October 11, 2020

WEEK 1

What is statistics?

→ Statistics is the art of learning from data. It is concerned with the collection of data, their subsequent description, and their analysis, which often leads to drawing of conclusions.

Descriptive statistics

- data collection
- organising data
- describing data

LEARNING OBJECTIVES

1. What is statistics?
   → descriptive statistics, inferential statistics
   → Distinguish b/w a sample and a population
2. Understand how data are collected
   → Identify variables and cases (observations) in a data set.
3. Types of data
   → classify data as categorical (qualitative) or numerical (quantitative) data.
   → Understand cross-sectional versus time-series data
   → Measurement scales.
4. Creating data sets; Downloading & manipulating data sets; working on subsets of data.
5. Framing Questions that can be answered from data.

Date
October 12, 2020

↳ Analysis → better

MAJOR BRANCH OF STATISTICS

1. DESCRIPTION - describing data
The part of statistics concerned with the description and summarization of data is called Descriptive Statistics.

2. INFERENCE
The part of statistics concerned with the drawing of conclusions from data is called Inferential Statistics.

» To be able to draw a conclusion from the data, we must take into account the possibility of chance - intro to probability.

POPULATION & SAMPLE

→ Population : The total collection of all the elements that we are interested in is called a population

→ Sample : A subgroup of the population that will be studied in detail is called a sample.

PURPOSE OF STATISTICAL ANALYSIS

* If the purpose of the analysis is to examine & explore information for its own intrinsic interest only, the study is descriptive.

* If the info. is obtained from a sample of a population and the purpose of the study is to use that info to draw conclusions about the population, the study is inferential.

* A descriptive study may be performed either on a sample or on a population.

* When an inference is made about the population, based on information obtained from the sample, does the study become inferential.

SUMMARY

Statistics - (descriptive and inferential) main branches.

Population & Sample

[Major Branch of Statistics]

Descriptive → (can be done on either sample or on population.)

Inferential → (can be done only on sample)

Understanding Data

October 10 2020

WHAT IS DATA ?

In order to learn something, we need to collect data.

* Data are the facts and figures collected, analysed and summarised for presentation and interpretation.

→ Statistics relies on data, information that is around us

WHY DO WE COLLECT DATA ?

* Interested in the characteristics of some group or groups of people, places, things or events.
* EXAMPLE : to know about temperatures in a particular month in Chennai, India.
* EXAMPLE : To know about the marks obtained by students in their class 12.
* to know how many people like a new song / product / video collected through comments.

DATA COLLECTION

* Data available : published data
* Data not available : need to collect, generate data

We assume data is available and our objective is to do a statistical analysis of available data. //

* Analyse → to look at or think about the different parts or details of something carefully in order to understand or explain it !
* Interpret → to explain or understand meaning of something

UNSTRUCTURED & STRUCTURED DATA

* For the information in a database to be useful, we must know the context of the numbers and text it holds.
* When they are scattered about with no structure, the information is of very little use.
* Hence, we need to organize data.

DATASET

* A structured collection of data → is Dataset !
* It is a collection of values - could be numbers, names, roll numbers.

VARIABLES & CASES

* Case (observation) : A unit from which data are collected.
  ↳ Rows represent cases
* Variable : → columns represents variables.
  - intuitive → A variable is that 'varies'
  - formally → A characteristic or attribute that varies across all units
* In our school data set :
  - Case : each student
  - Variable : names, marks obt., Board etc.

* Rows represent cases : for each case, same attribute is recorded.

* Columns represent variables : for each variables, same type of value for each case is recorded.

SUMMARY

We have organised data in a spreadsheet into a table.

→ Each variable must have its own column.
→ Each observation must have its own row.

Date October 12, 2020

CLASSIFICATION OF DATA

CATEGORICAL & NUMERICAL

Variable

DATA

- Categorical (Qualitative)
- Numerical (Quantitative)
  - Discrete (countable)
  - Continuous (measurable)

* Categorical data
  → also called qualitative variables
  → Identify group membership { gender, Board, Blood Grp, Jersey no., mobile no. }

* Numerical data
  → Also called quantitative variables
  → Describe numerical properties of cases
  → Have measurement units { marks, height, weight, temperature, score }

* Measurement Units : scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in cms etc.

  · The data that makeup a numerical ~~value~~ variable in a data table must share a common unit.

CROSS-SECTIONAL & TIME SERIES DATA

* Time series - data recorded over time (on a particular variable, example, quantity)

* Timeplot - graph of a time series showing values in chronological order

* Cross sectional - data observed at the same time

SUMMARY

- classify data as categorical or numerical
- For numerical data, find out unit of measurement
- Check whether data is collected at a point of time (cross-sectional data) or over time (time-series data)

Date October 12, 2020

SCALES OF MEASUREMENT

* Data collection requires one of the following scales of measurement : nominal, ordinal, interval or ratio.
  (nominal, ordinal — CATEGORICAL; interval, ratio — ORDINAL)

NOMINAL SCALE OF MEASUREMENT

→ When the data for a variable consists of labels or names used to identify the characteristics of an observation, the scale of measurement is considered a nominal scale.

→ Examples : Name, Board, Gender, Blood Group etc.

→ Sometimes nominal variables might be numerically coded.
  → For example, we might code Men as 1 & Women as 2 or Men as 3 & Women as 5. Both codes are valid.

→ There is no ordering in the variable.

→ NOMINAL : name categories without implying order.

ORDINAL SCALE OF MEASUREMENT

→ Data exhibits properties of nominal data and the order or rank of data is meaningful, the scale of measurement is considered an Ordinal scale.

→ Each customer who visits a restaurant provides a service rating of excellent, good or poor.

- The data obtained are the labels - excellent, good or poor - the data have the properties of nominal data.

- In addition the data can be ranked, or ordered, with respect to service quality.

