Summarizing Data

Md. Aminul Islam Shazid

Outline

1. Introduction
2. Tabular Summary
3. Graphical Summary

Introduction

Summarizing Data

- Raw data by itself is not useful
- Need to extract insight from data
- Data needs to be summarized to gain insights
- Can summarize in two ways:
  - Tabular summary
  - Graphical summary

Tabular Summary

Tabular Summary

Frequency Table

Frequency Table

- Organizes data from the sample by listing distinct values or classes
- Shows the frequency of each class in the sample
- Can be constructed for categorical or numerical data:
  - For categorical variable, it shows the number of observations in each category
  - For discrete numeric variable, it shows how many times each value has been observed
  - For continuous numeric variable, classes or bins are formed first, then the table shows how many values fall in each class
- Forms the basis for graphical summaries like bar charts and histograms

Example: Frequency Table for Categorical Variable

Suppose we have a discreete variable with four categories: A, B, C and D. The data in the sample is: A, B, A, D, A, A, B, C, C, B, A, A, C, D, D, B, D, C, C, B, C. The Frequency table would be:

Category	Tally	Frequency
A	𝍸	6
B	𝍸	5
C	𝍸	6
D	𝍸	4

The tally column is only used to keep track of the data when counting by hand.

One can add a percentage column in necessary.

Frequency Table for continuous Numeric Variable

- Data is grouped into classes or groups
- Each class has the same class interval (the difference between the upper limit and the lower limit of each class)

Before making the table, need to decide the value of:

- Either the number of classes or groups
- Or, the class interval for each class or group

Frequency Table for continuous Numeric Variable (cont.)

There is no hard and fast rules for setting the classes or the class interval.

It usually depends on the variable and its range (difference between the highest value and the lowest value).

- For example, if the range is 50, then with class interval of 10, there will be 5 classes
- Conversely, if 10 classes are wanted, then the class interval will be 5

Too many or too few classes may fail to reveal the basic shape of the data.

Frequency Table for continuous Numeric Variable (cont.)

Sometimes, the classes are determined through subject-matter considerations.

For example, in case of vitamin D levels:

- Values lower than 10 ng/mL would be classified as deficient
- 10 to 30 is insufficient
- 30 to 100 is sufficient
- Above 100 is considered toxic

Example

Make a frequency table with the following data: 30, 40, 5, 110, 11, 15, 55, 20, 130, 45, 30, 47, 52, 68, 105, 62, 52, 98, 76, 85, 83, 91, 49, 38, 57, 27, 23, 42, 9, 65

Solution:

The range in the sample is = highest value - lowest value = 130 - 5 = 125

Therefore, with 5 classes, the class interval would = 125/5=25

Example (cont.)

Category	Tally	Frequency
5 - 30	𝍸 |	7
30 - 55	𝍸 𝍸	10
55 - 80	𝍸 |	6
80 - 105	||||	4
105 - 130	|||	3

For each class, the upper limit is excluded and the lower limit is included, except for the last class in which the upper limit is also included. This is called upper limit exclusive and lower limit inclusive.

This is necessary because the variable is continuous, so there is no gap between the values of each class.

Tabular Summary

Contingency Table

Contingency Table

- A contingency table is used to summarize the relationship between two (or more) categorical variables
- Data are arranged in rows and columns, where each cell contains the frequency for a specific combination of categories
- Row totals, column totals, and a grand total are often included
- It helps identify patterns, associations, or possible dependence between variables

Example: Contingency Table

	Lung Cancer		
Smoking	**Yes**	**No**	**Total**
Smoker	55	45	100
Non-smoker	10	90	100
Total	65	135	200

In the above, it can be seen that lung cancer percentage is higher among the smokers compared to those who do not smoke.

Graphical Summary

Graphical Summary

Categorical Data

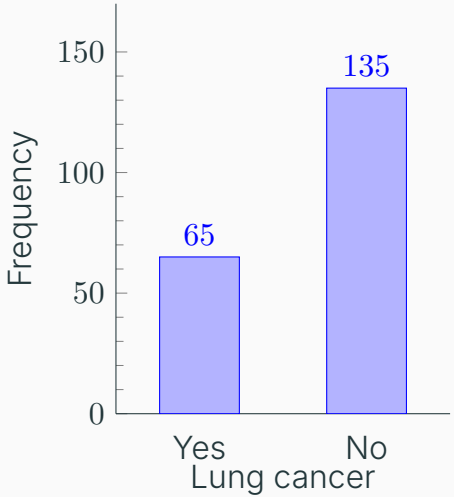
Graphs and Charts for Categorical variables

Graphs/charts/plots are a versatile method of summarizing data.

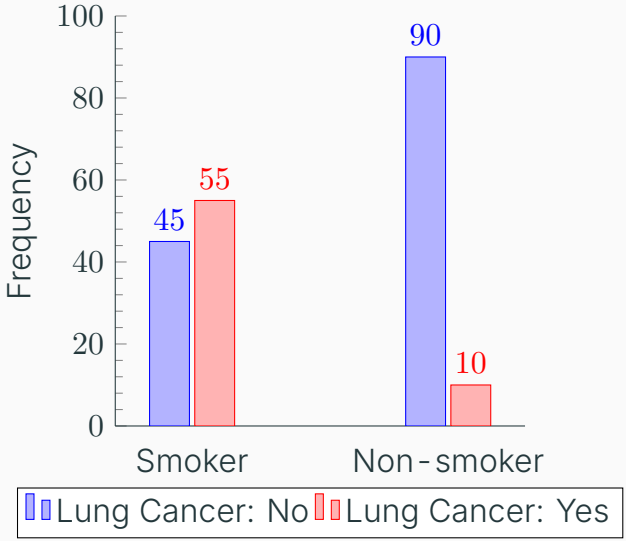
- **Bar diagram/plot**: graphically shows the frequencies of each of the categories of a single categorical variable
- **Clustered bar diagram, stacked bar diagram**: Graphically shows the frequencies of each combination of categories of multiple (usually two) categorical variables
- **Pie chart**: graphically shows the frequencies of each of the categories of a single categorical variable

These categorical variables can be of both nominal and ordinal scales of measurement.

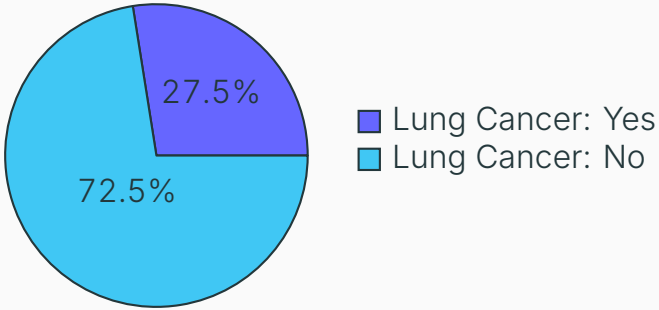
Bar Diagram



Clustered Bar Diagram



Pie Chart



Graphical Summary

Numeric Data

Graphs and Charts for Numeric variables

- Frequency Polygon
- Ogive curve
- Histogram
- Density Plot
- Scatterplot

Frequency Polygon

- Graphically shows the information provided by a frequency table
- Not used for categorical variables
- The information is laid out in a Cartesian plane (a 2D graph with two axes)
- For discreet numeric variables, the X axis shows the individual values of the variable
- For continuous numeric variables, the variable is grouped into a few classes first, just like in Frequency table. Then the class midpoints are plotted on the X axis
- The Y axis denotes the frequency of each value on the X axis
- Two empty classes with 0 frequency, are added at the start and the end of the data

Example: Frequency Polygon

Below is the frequency table from page 7:

Category	Class midpoint	Frequency
5-30	17.5	7
30-55	42.5	10
55-80	67.5	6
80-105	92.5	4
105-130	117.5	3

The class frequencies will be plotted against the class midpoints.

Example: Frequency Polygon (cont.)

Ogive Curve

- An ogive is a graph of cumulative frequency or cumulative relative frequency
- It is constructed by plotting cumulative frequencies against class boundaries
- The points are joined by a smooth curve or straight line segments
- Ogives are used to determine medians, quartiles, and percentiles graphically
- Two common types exist: less-than ogive and greater-than ogive

Less-than Ogive

- Plots less-than cumulative frequency on the Y axis
- Uses upper class boundaries on the X axis
- Cumulative frequency increases from 0 to total frequency
- Commonly used to estimate median, quartiles, and percentiles
- The curve slopes upward from left to right
- Each point (x, y) in the plot, represents that there are y observations in the sample that are less than x

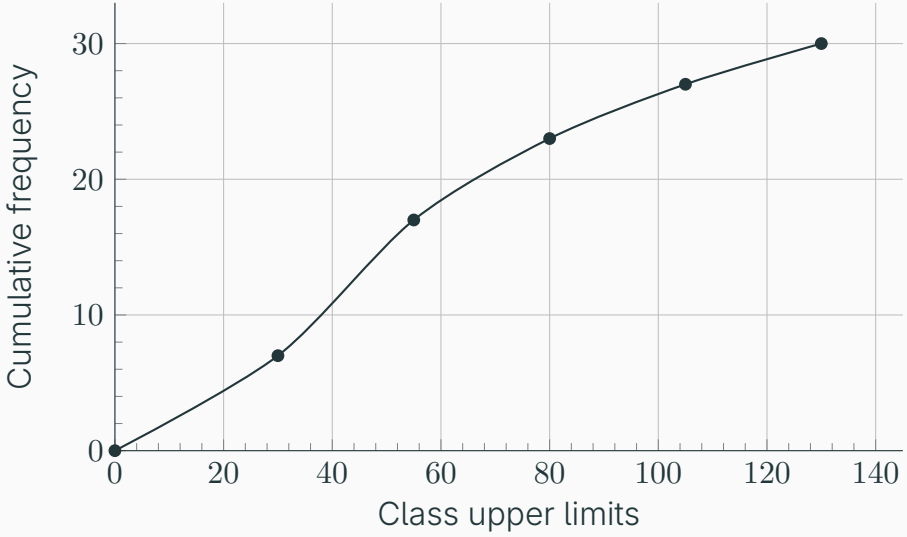
Example: Less-than Ogive

Below is the frequency table from page 7:

Category	Frequency	Cumulative Frequency
5-30	7	7
30-55	10	17
55-80	6	23
80-105	4	27
105-130	3	30

The cumulative frequencies will be plotted against the class upper-limits.

Example: Less-than Ogive (cont.)



Greater-than Ogive

- Plots greater-than cumulative frequency on the Y axis
- Uses lower class boundaries on the X axis
- Cumulative frequency decreases from total frequency to 0
- Useful for finding how many observations exceed a given value
- The curve slopes downward from left to right
- Each point (x, y) in the plot, represents that there are y observations in the sample that are greater than or equal to x

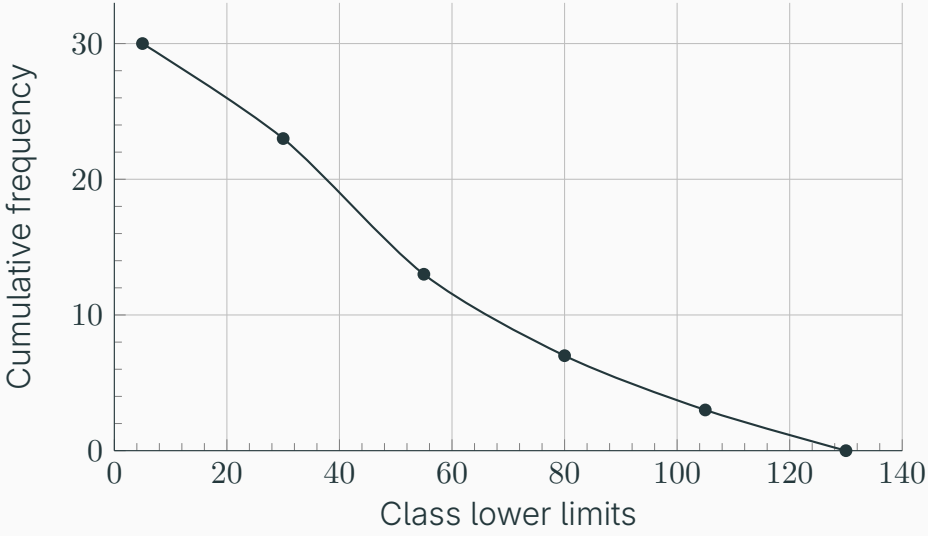
Example: Greater-than Ogive

Below is the frequency table from page 7:

Category	Frequency	Cumulative Frequency
5-30	7	30
30-55	10	23
55-80	6	13
80-105	4	7
105-130	3	3

The cumulative frequencies will be plotted against the class lower-limits.

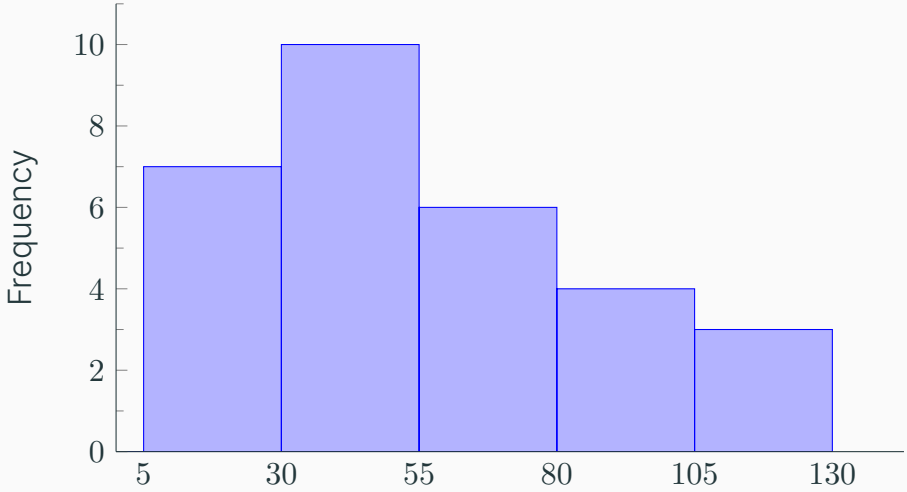
Example: Greater-than Ogive (cont.)



Histogram

- A histogram is a graphical display of the distribution of a continuous quantitative variable
- The horizontal axis (X axis) is divided into class intervals (bins) of equal or specified width
- The height of each bar represents frequency, relative frequency (also called density) within the corresponding bin or class
- There is no gap bewteen the bars, indicating continuity of the data
- It helps reveal the shape of the distribution (symmetric or skewed, unimodal or multimodal etc.)
- It is useful for identifying central tendency, dispersion, and potential outliers

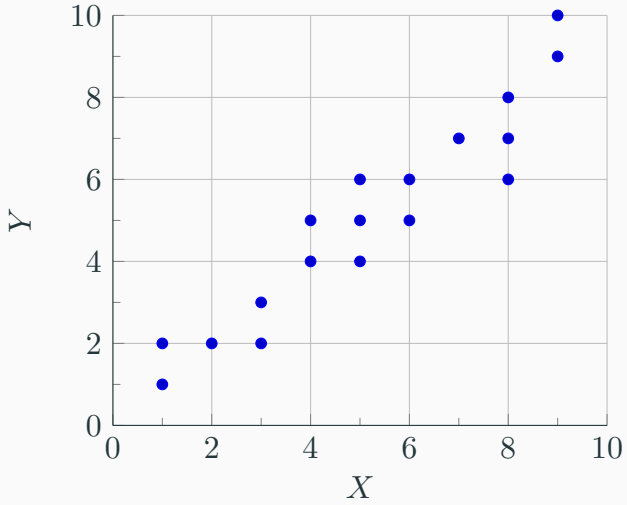
Example: Histogram



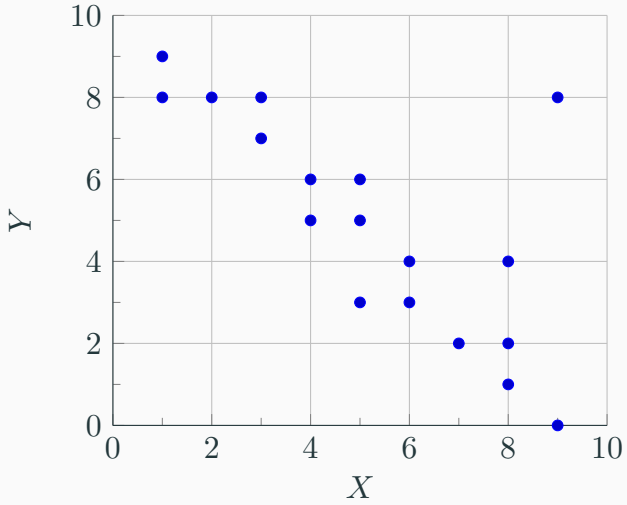
Scatterplot

- A scatterplot displays the relationship between two quantitative variables: one of them called the explanatory (independent) variable, the other called the outcome (dependent) variable
- Each point (x, y) represents an observation
- By convention, the X axis shows the explanatory variable, the Y axis shows the outcome variable
- It helps assess the relationship between the two variables, including direction (positive or negative) and form (linear or nonlinear)
- Patterns in a scatterplot can indicate strength of relationship, clusters, and outliers

Example: Scatterplot, Positive Relationship



Example: Scatterplot, Negative Relationship



Questions?