STATISTICS

We will discuss these below usecases after studying Statistics:

USECASE 1: (Asked in a Product Based Company)

Bank: HDFC

HDFC have already two ATM in Location A and B. Distance between location A and B is 30km. They are planning to open a new ATM C, Whether to open ATM in C location or not?

This task is on Data Analyst and Data Scientist.

USECASE 2: (Asked in Amazon for Data Analyst Position)

Find the average size of the shark throughout the world?

USECASE 3: (Asked in Intuit Product Based Company)

Which day of the month should be a Big Billion Day?

LIFE CYCLE OF DATA SCIENCE PROJECT:

Requirement Gathering

Business Analyst, Product Manager, Project Manager will be involved

Data Analytics

Requirements will be provided to data team-->Data Analyst, Data Scientist, Big Data Enggs, Cloud Enggs. To solve the problem, where should we get the data? Data Analyst, Data Scientist, Business Analyst, Product Managers will be involved in this process. They will have domain/business knowledge understanding. This data could be internal data, 3rd party API's, Live Streaming etc. Once we get all the data, data will be sent to Big Data Engg Team. They will save the data in databases (MySQL, NoSQL)

DATA SCIENCE PROJECT STEPS:

- EDA (Exploratory Data Analysis includes Cleaning of Data)

  In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.

- Feature Engineering (E.g. Handling Missing Values)
  - Feature engineering or feature extraction or feature discovery is the process of using domain knowledge to extract features from raw data.
  - Feature Engineering encapsulates various data engineering techniques such as selecting relevant features, **handling missing data, encoding the data, and normalizing** it. It is one of the most crucial tasks and plays a major role in determining the outcome of a model.
- Feature Selection
  - In machine learning and statistics, feature selection, also known as variable selection, attribute selection or variable subset selection, is the process of selecting a subset of relevant features for use in model construction.

- Feature selection is the process by which a subset of relevant features, or variables, are selected from a larger data set for constructing models. Variable selection, attribute selection or variable subset selection are all other names used for feature selection.

- Model Training

  Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.

- Hyperparameter Tuning (To improve the performance of the model)

  - Hyperparameter tuning consists of finding a set of optimal hyper parameter values for a learning algorithm while applying this optimized algorithm to any data set. That combination of hyperparameters maximizes the model's performance, minimizing a predefined loss function to produce better results with fewer errors.

- Deploy the Model

Note:
For Feature Engineering, Feature Selection, Model Training, Hyperparameter Tuning Statistics required everywhere. Hence it is important to know statistics.

WHAT IS STATISTICS

Statistics is the science of collecting, organizing and analyzing the data.

What is Data?

Data is Facts or Pieces of Information.

e.g. Salary of the employees, Designation of the Employees

TYPES OF STATISTICS:

DESCRIPTIVE STATISTICS:

It Consist of Organizing and Summarizing the Data using different kind of plots:

Mean-Median-Mode, Standard Deviation, Variance etc.

Measure of Central Tendency (Eg. Plotting Using Pie-Charts, Histogram, Bar Chart,

Distribution (We draw using Histogram, Candlestick, Box Plots, Whisker Plot, Skewed Plots, Scatter Plots, Violin Plots). This will be used extensively in EDA and Feature Engineering.

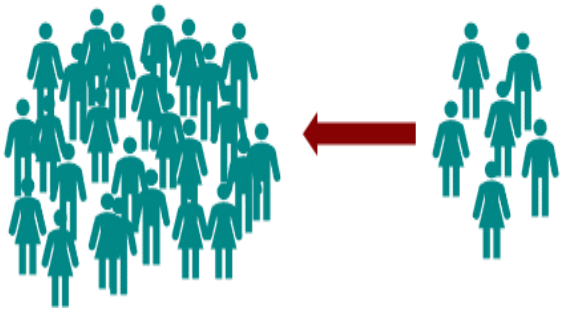
INFERENTIAL STATISTICS:

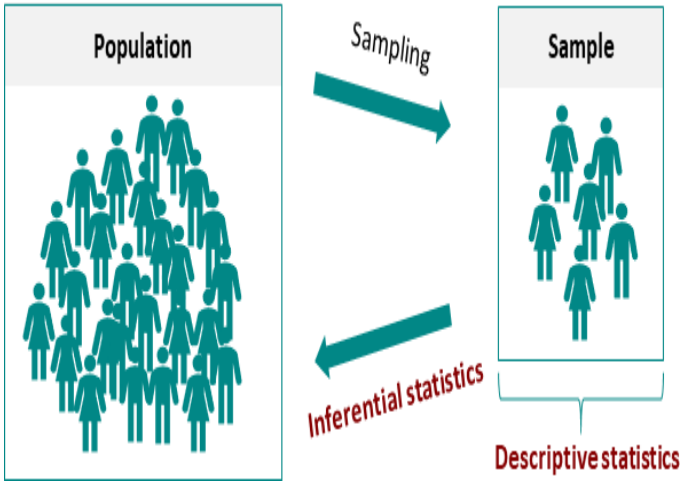
It Consist of Collecting Sample Data and **making conclusions** (Using Hypothesis Testing) about Population Data using some experiments.

Let's say there is a university with 500 students. Class A has 60 people. By using Age data of these 60 people (sample), can we make conclusions of Average age of entire university of 500 students ? (YES) We will be using Technique of Hyothesis Testing, Confidence Interval, P Value, Experiments like Z Test, T Test, Chi Square Test, Anova or F Test.

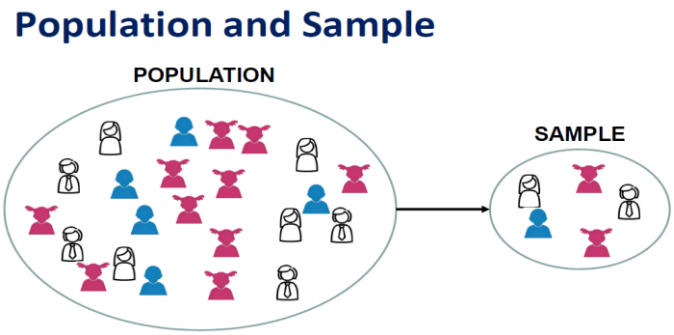
Inferential statistics:

Testing of statements about the population on the basis of sample characteristics.





Sample Data vs Population Data:



E.g. -Punjab Population 10 Crore

Exit Polls conducted: Asking people whom did you vote and make conclusions. They don't ask each and every people as it is not practically possible. So they need to take sample of data.

E.g. -Lets say there are 20 Classrooms in a University and you have collected Ages of students in 1 Classroom:

Ages: {21,20,18,34,17,22,24,25} Weight: {.....}

What Descriptive Stats Qns can come:

What is the Average Age for that particular Class?

Relationship between Age and Weight?

What Descriptive Inferential Stats Qns can come:

Are the average age of the students in the classroom less than the average age of the students in the university?

1000 Students->Class A 50 girls (95% Avg Marks) and Class B 50 boys (92% Avg Marks). Have girls performed better than boys in entire University ?

DIFFERENT SAMPLING TECHNIQUES:

- **Simple Random Sampling:** Every member of the population (N) has an equal chance of being selected for your sample (n)
  E.g. Exit Polls, General Survey, Lottery

- **Stratified Sampling:** Strata means Layers/Clusters/Groups: E.g. We can make groups basis Gender, Education Degree etc.

