12/06/24

Week - 2

L.2.1 :- Describing Categorical Data :- (Frequency Distribution)

*. Describing Categorical Data :- Single Variable :-

- **Frequency Distribution :-**

A Frequency distribution of qualitative data is a listing of the distinct values and their frequency / (count)

Each Row of a frequency table lists a category along with the number of cases in this category.
Count #

Ex :- (i). A, A, B, C, A, D, A, B, D, C

(ii). A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D.

*. Construct a Frequency Distribution :-

The Steps to Construct a frequency distribution.

Step1: List the distinct values of the observations in the data set in the first column of a table.

Step2: For each Observation, place a tally mark in the second column of the table in the Row of the appropriate distinct value.

Step3: Count the tallies for each distinct value and record the totals in the third column of the table.

(i)

Category	Tally mark	Frequency
A	IIII	4
B	II	2
C	II	2
D	II	2
Total		10

(ii).

Category	Tally mark	Frequency
A	HHT I	6
B	III	3
C	IIII	4
D	HHT	5
Total		18

* Frequency table in a googlesheet :-

Step1 Select / Highlight the cells having data you want to visualize.

Step2 In the formatting bar click on the Data option.

Step3 In the data Option go to Pivot Table option and create a new sheet.

New → Insert

Step4 After creating Pivot Table go in Pivot Table Editor and in that first add rows and then values.

Insert

Pivot

↓

Pivot table

↓

Copy Paste

↓

Name Chart → Formatting Chart Done.

* Frequency table gives the count of each variable, each Categorical Variable.

* Relative Frequency :- The Ratio of the frequency to the total Number of Observation is called relative frequency.

* The steps to construct a relative frequency distribution.

Step1. Obtain a frequency distribution of the data.

Step2. Divide each frequency by the total Number of Observations.

(I).

Category	Tally Mark	Fre	R.Fre
A	IIII	4	0.4
B	II	2	0.2
C	II	2	0.2
D	II	2	0.2
Total		10	1

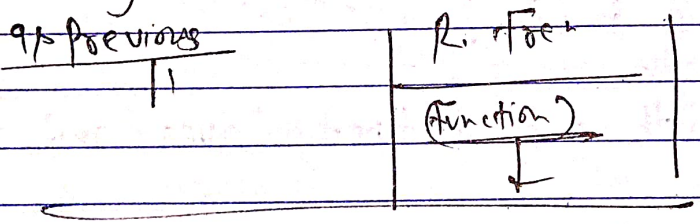
(II).

Category	T.M	Fre	R.Frequ
A	5IIII	6	0.4
B	3 III	3	0.2
C	3 III	3	0.2
D	3 III	3	0.2
Total	~~HH~~	15	1

*. Important of Relative frequency (Why?)

- For Comparing two data sets.
- Because relative frequencies always fall b/w 0 and 1, they provide a standard for Comparison.

* In Google Sheet :-



* Summary :-

1. Constructing a frequency table.
2. Notion of Relative frequency and Constructing a Relative frequency table.

13/06/24

1.2.2 - Describing Categorical Data - (Charts of Categorical Data)

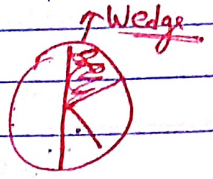
*. Charts for Categorical Data :-

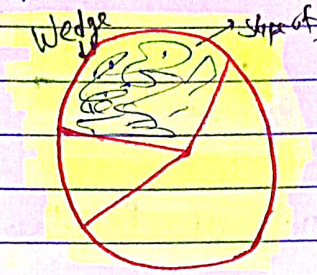
- The two most common displays of a Categorical Variable are a Bar chart and a Pie chart.
- Both describe a Categorical Variable by displaying its frequency table.

*. Pie charts :-

A pie chart is a circle divided into pieces proportional to the relative frequencies of the qualitative data.

WEDGES





* Construction of a Pie-Chart :-

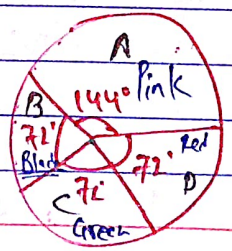
Step1 Obtain a relative-frequency distribution of the data.

Step2 Divide a circle into pieces proportional to the relative frequencies. (R.Frequency X 360°)

Step3 Label the slices with the distinct values and their relative frequencies.

Ex1. A, A, B, C, A, D, A, B, D, C.

