Quant

1 RATES AND RETURNS

1.1 Interest Rates and Return Measurement

- **Interpretation of Interest Rates**
  - **As required rates of return**: The equilibrium interest rate represents the return investors require to lend funds for a given risk profile.
  - **As discount rates**: Used to calculate present values of future cash flows. Example: If the borrowing rate is 10%, discount future payments at 10% to get their present value.
  - **As opportunity cost of consumption**: Choosing current consumption over saving forfeits the interest income that could have been earned. Example: If the market rate is 5%, that 5% is the cost of consuming today.
- **Real Risk-Free Rate of Interest**
  - Theoretical rate on a single-period loan with:
    - Zero inflation expectation.
    - Zero probability of default.
  - Reflects *time preference* — the degree to which current consumption is preferred over future consumption.
- **Nominal vs Real Risk-Free Rate**
  - Real rate of return: Increase in purchasing power, adjusted for inflation.
  - Nominal risk-free rate: Observed rate (e.g., on T-bills), includes inflation premium.
  - Relation:
$$(1 + \text{nominal}) = (1 + \text{real})(1 + \text{expected inflation})$$
    Approximation:
$$\text{nominal} \approx \text{real} + \text{expected inflation}$$
- **Risk Premium Components**

- **Default risk premium**: Compensation for risk of missed payments.
- **Liquidity risk premium**: Compensation for difficulty of selling quickly at fair value.
- **Maturity risk premium**: Compensation for greater price volatility of longer-maturity bonds.
- Total required rate of return:

$$\text{Nominal risk-free rate} + \text{Default RP} + \text{Liquidity RP} + \text{Maturity RP}$$

- **Return Measurement**
  - **Holding Period Return (HPR)**:

$$\text{HPR} = \frac{\text{Ending Value} + \text{Income}}{\text{Beginning Value}} - 1$$

    Example: Price from €20 to €22, dividend €1:

$$\text{HPR} = \frac{22+1}{20} - 1 = 15\%$$

  - Multi-period HPR: Multiply $(1 + \text{HPR}_t)$ across all periods, then subtract 1.

- **Average Return Measures**
  - **Arithmetic mean**:

$$\bar{R}_{\text{arith}} = \frac{\sum_{t=1}^{n} R_t}{n}$$

    * Unbiased estimator of the population mean.
    * Includes all observations, including outliers.
  - **Geometric mean**:

$$\bar{R}_{\text{geo}} = \left(\prod_{t=1}^{n}(1+R_t)\right)^{1/n} - 1$$

    * Compound growth rate over multiple periods.
    * Always $\leq$ arithmetic mean, with the difference increasing with return variability.
  - Example: Returns $= -9.34\%, 23.45\%, 8.92\%$ Step: Multiply $(1 + R_t)$, take cube root, subtract 1.
  - **Harmonic mean**:

$$\bar{X}_{\text{harm}} = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$

    * Common in average share cost when equal amounts invested periodically.
    * Example: Prices = \$8, \$9, \$10, with \$1,000 invested monthly:

$$\bar{X}_{\text{harm}} \approx 8.93$$

        * Relation:

$$\text{Arithmetic mean} \times \text{Harmonic mean} = (\text{Geometric mean})^2$$

        * For non-equal values: Harmonic < Geometric < Arithmetic.

- **Dealing with Outliers**
    - **Trimmed mean**: Remove a certain percentage of extreme values.
    - **Winsorized mean**: Replace extreme values with the nearest remaining value.
- **Appropriate Uses**
    - Arithmetic mean: Use for single-period return estimates, includes outliers.
    - Geometric mean: Use for multi-period compounding.
    - Harmonic mean: Use for average share cost with fixed periodic investments.
    - Trimmed/Winsorized mean: Use to reduce effect of outliers.
- **Module Quiz 1.1 — Key Answers**
    1. C — Required rate of return or opportunity cost of consumption.
    2. C — Real risk-free rate.
    3. A — Harmonic mean of $3, 4, 5$ is 3.74.
    4. C — Geometric mean for average stock return over multiple years.

