Question 1

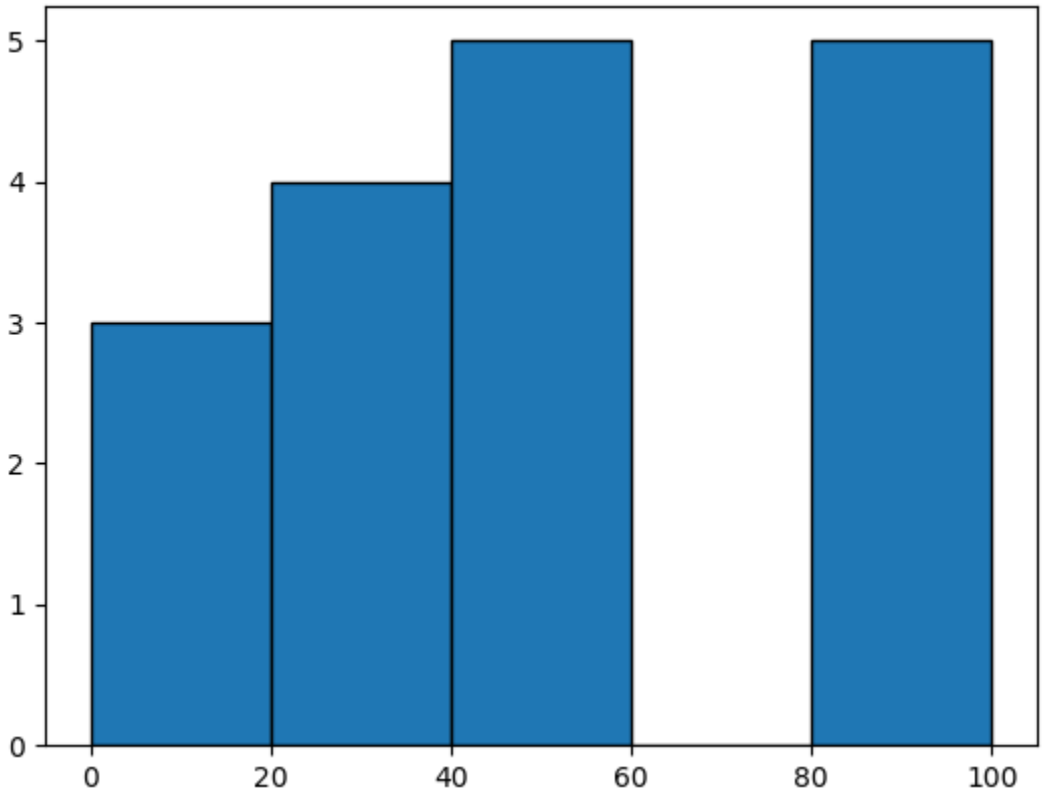
Plot a Histogram with the following: -

No. of Bins = 5

Bin Size = 20

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

Answer: -



Question 2

In the Quant test of the CAT exam, the population standard deviation is known to be 100. A sample of 25 test takers has a mean of 520. Construct an 80% Confidence Interval about the mean?

Answer

$\sigma = 100$, $\bar{x} = 520$, $n = 25$, Confidence Interval = 80%

Calculate Significance level

Significance Value(α) = 1 - Confidence Interval

$$\alpha = 1 - 0.80 = 0.20$$

$$\alpha = 0.20$$

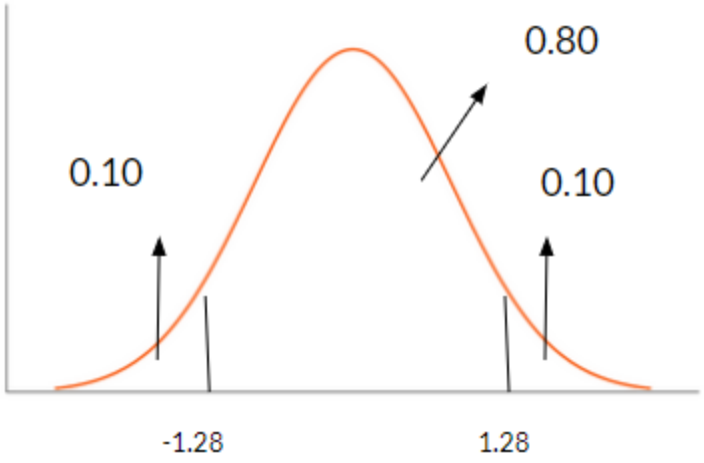
$z\ \alpha/2 = z\ 0.20/2 = 0.10$

$1 - 0.10 = 0.90$

Therefore, from the z table(if population

standard deviation is given)value is 1.28

$z\ \alpha/2 = 1.28$



Lower Fence = $\bar{x}$ - z α/2 ($\frac{\sigma}{\sqrt{n}}$)

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

= 520 - 1.28 x 20

= 520 - 25.6

= 494.4

Higher Fence = $\bar{x}$ + z α/2 ($\frac{\sigma}{\sqrt{n}}$)

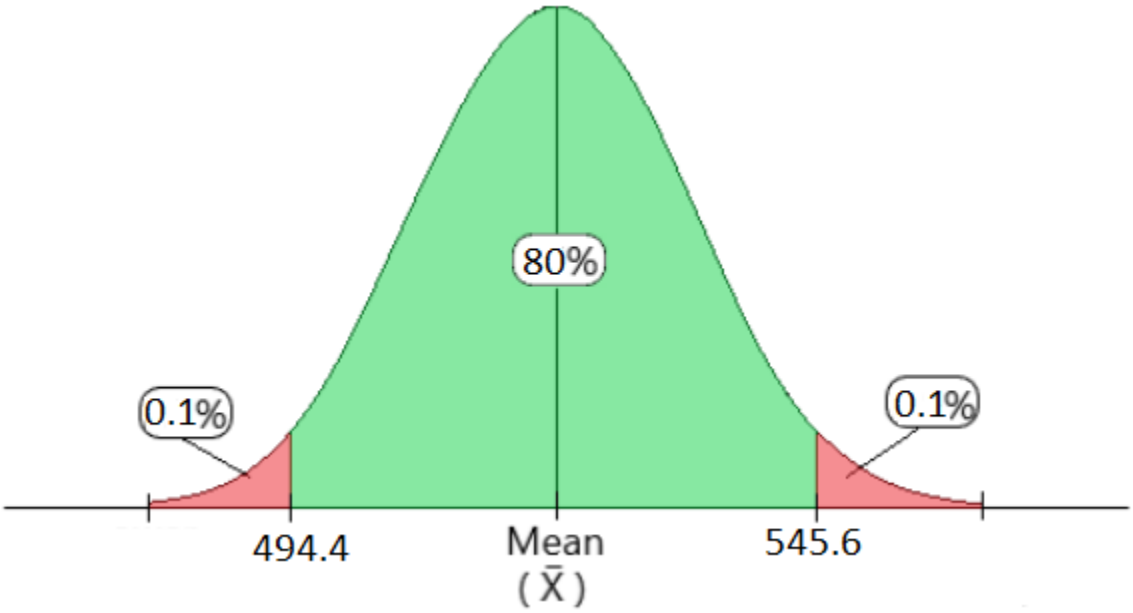
Higher Fence = 520 + 1.28 (100/$\sqrt{25}$)

= 520 + 1.28 x 20

= 520 + 25.6

= 545.6

The Lower and Higher Fence in Distribution



Question 3

A car company 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?