Category	Tally Mark	Fre	R. Fre	Degree
A	IIII	4	0.4	144°
B	II	2	0.2	72
C	II	2	0.2	72
D	II	2	0.2	72
Total		10	1	360°



* Pie-chart

*. Pie - Chart in a google sheet :-

Step1 Select / Highlight the cells having data you want to visualize.
Step2 Click the Insert chart option in Google sheets toolbar.
Step3 Change the visualization type in Chart editor
Step4 Select in Chart editor Chart type to Pie chart.

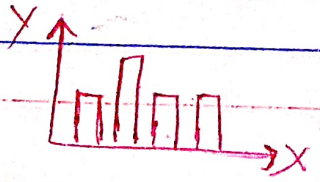
• Summary :-

① A Pie chart is used to show the proportions of a Categorical Variable.

② A pie chart is a good way to show that one category makes up more than Half of the total.

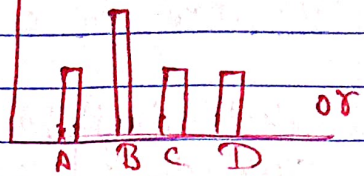
*. Bar Chart :-

A Bar chart Displays the distinct Values of the Qualitative Data on a horizontal axis and the relative frequencies (or frequencies or Percentages) of those Values on a vertical axis. The frequency / relative frequency of each distinct Value is represented by a Vertical Bar whose height is equal to the frenquency / relative frequency of that Value.
← the bars should be positioned so that they do not touch each other.



frequency — Vertical Bar
Value — horizontal axis

(a)



or

* Construction of a Bar chart :-

To Construct a Bar Chart.

Step1: Obtain a Frequency / relative frequency distribution of the Data.

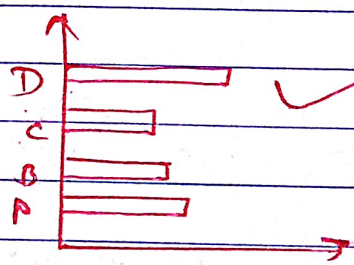
Step2: Draw a Horizontal axis on which to place the bars and a vertical axis on which to display the frequencies / relative frequencies.

Step3: For each distinct Value, Construct a Vertical Bar whose height equal the frequency / relative frequency of that Value.

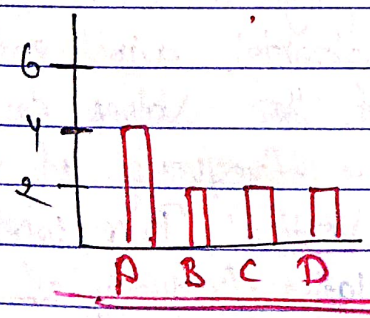
Step4: Label the bars with the distinct Value, the Horizontal axis with the name of the Variable, and the Vertical axis with "Frequency / relative frequency"

Both (a) and (b) Types are possible.

(b)



from Ex(1) —



* for Ordinal → Maintain Order in Pareto chart
* for Nominal → Name, Value.

* Bar chart in Google sheet :-

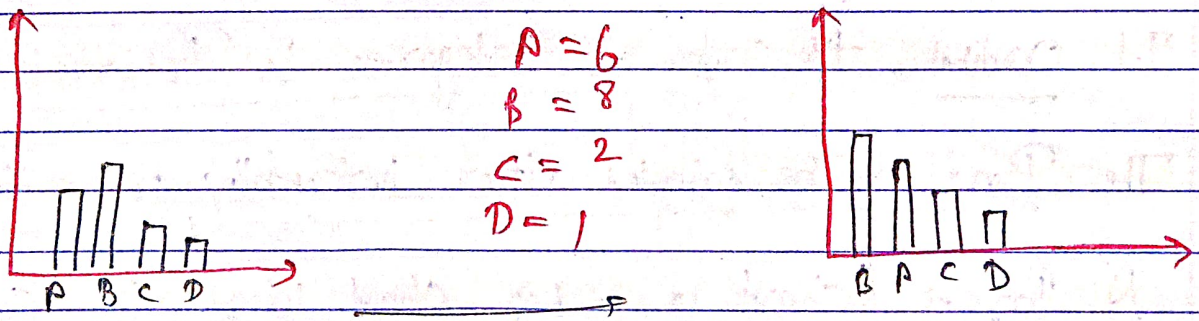
Step1 Select / highlight the cells having Data you want to visualize.
Step2 Click the insert Chart option in Google Sheets tool bar.
Step3 Change the visualization type in Chart editor
Step4 Select in Chart editor Chart type to Bar chart.

* Pareto Charts :-

When the Categories in a bar chart are sorted by frequency, the bar chart is sometimes called a Pareto Chart.

Pareto charts are popular in quality Control to identify problems in a business process. → (Nominal → Name, Value) (Ordinal → order)

- If the categorical variable is ordinal, then the bar Chart must preserve the Ordering.



Bar chart

Pareto Chart

* Ascending Order or Desending Order.

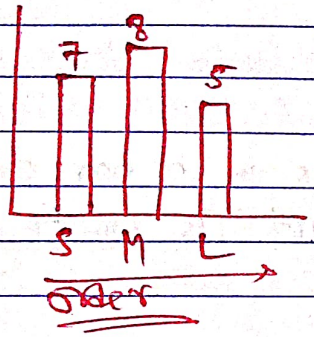
→ vertical for Category

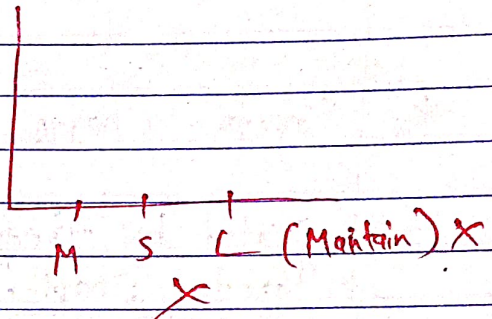
for | | | Category — Horizontal

Ex. Ordinal Data :

Size of T-Shirt - L, M, M, S, L, S, S, M, L, M, M, S, S, L, M, S, M, S, L, M.

Size	Tally Marks	Freq	Relative fre
Small	𝍸 II	7	0.35
Medium	𝍸 III	8	0.40
Large	𝍸	5	0.25
Total		20	1





If you have order then maintain the Order.

* Summary

1. A Bar Chart is used to show the frequencies/relative frequencies of a Categorical Variable.
2. If Ordinal, the Order of Categories is Preserved.
3. The Bars can be oriented either horizontally or vertically.
4. A Pareto chart is a bar chart where the Categories are sorted by frequency.

13/06/24

L.2.3. Describing Categorical Data (Best Practices while graphing Data-1)

Note. Know your purpose,

- Have a purpose for every table or graph you create.
- Choose the table/ Graph to serve the purpose

*. Pie charts are best to use when you are trying to compare Parts of a whole.

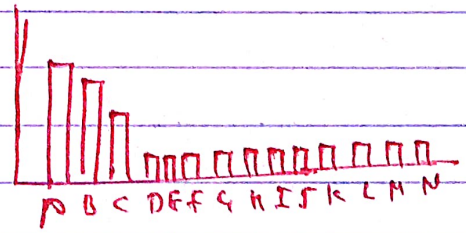
*. Bar graphs are used to compare things b/w different group.

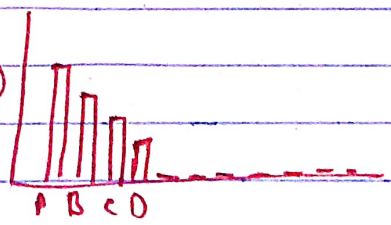
*. Label Your Data :- (Tutorial from sheets)

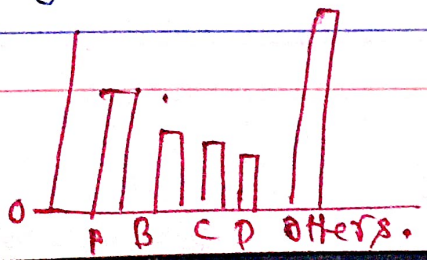
- Lable your chart to show the Categories and Indicate whether some have been Combined or omitted.
- Name the bars in a bar chart.
- Name the slices in a Pie chart.
- If You have omitted some of the cases, make sure the label of the plot defines the Collection that is summarized.

*. Many Categories.

A Bar chart or pie chart with too many categories might Conceal the more important Categories. In some case, grouping other Categories together might be done.







13/06/2024

L.2.4. D. C. Data

(Best Practices while Graphing Data-2)

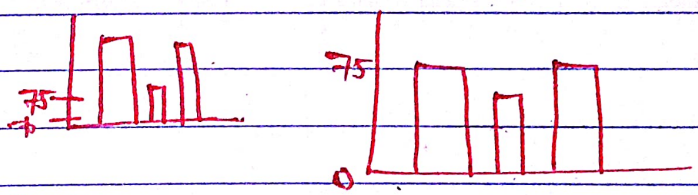
*. The area Principle :-

- Displays of Data must Obey a fundamental rule called the area principle.
- The area principle says that the area occupied by a part of the graph should correspond to the amount of data it represent.
- Violations of the area principle are a common way to mislead with statistics.

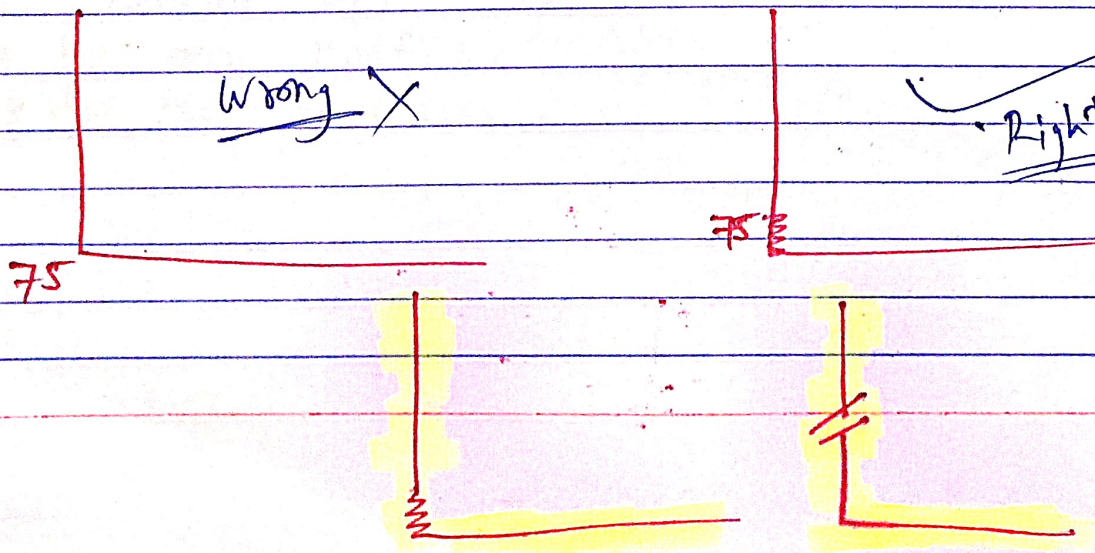
*. Misleading Graphs : ① Violating Area Principle.

- Decorated Graphics : Charts Decorated to Attract attention often Volate the area area Principal.

- ② Truncated Graph : → Because of loss of Data
- Another common violation is when the Baseline of a bar chart is not at zero.



*. ③ Indicating a y-axis break :-



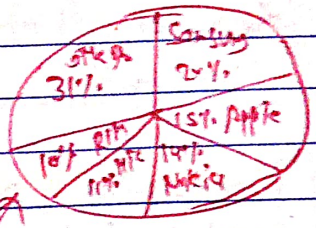
1.2.5. Describing Categorical Data: (Mode and Median).

14/06/2021

* Round-off errors :-

- Important to check for round-off errors.

		Before	
Ex →	A	22.5	23
	B	35.5	36
	C	12.5	13
	D	11	11
	E	18.50	19
		100	102



- When table entries are percentages or proportions, the total may sum to a value slightly different from 100% or 1. This might result in a pie chart where the total does not add up.

* Sectional Summary :-

1. Know your purpose and choose table/graph appropriately.
2. Lable Your Charts.
3. Handle multiple Categories appropriately.
4. Respect area principle.

4.1 Avoid Overly decorated Graphs.

4.2 Avoid truncated graphs - Use special symbols to indicate vertical axis has been modified.

4.3 Check for rand-off errors.

only counting ✓
→ Nominal Ordinal → order ✓

Fre → Bar chart → Barchart CAT

Pie chart

* . Summarizing Categorical Data :- Table | CAT | Fre | R.Fre

- Graphical Summaries of Categorical Data : Bar Chart and Pie Chart.
- Need for a Compact measure.
- Numbers that are used to describe data sets are called Descriptive measures.
- Descriptive measures that indicate where the Centre or most typical value of a data set lies are called measures of central tendency.

① . Mode :-

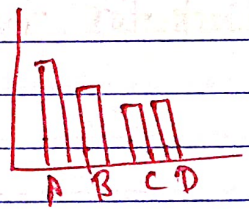
The Mode of a Categorical variable is the most Common category, the category with the highest frequency.

- The Mode labels :-
- The longest bar in a bar Chart.
- The widest slice in a pie chart.
- In a Pareto chart, the mode is the first category shown.

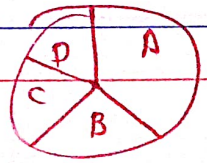
Ex ① . A, B, C, D, A, A, B, A, C, D, A

Sol → Mode = A (highest frequency).

①



②



* We can't define median for a Nominal Categorical Data. so for median we have ordinal Data we have median.
* Median is value that divides date set in 2 halfs.