1.2 Time-Weighted and Money-Weighted Returns

- **Definitions and Concepts**
    - **Money-Weighted Rate of Return (MWRR)**:
        * Equivalent to the **Internal Rate of Return (IRR)** for a portfolio.
        * IRR: The interest rate at which the present value (PV) of all cash inflows equals the PV of all cash outflows, i.e., NPV = 0.
        * Takes into account:
            · *Inflows*: Beginning portfolio value and all subsequent deposits.
            · *Outflows*: Withdrawals and ending value.
        * Cash flows into the portfolio and out of the portfolio must be entered with opposite signs for IRR calculation.
    - **Time-Weighted Rate of Return (TWRR)**:
        * Measures **compound growth** — the rate at which $1 grows over a specified performance horizon.
        * Removes the effect of cash flow timing; isolates investment performance.

      * Preferred in investment management since portfolio managers usually do not control deposit/withdrawal timing.

- **Example: MWRR Calculation**
  - $t = 0$: Buy 1 share at \$100.
  - $t = 1$: Dividend = \$2, buy 1 more share at \$120.
  - $t = 2$: Sell both shares at \$130 each, dividend = \$2/share.
  - Net cash flows:

$$CF_0 = +100, \quad CF_1 = +118, \quad CF_2 = -264$$

  - Solve IRR $\Rightarrow$ MWRR = 13.86%.

- **Steps for Time-Weighted Rate of Return**
  1. **Step 1:** Value the portfolio immediately before each significant addition or withdrawal to define *subperiods*.
  2. **Step 2:** Compute the *holding period return (HPR)* for each subperiod:

$$\text{HPR}_i = \frac{\text{Ending Value}_i - \text{Beginning Value}_i + \text{Income}_i}{\text{Beginning Value}_i}$$

  3. **Step 3:** Compound subperiod returns:

$$\text{Total Return} = \prod_{i=1}^{n}(1 + \text{HPR}_i) - 1$$

  4. **Step 4:** If total period > 1 year, compute the **geometric mean** to annualize:

$$\text{Annual TWRR} = \left[\prod_{i=1}^{n}(1 + \text{HPR}_i)\right]^{1/T} - 1$$

- **Example: TWRR Calculation**
  - Same investment as MWRR example.
  - $t = 0 \to t = 1$: $\text{HPR}_1 = 22\%$.
  - $t = 1 \to t = 2$: $\text{HPR}_2 = 10\%$.
  - TWRR:

$$\text{TWRR} = \sqrt{(1.22)(1.10)} - 1 = 15.84\%$$

  - Difference from MWRR due to greater weighting of Year 2 returns in MWRR.

- **Key Differences Between MWRR and TWRR**
  - **MWRR**:

    - Sensitive to **timing and magnitude** of cash flows.
    - If more money is invested before poor performance, MWRR < TWRR.
    - If more money is invested before strong performance, MWRR > TWRR.
    - Appropriate when manager controls cash flow timing.
  - **TWRR**:
    - Eliminates distortion from cash flow timing.
    - Best measure of investment selection skill when the manager does *not* control cash flows.

- **Module Quiz 1.2 — Key Answers**
  1. B — MWRR $\approx$ 23.0%.
  2. C — TWRR $\approx$ 26.8%.

1.3 Common Measures of Return

- **Annualized Return Measures**
  - Returns are typically expressed on an **annualized** basis, regardless of the actual holding period.
  - Formula for annualizing HPR over n days:

$$\text{Annualized Return} = (1 + \text{HPR})^{\frac{365}{n}} - 1$$

  - **Example (Shorter than 1 year)**: Deposit \$100 $\rightarrow$ \$100.75 after 90 days. Annualized Return:

$$(1 + 0.0075)^{\frac{365}{90}} - 1$$

  - **Example (Longer than 1 year)**: Buy a 500-day bill for \$970, maturity \$1,000:

$$(1 + 0.03093)^{\frac{365}{500}} - 1$$

- **Effect of Compounding Frequency**
  - More frequent compounding $\Rightarrow$ higher effective rate, higher future value (FV), and lower present value (PV).
  - Present value with compounding:

$$PV = \frac{FV}{\left(1 + \frac{r}{m}\right)^{mN}}$$

    where:
    - r = quoted annual interest rate.
    - N = number of years.
    - m = compounding periods/year.

