

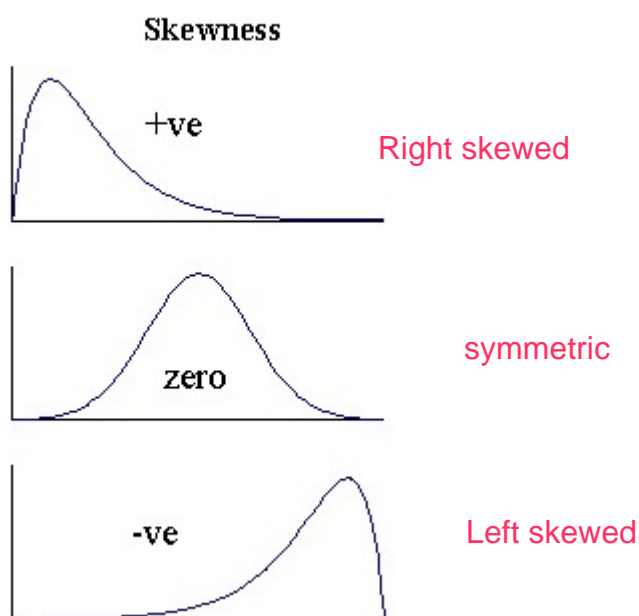
Five point summary

The dispersion gives the location and scale of the distribution.

A further characterization of the data includes skewness and kurtosis.

Skewness is a measure of lack of symmetry. A data set, is symmetric if it looks the same to the left and right of the center point.

The skewness for a Normal distribution is zero, and any symmetric data should have a skewness near zero.



Variance tells us about the amount of variability while skewness gives the direction of variability.

In business and economic series, measures of variation have greater practical application than measures of skewness.

However, in medical and life science field measures of skewness have greater practical applications than the variance.

Karl Pearson's Coefficient of Skewness

This method is most frequently used for measuring skewness.

The formula for measuring coefficient of skewness is given by

$$Sk = (\text{Mean} - \text{Mode}) / \sigma$$

The value of this coefficient would be zero in a symmetrical distribution.

If mean is greater than mode, coefficient of skewness would be positive otherwise negative.

The value of the Karl Pearson's coefficient of skewness usually lies between ± 1 for moderately skewed distribution.

If mode is not well defined, we use the formula

$$Sk = 3 (\text{Mean} - \text{Median}) / \sigma$$

By using the relationship

$$\text{Mode} = (3 \text{ Median} - 2 \text{ Mean})$$

Q. For a distribution Karl Pearson's coefficient of skewness is 0.64, standard deviation is 13 and mean is 59.2 Find mode and median

Solution: We have given

$$Sk = 0.64, \sigma = 13 \text{ and Mean} = 59.2$$

Therefore by using formulae

$$Sk = (\text{Mean} - \text{Mode}) / \sigma$$

$$0.64 = (59.2 - \text{Mode}) / 13$$

$$\text{Mode} = 59.20 - 8.32 = 50.88$$

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

$$50.88 = 3 \text{ Median} - 2 (59.2)$$

$$\text{Median} = (50.88 + 118.4) / 3 = 169.28 / 3 = 56.42.$$



UPDF
WWW.UPDF.COM

Q. Karl Pearson's coefficient of skewness is 1.28, its mean is 164 and mode 100, find the standard deviation.

$$Sk = (\text{Mean} - \text{Mode}) / \sigma$$

$$1.28 = (164 - 100) / \sigma$$

$$\sigma = 64 / 1.28 = 50.$$

Degree of skewness

In real world scenario , we will not get a perfect symmetrical data .

Thumb Rule .

If the degree of skewness is -1 to +1 data is considered fairly symmetrical



Kurtosis:-

Even If we have the knowledge of the measures of central tendency, dispersion and skewness, even then we cannot get a complete idea of a distribution.

In addition to these measures, we need to know another measure to get the complete idea about the **shape of the distribution which can be studied with the help of Kurtosis.**

Prof. Karl Pearson has called it the “Convexity of a Curve”. Kurtosis gives a measure of flatness of distribution or measure of the tailedness of a distribution

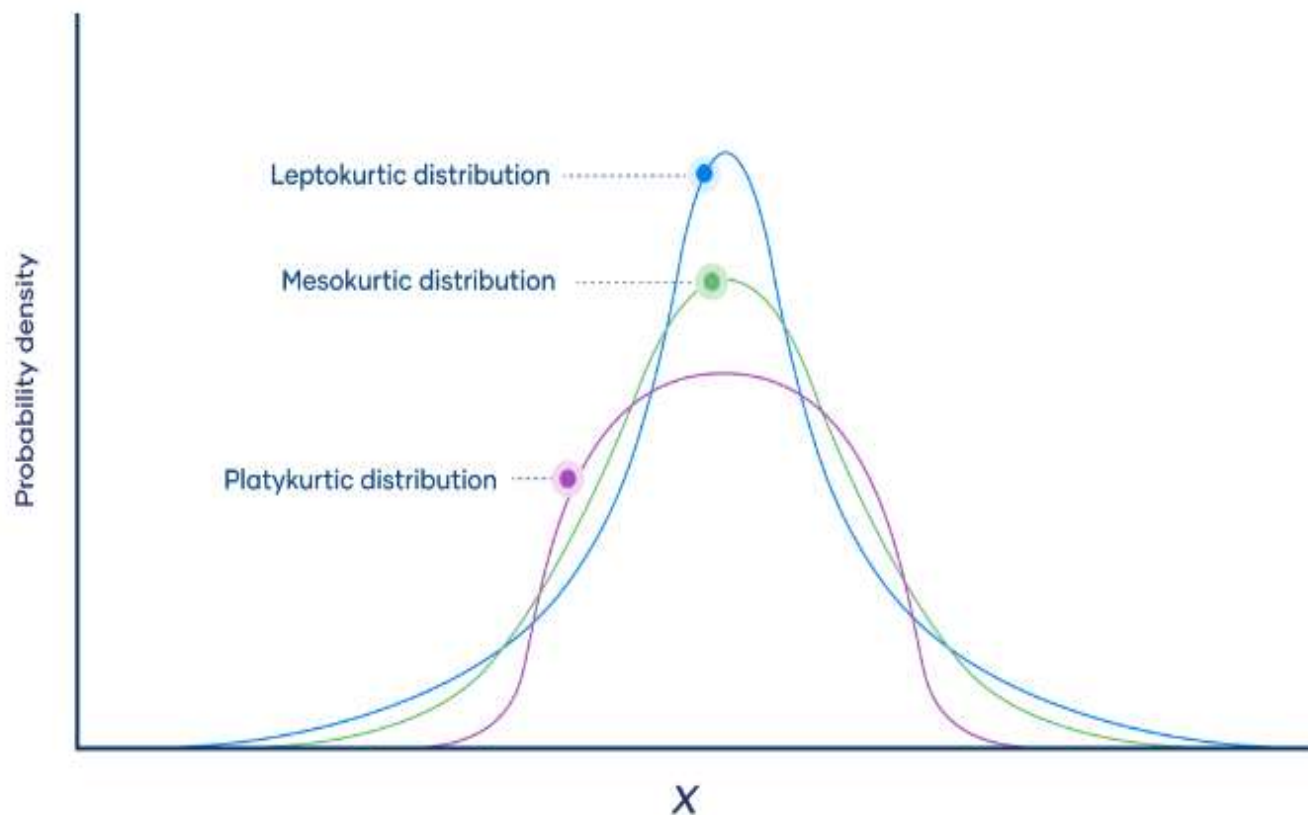
- Distributions with medium kurtosis (medium tails) are **mesokurtic..**
- Distributions with low kurtosis (thin tails) are **platykurtic.**
- Distributions with high kurtosis (fat tails) are **leptokurtic.**

Tails are the tapering ends on either side of a distribution. They represent the probability or frequency of values that are extremely high or low compared to the mean. In other words, **tails represent how often outliers occur.**



Types of kurtosis

Distributions can be categorized into three groups based on their kurtosis:



	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

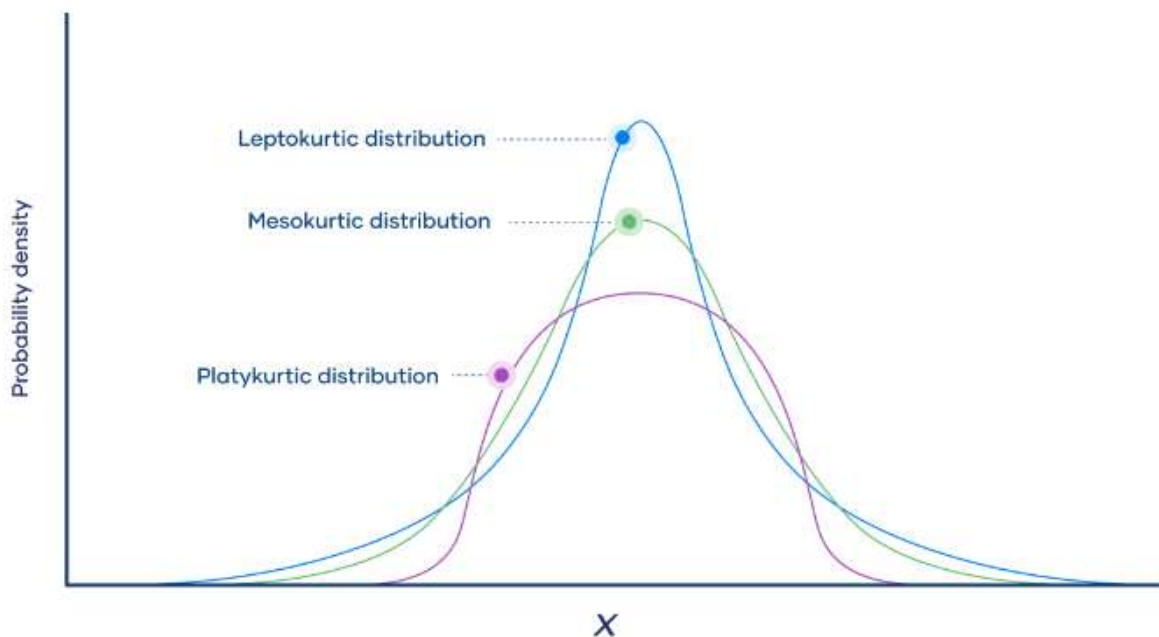


Frequently asked questions: -

What is the difference between skewness and kurtosis?

Skewness:- Measure of asymmetry of a distribution.

Kurtosis :- Measure of heaviness of a distribution's tail relative to normal distribution.



Q. what are the three categories of kurtosis?

Mesokurtosis: An excess kurtosis of 0. Normal distributions are mesokurtic.


- **Platykurtosis** : A negative excess kurtosis. Platykurtic distributions are thin-tailed, meaning that they have few outliers.
- **Lepokurtosis** : A positive excess kurtosis. Leptokurtic distributions are fat-tailed, meaning that they have many outliers.



Q. What is a normal distribution?

In a normal distribution, data are symmetrically distributed with no skew. Most values cluster around a central region, with values tapering off as they go further away from the center.

The measures of central tendency (mean, mode, and median) are exactly the same in a normal distribution.

Many Machine learning algorithms depend on normality assumptions. 

Significant skewness and kurtosis clearly indicate that data are not normal. (as indicated by a histogram or the numerical measures).. what can we do about it?

One approach is to apply some type of transformation to try to make the data normal, or more nearly normal. The Box-cox transformation is a useful technique for trying to normalize a data set. In particular, taking the log or square root of a data set is often useful for data that exhibit moderate right skewness.



5 point summary

Quantile

After arranging the dataset in ascending order, when we find the position of the

5 points: -

Minimum

Q1- first quantile

Q2-Median (2nd quantile)

Q3- third quantile

Maximum

Q1= Median position of the first half is first quantile.

Q3= Median position of the second half is Third quantile.

Q2= Median of the complete dataset.

Inter Quantile Range

$IQR = Q3 - Q1$.



df.describe():- function to summarize information

```
In [55]: df1.describe()
```

Out[55]:

	Profit	Unit Price	Shipping Cost
count	17.000000	18.000000	17.000000
mean	305.741382	75.768333	17.985294
std	884.219143	119.998788	23.028008
min	-481.041000	2.880000	0.700000
25%	-11.682000	4.987500	2.990000
50%	26.920000	39.415000	5.260000
75%	313.578000	99.722500	26.220000
max	3424.220000	500.980000	74.350000

- **Counts :-No. of observations**
- **Average value (mean)**
- **Std:-On average, how much each measurement deviates from the mean (standard deviation of the mean)**
- **Min:- minimum value of the observation.**
- **Max:- Maximum value of the observation.**
- **25,50,75 percentiles. (Quartile Range)**
- **Most frequently occurring value (mode)**
- **Midpoint between the lowest and highest value of the set (median)**



UPDF

WWW.UPDF.COM

Manual calculation:-

Find Q1, Q2 , Q3 , IQR?

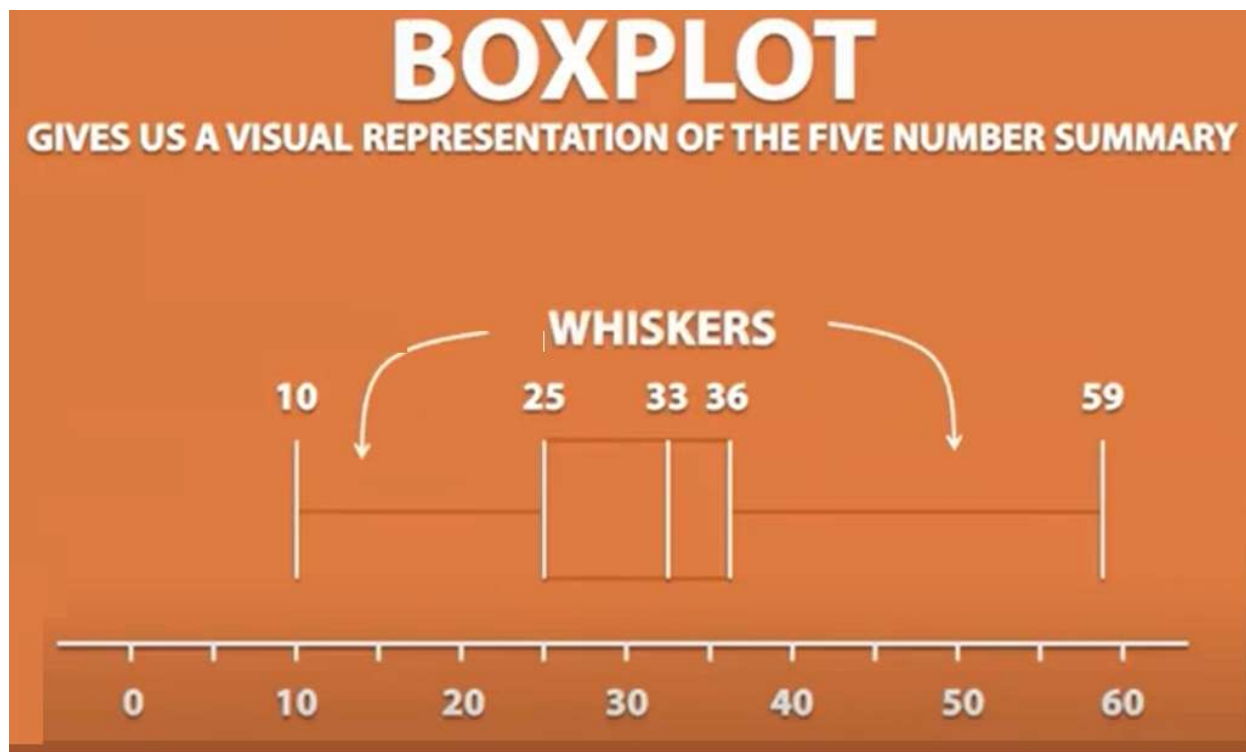
12 25 25 27 10 11 31 33 36 43 50 59 34 34 35

