R02 Organizing, Visualizing, and Describing Data

This document should be read in conjunction with the corresponding reading in the 2022 Level I CFA® Program curriculum. Some of the graphs, charts, tables, examples, and figures are copyright 2021, CFA Institute. Reproduced and republished with permission from CFA Institute. All rights reserved.

Required disclaimer: CFA Institute does not endorse, promote, or warrant the accuracy or quality of the products or services offered by IFT. CFA Institute, CFA®, and Chartered Financial Analyst® are trademarks owned by CFA Institute.

Ver 1.1

1. Introduction

This reading presents tools and techniques for organizing, visualizing and describing data. These tools and techniques can help us convert raw data into useful information for investment analysis.

2. Data Types

Data can be defined as a collection of numbers, characters, words and text - as well as images, audio, and video - in a raw or organized format to represent facts or information.

Data can be classified in three ways:

- Numerical versus categorical data
- Cross-sectional versus time-series versus panel data
- Structured versus unstructured data

2.1 Numerical versus Categorical Data

Based on a statistical perspective, data can be classified into numerical data and categorical data.

Numerical data: Numerical data (also called quantitative data) are values that represent measured or counted quantities as numbers. Numerical data can be further classified into two types:

- Continuous data: Data that can be measured and can take on any numerical value in a specified range of values. For example, the future value of a sum of money invested today. The FV can take on range of values depending on the investment period and interest rate.
- Discrete data: Data that can take numerical values that result from a counting process. The data is limited to a finite number of values. For example, the frequency of discrete compounding (m). The frequency could be monthly (m = 12), quarterly (m = 4), semi-yearly (m = 2), or yearly (m = 1).

Categorical data: Categorical data (also called qualitative data) are values that describe a quality or characteristic of a group of observations. It can usually take only a limited number of values that are mutually exclusive. Categorical data can be further classified into two types:

- Nominal data: Categorical values that cannot be organized in a logical order. For example, classification of publicly listed stocks into different sectors, such as: energy, information technology, financials, health care etc.
- Ordinal data: Categorical values that can be organized in a logical order or ranked. For example, Standard & Poor's star ratings for mutual funds. One star represents the group of mutual funds with the worst performance. Similarly, groups with two, three, four, and five stars represent groups with increasingly better performance.

Although the categories represented by ordinal data can be ranked, the numerical differences between the categories is not necessarily the same, and it cannot be used to draw inferences.

Example: Identifying Data Types

Identify the data type for each of the following items:

- Number of coupon payments for a bond
- Dividends paid by a stock
- Credit ratings for corporate bonds
- Hedge fund classification types

Solution:

Based on our above discussion, we can classify these items as follows:

- Number of coupon payments for a bond – Discrete data
- Dividends paid by a stock – Continuous data
- Credit ratings for corporate bonds – Ordinal data
- Hedge fund classification types – Nominal data

2.2 Cross-Sectional versus Time-Series versus Panel Data

Based on how data is collected, it can be classified into three types: cross-sectional, time-series, and panel.

Before we describe these data types, we need to understand two terms: ‘variable’ and ‘observation’.

- A variable (also called field, attribute, or feature) is characteristic or quantity that can be measured, counted, or categorized. A variable is subject to change. For example, the returns on Microsoft stock in a given quarter can be considered a variable.
- An observation is a value of a specific variable collected at a point in time or over a specified period of time. For example, if the returns on Microsoft stock in 2019 Q1 were 3%, then 3% is an observation.

Time-series data: Time-series data consists of observations for a single subject taken at specific and equally spaced intervals of time. For example, the quarterly returns of Microsoft stock from 2019 to 2020.

Cross-sectional data: Cross-sectional data consists of observations for multiple subjects taken at a specific point in time. For example, the quarterly returns in 2019 Q1 of a group of similar stocks – Microsoft, Oracle, and HP.

Panel data: Panel data is a combination of time-series and cross-sectional data. It consists of observations through time on one or more variables for multiple subjects. It is generally presented as a table. For example, the quarterly returns of Microsoft, Oracle, and HP from 2019 to 2020.

2.3 Structured versus Unstructured Data

Based on whether data is available in a highly organized form or not, it can be classified into structured and unstructured data.

Structured data: Structured data is highly organized in a pre-defined manner, usually with repeating patterns. It is easier to enter, store, query and analyze, without much manual processing. Examples:

- Market data: Daily closing stock prices and trading volumes.
- Fundamental data: Data contained in financial statement such as earnings per share.
- Analytical data: Data derived from analytics, such as cash flow projections.

Unstructured data: Unstructured data does not follow any conventionally organized forms. It is typically alternative data and is usually collected from unconventional sources. Based on the source, unstructured data can be classified into:

- Produced by individuals (i.e., via social media posts, web searches, etc.);
- Generated by business processes (i.e., via credit card transactions, corporate regulatory filings, etc.); and
- Generated by sensors (i.e., via satellite imagery, foot traffic by mobile devices, etc.).

2.4 Data Summarization

Raw data typically cannot be used by humans or computers directly to extract information and insights. The data usually has to be organized first. In the following sections we will discuss various techniques for organizing and summarizing data.

3. Organizing Data for Quantitative Analysis

Raw data is typically organized into either a one-dimensional array or a two-dimensional rectangular array (also called a data table) for quantitative analysis.

- A one-dimensional array is suitable for representing a single variable. For example, the closing price for the first 10 trading days for a company after it went public.
- A two-dimensional array consists of columns and rows to hold multiple variables and multiple observations, respectively. For example, quarterly revenue, EPS, and DPS for a company for the past two years.

4. Summarizing Data Using Frequency Distributions

A frequency distribution (also called a one-way table) is a tabular display of data summarized into a relatively small number of intervals.

Frequency distributions for categorical variables: The steps for constructing a frequency distribution for a categorical variable are:

1. Count the number of observations for each unique value of the variable.
2. Construct a table listing each unique value and the corresponding counts.
3. Sort the records by number of counts in descending or ascending order.

A sample frequency distribution of 200 companies across four sectors is presented below:

Sector	**Absolute Frequency**	**Relative Frequency**
Technology	22	11%
Healthcare	50	25%
Financial	58	29%
Industrial	70	35%
Total	**200**	**100%**

Points to note:

- Absolute frequency: The actual number of observations in a given interval is called the absolute frequency.
- Relative frequency: It is the absolute frequency of each interval divided by the total number of observations.

Frequency distributions for numerical variables: The steps for constructing a frequency distribution for numerical variables are:

1. Sort the data in ascending order.
2. Calculate the range of data.
3. Decide on the number of bins (k).
4. Determine bin width.
5. Determine bins.
6. Determine the number of observations in each bin.
7. Construct a table of the bins listed from smallest to largest.

A sample frequency distribution for 100 stocks with prices ranging between 45.00 and 65.00 is presented below.

Stock Price (Min - Max)	**Absolute Frequency**	**Cumulative Frequency**	**Relative Frequency**	**Cumulative Relative Frequency**
45.00 – 50.00	25	25	0.25	0.25
50.00 – 55.00	35	60	0.35	0.60
55.00 – 60.00	29	89	0.29	0.89
60.00 – 65.00	11	100	0.11	1.00

Points to note:

- Range of the data = Maximum value – Minimum value = 65.00 – 45.00 = 20
- We decided to have 4 bins
- Bin width = Range / Number of bins = 20/ 4 = 5
- The end points of each bin are determined as minimum value + bin width i.e. 45.00 + 5.00 = 50.00, 50.00 + 5.00 = 55.00, 55.00 + 5.00 = 60.00, 60.00 + 5.00 = 65.00. Thus, we get the bins listed in the table above.

- Minimum values are included in the bins whereas maximum values are excluded. For example, the observation 50.00 will fall in the 50.00 – 55.00 bin and not in the 45.00 – 50.00 bin. However, the last bin includes the maximum values. The observation 65.00 will both fall in the 60.00 – 65.00 bin, since it is the last bin.
- Cumulative frequency: For an interval, it is calculated as the sum of the absolute frequencies of all intervals lower than and including that interval.
- Cumulative relative frequency: For an interval, it is calculated as the sum of the relative frequencies of all intervals lower than and including that interval.

Instructor's Note: On the exam you are unlikely to be asked to construct a frequency distribution. However, you may be tested on the process and the terminology.

Example:

The actual number of observations in a given interval is called the:
A. absolute frequency.
B. relative frequency.
C. cumulative relative frequency.

Solution:

A is correct. The actual number of observations in a given interval is known as absolute frequency. Relative frequency is the absolute frequency of each interval divided by the total number of observations. Cumulative absolute frequency is the running total of all absolute frequencies.

Example:

Which of the following is *most likely* to be accurate?
A. An observation can fall in more than one interval.
B. The data is sorted in descending order for the construction of a frequency distribution.
C. The cumulative relative frequency tells the observer the fraction of the observations that are less than the upper limit of each interval.

Solution:

C is correct. The cumulative relative frequency tells the observer the fraction of the observations that are less than the upper limit of each interval. An observation cannot fall in more than one interval. The data is sorted in an ascending order for the construction of a frequency distribution.

5. Summarizing Data Using a Contingency Table

A contingency table is a tabular format that displays the frequency distributions of two or more categorical variables simultaneously. It can be used to find patterns between the variables.

Contingency tables are constructed by listing all levels of one variable as rows and all the levels of the other variables as columns in a table. For example, consider a contingency table created for a portfolio of 500 stocks based on two variables – sector and market capitalization.

	Market Capitalization Variable (3 Levels)			
Sector Variable (4 Levels)	Small	Mid	Large	**Total**
Financial	44	38	20	**102**
FMCG	130	54	46	**230**
Information Technology	57	34	21	**112**
Real estate	30	16	10	**56**
Total	**261**	**142**	**97**	**500**

Key points to note from the table are:

- Each cell shows the number of stocks of each sector with a given market cap level. For example, there are 130 small-cap FMCG stocks. This count is also called **joint frequencies**.
- The joint frequencies are added across rows and across columns, and the corresponding sub-totals are called **marginal frequencies**. For example, the marginal frequency of FMCG sector is 230, and the marginal frequency of small cap stocks is 261

Contingency tables can also be created using relative frequencies based on total count. Each number is expressed as percentage of the total number of stocks. For example, small cap FMCG stocks are 130 / 500 = 26% of the portfolio.

Applications

One application of contingency tables is for evaluating the performance of a classification model (using a confusion matrix). Suppose we have a model for classifying companies into two groups: those that default on their bond payments and those that do not default. The table below shows a confusion matrix for a sample of 1,000 non-investment-grade bonds.

Predicted Default	**Actual Default**		**Total**
	Yes	**No**	
Yes	150	10	160
No	6	834	840
Total	156	844	1,000

The table shows that the classification model incorrectly predicts default in 10 cases where an actual default did not occur. It also incorrectly predicts no default in 6 cases where a default did actually occur.

Another application of contingency tables is to investigate a potential association between two categorical variables. One way to test the potential association is to follow a three-step process:

1. Add the marginal frequencies and overall total to the contingency table.
2. Use the marginal frequencies to construct a table with expected values of the observations.
3. Compare with chi-square value for a given level of significance.

These steps are demonstrated in the following example.

Example: Contingency Tables and Association between Two Categorical Variables

Suppose we randomly pick 200 mutual funds and classify them based on two parameters:

- Fund style – Growth versus Value
- Risk level – Low risk versus High risk.

This data is summarized in a 2 x 2 contingency table shown below.

	Low Risk	**High Risk**
Growth	67	19
Value	98	16

1. Calculate the number of growth funds and the number of value funds.
2. Calculate the number of low-risk and high-risk funds.
3. Describe how the contingency table is used to set up a test for independence between fund style and risk level.

Solution to 1:

The marginal frequency for growth is 67 + 19 = 86

The marginal frequency for value is 98 + 16 = 114

Solution to 2:

The marginal frequency for low risk is 67 + 98 = 165

The marginal frequency for high risk is 19 + 16 = 35

Solution to 3:

To conduct a chi-square test of independence, we perform the following three steps.

Step 1: Add the marginal frequencies and overall total to the contingency table. We also show the relative frequency table for observed values.

Observed Values					**Observed Values**			
	Low Risk	**High Risk**				**Low Risk**	**High Risk**	
Growth	67	19	86		Growth	78%	22%	100%
Value	98	16	114		Value	86%	14%	100%
	165	35	200					

Step 2: Use the marginal frequencies to construct a table with expected values of the observations.

Expected Value$_{i,j}$ = (Total Row $_i$ × Total Column $_j$)/Overall Total

For example,
Expected value for Growth / Low Risk is: (86 x 165) / 200 = 70.95
Expected value for Value / High Risk is: (114 x 35) / 200 = 19.95`1 qA

The table of expected values and the corresponding relative frequency table is presented below:

Observed Values					Observed Values			
	Low Risk	High Risk				Low Risk	High Risk	
Growth	70.95	15.05	86		Growth	82.5%	17.5%	100%
Value	94.05	19.95	114		Value	82.5%	17.5%	100%
	165	35	200					

Step 3: The actual values and the expected values are used to derive the chi-square test statistic. This is then compared to a value from the chi-square distribution table for a given level of significance. If the test statistic is greater than the chi-square distribution value, then we can conclude that there is significant association between the categorical variables.

Instructor's Note: You will understand this step better when you go over the reading on 'Hypothesis Testing'.

6. Data Visualization

Visualization refers to the presentation of data in pictorial or graphical format to aid understanding of the data and for gaining insights into the data. There are multiple data visualization techniques, which are covered in the following sub-sections.

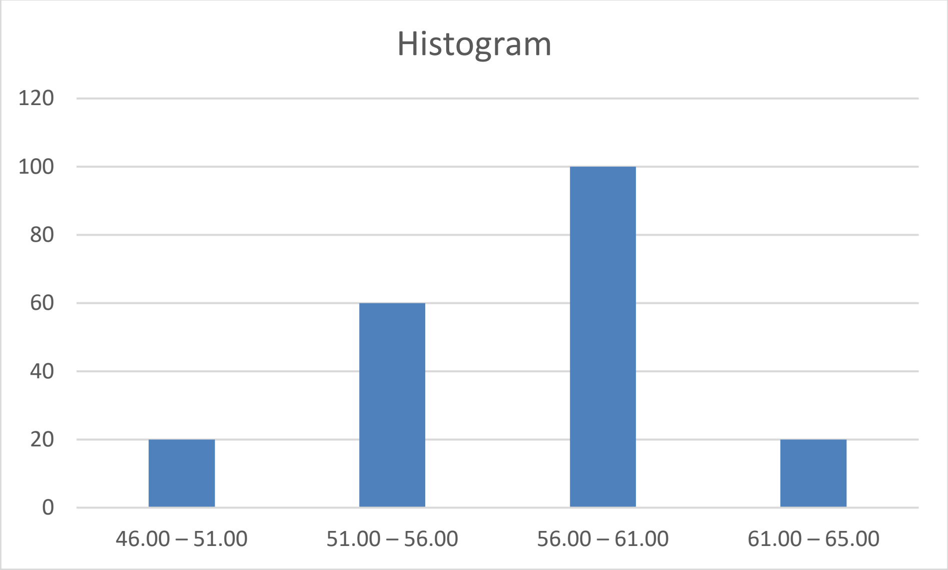
6.1 Histogram and Frequency Polygon

Histogram: A histogram presents the distribution of numerical data by using the height of a bar to represent the absolute frequency of each bin. The advantage of the visual display is that we can quickly see where most of the observations lie.

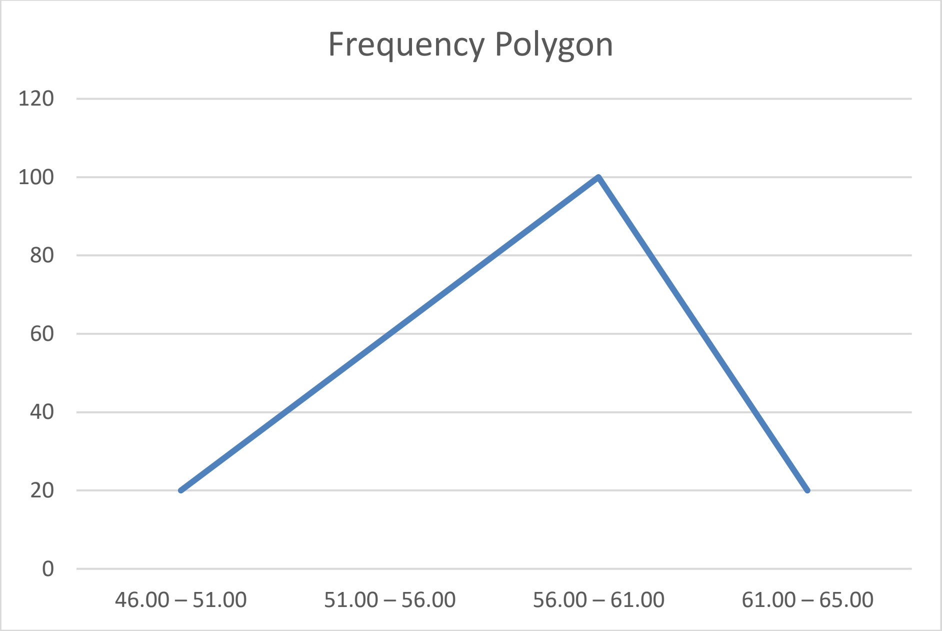
Suppose we are evaluating 200 stocks presented in the following frequency distribution table.

Price Range	Number of Stocks
46.00 – 51.00	20
51.00 – 56.00	60
56.00 – 61.00	100
61.00 – 65.00	20

We can depict this data graphically through a histogram.

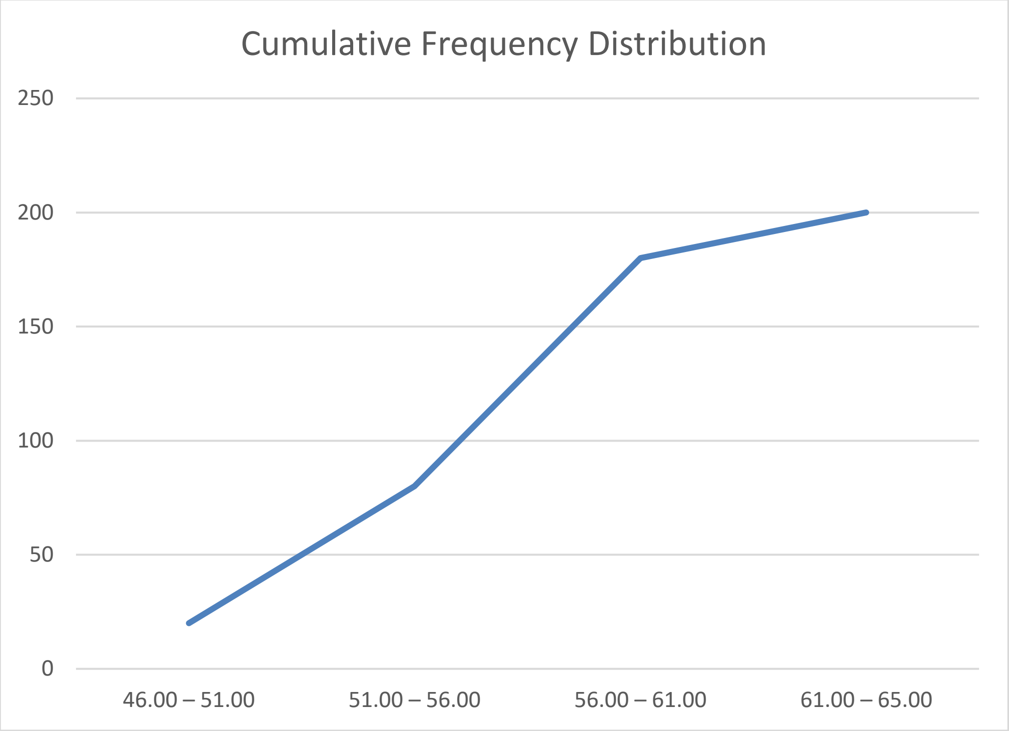


Frequency polygon:
A frequency polygon plots the midpoints of each interval on the X-axis and the absolute frequency of that interval on the Y-axis. Each point is then connected with a straight line.



Cumulative frequency distribution

Another graphical tool is the cumulative frequency distribution chart. Such a graph can plot either the cumulative frequency or cumulative relative frequency against the upper interval limit. The cumulative frequency distribution allows us to see how many or what percent of the observations lie below a certain value. The figure below is an example of a cumulative frequency distribution.



Notice that the slope is steep in the '51.00 -56.00' to '56.00 – 61.00' segment because a large number of stocks (100) are added. The slope flattens out in the last segment because only 20 stocks are added in the last segment.

Example:

Which of the following statements is *most likely* to be inaccurate about histograms?

A. A histogram is the graphical equivalent of a frequency distribution.

B. A histogram is a form of a bar chart.

C. In a histogram, the height represents the relative frequency for each interval.

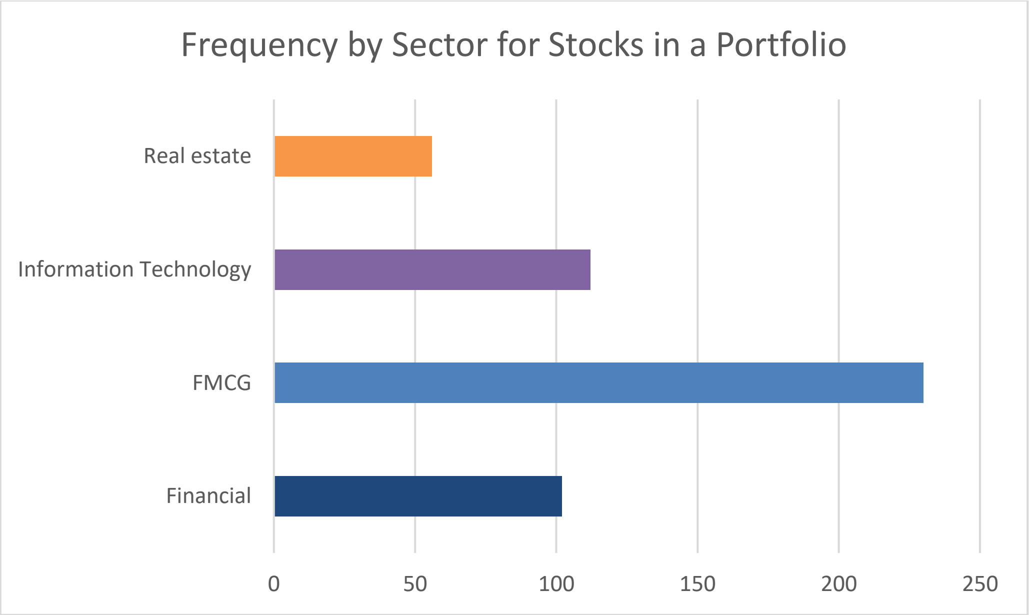
Solution:

C is correct. In a histogram, the height represents the absolute frequency for each interval.

6.2 Bar Chart

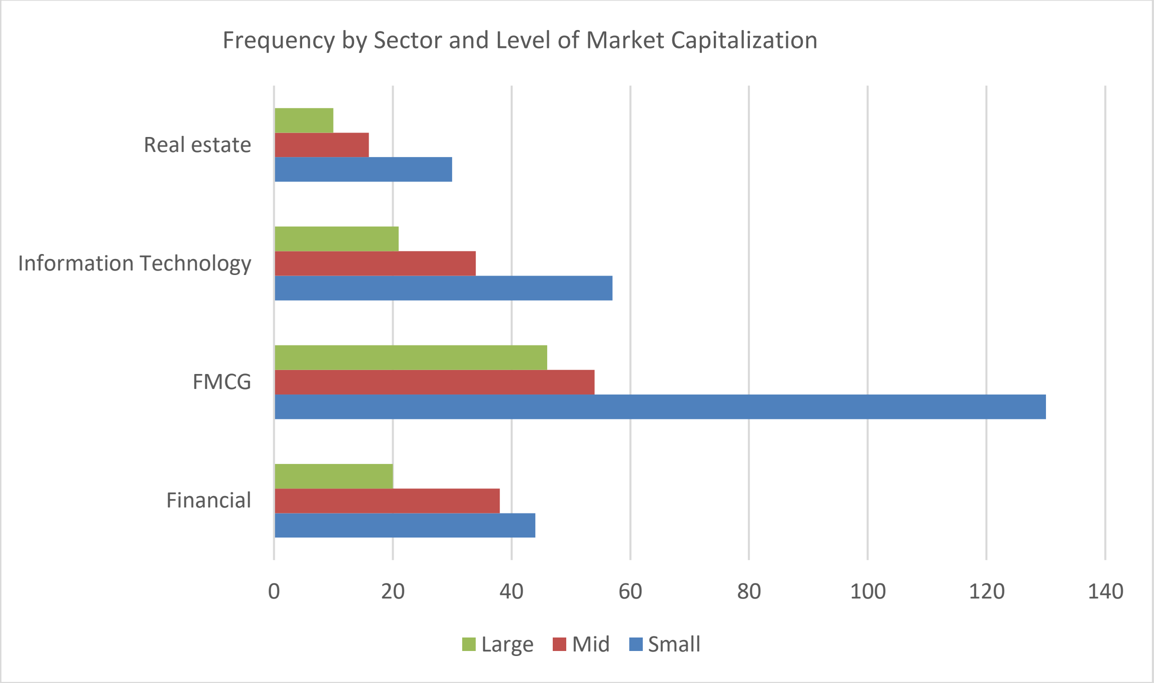
A **bar chart** is used to plot the frequency distribution of categorical data. Each bar represents a distinct category, and the bar's height is proportional to the frequency of that category.

The bar chart below shows that the sector in which the portfolio holds the most stocks is FMCG, with 230 stocks, followed by the IT sector, with 112 stocks.

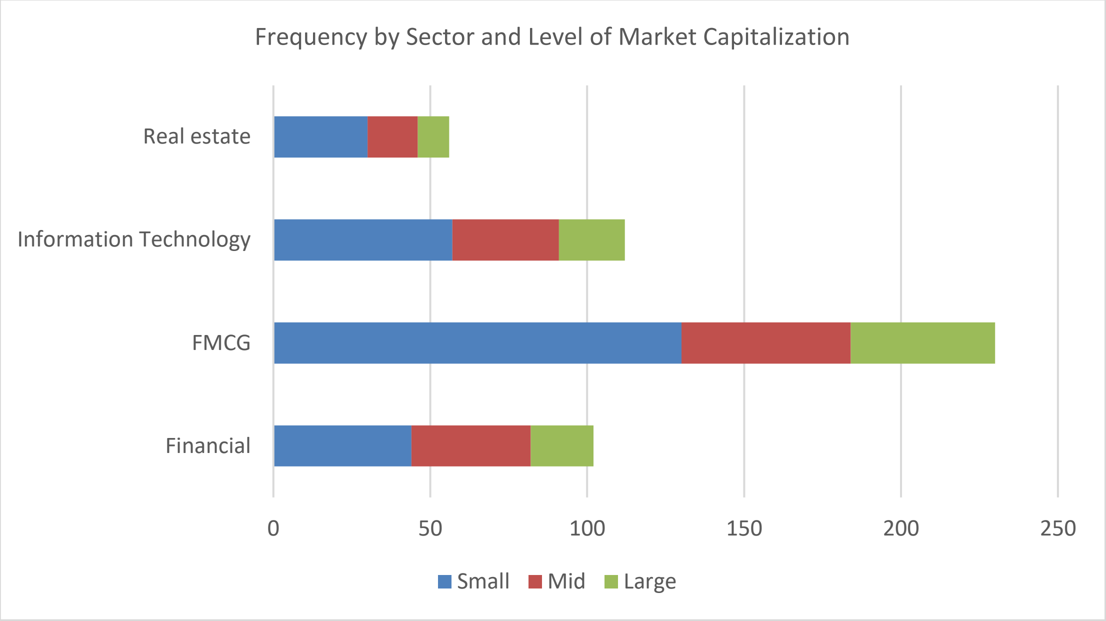


A **grouped bar chart** (also called a **clustered bar chart**) can be used to show the frequency distribution of multiple categorical variables simultaneously.

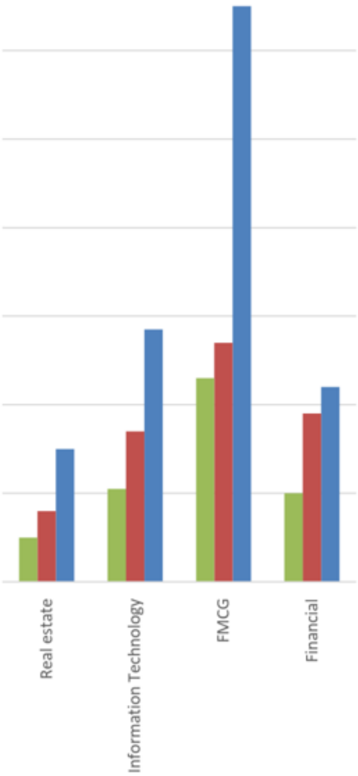
The chart below shows that small cap FMCG stocks have the highest frequency – 130. Also, we can easily observe that small cap stocks are the largest sub-group within each sector.

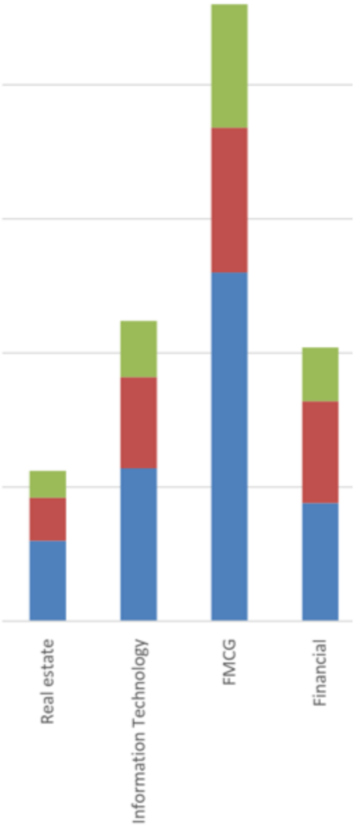


A **stacked bar chart** is an alternative form for presenting the frequency distribution of multiple categorical variables simultaneously.



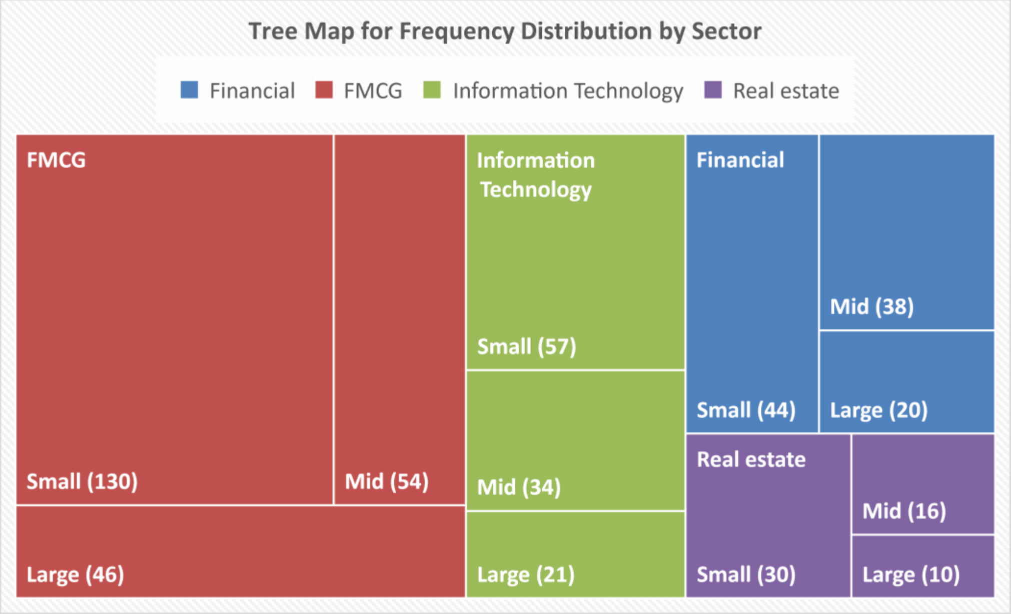
Bar charts can also be presented vertically instead of horizontally as shown below. Normally, the height of each bar is proportional to the value it depicts. However, sometimes the y-axis may be truncated, in which case the heights may not be proportional to the depicted values. In such cases, the graph needs to be evaluated more carefully.





6.3 Tree-Map

A **tree-map** is a graphical tool to display categorical data. It comprises of a set of colored rectangles to represent distinct groups. The area of each rectangle is proportional to the value of the corresponding group. Additional dimensions of categorical data can be displayed by a set of nested rectangles.



6.4 Word Cloud

A **word cloud** (also known as a **tag cloud**) is a visual device for representing textual data. The size of each word is proportional to the frequency of the word in the given text.

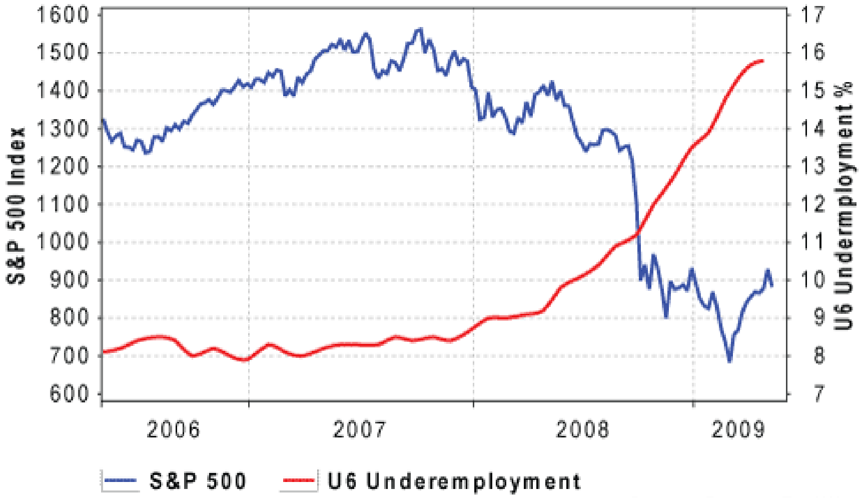
A sample word cloud is show below.

FINANCIAL REPORT
Finance
Trade
Data
Market
Commerce
Business
Stock
Money
Profit
Analysis
Success
Investment
Exchange
Economy
Management

Sometimes color can be used to add another dimension. For example, for a word cloud based on analyst reports related to a particular company, different colors can be used for positive, negative and neutral sentiment words. Positive sentiment can be depicted by the color 'green'. Negative sentiment can be depicted by the color 'red' and neutral sentiment can be depicted by the color 'blue'.

6.5 Line Chart

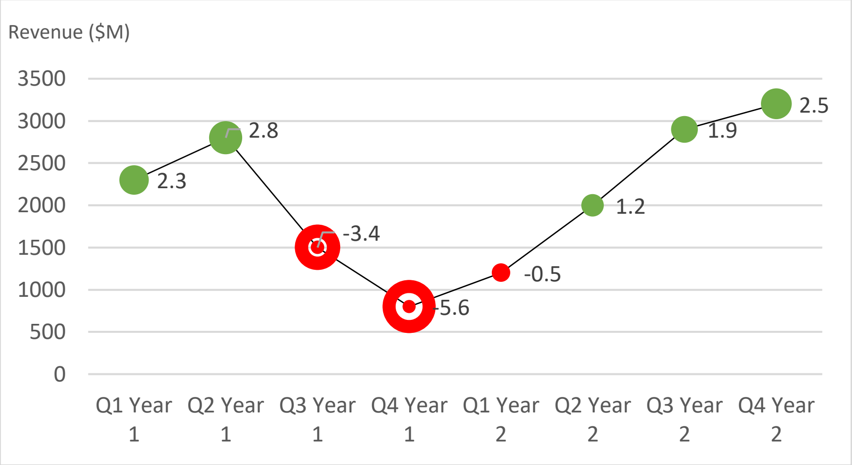
A **line chart** is a type of graph used to visualize ordered observations. It is often used to display the change of data series over time. A line chart can plot more than one set of data points, which helps in making comparisons. A sample line chart is shown below. After the 2008 crisis, stock prices dropped and unemployment rose.



Source: Reuters EcoWin

A **bubble line chart** is a special type of line chart that uses varying-sized bubbles as data points to represent an additional dimension of data.

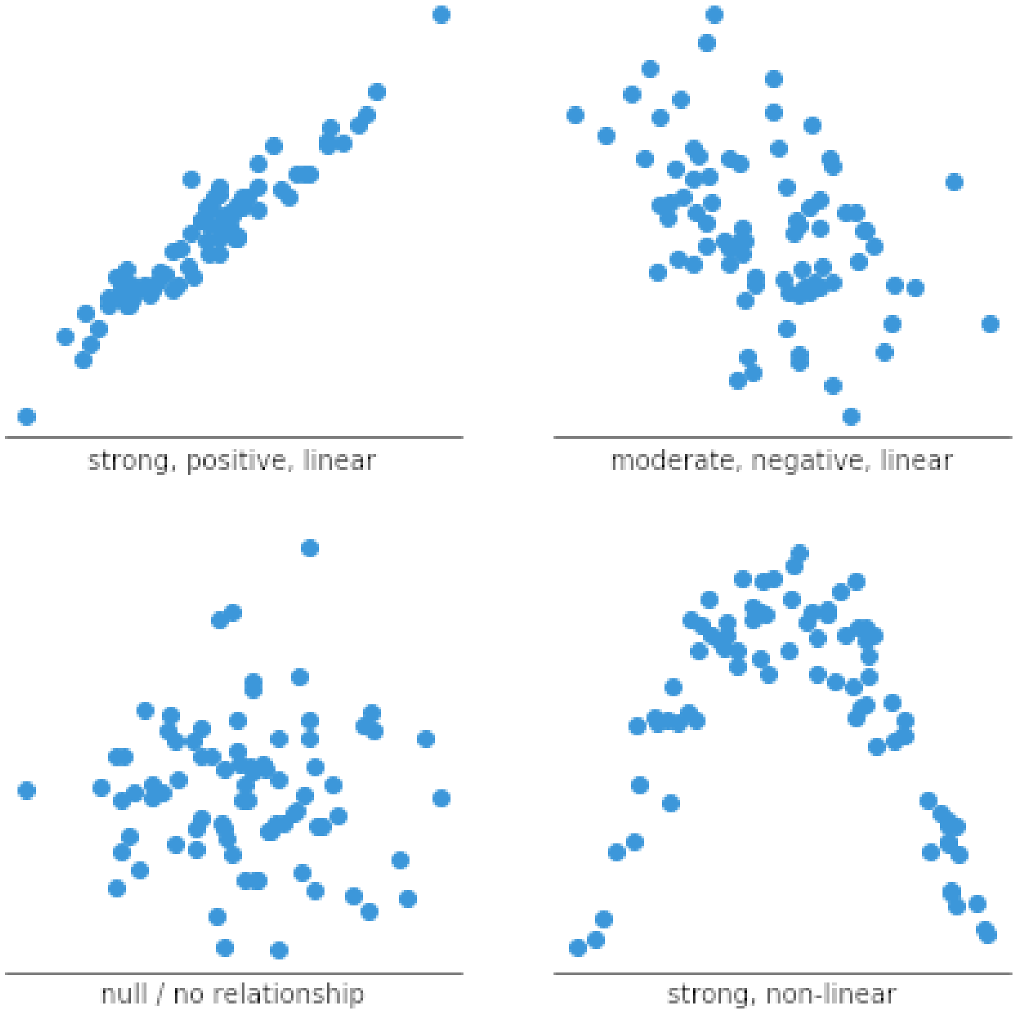
The following chart plots the quarterly revenue and EPS for a company over a two-year period. The x-axis represents time and the y-axis represents revenue. The line represents revenue. Each revenue data point is replaced by a circular bubble representing the EPS in the corresponding quarter. The size of the bubbles are proportional to the magnitude of the EPS. The bubbles are also color coded – red represents losses and green represents profits.



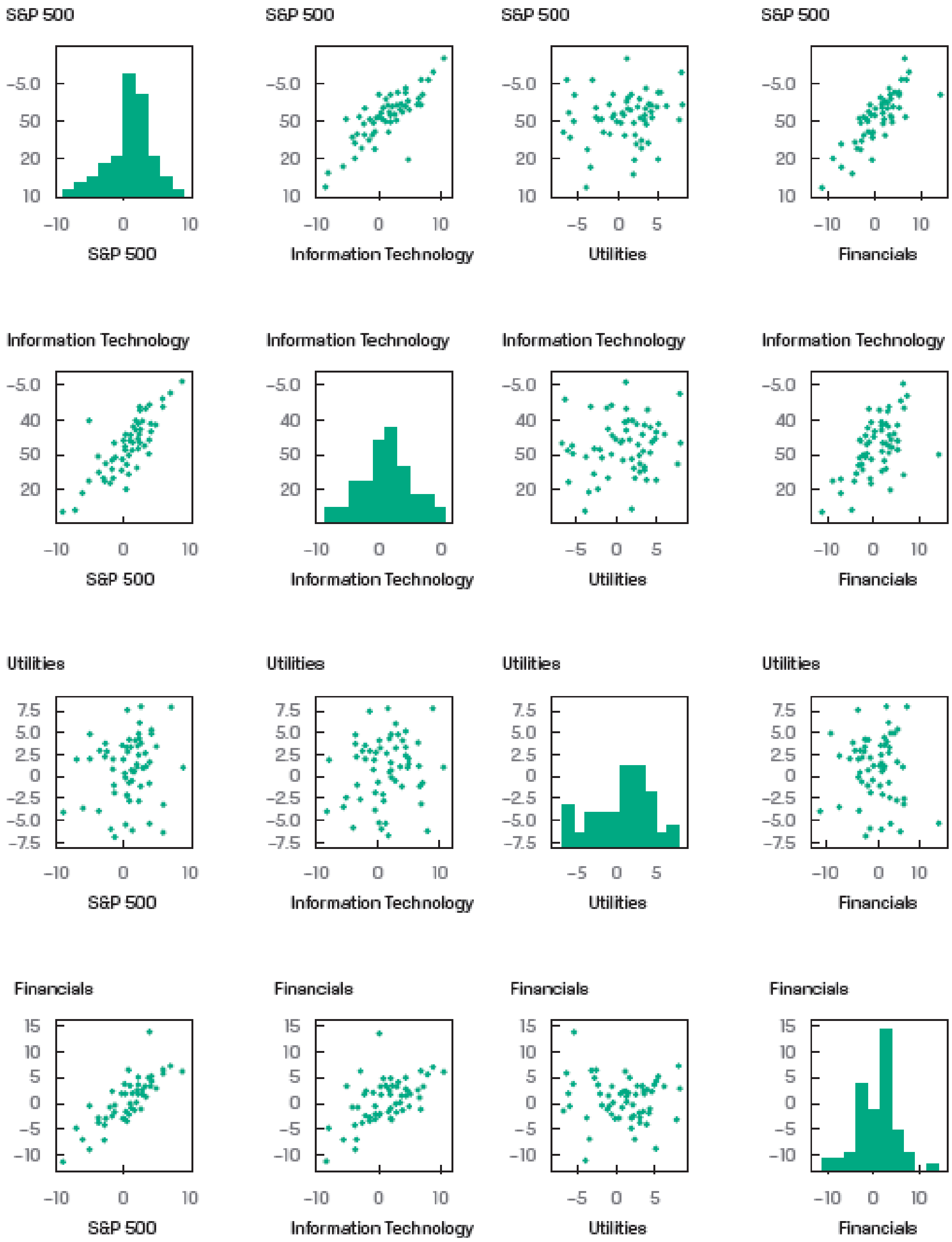
6.6 Scatter Plot

A scatter plot is a type of graph used to visualize the joint variation in two numerical variables. It is constructed with the x-axis representing one variable and the y-axis representing the other variable. Dots are drawn to indicate the values of the two variables at different points in time.

The pattern of a scatter plot may indicate no relationship, linear relationship or a non-linear relationship between the two variables. In case of a linear relationship, a positive slope indicates that the variables move in the same direction; whereas a negative slope indicates that the variables move in opposite directions.



A **scatter plot matrix** organizes scatter plots between pairs of variables into a matrix format. This makes it easy to inspect all pairwise relationships in one combined visual.

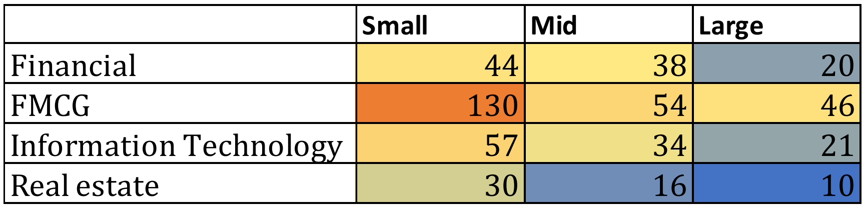


6.7 Heat Map

A heat map is a type of graphic that organizes and summarizes data in a tabular format and represents it using a color spectrum.

Heat maps are often used in displaying frequency distributions or visualizing the degree of correlation among different variables.

A sample heat map for a portfolio is shown below. Cells in the chart are color coded to differentiate high values from low values. Blue represents lower values whereas orange represents higher values.



	Small	Mid	Large
Financial	44	38	20
FMCG	130	54	46
Information Technology	57	34	21
Real estate	30	16	10

6.8 Guide to Selecting among Visualization Types

The intended purpose of visualizing data (i.e., whether it is for exploring/presenting distributions or relationships or for making comparisons) is the main factor that helps select the appropriate chart type.

To explore/present relationships between variable we can use the following visualization types:

- Two variables: Scatter plot
- More than two variables: Scatter plot matrix, heat map.

To explore/present distributions we can use the following visualization types:

- Numerical data: Histogram, frequency polygon, cumulative distribution chart.
- Categorical data: Bar chart, tree map, heat map
- Unstructured data: Word cloud.

To make comparisons we can use the following visualization types:

- Comparison among categories: Bar chart, tree map, heat map
- Comparison over time: Line chart for two variables, bubble line chart for three variables.

Common pitfalls

Four common pitfalls that should be avoided are:

- Improper chart type: To examine the correlation between two variables, a scatter plot should be used. If a line chart is used instead, then it will be difficult to examine the correlation.
- Selectively plotted data: Selecting an overly short time period may show the presence of a trend that is actually noise. For example, over a time period of a few days, it may appear that a stock is in a down trend, but when we consider a time period of the last two years, it is clear that the stock is a general uptrend with few days of consolidation in between.
- Improperly plotting data in a truncated graph: For example, suppose a vertical bar chart is used to compare EPS of two companies A and B. A has an EPS of $14 and B has an EPS of $15. If the y-axis starts at $13, then the bar heights would inaccurately imply that B's EPS is twice that of A.
- Improper scaling of axes: For example, consider a line chart of EPS of a company over time. The EPS is generally in the $10 – $20 range. If we set the Y-axis to plot numbers

up to $100, then the graph will be compressed and will appear to be less steep and less volatile than if we set the Y-axis to plot numbers only up to $25.

7. Measures of Central Tendency

A '**population'** is defined as all members of a specified group. A '**parameter'** describes the characteristics of a population.

A '**sample'** is a subset drawn from a population. A '**sample statistic'** describes the characteristic of a sample.

For example, all stocks listed on a country's exchange refers to a population. If 30 stocks are selected from the listed stocks, then this refers to a sample.

Sample statistics—such as measures of central tendency, measures of dispersion, skewness, and kurtosis—help make probabilistic statements about investment returns.

Measures of central tendency specify where data are centered.

Measures of location include not only measures of central tendency but other measures that explain the location or distribution of data.

7.1 The Arithmetic Mean

The arithmetic mean is the sum of the observations divided by the number of observations. It is the most frequently used measure of the middle or center of data.

The Sample Mean

The sample mean is the arithmetic mean calculated for a sample. It is expressed as:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

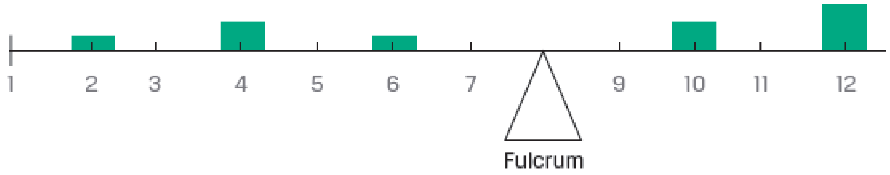
where: n is the number of observations in the sample.

If the sample data is: 2, 4, 4, 6, 10, 10, 12, 12, and 12 the sample mean can be calculated as:

$$\overline{X} = \frac{2 + 4 + 4 + 6 + 10 + 10 + 12 + 12 + 12}{9} = 8$$

Properties of the Arithmetic Mean

The arithmetic mean can be thought of as the center of gravity of an object. Exhibit 36 from the curriculum illustrates this concept by the above observations on a bar. When the bar is placed on a fulcrum and the fulcrum is located at the arithmetic mean, the bar balances. At any other point the bar will not balance.



A drawback of the arithmetic mean is that it is sensitive to extreme values (outliers). It can be pulled sharply upward or downward by extremely large or small observations, respectively.

Outliers

When data contains outliers, there are three options to deal with the extreme values:

Option 1: Do nothing; use the data without any adjustment.

Option 2: Delete all the outliers.

Option 3: Replace the outliers with another value.

Option 1 is appropriate in cases when the extreme values are genuine.

Option 2 excludes extreme observations. A **trimmed mean** excludes a stated percentage of the lowest and highest values and then calculates the arithmetic mean of the remaining values. For example, a 5% trimmed mean discards the lowest 2.5% and the highest 2.5% of values and computes the mean of the remaining 95% of values.

Option 3 replaces extreme observations with observations closest to them. A **winsorized mean** assigns a stated percentage of the lowest values equal to one specified low value and a stated percentage of the highest values equal to one specified high value, and then computes a mean from the restated data. For example, a 95% winsorized mean sets the bottom 2.5% of values equal to the value at or below which 2.5% of all the values lie (the "2.5th percentile" value) and the top 2.5% of values equal to the value at or below which 97.5% of all the values lie (the "97.5th percentile" value).

7.2 The Median

The median is the midpoint of a data set that has been sorted into ascending or descending order.

For odd number of observations: 2,5,7,11,14 → Median = 7

For even number of observations: 3, 9, 10, 20 → Median = (9 + 10)/2 = 9.5

As compared to a mean, a median is less affected by extreme values (outliers).

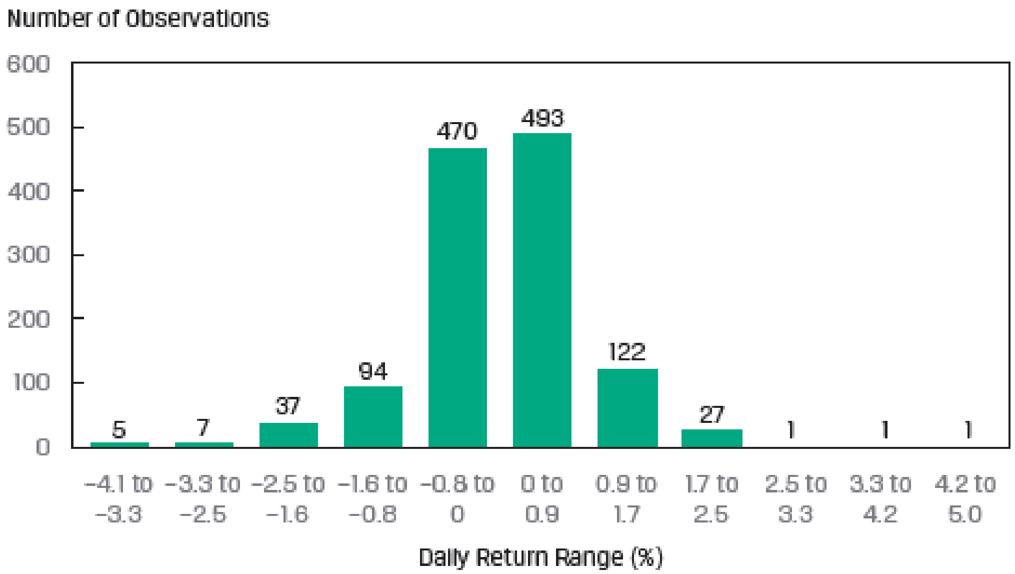
7.3 The Mode

The mode is the most frequently occurring value in a distribution.

For the following data set: 2, 4, 5, 5, 7, 8, 8, 8, 10, 12 → Mode = 8

A distribution can have more than one mode, or even no mode. When a distribution has one mode it is said to be unimodal. If a distribution has two or three modes, it is called bimodal or trimodal respectively.

When working with continuous data such as stock returns, '**modal interval**' is often used instead of a mode. The data is divided into bins and the bin with the highest frequency is considered the modal interval. Exhibit 39 from the curriculum demonstrates this concept by plotting a histogram of the daily returns on an index. The highest bar in the histogram '0.0 to 0.9%' is the modal interval.



7.4 Other Concepts of Mean

The Weighted Mean

In a weighted mean, instead of each data point contributing equally to the final mean, some data points contribute more "weight" than others. The formula for the weighted mean is:

$$\overline{X}_W = \sum_{i=1}^{n} w_i X_i$$

where: the sum of the weights equals 1; that is $\sum_{i=1}^{n} w_i = 1$

Example

Consider an investor with a portfolio of three stocks. $40 is invested in A, $60 in B, and $100 in C. If returns were 5% on A, 7% on B, and 9% on C, compute the portfolio return using the weighted mean.

Solution:

$$\left(\frac{40}{200}\right) \text{x } 5\% + \left(\frac{60}{200}\right) \text{x } 7\% + \left(\frac{100}{200}\right) \text{x } 9\% = 7.6\%$$

Example:

A portfolio manager wishes to compute the weighted mean of a portfolio that has the following asset allocation:

Local Equities:	25%

International Equities:	13%
Bonds:	27%
Mortgage:	18%
Gold:	17%

The returns on the above mentioned assets on December 31, 2012, were 5.4%, 8.9%, -2.5%, -7%, 11% respectively. What is the weighted mean for the portfolio?

Solution:

Weighted mean = (0.25 x 5.4) + (0.13 x 8.9) + (0.27 x -2.5) + (0.18 x -7) + (0.17 x 11) = 2.44%

An arithmetic mean is a special case of a weighted mean where all observations are equally weighted by the factor 1/n.

The Geometric Mean

The geometric mean is calculated as the nth root of a product of n numbers. The most common application of the geometric mean is to calculate the average return of an investment. The formula is:

$$R_G = [(1 + R_1)(1 + R_2) \dots (1 + R_n)]^{\frac{1}{n}} - 1$$

Example

The return over the last four periods for a given stock is: 10%, 8%, -5% and 2%. Calculate the geometric mean.

Solution:

$$[(1 + 0.10)(1 + 0.08)(1 - 0.05)(1 + 0.02)]^{\frac{1}{4}} - 1 = 0.0358 = 3.58\%$$

Given the returns shown above, $1 invested at the start of period 1 grew to:

$1.00 x 1.10 x 1.08 x 0.95 x 1.02 = $1.151. If the investment had grown at 3.58% every period, $1.00 invested at the start of period 1 would have increased to:

$1.00 x 1.0358 x 1.0358 x 1.0358 x 1.0358 = $1.151. As expected, both scenarios give the same answer. 3.58% is simply the average growth rate per period.

Other applications of the geometric mean involve the use of a second formula:

$$\ln\overline{X}_G = \frac{\sum_{i=1}^{n} \ln X_i}{n}$$

Instructor's Note: This formula is less testable.

Example:

The P/E ratio of a stock over the past four years has been: 10, 15, 14, 13. Calculate the geometric mean P/E.

Solution:

$$\ln\overline{X}_G = \frac{\sum_{i=1}^{n} \ln X_i}{n}$$

$$\ln\overline{X}_G = \frac{\ln 10 + \ln 15 + \ln 14 + \ln 13}{4} = 2.55$$

$$\overline{X}_G = e^{2.55} = 12.807$$

Using Geometric and Arithmetic Means

The geometric mean is appropriate to measure past performance over multiple periods.

Example
The portfolio returns for the past two years were 100% in year 1 and -50% in year 2. What was the mean return?

Solution:
Past return = geometric mean = $(2 \text{ x } 0.5)^{0.5} - 1 = 0\%$

The arithmetic mean is appropriate for forecasting single period returns.

Example
Two possible returns for the next year are 100% and -50%. What is the expected return?

Solution:
Expected return = Arithmetic mean = (100 – 50)/2 = 25%

The Harmonic Mean

The harmonic mean is a special type of weighted mean in which an observation’s weight is inversely proportional to its magnitude. The formula for a harmonic mean is:

$$X_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

where: $X_i > 0$ for i = 1, 2 … n, and n is the number of observations

The harmonic mean is used to find average purchase price for equal periodic investments.

Example
An investor purchases $1,000 of a security each month for three months. The share prices are $10, $15 and $20 at the three purchase dates. Calculate the average purchase price per share for the security purchased.

Solution:

The average purchase price is simply the harmonic mean of $10, $15 and $20.

The harmonic mean is:

$$\frac{3}{\frac{1}{\$10} + \frac{1}{\$15} + \frac{1}{\$20}} = \$13.85.$$

A more intuitive way of solving this is total money spent purchasing the shares divided by the total number of shares purchased.

Total money spent purchasing the shares = \$1,000 x 3 = \$3,000

Total shares purchased = sum of shares bought each month

$$= \frac{\$1{,}000}{10} + \frac{\$1{,}000}{15} + \frac{\$1{,}000}{20}$$

= 100 + 66.67 + 50 = 216.67

Average purchase price per share $= \frac{\$3{,}000}{216.67} = \$13.85$

Comparison of AM, GM and HM

- Arithmetic mean × Harmonic mean = Geometric mean^2
- If the returns are constant over time: AM = GM = HM.
- If the returns are variable: AM > GM > HM.
- The greater the variability of returns over time, the more the arithmetic mean will exceed the geometric mean.

Which mean to use?

- Arithmetic mean: Should be used with single period or cross-sectional data.
- Geometric mean: Should be used with time-series data.
- Weighted mean: Should be used when different observations have different weights.
- Harmonic mean: Should be used to find average purchase price for equal periodic investments.
- Trimmed mean: Should be used when the data has extreme outliers.
- Winsorized mean: Should be used when the data has extreme outliers.

8. Quantiles

8.1 Quartiles, Quintiles, Deciles, and Percentiles

A quantile is a value at or below which a stated fraction of the data lies. Some examples of quantiles include:

- **Quartiles:** The distribution is divided into quarters.
- **Quintiles:** The distribution is divided into fifths.
- **Deciles:** The distribution is divided into tenths.
- **Percentile:** The distribution is divided into hundredths.

The formula for the position of a percentile in a data set with n observations sorted in ascending order is:

$$L_y = \frac{(n + 1)y}{100}$$

where:
y is the percentage point at which we are dividing the distribution.
n is the number of observations.
L_y is the location (L) of the percentile (P_y) in an array sorted in ascending order.

Some important points to remember are:

- When the location, L_y, is a whole number, the location corresponds to an actual observation.
- When L_y is not a whole number or integer, L_y lies between the two closest integer numbers (one above and one below) and we use linear interpolation between those two places to determine P_y.
- Interquartile range is the difference between the third and the first quartiles.

Example
Consider the data set:

47 35 37 32 40 39 36 34 35 31 44

1. Find the 75th percentile point
2. Find the 1st quartile and 3rd quartile
3. Calculate the interquartile range
4. Find the 5th decile point
5. Find the 6th decile point.

Solution to 1:
First arrange the data in ascending order:

31, 32, 34, 35, 35, 36, 37, 39, 40, 44, 47

Location of the 75th percentile is the:
L_{75} = (11 + 1) (75/100) = 9th value. i.e. P_{75} = 40

With a small data set, such as this one, the location and the value is approximate. As the data set becomes larger, the location and percentile value estimates become more precise.

Solution to 2:
Location of the 1st quartile is:

L_{25} = (11 + 1) (25/100) = 3rd value. i.e. P_{25} = 34

Location of the 3rd quartile is:

L_{75} = (11 + 1) (75/100) = 9th value. i.e. P_{75} = 40

Solution to 3:
The interquartile range is the difference between the third and first quartiles, 40 – 34 = 6

Solution to 4:
Location of the 5th decile is:

L_{50} = (11 + 1) (50/100) = 6th value. i.e. P_{50} = 36

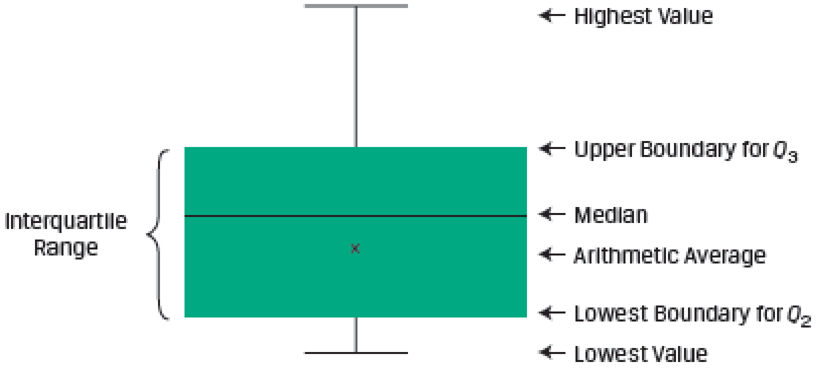
Solution to 5:

L_{60} = (11 + 1) (60/100) = 7.2

Use linear interpolation, which estimates an unknown value on the basis of two known values that surround it.

In this case, the 7th value is 37 and the 8th value is 39. The 6th decile is: P_{60} = 37+ 0.4 (0.2 times the linear distance between 37 and 39). P_{60} = 37.4

A **box and whiskers plot** is used to visualize the dispersion of data across quartiles. The box represents the interquartile range. The whiskers represent the highest and lowest values of the distribution. Exhibit 44 shows a sample box and whisker plot.



There are several variations of the box and whiskers plot. Sometimes the whiskers may be a function of the interquartile range instead of the highest and lowest values.

8.2 Quantiles in Investment Practice

Quantiles are used in:

- Portfolio performance evaluation: The performance of investment managers is often evaluated in terms of the percentile or quartile in which they fall relative to the performance of their peers.
- Investment research: For example, companies can be ranked based on their market capitalization and sorted into deciles. The first decile contains companies with smallest market values and the tenth decile contains companies with the largest market values. Such a classification allows analysts to compare the performance of small companies with large ones.

9. Measures of Dispersion

Measures of central tendency tell us where the investment results (expected returns) are centered. However, to evaluate an investment we also need to know how returns are dispersed around the mean. Measures of dispersion describe the variability of outcomes around the mean.

9.1 The Range

The range is the difference between the maximum and minimum values in a data set. It is expressed as:

Range = Max value – Min Value

If the annual returns data is: 10%, -5%, 10%, 25%. What is the range?

Here the maximum return is 25% and the minimum return is -5%. The range is 25% – (-5%) = 30%.

Another way to specify the range is to mention the actual minimum and maximum values. For example, for the above data the range is "from -5% to 25%".

The range is easy to compute; however, it does not tell us much about how the data is distributed.

9.2 The Mean Absolute Deviation

It is the average of the absolute values of deviations from the mean. It is expressed as:

$$\text{MAD} = \left[\sum_{i=1}^{n} |X_i - \overline{X}|\right] / n$$

where: $\overline{X}$ is the sample mean and n is the number of observations in the sample.

Example

Consider the following data set: 8, 12, 10, 8 and 5. Calculate the mean absolute deviation.

Solution:

$\overline{X}$= (8 + 12 + 10 + 8 + 5) / 5 = 8.6

$$\text{MAD} = \frac{|8 - 8.6| + |12 - 8.6| + |10 - 8.6| + |8 - 8.6| + |5 - 8.6|}{5}$$

$$\text{MAD} = \frac{0.6 \ + \ 3.4 \ + \ 1.4 \ + \ 0.6 \ + \ 3.6}{5} = \ 1.92$$

9.3 Sample Variance and Sample Standard Deviation

Variance is defined as the average of the squared deviations around the mean. **Standard deviation** is the positive square root of the variance.
Sample variance applies when we are dealing with a subset, or sample, of the total population. It is expressed as:

$$s^2 = \sum_{i=0}^{n} (X_i - \overline{X})^2 / (n - 1)$$

where: $\overline{X}$ is the sample mean and n is the number of observations in the sample.

Sample standard deviation is defined as the positive square root of the sample variance.

Example

Calculate the sample variance for the following data set: 8, 12, 10, 8 and 5.

Solution:

$$s^2 = \frac{[(8-8.6)^2 + (12-8.6)^2 + (10-8.6)^2 + (8-8.6)^2 + (5-8.6)^2]}{5-1}$$

$$s^2 = 6.80\%$$

The sample standard deviation is the positive square root of the sample variance. For the sample data given above, $s = \sqrt{6.80} = 2.61\%$

Using a financial calculator to calculate variance and standard deviations

The sample standard deviation can easily be computed using a financial calculator. Assume the following data set: 10%, -5%, 10%, 25%, the calculator key strokes are shown below:

Keystrokes	*Description*	*Display*
[2nd] [DATA]	Enters data entry mode	
[2nd] [CLR WRK]	Clears data register	X01
10 [ENTER]		X01 = 10
[↓] [↓] 5+/- [ENTER]		X02 = -5
[↓] [↓] 10 [ENTER]		X03 = 10
[↓] [↓] 25 [ENTER]		X04 = 25
[2nd] [STAT] [ENTER]	Puts calculator into stats mode	
[2nd] [SET]	Press repeatedly till you see →	1-V
[↓]	Number of data points	N = 4
[↓]	Mean	X = 10
[↓]	Sample standard deviation	Sx = 12.25
[↓]	Population standard deviation	σx = 10.61

Notice that the calculator gives both the sample and the population standard deviation. On the exam we will have to determine whether we are dealing with population or sample data and choose the appropriate value.

Dispersion and the relationship between the arithmetic and the geometric means

The sample standard deviation can be used to understand the gap between the arithmetic and geometric mean. The relationship between the arithmetic mean ($\overline{X}$) and geometric mean($\overline{X}_G$) is:

$$\overline{X}_G \approx \overline{X} - \frac{s^2}{2}$$

The larger the variance of the sample, the wider the difference between the geometric mean and the arithmetic mean.

Example:

The dividend yield for five hypothetical companies from a list of ten companies is given below. What is the sample variance?

Paknama	10.50%
Genie Ltd.	16.25%
Mirinda Corp.	27.00%
Tina Travels Ltd.	12.00%
Thomas Press Ltd.	7.80%

Solution:

$$\bar{X} = \frac{10.5 + 16.25 + 27 + 12 + 7.8}{5} = 14.71$$

Sample variance

$$= \frac{[(10.5 - 14.71)^2 + (16.25 - 14.71)^2 + (27 - 14.71)^2 + (12 - 14.71)^2 + (7.8 - 14.71)^2]}{5 - 1}$$

Sample variance = 56.49

10. Downside Deviation and Coefficient of Variation

Variance and standard deviation of returns take account of returns above and below the mean, but often investors are concerned only with downside risk, for example returns below the mean.

The target downside deviation, or target semideviation, is a measure of the risk of being below a given target. It is calculated as the square root of the average squared deviations from the target, but it includes only those observations below the target (B).
The sample target semideivation can be calculated as:

$$\mathrm{S_{Target}} = \sqrt{\sum_{\text{for all } \mathrm{X_i} \leq \mathrm{B}}^{\mathrm{n}} \frac{(\mathrm{X_i} - \mathrm{B})^2}{\mathrm{n} - 1}}$$

Example:
Suppose the monthly returns on a portfolio are as shown:

Month	**Return (%)**
Jan	6
Feb	4
Mar	-2
Apr	-5
May	5
Jun	2

Jul	1
Aug	0
Sep	4
Oct	3
Nov	0
Dec	2

Calculate the target downside deviation when the target return is 4%.

Solution:

Month	**Observation**	**Deviation from the 4% target**	**Deviation below the target**	**Squared deviations below the target**
Jan	6	2	-	-
Feb	4	0	-	-
Mar	-2	-6	-6	36
Apr	-5	-9	-9	81
May	5	1	-	-
Jun	2	-2	-2	4
Jul	1	-3	-3	9
Aug	0	-4	-4	16
Sep	4	0	-	-
Oct	3	-1	-1	1
Nov	0	-4	-4	16
Dec	2	-2	-2	4
Sum				**167**

$$\text{Target semideviation} = \sqrt{\frac{167}{11}} = 3.8964\%$$

The target downside deviation will be less than the standard deviation, because deviations above the target are ignored. As the target is increased, the target downside deviation will increase.

10.1 Coefficient of Variation

Coefficient of variation expresses how much dispersion exists relative to the mean of a distribution and allows for direct comparison of dispersion across different data sets, even if the means are drastically different from one another. It is used in investment analysis to compare relative risks. When evaluating investments, a lower value is better. Coefficient of variation is expressed as:

$$\text{CV} = \frac{s}{\overline{X}}$$

where: s = sample standard deviation of a set of observations and $\overline{X}$ = sample mean

Example

Investment A has a mean return of 7% and a standard deviation of 5%. Investment B has a mean return of 12% and a standard deviation of 7%. Calculate the coefficients of variation.

Solution

The coefficients of variation can be calculated as follows:

$$\mathrm{CV_A} = \frac{5\%}{7\%} = 0.71$$

$$\mathrm{CV_B} = \frac{7\%}{12\%} = 0.58$$

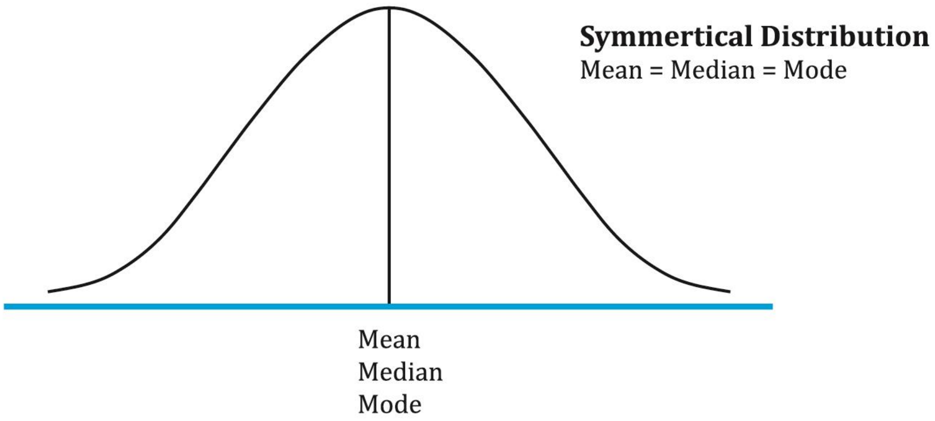
This metric shows that Investment A is riskier than Investment B.

11. The Shape of the Distributions

Symmetrical distribution

A distribution is said to be symmetrical when the distribution on either side of the mean is a mirror image of the other.

In a normal distribution, mean = median = mode.

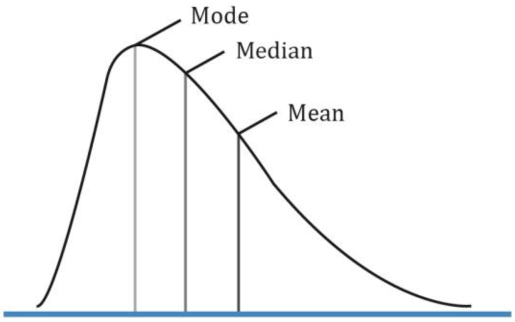


If a distribution is non-symmetrical, it is said to be skewed. Skewness is a measure of the asymmetry of the probability distribution. Skewness can be negative or positive.

Positively skewed distribution

A positively skewed distribution has a long tail on the right side, which means that there will be limited but frequent downside returns and unlimited but less frequent upside returns.

Here the mean > median > mode. The extreme values affect the mean the most which is pulled to the right. They affect the mode the least.

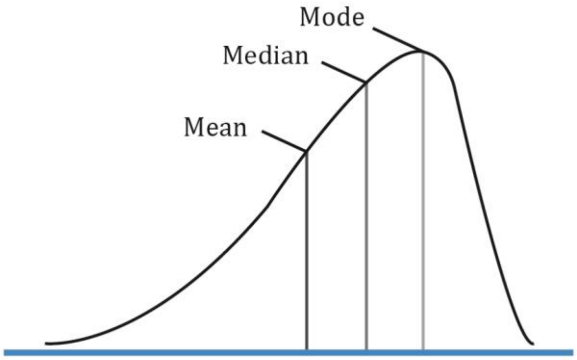


Right - Skewed (Positive Skewness)

Negatively skewed distribution

A negatively skewed distribution has a long tail on the left side, which means that there will be limited but frequent upside returns and unlimited but less frequent downside returns.

Here the mean < median< mode. The extreme values affect the mean the most which is pulled to the left. They affect the mode the least.



Left - Skewed (Negative Skewness)

Instructor's Note: Investors prefer positive skewness because it has a higher chance of very large returns and also because it has a higher mean return.

Example:

Which of the following distribution is most likely characterized by frequent small losses and a few extreme gains?

A. Normal distribution
B. Negatively skewed
C. Positively skewed

Solution:

C is correct. A positively skewed distribution is characterized by frequent small losses and a few extreme gains.

Example:

Which of the following is most likely to be true for a negatively skewed distribution?

A. Mean < Median < Mode
B. Mode < Median < Mean
C. Median < Mean < Mode

Solution:

A is correct. In a negatively skewed distribution, the mean < median < mode.

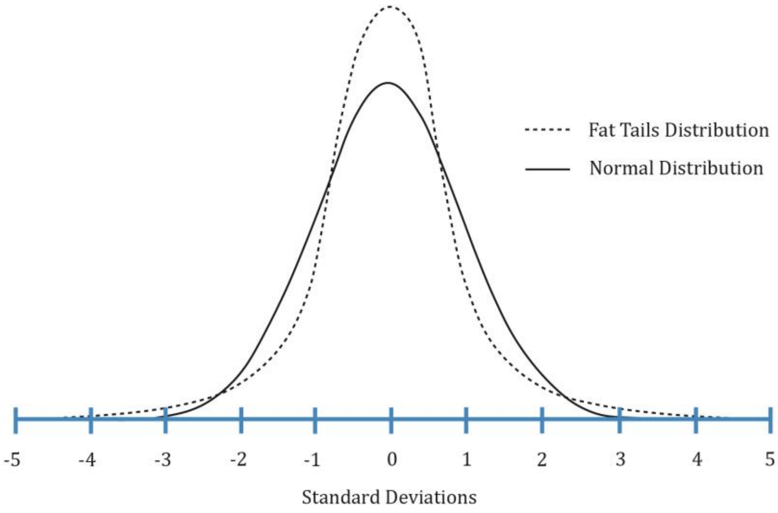
11.1 The Shape of the Distributions: Kurtosis

Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution.

Excess kurtosis = kurtosis - 3. An excess kurtosis with an absolute value greater than one is considered significant.

- A **leptokurtic** distribution has fatter tails than a normal distribution. It has an excess kurtosis greater than 0.
- A **platykurtic** distribution has thinner tails than a normal distribution. It has an excess kurtosis less than 0.
- A **mesokurtic** distribution is identical to a normal distribution. It has an excess kurtosis equal to 0.

The following figure shows a leptokurtic distribution. As compared to a normal distribution, a leptokurtic distribution is more likely to generate observations in the tail region. It is also more likely to generate observations near the mean. However, to have the total probabilities sum to 1, it will generate fewer observations in the remaining regions (i.e. regions between the central and the two tail regions)



12. Correlation Between Two Variables

Covariance

Covariance is a measure of how two variables move together. The formula for computing the **sample covariance** of X and Y is:

$$s_{XY} = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

The problem with covariance is that it can vary from negative infinity to positive infinity which makes it difficult to interpret. To address this problem, we use another measure called correlation.

Correlation

Correlation is a standardized measure of the linear relationship between two variables with values ranging between -1 and +1.

The **sample correlation coefficient** can be calculated as:

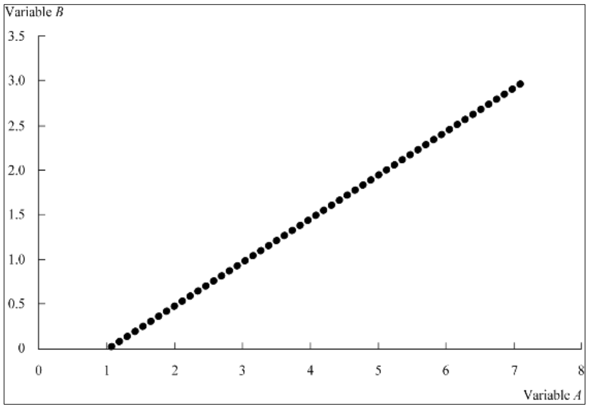
$$r_{XY} = \frac{s_{XY}}{s_x * s_y}$$

12.1 Properties of Correlation

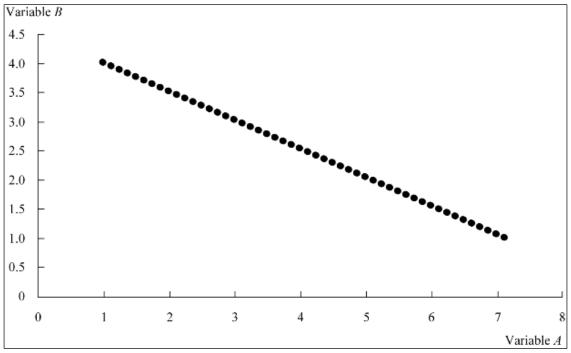
- Correlation ranges from -1 and +1.
- A correlation of 0 (uncorrelated variables) indicates an absence of any linear (straight-line) relationship between the variables.
- A correlation of +1 indicates a perfect positive relationship.
- A correlation of -1 indicates a perfect negative relationship.

The three scatter plots below show a positive linear, negative linear, and no linear relation between two variables A and B. They have correlation coefficients of +1, -1 and 0 respectively.

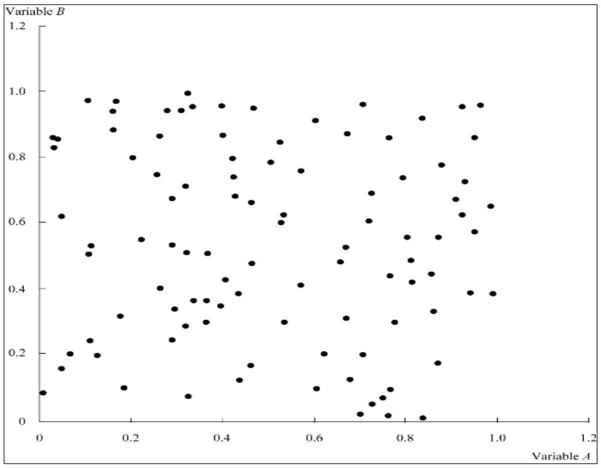
Variables with a correlation of 1.



Variables with a correlation of -1.



Variables with a correlation of 0.



12.2 Limitations of Correlation Analysis

The correlation analysis has certain limitations:

- Two variables can have a strong non-linear relation and still have a very low correlation.
- The correlation can be unreliable when outliers are present.
- The correlation may be spurious. **Spurious correlation** refers to the following situations:
    - The correlation between two variables that reflects chance relationships in a particular data set.
    - The correlation induced by a calculation that mixes each of two variables with a third variable.
    - The correlation between two variables arising not from a direct relation between them, but from their relation to a third variable. Ex: shoe size and vocabulary of school children. The third variable is age here. Older shoe sizes simply imply that they belong to older children who have a better vocabulary.

Summary

LO.a: Identify and compare data types.

Data can be defined as a collection of numbers, characters, words and text - as well as images, audio, and video - in a raw or organized format to represent facts or information.

Data can be classified in three ways:

- Numerical versus categorical data
- Cross-sectional versus time-series versus panel data
- Structured versus unstructured data

LO.b: Describe how data are organized for quantitative analysis.

Raw data is typically organized into either a one-dimensional array or a two-dimensional rectangular array (also called a data table) for quantitative analysis.

- A one-dimensional array is suitable for representing a single variable.
- A two-dimensional array consists of columns and rows to hold multiple variables and multiple observations, respectively.

LO.c: Interpret frequency and related distributions.

A frequency distribution is a tabular display of data summarized into a relatively small number of intervals.

The steps for constructing a frequency distribution for a categorical variable are:

1. Count the number of observations for each unique value of the variable.
2. Construct a table listing each unique value and the corresponding counts.
3. Sort the records by number of counts in descending or ascending order.

The steps for constructing a frequency distribution for numerical variables are:

1. Sort the data in ascending order.
2. Calculate the range of data.
3. Decide on the number of bins (k).
4. Determine bin width.
5. Determine bins.
6. Determine the number of observations in each bin.
7. Construct a table of the bins listed from smallest to largest.

Frequency distributions allow us to assess how data is distributed.

LO.d: Interpret a contingency table.

A contingency table is a tabular format that displays the frequency distributions of two or more categorical variables simultaneously. It can be used to find patterns between the variables. Contingency tables are constructed by listing all levels of one variable as rows and all the levels of the other variables as columns in a table.

One application of contingency tables is for evaluating the performance of a classification model (using a confusion matrix). Another application of contingency tables is to investigate a potential association between two categorical variables by performing a chi-square test of independence.

LO.e: Describe ways that data may be visualized and evaluate uses of specific visualizations.

Visualization refers to the presentation of data in pictorial or graphical format to aid understanding of the data and for gaining insights into the data. The different ways in which data can be visualized are:

- Histograms
- Frequency polygons
- Bar charts
- Tree maps
- Word clouds
- Line charts
- Scatter plots
- Heat maps

LO.f: Describe how to select among visualization types.

To explore/present relationships between variable we can use the following visualization types:

- Two variables: Scatter plot
- More than two variables: Scatter plot matrix, heat map.

To explore/present distributions we can use the following visualization types:

- Numerical data: Histogram, frequency polygon, cumulative distribution chart.
- Categorical data: Bar chart, tree map, hear map
- Unstructured data: Word cloud.

To make comparisons we can use the following visualization types:

- Comparison among categories: Bar chart, tree map, heat map
- Comparison over time: Line chart for two variables, bubble line chart for three variables.

LO.g: Calculate and interpret measures of central tendency.

Measures of central tendency specify where data are centered.

The arithmetic mean is the sum of the observations divided by the number of observations. It is the most frequently used measure of the middle or center of data.

The median is the midpoint of a data set that has been sorted into ascending or descending order.

The mode is the most frequently occurring value in a distribution.

In a weighted mean, instead of each data point contributing equally to the final mean, some data points contribute more “weight” than others.

The geometric mean is calculated as the n^{th} root of a product of n numbers.

The harmonic mean is a special type of weighted mean in which an observation’s weight is inversely proportional to its magnitude.

LO.h: Evaluate alternative definitions of mean to address an investment problem.

- Arithmetic mean: Should be used with single period or cross-sectional data.
- Geometric mean: Should be used with time-series data.
- Weighted mean: Should be used when different observations have different weights.
- Harmonic mean: Should be used to find average purchase price for equal periodic investments.
- Trimmed mean: Should be used when the data has extreme outliers.
- Winsorized mean: Should be used when the data has extreme outliers.

LO.i: Calculate quantiles and interpret related visualizations.

A quantile is a value at or below which a stated fraction of the data lies. Some examples of quantiles include:

- Quartiles: The distribution is divided into quarters.
- Quintiles: The distribution is divided into fifths.
- Deciles: The distribution is divided into tenths.
- Percentile: The distribution is divided into hundredths.

The formula for the position of a percentile in a data set with n observations sorted in ascending order is:

$$L_y = \frac{(n + 1)y}{100}$$

A box and whiskers plot is used to visualize the dispersion of data across quartiles. The box represents the interquartile range. The whiskers represent the highest and lowest values of the distribution. There are several variations of the box and whiskers plot. Sometimes the whiskers may be a function of the interquartile range instead of the highest and lowest values.

LO.j: Calculate and interpret measures of dispersion.

Measures of dispersion describe the variability of outcomes around the mean.

The range is the difference between the maximum and minimum values in a data set.

MAD is the average of the absolute values of deviations from the mean.

Variance is defined as the average of the squared deviations around the mean. Standard deviation is the positive square root of the variance.

Coefficient of variation expresses how much dispersion exists relative to the mean of a distribution and allows for direct comparison of dispersion across different data sets, even if the means are drastically different from one another.

LO.k: Calculate and interpret target downside deviation.

The target downside deviation, or target semideviation, is a measure of the risk of being below a given target. It is calculated as the square root of the average squared deviations from the target, but it includes only those observations below the target (B).

LO.l: Interpret skewness.

Skewness is a measure of the asymmetry of the probability distribution. If a distribution is non-symmetrical, it is said to be skewed. Skewness can be negative or positive.

A positively skewed distribution has a long tail on the right side, which means that there will be frequent small losses and few large gains.

A negatively skewed distribution has a long tail on the left side, which means that there will be frequent small gains and few large losses.

LO.m: Interpret kurtosis.

Kurtosis is a measure of the combined weight of the tails of a distribution relative to the rest of the distribution.

Excess kurtosis = kurtosis - 3. An excess kurtosis with an absolute value greater than one is considered significant.

- A leptokurtic distribution has fatter tails than a normal distribution. It has an excess kurtosis greater than 0.
- A platykurtic distribution has thinner tails than a normal distribution. It has an excess kurtosis less than 0.
- A mesokurtic distribution is identical to a normal distribution. It has an excess kurtosis equal to 0.

LO.n: Interpret correlation between two variables.

Correlation is a statistic that measures the degree to which two variables move in relation to each other.

- Correlation ranges from -1 and +1.
- A correlation of 0 (uncorrelated variables) indicates an absence of any linear (straight-line) relationship between the variables.
- A correlation of +1 indicates a perfect positive relationship.
- A correlation of -1 indicates a perfect negative relationship.

Practice Questions

1. Which of the following *best* describes a panel data?
   A. Daily market capitalization for 20 companies in S&P 500 in a given year by day.
   B. Market capitalization for 20 companies in S&P 500 on 30th June 2019.
   C. Daily market capitalization of the XYZ Company in S&P 500 during the calendar year 2019.

2. An analyst is analyzing different bonds and has classified them by type of issuers including financial, insurance, and corporate. This is *most likely* an example of which of the following measurement scale?
   A. Categorical and ordinal.
   B. Nominal and discrete.
   C. Categorical and nominal.

3. Earnings per share (EPS) of a company is *most likely* an example of:
   A. Ordinal data.
   B. Continuous data.
   C. Discrete data.

4. An analyst is comparing the ROE of five different companies belonging to Automobile Parts & Accessories sector for the year ended December 31, 2019. Which of the following would be *most likely* suitable for him for organizing this information?
   A. Two-dimensional rectangular array because the data he is using represents a panel data.
   B. One-dimensional rectangular array because the data he is using represents a time-series data.
   C. One-dimensional rectangular array because the data he is using represents a cross-sectional data.

5. For the following frequency distribution, the number of intervals, the sample size, and the relative frequency of the third interval are *closest* to:

Returns	**Frequency**
-10% up to 0%	4
0% up to 10%	13
10% up to 20%	5
20% up to 30%	3

   A. 4, 25, 20%.
   B. 1, 20, 10%.
   C. 3, 30, 15%.

6. Tristar, a mutual fund investment company, is studying the relationship between investor's self-perceived behavior toward investing and selection of mutual funds made by the investor. Tristar samples a total of 500 potential investors who have indicated their intention to invest in mutual funds. Two types of funds are specified for possible purchase. The results of the sample are shown below:

	Income Fund	Equity Fund
Risk-taker Investors	30	240
Risk-averse Investors	180	50

   The marginal frequency of risk-averse investors is *closest* to:
   A. 230.
   B. 210.
   C. 180.

7. Which of the following graphical tools for displaying data require the mid points to be plotted for each interval?
   A. Frequency polygon.
   B. Histogram.
   C. Cumulative frequency curve.

8. Which of the following is *most likely* true about the stacked bar chart?
   A. The overall height of the stacked bar represents the marginal frequency for the category.
   B. The overall height of the stacked bar represents the absolute frequency for the category.
   C. The overall height of the stacked bar represents the joint frequency for the category.

9. The visualization tool that can be used to depict the shape, center, and spread of the distribution of numerical data is *most likely*:
   A. Scatter plot.
   B. Heat map.
   C. Histogram.

10. The following ten observations are a sample drawn from a normal population: 20, 15, 12, 6, 4, -11, 19, 14, -3, and 19. The arithmetic mean of the sample is *closest* to:
   A. 8.6.
   B. 9.5.
   C. 10.2.

11. Considering the following set of numbers (which have been arranged in ascending order)

39	40	41	41	41
43	48	53	55	55

Which of the following statements is most accurate?
A. The mode is larger than the mean.
B. The median is smaller than the mean but larger than the mode.
C. The mean is smaller than both the mode and the median.

12. Over the past five years i.e. from 2015 to 2019, a portfolio gave returns of 16%, 10% -7%, -11% and 8%. The average return is best represented as the:
A. harmonic mean of 21.23%.
B. arithmetic average of 7.60%.
C. geometric mean of 14.06%.

13. The following table shows the annual returns of a Portfolio of an investor for the past 12 years.

Year	**Annual Return**
1	5.5
2	6.2
3	-2.2
4	-0.6
5	-0.6
6	1.3
7	2.7
8	3.0
9	3.3
10	4.1
11	4.9
12	5.4

The 7th decile percentage of the annual returns is *closest* to:
A. 5.85
B. 4.95
C. 7.60

14. The table below shows data on volatility of a series of funds:

Fund	**Volatility (%)**
Fund 1	5.05
Fund 2	6.20
Fund 3	6.93

Fund 4	7.56
Fund 5	8.25
Fund 6	10.11
Fund 7	11.36
Fund 8	14.52
Fund 9	15.02
Fund 10	15.66
Fund 11	15.98
Fund 12	16.01
Fund 13	19.25

The value of the second quintile is *closest* to:

A. 7.56%.
B. 9.37%.
C. 10.11%.

15. The second quartile represents which of the following?

I. Median
II. 50th percentile
III. 2nd quintile
IV. 5th decile

A. I and II only.
B. I, II, and IV only.
C. II, III, and IV only.

16. The annual returns of a portfolio are given below:

Year	**Portfolio return**
Year 1	6.5%
Year 2	8.2%
Year 3	10.5%
Year 4	-5.4%
Year 5	7.7%

The portfolio's range and mean absolute deviation for the five-year period are *closest* to:

A. 15.90% and 5.10%.
B. 5.10% and 4.36%.
C. 15.90% and 4.36%.

17. The dividend yield for five hypothetical companies is given below:

Company	Dividend Yield %
Paknama	10.5%
Genie Ltd.	16.25%
Mirinda Corp.	27.0%
Tina Travels Ltd.	12.0%
Thomas Press Ltd.	7.8%

The population variance is *closest* to:

A. 36.89.
B. 45.20.
C. 56.49.

18. George Baker, an equity fund manager has the following information about a common stock portfolio:

Arithmetic mean return	11.8%
Geometric mean return	10.6%
Portfolio beta	1.2
Risk-free rate of return	4.25%
Variance of returns	196

From the given information, the coefficient of variation is *closest* to:

A. 1.32.
B. 1.24.
C. 1.18.

19. Which of the following relationships best characterize a negatively skewed distribution?

A. Mean < median < mode.
B. Mode < median < mean.
C. Median < mean < mode.

20. Which of the following return distribution is most likely characterized by frequent small losses and a few large gains?

A. Normal distribution.
B. Negatively skewed.
C. Positively skewed.

21. A distribution more peaked than the normal distribution is best described as being:

A. platykurtic.

B. mesokurtic.
C. leptokurtic.

22. A correlation of 0.34 between two variables, *X* and *Y*, is *best* described as:
A. changes in X causing changes in Y.
B. a positive association between X and Y.
C. a curvilinear relationship between X and Y.

23. Which of the following is a potential problem with interpreting a correlation coefficient?
A. Outliers
B. Spurious correlation
C. Both outliers and spurious correlation

Solutions

1. A is correct. Daily market capitalization (i.e., the variable) for 20 companies (i.e., the observational units) in a given year by day is a panel data. Panel data is a combination of time-series and cross-sectional data. It consists of observations through time on one or more variables for multiple subjects. B is an example of cross-sectional data. Cross-sectional data consists of observations for multiple subjects taken at a specific point in time. C is an example of time-series data. Time-series data consists of observations for a single subject taken at specific and equally spaced intervals of time. LO. a

2. C is correct. Classification of bonds by type of issuers is an example of nominal and categorical values that cannot be organized in a logical order. A is incorrect because such classification cannot be organized in a logical order. B is incorrect because discrete is a type of numerical data. Discrete data can take numerical values that result from a counting process. LO. a.

3. B is correct. Continuous data are data that can be measured and can take on any numerical value in a specified range of values. A is incorrect because ordinal data are categorical values that can be logically ordered or ranked. C is incorrect because discrete data are numerical values that result from a counting process. LO. a.

4. C is correct. Cross-sectional data are a list of the observations of a specific variable (i.e. ROE) from multiple observational units (five different companies) at a given point in time (year ended December 31, 2019). A one-dimensional (not a two-dimensional rectangular) array would be most suitable for organizing a collection of cross-sectional data. Panel data consist of observations through time on one or more variables for multiple observational units. A two-dimensional rectangular array, or data table, would be suitable here as it is comprised of columns to hold the variable(s) for the observational units and rows to hold the observations through time. Hence, A is incorrect. B is incorrect because the analyst is not using time-series data. Time-series data are a sequence of observations for a single observational unit of a specific variable collected over time and at discrete and typically equally spaced intervals of time, such as daily, weekly, monthly, annually, or quarterly. LO. b.

5. A is correct. An interval is the set of return values that an observation falls within. There are four intervals. The sample size is the sum of all of the frequencies in the distribution: 4 + 13 + 5 + 3 = 25. The relative frequency is found by dividing the frequency of the interval by the total number of frequencies: 5/25 = 20%. LO. c.

6. A is correct. The marginal frequency of risk-averse investors is the sum of the joint frequencies across all two levels of bond rating, so 180 + 50 = 230. B is incorrect because

210 is the marginal frequency for income funds, i.e. 30 + 180 = 210. C is incorrect because 180 is the joint frequency. LO. d.

7. A is correct. For a frequency polygon, the mid points for each interval are plotted on the x-axis and the absolute frequency for that interval on the y-axis. On the other hand, histograms require the absolute frequency to be plotted on the y-axis and intervals on the x-axis. The cumulative frequency curve plots the cumulative frequency or cumulative relative frequency against the upper interval limit. LO.e.

8. A is correct. An alternative form for presenting the joint frequency distribution of two categorical variables is a stacked bar chart. In the vertical version of a stacked bar chart, the bars representing the sub-groups are placed on top of each other to form a single bar. Each subsection of the bar is shown in a different color to represent the contribution of each sub-group, and the overall height of the stacked bar represents the marginal frequency for the category. LO. e.

9. C is correct. A histogram depicts the shape, center, and spread of the distribution of numerical data. A histogram is a chart that presents the distribution of numerical data by using the height of a bar or column to represent the absolute frequency of each bin or interval in the distribution. A is incorrect. A scatter plot is used to visualize the joint variation in two numerical variables. B is incorrect. A heat map is commonly used for visualizing the degree of correlation between different variables. LO. f.

10. B is correct. The sum of the ten numbers is 95. Dividing by 10 gives the mean of 9.5. LO. g.

11. B is correct. The mode is the most frequent value in the set of items and thus is equal to 41.
The mean is the average value from the set of items and is computed as follows:
Mean = Sum of Observations / Number of Observations = 456 / 10 = 45.6
The median is the value of the middle item of a set of items. For even number of observations, the median is equal to the average of the middle two values. The median is thus the average of 41 and 43. Median = 42. Therefore, the median is smaller than the mean but larger than the mode. A is incorrect – the mode is not larger than the mean, since mode = 41 < 45.6 = mean. C is incorrect since the mean (45.6) is larger than both the mode (41) and the median (42). LO.g.

12. C is correct. The geometric mean is calculated as the n^th^ root of a product of n numbers. The most common application of the geometric mean is to calculate the average return of an investment. The formula is:

$$R_G = [(1 + R_1)(1 + R_2) \dots (1 + R_n)]^{\frac{1}{n}} - 1$$

R = {[(1 + 0.16) * (1 + 0.10) * (1 - 0.07)*(1 - 0.11) * (1 + 0.08)] } -1 = 0.1406 or 14.06%
A is incorrect. The harmonic mean is a special type of weighted mean in which an observation's weight is inversely proportional to its magnitude. The formula for a harmonic mean is:

$$X_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

where: $X_i > 0$ for i = 1, 2 … n, and n is the number of observations
The harmonic mean is used to find average purchase price for equal periodic investments.
B is incorrect. The arithmetic mean is the sum of the observations divided by the number of observations. It is the most frequently used measure of the middle or center of data. The arithmetic mean is appropriate for forecasting single period returns. LO.h.

13. B is correct. First arrange the data in ascending order:

Annual Return
-2.2
-0.6
-0.6
1.3
2.7
3
3.3
4.1
4.9
5.4
5.5
6.2

L70 = (12 + 1) (70/100) = 9.1
Use linear interpolation, which estimates an unknown value on the basis of two known values that surround it. In this case, the 9th value is 4.9 and the 10th value is 5.4. The 7th decile is: P70 = 4.9+ 0.05 (0.1 times the linear distance between 4.9 and 5.4). P70 = 4.95 LO.i.

14. B is correct. Quintiles divide data into five parts. Hence, the first quintile corresponds to the 20th percentile and the second quintile corresponds to the 40th percentile. The location can be determined using:

$$L_y = \frac{(n+1)y}{100}$$

$$L_y = \frac{(13+1)40}{100} = 5.6$$

The value corresponding to location 5 (Fund 5) is 8.25%. The value corresponding to location 6 (Fund 6) is 10.11%. The approximate value corresponding to location 5.6 can be estimated using linear interpolation: $ 8.25% + (0.6 * (10.11% - 8.25%))= 9.37% LO.i.

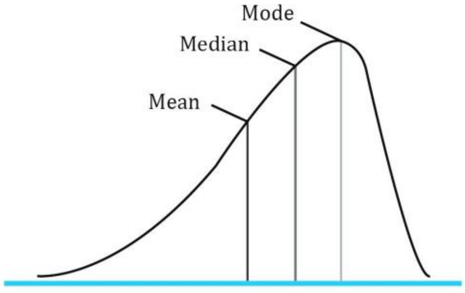
15. B is correct. A quartile is a quarter of the observations, or 25%, so the second quartile is the same as the 50th percentile. The second quartile is equivalent to the median (the middle most position or 50th percentile), the 50th percentile, and the 5th decile (the fifth portion if the data is divided into 10 portions, i.e. the 50th percentile). LO. i.

16. C is correct. Range = Highest value – Lowest value = 10.5% - (- 5.4%) = 15.9%
To calculate MAD, first compute the mean: (6.5 + 8.2 + 10.5 – 5.4 + 7.7) / 5 = 5.5% and compute MAD, (|6.5 – 5.5| + |8.2 – 5.5| + |10.5 – 5.5| + | –5.4 – 5.5| + |7.7 – 5.5|) / 5 = 4.36%. LO.j.

17. B is correct. Use the following keystrokes to calculate the population variance: [2nd] [DATA]
[2nd] [CLR WRK] X01 = 10.5
X02 = 16.25
X03 = 27
X04 = 12
X05 = 7.8
Choose the population standard deviation: σ = 6.723. Then square it, to get population variance as 45.2.
Be sure to read the question carefully so as not to confuse the sample variance with the population variance. Option C is the sample variance, but that is not the correct answer because the question asks for the population variance. LO.j.

18. C is correct. The coefficient of variation is: Standard deviation of return / Mean return = sqrt (196) / 11.8 = 1.18. LO. k.

19. A is correct. For a negatively skewed distribution, the mean is less than the median, which is less than the mode. LO.l.



Left - Skewed (Negative Skewness)

20. C is correct. A positively skewed distribution has frequent small losses and a few large gains. The result is that the extreme gains pull the mean to the right while the mode resides on the left with the bulk of the observations. A negatively skewed distribution has frequent small gains and a few large losses. The result is that the extreme losses pull the mean to the left while the mode resides on the right with the bulk of the observations. A normal distribution is symmetrical. LO.l.

21. C is correct. Leptokurtic describes a distribution that is more peaked than the normal distribution. Platykurtic is a distribution less peaked than a normal distribution. Mesokurtic is a distribution as peaked as the normal distribution. LO.m.

22. B is correct. The correlation coefficient is positive, indicating that the two series move together. LO.n.

23. C is correct. Both outliers and spurious correlation are potential problems with interpreting correlation coefficients. LO.n.