Answer

Step 1: - Null and Alternate Hypothesis

$H_0: P_0 <= 60$

$H_1: P_0 \neq 60$

Step 2: - Proportion

$\hat{P} = x/n$

$\hat{P} = 170/250$

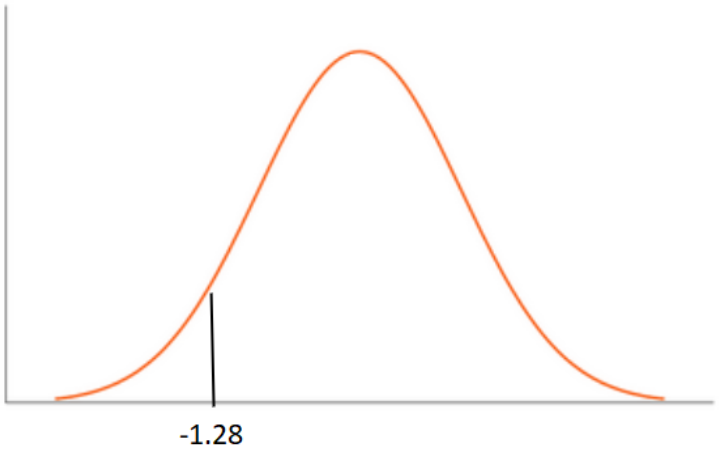
$\hat{P} = 0.68$

Step 3: - Significance Level and Confidence Interval

α = 0.10 i.e. α = 1 - 0.90 = 0.10, So the confidence interval is 0.90 or 90%

z α/1 as its a one tail test, 0.10 /1 = 0.10

From Z table for 0.10 we will get Standard Deviation as -1.28



Step 4: -

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}}$$

$$Z = \frac{0.68 - 0.60}{\sqrt{(0.60)\,(0.40)/250}}$$

Z = 0.08/0.03098 = 2.58

As 2.58 > -1.28, We accept the NULL Hypothesis

Step 5: - **Conclusion**

There is enough evidence to support the idea that vehicle owner in ABC city is 60% or less.

Question 4

What is the value of the 99 percentile?

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

$$\text{Index Value} = \frac{\text{Percentile}}{100} * (n+1)$$

$$\text{Index Value} = \frac{99}{100} \times (20+1)$$

$$\text{Index Value} = 20.79$$

As in the given array, there are only 20 values and the index which we got is 20.79. So we will take the 20th Value as our 99th Percentile i.e. 12

Question 5

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

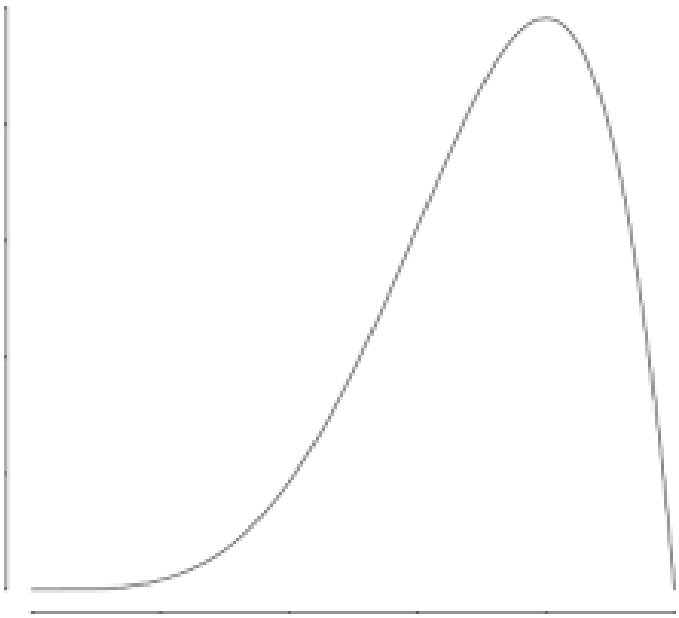
Draw the graph to represent the same.

Answer

What is **Skewness**?

Skewness is a **measure of the asymmetry of a distribution**. A distribution is asymmetrical when its left and right side do not mirror. A distribution can have right (or positive), left (or negative), or zero skewness

