Data gathering and quantitative analysis to studying a problem/issue

MD. MAZHARUL ISLAM

RESEARCH ASSOCIATE

BANGLADESH INSTITUTE OF GOVERNANCE AND MANAGEMENT

Data and Information

What is data?

- Data is a **raw** and **unorganized** fact that is required to be processed to make it meaningful.
- Data can be structured/tabular data or unstructured.
- Example: total marks of male and female students (in 500):
  male: 400, 390, 450, 410
  female: 395, 480, 460, 440

What is information?

- Information is a set of data that is **processed** in a **meaningful** way according to the given requirement.
- Information is language, ideas, and thoughts based on the given data.
- Example: average mark of male students is 412.5 and female students is 443.75. Female students do better in exam than male students.

You are 40 feet above from the ground.

Where am I?

Yes, but how could you know that!!?

Are you a researcher?

Are you a policy maker?

What you have said is 100% correct but it is useless!!

You don’t know where you are, and don’t know where to go, but you are blaming me!!

Yes, but how could you know that!!?

How to get information from data?

Through **Data Analysis**

What is data analysis?

Data analysis is defined as the process of cleaning, transforming, summarizing, modeling and presenting data to discover useful information for decision-making.

Can we imagine the relationship between decision-making and policy-making?

How to get information from data?

Through **Data Analysis**

What is data analysis?

Data analysis is defined as the process of cleaning, transforming, summarizing, modeling and presenting data to discover useful information for decision-making.

Can we imagine the relationship between decision-making and policy-making?

Through data analysis, we will get **some decisions**, and

through policy analysis, we will be able to pick the **most feasible and suitable decision**.

Types of Data

❑ Based on source of data

1. **Primary data**:
   Collected for the first time by an investigator for a specific purpose; may vary across time, space, subject and context.
2. **Secondary Data**:
   Sourced from somewhere that is originally collected and we are just using it.

❑ Based on attribute types

1. **Qualitative data:**
   Represents a particular quality or attribute.
   Mainly answers questions such as ‘why,’ ‘who’, ‘what’ or ‘how.’
   Eg: religion, Gender, Blood group etc.
2. **Quantitative data:**
   It is measured in terms of numbers.
   Mainly answers questions like ‘how much’, ‘how many’ or ‘what value’
   e.g. age, height, weight etc.

Types of Data

1. **Qualitative data**

1.1. Nominal data:
used for naming or labelling certain characteristics.
e.g. gender: male and female

1.2. Ordinal data:
type of categorical data with an order.
e.g. wealth categories: poorest, poorer, middle, richer, richest

2. **Quantitative data**

2.1 Discrete data:
numbers but discrete in nature.
e.g. number of students, family size etc.

2.2 Continuous data:
continuous numeric values.
e.g. Weight, length, height etc.

Types of Data

❑ Based on collection process

1. **Cross-sectional Data:**
   data is collected in a single time period and is characterized by individual units - people, companies, countries, etc.
2. **Time Series Data:**
   data is collected at a number of specific points in time.
3. **Panel (or Longitudinal) data:**
   combination of cross-sectional and time series data

❑ Based on format of data

1. **Structured data:**
   highly-organized, labeled and formatted in a way so it's easily searchable in a database.
2. **Unstructured data:**
   everything else. It may be textual or non-textual, and human- or machine-generated. e.g. social media data, text, images etc.

Sources of Secondary Data

Economics, Finance, Governance, politics and Development

- **The International Country Risk Guide (ICRG)** https://www.prsgroup.com/explore-our-products/international-country-risk-guide/
- **World Bank** https://data.worldbank.org/
- **Global economic databank** https://www.oxfordeconomics.com/global-economic-databank
- **Freedom House** https://freedomhouse.org/report/freedom-world
- **Penn World Table (PWT)** https://www.rug.nl/ggdc/productivity/pwt/?lang=en
- **World Development Indicators** https://databank.worldbank.org/source/world-development-indicators
- **Human Development Data** http://hdr.undp.org/en/data

Sources of Secondary Data

Health, Nutrition, Agriculture, Food and Safety

- **World Health Organization** https://www.who.int/data/collections
- **Unicef** https://data.unicef.org/
- **Demographic and Health Surveys (DHS)** https://dhsprogram.com/Data/
- **Food and Agriculture Organization (FAO)** http://www.fao.org/faostat/en/#home
- **International Labor Organization (ILO)** https://ilostat.ilo.org/data/
- **Our world Data and charts** https://ourworldindata.org/
- **Worldometer** https://www.worldometers.info/
- **IDF diabetes atlas** https://www.diabetesatlas.org/data/en/
- **Multiple Indicator Cluster Surveys** https://mics.unicef.org/

Sources of Secondary Data

Bangladesh

- **Bangladesh Bureau of Statistics**
- **Bangladesh Bank Open Data Initiative**
- **Bangladesh Open data**

- http://www.bangladeshstudies.org/resources-data.html

Variable and constant

- **Variable**

any characteristic of an individual or subject that varies across individuals or subjects.

e.g. Age of respondents, income, house rent etc.

- **Constant**

Value is unchanged over the study population.

e.g. value or pi.

Scale of Measurement

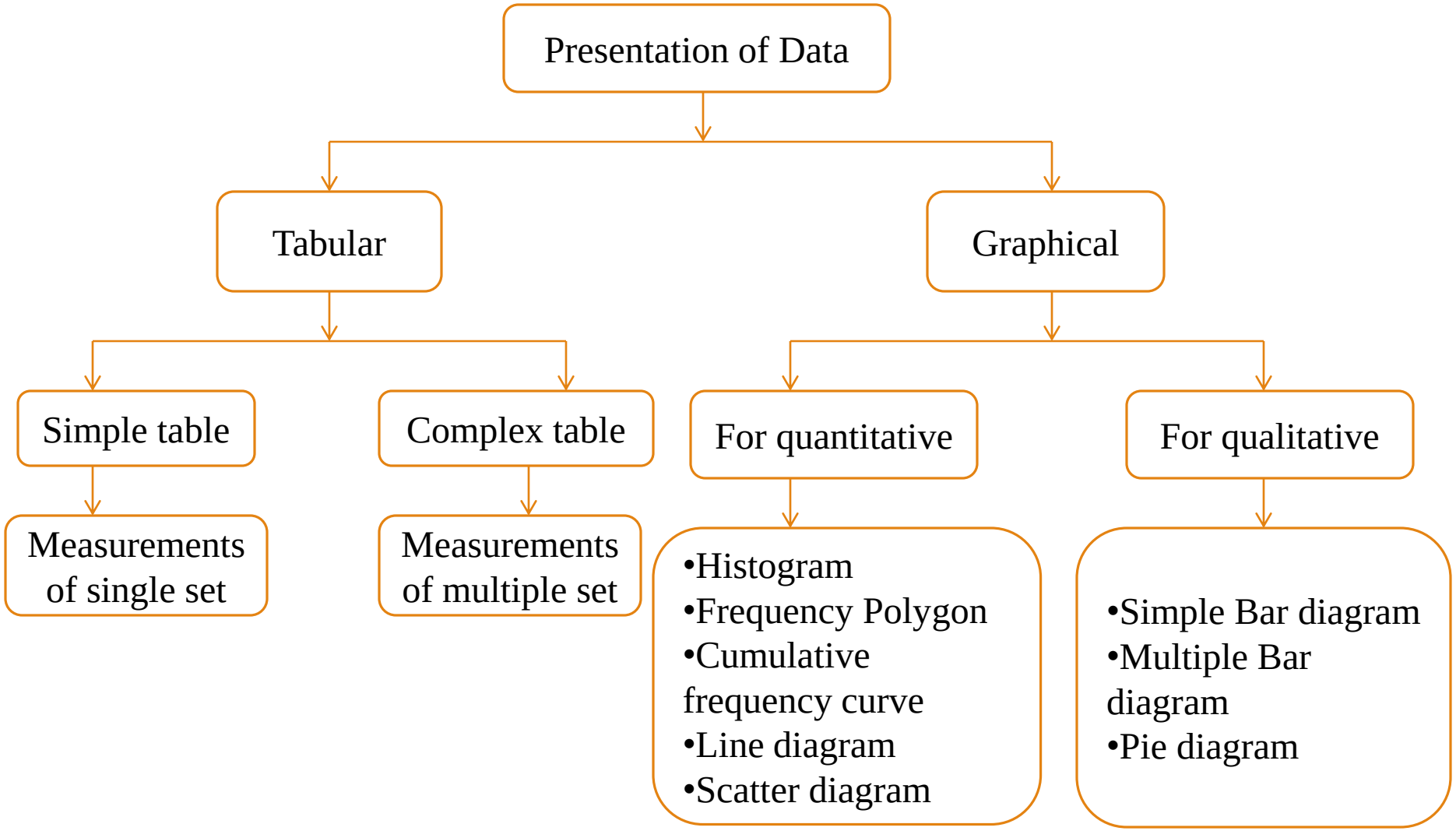
1. Nominal
   used to name, label or categorize particular attributes that are being measured. This variable has no numeric values and natural orders.
   e.g. name of respondents, gender, country, etc.

2. Ordinal
   a type of measurement variable that takes values with an order or rank.
   e.g. wealth category, Likert Scale

3. Interval
   values measured along a scale, with each point placed at an equal distance from one another. There is no absolute zero point.
   E.g. temperature.

4. Ratio
   exactly the same as the interval scale except that the zero on the scale means: does not exist.
   E.g. weight, height etc.

Data Presentation



Simple table

- When characteristics with values are presented in the simple form of table.
- For instance, Infant mortality rates in South Asian countries between 2010 and 2015 (infant deaths per 1,000 live births)
- Rule of thumb of presenting a table:
  1. Give a title with unique table number at the top of the table
  2. Avoid presenting *hanging* table (try to accommodate in single page)

Table 1: Infant mortality rates in South Asian

Country	Infant Mortality Rate
Pakistan	70
India	41
Bangladesh	33
Nepal	32
Sri lanka	8

Frequency distribution table

The data is first split up into convenient groups (class interval) and the number of items (frequency) which occur in each group is shown in adjacent columns.

Rule of thumb: maintain logical order of frequency groups.

From the above frequency distribution table, it can be said that prevalence of polio is higher in age group 0-4 where polio patients number is 35.

Table 2: Age distribution of polio patients

Age	Number of patients
0-4	35
5-9	18
10-14	11
15-19	8
20-24	6

Histogram

- used for Quantitative and Continuous variables.
- used to present variables which have no gaps e.g age, weight, height, blood pressure, blood sugar etc.
- the class intervals are given along horizontal axis and the frequency along the vertical axis.

Rule of thumb: add figure description and unique figure number at the bottom of the figure

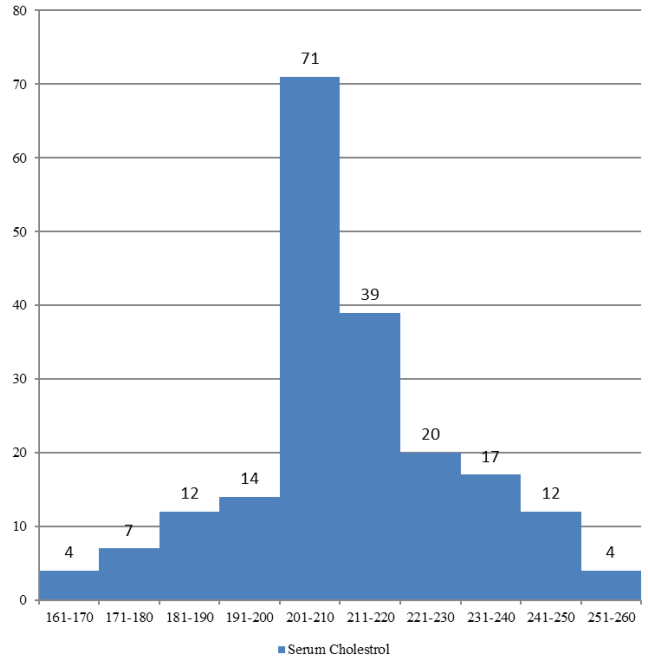
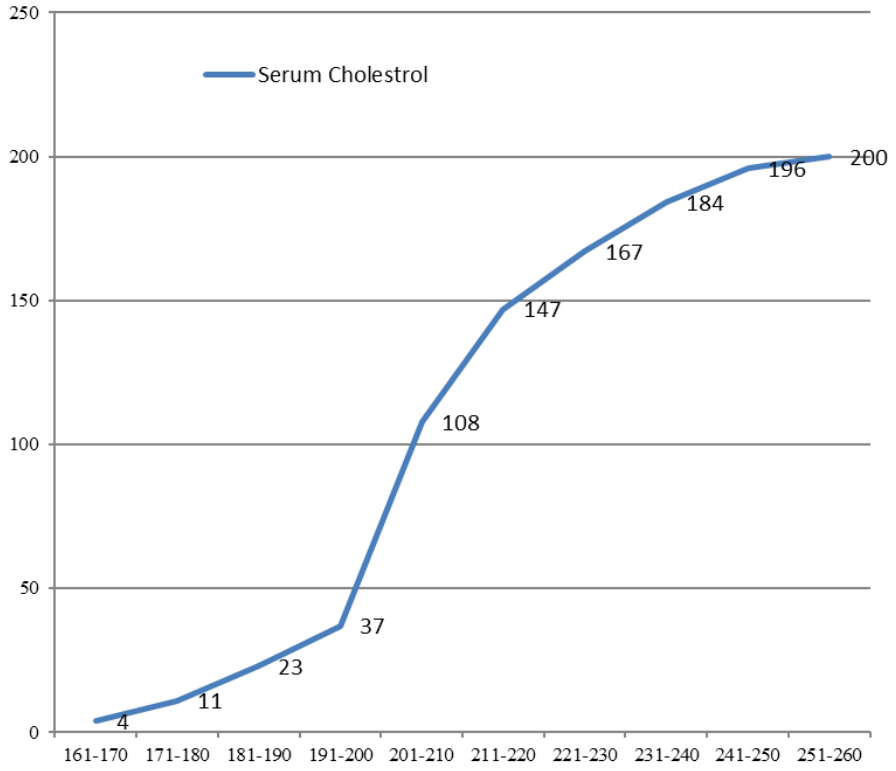