→ ORDINAL : name categories that can be ordered

INTERVAL SCALE OF MEASUREMENT

→ If the data have all the properties of ordinal data and the interval between values is expressed in terms of a fixed unit of measure, then the scale of measurement is interval scale.

→ Interval data are always numeric. Can find out difference between any 2 values.

→ Ratios of values have no meaning here because the value of zero is arbitrary.

→ INTERVAL : numerical values that can be added/subtracted (no absolute zero)

Example : TEMPERATURE

→ Suppose the response to a question on how hot the day is comfortable & uncomfortable, then the temperature is a variable is NOMINAL.

→ Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is ORDINAL.

→ Example : Consider an AC room where temperature is set at 20°C and the temperature outside the room is 40°C. It is correct to say that the difference in temp is 20°C, but it is incorrect to say that the outdoors is twice as hot as indoors.

→ Temperature in degrees Fahrenheit or degrees centigrade is an interval variable. No absolute zero.

	Celsius	Fahrenheit
Freezing pt.	0	32
Boiling pt.	100	212

but in Kelvin scale there is absolute 0.

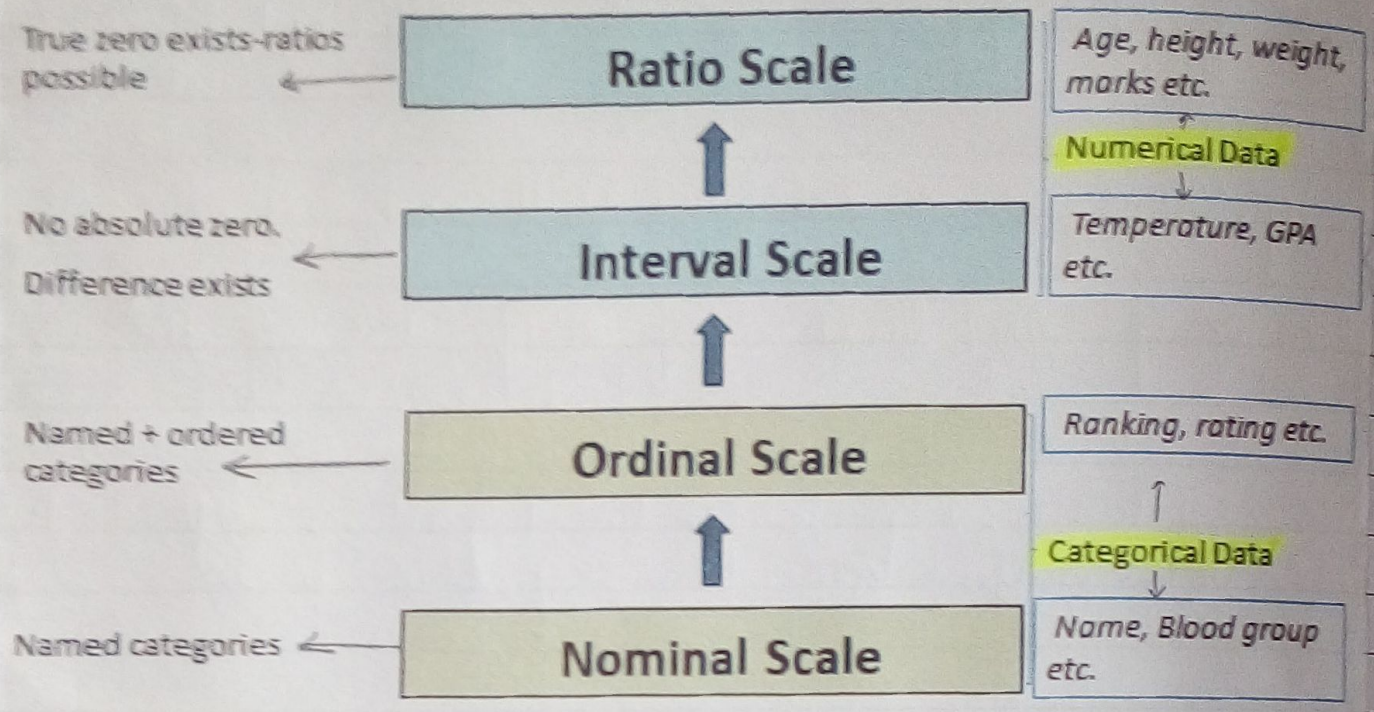
RATIO SCALE OF MEASUREMENT

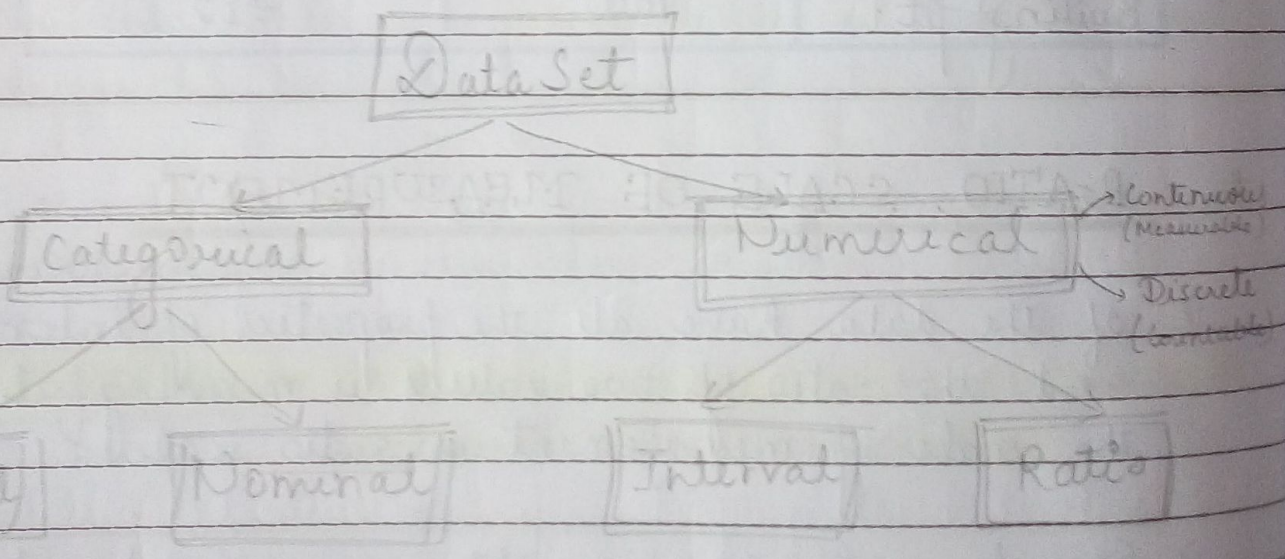
→ If the data have all the properties of interval data and the ratio of two values is meaningful, then the scale of measurement is ratio scale.

→ Example : height, weight, age, marks etc.

→ RATIO : numerical values that can be added, subtracted, multiplied or divided (make ratio comparisons possible)

SUMMARY


True zero exists-ratios possible
Ratio Scale
Age, height, weight, marks etc.
Numerical Data
No absolute zero.
Difference exists
Interval Scale
Temperature, GPA etc.
Named + ordered categories
Ordinal Scale
Ranking, rating etc.
Categorical Data
Named categories
Nominal Scale
Name, Blood group etc.


Data Set
Categorical
Numerical
Continuous
Discrete
Ordinal
Nominal
Interval
Ratio

Date
October 15, 2020

classmate
Date
Page

WEEK 2

Describing Categorical Data

FREQUENCY DISTRIBUTIONS

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

Each row of a frequency table lists a category along with the number of cases in this category. COUNT

CONSTRUCT A FREQUENCY DISTRIBUTION

The steps to construct a frequency distribution:

STEP 1: List the distinct values of the observations in the data set in the first column of a table.

STEP 2: For each observation, place a tally mark in the second column of the table in the row of the appropriate distinct value.

STEP 3: Count the tallies for each distinct value and record the totals in the third column of the table.

FREQUENCY TABLE IN A GOOGLESHEET

STEP 1: Select / Highlight the cells having data you want to visualize.

STEP 2: In the Formatting bar click on the Data option.

STEP 3: In the Data option go to Pivot Table option and create a new sheet

STEP 4: After creating Pivot Table, go in Pivot Table Editor and in that first add rows & then values.

RELATIVE FREQUENCY

The ratio of the frequency to the total no. of observation is called relative frequency.

→ The steps to construct a relative frequency distribution

Step 1: Obtain a frequency distribution of the data

Step 2: Divide each frequency by the total no. of observation.

Why relative frequency ?

→ for comparing two data sets.

→ Because relative frequencies always fall b/w 0 and 1, they provide a standard for comparison.

Date October 7, 2020

CHARTS OF 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.

The steps to construct a pie-chart -

Step 1: Obtain a relative frequency distribution of data.

Step 2: Divide a circle into pieces proportional to the relative frequencies.

Step 3: Label the slices with the distinct values and their relative frequencies.

EXAMPLE

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

Category	Tally mark	Frequency	R.F.	Degrees
A	𝍸 |	6	0.4	114
B	|||	3	0.2	72
C	|||	3	0.2	72
D	|||	3	0.2	72
Total		15	1	360°

Pie chart in a google sheet

STEP 1 : Select / Highlight the cells having data you want to visualize.

STEP 2 : Click the insert chart option in Google Sheets toolbar.

STEP 3 : Change the visualization type in chart Editor

STEP 4 : Sleet in Chart Editor, Chart Type to Pie Chart.

SECTIONAL SUMMARY

- A pie chart is used to show the proportions of the a categorical variable.
- 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 percents) of those values on a vertical axis.

The frequency / relative frequency of each distinct value is represented by a vertical bard whose height is equal to the frequency / relative frequency of that value. The bars should be positioned so that they do not touch each other.

The steps to construct a bar chart -

Step 1: Obtain a frequency / relative frequency distribution of the data.

Step 2: Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies / relative frequencies.

Step 3: For each distinct value, construct a vertical bar whose height equals the frequency / relative frequency of that value.

Step 4: Label the bars with the distinct values, the horizontal axis with the name of the variable, and the vertical axis with "Frequency" / "Relative Frequency"

Bar Chart In a Google Sheet

STEP 1: Select / Highlight the cells having data you want to visualize.

STEP 2: Click the Insert Chart option in Google Sheets Toolbar.

STEP 3: Change the visualization type in Chart Editor.

STEP 4: 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.

▶ If the categorical variable is ORDINAL, then the bar chart must preserve the ordering.

SECTIONAL SUMMARY

- A bar chart is used to show the frequencies / relative frequencies of a categorical variable.
- If ordinal, the order of categories is preserved.
- The bars can be oriented either horizontally or vertically.
- A Pareto chart is a bar chart where the categories are sorted by frequency.

Date October 17, 2020

BEST PRACTICES

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 between different groups.

LABEL YOUR DATA

→ Label 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 cases, grouping other categories together might be done.

October 17, 2020

MISLEADING GRAPHS

classmate
Date
Page

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 represents.

→ Violations of the area principle are a common way to mislead with statistics.

MISLEADING GRAPHS

Lec 2.3-2.4 Pg-18

(1) Violating Area Principle

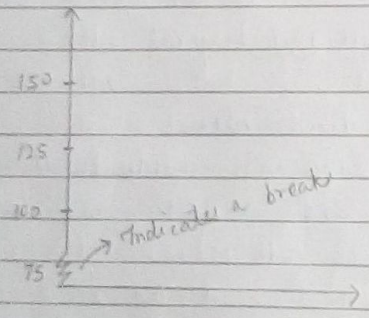
→ Decorated Graphics : charts decorated to attract attention often violate the area principle

(2) Truncated Graphs - Pg 23

→ Another common violation is when the baseline of the bar chart is not at zero.

