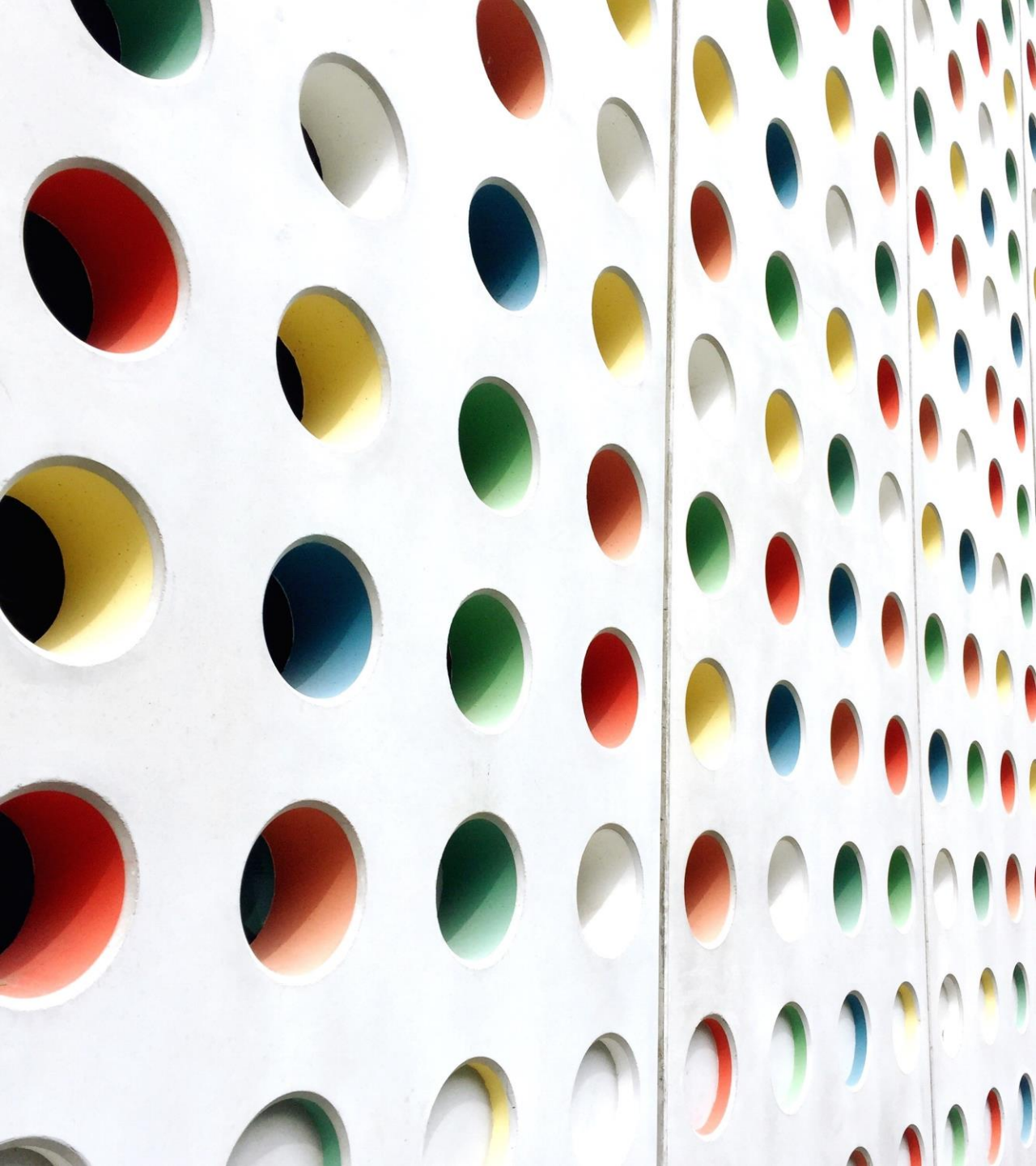




Descriptive Statistics for Business

BM 4419

BUSINESS ANALYTICS

Types of Variables

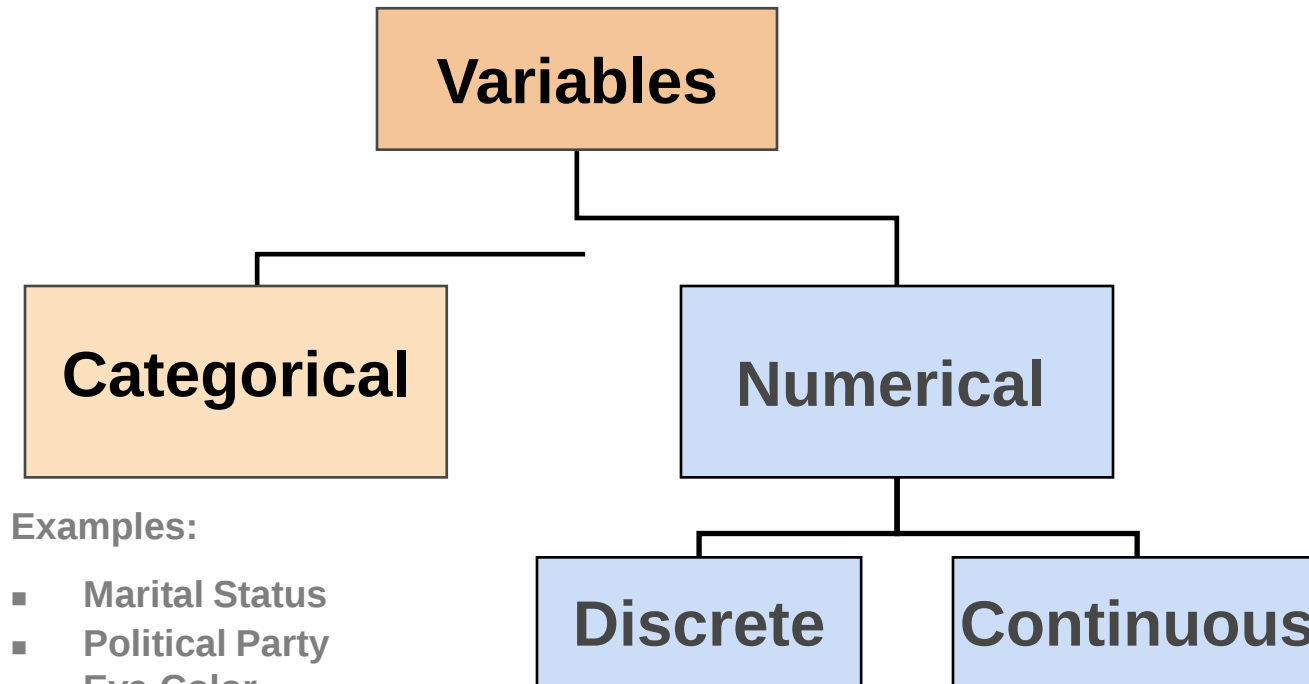
DCOVA

- **Categorical** (*qualitative*) variables have values that can only be placed into categories, such as “yes” and “no.”
- **Numerical** (*quantitative*) variables have values that represent quantities.
 - **Discrete** variables arise from a *counting process*
 - **Continuous** variables arise from a *measuring process*

May2024 _HK

Types of Variables

DCOVA



Examples:

- Marital Status
- Political Party
- Eye Color

(Defined categories)

Examples:

- Number of Children
- Defects per hour

(Counted items)

Examples:

- Weight
- Voltage

(Measured characteristics)

May2024 _HK

1-2 Discrete or Continuous Variables

Classify each variable as a discrete variable or a continuous variable.

- a. The highest wind speed of a hurricane.
Continuous
- b. The weight of baggage on an airplane.
Continuous
- c. The number of pages in a statistics book
Discrete
- d. The amount of money a person spends per year for online purchases.
Discrete

May2024 _HK

Levels of Measurement

DCOVA

A **nominal scale** classifies data into distinct categories in which no ranking is implied.

<i>Categorical Variables</i>		<i>Categories</i>
Personal Computer Ownership	←→	Yes / No
Type of Stocks Owned	←→	Growth / Value / Other
Internet Provider	←→	AT&T, Verizon, Time Warner Cable

May2024 _HK

Levels of Measurement (con' t.)

DCOVA

An **ordinal scale** classifies data into distinct categories in which ranking is implied

<i>Categorical Variable</i>	<i>Ordered Categories</i>
Student class designation	Freshman, Sophomore, Junior, Senior
Product satisfaction	Satisfied, Neutral, Unsatisfied
Faculty rank	Professor, Associate Professor, Assistant Professor, Instructor
Standard & Poor' s bond ratings	AAA, AA, A, BBB, BB, B, CCC, CC, C, DDD, DD, D
Student Grades	A, B, C, D, F

May2024 _HK

Levels of Measurement (con' t.)

DCOVA

- An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but the measurements do not have a true zero point.
- A **ratio scale** is an ordered scale in which the difference between the measurements is a meaningful quantity and the measurements have a true zero point.

May2024 _HK

Interval and Ratio Scales

DCOVA

<i>Numerical Variable</i>		<i>Level of Measurement</i>
Temperature (in degrees Celsius or Fahrenheit)		Interval
Standardized exam score (e.g., ACT or SAT)		Interval
Height (in inches or centimeters)		Ratio
Weight (in pounds or kilograms)		Ratio
Age (in years or days)		Ratio
Salary (in American dollars or Japanese yen)		Ratio

2024 _HK

What level of measurement would be used to measure each variable?

- a. The ages of patients in a local hospital
Ratio
- b. The ratings of movies released this month
Ordinal
- c. Colors of athletic shirts sold by Oak Park Health Club
Nominal
- d. Temperatures of hot tubs in local health clubs
Interval

May2024 _HK

Sources of Data

DCOVA

- **Primary Sources:** The data collector is the one using the data for analysis
 - Data from a political survey
 - Data collected from an experiment
 - Observed data
- **Secondary Sources:** The person performing data analysis is not the data collector
 - Analyzing census data
 - Examining data from print journals or data published on the internet.

May2024 _HK

Sources of data fall into five categories

DCOVA

- ❖ Data distributed by an organization or an individual
- ❖ A designed experiment
- ❖ A survey
- ❖ An observational study
- ❖ Data collected by ongoing business activities

May2024 _HK

Data Is Collected From Either A Population or A Sample

DCOVA

POPULATION

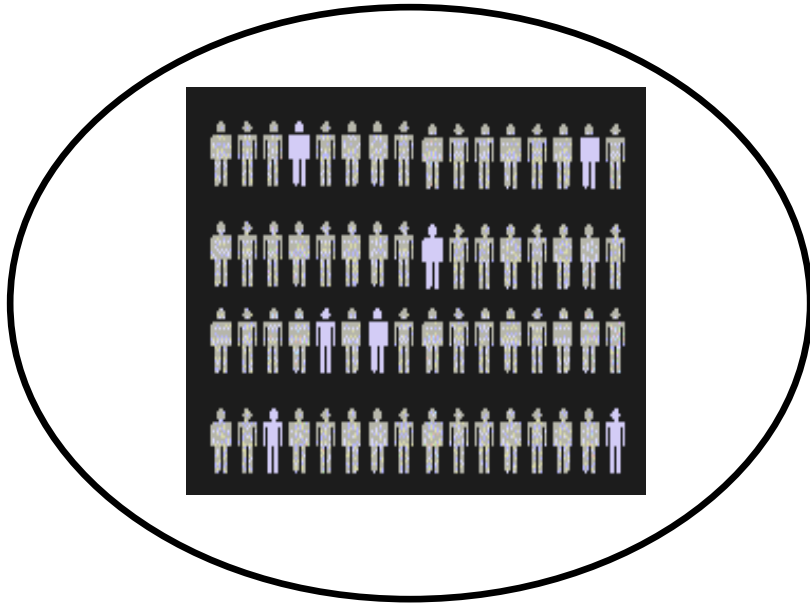
A **population** consists of all the items or individuals about which you want to draw a conclusion. The population is the “large group”

SAMPLE

A **sample** is the portion of a population selected for analysis. The sample is the “small group”

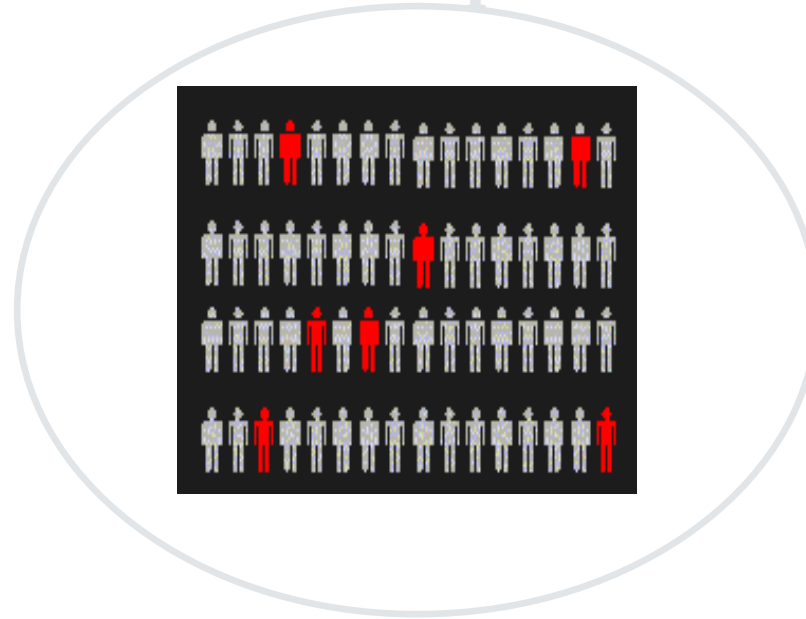
Population vs. Sample

Population



All the items or individuals about which you want to draw conclusion(s)

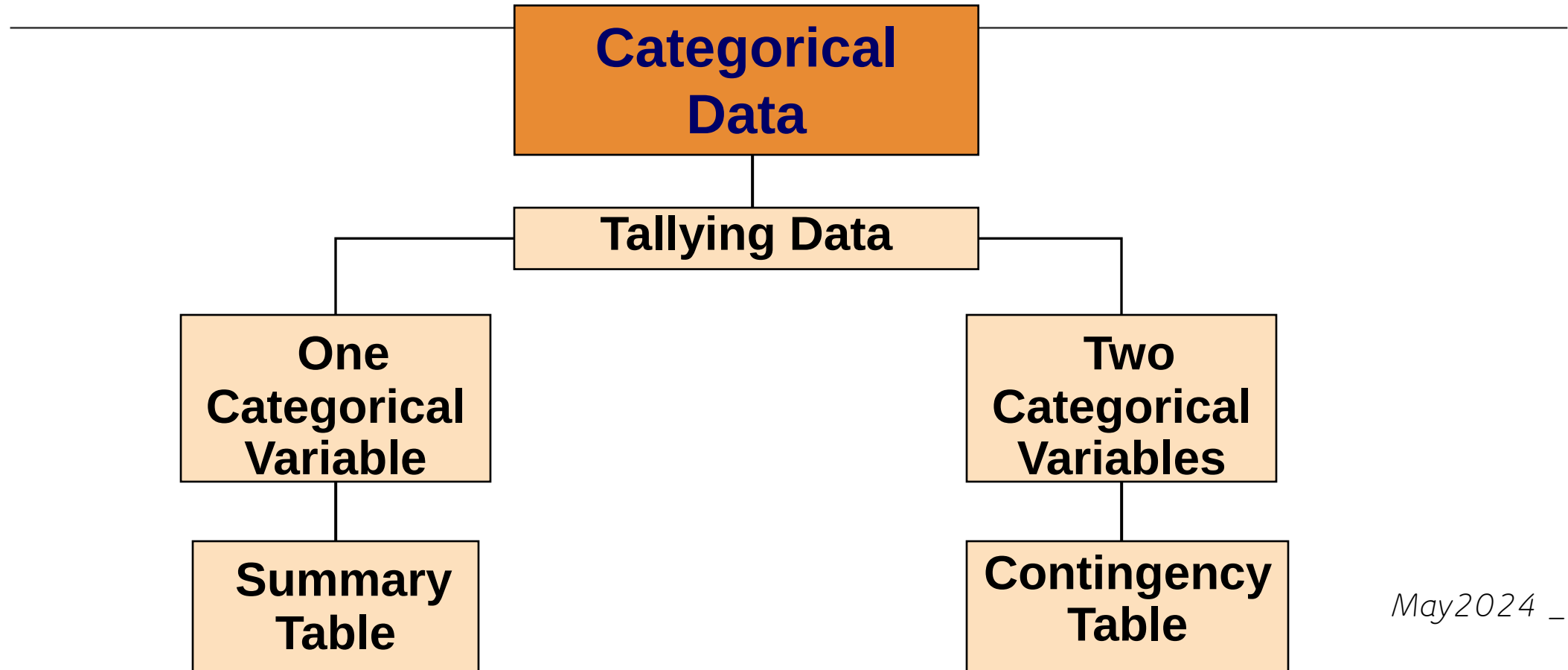
Sample



A portion of the population of *May2024 _HK* items or individuals

Categorical Data Are Organized By Utilizing Tables

DCOVA



May2024 _HK

Organizing Categorical Data: Summary Table

DCOVA

-
- A **summary table** tallies the frequencies or percentages of items in a set of categories so that you can see differences between categories.

Summary Table From A Survey of 1000 Banking Customers

Banking Preference?	Percent
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

A Contingency Table Helps Organize Two or More Categorical Variables

DCOVA

Used to study patterns that may exist between the responses of two or more categorical variables

Cross tabulates or tallies jointly the responses of the categorical variables

For two variables the tallies for one variable are located in the rows and the tallies for the second variable are located in the columns

May2024 _HK

Contingency Table - Example

DCOVA

A random sample of 400 invoices is drawn.

Each invoice is categorized as a small, medium, or large amount.

Each invoice is also examined to identify if there are any errors.

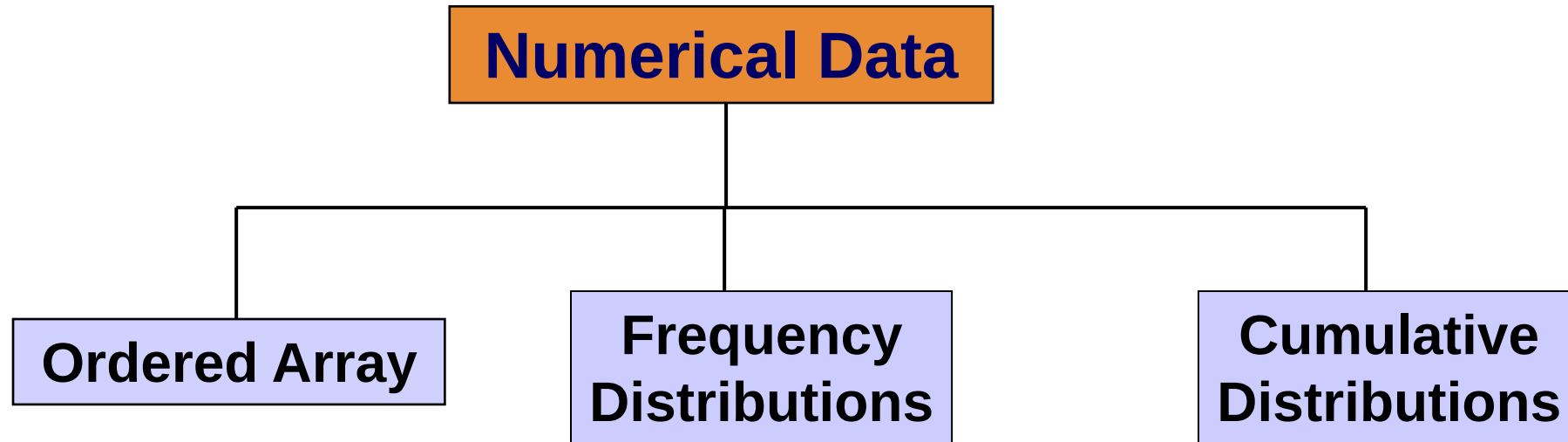
This data are then organized in the contingency table to the right.

Contingency Table Showing Frequency of Invoices Categorized By Size and The Presence Of Errors

	No Errors	Errors	Total
Small Amount	170	20	190
Medium Amount	100	40	140
Large Amount	65	5	70
Total	335	65	400

Tables Used For Organizing Numerical Data

DCOVA



May2024 _HK

Organizing Numerical Data: Ordered Array

- An **ordered array** is a sequence of data, in rank order, from the smallest value to the largest value.
- Shows range (minimum value to maximum value)
- May help identify outliers (unusual observations)

Age of Surveyed College Students	Day Students					
	16	17	17	18	18	18
	19	19	20	20	21	22
	22	25	27	32	38	42
	Night Students					
	18	18	19	19	20	21
	23	28	32	33	41	45

Organizing Numerical Data: Frequency Distribution

DCOVA

-
- The **frequency distribution** is a summary table in which the data are arranged into numerically ordered classes.
 - You must give attention to selecting the appropriate *number* of **class groupings** for the table, determining a suitable *width* of a class grouping, and establishing the *boundaries* of each class grouping to avoid overlapping.
 - The number of classes depends on the number of values in the data. With a larger number of values, typically there are more classes. In general, a frequency distribution should have at least 5 but no more than 15 classes.
 - To determine the **width of a class interval**, you divide the **range** (Highest value–Lowest value) of the data by the number of class groupings desired.

May2024 _HK

Organizing Numerical Data: Frequency Distribution Example

DCOVA

Example: A manufacturer of insulation randomly selects 20 winter days and records the daily high temperature

24, 35, 17, 21, 24, 37, 26, 46, 58, 30, 32, 13, 12, 38, 41, 43, 44, 27, 53, 27

May2024 _HK

Organizing Numerical Data: Frequency Distribution Example

DCOVA

-
- Sort raw data in ascending order:
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
 - Find range: $58 - 12 = 46$
 - Select number of classes: **5 (usually between 5 and 15)**
 - Compute class interval (width): **10 (46/5 then round up)**
 - Determine class boundaries (limits):
 - Class 1: 10 to less than 20
 - Class 2: 20 to less than 30
 - Class 3: 30 to less than 40
 - Class 4: 40 to less than 50
 - Class 5: 50 to less than 60
 - Compute class midpoints: **15, 25, 35, 45, 55**
 - Count observations & assign to classes

May2024 _HK

Organizing Numerical Data: Frequency Distribution Example

DCOVA

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Midpoints	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2
Total		20

May2024 _HK

Organizing Numerical Data: Relative & Percent Frequency Distribution Example

DCOVA

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15%
20 but less than 30	6	.30	30%
30 but less than 40	5	.25	25%
40 but less than 50	4	.20	20%
50 but less than 60	2	.10	10%
Total	20	1.00	100%

May2024 _HK

Organizing Numerical Data: Cumulative Frequency Distribution Example

DCOVA

Data in ordered array:

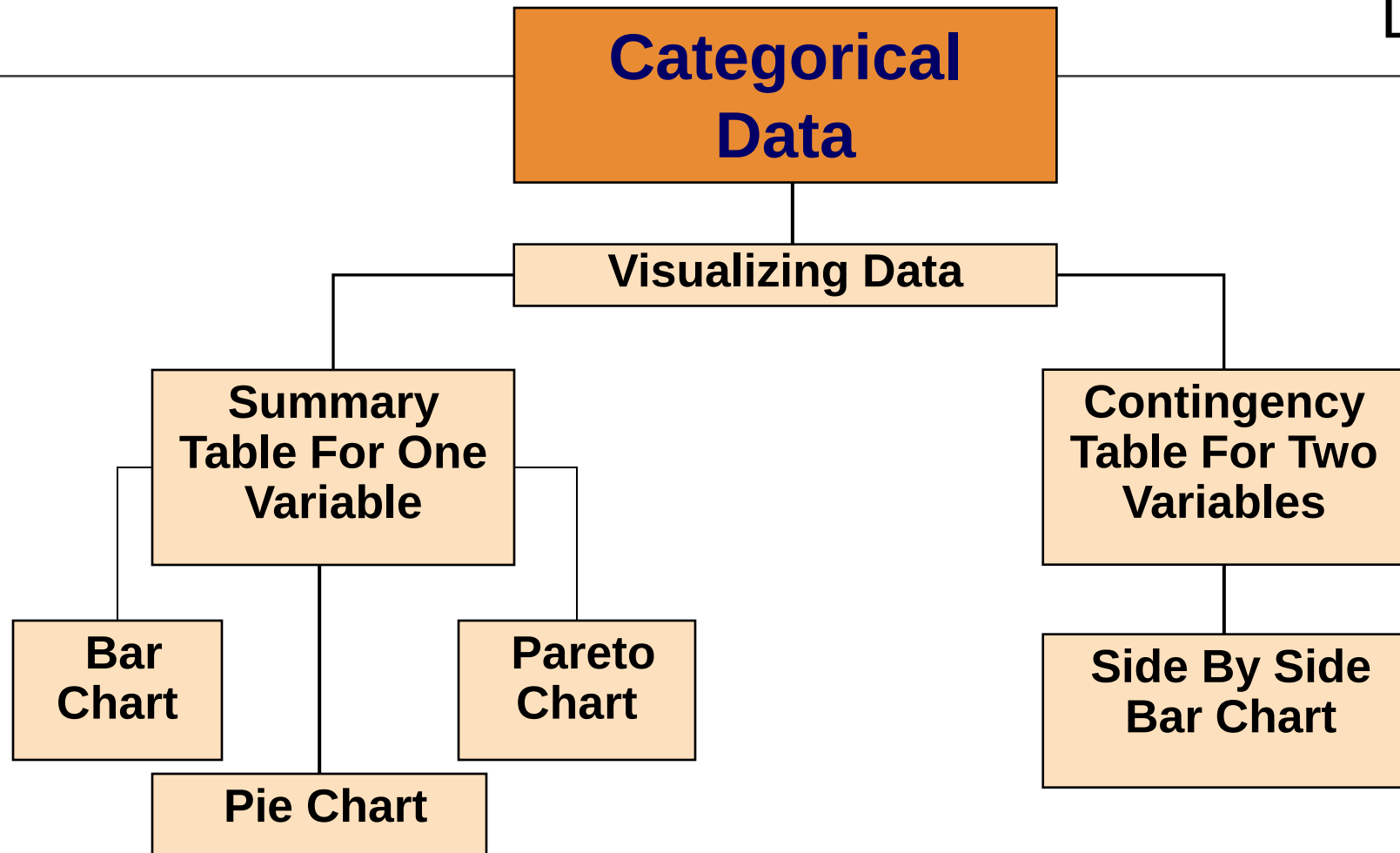
12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15%	3	15%
20 but less than 30	6	30%	9	45%
30 but less than 40	5	25%	14	70%
40 but less than 50	4	20%	18	90%
50 but less than 60	2	10%	20	100%
Total	20	100	20	100%

May2024 _HK

Visualizing Categorical Data Through Graphical Displays

DCOVA



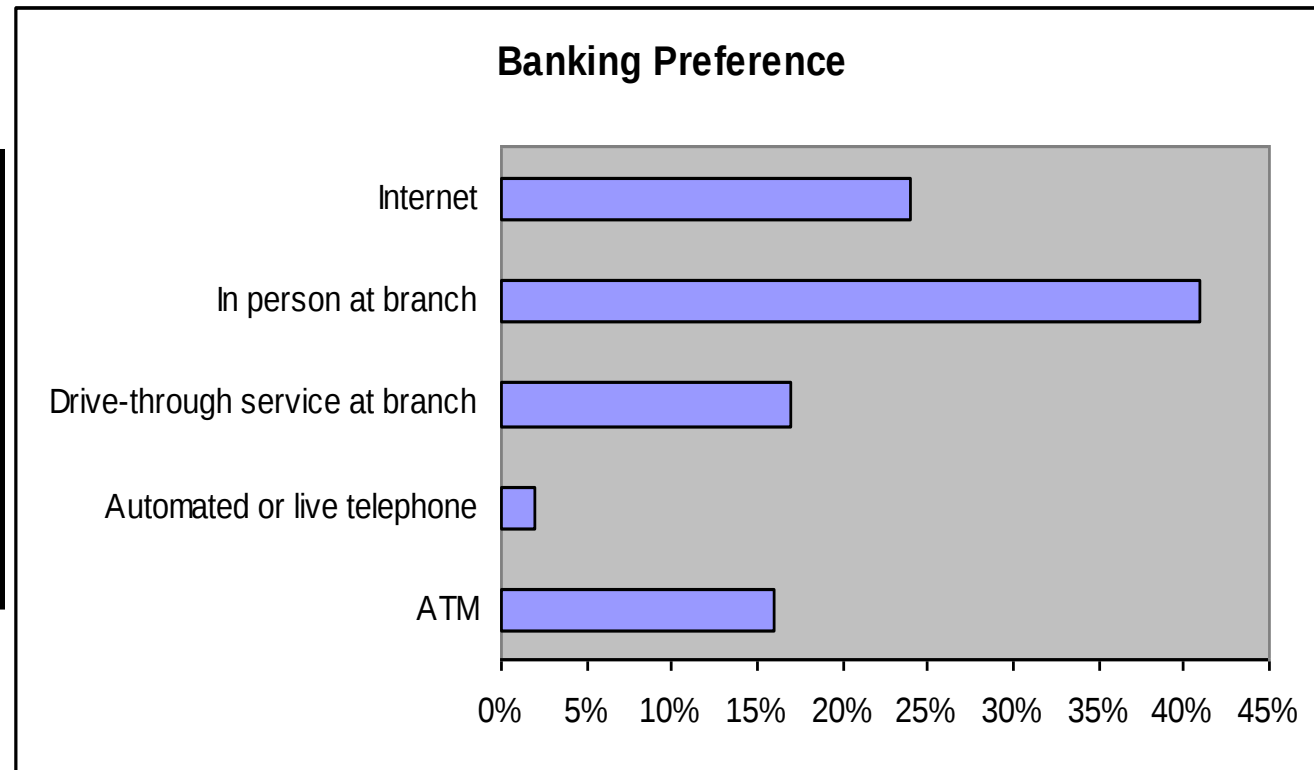
May2024 _HK

Visualizing Categorical Data: The Bar Chart

DCOVA

- In a **bar chart**, a bar shows each category, the length of which represents the amount, frequency or percentage of values falling into a category which come from the summary table of the variable.

Banking Preference?	%
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

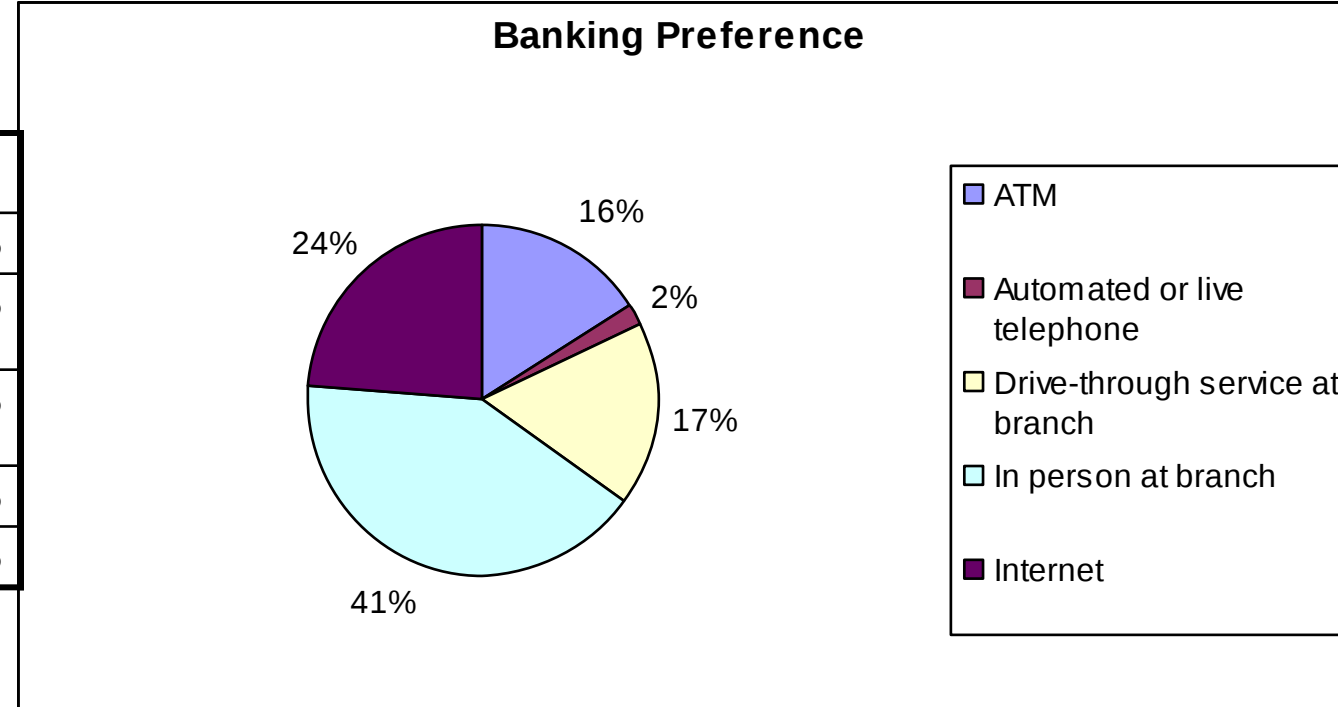


Visualizing Categorical Data: The Pie Chart

DCOVA

- The **pie chart** is a circle broken up into slices that represent categories. The size of each slice of the pie varies according to the percentage in each category.

Banking Preference?	%
ATM	16%
Automated or live telephone	2%
Drive-through service at branch	17%
In person at branch	41%
Internet	24%

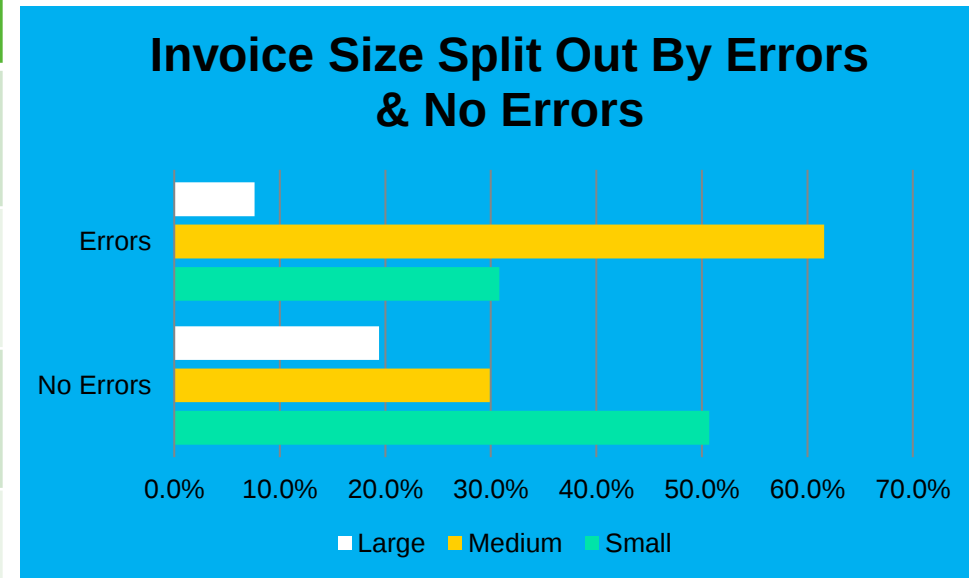


Visualizing Categorical Data: Side By Side Bar Charts

DCOVA

- The **side by side bar chart** represents the data from a contingency table.

	No Errors	Errors	Total
Small Amount	50.75%	30.77%	47.50%
Medium Amount	29.85%	61.54%	35.00%
Large Amount	19.40%	7.69%	17.50%
Total	100.0%	100.0%	100.0%

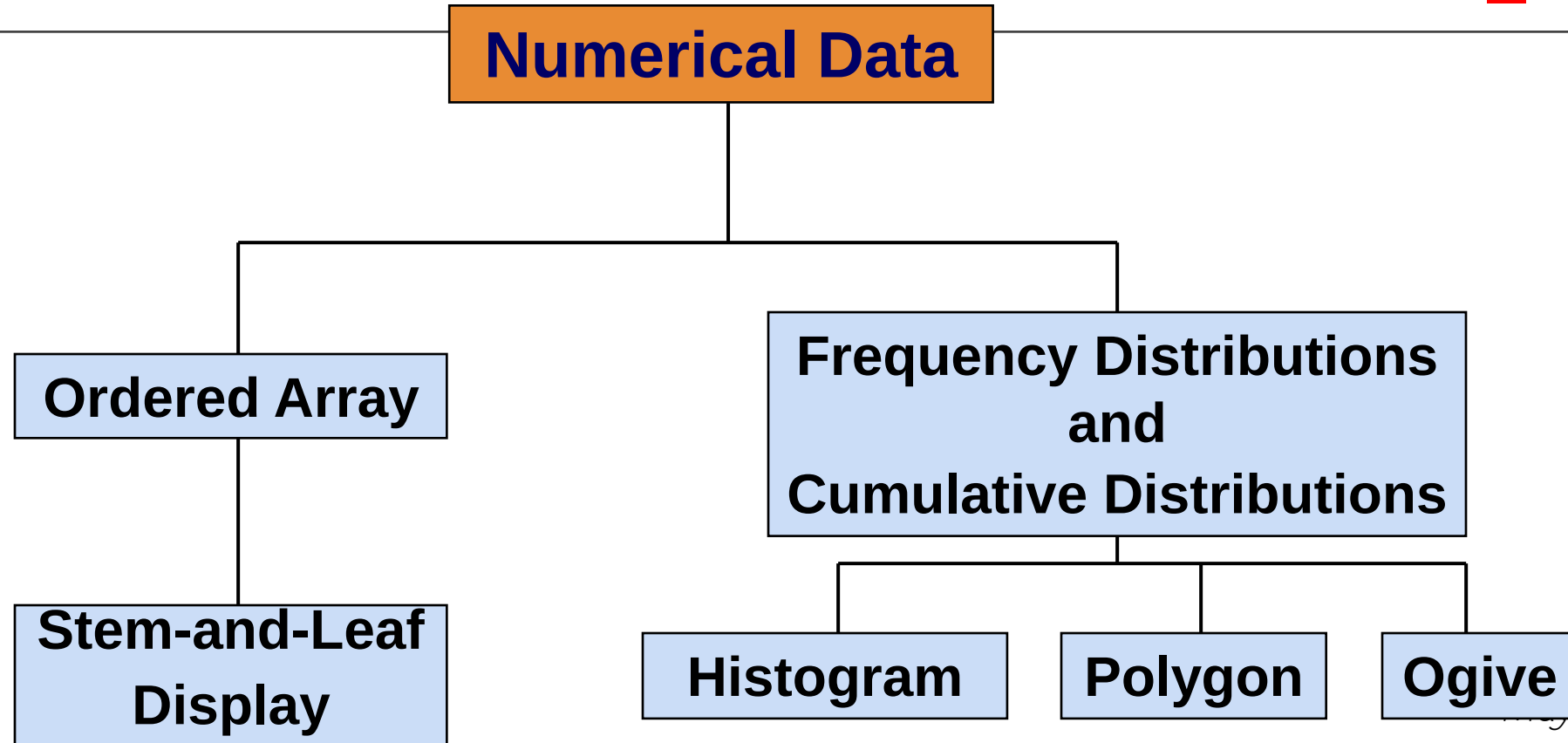


Invoices with errors are much more likely to be of medium size (61.54% vs 30.77% and 7.69%)

May2024 _HK

Visualizing Numerical Data By Using Graphical Displays

DCOVA



2024 _HK

Visualizing Numerical Data: The Histogram

DCOVA

- A vertical bar chart of the data in a frequency distribution is called a **histogram**.
- In a histogram there are no gaps between adjacent bars.
- The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
- The vertical axis is either **frequency**, **relative frequency**, or **percentage**.
- The height of the bars represent the frequency, relative frequency, or percentage.

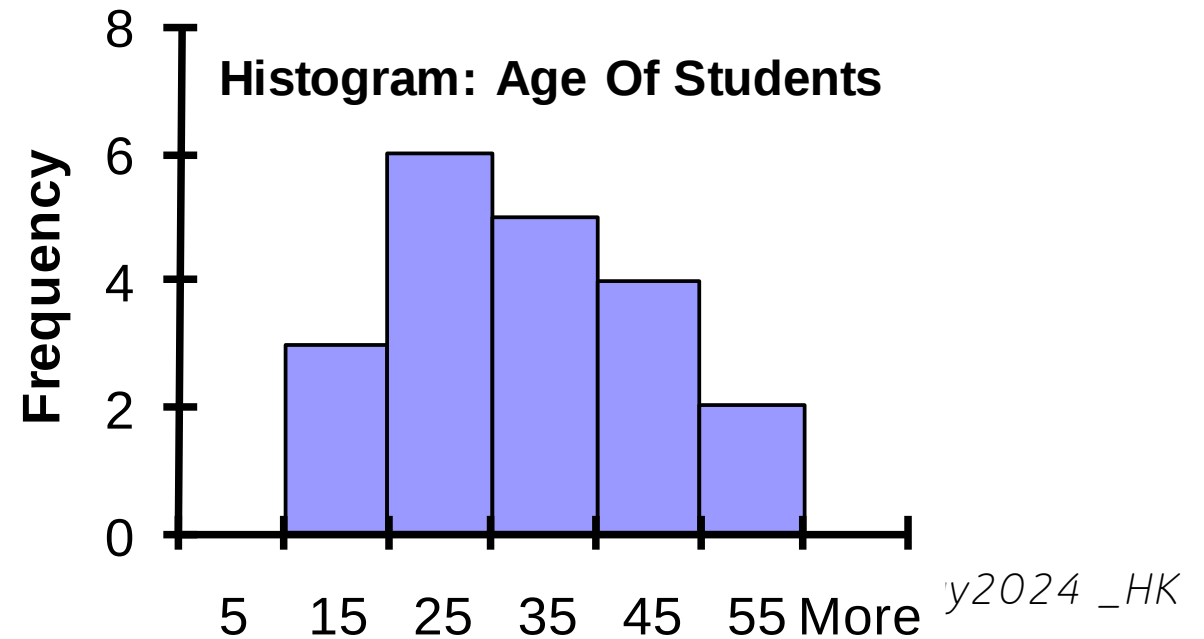
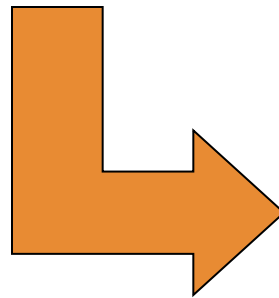
May 2024 _HK

Visualizing Numerical Data: The Histogram

DCOVA

Class	Frequency	Relative Frequency	Percentage
10 but less than 20	3	.15	15
20 but less than 30	6	.30	30
30 but less than 40	5	.25	25
40 but less than 50	4	.20	20
50 but less than 60	2	.10	10
Total	20	1.00	100

(In a percentage histogram the vertical axis would be defined to show the percentage of observations per class)



Visualizing Numerical Data: The Polygon

DCOVA

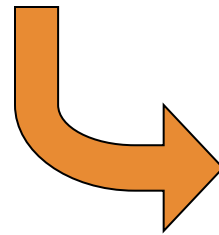
-
- A **percentage polygon** is formed by having the midpoint of each class represent the data in that class and then connecting the sequence of midpoints at their respective class percentages.
 - The **cumulative percentage polygon**, or **ogive**, displays the variable of interest along the X axis, and the cumulative percentages along the Y axis.
 - Useful when there are two or more groups to compare.

May2024 _HK

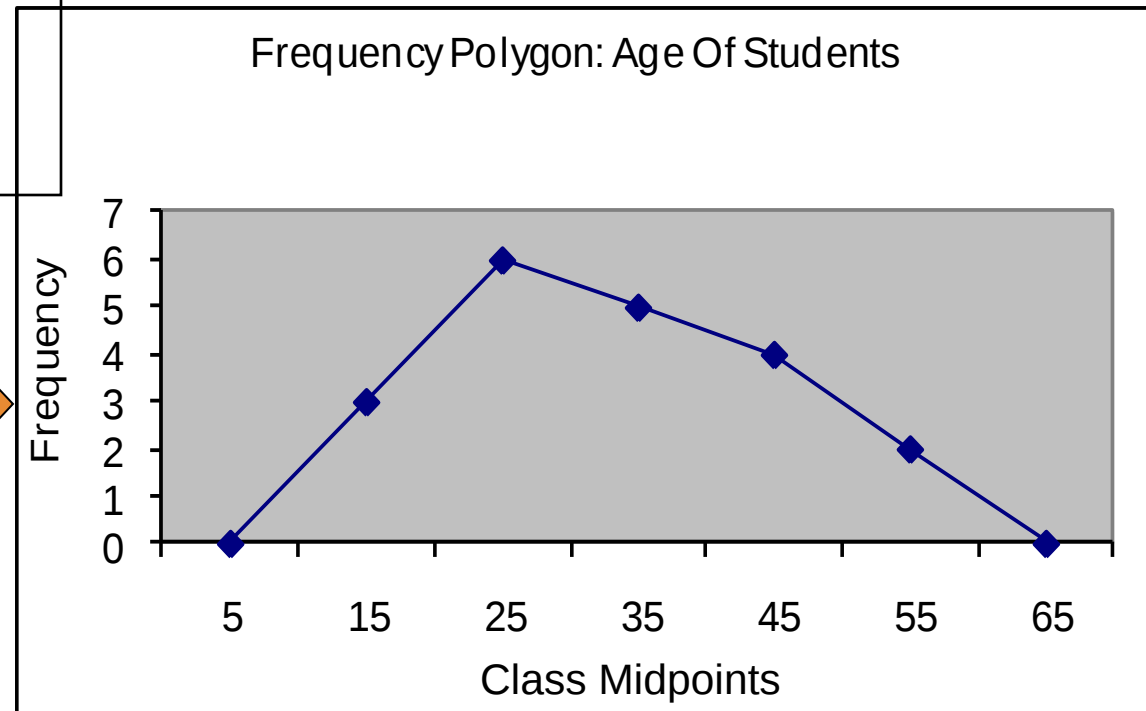
Visualizing Numerical Data: The Frequency Polygon

DCOVA

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2



(In a percentage polygon the vertical axis would be defined to show the percentage of observations per class)

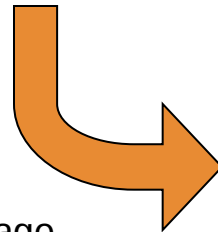


2024 _HK

Visualizing Numerical Data: The Ogive (Cumulative % Polygon)

DCOVA

Class	Lower class boundary	Cumulative percentage
10 but less than 20	10	15
20 but less than 30	20	45
30 but less than 40	30	70
40 but less than 50	40	90
50 but less than 60	50	100



(In an ogive the percentage of the observations less than each lower class boundary are plotted versus the lower class boundaries.)

