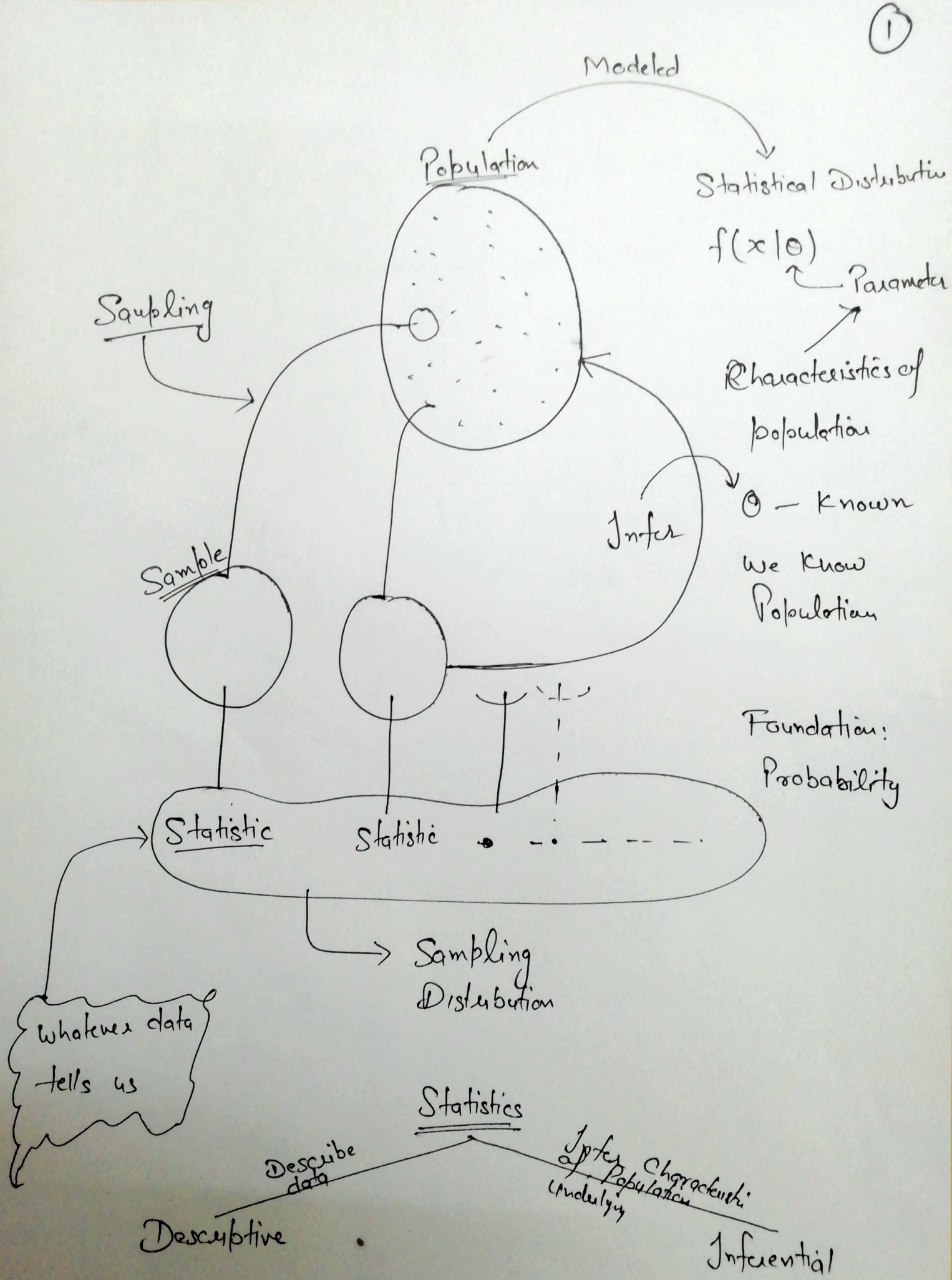

Modeled
Population
Statistical Distribution
f(x|θ)
Parameter
Characteristics of population
Sampling
θ – known
We know Population
Infer
Sample
Foundation: Probability
Statistic
Statistic
Sampling Distribution
Whatever data tells us
Statistics
Describe data
Infer Characteristics of Population underlying
Descriptive
Inferential

Descriptive Statistics

- Displaying Descriptive Statistics
  - Summarizing data in tables, charts and graphs
  - Frequency Distribution
    - A simple way to summarize raw data in tables and make the information more useful
- Calculating Descriptive Statistics
  - Measures of Central tendency
    - Mean
    - Median
    - Mode
    - — —
  - Measures of variation
    - Variance
    - Standard deviation
    - Range
    - — —

Frequency distributions

Showing your data in a table

Data set: The daily demand for hammers at a hardware store over the last 20 days

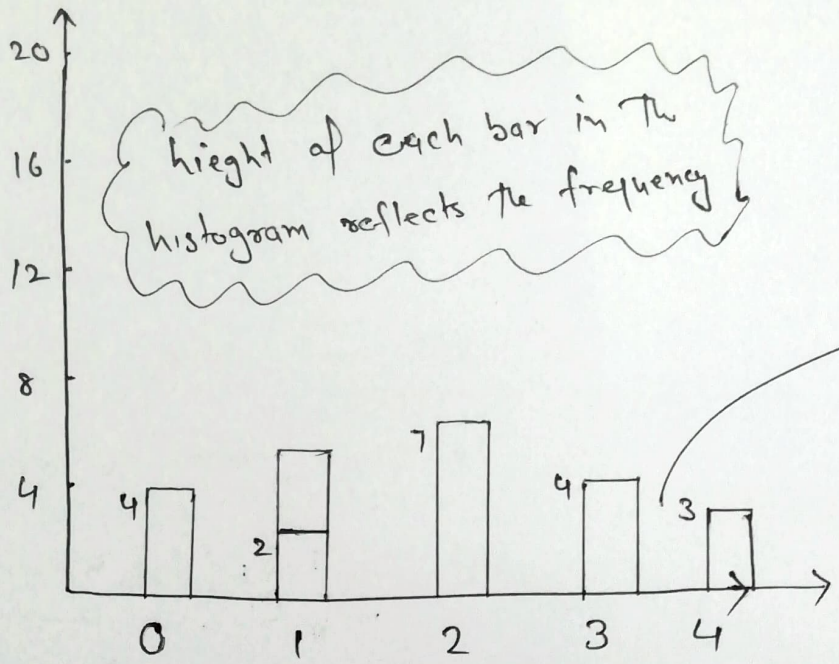
Daily demand				
2	1	0	2	1
3	0	2	4	0
3	2	3	4	2
2	2	4	3	0

A frequency distribution is a two-column table. In left column, list each values in the data set from least to greatest. Count the no. of times each value appears and record data in right column

Daily demand	Frequency	Cummulatin Frequency	Relatin freq.
0	4	4	$4/20 = 0.20$
1	2	6	0.10
2	7	13	0.35
3	4	17	0.20
4	3	20	0.15
	20		1

always equal to 1

Histograms — frequency distribution in a chart



The height of each bar in the histogram reflects the frequency.

A gap exists between columns because the data is discrete. Discrete data is data that can take on a countable no. of possible values.

Data: The mileage of a specific car with a full tank of gas.

Miles per tank

302	315	265	296	289	301	308
280	285	318	267	300	309	312
299	316	301	286	281	311	272
295	305	283	309	313	278	284
296	291	310	302	282	287	307
305	314	318	308	280		

$$W = \frac{\text{Largest value} - \text{Smallest value}}{\text{number of classes}}$$

Now data has many possible outcomes.

So group into classes

K – no. of classes

n – total no. of data points

$$2^K \geq n$$

least K satisfy

$$W = \frac{L - S}{K}$$

W – width of class, L – largest value, S – smallest value

So we have

Largest value — 318

Smallest value — 265

total data puts — 40

Therefore using $2^K \geq n$ rule

Set $K = 6$ $\because 2^6 \geq 40$

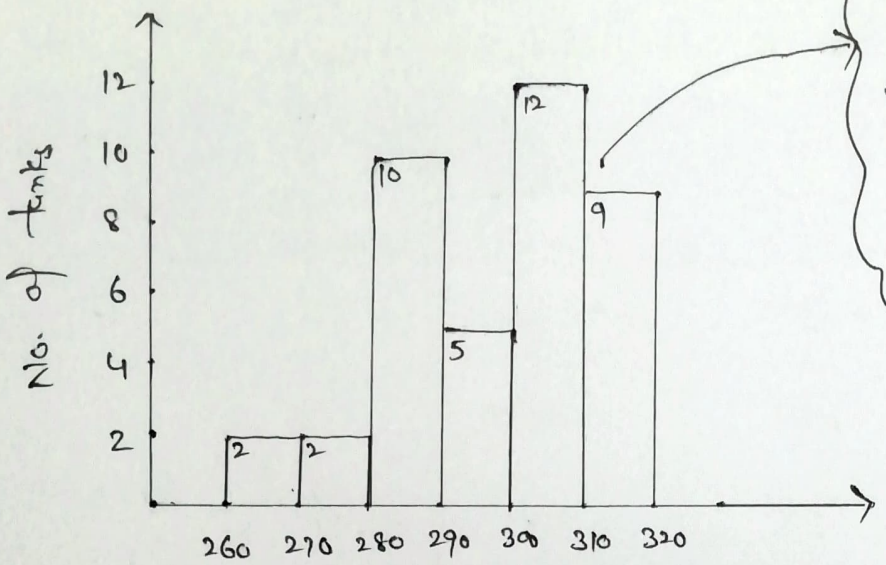
Subsequently

$$W = \frac{318 - 265}{6} = 8.833 \equiv 10$$

Set the size of each class equal to 10.

Count the no. of values contained in each class

Miles per tank	Frequency
under 260 — 270	2
270 — 280	2
280 — 290	10
290 — 300	5
300 — 310	12
310 — 320	9
	40



No gap exists because The data is continuous. Continuous data can assume any value in an interval

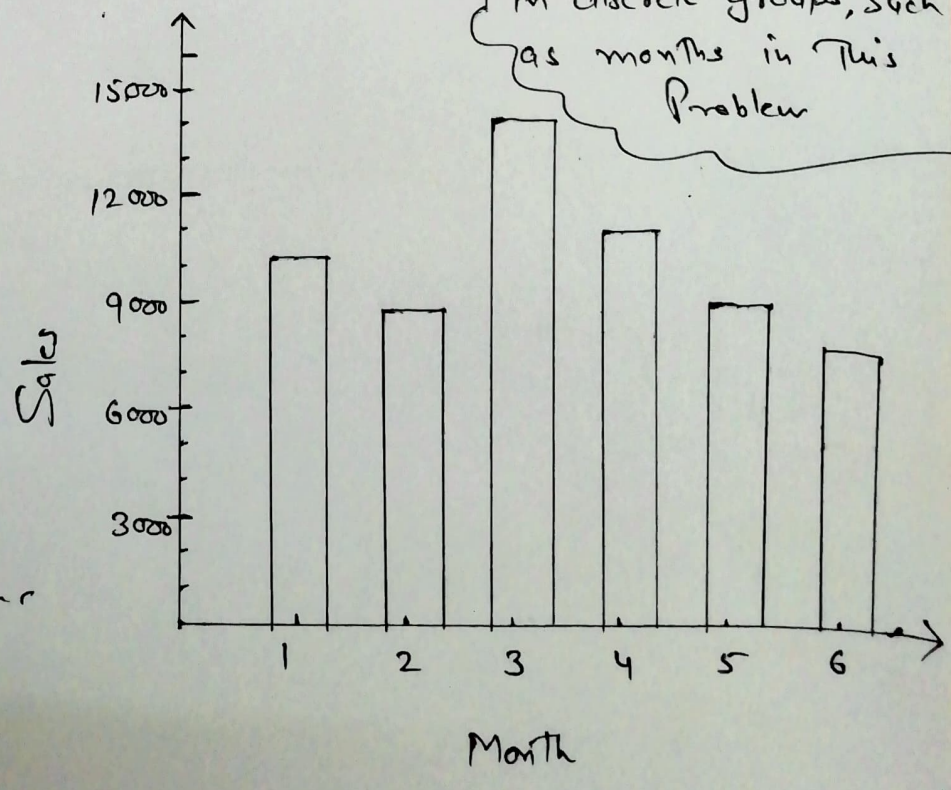
* for Categorical data

Bar Charts Setting The bar for visual data

Data: a Company's monthly sales totals

Categorical data is data That is organized in discrete groups, such as months in This Problem

Month	Sales
1	10734
2	8726
3	14387
4	11213
5	9008
6	8430

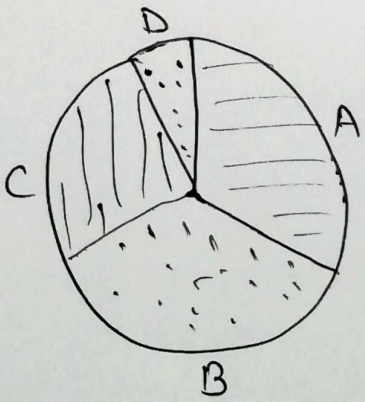


Pie Charts

Showing your categorical data in a circle

Data: a grade distribution for a college class

Grade	No. of Students	Relative frequency	Central angle
A	9	$9/30 = 0.30$	$0.30 \times 360^\circ = 108^\circ$
B	12	0.40	144°
C	7	0.23	83°
D	2	0.07	25°
Total	30	1.00	360°



Sorry I am very Poor in drawing

Plot in your Lab

Showing Categorical Data

- Bar chart
- Pie chart

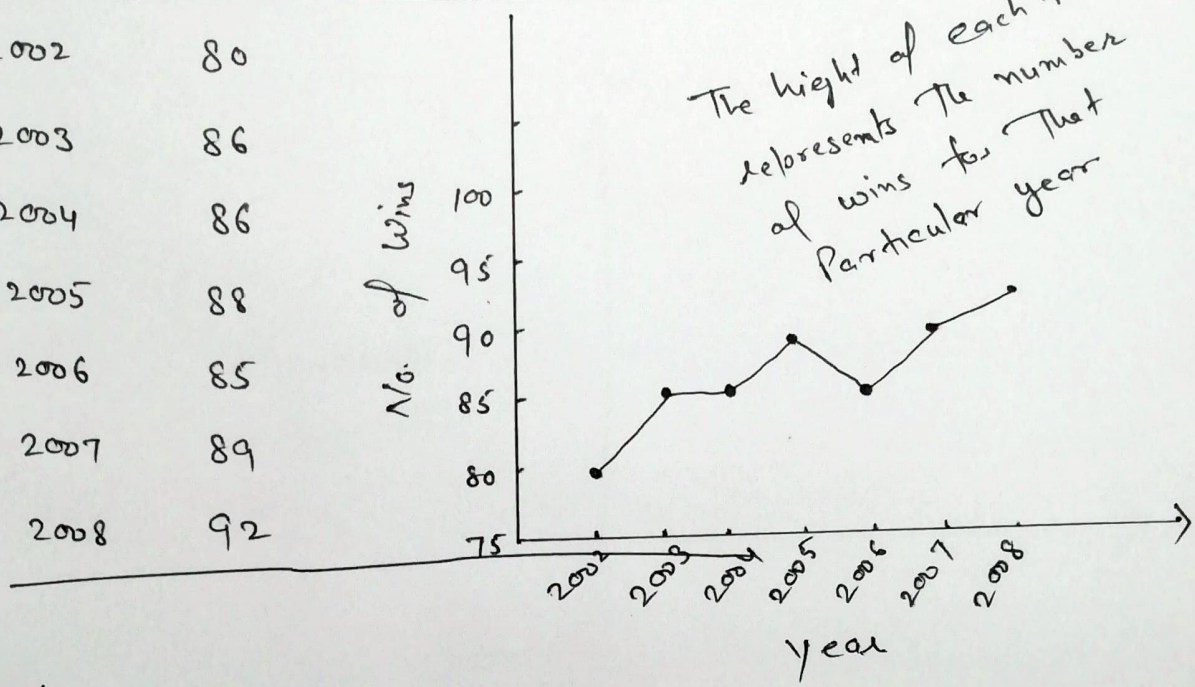
Histogram

- Discrete data
  - gap
- Continuous data
  - no gap

Line Charts : data over time in a chart

Data: The number of wins recorded by The India for seven seasons

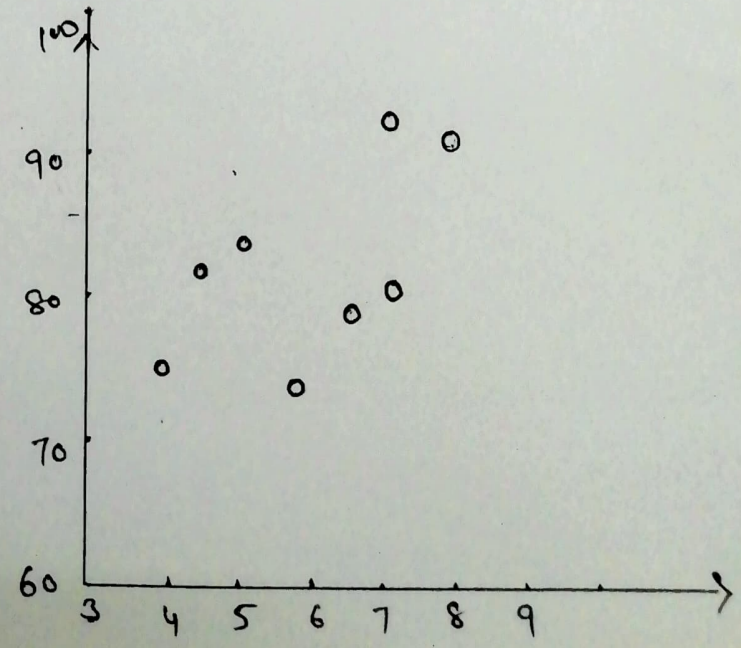
Year	No. of Wins
2002	80
2003	86
2004	86
2005	88
2006	85
2007	89
2008	92

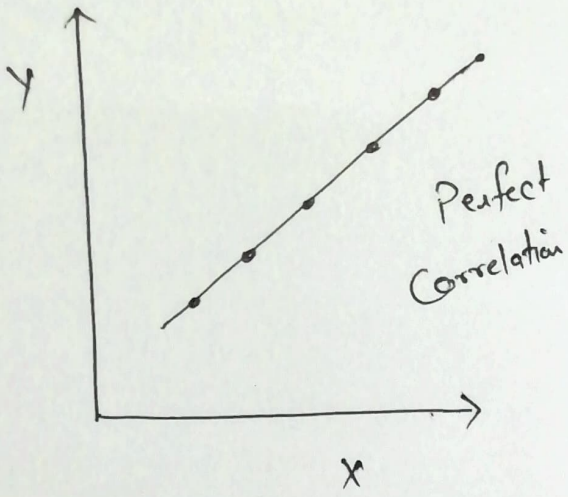


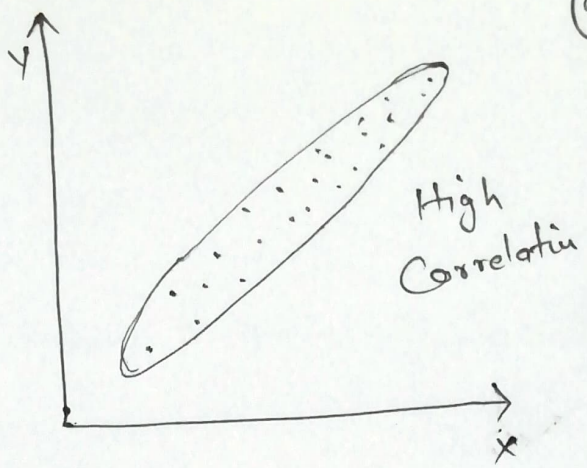
Scatter Charts Illustrate relationship between two variables

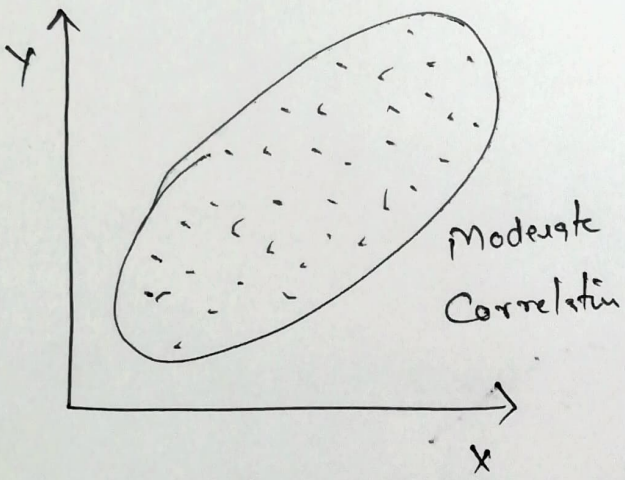
Data: The number of hours eight students studied for an exam and The scores They earned on The exam.

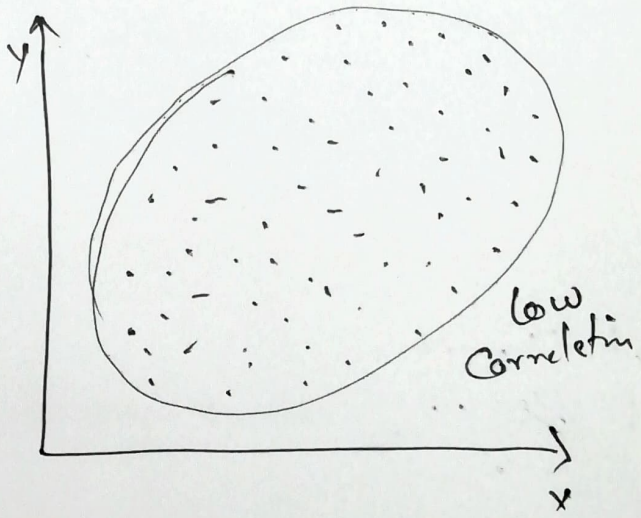
Study hours	Exam Score
5	84
7	92
4.5	82
7	80
8	90
6.5	78
5.5	74
4	75

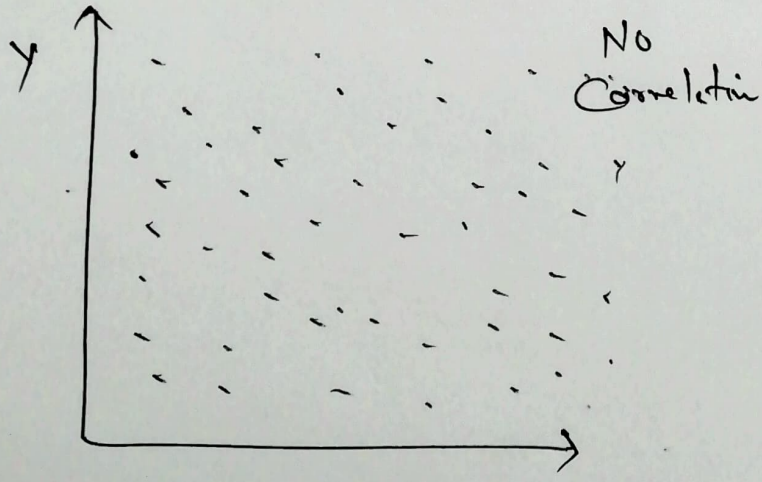


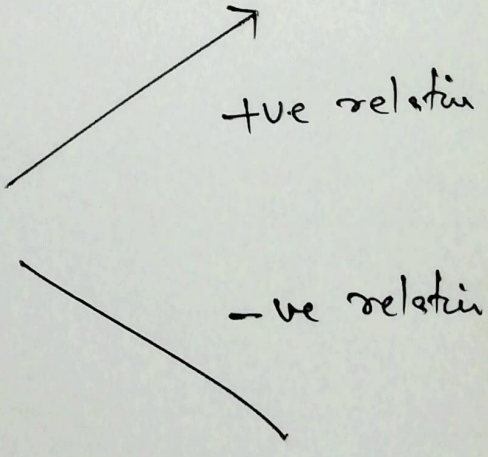

Y
Perfect
Correlation
X


Y
High
Correlation
X


Y
Moderate
Correlation
X


Y
Low
Correlation
X


Y
No
Correlation


+ve relation
-ve relation

Calculating Descriptive Statistics

→ Measures of central tendency

Center of data

→ Mean, Median, Mode, Percentile

Mean.

$X = (x_1, x_2, \dots - x_n)$ observations

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

Let $Y = X + a$, a is any number

$y_1 = x_1 + a$

$y_2 = x_2 + a$

- - - -

$y_n = x_n + a$

Then $$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} (x_i + a)$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{1}{n} \cdot na$$

$$\bar{y} = \bar{x} + a$$

Similarly $Y = X - a$ Then $\bar{y} = \bar{x} - a$

Let $Y = aX$

$y_1 = ax_1$

$y_2 = ax_2$

- - -

$y_n = ax_n$

Then $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} ax_i$

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

$= a\bar{x}$

Similarly $Y = X/a$ Then $\bar{Y} = \frac{1}{a}\bar{x}$

More Precisely

Let $Y = \frac{X - A}{h}$

Then $\bar{Y} = \frac{1}{h}X - \frac{A}{h}$

$\bar{Y} = \frac{1}{h}\bar{x} - \frac{A}{h}$

$h\bar{Y} = \bar{x} - A$

$\bar{x} = A + h\bar{Y}$

Let us take data having N data points/observations

$x_1, x_2, x_1, x_3, x_4, \ldots x_m$

$x_2, x_1, x_3, x_m, x_1, \ldots x_2$

$\ldots$

$x_3, x_4, x_m, x_3, x_1, \ldots x_4$

Here in the data (frequency)

x_1 is occuring, say f_1 times

x_2 is occuring, say f_2 times

$\ldots$

x_m is occuring, say f_m times

Therefore we can write data using frequency distribution

Data Points	frequency
x_1	f_1
x_2	f_2
$\ldots$	$\ldots$
x_m	f_m

$$N = \sum_{i=1}^{m} f_i$$

So $\bar{x}$ = Sum of all observations / Total no. of Data Points/observations

N

$x_1 + x_1 + \ldots f_1$ times

$+ x_2 + x_2 + \ldots f_2$ times

$\ldots$

$+ x_m + x_m + \ldots f_m$ times

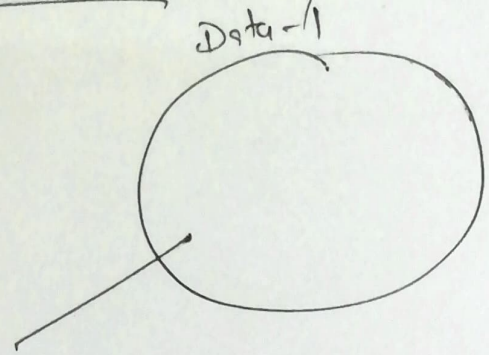
$= f_1 x_1 + f_2 x_2 + \ldots + f_m x_m$

$= \sum_{i=1}^{m} f_i x_i$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{m} f_i x_i$$

where $N = \sum_{i=1}^{m} f_i$

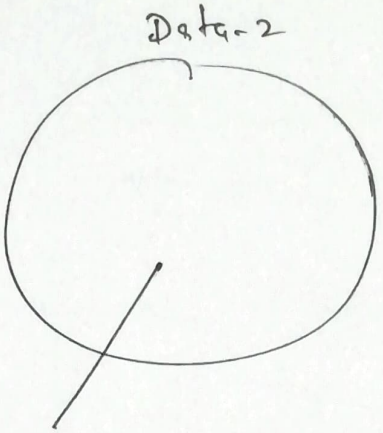
Composite Series / Combined Mean



$x_{11}, x_{12}, \dots x_{1n_1}$

$$\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}$$

$$\sum_{i=1}^{n_1} X_{1i} = n_1 \bar{X}_1$$



$x_{21}, x_{22}, \dots x_{2n_2}$

$$\bar{X}_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}$$

$$\sum_{i=1}^{n_2} X_{2i} = n_2 \bar{X}_2$$

Mean of data points together of data-1 and data-2

$$\bar{X} = \frac{\text{Sum of all data points}}{\text{Total no. of data pts}}$$

$$\bar{X} = \frac{n_1 \bar{X}_1 + n_2 \bar{X}_2}{n_1 + n_2}$$

① Data. Grouped data

Class/Interval	frequency	Take mid values.
under $a_1 - a_2$	f_1	$m_1 = \frac{a_1 + a_2}{2}$
$a_2 - a_3$	f_2	$m_2 = \frac{a_2 + a_3}{2}$
$a_3 - a_4$	f_3	$m_3 =$ — —
— — .		
$- a_{m+1}$	f_m	m_m - — -

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{m} f_i m_i$$

where $N = \sum_{i=1}^{m} f_i$

Weighted Mean

Averaging using different weights

① Data Points

X	Weights
x_1	w_1
x_2	w_2
|	|
x_n	w_n

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

Subject	Credit	Score/100
A	4	70
B	4.5	65
C	5	85
D	6.5	60

CGPA — ?

Median

median value devides the data into two halves, one is having values greater than median value and the other having less than median value.

Sort in increasing / decreasing and then look at the middle value.

n

odd → $x\left[\frac{n+1}{2}\right]$ th observation

even → $\frac{x\left[\frac{n}{2}\right] + x\left[\frac{n}{2}+1\right]}{2}$

Median

In grouped data

Median class $\frac{N}{2}$ (look cumulative frequency)

$$\text{Median value} = l + \frac{\left(\frac{N}{2} - cf\right)}{f} \times h$$

- l — Lower value of median class
- f — frequency of median class
- cf — cumulative frequency preceding the median class.
- h — length of class

Mode : value that occurs more frequently in the given data

- may not exist
  - if exist
    - may or may not be unique
- always from the data points
- least affected by extreme values
- visualized graphically, using Histogram

In grouped data

$$\text{Mode} = L + \frac{(f_1 - f_0)}{2f_1 - f_0 - f_2} * h$$

- L : lower value of modal class
- f_1 : frequency in modal class
- f_0 : frequency preceed the modal class
- h : difference b/w the class
- f_2 : frequency next to modal class

Partitioning the data

arrange data least to greatest value

2 halves	Median	$\frac{N}{2}$
4 equal Parts	Quartiles	$Q_i = i\frac{N}{4}$, $i = 1, 2, 3$ Q_2 = Median
10 equal parts	Deciles	$D_i = i\frac{N}{10}$, $i = 1, 2, \ldots 9$ $D_5 = Q_2$ = Median
100 equal Parts	Percentiles	$P_i = i\frac{N}{100}$, $i = 1, 2, \ldots 99$ $P_{50} = D_5 = Q_2$ = Median