Statistics

Stats is all about data

Raw data

- May have errors
- Not validated
- Unformatted
- Uninterpretable
- Not cleansed
- …

State	District	Jun	Jul	Aug
Andhra Pradesh	Adilabad	213.245	260.107	449.676
Andhra Pradesh	Vizianagaram	265.521	139.41	266.612
Arunachal Pradesh	Changlang	264.225	323.216	371.473
Arunachal Pradesh	DibangValley	214.674	316.565	336.434
Assam	Karimganj	164.937	270.082	589.803
Assam	Kokrajhar	580.901	312.208	564.161
Bihar	Supaul	173.777	152.199	200.673
Bihar	Vaishali	126.427	120.119	134.309
Chandigarh	Chandigarh	87.6	236.5	134.6
Chattisgarh	Bastar	318.126	255.674	366.698
Chattisgarh	Rajnandgaon	213.481	378.729	229.806
Chattisgarh	Surguja	227.882	210.418	159.516
Dadra & Nagar Haveli	Dadra & Nagar Haveli	341.727	603.201	234.86
Delhi	New Delhi	80.69	272.234	125.493
Gujarat	Ahmadabad	55.405	335.661	81.557
Gujarat	Amreli	55.892	376.289	103.858
Gujarat	The Dangs	280.156	585.72	242.533
Haryana	Ambala	93.162	237.152	141.453
Haryana	Bhiwani	58.104	219.684	75.383
Himachal Pradesh	Chamba	90.188	145.487	141.654
Himachal Pradesh	Hamirpur	96.383	201.116	147.078

Rainfall in Districts of Indian states during the monsoon season

Processed data

- No errors
- Validated
- Formatted
- Interpretable
- Cleansed
- …

State	Average (cm)
Andhra Pradesh	239.383
Arunachal Pradesh	239.4495
Assam	372.919
Bihar	150.102
Chandigarh	87.6
Chattisgarh	265.8035
Dadra & NagarHaveli	341.727
Delhi	80.69
Gujarat	55.6485
Haryana	75.633
Himachal Pradesh	93.2855

Average rainfall (in cms.) in Indian states

About data

Data collection

Data is collected in different ways:

- Census
- Observation
- Convenience sample
- Random samples
- Historical data (data collected over time)
- Any other

Data forms

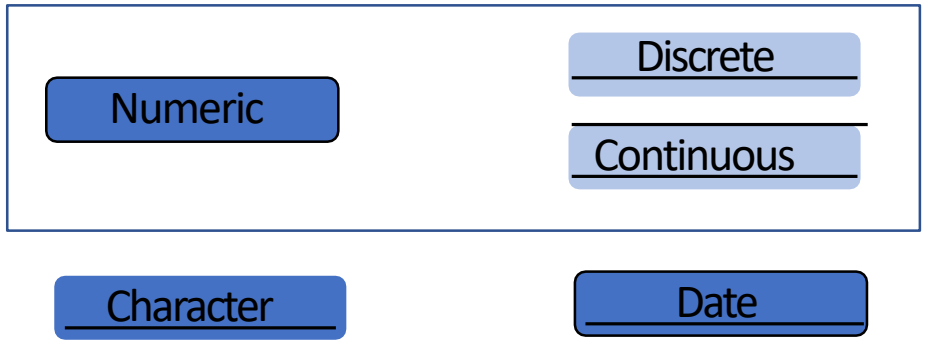
Data can be in any of these forms

- Structured (rows and columns)
- Semi-structured (XML / JSON)
- Unstructured (free text)

Data collected method

- Batch
- Real-time

Data types



day	Date	Open	High	Low	Close	Total Shares
1	01-Jan-15	171	171	167.5	168.35	35357
2	02-Jan-15	169.2	172.8	168.75	171.7	100909
3	05-Jan-15	171.3	171.95	167.45	168.85	95765
4	06-Jan-15	169	172.9	167.3	168.15	134474

Data Types

Numeric data can be of 2 types:

- **Discrete data**

Eg:
Year – 1972,1998,2005,2018 …
Age: 12,18,24,39,40 …

- **Continuous data**

Eg:
Weight – 43.1,55.4,76.9 …
Temperature: 31.1,33.4,90.5 …

Character data can be of 2 types:

- **Strings and literals**

Eg: “computer”, “Statistics” …

- **Factors**

Factor Data Types

Nominal	Ordinal	Interval	Ratio

Factor Data Types

Nominal	Ordinal	Interval	Ratio

- Used as names / labels without any quantitative measure
- No numerical significance

- Examples:

Gender

Male
Female

Marital Status

Single
Married
Divorced

Religion

Hindu Jain
Buddhism
Sikh
Christian

Factor Data Types

Nominal	Ordinal	Interval	Ratio

- Order is important, rather than the name
- Difference between 2 values is not really known

- Examples:

Income Level

1 = Low
2 = Middle
3 = High
4 = Very high

Feeling today

1 = Very unhappy 2
= Unhappy
3 = Ok
4 = Happy
5 = Very happy

Rating

3 = Very Good
2 = Good
1 = Bad

Factor Data Types

Nominal	Ordinal	Interval	Ratio

- Numerical scales where order and difference are known
- Do not have a true 0

•Examples:

Temperature

Time

Marks

Temp.
0
5
10
15
20
25

Marks	Freq.
90-100	2
80-89	3
70-79	7
60-69	11
50-59	15
40-49	3
30-39	4
20-29	5
0-19	0

Data Types

Nominal	Ordinal	Interval	Ratio

- Numerical scales where order and difference are known
- Has a true 0 (means "does not exist")
- Descriptive and Inferential statistical analysis performed

- Examples:

Height
Weight
Age
Income
Years of education

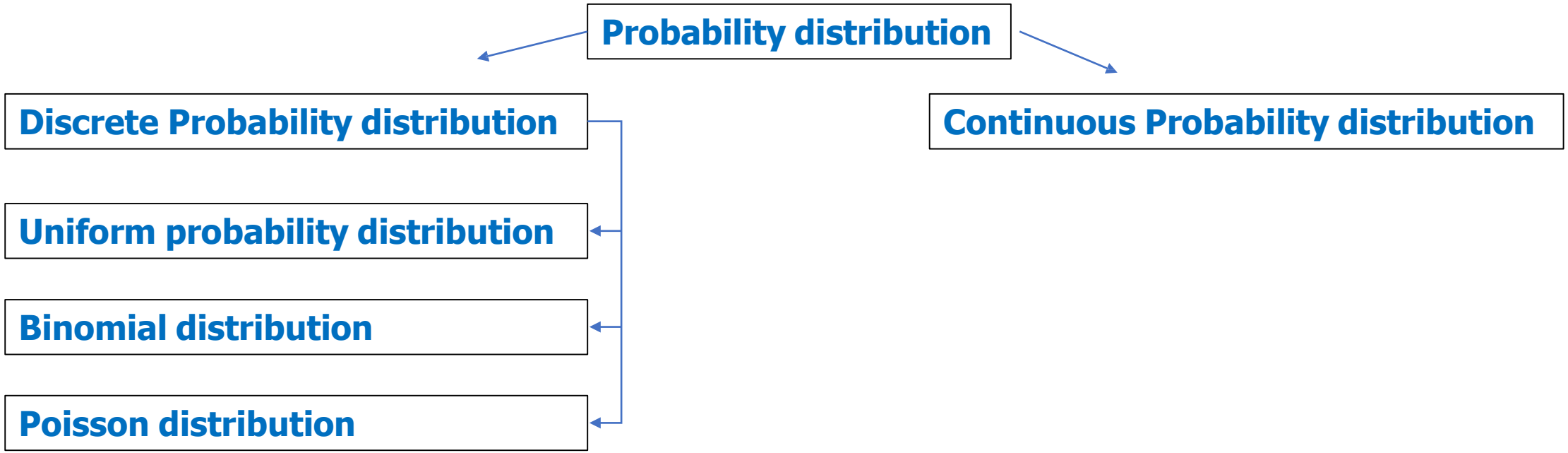
Data Distribution

Frequency distribution

Listing of the observed frequencies of all outcomes of an experiment that actually occurred when the experiment was done

Probability distribution

Listing of all the probabilities of all the possible outcomes that could result if the experiment were done



Data Mining

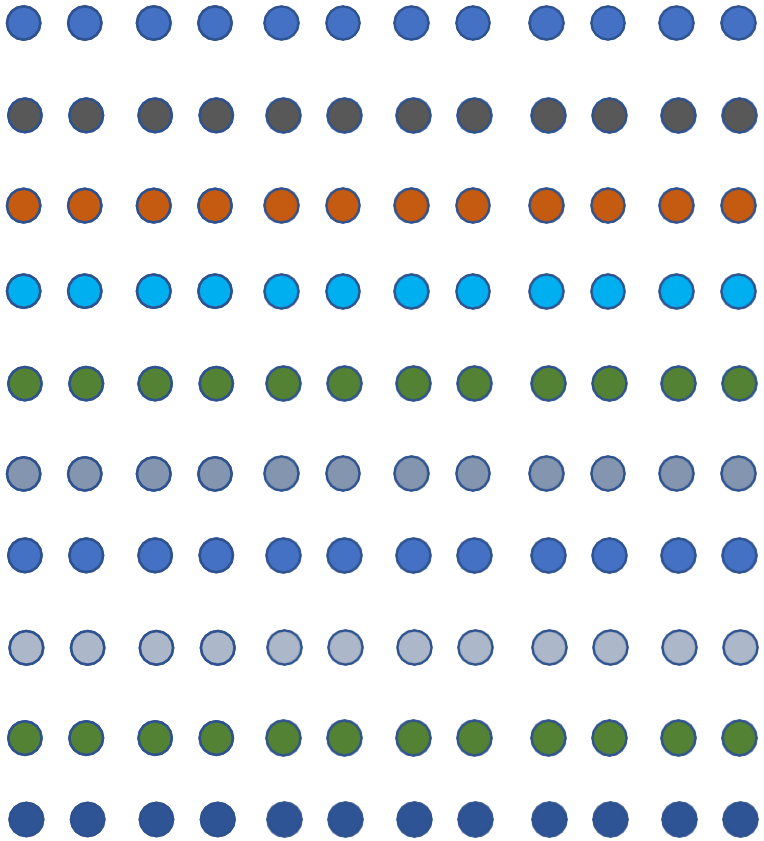
- To discover patterns in a dataset involving statistics, database concepts and machine learning
- Extract this information to transform them into useful interpretations
- Data Mining is an essential part of a process commonly known as KDD (Knowledge discovery in databases)

Some typical tasks in Data mining

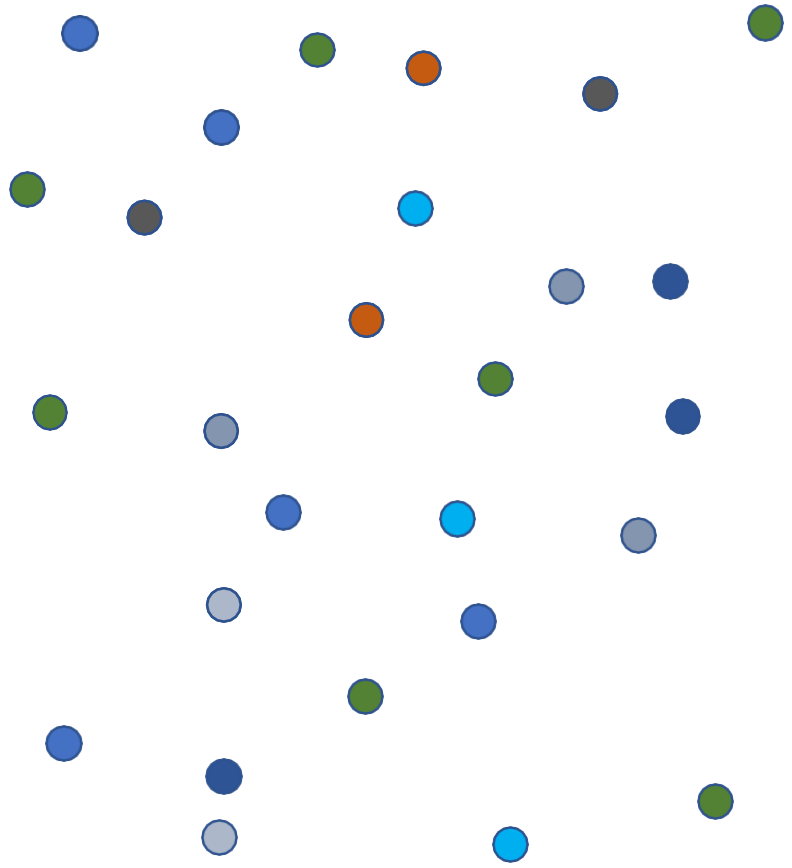
1) Anomaly detection in data
2) Association rule analysis (Apriori)
3) Clustering (Unsupervised machine learning)
4) Classification and Regression (Supervised machine learning)
5) Visualization