Left Skewed Distribution



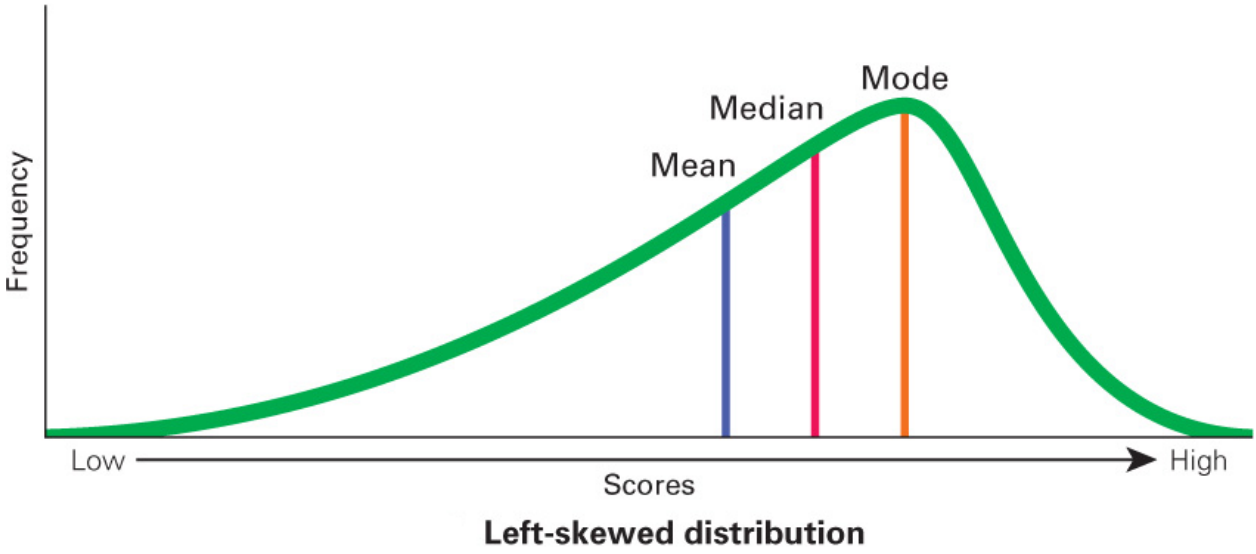
In this type of distribution more data falls on the right-hand side and the distribution becomes asymmetrical. In this type of distribution, the frequency of the data is higher on the right side. Hence, Mode will be more on the right-side of distribution as compared to the median and mean. It is also known as negative skewed.

While median is the middle or center value that lies between mean and mode

The mean is lower than the median and the mode i.e.the average of the data is on the lower side.

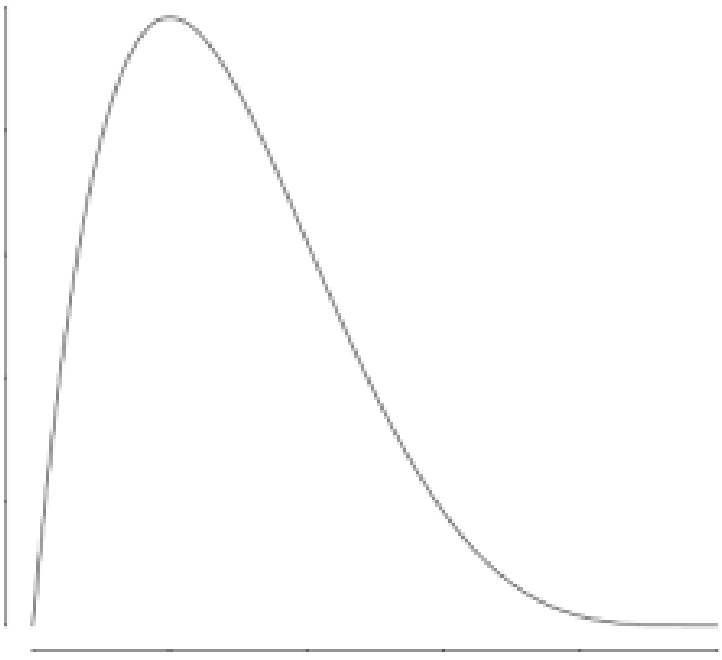
An example of this type of distribution is the Life span of people. Due to advancements in medical facilities, less number of people die at a young age. Hence the left is Skewed while the death rate gradually increases as age increases towards the right and the frequency of death gets more and more i.e. mode becomes highest at the tip of the curve as compared to the median and mean in this distribution.

Thus, Mode > Median > Mean



Left-skewed distribution

Right Skewed Distribution



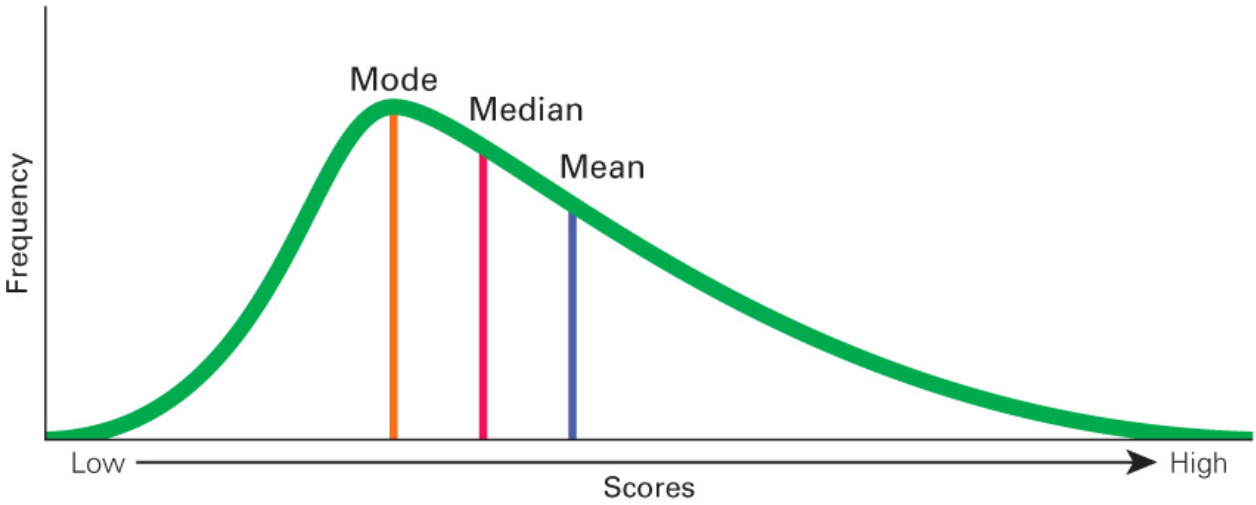
In this type of distribution more data falls on the left-hand side of the distribution and the distribution becomes asymmetrical. In this type of distribution, the frequency of the data is higher on the left side. Hence, Mode will be less as compared to the median and mean.

While median is the middle or center value that lies between mean and mode.

The mean is higher than the median and the mode i.e. the average of the data is on the higher side.

The best example of this type of distribution is wealth distribution. The world's richest people lie on the extreme right and as we move towards the left the income level of individuals decreases and the majority i.e. middle class and lower middle class come on the left tip of the curve while the poor would come on the extreme left of the distribution.

Thus, Mean > Median > Mode



Right-skewed distribution

Additional Questions

Question 6

Why do we use n-1 for sample variance?

Answer

Population dataset: - contains tons of data i.e. millions or even billions of data

Sample dataset: - By using the sampling technique we draw samples from the population dataset and put them in the sample dataset which we consider for analysis and to draw some inferences from it.

$$\sigma^2 = \frac{\sum^{n}_{i=1} (x_i - \mu)^2}{N}$$

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

If we use n in the denominator to calculate sample variance then we are underestimating the true population variance.

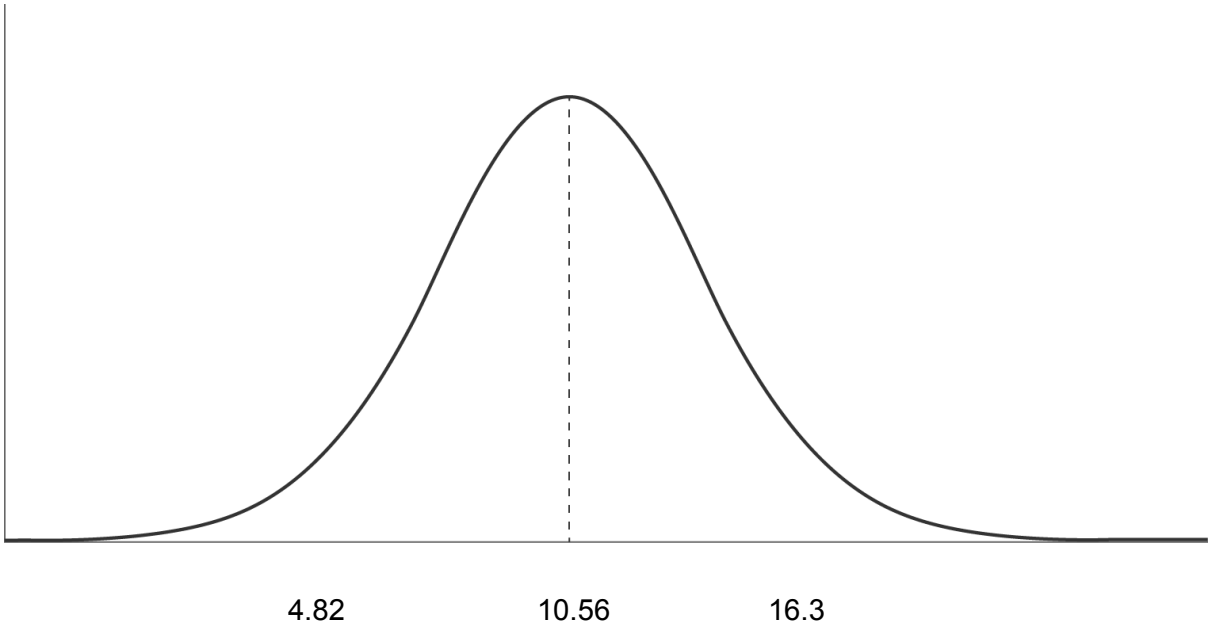
The inferences which will be drawn from the sample variance should be approximately relevant to the population. Thus,

research has proved that taking n-1 will give the nearest inferences.

Let's take an example

If x is a population dataset
x = {2,5,6,8,10,11,15,18,20}

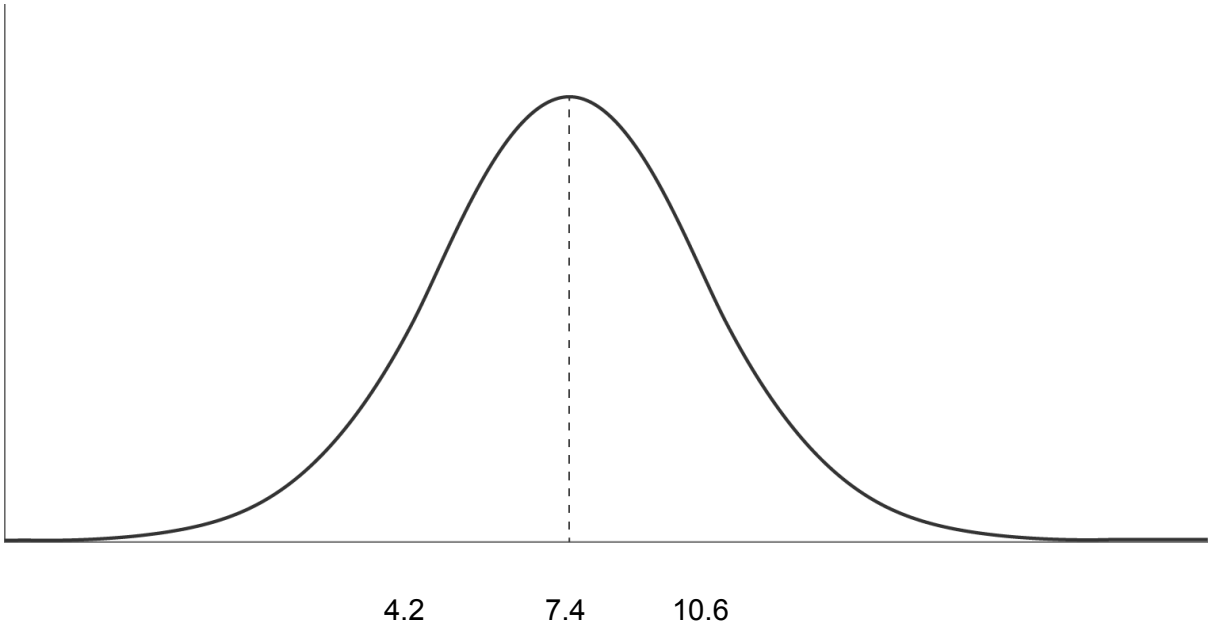
Population mean = 10.56
Population Standard Deviation = 5.74
PopulationVariance=32.91



