Statistics

Def: statistics is the science of collecting, organizing, interpreting, presenting and analyzing data.

Data: "Facts or pieces of information that can be measured".

Ex:) Heights of students in a classroom.

2) IQ of students.

3) Daily Activities.

4) Weight of people, Age.

Types of statistics :-

① Descriptive statistics

Def: It consists of organizing and summarizing the data.

① Measure of central tendency

{ Mean, Median, Mode }

② Measure of dispersion.

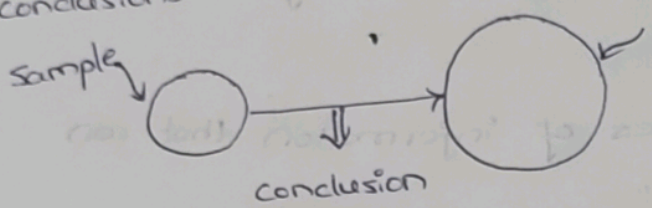
{ Variance, standard deviation }

③ Different types of Distribution of data.

Ex:- Histogram, pdf, pmf, cdf

② Inferential statistics

Def: It consists of data you have measured form conclusions.



⇓

Hypothesis Testing

① z-test
② t-test
③ chi-square Test
④ ANOVA
⑤ F test

} ⇒ conclusion of sample on population.

Ex: Let say there are 20 classes in your college. and you have collected the heights of student in the class.

Heights are recorded as [175 cm, 180 cm, 140 cm, 135 cm, 160 cm]

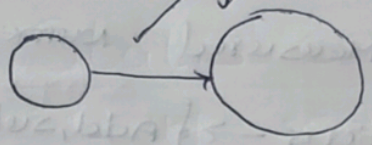
Descriptive:

"What is the average heights of the students in the classroom".

$$\text{Mean} = \frac{175+180+140+135+160}{5}$$

Inferential:

"Are the weight of the students in the classroom (sample.) similar to what you expect in the college! (population.)

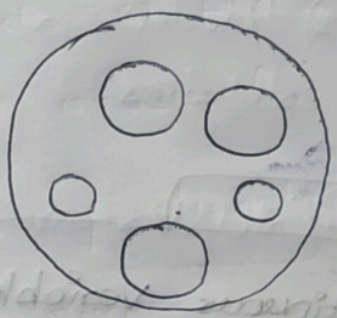


population(N) and sample Data(n)

Exit poll

population :- The entire group you want to study.

Sample :- The portion of the population you select to study.

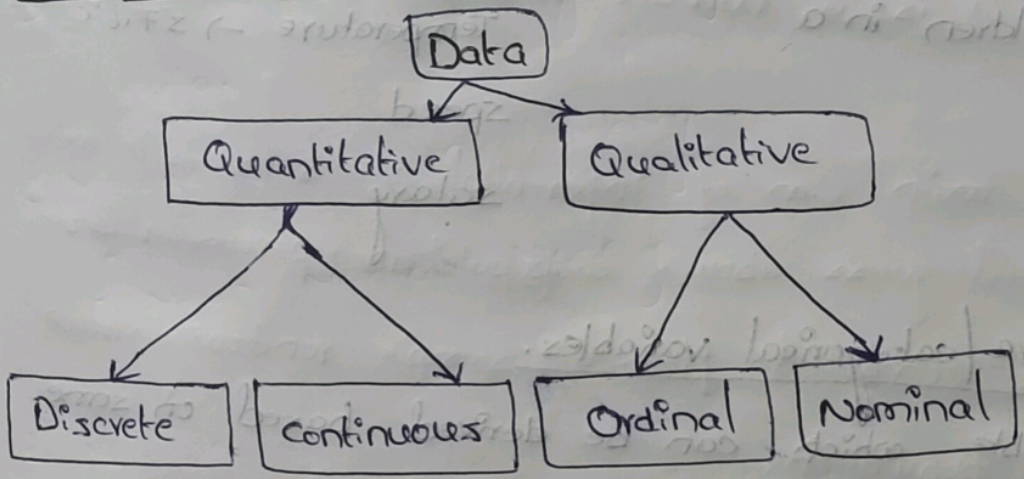


Sample size = 500

population = 10000

Media people will go and ask people in different areas whom they have voted for. Based on this sample of data, they will make conclusions for population, who will going to win in that election.

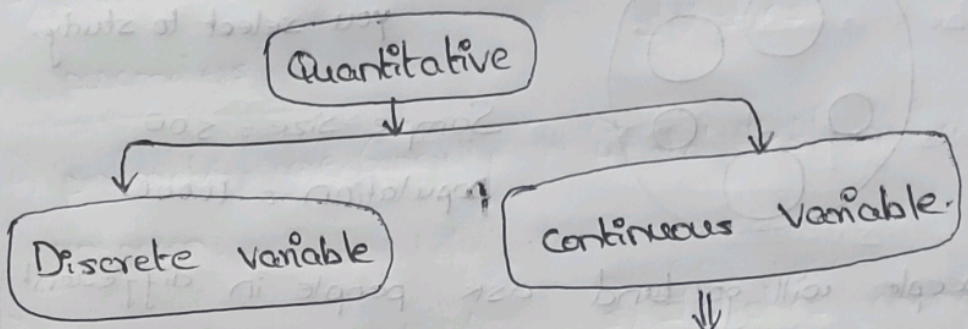
Types of data:



1) Quantitative Variables: Measured numerically

we can perform operations → {Add, substract, Multiply, divide}

Ex: Age
Weight
Height.



⇓

whole numbers with some range

Ex: No. of bank Accounts of people.
2, 3, 4, 5

→ It can't be 2.5, 3.5, 4.5

Ex: No. of children in a family.

⇓

Any value

Ex: Weight → 100 kg, 99.5, 99.75
Height → 172.5, 162.4 cm
Age → 22.5 years
Temperature → 57.6°C
speed
salary

② Qualitative / categorical variables.

→ The variable which can be derived based on some categorical variable.

Ex: Gender → M, F

Ex: Blood Group
A^{+ve}, A^{-ve}

Ex: IQ

0-10	10-50	50-100
⇓	⇓	⇓
low IQ	Medium IQ	Good IQ

Ex: T-shirt size
L, XL, M, S

Qualitative variable

Ordinal

→ Rank and order of data matters, but value doesn't.

Ex:- Ranks

- Good - 3
- Better - 2
- Best - 1

Ex:- students (Marks) Rank

100	1
96	2
57	4
85	3
44	5

Ordinal data

Nominal

→ is nothing but categorical data {Grouped into different classes}

→ Here Nominal means there is no such specific rank.

like Male is greater than femal, one colour is greater than other colour.

Ex:- Gender

M, F

→ Blood Group → A^+, A^-

→ color of hair
Black, Green

→ pincode

population	sample
1) whole Group	1) part of the Group
2) Group we want to know about.	2) Group we do know about.
3) characteristics are called parameters	3) characteristics are called statistics.
4) parameters are Generally unknown.	4) statistics are always known
5) parametres are fixed.	5) statistics change with the sample.

parameter:- It is a number describing a whole population

Ex:- population mean.

statistic:- It is a number describing a sample

Note:- The main aim of inferential statistics is to make an educated guess about a population parameter based on statistic computed from a sample randomly drawn from that population.

Scales of Measurement:

1. Nominal scale data
2. ordinal scale data.
3. Interval scale data.
4. Ratio scale data.

1. Nominal scale data.

1) Qualitative / categorical Data. (This type of data can be measured)

Ex: Gender, colors, Labels

2) Order does not matter.

Ex:- favourite colour

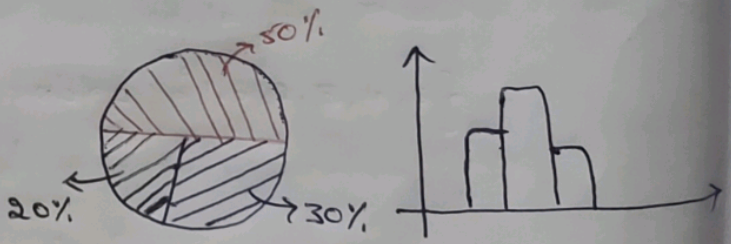
Red → 5 → 50%

Blue → 3 → 30%

Orange → 2 → 20%

10

This data can be represented through bar chart, pie chart.



2. Ordinal scale data:

1. categorical data.
2. Ranking and order matters.
3. Difference cannot be measured.

Ex:-
- Best → 1
- Good → 2
- Bad → 3

Instead focusing on these values we will mostly focus on these values while dealing with ordinal data.

- 1st Rank → 90
- 2nd Rank → 70
- 3rd Rank → 40

③ Interval data scale data :-

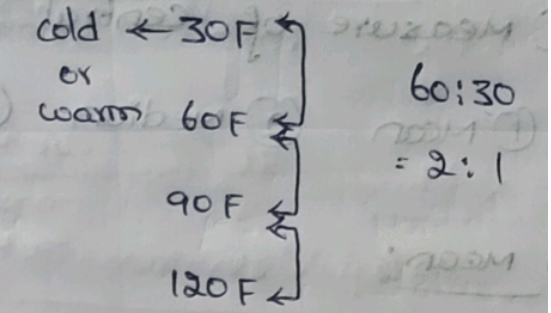
Ex:- Temperature.

① The order matters.

② Difference can be measured.

③ Ratio cannot be measured.

④ NO "0" starting points.



→ suppose room 1 with 30°F
room 2 with 60°F

This means that is this doesn't means that we will feel twice the temperature in room 2 compared to room 1. it will completely depends on external factors. so, Ratio cannot be measured for this type of data.

④ Ratio scale data :-

① The order matters

② Differences are measurable including Ratios.

③ Contain a "0" starting point.

Ex:- Student marks in class.

15

0, 30, 45, 60, 90, 95, 99
s1 s2 s3 s4 s5 s6 s7

Ex:-

① Marital status [Nominal scale data]

② Favourite food based on Gender ? [Nominal]

③ IQ measurements [Ratio scale]

Descriptive statistics

1. Measure of central tendency.
   1. Mean
   2. Median
   3. Mode.

i) Mean :-

population (N) | **sample (n)**

$x = \{1,1,2,2,3,3,4,5,5,6\}$

population Mean $(\mu) = \sum_{i=1}^{n} \frac{x_i}{N}$ | sample mean $(\bar{x}) = \sum_{i=1}^{n} \frac{x_i}{n}$

$$= \frac{1+1+2+2+3+3+4+5+5+6}{10} = 3.2$$

ii) Median :-

$x = \{4,5,2,3,2,1\}$

steps :-

1. sort the random variables $\{1,2,2,3,4,5\}$
2. No. of elements
3. if count == even

   1, 2, [2, 3] 4, 5

   ⇓

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

4. if count == odd

   1, 2, 2, [3], 4, 5, 6

   ⇓

   3 median.

Why median? → Means are affected by outliers.

$X = \{1,2,3,4,5\}$ $\quad X = \{1,2,3,4,5,100\}$ ← outliers

$\bar{X} = \frac{1+2+3+4+5}{5}$ $\quad \bar{X} = \frac{1+2+3+4+5+100}{6}$

$= 3$ $\quad = \frac{115}{6} \approx 19$

Median = 3

Incase of median:-

$X = \{1,2,3,4,5,100\}$

$\text{median} = \frac{3+4}{2} = 3.5$

Note:- Median is used to find the central tendency when outlier is present.

Outliers → completely different numbers compared to other distribution [due to these outliers, there is more difference or movement in the central distribution of data]

iii) **Mode** :- Maximum frequency occurring element.

$\{2,1,1,1,4,5,7,8,9,9,10\}$

Mode = 1

→ Mean, Median, Mode concepts used for EDA and Feature Engineering.

EDA and Feature Engineering

Mean, Median. | Mode → categorical feat.

Age	Weight	Salary	Gender	Degree
24	70	40k	M	B.E
25	80	70k	F	[-] B.E
27	95	45k	F	[-] B.E
24	[-]	50k	M	PHD
32	[-]	60k	[-] M	B.E
[-]	60	-	[-] M	Master
[-]	65	55k	[-] M	BSC
40	77	-	M.	B.E

Outliers

Median or Mode

- → Missing values
- → In case of Missing values if we drop in a particular row, if we drop that row means there will be a loss of data.
- → Let's say I want to handle the missing values in age, weight, salary.
  For that → we need to check whether there are any outliers or not.
  - → If there are outliers means we will replace missing values with median.
  - → If there are no outliers means we will go with Mean.
- → But wrt categorical features in order to handle the missing values we can specifically use the mode.

2) Measure of Dispersion?

i) variance (σ^2)

ii) standard deviation (σ)

i) Variance : It measures the spread of the data with an emphasis on extreme low and high value

Population variance

$$\sigma^2 = \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N}$$

x_i = Data points

μ = population mean

N = population size

Sample variance

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

$x_i \Rightarrow$ data points

$\bar{x} \Rightarrow$ sample mean

$n \Rightarrow$ sample size

Q) Why we devide sample variance by n-1?

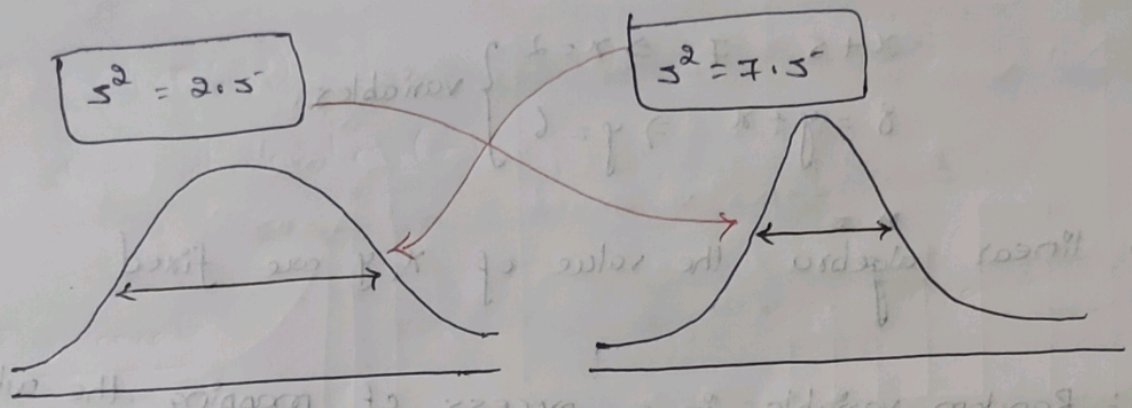
Ex:- $\{1, 2, 3, 4, 5\}$

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$$

x_i	$\bar{x}$	$(x_i - \bar{x})^2$
1	3	4
2	3	1
3	3	0
4	3	1
5	3	2
$\bar{x} = 3$		$\sum (x_i - \bar{x})^2 = 10$

$$s^2 = \frac{10}{4}$$

$x = \{\quad\}$ $y = \{\quad\}$



ii) Standard deviation :- It measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.

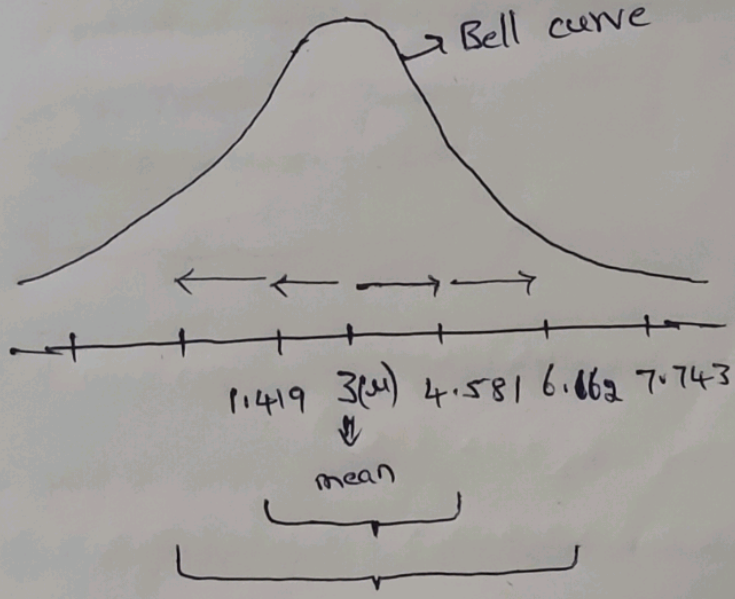
population standard deviation (σ) = $\sqrt{\text{variance}} = \sqrt{\sigma^2}$

Sample standard deviation (s) = $\sqrt{s^2}$

$X = \{1, 2, 3, 4, 5\}$

$\bar{x} = 3$

$\sigma = 1.581$



Random Variable :-

$$x + 5 = 7 \Rightarrow x = 2$$
$$8 = y + x \Rightarrow y = 6$$
} variables.

In linear algebra the value of x, y are fixed.

Def: Random variable is a process of mapping the outputs of a random process or experiment to a number.

Ex: Tossing a coin $\{Head, Tail\} \Rightarrow$ process

$$x = \begin{cases} 0 & \text{if Head} \\ 1 & \text{if Tail} \end{cases}$$

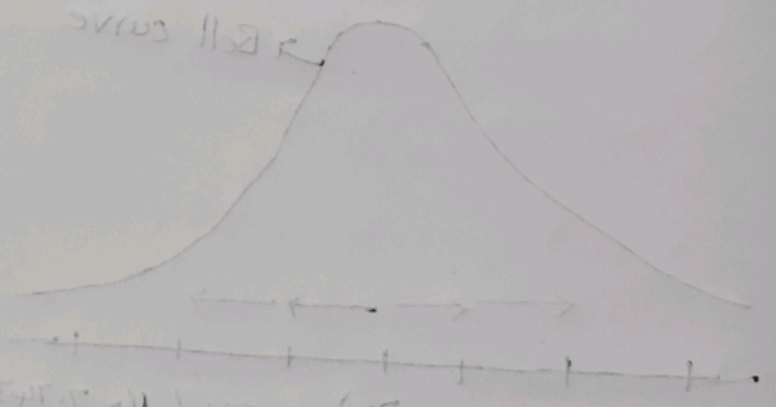
Ex: Rolling a dice $\{1, 2, 3, 4, 5, 6\}$

$$y = \{\text{sum of rolling a dice 7 times}\}$$

→ Random process will helps us to solve few questions like → $pr(y \geq 15)$?

→ $pr(y < 10)$?

→ $pr(10 \leq y \leq 15)$?

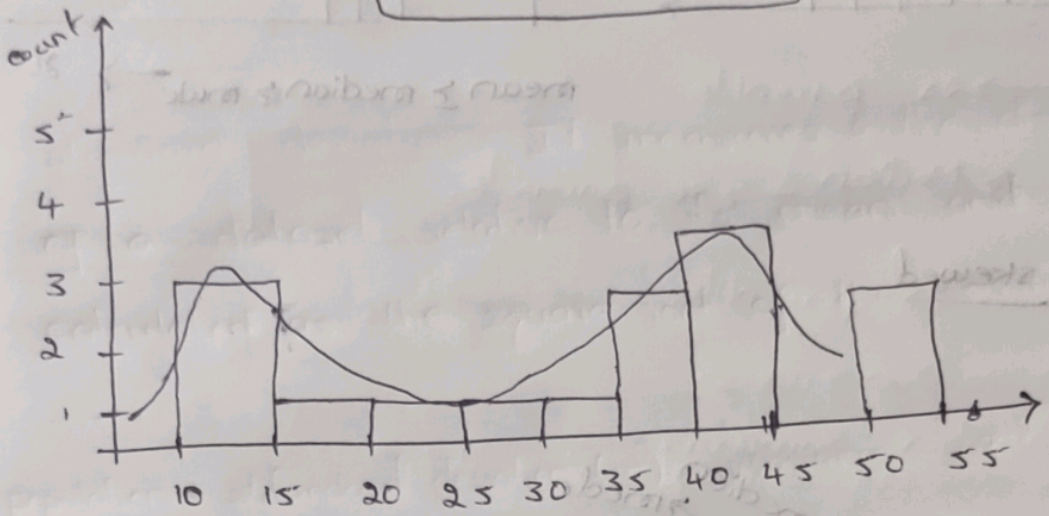


Histogram

Ages = {10, 12, 14, 18, 24, 26, 30, 35, 36, 37, 40, 41, 42, 43, 50, 51}

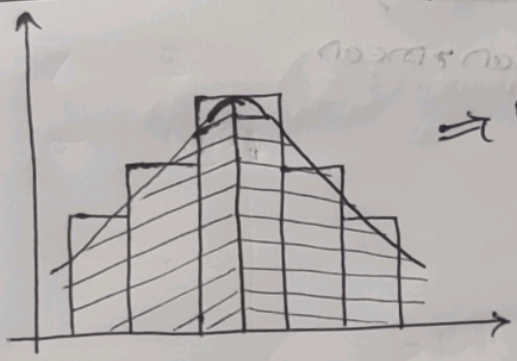
$\frac{50}{10} = 5$ ← bin size no. of bins = 10

$$\text{bin size} = \frac{\text{range}}{\text{no. of bins}}$$



Skewness :

(i) NO skewness



⇒ Normal distribution

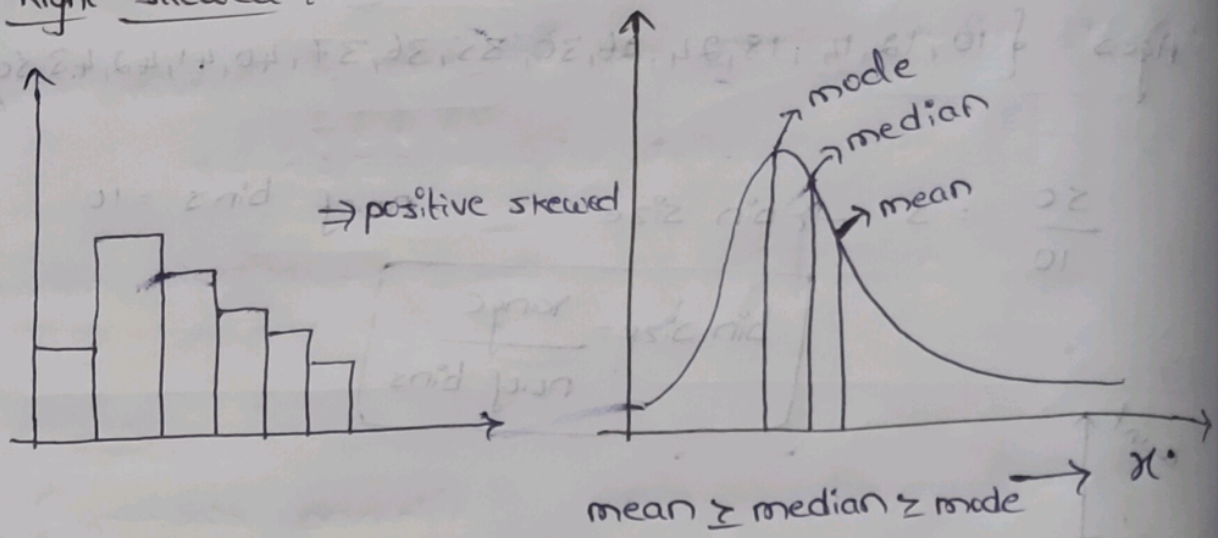
⇒ symmetrical distributio

⇓

NO skewness

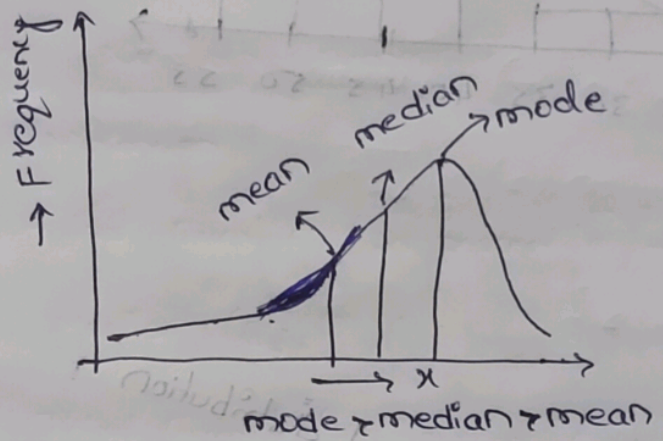
Median = Mean = Mode

② Right skewed :



Ex: Length of youtube comments
Income data set (most people having average incomes, smaller no. of people having higher incomes.

③ Left skewed



Ex: ~~Left~~ exam scores of a group of college students who took a relatively simple exam.

(Most of the students have high test scores, and a smaller number of people would have low scores that skew the curve toward the left of the group)