Population vs Sample

Population



Sample



Parameter

μ ⟵ Mean ⟶ $\bar{x}$

σ ⟵ Standard Deviation ⟶ s

Statistic

Types of Statistics

Descriptive Statistics

Describes the various aspects of dataset

- **Measure of Central tendency**
  - Mean Weighted Mean
  - Geometric Mean
  - Median
  - Mode
- **Measure of Dispersion**
  - Range
  - Interfractile Range
  - Quartiles
  - Interquartile Range
  - Standard Deviation
  - Variance
- **Measure of Association**
  - Correlation
  - Covariance
  - Coefficient of Covariation

Inferential Statistics

What conclusion can be drawn from the dataset

- Estimation
- Hypothesis Testing

Descriptive Statistics

I. Measure of Central tendency

Central tendency measures the centre value / middle value / average value of a given dataset

1. Mean
2. Median
3. Mode

1. Mean

- Arithmetic mean is the "average" of a range of data (numeric)
- Common examples: Test Marks, Temperature, Runs scored in cricket etc.
- Conventional symbols:
  - ✓ n = sample size
  - ✓ x = observation(s)
  - ✓ $\bar{x}$ = sample mean
  - ✓ μ = population mean
- Arithmetic Mean $\bar{x} = (\sum x / n)$

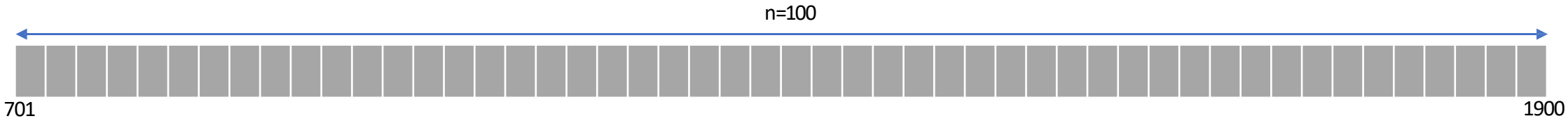
Advantages	Disadvantages
• A single number represents a whole dataset • Intuitively clear • Only 1 mean per dataset – easy for comparison	• Affected by extreme values – so not a reliable measure • Every value is taken for calculation (use grouped data) • Cannot compute mean for open-ended classes

Open-ended classes

One or more classes do not have a boundary

height	frequency
4.0 - 4.5	10
4.6 - 5.0	8
5.1 - 5.3	20
5.4 - 5.5	19
5.6 - 5.8	20
> 5.8	4

Grouped Data - Mean



Minimum score = 701
Maximum score = 1900
Interval = 100
Total classes = 12
Total observations = 100
N*x=125000

#	min	max	range	n		
1	701	800	701-800	4		
2	801	900	801-900	7		
3	901	1000	901-1000	8		
4	1001	1100	1001-1100	10		
5	1101	1200	1101-1200	12		
6	1201	1300	1201-1300	17		
7	1301	1400	1301-1400	13		
8	1401	1500	1401-1500	10		
9	1501	1600	1501-1600	9		
10	1601	1700	1601-1700	7		
11	1701	1800	1701-1800	2		
12	1801	1900	1801-1900	1		
				100		

Mean = $\Sigma(F_x/n)$ = 1250

Exercise

Frequency distribution represents the time in seconds to servecustomers at a local store. Compute the sample mean of the servingtime

time	frequency
20-29	6
30-39	16
40-49	21
50-59	29
60-69	25
70-79	22
80-89	11
90-99	7
100-109	4
110-119	0
120-129	2

min	max	x	n	n*x	avg
20	29	24	6	144	
30	39	34	16	544	
40	49	44	21	924	
50	59	54	29	1566	
60	69	64	25	1600	
70	79	74	22	1628	
80	89	84	11	924	
90	99	94	7	658	
100	109	104	4	416	
110	119	114	0	0	
120	129	124	2	248	
			143	8652	60.5

4. Median

- Position based single value that measures the central item in a dataset
- Middlemost / Centremost item in a dataset
- About half of the items lie above this point; and the other half below it
- To calculate Median, data needs to be sorted (Ascending / Descending)

- Formula for Median
  - ✓ For non-grouped data
    - $[(n+1) / 2]^{th}$ item, when ***n*** is odd
    - $[(n/2)^{th} + ((n/2)+1)^{th}) / 2]$ item, when ***n*** is even
  - ✓ For grouped data
    - $(n/2)^{th}$ item

Advantages
• Extreme values do not effect Median strongly
• Easy to calculate

Disadvantages
• Needs sorting of data before calculation
• Can be time consuming in large datasets

Median

Item	Time
1	10.2
2	10.3
3	10.7
4	10.8
5	11
6	11.1
7	15

n = 7
$[(n+1)/2]^{th} = 4^{th}$ 4^{th} element = 10.8
Median = 10.8

Exercise: Calculate the Median

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
42	53	90	81	120	41	42	29	87	11	35	69	40	77	97	63

1: Sort the dataset

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
11	29	35	40	41	42	42	53	63	69	77	81	87	90	97	120

n = 16
$[(n/2)^{th} + ((n/2)+1)^{th})] / 2$
$[8^{th} + 9^{th}] / 2 = (53 + 63) / 2 = 58$ ->Median

Median for Grouped data

Consider the monthly balance of 100 customers of a bank, calculate the Median monthly balance.

class	Balance	Freq	Cum freq
1	0-20	17	17
2	20-40	28	45
3	40-60	32	77
4	60-80	24	101
5	80-100	19	120

Formula
n = 120
(n/2) = 60 element

60 lies between 45 and 77
So it will be our median class

L= Lower limit of the median class
n= No. of observation
f = Frequency of the Median class
cf= Cumulative frequency of the class preceding the median class
h = class size

$$\text{Median} = L + \frac{n/2 - cf}{f} \times h$$

Solving for x.
x = 49.37

5. Mode

- A single value that is repeated most often
- Used for both qualitative and quantitative data
- Bimodal distribution – different values repeated same number of times

Unit produced / hour

1	3	7	10	20
1	3	7	11	20
2	5	9	12	20
2	5	9	16	24

Most popular colours for dresses

Black	White	Green	Blue	Blue
Blue	Green	Yellow	White	Green
Black	Pink	Black	Yellow	Green
Green	Green	Green	Blue	Blue

Frequency Distribution

0-2	4
3-9	9
10-20	7
>20	1

Advantages

- Not affected by extreme values
- Can be used even for open-ended classes

Disadvantages

- Datasets may not contain repeated values
- In case of many modes, interpretation may be difficult

Mode for Grouped data

class	Balance	Freq	
1	125-130	7	F0
2	130-135	14	F1-modal class
3	135-140	10	F2
4	140-145	10	
5	145-150	9	

Modal class- the class which has the highest frequecy

L= Lower limit

f1 = Frequency of the Modal class

h = class size

$$\text{Mode} = L + \frac{f1 - f0}{2f1 - f0 - f2} \times h$$

Grouped Mean

1. Consider the following frequency distribution. Calculate the mean weight of students.

Class Interval	0-10	10-20	20-30	30-40	40-50
Frequency	12	16	6	7	9

lass Interval	Frequency f_i	Class Mark x_i	$(f_i.x_i)$
0-10	12	5	60
10-20	16	15	240
20-30	6	25	150
30-40	7	35	245
40-50	9	45	405
	$\sum f_i=50$		$\sum f_{i.} x_i=1100$

Mean = $\sum(f_i.x_i)/\sum f_i$ = 1100/50 = 22

2. Calculate the median for the following frequency distribution.

Class Interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	10	16	24	15	7

Solution

lass	Frequency	Cumulative Frequency
0-8	8	8
8-16	10	18
16-24	16	34
24-32	24	58
32-40	15	73
40-48	7	80
	$N = \sum f_i = 80$	

Now, N = 80 = (N/2) = 40.

The cumulative frequency just greater than 40 is 58 and the corresponding class is 24-32.

Thus, the median class is 24-32.

l = 24, h = 8, f = 24, c_f = c.f. of preceding class = 34, and (N/2) = 40.

Median, M_e = l+ h{(N/2-cf)/f}