- **Systematic Sampling**: (In Airport, Credit Card Section wants to sell credit card: Suppose Two person are selling credit card, 1st individual said Every 5th person I will see I will approach for credit card, 2nd individual said Every 9th person I will see I will approach for credit card).
  Basically selecting every nth individual out of Population (N)
- **Convenience Sampling:** Only those who are interested in the survey will only participate.
  E.g. Data Science Survey

E.g-Survey Regarding New Technology - Convenience Sampling (bcz those who are interested in technology)

E.g.-RBI Survey:(specific survey for women who are taking care of entire house)- Stratified sampling(bcz survey for married women) then random sampling

E.g-Selling Credit Cards : We can apply Stratified Sampling (credit card directly based on salary) first and then on that sample, we can apply Random Sampling

Variable:

Variable is a Property that can take any values E.g. Age = 14

Variables will have multiple values (Collections): Ages = [24,25,26,27]

Two Different Type of Variables:

- **Quantitative Variable**: Measured Numerically. We can do mathematical operations: Eg. Age, Weight, Height, Rainfall in cm This can be further divided into Discrete and Continuous Variables:
  - Discrete (Cannot be Decimal): Number of Bank Account Person can have, No. of children in family
  - Continuous (Decimal allowed): Height, weight, ages, rainfall, temperature, distance
- **Qualitative Variable**: Categorical Variables: E.g. Based on Some Characteristics, they are grouped together.

What kind of Variable is :

- Marital Status? -Categorical Variable
- Ganga River Length? -Continuous
- Movie Duration? -Continuous
- Pin code? -Discrete
- IQ? - Discrete
- Gender? -Categorical

HISTOGRAM:

A frequency distribution shows how often each different value in a set of data occurs. A histogram is the most commonly used graph to **show frequency distributions**. It looks very much like a bar chart, but there are important differences between them.

Steps to plot Histogram:

1. Sort the Numbers
2. Create Bins - How many no. of groups you want to create
3. Bin size- size of bins

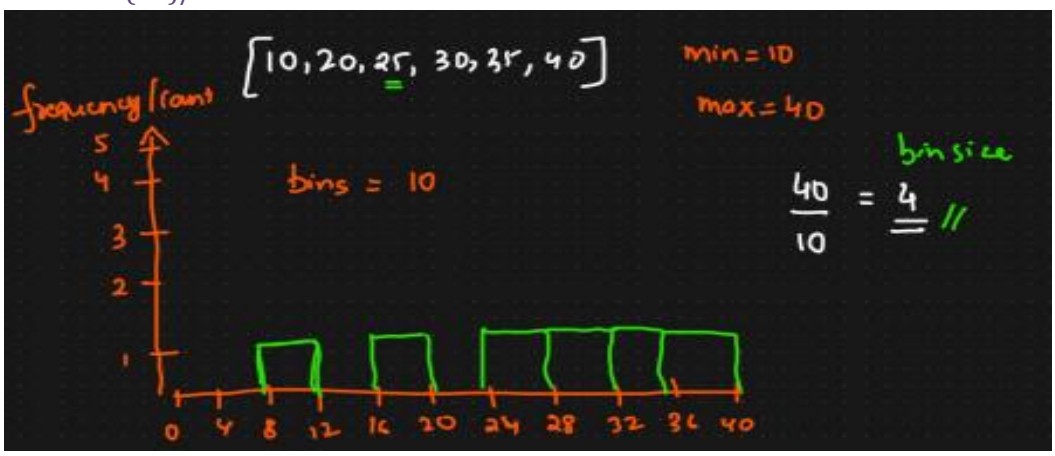
E.g - [10,20,25,30,35,40]

Min=10

Max=40

Bins=10 (I want to devide the group in to 10 division), depending upon us we can define the bin size)

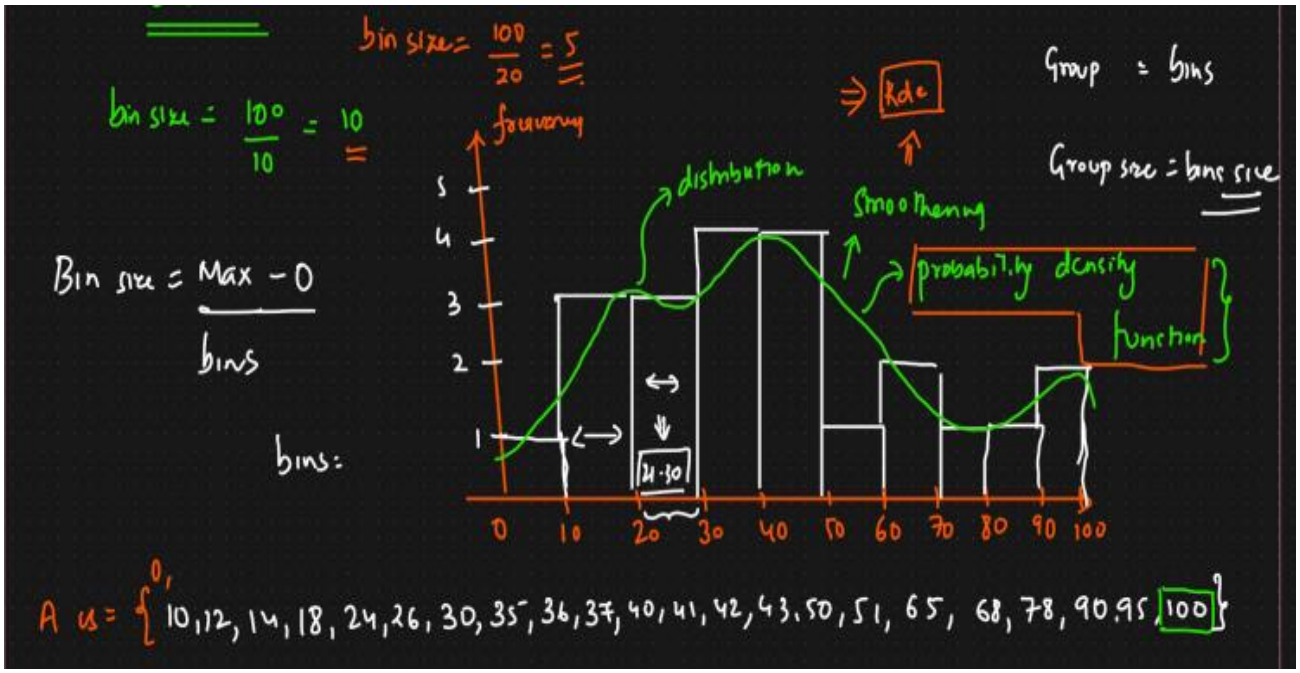
Bin size =(40)/10=4



E.g. - Ages={0,10,12,14,18,24,26,30,35,36,37,40,41,42,43,50,51,65,68,78,90,95,100} (sorted)

Bins=10 (depending upon us we can define the bin size)

Bin size=100/10=10



E.g. - Ages={0,10,12,14,18,24,26,30,35,36,37,40,41,42,43,50,51,65,68,78,90,95,100}

If the bin is 20 then what is the bin size

Bins=20 (depending upon us we can define the bin size)