Figure 1: histogram plot of serum cholesterol level

Cumulative frequency curve

A Cumulative Frequency Graph is a graph plotted from a cumulative frequency table.



Line Diagram

A line graph is particularly useful when we want to show the trend of a variable over time.

Time is displayed on the horizontal axis (x-axis) and the variable is displayed on the vertical axis (y- axis).

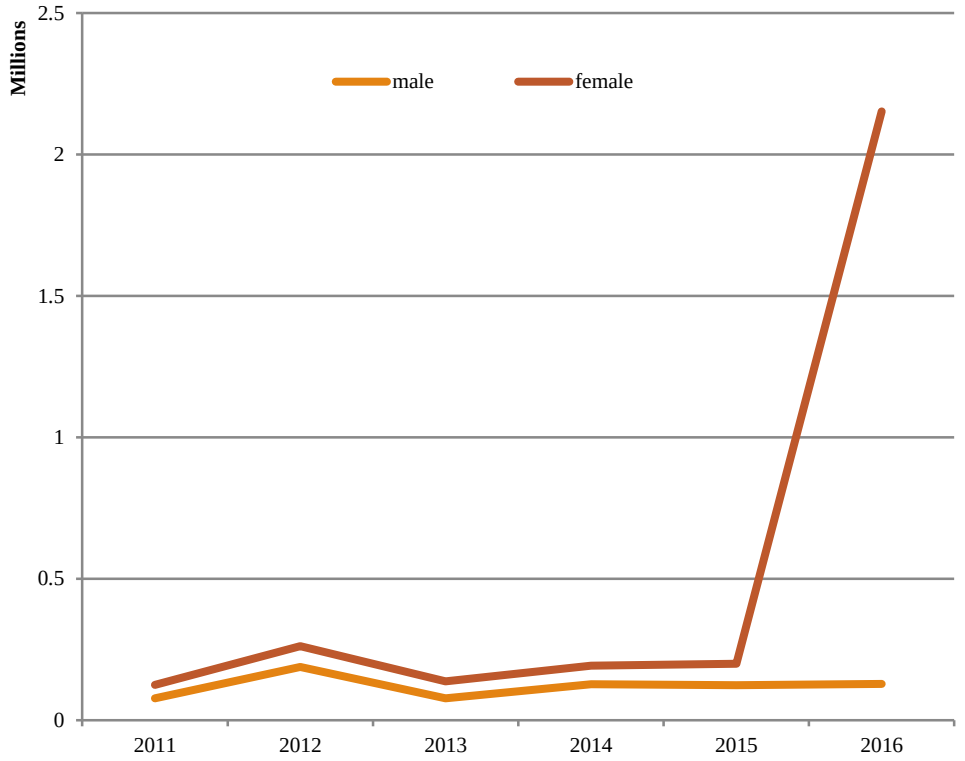


Figure: Number of teachers in primary school

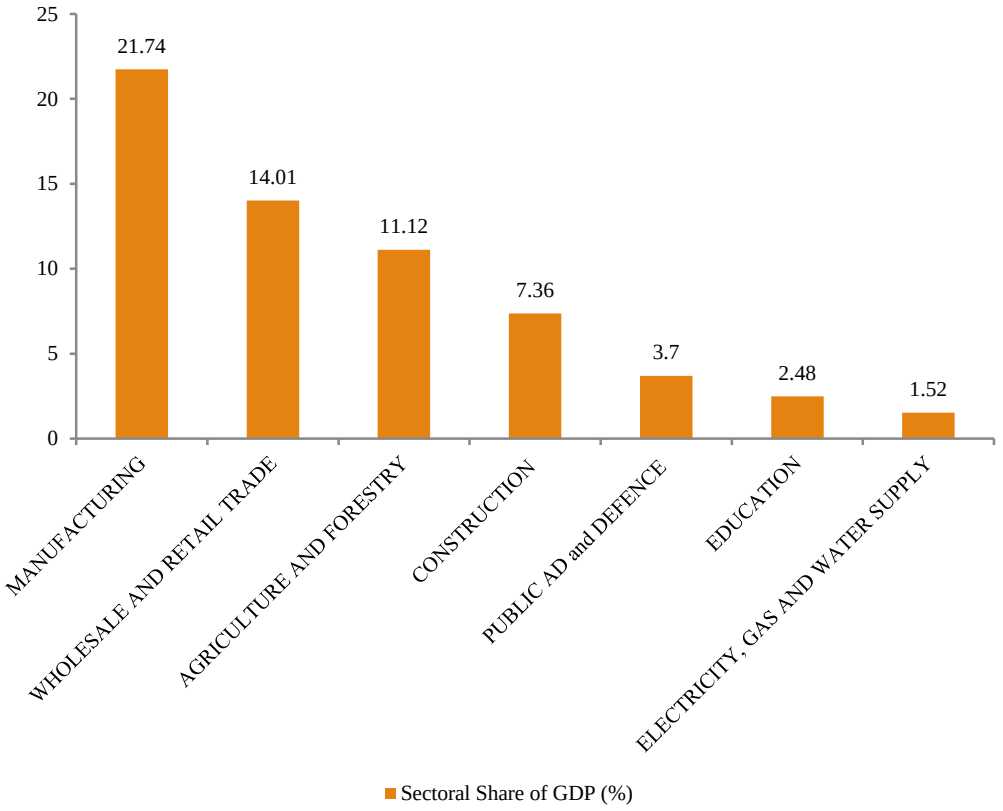
Simple Bar Diagram

The data presented is categorical.

Data is presented in the form of rectangular bar of equal distance.

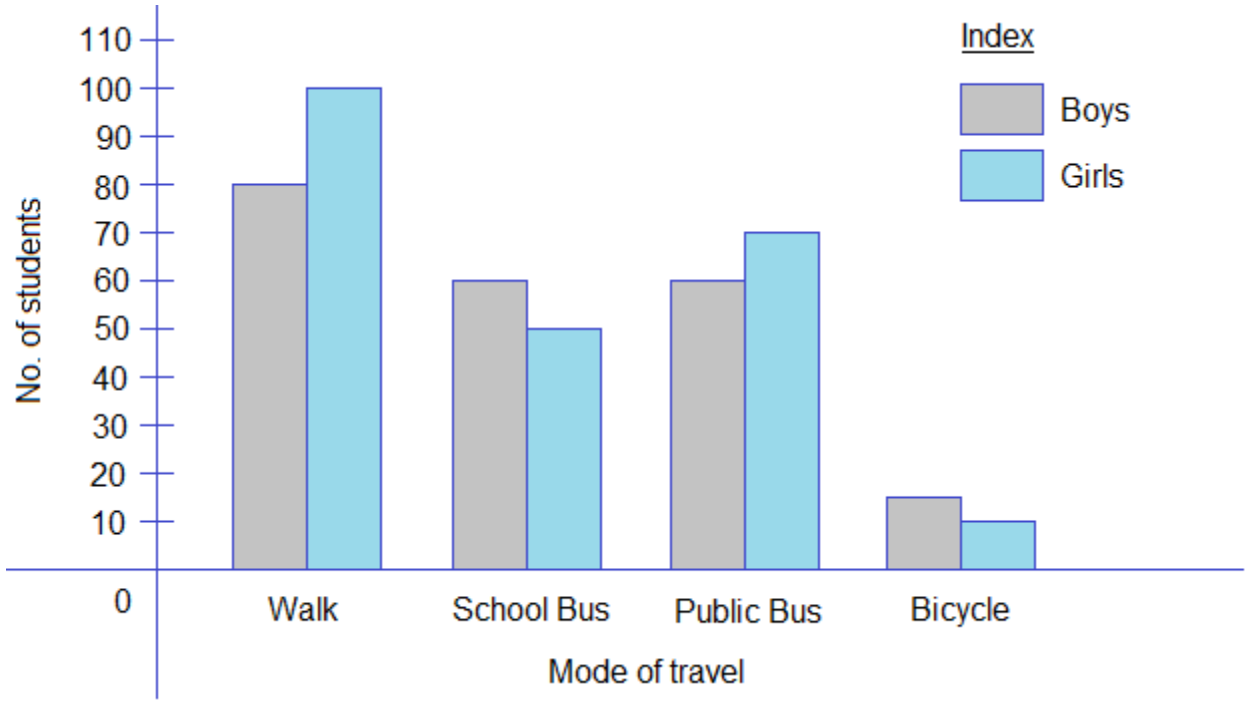
Each bar represent one attribute. The width of the bar and the gaps between the bars should be equal throughout.

The bars may be vertical or horizontal.



Multiple Bar Diagram

• More then one sub-attribute of variables can be expressed.



Data Summarization

1. Measures of center

1.1. The Mode

The mode is that value of the variable which occurs with the greatest frequency in a data set.

Example: 3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

In order these numbers are: 3, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23**, 29, 39, 40, 56

This makes it easy to see which numbers appear most often.

In this case the mode is **23**.

Data Summarization

1. Measures of center

1.2. The Median

Example: 3, 13, 7, 5, 21, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

When we put those numbers in order we have:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56

There are **fifteen** numbers. Our middle is the **eighth** number:

3, 5, 7, 12, 13, 14, 21, **23**, 23, 23, 23, 29, 39, 40, 56

The median value of this set of numbers is **23**.

Data Summarization

1. Measures of center

1.3. The Mean (μ, $\bar{x}$)

Example: 3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

The sum of these numbers is 330

There are fifteen numbers.

The mean is equal to 330 / 15 = 22

The mean of the above numbers is 22

Brainstorming

- What are the possible ways to present the data graphically?
- What is the most appropriate measure of center? Why?

Data Summarization

2. Measures of Variation

2.1. Range

The sample range of the variable is the difference between its maximum and minimum values in a data set:

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

Example: In **{4, 6, 9, 3, 7}** the lowest value is 3, and the highest is 9.

So the range is 9 − 3 = **6**.

Data Summarization

2. Measures of Variation

2.2. Variance

The range only involves the smallest and largest numbers, and it would be desirable to have a statistic which involved all of the data values.

The first attempt one might make at this is something they might call the average squared deviation from the mean and define it as:

$$Population\ \ Variance\ = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Data Summarization

2. Measures of Variation

2.3. Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$Population\ Standard\ Deviation = \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$Sample\ Standard\ Deviation = s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

McLaren

Average distance after pressing break: 850 m
Standard deviation: 25 m

Ferrary

Average distance after pressing break: 880 m
Standard deviation: 20 m

Which car would you like to buy?

Data Summarization

Coefficient of Variation

- Relative measure of variability that indicates the size of a standard deviation in relation to its mean
- Unit-less measure that allows to compare variability between groups
- CV = Standard deviation/mean; often presents as percentage
- Higher value indicates that the standard deviation is relatively large compared to the mean
- CV of McLaren = 25/850*100 = 2.94%
- CV of Ferrary = 20/880*100 = 2.27%