* Bimodal and Multimodal Data :-

- If Two or more categories tie for the highest frequency, the data are said to be bimodal (in case of two) or multimodal (more than two).

  * Bimodal —— 2 Modes
  * Multimodal —— More than 2 Modes.

* Median :-

Ordinal data offers another summary, the median, that is not available unless the data can be put into order.

- Definition :-

The median of an ordinal variable is the category of the middle observation of the sorted values.

- If there are an even number of observations, choose the category on either side of the middle of the sorted list as the median.
- If it odd then exactly that value.

Ex (1): A, A, B, A, C, D, A, B, B, D. (Ex Even)

- Median = B

(2) A, A, B, C, D, D.

- either B or C. (median) (3rd Observation or 4th Observation)

(3) A, A, B, C, D — Odd.

- Median = B — 3rd Observation.

- Mode — Nominal + Ordinal
- Median — only for Ordinal Data

*. Sectional Summary :-

*(Nominal ; ordinal)

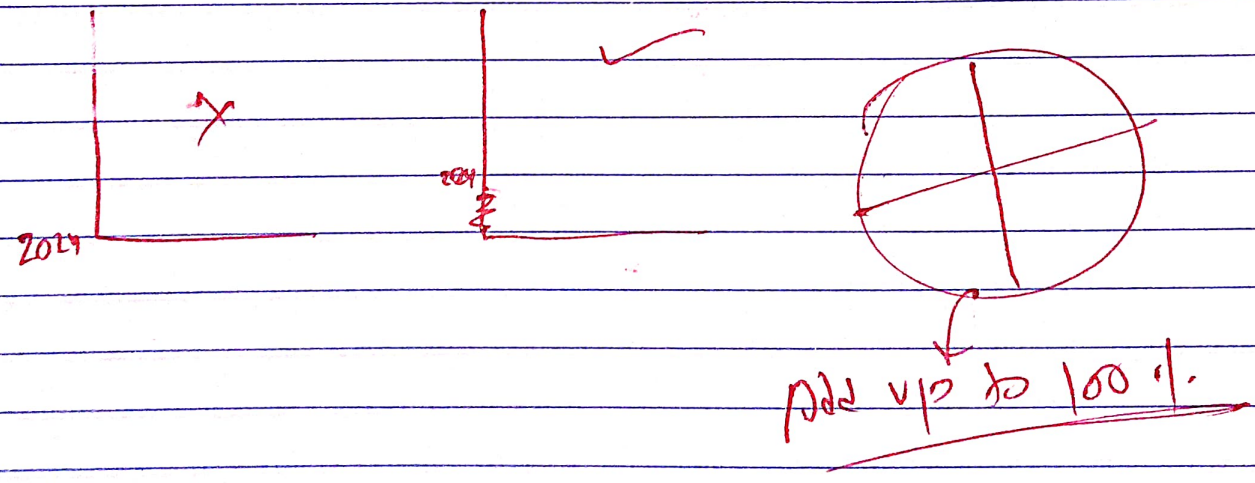
- The mode of a Categorical variable is the most common Category.
- The median of an ordinal variable is the Category of the middle observation of the sorted value. (~~Nominal~~ ordinal)

* Categorical Data ~~#-#~~

Week-2 Tutorial-1 (Problem Charts and table).

Q. Solving :-

Tutorial-2 - Problems Misleading Graph :-



Tutorial – 3 (SUMIF in Google Sheets)

* Syntax: (SUMIF (range, criterion, [sum_range])

- Range: The Range which is tested against criterion.
- Criterion: The Pattern or test to apply to range.
- Sum_range – The range to be summed, if different from Range.

Item	Category	Qty	Price	Cost
Earphones	Electronics	1	210	210
Phone cover	Accessories	1	140	140
Dongle	Electronics	1	790	790
A4 sheets	Stationary	200	1	200
Ball Pens	Stationary	2	12	24
				1364

= Sumif(B3:B7, "Stationary", E3:E7)

Tutorial - 4 - (VLOOKUP in Google sheet)

- **Vlookup** :- Vertical lookup. Searches down the First column of a range for a key and returns the value of a specified cell in the row found.

- Syntax. Vlookup (search_key, range, index, [is_sorted])
- Search_key :- The value to search for.
- Range :- The Range to consider for the search.
- index : The Column index of the value to be returned, where the 1st Column in range is numbered 1.
- is_sorted : (True by default) Indicates whether the Column to be searched (the first column of the specified range) is sorted. False is recommended in most cases.

= Vlookup("Place", A1 : A69, 9, False)