Bin size=100/5=20

Weight={30,35,38,42,46,58,59,62,63,68,75,77,80,90,95}

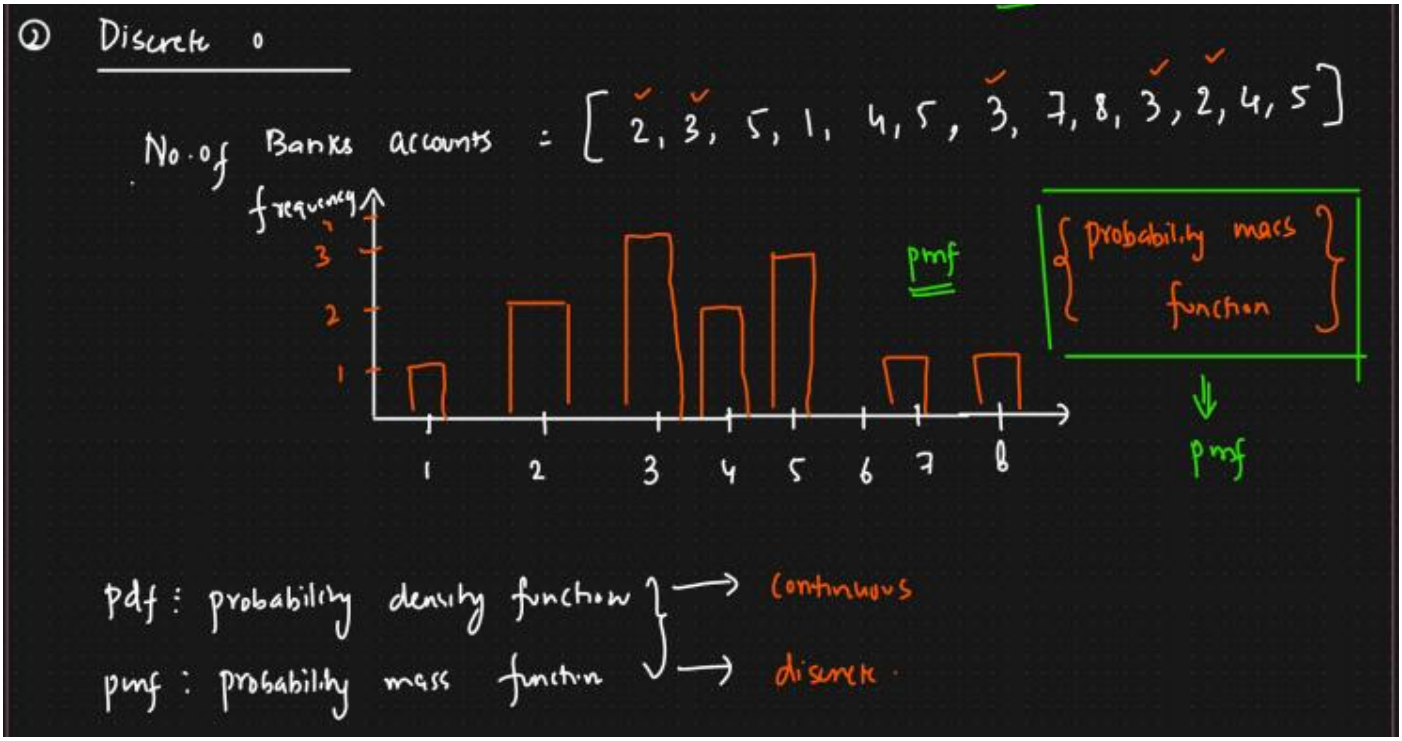
Bins=10

Bin size=95-30/10=65/10=6.5

Whatever we are constructing for histogram that is only for continuous value.

Discrete:

No. of bank accounts=[2,3,5,1,4,5,3,7,8,3,2,4,5]



Pdf: Probability density function- (Used for Continous Value)

To smoothen the histogram for Continous value we will use Probability Density Function.

Pmf: Probability Mass Function – (Used for Discrete Value)

To smoothen the histogram for discrete value we will use Probability Mass Function.

Measure of central Tendency:

A measure of central tendency is a single value that attempts to describe a set of data identifying the central position.

Mean:

E.g. X = {1, 2, 3, 4, 5}

Average/Mean = (1+2+3+4+5)/5 = 15/5=3

Some Notations:

Population (N), Sample Mean (n), Population Mean (μ), Sample Mean ($\bar{x}$)

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

E.g. Population Age = {24, 23, 2, 1, 28, 27}

N=6

Population mean=(24+23+2+1+28+27)/6 = 105/6=17.5

Let's consider randomly I have picked 4 no. of values from population age that is called sample age.

E.g. Sample age = {24, 2, 1, 27}

n =4

Sample mean = (24+2+1+27)/4 = 54/4=13.5

Note:

N>n always reverse is not possible.

But μ >= x̄ and x̄ >= μ both is possible.

> Note:
> N>n always reverse(n>N) is not possible. But μ >= x̄ and x̄ >= μ both is possible

Practical application: (Feature Engineering)

-Practical application of mean , median and mode actually used in feature engineering.

Mean:

- let's take Age, salary, family size. In this data there is so many nan values are there. If we will delete particular row that contain Nan value, there will be loss of info. So instead of doing that we can replace the Nan value of age /salary/family size with may be mean/median/mode. Mean basically represents the central position.

Age	Salary	Family Size
24	1000	3
28	2000	4
29	nan	5
nan	3000	nan
45	4000	4
nan	500	nan
23	nan	7

Let's take an another example

Age	Salary
24	45
28	50
29	Nan
nan	60
31	75
36	80
nan	Nan

-Here we can replace the nan with 29.6 and 62 for age and salary respectively. If we add one more in age=80 and salary = 200. Then the mean value will be changed to 38 for age and 85 in case of salary. Because of the outlier the mean value has changed. To prevent this we specifically used median.

Median:

Steps to find out median

1. Sort the number
2. Find the central no.
   -(if the no. of elements are even we find the average of central elements)
   -(If the no. of elements are odd we find the central element)

E.g.- {1,2,3,4,5,6,7,8,100,120}

n = 10(even)

so median= 5+6/2=5.5

E.g.- {0,1,2,3,4,5,6,7,8,100,120}

n = 11(odd)

so median =5

Mode: (The most frequent occurring element)

E.g. – {1, 2, 2, 3, 3, 3, 4, 5} hence mode=3

E.g. – {1, 2, 2, 2, 3, 3, 3, 4, 5} hence mode=2,3

Let's discuss about the practical usage of mode.

Suppose in my data set categorical values are there.

Types of Flower
Lily
Sunflower
Rose
Nan
Rose Sunflower
Rose
Nan

So in this case the nan value can be replace by rose. As rose is the most occurring value here. Mode basically used for categorical value

Note:

So whenever there is a outlier we replace the nan value with median.

When there is no outlier we will replace the nan value with mean.

Mode generally used for categorical value.

Measure of Dispersion:

1. Variance (σ2) – spread of data
2. Standard deviation (⍰)

Variance(σ2):

Population Variance (σ2) , Sample Variance(s^2)

Population Variance	Sample Variance
$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
σ^2 = population variance x_i = value of i^{th} element μ = population mean N = population size	s^2 = sample variance x_i = value of i^{th} element $\bar{x}$ = sample mean n = sample size

(xi−μ) means distance form mean.

Lets give an example:

X = {1,2,3,4,5,6,7,8,9,10}

Y= {1,2,3,4,50,60,70,100}

From above two in which of the data set variance is more?

The variance is more for y. as we know that variance is nothing but spread of the data and in 2nd data set it ranges between 1 to 100. If we plot the histogram for both the data set then the spread will be more for 2nd one only.

A={1,2,3,4,5} μ = 3

σ2 = [(1-3)²+(2-3)²+(3-3)²+(4-3)²+(5-3)²]/5 = 2

B={1,2,3,4,5,6,80} μ = 14.4

σ2 = [(1-14.4)²+(2-14.4)²+(3-14.4)²+(4-14.4)²+(5-14.4)²+(6-14.4)²+(80-14.4)²]/7 = 719.10

variance of B > variance of A

Note: Hence as variance increases spreads are also increases.

Standard Deviation (√σ2):

Standard deviation tells that how many Standard deviation away a no. in the distribution falls from the mean.

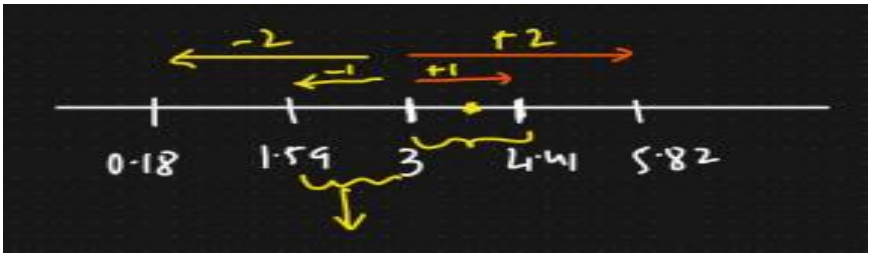
E.g. **{1,2,3,4,5}**

$\mu = 3$

$\sigma 2 = [(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2]/5 = 2$

$\sigma=\sqrt{2} = 1.41$

$\mu = 3$, $\mu+ \sigma = 4.41$, $\mu+ 2\sigma = 5.82$, $\mu- \sigma = 1.59$, $\mu- 2\sigma = 0.18$



Between two graphs for which graph variance and standard deviation will be high?

Ans : 2nd graph

Percentile and Quartiles:

Percentage = {1,2,3,4,5,6,7,8}

Percentage of even no. = No. of even numbers/ Total No. of Number =4/8 =0.5 = 50%

Percentile:

Percentile is a value below which a certain percentage of observations lie.

99 Percentile means: It means the person has got better marks than 99% of the entire students.

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

What is the percentile rank of this particular value 10?

Percentile rank of x = No. of value below x / n = 16/20 =80 Percentile

What is the percentile rank of this particular value 8?

Percentile rank of y = No. of value below y / n = 9/20 =45 Percentile (for all 8 the percentile will be same)

What is the percentile rank of this particular value 6?

Percentile rank of z = No. of value below z / n = 7/20 =35 Percentile (for all 6 the percentile will be same)

What is the percentile rank of this particular value 9?

Percentile rank of a = No. of value below a / n = 14/20 =70 Percentile (for all 9 the percentile will be same)

What is the value that exists at 25 Percentile?

Value = (Percentile/100) * (n+1)

= (25/100) * 21

= 5.25=5th index (index starts from oth position)

What is the value that exists at 95 Percentile?

Value = (Percentile/100) * (n+1)

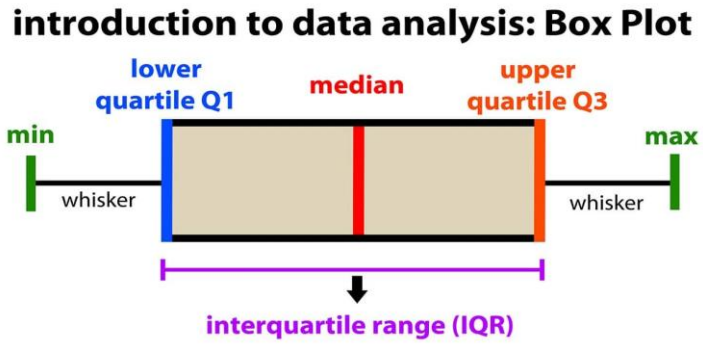
= (95/100) * 21

= 1995=20th index (index starts from 0th position)

5 No. Summary:

1. MINIMUM
2. FIRST QUARTILE (25th Percentile)(Q1)
3. MEDIAN (Central Element)
4. THIRD QUARTILE (75th Percentile)(Q3)
5. MAXIMUM

All the above calculation is done to remove the Outliers. For this we have to create a box plot. By using this we can create a box plot.



X = {1,2,2,2,3,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27}

To create the box plot we try to create a fence. Our value will be between **Lower Fence** and **Higher Fence.**

IQR = Q3 - Q1

Lower fence = Q1 – 1.5(IQR)

Higher fence = Q1 + 1.5(IQR)

Q1(25th Percentile) = (25/100)*(n+1)

Q3(75th Percentile) = (75/100)*(n+1)

Q1(25th Percentile) = (25/100)*(n+1)

= (25/100)*(20+1) = 5.25 Index = 3 (avg. of 5th and 6th index)

Q3(75th Percentile) = (75/100)*(n+1)

= (75/100)*(20+1) = 15.75 Index = 7.5 (avg. of 7th and 8th index)

Lower fence = Q1 – 1.5(IQR) = 3-1.5(4.5) = -3.65

Higher fence = Q1 + 1.5(IQR) = 7.5+1.5(4.5) = 14.25

5 No. summaries:

1. Minimum = 1
2. Q1= 3
3. Median =5
4. Q3 = 7.5
5. Maximum = 9