Left graph exaggerates the no. coming from the South and North. Graph on right shows same data with the baseline at zero.

INDICATING A Y-AXIS BREAK



ROUND-OFF ERRORS - Pg 26

→ Important to check for round off errors

→ When the 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

- Know your purpose & choose the table/graph appropriately
- Labels your charts
- Handle multiple categories appropriately
- Respect Area Principle
  - Avoid Overly decorated graphs
  - Avoid Truncated Graphs (use special symbols to indicate vertical axis has been modified)
  - Check for round off errors.

October 18, 2020

SUMMARIZING CATEGORICAL DATA

→ 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 center or most typical value of a data set lies are called MEASURES OF CENTRAL TENDENCY.

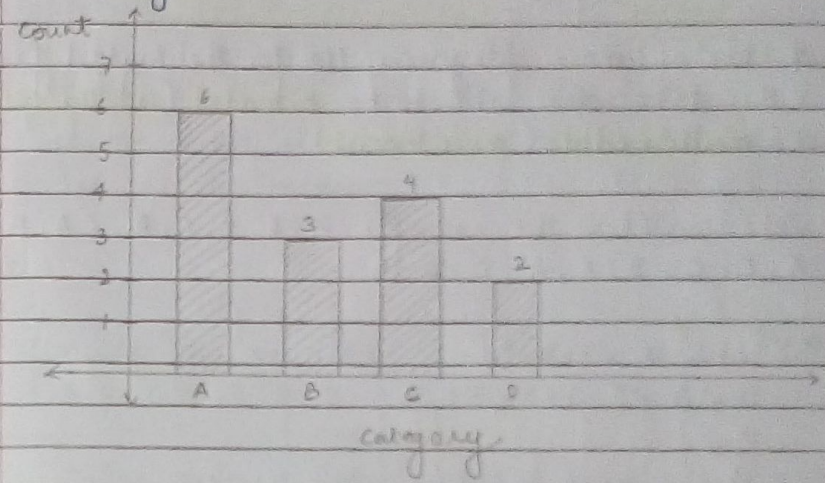
MODE

The mode of a categorical variable is the most common category, the category with the highest frequency.

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

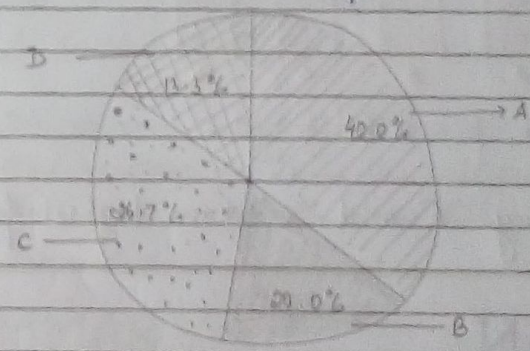
Example : Lets consider the example A, A, B, C, A, D, A, B, C, C, A, B, C, D, A

→ The longest bar in the bar chart



The most common category is "A".

→ The widest slice in a pie chart

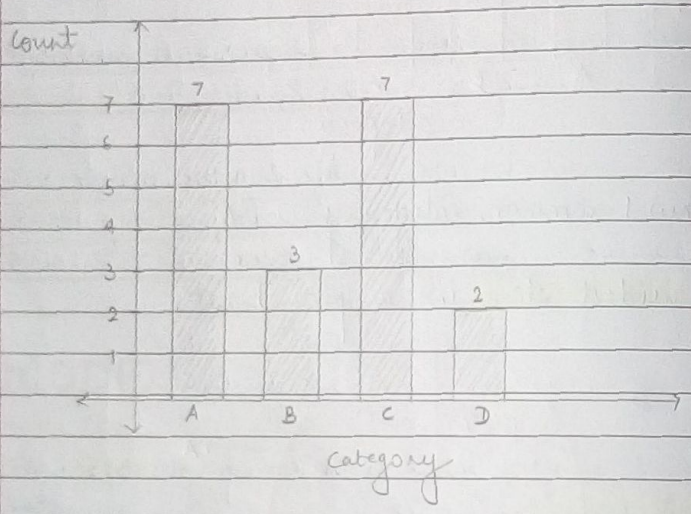


The most common category is "A".

BIMODAL & MULTIMODAL DATA

→ If two or more categories tie for the highest frequency, the data are said to be bimodal (in the case of 2) or multimodal (more than 2).

→ Lets consider the example A, A, B, C, A, C, A, B, C, C, A, C, C, D, A, A, C, D, B.



→ Both category "A" and "C" have highest frequency.

MEDIAN

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

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

→ If there are an even no. of observations, choose the category on either side of the middle of the sorted list as the median.

Example :

(1) Consider the grades of 15 students which is listed as A, B, B, C, A, D, B, B, A, C, B, B, C, D, A

Odd median = $(n+1) \div 2$ (n = no. of observation)

→ The ordered data is A, A, A, A, B, B, B, B, B, B, C, C, C, D, D
→ The median grade is the category "B", i.e, category associated with the 8th observation.

Even median = $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)$ (n = no. of observations)

(2) Consider the grade of 14 students which is listed as A, B, B, C, A, D, B, B, A, C, B, B, C, D.

→ The ordered data is A, A, A, B, B, B, B, B, B, C, C, C, D, D
→ The median grade is the category associated with the 7 or 8 observation which is "B".

SUMMARY

- 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 values.

Summary Of Week 2

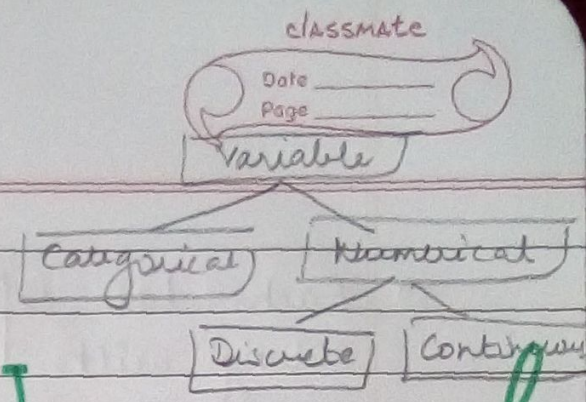
1. Tabulate Data : frequency and relative frequency

2. Charts of Categorical Data
   - Pie Charts
   - Bar Charts and Pareto Charts

3. Best Practices & Misleading Graphs
   - Label your Data
   - Dealing with multiple categories
   - Area Principle
   - Misleading Graphs
     - Decorated graphs
     - Truncated graphs (Missing Baseline)
     - Round off errors

4. Descriptive Measures
   - Mode (most frequency)
   - Median for ordinal data

October 29, 2020



WEEK 3

Describing Numerical Data

ORGANISING NUMERICAL DATA

- A discrete variable usually involves a count of something whereas a continuous variable usually involves a measurement of something.

- First, group the observations into classes (also known as categories or bins) and then treat the classes as the distinct values of qualitative data.

- Once we group the quantitative data into classes, we can construct frequency and relative frequency distribu-tions of the data in exactly the same way as we did for categorical data.

ORGANIZING DISCRETE DATA (SINGLE VALUE)

- If the data set contains only a relatively small number of distinct, or different, values, it is convinient to represent it in a frequency table

→ Each class represents a distinct value (single value) along with its frequency of occurrence.

EXAMPLE : Suppose the dataset reports the no. of people in a household. The following data is the response from 15 individuals.

→ 2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4

→ The distinct values the variable, no. of people in each household, takes is 1, 2, 3, 4, 5

→ The frequency distribution table is

Value	Tally Mark	Frequency	Relative Frequency
1	II	2	0.133
2	III	3	0.2
3	IIII	5	0.333
4	IIII	4	0.266
5	I	1	0.066
TOTAL		15	1

ORGANIZING CONTINUOUS DATA

Organize the data into a number of classes to make the data understandable. However, there are few guidelines that need to be followed. They are:

1. Number of classes → The appropriate no. is a subjective choice, the rule of thumb is to have between 5 and 20 classes.
2. Each Observation should belong to some class and no observation should belong to more than one class.
3. It is common, although not essential, to choose class intervals of equal length.

SOME NEW TERMS

1. Lower class Limit : The smallest value that could go in a class.
2. Upper class limit : The largest value that could go in a class.
3. Class Width : The difference between the lower limit of a class & the lower limit of the next higher class.
4. Class Mark : The average of the two class limits of a class.
5. A class interval contains its left-end but not its right-end boundary point.

EXAMPLE : The marks obtained by 50 students in a particular course.

→ 68, 79, 38, 68, 35, 70, 61, 47, 58, 66, 60, 45, 61, 60, 59, 45, 39, 80, 59, 62, 49, 76, 54, 60, 53, 55, 62, 58, 67, 55, 86, 56, 63, 64, 67, 50, 51, 78, 56, 62, 57, 69, 58, 52, 42, 66, 42, 36, 58

Class Interval	Tally Mark	Frequency	Relative Frequency
30 - 40	III	3	0.06
40 - 50	𝍸 I	6	0.12
50 - 60	𝍸 𝍸 𝍸 III	18	0.36
60 - 70	𝍸 𝍸 𝍸 II	17	0.34
70 - 80	IIII	4	0.08
80 - 90	II	2	0.04
TOTAL		50	1

SECTION SUMMARY

- Frequency table for discrete single value data.
- Frequency table for continuous data using class intervals.

GRAPHICAL SUMMARIES

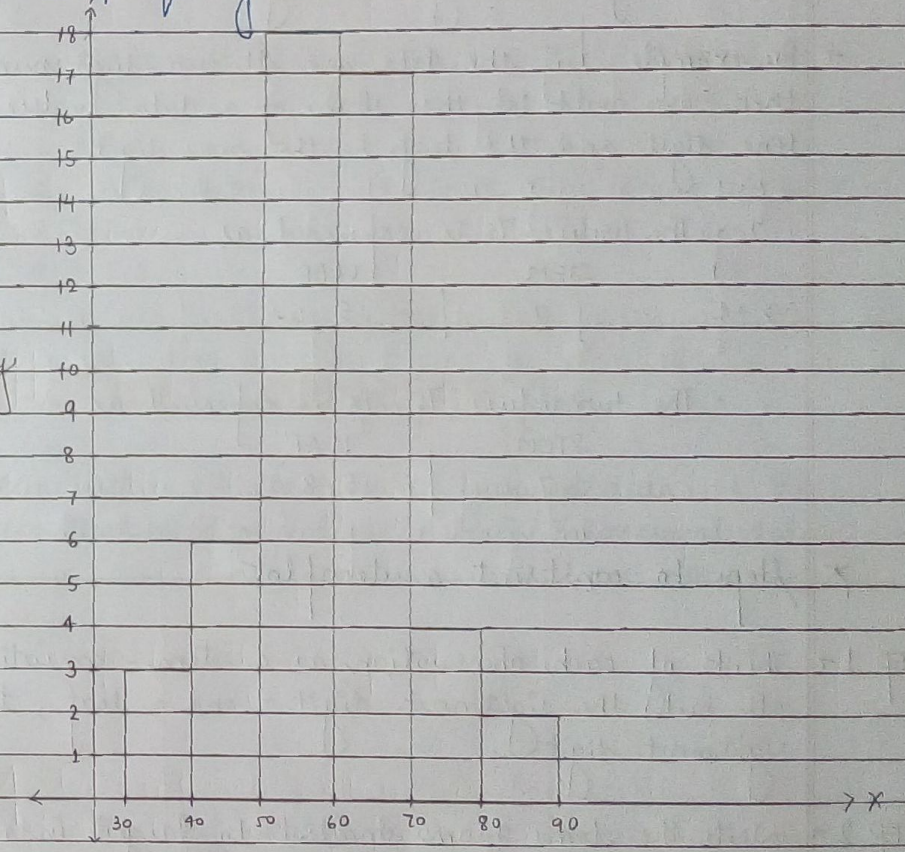
Steps to construct a histogram

STEP 1 → Obtain a frequency (relative frequency) distribution of the data.

STEP 2 → Draw a horizontal axis on which to place the classes and a vertical axis on which to display the frequencies

STEP 3 → For each class, construct a vertical bar whose height equals the frequency of that class.

STEP 4 → Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with "frequency".



HISTOGRAM

STEM AND LEAF DIAGRAM

In a stem-and-leaf diagram (or stemplot), each observation is separated into two parts, namely, a stem - consisting of all but the rightmost digit - and a leaf, the rightmost digit.

→ For example, if the data are all two-digit numbers, then we could let the stem of a data value be the tens digit and the leaf be the ones digit.

- The value 75 is expressed as

STEM	LEAF
7	5

- The two values 75, 78 is expressed as

STEM	LEAF
7	5, 8

→ Steps to construct a stemplot

STEP 1→ Think of each observation as a stem - consisting of all but the rightmost digit - and a leaf, the rightmost digit.

STEP 2→ Write the stems from smallest to largest in a vertical column to the left of a vertical rule.

STEP 3→ Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.

STEP 4→ Arrange the leaves in each row in ascending order.

EXAMPLE: The following are the ages, to the nearest year, of 11 patients admitted in a certain hospital:
15, 22, 29, 36, 31, 23, 45, 10, 25, 28, 48.

Draw a stem-and-leaf plot for this data set.

1	0, 5
2	2, 3, 5, 8, 9
3	1, 6
4	5, 8

SECTION SUMMARY

- construct a histogram for grouped data
- construct a stemplot to describe numerical data.

October 30, 2020

NUMERICAL SUMMARIES

DESCRIPTIVE MEASURES

- The objective is to develop measures that can be used to summarize a data set.
- These descriptive measures are quantities whose values are determined by the data.
- Most commonly used descriptive measures can be categorized as

→ Measures of Central Tendency : These are measures that indicate the most typical value or centre of a data set.

→ Measures of Dispersion : These measures indicate the variability or spread of a dataset.

MEASURES OF CENTRAL TENDENCY

1. MEAN

The most commonly used measure of central tendency is the mean.

Definition : The mean of a data set is the sum of the observations divided by the number of observations

→ The mean is usually referred to as 'Average'.

→ Arithmetic average ; divide the sum of the values by the number of values (another typical value)

→ For DISCRETE observations :

- Sample mean : $\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$
- Population mean : $\mu = \frac{x_1 + x_2 + \dots + x_N}{N}$

Example : The marks obtained by ten students in an exam is
68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ The sample mean is

$$\frac{68 + 79 + 38 + 68 + 35 + 70 + 61 + 47 + 58 + 66}{10} = \frac{590}{10} = 59$$

★ MEAN FOR GROUPED DATA (DISCRETE)

→ The following data is the response from 15 individuals
2, 1, 3, 4, 5, 2, 3, 3, 3, 4, 4, 1, 2, 3, 4.

→ $$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \dots + f_n x_n}{n}$$

where, f_i is frequency
x_i is value
n is total observations.

Value (x_i)	Tally Mark	Frequency (f_i)	$f_i x_i$
1	II	2	2
2	III	3	6
3	卌	5	15
4	IIII	4	16
5	I	1	5
TOTAL		15	44

$$\text{Mean} = \frac{\sum f_i x_i}{\sum f_i} = \frac{44}{15} = 2.93$$

★ MEAN FOR GROUPED DATA (CONTINUOUS)

$$\bar{x} = \frac{f_1 m_1 + f_2 m_2 + \ldots + f_n m_n}{n}$$

where m is MID POINT of class interval.

CLASS INTERVAL	TALLY MARK	FREQUENCY (f_i)	MID POINT (m_i)	$f_i m_i$
30-40	III	3	35	105
40-50	卌 I	6	45	270
50-60	卌 卌 卌 III	18	55	990
60-70	卌 卌 卌 II	17	65	1105
70-80	IIII	4	75	300
80-90	II	2	85	170
TOTAL		50		2940

→ AVERAGE $= \frac{2940}{50} = 58.8$

→ 58.8 is an approximate and not exact value of the mean.

★ ADDING A CONSTANT

- Let $y_i = x_i + c$ where c is a constant then $\bar{y} = \bar{x} + c$
- Example : Recall the marks of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ Suppose the teacher has decided to add 5 marks to each student.

→ Then the data becomes
73, 84, 43, 73, 40, 75, 66, 52, 63, 71

→ The mean of the new data set is $\frac{640}{10} = 64 = 59 + 5$
$\bar{y} = \bar{x} + c$

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$ where c is a constant then $\bar{y} = \bar{x}c$
- Example : Recall the marks of of students
  68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

→ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4

→ The mean of the new data set is $\frac{236}{10} = 23.6 = 59 \times 0.4$
$\bar{y} = \bar{x}c$

SECTION SUMMARY

1. Mean or average is a measure of central tendency
2. Compute sample mean for
   - → ungrouped data
   - → grouped discrete data
   - → grouped continuous data
3. Manipulating Data
   - → Adding a constant to each data pt.
   - → Multiplying each data point with a constant

Date October 31, 2020

NUMERICAL SUMMARIES

Measures of Central Tendency

2. MEDIAN

→ Another frequently used measure of center is the median.

