Level I of the CFA® 2025 Exam

Study Notes - Quantitative Methods

Offered by AnalystPrep

Last Updated: Aug 1, 2024

Table of Contents

Learning Module 1: Rate and Return

LOS 1a: interpret interest rates as required rates of return, discount rates, or opportunity costs and explain an interest rate as the sum of a real risk-free rate and premiums that compensate investors for bearing distinct types of risk

The time value of money is a concept that states that cash received today is more valuable than cash received in the future. If a person agrees to receive payment in the future, he foregoes the option of earning interest if he invests that amount of money today.

An interest rate or yield, usually denoted by *r,* is a rate of return that reflects the connection between cash flows dated at different times.

Assume you currently possess $100. Next, consider depositing this money into a savings account, expecting it to grow to $110 after one year. Intuitively, the compensation required for deferring the consumption of $100 now in favor of receiving $110 in one year is $10 (equal to 110 minus 100). This compensation is equivalent to a 10% rate of return (calculated as 10 divided by 100).

There are three ways to interpret interest rates:

1. **Required rate of return**: The minimum return an investor expects to earn to accept an investment.
2. **Discount rate**: The rate used to discount future cash flows to allow for the time value of money (to determine the present value **equivalent** of some money to be received sometime in the future). Discount rates and interest rates are used almost interchangeably.
3. **Opportunity cost**: The value of the **best-forgone alternative**; the most valuable alternative investors give up when they choose what to do with money.

Determinants of Interest Rates

Economics postulates that the forces of supply and demand determine interest rates. In this case, the investors (lenders) supply the money, and the borrowers demand money for their

consumption.

As such, interest is a reward a borrower pays for using an asset, usually capital, belonging to a lender. It is compensation for the loss or value depreciation occasioned by the use of the asset.

Therefore, an interest rate is composed of a real risk-free interest rate plus a set of four premiums that represent compensation for bearing distinct types of risk:

$$\begin{aligned} \text{Interest (r)} = \; & \text{Real risk-free interest rate} \\ & + \text{Inflation Premium} \\ & + \text{Default risk premium} \\ & + \text{Liquidity premium} \\ & + \text{Maturity premium} \end{aligned}$$

The Real Risk-free Interest Rate

The real risk-free interest rate is the single-period interest rate for a completely risk-free security if no inflation is expected. According to economic theory, the real risk-free rate reflects people's preferences for current compared to real future consumption.

Types of Risk Premiums

Inflation Risk Premium

Inflation risk is the loss of purchasing power of money as a result of the increase in prices of consumer goods.

The inflation premium compensates investors for expected inflation. It represents the average inflation rate expected over the maturity of the debt. The risk of a decrease in purchasing power validates the inflation risk premium.

Liquidity Risk Premium

Liquidity refers to the ease with which an investment can be converted into cash without significantly sacrificing market value.

The liquidity premium compensates investors for the risk of loss relative to an investment's fair value if the investment needs to be converted to cash quickly.

Default Risk Premium

Default risk describes a situation where a borrower may fail to repay borrowed funds as a result of bankruptcy. This might result in significant losses on the side of the lender.

The **default risk premium** compensates investors for the possibility that the borrower **will fail to make a promised payment** at the contracted time and in the contracted amount.

Maturity Risk Premium

The maturity risk premium is the additional return an investor requires for assuming interest rate and reinvestment risk resulting from a longer investment maturity timeline. Maturity risk premium increases with an increase in the maturity timeline. In other words, the longer the maturity timeline of an investment, the higher the maturity risk premium.

Nominal Risk-free Interest Rate

The nominal risk-free interest rate is defined as the sum of the real risk-free interest rate and the inflation premium. In other words, the nominal risk-free interest rate can be seen as the combination of the real risk-free rate plus an inflation premium, as shown by the following equation:

$$(1+\text{Nominal risk-free rate})=(1+\text{Real risk-free rate})(1+\text{Inflation premium})$$

The above equation is generally approximated as follows:

$$\text{Nominal risk-free rate}=\text{Real risk-free rate}+\text{Inflation premium}$$

Most rates quoted on short-term government debts can be taken as nominal risk-free interest rates over the respective maturity.

Question

Which of the following is *most likely* an interpretation of interest rate as a benefit foregone when investors spend money on current consumption instead of saving or investing?

A. Discount rate.

B. Opportunity cost.

C. Required rate of return.

Solution

The correct answer is **B**.

Opportunity cost is a key factor in interpreting interest rates. It refers to the interest foregone when investors opt for an alternate option, such as spending on current consumption instead of saving or investing.

A is incorrect. The discount rate is the interest rate used to discount future cash flows to reach the present value.

C is incorrect. The required rate of return is the minimum rate of return an investor would wish to earn to postpone current consumption.

LOS 1b: calculate and interpret different approaches to return measurement over time and describe their appropriate uses

Financial assets are primarily defined based on their return-risk characteristics. This helps when building a portfolio from all the assets available. Regarding returns, there are different ways of measuring returns.

Financial market assets generate two different streams of return: income through cash dividends or interest payments and capital gain or loss through financial asset price increases or decreases.

Some financial assets give only one stream of return. For instance, headline stock market indices typically report on price appreciation only. They do not include the dividend income unless the index specifies it is a “total return” series.

Holding Period Return

A holding period return is earned from holding an asset for a single specified period. The time period can be any specified period, such as a day, month, or ten years.

The general formula of the holding period return is given by:

$$R = \frac{(P_1 - P_0) + I_1}{P_0}$$

P_0 = Price of an asset at the beginning of the period (t=0).

P_1 = Price of an asset at the end of the period (t=1).

I_1 = Income received at the end of the period (t=1).

Example: Calculating Holding Period Return

An investor purchased 100 shares of a stock at $50 per share and held the investment for one year. During that period, the stock paid dividends of $2 per share. At the end of the year, the

investor sold all the shares for $60 per share.

The holding period return is *closest to*:

Solution

In this case, we have:

$$\begin{aligned} P_0 &= 100 \text{ shares} \times \$50 \text{ per share} = \$5,000 \\ I_1 &= 100 \text{ shares} \times \$2 \text{ per share} = \$200 \\ P_1 &= 100 \text{ shares} \times \$60 \text{ per share} = \$6,000 \end{aligned}$$

Therefore,

$$R = \frac{(P_1 - P_0) + I_1}{P_0} = \frac{6,000 - 5,000 + 200}{5,000} = 24\%$$

Holding period returns can also be calculated for periods longer than a year. For instance, if we need to calculate the holding period return for a five-year period, we should compound the five annual returns as follows:

$$R = \frac{(P_5 - P_0) + I_{(1-5)}}{P_0}$$

Arithmetic Return

When we have assets for multiple holding periods, it is necessary to aggregate the returns into one overall return.

Denoted by $\bar{R}_i$ arithmetic mean for an asset i is a simple process of finding the average holding period returns. It is given by:

$$\bar{R}_i = \frac{R_{i,1} + R_{i,2} + \ldots + R_{i,T-1} + R_{iT}}{T} = \frac{1}{T}\sum_{t=1}^{T} R_{it}$$

Where:

R_{it} = Return of asset i in period t.

T = Total number of periods.

For example, if a share has returned 15%, 10%, 12%, and 3% over the last four years, then the arithmetic mean is computed as follows:

$$\bar{R}_i = \frac{1}{T}\sum_{t=1}^{T} R_{it} = \frac{1}{4}(15\% + 10\% + 12\% + 3\%) = 10\%$$

Geometric Return

Computing a geometric mean follows a principle similar to the one used to compute compound interest. It involves compounding returns from the previous year to the initial investment's value at the start of the new period, allowing you to earn returns on your returns.

A geometric return provides a more accurate representation of the portfolio value growth than an arithmetic return.

Denoted by $\bar{R}_{Gi}$ the geometric return for asset i is given by:

$$\begin{aligned}\bar{R}_{Gi} &= \sqrt[T]{(1 + R_{i,1}) \times (1 + R_{i2}) \times \ldots \times (1 + R_{i,T-1}) \times (1 + R_{iT})} - 1 \\ &= \sqrt[T]{\prod_{t=1}^{T} (1 + R_t)} - 1\end{aligned}$$

Using the same annual returns of 15%, 10%, 12%, and 3% as shown above, we compute the geometric mean as follows:

$$\begin{aligned}\text{Geometric mean} &= [(1 + 15\%) \times (1 + 10\%) \times (1 + 12\%) \times (1 + 3\%)]^{\frac{1}{4}} - 1 \\ &= 9.9\%\end{aligned}$$

Note that the geometric return is slightly less than the arithmetic return. Arithmetic returns tend to be biased upwards unless the holding period returns are all equal.

Harmonic Mean

The harmonic mean is a measure of central tendency. It's especially useful for rates or ratios such as P/E ratios. Its formula is derived from the harmonic series, which is a specific mathematical sequence.

$$\bar{X}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}},\ X_i > 0 \text{ for all } i = 1, 2, \dots, n$$

The above formula is interpreted as the “harmonic mean of observations $X_1,\ X_2, \dots,\ X_n$.”

The harmonic mean is handy for averaging ratios when those ratios are consistently applied to a fixed quantity, resulting in varying unit numbers. For instance, it's applied in cost-averaging strategies where you invest a fixed amount of money at regular intervals.

Example: Calculating the Harmonic Mean

An investor is practicing cost averaging by investing in a particular stock over a period of three months. The investor decides to allocate different amounts of money each month. In the first month, the investor invests $2,000; in the second month, $3,000; and in the third month, $4,000. The share prices of the stock for these three months are $10, $12, and $15, respectively.

Calculate the average price paid per share for the three-month period.

Solution

Using the harmonic mean formula,

$$\bar{X}_H = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}} = \frac{3}{\frac{1}{10} + \frac{1}{12} + \frac{1}{15}} = 12$$

Trimmed and Winsorized Means

Trimmed and Winsorized means seek to lower the effect of outliers in a data set.

Trimmed Mean

The trimmed mean is a measure of central tendency in which we calculate the mean after excluding a small percentage of the lowest and highest values from the dataset.

For example, a data set consists of 10 observations: 12, 15, 18, 20, 22, 25, 27, 30, 35, and 40. We can calculate the trimmed mean after removing the highest and lowest values.

After removing these values, the remaining data set is: 15, 18, 20, 22, 25, 27, 30, and 35.

Now, let's calculate the trimmed mean by taking the average of these remaining values:

$$\frac{15+18+20+22+25+27+30+35}{8} = \frac{192}{8} = 24$$

Therefore, the trimmed mean of the given data set is 24.

Winsorized Mean

The Winsorized mean is a central tendency measure. It works by replacing extreme values at both ends of the data with the values of their closest observations. This process is similar to the trimmed mean. Essentially, it helps eliminate outliers in a dataset.

For example, consider a dataset of 12 observations: 8, 12, 15, 18, 20, 22, 25, 27, 30, 35, 40, and 50. We can calculate the Winsorized mean by replacing the lowest and highest values with those closest to the 10th and 90th percentiles, respectively. As such, the new values are **10**, 12, 15, 18, 20, 22, 25, 27, 30, 35, **37.5**, and 40, and the winsorized mean is:

$$\frac{10+12+15+18+20+22+25+27+30+35+37.5+40}{12} \approx 24.46$$

Question 2

What are the arithmetic mean and geometric mean, respectively, of an investment that returns 8%, -2%, and 6% each year for three years?

A. Arithmetic mean = 5.3%; Geometric mean = 5.2%.

B. Arithmetic mean = 4.0%; Geometric mean = 3.6%.

C. Arithmentic mean = 4.0%; Geometric mean = 3.9%.

Solution

The correct answer is **C**.

$$\text{Arithmetic mean} = \frac{8\% + (-2\%) + 6\%}{3} = 4\%$$

$$\text{Geometric mean} = [(1 + 8\%) \times (1 + (-2\%)) \times (1 + 6\%)]^{1/3} - 1 = 3.9\%$$

LOS 1c: Compare the money-weighted and time-weighted rates of return and evaluate the performance of portfolios based on these measures

Money-weighted Rate of Return

The money-weighted return considers the money invested and gives the investor information on the actual investment return. Calculating money-weighted return is similar to calculating an investment’s internal rate of return (IRR).

The money-weighted rate of return (MWRR) is like the portfolio's internal rate of return (IRR). It's the rate at which the present value of cash flows equals zero. In simple terms, it's a way to measure how well a portfolio is performing.

$$\sum_{t=0}^{T} \frac{CF_t}{(1+IRR)^t} = 0$$

Where:

T = Number of periods.

CF_t = Cash flow at time *t*.

IRR = Internal rate of return (or money-weighted rate of return).

The money-weighted rate of return (MWRR) looks at a fund’s starting and ending values and all the cash flows in between. In an investment portfolio, cash inflows are a part of it. These inflows could be from deposits or investments made during a certain period. The MWRR considers these inflows and calculates the overall rate of return for the portfolio:

- The beginning value.
- Dividends/interest reinvested.
- Contributions made.

Cash outflows, on the other hand, refer to:

- Withdrawals made.
- Dividends or interest received.
- The final value of the fund.

Example 1: Calculating Money-weighted Rate of Return

An investor makes the following investments in a portfolio over a two-year period:

- At the beginning of year one, the investor invests $10,000.
- At the end of the first year, after the portfolio's value increases to $12,000, the investor adds $5,000, making the total portfolio value $17,000.
- At the end of the second year, the portfolio value further increases to $25,000.

The money-weighted rate of return for the investor’s portfolio is *closest* to:

Solution

We need to calculate the internal rate of return (IRR) considering the following cash flows:

- $CF_0 = -\$10,000$ (Initial investment)
- $CF_1 = -\$5,000$ (Additional investment at the end of year one)
- $CF_2 = +\$25,000$ (Final portfolio value at the end of year two)

To find the money-weighted rate of return, solve the equation for IRR:

$$\frac{CF_0}{(1+IRR)^0} + \frac{CF_1}{(1+IRR)^1} + \frac{CF_2}{(1+IRR)^2} = \frac{-10,000}{1} + \frac{-5,000}{(1+IRR)} + \frac{25,000}{(1+IRR)^2} = 0$$

Using BA II Plus Calculator, $IRR \approx 35.08\%$.

Example: Calculating Money-weighted Return for a Dividend-paying Stock

Calvin Hair purchased a share of Superior Car Rental Company for $85 at the beginning of the first year. He bought an additional unit for $87 at the end of the first year. At the end of the second year, he sold both shares at $90. During both years, Hair received a dividend of $4 per share, which was not reinvested.

Calculate the money-weighted return.

Solution

To calculate the money-weighted return in this example, we need to consider the timing and amounts of cash flows and their respective investment periods.

Step 1: Calculate the total investment at the beginning (t=0):

$$\text{Initial investment} = -\$85$$

Step 2: Calculate the total investment at $t = 1$:

$$\begin{aligned}\text{Initial investment + Additional investment} &= \$87 - \$4(\text{ Dividend received at}\\ &\quad \text{the end of the first year, which is not reinvested})\\ &= -\$83\end{aligned}$$

Step 3: Calculate the final portfolio value at $t = 2$:

$$\begin{aligned}\text{Number of shares sold} \times \text{Selling price} &= 2 \text{ shares} \times \$90 = \$180 + 8(\text{ Dividend received for}\\ &\quad \text{both shares})\\ &= \$188\end{aligned}$$

As such, we have:

$CF_0 = -85$.

$CF_1 = -83$.

$CF_2 = 188$.

Using the BA II Plus calculator, you will get IRR = 7.71%, which is equivalent to the money-weighted rate of return.

Shortcomings of the Money-weighted Rate of Return

The money-weighted rate of return (MWRR) considers all cash flows, such as withdrawals or contributions. If an investment spans multiple periods, MWRR gives more importance to the fund's performance when the account is at its largest. This can be a problem for fund managers because it might make their performance seem worse due to factors they can't control.

Time-Weighted Rate of Return

The time-weighted rate of return (TWRR) calculates the compound growth of an investment. Unlike the money-weighted rate, it doesn't care about withdrawals or contributions. TWRR is like finding the average return of different time periods within your investment.

Steps of Calculating Time-weighted Rate of Return

Step 1: Value the portfolio immediately before any significant cash inflow or outflow of funds. Divide the evaluation period into subperiods based on dates of significant additions or withdrawals of funds.

Step 2: Compute the holding period return on the portfolio for each period.

Step 3: Compound or link the holding period returns to the annual rate of return, which is the time-weighted rate of return.

$$TWRR = (1 + HPR_1 \times (1 + HPR_2) \times (1 + HPR_3) \ldots \times (1 + HPR_{n-1}) \times (1 + HPR_n) - 1$$

If the evaluation period is more than one year, compute the geometric mean of the annual returns to get the time-weighted return for the investment period.

$$\begin{aligned} \bar{R}_{Gi} &= \sqrt[n]{(1 + HPR_1) \times (1 + HPR_1) \ldots \times (1 + HPR_n)} - 1 \\ &= [(1 + HPR_1) \times (1 + HPR_1) \ldots \times (1 + HPR_n)]^{\frac{1}{n}} - 1 \end{aligned}$$

Example: Calculating the Time-Weighted Rate of Return (Period More than one year)

An investor purchases a share of stock at t = 0 for $200. At the end of the year (at t = 1), the investor purchases an additional share of the same stock, this time for $220. She then sells both shares at the end of the second year for $230 each. She also received annual dividends of $3 per share at the end of each year. Calculate the annual time-weighted rate of return on her investment.

Solution

First, we break down the two years into two one-year periods.

Holding period 1:

Beginning value = 200.

Dividends paid = 3.

Ending value = 220.

Holding period 2:

Beginning value = 440 (2 shares × 220)

Dividends paid = 6 (2 shares × 3)

Ending value = 460 (2 shares × 230)

Secondly, we calculate the HPR for each period:

$$HPR_1 = \frac{(220 - 200 + 3)}{200} = 11.5\%$$
$$HPR_2 = \frac{(460 - 440 + 6)}{440} = 5.9\%$$

Lastly, we need to find the geometric mean of the HPRs since we are dealing with a period of more than a year.

$$TWRR = [(1 + HPR_1) \times (1 + HPR_1) \ldots \times (1 + HPR_n)]^{\frac{1}{n}} - 1$$
$$= (1.115 \times 1.059)^{0.5} - 1 = 8.7\%$$

Example: Calculating the Time-weighted Rate of Return (Period Less

than One Year)

The beginning value of a portfolio as of January 1, 2020, was $1,000,000. On February 10, the portfolio's value was $1,100,000, including an additional contribution of the $50,000 injected into the portfolio on this date. The portfolio's ending value at the beginning of April was $1,350,000.

The time-weighted rate of return is *closest to:*

Solution

The time-weighted return is calculated as follows:

$$\begin{aligned} HPR_1 &= \frac{V_1 - V_0}{V_0} = \frac{(1,100,000 - 50,000) - 1,000,000}{1,000,000} = 5\% \\ HPR_2 &= \frac{V_2 - V_1}{V_1} = \frac{1,350,000 - 1,100,000}{1,100,000} = 22.73\% \\ \Rightarrow TWRR &= (1 + HPR_1) \times (1 + HPR_1) - 1 \\ &= 1.05 \times 1.2273 - 1 = 28.87\% \end{aligned}$$

Question

A chartered analyst buys a share of stock at time t = 0 for $50. At t = 1, he purchases an extra share of the same stock for $53. The share gives a dividend of $0.50 per share for the first year and $0.60 per share for the second year. He sells the shares at the end of the second year for $55 per share. Calculate the annual time-weighted rate of return.

A. 5.90%.

B.12.24%.

C. 7.00%.

The correct answer is **A.**

We have two one-year holding periods:

$$\begin{array}{cc} HP_1 & HP_2 \\ P_0 = 50 & P_0 = 106 \\ D = 0.5 & D = 1.2 \\ P_1 = 53 & P_1 = 110 \end{array}$$

We now calculate the holding period returns:

$$HPR_1 = \frac{(53 - 50 + 0.5)}{50} = 7\%$$
$$HPR_2 = \frac{(110 - 106 + 1.2)}{106} = 4.9\%$$
$$\Rightarrow TWRR = 1.07 \times 1.049 - 1 = 12.24\%$$

Therefore,

$$\text{Annual TWRR} = (1 + 0.1224)^{0.5} - 1 = 5.9\%$$

LOS 1d: Calculate and interpret annualized return measures and continuously compounded returns and describe their appropriate uses

To compare returns over different timeframes, we need to annualize them. This means converting daily, weekly, monthly, or quarterly returns into annual figures.

Non-Annual Compounding

Interest may be paid semiannually, quarterly, monthly, or even daily – interest payments can be made more than once a year. Consequently, the present value formula can be expressed as follows when there are multiple compounding periods in a year:

$$PV = FV_N \left(1 + \frac{R_S}{m}\right)^{-mN}$$

Where:

m = Number of compounding periods in a year.

R_s = Quoted annual interest rate.

N = Number of years.

Example: Calculating the Present Value of a Lump Sum (More than One Compounding Period)

Jane Doe wants to invest money today and have it become $500,000 in five years. The annual interest rate is 8%, and it's compounded quarterly. How much should Jane invest right now?

Using the formula above:

$FV_N = \$500,000$.

$R_S = 8\%$.

$m = 4$.

$R_S/m = \frac{8\%}{4} = 2\% = 0.02$.

$N = 5$.

$mN = 4 \times 5 = 20$.

Therefore,

$$PV = FV_N (1 + \frac{R_S}{m})^{-mN} = \$500,000 \times (1.02)^{-20} = \$336,485.67$$

Using BA II *Plus* Calculator:

- Press the [2nd] button, then the [FV] button to clear the financial registers. The display should show "CLR TVM."
- Enter the future value (FV). This is the amount Jane wants to have in five years, which is $500,000. To do this, type "500000" and press the [FV] button.
- Enter the interest rate (I/Y). This is the annual interest rate, which is 8%. However, since interest is compounded quarterly, we need to divide this by 4. To do this, type "8", press the [÷] button, type "4", then press the [ENTER] button, and finally press the [I/Y] button.
- Enter the number of periods (N). This is the number of quarters in five years, which is 5*4 = 20. To do this, type "20" and press the [N] button.
- Compute the present value (PV). To do this, press the [CPT] and then the [PV] buttons. The display should show the amount Jane needs to invest today, approximately $336,485.49.

Annualized Returns

To annualize a return for a period shorter than a year, you need to account for how many times that period fits into a year. For example, if you have a weekly return, you would compound it 52 times because there are 52 weeks in a year.

Generally, we can annualize the returns using the following formula:

$$\text{Return}_{\text{annual}} = (1 + \text{Return}_{\text{period}})^{c} - 1$$

Where:

$\text{Return}_{\text{period}}$ = Quoted return for the period.

c = Number of periods in a year.

Example: Annualizing Returns

If the monthly return is 0.7%, then the compound annual return is:

$$\begin{aligned}\text{Return}_{\text{annual}} &= (1 + \text{Return}_{\text{monthly}})^{12} - 1 \\ &= (1.007)^{12} - 1 = 0.0873 = 8.73\%\end{aligned}$$

For a period of more than one year, for example, a 15-month return of 16% can be annualized as:

$$\begin{aligned}\text{Return}_{\text{annual}} &= (1 + \text{Return}_{15\text{ month}})^{\frac{12}{15}} - 1 \\ &= (1.16)^{\frac{4}{5}} - 1 = 12.61\%\end{aligned}$$

We may apply the same procedure to convert weekly returns to annual returns for comparison with weekly returns.

$$\text{Return}_{\text{annual}} = (1 + \text{Return}_{\text{weekly}})^{52} - 1$$

For comparison with weekly returns, we can convert annual returns to weekly returns by making $(\text{Return}_{\text{weekly}})^{52}$ the subject of the formula.

Example: Comparing Investments by Annualizing Returns

An investor is evaluating the returns of two recently formed bonds. Selected return information on the bonds is presented below:

Bond	Time Since Issuance	Return Since Issuance (%)
A	120 days	2.50
B	8 months	6.00

Annualized Return Calculation

To compare the annualized rate of return for both bonds, you can use the formula for annualizing returns based on different time periods:

Annualized Return = $(1 + \frac{\text{Return Since Issuance}}{100})^{\frac{365}{\text{Time Since Issuance}}} - 1$

Let's calculate the annualized returns for both bonds:

For Bond A:

Time Since Issuance = 120 days

Return Since Issuance = 2.50%

Annualized Return for Bond A = $(1 + \frac{2.50}{100})^{\frac{365}{120}} - 1$

Annualized Return for Bond A = $(1 + 0.025)^{3.0417} - 1$

Annualized Return for Bond A = 1.079847 - 1 = 0.079847 or 7.98%

For Bond B:

Time Since Issuance = 8 months = 240 days

Return Since Issuance = 6.00%

Annualized Return for Bond B = $(1 + \frac{6.00}{100})^{\frac{365}{240}} - 1$

Annualized Return for Bond B = $(1 + 0.06)^{1.5208} - 1$

Annualized Return for Bond B = 1.092751 - 1 = 0.092751 or 9.28%

Comparing the annualized returns:

Bond A has an annualized return of approximately 7.98%.

Bond B has an annualized return of approximately 9.28%.

Therefore, Bond B has a higher annualized rate of return compared to Bond A.

Continuously Compounded Returns

The continuously compounded return is calculated by taking the natural logarithm of one plus the holding period return. For example, if the monthly return is 1.2%, you'd calculate it as ln(1.012), which equals approximately 0.01192.

Generally, continuously compounded from t to $t+1$ is given by:

$$r_{t,t+1} = \ln\left(\frac{P_{t+1}}{P_t}\right) = \ln\left(1 + R_{t,t+1}\right)$$

Assume now that the investment horizon is from time $t = 0$ to time $t = T$ then the continuously compounded return is given by:

$$r_{0,T} = \ln\left(\frac{P_T}{P_0}\right)$$

If we apply the exponential function on both sides of the equation, we have the following:

$$P_T = P_0 e^{r_{0,T}}$$

Note that $\frac{P_T}{P_0}$ can be written as:

$$\frac{P_T}{P_0} = \left(\frac{P_T}{P_{T-1}}\right)\left(\frac{P_{T-1}}{P_{T-2}}\right)\ldots\left(\frac{P_1}{P_0}\right)$$

If we take natural logarithm on both sides of the above equation:

$$\ln\left(\frac{P_T}{P_0}\right) = \ln\left(\frac{P_T}{P_{T-1}}\right) + \ln\left(\frac{P_{T-1}}{P_{T-2}}\right) + \ldots + \ln\left(\frac{P_1}{P_0}\right)$$
$$\Rightarrow r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + \ldots + r_{0,1}$$

Therefore, the continuously compounded return to time T is equivalent to the sum of one-period continuously compounded returns.

Question

The weekly return of an investment that produces an annual compounded return of 23% is *closest to*:

A. 0.40%.

B. 0.92%.

C. 0.41%.

The correct answer is A.

Recall that:

$$\text{Return}_{\text{annual}} = (1 + \text{Return}_{\text{weekly}})^{52} - 1$$

We can rewrite the above equation as follows:

$$\begin{aligned} \text{Return}_{\text{weekly}} &= (1 + \text{Return}_{\text{annual}})^{\frac{1}{52}} - 1 \\ &= (1 + 0.23)^{\frac{1}{52}} - 1 \\ &\approx 0.40\% \end{aligned}$$

LOS 1e: calculate and interpret major return measures and describe their appropriate uses

Other Return Measures

Gross and Net Return

The gross return is what an asset manager earns before subtracting various costs such as management fees, custody fees, taxes, and other administrative expenses. However, it does account for trading costs such as commissions.

Gross return does not consider management or administrative costs. For this reason, it is a suitable metric for assessing and comparing the investment expertise of asset managers.

Net return is a metric for how much an investment has earned for the investor. It considers all administrative and management costs that reduce an investor's return.

Pre-tax and After-tax Nominal Return

Unless otherwise stated, all returns are nominal pre-tax returns in general. Depending on the jurisdiction, different rates apply to capital gains and income. Long-term and short-term taxes may also be applied to capital gains.

The after-tax nominal return is determined by subtracting any tax deductions applied to dividends, interest, and realized gains from the total return.

Real Returns

Returns are typically presented in nominal terms, which consist of three components: the real risk-free return as compensation for postponing consumption, inflation as compensation for the loss of purchasing power, and a risk premium. Real returns are useful in comparing returns over different periods, given that inflation rates vary over time.

Recall the relationship between the nominal rate and the real rate:

$$(1+\text{ Nominal Risk-free rate})=(1+\text{Real risk free rate})(1+\text{Inflation premium})$$

We can find the connection between nominal and real returns by considering the real risk-free rate of return and the inflation premium. This relationship can be expressed as:

$$(1+\text{Real Return})=\frac{(1+\text{Real risk-free rate})(1+\text{Risk premium})}{1+\text{Inflation premium}}$$

Real returns become particularly useful when you want to compare returns across various time periods and different countries. This is especially important when returns are shown in local currencies and when inflation rates vary from one country to another.

After-tax real return is the amount the investor receives as payment for delaying consumption and taking on risk after paying taxes on investment.

Leveraged Returns

If an investor uses derivative instruments within a portfolio or borrows money to invest, then leverage is introduced into the portfolio. The leverage amplifies the returns on the investor's capital, both upwards and downwards.

The leveraged return considers the actual return on the investment and the cost of the borrowed money. The cost of borrowing and financing fees are subtracted from the overall return produced by the investment to determine the leveraged return.

Using the borrowed capital (debt) increases the size of the leveraged position by the additional borrowed capital.

Intuitively, the leveraged return is given by:

$$\begin{aligned} R_L &= \frac{\text{Portfolio return}}{\text{Portfolio equity}} \\ &= \frac{[R_P \times (V_E + V_B) - (V_B \times r_D)]}{V_E} \\ &= R_P + \frac{V_B}{V_E}(R_P - r_D) \end{aligned}$$

Where:

R_L = Return earned on the leveraged portfolio.

R_P = Total investment return earned on the leveraged portfolio.

V_B = Value of debt in the portfolio.

V_E = Value of equity in the portfolio.

r_D = Borrowing cost on debt.

Example: Calculating Leveraged Return

For a $250,000 equity portfolio with an annual 9% total investment return, 40% financed by debt at 6%, the leveraged return would be:

$$R_L = R_P + \frac{V_B}{V_E}(R_P - r_D) = 9\% + \frac{\$100,000}{\$150,000}(9\% - 6\%) = 11\%$$

Question

A $7,500,000 equity portfolio is 35% financed by debt at a cost of 5% per annum. If the equity portfolio generates a 9% annual total investment return, the leverage return is *closest* to:

A. 11.15%.

B. 14.00%.

C. 8. 25%.

The correct answer is **A**.

$$R_L = R_P + \frac{V_B}{V_E}(R_P - r_D)$$
$$= 9\% + \frac{\$2,625,000}{\$4,875,000}(9\% - 5\%) = 11.15\%$$

Learning Module 2: The Time Value of Money in Finance

LOS 2a: calculate and interpret the present value (PV) of fixed-income and equity instruments based on expected future cash flows

The time value of money (TVM) is a fundamental financial concept. It emphasizes that a sum of money is worth more in the present than in the future. There are three key reasons supporting this principle:

- **The concept of opportunity** cost suggests that money available today can be invested and generate interest, increasing its value over time. By delaying the use of money, one forgoes potential investment opportunities and the growth they offer.
- **Inflation** poses a threat to the purchasing power of money in the future. Due to inflation, the same amount of money may buy fewer goods or services in the future compared to its present value. Consequently, having money now is advantageous since its purchasing power diminishes as time progresses.
- There is an element of **uncertainty regarding future cash flows**. Unexpected events or circumstances may prevent the receipt of money as planned, rendering it less reliable. Until the money is obtained, there is a level of uncertainty attached to its availability and utility.

Time value of money calculations allow us to establish the future value of a given amount of money.

Key Components of Time Value of Money

- **Discount rate or interest rate**: The rate of discounting or compounding that you apply to an amount of money to calculate its present or future value.
- **Time periods**: The whole number of time periods over which the present or future value of a sum is being calculated. These periods can be annually, semi-annually, quarterly, monthly, weekly, etc.

- **Present value (PV)**: The amount of money you have today (or at time T = 0) is referred to as the present value.
- **Future value (FV):** The accumulated amount of money you get after investing the original sum at a specific interest rate and for a given time period, say, two years.

Fundamental Formulas in Time Value of Money Calculations:

Let,

FV = Future value.

PV = Present value.

r = Stated discount rate per period.

N = Number of periods (Years).

Then the future value (FV) of an investment is given by:

$$FV_{=}PV(1+r)^{N}$$

If N is large such that $N \rightarrow \infty$ the initial cashflow is compounded continuously:

$$FV = PVe^{rN}$$

To find the present value of the investment, we rewrite the above formula so that:

$$PV = FV(1+r)^{-N}$$

And for the continuous compounding, we have,

$$PV = FV_{t}e^{-rN}$$

Example: Calculating the Present Value of continuously Compounded Cashflows

A fund continuously accumulates to $4,000 over ten years at a 10% annual interest rate.

Calculate the closest present value of this fund.

Solution

From the question, FV=4,000, r_s=10%, N=10

So,

$$PV = FVe^{-Nrs} = \$4,000 \times e^{-10\times0.1} = \$1,471.5178$$

Frequency of Compounding

When the frequency of compounding is more than once per year (quarterly, monthly, etc.), the formulas are analogously defined as:

$$FV_N = PV(1 + \frac{r_s}{m})^{mN}$$

Where:

m = Number of compounding periods per year.

N = Number of years.

r_s = Annual stated rate of interest.

Intuitively, the formula for the PV is given by:

$$PV = FV(1 + \frac{r_s}{m})^{-mN}$$

In the following discussion, we shall let $t = mN$ denote the number of compounding periods and $\frac{rs}{m} = r$ denote the stated discount rate per period.

Calculation using a Financial Calculator

For calculating FV and PV using the BA II Plus™ Financial Calculator, use the following keys:

N = Number of compounding periods.

I/Y = Rate per period.

PV = Present value.

FV = Future value.

PMT = Payment.

CPT = Compute.

It is important to note that the sign of PV and FV will be opposite. For example, if PV is negative, then FV will be positive. Generally, an inflow is entered with a positive sign, while an outflow is entered as a negative sign in the calculator.

Time Value of Money in Fixed-income Instruments

Fixed-income instruments are debt securities where an issuer borrows money from an investor (lender) in exchange for a promised future payment. Examples of fixed-income instruments are bonds, loans, and notes.

The market discount rate for fixed-income instruments is also known as yield-to-maturity (YTM). It's the interest rate investors require to invest in a specific fixed-income instrument.

Cash Flow Patterns Associated with Fixed-Income Instruments

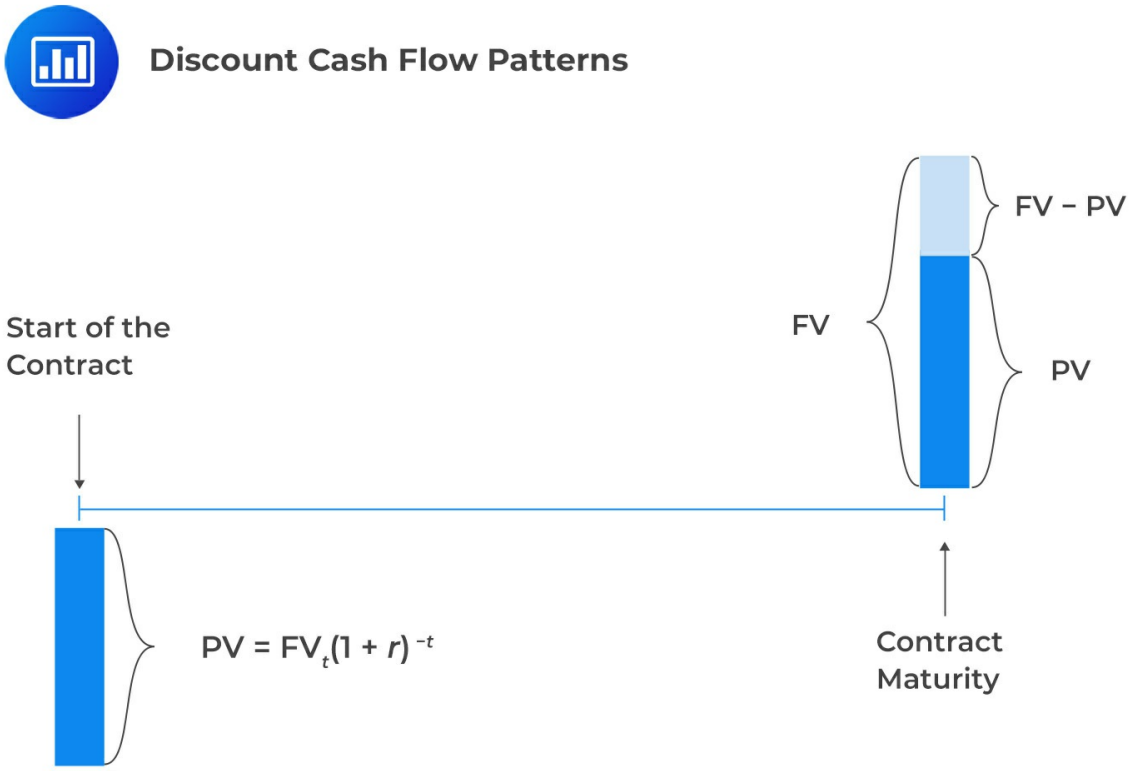
The cash flows in fixed-income instruments occur in three general patterns: Discount, periodic interest, and level payments.

Discount Cash Flow Patterns

For discount cashflow patterns, an investor pays an initial discounted price (PV) for the instrument (such as a bond or a loan) and gets one payment (FV) at the end maturity. The investor's return is the interest earned, that is, the difference between the initial price and principle (FV − PV).

The discount bonds are also called zero-coupon bonds – they do not have periodic interest

payments.



The price of a discount bond can be calculated using the formula for the present value (PV) of a single cash flow, which is as follows:

$$PV = FV_t(1 + r)^{-t}$$

Where:

FV = Future value.

PV = Present value.

r = Stated discount rate per period.

t = Number of compounding periods.

Example: Calculating the Future Value of a Zero-Coupon Bond

Assume Chad invests $8,000 in a zero-coupon bond that yields 8% annually and matures in four years. The maturity value of this bond is *closest to*:

Solution

Recall that:

$$FV = PV(1 + r)^t$$

In this case, we have PV=8,000, r=8%, t=4 so that:

$$FV = 8,000(1 + 8\%)^4 = 10,883.91$$

Using the BA II Plus™ Financial Calculator

Steps	Explanation	Display
[2nd][QUIT]	Return to standard calc Mode	0
[2nd][CLR TVM]	Clears TVM Worksheet	0
2[N]	Years/periods	N = 4
10[1/Y]	Set interest rate	PV = −8,000
0[PMT]	Set payment	PMT = 0
[CPT][FV]	Compute future value	FV = 10,883.91

Note that zero-coupon bonds can be issued at negative interest rates. In this case, the price (PV) of the bond is higher than the face value (FV).

Example: Calculating the Price of a Discount Bond Issued at Negative Interest Rates

In January 2018, the Swiss government issued 15-year sovereign bonds at a negative yield of -0.08%. The present value (PV) of the bond per CHF100 of principal (FV) at the time of issuance is *closet to*:

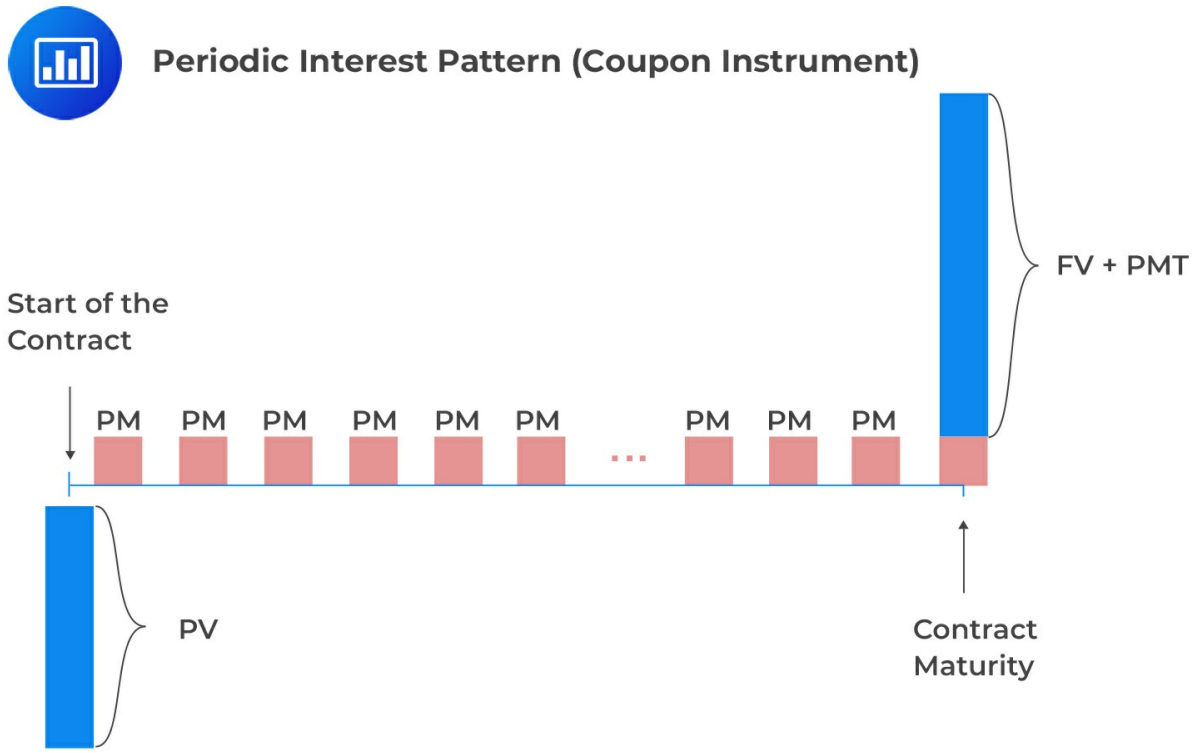
Solution

Recall for a zero-coupon bond,

$$\begin{aligned} PV &= FV_t(1 + r)^{-t} \\ &= 100(1 - 0.0008)^{-15} = 101.21 \end{aligned}$$

Periodic Interest Pattern (Coupon Instrument)

A coupon instrument is a fixed-income investment. It includes periodic cash flows called coupons and repays the principal at maturity. People often use these in coupon bond investments. These instruments have a set schedule with regular, equal payments.



The pricing of a coupon bond involves calculating its present value (PV) based on the market discount rate. The general formula for calculating the bond's price is derived from the discounted cash flow model. It considers the coupon payments (PMTs) and the final principal payment (FV) at maturity. The bond's price is determined by discounting each cash flow using the market discount rate (r).

The formula used to calculate the present value (PV) of a coupon bond is as follows:

$$\text{PV(Coupon Bond)} = \frac{\text{PMT}}{(1+\text{r})^1} + \frac{\text{PMT}}{(1+\text{r})^2} + \ldots + \frac{(\text{PMT}_\text{N} + \text{FV}_\text{N})}{(1+\text{r})^\text{N}}$$

Where:

PMT = Coupon payment.

FV = Future value.

r = Market discount rate (YTM).

N = Number of periods.

Example 1: Pricing a Coupon Bond on an Annual Basis

Suppose we have a 5-year bond with a face value of $1,000 and an annual coupon rate of 5%. The market discount rate is 6%. The bond's price is *closest to*:

Solution

$$\text{PV(Coupon Bond)} = \frac{\text{PMT}}{(1+\text{r})^1} + \frac{\text{PMT}}{(1+\text{r})^2} + \cdots + \frac{(\text{PMT}_\text{N} + \text{FV}_\text{N})}{(1+\text{r})^\text{N}}$$

In this case, we have PMT=5% of $1,000=$50, r=6%, N=5 years, FV=$1,000 so that:

$$\begin{aligned}\text{PV} &= \frac{\$50}{(1+0.06)^1} + \frac{\$50}{(1+0.06)^2} + \frac{\$50}{(1+0.06)^3} + \frac{\$50}{(1+0.06)^4} \\ &\quad + \frac{(\$50+\$1,000)}{(1+0.06)^5} \\ \text{PV} &= \$47.17 + \$44.50 + \$41.98 + \$39.60 + \$784.62 = \$957.88\end{aligned}$$

Therefore, the price of the bond would be $957.88

You could use a BA II Plus Calculator to solve the above question:

Steps	Explanation	Display
[2nd][QUIT]	Return to standard calc Mode	0
[2nd][CLR TVM]	Clears TVM Worksheet	0
5[N]	Years/periods	N = 5
6[1/Y]	Set interest rate	I/Y = 6.00
50[PMT]	Set payment	PMT = 50.00
1000[FV]	Set the face value	FV = 1000.00
[CPT][PV]	Compute the present value	PV = −957.88

Example 2: Pricing a Coupon Bond With a Single Cash Flow on a semi-annual Basis

Suppose an investor has a 2-year bond with a face value of $1000 and an annual coupon rate of 6%, paid semi-annually. The market discount rate is 5%. The price of the bond is *closest to*:

Solution

Recall that:

$$\text{PV(Coupon Bond)} = \frac{\text{PMT}}{(1+\text{r})^1} + \frac{\text{PMT}}{(1+\text{r})^2} + \ldots + \frac{(\text{PMT}_\text{N} + \text{FV}_\text{N})}{(1+\text{r})^\text{N}}$$

Where:

PMT = Coupon payments $(\$1,000 \times \frac{6\%}{2}) = \$30$ *in this case.*

FV = Future value ($1,000 *in this case*).

r = Market discount rate (YTM), $(\frac{(5\%)}{2} = 2.5\%)$ *per period in this case.*

N = Number of periods (4 *periods in this case*).

Plugging these values into the formula, we get:

$$\text{PV} = \frac{\$30}{(1.025)^1} + \frac{\$30}{(1.025)^2} + \frac{\$30}{(1.025)^3} + \frac{(\$30 + \$1,000)}{(1.025)^4}$$
$$\text{PV} = \$29.27 + \$28.55 + 27.85 + \$933.13 = \$1,018.81$$

Therefore, the bond's price is $\$1,018.81$

You can easily use the BA II Plus calculator (or any other allowed financial calculator) to solve the above question.

Steps	Explanation	Display
[2nd][QUIT]	Return to standard calc Mode	0
[2nd][CLR TVM]	Clears TVM Worksheet	0
4[N]	Years/periods	N = 4
2.5[1/Y]	Set interest rate	I/Y = 2.50
30[PMT]	Set payment	PMT = 30.00
1000[FV]	Set the face value	FV = 1000.00
[CPT][PV]	Compute the present value	$\text{PV} = -1,018.81$

Perpetual Bonds

Perpetual bonds are rare types of coupon bonds that do not have a stated date of maturity. They are generally issued by firms seeking equity-like financing and usually include redemption provisions.

The formula present value of perpetual bonds is obtained as follows: As $N \to \infty$, the formula for calculating PV of coupon changes as follows:

$$\begin{aligned}&\text{PV (perpetual bond)}\\&= \lim_{(N \to \infty)} \left[\frac{PMT}{(1+r)^1} + \frac{PMT}{(1+r)^2} + \cdots + \frac{(PMT_N + FV_N)}{(1+r)^N}\right]\\&= \frac{PMT}{r}\end{aligned}$$

So, the present value of a perpetuity is given by:

$$PV = \frac{PMT}{r}$$

Example: Perpetual Bond

In 2021, XYZ Financial (the holding company for XYZ Bank) issued $500 million in perpetual bonds with a 4.00 percent semi-annual coupon. Calculate the bond's yield to maturity (YTM) if the market price was $98.50 (per $100).

Solution

Recall,

$$PV = \frac{PMT}{r}$$

Hence,

$$r = \frac{PMT}{PV}$$

To solve this problem, we first need to calculate the semi-annual coupon payment, which is,

$$\text{PMT(semi-annual coupon payment)} = \frac{\$100 \times 4\%}{2} = \$2, \text{PV} = \$98.50$$

Therefore,

$$r = \frac{\$2}{\$98.50} = 0.0203 = 2.03\%$$

The annualized yield-to-maturity is:

$$r = 0.0203 \times 2 \approx 4.06\%$$

Level Payments (Annuity Instruments) Patterns

An annuity is a finite series of cash flows, all with the same value. **A fixed-income instrument** with annuity payments provides a stream of periodic equal cash inflows over a finite period.

The level payments consist of interest and principal payments. Fixed income instruments with level payments include fully amortizing loans such as mortgages.

There are two types of annuities: ordinary annuities and annuities due. Annuity due is a type of annuity where payments start immediately at the beginning of time, at time $t = 0$. In other words, payments are made at the beginning of each period.

On the other hand, an ordinary annuity is an annuity where the cashflows occur at the end of each period. Such payments are said to be made in arrears (beginning at time t = 1). We shall consider ordinary annuity in this section.

Ordinary Annuity

Remember that in an ordinary annuity, the series of payments does not begin immediately. Instead, payments are made at the end of each period. It is further worth noting that the present value of an annuity is equal to the sum of the current value of each annuity payment:

$$\begin{aligned} PV &= A(1+r)^{-1} + A(1+r)^{-2} + \ldots + A(1+r)^{-N-1} + A(1+r)^{-N} \\ &= A(1+r)^{-1} + (1+r)^{-2} + \ldots + (1+r)^{-(N-1)} + (1+r)^{-N} \\ PV &= A\frac{1-(1+r)^{N}}{r} \end{aligned}$$

Where:

A = Periodic cash flow.

r = Market interest rate per.

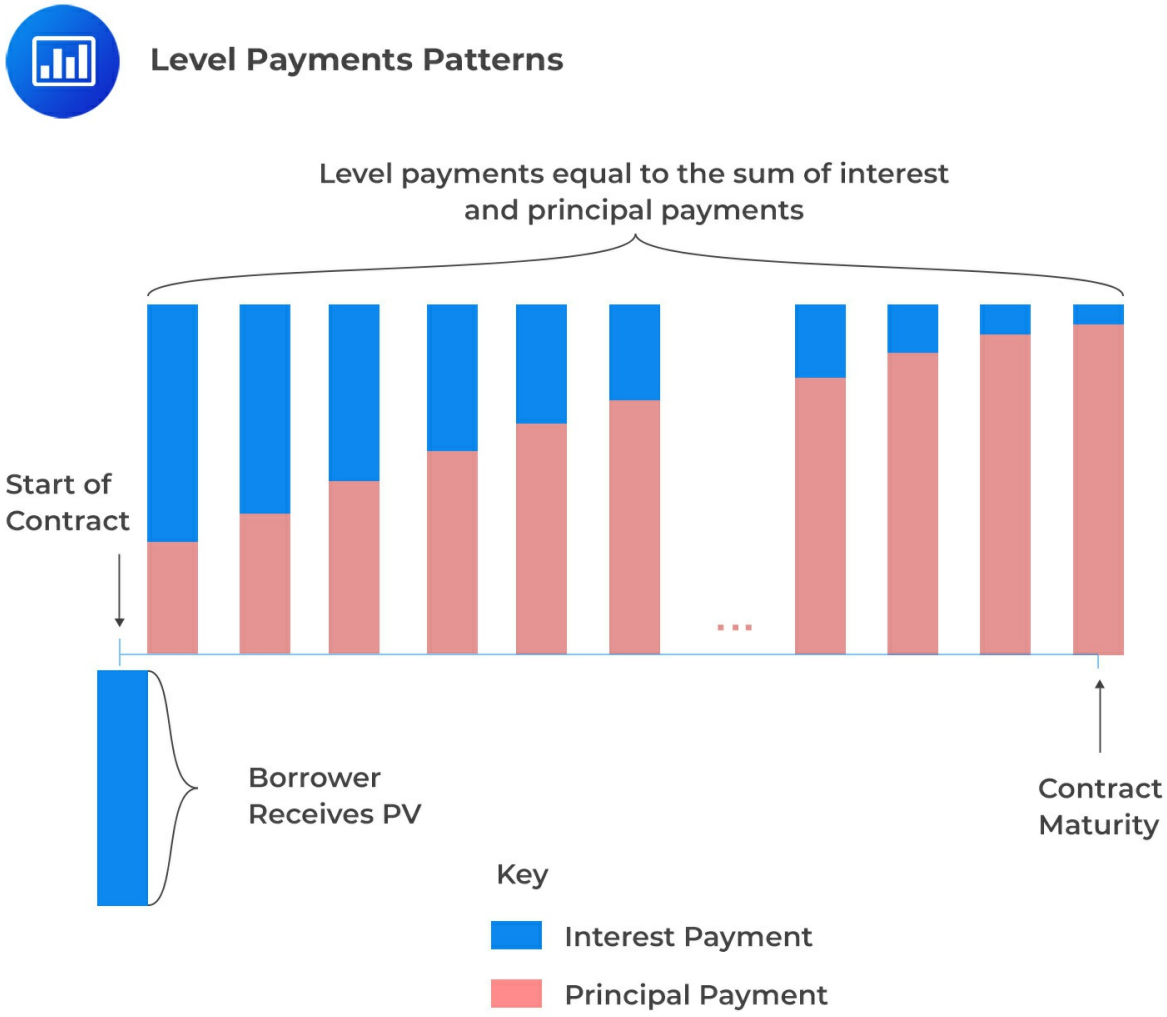
PV = Present value/ Principal Amount of the loan or bond.

N = Number of payment periods.

The periodic annuity is calculated as follows:

$$A = \frac{r(PV)}{(1-(1+r)^{-N}}$$

Consider a fully amortizing mortgage loan. In this case, the borrower receives the mortgage loan now and promises to make periodic payments equal to the sum of interest and principal payments.



Note that the periodic mortgage payment is constant, but the proportion of the interest payment decreases while the principal payment increases.

The cash flow pattern of a fully amortizing mortgage follows the pattern of an ordinary annuity with a series of equal cash flows. As such, the periodic annuity (periodic payment) of a fully amortizing mortgage is given by:

$$A = \frac{r(PV)}{1-(1+r)^{-t}}$$

Where:

A = Periodic cash flow.

r = Market interest rate per period.

PV = Present value or principal amount of loan or bond.

t = Number of payment periods.

Example: Calculating the Periodic Payment of a Mortgage

Jake is looking to secure a fixed-rate 25-year mortgage to finance 75% of the value of an $800,000 residential property. If the annual interest rate on the mortgage is 4.5%, Jake's monthly mortgage payment is *close to*:

Solution

Remember,

$$A = \frac{r(PV)}{1-(1+r)^{-t}}$$

Where:

A = Periodic cash flow.

r = Market interest rate per period.

PV = Present value/ Principal Amount of the loan or bond.

t = Number of payment periods.

In this case, we have:

- $r = 0.375\% \left(= \frac{4.5\%}{12}\right)$
- $N = 300 \text{ months} (= 25 \text{ years} \times 12 \text{ months/year})$
- $PV = \$600,000 (= 75\% \times \$800,000)$

Plugging these values into the formula, we get:

$$A = \frac{0.00375 \times \$600,000}{1-(1+0.00375)^{-300}} \overline{=} \$3,334.995 \approx \$3,335$$

Using a BA II Plus financial calculator:

Steps	Explanation	Display
[2nd][QUIT]	Return to standard calc Mode	0
[2nd][CLR TVM]	Clears TVM Worksheet	0
300[N]	Years/periods	N = 300
0.375[1/Y]	Set interest rate	I/Y = 0.375
−600,000[PV]	Set the present value of the mortgage	PV = −600,000.00
0[FV]	Set the face value	FV = 0.00
[CPT][PMT]	Compute the periodic payment	PMT = 3,334.99

Time Value of Money in Equity Instruments

Equity investments, such as stocks, enable an investor to acquire a fractional share/ownership by the issuing company. This gives investors the right to receive a share of the company's available cash flows as dividends.

In the context of equity instruments, the time value of money (TVM) is used to discount expected future cash flows to determine their present value. This allows investors to value the company shares.

The present value of expected future cash flows is calculated using a discount rate, r, which represents the expected rate of return on the investment.

Common Approaches for Valuing Equity Instruments

Valuing equity investments depends on the dividends cashflows which can take one of three forms: constant dividends, constant dividend growth rate, and changing dividend growth rate.

1. Valuing Equity Instruments based on Constant Dividend: The Constant Dividends model values stocks based on the assumption that dividends will remain constant over time. The

preferred or common share dividend cash flows are in the form of an infinite series that is valued like perpetuity. The formula for the constant dividends model is as follows:

$$PV_t = \sum_{i=1}^{\infty} \frac{D_t}{(1+r)^i} = \frac{D_t}{r}$$

Where: PV_t = Present value at time t

D_t = Dividend payment at time t

r = Discount rate.

Example: Valuing Equity Instruments based on Constant Dividend

Assuming we have a preferred stock with a dividend payment of $5 per year. The discount rate is 8%. The present value of the stock is *closest to*:

Solution

Recall,

$$PV_t = \frac{D_t}{r}$$

In this case, D=$5, r=8%, PV=?

So,

$$PV = \frac{5}{0.08} = \$62.5$$

This means that the present value of the stock is $62.5.

2. Valuing Equity Instruments Based on Constant Dividend Growth Rate The constant dividend growth model is a method used to estimate the value of a stock based on its future dividends. This model assumes that dividends will grow at a constant rate (g) forever. To derive

the formula for this model, we start by considering that the **present value of a stock is equal to the sum of its future dividends**, discounted by the required rate of return r. If dividends are assumed to grow at a constant rate, then each future dividend can be calculated by multiplying the previous dividend by (1 + g).

Let D_t represent the expected dividend in the next period. The present value of the stock can then be expressed as:

$$\begin{aligned} PV_t &= \frac{D_t}{(1+r)} + \frac{D_t(1+g)}{(1+r)^2} + \frac{D_t(1+g)^2}{(1+r)^3} + \dots \\ &= \sum_{i=1}^{\infty} \frac{D_t(1+g)}{(1+r)^i} \end{aligned}$$

This is an infinite geometric series with a common ratio of $\frac{(1+r)}{(1+g)}$. Using the formula for the sum of an infinite geometric series, we can simplify this equation to:

$$PV_t = \frac{D_t(1+g)}{r-g} = \frac{D_{t+1}}{r-g}$$

Where:

PV_t = Present value at time t.

D_{t+1} = Expected Dividend in the next period.

r = Required rate of return.

g = Constant growth rate.

$r - g > 0$

Therefore, this is the formula for calculating the present value of a stock using the constant dividend growth rate. This model can help estimate the value of a stock when its future dividends are expected to grow steadily.

Example: Valuing Equity Instruments based on Constant Dividend Growth Rate

Suppose a stock currently pays an annual dividend of $2.00 per share. The required rate of return for this stock is 10%, and the dividends are expected to grow at a constant rate of 5% per year indefinitely. Using the constant dividend growth model, the present value of this stock is *closest to*:

Solution

Recall that,

$$PV_t = \frac{D_t(1+g)}{r-g} = \frac{D_{t+1}}{r-g}$$

In this case, we know that D_t=$2.00, r=10%, g=5%

So,

$$PV = \frac{2 \times 1.05}{r-g} = \frac{2.10}{0.10-0.05} = \$42$$

Therefore, the present value of the stock is $42

3. Valuing Equity Instruments with Changing Dividend Growth Rates is a dynamic process. It begins with the investor buying a stock at an initial price and getting an initial dividend. The unique aspect is that the dividend is expected to grow at a rate that evolves as the company matures and shifts from high growth to slower growth. This valuation doesn't have a single formula because it relies on assumptions about future dividend growth. However, a common method is to use a multi-stage dividend discount model. This model assumes that dividends will grow at different rates during various stages of the company's growth. To find the stock's present value, you sum up the present values of dividends at each stage.
The Multi-Stage Dividend Discount Model builds on the Constant Dividend Growth Model. It accommodates a company's transition from high initial growth to lower, more stable growth.

Let's say a company has a high short-term growth rate g_s followed by a perpetual lower growth rate g_l. To find the present value (PV) of the stock at time t using this model, we compute it in

two stages:

I. **First Part**: The first part calculates the present value of dividends during the initial n periods of higher growth (g_s). This is done by discounting the dividends for each period by the required rate of return r using the following formula:

$$PV_t = \sum_{i=1}^{n} \frac{D_t(1+g_s)^i}{(1+r)^i}$$

Where:PV = Present value.n = Number of periods.D_t = Dividend at time (t).g_s = Initial higher dividend growth rate.r = Required rate of return.

II. **Second Part**: The second part calculates the present value of dividends after the initial n periods, assuming constant growth at a lower long-term rate (g_l). This can be simplified using the geometric series simplification, where $E(S_t + n)$ represents the terminal value or stock value in n periods:

$$PV_t = \frac{E(S_t + n)}{(1+r)^n}$$

Where:$E(S_t + n) = \frac{D_{t+n+1}}{(r-g_l)}$ and g_l is the lower, more stable dividend growth rate.

Example: Valuing Equity Instruments based on Changing Dividend Growth Rate

Assuming we have a stock with an expected dividend payment of $2 in one period. The discount rate is 10%. The stock is expected to have a high dividend growth rate of 20% for the first three years, followed by a slower growth rate of 5% thereafter. Calculate the present value of the stock.

Solution

First, we calculate the present value of the dividends during the high growth period:

Recall that,

$$PV_t = \sum_{i=1}^{n} \frac{D_t(1+g_s)^i}{(1+r)^i}$$

In this case, D_t=\$2, g_s=0.20, r=0.10, n=3

So,

$$PV_1 = \frac{2}{(1+0.10)^1} + \frac{2 \times (1+0.20)^1}{1+0.10)^2} + \frac{2 \times (1+0.20)^2}{1+0.10)^3}$$
$$PV_1 = 1.818 + 1.983 + 2.163 = 5.965$$
$$PV_1 = \$5.97$$

Next, we calculate the present value of the dividends during the slower growth period, assuming that dividends will grow at a constant rate of 5% thereafter:

Recall that,

$$E(S_t + n) = \frac{D_{t+n+1}}{(r - g_l)}$$

So,

$$E(S_t + n) = \frac{2(1+0.20)^3 \times (1+0.05)}{0.10 - 0.05} = \frac{3.629}{0.05} = \$72.578$$

Finally, we calculate the present value of P_4

$$PV_2 = \frac{\$72.578}{(1+0.10)^3} = \$54.527$$

The total present value of the stock is the sum of PV_1 and PV_2

$$PV_{total} = PV_1 + PV_2$$
$$= \$5.965 + 54.527$$
$$= \$60.493 \approx \$60.49$$

Question

Five years ago, Milton Inc. issued corporate bonds with a 15-year maturity. The bonds have a semi-annual coupon rate of 7.8% per annum, and the current yield to maturity is 8.5% per annum. The current price of Milton Inc's bonds (per CAD100 of par value) is *closest to*:

A. CAD91.23.
B. CAD95.35.
C. CAD96.15.

The correct answer is B.

Solution

Recall the formula for calculating the price of a bond:

$$\text{PV(Coupon Bond)} = \frac{\text{PMT}}{(1+\text{r})^1} + \frac{\text{PMT}}{(1+\text{r})^2} + \ldots + \frac{(\text{PMT}_\text{N} + \text{FV}_\text{N})}{(1+\text{r})^\text{N}}$$

First, let's get the semi-annual equivalent rates:

- The semi-annual coupon rate is $\frac{7.8\%}{2} = 3.9\%$.
- The semi-annual yield to maturity is $\frac{8.5\%}{2} = 4.25\%$.

Next, we find the number of periods remaining until the bond matures:

Since the bonds were issued 5 years ago and have a 15-year maturity, $10 (= 15 - 5)$ years remain to maturity. Since interest is paid semi-annually, this equates to $10 \times 2 = 20$ periods.

You can plug the above values into the general formula, consuming valuable time. Therefore, using BA II plus a financial calculator,

Steps	Explanation	Display
[2nd][QUIT]	Return to standard calc Mode	0
[2nd][CLR TVM]	Clears TVM Worksheet	0
20[N]	Years/periods	N = 20
4.25[1/Y]	Set interest rate	I/Y = 4.25
3.9[PMT]	Set the periodic coupon payment	PMT = 3.90
100[PV]	Set the face value of the bond	FV = 100.00
[CPT][PV]	Compute the present value	PV = −95.35

LOS 2b: Calculate and interpret the implied return of fixed-income instruments and the required return and implied growth of equity instruments given the present value (PV) and cash flows

Implied Return for Fixed-Income Instruments

The growth rate is the rate at which the market expects an asset to grow. On the other hand, implied return reflects a return based on the current price and future security cash flows.

Consider a fixed-income instrument. If we have its present value and assume all future cash flows happen as expected, the discount rate rrr, or yield-to-maturity (YTM), shows the implied return under these assumptions for the cash flow pattern.

Now, take an equity investment. If we have the present value, future value, and discount rate, we can find the implied growth rate that aligns with these values.

The implied return or growth rate provides a view of the market expectations incorporated into an asset's market price. It is useful for investors to understand these expectations when making investment decisions.

Calculating the Implied Return for Fixed-Income Instruments

Discount Bond

In the case of a discount bond or instrument, recall that an investor receives a single principal cash flow (FV) at maturity, with ($FV - PV$) representing the implied return.

To solve for the implied return earned over the life of an instrument (N periods), we can rearrange the single cash flow present value formula.

Recall that the single cash flow present value formula is:

$$PV = FV_t(1 + r)^{-t}$$

Where:

FV = Future value.

PV = Present value.

r = Stated discount rate per period.

t = Number of compounding periods.

To solve for r, we can rearrange this formula as follows:

$$r = \sqrt[t]{\frac{FV_t}{PV}} - 1 = \left(\frac{FV_t}{PV}\right)^{\frac{1}{t}} - 1$$

We use this formula to calculate the periodic return earned during the life of the instrument (t periods) based on the present value (or price) and future value of the instrument.

Example: Calculating the Implied Return for a Discount Bond

Consider a zero-coupon bond with price of $900, a future value of $1,000, and a maturity of 5 years. Calculate the implied annualized return r.

Solution

Recall that,

$$r = \left(\frac{FV_t}{PV}\right)^{\frac{1}{t}} - 1$$

In this case, $t=5$, $FV_t=\$1,000$, $PV=\$900$

So,

$$r = \left(\frac{1000}{900}\right)^{\frac{1}{5}} - 1 = 2.13\%$$

This means that an investor who purchases this zero-coupon bond at a price of $900 and holds it for five years would earn an annualized return of 2.13%.

Coupon Bonds

Recall that fixed-income instruments that pay periodic interest have cash flows throughout their life until maturity. The **yield-to-maturity (YTM)** is a single implied market discount rate for all cash flows, regardless of timing. It assumes an investor expects to receive all promised cash flows through maturity and reinvest any cash received at the same YTM.

The present value of a fixed-income instrument with periodic interest can be calculated using the following formula:

$$PV = \frac{PMT_1}{(1+r)^1} + \frac{PMT_2}{(1+r)^2} + \cdots + \frac{(PMT_N + FV_N)}{(1+r)^N}$$

Where:

PV = Present value (or price) of the instrument.

PMT = Periodic payment.

FV = Bond's principal.

N = Number of periods to maturity.

r = Discount rate (or internal rate of return) (YTM).

Example: Implied Return for Fixed-income Instruments With Periodic Interest

Consider a five-year corporate bond issued in 2023 with a 4.00 percent annual coupon and a price of USD110.00 per USD100 principal three years later. If Milka can reinvest periodic interest at the original YTM of 4.00 percent, the implied three-year return is *closest to*:

Solution

We can calculate the future value (FV) after three years, including the future price of 110.00 and all cash flows reinvested to that date:

$$\begin{aligned} FV_3 &= PMT_1(1+r)^2 + PMT_2(1+r) + PMT_3 + PV_3 \\ &= 4 \times (1.04)^2 + 4 \times (1.04) + 4 + 110.00 \\ &= \$122.49 \end{aligned}$$

We can then solve for Milka's annualized return r using the formula for implied return since we have PV=100, FV=122.49, N= 3 as follows:

$$r = \sqrt[t]{\frac{FV_t}{PV}} - 1 = \sqrt[3]{\frac{122.49}{110}} - 1 = 3.65\%$$

This means that Milka, who purchased the corporate bond at a price of 100 and held it for three years, would earn an annualized return of 3.65%.

Example: Calculating the Yield-to-Maturity of a Coupon Bond

CityGroup Corp. issued a corporate bond 7 years ago with a face value of $1,000 and a 20-year maturity. The bond pays annual interest at a coupon rate of 6%. Currently, the bond is trading at $1,120. The yield to maturity (YTM) of CityGroup Corp.'s bond is *closest to*:

Solution

Using the BA II Plus calculator, we solve the question as follows:

We have

Steps	Explanation	Display
[2nd][QUIT]	Return to standard calc Mode	0
[2nd][CLR TVM]	Clears TVM Worksheet	0
13[N]	Years/periods	N = 13
$-1,120$[PV]	Set the present value of the bond	PV = $-1,120$
60[PMT]	Set the periodic coupon payment	PMT = 3.90
$1,000$[PV]	Set the face value of the bond	FV = 1000.00
[CPT][I/Y]	Compute the YTM	I/Y = 4.74%

Therefore, the YTM is 4.74%.

Implied Return and Growth for Equity Instruments

The value of a stock is determined by both the expected return and the growth of its cash flows.

By assuming a constant growth rate for dividends, we can use the formula for the present value of an equity investment to calculate the stock's implied return or growth rate.

Implied Return

Recall that the present value of a stock for constant growth of dividends is given by:

$$PV_t = \frac{D_t(1+g)}{r-g} = \frac{D_{t+1}}{r-g}$$

Where:

PV_t = Present value at time t.

D_t = Expected Dividend in the next period.

r = Required rate of return.

g = Constant growth rate.

$r - g > 0$

Therefore, we can calculate the implied return on a stock given its expected dividend yield and implied growth by rearranging the above formula as follows:

$$r = \frac{D_t(1+g)}{PV_t} + g = \frac{D_{t+1}}{PV_t} + g$$

In simple terms, if we assume a stock's dividends will grow at a steady rate forever, the implied return is the combination of its expected dividend yield and the constant growth rate.

Example: Implied Return and Growth

Suppose Apple Inc. stock is trading at a share price of USD150.00, and its annualized expected dividend per share during the next year is USD2.00.

Moh, an analyst, projects that Apple's dividend per share will increase at a constant rate of 5% per year indefinitely. The required return expected by investors on the stock is *closest to*:

Solution

Recall that the implied return formula is,

$$r = \frac{D_t(1+g)}{PV_t} + g = \frac{D_{t+1}}{PV_t} + g$$

In this case, D=$2.00, PV=$150, g=5%

Therefore,

$$r = \frac{2.00(1.05)}{150} + 0.05 = 6.4\%$$

Implied Growth

We can also solve for a stock's implied growth rate, which is given by the following formula:

$$g = \frac{r \times PV_t - D_t}{PV_t} + D_t = \frac{r - D_{t+1}}{PV_t}$$

Example: Calculating the Implied

Consider the previous example. Suppose Moh believes that Apple stock investors should expect a return of 8%. Calculate the implied dividend growth rate for Apple Inc.

Solution

Recall that the formula for calculating implied growth is as follows.

$$g = \frac{r \times PV_t - D_t}{PV_t} + D_t = \frac{r - D_{t+1}}{PV_t}$$

So,

$$g = 0.08 - \frac{2.00 \times 1.05}{150} = 0.066 = 6.60\%$$

Price-to-Earnings Ratio

In equity instruments, it is common practice to compare the price-to-earnings ratio.

The **price-to-earnings (P/E) ratio** is a valuation metric that compares the current share price of a stock to its earnings per share. Investors and analysts use it to determine the relative value of a company's shares compared to other companies or the market.

A stock with a higher price-to-earnings ratio is more expensive than a lower one, as investors are willing to pay more for each unit of earnings. This ratio is also a valuation metric for stock indexes, such as S&P 500.

Relating P/E Ratio to Expected Future Cash Flows

The P/E ratio can relate to our earlier discussion on a stock's price (PV) to the expected future cash flow relationship. Recall the following equation:

$$PV_t = \frac{D_t \times (1 + g)}{r - g}$$

By dividing both sides of the equation by E_t, which represents earnings per share for period t, we get the following equation:

$$\frac{PV_t}{E_t} = \frac{\frac{D_t}{E_t} \times (1 - g)}{r - g}$$

Where:

$\frac{PV_t}{E_t}$ = Price-to-earnings (P/E) ratio.

$\frac{D_t}{E_t}$ = Dividend payout ratio.

g = Growth rate.

r = Required rate of return.

The dividend payout ratio represents the percentage of a company's earnings paid out to shareholders in the form of dividends.

Typically, the **forward P/E ratio**, which is based on a projection of a company's earnings per share for the next period $(t+1)$, is used. This ratio is positively correlated with higher expected dividend payouts and growth rates but negatively correlated with the required return.

Therefore, the equation:

$$\frac{PV_t}{E_t} = \frac{\frac{D_t}{E_t} \times (1-g)}{r-g}$$

Can be simplified as below to find the forward P/E ratio:

$$\frac{PV_t}{E_{t+1}} = \frac{\frac{D_{t+1}}{E_{t+1}}}{r-g} = \frac{D_{t+1}}{E_{t+1}} \times \frac{1}{r-g}$$

Example: Solving for Implied Dividend Growth Rate

Suppose a company has a forward P/E ratio of 15, a dividend payout ratio of 40%, and a required return of 10%. The implied dividend growth rate for this company is *closest to*:

Solution:

First, we can use the formula for the forward P/E ratio to solve for the implied dividend growth rate:

$$\frac{PV_t}{E_{t+1}} = \frac{D_{t+1}}{E_{t+1}} \times \frac{1}{r-g}$$

Where:

PV_t = Present value at time t.

E_{t+1}= Earnings per share for the next period.

D_{t+1}= Dividend payout for the next period.

r = Required return.

g = Implied dividend growth rate.

Substituting the given values into the formula, we get:

$$15 = \frac{0.4}{0.1 - g}$$

Solving for g, we get:

$$g = 0.1 - \frac{0.4}{15} = 0.0733$$

Therefore, the implied dividend growth rate for this company is 7.33%

Example: Solving for Required Return

Let's assume you are not given the required rate of return in the question above so that the company has a forward P/E ratio of 15, a dividend payout ratio of 40%, and an implied dividend growth rate of 7.33%. What is the required return for this company?

Solution

Recall the formula for the forward P/E ratio to solve for the required return:

$$\frac{PV_t}{E_{t+1}} = \frac{D_{t+1}}{E_{t+1}} \times \frac{1}{r - g}$$

Substituting the given values into the formula, we get:

$$15 = \frac{0.4}{r - 0.0733}$$

Solving for r, we get:

$$r = \frac{0.4}{15} + 0.0733 = 0.1000$$

Therefore, the required return for this company is 10%.

Question

Edmund company's stock trades at USD50.00. The company pays an annual dividend to its shareholders, and its most recent payment of USD 2.00 occurred yesterday. Analysts following the company expect its dividend to grow at a constant rate of 4 percent per year. What is the company's required return?

A. 8.16%.
B. 8.48%.
C. 9.16%.

The correct answer is A.

Solution

Recall that:

$$PV = \frac{D_{t+1} \times (1 + g)}{r - g}$$

Where:

PV = Current stock price.

D_{t+1} = Recent dividend payout.

g = Expected dividend growth rate.

r = Required return.

Substituting the given value into the formula, we get:

$$50 = \frac{2 \times (1 + 0.04)}{r - 0.04}$$

Solving for r, we get:

$$\frac{2 \times (1 + 0.04)}{}$$

$$r = \frac{2 \times (1 + 0.04)}{50} + 0.04 = 0.0816$$
$$r = 0.0816$$

Edmund's required return is 8.16%.

LOS 2c: explain the cash flow additivity principle, its importance for the no-arbitrage condition, and its use in calculating implied forward interest rates, forward exchange rates, and option values

A timeline is a physical illustration of the amounts and timing of cashflows associated with an investment project. For cashflows that are regular and of equal amounts, the standard annuity formula or the financial calculator can be used. However, a timeline is preferred for irregular, unequal, or both cashflows.

Remember that the general formula that relates the present value and the future value of an investment is given by:

$$FV_N = PV(1 + r)^N$$

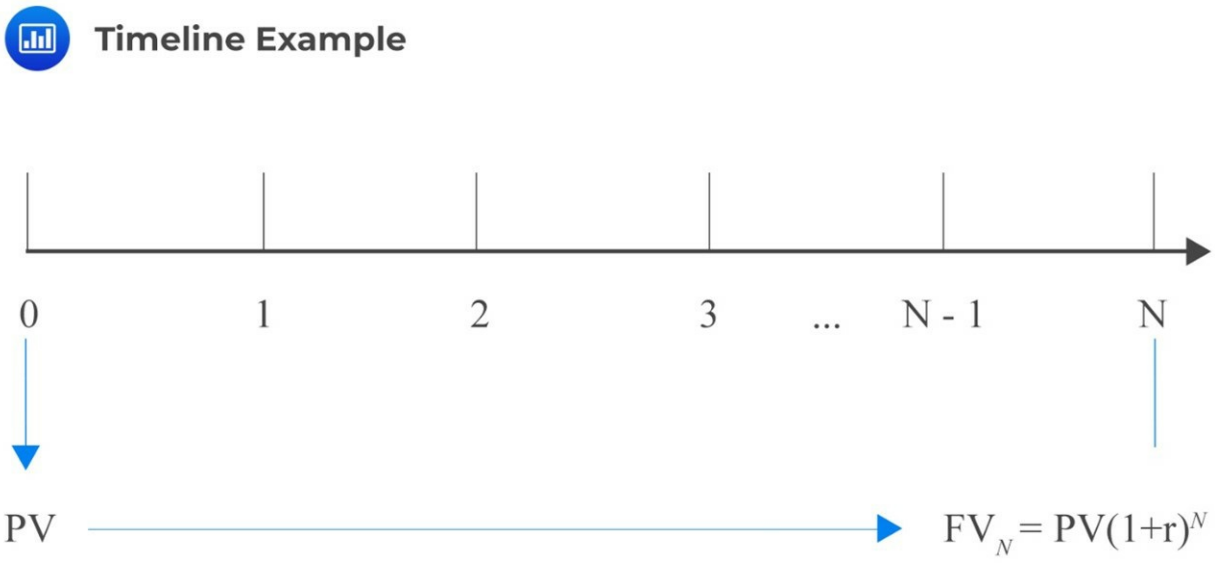
Where:

PV = Present value of the investment.

FV_N = Future value of the investment N periods from today.

r = Rate of interest per period.

We can represent this in a timeline:



In a particular timeline, a time index, t, represents a particular point in time, a specified number of periods from today. Therefore, the present value is the investment amount today (t = 0), and by using this amount, we can calculate the future value (t = N). Alternatively, we can use the future value to calculate the present value.

The above argument can be written in terms of the present value. That is:

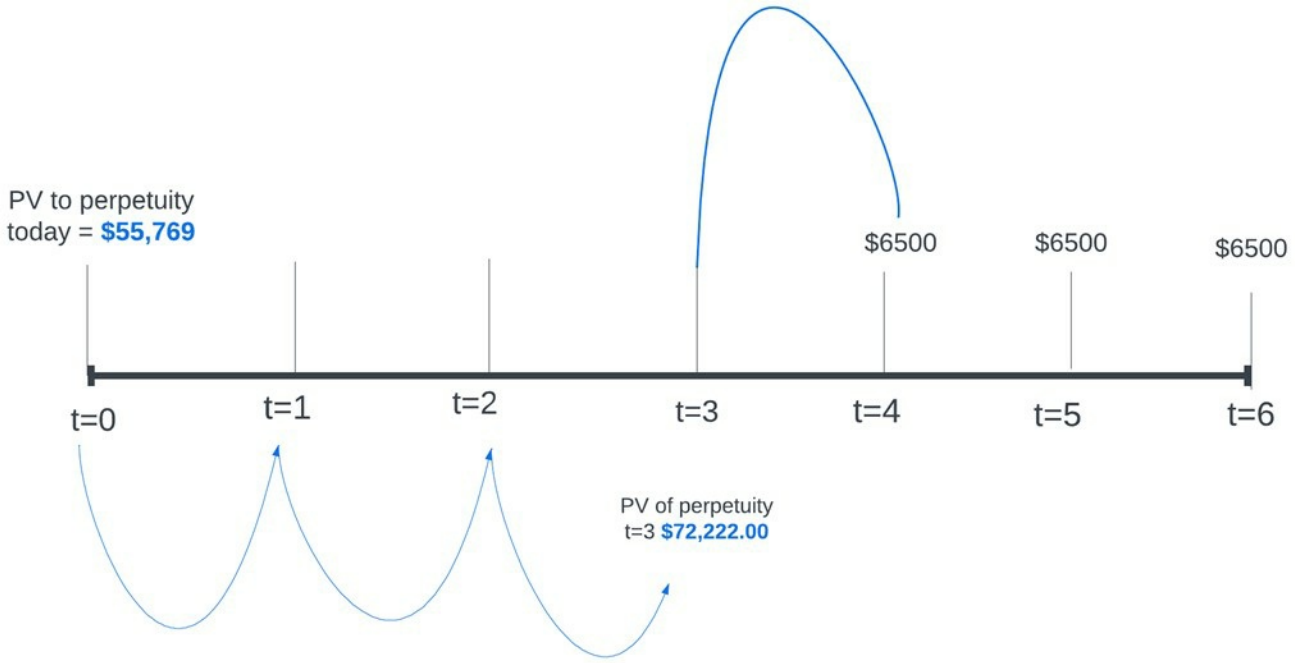
$$PV = FV_N(1 + r)^{-N}$$

Example: Applying a Timeline to Model Cashflows

A fixed-income investor receives a series of payments, each amounting to $6,500, set to be received in perpetuity. Payments are to be made at the end of each year, starting at the end of year 4. If the discount rate is 9%, then what is the present value of the perpetuity at t = 0?

Solution

We would then draw a timeline to understand the problem better:



Here, we can see that the investor is receiving $6,500 in perpetuity. Recall that the PV of aperpetuity is given by:

$$\text{PV of a perpetuity} = \frac{C}{r}$$

So, in this case:

$$PV_3 = \frac{\$6,500}{9\%} = \$72,222$$

This is the value of the perpetuity at $t = 3$, so we need to discount it for three more periods to get the value at $t = 0$. Using the formula:

$$PV_0 = FV_N(1 + r)^{-N}$$

The PV at time zero is $\frac{\$72,222}{(1+0.09)^3} = \$55,769$

Why is the Use of a Timeline Recommended?

There are many instances in real life when cashflows are uneven. A good example is a pension

contribution that varies with age. Applying one of the basic time value formulae is impossible in such cases. You are advised to draw a timeline even if the question appears relatively straightforward. It will help you understand the question structure better. A timeline also helps candidates add cashflows indexed to the same period and apply the value additivity principle.

Cashflow Additivity Principle

According to the cashflow additivity principle, the present value of any stream of cashflows indexed at the same point equals the sum of the present values of the cashflows. This principle has different applications in time value of money problems. Besides, this principle can be applied to in different economic scenarios.

Application of Cashflow Additivity

Investing in Different Currencies

The principle of cash flow additivity can be applied to scenarios involving different currencies by converting all cash flows to a common currency using the appropriate exchange rates. Doing so allows us to compare and combine cash flows from different currencies and make investment decisions based on their combined value.

For example, suppose we have two investment opportunities, one in US dollars and one in Japanese yen. In that case, we can convert the expected cash flows from the Japanese yen investment into US dollars using the appropriate exchange rates. Then, we can compare the combined value of the two investments and decide based on their relative values.

Dealing with different currencies assumes continuous compounding. Recall that the present value of a continuous compounding is given by:

$$PV = FV_N e^{-Nrs}$$

Example: Applying the Cash Flow Additivity Principle

Consider an investor with USD 2,000 who wants to invest it for three months. The investor can

choose between two options: investing in the US government debt or German government debt.

Option 1: Investing in US Government Debt

The investor can invest his USD 2,000 in a three-month US Treasury bill. This means that he lends the government USD 2,000, and it promises to pay him back with interest in three months. The interest is 3%, so after three months, he will receive:

Recall that,

$$\text{FV} = 2,000 \times e^{0.03 \times \frac{3}{12}} = \text{USD } 2,015$$

Option 2: Investing in German Government Debt

The investor chooses to invest in German government debt. To do this, the investor must convert his USD 2,000 into Euros at the current exchange of EUR/USD = 0.92 (1 USD =0.92EUR). This means that the investor will receive $(2,000 \times 0.92) = \text{EUR } 1840$. He can then lend this money to the German government by investing in a three-month German Treasury bill. Assuming the interest rate is 0.06 percent, after three months, the investor will receive:

$$\begin{aligned} \text{FV} &= \text{EUR } 1840 \times e^{0.06 \times \frac{3}{12}} \\ &= \text{EUR } 1,867.81 \end{aligned}$$

Assuming the investor wants his money in US dollars, we need to convert the EUR1867.81 back into USD at the forward exchange rate of USD/EUR = 1.0788. This means that the investor will receive:

$$\frac{\text{EUR } 1,867.81 \times \text{USD } 1.0788}{1 \text{ EUR}} = \$2,014.99 \approx 2,015$$

Both options give you the same amount of money after three months: USD 2,015. The difference is that one option involves investing in US dollars, and the other involves converting your money into Euros and back into US dollars.

The forward exchange rate of 1.0788 USD/EUR is important because it determines how much money you, the investor, will receive when converting your Euros back into US dollars. If this rate differs from 1.0788, there would be an arbitrage opportunity in converting Euros to dollars.

More on foreign exchange rates will be discussed later in the curriculum.

Implied Forward Rates

Consider two zero-coupon bonds. Bond A has a maturity of two years and a yield of 2% per annum, while bond B has a maturity of four years and a yield of 3%. An investor, who doesn't seek to take advantage of price differences and is risk-neutral, has $1,000 to invest. The investor has two investment options that earn the same return.

Option 1: The investor can put their money into bond B now, which has an annual yield of 3%, and will pay out at the end of the four years. The Future Value (FV) of this investment in four years, using the formula for compound interest, is:

$$FV_4 = PV_0(1 + r_4)^4 = 1000(1.03)^4 = 1,125.51$$

Option 2: Alternatively, the investor can initially invest in bond A and, after two years, reinvest the proceeds at a forward rate $F_{2,2}$ which represents a two-year forward rate starting in year two.

By the principle of cash flow additivity, a risk-neutral investor will not prefer one option over the other - they are indifferent between Options 1 and 2. This is because the Future Values of both investments at the end of four years should be the same:

$$FV_4 = PV_0(1 + r_4)^4 = PV_0(1 + r_2)(1 + F_{2,2})$$

In this scenario, this simplifies to:

$$1,125.51 = 1,000(1.02)^2(1 + F_{2,2})$$

Solving this equation for the forward rate gives:

$$\Rightarrow F_{2,2} = \frac{1,125.51}{1,000(1.02)^2} - 1 = 8.18\%$$

Therefore, to prevent arbitrage opportunities, the forward rate $F_{2,2}$ should be set to 8.18%. This

ensures that there is no potential for risk-free profits, maintaining market efficiency.

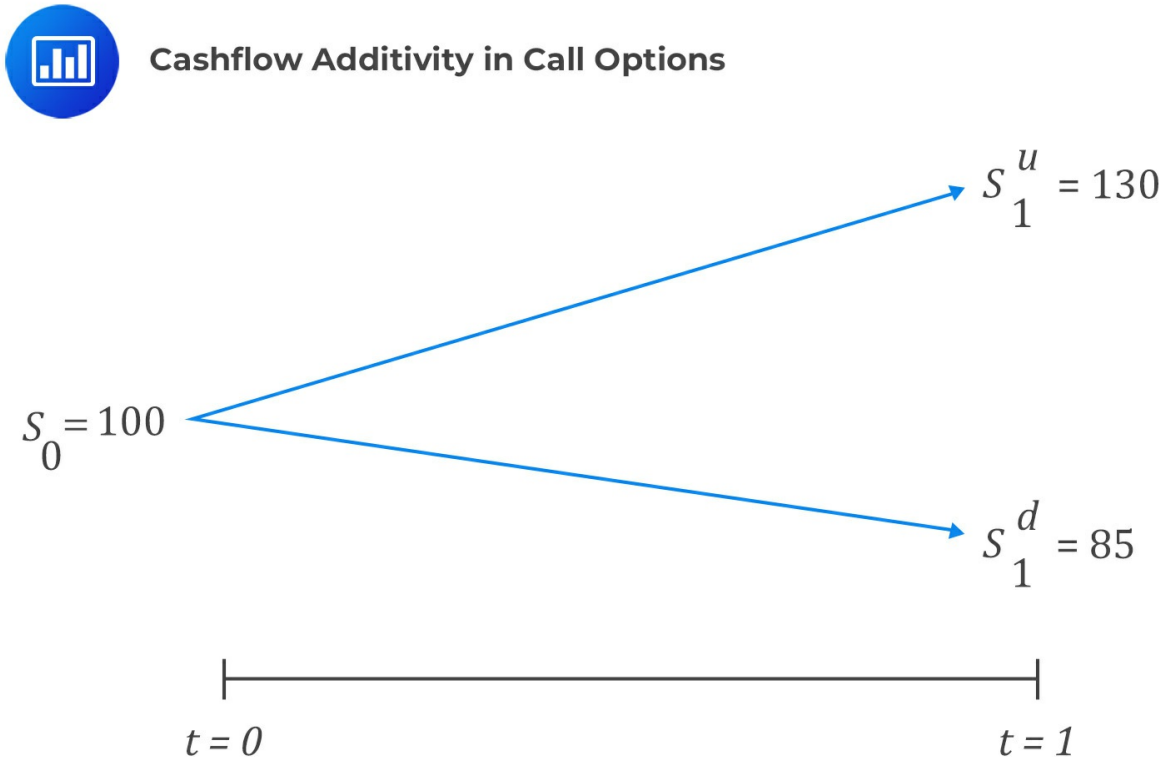
Cashflow Additivity and Option Pricing

Cash flow additivity can be used to determine the fair price of an option contract. An option contract gives the buyer the right, but not the obligation, to buy (call) or sell (put) an underlying asset at a specified price within a certain period.

Cash flow additivity allows investors to compare different strategies and determine a no-arbitrage price for a financial instrument.

Example: Illustrating Cashflow Additivity in Call Option

Consider a stock that costs $100 now. Its price might increase by 30% to $130 or decrease by 15% to $85 in one year.



Let's say an investor wants to sell a call on the stock that gives the buyer the right, but not the obligation, to purchase the asset for $120. The principle of cash flow additivity can be used to

determine the contract's no-arbitrage price.

If the stock price goes up, the contract is worth $c_1^u = 10$. That's because the buyer can use the contract to buy the asset for $120 and then sell it for $130, making a profit of $10. But if the price goes down, the contract is worth nothing. The buyer wouldn't want to use the contract to buy the asset for $120 when they could just buy it for $85 without the contract.

The underlying argument here is that the value of the option is each movement of the stock option may be used to construct a risk-free portfolio (the value of the portfolio is the same in both scenarios).

Denote the initial value of the call option by c_0, which we wish to determine using cash flow additivity and no-arbitrage pricing. Also, denote the value of the portfolio at $t = 0$ by V_0, when the stock price increases by V_1^u and when the stock price decreases by V_1^d

Assume that at $t = 0$ creates a risk-free portfolio by selling a call option at c_0 and buying 0.22 units of underlying assets. Then, the value of the portfolio at inception is:

$$V_0 = 0.22 \times 100 - c_0$$

This portfolio is called a **replicating portfolio** because it is designed to create a matching future cash flow stream to that of a risk-free asset.

Similarly, in each scenario of stock price decrease and increase:

$$V_1^u = 0.22 \times 130 - 10 = 18.89$$
$$V_1^d = 0.22 \times 85 - 0 = 18.89$$

Intuitively, the value replicating portfolio equals 18.89, whether the stock prices rise or decline. As such, the replicating portfolio is risk-free and can be discounted as a risk-free asset. Assuming that risk-free rate is r = 2.5%, then:

$$V_0 = V_1^u(1 + r)^{-1} = V_1^d(1 + r)^{-1}$$
$$= 18.89(1.025)^{-1} = 18.43$$

At this point, we can calculate the value of c_0 rearranging the initial portfolio value equation:

$$
\begin{aligned}
V_0 &= 0.22 \times 100 - c_0 \\
\Rightarrow 18.43 &= 0.22 \times 100 - c_0 \\
\therefore c_0 &= 3.57
\end{aligned}
$$

As such, the fair price of the call option is $3.57, which the seller expects to receive from the buyer.

Question

The current USD/CHF exchange rate is 0.9. The risk-free interest rates for one year are 2% for the US dollar and 1% for the Swiss franc. Which of the following one-year USD/CHF forward rates would best prevent arbitrage opportunities?

A. USD/CHF 0.909.
B. USD/CHF 0.099.
C. USD/CHF 0.891.

Solution

The correct Answer is A.

Dealing with different currencies assumes continuous compounding. Recall that the future value of a continuous compounding is given by:

$$FV = PV_N e^{Nr_s}$$

So,

In one year, a single unit of Swiss franc invested is:

$$e^{0.01} = \text{CHF } 1.0101$$

In one year, a single unit of Swiss Franc converted to US dollars and then invested risk-free is worth;

$$0.9e^{0.02} = \text{USD } 0.9182$$

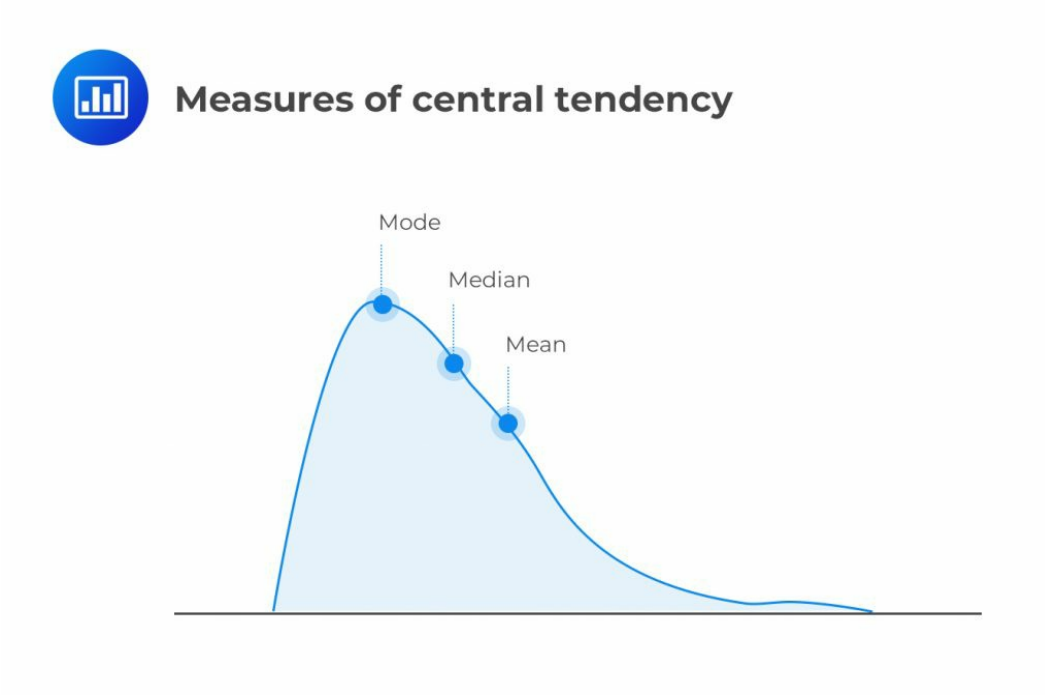
Therefore, to convert USD 0.9182 into CHF 1.0101 requires a forward exchange rate of:

$$\frac{0.9182}{1.0101} = \text{USD/CHF } 0.909$$

Learning Module 3: Statistical Measures of Asset Returns

LOS 3a: calculate, interpret, and evaluate measures of central tendency and location to address an investment problem

The center of any data is identified via a measure of central tendency. A measure of central tendency for a series of returns reveals the center of the empirical distribution of returns. They include mean, mode, and median.



Measures of location help us understand where data points tend to cluster. These measures include central tendency measures such as mean, median, and mode. There are also other measures that provide different insights into how the data is spread out or located within a distribution.

Measures of Central Tendency

Arithmetic Mean

The arithmetic mean is the sum of the values of the observations in a dataset divided by the

number of observations.

Recall the formula: denoted by $\bar{R}_i$ arithmetic mean for an asset *i* is a simple process of finding the average holding period returns. It is given by:

$$\bar{R}_i = \frac{R_{i,1} + R_{i,1} + \ldots + R_{i,T-1} + R_{iT}}{T} = \frac{1}{T}\sum_{t=1}^{T} R_{it}$$

Where:

R_{it} = Return of asset *i* in period *t*.

T = Total number of periods.

For example, if a share has returned 15%, 10%, 12%, and 3% over the last four years, then the arithmetic mean is computed as follows:

$$\bar{R}_i = \frac{1}{T}\sum_{t=1}^{T} R_{it} = \frac{1}{4}(15\% + 10\% + 12\% + 3\%) = 10\%$$

Population Mean and Sample Mean

The population mean is the summation of all the observed values in the population, $\sum X_i$ divided by the total number of observations, N. The population mean differs from the sample mean, which is based on a few observed values chosen from the population. Thus:

$$\text{Population mean} = \frac{\sum X_i}{N}$$

$$\text{Sample mean} = \frac{\sum X_i}{n}$$

Analysts use the sample mean to *estimate* the actual population mean.

The population mean and the sample mean are both arithmetic means. The arithmetic mean for any data set is unique and is computed using all the data values. Among all the measures of central tendency, it is the only measure for which the sum of the deviations from the mean is

zero.

Example: Calculating the Arithmetic Mean

The following are the annual returns realized from a given asset between 2005 and 2015.

{12% 13% 11.5% 14% 9.8% 17% 16.1% 13% 11% 14%}

Calculate the population mean.

Solution

$$\text{Population mean} = \frac{0.12 + 0.13 + 0.115 + 0.14 + 0.098 + 0.17 + 0.161 + 0.13 + 0.11 + 0.14}{10}$$
$$= 0.1314 = 13.14\%$$

Median

The median is the statistical value located at the center of a data set organized in ascending or descending order.

Consider a sample of n observations. For an odd-numbered of observations, the median is the observation that is located in $\frac{(n+1)}{2}$ position. On the other hand, if the number of observations is even, the media is the mean value of the observations located in $\frac{n}{2}$ and $\frac{n+2}{2}$ position.

Unlike the arithmetic mean, the median resists the effects of extreme observations. However, it only gives the relative position of the ranked observations without considering all observations relating to the size of the observations.

Example: Calculating the Median

The following are the annual returns on a given asset realized between 2005 and 2015.

{12% 13% 11.5% 14% 9.8% 17% 16.1% 13% 11% 14%}

The median is *closest* to:

Solution

First, we arrange the returns in ascending order:

{9.8% 11% 11.5% 12% 13% 13% 14% 14% 16.1% 17%}

Since the number of observations is even, the median return will be the middle point of the two middle values in the positions $\frac{n}{2}$ and $\frac{(n+2)}{2}$.

The value occupying $\frac{n}{2} = \frac{10}{5} = 5$th position is 13, and the value located in $\frac{(n+2)}{2} = \frac{12}{2} = 6$th position is 13, so that the mode is:

$$\frac{13\% + 13\%}{2} = 13\%$$

Mode

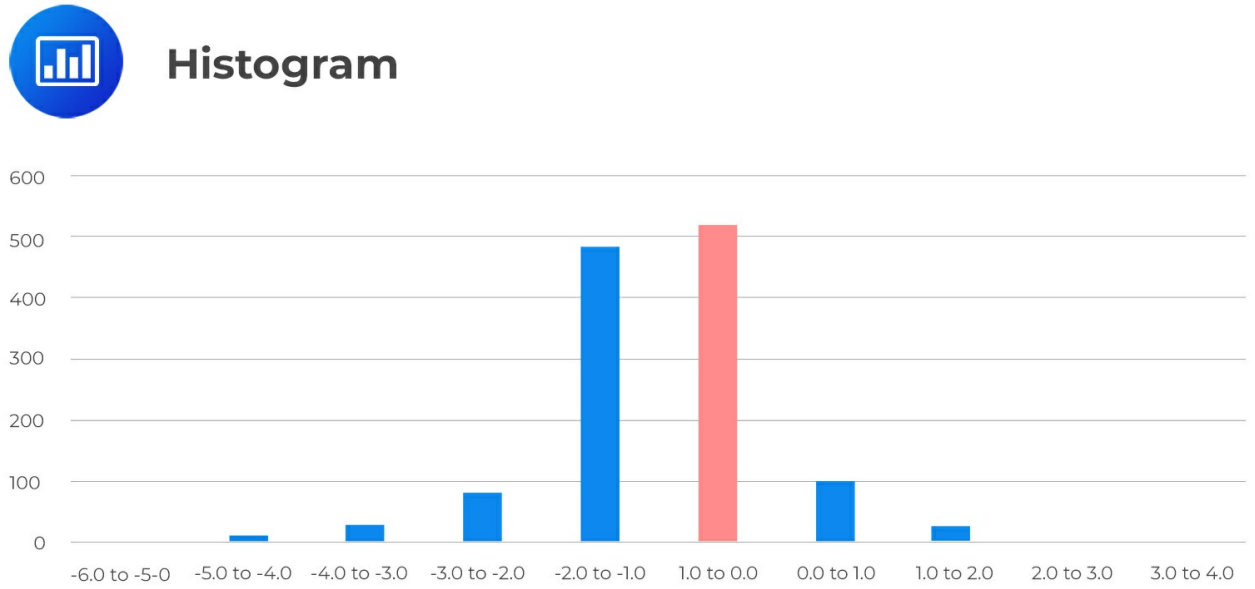
The mode is the value that appears most often in a dataset. Sometimes, a dataset has a mode; sometimes, it doesn't. If all the observations in a dataset are different and no value repeats more than others, then the dataset has no mode.

A dataset with one mode is called unimodal. When there are two modes, it's called bimodal. If the distribution has three frequently occurring values, it's termed trimodal.

An interval with the highest frequency is called the modal interval (or intervals) in a frequency distribution. For instance, in the frequency distribution below, the modal interval is -1.0 to 0.0 with an absolute frequency of 520.

Return Bin (%)	Absolute Frequency	Relative Frequency(%)	Cumulative Absolute Frequency	Cumulative] Relative Frequency (%)
-6.0 to -5.0	2	0.16	2	0.16
-5.0 to -4.0	8	0.64	10	0.80
-4.0 to -3.0	27	2.16	37	2.96
-3.0 to -2.0	80	6.40	117	9.36
-2.0 to -1.0	485	38.80	602	48.16
-1.0 to 0.0	520	41.60	1,122	89.76
0.0 to 1.0	100	8.00	1,222	97.76
1.0 to 2.0	24	1.92	1,246	99.68
2.0 to 3.0	3	0.24	1,249	99.92
3.0 to 4.0	1	0.08	1,250	100.00

In a histogram, the modal interval always has the highest bar.



The mode is the only measure of central tendency that can be used with nominal data. Nominal data refers to a type of data that is categorized into distinct categories or groups without any inherent order or numerical value. Examples of nominal data include gender (male, female), eye color (blue, brown, green), and marital status (single, married, divorced).

Example: Calculating the Mode

Determine the mode from the following data set:

{20% 23% 20% 16% 21% 20% 16% 23% 25% 27% 20%}

Solution

The mode is 20%. It occurs four times, a frequency higher than any other value in the data set. Clearly, this dataset is unimodal.

Dealing With Outliers

An outlier may represent a distinct value in a population. In addition, it may show that there was an error in recording the value, or it was generated from a different population.

When working with a sample with outliers, we can potentially transform the variable or choose another variable that achieves the same objective. If these observations prove to be impossible, there are three options:

Option 1: Take no action and use the data as it is. If these observations are accurate, then this is appropriate.

Option 2: Remove outliers using the trimmed mean. For example, when calculating the central tendency, a 4 percent trimmed mean excludes the lowest 2% and highest 2% of values.

Option 3: Substitute a different value for the outliers. The winsorized mean is an illustration of a central tendency that does this. For instance, when computing a 96% winsorized mean, the value at or above the lowest and highest 2% is assigned the lowest and highest 2% values.

Measures of Location

Quartiles, quintiles, deciles, and percentiles are values or cut points that partition a finite number of observations into nearly equal-sized subsets. The number of partitions depends on the type of cut point involved.

Quartiles

They divide data into **four** parts. The first quartile, Q_1, is referred to as the lower quartile, and the last quartile, $Q_{4,}$ is the upper quartile. Q_1 splits the data into the lower 25% and upper 75% values. Similarly, the upper quartile subdivides the data into the lower 75% of the values and the upper 25%. The difference between the upper and lower quartiles is known as the **interquartile range**, which indicates the spread of the middle 50% of the data.

Quintiles

Though rarely used in practice, quintiles split a set of data into **five** equal parts, i.e., fifths. Therefore, the second quintile splits data into the lower 40% of the values and the upper 60%.

Deciles

The deciles subdivide data into **ten** equal parts. There are 10 deciles in any data set. For example, the fourth decile splits data into the lower 40% of the values and the upper 60%.

Percentiles

Percentiles split data into **100** equal parts, i.e., hundredths. So, for instance, the 77^{th} percentile splits the data into the lower 77% of the values and the upper 23%.

Financial analysts commonly use the four types of subdivisions to rank investment performance. You should note that quartiles, quintiles, and deciles can all be expressed as percentiles. For instance, the first quartile is just the 25^{th} percentile. Similarly, the fourth decile is simply the 40^{th} percentile. This enables the application of the formula below.

$$\text{Position of percentile} = \frac{(n+1)\,y}{100}$$

Example: Calculating Quartiles

Given the following distribution of returns, calculate the lower quartile.

{10% 23% 12% 21% 14% 17% 16% 11% 15% 19%}

Solution

First, we have to arrange the values in ascending order:

{10% 11% 12% 14% 15% 16% 17% 19% 21% 23%}

Next, we establish the position of the first quartile. This is simply the 25^{th} percentile. Therefore:

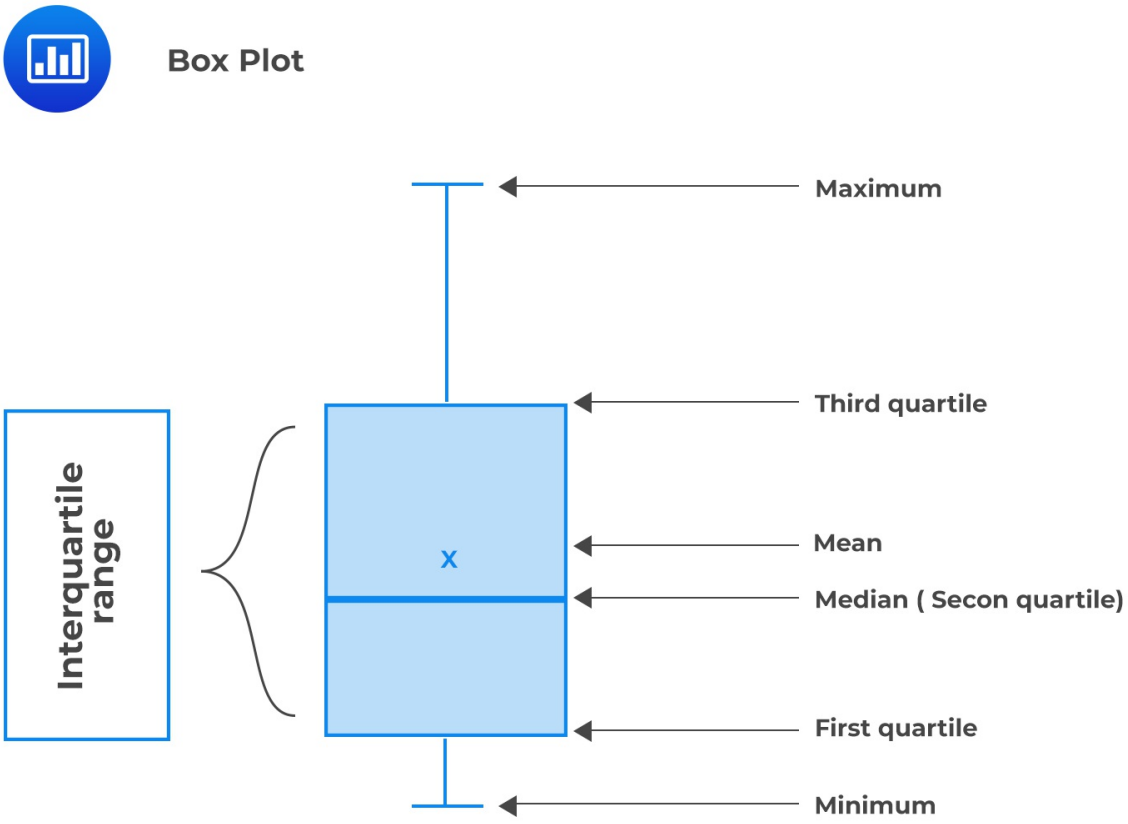
$$P_{25} = \frac{(10 + 1)\,25}{100} = 2.75^{th} \text{ value}$$

Since the value is not straightforward, we have to extrapolate between the 2^{nd} and the 3^{rd} data points. The 25^{th} percentile is three-fourths (0.75) of the way from the 2^{nd} data point (11%) to the 3^{rd} data point (12%):

$$11\% + 0.75 \times (12 - 11) = 11.75\%$$

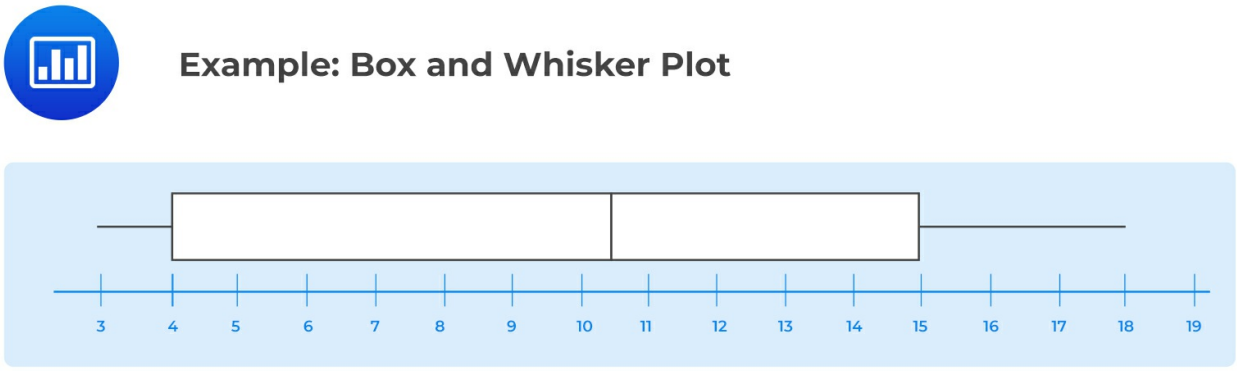
Box and Whisker Plot

Box and whisker plot is used to display the dispersion of data across quartiles. A box and whisker plot consists of a "box" with "whiskers" connected to the box. It shows the following five-number summary of a set of data.



Example: Box and Whisker Plot

Consider the following box and whisker plot:



1. Which of the following is *most likely* the median?

A. 10.

B. 10.5.

C. 11.

Solution

The correct answer is B.

$$\text{Median or Quartile 2 (Q2)} = \frac{(10 + 11)}{2} = 10.5$$

2. Which of the following is most likely the interquartile range?

A. 4.5.

B. 6.5.

C. 11.

Solution

The correct answer is C.

$$\text{Interquartile range is } Q_3 - Q_1 = 15 - 4 = 11$$

3. Which of the following is *most likely* the 3rd quartile?

A. 4.

B. 15.

C. 18.

Solution

$$\text{Quartile 3 (Q3)} = \frac{3 \times (n + 1)}{4} = \frac{3 \times (16 + 1)}{4} = 12.75\text{th term, which is } 14.75$$

Quantiles in Investment Practice

Quantiles have two main purposes in investment practice. First, quantiles are used to rank performance. Secondly, quantiles can be used in investment research for comparison purposes. For example, companies can be clustered into deciles to compare the performance of small companies with the large ones. In this case, the first decile will contain the portfolio of companies with the smallest market values, while the tenth decile will contain the companies with the largest market values.

Question

A mutual fund achieved the following rates of growth over an 11-month period:

{3% 2% 7% 8% 2% 4% 3% 7.5% 7.2% 2.7% 2.09%}

The 5^{th} decile from the data is *closest* to:

A. 2%.

B. 3%.

C. 4%.

Solution

The correct answer is B.

First, you should re-arrange the data in ascending order:

{2% 2% 2.09% 2.7% 3% 3% 4% 7% 7.2% 7.5% 8%}

Secondly, you should establish the 5^{th} decile. This is simply the 50^{th} percentile and is actually the median:

$$\begin{aligned} P_{50} &= \frac{(1 + 11)\,50}{100} \\ &= 12 \times 0.5 \\ &= 6, \text{ i.e., the 6th data point} \end{aligned}$$

Therefore, the 5th decile = 50th percentile = Median = 3%

LOS 3b: calculate, interpret, and evaluate measures of dispersion to address an investment problem

Measures of dispersion are used to describe the variability or spread in a sample or population. They are usually used in conjunction with measures of central tendency, such as the mean and the median. Specifically, measures of dispersion are the range, variance, absolute deviation, and standard deviation.

Measures of dispersion are essential because they give us an idea of how well the measures of central tendency represent the data. For example, if the standard deviation is large, then there are large differences between individual data points. Consequently, the mean may not be representative of the data.

Range

The range is the difference between the highest and the lowest values in a dataset, i.e.,

$$\text{Range} = \text{Maximum value} - \text{Minimum value}$$

Example: Calculating the Range

Consider the following scores of 10 level I candidates:

{78 56 67 51 43 89 57 67 78 50}

$$\text{Range} = 89 - 43 = 46$$

Advantage of the Range

- The range is easy to compute.

Disadvantages of the Range

- The range is not a reliable dispersion measure. It provides limited information about the distribution because it uses only two data points.

- The range is sensitive to outliers.

Mean Absolute Deviation (MAD)

MAD is a measure of dispersion representing the **average of the absolute values** of the deviations of individual observations from the arithmetic mean. Therefore,

$$\text{MAD} = \frac{\sum |X_i - \bar{X}|}{n}$$

Remember that the sum of deviations from the arithmetic means is always zero, which is why we use **absolute values**.

Example: Calculating Mean Absolute Deviation

Six financial analysts have reported the following returns on six different large-cap stocks over 2021:

{6% 7% 12% 2% 3% 11%}

Calculate the mean absolute deviation and interpret it.

Solution

First, we have to calculate the arithmetic mean:

$$\bar{X} = \frac{(6\% + 7\% + 12\% + 2\% + 3\% + 11\%)}{6} = 6.83\%$$

Next, we can now compute the MAD:

$$\begin{aligned}\text{MAD} &= \frac{\{|6\% - 6.83\%| + |7\% - 6.83\%| + |12\% - 6.83\%| + |2\% - 6.83\%| + |3\% - 6.83\%| + |11\% - 6.8}{6} \\ &= \frac{0.83 + 0.17 + 5.17 + 4.83 + 3.83 + 4.17}{6} \\ &= 3.17\%\end{aligned}$$

Interpretation: On average, an individual return deviates by 3.17% from the mean return of 6.83%.

Sample Variance and Sample Standard Deviation

The sample variance,s^2, is the measure of dispersion that applies when working with a sample instead of a population.

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

Where:

$\bar{X}$ = Sample mean.

n = Number of observations.

Note that we are dividing by n–1. This is necessary to remove **bias**.

The sample standard deviation, s, is simply the square root of the sample variance.

$$s = \sqrt{s^2} = \sqrt{\frac{(X_i - \bar{X})^2}{n - 1}}$$

Example: Calculating Sample Mean and Variance

Assume that the returns realized in the previous example were sampled from a population comprising 100 returns. The sample mean and the corresponding sample variance are *closest* to:

Solution

The sample mean will still be 6.83%.

Hence,

$$s^2 = \frac{\{(6\% - 6.83\%)^2 + (7\% - 6.83\%)^2 + (12\% - 6.83\%)^2 + (2\% - 6.83\%)^2 + (3\% - 6.83\%)^2 + (11\%}{5}$$
$$= 0.001656$$

Therefore,

$$s = 0.001656^{\frac{1}{2}}$$
$$= 0.0407$$

Downside Deviation and Coefficient of Variation

When trying to estimate downside risk (i.e., returns below the mean), we can use the following measures:

- **Semi-variance**: The average squared deviation below the mean.
- **Semi-deviation** (also known as semi-standard deviation): The positive square root of semi-variance.
- **Target semi-variance**: The sum of the squared deviations from a specific target return.
- **Target semi-deviation**: The square root of target semi-variance.

Sample Target Semi-Deviation

The target semi deviation, s_{Target} , is calculated as follows:

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

Where B is the target and n is the total number of sample observations.

Yearly returns of an equity mutual fund are provided as follows.

Month	Return %
2010	36%
2011	29%
2012	10%
2013	52%
2014	41%
2015	16%
2016	10%
2017	23%
2018	−10%
2019	−19%
2020	2%

What is the target downside deviation if the target return is 20%?

Solution

Month	Return %	Deviation from the 20% target	Deviation below the target	Squared deviations below the target
2010	36.00	16.00	−	−
2011	29.00	9.00	−	−
2012	10.00	(10.00)	(10.00)	100
2013	52.00	32.00	−	
2014	41.00	21.00	−	
2015	16.00	(4.00)	(4.00)	16
2016	10.00	(10.00)	(10.00)	100
2017	23.00	3.00	−	
2018	(10.00)	(30.00)	(30.00)	900
2019	(19.00)	(39.00)	(39.00)	1,521
2020	2.00	(18.00)	(18.00)	324
Sum				2,961

Here n = 11–1 = 10 so that:

$$\text{Target semi-deviation} = \left(\frac{2961}{10}\right)^{0.5} = 17.21\%$$

Coefficient of Variation

The coefficient of variation, CV, is a measure of spread that describes the amount of variability of data relative to its mean. It has **no units**, so we can use it as an alternative to the standard deviation to compare the variability of data sets that have different means. The coefficient of variation is given by:

$$CV = \frac{s}{\bar{X}}$$

Where:

s = Standard deviation of a sample.

$\bar{X}$ = Mean of the sample.

***Note*:** The formula can be replaced with $\frac{?}{?}$ when dealing with a population.

Procedure to Follow While Calculating the Coefficient of Variation:

1. Compute the mean of the data.
2. Calculate the sample standard deviation of the data set, s.
3. Find the ratio of s to the mean, x?.

Example: Coefficient of Variation

What is the relative variability for the samples 40, 46, 34, 35, and 45 of a population?

Solution

Step 1: Calculate the mean.

$$\text{Mean} = \frac{(40 + 46 + 34 + 35 + 45)}{5} = \frac{200}{5} = 40$$

Step 2: Calculate the sample standard deviation. (Start with the variance, s^2.)

$$s^2 = \frac{(40-40)^2+\ldots+(45-40)^2}{4}$$
$$= \frac{122}{4}$$
$$= 30.5$$

Note: Since it is the sample standard deviation (not the population standard deviation), we use $n-1$ as the denominator.

Therefore,

$$s = \sqrt{30.5} = 5.52268$$

Step 3: Calculate the ratio.

$$\frac{\text{Mean}}{s} = \frac{5.52268}{40} = 0.13806 \text{ or } 13.81\%$$

Interpreting the Coefficient of Variation

In finance, the coefficient of variation is used to measure the **risk per unit of return**. For example, imagine that the mean monthly return on a T-Bill is 0.5% with a standard deviation of 0.58%. Suppose we have another investment, say, Y, with a 1.5% mean monthly return and standard deviation of 6%; then,

$$CV_{T-Bill} = \frac{0.58}{0.5} = 1.16$$

$$CV_Y = \frac{6}{1.5} = 4$$

Interpretation: The dispersion per unit monthly return of T-Bills is less than that of Y. Therefore, investment Y is riskier than an investment in T-Bills.

Question 1

If a security has a mean expected return of 10% and a standard deviation of 5%, its coefficient of variation is *closest* to:

A. 0.005.

B. 0.500.

C. 2.000.

Solution

The correct answer is **B**.

$$\text{CV} = \frac{\text{S}}{\text{x?}} = \frac{0.05}{0.10} = 0.5$$

Where:

s = The standard deviation of the sample.

x? = The mean of the sample.

A is incorrect. It assumes the following calculation.

$$\text{CV} = \frac{0.05}{10} = 0.005$$

C is incorrect. It assumes the following calculation.

$$\text{CV} = \frac{10}{5} = 2$$

Question 2

You have been given the following data:

{12 13 54 56 25}

Assuming that this is a sample from a certain population, the sample standard deviation is *closest* to:

A. 21.62.

B. 374.00.

C. 1,870.00.

The correct answer is **A**.

$$\bar{X} = \frac{(12 + 13 + \cdots + 25)}{5} = \frac{160}{5} = 32$$

Hence,

$$s^2 = \frac{\{(12 - 32)^2 + (13 - 32)^2 + (54 - 32)^2 + (56 - 32)^2 + (25 - 32)^2\}}{4}$$
$$= \frac{1870}{4} = 468$$

Therefore,

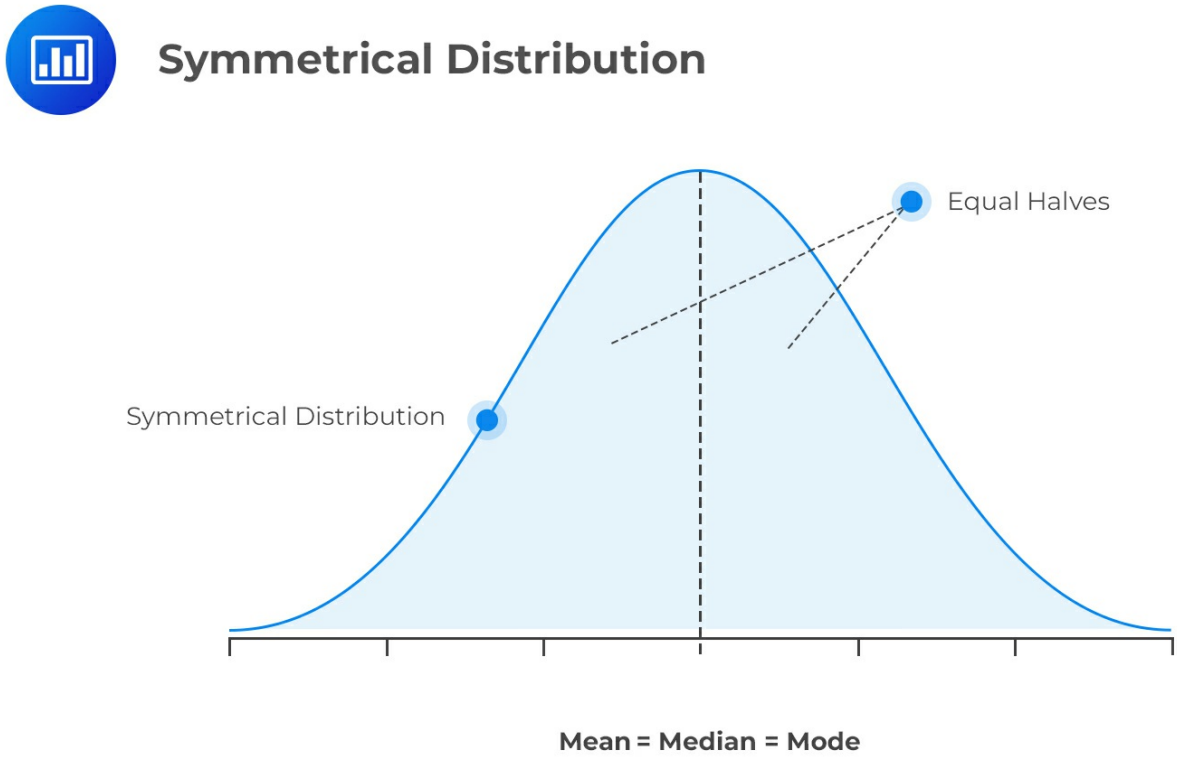
$$s = \sqrt{468} = 21.62$$

LOS 3c: interpret and evaluate measures of skewness and kurtosis to address an investment problem

Since the deviations from the mean are squared when calculating variance, we cannot determine whether significant deviations are more likely to be positive or negative. In order to identify other crucial distributional traits, we must look beyond measures of central tendency, location, and dispersion.

Skewness

Skewness refers to the degree of deviation from a symmetrical distribution, such as the normal distribution. A symmetrical distribution has identical shapes on either side of the mean.



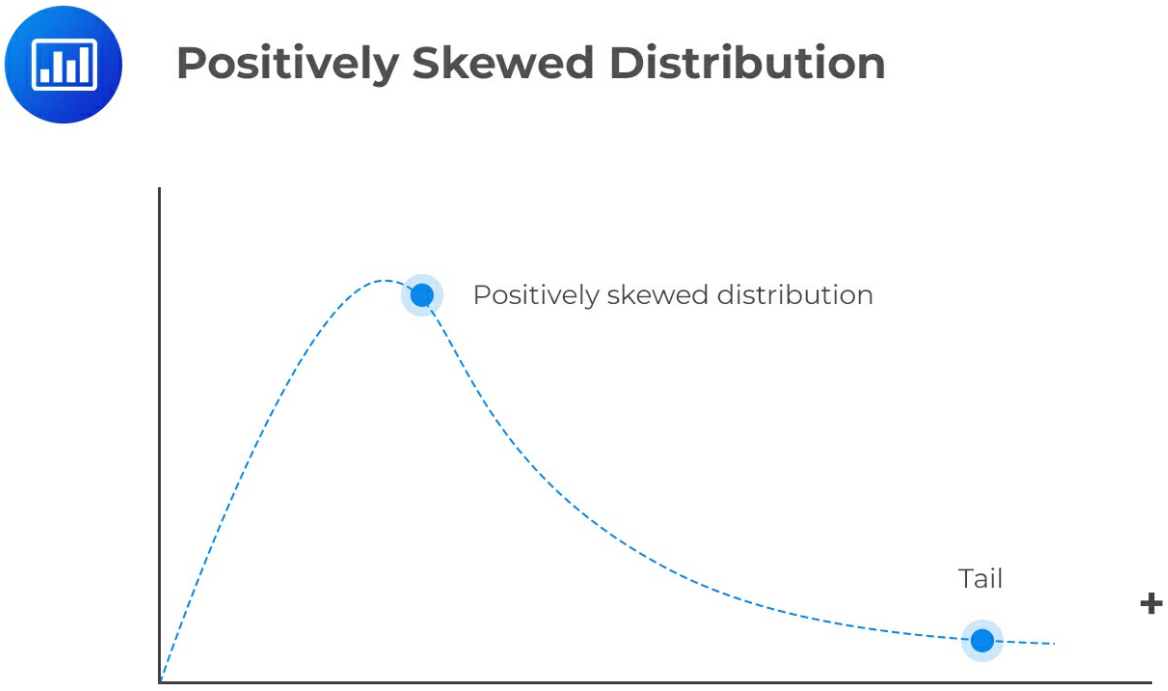
Distributions that are nonsymmetrical have unequal shapes on either side of the mean, leading to skewness. This is because nonsymmetrical distributions depart from the usual bell shape of the

normal distribution.

Skewness can be positive, negative, or, in some cases, undefined. The shape of a skewed distribution depends on outliers, which are extremely negative and positive observations.

Positive Skewness

A positively skewed distribution has a **long right tail** because of many outliers or extreme values on the right side. Perhaps the best way to remember its shape is to consider its points in a positive direction. Most data points are concentrated on the left side.

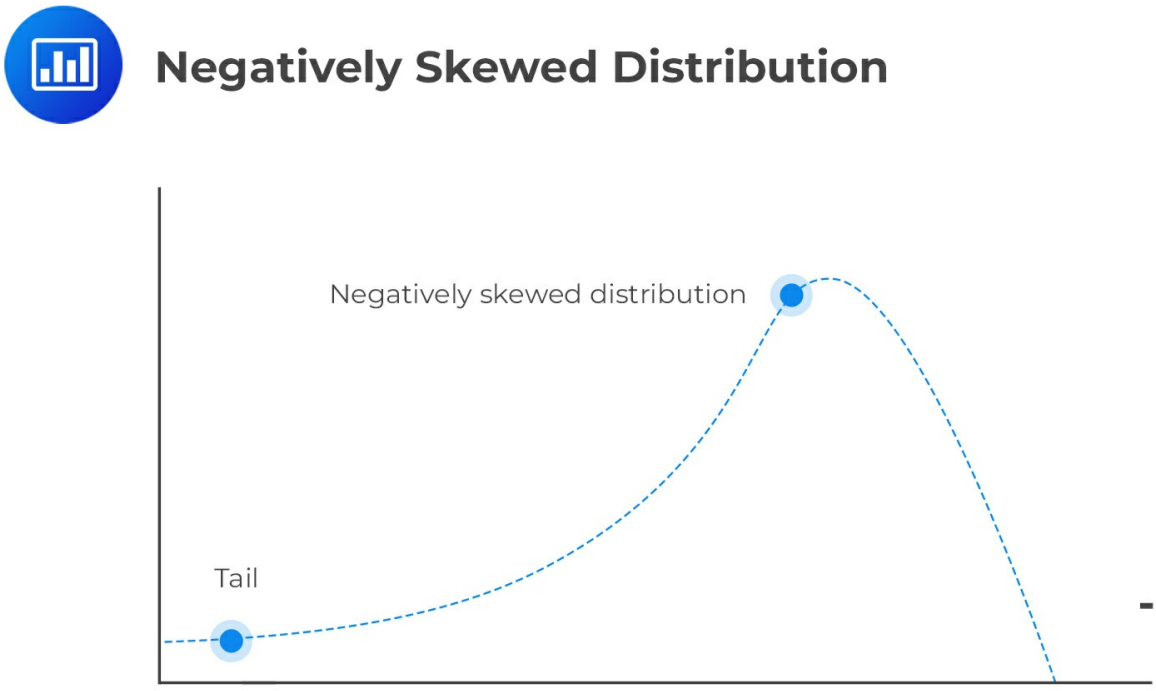


An example of a positively skewed distribution would be the income of individuals living in a specific country.

Negative Skewness

A negatively skewed distribution has a long left **tail** resulting from many outliers on the left side of the distribution. Therefore, we could say that it points in the negative direction. This is

because the right side harbors most of the data points.



Application of Skewness

Skewness matters in finance. Market data often show positive or negative skewness, like stock prices or mortgage costs. Investors can predict if future prices will be above or below the mean based on the skewness of the market segment.

Calculating Sample Skewness

The approximate sample skewness when sample is large ($n \geq 100$) is given by:

$$\text{Skewness} = \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$$

Where:

$\bar{X}$ = Sample mean.

s = Sample standard deviation.

n = Number of observations.

A positive value indicates positive skewness. A 'zero' value indicates that the data is not skewed. Lastly, a negative value indicates negative skewness or a negatively skewed distribution.

Example: Calculating Skewness

Suppose we have the following observations:

{12 13 54 56 25}

What is the skewness of the data?

Solution

First, we must determine the sample mean and the sample standard deviation:

$$\bar{X} = \frac{(12 + 13 + 54 + 56 + 25)}{5} = \frac{160}{5} = 32$$

$$s^2 = \frac{(12 - 32)^2 + (13 - 32)^2 + \ldots (25 - 32)^2}{4}$$

$$= 467.5$$

Therefore,

$$s = \sqrt{467.5} = 21.62$$

Now we can work out the skewness:

$$\text{Skewness} = \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n} (X_i - \bar{X})^3}{s^3}$$

$$\text{Skewness} = \left(\frac{1}{5}\right) \frac{(-20)^3 + (-19)^3 + 22^3 + 24^3 + (-7)^3}{21.62^3}$$

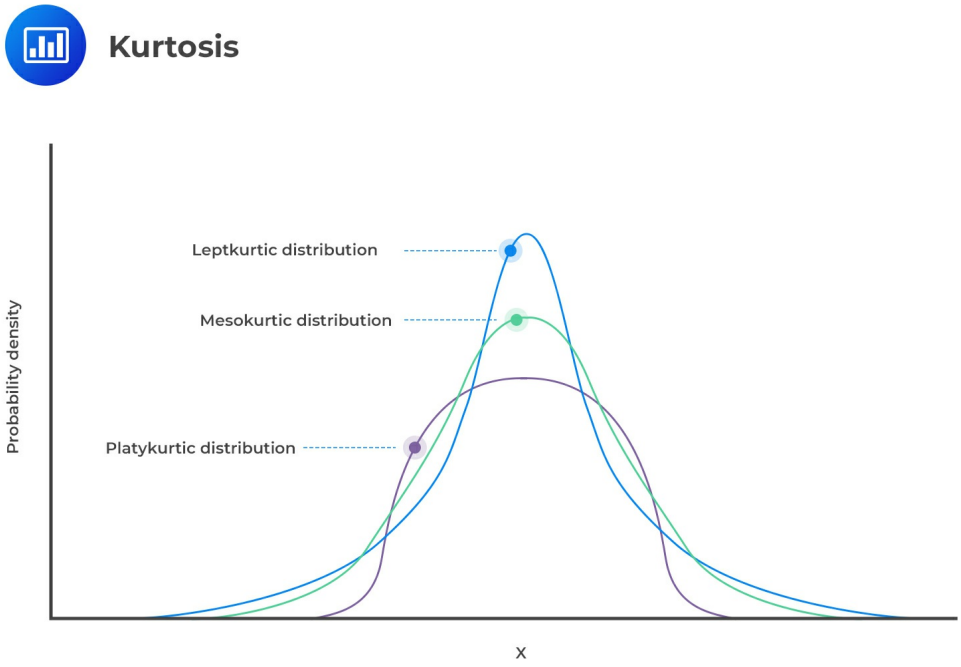
$$\text{Skewness} = 0.1835$$

Skewness is positive. Hence, the data has a positively skewed distribution.

Kurtosis

Kurtosis refers to the measurement of the degree to which a given distribution is more or less 'peaked' relative to the normal distribution. The concept of kurtosis is instrumental in decision-making. In this regard, we have three categories of distributions:

- Leptokurtic.
- Mesokurtic.
- Platykurtic.



Leptokurtic

A leptokurtic distribution is more peaked than the normal distribution. The higher peak results from the clustering of data points along the *x*-axis. The tails are also fatter than those of a normal distribution. The coefficient of kurtosis is usually more than 3.

The term "lepto" means thin or skinny. When analyzing historical returns, a leptokurtic distribution means that small changes are less frequent since historical values are clustered

around the mean. However, there are also large fluctuations represented by the fat tails.

Platykurtic

A platykurtic distribution has extremely dispersed points along the x-axis, resulting in a lower peak when compared to a normal distribution. "Platy" means broad. Hence, the prefix fits the distribution's shape, which is wide and flat. The points are less clustered around the mean compared to a leptokurtic distribution. The coefficient of kurtosis is usually less than 3.

Returns that follow this type of distribution have fewer major fluctuations compared to leptokurtic returns. However, you should note that fluctuations represent the riskiness of an asset. More fluctuations represent more risk and vice versa. Therefore, platykurtic returns are less risky than leptokurtic returns.

Mesokurtic

Lastly, mesokurtic distributions have a curve that is similar to that of a normal distribution. In other words, the distribution is mainly normal.

The majority of equity return series are found to have fat tails. Suppose a return distribution has fat tails, and we apply statistical models that do not consider distribution. In that case, we will overestimate the probability of either extremely poor or very favorable outcomes.

Investors often study a stock's daily trading volume distribution to assess its trading liquidity. It helps them see if the market can handle a large trade in that stock. This is useful for investors who want to make big investments or exit their positions in a particular stock.

Calculating Sample Kurtosis

Sample kurtosis is always measured relative to the kurtosis of a normal distribution, which is 3. Therefore, we are always interested in the "excess "kurtosis, i.e.,

$$\text{Excess kurtosis} = \text{Sample kurtosis} - 3$$

Where:

$$\text{Sample Excess Kurtosis} = \left(\frac{1}{n}\right)\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{s^4}$$

Positive excess kurtosis indicates a leptokurtic distribution. A zero value indicates a mesokurtic distribution. Lastly, a negative excess kurtosis represents a platykurtic distribution.

Example: Calculating Kurtosis

Using the data from the example above (12, 13, 54, 56, and 25), determine the type of kurtosis present.

$$\bar{X} = \frac{(12 + 13 + 54 + 56 + 25)}{5} = \frac{160}{5} = 32$$
$$s^2 = \frac{(12 - 32)^2 + (13 - 32)^2 + \ldots (25 - 32)^2}{4} = 467.5$$
$$s = \sqrt{467.5} = 21.62$$

Therefore,

$$\text{Excess Kurtosis} = \left(\frac{1}{n}\right)\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{s^4} - 3$$
$$\text{Excess Kurtosis} = \left(\frac{1}{5}\right)\frac{(-20)^4 + (-19)^4 + 22^4 + 24^4 + (-7)^4}{21.62^4} - 3$$
$$\text{Excess Kurtosis} = 2.2139$$

Since the excess kurtosis is negative, we have a platykurtic distribution.

Question 1

The skewness of the normal distribution is *most likely:*

A. Zero.
B. Positive.
C. Negative.

Solution

The correct answer is **A**.

Since the normal curve is symmetric about its mean, its skewness is zero.

B is incorrect because a positively skewed distribution has positive skewness.

C is incorrect because a negatively skewed distribution has negative skewness.

Question 2

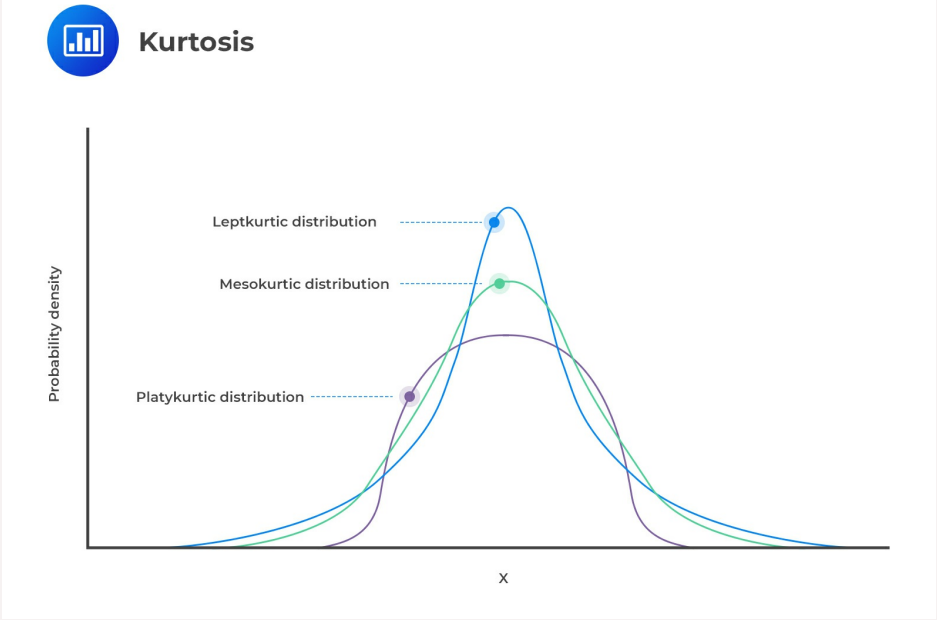
A frequency distribution in which there are too few scores at the extremes of the distribution is *most likely* called:

A. Platykurtic.
B. Leptokurtic.
C. Mesokurtic.

Solution

The correct answer is **A**.

A platykurtic distribution has "thin" tails and is flatter compared to a normal distribution. It implies that there are fewer scores at the extremes of the distribution, which aligns with the question's description.



Question 3

When most of the data are concentrated on the left of the distribution, it is *most likely* called:

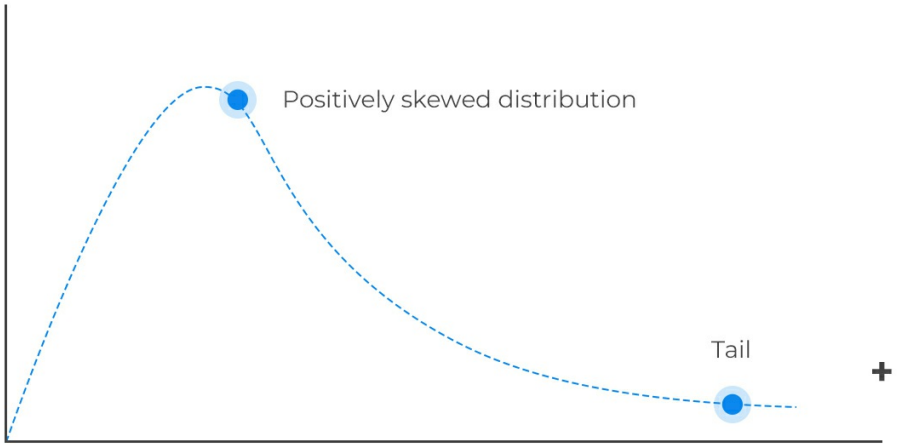
A. Symmetric distribution.
B. Positively skewed distribution.
C. Negatively skewed distribution.

Solution

The correct answer is **B**.

A distribution is said to be skewed to the right or positively skewed when most of the data are concentrated on the left of the distribution. A distribution is said to be skewed to the left or negatively skewed if most of the data are concentrated on the right of the distribution. The left tail clearly extends farther from the distribution's center than the right tail.

Positively Skewed Distribution



A is incorrect. A symmetric distribution is one in which the left and right sides mirror each other.

C is incorrect. A distribution is said to be skewed to the left or negatively skewed if most of the data are concentrated on the right of the distribution. The left tail extends farther away from the mean than the right tail.

LOS 3d: Interpret the correlation between two variables to address an investment problem

Covariance

Covariance is a measure of how two variables move together. The sample covariance of X and Y is calculated as follows:

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

The formula above implies that the sample covariance is the mean of the product of the deviations in the two random variables and from their sample means.

If the covariance between two random variables is positive, it means they move in the same direction. When one is below its mean, the other is also below its mean, and vice versa.

A major drawback of covariance is that it is difficult to interpret since its value can vary from negative infinity to positive infinity.

Correlation

Correlation is a standardized measure of the linear relationship between two variables. It takes the covariance and divides it by the product of the standard deviations of both variables. As a result, its value ranges between -1 and +1 and is easier to interpret.

The sample correlation coefficient is calculated as follows:

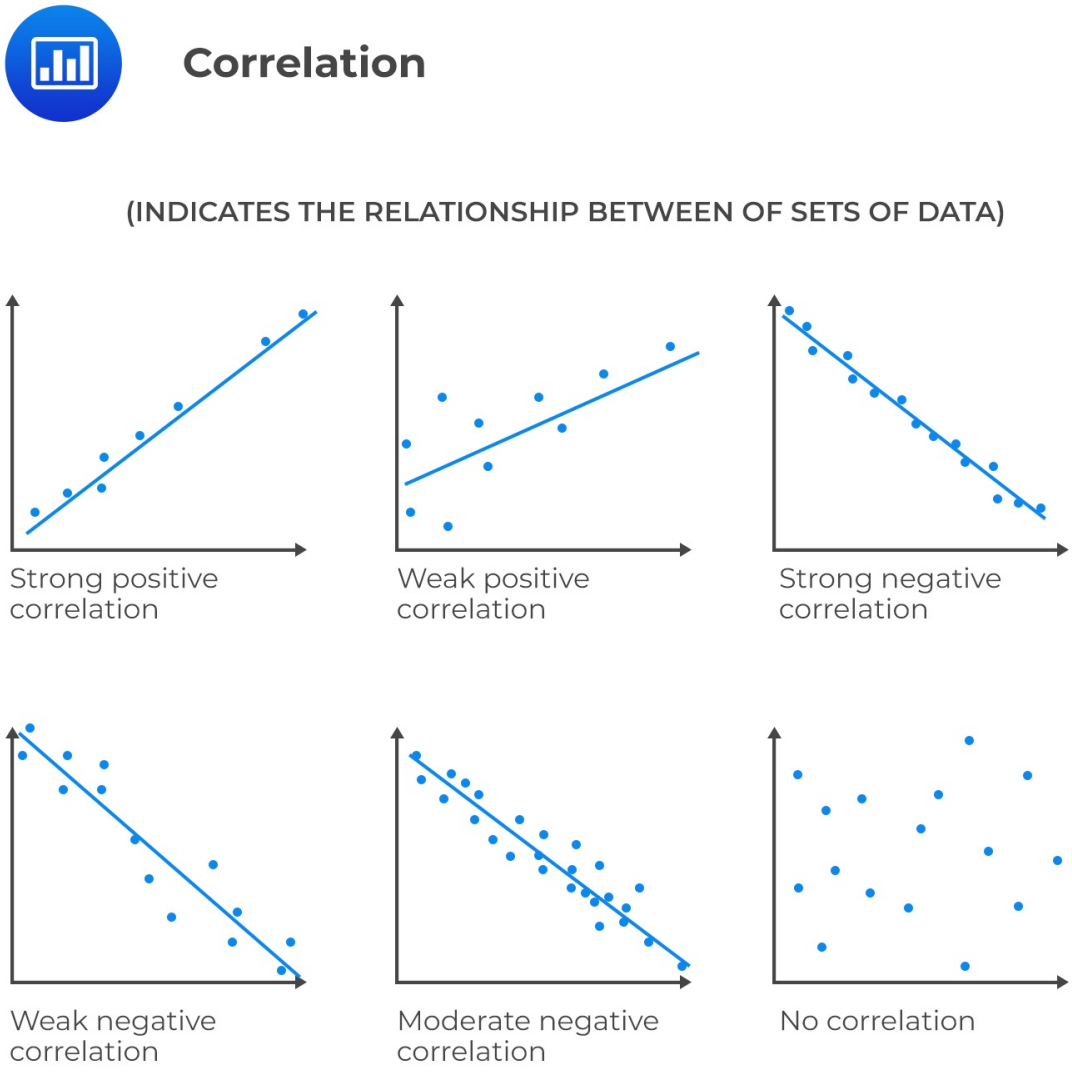
$$r_{xy} = \frac{S_{xy}}{S_x \times S_y}$$

Where:

s_{XY} = Covariance between variable X and Y.

s_X = Standard deviation of variable X.

s_Y = Standard deviation of variable Y.



Properties of Correlation

- Correlation ranges between -1 to +1 for two random variables, X and Y.
- A correlation of 0 (uncorrelated variables) indicates no linear (straight line) relationship exists between the variables.

- A positive correlation close to +1 indicates a strong positive linear relationship.
    - A correlation of 1 indicates a perfect linear relationship.
- A negative correlation close to -1 indicates a strong negative linear relationship.
    - A correlation of -1 indicates a perfect inverse linear relationship.

Limitations of Correlation Analysis

- Two variables can have a very low correlation despite having a strong ***nonlinear*** relationship.
- Correlation can be an unreliable measure when outliers are present in the data.
- Correlation does not imply causation. This implies that the correlation may be spurious. A spurious correlation refers to:
    - Correlation between two variables due to chance relationships in a particular dataset.
    - Correlation arising between variables when they are divided by a third variable.
    - Correlation between two variables arising from their relation to a third variable.

Question

The correlation coefficient between X and Y is 0.7, and the covariance is 29. If the variance of Y is 25, the variance of X is *closest* to:

A. 8.29.

B. 29.

C. 68.65.

Solution

The correct answer is **C**.

$$r_{XY} = \frac{s_{XY}}{s_X \times S_Y}$$
$$\Rightarrow 0.7 = \frac{29}{X \times 5}$$
$$\therefore X = 8.2857$$

$$\text{Variance} = 8.2857^2 = 68.65$$

Learning Module 4: Probability Trees and Conditional Expectations

LOS 4a: calculate expected values, variances, and standard deviations and demonstrate their application to investment problems

Expected Value

Expected value is an essential quantitative concept investors use to estimate investment returns and analyze any factor that may impact their financial position.

Mathematically, the expected value is the probability-weighted average of the possible outcomes of the random variable. For a random variable X, the expected value of X is denoted $E(X)$. More specifically,

$$\begin{aligned} E(X) &= P(X_1)X_1 + P(X_2)X_2 + \cdots + P(X_n)X_n \\ &= \sum_{i=1}^{n} P(X_i)X_i \end{aligned}$$

Where,

X_i= One of n possible outcomes of the discrete random variable X.

$P(X_i) = P(X_i = x_i)$ = Probability of X taking the value x.

Note that the expected can be a forecast (looking into the future) or the true value of the population mean.

The sample mean differs from the expected value. The sample mean is a central value for a specific set of observations, calculated as an equally weighted average of those observations.

Example: Calculating Expected Value

An analyst anticipates the following returns from an asset:

Return	Probability
5%	65%
7%	25%
8%	10%

The expected value of the investment is *closest to*:

Solution

Recall that,

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} P(X_i)X_i \\ &= 0.05 \times 0.65 + 0.07 \times 0.25 + 0.10 \times 0.08 \\ &= 0.0325 + 0.0175 + 0.008 \\ &= 0.058 = 5.8\% \end{aligned}$$

Variance and Standard Deviation

Consider expected value as a forecast of the outcome of an investment. Then, variance and standard deviation measure the risk of an investment. That is the dispersion of outcomes around the mean.

The variance of a random variable is the expected value (the probability-weighted average) of squared deviations from the random variable's expected value. Denoted by $\sigma^2(X)$ or $Var(X)$, its formula is given by:

$$\begin{aligned} Var(X) &= E[X - E(X)]^2 \\ &= P(X_1)[X_1 - E(X)]^2 + P(X_2)[X_2 - E(X)]^2 + \cdots \\ &\quad + P(X_n)[X_n - E(X)]^2 \\ \Rightarrow Var(X) &= \sum_{i=1}^{n} P(X_i)[X_i - E(X)]^2 \end{aligned}$$

Since variance is in squared terms, it can take any number greater than or equal to $0 (Var(X) \geq 0)$. Intuitively, if $Var(X) = 0$, there is no risk (dispersion). On the other hand, if $Var(X) > 0$, it signifies the dispersion of outcomes.

Moreover, $Var(X)$ is a quantity given in square units of X. That is, if the X is given in percentage, then $Var(X)$ is given in squared percentage.

The standard deviation is the square root of variance:

$$\sigma(X) = \sqrt{\sigma^2(X)} = \sqrt{Var(X)}$$

The standard deviation is given in the same units as the random variance; hence it is easy to interpret.

Question

An analyst anticipates the following returns from an asset:

Return	Probability
5%	65%
7%	25%
8%	10%

The variance and standard deviation of the investment are *closest to*:

Solution

We know that,

$$\begin{aligned} Var(X) &= \sum_{i=1}^{n} P(X_i)[X_i - E(X)]^2 \\ &= 0.65(0.05 - 0.058)^2 + 0.25(0.07 - 0.058)^2 \\ &+ 0.10(0.08 - 0.058)^2 \\ &= 0.000126 \end{aligned}$$

For standard deviation,

$$\sigma(X) = \sqrt{Var(X)} = \sqrt{0.000126} = 0.0112$$

LOS 4b: Formulate an investment problem as a probability tree and explain the use of conditional expectations in investment application

A tree diagram is a visual representation of all possible future outcomes and the associated probabilities of a random variable. Tree diagrams are handy when we have several possible outcomes.

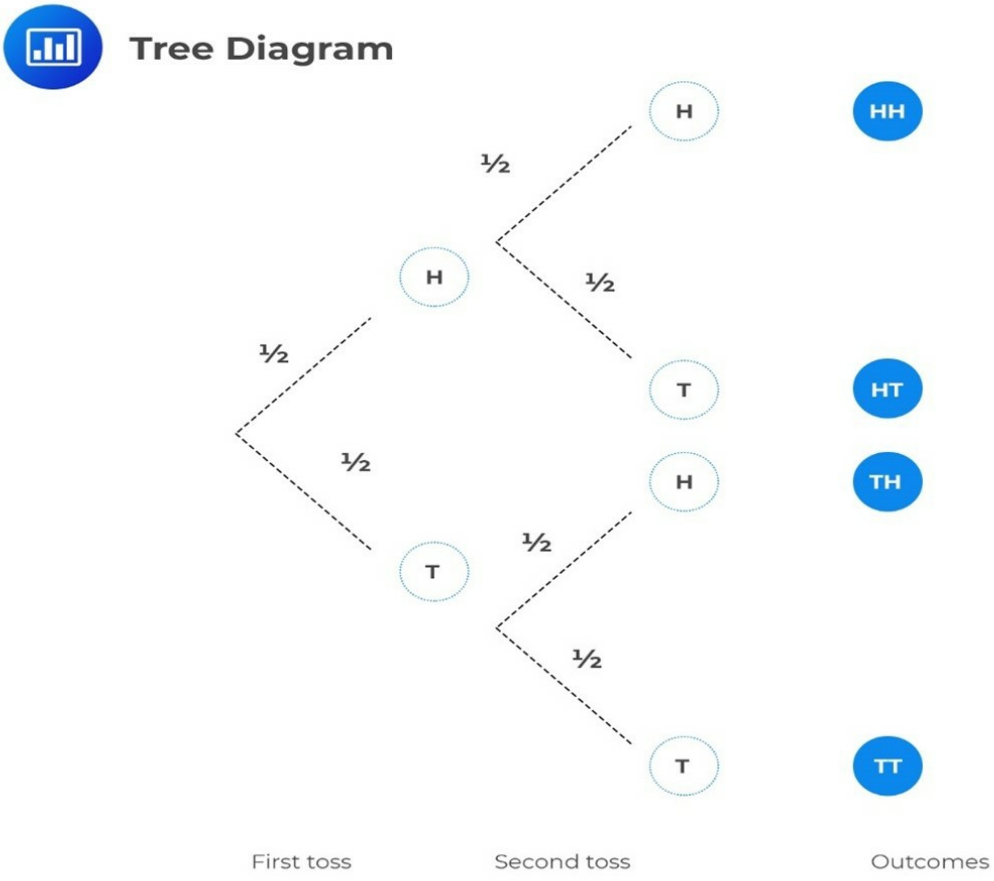
They facilitate the recording of all the possibilities in a clear, uncomplicated manner. Each branch in a tree diagram represents an outcome.

Example: Probability Tree

Let's consider a scenario where we toss a fair coin twice. The outcomes of these tosses are independent, meaning the first toss doesn't influence the second one.

For the first toss, we have two possibilities: It can result in either a head or a tail. Similarly, we still have the same two possibilities for the second toss: head or tail. Importantly, the outcome of the second toss is not affected by what happened in the first toss because coins don't have memory.

We can visualize these probabilities using a tree diagram like this:



Please, note the following:

- The tree diagram must include all the possible outcomes.
- The sum of the probabilities must add up to 1.
- The number of branches is the number of different possibilities.
- The numbers on the branches present probabilities.

To calculate probabilities, we follow the tree branches from left to right and multiply any probabilities we encounter.

So, to find the probability of getting two heads (HH), we multiply the probabilities along the path.

$$P(HH) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

If we sum up the probabilities of all possibilities, we get 1.

Conditional Expectations

In investments, conditional expectation means predicting the expected value of an investment based on specific real-world events. Analysts consider the probability and impact of future events when calculating this value.

Competitors, governments, and other financial institutions keep releasing new information. Such pieces of information may have a positive or a negative impact on investment. This means that a project's expected value must be based on real-world dynamics.

In statistics, the conditional expected value is the expected value of a random variable X given an event or scenario S, denoted by $E(X \mid S)$.

Now, assume that X can take on any of n different outcomes $X_1, X_2, \ldots, X_n$, which are outcomes from a set of mutually exclusive and exhaustive events. Then,

$$\begin{aligned} E(X \mid S) &= X_1 \cdot P(X_1 \mid S) + X_2 \cdot P(X_2 \mid S) + \cdots + X_n \cdot P(X_n \mid S) \\ &= \sum_{i=1}^{n} X_i \cdot P(X_i \mid S) \end{aligned}$$

Stating Unconditional Expected Value in Terms of Conditional Expected Value

To state the unconditional expected values in terms of conditional expected value, we use the total probability rule for expected value, which is built from the following formula:

$$E(X) = P(S) \cdot E(X \mid S) + P(S^C) \cdot E(X \mid S^C)$$

Where S^C is the complement (event or scenario “S” does not occur) of S.

Now assume that S can take on any of $S_1, S_2, \ldots, S_n$ mutually exclusive and exhaustive scenarios or events, then.

$$\begin{aligned} E(X) &= P(S_1) \cdot E(X \mid S_1) + P(S_2) \cdot E(X \mid S_2) + \cdots + P(S_n) \cdot E(X \mid S_n) \\ &= \sum_{i=1}^{n} P(S_i) \cdot E(X \mid S_i) \end{aligned}$$

Example: Conditional Expectation

The probability of relaxed trade restrictions in a given country is 40%. Therefore, shareholders of XYZ Company Limited expect a 5% share return if trade restrictions are maintained and a loss of 8% if they are relaxed. The expected change in return is *closest to*:

Solution

We must take every possibility into account. We have a 40% chance of relaxed trade restrictions in this case. Intuitively, this means there is a 60% chance that the current restrictions will be maintained. Therefore:

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} P(S_i) \cdot E(X \mid S_i) \\ &= 0.6(0.05) \times 0.4(-0.08) \\ &= -0.002 \end{aligned}$$

Example: Total Probability Rule for Expected Value

BlueChip Inc.'s profits are sensitive to economic growth, benefitting significantly during periods of high economic growth. Suppose there is a 0.70 probability that BlueChip Inc. will operate in a high-growth economic environment in the next fiscal year and a 0.30 probability that it will operate in a moderate-growth environment. Assume that the chance of a recession is negligible.

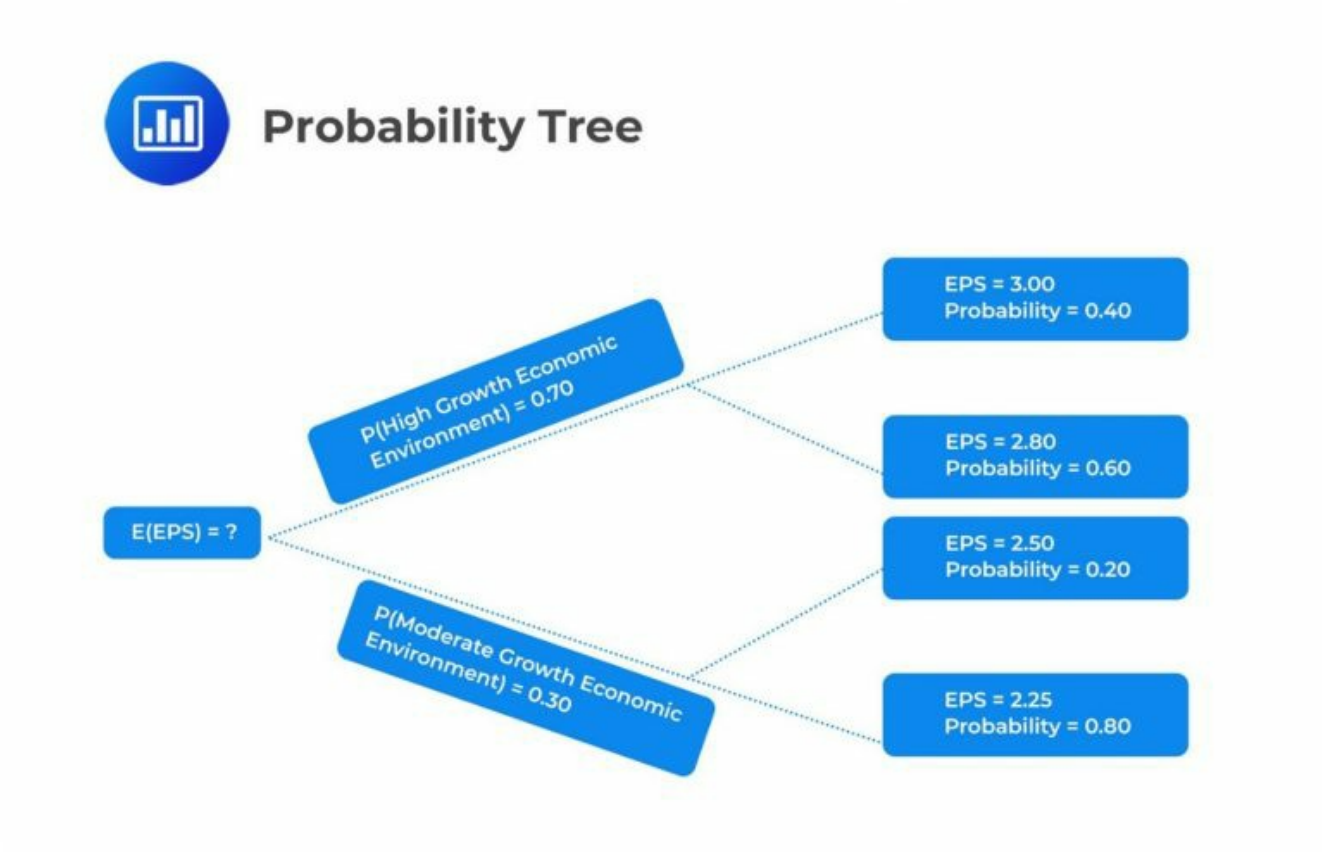
If a high-growth economic environment occurs, the probability that EPS will be USD 3.00 is estimated at 0.40, and the probability that EPS will be USD 2.80 is estimated at 0.60.

On the other hand, if the company operates in a moderate-growth environment, the probabilities that the EPS will be USD 2.50 and USD 2.25 are 20% and 80%, respectively.

Calculate the expected value of EPS for BlueChip Inc. in the next fiscal year.

Solution

We start by drawing a probability tree:



We first need to calculate the conditional expectations of EPS for each scenario: High-growth environment and moderate-growth environment. That is,

$$\begin{aligned} E(EPS \mid \text{High-growth environment}) &= 0.40 \times 3.00 + 0.60 \times 2.80 \\ &= \text{USD } 2.88 \\ E(EPS \mid \text{Moderate-growth environment}) &= 0.20 \times 2.50 + 0.80 \times 2.25 \\ &= \text{USD } 2.30 \end{aligned}$$

Using the total probability for the expected value, we have:

$$
\begin{aligned}
E(EPS) &= P(\text{High-growth environment}) \\
&\quad \cdot E(EPS \mid \text{High-growth environment}) \\
&\quad + P(\text{High-growth environment}) \\
&\quad \cdot E(EPS?\text{Moderate-growth environment}) \\
&\quad = 0.70 \times 2.88 + 0.30 \times 2.30 \\
&\quad = \text{USD } 2.71
\end{aligned}
$$

Question 1

Someone picks a card from an ordinary pack of 52 playing cards **without replacement**. He then picks another one. Draw a probability tree and use it to calculate the probability of picking two red cards.

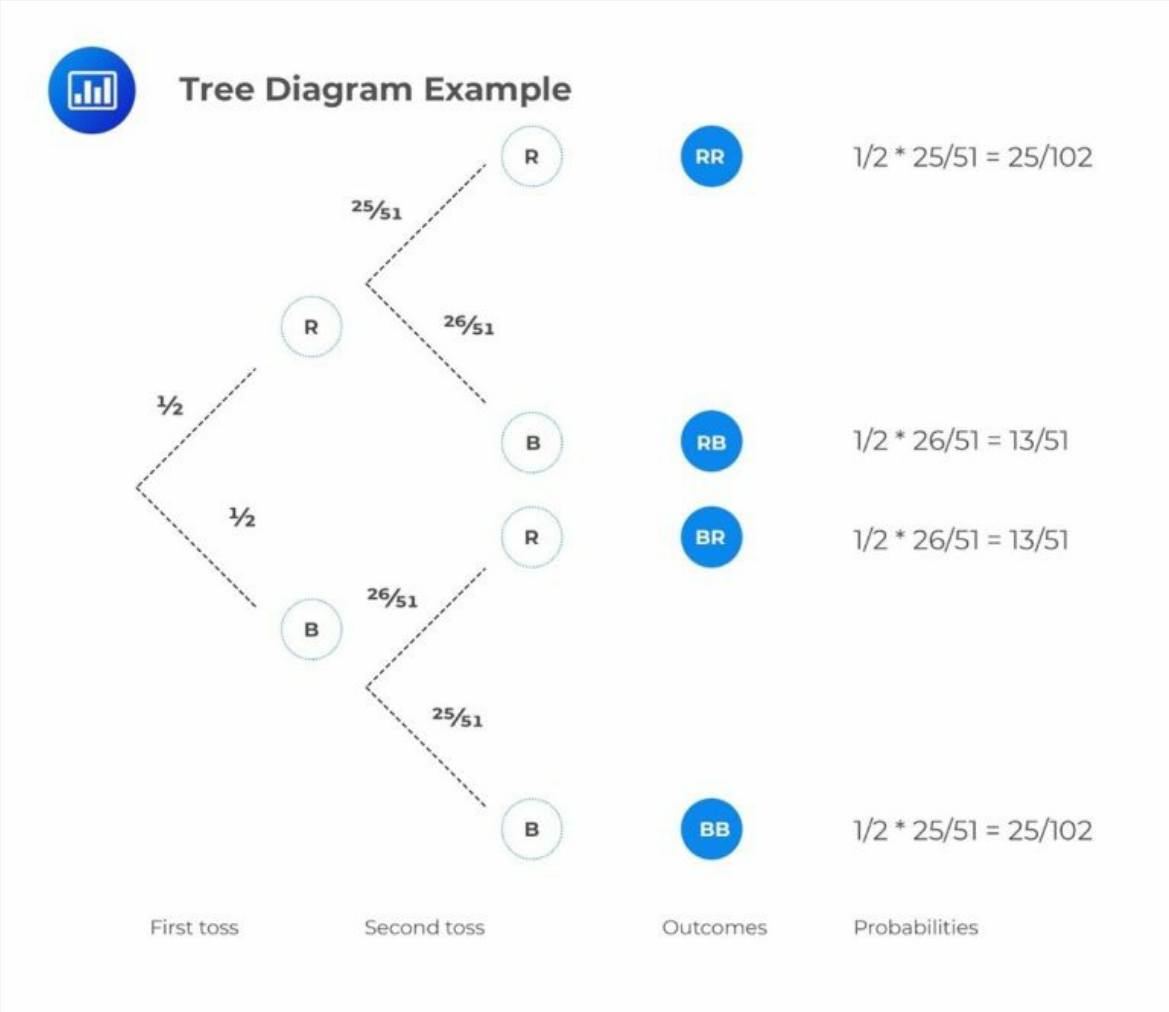
A. $\frac{25}{102}$.

B. $\frac{13}{51}$.

C. $\frac{26}{51}$.

Solution

The correct answer is A.



Question 2

There is a 20% chance that the government will impose a tariff on imported cars. A company that assembles cars locally expects returns of 14% if the tariff is imposed and returns of 11% otherwise. The (unconditional) expected return is *closest to*:

A. 11.6%.

B. 12.8%.

C. 12.5%.

Solution

The correct answer is A.

The unconditional expected return will be the sum of:

1. The expected return **given** no tariff times the probability that a tariff will not be imposed.
2. The expected return **given** tariff times the probability that the tariff will be imposed. Therefore,

$$\begin{aligned} E(X) &= \sum_{i=1}^{n} P(S_i) + E(X \mid S_i) \\ &= 0.11(0.8) + 0.14(0.2) \\ &= 0.116 = 11.6\% \end{aligned}$$

LOS 4c: calculate and interpret an updated probability in an investment setting using Bayes' formula

Investors make investment decisions based on their experience and expertise. Their decisions may change in the wake of new knowledge and observations.

Bayes' formula allows us to update our decisions as we receive new information. In other words, Bayes' formula is used to calculate an updated or posterior probability given a set of prior probabilities for a given event.

Given a set of prior probabilities for an event, if we receive new information, the updated probability is as follows:

$$\text{Updated probability of an event given the new information} = \frac{\text{Probability of the new information gi}}{\text{Unconditional probability of the new}} \times \text{Prior probability of event.}$$

The above equation can be written as:

$$P(\text{Event} \mid \text{Information}) = \frac{P(\text{Information} \mid \text{Event})}{P(\text{Information})} \cdot P(\text{Event})$$

Deriving Bayes' Formula

Let $B_1, B_2, B_3, \dots, B_n$ be a set of mutually exclusive and exhaustive events.

Using the conditional probability:

$$P(B_i \mid A) = \frac{P(B_i \cap A)}{P(A)} \dots\dots (1)$$

And also, the relationship:

$$P(B_i \cap A) = P(A \cap B_i) = P(B_i) \cdot P(A \mid B_i) \dots\dots (2)$$

Also, using the total probability rule:

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i) = \sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i) \ldots \ldots (3)$$

Substituting equations (2) and (3) in (1), we have:

$$P(B_i \mid A) = \frac{(P(A \mid B_i)}{\sum_{i=1}^{n} P(B_i) \cdot P(A \mid B_i)} \cdot P(B_i)$$

This is the Bayes' formula, and it allows us to ‘turnaround’ conditional probabilities, i.e., we can calculate $P(B_i \mid A)$ if given information only about $P(A \mid B_i)$.

Note that:

1. $P(B_i)$ are known as **prior probabilities**.
2. Event A is some event known **to have occurred**.
3. $P(B_i \mid A)$ is the **posterior probability**.

Example: Bayes' Formula

An Investment Analyst wishes to investigate the performance of stocks by considering a number of stocks listed on different exchanges. In the sample, 50% of stocks were listed on the New York Stock Exchange (NYSE), 30% on the London Stock Exchange (LSE), and 20% on the Tokyo Stock Exchange (TSE).

The probability of a stock posting a negative return on the NYSE, LSE, and TSE is 40%, 35%, and 25%, respectively.

If the Analyst picks a stock at random from this group, what is the probability that it has a negative return on the NYSE?

Solution

We are looking for P(NYSE | Negative Return).

Let’s define the following events:

NYSE is the event “A stock chosen at random is listed on the NYSE.”

LSE is the event “A stock chosen at random is listed on the LSE.”

TSE is the event “A stock chosen at random is listed on the TSE.”

Finally, let NR be the event “A randomly chosen stock posts a negative return.”

Therefore,

$$\begin{aligned} P(NYSE|NR) &= \frac{P(NYSE)\,P(NR|NYSE)}{P(NYSE)\,P(NR|NYSE) + P(LSE)\,P(NR|LSE) + P(TSE)\,P(NR|TSE)} \\ &= \frac{0.5 \times 0.4}{0.5 \times 0.4 + 0.3 \times 0.35 + 0.2 \times 0.25} \\ &= \frac{0.2}{0.355} \\ &= 0.5634 \approx 56.3\% \end{aligned}$$

Question

You have developed a set of criteria for assessing potential investments in growth-stage companies. Companies not meeting these criteria are predicted to be insolvent within 24 months. You gathered the following information when validating your criteria:

- Fifty percent of the companies that have been assessed will become insolvent within 24 months: $P(\text{insolvency}) = 0.50$.
- Sixty-five percent of the companies assessed meet the criteria: $P(\text{meet criteria}) = 0.65$.
- The probability that a company will meet the criteria given that it remains solvent for 24 months is 0.80: $P(\text{meet criteria} \mid \text{solvency}) = 0.80$.

The probability that a company will remain solvent, given that it meets the criteria, that is, $P(\text{solvency} \mid \text{meet criteria})$, is *closest to*:

A. 20%.
B. 50%.
C. 62%.

Solution

Using Bayes' formula, we have:

$$\begin{aligned} &P(\text{solvency} \mid \text{meet criteria}) \\ &= \frac{P(\text{meet criteria} \mid \text{solvency})P(\text{solvency})}{[P(\text{meet criteria} \mid \text{solvency})P(\text{solvency}) \\ &\quad + P(\text{meet criteria} \mid \text{insolvency})P(\text{insolvency})]} \\ &= \frac{0.80 \times 0.50}{0.80 \times 0.50 + P(\text{meet criteria} \mid \text{insolvency}) \times 0.50} \end{aligned}$$

Clearly, we need to calculate the $P(\text{meet criteria} \mid \text{insolvency})$. Using the total probability:

$$\begin{aligned} P(\text{meet criteria}) &= P(\text{meet criteria} \mid \text{solvency})P(\text{solvency}) \\ &\quad + P(\text{meet criteria} \mid \text{insolvency})P(\text{insolvency}) \\ \Rightarrow 0.65 &= 0.80 \times 0.50 + P(\text{meet criteria} \mid \text{insolvency}) \times 0.50 \end{aligned}$$

$$\therefore P(\text{meet criteria} \mid \text{insolvency}) = \frac{0.65 - 0.80 \times 0.50}{0.50} = 0.50$$

As such,

$$P(\text{solvency} \mid \text{meet criteria}) = \frac{0.80 \times 0.50}{0.80 \times 0.50 + 0.50 \times 0.50} = 0.6153 \approx 62\%$$

Learning Module 5: Portfolio Mathematics

LOS 5a: calculate and interpret the expected value, variance, standard deviation, covariances, and correlations of portfolio returns

A portfolio is a collection of investments a company, mutual fund, or individual investor holds. A portfolio consists of assets such as stocks, bonds, or cash equivalents. Financial professionals usually manage a portfolio.

Portfolio Expected Return

To calculate the portfolio's expected return, you take the expected returns of each security in the portfolio. Then, you multiply each security's expected return by its proportion in the portfolio and add them up. The formula below helps you find the portfolio's expected return:

$$E(R_p) = w_1E(R_1) + w_2E(R_2) + \cdots w_nE(R_n)$$

Where:

$w_1, w_2, \ldots, w_n$ = Weights (market value of asset/market value of the portfolio) attached to assets $1, 2, \ldots, n$.
$R_1, R_2, \ldots, R_n$ = Expected returns for assets $1, 2, \ldots, n$.

Example: Portfolio Expected Return

Assume we have a simple portfolio of two mutual funds, one invested in bonds and the other invested in stocks. Let us further assume that we expect a stock return of 8% and a bond return of 6%, and our allocation is equal in both funds. The expected return would be calculated as follows:

$$E(Rp) = (0.5 \times 0.08) + (0.5 \times 0.06) = 0.07 \text{ or } 7\%$$

Portfolio Variance

The variance of a portfolio's return is a function of the individual asset covariances as well as the

covariance between each of them.

Consider a portfolio with three assets: A, B, and C. The portfolio variance is given by:

$$\begin{aligned}&\text{Portfolio Variance}\\&\quad = W_A^2\sigma^2(R_A) + W_B^2\sigma^2(R_B) + W_C^2\sigma^2(R_C) + 2(W_A)(W_B)\text{Cov}(R_A, R_B)\\&\quad + 2(W_A)(W_C)\text{Cov}(R_A, R_C) + 2(W_B)(W_C)\text{Cov}(R_B, R_C)\end{aligned}$$

If we have two assets, A and B, then:

$$\text{Portfolio Variance} = W_A^2\sigma^2(R_A) + W_B^2\sigma^2(R_B) + 2(W_A)(W_B)\text{Cov}(R_A, R_B)$$

Where:

W_A = Weight of assets A in the portfolio.

W_B = Weight of assets B in the portfolio.

$\sigma^2(R_A)$ = Variance of the returns on assets A.

$\sigma^2(R_B)$ = Variance of the returns on assets B.

Portfolio variance is a measure of risk. The higher the variance, the higher the risk. Investors usually reduce the portfolio variance by choosing assets with low or negative covariance, e.g., stocks and bonds.

Portfolio Standard Deviation

Portfolio standard deviation is simply the square root of the portfolio variance. It is a measure of the riskiness of a portfolio.

Considering a portfolio with two assets, A and B, the portfolio standard deviation is given by:

$$\text{Standard deviation} = \sqrt{(W_A^2\sigma^2(R_A) + W_B^2\sigma^2(R_B) + 2(W_A)(W_B)\text{Cov}(R_A, R_B)}$$

Covariance

Covariance is a measure of the degree of co-movement between two random variables. The general formula used to calculate the covariance between two random variables, X and Y is:

$$\text{Cov}(X, Y) = \sigma(X, Y) = E[(X - E[X])(Y - E[Y])$$

Where:

$\text{Cov}(X, Y)$ = Covariance of X and Y.

$E[X]$ = Expected value of the random variable X.

$E[Y]$ = Expected values of the random variable Y.

This formula calculates the population covariance. It does this by taking the probability-weighted average of the cross-products of the random variables' deviations from their expected values for every possible outcome.

Sample Covariance

The sample covariance between two variables, X and Y, based on a sample data of size n is:

$$\text{Cov}(X, Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Where:

$\bar{X}$ = Sample mean of X.

$\bar{Y}$ = Sample mean of Y.

X_i and Y_i = i-th data points of X and Y, respectively.

The covariance between two random variables can be positive, negative, or zero.

- A positive number indicates co-movement. The variables tend to move in the same direction.
- A value of zero indicates no relationship.

- A negative value shows that the variables move in opposite directions.

Covariance Matrix

A covariance matrix displays a complete list of covariances between assets needed to calculate the portfolio variance. Consider a portfolio with three assets A, B, and C. The covariance matrix is as follows:

Asset	A	B	C
A	$Cov(R_A, R_A)$	$Cov(R_A, R_B)$	$Cov(R_A, R_C)$
B	$Cov(R_B, R_A)$	$Cov(R_B, R_B)$	$Cov(R_B, R_C)$
C	$Cov(R_C, R_A)$	$Cov(R_C, R_B)$	$Cov(R_C, R_C)$

The off-diagonal (bolded) terms represent variances since, for example:

$$Cov(R_A, R_A) = \rho(A, A)\sigma_A\sigma_A = 1?\sigma_A^2 = \sigma_A^2$$

As such, the table above transforms:

Asset	A	B	C
A	σ_A^2	$Cov(R_A, R_B)$	$Cov(R_A, R_C)$
B	$Cov(R_B, R_A)$	σ_B^2	$Cov(R_B, R_C)$
C	$Cov(R_C, R_A)$	$Cov(R_C, R_B)$	σ_C^2

Intuitively, a three-asset portfolio would have $3 \times 3 = 9$ entries of covariances. However, we do not count the off-diagonal terms since they contain the individual variances of the assets. As such, we have $6(= 9 - 3)$ covariances.

Note that:

$$\begin{aligned} Cov(R_B, R_A) &= Cov(R_A, R_B) \\ Cov(R_A, R_C) &= Cov(R_A, R_C) \\ Cov(R_C, R_B) &= Cov(R_B, R_C) \end{aligned}$$

Therefore, there are $\frac{6}{2} = 3$ distinct covariance terms in the above covariance matrix.

In general, if we have n securities in a portfolio, there are $\frac{n(n-1)}{2}$ distinct covariances and n

variances to estimate.

Correlation

Correlation is the covariance ratio between two random variables and the product of their two standard deviations. The correlation formula for random variables X and Y is:

$$\begin{aligned}\text{Correlation (X,Y)} &= \text{Corr}(X,Y) = \rho(X,Y) \\ &= \frac{\text{Cov}(X,Y)}{\text{Standard deviation}(X) \times \text{Standard deviation}(Y)} \\ &= \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}\end{aligned}$$

Correlation measures the strength of the linear relationship between two variables. While the covariance can take on any value between negative infinity and positive infinity, the correlation is always between -1 and +1:

- +1 indicates a perfect linear relationship (i.e., the two variables move in the same direction with equal unit changes).
- Zero indicates no linear relationship at all.
- -1 indicates a perfect inverse relationship, i.e., a unit change in one means that the other will have a unit change in the opposite direction.

Example: Calculating Correlation Coefficient from the Covariance Matrix #1

Harrison is a portfolio manager who oversees three assets: A, B, and C. The covariance matrix of these assets is shown below:

Asset	A	B	C
A	0.04	0.02	0.01
B	0.02	0.05	0.015
C	0.01	0.015	0.09

Using this information, what is the correlation coefficient between assets B and C?

Solution

Note:

$$\text{Correlation}(B, C) = \frac{\text{Cov}(B, C)}{\sigma_B \sigma_C} = \frac{0.015}{\sqrt{0.05 \times 0.09}} = 0.224$$

Example: Calculating the Correlation Coefficient #2

We expect a 15% chance that ABC Corp's stock returns for the next year will be 6%. There's a 60% probability that they will be 8% and a 25% probability of a 10% return. The expected return is 8.2%, and the standard deviation is 1.249%.

We also anticipate that the same probabilities and states are associated with a 4%, 5%, and 5.5% return for XYZ Corp. The expected value of returns is then 4.975%, and the standard deviation is 0.46%.

To calculate the covariance and the correlation between ABC and XYZ returns, then:

$$\begin{aligned}\text{Cov}(R_{ABC}, R_{XYZ}) &= 0.15(0.06 - 0.082)(0.04 - 0.04975) \\ &+ 0.6(0.08 - 0.082)(0.05 - 0.04975) \\ &+ 0.25(0.10 - 0.082)(0.055 - 0.04975) \\ &= 0.0000561\end{aligned}$$

$$\text{Correlation(Ri, Rj)} = \frac{\text{Covariance}(R_{ABC}, R_{XYZ})}{\text{Standard deviation(RABC)} \times \text{Standard deviation(RXYZ)}}$$

Therefore:

$$\text{Correlation} = \frac{0.0000561}{(0.01249 \times 0.0046)} = 0.976$$

The correlation between the returns of the two companies is very strong (almost +1), and the returns move linearly in the same direction.

Example: Calculating Correlation Coefficient #3

An analyst studied five years of historical data to examine how changes in Central Bank interest

rates affect the country's inflation rate. The covariance between the interest rate and inflation rate is -0.00075. The standard deviation of the interest rate is 5.5%, and the inflation rate is 12%. Now, let's calculate and interpret the correlation between these two variables.

Solution

$$\text{Correlation}_{\text{Interest rate, Inflation}} = \frac{\text{Covariance}_{\text{Interest Rate, Inflation}}}{\text{Standard deviation}_{\text{Interest rate}} \times \text{Standard deviation}_{\text{Inflation}}}$$

$$\text{Correlation}_{\text{Interest rate, Inflation}} = \frac{-0.00075}{(0.055 \times 0.12)} = -0.11364$$

A correlation of -0.11364 indicates a negative correlation between the interest rate and the inflation rate.

Note that if we consider, say, assets A and B, then:

$$\text{Corr}(A, B) = \rho(A, B) = \frac{\text{Cov}(A, B)}{\sigma_A \sigma_B}$$

$$\Rightarrow \text{Cov}(A, B) = \sigma_A \sigma_B \rho(A, B)$$

Consequently, in the formula for calculating portfolio variance, consisting of two assets, A and B, we substitute for Cov(A, B) so that:

$$\text{Portfolio Variance} = W_A^2 \sigma^2(R_A) + W_B^2 \sigma^2(R_B) + 2(W_A)(W_B)\sigma_A \sigma_B \rho(A, B)$$

Question

Assume that we have investments in two companies, ABC and XYZ. For ABC, there's a 15% chance of a 6% return, a 60% chance of an 8% return, and a 25% chance of a 10% return. The expected return for ABC is 8.2%, and the standard deviation is 1.249%. For XYZ, there are similar probabilities of 4%, 5%, and 5.5% returns. The expected return for XYZ is 4.975%, and the standard deviation is 0.46%.

Assuming equal weights, the portfolio standard deviation is *closest to*:

A. 0.0000561.

B. 0.00007234.

C. 0.00851.

The correct answer is C.

$$\text{Portfolio Variance} = W_A^2\sigma^2(R_A) + W_B^2\sigma^2(R_B) + 2(W_A)(W_B)\text{Cov}(R_A, R_B)$$

First, we must calculate the covariance between the two stocks:

$$\begin{aligned}\text{Cov}(R_{ABC}, R_{XYZ}) &= 0.15(0.06 - 0.082)(0.04 - 0.04975) \\ &+ 0.6(0.08 - 0.082)(0.05 - 0.04975) \\ &+ 0.25(0.10 - 0.082)(0.055 - 0.04975) \\ &= 0.0000561\end{aligned}$$

Since we already have the weight and the standard deviation of each asset, we can proceed and calculate the portfolio variance:

$$\begin{aligned}\text{Portfolio variance} &= 0.5^2 \times 0.01249^2 + 0.5^2 \times 0.0046^2 \\ &+ 2 \times 0.5 \times 0.5 \times 0.0000561 \\ &= 0.00007234\end{aligned}$$

Therefore, the standard deviation is:

$$\sqrt{0.00007234} = 0.00851$$

LOS 5b: Calculate and interpret the covariance and correlation of portfolio returns using a joint probability function for returns

Historical covariance or other techniques, such as market model regression with historical return data, can help us forecast return covariance and correlation. We use the joint probability function of the random variables for this estimation.

The probability that values of the two random variables X and Y will occur simultaneously is given by the joint probability function of X and Y, denoted as $P(X, Y)$. For instance, $P(X = 3, Y = 4)$ represents the likelihood that X and Y will be equal to 3 and 4, respectively.

Covariance can be defined as a probability-weighted average of the cross-products of each random variable's deviation from its own expected value. That is

$$\begin{aligned} Cov_(X_i Y_i) &= E[(X_i - \bar{X})(Y_j - \bar{Y})] \\ &= \sum_i \sum_j P(X = x_i, Y = y_j)(X_i - \bar{X})(Y_j - \bar{Y}) \end{aligned}$$

This formula calculates the covariance between random variables X and Y, such as portfolio returns.

To find it, we take the sum of the products of the deviations of X and Y from their expected values for all possible outcomes.

Each product is weighted by the probability of that specific outcome occurring.

Independence and Correlation

Two random variables, X and Y, are independent if $P(X, Y) = P(X) \cdot P(Y)$. That is, X and Y are independent. We find the product of independent probability to calculate joint probability.

The independence property is stronger than correlation because the correlation coefficient addresses linear relationships.

If random variables X and Y are uncorrelated (also holds for independent random variables),

then:

$$E(XY) = E(X) \cdot E(Y)$$

Example: Calculating the Covariance #1

Suppose we wish to find the variance of each asset and the covariance between the returns of ABC and XYZ, given that the amount invested in each company is $1,000.

This table is used to calculate the expected returns:

	Strong Economy	Normal Economy	Weak Economy
Probability	15%	60%	25%
ABC Returns	40%	20%	0%
XYZ Returns	20%	15%	4%

Solution

For us to find the covariance, we must calculate the expected return of each asset as well as their variances. The assets' weights are:

$$W_{ABC} = \frac{1000}{2000} = 0.5$$
$$W_{XYZ} = \frac{1000}{2000} = 0.5$$

Next, we should calculate the individual expected returns:

$$E(R_{ABC}) = 0.15 \times 0.40 + 0.60 \times 0.2 + 0.25 \times 0.00 = 0.18$$
$$E(R_{XYZ}) = 0.15 \times 0.2 + 0.60 \times 0.15 + 0.25 \times 0.04 = 0.13$$

Finally, we can compute the covariance between the returns of the two assets:

$$\begin{aligned} Cov(R_{ABC,XYZ}) &= 0.15(0.40 - 0.18)(0.20 - 0.13) \\ &+ 0.6(0.20 - 0.18)(0.15 - 0.13) \\ &+ 0.25(0.00 - 0.18)(0.04 - 0.13) \\ &= 0.0066 \end{aligned}$$

Example: Calculating the Covariance #2

A portfolio manager is considering the following two possible economic growth of a country and the joint variability of returns on two stocks in a portfolio:

Economic Growth	< 4%	> 4%
Probability	40%	60%
Return of Stock A	2.3%	8%
Return of Stock B	6.5%	3%

What is the covariance between the return of Stock A and Stock B?

Solution

$$\text{Expected return of Stock A} = (40\% \times 2.3\%) + (60\% \times 8\%) = 5.72\%$$
$$\text{Expected return of Stock B} = (40\% \times 6.5\%) + (60\% \times 3\%) = 4.40\%$$

Note: For the rest of the calculation, your curriculum sometimes ditches the percentage signs so that 4.40% becomes simply 4.40.

The deviations of returns at the economic growth of

$$< 4\% = (2.3 - 5.72) \times (6.5 - 4.40) = ?7.182$$

The deviations of returns at the economic growth of

$$> 4\% = (8 - 5.72) \times (3 - 4.40) = ?3.192$$

The covariance of returns between stock A and stock B is computed as follows:

$$\text{Cov}(R_{A,B}) = (-7.182 \times 0.40) + (-3.192 \times 0.60) = -4.788$$

Since covariance is negative, the two returns show some co-movement in opposite signs.

Question

The following table represents the estimated returns for two motor vehicle production brands – TY and Ford, in 3 industrial environments: strong (50% probability), average (30% probability), and weak (20% probability).

	TY Returns +6%	TY Returns +3%	Y Returns −1%
Ford Sales + 10%	Strong(0.5)		
Ford Sales + 4%		Average(0.3)	
Ford Sales − 4%			Weak(0.2)

Given the above joint probability function, the covariance between TY and Ford returns is *closest to*:

A. 0.054.

B. 0.1542.

C. 0.1442.

Solution

The correct answer is C.

First, we must start by calculating the expected return for each brand:

The expected return for TY:

$$= (0.5 \times 6\%) + (0.3 \times 3\%) + (0.2 \times (-1\%))$$
$$= 3\% + 0.9\% - 0.2\% = 3.7\%$$

The expected return for Ford:

$$= (0.5 \times 10\%) + (0.3 \times 4\%) + (0.2 \times (-4\%))$$
$$= 5\% + 1.2\% - 0.8\% = 5.4\%$$

Next, we can now compute the covariance:

$$\begin{aligned}\text{Covariance} &= 0.5(6\% - 3.7\%)(10\% - 5.4\%) \\ &+ 0.3(3\% - 3.7\%)(4\% - 5.4\%) \\ &+ 0.2(-1\% - 3.7\%)(-4\% - 5.4\%) \\ &= 5.29\% + 0.294\% + 8.836\% \\ &= 0.1442\end{aligned}$$

The covariance is positive. This means that the returns for the two brands show some co-movement in the same direction.

In real life, this scenario is highly likely because the companies belong to the same industry. As a result, they share similar systematic risks.

LOS 5c: define shortfall risk, calculate the safety-first ratio, and identify an optimal portfolio using Roy’s safety-first criterion

Modern Portfolio Theory (MPT) evaluates investment options based on mean return and return variance. This approach is applicable when investors are risk-averse, meaning they seek to maximize their expected satisfaction or utility from their investments.

Mean-return analysis holds under two assumptions:

i. Returns follow a normal distribution.
ii. Investors have quadratic utility functions, a mathematical model representing the balance between risk and return.

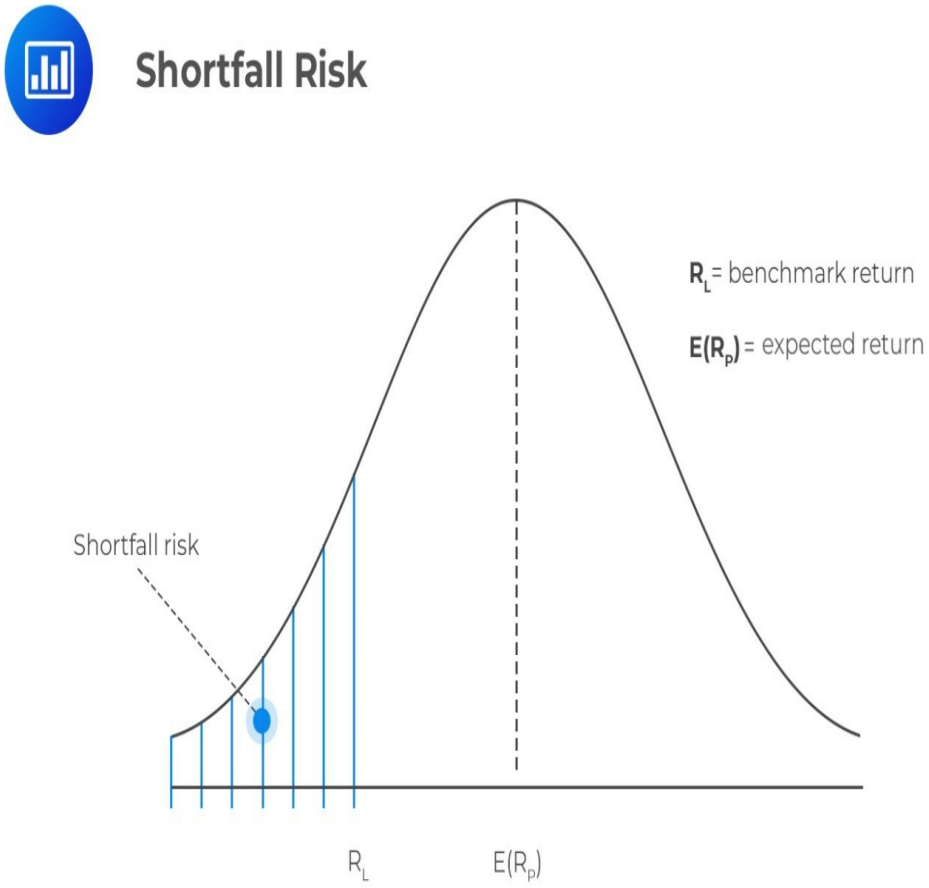
The mean-variance analysis can be reasonably accurate even if the two assumptions aren't entirely met. Professionals prefer using observable data, such as returns. The assumption that returns roughly follow a normal distribution has played a crucial role in applying MPT.

Mean-variance analysis only considers risk symmetrically. This implies that standard deviation reflects variability above and below the mean. An alternative strategy is focusing on downside risk. One such method is safety-first rules.

Before we dive into safety-first rules, we discuss Shortfall risk.

Shortfall Risk

Shortfall risk refers to the probability that a portfolio will not exceed the minimum (benchmark) return an investor sets. In other words, it is the risk that a portfolio will fall short of the level of return considered acceptable by an investor. As such, shortfall risks are downside risks. While a shortfall risk focuses on the downside economic risk, the standard deviation measures the overall volatility of a financial asset.



An illustration of shortfall risk

Safety-First Ratio

Roy's safety-first criterion states that the optimal portfolio is the one that minimizes the probability that a portfolio return, denoted by R_P, may fall below the threshold level of return, R_L. The optimal portfolio minimizes $P(R_P < R_L)$

As such, if returns are distributed normally, the optimal portfolio is the one with the highest safety-first ratio defined as:

$$\text{SFRatio} = \frac{E(R_P) - R_L}{\sigma_P}$$

The numerator, $E(R_P - R_L)$, represents the distance from the mean return to the threshold level, i.e., it measures the excess return over and above the threshold level of return per unit risk.

Intuitively, if the returns are normally distributed, the safety-first optimal portfolio maximizes the SFRatio.

Given a portfolio SFRatio, the probability that its return will be less than R_L is:

$$P(R_P < R_L) = N(-\text{SFRatio})$$

The safety-first optimal portfolio has the lowest $P(R_P < R_L)$.

Example: Safety-first Ratio

An investor sets a minimum threshold of 3%. There are three portfolios from which he is to choose one. The expected return and the standard deviation for each portfolio are given below:

	Portfolio A	Portfolio B	Portfolio C
Expected return	5%	10%	20%
Standard deviation	15%	20%	25%

What is the optimal portfolio for the investor?

Solution

Compute the safety-first ratio for each of the three portfolios and then compare them.

For portfolio A:

$$\text{SFRatio}_A = \frac{5 - 3}{15} = 0.1333$$

Similarly, for portfolio B:

$$\text{SFRatio}_B = \frac{10 - 3}{20} = 0.35$$

Lastly:

$$\text{SFRatio}_C = \frac{20 - 3}{25} = 0.68$$

The optimal portfolio should maximize the safety-first ratio. Comparing the three ratios, it is easy to notice that the safety-first ratio for portfolio C is the highest. Therefore, the investor should choose portfolio C.

Question

The returns on a fund are distributed normally. At the end of year t, the fund has a value of \$100,000. At the end of year $t+1$, the fund manager wishes to withdraw \$10,000 for further funding but is reluctant to tap into the \$100,000. There are two investment options:

	Portfolio A	Portfolio B
Expected return	14%	13%
Standard deviation	17%	20%

Which portfolio is preferable for the manager?

A. Portfolio A.
B. Portfolio B.
C. The manager is indifferent to the two portfolios.

Solution

The correct answer is A.

First, you should calculate the threshold return from the information given. Since there should be no tapping into the fund, the threshold return is:

$$\frac{10,000}{100,000} = 10\% \text{ or } 0.1$$

You should then calculate the safety-first ratio for each portfolio:

$$\text{SFRatio}_A = \frac{14-10}{17} = 0.24$$
$$\text{SFRatio}_B = \frac{13-10}{20} = 0.15$$

Portfolio A has the highest safety-first ratio. This is the reason it is the most desirable.

You can also go a step further and calculate $P(R_P < R_L)$. To do this, you would have to

negate each safety-first ratio and then find the CDF of the standard normal distribution for the resulting value. That is,

$$\begin{aligned} P(R_P < R_L) &= N(-\text{SFRatio}) \\ N(-0.24) &= 1 - N(0.24) \\ &= 1 - 0.5948 = 0.4052 \\ N(-0.15) &= 1 - N(0.15) \\ &= 1 - 0.5596 = 0.4404 \end{aligned}$$

(−where SFRatio is the z-value)

For portfolio A, there is approximately a 40% probability of obtaining a return below the threshold return. For portfolio B, this probability rises to 44%. Therefore, we choose the option for which the chance of not exceeding the benchmark return is lowest – portfolio A.

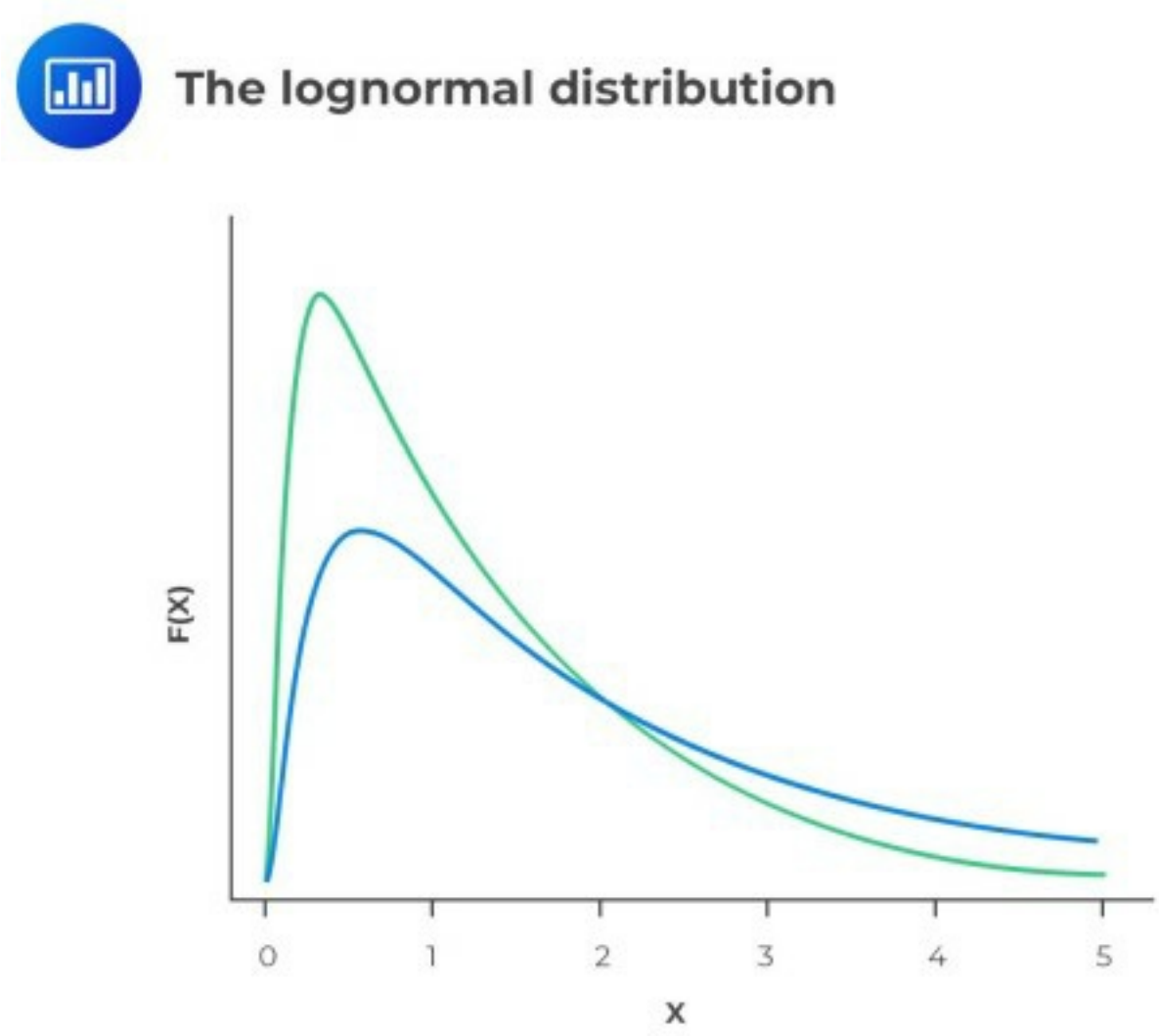
Learning Module 6: Simulation Methods

LOS 6a: explain the relationship between normal and lognormal distributions and why the lognormal distribution is used to model asset prices when using continuously compounded asset returns

A random variable Y is lognormally distributed if its natural logarithm, $\ln Y$, is normally distributed. The opposite is true. If $\ln Y$ is normally distributed, then Y is lognormally distributed.

The lognormal distribution is positively skewed, meaning it's skewed to the right and has a long right tail. In this distribution, values are bounded by 0. Typically, the mean is greater than the mode.

Consider the following graph of two probability density functions (pdfs) of two lognormal distributions.



Like the normal distribution, two parameters – the mean and variance of the associated normal distribution – fully describe the lognormal distribution.

Expressions for Mean and Variance of Lognormal Distribution

Assume that X is normally distributed with the mean μ and variance σ^2. Also, define the variable $Y = e^X$.

Then $\ln Y = \ln(e^X) = X$ is lognormally distributed with the following mean and variance expressions:

$$\text{Mean} = \mu_L = e^{(\mu + \frac{1}{2}\sigma^2)}$$

$$\text{Variance} = \sigma_L^2 = e^{2u+\sigma^2}\,(e^{\sigma^2} - 1)$$

Why the Lognormal Distribution is Used to Model Stock Prices

The lognormal distribution works well for modeling asset prices that cannot be negative because it has a lower bound at zero.

When the continuously compounded returns on a stock follow a normal distribution, the stock prices follow a lognormal distribution. Note that even if returns do not follow a normal distribution, the lognormal distribution is still the most appropriate for stock prices.

Continuously Compounded Rate of Return

Remember that given the investment horizon from time $t = 0$ to time $t = T$, the continuously compounded return of a stock is given by:

$$r_{0,T} = \ln\left(\frac{P_T}{P_0}\right)$$

If we apply the exponential function on both sides of the equation, we have the following:

$$P_T = P_0 e^{r_{0,T}}$$

Note that $\frac{P_T}{P_0}$ can be written as:

$$\frac{P_T}{P_0} = \left(\frac{P_T}{P_{T-1}}\right)\left(\frac{P_{T-1}}{P_{T-2}}\right)\ldots\left(\frac{P_1}{P_0}\right)$$

If we take natural logarithm on both sides of the above equation:

$$\ln\left(\frac{P_T}{P_0}\right) = \ln\left(\left(\frac{P_T}{P_{T-1}}\right)\left(\frac{P_{T-1}}{P_{T-2}}\right)\ldots\left(\frac{P_1}{P_0}\right)\right)$$
$$\Rightarrow r_{0,T} = r_{T-1,T} + r_{T-2,T-1} + \ldots + r_{0,1}$$

Therefore, the continuously compounded return to time T equals the sum of one-period continuously compounded returns.

Remember that a linear combination of normal random variables is also normal. Therefore, if the shorter period returns $r_{T-1,T}, r_{T-2,T-1}, \ldots, r_{0,1}$ are normally distributed or approximately normal, the $r_{0,T}$ is approximately normal.

As such, if we assume that the one-period continuously compounded returns $r_{T-1,T}, r_{T-2,T-1}, \ldots, r_{0,1}$ are independently and identically distributed (i.i.d) random variables, the mean of μ and variance of σ^2, then:

- The expected value of the continuously compounded return over a holding period of T periods is given by:

$$E(r_{0,T}) = E(r_{T-1,T}) + E(r_{T-2,T-1}) + \ldots + E(r_{0,1}) = \mu T$$

- The variance of the continuously compounded return over a holding period is given by:

$$\sigma^2(r_{0,T}) = \sigma^2 T$$

The standard deviation of the continuously compounded returns, also known as volatility, is given by:

$$\sigma(r_{0,T}) = \sigma\sqrt{T}$$

In other words, if $r_{T-1,T}, r_{T-2,T-1}, \ldots, r_{0,1}$ are normally distributed with the mean of μ and variance of σ^2 then $r_{0,T}$ is normally distributed with the mean of μT and variance of $\sigma^2 T$.

Let us go back to the formula:

$$P_T = P_0 e^{r_{0,T}}$$

If X is normally distributed with the mean μ and variance σ^2 and that $Y = e^X$ then, $\ln Y = \ln(e^X) = X$ is lognormally distributed. Assume we apply this intuition in the above formula. In that case, it is easy to see that we can model P_T as a lognormally distributed random variable

since $r_{0,T}$ is approximately normally distributed.

Volatility and Continuously Compounded Returns

Volatility measures the standard deviation of the continuously compounded returns on the underlying asset. Conventionally, it is usually annualized.

We calculate volatility using the historical series of continuously compounded returns. Another method is converting daily holding returns into continuously compounded daily returns and then calculating annualized volatility.

We base annualizing volatility on 250 trading days in a year, which is an estimate of the business days the financial markets operate. The formula we use for annualizing volatility is:

$$\sigma(r_{0,T}) = \sigma\sqrt{T}$$

For example, if the daily volatility is 0.05, then the annual volatility is:

$$\sigma(r_{0,T}) = 0.05 \times \sqrt{250} = 0.79$$

Example: Lognormal Distribution and Continuous Compounding

Jess Kasuku is analyzing the stock of ABC Company, which is listed on the London Stock Exchange under the ABC ticker symbol. Kasuku wants to understand how the stock's price changed during a particular week when significant developments in the global economy impacted the UK stock market. To do this, she calculates the stock's volatility for that week using the closing prices shown in Table 1.

Table 1: ABC Company Daily Closing Prices

Day	Closing Price (GBP)
Monday	75
Tuesday	78
Wednesday	72
Thursday	70
Friday	68

Using the information in Table 1, calculate the annualized volatility of ABC Company's stock for

that week, assuming 250 trading days in a year.

Solution

Step 1: Calculate the continuously compounded daily returns for each day using the formula $\ln(\frac{\text{Ending Price}}{\text{Beginning Price}})$:

$$r_1 = \ln(\frac{78}{75}) = 0.03922$$
$$r_2 = \ln(\frac{72}{78}) = -0.08004$$
$$r_3 = \ln(\frac{70}{72}) = -0.02817$$
$$r_4 = \ln(\frac{68}{70}) = -0.02899$$

Step 2: Calculate the mean of the continuously compounded daily returns:

$$\begin{aligned}\mu &= \frac{r_1 + r_2 + r_3 + r_4}{4} \\ &= \frac{0.03922 + (-0.08004) + (-0.02817) + (-0.02899)}{4} \\ &= -0.024495\end{aligned}$$

Step 3: Calculate the variance of the continuously compounded daily returns:

$$\begin{aligned}\sigma^2 &= \frac{(r_1 - \mu)^2 + (r_2 - \mu)^2 + (r_3 - \mu)^2 + (r_4 - \mu)^2}{4} \\ &= \frac{\begin{aligned}&[(0.03922 - (-0.24495))^2 + (-0.08004 - (-0.024495))^2 \\ &+ (-0.02817 - (-0.024495))^2 + (-0.02899 - (-0.024495))^2]\end{aligned}}{4} \\ &= \frac{0.007179}{4} = 0.001795\end{aligned}$$

Step 4: Calculate the standard deviation of the continuously compounded daily returns:

$$\begin{aligned}\sigma &= \sqrt{\sigma^2} \\ &= \sqrt{0.001795} = 0.042363\end{aligned}$$

Step 5: Annualize the volatility by multiplying the daily volatility by the square root of the

number of trading days in a year.

We know that:

$$\begin{aligned} \sigma(r_{0,T}) &= \sigma\sqrt{T} \\ \therefore \sigma_{\text{annualized}} &= \sigma_{\text{daily}} \times \sqrt{250} \\ &= 0.042363 \times \sqrt{250} \\ &= 0.6698 \approx 67\% \end{aligned}$$

So, the annualized volatility of ABC Company's stock for that week was 67.23 percent.

Question

Which of the following is true about lognormal distributions compared to normal distributions?

A. They are skewed to the right.
B. They can take on negative values.
C. They are less suitable for describing asset prices than asset returns.

Solution

The correct answer is A.

Lognormal distributions are continuous probability distributions that only take positive values and are often skewed to the right.

B is incorrect because lognormal distributions only take on positive values.

C is incorrect because there is no evidence to suggest that lognormal distributions are less suitable for describing asset prices than asset returns.

LOS 6b: describe Monte Carlo simulation and explain how it can be used in investment applications

Monte Carlo simulations are about producing many random variables based on specific probability distributions. This helps in estimating the probability of various results.

We will give an example to illustrate Monte Carlo Simulation implementation.

Steps Involved in Project Appraisal

Imagine an investor who wants to predict the results of a 70% stock and 30% bond portfolio over 20 years. This is how we set up a Monte Carlo simulation:

Specifying the Simulation:

Step 1: Specify the quantity of interest in terms of underlying variables.

The quantity of interest here could be the final portfolio value after 20 years, denoted as V_{iT}. In this case, this is the final portfolio value at time T resulting from ith simulation trial.

The underlying variable is the return on the portfolio. The starting portfolio value is $100,000, with 70% invested in stocks and 30% in bonds.

Step 2: Specify a time horizon.

Assume we're interested in yearly returns, so the time horizon is 20 years. Divide the calendar time into sub-periods. In this case, we will assume yearly returns so that the number of sub-periods is $K = 20$, and the time increment Δt is, therefore, one year.

Step 3: Specify the method for generating the data used in the simulation.

Here, we need to make distributional assumptions. We might assume that the annual portfolio return follows a normal distribution. Let's say we estimate an average return μ of 7% for stocks, 3% for bonds, a standard deviation σ of 15% for stocks, and 5% for bonds. We can model changes in the portfolio value using the formula below:

$$\begin{aligned}\Delta\text{Portfolio value} = 0.7 * (&\mu_{\text{stock}} \times \text{Prior portfolio value} \times \Delta t \\ &+ \sigma_{\text{stock}} \times \text{Prior portfolio value} \times Z_k) \\ &+ 0.3 * (\mu_{\text{bond}} \times \text{Prior portfolio value} \times \Delta t \\ &+ \sigma_{\text{bond}} \times \text{Prior portfolio value} \times Z_k)\end{aligned}$$

Here, Z_k is a standard normal random variable representing the uncertainty in the portfolio return (risk factor). We can use a computer program to draw 20 random values of Z_k.

Running the Simulation Over a Given Number of Trials:

Step 4: Use the simulated values to produce portfolio values.

This step involves converting the standard normal random numbers (Z_k) generated in step 3 into yearly changes in portfolio value (ΔPortfolio value) using our model from step 3. This gives us 20 observations of possible changes in portfolio value over the 20-year period. From these observations, we create a sequence of 20 portfolio values, starting with the initial value of $100,000.

Step 5: Calculate the final portfolio value.

The average portfolio value at the end of 20 years (V_{iT}) is calculated by summing up the portfolio values at the end of each year and dividing by 20. We then calculate the present value (V_{i0}) of this average value by discounting it to the present using an appropriate interest rate. The subscript i in V_{iT} and V_{i0} indicates that these values are from the ith simulation trial. This completes one simulation trial.

Step 6: Repeat steps 4 and 5 over the required number of trials.

Finally, we repeat steps 4 and 5 multiple times, say, 1,000 times. We then calculate summary statistics, such as the mean, median, and percentiles of the distribution of V_{i0} values. These summary statistics provide a range of potential outcomes for the portfolio value after 20 years, helping the investor understand the risks and rewards of the investment strategy.

Major Applications of Monte Carlo Simulations

- It can also be used to value complex securities such as American or European options.

Limitations of Monte Carlo Simulations

- It only provides us with statistical estimates of results, not exact figures.
- It is fairly complex and can only be carried out using specially designed software that may be expensive.
- The complexity of the process may cause errors, leading to wrong results that can be potentially misleading.

Question

Which of the following is a correct statement about the use of Monte Carlo simulations in finance and investment?

A. They provide exact valuations of call options.
B. They estimate a portfolio's potential returns by simulating its performance.
C. They assess how changes in assumptions, such as interest rates or market volatility, affect a financial model.

Solution

The correct answer is C.

Monte Carlo simulations can assess how changes in assumptions, such as interest rates or market volatility, affect a financial model. This allows analysts to understand the impact of these changes on the model's results.

A is incorrect because Monte Carlo simulations do not provide exact valuations of call options. Instead, they can estimate the value of these options by simulating their potential outcomes.

B is incorrect because while Monte Carlo simulations can estimate a portfolio's potential returns, they do not simply simulate its performance. Instead, they use probability distributions to model the uncertainty in the portfolio's returns.

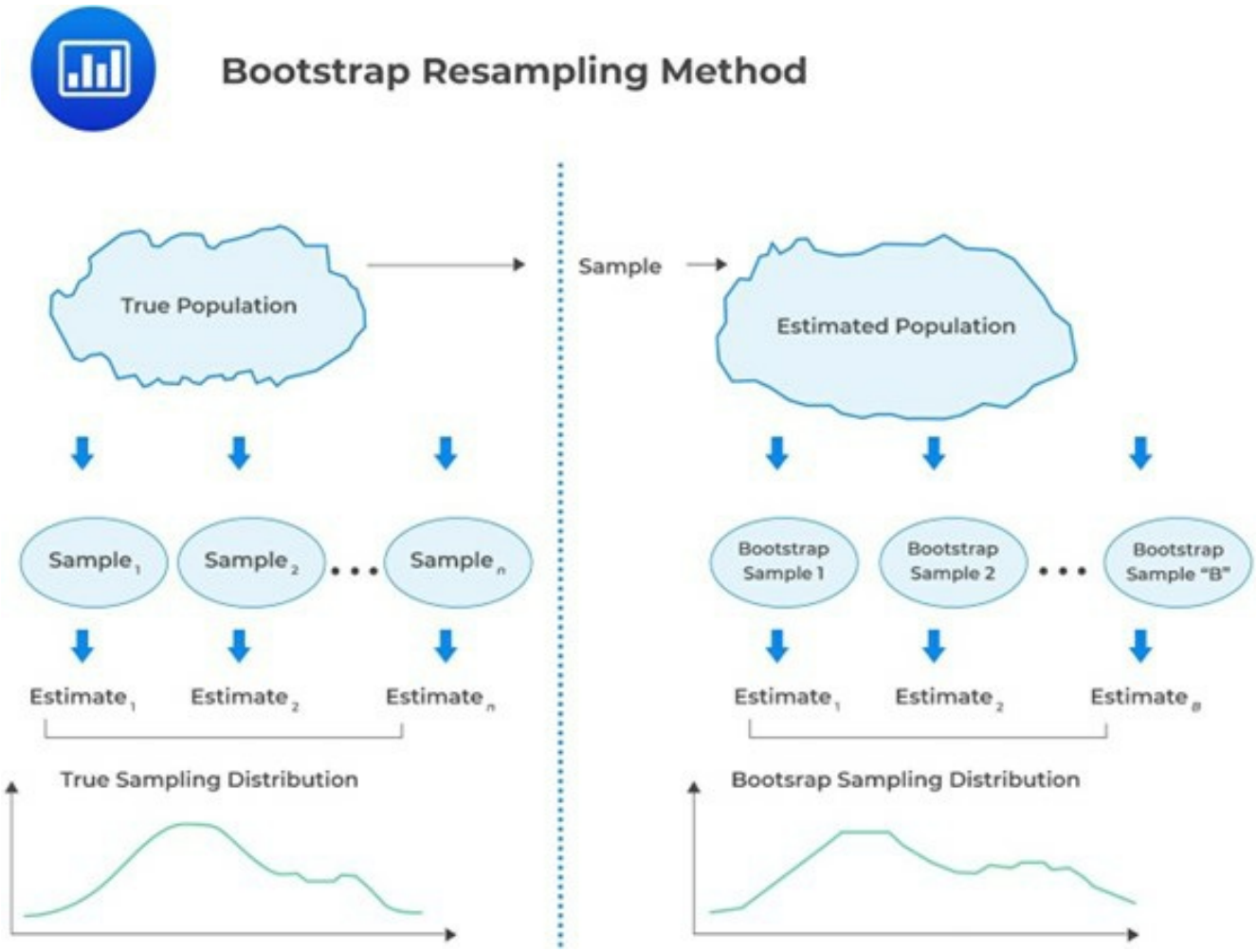
LOS 6c: describe the use of bootstrap resampling in conducting a simulation based on observed data in investment applications

Resampling

Resampling means repeatedly drawing samples from the original observed sample to make statistical inferences about population parameters. There are two common methods: Bootstrap and jackknife. Here, we'll focus on the Bootstrap method.

Bootstrap Resampling

Bootstrap resampling relies on computer simulations for statistical inferences, bypassing the need for conventional analytical formulas like z-statistics. The bootstrap technique is underpinned by a strategy that mirrors the random sampling process from a population to create a sampling distribution.



Note that in bootstrapping, we do not have information about the population. Our only insight comes from a sample of size n drawn from this “unknown population.”

The core concept is that a random sample can effectively stand in for the entire population. So, we can mimic drawing samples from the population by repeatedly resampling from the initial sample. Essentially, the bootstrap method treats the initially obtained sample as a stand-in for the entire population.

Bootstrapping vs Monte Carlo Simulation

Both bootstrap and Monte Carlo simulation techniques lean heavily on the concept of repetitive sampling. Bootstrap considers the resampled dataset as a proxy for the true population and infers the population parameters such as mean, variance, skewness, and kurtosis from the statistical distribution of these samples.

On the other hand, Monte Carlo simulation is centered on the generation of random data with pre-determined statistical distribution of parameter values.

Simulation Using Bootstrapping

Simulation using bootstrapping is similar to Monte Carlo Simulation except for the source of random variables. In bootstrapping, the random variables are taken from a bootstrap sample instead of a probability distribution.

Consider the previous example:

Let's say an investor wants to understand the potential outcomes of investing in a portfolio with a 70-30 split between stocks and bonds over a 20-year period. Here's how a Monte Carlo simulation could be set up:

The simulation steps using the bootstrap sampling distribution are as follows:

Specifying the Simulation:

Step 1: Specify the quantity of interest in terms of underlying variables.

The quantity of interest here could be the final portfolio value after 20 years, denoted as V_{iT}. The underlying variable is the return on the portfolio. The starting portfolio value is $100,000, with 70% invested in stocks and 30% in bonds.

Step 2: Specify a time horizon.

Assume we're interested in yearly returns, so the time horizon is 20 years. Divide the calendar time into sub-periods. In this case, we will assume yearly returns so that the number of subperiods is K = 20, and time increment Δt is, therefore, one year.

Step 3: Generate bootstrap samples from the empirical distribution of portfolio returns.

Here, we use the historical return data as our empirical distribution. Instead of assuming that the annual portfolio return follows a specific theoretical distribution, we will use the bootstrap procedure to draw the K = 20 yearly returns from the observed empirical distribution.

Running the Simulation Over a Given Number of Trials

Step 4: Use the bootstrap samples to produce portfolio values used to value the contingent claim.

This step involves using the bootstrap samples drawn in Step 3 to compute the yearly changes in portfolio value. From there, we create a sequence of 20 portfolio values, starting with the initial value of $100,000.

Step 5: Calculate the final portfolio value:

The average portfolio value at the end of 20 years (V_{iT}) is calculated by summing up the portfolio values at the end of each year and dividing by 20. We then calculate the present value (V_{i0}) of this average value by discounting it to the present using an appropriate interest rate. The subscript i in V_{iT} and V_{i0} indicates that these values are from the ith bootstrap sample. This completes one bootstrap sample.

Step 6: Repeat steps 4 and 5 over the required number of trials.

Finally, we repeat steps 4 and 5 multiple times, say, 1,000 times. We then calculate summary statistics, such as the mean, median, and percentiles of the distribution of V_{i0} values. These summary statistics provide a range of potential outcomes for the portfolio value after 20 years, helping the investor understand the risks and rewards of the investment strategy based on the observed empirical distribution of returns.

Question

Which of the following statements is *most likely* accurate in relation to bootstrap analysis?

A. Bootstrap analysis aims to deduce statistics about population parameters from a singular sample.
B. Bootstrap analysis involves the repeated extraction of samples of equal size, with replacement, from the initial population.
C. During bootstrap analysis, it is necessary for analysts to determine probability distributions for primary risk factors that govern the underlying random variables.

Solution

The correct answer is A.

The bootstrap analysis employs random sampling to generate an observed variable from a set of unknown population parameters. Although the actual distribution of the population is unknown to the analyst, the parameters of the population can be inferred through the sample produced via random sampling.

B is incorrect. In bootstrap analysis, the analyst repeatedly samples from the initial sample, not the entire population. Each resample has the same size as the original sample, and for each new draw, selected items go back into the sample.

C is incorrect. During bootstrap analysis, analysts simply utilize the empirical distribution of the observed underlying variables. In contrast, the analyst must establish probability distributions for the key risk factors that govern the underlying variables in a Monte Carlo simulation.

Learning Module 7: Estimation and Inference

LOS 7a: compare and contrast simple random, stratified random, cluster, convenience, and judgmental sampling and their implications for sampling error in an investment problem

Sampling refers to the systematic process of selecting a subset or sample from a larger population. Sampling is essential because it is costly and time-consuming to analyze the whole population.

Sampling methods can be broadly categorized into: probability sampling and non-probability sampling.

In probability sampling, every population member has an equal chance of being chosen for the sample, ensuring a representative sample. In contrast, non-probability sampling depends on factors such as the sampler's judgment or data accessibility, increasing the risk of an unrepresentative sample.

Probability Sampling Methods

Simple Random Sampling

Simple random sampling means selecting a sample from a population where each element has an equal chance of being chosen. This method aims to create an unbiased sample that accurately represents the population.

Simple random sampling is appropriate when applied to a homogeneous population.

Example: Simple Random Sampling

Imagine we wish to come up with a sample of 50 CFA level I candidates out of 100,000.

One approach may involve numbering each of the 100,000 candidates, placing them in a basket, and shaking the basket to jumble up the numbers. Next, we would randomly draw 50 numbers from the basket, one after the other, without replacement.

A more scientific approach may also involve the use of random numbers where all the 100,000 candidates are numbered in a sequence (from 1 to 100,000). We may then use a computer to randomly generate 50 numbers between 1 and 100,000, where a given number represents a particular candidate who can be identified by their name or admission number.

The underlying feature of random sampling is that all elements in the population must have equal chances of being chosen.

Stratified Random Sampling

In stratified random sampling, analysts subdivide the population into separate groups known as strata (singular stratum). Each stratum comprises elements with a common characteristic (attribute) that distinguishes them from all the others. The method is most appropriate for large **heterogeneous** populations.

A simple random sample is then drawn from within each stratum and combined to form the overall, final sample that takes heterogeneity into account. The number of members chosen from any one stratum depends on its size relative to the population as a whole.

Example: Stratified Random Sampling

An advertising firm wants to determine the extent to which it needs to invigorate television advertisements in a district. The company decides to conduct a survey to estimate the mean number of hours households spend watching TV per week. The district has three distinct towns – A, B, which are urbanized, and C, located in a rural area. Town A is adjacent to a major factory where most residents work, with most having kids of school-going age. Town B mainly harbors retirees, while most people in town C practice agriculture.

There are 160 households in town A, 60 in town B, and 80 in C. Given the differences in the composition of each region, the firm decides to draw a sample of 50 households, considering the total number of families in each.

What is the number of homes that have been sampled in each town?

Solution

We have three strata: towns A, B, and C. We use the following formula to determine the number of households from each region to be included in the sample:

$$\text{Number of households in sample} = \left(\frac{\text{Number of households in the region}}{\text{Total number of households}}\right) \times \text{Required sample size}$$

Therefore, the number of households to be sampled in town A,

$$= \frac{160}{300} \times 50 = 27 \,(\text{approximately})$$

Similarly, the number of households to be sampled in town B,

$$= \frac{60}{300} \times 50 = 10$$

Finally, the firm would need $\left(\frac{80}{300} \times 50\right) = 13$ households in town C.

Advantages of Stratified Sampling over Simple Random Sampling

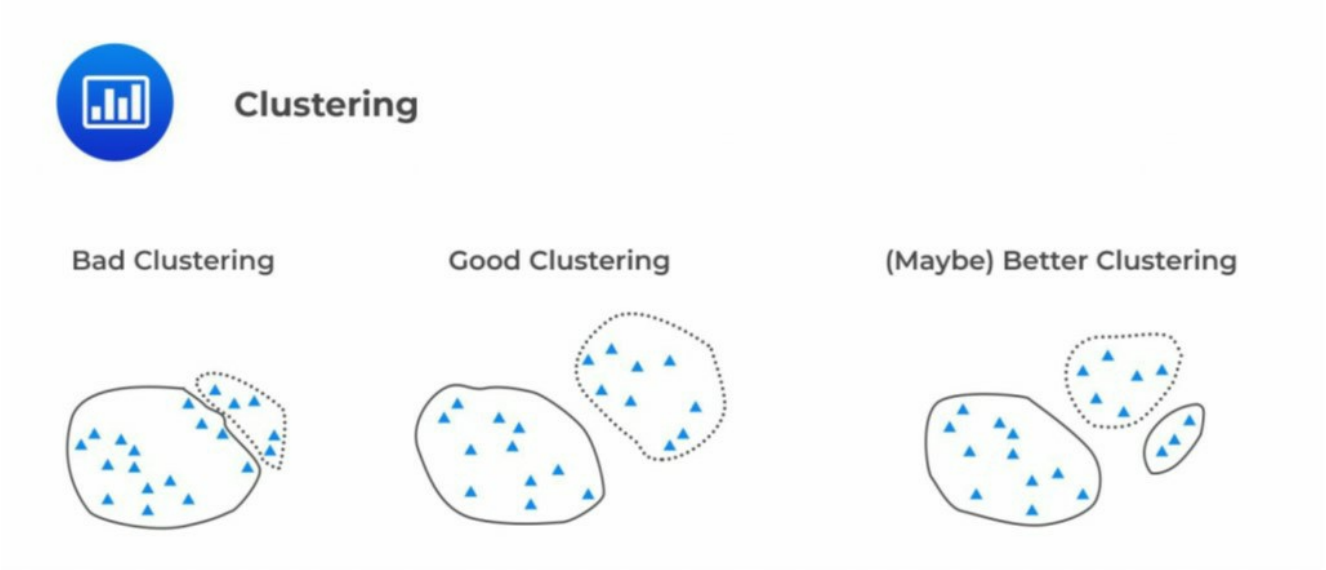
- Stratification is associated with a smaller estimation error compared to simple random sampling, especially when each stratum is homogeneous.
- Stratification enables analysts to estimate the population parameter, say, the mean for all the subgroups of the entire population.

Cluster Sampling

Cluster sampling involves categorizing all population elements into distinct and all-encompassing groups called clusters. Then, you can either choose a random sample of entire clusters or select a random subset from each cluster. So, there are two cluster sampling approaches:

- One-stage (or single-stage) cluster sampling: All the members in each sampled cluster are sampled.
- Two-stage cluster sampling: A simple random sub-sample of members is selected from

each cluster.



Key Point Difference Between Stratified Sampling and Cluster Sampling

- In cluster sampling, a cluster serves as a single sampling unit, and only specific clusters are sampled.
- In stratified sampling, you select members from within each stratum and then draw a random sample from each stratum.

Non-Probability Sampling Techniques

Non-probability samples are selected based on judgment or the convenience of accessing data. As such, non-probability sampling depends on the researchers' sample selection skills. There are two types of non-probability sampling methods:

i. **Convenience sampling**: Researchers choose population elements based on ease of access. This method may not provide a fully representative sample, limiting sampling accuracy.
ii. **Judgmental sampling**: Researchers select elements subjectively, often based on their own knowledge and expertise. However, this approach can introduce bias and result in a

non-representative sample.

Judgmental sampling is preferred when a restricted number of people in a population possess qualities that the researcher expects from the target population.

Comparison Between Probability Sampling and Nonprobability Sampling

Method	Strengths	Weaknesses
Probability Sampling		
Simple random sampling	Easy to use	Lower precision; no assurance of representativeness
Stratified sampling	Higher precision relative to simple random sampling	Difficult to choose relevant stratification; expensive
Cluster sampling	Cost-effective and efficient	Lower precision
Non-probability Sampling		
Convenience sampling	Cost-effective and saves time; easy to use	Selection bias, sample may not accurately represent population
Judgmental sampling	Cost-effective, convenient, less time consuming	Subjective method. Selection bias, sample may not accurately represent the population.

Sampling Error and Its Implications on Investment

Sampling error refers to the difference between the observed value (results obtained from analyzing a sample of investment data) and the true values that would have been obtained from analyzing the entire population of investments.

For instance, when we take a sample to estimate a population's mean, there's typically a difference between the sample mean and the true population mean. This difference, known as sampling error, emerges due to natural variation in sampling and because we work with data from only a part of the full population.

Therefore, any conclusions or predictions drawn based on the sample data may deviate from the actual performance or characteristics of the entire investment population.

Question 1

An analyst is analyzing the spending habits of people belonging to different annual income categories. In his analysis, he creates the following different groups according to the annual family income: Less than $30,000, $31,000 – $40,000, $41,000 to $50,000, and $51,000 to $60,000. He then selects a sample from each distinct group to form a whole sample. The sampling method used by the analyst is most likely:

A. Cluster sampling.
B. Stratified sampling.
C. Simple random sampling.

Solution

The correct answer is B.

Dividing the population into different strata/groups and selecting a sample from each group is called the stratified sampling technique.

A is incorrect. In cluster sampling, each cluster is considered a sampling unit, and only selected clusters are sampled.

C is incorrect. Simple random sampling involves the selection of a sample from an entire population such that each member or element of the population has an equal probability of being picked.

Question 2

A Ph.D. student is conducting research related to her thesis, and for this purpose, she uses some students from her university to constitute a sample. The sampling method used by the analyst is *most likely*:

A. Simple random sampling.
B. Convenience sampling.
C. Judgmental sampling.

Solution

The correct answer is B.

The researcher has selected the students from her university because she can conveniently access them.

A is incorrect. Simple random sampling involves the selection of a sample from an entire population such that each member or element of the population has an equal probability of being picked.

C is incorrect. Judgmental sampling involves handpicking elements from a sample based on the researcher's knowledge and expertise.

Question 3

An analyst wants to estimate the downtime of ABC Bank's ATMs in a city for the last six months. For this purpose, he selects 20 locations or areas within the city and then selects 50% of the ATMs in each area. The sampling method used by the analyst is *most likely*:

A. Cluster sampling.
B. Stratified random sampling.
C. Simple random sampling.

Solution

The correct answer is A.

In cluster sampling, all population elements are categorized into mutually exclusive and exhaustive groups called clusters. A simple random sample of the cluster is

selected, and then the elements in each of these clusters are sampled.

B is incorrect. In stratified random sampling, analysts subdivide the population into separate groups known as strata (singular–stratum), and each stratum is composed of elements that have a common characteristic (attribute) that distinguishes them from all the others.

C is incorrect. Simple random sampling involves the selection of a sample from an entire population such that each member or element of the population has an equal probability of being picked.

LOS 7b: explain the central limit theorem and its importance for the distribution and standard error of the sample mean

The central limit theorem asserts that "given a population described by any probability distribution having mean μ and finite variance σ^2, the sampling distribution of the sample mean $\bar{X}$ computed from random samples of size n from this population will be approximately normal with mean μ (the population mean) and variance $\frac{\sigma^2}{n}$ (the population variance divided by n) when the sample size n is large".

What is a Large Enough n?

The answer to this question might not be straightforward. Nevertheless, the widely accepted value is $n \geq 30$. The truth is that the value of n depends on the shape of the population involved, i.e., the distribution of X_i and its skewness.

In a non-normal but fairly symmetric distribution, n = 10 can be considered large enough. With a very skewed distribution, the value of n can be 50 or even more.

Standard Error of the Sample Mean

Remember that from the central limit theorem, the variance of the sample mean distribution is given by:

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

Where n is the sample size.

The standard error is the standard deviation of the statistic (sample mean).

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Formally defined, for a sample mean $\bar{X}$ computed from a sample generated by a population with

standard deviation σ, the standard error of the sample mean is given by:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Where σ = Known population standard deviation.

When the population standard deviation, σ, is unknown, the following formula is used to estimate the standard error of the sample mean, also denoted as $s_{\bar{X}}$:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Where: s = Sample standard deviation.

The formula above is applicable where we do not know the population standard deviation. Note that the sample standard deviation is the square root of the sample variance, s^2, given by:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

$$\Rightarrow s = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}$$

The standard error of the sample mean estimates the variation that would occur if you took multiple samples from the same population. While the standard deviation measures variation within one sample, the standard error estimates variation across many samples. So, standard deviation and standard error are distinct concepts.

The standard error of the sample mean gives analysts an idea of how **precisely** the sample mean estimates the population mean. A lower standard error value indicates a more precise estimation of the population mean. On the other hand, a larger standard error value indicates a less precise estimate of the population mean.

It is also important to note that the standard error becomes smaller as the sample size increases. This can be seen from its formula. This happens because increasing the sample size ultimately brings the sample mean closer to the true value of the population mean.

Example 1

In a certain property investment company with an international presence, workers have a mean hourly wage of $12 with a population standard deviation of $3. Given a sample size of 30, the standard error of the sample mean is closest to:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$
$$= \frac{3}{\sqrt{30}} = \$0.55$$

If we were to draw several samples of size 30 from the employee population and construct a sampling distribution of the sample means, we would end up with a mean of $12 and a standard error of $0.55.

Example 2

A sample of 30 latest returns on XYZ stock reveals a mean return of $4 with a sample standard deviation of $0.13. The standard error of the sample mean is closest to:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$
$$\frac{0.13}{\sqrt{30}} = \$0.02$$

If we were to draw more samples from the population of yearly returns on XYZ stock and construct a sample mean distribution, we would end up with a mean of $4 and a standard error of $0.02.

Question

Emma Johnson wants to know how finance analysts performed last year. Johnson assumes that the population cross-sectional standard deviation of finance analyst returns is 8 percent and that the returns are independent across analysts.

The random sample size that Johnson needs if she wants the standard deviation of the sample means to be 2% is *closest to*:

A. 4.

B. 16.

C. 72.

Solution

The correct answer is B.

Remember that,

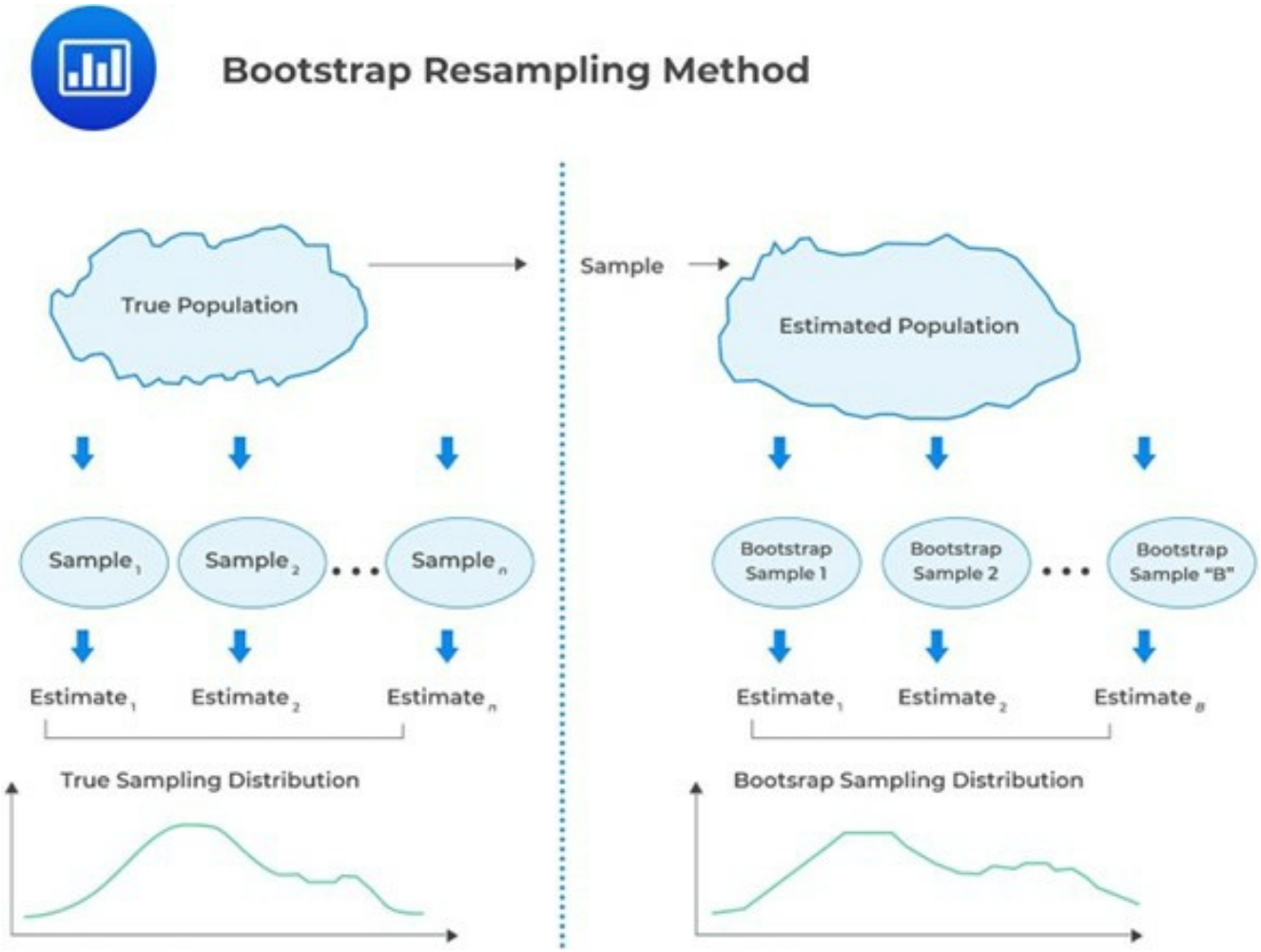
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$
$$\Rightarrow 0.02 = \frac{0.08}{\sqrt{n}}$$
$$\therefore n = 16$$

LOS 7c: describe the use of resampling (bootstrap, jackknife) to estimate the sampling distribution of a statistic

Resampling refers to the act of repeatedly drawing samples from the original observed data sample for the statistical inference of population parameters. The two commonly used methods of resampling are bootstrap and jackknife.

Bootstrap

Using a computer, the bootstrap resampling method simulates drawing multiple random samples from the original sample. Each resample is the same size as the original sample. These resamples are used to create a sampling distribution.



In the bootstrap method, the number of repeated samples to be drawn is at the the researcher's discretion. Note that bootstrap resampling is done with replacement.

Furthermore, we can calculate the standard error of the sample mean. This is done by resampling and calculating the mean of each sample. The following formula is used to estimate the standard error.

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\theta})^2}$$

Where:

$s_{\bar{X}}$ = Estimate of the standard error of the sample mean.

B = Number of resamples drawn from the original sample.

$\hat{\theta}_b$ = Mean of a resample.

$\bar{\theta}$= Mean across all the resample means.

The bootstrap resampling method can also be applied in estimating the confidence intervals for the statistic of other population parameters such as median.

Advantages of Bootstrap Resampling

1. **No Reliance on Analytical Formulas**: Bootstrap differs from traditional statistics because it doesn't rely on an analytical formula for estimating distributions. This makes it versatile for complex estimators and especially useful when analytical formulas are unavailable.
2. **Applicability to Complicated Estimators**: Bootstrap is a simple yet powerful method that can handle complicated estimators effectively. It can handle a wide range of statistical models, making it suitable for various applications in finance where complex estimations are common.
3. **Increased Accuracy**: Bootstrap can enhance accuracy by creating multiple resampled datasets and estimating population parameters on each. This helps understand estimator

variability and robustness, ultimately improving result accuracy.

Jackknife

Jackknife is a resampling method in which samples are drawn by omitting one observation at a time from the original data sample. This process involves drawing samples without replacement. For a sample size of n, we need n repeated samples. This method can be used to reduce the bias of an estimator or to estimate the standard error and the confidence interval of an estimator.

Question

Assume that you are studying the median height of 100 students in a university. You draw a sample of 1000 students and obtain 1000 median heights. The mean across all resample means is 5.8. The sum of squares of the differences between each sample mean, and the mean across all resample means $\sum_{b=1}^{B} (\hat{\theta}_b - \bar{\theta})^2$ is 2.3.

The Estimate of the standard error of the sample mean is *closest to*:

A. 0.05.

B. 0.08.

C. 0.10.

Solution

The correct answer is A.

$$s_{\bar{X}} = \sqrt{\frac{1}{B-1} \sum_{b=1}^{B} (\hat{\theta}_b - \bar{\theta})^2}$$

$$= \sqrt{\frac{1}{1000-1} \times 2.3} = 0.04798 \approx 0.05$$

Learning Module 8: Hypothesis Testing

LOS 8a: Explain hypothesis testing and its components, including statistical significance, Type I and Type II errors, and the power of a test

A hypothesis is an assumed statement about a population's characteristics, often considered an opinion or claim about an issue. To determine if a hypothesis is accurate, statistical tests are used. Hypothesis testing uses sample data to evaluate if a sample statistic reflects a population with the hypothesized value of the population parameter.

Below is an example of a hypothesis:

“The mean return of small-cap stock is higher than that of large-cap stock.”

Hypothesis testing involves collecting and examining a representative sample to verify the accuracy of a hypothesis. Hypothesis tests help analysts to answer questions such as:

- Is bond type A more profitable than type B?
- Does staff training lead to improved efficiency at the workplace?
- Are motor vehicle insurance claims consistent with a lognormal distribution?

Procedure Followed During Hypothesis Testing

Whenever a statistical test is being performed, the following procedure is generally considered ideal:

1. Statement of both the null and the alternative hypotheses.
2. Selection of the appropriate test statistic, i.e., what's being tested, e.g., the population mean, the difference between sample means, or variance.
3. Specification of the level of significance.
4. A clear statement of the decision rule to guide the choice of whether to reject or approve the null hypothesis.
5. Calculation of the sample statistic.

6. Arrival at a decision based on the sample results.

Step 1: Stating the Hypotheses

The Null vs. Alternative Hypothesis

The **null hypothesis**, denoted as H_0, signifies the existing knowledge regarding the population parameter under examination, essentially representing the "status quo." For example, when the U.S. Food and Drug Administration inspects a cooking oil manufacturing plant to confirm that the cholesterol content in 1 kg oil packages doesn't exceed 0.15%, they might create a hypothesis like:

H_0: Each 1 kg package has 0.15% cholesterol.

A test would then be carried out to confirm or reject the null hypothesis.

Typical statements of H_0 include:

$$H_0 : \mu = \mu_0$$
$$H_0 : \mu \leq \mu_0$$
$$H_0 : \mu \geq \mu_0$$

Where:

μ = True population mean.

μ_0 = Hypothesized population mean.

The **alternative hypothesis**, denoted as H_a, is a contradiction of the null hypothesis. Therefore, rejecting the H_0 makes H_a valid. We accept the alternative hypothesis when the “status quo” is discredited and found to be false.

Using our FDA example above, the alternative hypothesis would be:

H_a: Each 1 kg package does not have 0.15% cholesterol.

One-tailed vs. Two-tailed Hypothesis Testing

One-tailed Test

A one-tailed test (one-sided test) is a statistical test that considers a change in only one direction. In such a test, the alternative hypothesis either has a < (less than sign) or > (greater than sign), i.e., we consider either an increase or reduction, but not both.

A one-tailed test directs all the significance levels (α) to test statistical significance in one direction. In other words, we aim to test the possibility of a change in one direction and completely disregard the possibility of a change in the other direction.

If we have a 5% significance level, we shall allot 0.05 of the total area in one tail of the distribution of our test statistic.

Examples: Hypothesis Testing

Let us assume that we are using the standardized normal distribution to test the hypothesis that the population mean is equal to a given value X. Further, let us assume that we are using data from a sample drawn from the population of interest. Our null hypothesis can be expressed as:

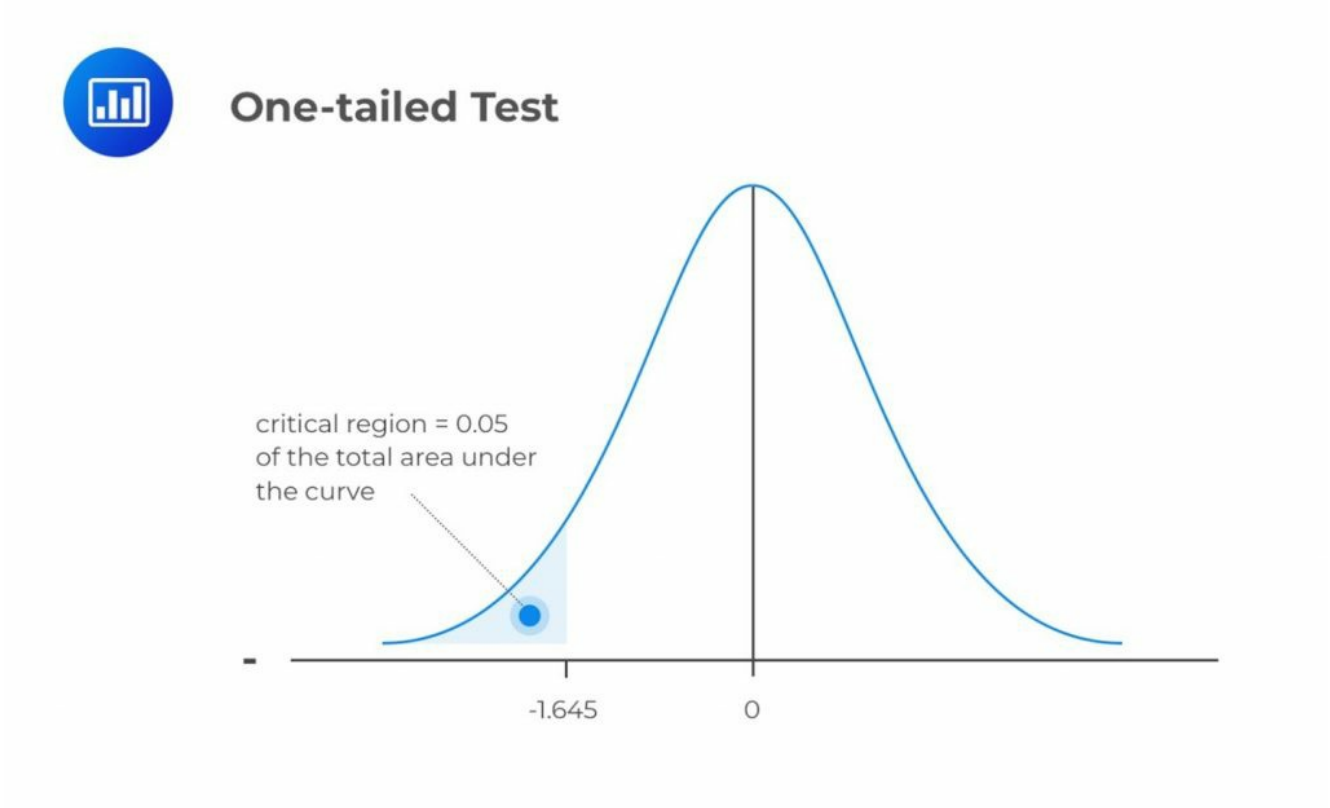
$$H_0 : \mu = X$$

If our test is one-tailed, the alternative hypothesis will test if the mean is either significantly greater than X or significantly less than X, but NOT both.

Case 1: At the 95% Confidence Level

$$H_a : \mu < X$$

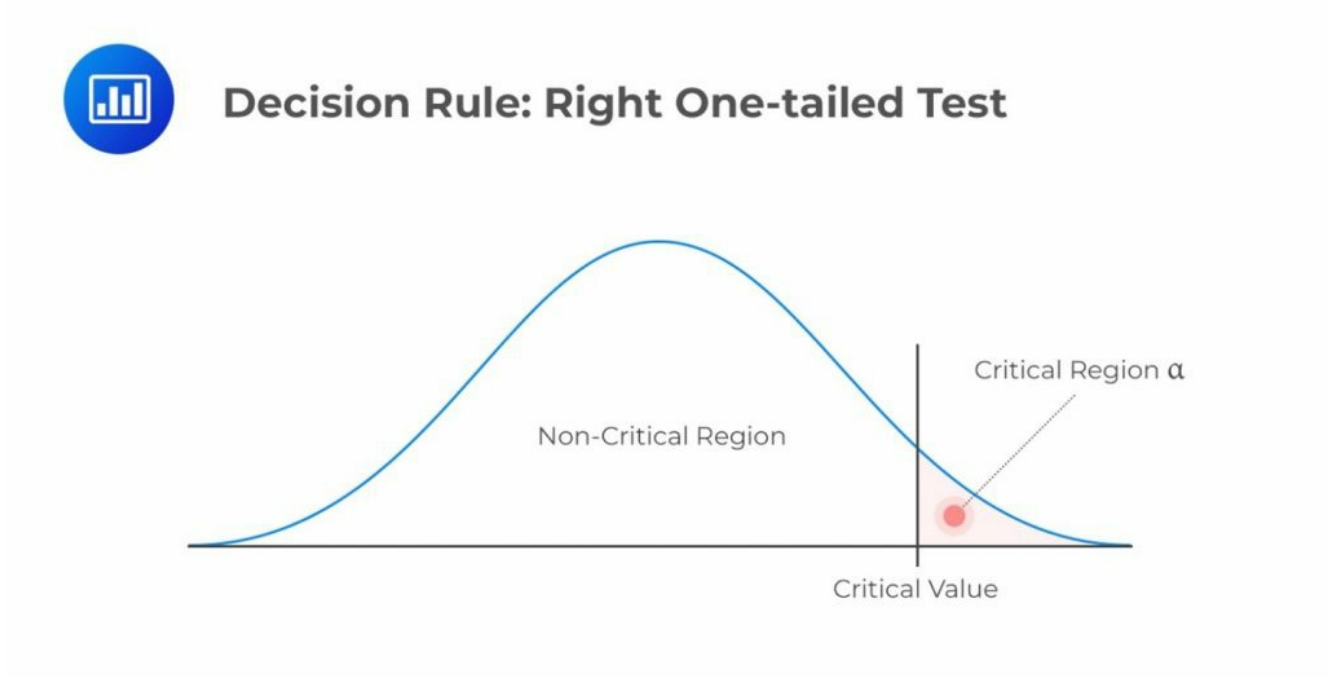
The mean is significantly less than X if the test statistic is in the bottom 5% of the probability distribution. This bottom area is known as the critical region (rejection region). We will reject the null hypothesis if the test statistic is less than -1.645.



Case 2: Still at the 95% Confidence Level

$$H_{a1} : \mu > X$$

We would reject the null hypothesis only if the test statistic is greater than the upper 5% point of the distribution. In other words, we would reject H_0 if the test statistic is greater than 1.645.

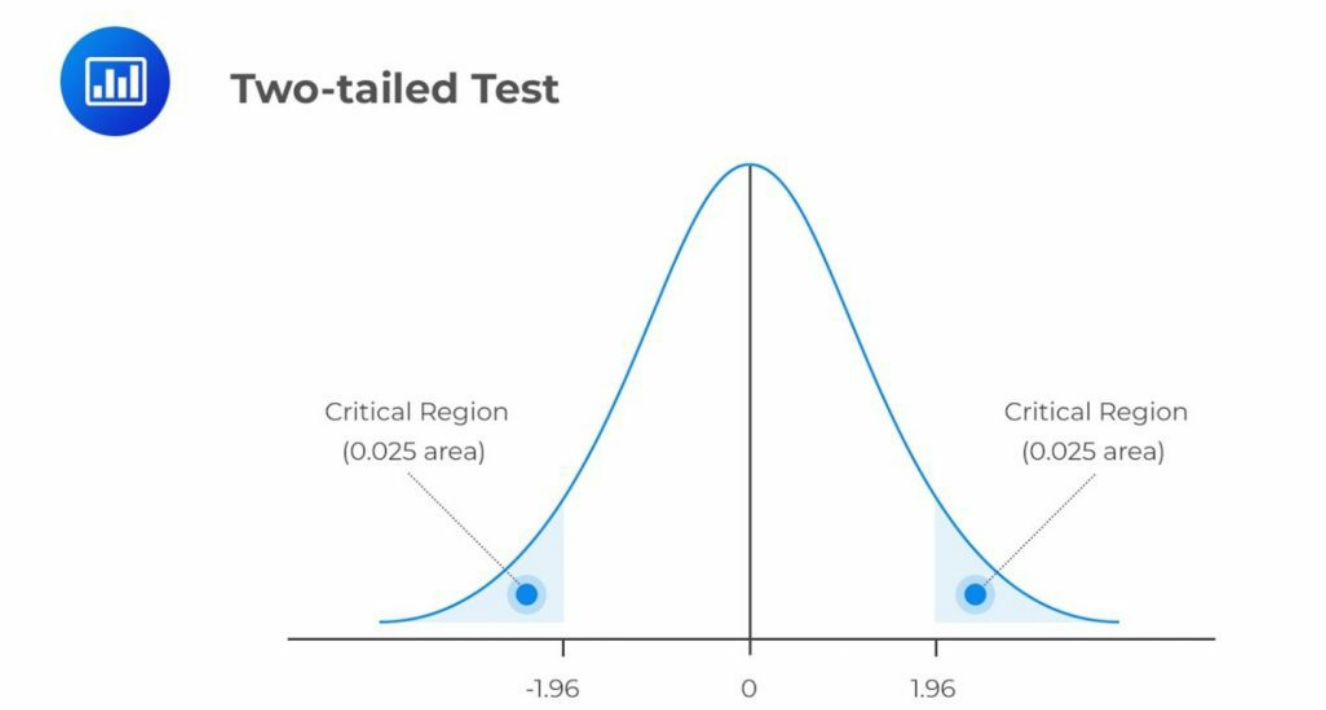


A Two-tailed Test

A two-tailed test considers the possibility of a change in either direction. It looks for a statistical relationship in both a distribution's positive and negative directions. Therefore, it allows half the value of α to test statistical significance in one direction and the other half to test the same in the opposite direction. A two-tailed test may have the following set of hypotheses:

$$H_0 : \mu = X$$
$$H_1 : \mu \neq X$$

Refer to our earlier example. If we were to carry out a two-tailed test, we would reject H_0 if the test statistic turned out to be less than the lower 2.5% point or greater than the upper 2.5% point of the normal distribution.



Step 2: Identify the Appropriate Test Statistic and Distribution

Test Statistic

A test statistic is a standardized value computed from sample information when testing hypotheses. It compares the given data with what an analyst would expect under a null hypothesis. As such, the null hypothesis is a major determinant of the decision to accept or reject H_0, the null hypothesis.

We use test statistic to gauge the degree of agreement between sample data and the null hypothesis. Analysts use the following formula when calculating the test statistic for most tests:

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Hypothesized value}}{\text{Standard error}}$$

The test statistic is a random variable that varies with each sample. The table below provides an overview of commonly used test statistics, depending on the presumed data distribution:

Hypothesis Test	Test Statistic
Z-test	Z- statistic (Normal distribution)
Chi-Square Test	Chi-square statistic
t-test	t-statistics
ANOVA	F-statistic

We can subdivide the set of values that the test statistic can take into two regions: The non-rejection region, which is consistent with the H_0, and the rejection region (critical region), which is inconsistent with the H_0. If the test statistic has a value found within the critical region, we reject the H_0.

As is the case with any other statistic, the distribution of the test statistic must be completely specified under the H_0 when the H_0 is true.

The following is the list of test statistics and their distributions:

Test Subject	Test Statistic Formula	Test Statistic Distribution	Number of Degrees of Freedom
Single Mean	$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$	$t-$ distribution	$n-1$
Difference in Means	$t = \frac{(\bar{X_1} - \bar{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$	$t-$ distribution	$n_1 + n_2 - 1$
Mean of Differences	$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$	$t-$ distribution	$n-1$
Single Variance	$\chi^2 = \frac{s^2(n-1)}{\sigma_0^2}$	Chi-square Distribution	$n-1$
Difference in variances	$F = \frac{S_1^2}{S_2^2}$	F-distribution	$n_1 - 1, n_2 - 1$
Correlation	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	t-distribution	$n-2$
Independence (categorical data)	$\chi^2 = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$	Chi-square Distribution	$(r-1)(c-1)$

Where:

μ_0, μ_{d_0}, and σ_0^2 denote hypothesized values of the mean, mean difference, and variance in that

order.

$\bar{X}$, $\bar{b}$ s^2, s and r denote the sample mean of the differences, sample variance, sample standard deviation, and correlation, in that order.

O_{ij} and E_{ij} are observed and expected frequencies, respectively, with r indicating the number of rows and c indicating the number of columns in the contingency table.

Step 3: Specify the Level of Significance

The significance level represents the amount of sample proof needed to reject the null hypothesis. First, let us look at type I and type II errors.

Type I and Type II Errors

When using sample statistics to draw conclusions about an entire population, the sample might not accurately represent the population. This can result in statistical tests giving incorrect results, leading to either erroneous rejection or acceptance of the null hypothesis. This introduces the two errors discussed below.

Type I Error

Type I error occurs when we reject a true null hypothesis. For example, a type I error would manifest in the rejection of $H_0 = 0$ when it is zero.

Type II Error

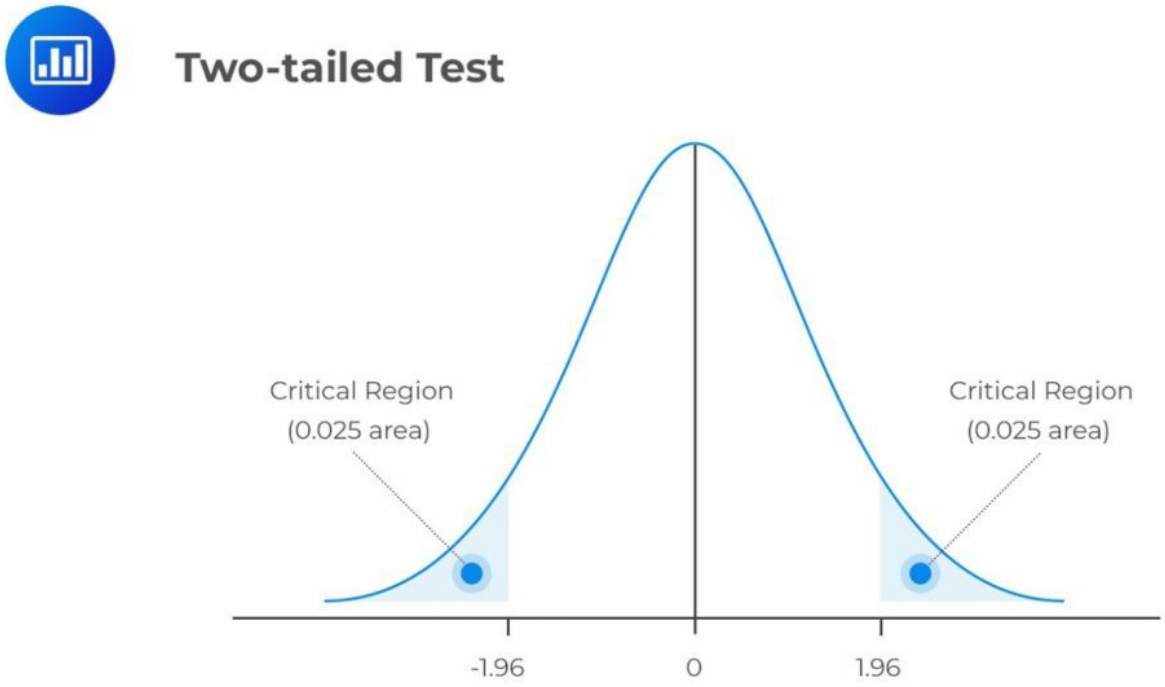
Type II error occurs when we fail to reject a false null hypothesis. In such a scenario, the evidence the test provides is insufficient and, as such, cannot justify the rejection of the null hypothesis when it is false.

Consider the following table:

Decision	True Null Hypothesis (H_0)	False Null Hypothesis (H_0)
Fail to reject the null hypothesis	Correct decision	Type II error
Reject null hypothesis	Type I error	Correct decision

The level of significance, denoted by α, represents the probability of making a type I error, i.e., rejecting the null hypothesis when it is true. The **confidence level** complements the significance level, $(1 - \alpha)$.

We use α to determine critical values that subdivide a distribution into the rejection and the non-rejection regions. The figure below gives an example of the critical regions under a two-tailed normal distribution and 5% significance level:



Consequently, β, the direct opposite of α, is the probability of making a type II error within the bounds of statistical testing. The ideal but practically impossible statistical test would be one that **simultaneously** minimizes α and β.

The Power of a Test

The power of a test is the direct opposite of the significance level. The level of significance gives us the probability of rejecting the null hypothesis when it is, in fact, true. On the other hand, the power of a test gives us the probability of correctly discrediting and rejecting the null hypothesis when it is false. In other words, it gives the likelihood of rejecting H_0 when, indeed, it is false. Expressed mathematically,

$$\text{Power of a test} = 1 - \beta = 1 - P\,(\text{type II error})$$

In a scenario with multiple test results for the same purpose, the test with the highest power is considered the best.

Steps 4, 5, 6: State the Decision Rule, Calculate the Test Statistic, and Make a Decision

The decision rule is the procedure that analysts and researchers follow when deciding whether to reject or not reject a null hypothesis. We use the phrase "not to reject" because it's statistically incorrect to "accept" a null hypothesis. Instead, we can only gather enough evidence to support it.

Breaking Down the Decision Rule

The decision to reject or not reject a null hypothesis relies on the distribution of the test statistic. The decision rule compares the calculated test statistic to the critical value.

If we reject the null hypothesis, the test is considered statistically significant. If not, we fail to reject the null hypothesis, indicating insufficient evidence for rejection.

We use the test's significance level if a variable follows a normal distribution. This helps us find critical values corresponding to specific points on the standard normal distribution. These critical values guide the decision-making process for rejecting or not rejecting a null hypothesis.

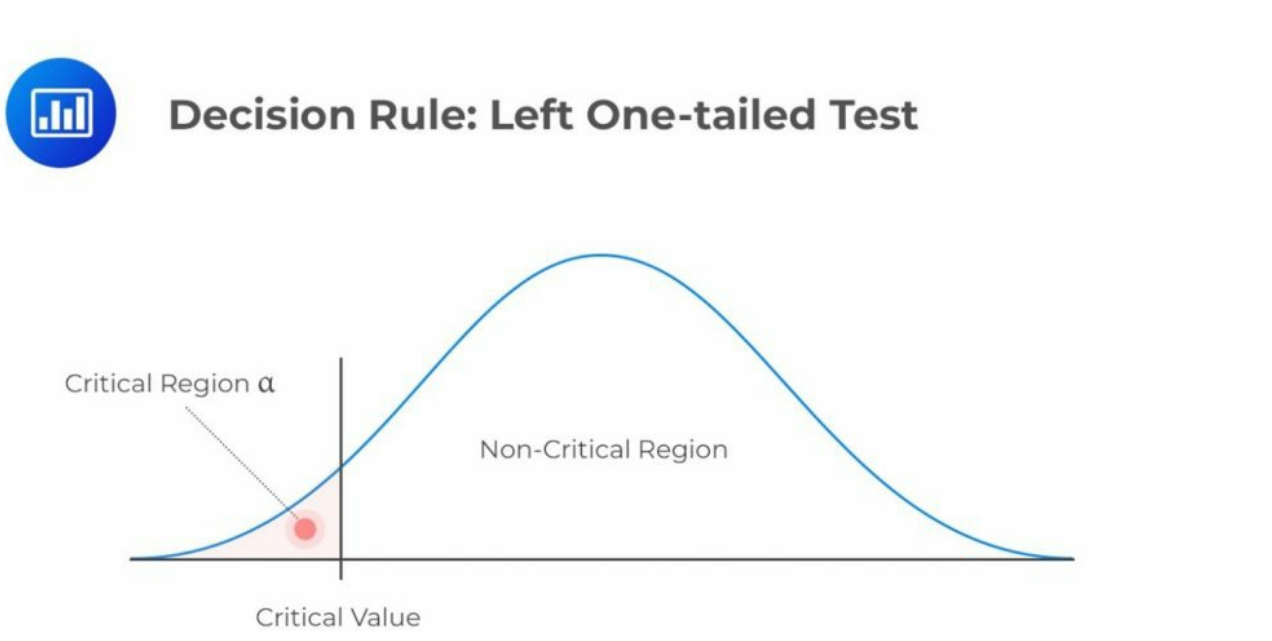
Before deciding whether to reject or not reject a null hypothesis, it's crucial to determine

whether the test should be one-tailed or two-tailed. This choice depends on the nature of the research question and the direction of the expected effect. Notably, the number of tails determines the value of α (significance level). The following is a summary of the decision rules under different scenarios.

Left One-tailed Test

H_a : Parameter $< X$

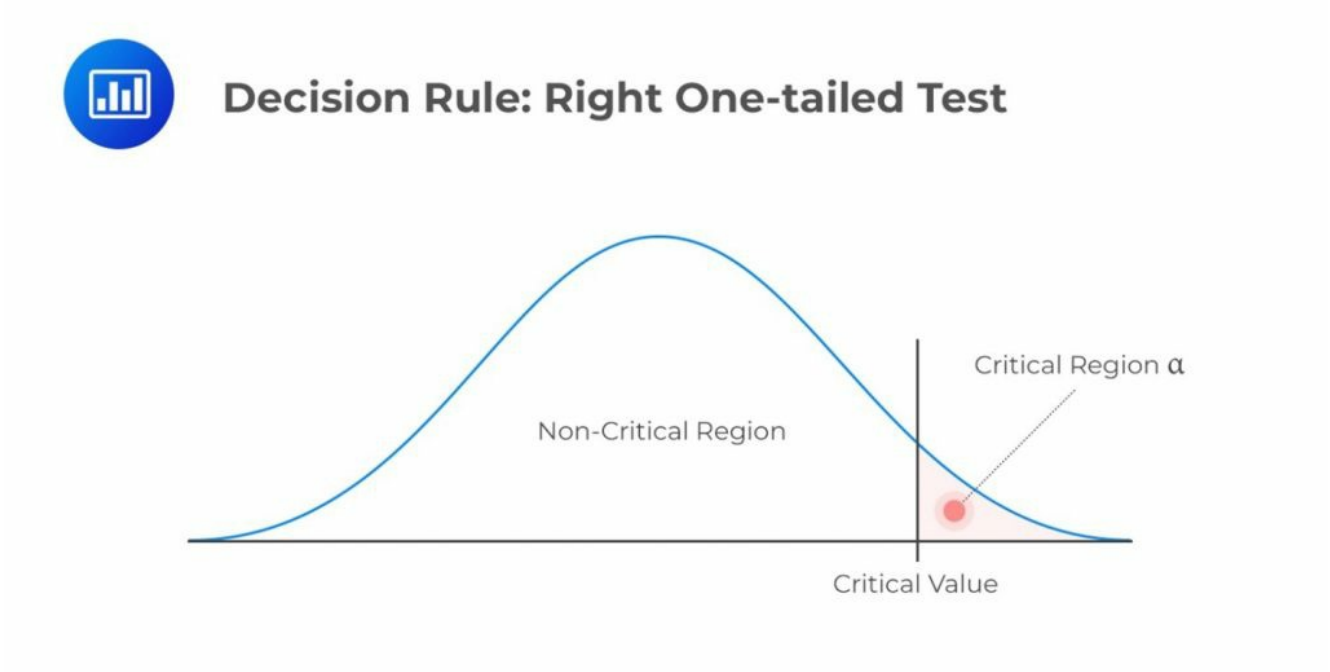
Decision rule: Reject H_0 if the test statistic is less than the critical value. Otherwise, **do not reject** H_0.



Right One-tailed Test

H_a : Parameter $> X$

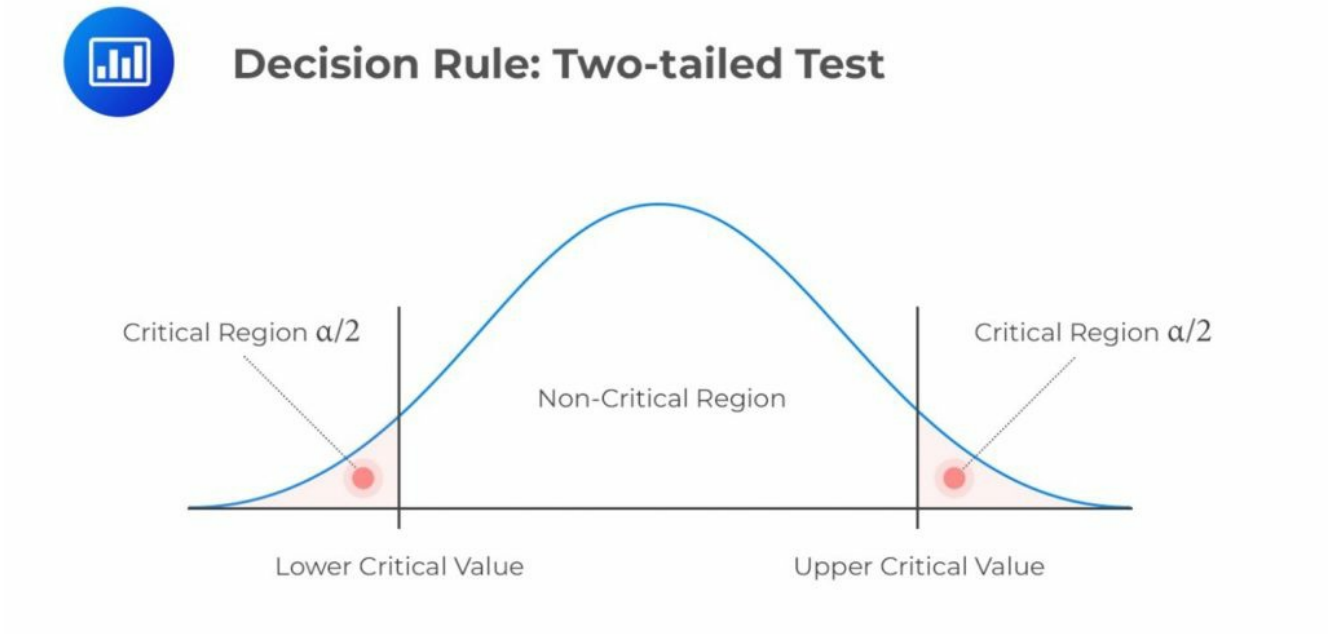
Decision rule: Reject H_0 if the test statistic is greater than the critical value. Otherwise, **do not reject** H_0.



Two-tailed Test

H_a : Parameter $\neq$ X (not equal to X)

Decision rule: Reject H_0 if the test statistic is greater than the upper critical value or less than the lower critical value.



The p-Value in Hypothesis Testing

The p-value is the lowest level of significance at which we can reject a null hypothesis. The probability of coming up with a test statistic would justify our rejection of a null hypothesis, assuming that the null hypothesis is indeed true.

Breaking Down the p-value

When carrying out a statistical test with a fixed value of the significance level (?), we merely compare the observed test statistic with some critical value. For example, we might “reject an H_0 using a 5% test” or “reject an H_0 at a 1% significance level.” The problem with this ‘classical’ approach is that it does not give us details about the **strength of the evidence** against the null hypothesis.

Determination of the p-value gives statisticians a more informative approach to hypothesis testing. The p-value is the lowest level at which we can reject an H_0. This means that the strength of the evidence against an H_0 increases as the p-value becomes smaller.

In one-tailed tests, the p-value is the probability below the calculated test statistic for left-tailed tests or above the test statistic for right-tailed tests. For two-tailed tests, we find the probability below the negative test statistic and add it to the probability above the positive test statistic. This combines both tails for the p-value calculation.

Example: p-value

θ represents the probability of obtaining a head when a coin is tossed. Assume we tossed a coin 200 times, and the head came up in 85 out of the 200 trials. Test the following hypothesis at a 5% level of significance.

$H_0: \ \theta = 0.5$

$H_1: \ \theta < 0.5$

Solution

First, note that repeatedly tossing a coin follows a binomial distribution.

Our p-value will be given by $P(X < 85)$, where X follows a binomial (200,0.5), assuming the H_0 is true.

$$= P\left[Z < \frac{(85 - 100)}{\sqrt{50}}\right]$$
$$= P(Z < -2.12) = 1 - 0.9834 = 0.01660$$

(We have applied the Central Limit Theorem by taking the binomial distribution as approximately normal.)

Since the probability is less than 0.05, the H0 is extremely unlikely, and we have strong evidence against an H_0 that favors H_1. Therefore, clearly expressing this result, we could say:

“There is very strong evidence against the hypothesis that the coin is fair. We, therefore, conclude that the coin is biased against heads.”

Remember, failure to reject a H_0 does not mean it is true. It means there is insufficient evidence to justify the rejection of the H_0 given a certain level of significance.

Question

A CFA candidate conducts a statistical test about the mean value of a random variable X.

$H_0 : \mu = \mu_0$ vs $H_1 : \mu \neq \mu_0$

She obtains a test statistic of 2.2. Given a 5% significance level, determine the p-value.

A. 1.39%.

B. 2.78.

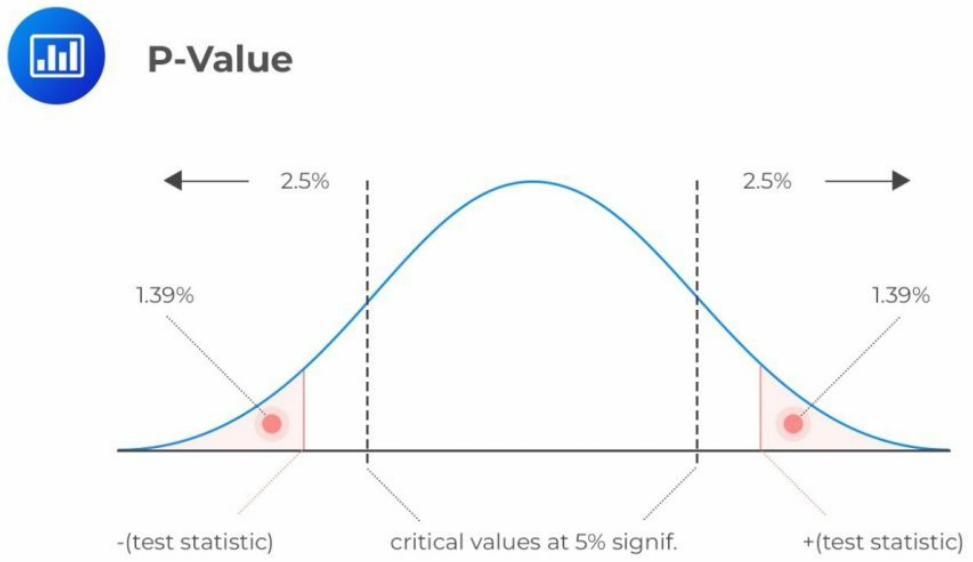
C. 2.78%.

Solution

The correct answer is C.

$$P - \text{value} = P(Z > 2.2) = 1 - P(Z < 2.2) = 1.39\% \times 2 = 2.78\%$$

(We have multiplied by two since this is a two-tailed test.)



Interpretation: The p-value (2.78%) is less than the significance level (5%). Therefore, we have sufficient evidence to reject the H_0. In fact, the evidence is so strong that we would also reject the H_0 at significance levels of 4% and 3%. However, at significance levels of 2% or 1%, we would not reject the H0 since the p-value surpasses these values.

LOS 8b: Construct hypothesis tests and determine their statistical significance, the associated Type I and Type II errors, and the power of the test given a significance level

Hypothesis Test Concerning Single Mean

The z-test is the ideal hypothesis test when the sampling distribution of the sample is normally distributed or when the standard deviation is known.

The z-statistic is the test statistic used in hypothesis testing.

Testing $H_0 : \mu = \mu_0$ Using the z-test

Given a random sample of size n from a normally distributed population with mean μ, variance σ^2, and a sample mean $\bar{X}$, we can compute the z-statistic as follows:

$$z - \text{statistic} = \frac{(\bar{X} - \mu_0)}{(\frac{\sigma}{\sqrt{n}})}$$

Where:

$\bar{X}$ is the sample mean.

$?_0$ is the hypothesized mean of the population.

σ is the standard deviation of the population.

n is the sample size.

Once computed, the z-statistic is compared to the critical value that corresponds to the level of significance of the test. For example, if the significance level is 5%, the z-statistic is screened against the upper or lower 95% point of the normal distribution (±1.96). The decision rule is to reject the H0 if the z-statistic falls within the critical or rejection region.

Example: z-test

Academics carried out a study on 50 former United States presidents and found an average IQ of 135. You are required to carry out a 5% statistical test to determine whether the average IQ of presidents is greater than 130. (IQs are distributed normally, and previous studies indicate that $\sigma = 25$.)

Solution

Step 1: State the hypothesis:

$H_0 : \mu \leq 130$

$H_1 : \mu > 130$

Step 2: Identify the appropriate t-statistic:

$$z - \text{statistic} = \frac{(\bar{X} - \mu_0)}{(\frac{\sigma}{\sqrt{n}})}$$

Assuming the H_0 is true, $\frac{(\bar{X}-130)}{(\frac{\sigma}{\sqrt{n}})} \sim N(0,1)$

Step 3: Specify the level of significance:

This is a right-tailed test. Therefore, we compare our test statistic to the upper 95% point of the standard normal distribution (1.645).

Step 4: State the decision rule:

Reject the null hypothesis if the z-statistic is greater than 1.645.

Step 5: Calculate the test statistic:

The z-statistic is $\frac{(135-130)}{(\frac{25}{\sqrt{50}})} = 1.414$

Step 6: Make a decision:

Since 1.414 is less than 1.645, we **do not have** sufficient evidence to reject the H_0. As such, it would be **reasonable** to conclude that the average IQ of U.S. presidents is not more than 130.

The t-test

The t-test is based on the t-distribution. The test is appropriate for testing the value of a population mean when:

- σ is unknown.
- The sample size is large (n ? 30), and if n < 30, the distribution must be normal or approximately normal.

Testing $H_0 : \mu = \mu_0$ Using the t-test

We compute a t-statistic with $n - 1$ degrees of freedom as follows:

$$t_{n-1} = \frac{(\bar{X} - \mu_0)}{(\frac{s}{\sqrt{n}})}$$

Where:

$\bar{X}$ is the sample mean.

μ_0 is the hypothesized mean of the population.

s is the standard deviation of the sample.

n is the sample size.

Example: t-Test

Financial analysts in a certain equatorial country are interested in evaluating the potential impact of rainfall on agricultural investments. They have gathered data on the annual rate of rainfall (cm) over the last 10 years, as shown below:

{ 25, 26, 25, 27, 28, 29, 28, 27, 26, 25 }

Previously, the recorded average rainfall was 23 cm. The analysts want to find out if there's been an increase in the average rainfall rate, which could impact agricultural investment. Conduct a

statistical test at a 5% significance level to investigate this.

Solution

Follow the steps outlined above.

Step 1: State the hypothesis:

As always, you should begin by stating the hypothesis:

$H_0 : \mu \leq 23$

$H_1 : \mu > 23$

Step 2: Identify the appropriate t-statistic:

If we assume that the annual rainfall quantities are distributed normally and recorded independently, then:

$$\frac{(\bar{X} - \mu_0)}{(\frac{s}{\sqrt{n}})} \sim t_{n-1}$$

Please, confirm that $\bar{X} = 26.6$ and $s = 1.43$

Step 3: Specify the level of significance:

$$\alpha = 5\% \, (\text{right} - \text{tailed})$$

Step 4: State the decision rule:

Reject the null hypothesis if the t-statistic is greater than $t_{0.05,9} = 1.833$

Step 5: Calculate the test statistic:

Therefore, our t-statistic $= \frac{(26.6 - 23)}{(\frac{1.43}{\sqrt{10}})} = 7.96$

Step 6: Make a decision:

Our test statistic (7.96) is greater than the upper 95% point of the $t_{0.05,9}$ distribution (1.833).

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.31 752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.91 986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.35 363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.13 847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.01 048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.94 180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.89 579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.85 548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869

Therefore, we have **sufficient evidence** to reject the H0. As such, it is **reasonable** to conclude that the average annual rainfall has increased from its former long-term average of 23.

Question

What is the value of t in the example above if the significance level is reduced from 5% to 0.5%, and does this change the decision rule?

A. 2.02; it does not change the decision rule.

B. 3.25; it does not change the decision rule.

C. 3.25; it changes the decision rule.

Solution

The correct answer is B.

A quick glance at the $t_{0.005,9}$ distribution when $\alpha = 0.5\%$ gives a value of 3.25.

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869

However, the evidence against the H_0 is overwhelming since our test statistic (7.96) is still greater than 3.25. As such, the conclusion would remain unchanged.

Hypothesis Test Concerning the Equality of the Population Means

Analysts are often interested in establishing whether there exists a significant difference between the means of two different populations. For instance, they might want to know whether

the average returns for two subsidiaries of a given company exhibit a **significant** variance. Such a test may then be used to make decisions regarding resource allocation or the reward of the directors. Before embarking on such an exercise, it is paramount to ensure that the samples taken are independent and sourced from normally distributed populations. It can either be assumed that the population variances are equal or unequal. In this reading, we will assume that the population variances are equal.

Assume that μ_1 is the mean of the first population while μ_2 is the mean of the second population. In testing the equality of two population means, we wish to determine if they are equal or not. As such, the hypotheses can be any of the following:

I. **Two-sided**:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ vs. } H_a ? \mu_1 - \mu_2 \neq 0,$$

This can be stated as follows:

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_a ? \mu_1 \neq \mu_2$$

II. **One-sided (right)**:

$$H_0 : \mu_1 - \mu_2 \leq 0 \text{ vs. } H_a ? \mu_1 - \mu_2 > 0,$$

This can be stated as follows:

$$H_0 : \mu_1 \leq \mu_2 \text{ vs. } H_a ? \mu_1 > \mu_2$$

III. **One-sided (Left)**:

$$H_0 : \mu_1 - \mu_2 \geq 0 \text{ vs. } H_a ? \mu_1 - \mu_2 < 0,$$

This can be stated as follows:

$$H_0 : \mu_1 \geq \mu_2 \text{ vs. } H_a ? \mu_1 < \mu_2$$

However, note that tests such as $H_0 : \mu_1 - \mu_2 = 3 \text{ vs. } H_a ? \mu_1 - \mu_2 \neq 3$ are valid. The process is similar in both cases.

Hypothesis Test Concerning Differences between Means With Independent Samples

When testing for the difference between two population means, we assume that the two populations are distributed normally. Further, we assume that they have equal and unknown variances. We always make use of the student's t-distribution where the test statistic is given by:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p \left(\sqrt{\frac{1}{n1} + \frac{1}{n2}}\right)}$$

Where s_p^2 is the pooled estimator of the common variance and is given by:

$$s_p^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2}$$

Also, the variables are defined as follows:

$\bar{X}_1$ = Mean of the first sample.

$\bar{X}_2$ = Mean of the second sample.

s_1^2 = variance of the first sample.

s_2^2 = variance of the second sample.

n_1 = Sample size of the first sample.

n_2 = Sample size of the second sample.

Example 1: Hypothesis Test Concerning the Equality of the Population Means

Nutritionists want to establish whether obese patients on a new special diet have a lower weight than the control group. After six weeks, the average weight of 10 patients (group A) on the special diet is 75kg, while that of 10 more patients of the control group (B) is 72kg. Carry out a 5% test to determine if the patients on the special diet have a lower weight.

Additional information: $\sum A^2 = 59520$ and $\sum B^2 = 56430$

Solution

As is the norm, start by stating the hypothesis:

$$H_0 : \mu_1 - \mu_2 = 0 \text{ Vs } H_a ? \mu_1 - \mu_2 \neq 0,$$

We assume that the two samples have equal variances, are independent, and are normally distributed. Then, under H_0,

$$\frac{\bar{B} - \bar{A}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Note that the sample variance is given by:

$$s^2 = \frac{\sum X^2 - n\bar{X}^2}{n-1}$$

So,

$$S_A^2 = \frac{\{59520 - (10 * 75^2)\}}{9} = 363.33$$
$$S_B^2 = \frac{\{56430 - (10 * 72^2)\}}{9} = 510$$

Therefore,

$$S_2^p = \frac{(9 \times 363.33 + 9 \times 510)}{(10 + 10 - 2)} = 436.665$$

And

$$\text{Test statistic} = \frac{(75-72)}{\{\sqrt{439.665} \times \sqrt{(\frac{1}{10}+\frac{1}{10})}\}} = 0.3210$$

Our test statistic (0.3210) is less than the upper 5% point (1.734) of the t-distribution with 18 degrees of freedom.

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.31 752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.91 986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.35 363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.13 847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.01 048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.94 180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.89 579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.85 548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.83 113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.81 461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.79 885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.78 288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.77 933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.76 310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.75 050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.74 884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.73 607	2.10982	2.56693	2.89823	3.9651
18	[illegible]	[illegible]	[illegible]	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495

Therefore, we **do not have sufficient** evidence to reject the H_0 at 5% significance. As such, it is **reasonable** to conclude that the special diet has the same effect on body weight as the placebo.

Note to candidates: You could choose to work with the p-value and determine $P(t_{18} > 0.937)$ and then establish whether this probability is less than 0.05. Working out the problem this way would lead to the same conclusion as above.

Example 2: **Hypothesis Test Concerning the Equality of the Population Means**

Suppose we replace '>' with '≠' in H_1 in the example above, would the decision rule change?

Replacing '>' with '≠' in H_1 would change the test from a one-tailed one to a two-tailed test. We would compute the test statistic just as demonstrated above. However, we would have to divide the level of significance by two and compare the test statistic to both the lower and upper 2.5% points of the t18-distribution (±2.101).

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.7[illegible]620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30[illegible]265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18[illegible]245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77[illegible]645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57[illegible]058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44[illegible]691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36[illegible]462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30[illegible]600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26[illegible]216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22[illegible]814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20[illegible]099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17[illegible]881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16[illegible]037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14[illegible]479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13[illegible]145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11[illegible]991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10[illegible]982	2.56693	2.89823	3.9651
18	[illegible]	[illegible]	[illegible]	[illegible]	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495

Since our test statistic lies within these limits (non-rejection region), the decision rule would remain unchanged.

Hypothesis Test Concerning Differences between Means With Dependent Samples

There are some challenges when testing the difference between two population means using

independent samples. Variability within each sample, caused by factors unrelated to the research, can obscure the real difference of interest. Random variation within a sample might be so substantial that it obscures the actual difference caused by the specific phenomenon the analyst is studying.

When we want to test the differences between means with dependent samples, we use the paired comparison test (test of the mean of the differences).

Paired Comparisons Test

Assume that we have observations for the random variables X_A and X_B and that the samples are dependent.

Organize the observations in pairs and denote the differences between the two paired observations by d_i. That is $d_i = x_{Ai} - x_{Bi}$ where x_{Ai} and x_{Bi} are the ith pair of observations $1, 2, \dots, n$.

Also, let μ_d be the population mean difference and μ_{d0} be the hypothesized value for the population mean difference. At this point, we can state the hypotheses:

- **Two-sided**: $H_0 : \mu_d = \mu_{d0}$ versus $H_a : \mu_d \neq \mu_{d0}$
- **One-sided (right side)**: $H_0 : \mu_d \leq \mu_{d0}$ versus $H_a : \mu_d > \mu_{d0}$
- **One-sided (left side)**: $H_0 : \mu_d \geq \mu_{d0}$ versus $H_a : \mu_d < \mu_{d0}$

Practically, $\mu_{d0} = 0$.

We are considering normally distributed populations with unknown population variances. As such, we will use the t-distributed statistic given by:

$$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$$

Where:

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$$

$$s_{\bar{d}} = s_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \text{standard error of the mean differences}$$

$$s_d = \text{standard deviation of the differences}$$

Note that the degree of freedom is $n - 1$ where n is the number of the paired observations.

Example 2: **Paired Comparison Test**

An analyst aims to compare the performance of the BCD High Growth Index and the BCD Investment Grade Index. They collect data for both indexes over 2,050 days and calculate the means and standard deviations, as shown in the table below:

	BCD High Growth Index (%)	BCD Investment Grade Index (%)	Difference (%)
Mean return	0.0183	0.0161	-0.0022
Standard deviation	0.2789	0.3298	0.3321

Using a 5% significance level, determine whether the mean of the differences is different from zero.

Solution

Step 1: State the hypothesis:

$$H_0 : \mu_{d0} = 0 \text{ vs } H_a : \mu_{d0} \neq 0$$

Step 2: Identify the appropriate t-statistic:

We use the t-statistic, calculated by:

$$t = \frac{\bar{d} - \mu_{d0}}{s_{\bar{d}}}$$

Step 3: Specify the level of significance:

$$\alpha = 5\% \text{ (two-tailed test)}$$

Step 4: State the decision rule:

The degrees of freedom amount to $n-1=2,050-1=2,049$; thus, the critical values are ±1.960. Therefore, we will reject the null hypothesis if the calculated t-statistic is less than -1.96 or greater than 1.96.

Step 5: Calculate the test statistic:

Note that from the table $\bar{d}=-0.0022$ and $s_d=0.3321$ so that the t-statistic is given:

$$t=\frac{\bar{d}-\mu_{d0}}{s_{\bar{d}}}=\frac{-0.0022-0}{\frac{0.3321}{\sqrt{2050}}}=-0.30$$

Step 6: Make a Decision:

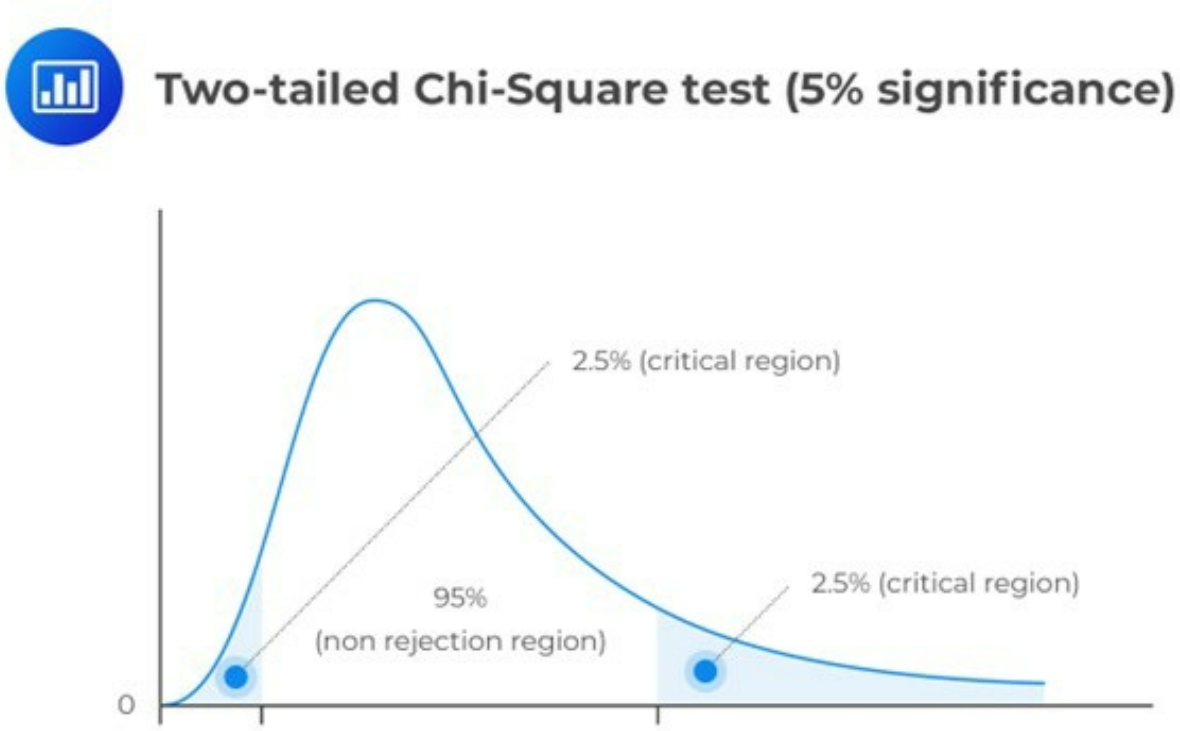
-0.30 falls within the bounds of the critical values of ±1.960. As such, there is insufficient evidence to show that the mean of the differences in returns differs from zero.

ν \ α	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
1	3.078	6.314	12.076	31.821	63.657	318.310	636.620
2	1.886	2.920	4.303	6.965	9.925	22.326	31.598
3	1.638	2.353	3.182	4.541	5.841	10.213	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	1.296	1.671	2.000	2.390	2.660	3.232	3.460
120	1.289	1.658	1.980	2.358	2.617	3.160	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Test Variance and Equality of Variance

Testing of a Single Variance

A chi-square test helps determine if a hypothesized variance value matches the true population variance. Unlike other distributions in the CFA® Program, the chi-square distribution is asymmetrical. Yet, as degrees of freedom increase, it approaches a more normal distribution.



As a natural consequence, the chi-square distribution has no negative values and is bounded by zero. A chi-square statistic with $(n-1)$ degrees of freedom is computed as:

$$\chi^2_{n-1} = \frac{(n-1)\,S^2}{\sigma_0^2}$$

Where:

n = Sample size.

S^2 = Sample variance.

σ_0^2 = Hypothesized population variance.

Example: Chi-square Test

For the 15-year period between 1995 and 2010, ABC's monthly return had a standard deviation of 5%. John Matthew, CFA, wishes to establish whether the standard deviation witnessed during that period still adequately describes the long-term standard deviation of the company's return. To achieve this end, he collects data on the monthly returns recorded between January 1, 2015, and December 31, 2016, and computes a monthly standard deviation of 4%.

Carry out a 5% test to determine if the standard deviation computed in the latter period differs from the 15-year value.

Solution

As is the norm, start by writing down the hypothesis.

$H_0 : \sigma_0^2 = 0.0025$

$H_1 : \sigma^2 \neq 0.0025$

Since the latter period has 24 months, n = 24, the test statistic is:

$$\chi^2_{24-1} = \frac{(24-1)\,0.0016}{0.0025} = 14.72$$

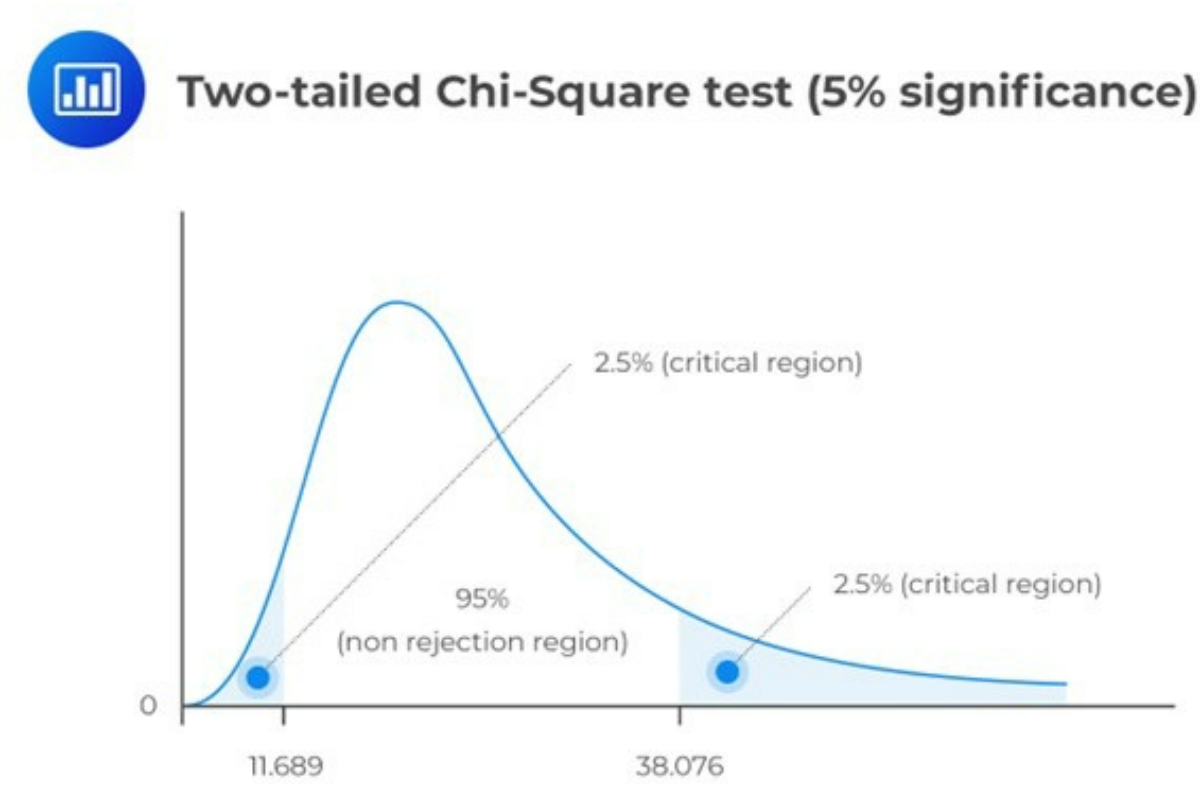
This is a two-tailed test. As such, we have to divide the significance level by two and screen our test statistic against the lower and upper 2.5% points of χ^2_{23}.

Consulting the chi-square table, the test statistic (14.72) lies between the lower (11.689) and the upper (38.076) 2.5% points of the chi-square distribution.

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Note that you will be given a simplified critical value table in the exam situation.



Evidently, we have insufficient evidence to reject the H_0. It is, therefore, reasonable to conclude that the latter standard deviation value is close enough to the 15-year value.

Test Concerning the Equality of the Variances

To test the equality concerning the variances, we use the F-test. Assume that we have 2 independent random samples of sizes n1 and n2 from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$.

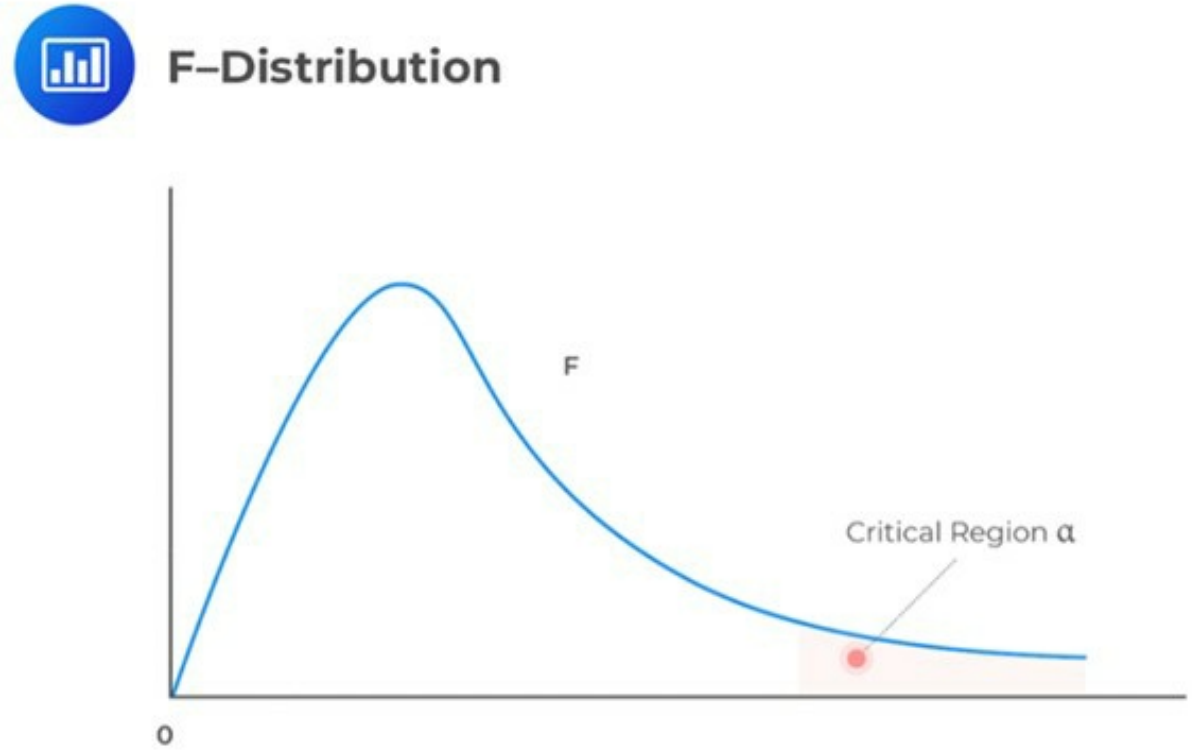
Also, let us consider a scenario where we have the sample variances as S_1^2 and S_2^2. The basic situation usually involves the following hypothesis:

$H_0 : \sigma_1^2 = \sigma_2^2$

$H_a : \sigma_1^2 \neq \sigma_2^2$

The test statistic is $\frac{S_1^2}{S_2^2} \sim F_{n1-1,n2-1}$ under H_0.

The decision rule is to reject the null hypothesis if the test statistic falls within the critical region of the F-distribution.



Example: F-test

An analyst is studying whether the population variance of returns on a commodity index changed after the introduction of new trading guidelines. The first 320 weeks elapsed before the guidelines were introduced, and the second 320 weeks came after the introduction. The analyst gathers the data in the table below for 320 weeks of returns both before and after the change in guidelines.

	Mean Weekly Return (%)	Variance of Returns
Before guidelines change	0.180	3.520
After guidelines change	0.090	2.967

Do the variances of returns differ before and after the guideline change? Employ a 5 percent significance level.

Solution

Step 1: State the hypothesis:

$$H_0 : \sigma^2_{Before} = \sigma^2_{After} \text{ vs } H_a : \sigma^2_{Before} \neq \sigma^2_{After}$$

Step 2: Identify the appropriate t-statistic:

$$t = \frac{s^2_{Before}}{s^2_{After}}$$

Step 3: Specify the level of significance:

$$\alpha = 5\% \text{ (two-tailed)}$$

Step 4: State the decision rule:

$$\text{Left side} = 0.803$$
$$\text{Right side} = 1.246$$

Reject the null if the calculated t-statistic is less than 0.803 and reject the null if the calculated t-statistic is greater than 1.246.

Step 5: Calculate the test statistic:

$$t = \frac{s^2_{Before}}{s^2_{After}} = \frac{3.520}{2.967} = 1.1864$$

Step 6: Make a decision:

Fail to reject the null hypothesis because 1.1864 falls within the bounds of the critical values of [0.80,1.246]. There is not sufficient evidence to indicate that the weekly variances of returns are different in the periods before and after the guidelines change.

LOS 8c: Compare and contrast parametric and nonparametric tests, and describe situations where each is the more appropriate type of test

Parametric Tests

Parametric tests are statistical tests in which we make assumptions regarding population distribution. Such tests involve the estimation of the key parameters of a distribution. For example, we may wish to estimate the mean or compare population proportions.

When conducting statistical tests, the choice of distribution directly influences how the test statistic is calculated. The tests we've discussed are considered parametric tests. For example, assuming a parameter follows a normal distribution leads to the computation of the z-statistic.

During parametric testing, approximating normal distributions for non-normal data may be required. This approximation is valid due to the central limit theorem, which states that as sample sizes increase, non-normal distributions tend to become more normal.

Parametric tests are generally considered to be stronger than nonparametric ones.

Nonparametric Tests

Nonparametric tests, sometimes known as distribution-free tests, don't assume anything about the parameter's distribution being studied. Researchers turn to nonparametric testing when they have concerns about factors other than the parameter's distribution.

The following table gives the alternative nonparametric tests for the parametric tests.

	Parametric Test	Non-parametric Test
Test concerning single mean	t-distributed test	Wilcoxon signed-rank test
Tests concerning differences between means	z-distributed test	Mann–Whitney U test (Wilcoxon rank sum test)
Test concerning mean differences (pair comparison test)	t-distributed test	Wilcoxon signed-rank test Sign test

Situations Where Nonparametric Tests are Appropriate

The data do not meet distributional assumptions:

This happens when the distributional assumptions of the parametric tests are not met. For instance, we may find parametric tests such as t-test are inappropriate because the sample size is small and may be drawn from non-normally distributed. As such, a nonparametric test is appropriate.

When there are outliers:

Outliers can affect the parametric statistics. On the other hand, outliers do not affect parametric tests.

Consider a situation where we want to establish the center of a rather skewed distribution, such as that of the income of the residents of a given city. While the majority of the residents could be categorized as the middle class, the presence of just a few billionaires in a sample can greatly increase the mean income. Such a mean, therefore, may not provide a very reliable or realistic measure of income.

Instead, it may be more appropriate to use the median. Compared to the mean, the median can better represent the center of the income distribution. This is due to the fact that 50% of the residents will be above the median and the remaining 50% below it.

In summary, “outliers” affect the mean when dealing with skewed data. The median, on the other hand, sticks closer to the center of the distribution.

When data is given in the form of ranks or uses an ordinal scale:

Although nonparametric tests are usually easier to conduct than parametric ones, they do not have as much statistical power. Nonetheless, they provide an efficient tool for analyzing ordinal, ranked, or very skewed data.

When the hypotheses do not concern a parameter:

We often use nonparametric tests, such as the runs test, when our goal is to determine if a sample from a population isn't random. Since randomness isn't a parameter, nonparametric tests are the right choice in such cases.

Nonparametric inference:

Nonparametric methods make our statistical analysis broader. They work with limited assumptions and can be used for ordered data. Plus, they handle questions that aren't tied to specific parameters.

Nonparametric tests are commonly used alongside parametric tests. They help analysts understand how sensitive the statistical results are to the assumptions of parametric tests. But when the conditions for a parametric test are met, we usually choose it over nonparametric tests. We prefer parametric tests because they often have more statistical power, which means they are better at detecting when the null hypothesis is incorrect.

Learning Module 9: Parametric and Non Parametric Tests of Independence

LOS 9a: explain parametric and non-parametric tests of the hypothesis that the population correlation coefficient equals zero and determine whether the hypothesis is rejected at a given level of significance

Parametric versus Non-parametric Tests of Independence

A parametric test is a hypothesis test concerning a population parameter used when the data has **specific distribution assumptions**. If these assumptions are not met, **non-parametric tests** are used.

In summary, researchers use non-parametric testing when:

- Data do not meet distributional assumptions.
- There are outliers.
- Data is given in the form of ranks.
- The hypothesis test objective does not concern a parameter.

Hypotheses Concerning Population Correlation Coefficient

We frequently compare the population correlation coefficient to zero when testing for correlation. This helps us determine whether there's a relationship between the variables. The population correlation coefficient, represented by ρ, is used to test the relationship. There are three possible hypotheses:

- Two-sided; $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$.
- One-sided right side; $H_0 : \rho \leq 0$ versus $H_a : \rho > 0$.
- One-sided left side; $H_0 : \rho \geq 0$ versus $H_a : \rho < 0$.

Let's assume that we have variables X and Y. The sample correlation, r_{XY}, tests the above

hypotheses.

Parametric Test of a Correlation

The **parametric pairwise correlation coefficient**, also known as **Pearson correlation**, is used to test the correlation in a parametric test. The formula for the sample correlation involves the sample covariance between the X and Y variables and their respective standard deviations, which is expressed as:

$$r = \frac{S_{XY}}{S_X S_Y}$$

Where:

S_{XY} = Sample covariance between the X and Y variables.

S_X = Standard deviation of the X variable.

S_Y = Standard deviation of the Y variable.

A **t-test** can determine if the null hypothesis should be rejected using the sample correlation, r if the two variables are **normally distributed**. The formula for the t-test is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}}$$

Where:

r= Sample correlation.

n= Sample size.

$(n-2)$= Degrees of freedom.

The test statistic follows a t-distribution with $n-2$ degrees of freedom. From the equation above, it is easy to see that the sample size, n, increases, and the degrees of freedom increase. In other words, as the sample size n increases, the power of the test increases. This implies that a false

null hypothesis is more likely to be rejected as the sample size increases.

Example: Parametric Test of a Correlation

The table below shows the sample correlations between the monthly returns of five different sector-specific exchange-traded funds (ETFs) and the overall market index (Market 1). There are 48 monthly observations, and the following ETFs are included in the analysis:

	ETF 1	ETF 2	ETF 3	ETF 4	ETF 5	Market 1
ETF 1	1					
ETF 2	0.8214	1				
ETF 3	0.5672	0.6438	1			
ETF 4	0.4276	0.5789	0.4123	1		
ETF 5	0.7121	0.7942	0.6896	0.5614	1	
Market 1	0.8375	0.9096	0.7223	0.6954	0.7919	1

Using a 1% significance level and the following hypotheses: $H_0 : \rho = 0$ versus $H_a : \rho \neq 0$, calculate the t-statistic for the correlation between ETF 2 and ETF 4. Based on the calculated t-statistic, draw a conclusion about the significance of the correlation using the following sample t-table:

df	p = 0.10	p = 0.05	p = 0.025	p = 0.01	p = 0.005
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
36	1.306	1.688	2.028	2.434	2.719
37	1.305	1.687	2.026	2.431	2.715
38	1.304	1.686	2.024	2.429	2.712
39	1.304	1.685	2.023	2.426	2.708
40	1.303	1.684	2.021	2.423	2.704
41	1.303	1.683	2.020	2.421	2.701
42	1.302	1.682	2.018	2.418	2.698
43	1.302	1.681	2.017	2.416	2.695
44	1.301	1.680	2.015	2.414	2.692
45	1.301	1.679	2.014	2.412	2.690
46	1.300	1.679	2.013	2.410	2.687
47	1.300	1.678	2.012	2.408	2.685
48	1.299	1.677	2.011	2.407	2.682

Solution

To test the significance of the correlation between ETF 2 and ETF 4, we will use the t-test formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}}$$

Where:

r = Sample correlation coefficient (in this case, $r_{EFT2,ETF4} = 0.5789$).

n = Number of observations (48 in this case).

Now, let's calculate the t-statistic:

$$t = \frac{r\sqrt{n-2}}{\sqrt{(1-r^2)}} = \frac{0.5789\sqrt{48-2}}{\sqrt{1-0.5789^2}} = 4.815$$

The calculated t-statistic for the correlation between ETF2 and ETF4 is 4.815.

At the 1% significance level, with a two-tailed test and degrees of freedom,

df = $n-2 = 46$, the critical t-value is approximately ± 2.687.

Conclusion: We reject the null hypothesis since our calculated t-statistic (4.815) is greater than the critical value (+2.687). This indicates sufficient evidence to suggest that the correlation between ETF 2 and ETF 4 significantly differs from zero.

Non-Parametric Test of Correlation: The Spearman Rank Correlation Coefficient

The Spearman rank correlation coefficient, r_S, is a non-parametric test used to examine the relationship between two data sets when the population deviates from normality.

The Spearman rank correlation coefficient is like the Pearson correlation coefficient. The difference is that the Spearman coefficient is calculated based on the ranks of variables in the samples.

Consider two variables, X and Y. We need to calculate Spearman's Rank Correlation r_S.

Steps of Calculating Spearman's Rank Correlation Coefficient, r_S

- Rank the observations of each variable X and Y in descending order. Note that when there are tied values in the data, their ranks are calculated by taking the average of the ranks that would have been assigned to those values if they were not tied.
- Find the difference between the ranks for each pair of observations.
- Square the difference and calculate the sum of the difference, that is $\sum d_i$.
- Use the following formula to find r_S:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

Where;d_i=The difference between the ranks for each pair of observationsn= Sample size.

Example: Calculating Spearman's Rank Correlation Coefficient

An analyst is studying the relationship between returns for two sectors, steel and cement, over the past 5 years by using Spearman's rank correlation coefficient. The hypotheses are $H_0 : r_S = 0$ and $H_a : r_S \neq 0$. The returns of both sectors are provided below.

Year	Steel sector returns	Cement sector returns
1	10%	8%
2	6%	7%
3	9%	5%
4	12%	6%
5	8%	9%

The Spearman's rank correlation coefficient is *closest to*:

Solution

Year	Steel sector returns (X)	Cement sector returns (Y)	Rank order for X	Rank order for Y	D	d^2
1	10%	8%	2	2	0	0
2	6%	7%	5	3	2	4
3	9%	5%	3	5	−2	4
4	12%	6%	1	4	−3	9
5	8%	9%	4	1	3	9
					Sum=	26

We can now use the formula:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - [\frac{(6 \times 26)}{5 \times (5^2-1)}] = 1 - 1.3$$
$$r_s = -0.3$$

This indicates a very weak negative correlation between the returns of the steel and cement sectors.

Hypothesis Test for the Spearman Rank Correlation

The hypothesis test on the Spearman Rank depends on the sample size. If the sample size is small ($n \leq 30$), we would need a specialized table of critical value. On the other hand, if the sample size is large ($n > 30$), we can perform a t-test using the test statistic similar to that of Pearson correlation:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{(1-r_s^2)}}$$

Consider the above example. Assume we want to conduct a hypothesis test at a 5% significance level. The hypotheses statement is $H_0 : r_S = 0$ and $H_a : r_S \neq 0$

Question

Assume an investment analyst, John Smith, is studying the relationship between two stocks, X and Y. Based on 100 observations, he has found that $S_{XY} = 10, S_X = 2$, and $S_Y = 8$. Smith needs to find the sample correlation r_{XY} and use it to perform a t-test to determine if there is a significant correlation between the returns of stocks X and Y. The critical value for the test statistic at the 0.05 level of significance is approximately 1.96. He should conclude that the statistical relationship between X and Y is:

A. Significant because the test statistic falls outside the range of the critical values.
B. Significant, because the absolute value of the test statistic is less than the critical value.
C. Insignificant because the test statistic falls outside the range of the critical values.

Solution

The correct answer is A.

Note that the sample correlation coefficient, r_{XY} is calculated using the following formula:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

Substituting the given values in this formula, we get:

$$r_{XY} = \frac{10}{2 \times 8} = 0.625$$

To test the significance of the sample correlation, we can use a t-test with the following null and alternative hypotheses: $H_0 = \rho = 0$ and $H_\propto = \rho \neq 0$

The test statistic for this test is calculated using the following formula:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where:

r = Sample correlation coefficient.

n = The Number of observations.

Substituting the given values into this formula, we get:

$$t = \frac{0.625\sqrt{100-2}}{\sqrt{1-0.625^2}} \approx \frac{6.1872}{0.7806} = 7.9262$$

The critical value for the test statistic at the 0.05 level of significance is approximately 1.96.

Since our calculated test statistic (7.9262) is greater than the upper bound of the critical values for the test statistic (1.96), we reject the null hypothesis. This indicates sufficient evidence to suggest that the correlation between X and Y is significantly different from zero.

Therefore, John Smith should conclude that the statistical relationship between X and Y is significant because the test statistic falls outside the range of the critical values (**Option A**).

LOS 9b: Explain tests of independence based on contingency table data

With categorical or discrete data, correlation is not suitable for assessing relationships between variables. Instead, we use a non-parametric test called the chi-square test of independence, which employs a chi-square distributed test statistic.

We employ a contingency table to structure the data when examining the connection between two categorical variables. Subsequently, we apply a test of independence utilizing a chi-square distribution to assess whether a noteworthy relationship exists between these variables. The test statistic is calculated as follows:

$$\chi = \sum_{i=1}^{m} \frac{(O_{ij} - E_{ij})^2}{(E_{ij})}$$

Where:

$$E_{ij} = \frac{(\text{Total row i}) \times (\text{Total column j})}{\text{Overall Total}}$$

m= Number of cells in the table, the Number of groups in the first class, multiplied by the number of groups in the second class.

O_{ij}= Number of observations in each cell of row i and column j (i.e., observed frequency).

E_{ij}= Expected number of observations in each cell of row i and column j, assuming independence (i.e., expected frequency).

The degrees of freedom are given by:

$$\text{Degrees of freedom} = (r - 1)(c - 1)$$

Where:

r= Number of rows.

c= Number of columns.

Example: Testing Independence Based on Contingency Table Data

The following contingency table shows the responses of two categories of investors (employed vs. retired) with regard to their primary investment objectives (growth, income, or both). The total sample size is 173.

	Growth	Income	Both	Total
Employed	52	25	10	87
Retired	32	47	7	86
Total	84	72	17	173

Use a 95% significance level to test whether there is any significant difference between employed and retired investors concerning primary investment objectives.

Solution

H_0: There is no significant difference between employed and retired investors with regard to primary investment objectives.

H_α: There is a significant difference between employed and retired investors with regard to primary investment objectives.

Step 1: We calculate the expected frequency of investors by their category (employed vs. retired) and investment objective using the following formula:

$$E_{ij} = \frac{\text{(Total row i)} \times \text{(Total column j)}}{\text{Overall Total}}$$

	Growth	Income	Both	Total
Employed	$\frac{(87\times84)}{173} = 42.24$	$\frac{(87\times72)}{173} = 36.20$	$\frac{(87\times17)}{173} = 8.55$	87
Retired	$\frac{(86\times84)}{173} = 41.75$	$\frac{(86\times72)}{173} = 35.79$	$\frac{(86\times17)}{173} = 8.45$	86
Total	84	72	17	173

Step 2: We calculate the scaled squared deviation for each combination of investor category and investment objective as follows:

	Growth	Income	Both
Employed	$\frac{(52-42)^2}{42} = 2.254$	$\frac{(25-36)^2}{36} = 0.469$	$\frac{(10-9)^2}{9} = 0.246$
Retired	$\frac{(32-42)^2}{42} = 2.280$	$\frac{(47\times36)^2}{36} = 3.510$	$\frac{(7-8)^2}{8} = 0.349$
Total	4.534	6.979	0.495

Step 3: We calculate the value of χ^2:

$$\chi^2 = 4.534 + 6.979 + 0.495 = 12.008$$

Step 4: The critical value of X^2 is 5.99. It is determined as follows:

- There are $(r-1)(c-1) = (2-1) \times (3-1) = 2$ degrees of freedom.
- It is a one-sided test with a 5% level of significance.

Chi-Square (χ^2) Distribution

Degrees of Freedom	Area to the Right of Critical Value							
	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
1	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635
2	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210
3	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345
4	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277
5	0.554	0.831	1.145	1.610	9.236	11.071	12.833	15.086
6	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812
7	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475
8	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090
9	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666
10	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209
11	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725
12	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217
13	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688
14	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141
15	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578
16	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000
17	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409
18	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805
19	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191
20	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566
21	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932
22	9.542	10.982	12.338	14.042	30.813	33.924	36.781	40.289
23	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638
24	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980
25	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314
26	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642
27	12.879	14.573	16.151	18.114	36.741	40.113	43.194	46.963
28	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278
29	14.257	16.047	17.708	19.768	39.087	42.557	45.722	49.588
30	14.954	16.791	18.493	20.599	40.256	43.773	46.979	50.892

Decision rule: The calculated value of $\chi^2 = 12.008$ is greater than the critical value of 5.99. As such, there is sufficient evidence to support the conclusion that retired investors and employed investors have different primary investment objectives.

Question

Regarding the chi-square test of independence, which statement is accurate? The chi-square test of independence is:

A. A parametric hypothesis test.
B. Used to test whether two categorical variables are related to each other.
C. Used to test whether two continuous variables are related to each other.

Solution

The correct answer is B. The chi-square test of independence is a non-parametric hypothesis test that can be used to test whether two categorical variables are related.

A is incorrect because the chi-square test of independence is non-parametric, not parametric.

C is incorrect because the chi-square test of independence is used for categorical variables, not continuous variables.

Learning Module 10: Simple Linear Regression

LOS 10a: describe a simple linear regression model, how the least squares criterion is used to estimate regression coefficients, and the interpretation of these coefficients

Linear regression is a mathematical method used for analyzing how the variation in one variable can explain the variation in another variable.

Let Y be the variable we wish to explain. As such, the observation of this variable is Y_i, and $\bar{Y}$ is the mean of the sample size n. The variation of Y is given by:

$$\text{Variation of } Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

Our main objective is to explain what causes this variation, usually called the sum of squares total (SST).

By definition of the regression, we need to explain the variation of Y with another variable. Let X be the explanatory variable. As such, the observations of X will be denoted by X_i and $\bar{X}$ sample mean of size n. The variation of X is given by:

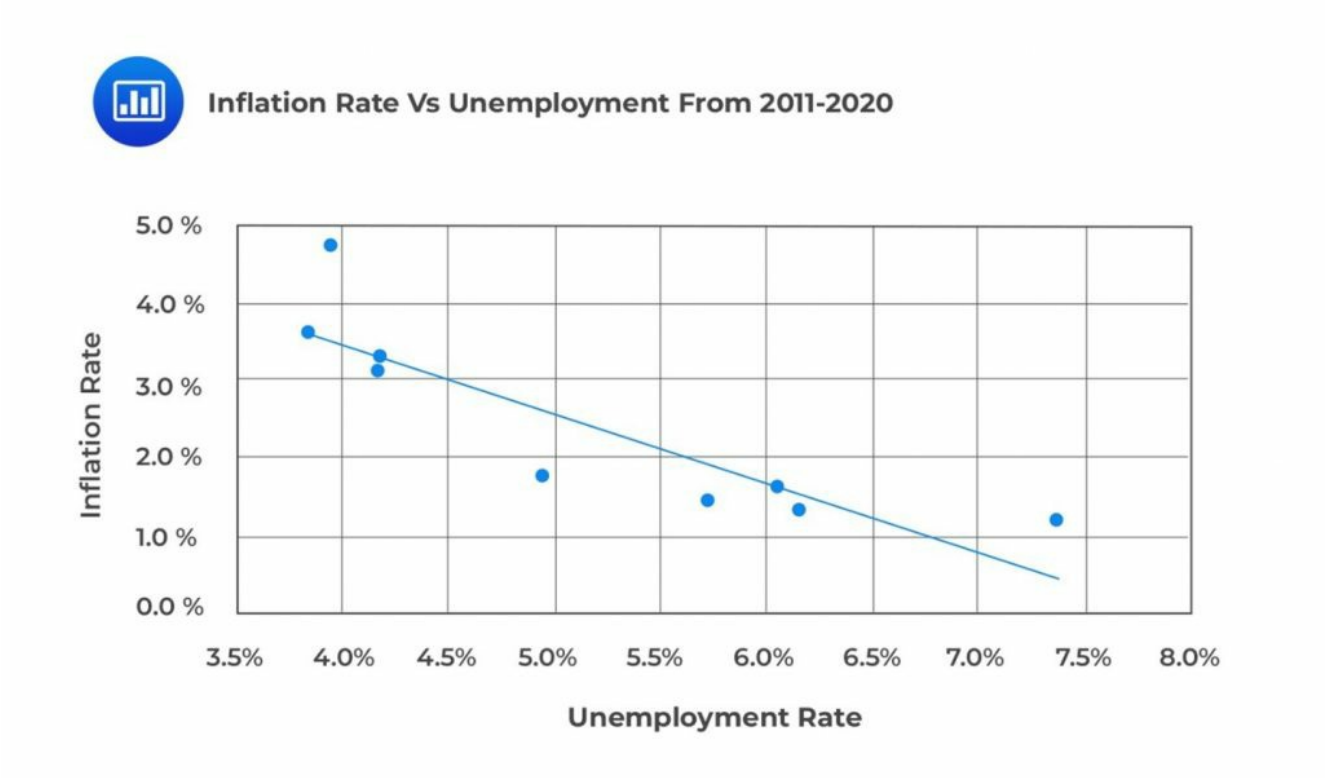
$$\text{Variation of } X = \sum_{i=1}^{n} (X_i - \bar{X})^2$$

To visualize the relationship between variables X and Y, you can use a scatter plot, also known as a scattergram. In this type of plot, the variable you want to explain (Y) is usually plotted on the vertical axis. In contrast, the explanatory variable (X) is placed on the horizontal axis to show the relationship between their variations.

For example, consider the following table. We wish to use linear regression analysis to forecast inflation, given unemployment data from 2011 to 2020.

Year	Unemployment Rate	Inflation Rate
2011	6.1%	1.7%
2012	7.4%	1.2%
2013	6.2%	1.3%
2014	6.2%	1.3%
2015	5.7%	1.4%
2016	5.0%	1.8%
2017	4.2%	3.3%
2018	4.2%	3.1%
2019	4.0%	4.7%
2020	3.9%	3.6%

In this scenario, the Y variable is the inflation rate, and the X axis is the unemployment rate. A scatter plot of the inflation rates against unemployment rates from 2011 to 2020 is shown in the following figure.



Dependent and Independent Variables

A dependent variable, often denoted as YYY, is the variable we want to explain. In contrast, an independent variable, typically denoted as XXX, is used to explain variations in the dependent

variable. The independent variable is also referred to as the exogenous, explanatory, or predicting variable.

In our example above, the dependent variable is the inflation rate, and the independent variable is the unemployment rate.

To understand the relationship between dependent and independent variables, we estimate a linear relationship, usually a straight line. When there's one independent variable, we use simple linear regression. If there are multiple independent variables, we use multiple regression.

This reading focuses on linear regression.

Least Squares Criterion

In simple linear regression, we assume linear relationships exist between the dependent and independent variables. The aim is to fit a line to the observations of X (X_is) and Y (Y_is) to minimize the squared deviations from the line. To accomplish this, we use the least squares criterion.

The following is a simple linear regression equation:

$$Y = b_0 + b_1X_1 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

Where:

Y = Dependent variable.

b_0 = Intercept.

b_1 = Slope coefficient.

X = Independent variable.

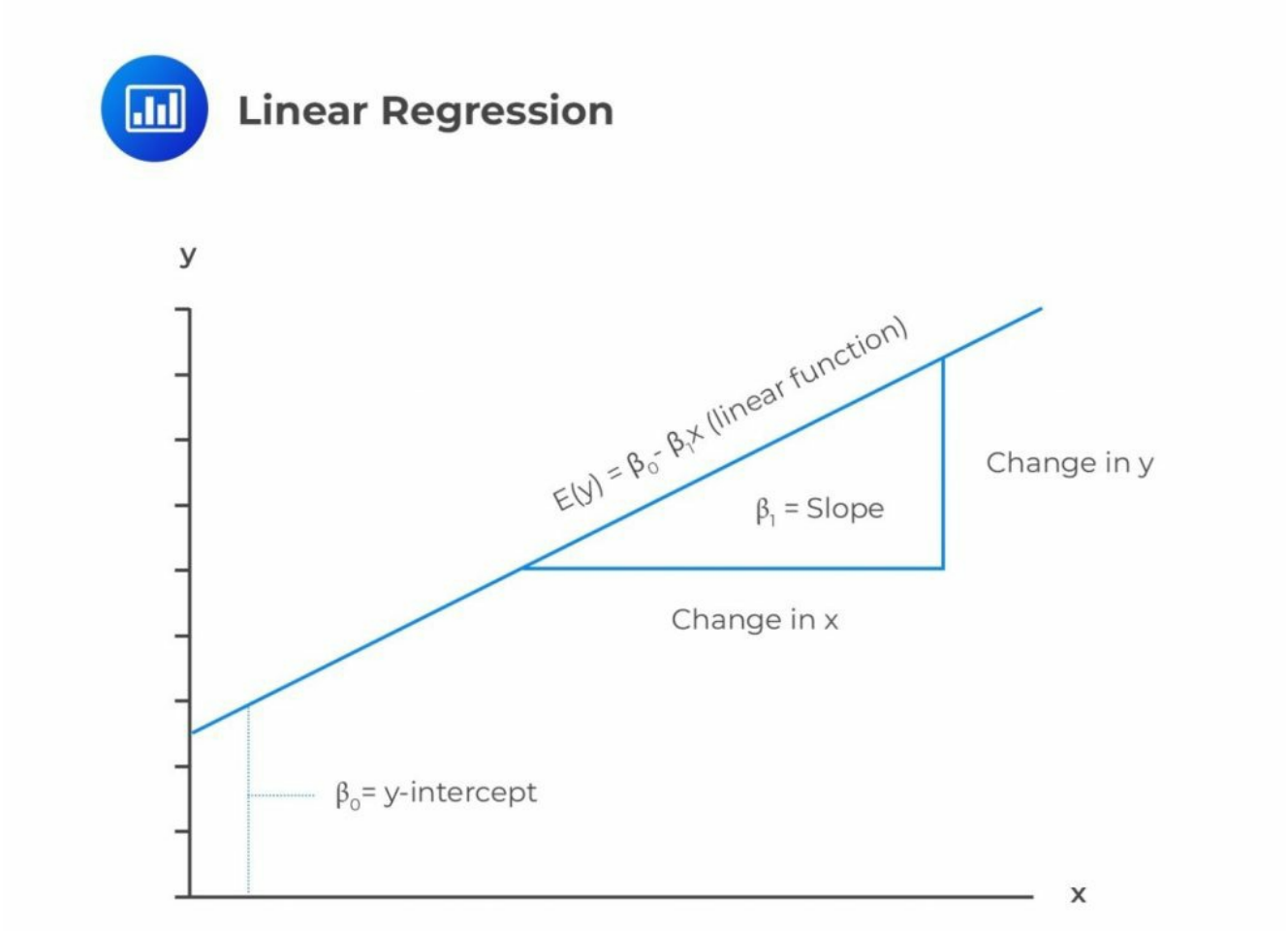
ε = Error term (Noise).

b_0 and b_1 are known as **regression coefficients**. The equation above implies that the dependent

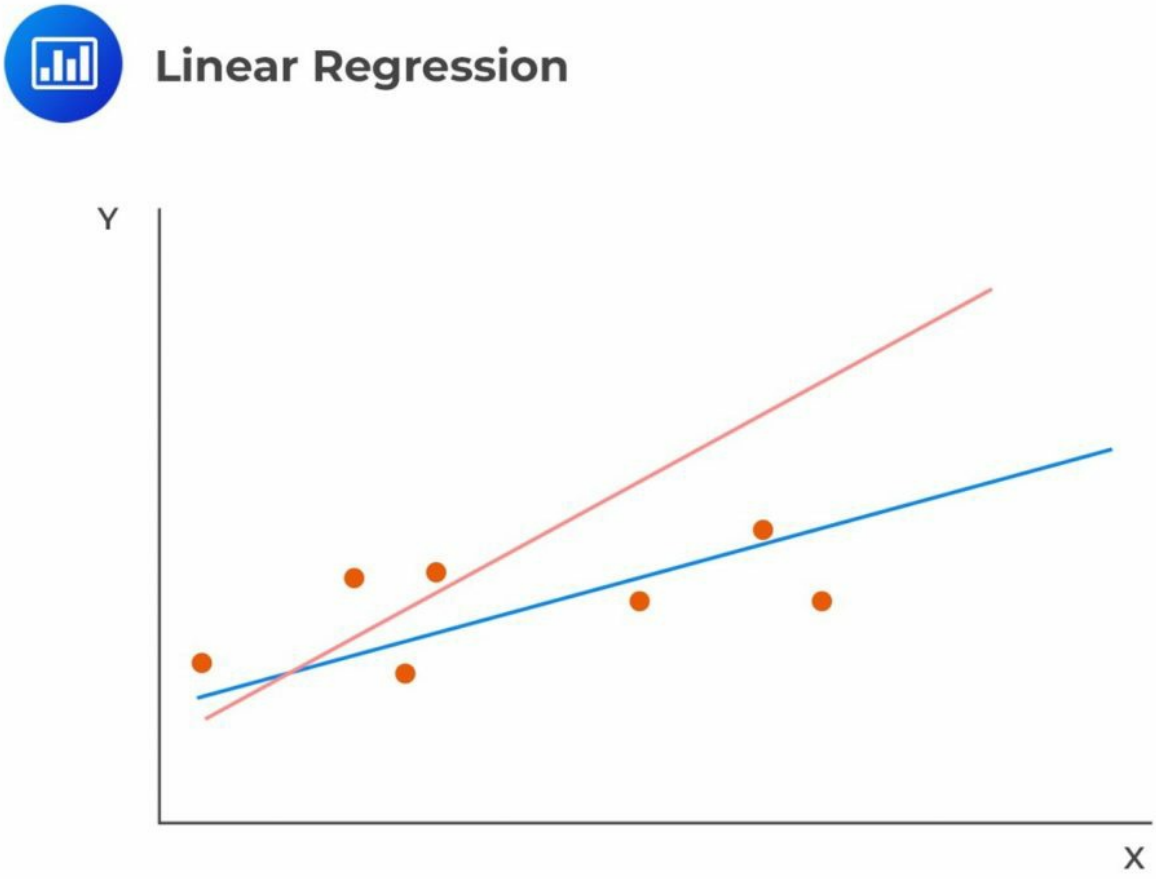
is equivalent to the intercept (b_0) plus the product of the slope coefficient (b_1) and the independent variable plus the error term.

The error term is equal to the difference between the observed value of Y and the one expected from the underlying population relation between X and Y

Below is an illustration of a simple linear regression model.



As stated earlier, linear regression calculates a line that best fits the observations. In the following image, the line that best fits the regression is clearly the blue one:



Note that we cannot directly observe the population parameters b_0 and b_1. As such, we observe their estimates, $\hat{b}_0$ and $\hat{b}_1$. They are the estimated parameters of the population using a sample. In simple linear regression, $\hat{b}_0$ and $\hat{b}_1$ are such that the sum of squared vertical distances is minimized.

Specifically, we concentrate on the sum of the squared differences between observations Y_i and the respective estimated value $\hat{Y}_i$ on the regression line, also called the sum of squares error (SSE).

Note that,

$$\hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i + e_i^2$$

As such,

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{b}_0 - \hat{b}_1 X_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Note that the residual for the ith observation ($e_i = Y_i - \hat{Y}_i$) is different from the error term (ε_i). The error term is based on the underlying population, while the residual term results from regression analysis on a sample.

Conventionally, the sum of the residuals is zero. As such, the aim is to fit the regression line in a simple linear regression that minimizes the sum of squared residual terms.

Estimation and Interpretation of Regression Coefficients

The Slope Coefficient $\hat{\beta}_1$

For a simple linear regression, the slope coefficient is estimated as the ratio of the $\text{Cov}(X, Y)$ and $\text{Var}(X)$:

$$\hat{b}_1 = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{n-1}}{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

The slope coefficient is defined as the change in the dependent variable caused by a one-unit change in the value of the independent variable.

The Intercept $\hat{\beta}_0$

The intercept is estimated using the mean of X and Y as follows:

$$\hat{b}_0 = \bar{Y} - \hat{b}_1 \bar{X}$$

Where:

$\bar{Y}$ = Mean of Y.

$\bar{X}$ = Mean of X.

The intercept is the estimated value of the dependent variable when the independent variable is zero. The fitted regression line passes through the point equivalent to the means of the dependent and the independent variables in a linear regression model.

Example: Estimating Regression Line

Let us consider the following table. We wish to estimate a regression line to forecast inflation given unemployment data from 2011 to 2020.

Year	Unemployment Rate% (X_is)	Inflation Rate% (Y_is)
2011	6.1	1.7
2012	7.4	1.2
2013	6.2	1.3
2014	6.2	1.3
2015	5.7	1.4
2016	5.0	1.8
2017	4.2	3.3
2018	4.2	3.1
2019	4.0	4.7
2020	3.9	3.6

We can create the following table:

Year	Unemployment Rate% (X_is)	Inflation Rate% (Y_is)	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})$ $(X_i - \bar{X})$
2011	6.1	1.7	0.410	0.656	−0.518
2012	7.4	1.2	1.300	4.452	−2.405
2013	6.2	1.3	1.082	0.828	−0.946
2014	6.2	1.3	1.082	0.828	−0.946
2015	5.7	1.4	0.884	0.168	−0.385
2016	5.0	1.8	0.292	0.084	0.157
2017	4.2	3.3	0.922	1.188	−1.046
2018	4.2	3.1	0.578	1.188	−0.828
2019	4.0	4.7	5.570	1.664	−3.044
2020	3.9	3.6	1.588	1.932	−1.751
Sum	52.90	23.4	13.704	12.989	−11.716
Arithmetic Mean	5.29	2.34			

From the table above, we estimate the regression coefficients:

$$\hat{b}_1 = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \frac{\sum_{i=1}^{n}(Y_i - Y)(X_i - X)}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{-11.716}{12.989} = -0.9020$$

$$\hat{b}_0 = \bar{Y} - \hat{b}_1\bar{X} = 2.34 - (-0.9020) \times 5.29 = 7.112$$

As such, the regression model is given by:

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

From the above regression model, we can note the following:

- The inflation rate is 7.112% if the unemployment rate is 0% (theoretically speaking).
- If the unemployment rate increases (decreases) by one unit–say, from 2% to 3%–the inflation rate decreases(increases) by 0.9020%.

In general,

- If the slope is positive, a unit increase(decrease) in the independent variable results in an increase(decrease) in the dependent variable.
- If the slope is negative, a one-unit increase(decrease) in the independent variable results in a decrease(increase) in the dependent variable.

Furthermore, with the estimated regression model, we can predict the values of the dependent variable based on the value of the independent variable. For instance, if the unemployment rate is 4.5%, then the predicted value of the dependent variable is:

$$\hat{Y} = 7.112 - 0.9020 \times 4.5 = 3.05\%$$

In practice, analysts use statistical functions in software like Excel, statistical tools like R, or programming languages such as Python to perform regression analysis.

Cross-sectional and Time Series Regressions

Regression analysis is commonly used with cross-sectional and time series data. In cross-

sectional analysis, you compare X and Y observations from different entities, like various companies in the same time period. For instance, you might analyze the link between a company's R&D spending and its stock returns across multiple firms in a single year.

Time-series regression analysis involves using data from various time periods for the same entity, like a company or an asset class. For instance, an analyst might examine how a company's quarterly dividend payouts relate to its stock price over multiple years.

Question

The independent variable in a regression model is *most likely* the:

A. Predicted variable.
B. Predicting variable.
C. Endogenous variable.

Solution

The correct answer is B.

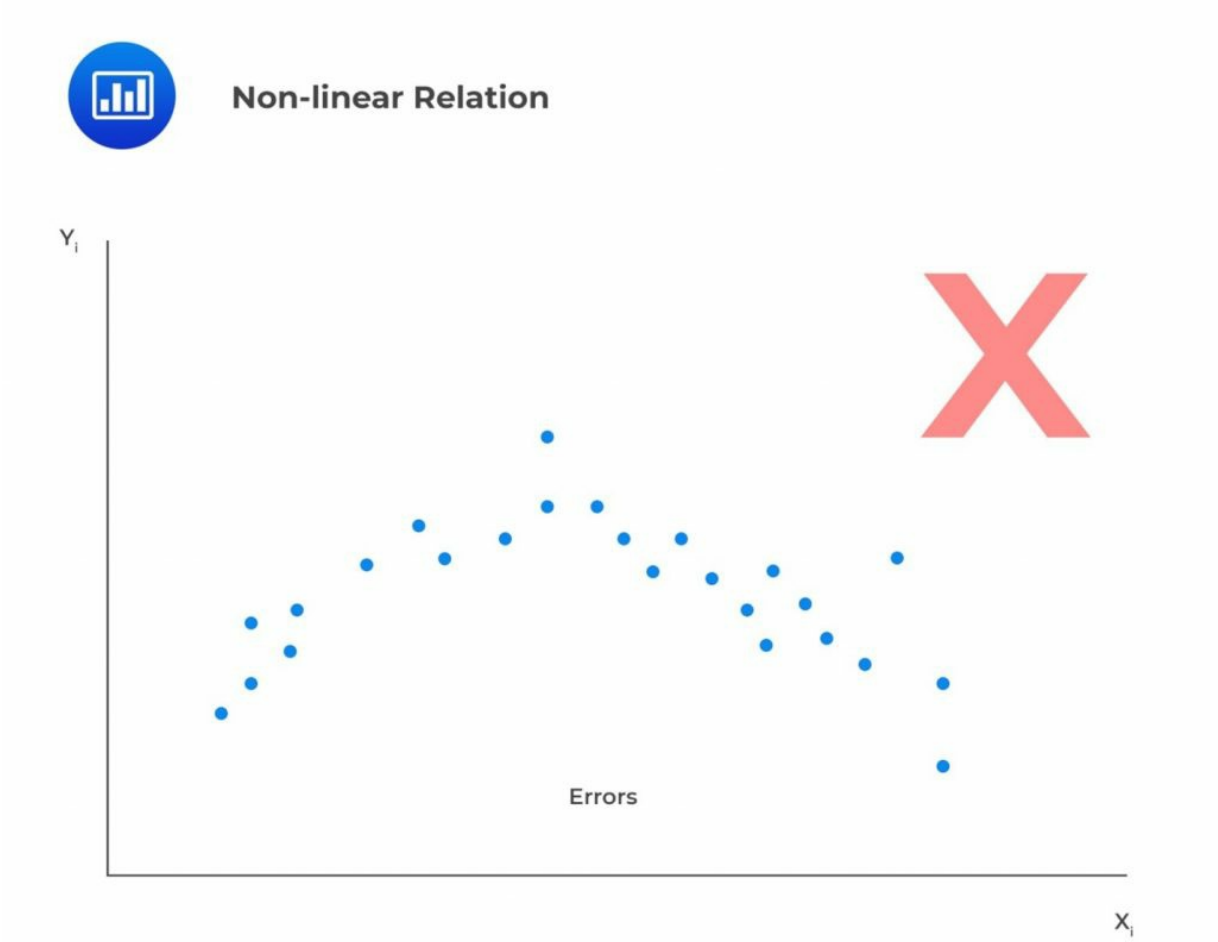
An independent variable explains the variation of the dependent variable. It is also called the explanatory variable, exogenous variable, or the **predicting variable**.

A and C are incorrect. A dependent variable is a variable predicted by the independent variable. It is also known as the **predicted variable**, explained variable, or **endogenous variable**.

LOS 10b: explain the assumptions underlying the simple linear regression model, and describe how residuals and residual plots indicate if these assumptions may have been violated

Assume that we have samples of size n for dependent variable Y and independent variable X. We wish to estimate the simple regression of Y and X. The classic normal linear regression model assumptions are as follows:

I. **Linearity**: A linear relationship implies that the change in Y due to a one-unit change in X is constant, regardless of the value X takes. If the relationship between the two is not linear, the regression model will not accurately capture the trend, resulting in inaccurate predictions. The model will be biased and underestimate or overestimate Y at various points. For example, the model $Y = b_0 + b_1 e^{b_1 x}$ is nonlinear in b_1. For this reason, we should not attempt to fit a linear model between X and Y. It also follows that the independent variable, X, must be non-stochastic (must not be random). A random independent variable rules out a linear relationship between the dependent and independent variables.In addition, linearity means the residuals should not exhibit an observable pattern when plotted against the independent variable. Instead, they should be completely random. In the example below, we're looking at a scenario where the residuals appear to show a pattern when plotted against the independent variable, X. This effectively indicates a nonlinear relation.



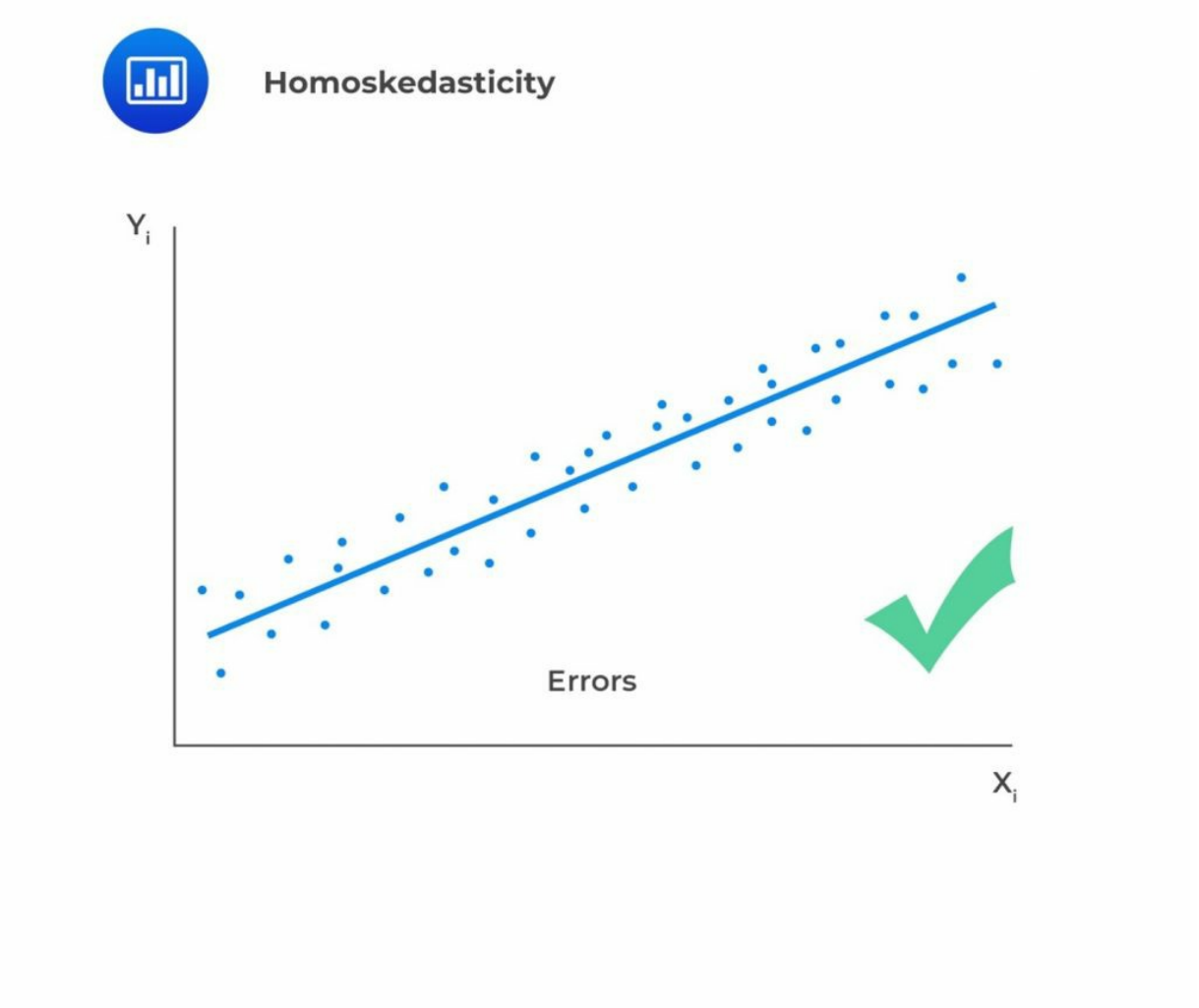
II. **Normality Assumption**: This assumption implies that the error terms (residuals) must follow a normal distribution. It's important to note that this doesn't mean the dependent and independent variables must be normally distributed. However, it's crucial to check the distribution of the dependent and independent variables to identify any outliers. A histogram of the residuals can be used to detect if the error term is normally distributed. A symmetric bell-shaped histogram indicates that the normality assumption is likely to be true.

III. **Homoskedasticity**: Homoskedasticity implies that the variance of the error terms is **constant** across all observations. Mathematically, this is expressed as:

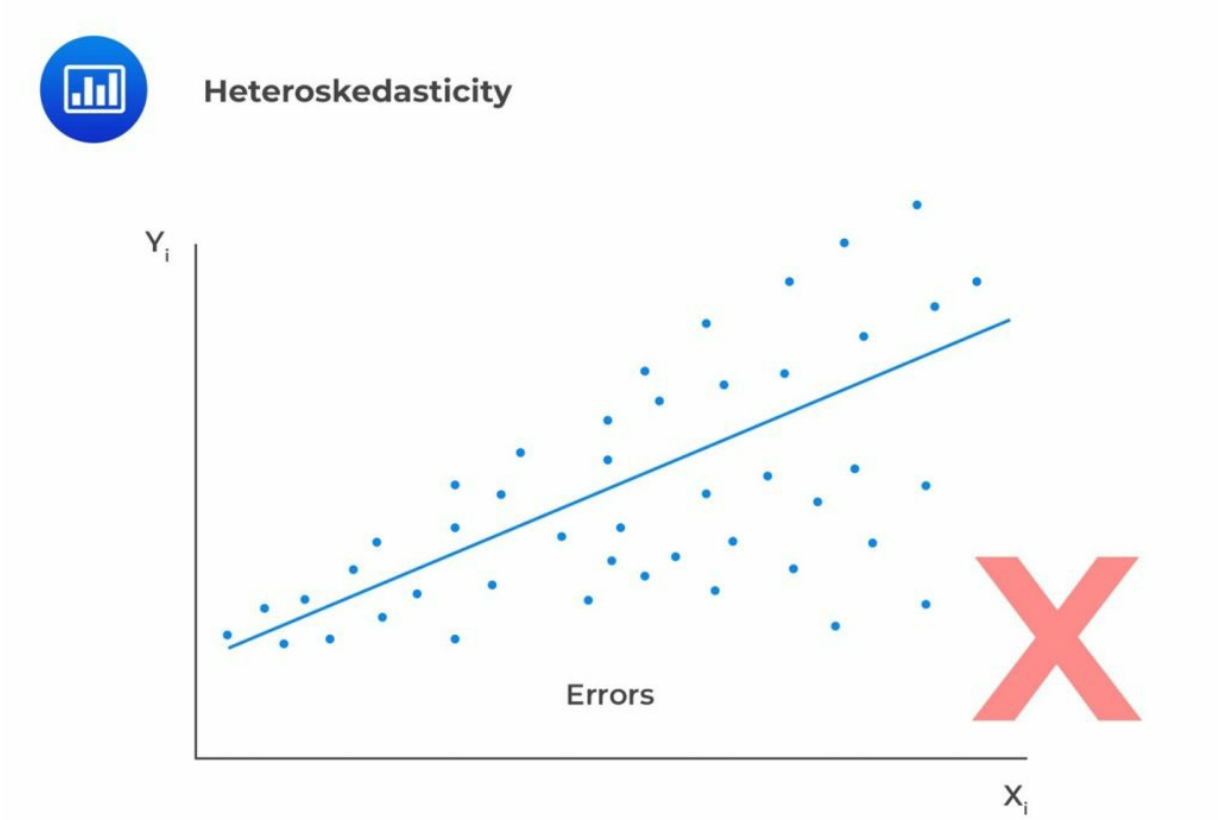
$$E(\epsilon_i^2) = \sigma_\epsilon^2, \quad i = 1, 2, \ldots, n$$

If the variance of residuals varies across observations, then we refer to this as heteroskedasticity (not homoscedasticity). We plot the least square residuals against the

independent variable to test for heteroscedasticity. If there is an evident pattern in the plot, that is a manifestation of heteroskedasticity.



In case residuals and the predicted values increase simultaneously, then such a situation is known as **heteroscedasticity** (or heteroskedasticity).



IV. **Independence Assumption**: The independence assumption implies that the observations X_i and Y_i are independent of each other. Failure to satisfy this assumption implies the variables are not independent, and thus, residuals will be correlated. To ascertain this assumption, we visually and statistically analyze the residuals to check whether residual shows exhibit a pattern.

Question

A regression model with one independent variable requires several assumptions for valid conclusions. Which of the following statements *most likely* violates those assumptions?

A. The independent variable is random.
B. The error term is distributed normally.
C. There exists a linear relationship between the dependent variable and the independent variable.

Solution

The correct answer is A.

Linear regression assumes that the independent variable, X, is NOT random. This ensures that the model produces the correct estimates of the regression coefficients.

B is incorrect. The assumption that the error term is distributed normally allows us to easily test a particular hypothesis about a linear regression model.

C is incorrect. Essentially, the assumption that the dependent and independent variables have a linear relationship is the key to a valid linear regression. If the parameters of the dependent and independent variables are not linear, then the estimation of that relation can yield invalid results.

LOS 10c: calculate and interpret measures of fit and formulate and evaluate tests of fit and of regression coefficients in a simple linear regression

The sum of Squares Total (SST) and Its Components

The sum of Squares Total (total variation) is a measure of the total variation of the dependent variable. It is the sum of the squared differences of the **actual** y-value and **mean** of y-observations.

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

The Sum of Squares Total contains two parts:

i. The Sum of Square Regression (SSR).
ii. The sum of Squares Error (SSE).

i. **The sum of Squares Regression (SSR)**: The sum of squares regression measures the **explained** variation in the dependent variable. It is given by the sum of the squared differences of the **predicted** y-value $\hat{Y}_i$, and **mean** of y-observations, $\bar{Y}$:

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

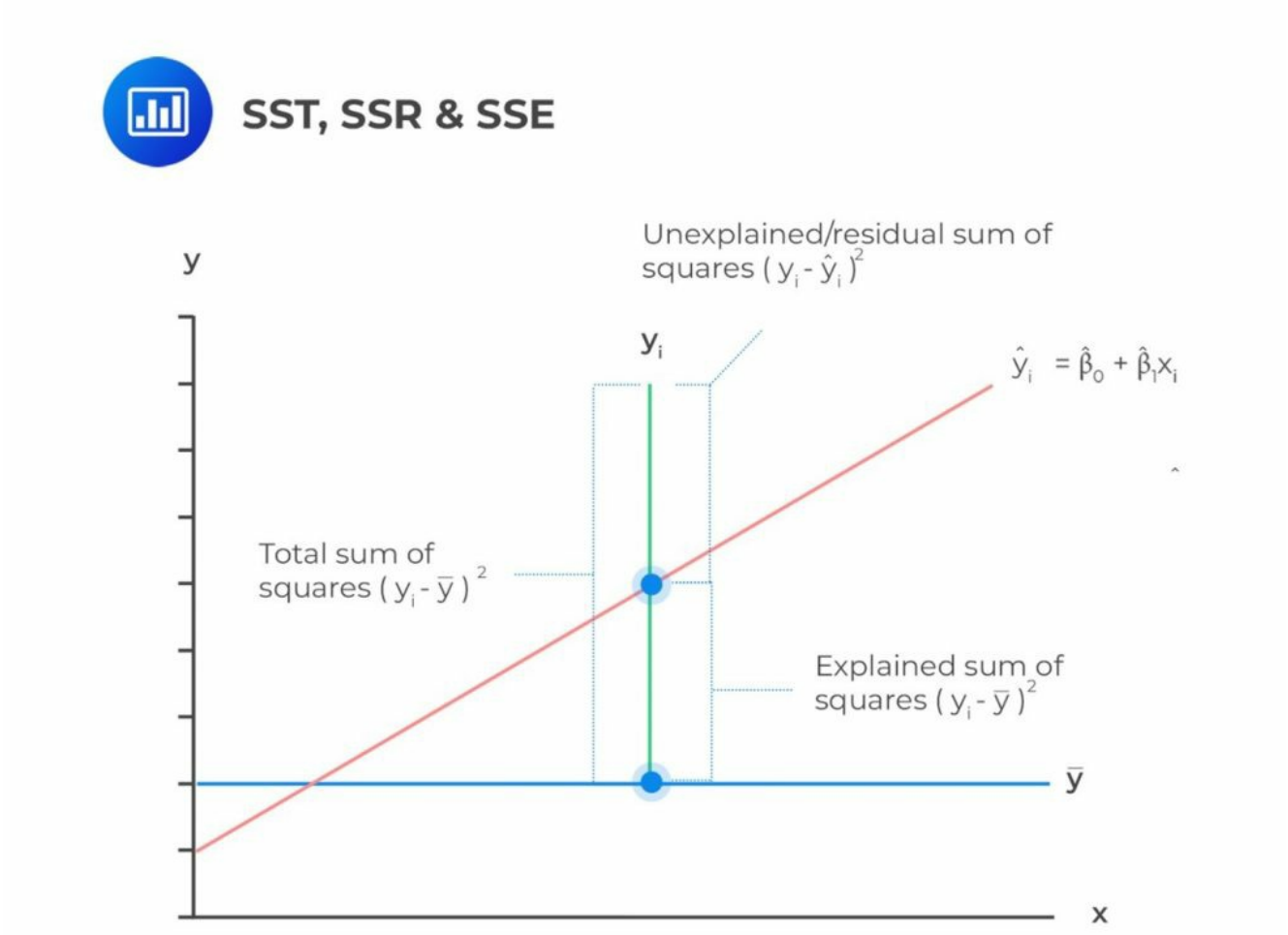
ii. **The Sum of Squared Errors (SSE)**: The sum of squared errors is also called the residual sum of squares. It is defined as the variation of the dependent variable **unexplained** by the independent variable. SSE is given by the sum of the squared differences of the **actual** y-value (Y_i) and the **predicted** y-values, $\hat{Y}_i$.

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Therefore, the sum of squares total is given by:

$$\text{Sum of Squares Total} = \text{Explained Variation} + \text{Unexplained Variation}$$
$$= \text{SSR} + \text{SSE}$$

The components of the total variation are shown in the following figure.



For example, consider the following table. We wish to use linear regression analysis to forecast inflation, given unemployment data from 2011 to 2020.

Year	Unemployment Rate (%)	Inflation Rate (%)
2011	6.1	1.7
2012	7.4	1.2
2013	6.2	1.3
2014	6.2	1.3
2015	5.7	1.4
2016	5.0	1.8
2017	4.2	3.3
2018	4.2	3.1
2019	4.0	4.7
2020	3.9	3.6

Remember that we had estimated the regression line to be $\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$. As such, we can create the following table:

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
2013	6.2	1.3	1.520	1.082	0.048	0.673
2014	6.2	1.3	1.520	1.082	0.048	0.673
2015	5.7	1.4	1.971	0.884	0.326	0.136
2016	5.0	1.8	2.602	0.292	0.643	0.069
2017	4.2	3.3	3.324	0.922	0.001	0.967
2018	4.2	3.1	3.324	0.578	0.050	0.967
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

From the table above, we can calculate the following:

$$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 13.704$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 10.568$$

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 3.136$$

Measures of Goodness of Fit

We use the following measures to analyze the goodness of fit of simple linear regression:

I. Coefficient of determination.
II. F-statistic for the test of fit.
III. Standard error of the regression.

Coefficient of Determination

The coefficient of determination (R^2) measures the proportion of the total variability of the dependent variable explained by the independent variable. It is calculated using the formula below:

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{\text{Sum of Squares Regression (SSR)}}{\text{Sum of Squares Total (SST)}} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Intuitively, we can think of the above formula as:

$$R^2 = \frac{\text{Total Variation} - \text{Unexplained Variation}}{\text{Total Variation}} = \frac{\text{Sum of Squares Total (SST)} - \text{Sum of Squared Errors (SSE)}}{\text{Sum of Squares Total}}$$

Simplifying the above formula gives:

$$R^2 = 1 - \frac{\text{Sum of Squared Errors (SSE)}}{\text{Sum of Squares Total (SST)}}$$

In the above example, the coefficient of determination is:

Explained Variation

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}$$
$$= \frac{\text{Sum of Squares Regression (SSR)}}{\text{Sum of Squares Total (SST)}}$$
$$= \frac{10.568}{13.794} = 76.61\%$$

Features of Coefficient of Determination (R^2)

R^2 lies between 0% and 100%. A high R^2 explains variability better than a low R^2. If R^2=1%, only 1% of the total variability can be explained. On the other hand, if R^2=90%, over 90% of the total variability can be explained. In a nutshell, the higher the R^2, the higher the model's explanatory power.

For simple linear regression (R^2) is calculated by squaring the correlation coefficient between the dependent and the independent variables:

$$r^2 = R^2 = \left(\frac{Cov(X,Y)}{\sigma_X \sigma_Y}\right)^2 = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Where:

$(Cov(X,Y))$ = Covariance between two variables, X and Y.

(σ_X) = Standard deviation of X.

(σ_Y) = Standard deviation of Y.

Example: Calculating Coefficient of Determination (R^2)

An analyst determines that $(\sum_{i=1}^{6} (Y_i - \bar{Y})^2 = 13.704)$ and $(\sum_{i=1}^{6} (Y_i - \hat{Y}_i)^2 = 3.136)$ from the regression analysis of inflation rates on unemployment rates. The coefficient of determination $((R^2))$ is *closest to*:

Solution

Sum of Squares Total (SST) Sum of Squared Errors (SSE)

$$R^2 = \frac{\text{Sum of Squares Total (SST)} - \text{Sum of Squared Errors (SSE)}}{\text{Sum of Squares Total (SST)}}$$

$$= \frac{\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \sum_{i=1}^{n}(Y_i - \hat{Y})^2\right)}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = \frac{13.704 - 3.136}{13.704}$$

$$= 0.7712 = 77.12\%$$

F-statistic in Simple Regression Model

Note that the coefficient of determination discussed above is just a descriptive value. To check the statistical significance of a regression model, we use the F-test. The F-test requires us to calculate the F-statistic.

In simple linear regression, the F-test confirms whether the slope (denoted by (b_1)) in a regression model is equal to zero. In a typical simple linear regression hypothesis, the null hypothesis is formulated as: $(H_0 : b_1 = 0)$ against the alternative hypothesis $(H_1 : b_1 \neq 0)$. The null hypothesis is rejected if the confidence interval at the desired significance level excludes zero.

The Sum of Squares Regression (SSR) and Sum of Squares Error (SSE) are employed to calculate the F-statistic. In the calculation, the Sum of Squares Regression (SSR) and Sum of Squares Error (SSE) are adjusted for the degrees of freedom.

The Sum of Squares Regression(SSR) is divided by the number of independent variables (k) to get the Mean Square Regression (MSR). That is:

$$MSR = \frac{SSR}{k} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{k}$$

Since we only have $(k = 1)$, in a simple linear regression model, the above formula changes to:

$$MSR = \frac{SSR}{1} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{1} = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$$

Therefore, in the Simple Linear Regression Model, MSR = SSR.

Also, the Sum of Squares Error (SSE) is divided by degrees of freedom given by $(n - k - 1)$ (this

translates to $(n-2)$ for simple linear regression) to arrive at Mean Square Error (MSE). That is,

$$MSE = \frac{\text{Sum of Squares Error (SSE)}}{n-k-1} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-k-1}$$

For a simple linear regression model,

$$MSE = \frac{\text{Sum of Squares Error(SSE)}}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}$$

Finally, to calculate the F-statistic for the linear regression, we find the ratio of MSR to MSE. That is,

$$F-\text{statistic} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{\frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{k}}{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-k-1}}$$

For simple linear regression, this translates to:

$$F-\text{statistic} = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y})^2}{n-2}}$$

The F-statistic in simple linear regression is F-distributed with (1) and $(n-2)$ degrees of freedom. That is,

$$\frac{MSR}{MSE} \sim F_{1,n-2}$$

Note that the F-test regression analysis is a one-side test, with the rejection region on the right side. This is due to the fact that the objective is to test whether the variation in Y explained (the numerator) is larger than the variation in Y unexplained (the denominator).

Interpretation of F-test Statistic

A large F-statistic value proves that the regression model effectively explains the variation in the dependent variable and vice versa. On the contrary, an F-statistic of 0 indicates that the independent variable does not explain the variation in the dependent variable.

We reject the null hypothesis if the calculated value of the F-statistic is greater than the critical F-value.

It is worth mentioning that F-statistics are not commonly used in regressions with one independent variable. This is because the F-statistic is equal to the square of the t-statistic for the slope coefficient, which implies the same thing as the t-test.

Standard Error of Estimate

Standard Error of Estimate, S_e or SEE, is alternatively referred to as the root mean square error or standard error of the regression. It measures the distance between the observed dependent variables and the dependent variables the regression model predicts. It is calculated as follows:

$$\text{Standard Error of Estimate}\,(S_e) = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

The standard error of estimate, coefficient of determination, and F-statistic are the measures that can be used to gauge the goodness of fit of a regression model. In other words, these measures tell the extent to which a regression model syncs with data.

The smaller the Standard Error of Estimate is, the better the fit of the regression line. However, the Standard Error of Estimate does not tell us ***how well*** the independent variable explains the variation in the dependent variable.

Hypothesis Tests of Regression Coefficients

Hypothesis Test on the Slope Coefficient

Note that the F-statistic discussed above is used to test whether the slope coefficient is

significantly different from 0. However, we may also wish to test whether the population slope differs from a specific value or is positive. To accomplish this, we use the t-distributed test.

The process of performing the t-distributed test is as follows:

1. **State the hypothesis**: For instance, typical hypothesis statements include:
    - $H_0 : b_1 = 0$ versus $H_a : b_1 \neq 0$
    - $H_0 : b_1 \leq 0$ versus $H_a : b_1 > 0$
2. **Identify the appropriate test statistic**: The test statistic for the t-distributed test on slope coefficient is given by:

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$$

Where:B_1 = Hypothesized slope coefficient.$\hat{b}_1$ = Point estimate for $b_1 s_{\hat{b}_1}$ = Standard error of the slope coefficient.The test statistic is t-distributed with $n - k - 1$ degrees of freedom. Since we are dealing with simple linear regression, we will deal with $n - 2$ degrees of freedom. The standard error of the slope coefficient ($s_{\hat{b}_1}$) is calculated as the ratio of the standard error of estimate (s_e) and the square root of the variation of the independent variable:

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

Where:

$$s_e = \sqrt{MSE}$$

3. **Specify the level of significance**: Note the level of significance level, usually denoted by alpha, α. A typical significance level might be α = 5%
4. **State the decision rule**: Using the significance level, find the critical values. You can use the t-table or spreadsheets such as Excel, statistical software such as R, or programming languages such as Python. In an exam situation, such critical values will be

provided. Compare the t-statistic value to the critical t-value (t_c). Reject the null hypothesis if the absolute t-statistic value is greater than the upper critical t-value or less than the lower critical value, i.e., $t > +t_{critical}$ or $t < -t_{critical}$

5. **Calculate the test statistic**: Using the formula above, calculate the test statistic. Intuitively, you might need to calculate the standard error of the slope coefficient ($s_{\hat{b}_1}$) first.
6. **Make a decision**: Make a decision whether to reject or fail to reject the null hypothesis.

Example: Hypothesis Test Concerning Slope Coefficient

Recall the example where we regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

The estimated regression model is

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

Assume that we need to test whether the slope coefficient of the unemployment rates is positive at a 5% significance level.

The hypotheses are as follows:

- $H_0 : b_1 < 0$ versus $H_a : b_1 \geq 0$

Next, we need to calculate the test statistic given by:

- $t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}}$

Where:

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

Recall that,

$$s_e = \sqrt{MSE} = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n-2}} = \sqrt{\frac{3.136}{8}} = 0.6261$$

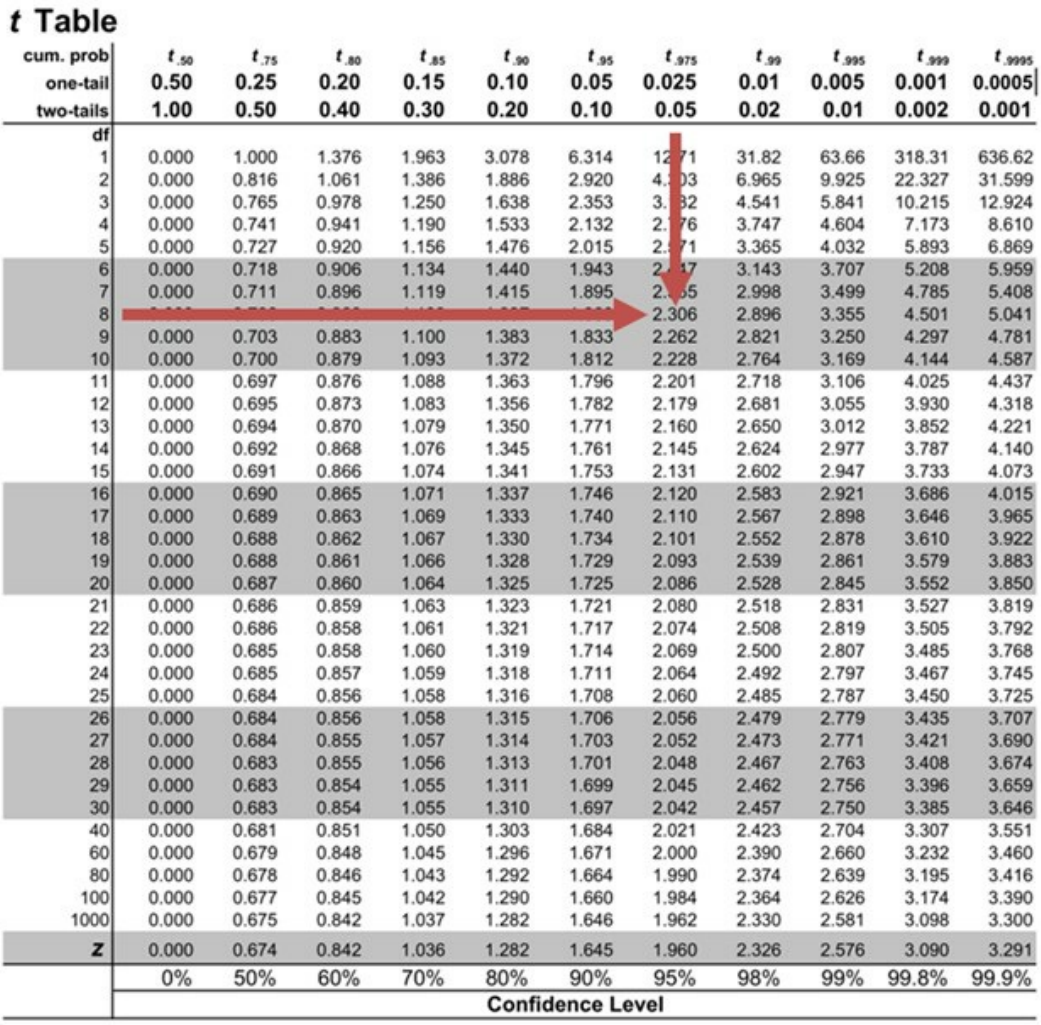
So that,

$$s_{\hat{b}_1} = \frac{s_e}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = \frac{0.6261}{\sqrt{12.989}} = 0.1737$$

Therefore,

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \frac{-0.9020 - 0}{0.1737} = -5.193$$

Next, we need to find critical t-values. Note that this is a one-sided test. As such, we need to find $t_{8}, 0.05$. We will use the t-table:



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8							2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

From the table, $t_{8,}0.05 = 1.860$. We fail to reject the null hypothesis since the calculated test statistic is less than the critical t-value (?$5.193 < 1.860$). There is sufficient evidence to indicate that the slope coefficient is not positive.

Relationship between the Hypothesis Test of Correlation and Slope Coefficient

In simple linear regression, a distinct characteristic exists: the t-test statistic checks if the slope coefficient equals zero. This t-test statistic is the same as the test-statistic used to determine if the pairwise correlation is zero.

This feature is true for two-sided tests ($H_0 : \rho = 0$ versus $H_a : \rho \neq 0$ and $H_0 : b_1 = 0$ versus $H_a : \rho \neq 0$) and one-sided test ($H_0 : \rho \leq 0$ versus Ha : $\rho > 0$ and

$H_0 : b_1 \leq 0$ versus $H_a : \rho > 0$ or $H_0 : \rho > 0$ versus $H_a : \rho \leq 0$ and $H_0 : b_1 > 0$ versus $H_a : \rho \leq 0$).

Note that the test -statistic to test whether the correlation is equal to zero is given by:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

The above test statistic is t-distributed with $(n-2)$ degrees of freedom.

Consider our previous example, where we regressed inflation rates against unemployment rates from 2011 to 2020. Assume we want to test whether the pairwise correlation between the unemployment and inflation rates equals zero.

In the example, the correlation between the unemployment rates and inflation rates is -0.8782. As such, the test- statistic to test whether the correlation is equal to zero is

$$t = \frac{-0.8782\sqrt{10-2}}{\sqrt{1-(-0.8782)^2}} \approx -5.19$$

Note this is equal to the test statistic t-test statistic used to perform the hypothesis test whether the slope coefficient is zero:

$$t = \frac{\hat{b}_1 - B_1}{s_{\hat{b}_1}} = \frac{-0.9020 - 0}{0.1737} = -5.193$$

Hypothesis Test of the Intercept Coefficient

Similar to the slope coefficient, we may also want to test whether the population intercept is equal to a certain value. The process is similar to that of the slope coefficient. However, the test statistic for t-distributed test on slope coefficient is given by:

$$t = \frac{\hat{b}_0 - B_0}{s_{\hat{b}_0}}$$

Where:

B_1 = Hypothesized intercept coefficient.

$\hat{b}_1$ = Point estimate for b_1.

$s_{\hat{b}_0}$ = Standard error of the intercept.

The formula for the standard error of the intercept $s_{\hat{b}_0}$ is given by:

$$s_{\hat{b}_0} = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

Recall the example where regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

The estimated regression model is

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

Assume that we need to test whether the intercept is greater than 1 at a 5% significance level.

The hypotheses are as follows:

$$H_0 : b_0 \leq 1 \text{ versus } H_a : b_0 > 1$$

Next, we need to calculate the test statistic given by:

$$t = \frac{\hat{b}_0 - B_0}{s_{\hat{b}_0}}$$

Where:

$$s_{\hat{b}_0} = \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}} = \sqrt{\frac{1}{10} + \frac{5.29^2}{12.989}} = 1.501$$

Therefore,

$$t = \frac{7.112 - 1}{1.501} = 4.0719$$

Note that this is a one-sided test. From the table, $t_8, 0.05 = 1.860$. Since the calculated test statistic is less than the critical t-value ($4.0179 > 1.860$), we reject the null hypothesis. There is sufficient evidence to indicate that the intercept is greater than 1.

Hypothesis Tests Concerning Slope Coefficient When Independent Variable is an Indicator Variable

Dummy variables, also known as indicator variables or binary variables, are used in regression analysis to represent categorical data with two or more categories. They are particularly useful for including qualitative information in a model that requires numerical input variables.

Example: Regression Analysis With Indicator Variables

Assume we aim to investigate if a stock's inclusion in an Environmental, Social, and Governance (ESG) focused fund affects its monthly stock returns. In this case, we'll analyze the monthly returns of a stock over a 48-month period.

We can use a simple linear regression model to explore this. In the model, we regress monthly returns, denoted as R, on an indicator variable, ESG. This indicator takes the value of 0 if the stock isn't part of an ESG-focused fund and 1 if it is.

$$R = b_0 + b_1 ESG + \varepsilon_i$$

Note that we estimate the simple linear regression in a way similar to if the independent variable was continuous.

The intercept β_0 is the predicted value when the indicator variable is 0. On the other hand, the slope when the indicator variable is 1 is the difference in the means if we grouped the observations by the indicator variable.

Assume that the following table is the results of the above regression analysis:

	Estimated Coefficients	Standard Error of Coefficients	Calculated Test Statistic
Intercept	0.5468	0.0456	9.5623
ESG	1.1052	0.1356	9.9532

Additionally, we have the following information regarding the means and variances of the variables.

	Monthly returns of ESG Focused Stocks	Monthly Returns of Non-ESG Stocks	Difference in Means
Mean	1.6520	0.5468	1.1052
Variance	1.1052	0.1356	
Observations	10	38	

From the above tables, we can see that:

- The intercept (0.5468) is equal to the mean of the returns for the non-ESG stocks.
- The slope coefficient (1.1052) is the difference in means of returns between ESG-focused stocks and non-ESG stocks.

Now, assume that we want to test whether the slope coefficient is equal to 0 at a 5% significance level. Therefore, the hypothesis is $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. Note that the degrees of freedom in $48 - 2 = 46$. As such, the critical t-values (usually given in the table above) is $t_{46, 0.025} = \pm 2.013$.

From the first table above, the calculated test statistic for the slope is greater than the critical t-value ($9.9532 > 2.013$). As a result, we reject the null hypothesis that the slope coefficient is equal to zero.

p-Values and Level of Significance

The p-value is the smallest level of significance level at which the null hypothesis is rejected. Therefore, the smaller the p-value, the smaller the probability of rejecting the true null hypothesis (type I error) and, hence, the greater the validity of the regression model.

Software packages commonly offer p-values for regression coefficients. These p-values help test a null hypothesis that the true parameter equals 0 versus the alternative that it's not equal to zero.

We reject the null hypothesis if the p-value corresponding to the calculated test statistic is less than the significance level.

Example: Hypothesis Testing of Slope Coefficients

An analyst generates the following output from the regression analysis of inflation on unemployment:

	Regression Statistics		
	R Square	0.7684	
	Standard Error	0.0063	
	Observations	10	
	Coefficients	Standard Error	t-Stat
Intercept	0.0710	0.0094	7.5160
Forecast (Slope)	−0.9041	0.1755	−5.1516

At the 5% significant level, test the null hypothesis that the slope coefficient is significantly different from one, that is,

$$H_0 : b_1 = 1 \text{ vs. } H_a : b_1 \neq 1$$

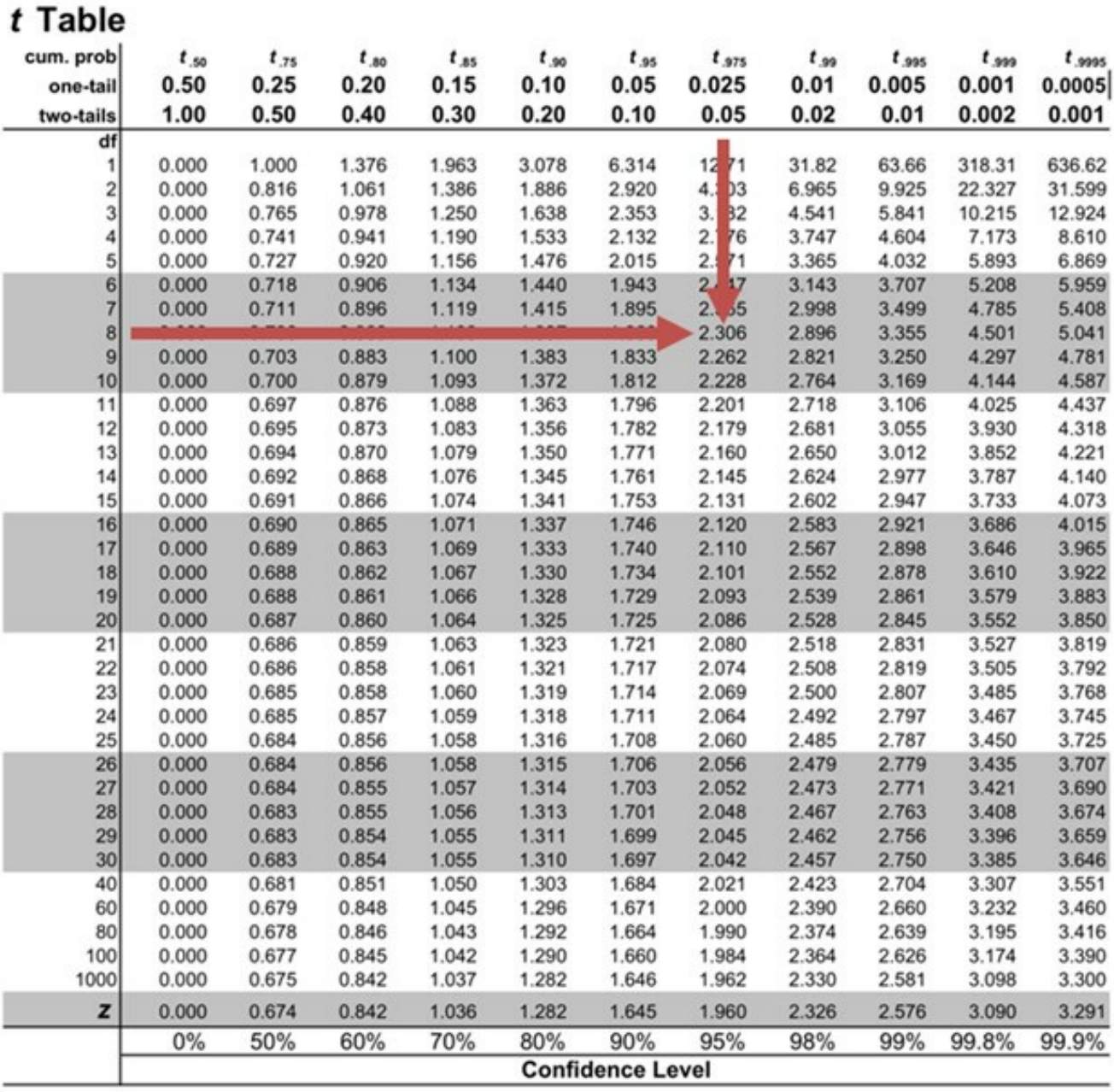
Solution

The calculated t-statistic, $t = \frac{\hat{b}_1 - b_1}{\hat{S}_{b_1}}$ is equal to:

$$t = \frac{-0.9041 - 1}{0.1755} = -10.85$$

The critical two-tail t-values from the table with $n - 2 = 8$ degrees of freedom are:

$$t_c = \pm 2.306$$



t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8							2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Notice that $|t| > t_c$ i.e., $(10.85 > 2.306)$

Therefore, we reject the null hypothesis and conclude that the estimated slope coefficient is statistically different from one.

Note that we used the confidence interval approach and arrived at the same conclusion.

Question 1

Samantha Lee, an investment analyst, is studying monthly stock returns. She focuses on companies listed in a Renewable Energy Index across various economic conditions. In her analysis, she performed a simple regression. This regression explains how stock returns vary concerning the indicator variable RENEW. RENEW equals 1 when there's a positive policy change towards renewable energy during that month, and 0 if not. The total variation in the dependent variable amounted to 220.34. Of this, 94.75 is the part explained by the model. Samantha's dataset includes 36 monthly observations.

Calculate the coefficient of determination, F-statistic, and standard deviation of monthly stock returns of companies listed in a Renewable Energy Index.

A. R^2=43.00%;F=26.07;Standard deviation=2.51.
B. R^2=53.00%;F=26.41;Standard deviation=2.55.
C. R^2=33.00%;F=36.07;Standard deviation=3.55.

Solution

The correct answer is A.

Coefficient of determination:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = \frac{94.75}{220.34} \approx 43\%$$

F-statistic:

$$F = \frac{\frac{\text{Explained variation}}{k}}{\frac{\text{Unexplained variation}}{n-2}} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-2}} = \frac{\frac{94.75}{1}}{\frac{220.34-94.75}{34}} = 26.07$$

Standard deviation:

Note that,

$$\text{Total Variation} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 220.34$$

And the standard deviation is given by:

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}}$$

As such,

$$\text{Standard deviation} = \sqrt{\frac{\text{Total variation}}{n-1}} = \sqrt{\frac{220.34}{n-1}} = 2.509$$

Question 2

Neeth Shinu, CFA, is forecasting the price elasticity of supply for a specific product. Shinu uses the quantity of the product supplied for the past 5months as the dependent variable and the price per unit of the product as the independent variable. The regression results are shown below.

Regression Statistics				
R Square	0.9941			
Standard Error	3.6515			
Observations	5			
	Coefficients	Standard Error	t Stat	P-value
Intercept	−159	10.520	(15.114)	0.001
Slope	0.26	0.012	22.517	0.000

Which of the following most likely reports the correct value of the t-statistic for the slope and most accurately evaluates its statistical significance with 95% confidence?

A. t = 21.67; the slope is significantly different from zero.

B. t = 3.18; the slope is significantly different from zero.

C. t = 22.57; the slope is not significantly different from zero.

Solution

The correct answer is A.

The t-statistic is calculated using the formula:

$$t = \frac{\hat{b}_1 - b_1}{\hat{S}_{b_1}}$$

Where:

- b_1 = True slope coefficient.
- $\hat{b}_1$ = Point estimator for B_1.
- $\hat{S}_{b_1}$ = Standard error of the regression coefficient.

$$t = \frac{0.26 - 0}{0.012} = 21.67$$

The critical two-tail t-values from the t-table with $n - 2 = 3$ degrees of freedom are:

$$t_c = \pm 3.18$$

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3							3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

Notice that $|t| > t_c$ (i.e., 21.67 > 3.18).

Therefore, the null hypothesis can be rejected. Further, we can conclude that the estimated slope coefficient is statistically different from zero.

LOS 10d: describe the use of analysis of variance (ANOVA) in regression analysis, interpret ANOVA results, and calculate and interpret the standard error of estimate in a simple linear regression

The sum of squares of a regression model is usually represented in the Analysis of Variance (ANOVA) table. The ANOVA table contains the sum of squares (SST, SSE, and SSR), the degrees of freedom, the mean squares (MSR and MSE), and F-statistics.

The typical format of ANOVA is as shown below:

Source	Sum of Squares	Degrees of Freedom	Mean square	F-statistic
Regression (Explained)	$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Residual (explained)	$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$	
Total	$SST = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$n - 1$		

Standard Error of EstimateStandard Error of Estimate, S_e or SEE, is referred to as the root mean square error or standard error of the regression. It measures the distance between the observed and dependent variables predicted by the regression model. The Standard Error of Estimate is easily calculated from the ANOVA table using the following formula:

$$\text{Standard Error of Estimate } (S_e) = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_i - \hat{Y})^2}{n - 2}}$$

The standard error of estimate, coefficient of determination, and F-statistic are the measures that can be used to gauge the goodness of fit of a regression model. In other words, these measures are used to tell the extent to which a regression model syncs with data.

The smaller the Standard Error of Estimate is, the better the fit of the regression line. However, the Standard Error of Estimate does not tell us how well the independent variable explains the variation in the dependent variable.

Example: Calculating and Interpreting F-Statistic

The completed ANOVA table for the regression model of the inflation rate against the unemployment rate over 10 years is given below:

Source	Sum of Squares	Degrees of Freedom	Mean Sum of Squares	F-Statistic
Regression	10.568	1	10.568	?
Error	3.136	8	0.392	
Total	13.704	9		

a. Use the above ANOVA table to calculate the F-statistic.

b. Test the hypothesis that the slope coefficient is equal to a 5% significance level.

Solution

a. $= \frac{\text{Mean Regression Sum of Squares (MSR)}}{\text{Mean Squared Error(MSE)}} = \frac{10.568}{0.392} = 26.960$

b. We are testing the null hypothesis $H_0 : b_1 = 0$ against the alternative hypothesis $H_1 : b_1 \neq 0$. The critical F-value for $k = 1$ and $n - 2 = 8$ degrees of freedom at a 5% significance level is roughly 5.32. Note that this is a one-tail test, and therefore, we use the 5% F-table.

	α = 0.050	F-table								
	$dF_1(v_1)$									
$dF_2(v_2)$	1	2	3	4	5	6	7	8	9	10	11
1	16.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.87

Remember that the null hypothesis is rejected if the calculated value of the F-statistic is greater than the critical value of F. Since $26.960 > 5.32$, we reject the null hypothesis and conclude that the slope coefficient is significantly different from zero.Notice that we also rejected the null hypothesis in the previous examples. We did so because the 95% confidence interval did not include zero.

An F-test duplicates the t-test in regard to the slope coefficient significance for a linear regression model with one independent variable. In this case, $t^2 = 2.306^2 \approx 5.32$. Since the F-statistic is the square of the t-statistic for the slope coefficient, its inferences are the same as the t-test. However, this is not the case for multiple regressions.

Question

Consider the following analysis of variance (ANOVA) table:

Source	Degrees of Freedom	Sum of Squares	Mean Sum of Squares
Regression	1	$1,701,563$	$1,701,563$
Error (Unexplained)	3	$106,800$	$13,350$
Total	4	$1,808,363$	

The value of R^2 and the F-statistic for the test of fit of the regression model are *closest to*:

A. 6% and 16.

B. 94% and 127.

C. 99% and 127.

Solution

The correct answer is B.

$$R^2 = \frac{\text{Sum of Squares Regression (SSR)}}{\text{Sum of Squares Total (SST)}} = \frac{1,701,563}{1,808,363} = 0.94 = 94\%$$

$$F = \frac{\text{Mean Regression Sum of Squares (MSR)}}{\text{Mean Squared Error (MSE)}}$$
$$= \frac{1,701,563}{13,350} = 127.46 \approx 127$$

LOS 10e: calculate and interpret the predicted value for the dependent variable, and a prediction interval for it, given an estimated linear regression model and a value for the independent variable

We calculate the predicted value of the dependent variable, Y, by inserting the estimated value of the independent variable, X, into the regression equation. The predicted value of the dependent variable, Y, is determined using the following formula:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X$$

Where:

$\hat{Y}$ = Predicted value of the dependent variable.

X = Estimated value of the independent variable.

Example: Calculating the Predicted Value of a Dependent Variable

Refer to the example of regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

The estimated regression model is illustrated below.

$$\hat{Y} = 7.112 - 0.9020X_i + \varepsilon_i$$

Calculate the predicted inflation rate value if the forecasted value of the unemployment rate is

4.5%.

Solution

The predicted value of the inflation rate is determined as follows:

$$\hat{Y} = 7.112 - 0.9020 \times 4.5 = 3.053\%$$

Confidence Interval for Predicted Values

The calculation of the confidence interval for the predicted value of a dependent variable is the same as that of the confidence interval for regression coefficients. The confidence interval for a predicted value of the dependent variable is given by:

$$\text{Prediction Interval} = \hat{Y} \pm t_c s_f$$

Where:

t_c= Two-tailed critical t-value at the given significance level with $n - 2$ df.

$\hat{Y}$ = Predicted value of a dependent variable.

s_f^2= The estimated variance of the prediction error.

$$s_f^2 = s_e^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)\, s_x^2}\right] = s_e^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}\right]$$

Where:

s_e^2 = The squared standard error of the estimate.

n = Number of observations.

s_X^2= Variance of the independent variable.

X_f = Value of the independent variable.

We can, therefore, calculate the standard error of forecast as shown below:

$$s_f = s_e \sqrt{1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

From the formula above, we can observe that:

- A better fit of the regression analysis leads to a smaller standard error of the estimate (s_e), subsequently resulting in a lower standard error of the forecast.
- When the sample size (n) in the regression calculation increases, it directly corresponds to a reduction in the standard error of the forecast.
- If the forecasted independent variable (X_f) approaches the mean of the independent variable ($\bar{X}$) utilized in the regression analysis, it decreases the standard error of the forecast.

Example: Calculating the Confidence Interval of the Predicted Value

Refer to the example of regressed inflation rates against unemployment rates from 2011 to 2020.

Year	Unemployment Rate %(X_i)	Inflation Rate % (Y_i)	Predicted Unemployment rate ($\hat{Y}_i$)	Variation to be Explained. $(Y_i - \bar{Y})^2$	Variation Unexplained $(Y_i - \hat{Y}_i)^2$	Variation Explained $(\hat{Y}_i - \bar{Y})^2$
2011	6.1	1.7	1.610	0.410	0.008	0.533
2012	7.4	1.2	0.437	1.300	0.582	3.621
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2019	4.0	4.7	3.504	5.570	1.430	1.355
2020	3.9	3.6	3.594	1.588	0.000	1.573
Sum	52.90	23.4		13.704	3.136	10.568
Arithmetic Mean	5.29	2.34				

Consider the results of the regression analysis of inflation rates on unemployment rates:

Regression Statistics	
R Square	0.7711
Standard Error	0.6261
Observations	10

ANOVA				
	df	Sum of Squares	Mean Square	F
Regression	1	10.568	10.568	26.9565
Residual	8	3.136	0.392	
Total	9	13.704		

	Coefficients	Standard Error	t Stat	p-value
Intercept	7.112	0.940	7.565	0.000
Unemployment rate (%)	−0.902	0.174	−5.192	0.001

Calculate the 95% confidence interval of the predicted value of the inflation rate, given that the forecasted unemployment rate is 4.5%.

Solution

$$\text{Prediction Interval} = \hat{Y} \pm t_c s_f$$

The estimated variance of the prediction error is:

$$\begin{aligned} s_f^2 &= s_e^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{(n-1)\, s_X^2} \right] \\ &= s_e^2 \left[1 + \frac{1}{n} + \frac{(X_f - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right] \\ &= 0.6261^2 \left[1 + \frac{1}{10} + \frac{(4.5 - 5.29)^2}{12.989} \right] = 0.450 \end{aligned}$$

As such, the standard error of forecast is:

$$s_f = \sqrt{0.450} = 0.6708$$

The predicted value of the inflation rate given an unemployment rate of 4.5% is:

$$\hat{Y} = 7.112 - 0.9020 \times 4.5 = 3.05\%$$

The two-tailed critical t-value with 8 $(n - 2)$ degrees of freedom at the 5% significance level is 2.306.

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8							2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

The prediction interval at the 95% confidence level is:

$$\text{Prediction Interval (PI)} = \hat{Y} \pm t_c s_f$$

$$PI = 3.05 \pm 2.306 \times 0.6708 = 1.50\% \text{ to } 4.60\%$$

Interpretation

Given an unemployment rate of 4.5%, we are **95% confident** that **the inflation rate will lie between 1.50%** and **4.60%**.

Question 1

The regression equation of the quantity of goods against the price is given by:

$$Y = -159 + 0.26X$$

Where:

Y = Quantity supplied.

X = Price per unit of the product.

The predicted value of the quantity supplied when the price equals 1,200 is *closest to*:

A. 153.

B. 155.

C. 471.

The correct answer is A.

$$Y = -159 + 0.26 \times 1,200 = 153$$

LOS 10f: Describe different functional forms of simple linear regressions

To address non-linear relationships, we employ various functional forms to potentially convert the data for linear regression. Here are three commonly used log transformation functional forms:

1. **Log-lin model:** In this log transformation, the dependent variable is logarithmic, while the independent variable is linear. It is represented as shown below.

   $$\ln Y = b_0 + b_1 X_i.$$

   The slope coefficient in the log-lin model is the relative change in the dependent variable for an absolute change in the independent variable.

   When utilizing a log-lin model, caution must be exercised when making forecasts. For example, in the predicted regression equation like $Y = -3 + 5X$, if X is equal to 1, the $\ln Y = -3$, then,

   $$Y = e^{-3} = 0.0498$$

   Moreover, the lin-lin model cannot be compared with the log-lin model without the transformation. As such, we need to transform R^2 and F-statistic.

2. **Lin-log model:** In this case, the dependent variable is linear, while the independent variable is logarithmic. It is represented as follows:
   $Y_i = b_0 + b_1 \ln X_i$.

   The slope coefficient in the lin-log model is responsible for the absolute change in the dependent variable for a relative change in the independent variable.

3. **Log-log model:** In this log transformation, both the dependent and independent variables are logarithmic. It is represented as $\ln Y_i = b_0 + b_1 \ln X_i$. The slope coefficient in the log-log model is the relative change in the dependent variable for a relative change in the independent variable. In other words, if X increases by 1%, Y will change by b_1.

Selecting the Correct Functional Form

To settle on the correct functional form, consider the following goodness of fit measures:

I. Coefficient of determination (R^2). A high value is better.
II. F-statistic. The high value of the F-statistic is better.
III. Standard error of the estimate (S_e). A low value of S_e is better.

Aside from the factors cited above, the patterns in residuals can also be analyzed when evaluating a model. Residuals are random and uncorrelated in a good model.

Question 1

Which of the following statements about the log-lin model is *most likely* correct:

A. The dependent variable is linear, while the independent variable is logarithmic.
B. Both the dependent and independent variables are logarithmic
C. The dependent variable is logarithmic, while the independent variable is linear.

The correct answer is c.

In the log-lin model, the dependent variable (Y) is logarithmic, as represented by

$$\ln Y = b_0 + b_1 X_i$$

While the independent variable (X) is linear.

A is incorrect. It describes the lin-log model, where the dependent variable is linear and the independent variable is logarithmic.

B is incorrect. It describes the log-log model, where both the dependent and independent variables are logarithmic.

Learning Module 11: Introduction to Big Data Techniques

LOS 11b: describe Big Data, artificial intelligence, and machine learning

Big data is a term that describes large, complex datasets. These datasets are analyzed with computers to uncover patterns and trends, particularly those related to human behavior. Big data includes **traditional sources** like company reports and government data and **non-traditional** sources like social media, sensors, electronic devices, and data generated as a byproduct of a company's operations.

Characteristics of Big Data

Volume: The amount of data collected in various forms, including files, records, tables, etc. Quantities of data reach almost incomprehensible proportions.

Velocity: The speed of data processing can be extremely high. In most cases, we deal with real-time data.

Variety: The number of types/formats of data. The data could be structured (e.g., SQL tables or CSV files), semi-structured (e.g., HT ML code), or unstructured (e.g., video messages).

Veracity: This is the trustworthiness and reliability of data sources. Veracity is crucial when using big data for making predictions or drawing conclusions. Big data makes it challenging to distinguish between data quality and quantity.

Types of Big Data

Big Data can be structured, unstructured, or semi-structured:

Structured data refers to information with a high degree of organization. Items can be organized in tables and stored in a database where each field represents the same type of information.

Unstructured data refers to information with a low degree of organization. Items are unorganized and cannot be presented in tabular form, such as text messages, tweets, emails, voice recordings, pictures, blogs, scanners, and sensors.

Semi-structured data may have the qualities of both structured and unstructured data.

Sources of Big Data

- **Financial markets**: Equity, swaps, futures, options, and other derivatives.
- **Businesses**: Financial statements, credit card purchases, and commercial transactions.
- **Governments**: Payroll, economic, trade, employment data, etc.
- **Individuals**: Product reviews, credit card purchases, social media posts, etc.
- **Sensors**: Shipping cargo information, traffic data, and satellite imagery.
- **The Internet of Things**: data generated by ‘smart’ buildings through fittings such as CCT V cameras, vehicles, home appliances, etc.

Professional investors, particularly quantitative ones, use alternative data sources in their financial analysis and decision-making. These sources significantly influence how they conduct their processes. They use alternative data to support data-driven investment models and decisions.

The following are the top three alternative data sources:

- **People-generated data**: This data is unstructured and is primarily accessed through website clicks and page visits
- **Commercial operations data**: This includes data on credit cards and corporate exhaust. It includes information from business transactions like point-of-sale records and banking activities. This data is typically structured.

- **Data produced by sensors**: This data is typically unstructured and is gathered through satellites, smartphones, cameras, RFID chips, and webcams.

Investment professionals must consider legal and ethical aspects when they use non-public information. Web data scraping can gather personal data that might be legally protected or disclosed without people's knowledge or consent.

Big Data Challenges

- **Quality**: Important questions include, but are not limited to: Does the dataset contain selection bias, missing data, or outliers?
- **Volume**: Is the quantity of data gathered adequate?
- **Appropriateness:** Is the dataset suitable for the chosen analysis method?

Experts have created artificial intelligence (AI) and machine learning methods to handle large and intricate alternative datasets. These technologies help in understanding and evaluating this vast and complex data.

Artificial Intelligence (AI) and Machine Learning (ML)

Artificial Intelligence

In broad terms, **artificial intelligence** refers to machines that can perform tasks in “intelligent” ways. It has much to do with developing computer systems that exhibit cognitive and decision-making abilities comparable to or superior to humans. It is the broader concept of machines being able to carry out tasks in a way that we would consider “smart.”

Early AI took the shape of expert systems, using "if-then" computer programming to mimic human knowledge and analysis. Neural networks, another early form, mimicked human brain functions in learning and processing information.

Machine Learning

Machine learning is a current application of AI that revolves around the idea that we should really just give machines access to data and let them learn by themselves without making any assumptions about the underlying probability distribution.

The idea is that when exposed to more data, machines can make changes on their own and come up with solutions to problems without reliance on human expertise – find and apply the pattern.

In the context of investment, machine learning requires big data for training. The growth of big data has enabled AI algorithms to improve modeling and predictive accuracy.

In machine learning (ML), a computer algorithm receives inputs, which can be datasets or variables, as well as outputs, representing the target data. The algorithm then learns how to model inputs into outputs or describe a data structure effectively. It learns by identifying data relationships and using this knowledge to enhance its learning process.

The ML divides the dataset into three unique types: a **training dataset**, a **validation dataset**, and a **test dataset**. A training dataset allows the algorithm to identify the link between inputs and outputs based on the historical pattern in the data. These relationships are then validated, and the model is adjusted using the validation dataset.

As the name suggests, the test dataset is used to test the model's strength in predicting well on the new data. Note that machine learning still needs human intervention to understand the underlying data and choose suitable techniques for data analysis. In other words, before data is utilized, it must be cleaned and free from bias and spurious data.

Causes of Errors in Machine Learning

Overfitting the Data

The model overfits the data when it discovers “false” associations or “unsubstantiated” patterns that cause prediction errors and wrong forecasts. In other words, overfitting happens when the ML model is overtrained on the data and considers the noise in the data as true parameters.

Underfitting the Data

Underfitting of data occurs when the model considers the true parameters as noise and is unable to identify the relationship within the training data. In other words, the model is too simple to recognize patterns in the data.

Black Box Problem

Machine learning models don't use explicit rules like traditional software. They learn from lots of data during training. This makes ML models, such as black boxes, sometimes give results that are hard to understand or describe.

Types of Machine Learning

Supervised Learning

Under supervised learning, computers learn to model data based on labeled training data containing inputs and the desired outputs. After “learning” how best to model the relationships for the labeled data, the algorithms are employed to predict the results for the new datasets.

Unsupervised Learning

In unsupervised learning, computers get input data without labels and have to describe it, often by grouping data points. They learn from unlabeled data and react based on commonalities. For example, grouping companies based on their financial, not geographical or industrial, characteristics is unsupervised learning.

Deep Learning

Deep learning involves computers using neural networks to process data in multiple stages, identifying complex patterns. It employs both supervised and unsupervised machine learning methods.

Question

Machine learning refers to one of the following:

A. Autonomous acquisition of knowledge through the use of computer programs.
B. Ability of machines to execute coded instructions.
C. Selective acquisition of knowledge through the use of computer programs.

Solution

The correct answer is A.

Machine learning means computers independently acquire knowledge through programs, enabling them to solve problems without human input. It's about computers learning and making decisions on their own.

LOS 11c: Describe applications of Big Data and Data Science to investment management

Data science is an interdisciplinary field that uses developments in computer science, statistics, and other fields to extract information from Big Data or data in general.

Data Processing Methods

Data analysts and scientists in big data analysis use different data management approaches. They consist of capture, curation, storage, search, and transfer.

- **Capture:** describes the method by which data is gathered and put into a form that may be used by the analytical process.
- **Curation:** By undertaking a data cleaning activity, data curation ensures the quality and accuracy of the data. Data inaccuracies are found in this procedure, and any missing data is made up for.
- **Storage:** Process of recording, archiving, and accessing data, as well as the fundamental structure of the underlying database:
- **Search:** Involves querying data to locate specific information. With big data, sophisticated techniques are necessary to efficiently retrieve the requested data content.
- **Transfer:** Describes the process of transferring data from the underlying data source or storage place to the underlying analytical instrument.

Data Visualization

Visualization encompasses data formatting, display, and summarization through graphical representations. For traditional structured data, tables, charts, and trends are commonly used, while non-traditional unstructured data demand innovative techniques like interactive three-

dimensional (3D) graphics, tag clouds, and mind maps.

Fintech is applied in investment management, including text analytics and natural language processing, risk assessment, and algorithmic trading.

Text Analytics and Natural Language

Text analytics employs computer programs to analyze and extract insights, primarily from unstructured text- or voice-based datasets like company filings, written reports, quarterly earnings calls, and social media content. Text analytics can be utilized in predictive analysis to identify potential indicators of future performance, such as consumer sentiment.

Natural language processing (NLP) is an area of study that involves creating computer programs to decipher and analyze human language. Essentially, NLP combines computer science, AI, and linguistics.

Translation, speech recognition, text mining, sentiment analysis, and topic analysis are examples of automated tasks that use NLP. Annual reports, call transcripts, news articles, social media posts, and other text- and audio-based data may all be analyzed using natural language processing (NLP), allowing NLP to discover trends more quickly and accurately than is humanly possible.

Using natural language processing data, earnings projections for a company's near-term prospects can be created. Twitter sentiments have also been used to gauge an initial public offering (IPO) success.

Python, R, and Excel VBA are frequently used programming languages, whereas SQL, SQLite, and NoSQL are prominent database systems.

Question

Which of the five data processing methods refers to the process of ensuring data quality and accuracy through a data cleaning exercise?

A. Data search
B. Data storage
C. Data curation

The correct answer is C.

Data curation refers to the process of ensuring data quality and accuracy through a data cleaning exercise. It involves uncovering data errors and adjusting for missing data.

A is incorrect. Data search refers to how to query data. Big data requires advanced techniques to locate requested data content.

B is incorrect. Data storage refers to how the data will be recorded, archived, and accessed and the underlying database design.

LOS 11a: describe aspects of “fintech” that are directly relevant for the gathering and analyzing of financial data

Fintech refers to technological innovation in designing and delivering financial services and products. At its core, fintech has helped companies, business owners, and investment managers to manage their operations better thanks to specialized software and algorithms.

Note that the term fintech is commonly used to refer to companies that develop new technologies and their applications and also the business sector that encompasses such companies.

Fintech in Gathering and Analyzing Financial Data

Initially, financial innovation was limited to simple tasks such as data processing and automation of routine tasks. Today, fintech encompasses more advanced systems that can analyze information and make decisions based on machine-learning logic. Machines have been developed to “learn” how to perform tasks over time. Using such systems has brought about high levels of efficiency that surpass human capabilities. Fintech covers a broader range of services and applications. As such, services and applications of fintech relevant to the investment industry include:

1. **Analysis of Large Datasets**Apart from traditional data such as corporate financial statements and economic indicators, fintech development has helped integrate alternative data, such as social media, into investment decision-making.
2. **Analytical Tools**Artificial Intelligence (AI) can identify complex, non-linear relationships compared to traditional quantitative methods by enabling different data analysis techniques. Diverse approaches to data analysis are now possible because of advancements in AI-based methodologies. As an illustration, analysts use AI to sift through the vast volumes of data from corporate filings and annual reports to produce insights.

Question

A characteristic of fintech is that it is:

A. at its most advanced state, using systems that follow specified rules and instructions.

B. limited to simple tasks such as automating routine processes and data processing.

C. primarily driven by the increased availability of data and technological advancement.

The correct answer is C.

The availability of vast amounts of data and technological advancements have been the primary drivers of fintech's expansion. The rapid growth in data, including diverse types, large quantities, and improved quality, has provided valuable insights for financial institutions and fintech companies to develop data-driven solutions. Technological advancements, such as artificial intelligence and big data analytics, have made it possible to analyze and interpret this data effectively, creating innovative financial products and services.

A is incorrect. While this may be true for some aspects of fintech, it doesn't directly address the two most important reasons behind its growth - the rapid growth in data and technological advancements. The advanced state of fintech is more a result of leveraging data and technological innovations to develop sophisticated and efficient financial solutions.

B is incorrect. Fintech is not limited to simple tasks; it has expanded to encompass various complex financial activities. While it does automate routine processes and data processing, it does so through advanced technologies that enable the handling of massive amounts of data efficiently.