If we take n in the denominator of the sample dataset, let's see the result
If y is the sample dataset
Y = {2,6,8,10,11}

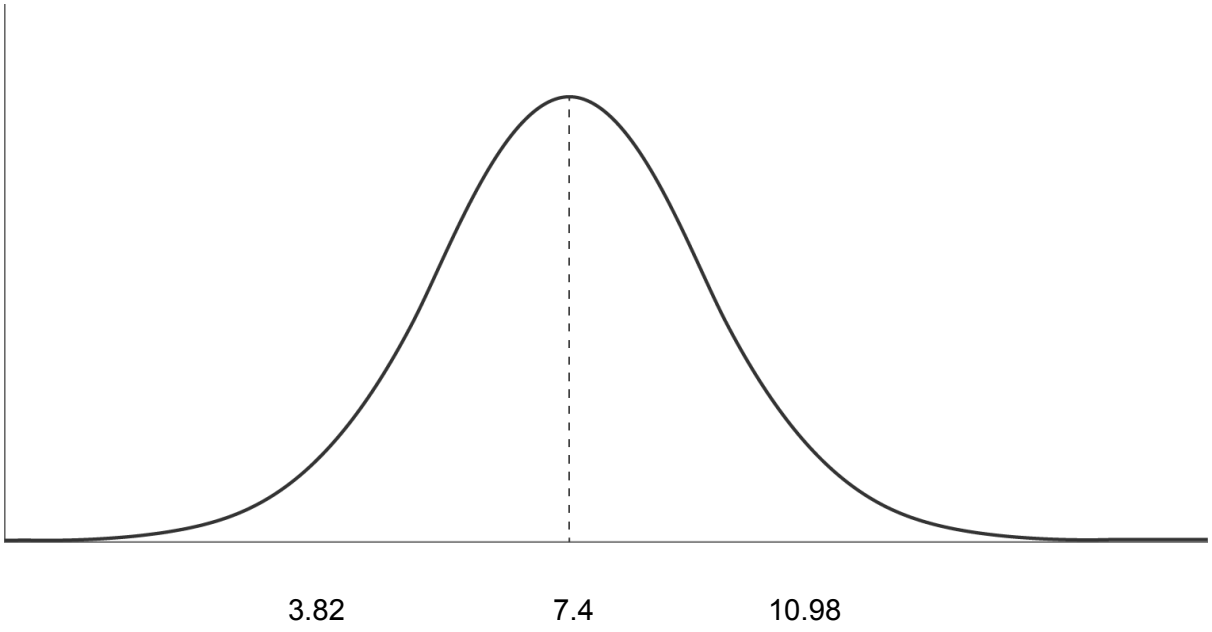
Sample mean = 7.4

Sample Standard Deviation = 3.2
Sample Variance = 10.24



Lets take n-1

Sample mean = 7.4
Sample Standard Deviation = 3.58
Sample Variance = 12.8



Thus, the above example, proved that when we take n-1 we get closer inferences to the population dataset which helps to analyze the data more effectively. In other words, the reason we use n-1 rather than n is so that the sample variance will be what is called an unbiased estimator of the population variance.

Question 7

In an organization, there is a total of 1,00,000 employees. HR Department wants to order XL and L T-shirts for their employees. HR has taken a sample of 500 employees out of which 300 want XL and 200 want L. From this given data HR wants us as a data analyst to tell them the total how many XL and L T-shirts they have to order with a 95% confidence interval?