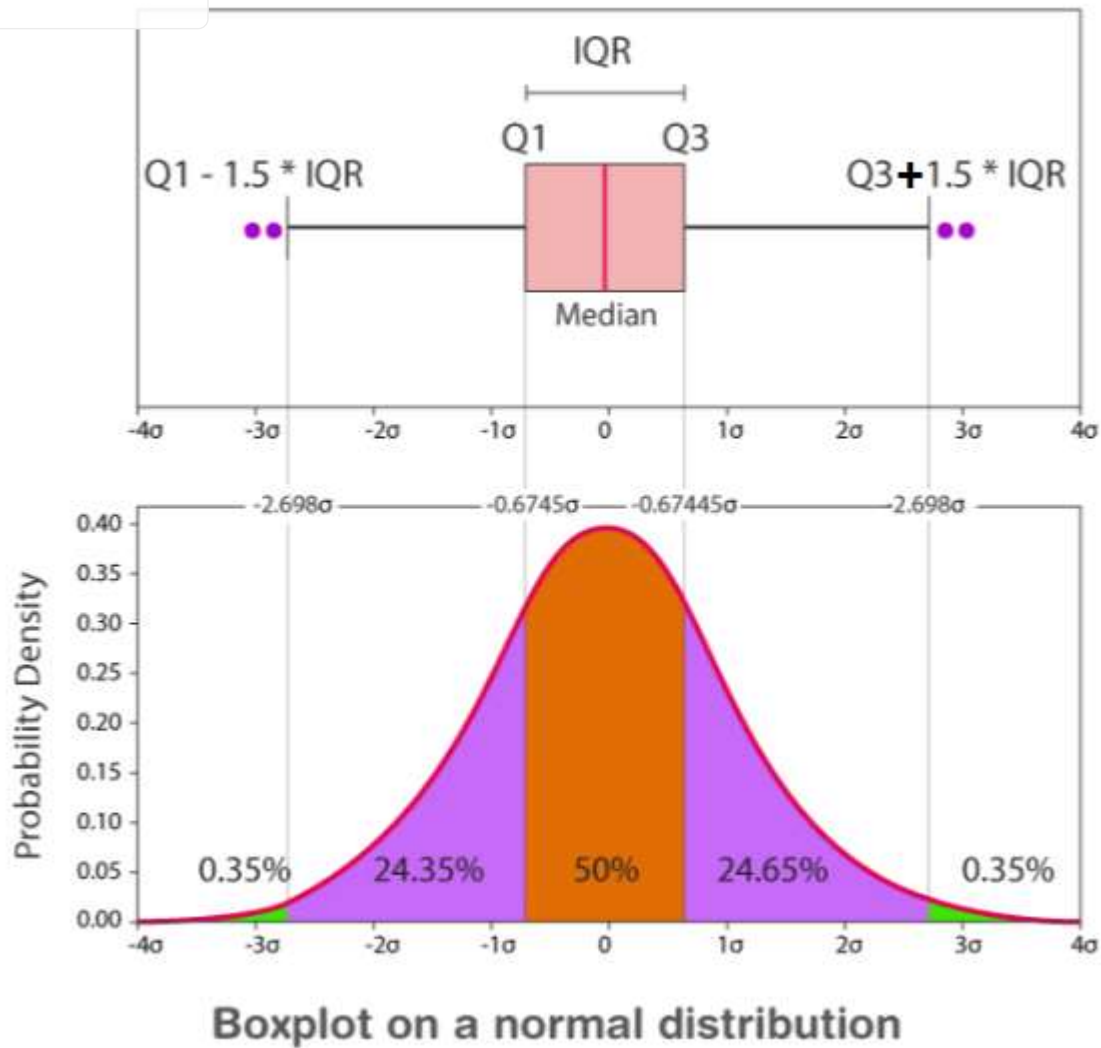
Data point arranged in ascending order,

10 11 12 25 25 27 31 33 34 34 35 36 43 50 59



Representation on Boxplot:-



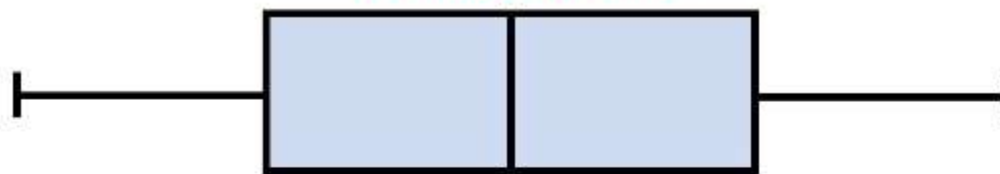


Also called Gaussian Distribution

Similar to bell shape curve....



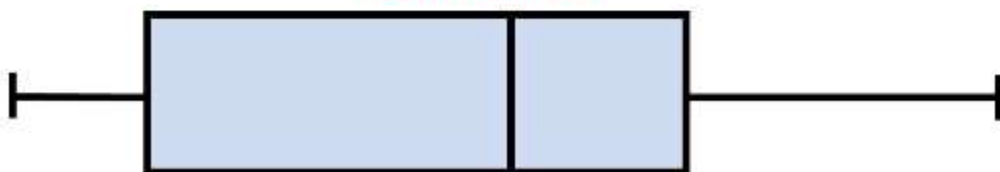
Normal Distribution



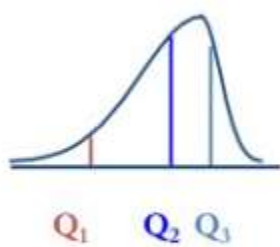
Positive Skew



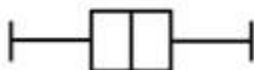
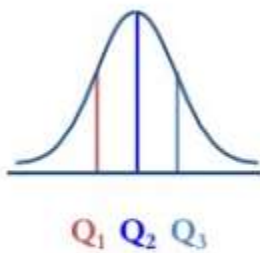
Negative Skew



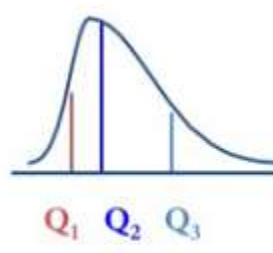
Left-Skewed



Symmetric

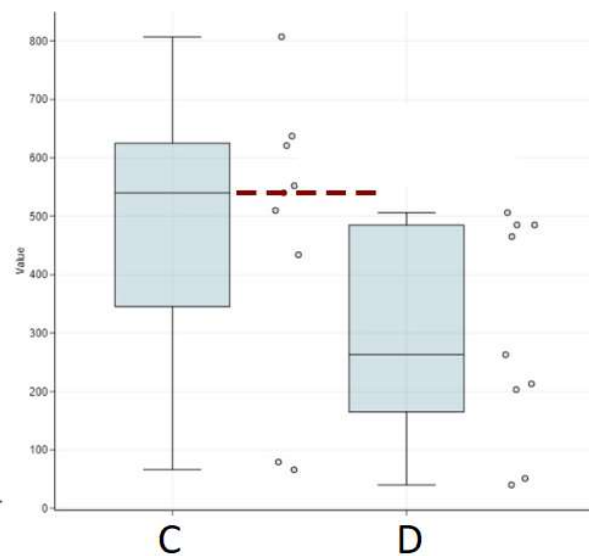
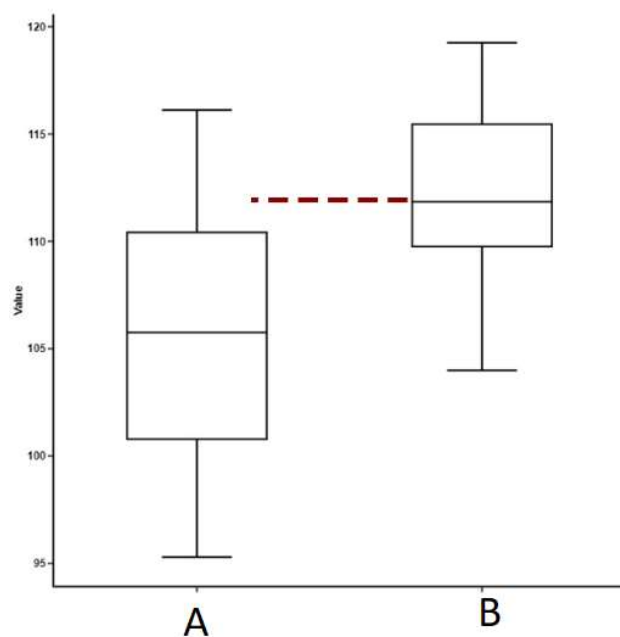


Right-Skewed





Compare between different variable BOX PLOTS

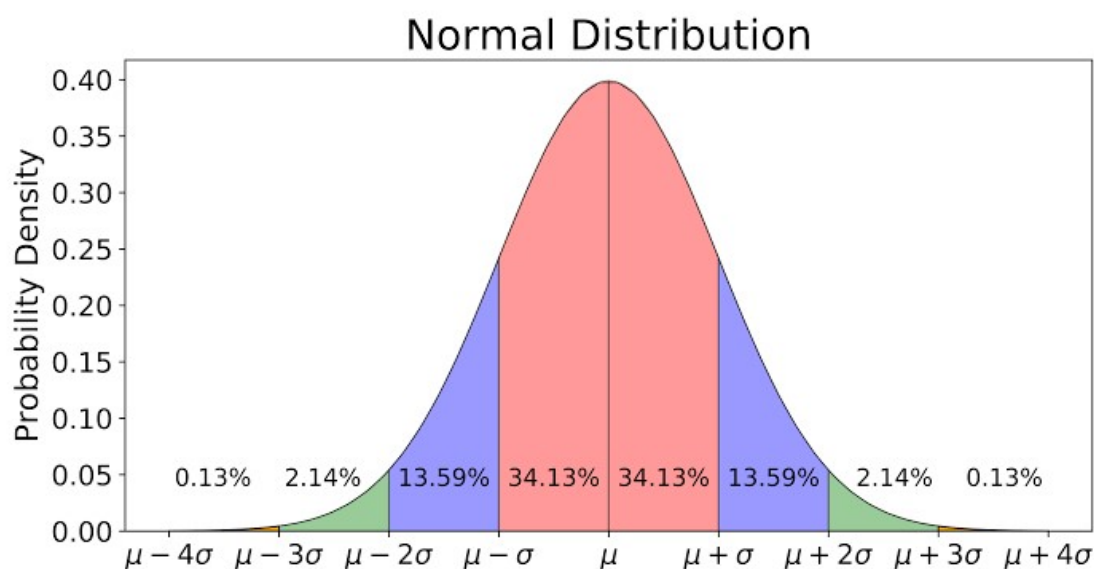




Empirical Rule.

BELL shape and symmetric.

A useful continuous distribution is the **standard normal distribution** with mean equal to 0 and standard deviation equal to 1



Normal distribution is a density curve total area equals to 100%.

There are certain observations which could be inferred from this Rule 68,95,99. Called **Empirical Rule**.

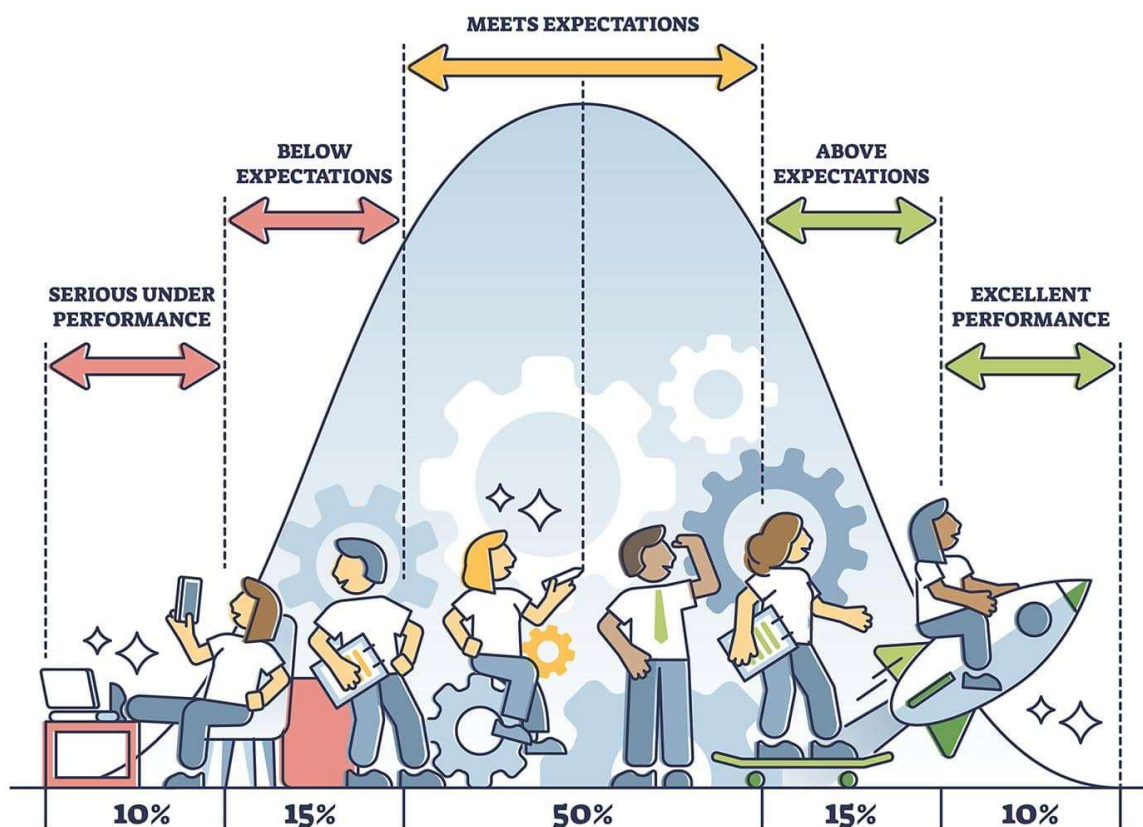
About 68.26% data lies within 1 SD ($<\sigma$) of the mean (μ),
About 95.44% data lies within 2 s d ($<\sigma$) of the mean (μ),
About 99.72% data lies within 3 s d ($<\sigma$) of the mean (μ)
& the rest 0.28% of the data lies outside 3 SD ($>3\sigma$) of the mean (μ), And this part of the data is considered as outliers.



UPDF

WWW.UPDF.COM

BELL CURVE

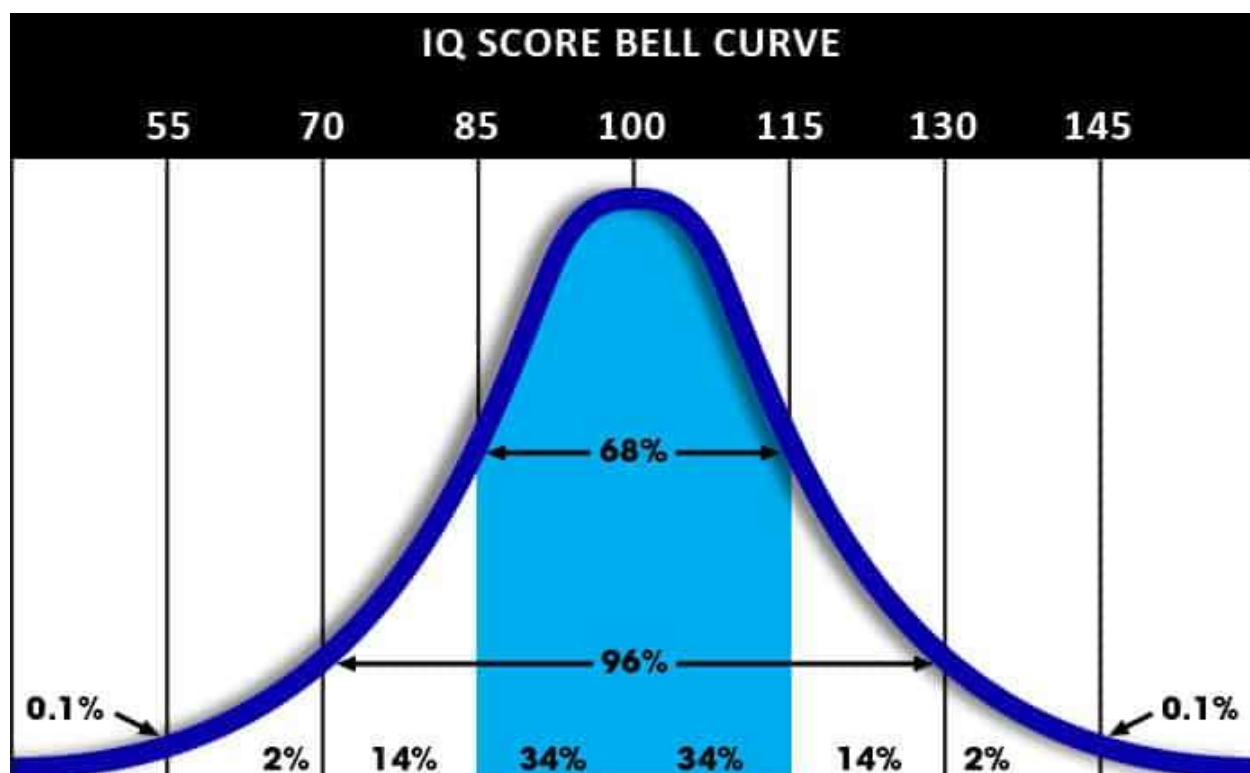


Why Is The Normal Distribution Important?

The bell-shaped curve is a common feature of nature and psychology

The normal distribution is the most important probability distribution in statistics because many continuous data in nature and psychology display this bell-shaped curve when compiled and graphed.

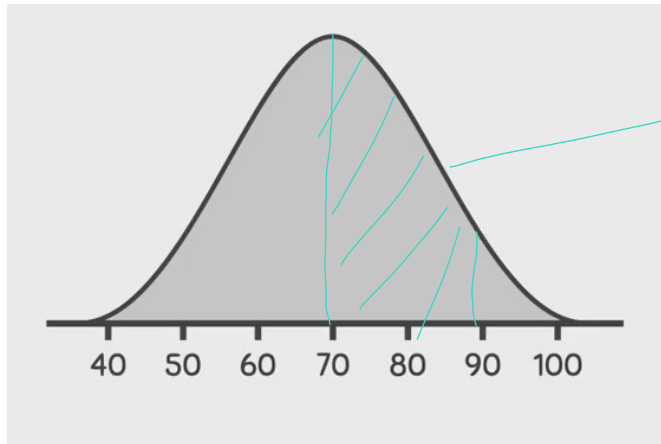
For example, if we randomly sampled 100 individuals, we would expect to see a normal distribution frequency curve for many continuous variables, **such as IQ, height, weight, and blood pressure.**



		Classification
130 and above		Very superior
120–129		Superior
110–119		High average
90–109		Average
80–89		Low average
70–79		Borderline
69 and below		Extremely low



(1) The Normal distribution has a standard deviation of 10 and mean 70. Approximately what area is contained between 70 and 90?



$95/2$

Hint:- use Empirical Rule 68,95,99.



Q2. The mean life of a tire is 30,000 km. The standard deviation is 2000 km.

Then, 68% of all tires will have a life between
28k km and 32k km.



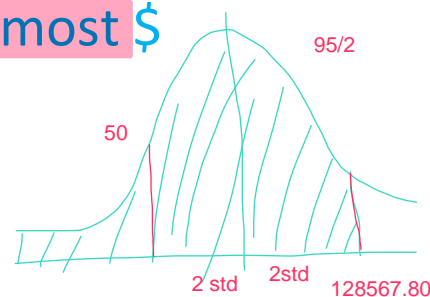
UPDF
WWW.UPDF.COM

(3) Q. In 2019 base salary of NYC employees between \$ 1000 and \$ 150000.

Mean = \$ 73,555.88

SD = \$ 27,505.98

What proportion of population makes at most \$ 128,567.80?



so total area = 50 + (95/2)



$$128567.80 = 73555.88 + n * 27505.98$$

$$n = (128567.80 - 73555.88) / 27505.98$$

$$n = 2, \text{ so } (95/2)\%$$

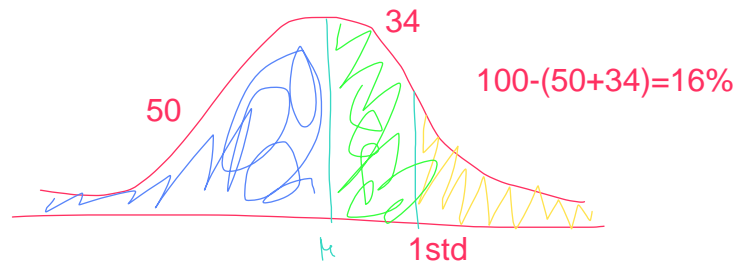
$$\text{Total area} = 50 + (95/2)$$



(4) what proportion of population makes **more than \$ 101,061.90?**

Mean = \$ 73,555.88

SD = \$ 27,505.98



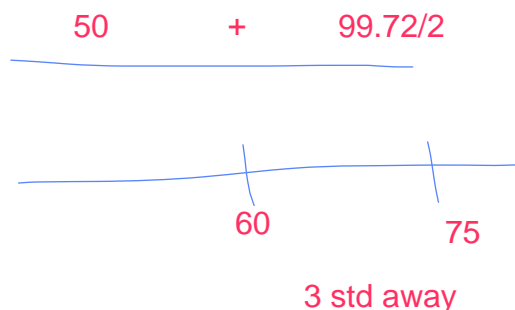
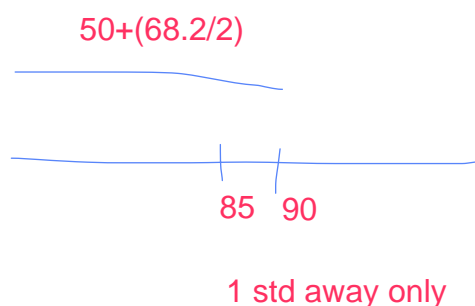


Q5 . One student score 90 marks in Mathematics and 75 Marks in English.

Consider the average class score of Mathematics is 85, SD is 5. Average class score of English is 60 , SD is 5.

Which performance is better Mathematics or English?

when performance comparing to class marks in english is better, coz both eng and maths have same std.and avg english is very low compaired to maths and 90 in maths close to its avg but 75 (marks in eng) is away to its avg 60, so marks in eng is better when compairing to class.

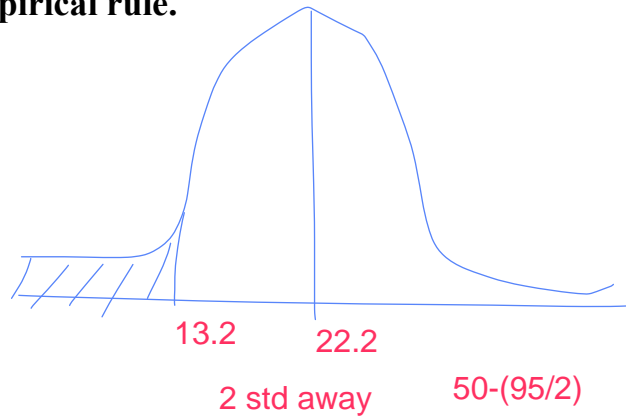




(6) In a Garden the heights of plants are normally distributed, the mean of the plants are 22.2 inches and the SD of 4.5 inches.

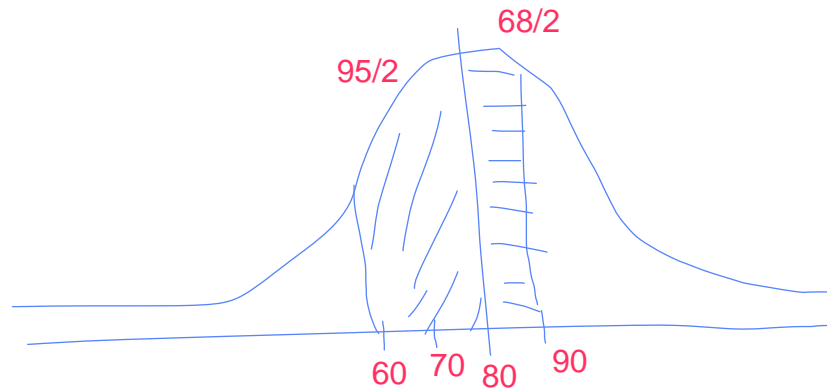
Estimate the percentage of plants that are less than 13.2 inches tall.

Hint: - Use Empirical rule.





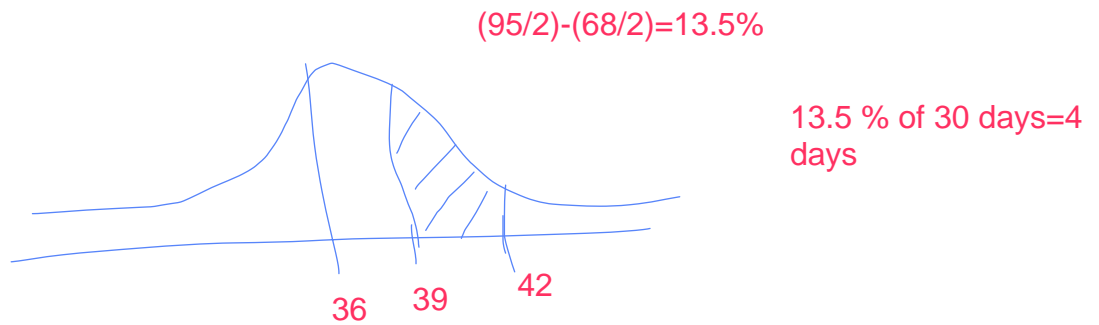
Q7. A competency test has scores with a mean of 80 and a standard deviation of 10. A histogram of the data shows that the distribution is normal. Use the Empirical Rule to find the percentage of scores between 60 and 90.





Q.8 The mean June midday temperature in Delhi is 36°C and the standard deviation is 3°C

Assuming this data is normally distributed, how many days in June would you expect the midday temperature to be between 39°C and 42°C ?





Q9. Mean demand of an oil is 1000 ltr per month with SD Of 250 ltr.

(a) if 1200 ltrs are stocked, what is the satisfaction level? less than 50+34 %

(b) For an assurance of 95%. what stock must be kept? less than 1500 ,1500 will be 50+47.5 %

