

Descriptive Statistics and Data Analysis

Prepared by Dr. Mohamed Sleem

Subject code : FNT 631
Data Analysis

Lecture 3



Descriptive Statistics and Data Analysis

Lecture 3

Subject code : FNT 631
Data Analysis

Prepared by
Dr. Mohamed Sleem

Chapter

01

Introduction to Descriptive Statistics and Data Analysis



Answer Chapter 2 Revision Questions

Compare and contrast three different data visualization programs, highlighting their key features and capabilities.

1. Tableau:

- User-Friendly Interface: Easy to learn and use without extensive programming knowledge.
- Data Integration: Connects to various data sources like databases, Excel, and cloud data.
- Real-Time Data Updates: Automatically updates data from multiple sources.
- Advanced Analytics: Supports predictive analytics, statistical analysis, and forecasting.
- Interactive Dashboards: Create and share interactive visualizations easily.
- Security: Offers advanced security features to protect sensitive data.

2. Microsoft Power BI:

- Integration with Microsoft Ecosystem: Integrates smoothly with other Microsoft products like Excel, Azure, and SharePoint.
- Ease of Use: Intuitive and easy-to-navigate interface with drag-and-drop functionality.
- Real-Time Data Access: Provides real-time dashboards and automatic data refreshes.
- Advanced Data Modeling: Uses powerful DAX language for complex calculations and built-in Power Query for data transformation.
- Interactive Visualizations: Dynamic visuals with drill-through capabilities.

3. Google Data Studio:

- Free to Use: Completely free with a Google account.
- Google Ecosystem Integration: Seamlessly integrates with Google Analytics, Google Ads, and BigQuery.
- Customizable Reports: Highly customizable and visually appealing reports and dashboards.
- Real-Time Data: Offers real-time data updates and live data connections.
- Collaboration: Features easy sharing and collaboration.

Answer Chapter 2 Revision Questions

2. Question: What are the four main strategies of data analysis and briefly explain each?

Answer:

- i. Descriptive Analysis: Summarizes data to show patterns.
- ii. Diagnostic Analysis: Explains why something happened.
- iii. Predictive Analysis: Predicts future trends based on data.
- iv. Prescriptive Analysis: Suggests actions to achieve desired outcomes.

3. Question: Outline the steps involved in the data visualization process.

Answer:

- Data Collection: Gather relevant data from various sources.
- Data Preparation: Clean and preprocess the data to ensure accuracy and consistency.
- Data Analysis: Explore the dataset to identify patterns, trends, and relationships using statistical methods and visualization techniques.
- Visualization Design: Select appropriate visualization types based on the data characteristics and analysis goals.
- Visualization Creation: Create visualizations using data visualization tools or programming languages.
- Interpretation: Interpret the visualizations to extract meaningful insights and actionable recommendations.
- Communication: Present the visualizations and findings to stakeholders through reports, dashboards, presentations, or interactive tools, ensuring clear and effective communication of insights.

Chapter
03

Chapter's Objectives

1. Understand the concept and importance of descriptive statistics in data analysis.
2. Identify key descriptive statistical measures and their role in summarizing datasets.
3. Explore real-life applications of descriptive statistics in business, healthcare, and education.
4. Integrate descriptive statistics into the broader data analysis process to make informed decisions.

Introduction to Descriptive Statistics and Data Analysis

1. Descriptive statistics serve as fundamental tools in analyzing and interpreting data.
2. They play a crucial role in various fields, including business, healthcare, and social sciences.
3. Descriptive statistics help in summarizing and presenting data in a meaningful way.

Introduction to Descriptive Statistics and Data Analysis

4. Understanding descriptive statistics is essential for making informed decisions based on data.

5. This presentation will explore the significance of descriptive statistics and their relationship with data analysis across different sectors.

What are Descriptive Statistics?

- **Definition:** Descriptive statistics are numerical and graphical techniques used to summarize and describe the main features of a dataset.
- Descriptive statistics provide concise summaries about the characteristics of data samples.

What are Descriptive Statistics?

Examples:

- **Mean:** The average value of a dataset.
- **Median:** The middle value of a dataset when arranged in ascending order.
- **Mode:** The value that appears most frequently in a dataset.
- **Standard Deviation:** A measure of the dispersion or spread of data around the mean.
- **Range:** The difference between the maximum and minimum values in a dataset.

