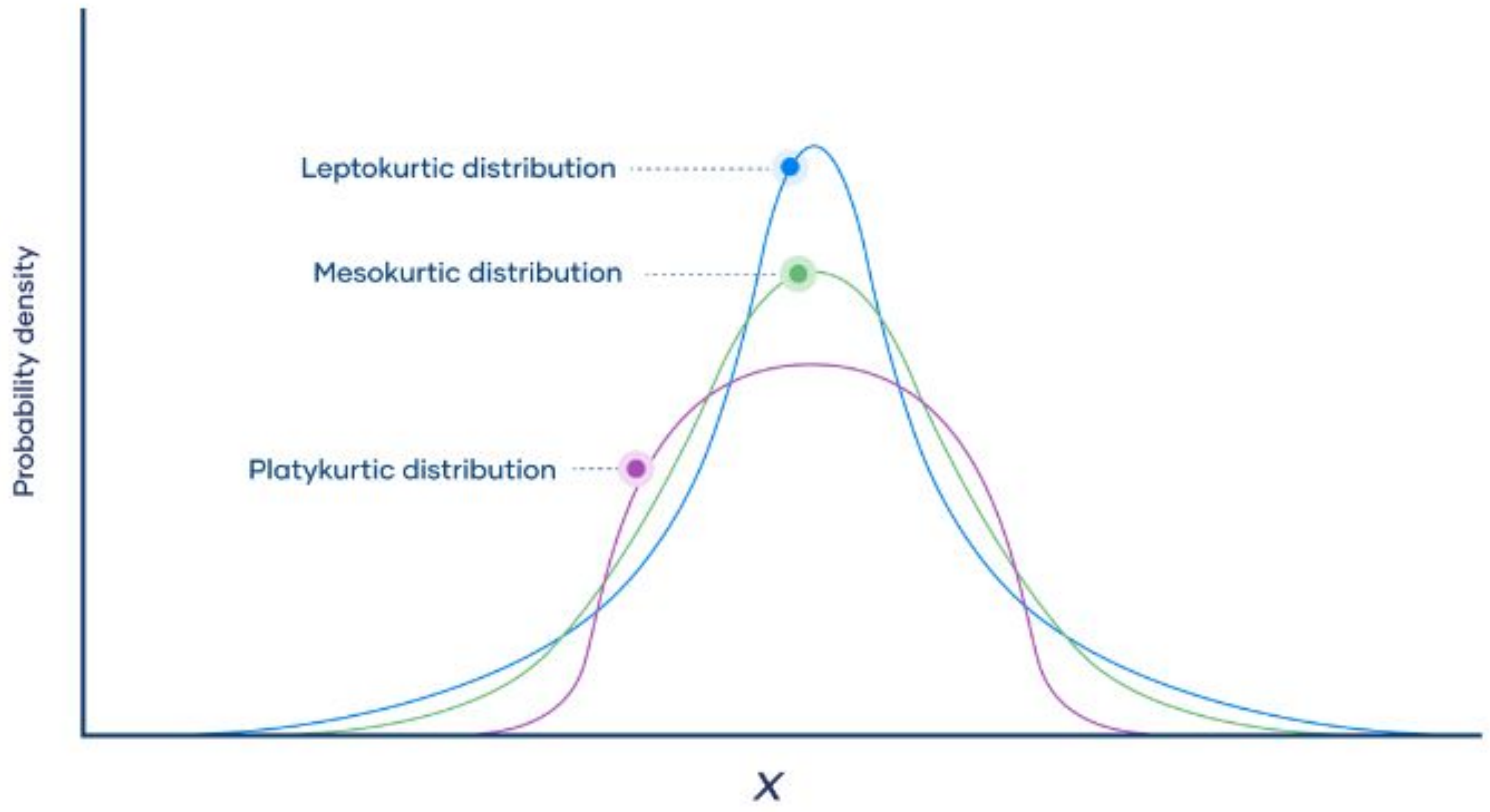
-1.5 | -1 | -0.5 | 0 | 0.5 | 1 | 1.5

Kurtosis

- Kurtosis refers to measuring the degree to which a given distribution is more or less ‘peaked’ relative to the normal distribution.
- The concept of kurtosis is very useful in decision-making. In this regard, we have 3 categories of distributions:

In this regard, we have 3 categories of distributions:

- Leptokurtic
- Mesokurtic
- Platykurtic


Leptokurtic distribution
Mesokurtic distribution
Platykurtic distribution
Probability density
X

Leptokurtic

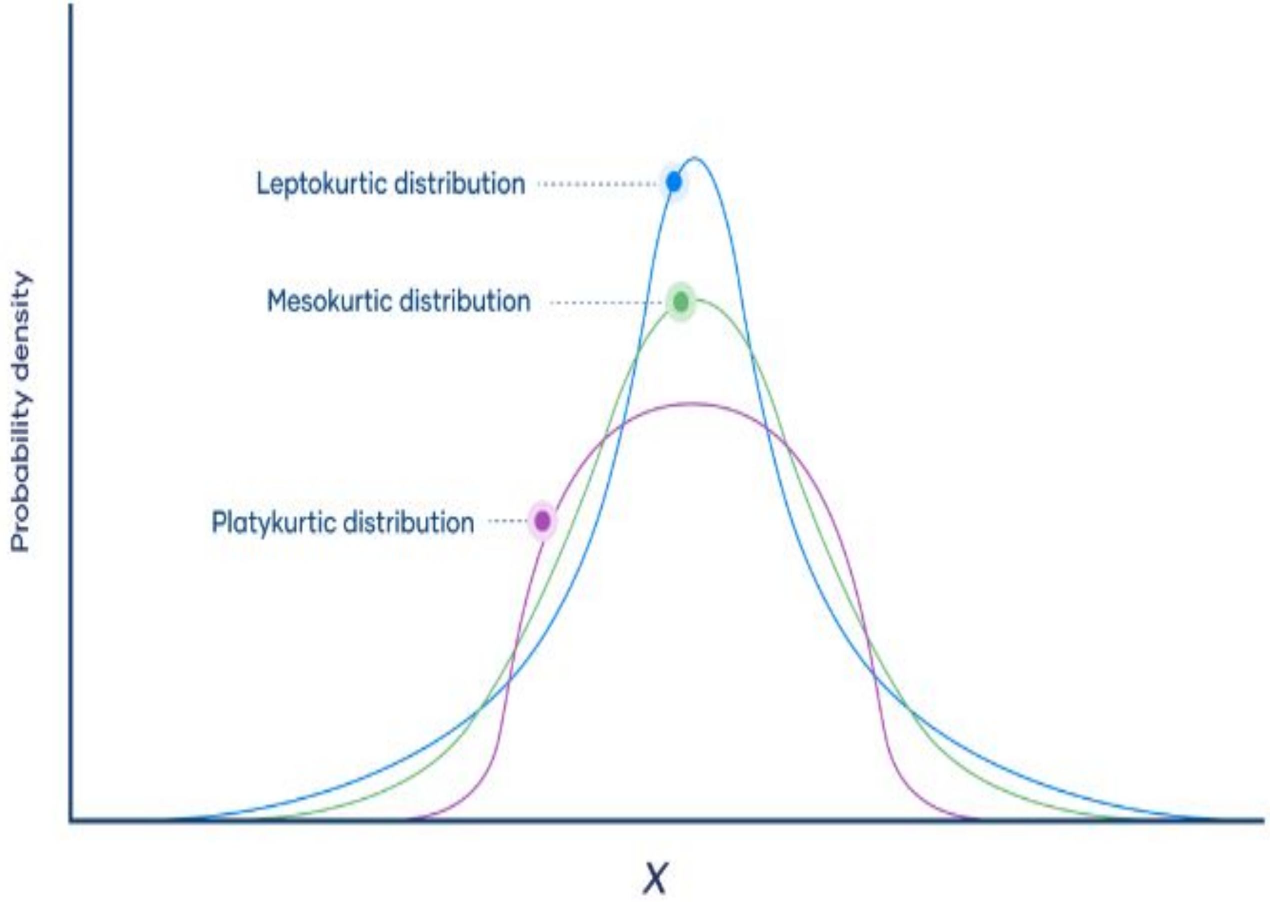
A leptokurtic distribution is more peaked than the normal distribution.

The higher peak results from the clustering of data points along the X-axis.

“Lepto” means thin.

The tails are also fatter than those of a normal distribution.

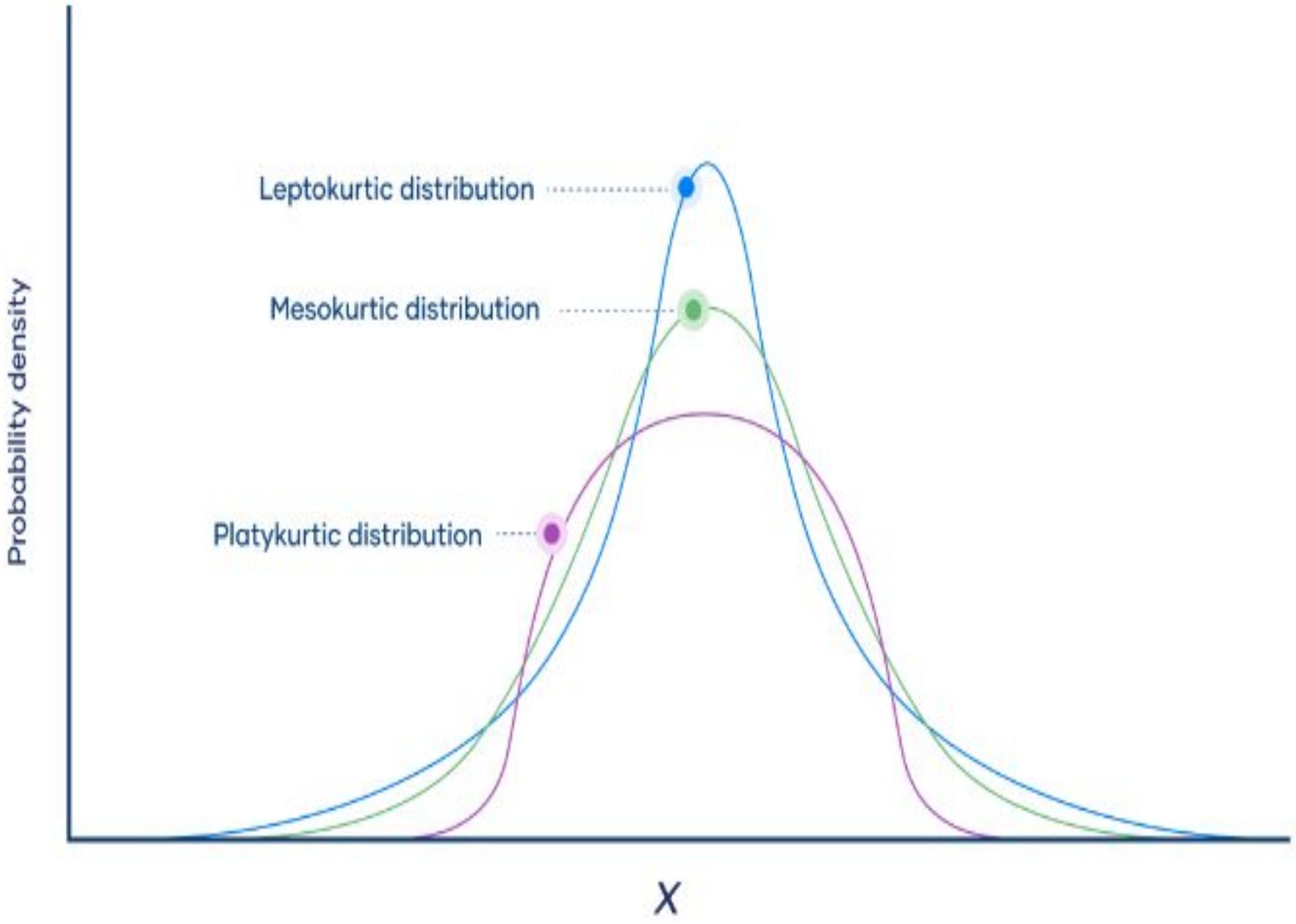
The coefficient of kurtosis is usually found to be more than 3.



Interpretation

When analyzing historical returns, a leptokurtic distribution means that changes are less frequent since historical values are clustered around the mean.

However, there are also large fluctuations represented by the fat tails.



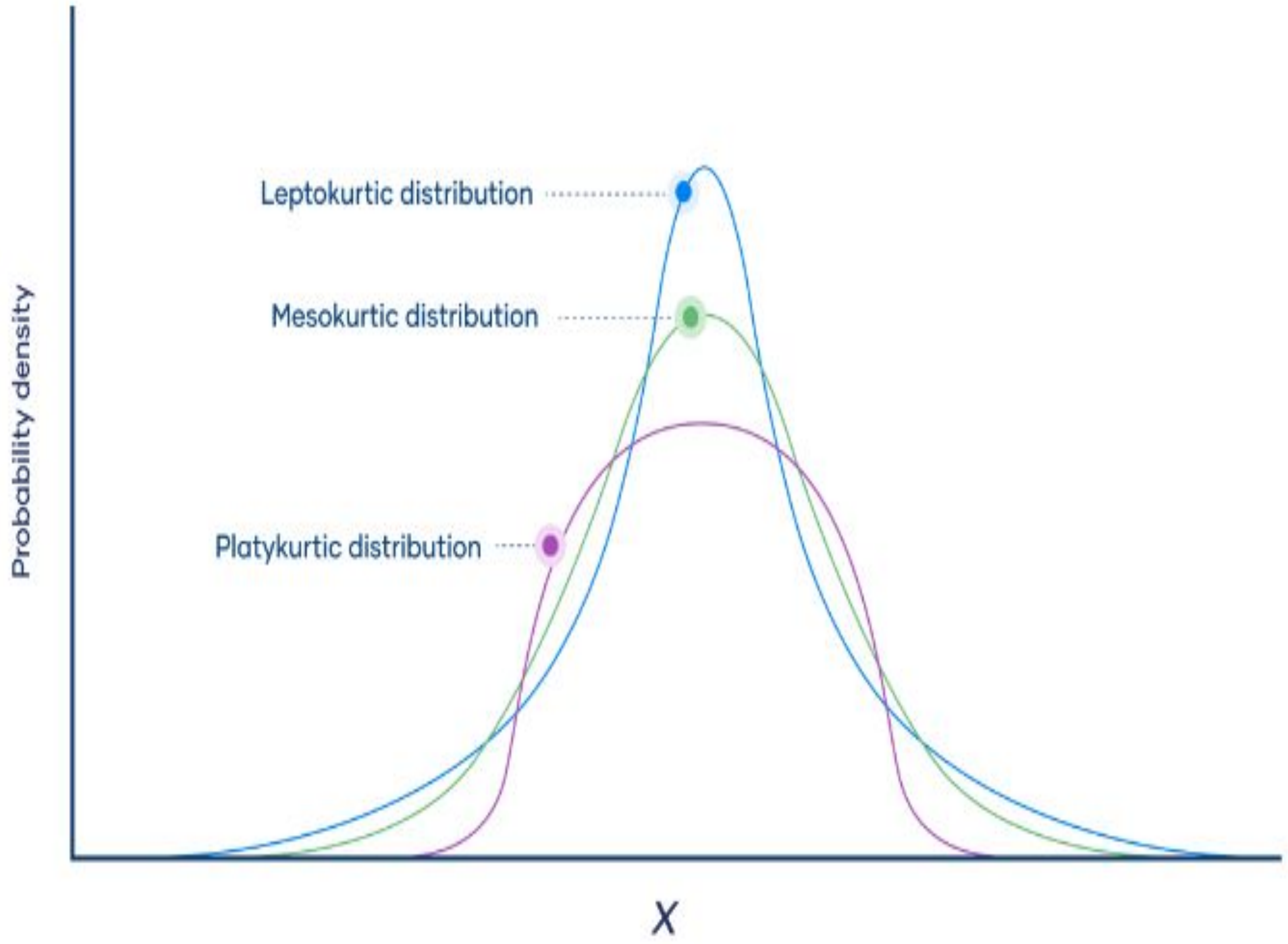
Platykurtic

A platykurtic distribution has extremely dispersed points along the X-axis, resulting in a lower peak when compared to the normal distribution.

“Platy” means broad.

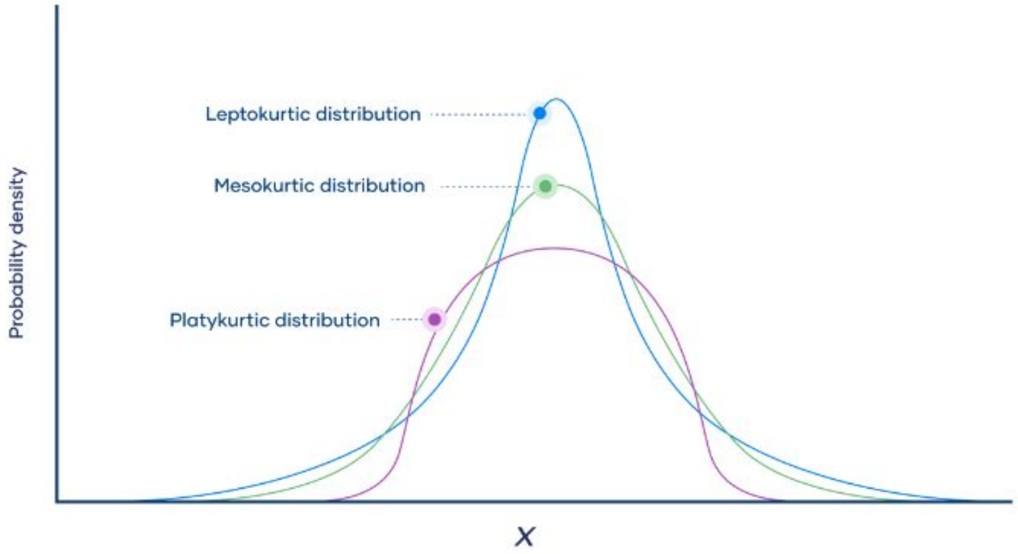
Hence, the prefix fits the distribution’s shape, which is wide and flat.

The points are less clustered around the mean compared to the leptokurtic distribution.



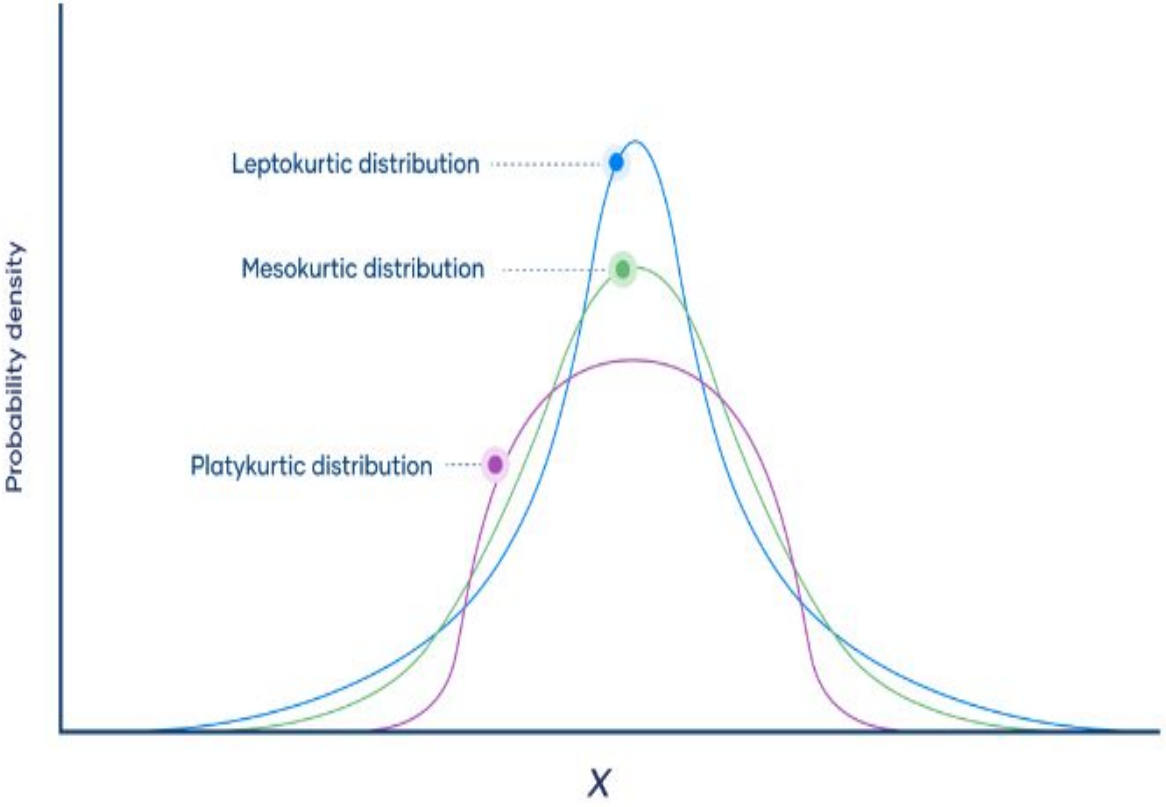
Platykurtic

- Returns that follow this type of distribution have fewer major fluctuations compared to leptokurtic returns.
- However, you should note that fluctuations represent the riskiness of an asset.
- More fluctuations represent more risk and vice versa.
- Therefore, platykurtic returns are less risky than leptokurtic returns.



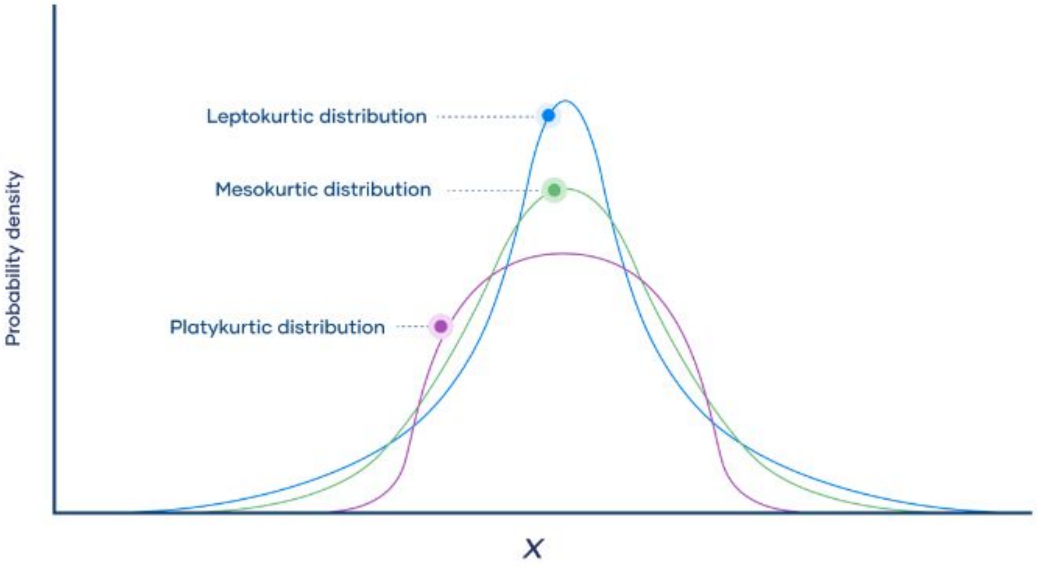
Mesokurtic

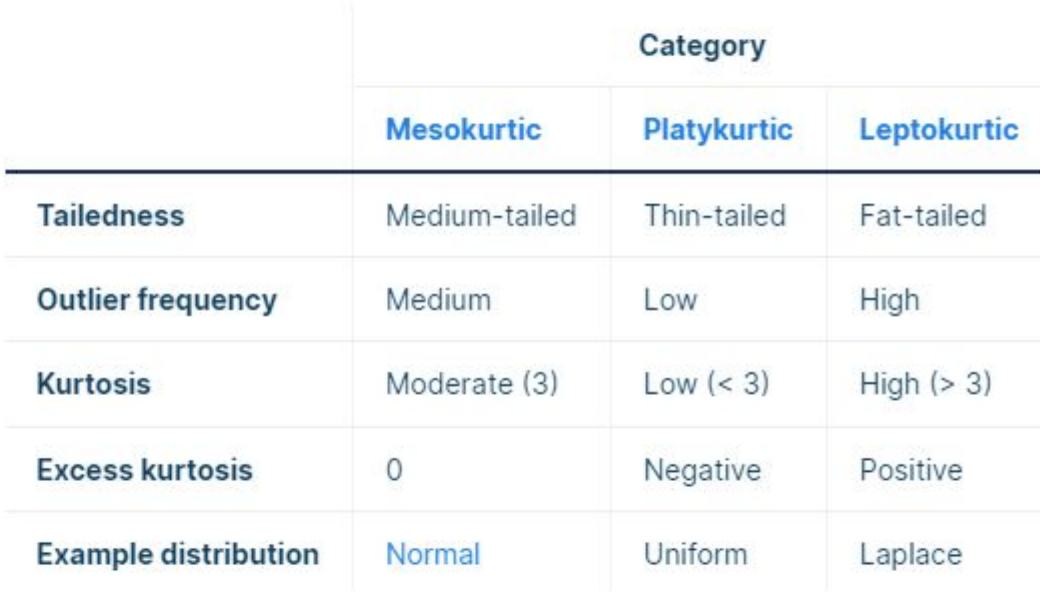
- Meso = middle; intermediate.
- This means the distribution is a normal distribution.



Kurtosis

- **Leptokurtic** or heavy-tailed distribution (kurtosis more than normal distribution)
- **Mesokurtic** (kurtosis same as the normal distribution)
- **Platykurtic** or short-tailed distribution (kurtosis less than normal distribution)

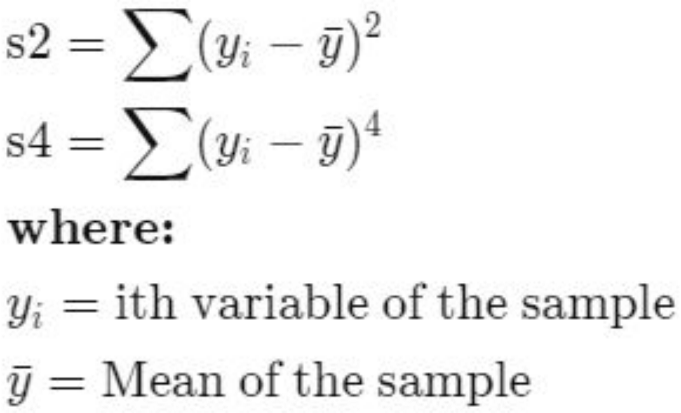




	Category		
	Mesokurtic	**Platykurtic**	**Leptokurtic**
Tailedness	Medium-tailed	Thin-tailed	Fat-tailed
Outlier frequency	Medium	Low	High
Kurtosis	Moderate (3)	Low (< 3)	High (> 3)
Excess kurtosis	0	Negative	Positive
Example distribution	Normal	Uniform	Laplace

Calculating Kurtosis

The new data points are 27, 13, 17, 57, 113, and 25.



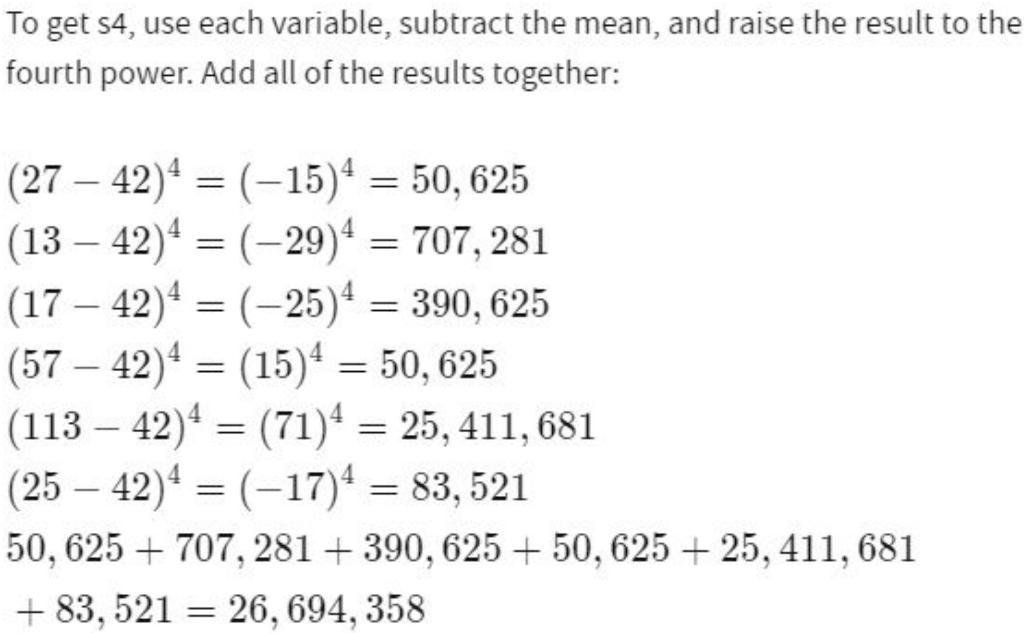
$$\text{s2} = \sum (y_i - \bar{y})^2$$

$$\text{s4} = \sum (y_i - \bar{y})^4$$

where:

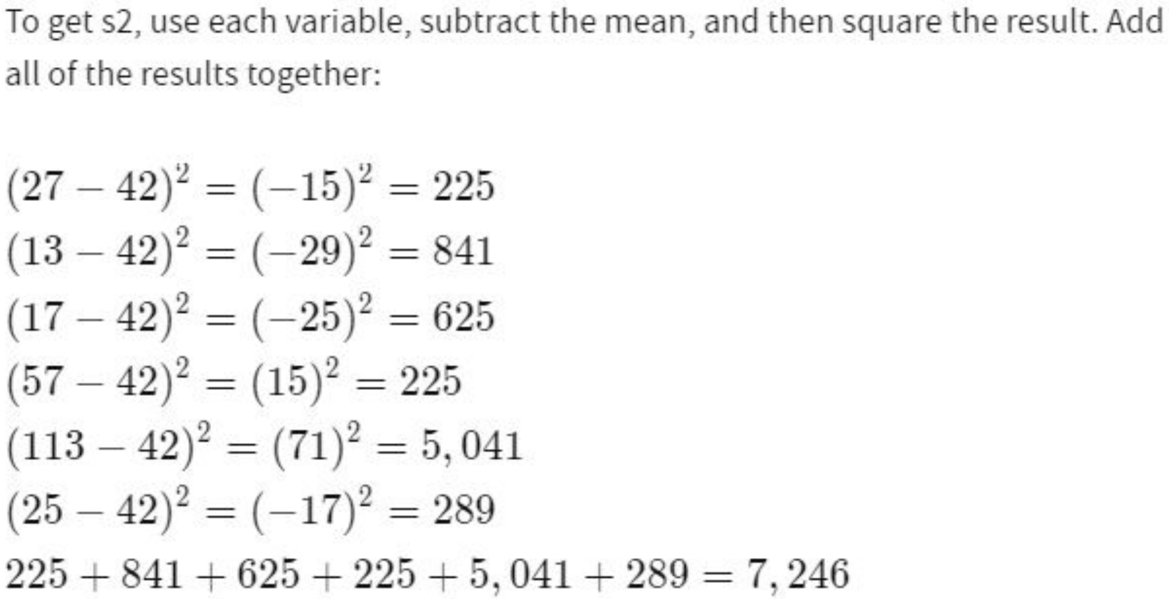
y_i = ith variable of the sample

$\bar{y}$ = Mean of the sample



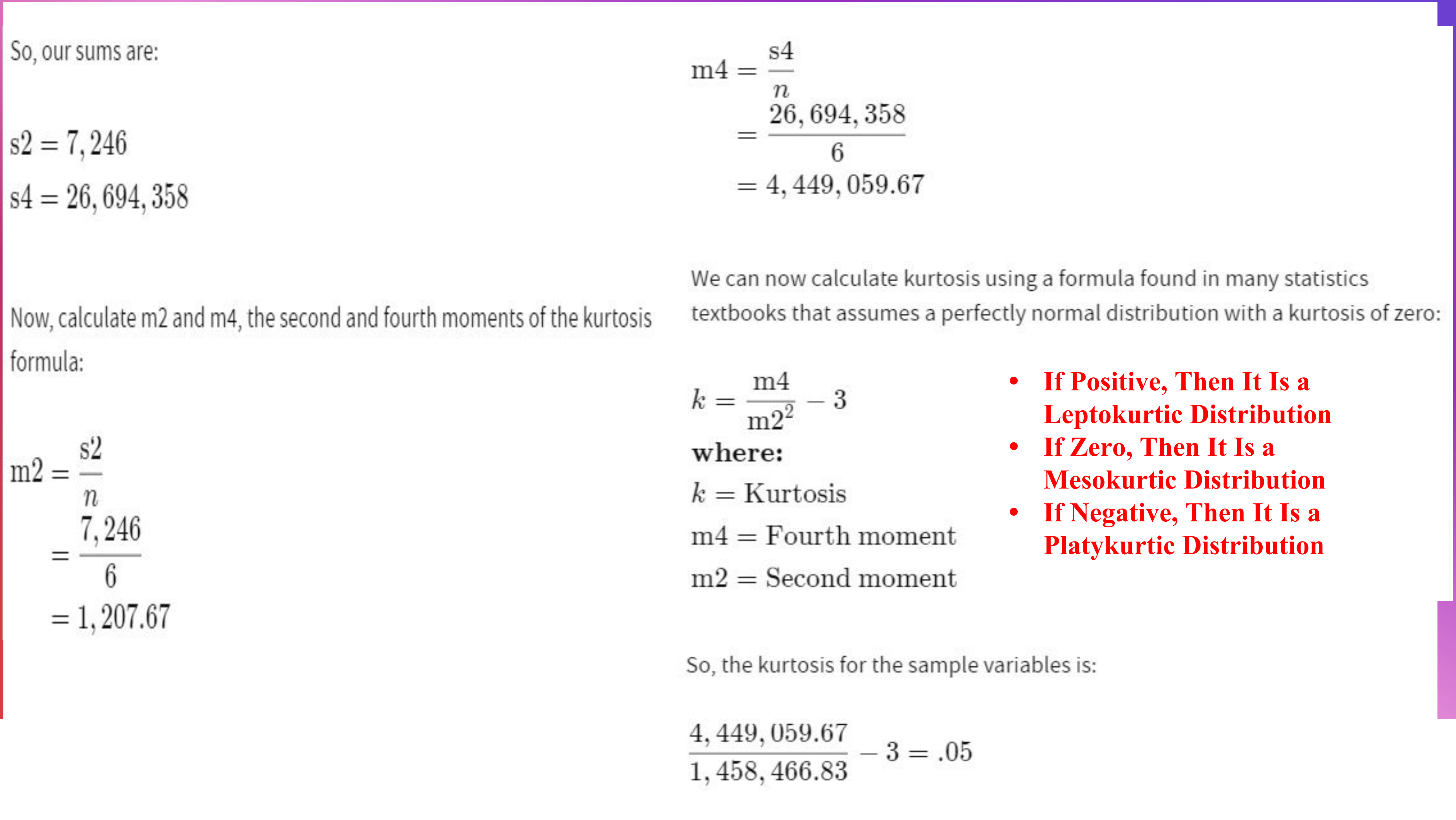
To get s4, use each variable, subtract the mean, and raise the result to the fourth power. Add all of the results together:

$$(27 - 42)^4 = (-15)^4 = 50,625$$
$$(13 - 42)^4 = (-29)^4 = 707,281$$
$$(17 - 42)^4 = (-25)^4 = 390,625$$
$$(57 - 42)^4 = (15)^4 = 50,625$$
$$(113 - 42)^4 = (71)^4 = 25,411,681$$
$$(25 - 42)^4 = (-17)^4 = 83,521$$
$$50,625 + 707,281 + 390,625 + 50,625 + 25,411,681$$
$$+ 83,521 = 26,694,358$$



To get s2, use each variable, subtract the mean, and then square the result. Add all of the results together:

$$(27 - 42)^2 = (-15)^2 = 225$$
$$(13 - 42)^2 = (-29)^2 = 841$$
$$(17 - 42)^2 = (-25)^2 = 625$$
$$(57 - 42)^2 = (15)^2 = 225$$
$$(113 - 42)^2 = (71)^2 = 5,041$$
$$(25 - 42)^2 = (-17)^2 = 289$$
$$225 + 841 + 625 + 225 + 5,041 + 289 = 7,246$$



So, our sums are:

$$s2 = 7,246$$

$$s4 = 26,694,358$$

Now, calculate m2 and m4, the second and fourth moments of the kurtosis formula:

$$\begin{aligned} m2 &= \frac{s2}{n} \\ &= \frac{7,246}{6} \\ &= 1,207.67 \end{aligned}$$

$$\begin{aligned} m4 &= \frac{s4}{n} \\ &= \frac{26,694,358}{6} \\ &= 4,449,059.67 \end{aligned}$$

We can now calculate kurtosis using a formula found in many statistics textbooks that assumes a perfectly normal distribution with a kurtosis of zero:

$$k = \frac{m4}{m2^2} - 3$$

where:

k = Kurtosis

m4 = Fourth moment

m2 = Second moment

- **If Positive, Then It Is a Leptokurtic Distribution**
- **If Zero, Then It Is a Mesokurtic Distribution**
- **If Negative, Then It Is a Platykurtic Distribution**

So, the kurtosis for the sample variables is:

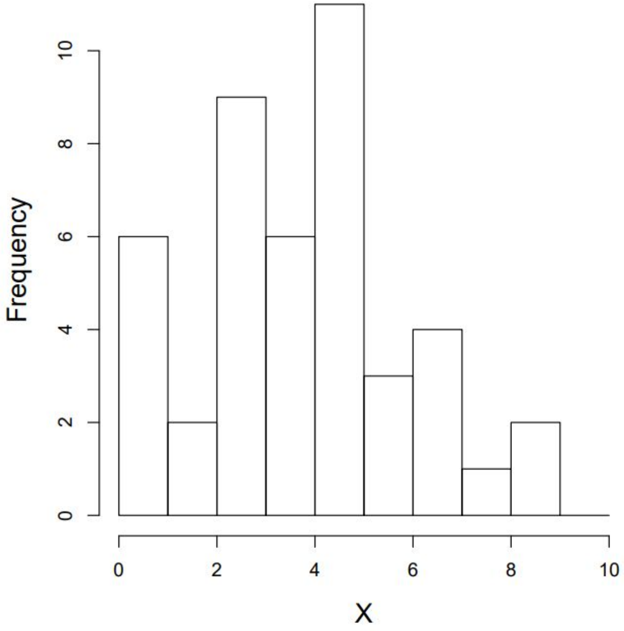
$$\frac{4,449,059.67}{1,458,466.83} - 3 = .05$$

Histograms

- The most basic graph is the histogram, which is a barplot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. Typically the bars run vertically with the count (or proportion) axis running vertically. To manually construct a histogram, define the range of data for each bar (called a bin), count how many cases fall in each bin, and draw the bars high enough to indicate the count.

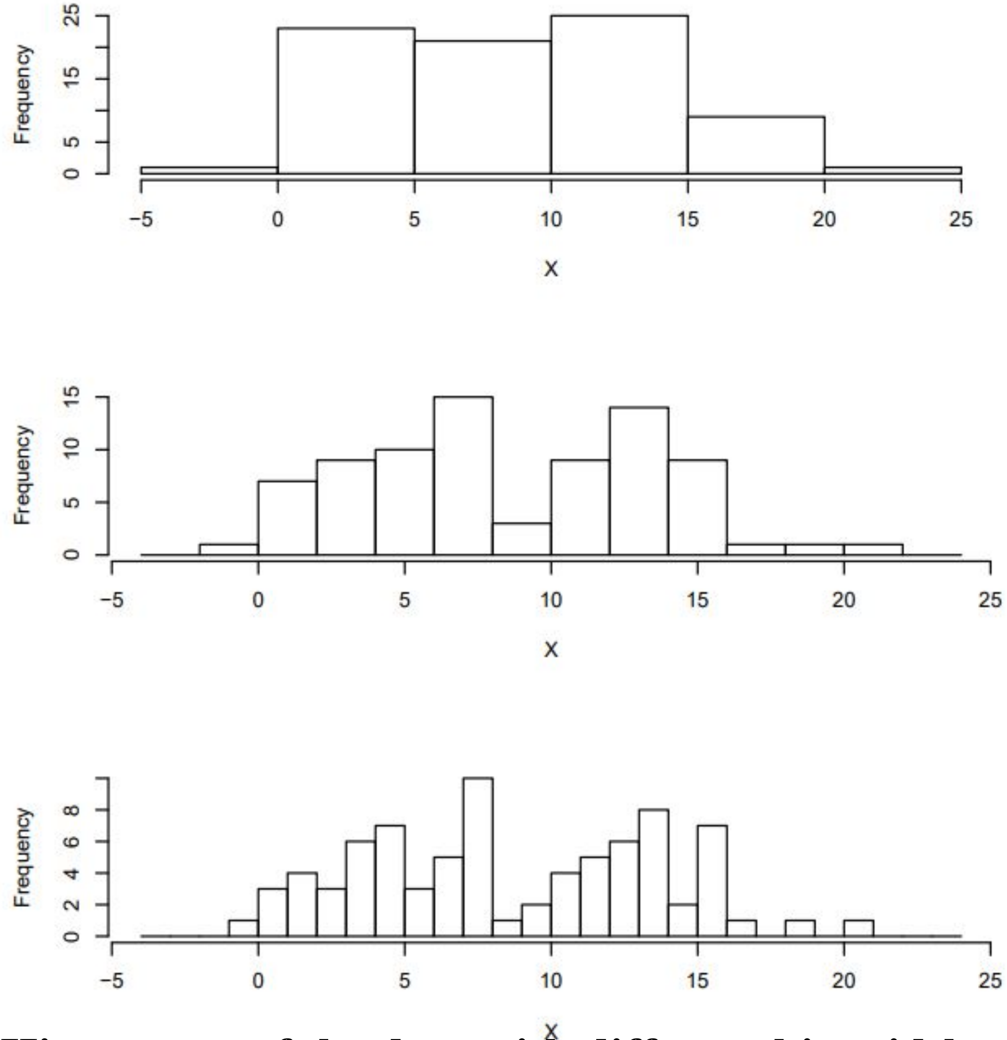
Sample Data:

3 5 4 9 3 5 3 7 6 4 8 5 1 4 5
4 3 4 1 3 5 5 5 7 1 5 6 2 1 7
5 9 1 2 1 5 3 3 3 5 4 6 7 3



Histogram

4.12 7.05 1.31 6.64 3.46 4.88 7.56 7.25 1.11 4.19 10.72 -0.95 0.29 3.05 0.64 2.83 2.41 4.97 7 5.8 11.73 1.94 7.99 3.84 4.87 3.13 2.17
7.87 4.52 7.32 3.3 1.79 8.38 0.42 5.28 6.94 3.55 6.61 5.44 7.74 12.71 13.29 12.29 12.58 10.63 15.53 18.28 13.32 10.06 20.4 11.12
12.61 4.97 7.96 13.27 12.01 13.79 15.04 13.88 15.51 11.66 16.23 13.58 13.53 11.4 11.86 13.68 15.53 9.69 9.06 14.65 7.71 15.45 6.77
15.72 15.15 7.59 14.7 12.94 10.01

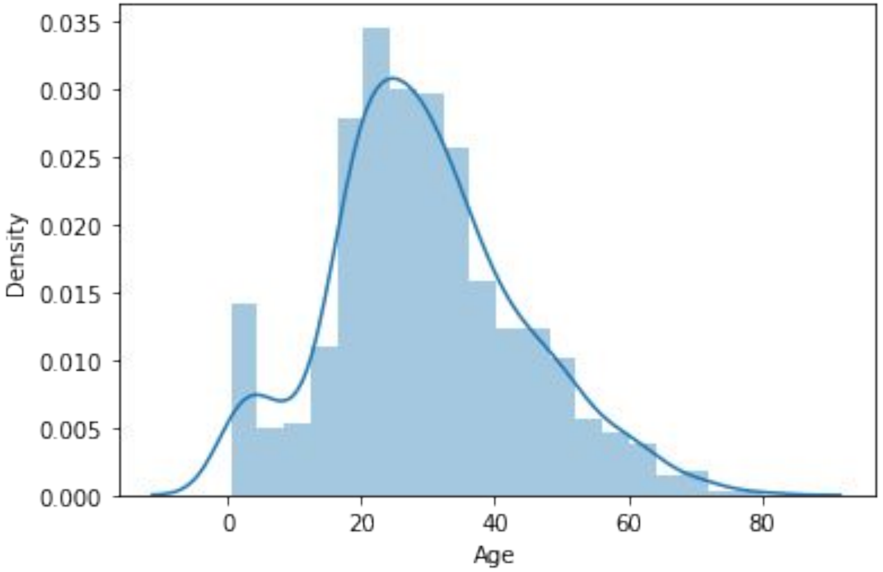


Histograms of the data with different bin widths (5,2,1)

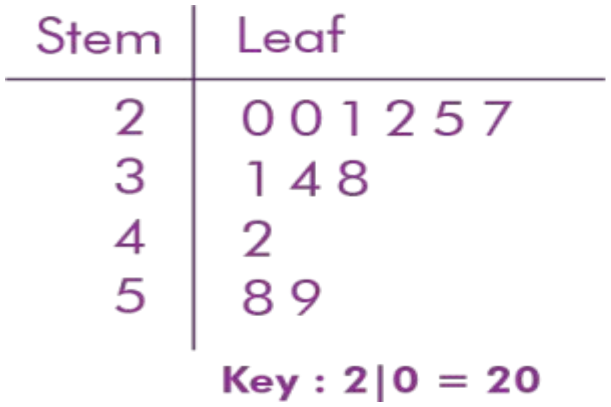
Distplot

```
sns.distplot(data['Age'])
plt.show()
```

- Distplot is also known as the second Histogram because it is a slight improvement version of the Histogram. Distplot gives us a KDE(Kernel Density Estimation) over histogram which explains PDF(Probability Density Function) which means what is the probability of each value occurring in this column. If you have study statistics before then definitely you should know about PDF function.



Stem-and-leaf plots



Stem	Leaf
2	0 0 1 2 5 7
3	1 4 8
4	2
5	8 9

Key : 2|0 = 20

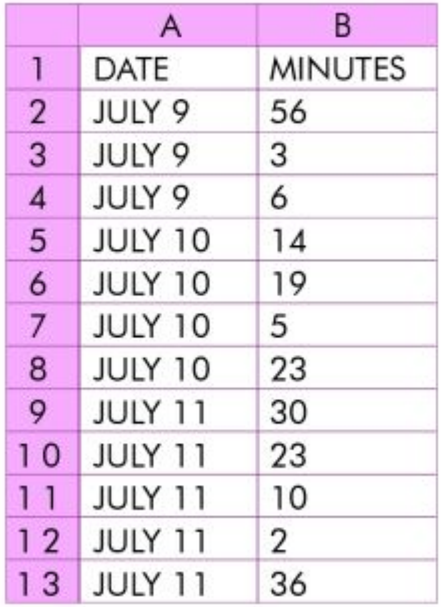
- A stem and leaf plot, also known as a stem and leaf diagram, is a way to arrange and represent data so that it is simple to see how frequently various data values occur. It is a plot that displays ordered numerical data.
- A stem and leaf plot is shown as a special table where the digits of a data value are divided into a stem (first few digits) and a leaf (usually the last digit). The symbol '|' is used to split and illustrate the stem and leaf values. For instance, 105 is written as 10 on the stem and 5 on the leaf. This can be written as **10 | 5.** Here, **10 | 5 = 105** is called the key.

How do we Construct a Stem and Leaf Plot?

- **Step 1:** Classify the data values in terms of the number of digits in each value, such as 2 digit numbers or 3 digit numbers.
- **Step 2:** Fix the key for the stem and leaf plot. For example, 2 | 5 = 25, 3 | 2 = 3.2 or 19 | 2 is 192.
- **Step 3:** Consider the first digits as stems and the last digit as leaves.
- **Step 4:** Find the range of the data, that is the lowest and the highest values among the data.
- **Step 5:** Draw a vertical line. Place the stem on the left and the leaf on the right of the vertical line.
- **Step 6:** List the stems in the stem column. Sort them in ascending order.
- **Step 7:** List the leaf values in the column against the stem from lowest to the highest horizontally.

Example 1:

The table below shows the duration of calls that Rosy makes each day. Represent the given data using a stem and leaf plot.



	A	B
1	DATE	MINUTES
2	JULY 9	56
3	JULY 9	3
4	JULY 9	6
5	JULY 10	14
6	JULY 10	19
7	JULY 10	5
8	JULY 10	23
9	JULY 11	30
10	JULY 11	23
11	JULY 11	10
12	JULY 11	2
13	JULY 11	36

Solution:

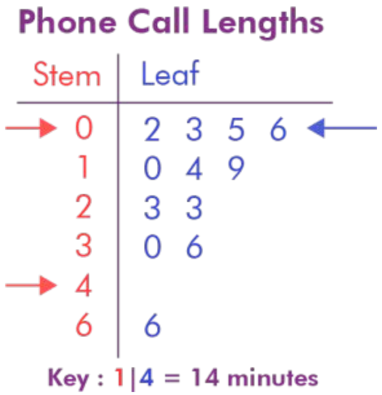
Step 1: Sort the data (number of minutes).

2, 3, 5, 6, 10, 14, 19, 23, 23, 30, 36, 56

Step 2: Choose the stems and the leaves. Just because the data values range from 2 to 56, use the tens digit for the stem and the ones digit for the leaf. Also, include the key.

Step 3: Write down the stems on the left of the vertical line.

Step 4: Write down the leaves for each stem on the right of the vertical line.

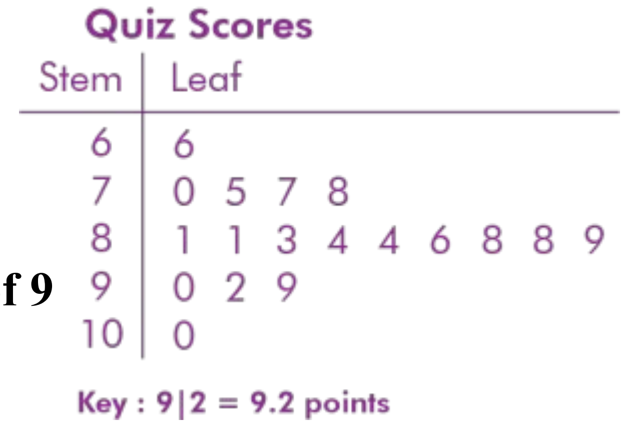


Phone Call Lengths

Stem	Leaf
0	2 3 5 6
1	0 4 9
2	3 3
3	0 6
4	
6	6

Key : 1|4 = 14 minutes

Example 2
The stem-and-leaf plot below shows the quiz scores of students.
(a) Find the number of students who scored less than 9 points?
(b) Find the number of students who scored a minimum of 9 points?



Quiz Scores

Stem	Leaf
6	6
7	0 5 7 8
8	1 1 3 4 4 6 8 8 9
9	0 2 9
10	0

Key : 9|2 = 9.2 points

Solution:

a) There are fourteen scores less than 9 points.

They are 6.6, 7.0, 7.5, 7.7, 7.8, 8.1, 8.1, 8.3, 8.4, 8.4, 8.6, 8.8, 8.8 and 8.9.

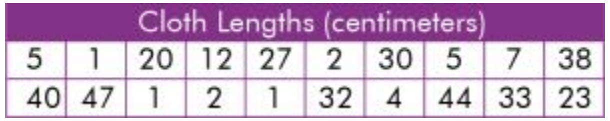
So, fourteen students scored less than 9 points.

b) There are two scores which are at least 9 points.

They are 9.0, 9.2, 9.9, and 10.0.

So, four students scored a minimum of 9 points.

Example 3:
Construct a stem-and-leaf plot for the data in the table.



Cloth Lengths (centimeters)									
5	1	20	12	27	2	30	5	7	38
40	47	1	2	1	32	4	44	33	23

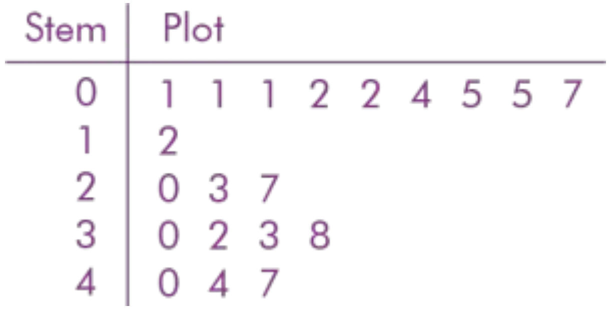
Solution:

Step 1: Sort the data values: 1, 1, 1, 2, 2, 4, 5, 5, 7, 12, 20, 23, 27, 30, 32, 33, 38, 40, 44, 47

Step 2: Choose the stems and the leaves. As the data values range from 1 to 47, use the tens digits for the stems and the ones digits for the leaves. Be sure to include the key.

Step 3: Write the stems to the left of the vertical line from the top to bottom.

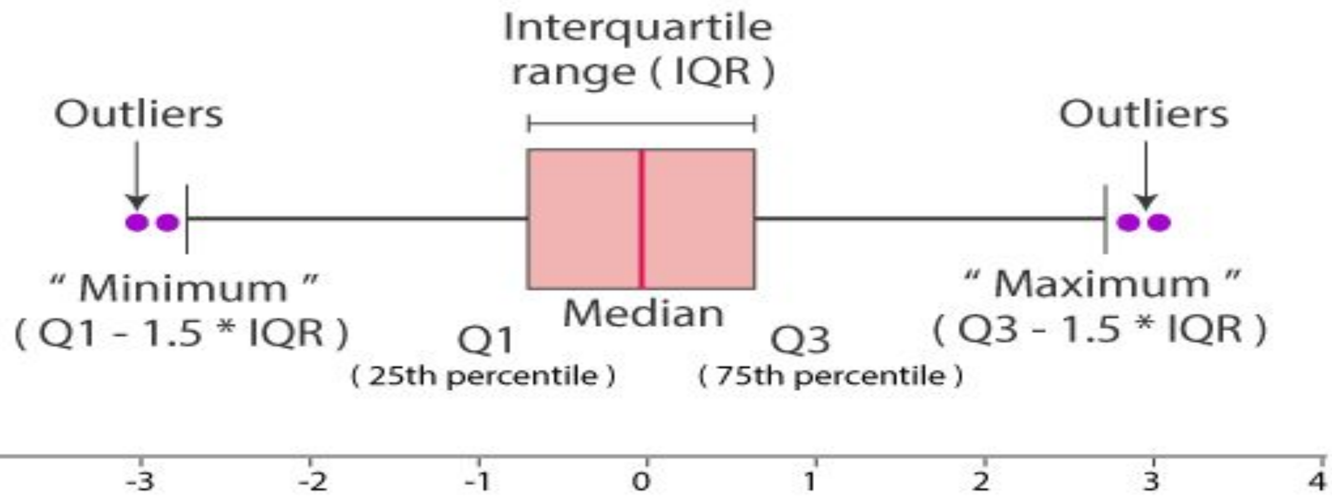
Step 4: Write the leaf values corresponding to each stem to the right of the vertical line.



Stem	Plot
0	1 1 1 2 2 4 5 5 7
1	2
2	0 3 7
3	0 2 3 8
4	0 4 7

Boxplot

- When we display the data distribution in a standardized way using 5 summary – minimum, Q1 (First Quartile), median, Q3(third Quartile), and maximum, it is called a **Box plot**. It is also termed as box and whisker plot.



Different parts of boxplot

© Byjus.com

Data:

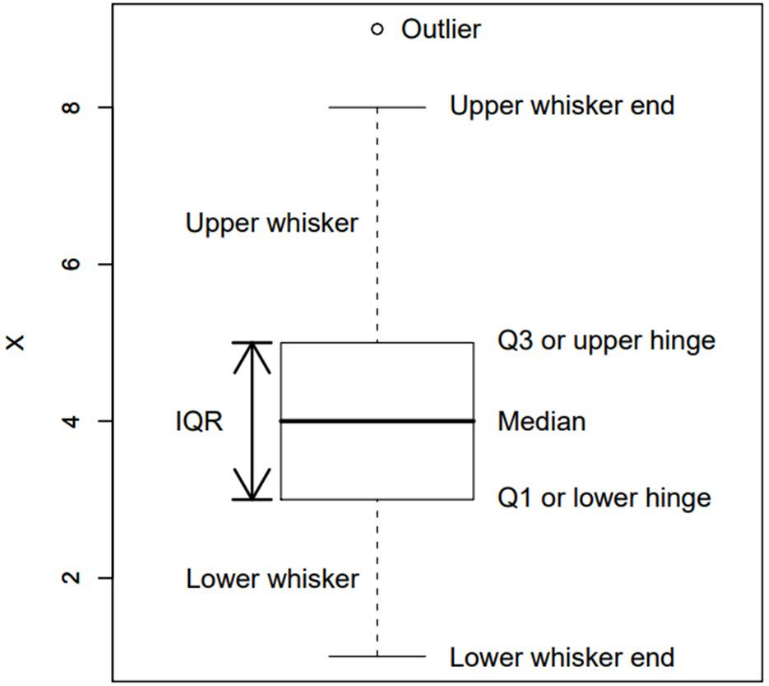
3, 5, 4, 9, 3, 5, 3, 7, 6, 4, 8, 5, 1, 4, 5, 4, 3, 4, 1, 3, 5, 5, 5, 7, 1, 5, 6, 2, 1, 7, 5, 9, 1, 2, 1, 5, 3, 3, 3, 5, 4, 6, 7, 3

Represents the number of eggs that each hen laid during the experiment.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44
1	1	1	1	1	1	2	2	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	5	5	6	6	6	7	7	7	7	8	9	9
										Q1=3+3/2											Median=4+4/2											Q3=5+5/2											
										3											4											5											

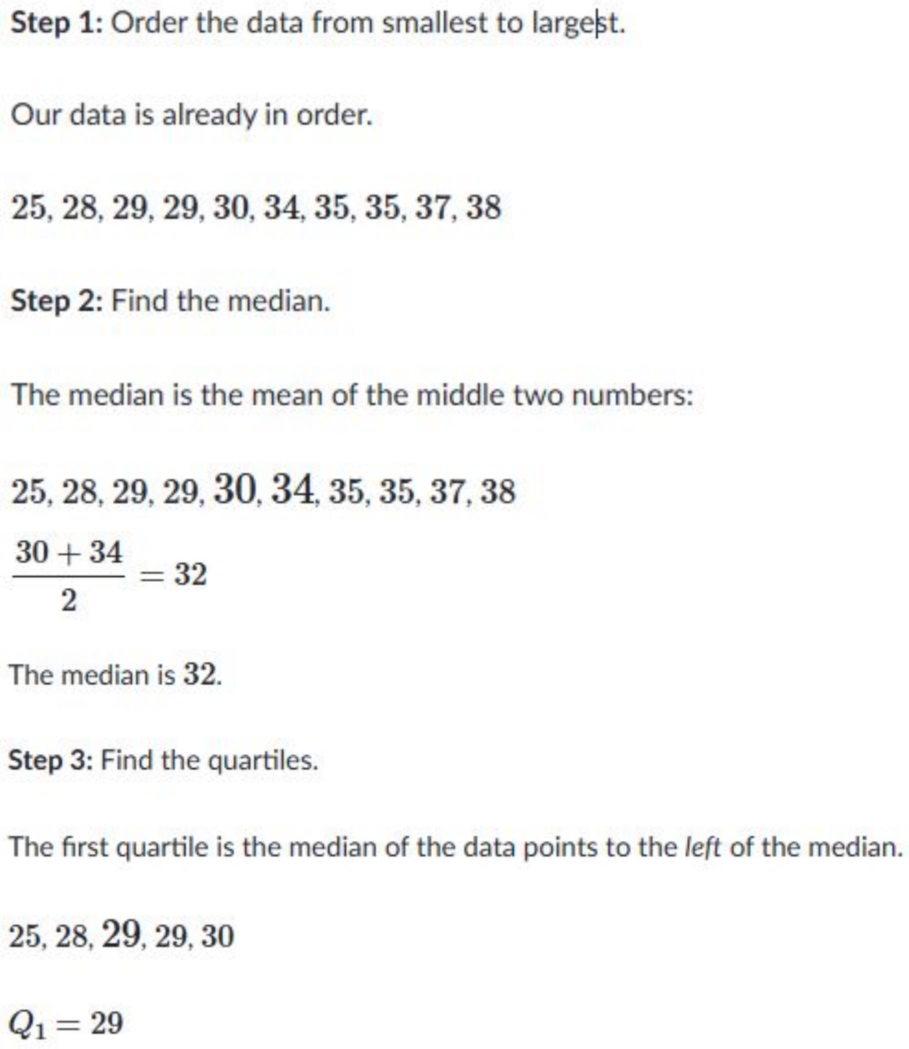
Median= 4
Min= 1
Max= 9
IQR=Q3-Q1=2
Lower Limit = Q1 - 1.5*IQR =0
Upper Limit = Q3 + 1.5*IQR=8

- The “outliers” identified by a boxplot, which could be called “boxplot outliers” are defined as any points more than 1.5 IQRs above Q3 or more than 1.5 IQRs below Q1.



Example: A sample of 10 boxes of raisins has these weights (in grams):

25,28, 29,29, 30,34, 35,35, 37, 38 **Make a box plot of the data.**



Step 1: Order the data from smallest to largest.

Our data is already in order.

$25, 28, 29, 29, 30, 34, 35, 35, 37, 38$

Step 2: Find the median.

The median is the mean of the middle two numbers:

$25, 28, 29, 29, \mathbf{30}, \mathbf{34}, 35, 35, 37, 38$

$$\frac{30 + 34}{2} = 32$$

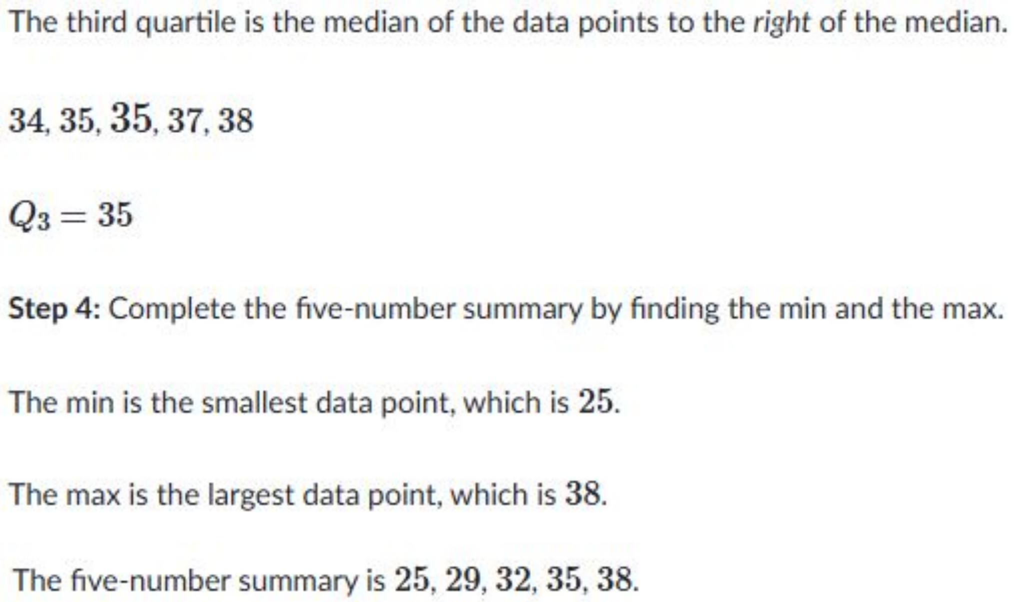
The median is 32.

Step 3: Find the quartiles.

The first quartile is the median of the data points to the *left* of the median.

$25, 28, \mathbf{29}, 29, 30$

$Q_1 = 29$



The third quartile is the median of the data points to the *right* of the median.

$34, 35, \mathbf{35}, 37, 38$

$Q_3 = 35$

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

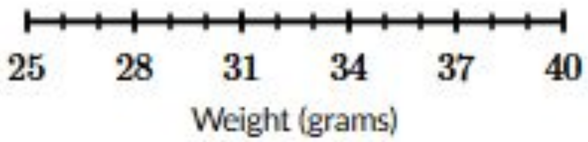
The min is the smallest data point, which is 25.

The max is the largest data point, which is 38.

The five-number summary is $25, 29, 32, 35, 38$.

Let's make a box plot for the same dataset

Step 1: Scale and label an axis that fits the five-number summary.

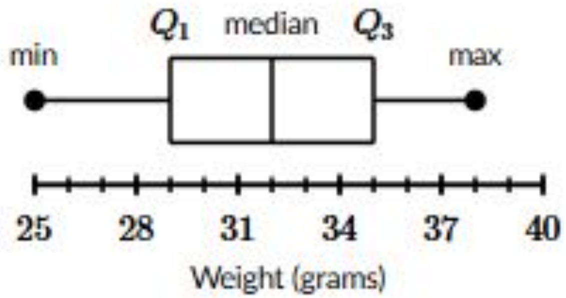


Step 2: Draw a box from Q_1 to Q_3 with a vertical line through the median.

Recall that $Q_1 = 29$, the median is 32, and $Q_3 = 35$.

Step 3: Draw a whisker from Q_1 to the min and from Q_3 to the max.

Recall that the min is 25 and the max is 38.


Q_1
median
Q_3
min
max
Weight (grams)