1/7/2022

: DATASCIENCE-FOR-BEGINNERS :

Population

collection of all items of interest

→ N

→ The numbers we've obtained when using population is called parameters.

→ Hard to observe and contact.

Sample.

A subset of the population.

→ n

→ The numbers we've obtained using samples is called statistics.

Adv :-

1) Less time consuming
2) Less costly.

SAMPLE :-

Two defining characteristics

1) Randomness → A random sample is collected when each member of the sample is chosen from popula strictly by chance.

2) Representativeness

↳ A representative sample is a subset of the population that accurately reflects the members of the population entire.

Types of Data:-

- categorical
  - categories or groups
  - ex:- car brands
  - yes/No questions.
- Numerical.
  - represents number
    - discrete
      - finite
      - eg:- no. of children.
      - S.A.T score.
      - ↳ we can imagine each member of data set
      - <integer values>
    - continuous
      - infinite
      - eg:- 68.0434kg weight.
      - ⇒ changes by smaller amounts.

Measurement Levels:-

- Qualitative
  - Nominal
    - eg:- car brands/ seasons
    - → they aren't numbers so can't
  - ordinal.
    - Eg:- rate your lunch
    - → has proper or strict order
- Quantitative
  - interval
    - → doesn't have 0
    - eg:- temperature
  - ratio.
    - → has a 0

Visualizing techniques for categorical variables :-

1) Frequency distribution tables
2) Barcharts
3) Piecharts
4) Pareto Diagrams.
   ↳ categories are shown in deccending order of frequency.
   → Shows the line graph of cumulative frequency.

	Frequency	relative frequency	cummulative frequency.
Audi	124	37%	37%
Mercedes	113	33%	71%
BMW	98	29%	100%
	335	99	

$$\text{Relative frequency} = \frac{124}{335} \times 100\%$$

$$= 37$$

$$= \frac{113}{335} \times 100 = 33\%$$

$$= \frac{98}{335} \times 100 = 29\%$$

$$\text{cummulative frequency} = \frac{\text{subgroup relative frequency}}{\text{Total of relative frequency}}$$

↳ Sum of relative frequency.

Visualizing techniques for numerical data

① Frequency Distribution Data with intervals

Desired intervals = $\frac{\text{largest number} - \text{smallest number}}{\text{number of desired intervals}}$.

eg:- $\frac{100-1}{5} = 19.8 \approx 20$.

1-21, 21-41, 41-61, 61-81, 81-101.

A number is included in an interval if the number:

1) is greater than lower bound
2) is lower or equal to the upper bound.

interval	frequency	relative frequency
1-21	2	0.10
21-41	4	0.20
41-61	3	0.15
61-81	6	0.30
81-101	5	0.25
	20	

② Histogram charts :-

⇒ Histogram with unequal intervals.

Graphs to represent relationship between two variables :-

1. Cross tables

2. Scatter plots
   ↳ used when we are representing two numerical

Meaures of Central tendency:-

1/19/2022

Mean:-

Also known as simple average.

denoted by $\mu \rightarrow$ population

$\bar{x} \rightarrow$ sample.

1. Trimmed mean
2. weighted mean
3. Harmonic mean
4. Geometic mean

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

} By adding up all the components and then dividing by the number of components.

$$(\text{or}) \quad \frac{x_1 + x_2 + x_3 + x_4 + \cdots + x_n}{N}$$

Disadvantages

⇒ easily affected by outliers.

⇒ Mean is not enough to make definite conclusions

Median :-

Median is middle number in ordered dataset.

⇒ Order data in ascending order

⇒

Median is the number at position $(n+1)/2$ in the ordered list

Mode :- → it is mostly useful for categorical variable.

The mode is the value that occurs most often.

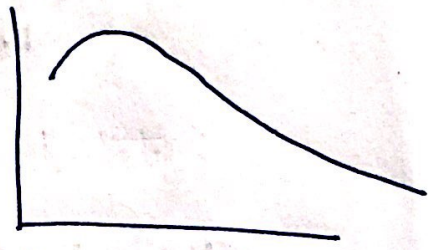
Skewness :—

Skewness indicates whether the data is concentrated on one side.

① Right skewness → outliers are on the rightside

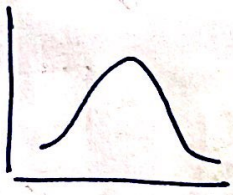
↳ mean > median

↳ positive skew.



② Zero skew.

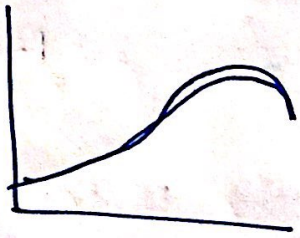
Mean = Median = Mode



③ Left skewness → outliers are to the left.

↳ mean < median

↳ Negative skewness



⇒ Skewness tells us where the data is situated.

⇒