IIT MADRAS

BSc IN PROGRAMMING AND DATA SCIENCE

Statistics 1 Notes W1-3

Aditya Dhar Dwivedi

STATISTICS FOR DATA SCIENCE – 1

PROFESSOR USHA MOHAN

INDIAN INSTITUTE OF TECHNOLOGY

MADRAS

❖INTRODUCTION

What are 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 the drawing of conclusions.

Major branches of statistics

1. Descriptive Statistics

The part of statistics concerned with the description and summarization of data is called descriptive statistics.

2.Inferencial Statistics

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- introduction to probability.

Population and Sample

- The total collection of all the elements that we are interested in is called a population.
- 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 and explore information for its own intrinsic interest only, the study is descriptive.
- If the information is obtained from a sample of a population and the purpose of the study is to use that information 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.

What is Data?

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

Data are the facts and figures collected, analysed, and summarized 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.

Unstructured and 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

Variables and cases

- Case (observation): A unit from which data are collected
- Variable:
  - Intuitive: A variable is that “varies”. I
  - Formally: A characteristic or attribute that varies across all units.

- In our school data set:
  - Case: each student
  - Variable: Name, marks obtained, Board etc.
- Rows represent cases: for each case, same attribute is recorded
- Columns represent variables: For each variable, same type of value for each case is recorded.

Summary-- We have organized data in a spreadsheet into a table.

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

Understanding Data—Classification of Data

♦ Categorical and Numerical Data

- Categorical data
  - Also called qualitative variables.
  - Identify group membership
- Numerical data
  - Also called quantitative variables.
  - Describe numerical properties of cases
  - Have measurement units
- Measurement units: Scale that defines the meaning of numerical data, such as weights measured in kilograms, prices in rupees, heights in centimetres, etc.
  - The data that make up a numerical variable in a data table must share a common unit.

♦ Cross-sectional and Time-series Data

- Time series - data recorded over time
- Time plot – 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).

♦ Scales of Measurement

➢ Nominal scale of measurement

- When the data for a variable consist of labels or names used to identify the characteristic of an observation, the scale of measurement is considered a nominal scale. I
- Examples: Name, Board, Gender, Blood group etc. I
- Sometimes nominal variables might be numerically coded. I

#For example: We might code Men as 1 and Women as 2. Or Code Men as 3 and Women as 1. Both codes are valid.

- There is no ordering in the variable. I

✓ 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. I

#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 the 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 two 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 and uncomfortable, then the temperature as a variable is nominal. I

#Suppose the answer to measuring the temperature of a liquid is cold, warm, hot - the variable is ordinal. I

#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 temperature is 20°C, but it is incorrect to say that the outdoors is twice as hot as indoors. I

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

	Celsius	Fahrenheit
Freezing Point	0	32
Boiling Point	100	212

➢ 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 (makes ratio comparisons possible)

➢ **Summary**

True zero exists possible	**Ratio Scale**	Age, Height, Weight, Marks Etc.
No absolute zero, Difference exists	**Interval Scale**	Temperature, GPA Etc.
Named + ordered categories	**Ordinal Scale**	Ranking, Rating Etc.
Named categories	**Nominal Scale**	Name, Blood group Etc.

Describing Categorical Data-Single Variable

Frequency distributions

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

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

♦ Relative frequency

➢ The ratio of the frequency to the total number of observations 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 number of observations.

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.

Why relative frequency?

