Statistics: it is the discipline that deals with the collection , organization, analysis, representation and presentation of data.

Problem:

You want to identify the loan defaulter:

1. Gather the data

2. Preprocess the data

3. Analysis

4. Representation

5. Presentation

Data is divided into two types:

* Categorical data
  + Also called Qualitative Data

Example – Name, City, Gender

* Numerical Data
  + Aka Quantitative Data

Example-Salary, age, distance

Numerical data:

1. Continuous Data
   1. Float type- 60.5kgs, height-153.5cm

Measurable data

1. Discrete Data
   1. Int type- phone number, age, num of kids, books

Countable data

Example:

Emp working in an org: Categorical data

Num of emp working in an org: numerical (discrete data)

Emp work exp in an org: numerical (continuous data)

* In data set we have columns and rows available
* Cols: variables, features, predictors, dependent and independent variable
* Rows: values, samples, tuples, fields, observations

Data is divided into 4 parts or level:

Nominal level [Categorical]

Ordinal level [Categorical]

Interval level [Numerical]

Ratio level [Numerical]

Nominal level:

* It is under categorical data.
* It is called as lowest level of data.
* It means simply names (there is no relation between the names and there is no meaning between the names Ex- Ramesh, ram, hydrabad , bbsr )

Ordinal level:

* It is also under categorical data.
* Ordinal means we can see some order between the names
  + Ex- flop->avg->hit->superhit->blockbuster
  + Small->medium->large
  + Fail->pass
  + Single->married->divorced
  + Primary->secondary->tertiary

Interval level:

* It is under numerical data.
* It does not have zero scale.
  + It means it has positive and negative values (starts from -infinity to +infinity).
  + Ex- temperature, marks

Ratio level:

* It is under numeric type data.
* It has zero scale.
  + It means it has positive data.
  + Ex= age, length, height, weight

Question:

The temp in hyd. is =50 C , blr = 25 C

The temp hyd. =2\*25 C

* hydtemp / blrtemp = 2/1

50 C= 122F

25 C=77 F

Now 50c/25c ≠ 122f/77f

Since the ratio is different that’s why we can’t categorise temp in ratio level , we have to put it under interval level.

To become ratio level data it needs to be of same ratio in every scales.

Weight is a ratio level data.

W1= 5kg, w2=10kg

If we change some unit

5kg/10kg=5000g/10000g=1/2

Since ratio is same for all units we can call weight in ratio level data.

|  |  |
| --- | --- |
| Numerical | Categorical |
| Quantitative | Qualitative |
| Continuous, Discrete | --------- |
| Interval, Ratio | Nominal, Ordinal |
| ex-int, float | ex-string, objective(yes/no) |

Population vs sample

* Populations- whole data
* Sample – some part of the data
* Working with whole data is tough, it increases the time complexity, resource, money that’s why we choose to work on sample data.
* We work on sample and will estimate the output on population this is called inferential statistics.
  + Infer = estimation

Data representation: categorical, numerical

Categorical:

Statement: in a class 20 boys 30 girls how can we display it in tabular representation

Table:

|  |  |
| --- | --- |
| Gender | Count |
| Boys | 20 |
| Girls | 30 |

Gender === class === category

Count=== class frequency=== num of ocrence

Frequency table

|  |  |
| --- | --- |
| class | Class frequency |
| Boys | 20 |
| Girls | 30 |

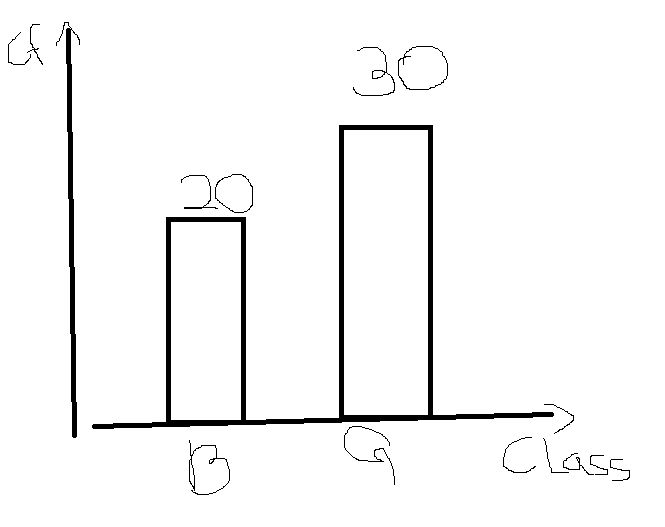
* Frequency table has 2 cols:
  + class-> categorical col
  + class frequency->numerical col

Graphical representation:

* in order to plot frequency table we will use bar chart or bar plot
* bar chat will take x-class(categorical), y-frequency(numerical).

NOTE:

To plot a bar plot we need a categorical col and numerical col.



RELATIVE FREQUENCY TABLE:

* relative frequency will give the percentage of the class members.
* In the above ex- boys are 20 out of 50 and girls are 30 out of 50
* So, the percentage of boys=40% and girls=60%

|  |  |  |
| --- | --- | --- |
| class | Class frequency | Relative Frequency |
| Boys | 20 | 0.4=40% |
| Girls | 30 | 0.6=60% |
| total | 50 | 1=100% |

GRAPHICAL REPRESENTATION:

To draw relative frequency we will use pie chart.

|  |  |
| --- | --- |
| Tabular | Graphical |
| Frequency table | Bar chart(x-categorical,y-numerical) |
| Relative frequency | Pie chart |

Numerical col representation:

In class 10 members are there

Raw data: marks

5,8,12,15,20

6,9,11,4,17

* We can make raw data into some grp and
* we can get count of each grp.

Tabular representation:

Frequency distribution table.

* It has two col.
* 1col== class interval
* Another col== class interval freq

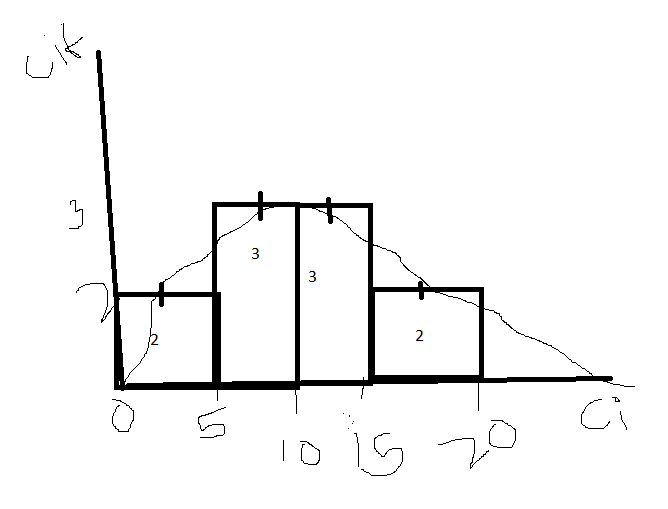
|  |  |
| --- | --- |
| Class interval | Interval freq |
| 0-5 | 2 |
| 5-10 | 3 |
| 10-15 | 3 |
| 15-20 | 2 |

If u want to add 5 in first interval, then don’t add in second interval.

Graphical representation:

Histogram:

* It is graphical representation of numerical data
* X- class interval
* Y- interval frequency
* Numerical vs numerical



* Whenever u see any distribution plot
  + Immediately u need to recall histogram
* Whenever u think of histogram
  + U need to recall CI vs CIF (frequency distribution table)
* Whenever u think of frequency distribution table
  + Immediately u need to recall raw data.

Distribution === histogram=== fdt=== ci vs cif=== raw data

How to choose intervals:

* What happens if I choose less intervals.
  + The minute info will miss.
  + We cannot discriminate the exact categories
* What happens if I choose more intervals.
  + Unwanted information will come.

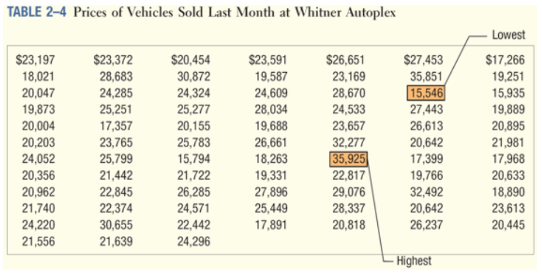
Then refer statistic 2nd ppt.

How to create class interval and class width.

From raw data === f d t

Class interval and class frq

It is very imp how to choose class interval and width of the interval.



|  |  |
| --- | --- |
| K | 2k |
| 0 | 20 =1 |
| 1 | 21 =2 |
| 2 | 4 |
| 3 | 8 |
| 4 | 16 |
| 5 | 32 |
| 6 | 64 |
| 7 | 128 |

In which values 80 occurs

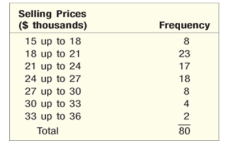
Bet k=6 and k=7

We can not take 6 because 64<80

So we need to consider 7

The number of intervals are k=7

Interval width = =



23+17+18=(58\*80)/100=72%

That means between 18-27 k 72% of car price range

Data measurements==:

* Central tendency
  + Mean
  + Median
  + Mode
* Data dispersion
  + Ragne
  + Mean deviation
  + Absolute mean deviation
  + Variance
  + Standard deviation

Central tendency:

Mean:

* Median
  + Median also a metric tell about mid point of the data
  + 50 percentile of data

1,2,3,4,5====3 is the median

Mean vs median=

1,2,3,4,5

Mean=15/5=3

Median=3

1,2,3,4,5,200

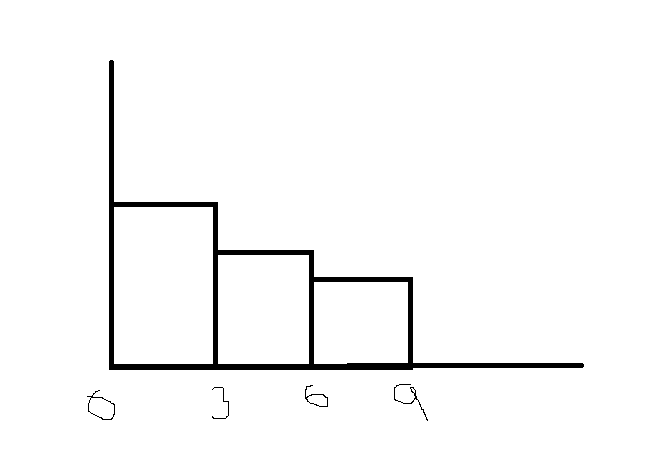
Avg=215/6=35.8

Median=3.5

If a data has very huge value or very less value mean will be affected but median is not effected. Those very high or low values are called outliers.

* Mode: most repeated value.
  + Most frequently occurred value.
  + 1,5,6,7,1,6,1,1,8🡺 mode=1
  + Data distribution:

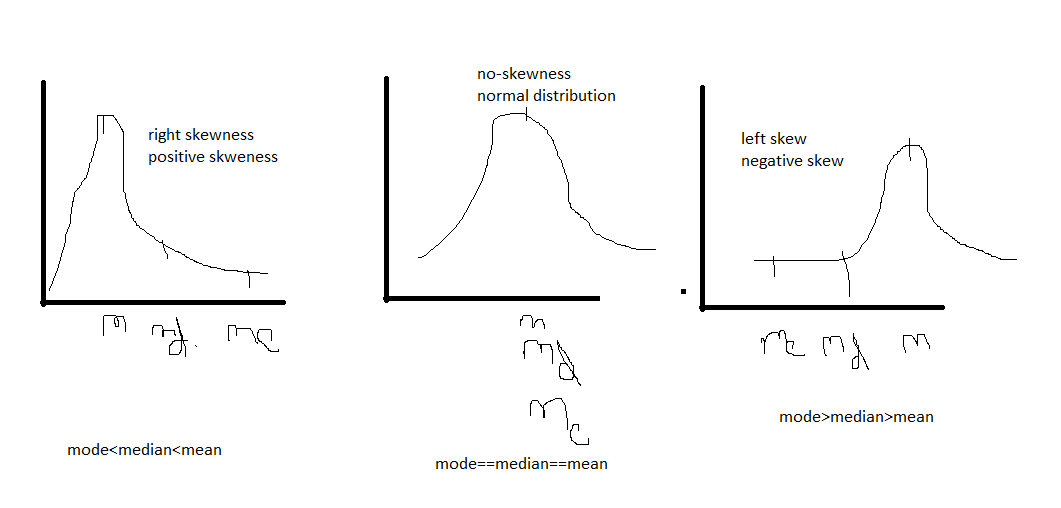
|  |  |
| --- | --- |
| * + Class interval | * + Frequency |
| * + 0-3 | * + 6 |
| * + 3-6 | * + 4 |
| * + 6-9 | * + 2 |

* + Data distribution highest peak is called mode 🡺 mode is available at that point. we cannot say what is the mode only by looking at the diagram.
  + 
  + We know that distribution forms histogram
  + And histogram formed from interval
  + If u see highest peak in the distribution means that corresponding interval has the mode value.

Mean-will give avg of the data

Median – will give middle value, 50 percent of the data

Mode- will give the highest peak in the distribution.

**

Left skewed or neg skewed:

* Because of negative outliers
* Mode>median>mean
* Assume that data ranges 0,100.
* Neg side data is pulling which means= mean 0 side
* So that mean value is low.

Right skewed or pos skewed:

* Because of positive outliers
* Mode<median<mean
* Assume that data ranges 0,100.
* Pos side data is pulling which means= mean 100 side
* So that mean value is high.

No skew or normal distribution:

* No outliers
* Mean==mode==median
* Equally distributed data
* Bell shape curve.

What is the meaning of skew ==== pulling

Reason for skewness ==== outliers

The outlier is which side:

Right side means (max values)

Left side means (min values)

Are you feeling low???? == keep calm trust omkar sir.

Data dispersion: