ISIMS - UNIVERSITY OF SFAX
LSI ADBD - BACHELOR 2
2023/2024

***

DESCRIPTIVE AND INFERENTIAL STATISTICS

***

LECTURE I

ORGANIZING AND VISUALIZING VARIABLES

Lecture Content

1. Introduction
   - General Definitions
   - Data Sources and data types
2. Categorical data
   - Organizing Categorical Variables
   - Visualizing Categorical Variables
3. Numerical data
   - Organizing Numerical Variables
   - Visualizing Numerical Variables
   - Visualizing two Numerical Variables

General Definitions

Statistics: Statistics is the discipline that provides formal basis to summarize and visualize data, reach conclusions about the data, make reliable predictions and improve decision process. Statistics are specifically the methods that allow you to work with data effectively.

Data: Data are the facts about the world that one seeks to study and explore. Some data are unsummarized whereas other are summarized. Data can be numbers or non-numerical.

A variable in Statistics: Defines a characteristic, or property of an item or individual that can vary among the occurrences of those items or individuals. For example for the item ”employee” variables would include age, wage, occupation, number of schooling years, marital status,..

A Statistic: A function of data that summarizes a particular variable.
Example: The mean size of LSI Students.

General Definitions

Population: The population contains all the items or individuals of interest that you seek to study. Examples:

- To study the wages in Tunisia your population is all the Active Tunisians
- To study the school performance of High School students in UK, you population contains all the students enrolled in High School in UK at that time.
- To Study the election statistics in Ohio, your population are all the registred voters

Sample: Contains only a portion of population of interest. You analyse a sample to estimate charactertics of an entire population. You may select a sample of 1000 Tunisian workers, a sample of 200 High School students or a sample of 500 registrered voters in Ohio, instead of analysing the entire corresponding populations.

General Definitions

A Statistical Individual: An individual is an item in the data. It can be a person (a student, an employee, a citizen, a voter, a player,..), a company, a country, a good, a financial stock, a team, an advertisment,.. In a data table an individual corresponds to a row. All the information about the same statistical individual is given in a row. Columns correspond to variables.

Descriptive Statistics: The methods that primely help organizing and summarizing the data in ways that facilitate its interpretation and subsequent analysis.

Inferential Statistics: Statistical methods that are used to make decisions and draw conclusions about populations. These techniques utilize the information in a sample for drawing conclusions.

Data Sources and Data Types

Broadly, all variables are either:

- **Numerical:** Variables representing a counted or measured quantity: age, wage, number of children, price, sales, cost, brand fame index, investment amount, GDP, imports,..
- **Categorical:** Variables representing Categories or indexing an event: gender, marital status, transport mode, presence of a star in advertising campaign, level of satisfaction of customs,..

Numerical variable can either be:

- **Discrete:** Data that arise from a counting process, i.e. number of something: monthly number of smartphones sold, number of children, occurrences of an event, Number of Passengers,.. The values of a discrete variable are in a finite and countable set.
- or **Continuous:** Data that arise from a measuring process: Time spent waiting on a line, Life Duration of an Electronic Component, volume of sales, price, GDP, investment, stock price, rate of return, market value of a firm,..

Data Sources and Data Types

Categorical variable can either be:

- **Nominal:** Category variables express no order or ranking: Gender, Transport Mode, Device used to watch movies, Industry or Business Activity,..
- or **Ordinal:** An ordering or ranking of category values are implied: Level of Satisfaction (from 1 to 5), Risk Level (High, Average, Low), Level of Protection of a Face Mask,..

Data Sources and Data Types

Variable	**Set of values**	**Type**
Cellular Provider	Orange, Tunisie Telecom, Oreedoo	Categorical, Nominal
Student's Excel Skills	Weak, Average, Good, Outstanding	Categorical, ordinal
Age of students in years	A number between 17 and 28	Numerical, Discrete
Sales in Millions	A positive real number	Numerical, Continuous
Gender	Male, Female	Categorical, Nominal
Pound per Capita in Fresh food	A positive Real number	Numerical, Continuous
Industry of a Firm	Automobiles, Banks, Capital Goods, Consumer Services, Energy,..	Categorical, Nominal

Data Sources

Data sources arise from the following activities:

- Capturing data generated by ongoing business activities
- Distributing data compiled by an organization or individual
- Compiling the response from a survey
- Conducting a designed experiment and recording the outcomes.
- Conducting an observational study and recording its results.

CATEGORICAL DATA

Organizing Categorical Variables - The Summary table

The **summary table** helps you see the difference among the categories by displaying the frequencies, amount or percentage of items in a set of categories.

Example 1: The table below gives responses to a recent survey that asked those born between 1982 and 2001, which devices they used to watch movies or TV shows.

Device	**Percentage**
Laptop/desktop	32%
Smartphone	10%
Tablet	9%
Television set	49%

Exercise 1: A survey asked 40 college students majoring in business: What is your major? (A= Accounting; C= Computer Information Systems; M= Marketing). Answers are: A C C M A C A A A C C A A A M C M A A A C C C A A M M C A A A C C A A A A C C A C Give the summary table of frequencies.

Organizing Categorical Variables - The Contingency Table

The **Contingency tables** cross-tabulates (or tallies jointly), the data of two or more categorical variables, allowing you to study the patterns that may exist between the variables. Tallies can be shown as a frequency, a percentage of overall total, a percentage of the row total or a percentage of the column total

Example 2: Investors classify the financial market stocks into two types: Growth Stock and Value Stock. They also rank the assets risk into three categories: Low, Average and High. The table below tallies a sample of 479 retirement funds on the basis of two categorical variables: Fund Type and Risk Level.

	Risk Level			
Fund Type	**Low**	**Average**	**High**	**Total**
Growth	63	152	91	306
Value	84	72	17	173
All	147	224	108	479

Organizing Categorical Variables - The Contingency Table

Exercise 2: The following data represent the responses to two questions asked in a survey of 40 college students majoring in business: What is your gender ? (M=Male; F=Female) and what is your major?

Gender	M	M	M	F	M	F	F	M	F	M	F	M	M	M	M	F	F	M	F	F
Major	A	C	C	M	A	C	A	A	C	C	A	A	A	M	C	M	A	A	A	C
Gender	M	M	M	M	F	M	F	F	M	M	F	M	M	M	M	F	M	F	M	M
Major	C	C	A	A	M	M	C	A	A	A	C	C	A	A	A	A	C	C	A	C

1. Tally the data into a contingency table where rows represent the gender categories and columns represent the academic major categories.
2. Construct a contingency table based on percentages of all 40 students responses.

Organizing Categorical Variables - The Contingency Table

Exercise 3: A survey of 1,520 Americans adults asked "Do you feel overloaded with too much information?". The results are given in this table.

Overloaded	**Male**	**Female**	**Total**
Yes	134	170	304
No	651	565	1,216
Total	785	735	1,520

1. Construct contingency tables based on total percentages, row percentages and column percentages.
2. What conclusions can you reach from these analyses?

Visualizing Categorical Variables - The Bar Chart

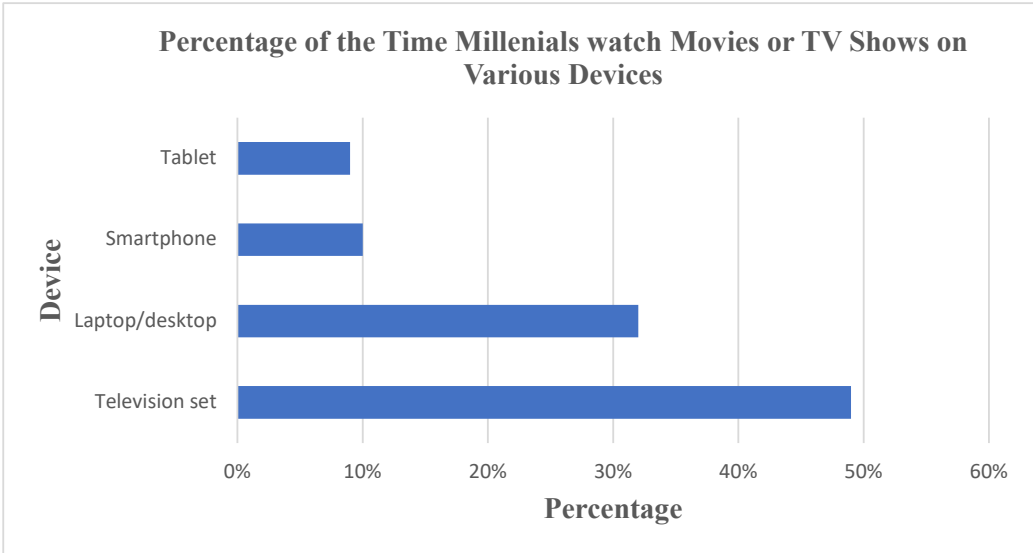
When you visualize a single categorical variable, you must think about what you want to highlight and whether your data are concentrated in only a few of your categories.

- To highlight how categories directly compare to each other, you use a **bar chart**.
- To highlight how categories form parts of a whole, you use a ***pie chart***.
- To present data that are concentrated in only a few of your categories, you use a **Pareto chart**.

The Bar Chart: The bar chart visualize a categorical variable as a series of bars, with each bar representing the tallies for a single category. The length of each bar represents either the frequency or percentage of values for a category and each bar is represented by a space called a *gap*.

Visualizing Categorical Variables - The Bar Chart

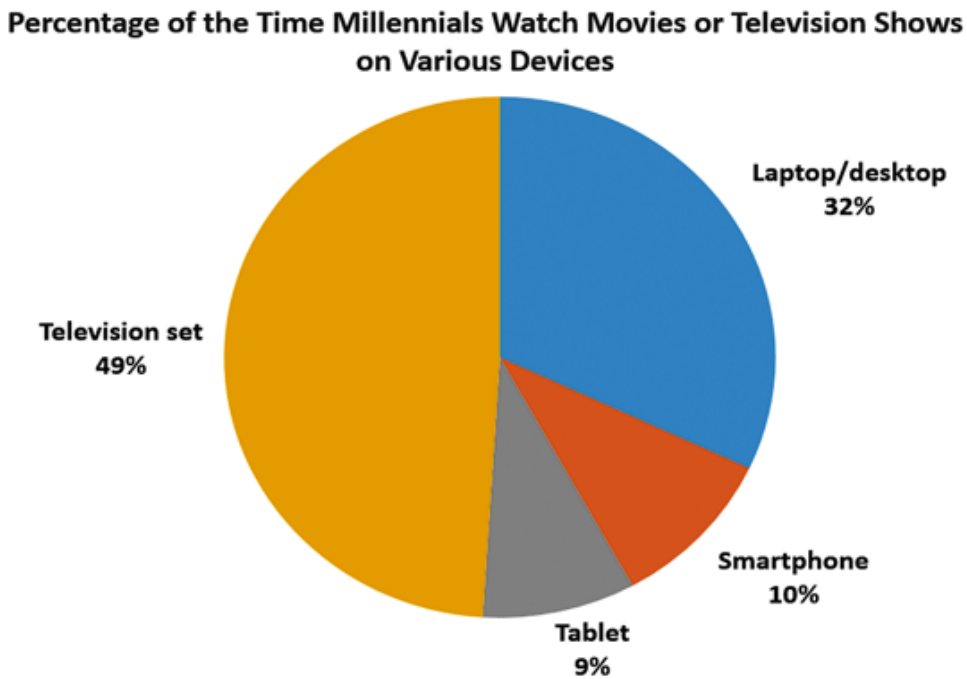
Example 3: The figure below represents bar charts of the table displayed in example 1.



By viewing this bar chart, we can make the same conclusion as reviewing the summary table: About half of the individuals watch movies and TV shows on television set and half do not. With complex data, visualization will generally allow you to discover relationship among items faster than the equivalent tabular summaries.

Visualizing Categorical Variables - The Pie Chart

The pie chart represent the tallies of each category of a categorical variable as a part of a circle. These parts, or ***slices***, vary by the percentages of the whole for each category.
Multiplying category percentages by 360 gives the number of degrees in a circle, which determines the size of each slice.
Example 4: The figure below represents pie charts of the table displayed in example 1.



Visualizing Categorical Variables - The Pareto Chart

Pareto charts help identify the categories that contain the largest tallies from the categories that contain the smallest. The economist Pareto used these charts to visualize the principle that 80% of the consequences result from 20% of the causes.
In quality management efforts, Pareto charts are very useful tools for prioritizing improvement efforts, such as when data that identify defective or non conforming items are collected.

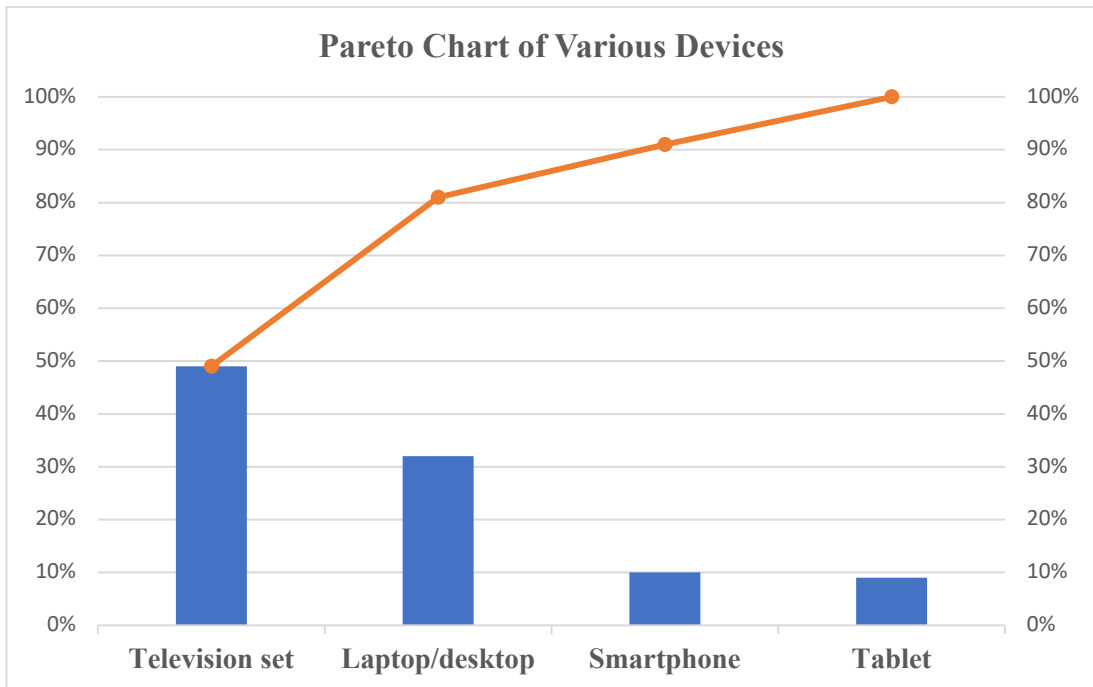
Pareto charts combine two different visualizations: a vertical bar chart and a line graph connecting points.

- the vertical bars represent the tallies for each category arranged in the descending order
- the line graph represent a cumulative percentage of the tallies from the first category through the last one.

Visualizing Categorical Variables - The Pareto Chart

Example 5: Pareto chart from the summary table of example 1.

- First we construct a new table in which the categories are ordered by descending frequencies, including the columns for percentages and cumulative percentages.
- Create a Pareto chart.



We conclude the televisions and computers together account for over four-fifths of all such viewing.

Visualizing Categorical Variables - The Pareto Chart

Exercise 4: Consider a bank study team that wants to enhance the user experience of automated teller machines (ATMs). The team identifies incomplete ATM transactions as a significant issue. It decides to collect data about the causes of such transactions using the bank's own processing systems. Data are organized in the table below.

Cause	**Frequency**
ATM malfunctions	32
ATM out of cash	28
Invalid amount requested	23
Lack of funds in account	19
card unreadable	234
Warped card jammed	365
Wrong keystroke	23
Total	**724**

Using Excel, construct a Pareto chart of causes of incomplete ATM transactions.

Visualizing Categorical Variables - Exercises

Exercise 5: The Consumer Financial Protection Bureau reports complaints received from Louisiana consumers. The **Table1** of Excel Sheet **DataComplaints** gives the number of complaints by category for 2016.

1. Construct a Pareto chart for the categories of complaints.
2. Discuss the "vital few" and "trivial many" reasons for the categories of complaints.

The **Table2** of the same Excel Sheet gives the tally of the complaints received from Louisiana consumers by most-complained-about companies for 2016.

3. Construct a bar chart and a pie chart for the complaints by company.
4. What graphical method (Pareto, bar, or pie chart) do you think is best for portraying these data.

Visualizing Categorical Variables - Exercises

Exercise 6: Table od Excel Sheet **DataElectricity** indicates the percentage of residential electricity consumption in the Unites States, in a recent year organized by type of use.

1. Construct a bar chart, a pie chart, and a Pareto chart.
2. Which graphical method do you think is best for portraying these data ?
3. What conclusions can you reach concerning residential electricity consumption in the United States?

NUMERICAL DATA

Organizing Numerical Variables - The Frequency Distribution

A **frequency distribution** tallies the values of a numerical variable into a set of numerically ordered classes. Each class groups a mutually exclusive range of values, called a **class interval**.
To create a useful frequency distribution, you must consider:

- How many classes would be appropriate for your data: Generally at least 5 and no more than 15.
- What is the suitable ***width*** for each class interval.

We determine the class Interval width by using this formula:

$$\text{Interval width} = \frac{\text{highest value} - \text{lowest value}}{\text{number of classes}}$$

Example 6: Excel Sheet **Data3** gives the meal cost data collected from a sample of 50 center city restaurants and 50 metro area restaurants. Organize your data in a frequency distribution table and interpret the main features shown by this table.

Organizing Numerical Variables - The Relative Frequency Distribution

A **relative frequency distribution** presents the relative frequency, or proportion, of the total for each group that each class represents. It can also be given by a **percentage distribution**.
We determine the proportion and the percentage by using this formula:

$$\text{Proportion} = \text{relative frequency} = \frac{\text{number of values of each class}}{\text{total number of values}}$$

$$\text{Percentage distribution} = \frac{\text{number of values of each class}}{\text{total number of values}} \times 100$$

Example 7: Using Excel Sheet **Data3**, give the relative frequency and percentage for each group of meal cost in the center city and metro area restaurants.

Organizing Numerical Variables - The Cumulative Distribution

The **cumulative percentage distribution** gives the percentage of values that are less than a specific amount. To construct a cumulative percentage distribution you need the percentage distribution.

Example 8:

- Use the percentage distribution of the previous example in order to construct for each class the percentage of meal costs that are less than the class interval lower boundary.
- Give the cumulative percentage distribution corresponding to the values of the lower class boundaries from the class intervals.

Organizing Numerical Variables - Exercises

Exercise 7: The Excel Sheet **Data5** contains the time in seconds to answer 50 incoming calls to a financial services call center.

1. Construct a frequency distribution and a percentage distribution.
2. Construct a cumulative percentage distribution.
3. The service target level is set at "80% of calls is answered within 20 seconds". What do you conclude about call center performance ?

Exercise 8: The file **Data6** contains average age of the players (years, in 2018) of the 32 teams that qualified for the FIFA 2018 World Cup.

1. Organize these mean ages in an ordered array.
2. Construct a frequency distribution and a percentage distribution for these mean ages.
3. Around which class grouping, if any, are these mean ages concentrated? Explain.

Organizing Numerical Variables - Exercises

Exercise 9: The Excel Sheet **Data7** contains the cost of electricity (in euros) during July 2017 for a random sample of 50 one-bedroom appartments in a large city.

1. Construct a frequency distribution and a percentage distribution that have class intervalls with the upper class boundaries 99,119, and so on.
2. Construct a cumulative percentage distribution.
3. Around what amount does the electricity cost seem to be concentrated ?

Visualizing Numerical Variables - Stem-and-Leaf Display

The **Stem-and-Leaf** display (or diagram) is a good way to obtain an informative visual display of a data set where each number consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps.

- Divide each number into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
- List the stem values in a vertical column.
- Record the leaf for each observation beside its stem.
- Write the units for stems and leaves on the display.

Stem-and-leaf displays allows you to see how the data are distributed and where concentrations of data exist. Leaves typically present the last significant digit of each value, but sometimes you round values.

Visualizing Numerical Variables - Stem-and-Leaf Display

Example 9: Suppose you collect the following meal costs (in euros) for 16 classmates who had lunch at a fast food restaurant:

7.42	6.29	5.83	6.50	8.34	9.51	7.10	6.80
5.90	4.89	6.50	5.52	7.90	8.30	9.60	6.87

The complete Stem-and-leaf display of these data with the leaves ordered within each stem is:

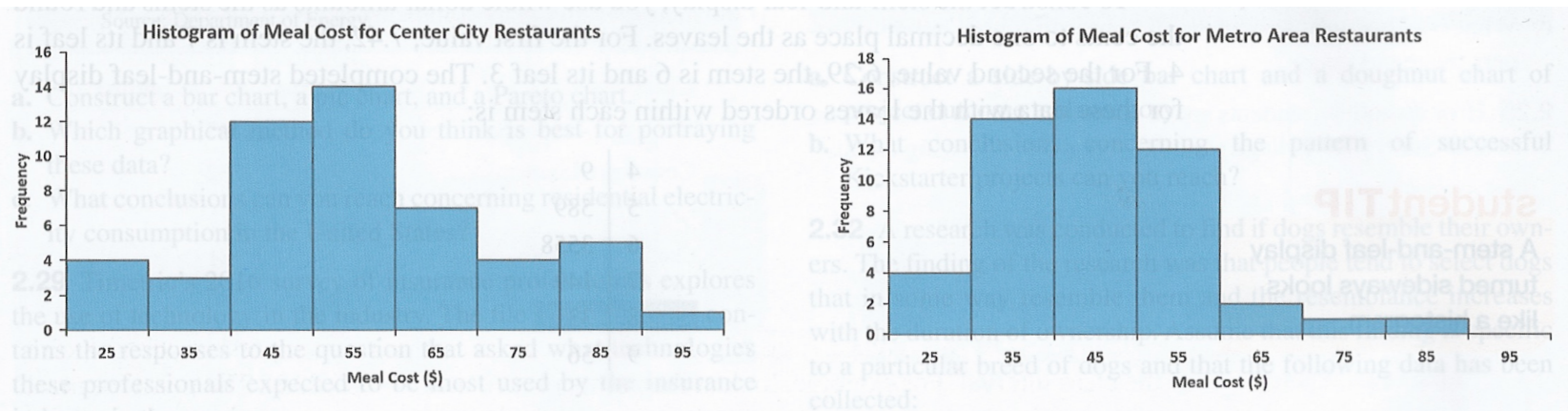
4	9
5	589
6	35589
7	149
8	33
9	56

Exercise 11: Consider the compressive strength data of 80 aluminum specimens given in data sheet **Data8**. Illustrate these data in a Stem-and-leaf diagram.

Visualizing Numerical Variables - The Histogram

- A **Histogram** visualizes data as a vertical bar chart in which each bar represents a class interval from a frequency or percentage distribution.
- In a histogram you display the numerical variable along the horizontal (X) axis and the vertical (Y) axis to represent either the frequency or the percentages of values per class interval.
- There are never any gaps between adjacent bars in a histogram

Example 10: Meal Costs Histograms

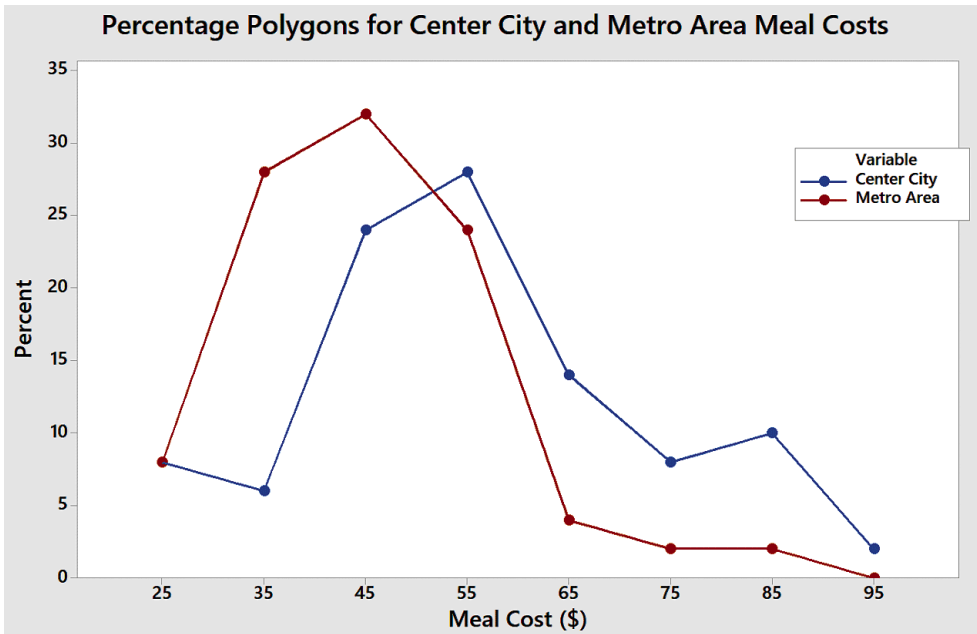


Visualizing Numerical Variables - The Percentage Polygon

The **percentage polygon**

- Midpoint of each class interval represents the data of each class.
- Plot the midpoint at their class percentage
- Connect the points with a Polygon line

Example 11: Meal Costs Percentage Polygon

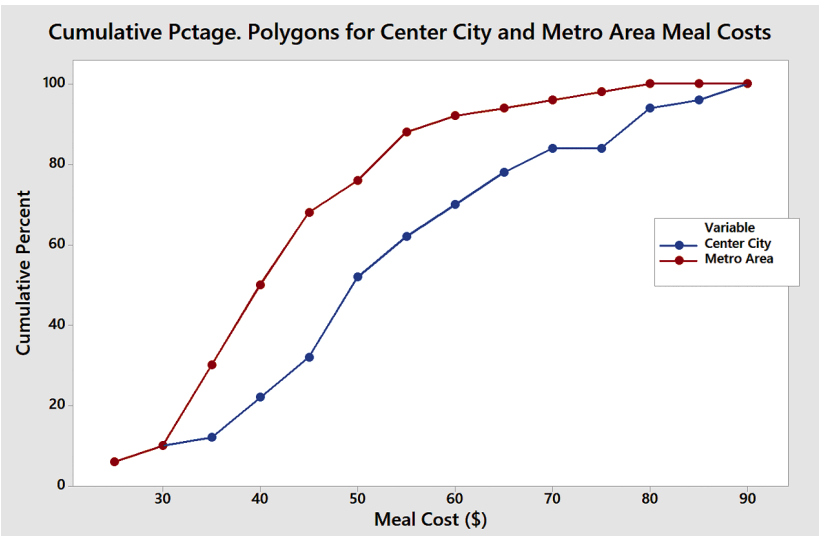


Visualizing Numerical Variables - The Cumulative Percentage Polygon

The **cumulative percentage polygon** uses the cumulative percentage distribution discussed in the previous section.

- Lower boundaries of the class interval for the numerical variable are plotted (Unlike the percentage polygon).
- Plot the lower boundaries at their respective class cumulative percentages.
- Connect the points with a polygon line.

Example 12: Meal Costs Cumutative Percentage Polygon



Visualizing Numerical Variables - Exercises

Exercise 12: Use again the data on three-year return percentage variable from the sample of 479 retirement funds, given in the Excel Sheet **Data4**. Display for each of the growth and the value funds:

1. Frequency histograms.
2. Percentage polygons.
3. Cumulative percentage polygons

Exercise 13: Take again the Excel Sheet **Data5** containing time to answer 50 incoming calls.

1. Construct a percentage histogram and a percentage polygon.
2. Construct a cumulative percentage polygon.
3. The service target level is set at "80% of calls is answered within 20 seconds". What do you conclude about call center performance ?

Visualizing Numerical Variables - Exercises

Exercise 14: Take again the Excel Sheet **Data6** on the average age of players.

1. Construct a stem-and-leaf display.
2. Around which value, if any, are the mean ages of teams concentrated? Explain.

Exercise 15: Take again the Excel sheet **Data7** on the cost of electricity during July 2017.

1. Construct a histogram and a percentage polygon.
2. Construct a cumulative percentage polygon.
3. Aroun what amount does the monthly electricuty cost seem to be concentrated ?

TWO NUMERICAL VARIABLES

Visualizing Two Numerical Variables - The Scatter Plot

To visualize two numerical variables you use a **scatter plot**. For the special case in which one of the two variables represents the passage of time, you use a **time series plot**.

A scatter plot explores the possible relationship between two numerical variables by plotting the values of one numerical variable on the X-axis and the values of a second numerical variable on the Y-axis.

Example 13: You seek to know if the value of a professional NBA basketball team reflects its revenues. Excel Sheet **Data9** gives revenues and valuation data (both in $millions) for all 30 NBA teams. To quickly visualize a possible relationship between team revenues and valuations, construct a scatter plot in which you plot the revenues on the X-axis and the valuations on the Y-axis.

Visualizing Two Numerical Variables - The Scatter Plot

Exercise 16: Movies companies need to predict the gross receipts of individual movies once a movie has debutes. Table in **Data10** gives the first weekend gross, the US gross, and the worldwide gross (in $millions) of the eigth Harry Potter movies.

- Construct a scatter plot with first weekend gross on the X-axis and U.S gross on the Y-axis.
- Construct a scatter plot with first weekend gross on the X-axis and worldwide gross on the Y-axis.
- What can you say about the relationship between first weekend gross and U.S gross, and between first weekend gross and worldwide gross.

Visualizing Two Numerical Variables - The Time-Series Plot

A **time-series plot** plots the values of a numerical variable on the Y-axis and plots the time period associated with each numerical value on the X-axis. A time-series plot can help you visualize trends in data that occur over time.

Example 13: As an investment analyst who specializes in the entertainment industry, you are interested in discovering any long-term trends in movie revenues. You collect the annual revenues (in $billions) for movies released from 1995 to 2016, organize the data as table in Excel Sheet **Data11**.
To see if there is a trend over time, construct the time-series plot of these data.

Visualizing Two Numerical Variables - The Time-Series Plot

Exercise 17: Excel Sheet **Data12** shows the performance of a broad measure of stocks (by percentage) for each decade from 1930s through the 2000s.

1. Construct a time-series plot of the stock performance from the 1830s to the 2000s.
2. Does there appear to be any pattern in the data ?

Exercise 18: The table in Excel Sheet **Data13** contains the yearly movie attendance (in billions) from 2001 to 2016.

1. Construct a time-series plot for the movie attendance (in billions).
2. What pattern, if any, is present in the data?