Answer

n= 500
Confidence Interval = 0.95

Calculate Significance level

Significance Value(α) = 1 - Confidence Interval

= 1- 0.95 = 0.05

z α/2 = z 0.05/2 = z 0.025

1 - 0.025 = 0.975
Therefore, from the z table value is 1.96

z α/2 = 1.96

Proportion for XL T-Shirts

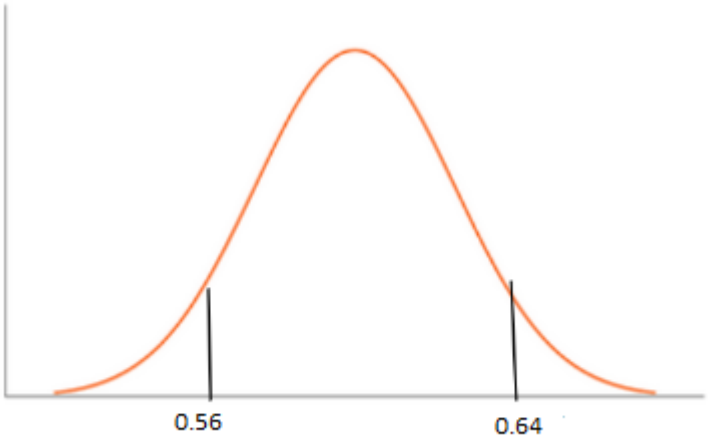
$\widehat{p}$= 300/500 = 0.6

$\widehat{q}$ = 1 - $\widehat{p}$ = 0.4

Confidence Interval of the proportions for XL T-Shirts

Lower Fence = $\widehat{p}$ - z α/2 ($\sqrt{\widehat{pq}}$/n)
= 0.6 - 1.96(0.0219)
= 0.6 - 0.0429
= 0.56

Higher Fence = $\widehat{p}$ + z α/2 ($\sqrt{\widehat{pq}}$/n)
= 0.6 + 1.96(0.0219)
= 0.6 + 0.0429
= 0.64



Proportion for L T-Shirts

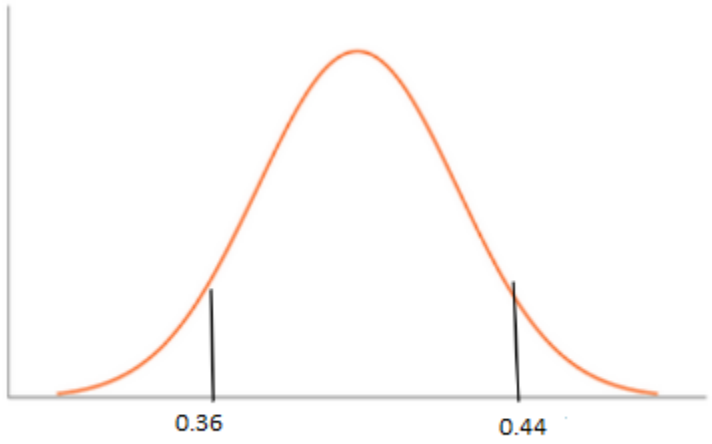
$\widehat{p}$= 200/500 = 0.4

$\widehat{q}$ = 1 - $\widehat{p}$ = 0.6

Confidence Interval of the proportions for L T-Shirts

Lower Fence = $\widehat{p}$ - z α/2 ($\sqrt{\widehat{pq}}$/n)
= 0.4 - 1.96(0.0219)
= 0.4 - 0.0429
= 0.36

Higher Fence = $\widehat{p}$ + z α/2 ($\sqrt{\widehat{pq}}$/n)
= 0.4 + 1.96(0.0219)
= 0.4 + 0.0429
= 0.44



We have found Lower Fence and Higher fence for each type of T-Shirts. In sample the highest number of employees said XL so we will consider the higher fence of XL and lower fence of L

XL T-Shirts to be ordered are
0.64 x total number of employees
0.64 x 1,00,000 = 64000 units

L t-Shirts to be ordered are
0.36 x total number of employees
0.36 x 1,00,000 = 36000 units

Conclusion
As a data analyst, we can say to HR Department that the quantities of XL T-Shirts to be ordered are 64000 units and L T-Shirts to be ordered are 36000 units.