

Statistics Consolidated Assignment 1

Que 1) Plot a histogram,

10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99

Problem Statement:

Plot a Histogram for the below values

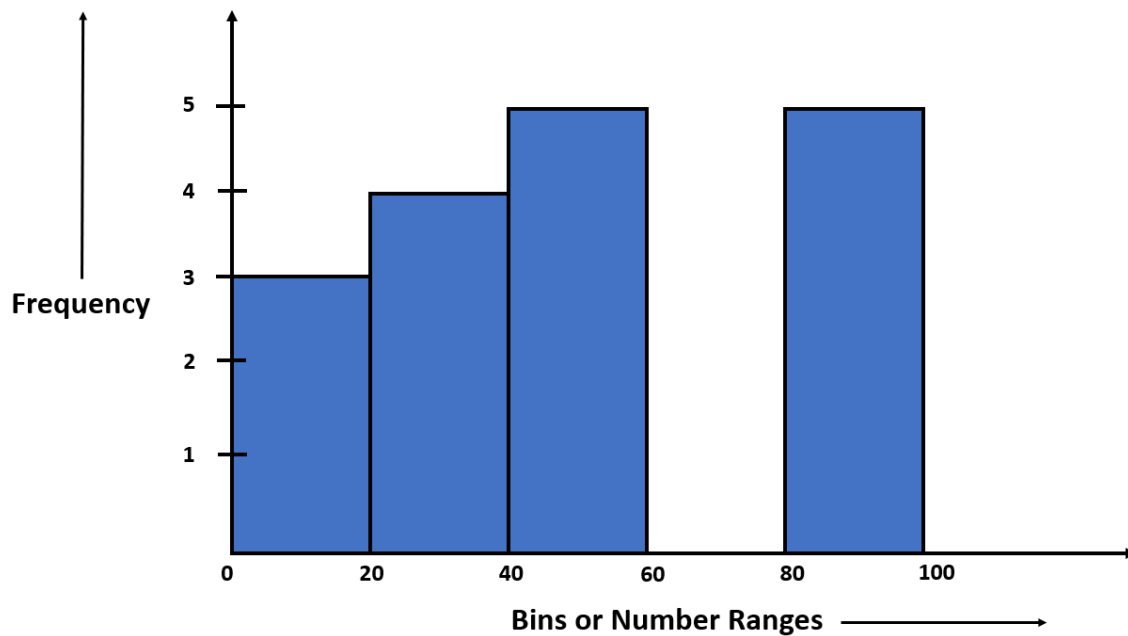
{10, 13, 18, 22, 27, 32, 38, 40, 45, 51, 56, 57, 88, 90, 92, 94, 99}

Bin Size to be used is 20

Number of Bins to be used is 5

Solution:

Given below is Histogram:



NOTE: The 4th Bucket i.e., 60 to 80 will not have any values.

Que 2) In a quant test of the CAT Exam, the population standard deviation is known to be 100. A sample of 25 tests taken has a mean of 520. Construct an 80% CI about the mean.

Solution:

Since we have been given the Population Standard deviation we can make use of the Z-Score Table to solve this problem.

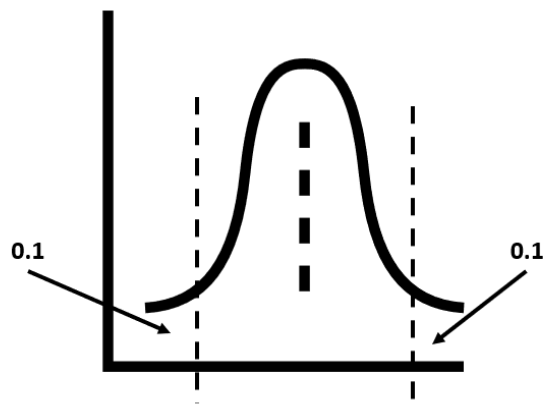
Population Standard Deviation = 100

Sample Mean = 520

Sample Size = 25

Since we need to find an 80% Confidence Interval about the mean we can determine the significance value as $1 - (80/100) = 1 - 0.8 = 0.2$

Since the distribution is symmetric, we have 0.1 on both sides of the curve.



The Lower Fence of the Confidence Interval can be calculated using the below formula

$$\text{Lower Fence} = \text{Point Estimate} - Z (\alpha/2) * (\sigma / \sqrt{n})$$

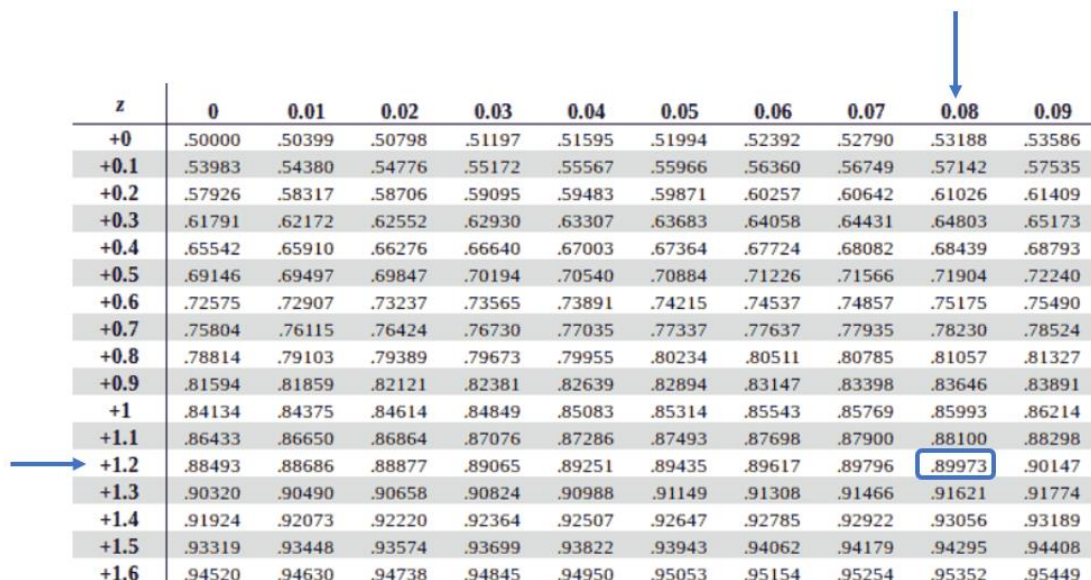
Here the Point Estimate is the Sample Mean = 520

α is the Significance value = 0.2

We need to find the Z score for 0.1

We can find it by subtracting 1 from 0.1 i.e., 0.9. We need to find the Z value corresponding to area value of 0.9.

The value is determined to be 1.28. Since it is symmetric for the negative side it will be – 1.28.



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449

$$\text{Lower Fence} = \text{Point Estimate} - Z (\alpha/2) * (\sigma / \sqrt{n})$$

$$= 520 - 1.28 * (100 / \sqrt{25})$$

$$= 520 - 1.28 * 20$$

$$= 494.4$$

The Upper Fence of the Confidence Interval can be calculated using the below formula

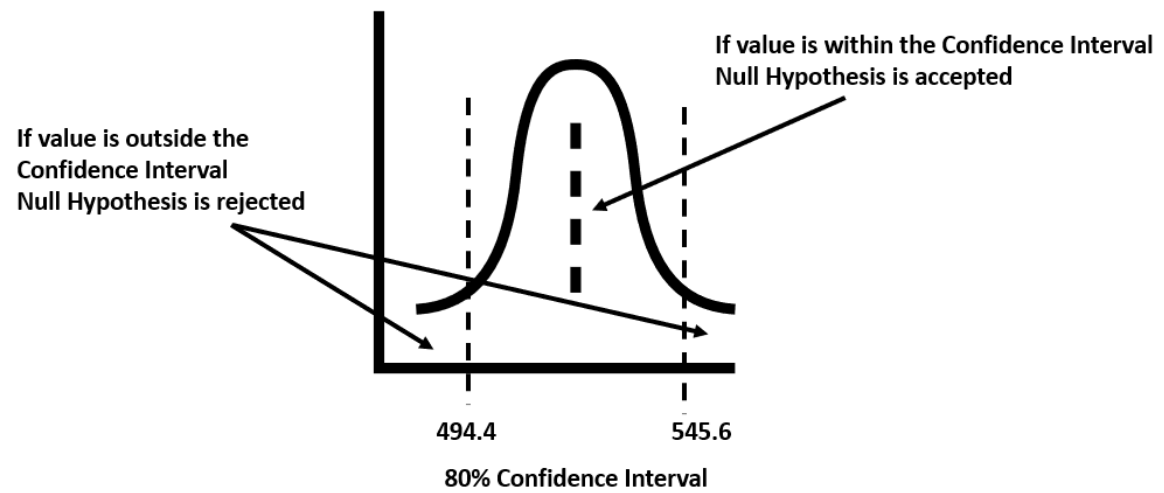
$$\text{Upper Fence} = \text{Point Estimate} + Z (\alpha/2) * (\sigma / \sqrt{n})$$

$$= 520 + 1.28 * (100 / \sqrt{25})$$

$$= 520 + 1.28 * 20$$

$$= 545.6$$

Given below is the Distribution with the 80% confidence interval values.



Que 3) A car believes that the percentage of citizens in city ABC that owns a vehicle is 60% or less. A sales manager disagrees with this. He conducted a hypothesis testing surveying 250 residents & found that 170 residents responded yes to owning a vehicle.

- a) State the null & alternate hypothesis.
- b) At a 10% significance level, is there enough evidence to support the idea that vehicle owner in ABC city is 60% or less.

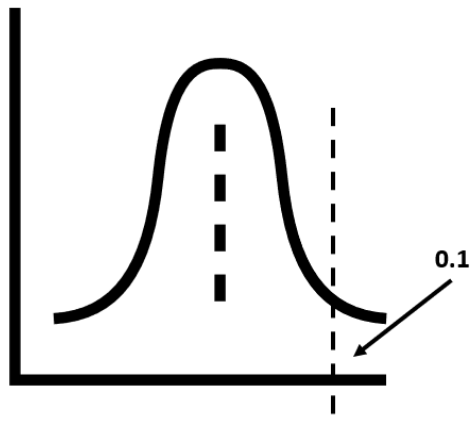
Solution:

- a) Null Hypothesis: H_0 : Percentage of citizens in city ABC that own a vehicle is $\leq 60\%$
Alternate Hypotheses: H_1 : Percentage of citizens in city ABC that own a vehicle is $> 60\%$

Since the number of samples is greater than 30, we need to make use of Z-test.

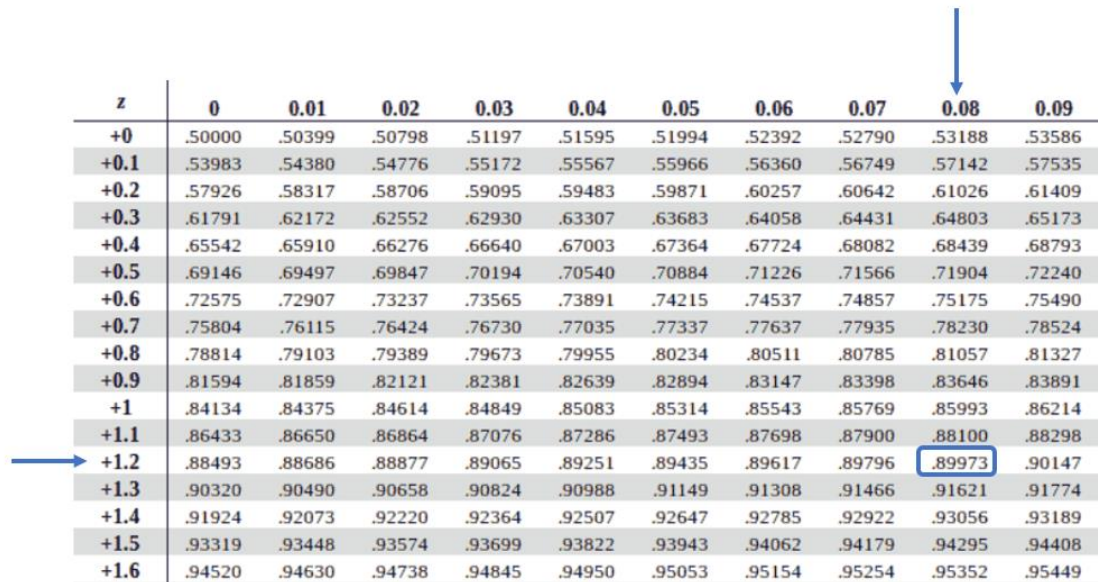
α is the Significance value = 0.1

We need to find the Z score for 0.1 since this is a one-tail test.



We can find it by subtracting 1 from 0.1 i.e., 0.9. We need to find the Z value corresponding to area value of 0.9.

The value is determined to be 1.28.



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449

Z score with Proportions = $(\hat{p} - p_0) / (\sqrt{p_0 * q_0 / N})$

Where $\hat{p} = 170 / 250 = 0.68$

$p_0 = 0.6$

$q_0 = 1 - 0.6 = 0.4$

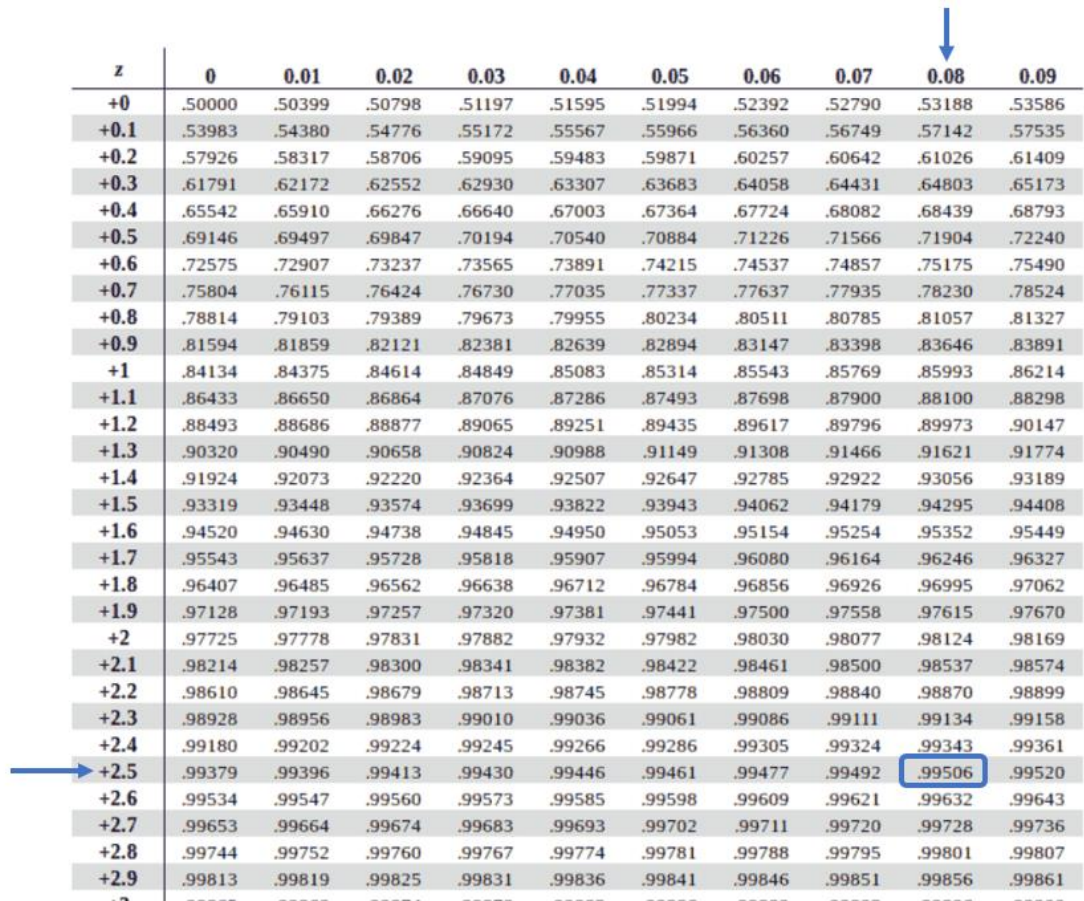
$N = 250$

Z score = $0.68 - 0.6 / (\sqrt{0.24/250}) = 2.58$

Conclusion: Since the calculated Z score 2.58 is greater than 1.28 we reject the Null Hypothesis. Hence, we can conclude that there is NOT enough evidence to support the idea that vehicle owner in ABC city is 60% or less

Alternate method to decide on the conclusion is to determine the p value for the calculated Z score of 2.58.

The value for 2.58 is 0.99506. Since we need to find the area of the tail, we need to subtract 1 from 0.99506. p value is 0.00494



z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
+0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
+0.1	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
+0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
+0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
+0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
+0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
+0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
+0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
+0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
+0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
+1	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
+1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
+1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
+1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91308	.91466	.91621	.91774
+1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
+1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
+1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
+1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
+1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
+1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
+2	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
+2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
+2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
+2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
+2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
+2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
+2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
+2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
+2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
+2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861

As the p value of 0.00494 is less than the significance value of 0.1 we reject the Null Hypothesis. Hence, we can conclude that there is NOT enough evidence to support the idea that vehicle owner in ABC city is 60% or less

Que 4) What is the value of the 99 percentile?

2,2,3,4,5,5,5,6,7,8,8,8,8,9,9,10,11,11,12

Solution:

Formula to determine the Index Position for Percentile Rank:

$(\text{Required Percentile Rank}/100) * (N + 1)$

where N is the number of values in the distribution

Index Position of the 99 Percentile = $(99/100) * 21 = 20.79$

Since there are only 20 values in the distribution we need to consider the value at the 20th Index.

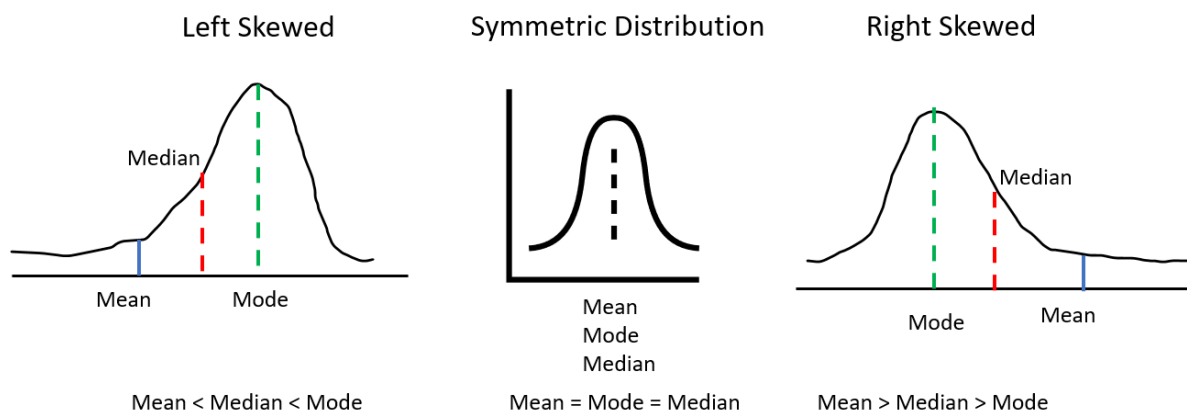
Hence the value of the 99 percentile for this distribution is 12.

Que 5) In left & right-skewed data, what is the relationship between mean, median & mode?

Draw the graph to represent the same.

Solution:

Given below are the diagrams for a Symmetric, Left Skewed and Right Skewed distribution



For a **Symmetric** distribution the **Mean, Mode and Median** are **equal** and will be located at the centre of the distribution.

For a **Left Skewed** distribution, the **Mean** is pulled towards the **left side** due to the outliers in the left side, but the **Median** is only slightly affected by the outlier hence typically the **Mean < Median < Mode**

For a **Right Skewed** distribution, the **Mean** is pulled towards the right side due to the outliers in the right side, but the **Median** is only slightly affected by the outlier hence typically the **Mean > Median > Mode**

For all cases the **Mode** is the **most repeated value** and where the highest peak of the curve occurs and is always at the same position.

The Left or Right Skewed is called a such to indicate the direction that the data is getting skewed i.e., Left Skewed distribution is skewed in the left side and Right Skewed distribution is skewed in the right side.