Descriptive Statistics

Collect
↓
Organize
↓
Summarize (analyse/describe)

methods — mean, mode, median, Standard deviation, Variation, Range, percentiles.

exc → 40 students
avg score, spread of variation of scores

Organization — population — (mean, median) Parameter
1000 students
9th class Student 120 — sample — (mean, median) — statistic

Inferential Statistics — inferences or prediction

sample
9th class
↓ 120 students
avg score 70

or
school — population
1050 students
9th class
120 — samples

12th class — 160 students

Sample size = 120
no. of data points in Sample

population 1050 students
avg score

methods: Regression analysis, confidence interval, Hypothesis

BHK	Plot size	location	furniture	Price
2	500	Pune	Kitchen	10 L
1	450	Bang.	Doors	8 L
3	600	mumbai	—	20 L

variable column
target column
computer
manual data

Price — Target Colum, label
BHK, Plot size, loc, furn. — independent columns

4 BHK, 1000 m², Delhi, Doors — price
Kitchen

Accts

Company

2000 Population

Finance, HR, Sales, Account

Sample

150 employees

Parameter

Avg members - 5

Population

Total no of people country

1 L

10 lakhs Records

Population

Females samples

10000

male

20000 Records

Average members in family

City division

Etc

Cities

10000 families

Avg. mansa - 5 people/family

Statistic

parameter	Statistics
1) numerical characteristics of population	numerical characteristics (Avg, mean, median ...) of a sample
2) population	Statistics
3) fixed value for entire Pulation	

School 1050 students

Avg Height - 160 cm - Parameter

9th class student 120

Avg Height of 120 students - 155 cm - Statistics

160 - 155 = 5 cm

(9th class)

12th class

Avg Height of 180 students - 170 cm - Statistics

160 - 170 = -10 cm

parameter

Statistic

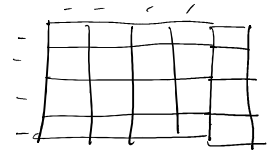
3) fixed Value for population

Variable and vary from one sample to another

1) fixed value for population | Variable and vary from one sample to another

dictionary: {key: value}

Structured
Tabular form
4 Rows 5 columns



name	id	subject	Adj
Ram	E01	math	pune

excel, csv

Semi Structured

name: Ram
age: 19
id: E01
subject: ["math", "bio", "SE", "english"]
Add: {country: India, state: "MH", city: "pune"}

JSON, XML

Unstructured
Images, videos, Audios, text

(Quantitative Data)
Numerical Data: made of numbers

Age	Weight	Height
12	20 Kg	100 cm
15	14.5	160 cm
20		
29	17.6	
6.9	50.	
20.10	100.50 Kg	

Categorical data: made of words
(qualitative Data)

Ex Colour	name	car brands
BMe	Ram	maruti
Green	Shyam	BMW
Red		Skoda

Numerical Data

1) continuous Data: it can take any value within a range
2) infinite no of possible values
3) decimal value
4) that's not fixed value

195.10 marks
12 days, 10 hr, 10 min

Age
0 - 300
0.5

Height
0 - 250 cm

weight | temp.

0 – 300	Height (fixed value)	weight	temp.
0.5	0 – 250 cm	20.4 kg	127.2 °C
1.5	0.5	40.5 kg	35.5 °C
30.5	1 m		
19.5.2	162. cm		
	165.5 cm		
	165. c m		

2) Discrete data:- 1) can take only distinct separate values
2) counted in whole no.
3) finite no. of possible value

(12.5) ×

ex.

no. of student	Shoe size
120	2
190	3
90	4
	5
	7
	8
	9

7.5
9.5 ×

Categorical Data:

1) Nominal Data: no hierarchy or order inside data

ex.:-

colours	fruits	name	gender
Blue	Apple	Ram	Female
green	banana	Shyam	male
Red	Pineapple	mahol	

2) ordinal data: when we get hierarchy or orders or ranking inside data.

education level	class	customer satisfaction
high school	8th	excellent – 5
college	9th	good – 4
graduate	5th	fair – 3
masters	10th	poor – 2
phd		bad –

1) business problem - To find out best comfortable weather location all over world

2) longitude / lattitude / Highest temp. / lowest temp. / forest % / population / air quality

3) Sample size = 10000 cities

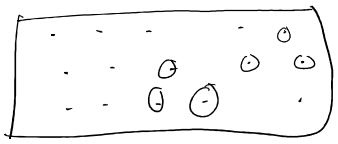
4) max. temp. / min. temp. — descriptive statistics

Probability sampling - It randomly selects a sample from the population, ensuring that each member of the population has an equal chance of being selected in sample.

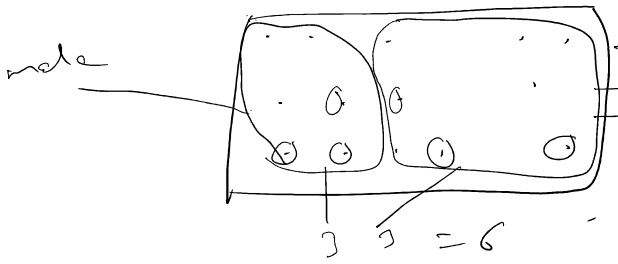
1) Simple Random sampling - every data points has an equal chances of being selected

15 Players cricket team

Sample size = 6



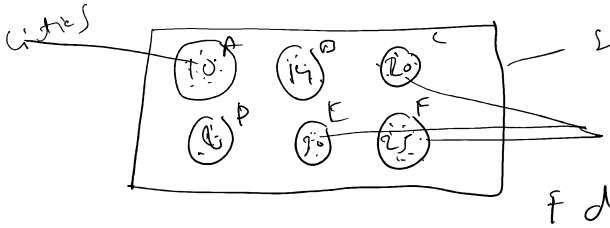
2) Stratified sampling :- Population is diveded into groups (Strata) and samples are randomly selected from each Stratum.



Sample size = 6

$\frac{6}{\text{no. of groups}} = \frac{6}{2} = 3$

3) Clustering sampling - populatn is divided into clusters, and entire clusters are randomly selected.



If decision

Sample size = 25

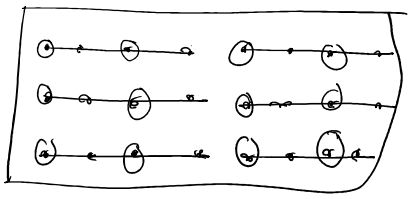
madhya pradesh - 9 division

Jabalpur

(12 cities)

4) systematic sampling - every nth datapoints selected from a list after a random starting point.

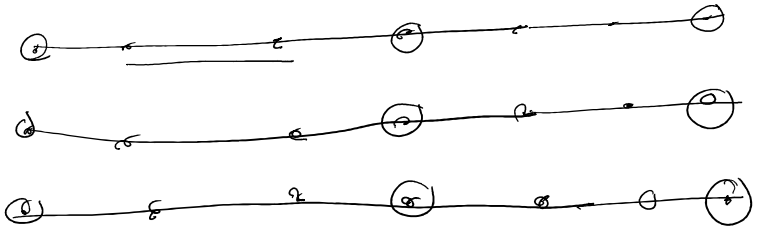
ex 1)



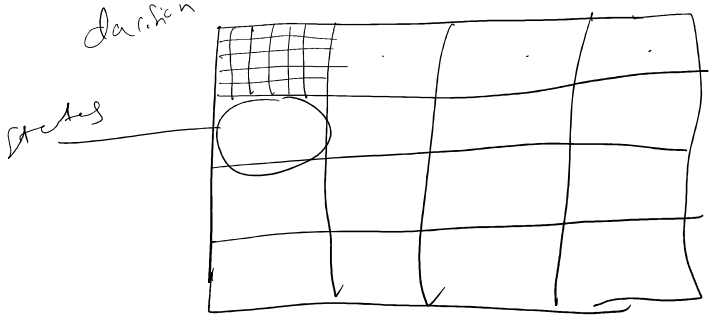
dance team = 12 students

example

restaurant feedback



5) multistage sampling -



Country
↓
states
↓
division
↓
cities - people
S IC

Non-probability sample - not every data points of the population has an equal chances of being selected.

1) convenience sampling - select sample on availability accessibility
2) Judgement sampling - knowledge or judgements
3) purposive sampling -

Research wants to identify this disease in dogs.

dog horse horse

dogs — sample

Variance = $\sqrt{9.01\ min^2}$

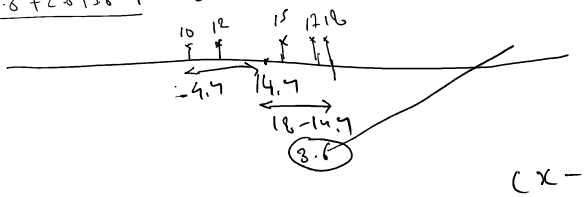
Standard deviation σ = 3. min

A = [10, 18, 12, 15, 17]

(10) (12) (15)

$\frac{-4.4 + (-2.4) + 0.6 + 2.6 + 3.6}{5}$ Range = 18 − 10 = 8

mean = 14.4



$= \frac{0}{5} = 0$

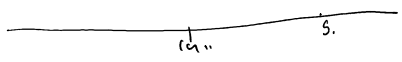
mean

Avg variation = Variance = $\frac{(x-\bar{x})^2}{n}$ — n = total no of datapoint

$= \frac{(10-14.4)^2 + (18-14.4)^2 + (12-14.4)^2 + \dots}{5}$

Variance = $\sqrt{9.01\ min^2}$ = 9.01

Std dvn σ = 3 min



Percentile : = $\frac{\text{number of value below x (determined)}}{\text{total no. of value}} \times 100$

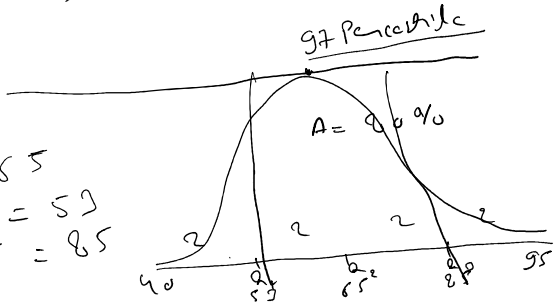
original [45, 49, 40, 53, 65, (71), 61, 79, 85, 91]

sorted = [40, 45, 49, 53, 61, 65 (71), 75, 85, 91]

95

Percentile of 71 = $\frac{6}{10} \times 100$ = 60 Percentile

1000 students



Q_2 = 50th Percentile = 65

Q_1 = 25th Percentile = 53

Q_3 = 75th Percentile = 85

10K Km/year

[10K, 20K, 25K, 16K, 30K, (10K)]

10K Km/sec

[10K, 20K, 25K, 16K, 30K, 10K]

Reasons for outlier
1) human mistake
2) machinery/measuring tool mistake
3) experimental mistake

A = [10, 12, 15, 17, 18, 100]

before
$\mu = 14.4$
$\sigma = 3$
range = 18 − 10 = 8

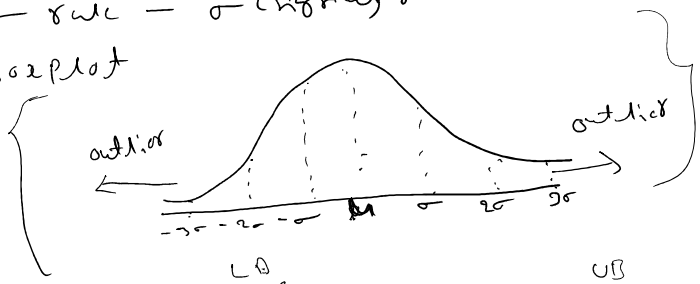
after
$\mu = 57.2$
$\sigma = 9$
range = 100 − 10 = 90

57.2 / 14.4 / 3c

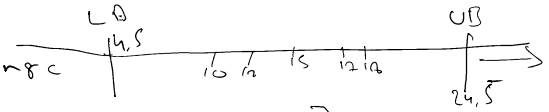
methods for identify outlier - $\sigma = \sqrt{\frac{(x-\mu)^2}{n}}$

1) Interquartile range
2) z-score method
3) σ rule — σ (sigma) rule
4) Boxplot

σ - rule



Interquartile range



[10, 12, 15, 17, 18]

Q_1 = 25th percentile = $\frac{25}{100} \times 5 = 1.25 \approx 2^{nd}$

IQR = $Q_3 - Q_1$

= 12 − 2.5

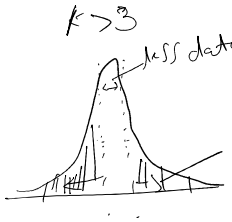
= 12

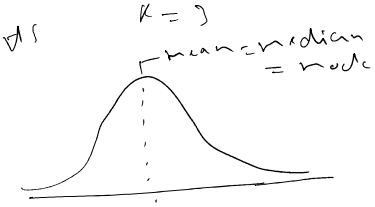
Q_3 = 75th percentile = 17 $\quad \frac{75}{100} \times 5 = 3.75 \approx 4^{th}$

Upper boundary $UB = Q_3 + 1.5 (Q_3 - Q_1)$ (1.5 = threshold value)
$= 17 + 1.5 (17 - 12) = 17 + 7.5$
$= 17 + 1.5 (5)$
$= 24.5$

Lower boundary $LB = Q_1 - 1.5 (IQR) = 12 - 1.5 (7.5)$
$= 4.5$

$UB = Q_3 + 1.5 (IQR)$
$LB = Q_1 - 1.5 (IQR)$





mean ≠ median ≠ mode

$\neq$ mode

$$\text{Kurtosis} = \frac{\sum (x-\bar{x})^4}{n s^4}$$

$K < 3$ multiple mode

most outlier

$\bar{x}$ = mean
s = std dvn
n = no. of data points

Central Limit theorem

sample size > 30

Sample 1 = 40 data points $= \bar{x}_1 \approx 20$

Total no. of data points

Sample 2 = 50 data points $= \bar{x}_2 \approx 30$

3 = 30 $= \bar{x}_3$

4 = 60 $\bar{x}_3$

Sample 10 = 150 $= \bar{x}_{10}$

[10, 12, 15, 17, 18, 21]

(6)

mean $\mu = 14.4$

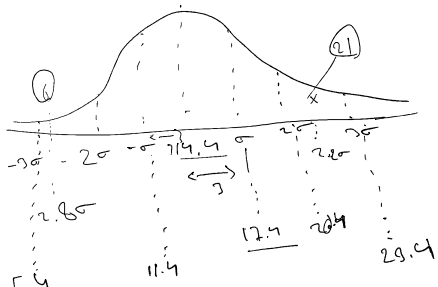
std dvn $\sigma = 3$

$$z = \frac{6-14.4}{3} = \frac{-8.4}{3} = -2.8$$

$$z \text{ score} = \frac{x-\mu}{\sigma}$$

$$= \frac{21-14.4}{3} = \frac{6.6}{3} = 2.2\sigma$$

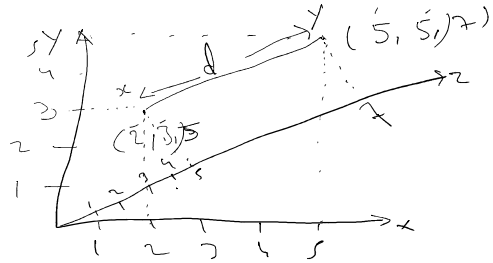
$$\frac{18-14.4}{3} = \frac{3.6}{3} = 1.2$$



$$z \text{ score}(10) = \frac{10-14.4}{3} = \frac{-4.4}{3} = -1.33$$

$\sigma = 17.4 - 14.4 = 3$

Euclidian Distance



$d = 7.6$

$N(x_1, s_1)$ $N(x_2, s_2)$

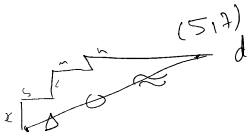
$$d = \sqrt{(x_2-x_1)^2 + (s_2-s_1)^2} = \sqrt{9+4}$$

$$\sqrt{-2^2 + (5-3)^2}$$

$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{9 + 4}$

$d = \sqrt{(5-2)^2 + (5-3)^2}$

$= \sqrt{13} = 3.6$



$d = |x_2 - x_2| + |y_2 - y_1|$

$= x + 5 + 2 + m + n$

$(3, 2)$

$(4, -2)$

$(-3, -2)$

5

$[-1-(-3)] + [2-(-2)] = 2 + 4 = -2$

$5 - (-3) + (7 - 2)$