- Example: PV of \$1,000, $r = 6\%$:
  * Semiannual ($m = 2$): $PV = \frac{1{,}000}{(1.03)^2}$
  * Quarterly ($m = 4$): $PV = \frac{1{,}000}{(1.015)^4}$
  * Monthly ($m = 12$): $PV = \frac{1{,}000}{(1.005)^{12}}$
  * Daily ($m = 365$): $PV = \frac{1{,}000}{(1+0.06/365)^{365}}$

- **Continuous Compounding**
  - Limit as compounding frequency → infinity.
  - Continuously compounded return from HPR:

$$R_c = \ln(1 + \text{HPR}) = \ln\left(\frac{\text{End Value}}{\text{Start Value}}\right)$$

  - **Example**: Buy at \$100, sell at \$120:

$$R_c = \ln\left(\frac{120}{100}\right) = 0.18232 = 18.232\%$$

  - **Additivity property**:

$$R_{c,0\to 2} = R_{c,0\to 1} + R_{c,1\to 2}$$

- **Major Return Measures and Uses**
  - **Gross return**: Total return before deducting management/administration fees; trading commissions are deducted in both gross and net returns.
  - **Net return**: Return after deducting all management and administration fees.
  - **Pretax nominal return**: Before taxes; may differ depending on how dividends, interest, and capital gains are taxed.
  - **After-tax nominal return**: Net of tax liabilities.
  - **Real return**:
    * Adjusts nominal return for inflation.
    * Approximation:

$$\text{Real} \approx \text{Nominal} - \text{Inflation}$$

    * Exact:

$$\text{Real} = \frac{1 + \text{Nominal}}{1 + \text{Inflation}} - 1$$

    * Example: Nominal 7%, Inflation 2% $\Rightarrow$ Real $\approx 4.9\%$.
  - **Leveraged return**:
    * Uses borrowed funds to magnify returns.

    * Unleveraged return (money amount):

$$r \times V_0$$

    * Leveraged return (money amount):

$$r \times (V_0 + V_B) - r_B \times V_B$$

    * Leveraged rate of return:

$$R_{\text{leveraged}} = \frac{r(V_0 + V_B) - r_B V_B}{V_0}$$

- **Appropriate Uses**
  - **Annualized return**: Compare investments over different periods.
  - **Continuous compounding**: Used in advanced finance/math modeling; additive across periods.
  - **Gross/Net return**: Evaluate portfolio performance pre- or post-fee.
  - **Real return**: Assess purchasing power changes.
  - **Leveraged return**: Analyze impact of debt on investment outcomes.
- **Module Quiz 1.3 — Key Answers**
  1. B — Annualized return $\approx -8.5\%$.
  2. A — Continuously compounded return $\approx 13.64\%$.
  3. B — The 5% increase before fees = **gross return**.

Key Concepts — Modul 1

- **LOS 1.a: Interpretation of Interest Rates**
  - An **interest rate** can be interpreted as:
    * The *required rate of return* in equilibrium for a specific investment.
    * The *discount rate* for calculating the present value of future cash flows.
    * The *opportunity cost* of consuming now rather than saving and investing.
  - **Real risk-free rate**:
    * Reflects *time preference* for present goods over future goods.
  - Relationship:

$$\text{Nominal risk-free rate} \approx \text{Real risk-free rate} + \text{Expected inflation}$$

  - **Risk premiums** increase required return:

    - *Default risk premium.*
    - *Liquidity premium.*
    - *Maturity premium.*
  - Nominal interest rate components:

    Nominal Rate = Real Risk-Free Rate+Expected Inflation+Default RP+Liquidity RP+Maturit

- **LOS 1.b: Return Measurement**
  - **Holding Period Return (HPR)**: Return over a specific period.
  - **Arithmetic mean return**:
    - Simple average of periodic returns.
    - Includes all observations, including outliers.
  - **Geometric mean return**:
    - Compound annual growth rate over multiple periods.
    - Used for compounding returns.
  - **Harmonic mean**:
    - Used to calculate the average price paid when equal amounts are invested periodically.
  - **Trimmed mean / Winsorized mean**:
    - Reduce the influence of outliers.
- **LOS 1.c: Money-Weighted vs. Time-Weighted Returns**
  - **Money-weighted rate of return (MWRR)**:
    - IRR of the portfolio.
    - Discount rate that equates PV of cash inflows to PV of cash outflows.
  - **Time-weighted rate of return (TWRR)**:
    - Measures compound growth over a specified horizon.
    - Removes the effect of cash flow timing.
  - Comparison:
    - If funds are added before poor performance $\Rightarrow$ MWRR $<$ TWRR.
    - If funds are added before strong performance $\Rightarrow$ MWRR $>$ TWRR.
    - TWRR is preferred for assessing *manager skill* when they don't control cash flows.
    - MWRR is appropriate if the manager controls cash flows.
- **LOS 1.d: Annualized and Continuously Compounded Returns**
  - Returns are often stated on an annualized basis.

- Continuously compounded return from HPR:

$$R_c = \ln(1 + \text{HPR})$$

- **LOS 1.e: Major Return Measures**
  - **Gross return**:
    * Total return after deducting commissions/trading costs but *before* management and administration fees.
  - **Net return**:
    * Return after deducting management and administration fees.
  - **Pretax nominal return**:
    * Before taxes; ignores inflation.
  - **After-tax nominal return**:
    * After taxes; ignores inflation.
  - **Real return**:
    * Adjusts nominal return for inflation.
    * Approximation: Nominal − Inflation.
  - **Leveraged return**:
    * Gain/loss as a percentage of investor's cash investment.
    * Common in derivatives and real estate.

2 THE TIME VALUE OF MONEY IN FINANCE

2.1 Discounted Cash Flow Valuation

- **Present Value and Future Value Relationship**
  - Discrete compounding:

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

    where:
    * r = interest rate per compounding period.
    * t = number of compounding periods.
  - Continuous compounding:

$$PV = FV \cdot e^{-rt}$$

- **Fixed-Income Securities**
  - *Zero-coupon bond*:

        * Price = PV of face value:

$$P = \frac{\text{Face Value}}{(1+r)^t}$$

        * Interest earned = Face Value − Purchase Price.
        * Example: FV = \$1,000, $t = 15$, $r = 4\%$ (annual):

$$P = \frac{1000}{(1.04)^{15}} = 555.26$$

        * Negative yields ⇒ Price > Face Value.

    - *Coupon bond*:
        * Coupon payment = Coupon Rate × Par Value.
        * Price = PV of coupon payments + PV of face value.
        * Example: 10-year, 10% coupon (\$100/year), Par = \$1,000, YTM = 8%:

$$P = \sum_{t=1}^{10} \frac{100}{(1.08)^t} + \frac{1000}{(1.08)^{10}} = 1{,}134.20$$

    - *Perpetual bond (perpetuity)*:

$$PV = \frac{\text{Payment}}{\text{Discount Rate}}$$

    - *Amortizing bond / Loan payment* (annuity formula):

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

    - Example: Loan \$2,000, $t = 13$, $r = 6\%$ (annual), FV=0:

$$PMT = 225.92$$

- **Equity Securities**
    - *Preferred stock*:
        * Pays fixed dividend D_p forever.
        * Value:

$$P = \frac{D_p}{k_p}$$

        * Example: Par = \$100, Dividend = \$5, $k_p = 8\%$:

$$P = \frac{5}{0.08} = 62.50$$

    - *Common stock*:
        * Residual claim on assets.

- Dividend Discount Models (DDMs):
  1. **Constant dividend** (perpetuity):

$$P_0 = \frac{D}{k_e}$$

  2. **Constant growth** (Gordon Growth Model):

$$P_0 = \frac{D_1}{k_e - g_c}$$

  where $k_e > g_c$.

  3. **Multistage growth**:
     - Step 1: Forecast dividends during high-growth phase individually.
     - Step 2: Apply constant growth model to first dividend in constant-growth phase to get terminal value:

$$P_{n-1} = \frac{D_n}{k_e - g_c}$$

     - Step 3: Discount all cash flows and terminal value to present.

- **Example: Gordon Growth Model**:
  - $D_1 = 1.62$, $g_c = 8\%$, $k_e = 12\%$:

$$P_0 = \frac{1.62}{0.12 - 0.08} = 40.50$$

- **Example: Multistage Growth**:
  - $D_0 = 1.00$, High growth $g_h = 15\%$ for 2 years, then $g_c = 5\%$ forever, $k_e = 11\%$.
  - $D_1 = 1.15$, $D_2 = 1.15 \times 1.15 = 1.3225$.
  - Terminal value at $t = 1$:

$$P_1 = \frac{D_2}{0.11 - 0.05} = \frac{1.3225}{0.06} = 22.0417$$

  - Present value:

$$P_0 = \frac{1.15}{1.11} + \frac{22.0417}{1.11} = \dots$$

- **Module Quiz 2.1 — Key Answers**

1. A — $P = \frac{9}{0.11} = 81.82$.
2. A — Required yield > coupon rate $\Rightarrow$ Price < Par Value.

2.2 Implied Returns and Cash Flow Additivity

- **LOS 2.b: Implied Returns for Fixed-Income and Equity**
  - Rearrange PV–FV–rate relationships to solve for the **required rate of return** given the price and expected cash flows.
  - **Pure discount bond**:
  $$r = \left(\frac{FV}{PV}\right)^{1/t} - 1$$
  - **Coupon bond**:
    * Yield to maturity (YTM) is solved by equating price to PV of coupons and face value.
    * Relationship: Price $\uparrow \Rightarrow$ Yield $\downarrow$ and vice versa.
  - **Equity: Constant Growth Dividend Discount Model (DDM)**
    * Value:
    $$P_0 = \frac{D_1}{k_e - g_c}$$
    * Required return:
    $$k_e = \frac{D_1}{P_0} + g_c$$
    * Implied growth:
    $$g_c = k_e - \frac{D_1}{P_0}$$
    * Dividend yield $= D_1/P_0$.
- **LOS 2.c: Cash Flow Additivity Principle**
  - PV of a stream of cash flows = sum of PVs of each cash flow.
  - Can split or combine streams without changing total PV.
  - Basis for **no-arbitrage principle** ("law of one price"): identical cash flows $\Rightarrow$ identical prices.
  - **Example:** Payments: 100, 100, 400, 100 at 10%.
    * Series 1: 4-year annuity of 100 $\Rightarrow$ PV = 316.99.
    * Series 2: 3-year zero-coupon of 300 $\Rightarrow$ PV = 225.39.
    * Total PV = 542.38, equal to PV of original stream.
- **Applications of No-Arbitrage**
  - **Forward interest rates**:
    * Spot rate notation: $S_n = n$-year rate today.
    * Forward rate notation: $myny = n$-year loan starting m years from now.
    * Relation:
    $$(1 + S_3)^3 = (1 + S_1)(1 + 1y1y)(1 + 2y1y)$$

    * Example: $S_2 = 8\%$, $S_1 = 4\%$:

$$(1.08)^2 = (1.04)(1 + 1y1y) \Rightarrow 1y1y = 12.154\%$$

  - **Forward currency exchange rates**:
    * Price currency / Base currency quotation.
    * No-arbitrage formula:

$$F_{A/B} = S_{A/B} \cdot \frac{(1 + r_A)}{(1 + r_B)}$$

    * Example: Spot = 4.5671 ABE/DUB, $r_{ABE} = 5\%$, $r_{DUB} = 3\%$:

$$F = 4.5671 \cdot \frac{1.05}{1.03} = 4.6558$$

    * Forward premium $\approx r_A - r_B$.
  - **Option pricing (binomial model)**:
    * One-period binomial tree: asset price can go up (S_u) or down (S_d).
    * Hedge ratio:

$$h = \frac{C_u - C_d}{S_u - S_d}$$

      where C_u = option payoff if up, C_d = payoff if down.
    * Portfolio replicates option payoff in both states $\Rightarrow$ price via risk-free discounting.
    * Example:
      · $S_0 = 50$, $S_u = 60$, $S_d = 42$, $X = 55$ (call option).
      · Payoffs: $C_u = 5$, $C_d = 0$.
      · Hedge ratio:

$$h = \frac{5 - 0}{60 - 42} = 0.278$$

      · Portfolio value in 1 period: 11.68 in both states.
      · Risk-free rate = 3%:

$$V_0 = \frac{11.68}{1.03} = 11.34$$

      · Option value:

$$c_0 = hS_0 - V_0 = 0.278 \cdot 50 - 11.34 = 2.56$$

- **Module Quiz 2.2 — Key Answers**

  1. C — Required return = dividend yield + growth rate.
  2. B — Annual return $\approx 13\%$.

Key Concepts — Module 2

- **LOS 2.a: Valuation of Fixed-Income and Equity Securities**
  - The **value** of a fixed-income instrument or equity security = **PV of future cash flows** discounted at the investor's required rate of return.
  - General formula (discrete compounding):

$$PV = \sum_{t=1}^{n} \frac{CF_t}{(1+r)^t}$$

  - **Perpetual bond / Preferred stock**:

$$PV = \frac{\text{Payment}}{r}$$

    where r = required rate of return.
  - **Common stock with constant growth rate of dividends** (Gordon Growth Model):

$$P_0 = \frac{D_1}{k_e - g_c}$$

    where:
    * D_1 = dividend expected next period.
    * k_e = required return on equity.
    * g_c = constant dividend growth rate.

- **LOS 2.b: Implied Returns and Growth Rates**
  - Rearrange PV formulas to calculate:
    * **Required rate of return** given price and expected cash flows.
    * **Implied growth rate** given price, expected dividend, and required return.
  - Price–return relationship is **inverse**:
    * Price $\uparrow \Rightarrow$ Required return $\downarrow$.
    * Price $\downarrow \Rightarrow$ Required return $\uparrow$.
  - For constant dividend growth stock:

$$k_e = \frac{D_1}{P_0} + g_c \quad \text{(Dividend Yield + Growth Rate)}$$

$$g_c = k_e - \frac{D_1}{P_0} \quad \text{(Required Return - Dividend Yield)}$$

- **LOS 2.c: Cash Flow Additivity Principle**

- PV of a stream of cash flows = sum of PVs of each individual cash flow:

$$PV_{\text{total}} = \sum_{t=1}^{n} PV(CF_t)$$

- The series can be split into components, and the sum of the PVs of the components = PV of the original stream.
- Basis for **no-arbitrage condition**:
  * If two sets of future cash flows are identical ⇒ They must have the same PV.
  * If prices differ, arbitrage opportunities exist (buy cheaper, sell more expensive until prices converge).

3 STATISTICAL MEASURES OF ASSET RETURNS

3.1 Central Tendency and Dispersion

- **LOS 3.a: Measures of Central Tendency and Location**
  - **Purpose:** Identify the center (average) of a dataset to represent the typical or expected value.
  - **Arithmetic Mean:**

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

    Example: Returns [30%, 15%, 25%, 21%, 23%]:

    $$\bar{X} = \frac{30 + 15 + 25 + 21 + 23}{5} = 22\%$$

  - **Median:**
    * Odd n: middle value after sorting.
    * Even n: mean of two middle values.

    Example (odd n): [30%, 25%, 23%, 21%, 15%] ⇒ Median = 23%. Example (even n): [30%, 28%, 25%, 23%, 21%, 15%] ⇒ Median = $\frac{25+23}{2}$ = 24%.
  - **Mode:**
    * