Statistics = 1

Use Case

Product Based Company

HDFC Bank

(A) Atm (C) Atm (B) Atm

Q. Should we need to open new Atm (C) or not?

2) Find the average size of the Shark throughout the world?

3) Amazon Big Billion Day Sale ~~Init~~ In it?
which month Should you select?

Life Cycle of Data Science Project?

1) Requirement gathering
   - ① Product managers / Project managers
   - ② Business Analyst

2) Data Analysts team
   1. Data Analyst
   2. Data scientist
   3. Big Data engineers
   4. Cloud Engineers

Where we get data

1) Internal DataBase
2) 3rd Party API's
3) web scraping

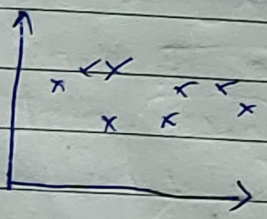
Sent Data to

Big Data Eng → (1) mysql data base
(2) No sql data base

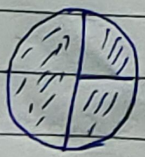
Now start life cycle of DS project

1) EDA
2) FE
3) feature selection
4) Model Training
5) Hyper parameter tuning.
6) Deploy

Analysis of Data



Descriptive stats
=> Summarising of Data



Descriphive Stats

Age : [12, 13, 14, 18, 20, 25] = Avg of Age

↓

Descriptive stats

↓

Measure of Central Tend

Statistics ① Definition => Statistics is the science of collecting, organising and analysing the data.

Data = Facts or Pieces of information.

Eg: Ages of students in Class room

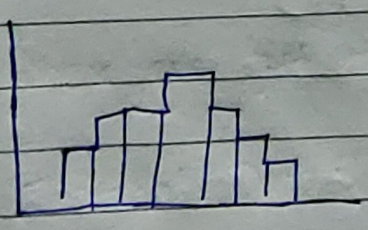
[24, 25, 32, 29, 28] => mean, median, r
standard deviatio

Types of statistics

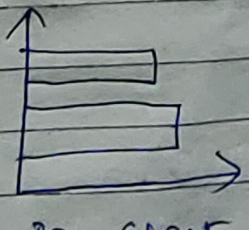
Descriptive stats

Inferential statics

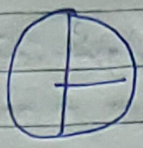
① It consists of organising and Summarizing the data



Hist

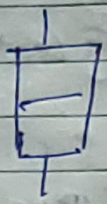


Bar Chart



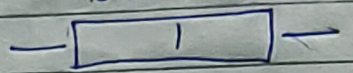
Pie

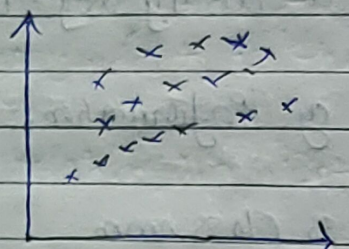
Distribution



Candle plot

Box plot





Scatter plot

(2) Inferential Stat

(1) It Consists of Counting Sample data and making Conclusion about population data using some experiments

→ Hypothesis testing

Example

University

Class A → 60 people

↓↓

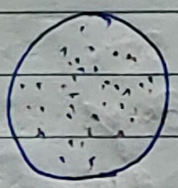
Sample data => Age => Average age of the uni

C.I = Confidence Interval

1) Z -est
2) t test
3) Chi square test
4) F test

Sample data → Conclusion By Hypothesis testing → Population Data

Sample Data vs Population Data



Population Data
total data is population

Sample data
Some ~~of~~ random sample we will take
population.

Ex: Let's say there are 20 classrooms in a university and you have collected the age of students in one classroom.

Age [21, 20, 18, 34, 17, 22, 25, 26, 23, 22]

weight [. . . . .]

Descriptive stats :- 1) What is the average age of student in the classroom?

2) Relationship Between Age and weight?

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

Greater & Equal

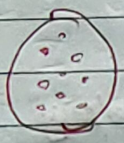
1000 students	50% Girls	50% Boys
	95%	90%

Population (N) Sample (n)

Date
Page

Q Sampling techniques : (7)

1) Simple Random sampling : Every number of the population (N) has an equal chance of being selected for your sample (n)

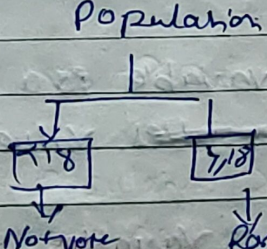


Exit poll → Random Sampling
General Survey

2) Stratified sampling
Strata → layer → clusters → Groups

Gender [male, female]

Education [High School, master, phd]



3) Systematic sampling → [Airport]
[Credit Card] Sale persons

→ select every nth individual out of population (N)

Convenience Sampling => only those who are interested in the survey will only participate.

{Data science survey → General AI survey}

① Survey Regarding New technology
↓
convenience Sampling

② RBI ~~-~~ survey → [married women] => Stratified + Random Sampling

③ Credit Card => Stratified + Random Sampling

① Variable :- A variable is a property that can take any ~~variables~~ values.

Eg : age = 14
age = 25
age = 68

Variables
Ages = [24, 25, 26, 27]
↓
Collection

Two different types of variable

① Quantitative variables → measured numerically [mathematical operation]

Eg: Age, weight, height, rainfall (cm), temp

2) Qualitative variables → Categorical variables
Because ~~~~ Based on some characteristics they are group together.

Eg: Gender, types of flower, types of movies.

Quantitative variables
(1) Discrete variables → whole number
(2) continuous variables → continuous number

Discrete variables
Eg: No of children → whole
Eg: No of Bank Account.
[1, 2, 3, 4)

Continuous variables
Eg: Height, weight, money in account.

Assessment in variables?

1) what kind of variables is marital status?
→ Categorical variables

2) what kind of variables is Ganga river levels?
→ Continuous variable

3) What kind of variables is movie duration?
→ Continuous variable

4) What kind of variable is Pincode?
→ Discrete variable

5) What kind of variable is IQ?
→ ~~Continuous variable~~. discrete variable

Statistics 2

Agenda

1. Histogram
2. Measure of central Tendency
3. Measure of Dispersion
4. Percentiles And Quartiles
5. 5 Number Summary (Box plot)

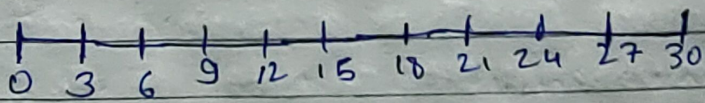
1. Histogram

Ages : [10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40]

Min = 10 Max = 40

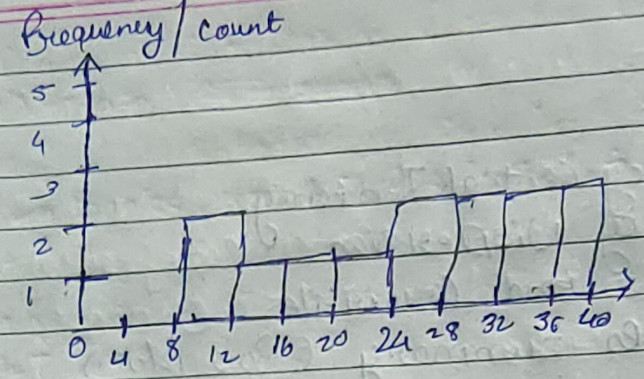
1. Sort the Numbers
2. Bins → no. of groups
3. Bins size → Size of Bins formula = $\frac{Max}{Bins}$

Bin = 10 Included 1st no. $= \frac{40-10}{10} = \frac{30}{10} = 3$



while calculating the Binsize do not include 1st number

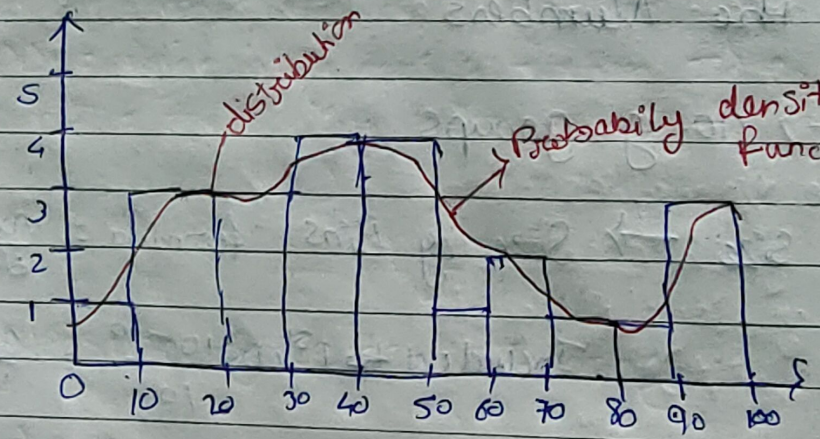
$\frac{40}{10} = 4$



Ages: [10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 55, 65, 67, 78, 90, 95, 100]

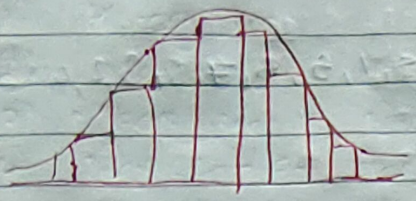
Bin: 10

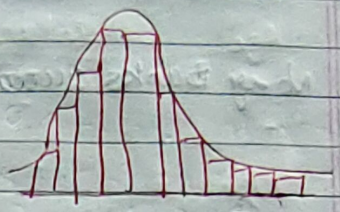
Bin Size $\frac{100}{10}$ ← no of Bins

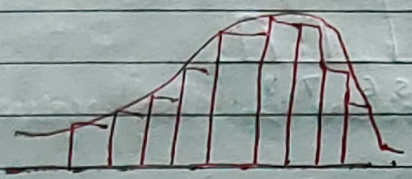


Date
Page

Types of histogram





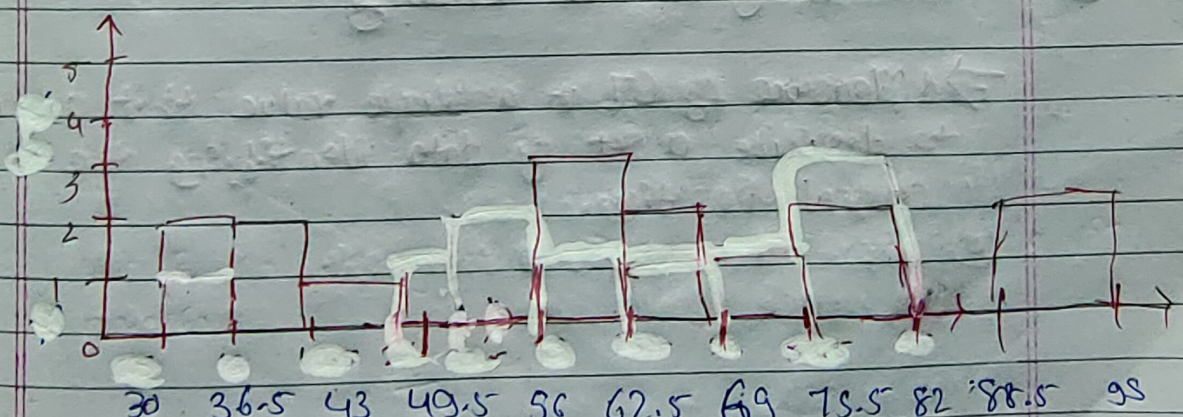


Ex:2

weight = [30, 35, 38, 42, 46, 58, 59, 62, 63, 68, 75, 77, 80, 90, 95]

Bins = 10

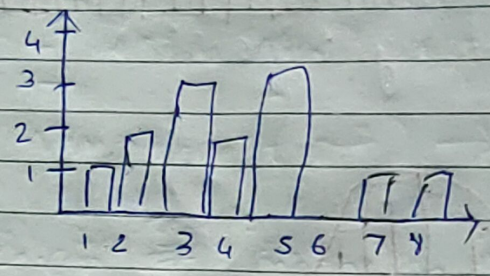
Bins size = $\frac{95-30}{10} = \frac{65}{10} = 6.5$



30 36.5 43 49.5 56 62.5 69 75.5 82 88.5 95

② Discrete ~~Continuous~~

No. of Banks' account = [2, 3, 4, 5, 1, 5, 3, ~~4~~, 8, 3, 2, 4, 5]



Probability mass function we use in Discrete

PDF = Probability density function → continuous
pmf = Probability mass function → Discrete.

② Measure of central Tendency

⇒ A Measure of CT is a single value that attempts to describe a set of data identifying the central position.

Let $x = [1, 2, 3, 4, 5]$

2) Mean : Average / mean $= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

Thus, 3 specify the central position of the set of x

Population mean (μ) $= \frac{\sum_{i=1}^{N} x_i}{N}$

Sample mean ($\bar{x}$) $= \frac{\sum_{i=1}^{n} x_i}{n}$

Population Age : $[24, 23, 2, 1, 28, 27]$

$\therefore N = 6$

Population mean (μ) $= \frac{24+23+2+1+28+27}{6}$

$= \frac{105}{6} \quad \mu = 17.5$

Sample mean ($\bar{x}$) : $[\frac{24, 2, 1, 27}{4}]$

$\bar{x} = 13.5$

$\mu > \bar{x}$ $\quad \bar{x} > \mu$

Practical Application (feature Engineering)

Age	Salary	family Size
-	-	Nan
-	-	-
Nan	-	-
-	Nan	-
-	-	Nan
Nan	-	-

We can ~~fest~~ replace mean on Nan values.

② median : we use when there is outlies in Data.

Ex [1,2,3,4,5] [1,2,3,4,5,100]

$\bar{x} = 3$ $\bar{x} = 19.16$

Steps to find out median

① Sort the Numbers
② Find the central Number.

Date
Page

(1) If the no. of elements are even we find the average of central elements

(2) If the no. of elements are odd we find the central element

Ex : Even so we find average of central element

[1, 2, 3, 4, 5, 6, 7, 8, 100, 120]

median $= \frac{5+6}{2} = 5.5$

(3) MODE : Most frequent occuring elements

[1, 2, 2, 3, 3, 3, 4, 5]
Mode = 3

[1, 2, 3, 2, 2, 3, 3, 4, 5]
mode = 2, 3

Data sets

Type of flower

(1) Lily
(2) Sunflower
(3) Rose
(4) Nan
(5) Rose
(6) Sunflower
(7) Rose
(8) Nan

Note = MODE most of times we in Categorical Data or variables

③ Measure of Dispersion

① variance (σ^2) → talks about spread of Data

② Standard deviation (σ)

① Variance

Population variance (σ^2)

$$\Rightarrow \sigma^2 = \sum_{i=1}^{N} \left(\frac{x_i - \mu}{N} \right)^2$$

Lets take a example

① [1,2,3,4,5,6,7,8,9,10] ② [1,2,3,4,50,60,70]

If we find variance of Both data the variance will high on 2nd data set

How to calculate [1,2,3,4,5]

$\mu = 3$

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

$= 4+1+0+1+4 = \frac{10}{5} = 2$

Date
Page

⑥ Sample Variance (S^2)

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{N} (xi - \bar{x})^2}{n-1}$$

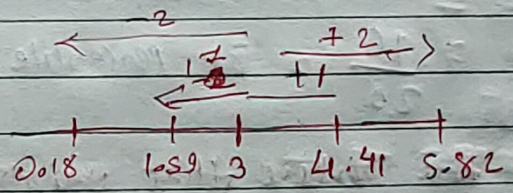
② standard deviation ($\sqrt{\sigma^2}$) = (σ)

[1, 2, 3, 4, 5]

$\mu = 3$

$\sigma^2 = 2$

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



→ Percentiles and Quantiles

Percentage : [1, 2, 3, 4, 5, 6, 7, 8]

Percentage of even no: $\frac{\text{No. of even numbers}}{\text{Total no. of Number}}$

$= \frac{4}{8} = 0.5 = 50\%$

Percentiles: A percentile is a value below which a certain percentage of observation lie.

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

Ex:
Datasets: [2, 2, 3, 4, 5, 5, 6, 7, 8, 8, 8, 8, 8, 9, 9, 10, 11]

→ what is the percentile rank of 10?

Percentile Rank of x_i = $\frac{\#\text{no. of value below } x}{n}$

$= \frac{16}{20} = 80$ Percentile.

→ what is the value that exists at 25 Percentile

→ Value = $\frac{\text{Percentile}}{100} \times n$

$= \frac{25}{100} \times 20 = 5^{th}$ Index

Index always start with 0^{th}

Formula.

Value = $\frac{\text{Percentile} \times n}{100}$

Date
Page

→ 5 number Summary

① Minimum
② first Quartile (25 percentile) Q1
③ median (50 Percentile) Q2
④ third Quartile (75 Percentile) Q3
⑤ maximum

} we use to Remove Outliers.

[1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27]

we create fence were value ranging between for finding outlier
[lower fence ⟷ higher fence]

Lower fence = Q1 − 1.5 [IQR], IQR = Q3 − Q1
↓
Inter quartile Range

Higher Fence = Q3 + 1.5 (IQR]

Q1 = $\frac{25}{100} \times n+1 = \frac{25}{100} \times 21 = 5.25$
= 3

There is no 5.25 index so we will take average of 5th and 6th index
= $\frac{3+3}{2} = 3$