5CS037 - Concepts and Technologies of Artificial Intelligence
Herald College
University of Wolverhampton

Tutorial - 02
When Data Meets Statistics!!!

Siman Giri {Module Leader - 5CS037}

November - 30 - 2024

Instructions to the Instructor:

- Please Complete all the exercises.

A statistician drowned crossing a river that was, on average, three feet deep.{The Big Short}.

Contents

1 About a Data.

1.1 Thing or Two about a Data:

1. Definition of Data:

Contrary to the popular belief that data is merely a **collection of raw facts**, I define data as a structured collection of factual information about **specific objects or phenomena**. In this definition, data is viewed as a combination of variables and measurements, where variables represent the entities or phenomena being observed, and measurements provide the specific values or characteristics associated with these variables. The measurements can be of two types:

1. Quantitative Data: Quantitative data represents information that can be measured and expressed numerically. It deals with quantities and is often used for mathematical calculations and statistical analysis. Quantitative data is divided into two types:
   - Discrete Data: Whole numbers that cannot be subdivided (e.g., the number of students in a class, the number of cars in a parking lot).
   - Continuous Data: Values that can be any number within a range and can include fractions or decimals (e.g., height, weight, temperature).
2. Qualitative Data: Qualitative data describes qualities or characteristics that cannot be measured with numbers. Instead, it is descriptive and is used to categorize or label elements {Hence also known as Categorical Data}. This type of data is often subjective and categorized based on attributes, properties, or labels rather than numerical values.
   - Nominal Data: Categories with no specific order (e.g., colors, types of food, gender).
   - Ordinal Data: Categories that follow a specific order but don't have precise differences between values (e.g., rankings, satisfaction levels like "satisfied," "neutral," "dissatisfied").

2. Data Forms and Formats:

Data in general are organized in two way:

- **Structured Data**: Organized in a predefined format, typically in rows and columns with clear labels and consistent types (e.g., relational databases, spreadsheets).
  - Easily searchable and analyzable due to its fixed schema and organization, ideal for numerical or categorical data (like sales records).
- **Unstructured Data**: Lacks a predefined format or consistent structure, making it harder to organize (e.g., text files, emails, images, videos).
  - Requires specialized tools to analyze since it doesn't fit into traditional databases, common in applications like NLP, computer vision, and big data analytics.

{Disclaimer!! This module will only talk about dataset in structured format.}

2.1 Data Forms:

Structured Data can be put into in various forms based on its type and structure. Here are some common forms:

1. Tabular Data:
   - Structure: Rows and columns (like in spreadsheets or databases).
   - Examples: CSV, Excel, SQL tables.
   - Usage: Ideal for structured data with a consistent schema, such as survey responses or transaction records.
2. Time Series Data:
   - Structure: Indexed by time, often with one column for the timestamp and others for variables measured at that time.
   - Examples: Stock price data, weather data.
   - Usage: Analyzing trends and patterns over time, like forecasting.
3. Spatial Data:
   - Structure: Data associated with geographical locations (e.g., latitude and longitude).
   - Examples: GIS data files, coordinates for city locations.
   - Usage: Mapping and geographic analysis, such as visualizing population density or weather patterns across regions.
4. Text Data:
   - Structure: Unstructured or semi-structured, often in string format.
   - Examples: Documents, tweets, web pages.
   - Usage: Natural Language Processing (NLP) tasks like sentiment analysis, topic modeling, and text summarization.
5. Hierarchical Data:
   - Structure: Data arranged in a nested or tree-like format.
   - Examples: XML, JSON, and HTML files.
   - Usage: Representing complex, nested relationships, often found in configurations or multi-level data such as family trees.

2.2 Data Formats:

Data formats refer to how data is encoded and stored in files, databases, or other structures. Here are some common data formats:

{Cautions!! Data formats are not limited to list describe below, list focuses on format that is mostly structured and stored in tabular formats.}

1. CSV (Comma-Separated Values):
   - Description: Stores data in plain text with each row as a line and columns separated by commas.
   - Advantages: Simple, widely supported, human-readable.

- Limitations: Limited to tabular data, no support for complex data structures.

2. JSON (JavaScript Object Notation):
   - Description: Lightweight, text-based format for representing structured data, ideal for nested structures.
   - Advantages: Supports hierarchical data, widely used in web APIs, easy to parse in most programming languages.
   - Limitations: Can become hard to read for large or deeply nested data.
3. Excel (XLSX):
   - Description: A spreadsheet format used by Microsoft Excel, supporting both data and metadata (like formatting).
   - Advantages: Familiar, good for data visualization and formula-based manipulation.
   - Limitations: Proprietary, not ideal for large datasets or web-based applications.
4. SQL Databases (Relational Format):
   - Description: Organizes data into related tables, ideal for structured data with predefined schemas.
   - Advantages: Supports complex queries and large datasets, ACID-compliant (reliable for transactions).
   - Limitations: Fixed schema, less flexible for unstructured or hierarchical data.

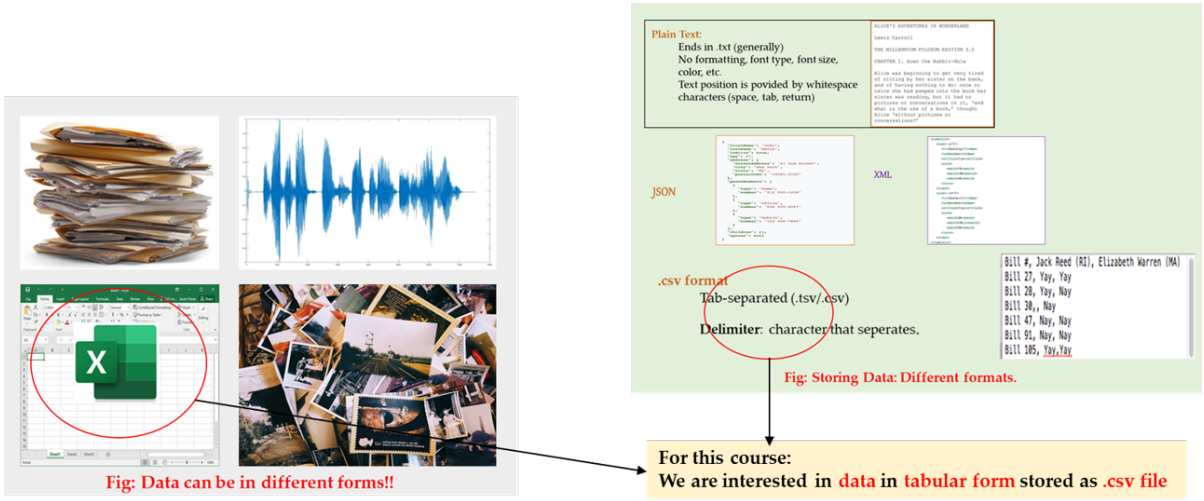


Figure 1: Data Format for this Module - {Slide - 6}.

3. Elements of a Data Science Process:

Every data science process begins with **data collection**, typically in the form of datasets. These datasets can either be gathered specifically to answer a research question, or the question itself may be adapted based on the available data.Although a comprehensive dataset covering the entire population is ideal, practical constraints mean that data scientists almost always work with a sample of the data. Consequently, the quality and

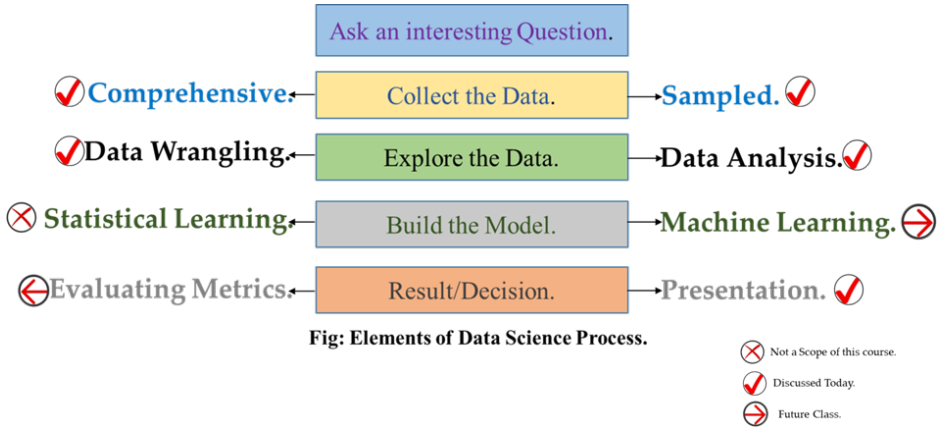


Figure 2: Elements of a Data Science Process. {Form Slide 75}

reliability of the final insights depend heavily on the sampling method used and any biases that may have been introduced in the process.{For details on biases introduced by sampling process refer to lecture slide 26 - 32}.

3.1 Collecting Data:

{For more details in Searching and Collecting data, refer to lecture slide 10 - 17}

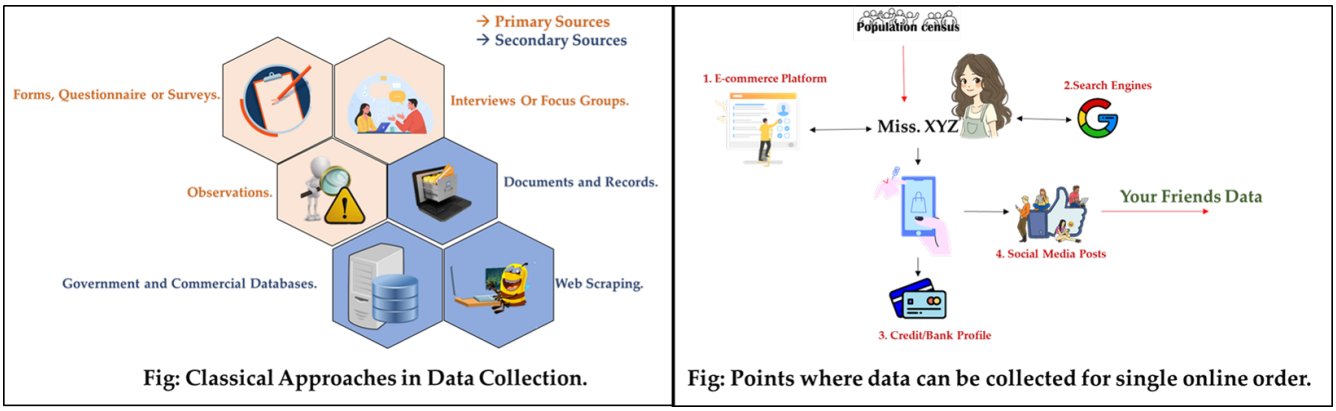


Figure 3: Techniques for Collecting Data.{Slide - 15 and 16}.

3.2 Population Vs. Sampled Data:

- Population or Comprehensive Data: A population is the complete set of all individuals, items, or data points of interest in a particular study. It includes every possible data point that could be observed. The population can be large or small, depending on the research question. For example:
  - In a study on average income in a country, the population would be all individuals in that country.
  - In a clinical trial, the population could be all patients who have a certain disease.

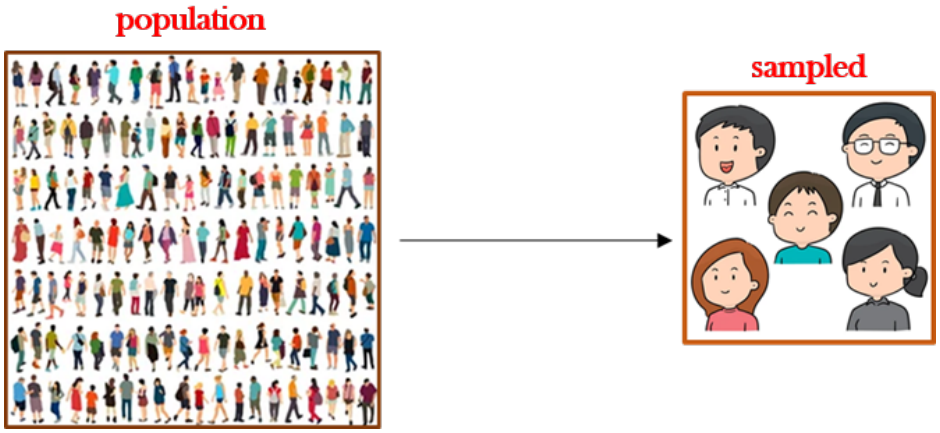


Figure 4: Population Vs. Sampled Data. {Form Slide 19}.

The main characteristic of the population is that it is complete, meaning every possible subject or data point is included.

- Sample Data: A sample is a subset of the population that is selected for analysis. Since it is often impractical to collect data from an entire population (due to time, cost, or logistical constraints), researchers collect a sample that represents the population as accurately as possible. The sample should ideally be random and representative of the population.For Example:
  - Instead of surveying every person in a city, a sample might be selected to represent different age groups, genders, and income levels to make conclusions about the population.
  - In a clinical trial, a sample of patients from the population of interest might be selected to test a new drug.

1.2 A Simple Random Sampling:

Simple Random Sampling (SRS) is a fundamental method of sampling where each member of the population has an equal chance of being selected for the sample. It is a type of probability sampling and is often considered the most straightforward way to ensure that the sample is unbiased and representative of the population. Sampling with or without Replacement:

- With Replacement: After an individual is selected, they are returned to the population and can be selected again.
- Without Replacement: Once an individual is selected, they are not returned to the population, meaning they cannot be selected again.

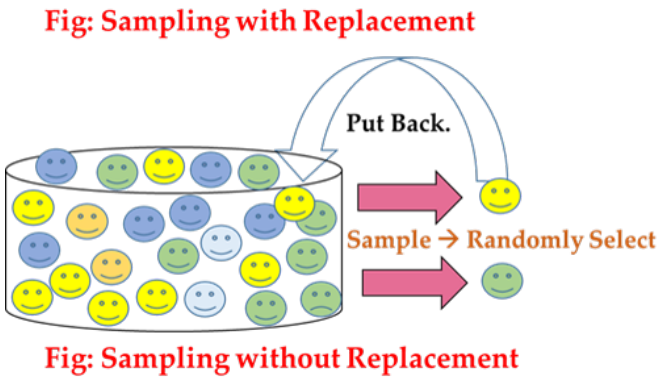


Figure 5: Sampling with and without replacements.

1. Key Characteristics:

- **Equal Probability**: Every individual or item in the population has the same likelihood of being selected.
- **Independence**: The selection of one member does not affect the selection of another. Each selection is independent of the others.
- **Randomness**: Selections are made randomly, often using tools like random number generators or lottery-like methods.

2. Process of Simple Random Sampling:

1. Define the Population: Identify the entire set of individuals or items from which the sample will be drawn. This could be a group of people, products, or observations.
2. Determine Sample Size: Decide how many individuals (or items) you want to include in the sample. This depends on factors like the research objectives, budget, and desired precision.
3. Assign Numbers to the Population: Each member of the population is assigned a unique number, typically from 1 to N, where N is the total population size.
4. Random Selection: Use a random method (e.g., random number generator, drawing lots, or random number table) to select the individuals from the population. Each selection is made independently and

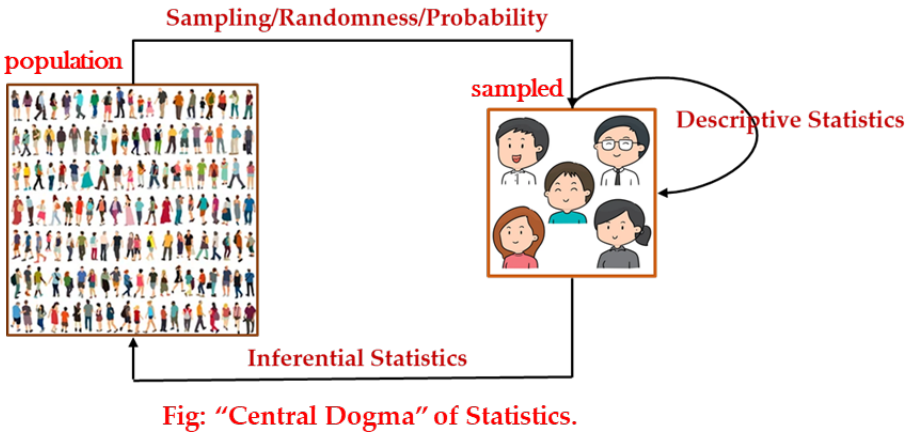


Figure 6: A statistics is: {Slide - 47}.

has the same probability.

5. Collect Data: The selected individuals make up the sample. Data is then collected from these individuals.

3. Example of Simple Random Sampling:

Imagine a study where you want to survey 100 students about their satisfaction with campus facilities. The population is all 1,000 students at the university.

1. Assign numbers: Number the students from 1 to 1000.
2. Random selection: Use a random number generator to select 100 unique numbers between 1 and 1000. The students corresponding to those numbers will form your sample.
3. Survey: Collect data from those 100 students.

4. Sampling Techniques Related to Simple Random Sampling:

- Stratified Sampling: In stratified sampling, the population is divided into subgroups (or strata) based on certain characteristics, and a simple random sample is taken from each stratum.
- {Cautions: Stratified is not the only techniques in simple random sampling, others are not in scope of this course.}

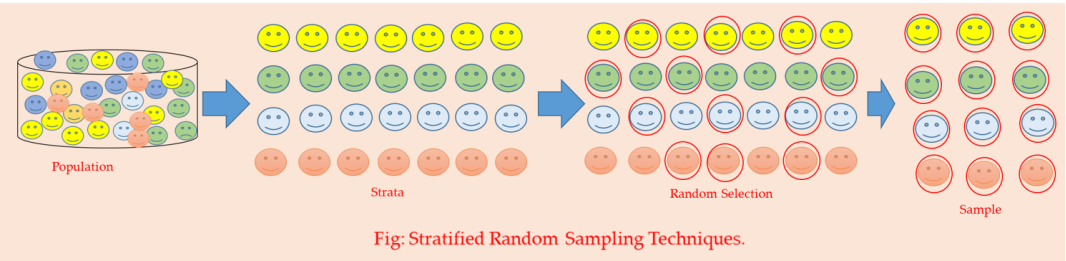


Figure 7: Stratified Random Sampling {Slide - 24}.

1.3 Exercises: About a Data.

Problem 1:

Classify each of the following attributes as either categorical or numerical. For those that are numerical, determine whether they are discrete or continuous.

1. Brand of computer purchased by a customer.
2. Province of Birth for someone born in Nepal.
3. Price of textbook.
4. Concentration of a contaminant(micrograms per cubic centimeter)in a water sample.
5. Zip Code.
6. Actual weight of coffee in a 1-lb can.
7. The length of 1 year old cat.
8. Brand of a motorcycle purchased

Problem 2:

For each of the following situations, identify the population of interest, the inferential objective, and how you might go about collecting a sample.

1. A researcher want to estimate the Average Time until Disease Recurrence among patients treated in Nepali Hospitals.

   Sample Solution

   - Population of Interest: Patients treated for the disease in Nepali hospitals.
   - Inferential Objective: To estimate the average length of time until the recurrence of the disease among treated patients.
   - Sampling Method: Track a random sample of patients treated for the disease in hospitals across Nepal. Follow up over time to record when, or if, the disease recurs, allowing for a calculation of the average time to recurrence.

2. A marketing consultant wants to estimate the average monthly advertising expenditure for e-commerce companies in Nepal.
3. A retail company wants to evaluate customer satisfaction with a recently launched product line to inform future product development.
4. A bank wants to estimate the proportion of its customers who are interested in transitioning to digital banking services.
5. A company wants to determine whether a majority of its employees are satisfied with the health benefits provided.

2 Exploratory Data Analysis - Making Data Usable.

Data Exploration is the very first and most important stage in any data science process or project. It is a process where we examine datasets to summarize their main characteristics and gain insights before applying more complex modeling techniques.In general exploratory data analysis consist of two primary task:

- Data Wrangling: The act of tidying and cleaning data.
- Data Analysis: The act of making sense of data.

Figure 8: Primary Tasks in Exploratory Data Analysis: Data Wrangling and Data Analysis {slide - 35}.

Data Wrangling:

{Cautions!!! The word data wrangling can be defined differently dependent on the disciplinary context in which it is applied to.For the purpose of this module and in general process of data science I like to define data wrangling as the step of cleaning and tidying dataset.}

The process of transforming raw data into data that can be consolidated for it's analytical use case.In general,this constitutes of all the pre-processing steps we applied to a raw data but can be grouped in following:

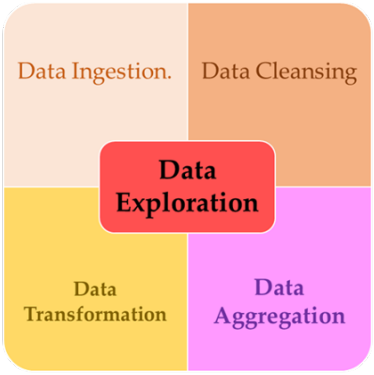


Fig: Major Steps in Data Wrangling Process

Figure 9: What we do in to Tidy Our Data {slide - 36 - 40}.

{Disclaimer!!!Data wrangling is mostly carried out with computer using programming language. In this module we will perform data wrangling in python with pandas framework.Thus how to perform data wrangling will be discussed in depth in your workshop session.}

Data Analysis:

In this step we mostly used statistics to make sense out of data.This module will discuss one branch of statistics called descriptive statistics and how it can be used to further explore and describe data.This is covered along with pen and paper exercise through out the upcoming sections.

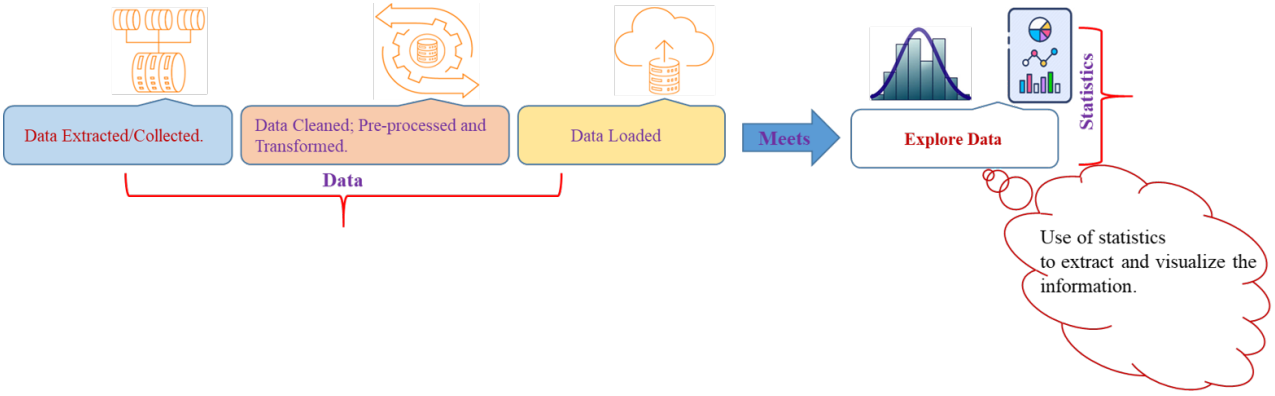


Figure 10: When Data Meets Statistics!!! {slide - 43}.

3 Describing Data with Statistics.

1. About a Statistics:

Statistics is a science whose focus is on collecting, analyzing and drawing conclusion from data.

- "The objective of statistics is to make an inference about a population based on information contained in a sample from that population and to provide an associated measure of goodness for the inference." from Mathematical Statistics by Wackerly et. al

Types of statistics analysis:

- Univariate Statistical Analysis: Analysis made only on one variable per subject.
- Bivariate Statistical Analysis: Analysis made among two variables per subject.
- Multivariate Statistical Analysis: Analysis made among more than two variables per subject.

Introduction to Descriptive Statistics:

Descriptive statistics is a branch of statistics focused on summarizing, organizing, and presenting data in a way that makes it easier to understand and interpret. This can be achieved through numerical, graphical, or tabular methods.Descriptive statistics is a univariate statistical analysis technique, meaning it focuses on only one variable per subject. Descriptive statistics can be further subdivided into:

1. Graphical Methods:
2. Numerical Methods:

3.1 Numerical Methods for Describing Data:

In this section, we demonstrate how to calculate numerical summary measures that describe the center and the extent of spread in a dataset. We will discuss several techniques for using these summary measures to characterize various aspects of the data.

3.1.1 Measure of Central Tendency:

The central tendency of a distribution is measure of the middle of the distribution.

About a Distribution!!!

Distribution is a Collection of all the observation/ data-points for a particular variable.
For example - Height Distribution:
The distribution of measured height of sample students at Herald College could be represented as X such that $X \in \{x_1, x_2, \dots x_n\}$ n is the total data points or observation in X and x_1 are the individual observations.

Limitations of Measurement of Central Tendency:

The central tendency of a distribution of data gives an idea of the midpoint of the distribution:

- but doesn't tell you anything about the spread of a distribution,
- or the shape of a distribution,
- or how concentrated the distribution is in different places.

Techniques of Measuring center:

There are several ways to define the middle, center or typical value of distribution or data.Each measure is a different way to tell the story of the center aspect of a distribution. The three most common measures for central tendency are:

1. The Mean.
2. The Median.
3. The Mode.

The Mean:

The mean or arithmetic mean is an average of all the observations.

Simple Arithmetic Mean:

The sample mean of a sample of numerical observations:

$$x_1, x_2, \ldots, x_n$$

denoted by $\bar{x}$ is:

$$\begin{aligned} \bar{x} &= \frac{\texttt{sum of all observations in sample}}{\texttt{number of observations in sample.}} \\ &= \frac{x_1 + x_2 + \cdots + x_n}{n} \\ &= \frac{\sum_{i=1}^{n} x_i}{n} \end{aligned} \tag{1}$$

The **population mean** denoted by μ, is the average of all x values in the entire population.

Weigthed Average:

Unlike a simple arithmetic average, where each value is treated equally, a weighted average assigns different weights to different values, allowing some values to have more influence on the final result. Formula for weighted average:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i \cdot x_i}{\sum_{i=1}^{n} w_i} \tag{2}$$

The Median:

The median is a measure of central tendency that represents the middle value in a dataset when the values are ordered from least to greatest (or vice versa). It is especially useful when the data contains outliers or is skewed, as it is not affected by extremely large or small values, unlike the mean.

How to find the Median:

1. Arrange the data by ordering the n observations from smallest to largest.{with any repeated values included.}.
2. For odd number of data points:
   If the dataset has n data points and where n is odd, then the **median** is the value positioned at:

$$\texttt{Median} = (x_{\frac{n+1}{2}})^{th} \quad \texttt{position}$$

3. For even number of data points:
   If the dataset has n data points and where n is even, then the **median** is the value positioned at:

$$\texttt{Median} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

Key Points about the Median:

- The median divides the dataset into two equal halves.
- It is not influenced by extreme values (outliers), making it a better measure of central tendency when the data is skewed or contains outliers.
- It is commonly used in statistics when dealing with income distributions, house prices, and other datasets that may have significant skewness.

The Mode:

The mode is a measure of central tendency that identifies the most frequently occurring value(s) in a dataset. Unlike the mean and median, which focus on central location, the mode reflects the most common or frequent observation.

How to find the Mode:

There is no specific formula for calculating the mode, as it simply refers to the value(s) that appear most frequently in the dataset.

1. List the frequency of each value in the dataset.
2. Identify the value(s) with the highest frequency.

The mode is often useful for categorical data or datasets where the most frequent occurrence is of interest.

Key Points about the Mode:

- Unimodal: If there is only one value that occurs most frequently.

- Bimodal: If there are two values that occur with the same highest frequency.
- Multimodal: If there are more than two values with the same highest frequency.
- No Mode: If no value repeats, meaning each value occurs only once.

The mean Vs The median Vs The Mode:

- For symmetric distributions, mean = median.
- For skewed distributions, mean is drawn in direction of longer tail, relative to median.
- Mean sensitive to "outliers" (median often preferred for highly skewed distributions).
- When distribution symmetric or mildly skewed or discrete with few values, mean preferred because uses numerical values of observations.

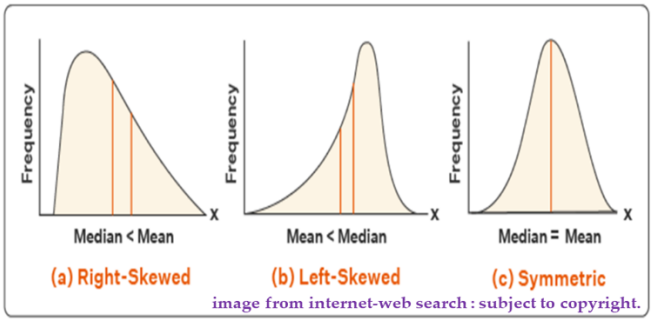


Figure 11: Measure of Central Tendency and Skewness.

3.1.2 Exercises - Mean, Median, Mode:

1. Prove that the sum of the deviations of a set of measurements about their mean is equal to zero:

$$\sum_{1=1}^{n}(y-\bar{y})=0$$

The Proof

To prove that the sum of the deviations of a set of measurements about their mean is equal to zero, we start with a set of n measurements, denoted $y_1, y_2, \ldots, y_n$, and let $\bar{y}$ represent their mean. The mean $\bar{y}$ is defined as:

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

We want to show that:

$$\sum_{i=1}^{n}(y_i-\bar{y})=0$$

Proof:
Begin with the sum of the deviations of each y_i from the mean $\bar{y}$:

$$\sum_{i=1}^{n}(y_i - \bar{y})$$

Using the properties of summation, we can expand this as:

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \bar{y}$$

Notice that $\bar{y}$ is a constant (the mean), so we can rewrite $\sum_{i=1}^{n} \bar{y}$ as $n \cdot \bar{y}$:

$$= \sum_{i=1}^{n} y_i - n \cdot \bar{y}$$

Substitute the definition of $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$:

$$= \sum_{i=1}^{n} y_i - n \cdot \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$$

Simplify by distributing $n \cdot \frac{1}{n} = 1$:

$$= \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} y_i$$

Finally, subtracting the same quantity from itself gives:

$$= 0$$

Thus, we have shown that:

$$\sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

2. The lengths of time (in minutes) that ten patients at a doctor's clinic wait to see their doctor are as follows: 5, 17, 8, 2, 55, 9, 22, 11, 16, 5.
   (a) What are the mean, median and mode for these datasets?
3. Given a provided collection D with a mean of 14, find the missing value of y.

$$\mathbf{D} = [\mathbf{12}, \mathbf{14}, \mathbf{10}, \mathbf{15}, \mathbf{10}, \mathbf{10}, \mathbf{y}, \mathbf{14}, \mathbf{13}, \mathbf{11}, \mathbf{10}, \mathbf{10}]$$

4. Find the mean of 30 given numbers, when it is given that the mean of 10 of them is 12 and the mean of the remaining 20 is 9.[**10**]

Solution:

To find the mean of all 30 numbers, we can use the given information about the means of the two groups.

Given Information:

- The mean of the first 10 numbers is 12.
- The mean of the remaining 20 numbers is 9.

Steps:

(a) Calculate the total sum of the first 10 numbers:

$$\text{Sum of first 10 numbers} = 10 \times 12 = 120$$

(b) Calculate the total sum of the remaining 20 numbers:

$$\text{Sum of remaining 20 numbers} = 20 \times 9 = 180$$

(c) Add these two sums to get the total sum of all 30 numbers:

$$\text{Total sum of all 30 numbers} = 120 + 180 = 300$$

(d) Calculate the mean of all 30 numbers:

$$\text{Mean of all 30 numbers} = \frac{\text{Total sum of all 30 numbers}}{30} = \frac{300}{30} = 10$$

Answer:

The mean of all 30 numbers is 10.

5. A teacher wants to calculate the overall performance of a student who scored 70 in the midterm (weighted 40%) and 90 in the final exam (weighted 60%). Calculate the weighted mean of the student's scores.[**82**]
6. Find out the median from the following data:

Daily wages(Rs)	5	7	8	10	11
No.of Workers	20	15	12	15	18

Solution:

Steps:

(a) calculate the Cumulative Frequency:

Daily wages (Rs)	No. of Workers	Cumulative Frequency
5	20	20
7	15	$20 + 15 = 35$
8	12	$35 + 12 = 47$
10	15	$47 + 15 = 62$
11	18	$62 + 18 = 80$

(b) Determine the Median Class:
The total number of workers is $N = 80$. The median is located at the $\frac{N}{2}$-th value, which is:

$$\frac{N}{2} = \frac{80}{2} = 40$$

Looking at the cumulative frequency, the median class is the one where the cumulative frequency just exceeds 40. This occurs in the 8 wage category.

(c) Apply the Median Formula:
The formula for the median in a frequency distribution is:

$$\text{Median} = L + \left(\frac{\frac{N}{2} - F}{f} \right) \times h$$

where:

- L is the lower boundary of the median class (7.5 for the class 8),
- F is the cumulative frequency of the class before the median class (35),
- f is the frequency of the median class (12),
- h is the class width (assuming a width of 1).

Substituting in the values:

$$\text{Median} = 7.5 + \left(\frac{40 - 35}{12} \right) \times 1$$

$$= 7.5 + \frac{5}{12}$$

$$= 7.5 + 0.4167 = 7.9167$$

The median daily wage is approximately 7.92 Rs.

7. In a class of 100 students, the average amount of pocket money is Rs 35 per student, if the average is Rs 25 for girls and Rs 50 for boys, then the number of girls in the class will be?[**60**].

For Sample

In a class of 100 students, the average amount of pocket money is Rs 35 per student. Given:

- The average pocket money for girls is Rs 25.
- The average pocket money for boys is Rs 50.

Let:

- x be the number of girls.
- Then, the number of boys will be $100 - x$ (since the total number of students is 100).

Steps:

(a) Calculate the total pocket money: Since the average pocket money for the class is Rs 35 per student, the total pocket money for all students is:

$$\text{Total pocket money} = 100 \times 35 = 3500$$

(b) Set up the equation based on the pocket money for girls and boys: The total pocket money can also be calculated as the sum of the total pocket money for girls and boys.
So:

$$\boxed{\begin{aligned}\text{Total pocket money} = &(\text{Pocket money per girl} \times \text{Number of girls}) \\ &+ (\text{Pocket money per boy} \times \text{Number of boys})\end{aligned}}$$

Substituting the values, we get:

$$3500 = (25 \times x) + (50 \times (100 - x))$$

(c) Expand and solve for x:

$$3500 = 25x + 50 \times 100 - 50x$$

$$3500 = 25x + 5000 - 50x$$

Combine like terms:

$$3500 = 5000 - 25x$$

Rearranging to isolate x:

$$25x = 5000 - 3500$$

$$25x = 1500$$

$$x = \frac{1500}{25} = 60$$

The number of girls in the class is **60**.

8. A sample consisting of four pieces of luggage was selected from among those checked at an airline counter, yielding the following data on `x = weight in pounds`: Suppose that one more piece is selected: denote

Selected Baggage	x_1	x_2	x_3	x_4
Weights	33.5	27.3	36.7	30.5

its wight by x_5.Find a value of x_5, such that:

$$\bar{x} = \texttt{sample median.}$$

[**32**]

9. An experiment to study the lifetime (in hours) for a certain type of component involved putting 10 components into operation and observing them for 100 hr.Eight of the components failed during that period, and those lifetimes were recorded. The lifetime of the two components still functioning after 100 hr are recorded as $100+$. The resulting sample observations were:

$$[\mathbf{48}, \mathbf{79}, \mathbf{100}+, \mathbf{35}, \mathbf{92}, \mathbf{86}, \mathbf{57}, \mathbf{100}+, \mathbf{17}, \mathbf{29}]$$

   Which of the measures of central tendency discussed in this section can be calculated, and what are the values of those measures?

10. An instructor has graded 19 exam papers submitted by students in a class of 20 students, and the average so far is 70.)(The maximum possible score is 100.)How high would the score on the last paper have to be to raise the class average by 1 point?

3.1.3 Measure of Spread - Describing the variability in a Dataset:

Reporting a measure of center provides only partial information about a dataset. It is also essential to describe the spread of values around the center. This section explores the following techniques used to measure the spread of a dataset:

1. Range and Inter-quartile Range{IQR}.
2. Variance and Standard Deviation.
3. Coefficient of Variance.

1. The Range:

Definition: The Range is the simplest measure of spread. It is calculated as the difference between the largest and smallest values in a dataset, providing a sense of how spread out the data is.

Formula - Range:

$$\texttt{Range = largest observation - lowest observation.} \tag{1}$$

Example: Determine the Range for the dataset:

$$[\mathbf{3}, \mathbf{7}, \mathbf{2}, \mathbf{9}, \mathbf{5}]$$

$$\texttt{Largest Observation} = 9$$

.

$$\texttt{Lowest Observation} = 2$$

$$\texttt{Range} = 9 - 2 = 7 \quad \square$$

Limitations: The range is highly sensitive to outliers or extreme values. Even a single extreme value can significantly impact the range.

2. Interquartile Range(IQR):

Definition: The inter-quartile range (IQR) is a measure of statistical dispersion, which is resistant to the effects of outliers. It is defined as the range between the first quartile Q_1 and the third quartile Q_3 of a dataset.

About a Quartile: Quartiles break down the data into specific percentiles.

- Lower quartile (Q_1): This is the 25th percentile of the data. It separates the bottom 25% of the data from the top 75%.
- Upper quartile (Q_3): This is the 75th percentile of the data. It separates the top 25% of the data from the bottom 75%.
- Middle quartile (Median or (Q_2): This is the median, which divides the dataset into two halves, separating the bottom 50% from the top 50%.

Formula - IQR:

The inter-quartile range (IQR) itself is given by the difference between the third and first quartiles:

$$\texttt{IQR} = Q_3 - Q_1 \tag{1}$$

Steps to calculate IQR:

- Sort the data in ascending order.
- Find the median of the dataset(Q_2).
- Find(Q_1 and (Q_3:
  - Q_1 is the median of the lower half of the data.
  - Q_3 is the median of the upper of the data.
- Subtract Q_1 from Q_3 to get the IQR.

Example - IQR:

Determine the IQR for the dataset:

$$[\mathbf{3}, \mathbf{7}, \mathbf{2}, \mathbf{9}, \mathbf{5}, \mathbf{8}, \mathbf{6}, \mathbf{10}, \mathbf{4}, \mathbf{11}]$$

Solution:

1. Sort the data:
$$[\mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{10}, \mathbf{11}]$$
2. Find the median(Q_2): The middle values are 6 and 7, so:
$$Q_2 = \frac{6+7}{2} = 6.5$$
3. Find Q_1 i.e. median of the lower half:
$$[\mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}] \quad \text{so,} \quad Q_1 = 4 \quad \square$$
4. Find Q_3 i.e. median of the upper half:
$$[\mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{10}, \mathbf{11}] \quad \text{so,} \quad Q_3 = 9 \quad \square$$
5. Calculate IQR:
$$\texttt{IQR} = Q_3 - Q_1 = 9 - 4 = 5$$

Interpretation: The IQR indicates that the middle 50% of the data values lie with in a range of 5 units.

3. Outliers:

Definition: An outlier is a data point that is significantly different from the rest of the data. Outliers can be much higher or lower than the other observations, and they can skew the results of statistical analysis.

Identifying Outlier Using IQR:

One common method for detecting outliers is based on the IQR. A data point is considered an outlier if it is:

$$\texttt{Lower Bound} = Q_1 - 1.5 \times IQR$$
$$\texttt{Upper Bound} = Q_3 + 1.5 \times IQR$$

If a data point lies below the lower bound or above the upper bound, it is considered an outlier.

Example - Identifying Outlier with IQR:

Using the earlier dataset $[\mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{5}, \mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{10}, \mathbf{11}]$ where $Q_1 = 4$, $Q_3 = 9$ and IQR $= 5$. Identify the Outliers.
Solution:

$$\texttt{Lower Bound} = 4 - 1.5 \times 5$$
$$\texttt{Upper Bound} = 9 + 1.5 \times 5$$

Any value less than -3.5 or greater than 16.5 would be an outlier. In this case, all values in the dataset are within the bounds, so there are no outliers.

4. Variance:

Definition: Variance measures the spread of a set of data points around their mean. It is the average of the squared deviations from the mean, and it provides an idea of how much the data varies.
Key Points:

- Variance tells you how far each data point is from the mean on average, but it does so in squared units, which makes it less interpretable in the same units as the data itself.
- The larger the variance, the more spread out the data is.
- The smaller the variance, the more concentrated the data is around the mean.

Formula for Variance:

- Population Variance denoted as σ^2 and calculated as:

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

  Where:
  - N is the number of data points in the population.

  - x_i is each individual data point.
  - μ is the population mean.
- Sample variance denoted as s^2 and calculated as:

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

Where:
  - n is the number of data points in the sample.
  - x_i is each individual data point.
  - $\bar{x}$ is the sample mean.

Example - Sample Variance:

Consider the sample $[4, 8, 10]$.
The mean $\bar{x}$ is:

$$\bar{x} = \frac{4+8+10}{3} = 7.33$$

Squared Deviation from mean:

Data Point	Squared Deviation from Mean	Value
4	$(4-7.33)^2 = (-3.33)^2$	11.09
8	$(8-7.33)^2 = (0.67)^2$	0.45
10	$(10-7.33)^2 = (2.67)^2$	7.13
Sum of Squared Deviations		**18.66**

Finally, the sample variance s^2 is calculated using $n-1$:

$$s^2 = \frac{\text{Sum of squared deviations}}{n-1} = \frac{18.66}{3-1} = \frac{18.66}{2} = 9.33$$

5. Standard Deviation:

Definition: Standard deviation is the square root of the variance. It provides a measure of the spread of data in the original units of the data, which makes it more interpretable than variance.
Key points:

- Standard deviation is the average distance of the data points from the mean, but unlike variance, it is expressed in the same units as the original data.
- A high standard deviation means the data points are spread out over a wide range of values.
- A low standard deviation means the data points are clustered closely around the mean.

Formula for Standard Deviation:

- Population Standard Deviation denoted as σ and calculated as:

$$\sigma = \sqrt{\frac{1}{N}(\sum_{i=1}^{N}(x_i - \mu)^2} \tag{3}$$

  Where:
  - N is the number of data points in the population.
  - x_i is each individual data point.
  - μ is the population mean.
- Sample variance denoted as s and calculated as:

$$S = \sqrt{\frac{1}{n-1}(\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{4}$$

  Where:
  - n is the number of data points in the sample.
  - x_i is each individual data point.
  - $\bar{x}$ is the sample mean.

Example - Sample Standard Deviation:

From above example where variance was:

$$S^2 = 9.33$$

$$S = \sqrt{9.33} = 3.06$$

Key Takeaway from Variance and Standard Deviation:

- Variance measures how data points differ from the mean but in squared units, making it less directly interpretable in real-world terms.
- Standard Deviation is the square root of variance, and it is more interpretable because it has the same units as the data, giving a direct sense of spread.
- Both measures are essential for understanding how spread out or dispersed a dataset is.
- Sample variance uses $(n-1)$ to correct for the bias introduced by using the sample mean, while population variance uses n.

Why $n-1$? {Optional}

The use of $(n-1)$in the formula for sample variance (instead of n)is known as Bessel's correction, and it is used to make the variance an unbiased estimator of the population variance. Here's the intuition behind it:

- Degrees of Freedom:
  - The concept of degrees of freedom helps explain this adjustment. When we calculate the sample

mean, we lose one degree of freedom because the sample mean is a fixed value that depends on all the data points.

  - After calculating the mean, only $n-1$ values in the dataset are free to vary; the last one is constrained by the others (since the sum of deviations from the mean must be zero). This means that the variance should be calculated based on $(n1)$ independent pieces of information, not n.

- Unbiased Estimator:
  - Dividing by $(n1)$ corrects for the systematic underestimation of the variance that occurs when we use the sample mean. It ensures that the sample variance is an unbiased estimator of the population variance. That means, on average, the sample variance will be equal to the population variance when using $(n1)$.

Example - Why $(n-1)$?

Going back to original example i.e. with sample dataset:

$$[\mathbf{4}, \mathbf{8}, \mathbf{10}]$$

We calculated sample variance as:

$$\sigma^2 = \frac{\texttt{sum of squared deviation}}{n-1} = 9.33 \quad \square$$

Instead of $n-1$ let's use n i.e. 3:

$$\sigma^2 = \frac{\texttt{sum of squared deviation}}{n} = \frac{18.66}{3} = 6.22 \quad \square$$

Which value better represents the variance for given dataset?

6. Coefficient of Variation (CV):

Definition: The Coefficient of Variation (CV) is a relative measure of the spread of data, defined as the ratio of the standard deviation to the mean, often expressed as a percentage. It is useful for comparing the variability of datasets that have different units or different means.

Key points:

- The coefficient of variation measures the relative variability of the data. It tells us how large the standard deviation is in relation to the mean.
- It is especially useful for comparing the variability of datasets with different units of measurement or vastly different scales.
- A higher CV indicates greater variability relative to the mean, while a lower CV indicates less variability.

Formula for Coefficient of Variation:

- For Population:

$$\texttt{CV} = \frac{\sigma}{\mu} \times 100 \tag{5}$$

Where:
- σ is the population standard deviation.
- μ is the mean of the dataset.

- For Sample:

$$\texttt{CV} = \frac{S}{\bar{x}} \times 100 \tag{6}$$

Where:
- § is the sample standard deviation.
- $\bar{x}$ is the sample mean of the dataset.

Example - Coefficient of Variation:

Consider two datasets:
- Dataset 1: $[10, 12, 14, 16, 18]$
- Dataset 2: $[100, 200, 300, 400, 500]$

Let's calculate the Coefficient of Variation (CV) for both datasets.

- For Dataset - 1 $\Rightarrow X$:
  - Mean, $\bar{X}$:
$$\bar{X} = \frac{10 + 12 + 14 + 16 + 18}{5} = 14$$
  - Standard Deviation S:
$$S = 3.162$$
  - Coefficient of Variation (CV):
$$\texttt{CV} = \frac{S}{\bar{x}} \times 100 = 22.6\%$$
- For Dataset - 2 $\Rightarrow Y$:
  - Mean, $\bar{y}$:
$$\bar{y} = \frac{10 + 12 + 14 + 16 + 18}{5} = 14$$
  - Standard Deviation S:
$$S = 158.11$$
  - Coefficient of Variation (CV):
$$\texttt{CV} = \frac{S}{\bar{x}} \times 100 = 52.7\%$$
- Interpretation:
  - Dataset 1 has a lower coefficient of variation (22.6%), indicating less relative variability.
  - Dataset 2 has a higher coefficient of variation (52.7%), indicating more relative variability, even though the actual values are much larger.

Key Takeaways:

- Coefficient of Variation (CV) is a dimensionless measure of variability, allowing comparison of the relative spread of data across different datasets or different units.
- CV is particularly useful when comparing datasets that have different means or units, as it expresses the standard deviation as a percentage of the mean.
- A higher CV indicates a greater level of variation relative to the mean, and a lower CV indicates less variation.
- It is commonly used in fields like finance, where comparing the relative volatility of different assets is essential, or in quality control where the precision of processes is being compared.

3.1.4 Exercises - Variance, Standard Deviation and CV:

Problem - 1 - Range , IQR and Outliers:

Complete all the Tasks:

- For the Following Sample Data: Find Range, IQR, and Outliers if exist.
  1. data: $[\mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{7}, \mathbf{12}, \mathbf{15}, \mathbf{8}, \mathbf{20}]$
  2. data: $[\mathbf{2}, \mathbf{3}, \mathbf{5}, \mathbf{5}, \mathbf{5}, \mathbf{7}, \mathbf{14}, \mathbf{18}, \mathbf{6}, \mathbf{25}]$
  3. Frequency Table:

X	15-20	20-25	25-30	30-35
f	8	21	15	4

  4. data: $[\mathbf{56}, \mathbf{14}, \mathbf{84}, \mathbf{21}, \mathbf{85}, \mathbf{2}, \mathbf{35}, \mathbf{74}, \mathbf{66}, \mathbf{52}, \mathbf{45}]$.
  5. data: $[\mathbf{1}, \mathbf{3}, \mathbf{4}, \mathbf{6}, \mathbf{13}, \mathbf{20}, \mathbf{25}, \mathbf{26}, \mathbf{28}, \mathbf{62}, \mathbf{95}]$.

Sample Solutions - 3.

- **Frequency Table:**

X	15-20	20-25	25-30	30-35
f	8	21	15	4

  Solutions:
- **Range:** The **Range** is the difference between the maximum and minimum observations in the dataset. In this table, the lowest interval is 15-20 and the highest is 30-35. Therefore:

$$\text{Range} = 35 - 15 = 20$$

- **Interquartile Range (IQR):**
  To find the IQR, we need the first quartile (Q1) and the third quartile (Q3). We can estimate these values for grouped data using cumulative frequency.

Step 1: Calculate Cumulative Frequencies

Class Interval	f	Cumulative Frequency
$15-20$	8	8
$20-25$	21	$8+21=29$
$25-30$	15	$29+15=44$
$30-35$	4	$44+4=48$

Step 2: Find Q1 and Q3 Positions

Since there are 48 total frequencies, we find the positions of Q1 and Q3:

$$Q1 \text{ position} = \frac{48}{4} = 12$$

$$Q3 \text{ position} = \frac{3 \times 48}{4} = 36$$

Step 3: Locate Q1 and Q3 within Intervals

* **Q1** falls at the 12th observation, which lies in the 20-25 class interval.
* **Q3** falls at the 36th observation, which lies in the 25-30 class interval.

Using linear interpolation within each interval to find Q1 and Q3:

For Q1 (in the 20-25 interval):

* Start of interval = 20
* Frequency in this interval = 21
* Cumulative frequency before this interval = 8
* $Q1$ falls at position 12, so the number of observations into this interval is $12-8=4$.

Using the formula for linear interpolation:

$$Q1 = 20 + \frac{4}{21} \times 5 \approx 20.95$$

For Q3 (in the 25-30 interval):

* Start of interval = 25
* Frequency in this interval = 15
* Cumulative frequency before this interval = 29
* $Q3$ falls at position 36, so the number of observations into this interval is $36-29=7$.

Using linear interpolation:

$$Q3 = 25 + \frac{7}{15} \times 5 \approx 27.33$$

Step 4: Calculate IQR

$$\text{IQR} = Q3 - Q1 = 27.33 - 20.95 \approx 6.38$$

Problem - 2:

For the following set of sample data compute sample variance and Standard deviation:

1. data:
$$\mathbf{D} = [\mathbf{27}, \mathbf{41}, \mathbf{23}, \mathbf{56}, \mathbf{76}, \mathbf{54}, \mathbf{53}, \mathbf{49}, \mathbf{50}, \mathbf{92}, \mathbf{47}, \mathbf{23}, \mathbf{56}, \mathbf{65}, \mathbf{71}, \mathbf{73}, \mathbf{76}, \mathbf{77}]$$
2. data:
$$\mathbf{T} = [\mathbf{92}, \mathbf{95}, \mathbf{85}, \mathbf{80}, \mathbf{75}, \mathbf{50}]$$
3. data:
$$\mathbf{A} = [\mathbf{18}, \mathbf{22}, \mathbf{19}, \mathbf{25}, \mathbf{12}]$$
4. data:
$$\mathbf{C} = [\mathbf{112}, \mathbf{100}, \mathbf{127}, \mathbf{120}, \mathbf{134}, \mathbf{118}, \mathbf{105}, \mathbf{110}]$$
5. data:
$$\mathbf{W} = [\mathbf{7.7}, \mathbf{7.4}, \mathbf{7.3}, \mathbf{7.9}]$$
6. data:
$$\mathbf{O} = [\mathbf{18}, \mathbf{22}, \mathbf{19}, \mathbf{25}, \mathbf{12}]$$

Answer key:

Q:	1	2	3	4	5
Var:	≈ 385.58	≈ 263.5	≈ 23.7	≈ 127.6	≈ 0.076
SD:	≈ 19.64	≈ 16.23	≈ 4.87	≈ 11.3	≈ 0.275

Sample Solutions - 6.

To compute the sample variance and standard deviation for the dataset

$$[18, 22, 19, 25, 12]$$

, we follow these steps:

1. Find the Mean: The mean $\bar{O}$ is:
$$\bar{O} = \frac{\sum X}{n} = \frac{18 + 22 + 19 + 25 + 12}{5} = \frac{96}{5} = 19.2$$
2. Compute the Squared Differences from the Mean:

X_i	$X_i - \mu$	$(X_i - \mu)^2$
18	$18 - 19.2 = -1.2$	$(-1.2)^2 = 1.44$
22	$22 - 19.2 = 2.8$	$(2.8)^2 = 7.84$
19	$19 - 19.2 = -0.2$	$(-0.2)^2 = 0.04$
25	$25 - 19.2 = 5.8$	$(5.8)^2 = 33.64$
12	$12 - 19.2 = -7.2$	$(-7.2)^2 = 51.84$

3. Calculate the Sample Variance: The sample variance s^2 is the average of the squared differences,

adjusted for the sample size:

$$s^2 = \frac{\sum(X_i - \mu)^2}{n-1} = \frac{1.44 + 7.84 + 0.04 + 33.64 + 51.84}{4} = \frac{94.8}{4} = 23.7$$

4. Compute the Sample Standard Deviation: The sample standard deviation s is the square root of the sample variance:
$$s = \sqrt{23.7} \approx 4.87$$
5. Final Answer:
   - **Sample Variance** $= 23.7$
   - **Sample Standard Deviation** ≈ 4.87

Problem - 3:

Following table shows the distribution of Nepalese women in economic activities:
Find Coefficient of Variation and interpret the result.

Economic activity	1981	1991
Labour Force	34.6	40.4
Agricultural Work	36.4	45.0
Non-agricultural Work	16.6	20.3
Manufacturing	14.9	22.9
Various	46.2	45.8

Sample Solution:

To find the coefficient of variation(CV), follow these steps:

- Calculate the mean for each year:

$$\bar{\mathbf{X}}_{\mathbf{1981}} = \frac{34.6 + 36.4 + 16.6 + 14.9 + 46.2}{5} = \mathbf{30.74}$$
$$\bar{\mathbf{X}}_{\mathbf{1991}} = \frac{40.4 + 45.0 + 20.3 + 22.9 + 45.8}{5} = \mathbf{34.8}$$

- Calculate the standard deviation for each year: Formula for Standard Deviation is:

$$S = \sqrt{\frac{\sum(X_i - \bar{x})^2}{n-1}}$$

For Year - 1981:

$\mathbf{X}_i$	$\mathbf{X}_i$ - $\bar{X}_{1981}$	$(\mathbf{X}_i$ - $\bar{X}_{1991})^2$
34.6	34.6 - 30.74 = 3.86	$(3.86)^2 = 14.8996$
36.4	36.4 - 30.74 = 5.66	$(5.66)^2 = 32.0356$
16.6	16.6 - 30.74 = -14.14	$(-14.14)^2 = 199.9796$
14.9	14.9 - 30.74 = -15.84	$(-15.84)^2 = 251.7796$
46.2	46.2 - 30.74 = 15.46	$(15.46)^2 = 239.8116$
Sum		738.506

$$S_{1981} = \sqrt{\frac{738.506}{5-1}} = \sqrt{184.6265} \approx 13.58$$

For Year - 1991:

$\mathbf{X}_i$	$\mathbf{X}_i$ - $\bar{X}_{1991}$	$(\mathbf{X}_i$ - $\bar{X}_{1991})^2$
40.4	40.4 - 34.8 = 5.6	$(5.6)^2 = 31.36$
45.0	45.0 - 34.8 = 10.2	$(10.2)^2 = 104.04$
20.3	20.3 - 34.8 = -14.5	$(-14.5)^2 = 210.25$
22.9	22.9 - 34.8 = -11.9	$(-11.9)^2 = 141.61$
45.8	45.8 - 34.8 = 11.0	$(11.0)^2 = 121.00$
Sum		608.388

$$S_{1991} = \sqrt{\frac{608.388}{5-1}} = \sqrt{152.097} \approx 12.34$$

- Calculate the Coefficient of Variation (CV):
  - For 1981:
$$CV_{1981} = \frac{S_{1981}}{\bar{X}_{1981}} \times 100 = \frac{13.58}{30.74} \times 100 \approx 44.18\%$$
  - For 1991:
$$CV_{1991} = \frac{S_{1991}}{\bar{X}_{1991}} \times 100 = \frac{12.34}{34.88} \times 100 \approx 35.39\%$$
- Final Result:

Year	Mean ($\bar{X}$)	Standard Deviation (σ)	Coefficient of Variation (CV)
1981	30.74	13.58	44.18%
1991	34.88	12.34	35.39%

- Interpretation of the Result:
  - For 1981: The CV of 44.18% indicates a relatively higher variability in the distribution of Nepalese women across economic activities. This suggests that the involvement of women in different sectors was more spread out, with greater fluctuations in proportions.
  - For 1991: The CV of 35.39% shows a reduced variability compared to 1981, indicating that the distribution of women across economic activities became more consistent in 1991.

  Thus, the lower CV in 1991 compared to 1981 suggests more stability in the proportions of women involved in different economic activities by 1991.

Problem - 4:

Suppose you are analyzing the annual returns of two investment portfolios, Portfolio A and Portfolio B, over the past 5 year. The returns are as follows (in percentage):{**Hint: Use Population Variance and SD**}

$$\mathbf{PortfolioA} : [\mathbf{10}, \mathbf{8}, \mathbf{12}, \mathbf{15}, \mathbf{11}]$$

$$\mathbf{PortfolioB} : [\mathbf{5}, \mathbf{9}, \mathbf{11}, \mathbf{14}, \mathbf{10}]$$

Find the Coefficient of Variance and Interpret the result.

Problem - 5:

Two plants C and D of a factory show the following results about the number of workers and the wages paid to them.

Plant	C	D
No. of Workers	5000	6000
Average Monthly Wages ($)	2500	2500
Standard Deviation	9	10

Using coefficients of variation formulas, find in which plant, C or D is there greater variability in individual wages. Interpret the result.[$\mathbf{CV_C} \approx \mathbf{0.36}\%$, $\mathbf{CV_D} \approx \mathbf{0.40}\%$].

Problem - 6:

If the coefficient of variation of two distributions are 60 and 70, and their standard deviations are 25 and 16, respectively, find their arithmetic means.[$\mathbf{am_1 = 41.77}; \mathbf{am_2 = 22.86}$]

3.2 Graphical Methods for Describing Data:

An appropriate graphical or tabular display of data can be an effective way to summarize and communicate information. In this section we introduce some techniques for organizing and describing data using tables and graphs. {Recommendation!!For exercise in this section feel free to use any spreadsheet application like Excel or Google Sheet}

3.2.1 Displaying Categorical Data:

1. Frequency Distributions and Bar Chart:

Frequency Distributions Table

A Frequency distribution table for categorical data is a table that displays the possible categories along with associated frequencies or relative frequencies.

- Frequency: number of time particular category appear in the dataset.
- Relative Frequency: For a particular category relative frequency is the fraction or proportion of the observations resulting in the category given by:

$$\texttt{relative frequency} = \frac{\texttt{Frequency of the Category}}{\texttt{Total number of observations in the dataset}}$$

Where:

- `Frequency:` It is the count of occurrences for the specific category.
- `Total number:` It is the sum of all frequencies in the dataset.

Example - Frequency Distributions Table

Following is the Pet owned by tenants in a apartment building, present the information in frequency table.

`[Dog, Cat, Dog, Hamster, Snake, Cat, Dog, Parrot, Hamster, Dog]`

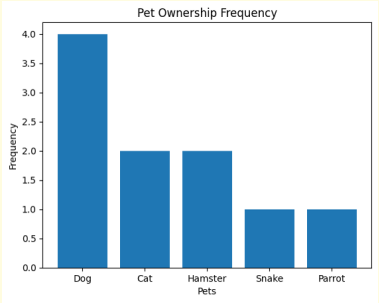
A sample frequency table may look like:

Pet	Frequency	Relative Frequency
Dog	4	$\frac{4}{10} = 0.4$
Cat	2	$\frac{2}{10} = 0.2$
Hamster	2	$\frac{2}{10} = 0.2$
Snake	1	$\frac{1}{10} = 0.1$
Parrot	1	$\frac{1}{10} = 0.1$

A Bar Chart:

- **Definition:** A bar chart is a graph of the frequency distribution of categorical data.Each category in the frequency distribution is represented by a bar or rectangle, and the graph is constructed in such a way that the area of each bar is proportional to the corresponding frequency or relative frequency.
- **When to Use?** For Categorical Data.
- **How to construct?**
  1. Collect and Organize your Data:
     - The first step in creating a bar chart is to gather your data and organize it into categories. You should have a list of categories (x-axis) and their corresponding values (y-axis).
  2. Choose the Bar Chart Type:
     - Choose the appropriate type of bar chart. For most simple data, a vertical bar chart (columns) or horizontal bar chart (bars) will work well. In a vertical bar chart:
       * he categories (such as types of pets) go on the x-axis.
       * The values (frequencies) go on the y-axis.

       For a horizontal bar chart, it's just the opposite: categories are on the y-axis and values on the x-axis.
  3. Plot the Bar Chart: Plot the bars based on your data. Each bar's height (for a vertical bar chart) or width (for a horizontal one) will represent the value of the corresponding category.



Grouped Frequency Table - {Optional}:

- **What is Grouped Frequency Table?**
  - A grouped frequency table organizes data into intervals (called "classes") and records the frequency, or count, of data points within each interval.
- **Why Grouped Frequency Table?**
  1. This method is useful for summarizing large datasets or continuous data, as it allows us to see patterns and distributions more clearly by grouping similar values together.
  2. Grouped frequency tables help simplify complex data, making it easier to analyze trends across ranges instead of individual values.
- **Constructing Grouped Frequency Table:**To create a grouped frequency table, we first need to decide on the class intervals. Determining an appropriate class interval involves considering the range of the data and the number of data points. The goal is to create intervals that:
  1. Capture the variability in the data

2. Maintain a reasonable number of classes for clarity in presentation.

There are various ways to determine the class intervals, In this section we will discuss **Square Root Rule** due to its simplicity.

- Square Root Rule: The Square Root Rule is often used to determine the number of classes and class width.
  1. **Number of classes:**
     - Purpose: To determine the appropriate number of intervals (or classes) to group data effectively.
     - Method: Approximate the number of classes by taking the square root of the total number of data points.
     - Formula:
     $$\texttt{Number of Classes} \approx \sqrt{\texttt{Total Number of Data Points}}$$
  2. **Range:**
     - Purpose: The range provides the spread of the data by calculating the difference between the maximum and minimum values.
     - Formula:
     $$\texttt{Range} = \texttt{Maximum value - Minimum Value}$$
  3. **Class Width:**
     - Purpose: To determine the size of each interval (or class) in the grouped frequency table.
     - Method: Divide the range by the number of classes, then round to the nearest whole number for simplicity.
     - Formula:
     $$\texttt{Class Width} = \frac{\texttt{Range}}{\texttt{Number of Classes}}$$
     - Inclusion and Exclusion: Class intervals are typically inclusive at the lower boundary and exclusive at the upper boundary (e.g., 60–69 includes 60 but not 69).
     - Dealing with Decimals: If the calculated class width is a decimal (e.g., 3.6), you generally round to the nearest whole number (e.g., 4). However, if you need more precision or the dataset contains precise decimal values, you might keep one or two decimal places in the class width to avoid losing detail.

Example - Grouped Frequency Table:

Suppose we have the following test scores (out of 100) for a class of 40 students:

$$72, 88, 62, 95, 78, 90, 85, 68, 77, 82, 93, 75, 88, 72, 98, 65, 88, 72, 85, 92, 80, 78, 88, 69, 77, 84$$
$$75, 90, 68, 82, 76, 88, 73, 81, 89, 70, 78, 87, 94, 79$$

Create a grouped frequency table with class intervals and frequencies.
Solution:

1. Determine the Number of Classes:

$$\texttt{Number of Classes} = \sqrt{100} = 10 \quad \square$$

2. Determine the Range:

$$\texttt{Range} = \texttt{Maximum Value} - \texttt{Minimum Value} = 98 - 62 = 36 \quad \square$$

3. Determine the Class Width:

$$\texttt{Class Width} = \frac{\texttt{Range}}{\texttt{Number of Classes}} = \frac{36}{10} = 3.6 \approx 4 \quad \square$$

4. Construct the Grouped Frequency Table:

Grouped Frequency Table:

Class Interval	Frequency
60 - 69	6
70 - 79	10
80 - 89	12
90 - 99	12

2. A Pie Chart:

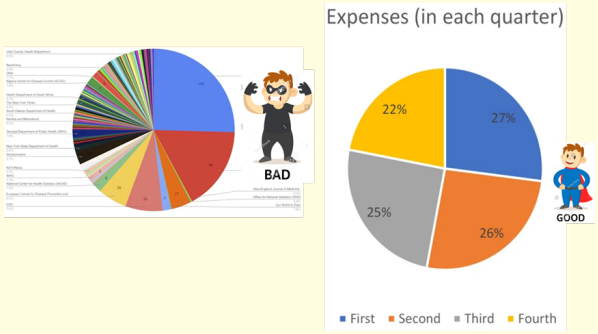
Definition: A categorical dataset can also be summarized using a pie chart. In a pie chart, a circle is used to represent the whole dataset, with slices of the pie representing possible categories. The size of the slice for a particular category is proportional to the corresponding frequency or relative frequency.

A Pie Chart:

- **When to Use?**
  1. Categorical Data with a relatively small number of possible categories.
  2. Pie charts are most useful for illustrating proportions of the whole data set for various categories.
- **How to Construct?**
  1. Draw a circle to represent the entire dataset.
  2. For each category, calculate the "slice" size. Because there are 360 degrees in a circle.

$$\texttt{slice Size} = 360 \times (\texttt{category relative frequency})$$

  3. Draw a slice of appropriate size for each category.This can be tricky thus use computing tools.
- **What to look For?**
  - Categories that form large and small proportions of the dataset.
- Warnings!!Avoid the pie chart when there are too many categories to be represented. See the example below.



3.2.2 Displaying Numerical Data:

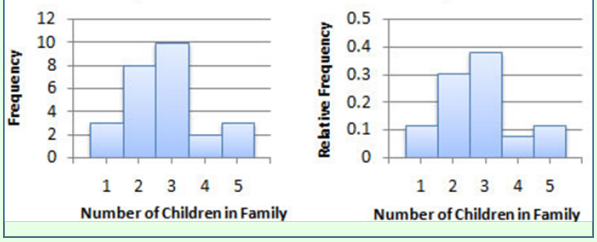
1. Frequency Distributions and Histograms for Discrete Numerical Data:

Definition: A histogram for discrete numerical data is a graph of the frequency distribution and it is similar to the bar chart for categorical data. Each frequency or relative frequency is represented by a rectangle centered over the corresponding value or range of values and the area fo the rectangle si proportional to the corresponding frequency or relative frequency.

Histogram for Discrete Numerical Data:

- **When to Use:** Discrete Numerical Data. Works well even for large datasets.
- **How to construct?**
  1. Draw a horizontal scale, and mark the possible values of the variable.
  2. Draw a vertical scale, and mark it with either frequency or relative frequency.
  3. Above each possible value on vertical scale, draw a rectangle centered at that value.(so that the rectangle for 1 is centered at 1, the rectangle for 5 is centered at 5, and so on.
  4. The height of each rectangle is determined by the corresponding frequency or relative frequency.
- **What to look for?**
  - Central or typical value.
  - Extent of spread or variation.
  - General shape.
  - Location and number of peaks.
  - Presence of gaps and outliers.
- Sample:

Children	frequency	relative frequency
1	3	3/26 ≈ 0.12
2	8	8/26 ≈ 0.31
3	10	10/26 ≈ 0.38
4	2	2/26 ≈ 0.08
5	3	3/26 ≈ 0.12



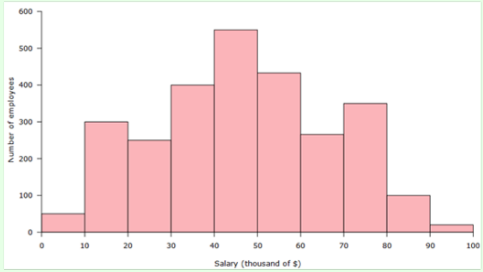
1. Frequency Distributions and Histograms for Continuous Numerical Data:

- When class interval has equal widths:

Histogram for Continuous Numerical Data - For equal Widths:

- **When to Use:** Continuous numerical data. Works well even for large datasets.
- **How to construct?**
  1. Mark the boundaries of the class intervals on a horizontal axis.
  2. Mark either with frequency or relative frequency on a vertical axis.
  3. Above each possible value on vertical scale, draw a rectangle on the corresponding interval such that edges are at the class boundaries.
- **What to look for?**
  - Central or typical value.
  - Extent of spread or variation.
  - General shape.
  - Location and number of peaks.
  - Presence of gaps and outliers.
- Sample:

Salary (thousands)	No of Employees
0-10	50
11-20	300
21-30	250
……	……
91-	20



- When Class Interval has Un-equal Widths:

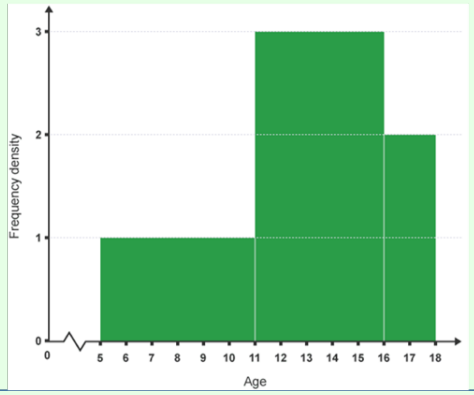
Histogram for Continuous Numerical Data - For equal Widths:

- **When to Use:** Continuous numerical data. Works well even for large datasets.
- **How to construct?**
  1. In this case, frequencies or relative frequencies should not be used on the vertical axis.
  2. Instead, the height of each rectangle, called the density for the class, is given by:

$$\texttt{Density} = \texttt{rectangle height} = \frac{\texttt{relative frequency of class interval}}{\texttt{Class interval width}}$$

  The use of the density scale to construct the histogram ensures the area of each rectangle in the histogram will be proportional to corresponding relative frequency.
- Sample:

Age	frequency	Class width	Frequency density
5-10	6	6	6/6 = 1
11-15	15	5	15/5 = 3
16-17	4	2	4/2 = 2



2. Interpretation of Histogram:

In this section, I have aimed to explain how to interpret histograms based on their shape and spread. This provides only a high-level overview, which meets the needs of this module.

- **1. Shape of a Histogram:**
  - Modality: Modality refers to the number of peaks, or high points, in a data distribution.
    1. Unimodal: A single peak indicates that most data points are centered around one value, often suggesting a single dominant trend.
    2. Bimodal: Two distinct peaks suggest the data may contain two subgroups or clusters.
    3. Multimodal: Three or more peaks indicate multiple groups or clusters, suggesting varied subpopulations within the data.
  - Uniform: No peaks; all bars are roughly the same height, indicating a uniform distribution where each value has a similar frequency.
  - Symmetry: Symmetry in a distribution refers to a balanced shape where both sides of the central peak are mirror images of each other, with the mean located at the center, indicating no skewness in the data.

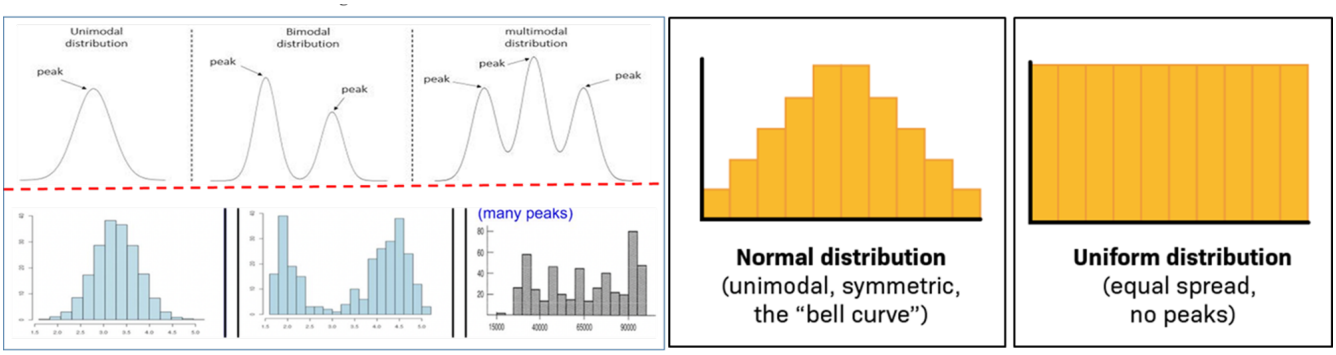


Figure 12: Interpretation of Histograms - Shapes {Slides - 62}.

- **2. Spread of Histogram:**
  - Skewness: It represents the asymmetry in data. In general any Unimodal histogram that is not symmetric are called skewed histogram.
    1. Right-Skewed (Positively Skewed): The data has a long tail on the right side. Most values are concentrated on the left, with fewer higher values. This indicates that higher values are less frequent.
    2. Left-Skewed (Negatively Skewed): The data has a long tail on the left side, with most values concentrated on the right. Lower values are less frequent.
  - The spread also shows the data's variability: wider distributions imply more spread-out data, while narrower distributions suggest that data points are closer to the center.

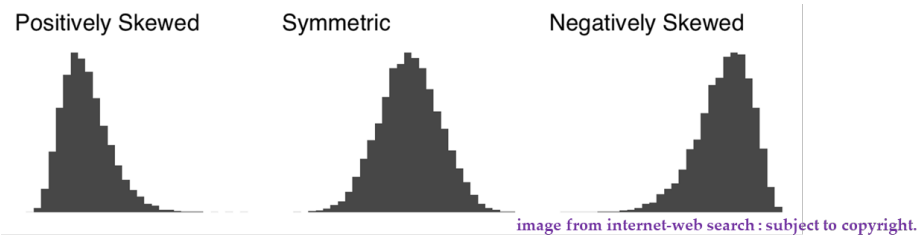


Fig: Skewness of Distribution

Figure 13: Spread of Histogram - Skewness {Slide - 63}.

3.2.3 Summarizing Data Graphically - Boxplot:

A boxplot (also known as a box-and-whisker plot) is a powerful visualization tool that summarizes the key aspects of a dataset's distribution. It provides a clear overview of the central tendency, spread, and outliers. A boxplot is especially useful for comparing the distributions of multiple datasets side by side.

- **Key Components of Boxplot:**
  1. Minimum: The smallest data point within a specified range, not considered an outlier.
  2. First Quartile (Q1): The median of the lower half of the data, marking the 25th percentile.
  3. Median (Q2): The middle value of the dataset, representing the 50th percentile.
  4. Third Quartile (Q3): The median of the upper half of the data, marking the 75th percentile.
  5. Maximum: The largest data point within a specified range, not considered an outlier.
- **Shape of Boxplot:**
  1. The box: The box spans from the first quartile (Q1) to the third quartile (Q3), containing the middle 50% of the data (inter-quartile range or IQR). A line inside the box indicates the median (Q2).
  2. The whiskers: Lines extending from the box to the minimum and maximum values within a set range (usually 1.5 times the IQR from Q1 and Q3).
  3. Outliers: Data points outside the whiskers, typically considered outliers, are shown as individual points.
- **Interpreting Boxplot:**
  1. Central Tendency: The median line inside the box shows the dataset's center.
  2. Spread: The width of the box (IQR) shows the spread of the middle 50% of the data.
  3. Symmetry: If the box and whiskers are symmetric, the data is likely to be symmetrically distributed around the median. If one whisker is longer than the other, the data may be skewed.
  4. Outliers: Points outside the whiskers indicate potential outliers that differ significantly from the rest of the data.
- **Advantages:**
  1. Provides a clear summary of the distribution, highlighting the center, spread, and outliers.
  2. Easy to compare distributions between multiple datasets.

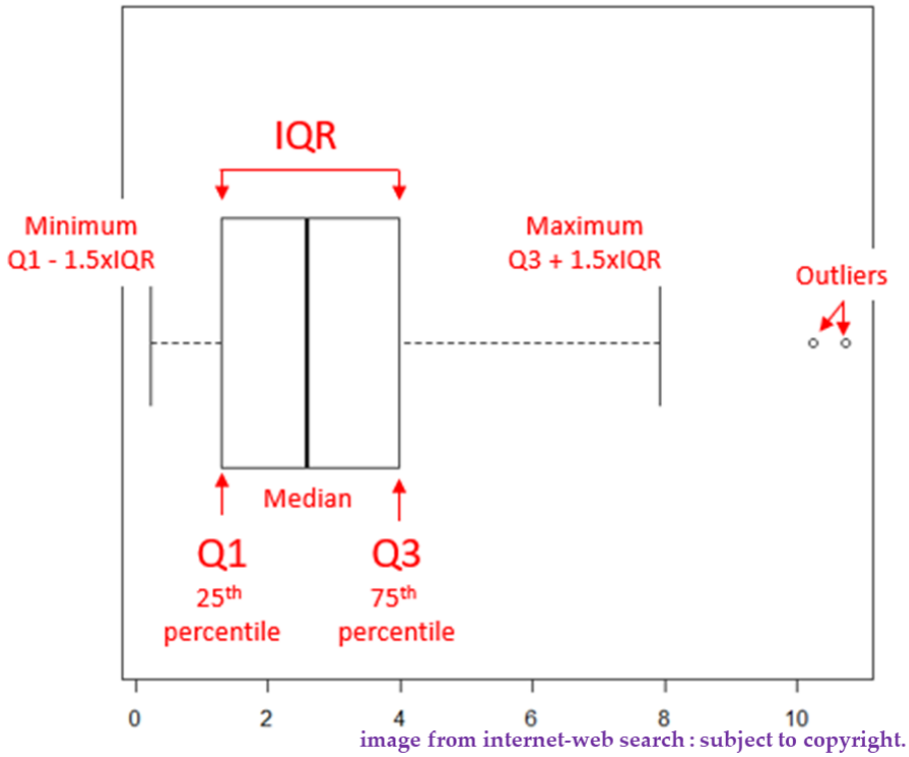


Figure 14: Shape and Element of Boxplot.

3. Visualizes skewness and potential outliers effectively.

A boxplot is a concise tool for quickly understanding the shape and variability of the data, making it invaluable for exploratory data analysis.

3.3 Exercises - Summarizing Data Graphically:

{Recommendation!!For exercise in this section feel free to use any spreadsheet application (Excel or Google Sheet) or python with Matplotlib to plot a graph only.}

Problem - 1:

For the following dataset construct:

- Frequency Table with Relative frequency and Cumulative frequency.
- Grouped Frequency Table using square root method to determine the class interval.

Datasets are:

1. The scores (out of 100) obtained by 33 students in a mathematics test are as follows: 69, 48, 84, 58, 48, 73, 83, 48, 66, 58, 84, 66, 64, 71, 64, 66, 69, 66, 83, 66, 69, 71, 81, 71, 73, 69, 66, 66, 64, 58, 64, 69, 69.
2. Following is recorded weight of apples:

$$[106, 107, 123, 70, 139, 186, 111, 115, 107, 104107, 115, 125, 126, 119, 84, 141, 98, 81, 110$$
$$76, 82, 109, 93, 187, 95, 111, 92, 86, 68, 130129, 115, 128, 100, 99, 113, 204, 90, 123, 136$$
$$110, 131, 80, 78, 75, 118, 90, 84, 82]$$

Problem - 2:

You conducted a survey to collect the ages of 50 participants in a fitness program. The ages were recorded as follows:

$$32, 40, 38, 45, 50, 34, 42, 36, 48, 55,$$
$$39, 41, 37, 44, 49, 43, 35, 52, 46, 40,$$
$$33, 47, 39, 51, 38, 44, 42, 36, 53, 41,$$
$$37, 45, 39, 54, 40, 38, 36, 43, 50, 42,$$
$$38, 55, 39, 37, 49, 46, 44, 40, 35, 52$$

Construct a Histogram and Discuss any patterns or insights you observe from the histogram {shape or spread}.

Hint

To construct a histogram:
1. Create a Grouped Frequency Table using square root method with equal intervals.
2. Draw a histogram representing the age distribution using the calculated class width and intervals.

Problem - 3:

Consider the data given in the table below: Construct a Histogram and Interpret the Graph.

Weight(gms.)	Frequency
$0 \leq w < 40$	5
$40 \leq w < 50$	6
$50 \leq w < 60$	8
$60 \leq w < 70$	4
$70 \leq w < 100$	2

Hint

The widths are not equal in the provided Frequency Table,Thus in order to keep the histogram fair, the area of the bars rather than the height, must be proportional to the frequency. So on the vertical scale or axis we plot density instead of frequency, where

$$\texttt{Density} = \frac{\texttt{Frequency}}{\texttt{Class Width}}$$

Problem - 4:

Construct the Histogram and Interpret:

1. The value of π up to 50 decimal places is given below:

   $$3.14159265358979323846264338327950288419716939937510$$

   Make a frequency distribution of the digits from 0 to 9 after the decimal point.

4 Summarizing Bivariate Data.

1. What are Bivariate Data?

Bivariate data consists of paired observations, where each data point includes two variables measured on the same individual, object, or event. This type of data allows us to analyze the relationship or association between the two variables. These two variables are typically analyzed together to understand the relationship

Graduation Rate	Student Related Expenditure
64.6	8011
53.0	7323
46.3	8735
38.5	7071

Figure 15: Example of Bivariate Data - Graduate Rate and Student Related Expenditure -{Slide - 65}.

or association between them.{cause-relationship.}

2. Bivariate Data Analysis:

Bivariate data analysis explores the relationship between two variables, helping us understand how changes in one variable are associated with changes in another. This analysis is essential for identifying correlations, trends, and patterns within paired data.

3. Correlation:

Correlation is a measure of a monotonic association between two variables in which either:

- As the value of one variable increases, the value of the other variable consistently increases (positive correlation) or
- As the value of one variable increases, the value of the other variable consistently decreases (negative correlation).

In a monotonic relationship, the direction of change is consistent, but the rate of change does not necessarily have to be constant.

4.1 Graphical Summary of Bivariate Data:

Scatter Plot is a tool for visualizing the relationship between two variables graphically or pictorially. Each point on a scatter plot represents an observations, with values of one variable plotted on the x-axis and the other on the y-axis.Patterns in the scatter plot can indicate the nature of the relationships i.e. positive, negative, or no correlation. In the figure:

- **a): Positive linear relationships.**
- **b): another positive linear relationships.**
- **c): Negative linear relationships.**

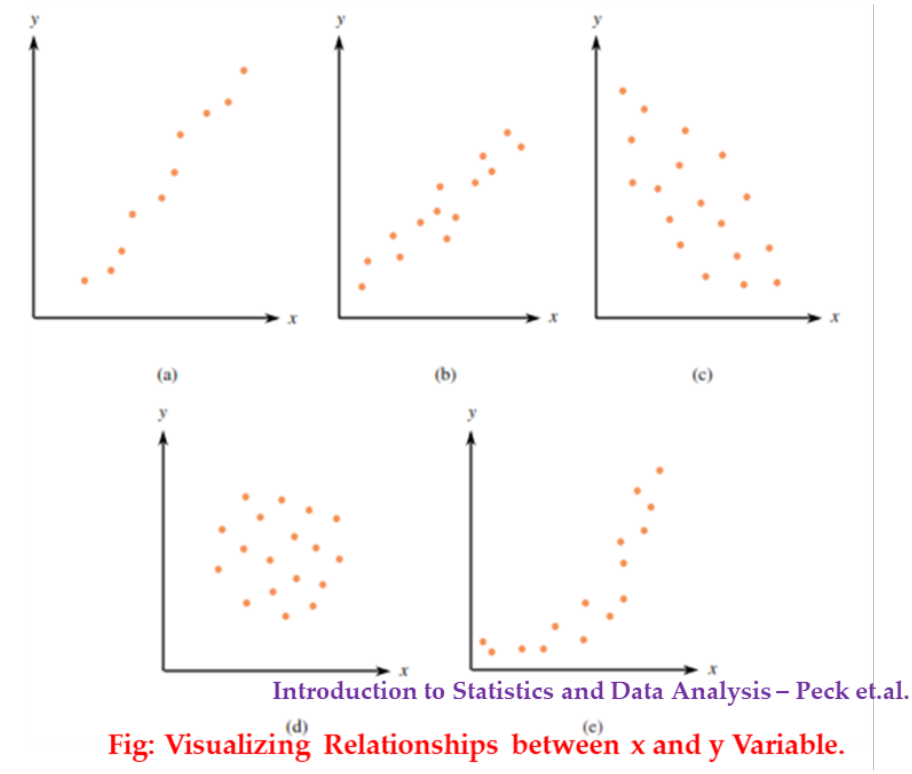


Figure 16: Scatter plots illustrating various types of relationships -{Slides - 67 - 68}.

- **d): No relationships observed.**
- **e): Curved linear relationships.**

4.2 Numerical Summary of Bivariate Data:

For two variable `X` and `Y` correlation coefficient or simply called as correlation {from **co** and **relation**} is a numerical assessment between them. Some popular techniques for calculating correlation coefficients:

1. Pearson Correlation Coefficient
2. Spearman Rank Correlation Coefficient
3. Kendall's Tau

{Disclaimer!!! In this module we will only discuss about Pearson Correlation Coefficient}

1. Pearson Sample Correlation Coefficient:

The Pearson sample correlation coefficient is a statistical measure that quantifies the strength and direction of a linear relationship between two variables, ranging from $-1 \text{to} +1$. Pearson sample correlation coefficient denoted by `r` is given by:

Formula Pearson Sample Correlation Coefficient:

For a pair of sample data:

$$(x_1, y_1), \ldots, (x_n, y_n) \in (X, Y)$$

Pearson sample correlation coefficient "`r`" is:

$$\mathtt{r}_{X,Y} = \frac{\sum((x_i - \bar{X}) \cdot (y_i - \bar{Y}))}{\sqrt{\sum(x_i - \bar{X})^2 \cdot \sum(y_i - \bar{Y})^2}}$$

where:

- x_i and y_i are the individual data points,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y.

An alternative Representations:

An alternative representation for Pearson sample correlation coefficient "`r`" is:

$$\mathtt{r}_{x,y} = \frac{\sum z_x z_y}{n-1}$$

Here:

$$z_x = \frac{X_i - \bar{X}}{s_x} \quad ; \quad z_y = \frac{Y_i - \bar{Y}}{s_y}$$

This represents the correlation formula in terms of standardized scores (z-scores) for X and Y, with Where:

- z_x and z_Y as standardized values of X and Y,
- $\bar{X}$ and $\bar{Y}$ as the means,
- s_x and s_y as the standard deviations of X and Y, respectively.

2. Interpreting the value of `r`:

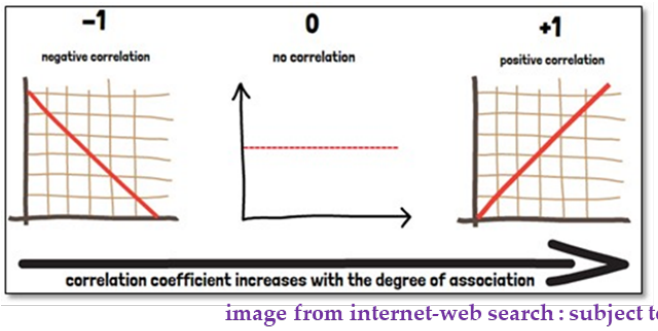


image from internet-web search : subject to copyright.

Fig: Correlation Coefficient Value and it's interpretation.

Figure 17: Understanding the value of `r` - {slide 70}.

1. The value of "r" does not depend on the unit of measurement for either variable.
2. The value of "r" does not depend on which of the two variables is considered x.
3. The value of "r" is between -1 and +1 . A value near the upper limit: +1 indicates a substantial positive relationship and value near lower limit -1 suggests a substantial negative relationship.
4. The value of "r" = 1 only when all the points in scatter-plot of the data lie exactly on a straight line that slopes upward. Similarly, "r" = -1 only when all the points lie exactly on a downward sloping line.
5. The value of "r" is a measure of the extent to which x and y are linearly related. A value close to 0 does not rule out any strong relationship between x and y, there may exist a non linear relationship.

3. Cautions!!! - "Correlation: Not Always What it Seems!":

1. Association vs. Causation:
   - Association refers to a relationship between two variables where they seem to move together, either in a positive or negative direction. However, association does not imply causation.
   - Just because two variables are correlated does not mean that one variable causes the other to change. There could be other factors at play, or the correlation could be due to chance.

   Example: There may be a positive correlation between ice cream sales and drowning incidents, but this does not mean that eating ice cream causes drowning. The common cause could be hot weather, which increases both activities.
2. Correlation Does Not Measure Strength of Agreement:
   - Correlation measures the direction and strength of a linear relationship between two variables, but it does not describe the strength of agreement between them. For example, even with a high correlation, the data points might be widely scattered, meaning that the relationship is weak in practical terms.
   - A high correlation value (close to 1 or -1) only means that there is a strong linear relationship, not necessarily a strong agreement between the variables.
3. Correlation Coefficient and Zero Correlation
   - A correlation coefficient close to zero means that there is no linear relationship between the two variables, but this does not imply that the variables are entirely unrelated. There may still be a nonlinear or non-monotonic relationship that correlation cannot detect.
   - Misconception: A zero correlation coefficient does not necessarily indicate no relationship at all. For example, a quadratic or other types of non-linear relationships may exist even when correlation is zero.
4. Key Takeaways are:
   - Correlation Causation: Do not jump to conclusions about cause and effect just because two variables are correlated.
   - Zero correlation does not mean no relationship: The variables could still be related in a nonlinear way.
   - Correlation fails for non-linear relationships: Use other methods (like regression analysis or non-parametric tests) to understand more complex relationships.

4.3 Exercises - Bivariate Analysis:

Problem - 1:

You are investigating the relationship between the number of hours students spend studying and their exam scores.You collect data from 10 students and want to explore if there is a correlation between study hours and exam scores.

$$\mathbf{StudyHours(X)} : [\mathbf{3, 5, 2, 7, 4, 6, 1, 8, 5, 6}]$$

$$\mathbf{ExamScores(Y)} : [\mathbf{70, 75, 60, 85, 72, 80, 55, 90, 78, 82}]$$

Find the correlation coefficients and interpret the result.

Solution:

Formula Review:

$$\textbf{Correlation Coefiicient}(\mathbf{r_{XY}}) = \frac{\sum(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{Y} - \bar{\mathbf{Y}})}{\sqrt{\sum(\mathbf{X} - \bar{\mathbf{X}})^{\mathbf{2}} \times \sum(\mathbf{Y} - \bar{\mathbf{Y}})^{\mathbf{2}}}} \tag{1}$$

Solution:

- Step - 1: Find the Mean:

$$\bar{\mathbf{X}} = \frac{\sum X}{n} = \frac{47}{10} = 4.7$$
$$\bar{\mathbf{Y}} = \frac{\sum Y}{n} = \frac{767}{10} = 76.7$$

- Step - 2: Calculating all the elements of the formula:

$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$	$(X - \bar{X}) \times (Y - \bar{Y})$
-1.70	-4.70	2.890	22.09	7.99
0.30	0.30	0.09	0.09	0.09
-2.70	-14.70	7.29	216.09	39.69
2.30	10.30	5.29	10.09	23.69
-0.70	-2.70	0.49	7.29	1.89
1.30	5.30	1.69	28.09	6.89
-3.70	19.70	13.69	388.09	72.89
3.30	15.30	10.89	234.09	50.49
0.30	3.30	0.09	10.89	0.99
1.30	7.30	1.69	53.29	9.49
		$\sum = 44.10$	$\sum = 1066.10$	$\sum = 214.10$

- Step - 3: Substituting in the equation{1}:

$$\mathbf{r_{XY}} = \frac{214.10}{\sqrt{44.10 \times 1066.10}} \approx 0.9874$$

- Step - 4: Interpretation.
  - This indicates a strong positive linear relationship between study hours and exam scores.

Problem -2:

Suppose you are studying the relationship between the time spent on homework,the number of extracurricular activities and students' overall grades. You collect data from 10 students; and want to explore the correlation among variables.{Hint: Explore the correlation between each pair of variables. i.e $\mathbf{r_{XY}}$; $\mathbf{r_{XZ}}$; $\mathbf{r_{YZ}}$}

$X : [2, 3, 1, 4, 2, 3, 1, 5, 3, 4]$ (hours spent on homework)

$Y : [3, 2, 1, 4, 3, 2, 1, 5, 4, 3]$ (number of extracurricular activities)

$Z : [75, 80, 70, 85, 78, 82, 68, 90, 85, 88]$ (overall grades)

Problem -3:

Look at the following bi-variate data table. It represents the age and average height of a group of babies and kids. Find the correlation coefficient and interpret the result:

Age (mnths.)	Height(cms.)
3	58.5
6	64
9	68.5
12	74
24	81.2
36	89.1
48	95
60	102.5

5 Multivariate Analysis with Covariance Matrix.

Optional but Highly Recommended!

Multivariate Analysis refers to statistical methods used to analyze data that involves more than one variable at a time. It allows us to examine the relationships between multiple variables simultaneously, rather than in isolation.

1. **Covariance Matrix in Multivariate Analysis:**
   - The Covariance Matrix is a key concept in multivariate analysis. It is a square matrix that describes the covariance (the measure of how much two variables change together) between multiple variables.
   - If you have p variables, the covariance matrix is $p \times p$, where each element σ_{ij}represents the covariance between variables X_i and X_j.

Formula for Covariance

– The General Textbook formula is:

$$\texttt{Cov}(X_i, X_j) = \frac{1}{n-1}(\sum_{k=1}^{n}(X_{ik} - \bar{X})(X_j k - \bar{X_j})$$

where:

* X_{ik} and X_{jk} are individual data points of variables X_i and X_j.
* $\bar{X}_i$ and $\bar{X}_j$ are the means of the respective variables.

– Based on Intution: The covariance matrix for two variables X and Y is given by:

$$\textbf{CovarianceMatrix} = \begin{bmatrix} \textbf{Cov}(\mathbf{X},\mathbf{X}) & \textbf{Cov}(\mathbf{X},\mathbf{X}) \\ \textbf{Cov}(\mathbf{Y},\mathbf{X}) & \textbf{Cov}(\mathbf{Y},\mathbf{Y}) \end{bmatrix} = \begin{bmatrix} \textbf{Var}(\mathbf{X}) & \textbf{Cov}(\mathbf{X},\mathbf{X}) \\ \textbf{Cov}(\mathbf{Y},\mathbf{X}) & \textbf{Var}(\mathbf{Y}) \end{bmatrix}$$

$$\texttt{COV}(XY) = \frac{\sum(X-\bar{X})\cdot(Y-\bar{Y})}{n-1}$$

2. **Understanding the Covariance Matrix:**
   - The covariance matrix is symmetric, meaning that the covariance between X_i and X_j is equal to the covariance between X_j and X_i.
   - Diagonal elements of the covariance matrix represent the variance of each variable (covariance of a variable with itself).
   - Off-diagonal elements represent the covariance between pairs of variables.
3. **Interpreting the Covariance Matrix:**
   - Positive Covariance: If $\sigma_{ij} > 0$, Variables X_i and X_j tend to increase or decrease together (they have a positive relationship).
   - Negative Covariance: If $\sigma_{ij} < 0$, one variable tends to increase while the other decreases (they have an inverse relationship).
   - Zero Covariance: If $\sigma_{ij} = 0$, there is no linear relationship between the two variables.

5.1 Exercise - Covariance Matrix!!

Problem -1:

Let's say you have to study the relationship between the age and the systolic blood pressure in a company. You have a sample of 10 workers aged thirty to fifty-five years. The results are presented in the following bi-variate data table:

S.No	Age	Systolic Blood Pressure
1	37	130
2	38	140
3	40	132
4	42	149
5	45	144
6	48	157
7	50	161
8	52	145
9	53	165
10	55	162

Construct a Covariance Matrix.

Constructing Covariance Matrix:

We Know: The covariance matrix for two variables X and Y is given by:

$$\textbf{CovarianceMatrix} = \begin{bmatrix} \textbf{Cov}(\mathbf{X},\mathbf{X}) & \textbf{Cov}(\mathbf{X},\mathbf{X}) \\ \textbf{Cov}(\mathbf{Y},\mathbf{X}) & \textbf{Cov}(\mathbf{Y},\mathbf{Y}) \end{bmatrix} = \begin{bmatrix} \textbf{Var}(\mathbf{X}) & \textbf{Cov}(\mathbf{X},\mathbf{X}) \\ \textbf{Cov}(\mathbf{Y},\mathbf{X}) & \textbf{Var}(\mathbf{Y}) \end{bmatrix}$$

Let's compute the individual element of the above matrix:

- Step - 1: Compute the Mean:
  1. Mean of Variable - Age:
  $$\bar{Age} = \frac{\sum(Age)}{n} = \frac{460}{10} = 46 \quad \square$$
  2. Mean of Variable - Pressure:
  $$\bar{Pressure} = \frac{\sum(Pressure)}{n} = \frac{1505}{10} = 150.5 \quad \square$$
- Step - 2: Covariance Calculation:
  1. `cov(Age, Age)` $==$ `var(Age, Age)`:
  $$\texttt{Cov}(\texttt{Age},\texttt{Age}) = \frac{\sum(\texttt{Age}_i - \bar{\texttt{Age}})^2}{n-1} \approx 2.8 \quad \square$$
  2. `cov(Age, Pressure)` $==$ `cov(Age, Press )`{property of Symmetry}:
  $$\texttt{Cov}(\texttt{Age},\texttt{Pressure}) = \frac{\sum(\texttt{Age}_i - \bar{\texttt{Age}})(\texttt{Pressure}_i - \bar{\texttt{Pressure}})}{n-1} \approx 9.31 \quad \square$$
  also
  $$\texttt{Cov}(\texttt{Pressure},\texttt{Age}) \approx 9.31 \quad \square$$
  3. `cov(Pressure, Pressure)` $==$ `var(Pressure, Pressure)`:
  $$\texttt{Cov}(\texttt{Pressure},\texttt{Pressure}) = \frac{\sum(\texttt{Pressure}_i - \bar{\texttt{Pressure}})^2}{n-1} \approx 60.74 \quad \square$$
  4. Substituting back to the covariance matrix:
  $$\text{Covariance Matrix} = \begin{bmatrix} \text{Cov(Age, Age)} & \text{Cov(Age, Pressure)} \\ \text{Cov(Pressure, Age)} & \text{Cov(Pressure, Pressure)} \end{bmatrix} = \begin{bmatrix} 2.8 & 9.31 \\ 9.31 & 60.74 \end{bmatrix} \quad \square$$

Problem - 2:

Construct a Covariance Matrix for Problem [2] from previous section.
Hint:

$$\textbf{Covaraince Matrix}(\Sigma) = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) & \text{Cov}(X,Z) \\ \text{Cov}(Y,X) & \text{Var}(Y) & \text{Cov}(Y,Z) \\ \text{Cov}(Z,X) & \text{Cov}(Z,Y) & \text{Var}(Z) \end{bmatrix}$$