What are Descriptive Statistics?

Descriptive Statistics

- Organise
- Summarise
- Simplify
- Describe and present data

Inferential Statistics

- Generalise from samples to populations
- Hypothesis testing
- Make predictions

Visual Tools: **Descriptive** statistics are often represented visually through graphs, charts, and tables.

Graphs: Histograms, bar charts, pie charts, box plots, and scatter plots are commonly used to visually represent descriptive statistics.

Charts and Tables: Tables can provide a detailed breakdown of descriptive statistics, while charts offer visual representations for easier interpretation.

1. Trend Identification:

Example: Using moving averages in algorithmic trading to identify long-term market trends and make informed trading decisions.

2. Credit Risk Assessment:

Example: Calculating the average credit score to determine the creditworthiness of borrowers, minimizing the risk of defaults.

3. Performance Benchmarking:

Example: Comparing the average return of a fintech product to market indices like the S&P 500 to gauge performance.

4. Portfolio Risk Management:

Example: Analyzing the standard deviation of returns in a portfolio to understand the average risk level and adjust asset allocation accordingly.

5. Customer Segmentation:

Example: Using mean and median income levels to segment customers for targeted financial products and services, improving marketing efficiency.

Key Descriptive Statistical Measures:

Central Tendency:

Mean (μ): The mean, also known as the average, is the sum of all values divided by the number of values.

$$\text{Mean}(\mu) = \frac{\sum X}{N}$$

Example in FinTech:

Algorithmic Trading:

Application: Mean is used to calculate the average price of a stock over a specific period.

Example: For stock prices over five days: \$100, \$105, \$102, \$108, \$110, the mean price is: $\text{Mean} = \{100 + 105 + 102 + 108 + 110\} / 5 =$

105
 $\text{Mean} = (100 + 105 + 102 + 108 + 110) / 5 = 105$

Importance: This average helps in identifying trends and making buy/sell decisions.

Central Tendency:

Find Mean in Excel

AVERAGE Function
(Calculates Average/Mean)

Data
5
10
15
20
25
30
35
40
45
50

AVERAGE Function
=AVERAGE(A2:A11)
27.5

AVERAGEIF Function
(Calculates Average with Specific Condition)

Product	Zone	Total Sales (\$)
Rice	P	24.38
Wheat	A	21.65
Sugar	C	36.12
Wheat	H	22.26
Sugar	P	40.25
Wheat	P	19.56
Sugar	A	45.32
Sugar	H	42.02
Rice	C	25.85

AVERAGEIF Function
=AVERAGEIF(C2:C10,"Sugar",E2:E10)
40.93

AVERAGEA Function
(Calculates Average with Text String)

Product	Total Sales (\$)
Rice	24.38
Wheat	21.65
Sugar	36.12
Wheat	Not Sold
Sugar	40.25
Wheat	19.56
Sugar	45.32
Sugar	42.02
Rice	Not Sold

AVERAGEA Function
=AVERAGEA(H2:H10)
25.48

Key Descriptive Statistical Measures:

Central Tendency:

Median: The middle value of a dataset when arranged in ascending order.

- **Less Sensitive to Outliers:** Unlike the mean, the median is not affected by extreme values.
- **Realistic Data Representation:** Provides a better central tendency measure for skewed distributions.

Example in Finance:

Income Analysis:

- **Scenario:** A financial firm analyzes client incomes.
- **Data:** \$30,000, \$40,000, \$50,000, \$60,000, \$100,000.
- **Calculation:** Arranged incomes: \$30,000, \$40,000, \$50,000, \$60,000, \$100,000.
 - **Median:** \$50,000.
- **Explanation:** The median income (\$50,000) provides a more accurate picture of a typical client's income than the mean, which could be skewed by the high \$100,000 value.

Median:

Real-Life Application:

- In **Banking**: Identifying the median credit score of loan applicants helps banks understand the creditworthiness profile of their typical client base.
- In **Banking**: Finding the median loan amount requested by customers provides a better representation of the typical loan demand compared to the mean, especially in cases where there are a few extremely large loan requests.

Key Descriptive Statistical Measures:

Median Formula in Excel

File Home Analyze Insert Page Layout Formulas Data Review View Dev

=MEDIAN(

MEDIAN(number1, [number2], ...)

	A	B
1	Car Type	Cost in lakhs
2	Sedan	10
3	SUV	22
4	Sedan	12
5	Hatchback	25
6	SUV	18
7	Crossover	30
8	Convertible	50



Text



Date & Time



Lookup & Reference



Math & Trig



More Functions



Name Manager



Define



Use in



Create



Defined

MAX

MAXA

MAXIFS

MEDIAN

MEDIAN(number1,number2,)

Returns the median, or the number in the middle of the set of given numbers.

[Tell me more](#)

NEGBINOM.DIST

NORM.DIST

Statistical

Engineering

Cube

Information

Compatibility

Web

Key Descriptive Statistical Measures:

Central Tendency:

Mode: The value that appears most frequently in a dataset.

- Identifies Common Data Points: Highlights the most frequent occurrences in the dataset.
- Useful for Categorical Data: Especially valuable for identifying the most common category.

Example in Finance:

Customer Transaction Analysis:

Scenario: A fintech company analyzes transaction types. Data: "Deposit", "Withdrawal", "Deposit", "Transfer", "Deposit". Calculation: The most frequent transaction type is "Deposit". Explanation: Knowing that "Deposit" is the mode helps the company focus on enhancing deposit-related services and promotions.

Mode:

- Real-Life Application:
 - In **Marketing**: Identifying the most frequently purchased product in a store helps retailers understand consumer preferences and optimize inventory management.
 - In **Transportation**: Determining the most common mode of transportation used by commuters in a city aids urban planners in improving transportation infrastructure.

Key Descriptive Statistical Measures:

MODE Function in Excel

The screenshot shows the Excel ribbon with the 'Formulas' tab selected. The 'More Functions' button is highlighted, and the 'Statistical' category is expanded. The 'MODE.SNGL' function is selected from the list. A red arrow points from the 'Statistical' category to the 'MODE.SNGL' function. Another red arrow points from the 'MODE.SNGL' function to the formula bar. The formula bar shows '=MODE(' and a tooltip for the function is displayed below it.

Function Library

Financial Logical Text Date & Time Lookup & Reference Math & Trig More Functions Name Manager Define Name Use in Formula Create from Selection Trace Precedents Trace Dependents Remove Arrows Show Error Evaluate Formula Auditing

Statistical Engineering Cube Information Compatibility

MAX
MAXA
MEDIAN
MIN
MINA
MODE.MULT
MODE.SNGL
NEGBINOM.DIST
NORM.DIST
NORM.INV
NORM.S.DIST
NORM.S.INV
PEARSON

=MODE(

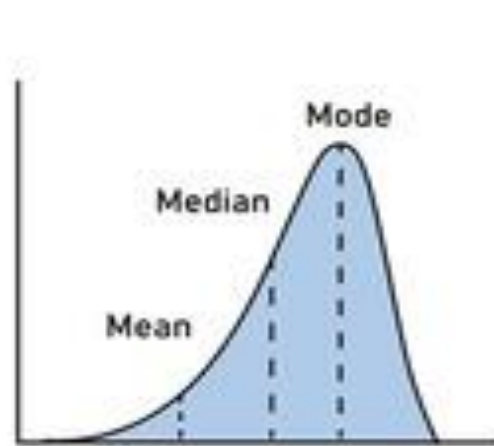
MODE(number1, [number2], ...)

MODE.SNGL(number1,number2,)
Returns the most frequently occurring, or repetitive, value in an array or range of data.
Press F1 for more help.

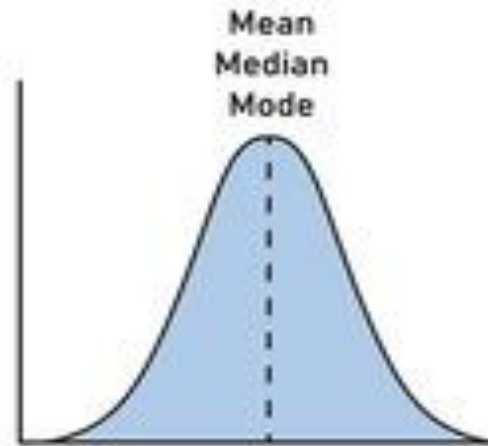
Key Descriptive Statistical Measures:

Central Tendency:

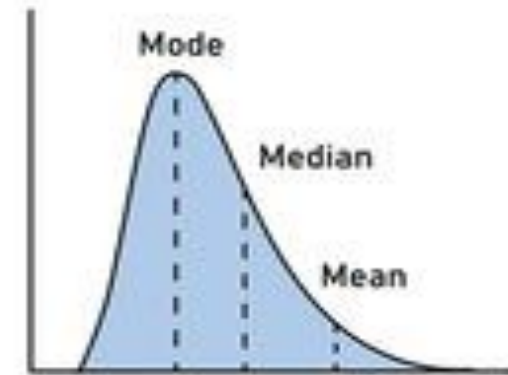
Mean, Median and Mode



Left skew



Normal distribution



Right skew

Key Descriptive Statistical Measures:

Measures of Dispersion:

- Range:** The difference between the maximum and minimum values in a dataset.
- Variance (σ^2):** The average of the squared differences from the mean.
- Standard Deviation (σ):** The square root of the variance.

Key Descriptive Statistical Measures:

Dispersion:

Dispersion: Real Applications

Range:

•Real-Life Application:

- In **Weather Forecasting**: Calculating the range of temperatures within a day helps meteorologists understand the variation in weather conditions and predict temperature fluctuations.
- In **Sports Analytics**: Determining the range of scores in a game provides insight into the competitiveness and variability of team performances.

Key Descriptive Statistical Measures:

Dispersion :

Find Range in Excel

B58					=MAX(C42:C57)-MIN(C42:C57)
	A	B	C	D	E
41	Employee ID	Name	Sales Amount		
42	I3000	Adri			
43	I3001	Nick			
44	I3002	N			
45	I3003	San			
46	I3004	Patrick	14872		
47	I3005	Davind	26254		
48	I3006	David	27153		
54	I3012	John Smith	23192		
55	I3013	Martin	22905		
56	I3014	Christine	20970		
57	I3015	Samson	29428		
58	Range	19098			

**Range = Maximum Value -
Minimum Value**

Key Descriptive Statistical Measures:

Dispersion:

Dispersion: Real Applications

Variance (σ^2):

Real-Life Application:

- In Financial Risk Management: Assessing the variance of stock prices helps investors understand the volatility of a stock, which is crucial for risk management and portfolio diversification.
- In Quality Control: Analyzing the variance of product dimensions in manufacturing processes helps identify potential issues and improve product consistency.

Key Descriptive Statistical Measures:

Dispersion :

Variance in Excel

C16	✓	<i>f_x</i>	=VAR.P(A2:A14)	C16	✓	<i>f_x</i>	=VAR.S(A2:A14)
	A	B	C		A	B	C
1	Set A	Set B		1	Set A	Set B	
2	16	12		2	34	48	
3	18	15		3	48	34	
4	10	12		4	38	31	
5	16	20		5	30	20	
6	14			6	20	29	
7	17			15			
8	19			16	Variance of Set A	93.333	
15				17	Variance of Set B	87.974	
16	Variance of Set A	8.746					
17	Variance of Set B	16.438					

Key Descriptive Statistical Measures:

Dispersion:

Dispersion: Real Applications

Standard Deviation (σ):

Real-Life Application:

- In **Healthcare**: Calculating the standard deviation of patient wait times in a hospital emergency room helps administrators optimize staffing levels and improve efficiency.
- In **Education**: Determining the standard deviation of test scores within a classroom provides insight into the level of academic achievement and helps identify students who may need additional support.

Standard Deviation Formula in Excel

The screenshot shows an Excel spreadsheet with the following data and calculations:

	A	B	C	D	E	F	G
1	Date of Delivery	Quantity		Standard Deviation			
2	25/2/2019	10		5.7446			
3	26/2/2019	12					
4	27/2/2019	23		Mean			
5	28/2/2019	23		43513			
6	03-01-2019	27					
7	03-02-2019	17					
8	03-03-2019	22					
9	03-04-2019	20					
10	03-05-2019	14					

Formulas shown in the image:

- Formula Bar:** `=STDEV(B2:B10)`
- Cell D2:** `=AVERAGE(`
- Cell D5:** `=STDEV(`
- Yellow Highlighted Text:** `AVERAGE(number1, [number2], ...)` and `STDEV(number1, [number2], ...)`
- Yellow Highlighted Text:** "Calculates standard deviation based on the entire population given as arguments (ignores logical values and text)"

Understanding Skewness in Statistics

- Skewness measures the asymmetry of a data distribution.
- It indicates whether the data points are more concentrated on one side of the mean.
- Understanding skewness helps in identifying the direction and extent of asymmetry in the data, which is crucial for statistical analysis and decision-making.

Key Descriptive Statistical Measures:

Understanding Skewness in Statistics

Types of Skewness

1. Positive Skewness (Right-Skewed)

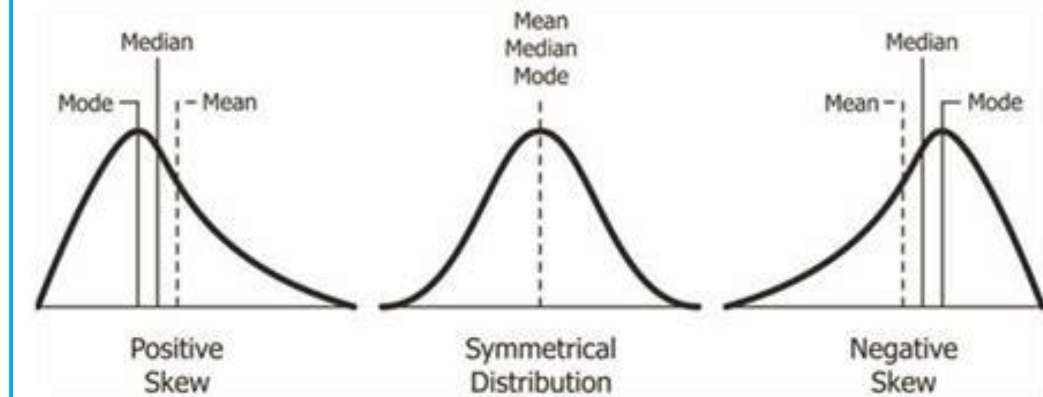
- Description: The tail on the right side of the distribution is longer or fatter. Most values are concentrated on the left side of the mean.
- Implication: Indicates that there are a few exceptionally high values.
- Example: Income distribution, where a small number of people earn significantly more than the average.

2. Negative Skewness (Left-Skewed)

- Description: The tail on the left side of the distribution is longer or fatter. Most values are concentrated on the right side of the mean.
- Implication: Indicates that there are a few exceptionally low values.
- Example: Age at retirement, where most people retire around a common age, but a few retire very early.

3. No Skewness (Symmetrical)

- Description: The distribution is symmetrical, with data points evenly distributed around the mean.
- Implication: Indicates that the data is balanced, with no extreme values skewing the distribution.
- Example: Heights of adults in a population where the majority of values cluster around the mean.



Key Descriptive Statistical Measures:

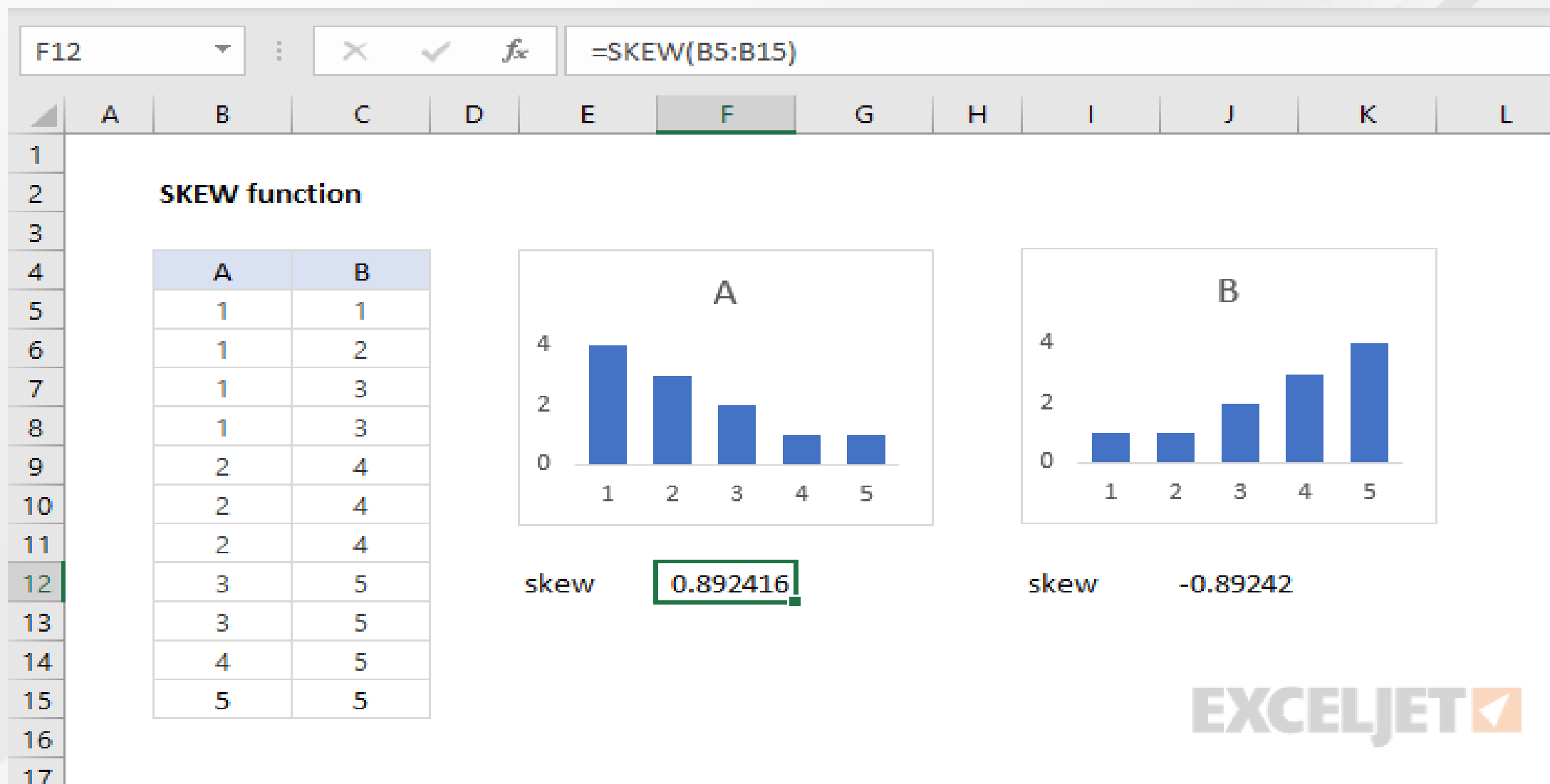
Distribution Shape: Real Applications:

Real-Life Application:

- In **Finance**: Evaluating the skewness of interest rate movements aids financial institutions in assessing interest rate risk and optimizing their investment strategies. Positive skewness may suggest a higher probability of large interest rate increases, while negative skewness may indicate a higher likelihood of significant interest rate decreases.

Key Descriptive Statistical Measures:

Distribution Shape: Real Applications:



Distribution Shape:

Kurtosis:

- Real-Life Application:
- In **Insurance**: Assessing the kurtosis of insurance claim amounts helps insurers understand the likelihood of extreme claim events. High kurtosis indicates a higher probability of rare, high-impact claims, which is essential for pricing and risk management.

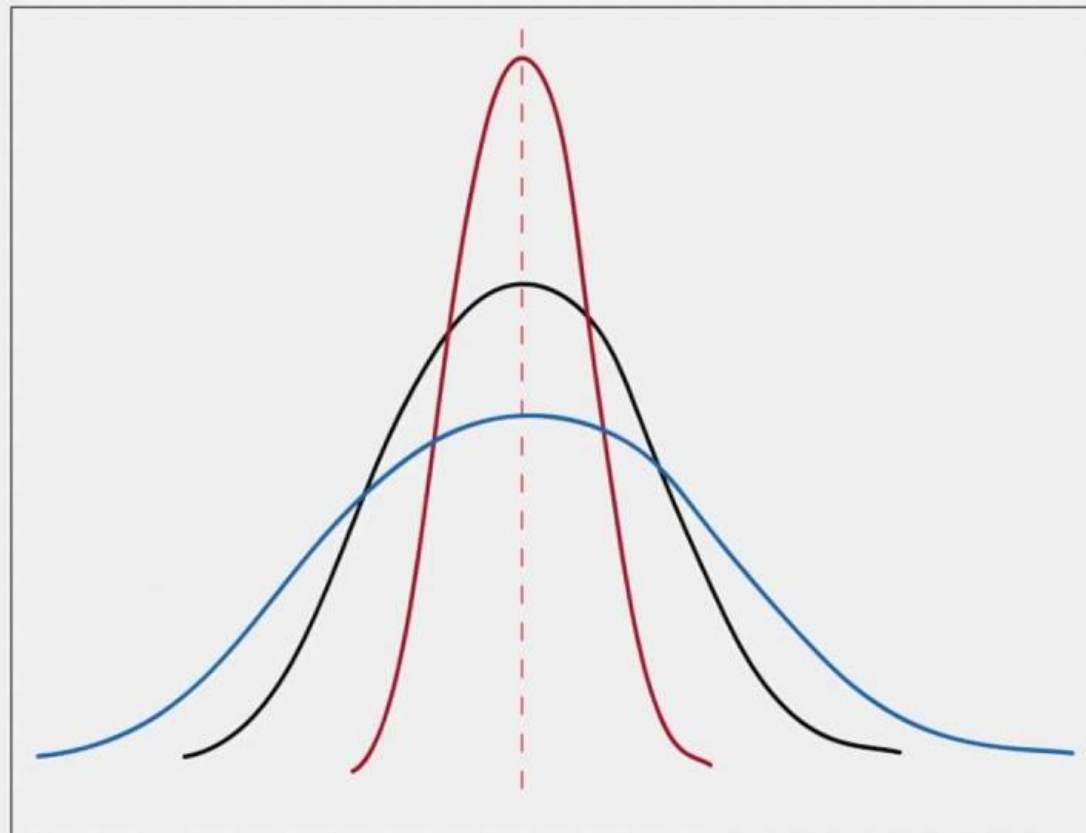
Key Descriptive Statistical Measures:

Distribution Shape:

Kurtosis

- It is all about the size of the tail – more concern about the outliers

$CK = 3$	Mesokurtic
$CK < 3$	Platykurtic
$CK > 3$	Leptokurtic



Real-Life Application: Customer Data Analysis in a Bank:

Banks collect extensive data on customers, including demographics, transaction history, loan repayments, and service usage.

Descriptive Statistics Used:

- **Average Account Balance:** Helps understand the financial profile of customers.

Real-Life Application: Customer Data Analysis in a Bank:

Example: Calculating the mean account balance of customers to tailor financial products to different customer segments.

- **Application:** A bank may find that the average account balance of its customers is \$10,000. With this information, they could develop premium banking services for customers with balances above this average, offering personalized investment advice and higher interest rates for higher balances.

Real-Life Application: Customer Data Analysis in a Bank:

- **Loan Repayment Times:** Measures the time taken for customers to repay their loans.
- **Example:** Analyzing the median repayment time for different types of loans (e.g., mortgages, personal loans, car loans) to assess credit risk and adjust lending policies.
- **Application:** If the median repayment time for personal loans is found to be 3 years, the bank could design repayment plans and interest rates that align with this typical repayment period, ensuring better cash flow management and risk assessment.

Real-Life Application: Customer Data Analysis in a Bank:

- **Transaction Frequency:** Measures how often customers engage in transactions within a specific period.
- **Example:** Calculating the average number of transactions per month per customer to identify active users and tailor marketing strategies.
- **Application:** A bank might find that customers on average make 25 transactions per month. They could then target customers making fewer transactions with incentives to increase their usage, such as cashback offers or reduced fees for increased activity.

Real-Life Application: Customer Data Analysis in a Bank:

- **Customer Demographics:** Analyzing age, income, employment status, and other demographic factors to understand customer segments.
- **Example:** Determining the average age and income of customers to develop targeted financial products and services.

Real-Life Application: Customer Data Analysis in a Bank:

- **Application:** If the bank finds that a significant portion of their customers are young professionals with an average income of \$70,000, they might create tailored financial planning services, investment options, and loan products that cater to the needs and aspirations of this demographic.

These descriptive statistics help banks tailor their products and services to better meet the needs of their customers, manage risk, and enhance customer satisfaction and loyalty.

Conclusion..:

In education, analyzing student performance data through descriptive statistics like mean test scores, grade distributions, and outlier identification enables tailored educational programs to meet individual student needs, fostering academic success and personal growth.



Chapter

03

Summary of chapter

Review Questions

Question 1:

Explain the concepts of mean, median, and mode, and provide an example of how each is used in financial data analysis.

Question 2:

What are the measures of dispersion, and why are they important in statistical analysis? Provide an example in finance.

Question 3:

Describe skewness and kurtosis and their significance in data analysis. Provide an example of how each is applied in financial analysis.

Thank You