→ Essentially, the median of a data set is the number that divides the bottom 50% of the data from the top 50%.

Definition : The median of a data set is the middle value in its ordered list.

Steps to obtain median

Arrange the data in increasing order. Let n be the total no. of observations in the dataset.

1. If the no. of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e, $(n+1)/2$ observation.

2. If the no. of observation is even, then the median is the mean of the two middle observations in the ordered list, i.e, mean of $n/2$ and $n/2 + 1$ observation.

EXAMPLES : (1) 2, 12, 5, 6, 7, 3

- Arrange data in increasing order → 2, 3, 5, 6, 7, 12
- $n = 7$ (odd) $\therefore$ median = $\left(\frac{n+1}{2}\right)^{th}$ observation = $\frac{8}{2}$ = 4th observation = "6"

(2) 2, 105, 5, 7, 6, 7, 3

- Arrange in ↑ order : 2, 3, 5, 6, 7, 7, 105
- n = 7 (odd)

∴ Median = $\left(\frac{n+1}{2}\right)^{th}$ observation = $\left(\frac{7+1}{2}\right) = \frac{8}{2} = 4^{th}$ obs. = "6"

(3) 2, 105, 5, 7, 6, 3

- Arrange in ↑ order : 2, 3, 5, 6, 7, 105
- n = 6 (even)

∴ Median is avg. of $\left(\frac{n}{2}\right)^{th}$ & $\left(\frac{n}{2}\right)+1^{th}$ observation

⇒ $\frac{5+6}{2}$ = 5.5 is median.

Difference Between Mean & Median

EXAMPLE : (1) 2, 12, 5, 6, 7, 3, 7

Sample mean = $\frac{2+3+6+5+7+7+12}{7}$ = 6

Sample median = 6

(2) 2, 117, 5, 7, 6, 7, 3

Sample mean = $\frac{2+3+5+6+7+7+117}{7}$ = 21

Sample median = 6

Note → The sample mean is sensitive to outliers, whereas the sample median is not sensitive to outliers.

★ ADDING A CONSTANT

- Let $y_i = x_i + c$, where c is the constant, then

new median = old median + c

EXAMPLE, Recall the marks of students

68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ Arranging in ascending order, 35, 38, 47, 58, 61, 66, 68, 68, 70, 79

→ The median for the data is avg. of (n/2) & (n/2 + 1) observation which is (61+66)/2 = 127/2 = 63.5

→ Suppose the teacher has decided to add 5 marks to each student.

→ Then data in ascending order is

40, 43, 52, 63, 66, 71, 73, 73, 75, 84

→ The median of the new data set is $\frac{66+71}{2} = \frac{137}{2}$ = 68.5

→ 68.5 = 63.5 + 5

new median = old median + constant

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$ where c is a constant then

new median = old median × c

EXAMPLE, Recall the marks of students

68, 79, 38, 68, 35, 70, 61, 47, 58, 66

We already know median for this data is 63.5.

→ Suppose the teacher has decided to scale down each mark by 40%, in other words each mark is multiplied by 0.4.

→ Then the data becomes
27.2, 31.6, 15.2, 27.2, 14, 28, 24.4, 18.8, 23.2, 26.4
The ascending order is
14, 15.2, 18.8, 23.2, 24.4, 26.4, 27.2, 28, 31.6

The median of the new data set is $(24.4 + 26.4) \div 2$
$= 50.8 \div 2 = 25.4$

→ Note, $25.4 = 0.4 \times 63.5$
new median = constant × old median

3. MODE

Another measure of central tendency is the sample mode

Definition : The mode of the data set is its most frequently occurring value.

Steps To Obtain Mode

- If no value occurs more than once, then the data set has no mode.
- Else, the value that occurs with the greatest frequency is a mode of the data set.

★ ADDING A CONSTANT

- Let $y_i = x_i + c$ where c is a constant then
new mode = old mode + c

EXAMPLE, Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66
The mode for this data is 68.

→ Suppose the teacher has decided to add 5 marks to each student.
Then the data in ascending order is
40, 43, 52, 63, 66, 71, 73, 73, 75, 84

→ The mode of the new data set is 73

→ $73 = 68 + 5$
$\text{new}_{\text{mode}} = \text{old}_{\text{mode}} + \text{constant}$

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$, where c is a constant then

new mode = old mode × c

SUMMARY

→ Measure of Central Tendency : Mean, Median, Mode
→ Impact of adding a constant or multiplying with constant on the measures.

Date: October 31 2020

Measures Of Dispersion

Why Do We Need a Measure of Dispersion?

→ Consider the two data sets given below :
- Dataset 1 → 3, 3, 3, 3, 3
- Dataset 2 → 1, 2, 3, 4, 5

→ The measures of central tendency for both datasets are:

	Dataset 1	Dataset 2
MEAN	3	3
MEDIAN	3	3
MODE	3	not available

→ The mean median are same for both datasets. However the datasets are not same. They are different.

→ To describe that difference quantitatively, we use a descriptive measure, that indicates the amount of variation, or spread, in a data set.

→ Such descriptive measures are refered to as
- measures of dispersion, or
- measures of variation, or
- measures of spread

→ In this course, we will be discussing about the following measures of dispersion.

1. Range
2. Variance
3. Standard Deviation
4. Interquartile range

1. RANGE

Definition : The range of a dataset is the difference b/w its largest and smallest values.

→ The range of a dataset is given by the formula

$$\text{Range} = \text{Max} - \text{Min}$$

where max and min denote the maximum & minimum observations, respectively.

		Dataset 1	Dataset 2
		3, 3, 3, 3, 3	1, 2, 3, 4, 5
→	Max	3	5
	Min	3	1
	Range	0	4

∴ Range is sensitive to outliers. For example, consider two datasets as given below,

	Dataset 1	Dataset 2
	1, 2, 3, 4, 5	1, 2, 3, 4, 15
Max	5	15
Min	1	1
Range	4	14

→ Though the two datasets differ only in one datapoint, we can see that this contributed to the value of range significantly. This happens because the range takes into consideration only the Min & Max of the dataset.

2. VARIANCE

- In contrast to the range, the variance takes into account all the observations.
- One way of measuring the variability of a dataset is to consider the deviations of the data values from a central value.

POPULATION VARIANCE & SAMPLE VARIANCE

Recall when we refer to a dataset from a population, we assume the dataset has N observations, whereas, when refer to a dataset from a sample, we assume the datbset has n observations.

- The variance is computed using the following formulae-

→ Population Variance : $\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \dots + (x_N-\mu)^2}{N}$

→ Sample Variance : $s^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \dots + (x_n-\bar{x})^2}{n-1}$

- The numerator is the sum of squared deviations of every observation from its mean.
- The denominator for computing population variance is N, the total no. of observations.
- The denominator for computing sample variance is (n-1)

EXAMPLE : Recall marks of students obtained by 10 students in an exam is 68, 79, 38, 68, 35, 70, 61, 47, 58, 66

→ The mean was computed to be 59.

→ The deviations of each data pt. from its mean is given in the table below :

	Data	Deviation $(x_i - \bar{x})$	Squared Deviations $(x_i - \bar{x})^2$
1	68	9	81
2	79	20	400
3	38	-21	441
4	68	9	81
5	35	-24	576
6	70	11	121
7	61	2	4
8	47	-12	144
9	58	-1	1
10	66	-7	49
TOTAL	590	0	1898

1. Population Variance = $\frac{1898}{10}$ = 189.8

2. Sample Variance = $\frac{1898}{9}$ = 210.88

★ ADDING A CONSTANT

→ Let $y_i = x_i + c$, where c is constant then,

new variance = old variance

EXAMPLE, Recall the marks of students
68, 79, 38, 68, 35, 70, 61, 47, 58, 66
Sample Variance = 210.88

→ Suppose the teacher has decided to add 5 marks to each student.
Then the data is 73, 84, 43, 73, 40, 75, 66, 52, 63, 71

→ The variance of new dataset is $\frac{1898}{9}$ = 210.88

→ In general, adding a constant does not change variability of a dataset, and hence it is the same.

★ MULTIPLYING A CONSTANT

• Let $y_i = x_i c$, where c is a constant then

new variance = c^2 x old variance

3. STANDARD DEVIATION

Another very useful measure of dispersion is the standard deviation

Definition : The quantity

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1}}$$

which is the square root of sample variance is the sample standard deviation.

UNITS OF STANDARD DEVIATION

→ The sample variance is expressed in units of square units if original variable. For example, instead of marks if the data were weights of 10 students measured in kg. Then the unit of variance would be kg^2.

→ The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in kg, then the units of std. deviation are also in kg.

★ ADDING A CONSTANT

• Let $y_i = x_i + c$, where c is a constant then,

new variance = old variance

★ MULTIPLYING A CONSTANT

- Let $y_i = x_i c$ where c is a constant then,

  new variance = c^2 x old variance

SECTION SUMMARY

- Measures of Dispersion
  1. Range
  2. Variance
  3. Standard Deviation
- Impact of Adding a constant or multiplying with a constant on the measures.

Date October 31, 2020

PERCENTILES

→ The sample 100p percentile is that data value having the property that at least 100p percent of the data are less than or equal to it and at least $100(1-p)$ percent of the data values are greater than or equal to it.

→ If two data values satisfy this condition, then the sample 100p percentile is the arithmetic avg of these values.

→ Median is the 50^{th} percentile.

Computing Percentile

- To find the sample 100p percentile of a data set of size n.

1. Arrange the data in increasing order.
2. If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample 100p percentile.
3. If np is an integer, then the avg. of the values in positions np & np + 1 is the sample 100p percentile.

Example : Let n = 10

→ Arrange data in ascending order

35, 38, 47, 58, 61, 66, 68, 68, 70, 79

p	np	Percentile Value
0.1	1	36.5
0.25	2.5	47
0.5	5	63.5
0.75	7.5	68
1	10	79

QUARTILES

Definition : The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or second quartile. The sample 75th percentile is called the third quartile.

In other words,

the quartiles break up a data set into four parts with about 25 percent of the data values being less than the first (lower) quartile, about 25 percent being b/w the first & second quartiles, about 25 percent being b/w the second & third (upper) quartiles, and about 25 percent being larger than the third quartile.

The Five Number Summary

→ Minimum
→ Q_1 : First Quartile or Lower Quartile
→ Q_2 : Second Quartile or Median
→ Q_3 : Third Quartile or Upper Quartile
→ Maximum

4. THE INTERQUARTILE RANGE

Definition : The interquartile range, IQR, is the difference between the first and third quartiles; that is

$$IQR = Q_3 - Q_1$$

→ IQR for the example

- First Quartile, $Q_1 = 49.75$
- Third Quartile, $Q_3 = 68$
- $IQR = Q_3 - Q_1 = 18.25$

SECTION SUMMARY

- Definition of Percentiles
- How to compute percentiles
- Definition of quartile
- Five no. summary
- Interquartile range as a measure of dispersion

SUMMARY OF WEEK 3

1. Frequency Tables
   - Frequency tables for discrete data
   - Frequency table for continuous data

2. Graphical Summaries
   - Histograms
   - Stem-and-leaf plot

3. Numerical Summaries
   - Measures of central Tendency
     → Mean, Median, Mode
   - Measures of Dispersion
     → Range, Variance, Std. Deviation
   - Percentiles
     → Interquartile range as a measure of dispersion