Statistics

Data Analyst → Amazon wants to decide when is the Next Big Billion Sale?

Month and on which day?

Domain Experts

[Tool]

① Requirement Gathering → DATA ANALYST ← PRODUCT MANAGER

① What all data they require???

② 3rd party, Internal Database, Paid Data

Business Analyst

Product Manager

Domain Knowledge

Problem → { SQL → Database }

Statistical Analysis ⇒ conclusions

EDA { Exploratory Data Analysis } ✓

Reporting [Power BI, Tableau] (Understanding Data)

3rd Month

4th Month

{ Hypothesis Testing }

Stakeholders

Key Performance Indicators (KPI's).

DIRECTORS OF { AMAZONS }

Decisions

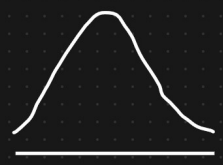
Statistics ÷ Defn

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

Data ÷ "facts or pieces of information"

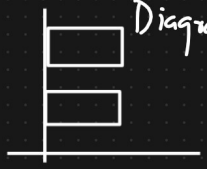
Eg: Ages of students in a classrooms ⇓ Histograms

{24, 30, 21, 34, 20} ⇒ Mean, Median ⇒ Analysis



Eg: Weights of students in a classroom ↓ Diagrams

{75, 45, 50, 55, 60} ⇒



2 types

Statistics

Descriptive Stats

Influential Stats

- It consists of organizing and summarizing of data
- It consists of a technique to form some conclusions

Histograms

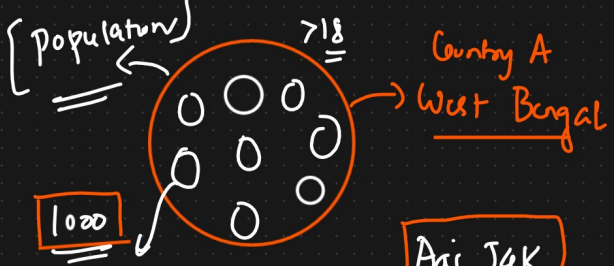
BAR CHART

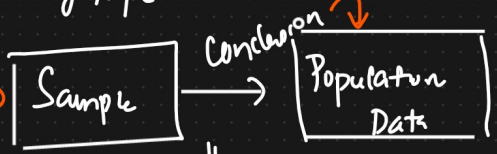
Box plot

pie chart

CRYPTO

CANDLESTICK





⇓

Hypothesis Testing

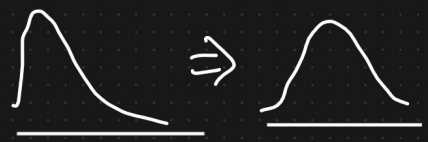
Eg: P value, z test, t-test, Chi Square test Annova,

Interview Amazon

What is the average size of All Sharks In the world??

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

Ages {21, 20, 18, 34, 17, 22, 24, 25, 26, 23, 22}



Descriptive Stats : What is the average age of the students in the classroom??

{1, 3, 4, 100} ← outlier

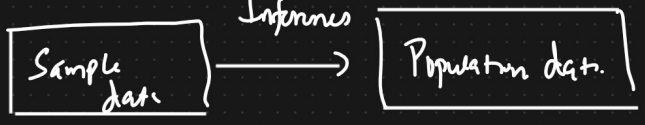
Minimum Age, Outliers, Distribution

Maximum Age

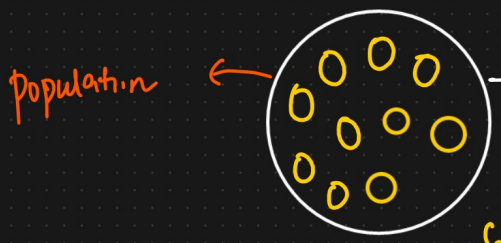
Inferential question : Are the average age of the student in the classroom equal to the average ages of the entire university

Sample ⇑

⇓ population data.



(N) Population And (n) Sample



1 Million → State A

Sample → Population

Avg height

Sample Size = 1000

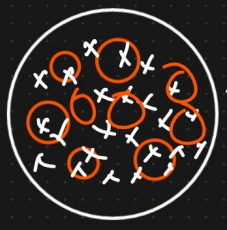
EXIT POLL { Average height of all the people in the state??

Sampling Techniques

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

1000000

1 / 1000000



→ population

Tossing a Coin

$\frac{1}{2}$ $\frac{1}{2}$

② Stratified Sampling

Strata → layers → clusters

⇓

Non overlapping

Gender → Male, Female

Education Degree → BE, Masters, PHD, High School

Blood Groups → A+, B+, O+

③ Systematic Sampling :

BANGALORE

AIRPORT

(N) → select every nth individual

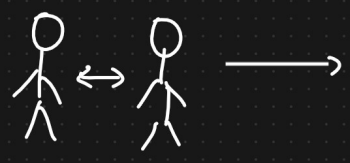
nth person

Every 5th person

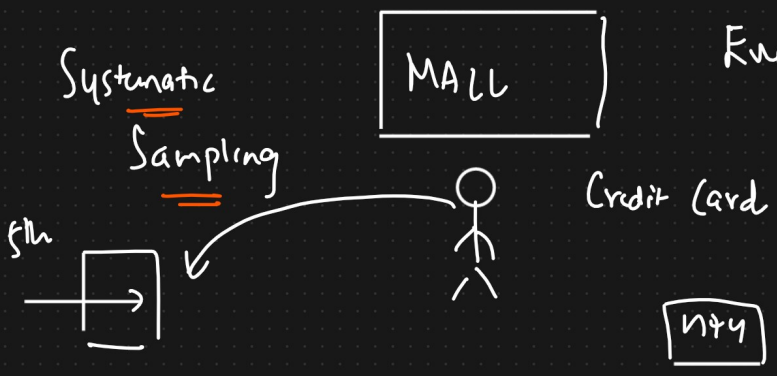
5th person

{ SBI Credit Card Stall }

7th person



Systematic Sampling

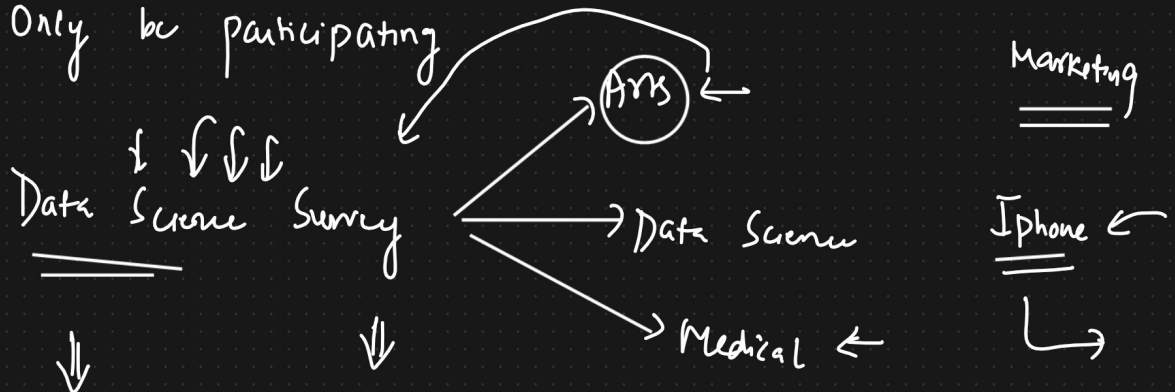


Every nth individual

n = 5

Credit Card

④ Convenience Sampling : Only those people who are interested will only be participating



Accurate

Target Audience.

Teams → Calls

① Sells Credit Card → Airport → Systematic / Random

② Exit Polls → Stratified + Random Sampling (age ≥18, >18)

③ Survey Regarding New Technology

Variable

A variable is a property that can take on any value

Eg: Age = 24

Age = 25

Age = 26

Ages = {24, 25, 28, 29, 30} => Collection

Two Kinds of variable

① Quantitative Variable → Measured Numerically {Add, Subtract, * & ÷}

② Qualitative Variables → Categorical Variable

↳ Eg: Gender → Male, Female

{Based on some characteristics we can group categorical variables}

Type of flowers → Rose, Lily, Sunflower

Types of Movies → Action, Comedy, Action & Comedy

Quantitative Variable

- Discrete Variable
  - Eg: Whole numbers
- Continuous Variable
  - Eg: Height = 172.5cm, 170.5cm, 160.62cm

Eg: No. of Bank Account

{1, 2, 3, 4, 0, 5, }

2.5 ~~2.5~~ 3.75 ~~3.75~~

Rainfall : 1.35cm, 1.25, 1.3589cm

Weight, Temperature.

Eg: No. of children in a family

{0, 1, 2, 3, 4, 5, }

~~2.5~~ children

~~3.5~~

Whole Number Numerical

Widow
Married
Not Married

0 → Categorical variable
1
2

100.0+5 km

17 years 5 month 24 day

Age ?: 20 seconds.

Assignment

① What kind of variable is Marital Status? [Categorical]

② What kind of " is Ganga River length? [Continuous]

③ What kind of " is Movie duration? [continuous]

④ What kind of variable IQ is? [continuous]

{[90-100] [100-120] [120-150]} → Categorical variable

Histograms {Construct a histogram}

Discrete → 10 unique

2.5

{Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51, 65, 68, 78, 90, 95, 100}} → Continuous value

Probability density function

① Sort the Numbers {Ascending} Frequency

② Bins → No. of groups

③ Bins size → Size of Bins

[0-100] Bins = 10 ✓

Bins size = $\frac{100}{10} = 10$

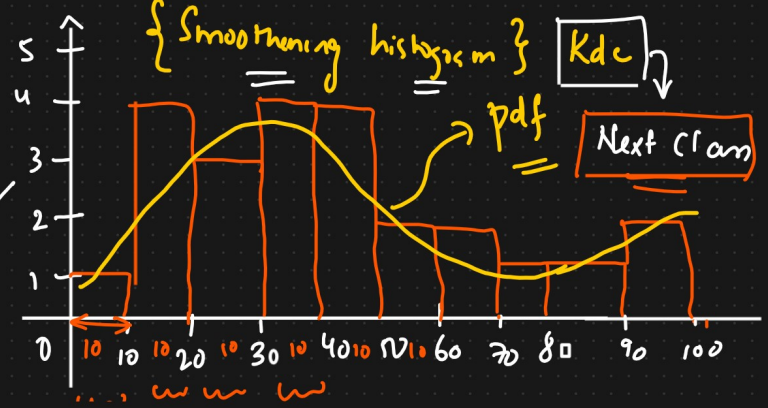
0-10

<10

{Smoothening histogram} Kde

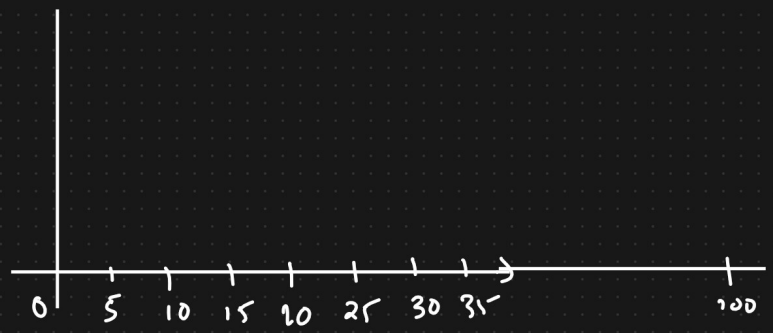
pdf

Next class



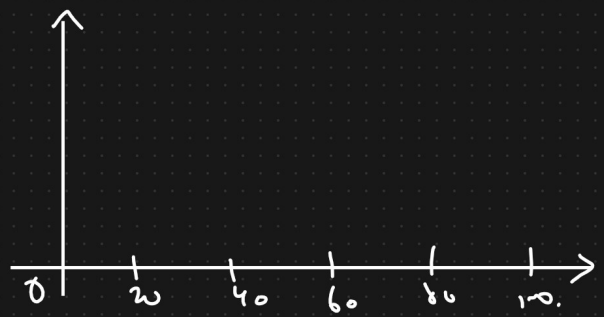
[0-100] Bins = 20

Bin size = $\frac{100}{20}$ = 5



[0-100] Bin = 5

Bin size = $\frac{100}{5}$ = 20

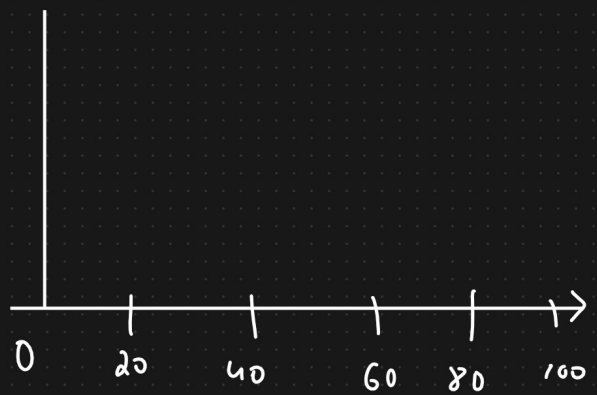


Assignment :- Data Science [jayant@ineuron.ai] →

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

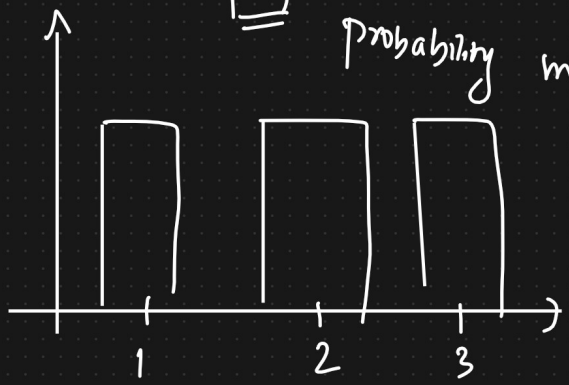
bins = 5

bin size = 20



pmf

Probability mass function



100.5

99.5 78.5

120.5

Age

IQ

+

No. of house (Discrete) → Whole number

1
2
3

2
2

[ ]

4
5
6
10

so 21 5 22day 10 m

No of years

Continuous

→ 2
→ 4
→ 6
→ 8
→ 10
→ 12
→ 14

>= 25 ⇒

< 25

↓

discrete

{ whole number }

Age of house

2
4
6
2
4
6
10
12

2.5
28
29
3.1

No.of

⇒ Continuous ✗

⇒ Restricted limit

limit

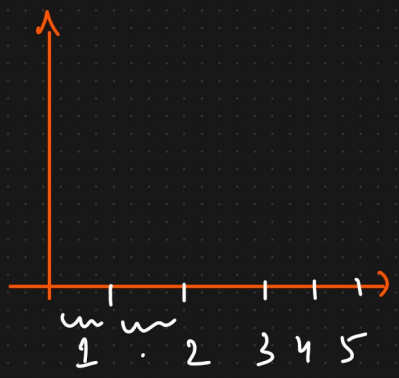
⇒ Discrete

{ 1, 2, 3, 4, 5 }

bins = 5

0-5

bin size = 5/5 = 1



Assignments : Need to be Submitted

✓✓ [Email : jayant @ineuron.ai]