- For comparing two data sets.
- Because relative frequencies always fall between 0 and 1, they provide a standard for comparison

Example

- Construct a frequency table for the given data
- 1. A, A, B, C, A, D, A, B, D, C
- 2. A, A, B, C, A, D, A, B, D, C, A, B, C, D, A
- 3. A, A, B, C, A, A, B, B, D, C, A, B, C, D, B
- 4. A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D,

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

Category	Tally mark	Frequency	Relative frequency
A	IIII	4	0.4
B	II	2	0.2
C	II	2	0.2
D	II	2	0.2
Total		10	1

2. A, A, B, C, A, D, A, B, D, C, A, B, C, D, A

Category	Tally mark	Frequency	Relative frequency
A		6	0.4
B		3	0.2
C		3	0.2
D		3	0.2
Total		15	1

3. A, B, B, C, A, D, B, B, D, C, A, B, C, D, B

Category	Tally mark	Frequency	Relative frequency
A		3	0.2
B		6	0.4
C		3	0.2
D		3	0.2
Total		15	1

4. A, A, B, C, A, D, A, B, D, C, A, B, C, D, A, C, D, D

Category	Tally mark	Frequency	Relative frequency
A		6	0.33
B		3	0.17
C		4	0.23
D		5	0.27
Total		18	1

Charts of categorical data

- The two most common displays of a categorical variable are a bar chart and a pie chart. I
- Both describe a categorical variable by displaying its frequency table.

1-Pie Chart

- 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 the 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

- Use a protractor and the fact that there are 360◦ in a circle. Thus, for example, the first slice of the circle is obtained by marking off 0.4×360 = 144◦.

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

Category	Tally mark	frequency	Relative freq.	Degree
A		4	0.4	144
B		2	0.2	72
C		2	0.2	72
D		2	0.2	72
Total		10	1	360*

2-Bar Chart

- A bar chart displays the distinct values of the qualitative data on a horizontal axis and the relative frequencies (or frequencies or precents) 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 frequency/relative frequency of that value. The bars should be positioned so that they do not touch each other.

➢ Steps to construct a bar chart

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."

3-Pareto Chart

➢ 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.

Example- ordinal data

- The T-shirt sizes (Small-S, Medium-M, Large-L) of twenty students is listed below:

L, M, M, S, L, S, S, M, L, M, M, S, S, L, M, S, M, S, L, M

Size	Tally mark	frequency	Relative freq.
Small		7	0.35
Medium		8	0.40
Large		5	0.25
Total		20	1

Sectional summary

- A pie chart is used to show the proportions of a categorical variable.
- A pie chart is a good way to show that one category makes up more than half of the total.

- 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.

Best Practise and Misleading graph

➢ 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 case, grouping other categories together might be done.

➢ 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.

Frequency table in a google sheet

- 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 and then values.

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 Select in Chart Editor Chart type to Pie chart.

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.

Misleading graphs

- Violating area principal
  - Decorated graphics: Charts decorated to attract attention often violate the area principle
- Truncated graphs
  - Another common violation is when the baseline of a bar chart is not at zero.
- Indicating a y- axis break
- Round-off errors
  - Important to check for round-off errors.
  - 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

- Know your purpose and choose table/graph appropriately
- Label 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

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

Example

- The longest bar in a bar chart.
- The widest slice in a pie chart.

Bimodal and multimodal data

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

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 number of observations, choose the category on either side of the middle of the sorted list as the median.

Example

- 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
  - 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 associated with the 8 observation which is" B".
- Consider the grades 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".

- 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
  - 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 associated with the 8 observation which is" B".
  - The most common grade is" B", hence mode is" B"
  - In this example both mode and median are the same.
- Consider the grades of 15 students which is listed as A, B, B, C, A, D, A, B, A, C, B, A, C, D, A
  - The ordered data is A, A, A, A, A, A, B, B, B, B, C, C, C, D, D
  - The median grade is the category associated with the 8 observation which is" B".
  - The most common grade is" A", hence mode is" A"
  - In this example both mode and median are the different

Sectional 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

- Tabulate data: frequency and relative frequency.
- Charts of categorical data
  - Pie charts
  - Bar charts and Pareto charts
- Best practices and misleading graphs
  - Label your data.
  - Dealing with multiple categories.
  - Area principle
  - Misleading graphs
    - Decorated graphs
    - Truncated graphs.
    - Round-off errors.
- Descriptive measures
  - Mode.
  - Median for ordinal data.

Describing Numerical Data

• Frequency tables for numerical data

Organizing numerical data

- Recall, 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 distributions 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 convenient 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 number 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, number of people in each household, takes is 1,2,3,4,5.

➢ The frequency distribution table is

Value	Tally mark	Frequency	Relative frequency
1		2	0.13
2		3	0.2
3		5	0.33
4		4	0.27
5		1	0.07
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
- Number of classes: The appropriate number is a subjective choice; the rule of thumb is to have between 5 and 20 classes.
- Each observation should belong to some class and no observation should belong to more than one class.
- It is common, although not essential, to choose class intervals of equal length.

Some new terms

- Lower class limit: The smallest value that could go in a class.
- Upper class limit: The largest value that could go in a class.
- Class width: The difference between the lower limit of a class and the lower limit of the next-higher class.
- Class mark: The average of the two class limits of a class.

✓ 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, 56, 58.

Class interval	Tally mark	Frequency	Relative frequency
30-40		3	0.06
40-50		6	0.12
50-60		18	0.36
60-70		17	0.34
70-80		4	0.08
80-90		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 (relative frequencies).
- **Step 3** For each class, construct a vertical bar whose height equals the frequency (relative 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" ("Relative frequency").

➢ Stem-and-leaf diagram

In a stem-and-leaf diagram (or stem plot)1, 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 stem plot

- 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 stem plot to describe numerical data.

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

• The Mean

The most commonly used measure of central tendency is the mean.
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: $x = \frac{x_1+x_2+\cdots+x_n}{n-1}$
  - Population mean: $\mu = \frac{x_1+x_2+\cdots+x_N}{N}$

Example

- 2, 12, 5, 7, 6, 7, 3; x = (2+12+5+7+6+7+3)/ 7 = 42/ 7 = 6
- 2, 105, 5, 7, 6, 7, 3; x = (2+105+5+7+6+7+3)/ 7 = 135/ 7 = 19.285
- 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 single value data

- The following data is the response from 15 individuals.

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

- $\bar{X} = \frac{f_1x_1+f_2x_2+\cdots+f_nx_n}{n}$

Value (xi)	Tally mark	Frequency	fixi
1		2	2
3		3	6
3		5	15
4		4	16
5		1	5
Total		15	44

Mean=44/ 15=2.93

Mean for grouped data: continuous data

- $\bar{X} = \frac{f_1m_1 + f_2m_2 + \ldots + f_nm_n}{n}$

Class interval	Tally mark	Frequency (fi)	Mid-point (mi)	fimi
30-40		3	35	105
40-50		6	45	270
50-60		18	55	990
60-70		17	65	1105
70-80		4	75	300
80-90		2	85	170
Total		50		2940

Average = 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 $y_i = x_i + c$
- 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 640 /10 = 64 = 59 + 5

Multiplying a constant

- Let $y_i = x_i c$ where c is a constant then $y_i = x_i c$
- Example: Recall the marks 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 236/ 10 = 23.6 = 59×0.4

Section summary

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

Median

- Another frequently used measure of centre 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%.

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 number of observations in the dataset.

- If the number of observations is odd, then the median is the observation exactly in the middle of the ordered list, i.e. $\frac{n+1}{2}$th observation
- If the number of observations is even, then the median is the mean of the two middle observations in the ordered list, i.e. mean of $\frac{n}{2}$th and $\frac{n}{2}$ +1th observation

Example

- 2, 12, 5, 7, 6, 7, 3
  - Arrange the data in increasing order --2, 3, 5, 6, 7, 7, 12
  - n = 7 odd, median is the n+1/ 2 = 8/ 2 = 4th observation, "6".
- 2, 105, 5, 7, 6, 7, 3
  - Arrange the data in increasing order --2, 3, 5, 6, 7, 7, 105
  - n = 7 odd, median is the n+1/ 2 = 8/ 2 = 4th observation, "6".
- 2, 105, 5, 7, 6, 3
  - Arrange the data in increasing order --2, 3, 5, 6, 7, 105
  - n = 6 even, median is the average of n 2 and n/ 2 + 1 observation = 5+6/2 = 5.5

Example

- 2, 12, 5, 7, 6, 7, 3 1.1
  - Sample mean= (2+3+5+6+7+7+12)/ 7 = 6
  - Sample median = 6
- 2, 117, 5, 7, 6, 7, 3 2.1
  - Sample mean= (2+3+5+6+7+7+117)/ 7 = 21
  - Sample median = 6

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 a 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 this data is the average of n/ 2 and 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 the data in ascending order is 40,43,52,63,66,71,73,73,75,84
- The median of the new dataset is 66+71/ 2 = 137/ 2 = 68.5
- Note 68.5=63.5+5

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 new dataset is 24.4+26.4/ 2 = 50.8/ 2 = 25.4
- Note 25.4 = 0.4×63.5

- **Mode**
  - Another measure of central tendency is the sample mode.

The mode of a 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.

Example

- 2, 12, 5, 7, 6, 7, 3; -- 7 occurs twice, hence 7 is mode
- 2, 105, 5, 7, 6, 3-- no mode

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 dataset is 73
- Note 73 = 68 + 5

Multiplying 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.
  We already know mode for this data is 68
- 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 mode of new dataset is 27.2
- Note 27.2 = 0.4×68

Section summary

- Measures of central tendency
  1. Mean

2. Median
3. Mode

- Impact of adding a constant or multiplying with a constant on the measures.

❖ Measures of Dispersion

Introduction- 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 the data sets are

	Dataset 1	Dataset 2
Mean	3	3
Median	3	3
Mode	3	Not available

- The mean, median is same for both the 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 referred 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.

- **Range**

The range of a data set is the difference between its largest and smallest values.

- The range of a data set is given by the formula
  Range = Max - Min
  where Max and Min denote the maximum and 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 sensitive to outliers

- 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 contributes to the value of Range significantly. This happens because the range takes into consideration only the Min and Max of the dataset

- **Variance**

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

Population variance and 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 dataset has n observations.

- The variance is computed using the following formulae
  - Population variance: $\sigma^2 = \frac{(x_1-\mu)^2+(x_2-\mu)^2+\ldots+(x_N-\mu)^2}{N}$
  - Sample variance: $s^2 = \frac{(x_1-x)^2+(x_2-x)^2+\ldots+(x_n-x)^2}{n-1}$
- The numerator is the sum of squared deviations of every observation from its mean. I
- The denominator for computing population variance is N, the total number of observations. I The denominator for computing sample variance is (n−1). The reason for this will be clear in forthcoming courses on statistics.

Example

- Recall marks of students obtained by ten 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 point from its mean is given in the table below:

	Data	Deviation from mean $(x_i - \bar{x})$	Squared deviation $(x_i - \bar{x})^2$
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

- Population variance = 1898/ 10 = 189.8
- Sample variance = 1898/ 9 = 210.88

Adding a constant

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

- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has 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 the new dataset is 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× old variance
- Example: Recall the marks of students-- 68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- 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 new dataset is 23.6
- The sum of squared deviations from mean = 303.68 and the
  variance = 303.68/ 9 = 33.74. We can verify that 33.74 = 0.42 ×210.88.

• Standard deviation

- Another very useful measure of dispersion is the standard deviation.
  The quantity

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

which is the square root of sample variance being 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 kilograms. Then the unit of variance would be (kilogram)2
- The sample standard deviation is measured in the same units as the original data. That is, for instance, if the data are in kilograms, then the units of standard deviation are also in kilograms.

Adding a constant

- Let $y_i = x_i + c$ where c is a constant then
  new variance = old variance
- Example: Recall the marks of students
  68,79,38,68,35,70,61,47,58,66. has 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 the new dataset is 1898/ 9 = 210.88
- The standard deviation of the new dataset is √210.88 = 14.522

- 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 × old variance
- Example: Recall the marks of students 68,79,38,68,35,70,61,47,58,66.
  We already know variance for this data is 210.88
- 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 new dataset is 23.6
- The sum of squared deviations from mean = 303.68 and the variance = 303.68/ 9 = 33.74. I The standard deviation of the new data set is √33.74 = 5.808.
  We can verify 5.808 = 0.4×14.522.

Section summary

- Measures of dispersion
  1. Range
  2. Variance: population variance and sample variance.
  3. Standard deviation.
- Impact of adding a constant or multiplying with a constant on the measures.

♦ 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 average of these values.
- Median is the 50th percentile.

Computing Percentile

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

- Arrange the data in increasing order.
- If np is not an integer, determine the smallest integer greater than np. The data value in that position is the sample 100p percentile.
- If np is an integer, then the average of the values in positions np and 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	
0.1	1	(35+38)/2=36.5
0.25	2.5	47
0.5	5	(61+66)/2=63.5
0.75	7.5	68
1	10	79

Computing percentile using google sheets-PERCENTILE function

- **Step 1** Paste the dataset in a column.
- **Step 2** In a blank cell enter PERCENTILE (data, percentile), where data indicates the range of data for which percentile needs to be computed, and percentile is the decimal form of the desired percentile.
  - For example, if the data is in cell A1:A10, and we are interested in computing the 90th percentile, then enter

    PERCENTILE (A1:A10,0.9) in a blank cell.

Computing percentile using google sheets-algorithm

- **Step 1** Arrange data in increasing order

Order	1	2	3	4	5	6	7	8	9	10
$x_{[i]}$	$x_{[1]}$	$x_{[2]}$	$x_{[3]}$	$x_{[4]}$	$x_{[5]}$	$x_{[6]}$	$x_{[7]}$	$x_{[8]}$	$x_{[9]}$	$x_{[10]}$
Data	35	38	47	58	61	66	68	68	70	79

Let $x_{[i]}$ denote the i^{th} ordered value of the dataset.

- **Step** 2 Find rank using the following formula.

  rank = percentile ×(n−1) + 1 where n is total number of observations in the dataset
  - Example: to compute 25 percentiles of a set of n = 10 observations,
    rank = 0.25× (10−1) + 1 = 3.25
- **Step 3** Split the rank into integer part and fractional part.
  - Integer part of 3.25 =3; fractional part is 0.25.
- **Step 4** Compute the ordered data value $x_{[i]}$ corresponding to the integer part rank.
  - The ordered data value corresponding to integer part rank of 3, $x_{[3]}$ is 47.
- **Step 5** The percentile value is given by the formula

  Percentile = $x_{[i]}$ + fractional part × $[x_{[i-1]} - x_{[i]}]$
  - Percentile = 47 + 0.25× [58−47] = 47 + 0.25×11 = 47 + 2.75 = 49.75

Quartiles

- The sample 25th percentile is called the first quartile. The sample 50th percentile is called the median or the 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 between the first and second quartiles, about 25 percent being between the second and third(upper) quartiles, and about 25 percent being larger than the third quartile.

The Five Number Summary

- Minimum
- Q1: First Quartile or lower quartile
- Q2: Second Quartile of Median
- Q3: Third Quartile or upper quartile
- Maximum

The Interquartile Range (IQR)

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

  $$IQR = Q3 - Q1$$

- IQR for the example I
  - First quartile, Q1 = 49.75
  - Third quartile, Q3 = 68
  - IQR = Q3 −Q1 = 18.25

Section summary

- Definition of percentiles.
- How to compute percentiles.
- Definition of quartile.
- Five-number summary.
- Interquartile range as a measure of dispersion.

Summary

- Frequency tables
  - Frequency table for discrete data.
  - Frequency table for continuous data.
- Graphical summaries
  - Histograms.
  - Stem-an-leaf plot.
- Numerical summaries
  - Measures of central tendency
    - Mean, Median, Mode
  - Measures of dispersion
    - Range, Variance, Standard deviation
  - Percentiles
    - Interquartile range as a measure of dispersion.