Introduction to statistics

Definition :- Statistics is a mathematical science including methods of collecting, organizing and analysing data in a such a way that meaningful conclusions can be drawn from them.

data : facts or pieces of information that can be stored and measured.

Example1: Scores made by Virat Kohli in last five matches

[102, 33, 0, 45, 46]

Example2 - Heights of students in the classrooms

[110cm, 150cm, 120 cm - - -]

Motivation | Uses of Statistics

1. Weather forecast
2. Sports Analytics — [102, 33, 0, 45, 45] — min = 0, max = 102, Avg =
3. Election Campaign
4. FMCG | e-commerce
5. Medical | Genetics

Introduction to statistics

Definition :- Statistics is a mathematical science including methods of collecting, organizing and analysing data in a such a way that meaningful conclusions can be drawn from them.

data : facts or pieces of information that can be stored and measured.

Example1: Scores made by Virat Kohli in last five matches
[102, 33, 0, 45, 46]

Example2 - Heights of students in the classrooms
[110cm, 150cm, 170 cm - - -]

Motivation | Uses of Statistics

1. Weather forecast
2. Sports Analytics — [102, 33, 0, 45, 45] — min = 0, max = 102, Avg =
3. Election Campaign
4. FMCG | e-commerce
5. Medical | Genetics

Types of statistics

① Descriptive statistics

↓

Describe

Defn → It consists of organising and summarizing the data | population.

Virat Kohli — [45, 62, 65, 81 - - -]
strike rate

② Inferential Statistics

↓

Inference.

Defn

It consists of using data that has been measured to form conclusion about population.

→ With the given sample data can we conclude something about the population.

① Descriptive Statistics

Example-1 – Virat kohli scores

[45, 60, 80, 80, 70, 80, 91 - - -]

Avg score / strike rate

Example-2 Avg Height / weight of students in class

[150 cm, 130 cm, 130 cm, - - - -]

Q. What is Avg height of students in class? → $\frac{150+130+\cdots}{50}$ → some avg value.

Q. What is the most frequent height of students in class? → 130 cm.

Q What is avg score scored by Vk in his career?

$$\frac{45+60+\cdots}{n} = \text{avg score}$$

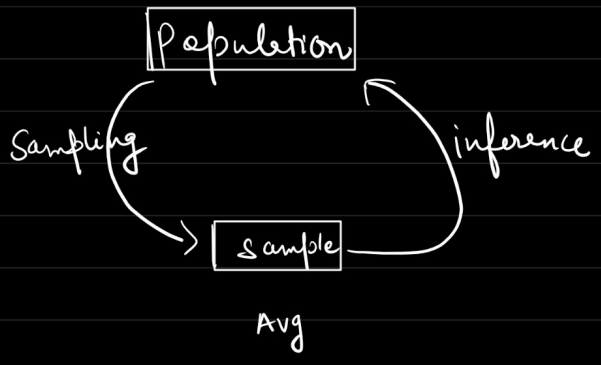
Techniques of Descriptive statistics

① Measures of central tendency (mean, median, mode)
② Measures of Symmetricity (skewness, kurtosis)
③ Measures of dispersion (standard deviation, Variance)

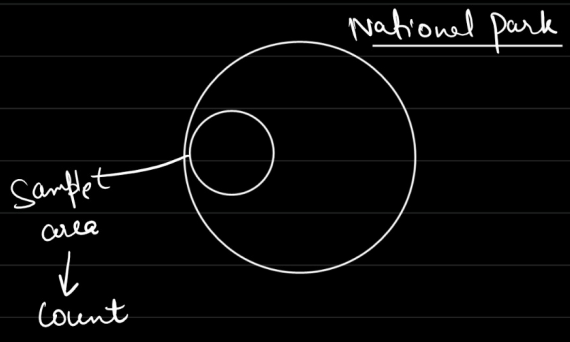
② Inferential Statistics

Why? → Population is large
→ time and resource constraints

Example-1 Avg age of population of India.



Example-2 - No of trees in a National Park



10% of area → 2000

100% of National park ⟹ 20,000

Descriptive Question

* Classroom

→ What is Avg age/height of students in class.?

Inferential Question

→ Are the avg height of students in the class is what you expect in the entire school?

Introduction to statistics

Definition :- Statistics is a mathematical science including methods of collecting, organizing and analysing data in a such a way that meaningful conclusions can be drawn from them.

data : facts or pieces of information that can be stored and measured.

Example1: Scores made by Virat Kohli in last five matches
[102, 33, 0, 45, 46]

Example2 - Heights of students in the classrooms
[110cm, 150cm, 170 cm - - -]

Motivation | Uses of Statistics

1. Weather forecast
2. Sports Analytics — [102, 33, 0, 45, 45] — min = 0, max = 102, Avg =
3. Election Campaign
4. FMCG | e-commerce
5. Medical | Genetics

Types of statistics

① Descriptive statistics

↓

Describe

Defn → It consists of organising and summarizing the data | population.

Virat Kohli — [45, 62, 65, 81 - - -]
strike rate

② Inferential Statistics

↓

Inference.

Defn

It consists of using data that has been measured to form conclusion about population.

→ With the given sample data can we conclude something about the population.

① Descriptive Statistics

Example-1 – Virat kohli scores

[45, 60, 80, 80, 70, 80, 91 - - -]

Avg score / strike rate

Example-2 Avg Height / weight of students in class

[150 cm, 130 cm, 130 cm, - - - -]

Q. What is Avg height of students in class? → $\frac{150+130+\cdots}{50}$ → some avg value.

Q. What is the most frequent height of students in class? → 130 cm.

Q What is avg score scored by Vk in his career?

$$\frac{45+60+\cdots}{n} = \text{avg score}$$

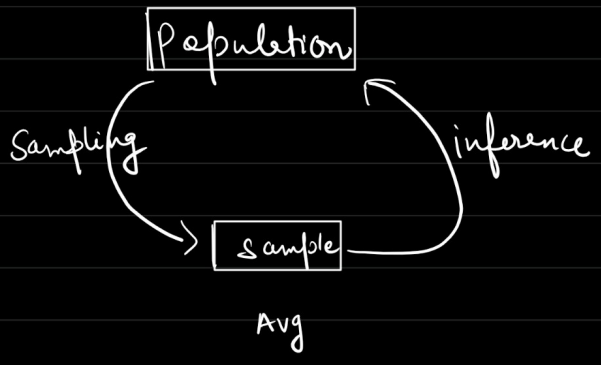
Techniques of Descriptive statistics

① Measures of central tendency (mean, median, mode)
② Measures of Symmetricity (skewness, kurtosis)
③ Measures of dispersion (standard deviation, Variance)

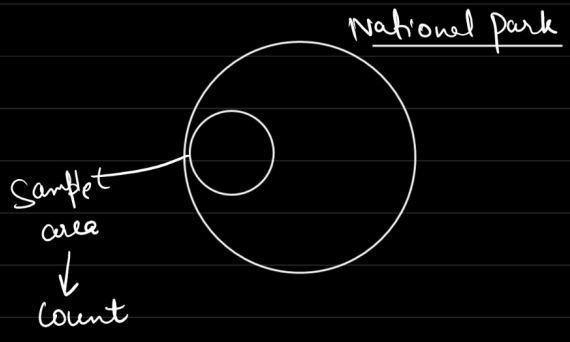
② Inferential Statistics

Why? → Population is large
→ time and resource constraints

Example-1 Avg age of population of India.



Example-2 - No of trees in a National Park



10% of area → 2000

100% of National park ⟹ 20,000

Descriptive Question

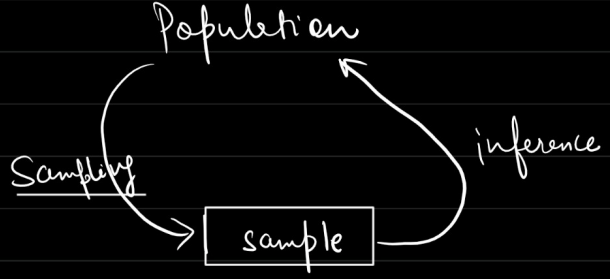
* Classroom

→ What is Avg age/height of students in class.?

Inferential Question

→ Are the avg height of students in the class is what you expect in the entire school?

Types of Sampling



ex. income of all households in India.

⇓

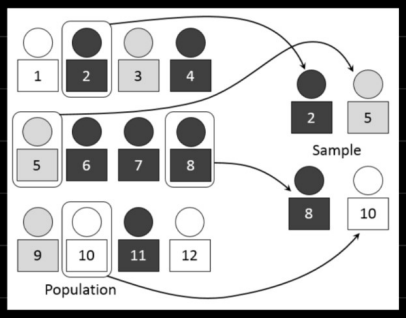
Sampling technique

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. systematic sampling

① Simple Random sampling

→ Every member of the population (N) has an equal chance of being selected for your sample.

→ Each person has 1/12 of being selected in the sample.

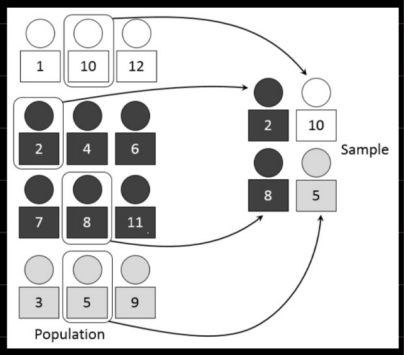


② Stratified Sampling

⇓

Strata → Layers / groups

Avg height of Population of India :- In SRS the people of North East India might have lesser avg height as compared to rest of India → Stratified Sampling should be used.



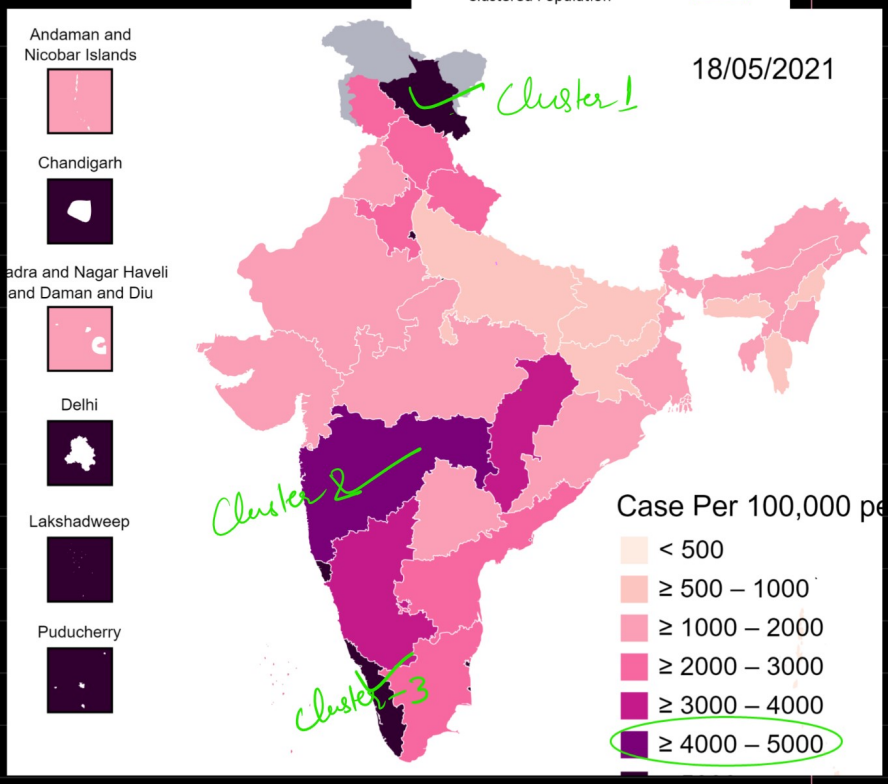
→ different distinct categories are there.

→ A simple sample would be chosen from each strata or layer.

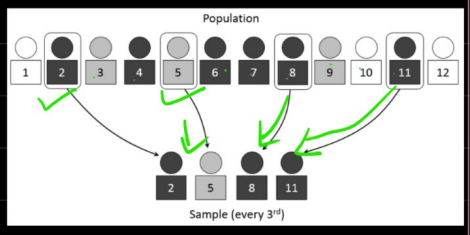
③ Cluster Sampling

→ divides the population into groups or clusters, some of these clusters are randomly selected.

→ Then all individual in the chosen cluster are selected in the sample.



④ Systematic Sampling

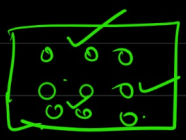


→ Every n^{th} element will be selected.

Examples → odd roll no
→ People born on odd years.

Recap

→ SRS



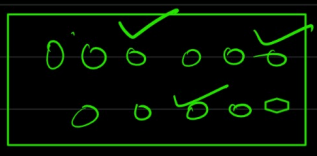
→ Quota Sampling
→ Min-max Sampling
→ Convenience Sampling
→ Accidental Sampling

→ Stratified Sampling

eg. Students of a school — Class 1, 2, 3, ---- 10
eg. Population of India
eg. Gender - Male / Female

→ Cluster Sampling
All the data of randomly selected clusters as sample.

→ Systematic Sampling
every nth member



ex. Students odd robl no
eg people born in leap years

Introduction to statistics

Definition :- Statistics is a mathematical science including methods of collecting, organizing and analysing data in a such a way that meaningful conclusions can be drawn from them.

data : facts or pieces of information that can be stored and measured.

Example1: Scores made by Virat Kohli in last five matches
[102, 33, 0, 45, 46]

Example2 - Heights of students in the classrooms
[110cm, 150cm, 170 cm - - -]

Motivation | Uses of Statistics

1. Weather forecast
2. Sports Analytics — [102, 33, 0, 45, 45] — min = 0, max = 102, Avg =
3. Election Campaign
4. FMCG | e-commerce
5. Medical | Genetics

Types of statistics

① Descriptive statistics

↓

Describe

Defn → It consists of organising and summarizing the data | population.

Virat Kohli — [45, 62, 65, 81 - - -]
strike rate

② Inferential Statistics

↓

Inference.

Defn

It consists of using data that has been measured to form conclusion about population.

→ With the given sample data can we conclude something about the population.

① Descriptive Statistics

Example-1 – Virat kohli scores

[45, 60, 80, 80, 70, 80, 91 - - -]

Avg score / strike rate

Example-2 Avg Height / weight of students in class

[150 cm, 130 cm, 130 cm, - - - -]

Q. What is Avg height of students in class? → $\frac{150+130+\cdots}{50}$ → some avg value.

Q. What is the most frequent height of students in class? → 130 cm.

Q What is avg score scored by Vk in his career?

$$\frac{45+60+\cdots}{n} = \text{avg score}$$

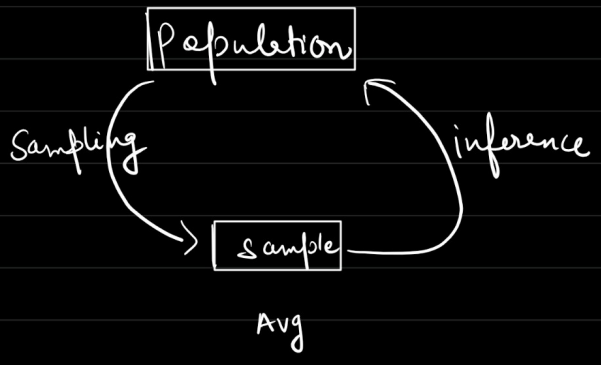
Techniques of Descriptive statistics

① Measures of central tendency (mean, median, mode)
② Measures of Symmetricity (skewness, kurtosis)
③ Measures of dispersion (standard deviation, Variance)

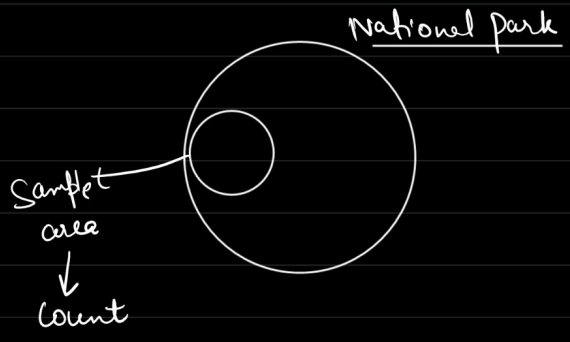
② Inferential Statistics

Why? → Population is large
→ time and resource constraints

Example-1 Avg age of population of India.



Example-2 - No of trees in a National Park



10% of area → 2000

100% of National park ⟹ 20,000

Descriptive Question

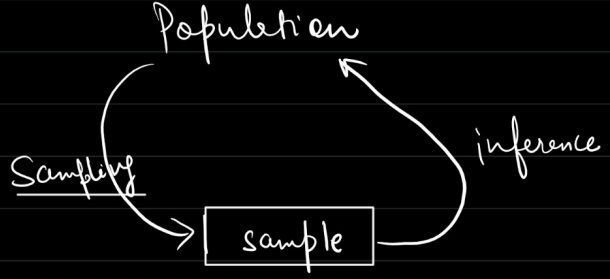
* Classroom

→ What is Avg age/height of students in class.?

Inferential Question

→ Are the avg height of students in the class is what you expect in the entire school?

Types of Sampling



ex. income of all households in India.

⇓

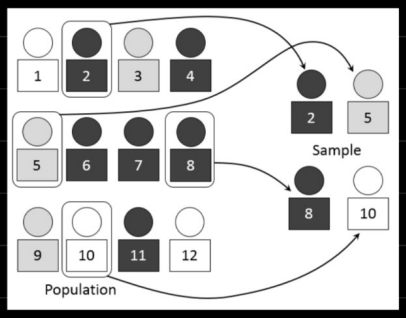
Sampling technique

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. systematic sampling

① Simple Random sampling

→ Every member of the population (N) has an equal chance of being selected for your sample.

→ Each person has 1/12 of being selected in the sample.

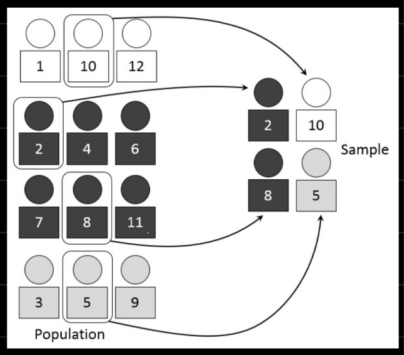


② Stratified Sampling

⇓

Strata → Layers / groups

Avg height of Population of India :- In SRS the people of North East India might have lesser avg height as compared to rest of India → Stratified Sampling should be used.



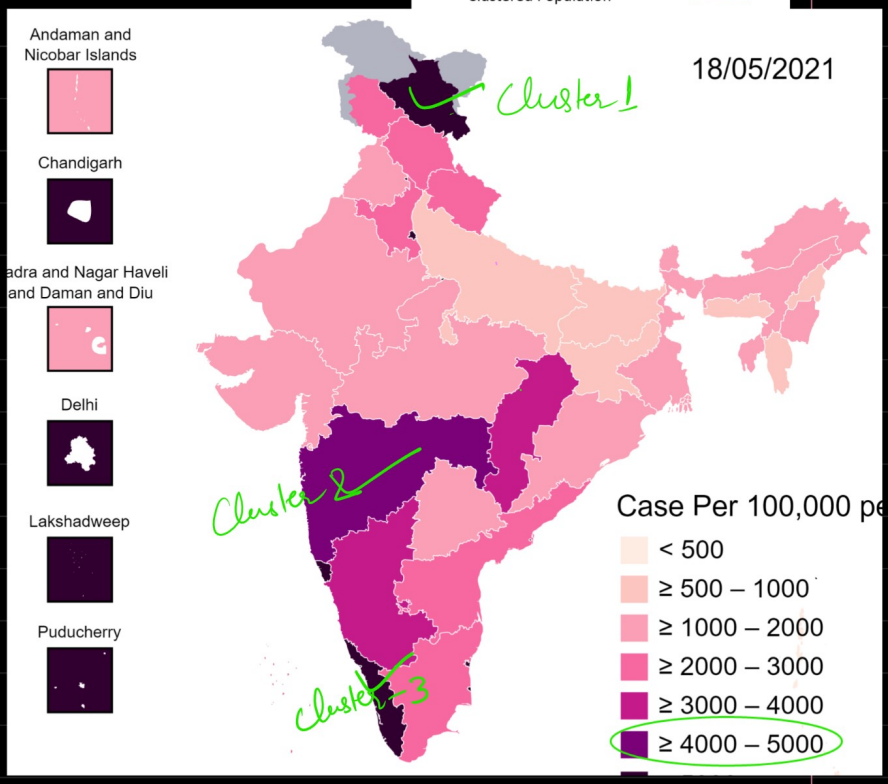
→ different distinct categories are there.

→ A simple sample would be chosen from each strata or layer.

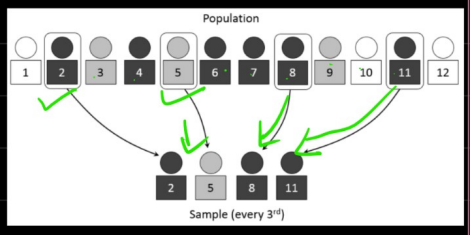
③ Cluster Sampling

→ divides the population into groups or clusters, some of these clusters are randomly selected.

→ Then all individual in the chosen cluster are selected in the sample.



④ Systematic Sampling

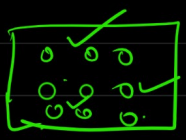


→ Every n^{th} element will be selected.

Examples → odd roll no
→ People born on odd years.

Recap

→ SRS



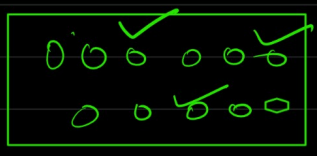
→ Quota Sampling
→ Min-max Sampling
→ Convenience Sampling
→ Accidental Sampling

→ Stratified Sampling

eg. Students of a school — Class 1, 2, 3, ---- 10
eg. Population of India
eg. Gender - Male / Female

→ Cluster Sampling
All the data of randomly selected clusters as sample.

→ Systematic Sampling
every nth member



ex. Students odd robl no
eg people born in leap years

Types of data

Age	TypeofContact	CityTier	DurationOfPitch	Occupation	Gender	NumberOfPersonVisi	NumberOfFollowups	ProductPitched	PreferredPropertySta	MaritalStatus
41	Self Enquiry	3	6	Salaried	Female	3	3	Deluxe	3	Single
49	Company Invited	1	14	Salaried	Male	3	4	Deluxe	4	Divorced
37	Self Enquiry	1	8	Free Lancer	Male	3	4	Basic	3	Single
33	Company Invited	1	9	Salaried	Female	2	3	Basic	3	Divorced
	Self Enquiry	1	8	Small Business	Male	2	3	Basic	4	Divorced
32	Company Invited	1	8	Salaried	Male	3	3	Basic	3	Single
59	Self Enquiry	1	9	Small Business	Female	2	2	Basic	5	Divorced
30	Self Enquiry	1	30	Salaried	Male	3	3	Basic	3	Married
38	Company Invited	1	29	Salaried	Male	2	4	Standard	3	Unmarried
36	Self Enquiry	1	33	Small Business	Male	3	3	Deluxe	3	Divorced
35	Self Enquiry	1	22	Small Business	Male	2	2	Basic	4	Divorced
	Self Enquiry	1	21	Salaried	Female	2	4	Deluxe	3	Single
31	Self Enquiry	1	32	Salaried	Male	2	3	Basic	3	Married
34	Self Enquiry	1	25	Small Business	Male	3	3	Basic	3	Married
28	Self Enquiry	1	30	Salaried	Male	2	4	Basic	3	Single
29	Self Enquiry	1	27	Salaried	Female	2	2	Standard	5	Married
32	Self Enquiry	1	11	Salaried	Male	3	2	Basic	4	Married
22	Company Invited	1	22	Small Business	Male	3	2	Basic	3	Married
53	Self Enquiry	3	8	Salaried	Female	3	4	Super Deluxe	3	Divorced
	Self Enquiry	1	8	Salaried	Male	2	3	Basic	3	Single
	Company Invited	1	17	Salaried	Female	3	2	Deluxe	3	Married
	Self Enquiry	3	15	Salaried	Male	2	4	Deluxe	5	Single
34	Self Enquiry	1	13	Salaried	Fe Male	2	3	Standard	4	Unmarried
21	Self Enquiry	1	21	Salaried	Male	3	3	Basic	3	Single
34	Self Enquiry	1	12	Small Business	Male	2	3	Basic	5	Single
53	Self Enquiry	1	11	Salaried	Female	2	3	King	3	Married

CityTier: best, good, better

Data

Numerical

+, −, ÷, %

Quantitative

0 1 2 3 10

Discrete → 0, 1, 2

Continuous → 1·2, 2·3

Whole no

eg: No of children
1, 2, 3

eg: No of mobile phones

eg: No of Bank account
1, 2,

eg: Total no of employees in a team.

Any values in a range

eg: Height
Weight
Speed
temperature
IQ
movie duration
length

Categorical

Qualitative data

Nominal
(No order in Categorial data)

eg: Gender
eg: Blood group
eg: Colour
eg: race/ethnicity
eg: Marital Status

Ordinal
(Order)
⇓
Rank.

eg: good, better best

eg: Grade
A, B, C, D

eg: Salary
high, medium, low

→ It will take
specific finite
whole no. values.

Introduction to statistics

Definition :- Statistics is a mathematical science including methods of collecting, organizing and analysing data in a such a way that meaningful conclusions can be drawn from them.

data : facts or pieces of information that can be stored and measured.

Example1: Scores made by Virat Kohli in last five matches
[102, 33, 0, 45, 46]

Example2 - Heights of students in the classrooms
[110cm, 150cm, 170 cm - - -]

Motivation | Uses of Statistics

1. Weather forecast
2. Sports Analytics — [102, 33, 0, 45, 45] — min = 0, max = 102, Avg =
3. Election Campaign
4. FMCG | e-commerce
5. Medical | Genetics

Types of statistics

① Descriptive statistics

↓

Describe

Defn → It consists of organising and summarizing the data | population.

Virat Kohli — [45, 62, 65, 81 - - -]
strike rate

② Inferential Statistics

↓

Inference.

Defn

It consists of using data that has been measured to form conclusion about population.

→ With the given sample data can we conclude something about the population.

① Descriptive Statistics

Example-1 – Virat kohli scores

[45, 60, 80, 80, 70, 80, 91 - - -]

Avg score / strike rate

Example-2 Avg Height / weight of students in class

[150 cm, 130 cm, 130 cm, - - - -]

Q. What is Avg height of students in class? → $\frac{150+130+\cdots}{50}$ → some avg value.

Q. What is the most frequent height of students in class? → 130 cm.

Q What is avg score scored by Vk in his career?

$$\frac{45+60+\cdots}{n} = \text{avg score}$$

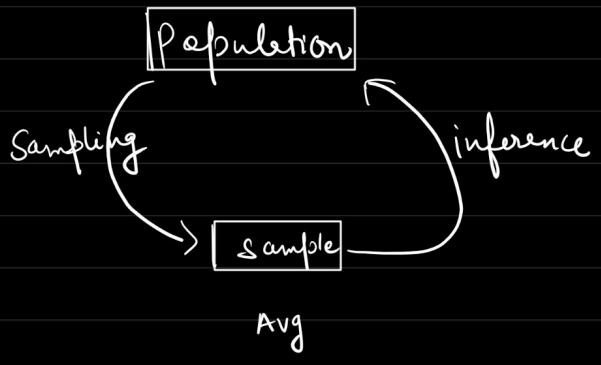
Techniques of Descriptive statistics

① Measures of central tendency (mean, median, mode)
② Measures of Symmetricity (skewness, kurtosis)
③ Measures of dispersion (standard deviation, Variance)

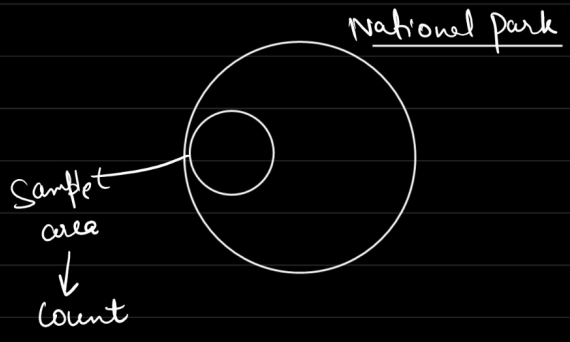
② Inferential Statistics

Why? → Population is large
→ time and resource constraints

Example-1 Avg age of population of India.



Example-2 - No of trees in a National Park



10% of area → 2000

100% of National park ⟹ 20,000

Descriptive Question

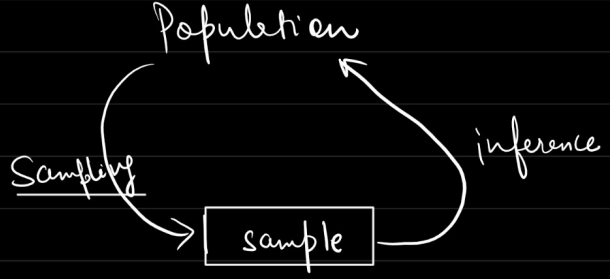
* Classroom

→ What is Avg age/height of students in class.?

Inferential Question

→ Are the avg height of students in the class is what you expect in the entire school?

Types of Sampling



ex. income of all households in India.

⇓

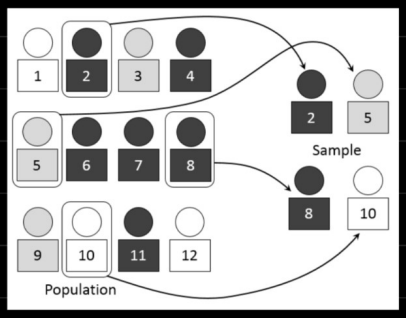
Sampling technique

1. Simple Random Sampling
2. Stratified Sampling
3. Cluster Sampling
4. systematic sampling

① Simple Random sampling

→ Every member of the population (N) has an equal chance of being selected for your sample.

→ Each person has 1/12 of being selected in the sample.

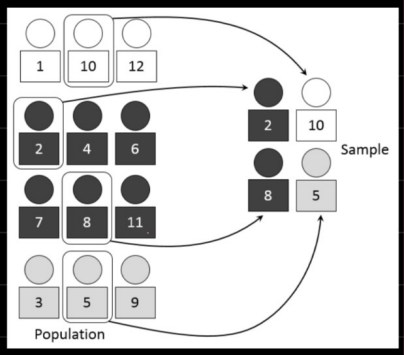


② Stratified Sampling

⇓

Strata → Layers / groups

Avg height of Population of India :- In SRS the people of North East India might have lesser avg height as compared to rest of India → Stratified Sampling should be used.



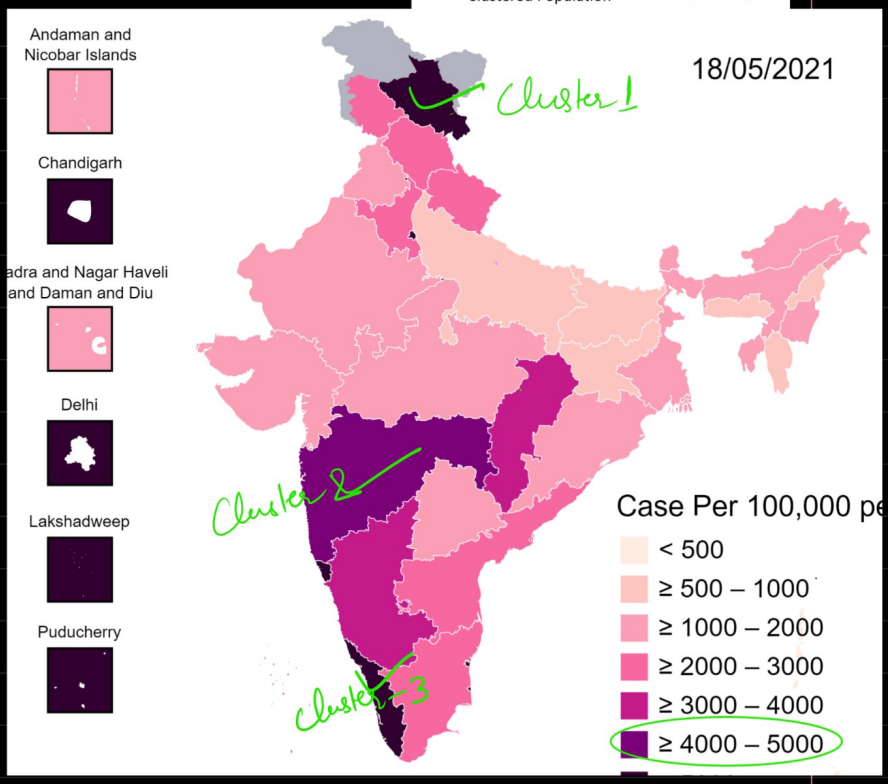
→ different distinct categories are there.

→ A simple sample would be chosen from each strata or layer.

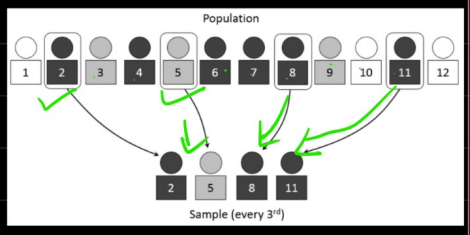
③ Cluster Sampling

→ divides the population into groups or clusters, some of these clusters are randomly selected.

→ Then all individual in the chosen cluster are selected in the sample.



④ Systematic Sampling

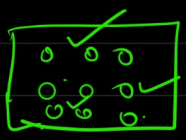


→ Every n^{th} element will be selected.

Examples → odd roll no
→ People born on odd years.

Recap

→ SRS



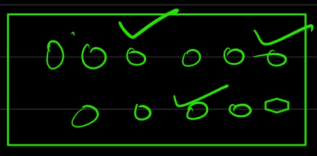
→ Quota Sampling
→ Min-max Sampling
→ Convenience Sampling
→ Accidental Sampling

→ Stratified Sampling

eg. Students of a school — Class 1, 2, 3, ---- 10
eg. Population of India
eg. Gender - Male / Female

→ Cluster Sampling
All the data of randomly selected clusters as sample.

→ Systematic Sampling
every nth member



ex. Students odd robl no
eg people born in leap years

Types of data

Age	TypeofContact	CityTier	DurationOfPitch	Occupation	Gender	NumberOfPersonVisi	NumberOfFollowups	ProductPitched	PreferredPropertySta	MaritalStatus
41	Self Enquiry	3	6	Salaried	Female	3	3	Deluxe	3	Single
49	Company Invited	1	14	Salaried	Male	3	4	Deluxe	4	Divorced
37	Self Enquiry	1	8	Free Lancer	Male	3	4	Basic	3	Single
33	Company Invited	1	9	Salaried	Female	2	3	Basic	3	Divorced
	Self Enquiry	1	8	Small Business	Male	2	3	Basic	4	Divorced
32	Company Invited	1	8	Salaried	Male	3	3	Basic	3	Single
59	Self Enquiry	1	9	Small Business	Female	2	2	Basic	5	Divorced
30	Self Enquiry	1	30	Salaried	Male	3	3	Basic	3	Married
38	Company Invited	1	29	Salaried	Male	2	4	Standard	3	Unmarried
36	Self Enquiry	1	33	Small Business	Male	3	3	Deluxe	3	Divorced
35	Self Enquiry	1	22	Small Business	Male	2	2	Basic	4	Divorced
	Self Enquiry	1	21	Salaried	Female	2	4	Deluxe	3	Single
31	Self Enquiry	1	32	Salaried	Male	2	3	Basic	3	Married
34	Self Enquiry	1	25	Small Business	Male	3	3	Basic	3	Married
28	Self Enquiry	1	30	Salaried	Male	2	4	Basic	3	Single
29	Self Enquiry	1	27	Salaried	Female	2	2	Standard	5	Married
32	Self Enquiry	1	11	Salaried	Male	3	2	Basic	4	Married
22	Company Invited	1	22	Small Business	Male	3	2	Basic	3	Married
53	Self Enquiry	3	8	Salaried	Female	3	4	Super Deluxe	3	Divorced
	Self Enquiry	1	8	Salaried	Male	2	3	Basic	3	Single
	Company Invited	1	17	Salaried	Female	3	2	Deluxe	3	Married
	Self Enquiry	3	15	Salaried	Male	2	4	Deluxe	5	Single
34	Self Enquiry	1	13	Salaried	Fe Male	2	3	Standard	4	Unmarried
21	Self Enquiry	1	21	Salaried	Male	3	3	Basic	3	Single
34	Self Enquiry	1	12	Small Business	Male	2	3	Basic	5	Single
53	Self Enquiry	1	11	Salaried	Female	2	3	King	3	Married

CityTier: best, good, better

Data

Numerical

+, −, ÷, %

Quantitative

0 1 2 3 10

Discrete → 0, 1, 2

Continuous → 1·2, 2·3

Whole no

eg: No of children
1, 2, 3

eg: No of mobile phones

eg: No of Bank account
1, 2,

eg: Total no of employees in a team.

Any values in a range

eg: Height
Weight
Speed
temperature
IQ
movie duration
length

Categorical

Qualitative data

Nominal
(No order in Categorial data)

eg: Gender
eg: Blood group
eg: Colour
eg: race/ethnicity
eg: Marital Status

Ordinal
(Order)
⇓
Rank.

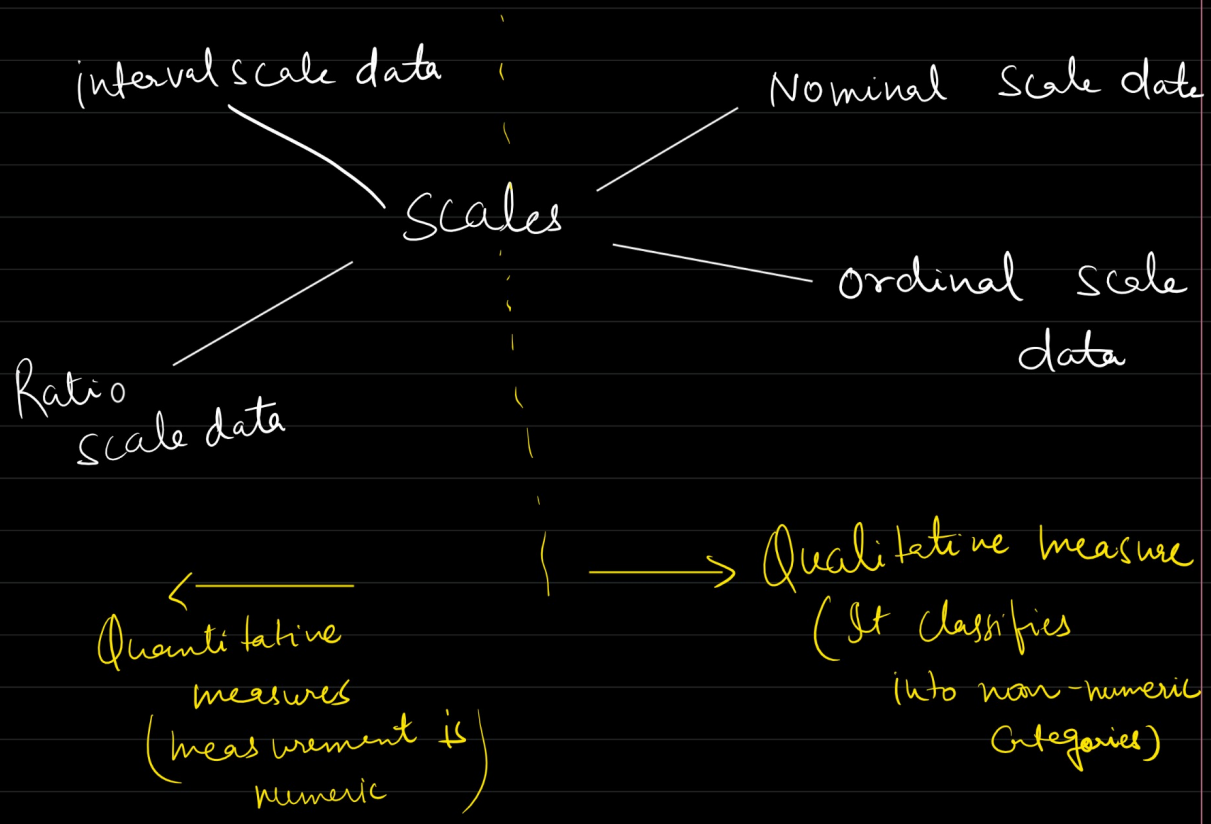
eg: good, better best

eg: Grade
A, B, C, D

eg: Salary
high, medium, low

→ It will take specific finite whole no. values.

Scales of measurement



① Nominal Scale data

→ Qualitative or categorical data
→ Gender, ethnicity, colour, location
→ No Order in the data

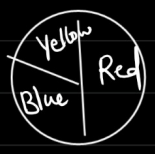
Example:-

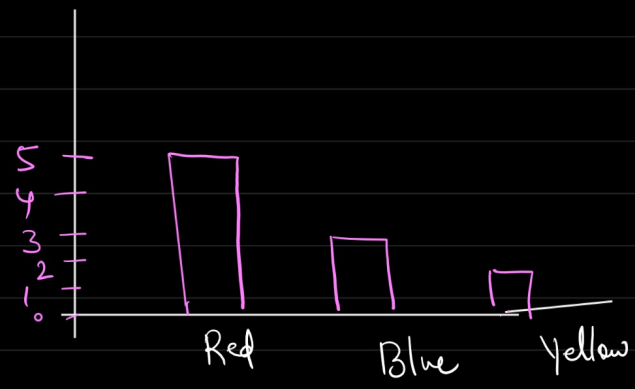
Employees
M
F
M
F
M
M
M

M – 5
F – 2

Example Red – 5 —— 50%.

Blue - 3 — 30%
Yellow - 2 — 20%





② Ordinal Scaled data

→ Order and rank matter.
→ Difference can not be measured

Performance of students
1st
2nd
3rd
4th rank

ex - Salary (high, median, low)

ex - Educational qualification

Rank
1st
2nd
3rd
4th

1st — 2nd

③ Interval Scale data

→ The rank and order has a meaning
→ Difference can be measured (excluding ratio)
→ It doesn't have 0 starting value

Example → Score 90, 85 (5), 70 (5), 50 (20)

Temperature.
inside - 30°C — 0°C
outside - 60°C — -10°C

$\frac{30^\circ C}{60^\circ C} = 1:2$

example - Length
Weight

height
temperature
Revenue of a company

④ Ratio Scaled data

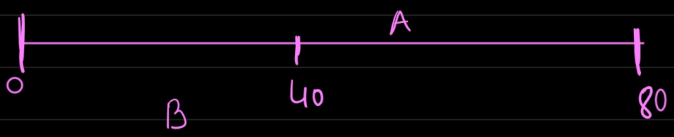
— order and rank has a meaning
— Differences and ratio are measurable
— It does have a 0 starting point

Ex:

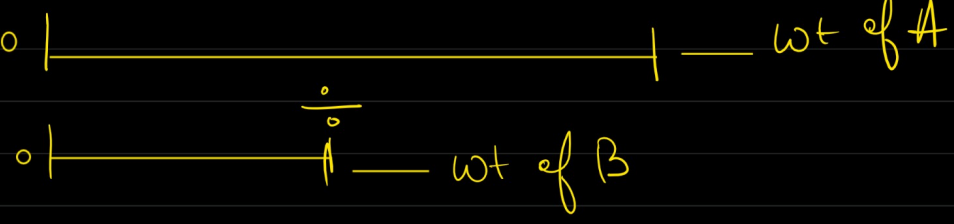
A → 80kg
B – 40kg.

Person A is twice the weight of B

$\frac{80}{40} = 2:1$



→ 0 as true Starting Point



eg ⇒ height
weight
Time
Age

Scale of Measurement

Data	Nominal	Ordinal	Interval	Ratio
Labelled	✓	✓	✓	✓
Meaningful order	X	✓	✓	✓
Measurable difference	X	X	✓	✓
True zero starting Point	X	X	X	✓
Example	Gender Religion Post office code Location	Satisfaction Rating Grade Rank	IQ Temp Score Wt Wt	Height Wt Time Age

* Descriptive Statistic — Summarization of data without adding or subtracting anything at a specific instance.

① Measures of Central tendency

② Measures of dispersion

③ Measures of symmetricity

① Measures of Central tendency

Central → 1, 2, 3, 4, 5

What is one value around which all the data is revolving? — 3

* CT represents the center point of a dataset.

① mean
② median
③ Mode

} EDA and feature Engineering

① Mean (Arithmetic mid value of data)

Population - {1, 2, 3, 4, 5}

$$\mu = \sum_{i=1}^{n} \frac{x_i}{N}$$

Sample (n)

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

Σ → summation $\Rightarrow \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

Summing up all the observations and dividing by no of observation.

② Median (Physical mid point of data)

4, 5, 2, 3, 1, 2

→ Sort the data – 1, 2, 2, 3, 4, 5

→ Count the no. of elements – 6

if count is even

1, 2, 2, 3, 4, 5

median = avg of two middle most element

$= \frac{2+3}{2} \Rightarrow 2.5$

4, 5, 2, 3, 1

→ Sort – 1, 2, 3, 4, 5

→ count – 5

↓

odd

median = the middle most element

= 2

Scenario 1 (mean)

1, 2, 3, 4, 5

$\text{mean} = \frac{1+2+3+4+5}{5} = 3$

Scenario 2.

1, 2, 3, 4, 5

median = 3

1, 2, 3, 4, 1000

→ Here 1000 is an outlier

↓

no which is much higher or lower as compared to other numbers. (Extreme Values)

$\text{mean} = \frac{1+2+3+4+1000}{5} = \frac{1010}{5} = 202$

1, 2, (3), 4, 1000

median = 3

* The mean is affected by outliers whereas median is not affected by outliers.

③ mode — frequency maximum

{2, 3, 1, 1, 4, 4, 4, 3, 4, 2}

mode = 4

Use cases of Central tendency

Age	Gender	Weight	Salary (k)
25	M	80	—
26	— → M	70	50
→ 24.75	M	30	60
23	M	—	~~70~~
25	F	60	45

→ Age is continuous variable

→ impute the missing/null value with mean

$$\frac{25+26+23+25}{4} = 24.75$$

→ Gender → Categorical data

→ Highest frequency

→ M, M, M, F ⇒ Mode — M

* Median is not affected by outlier.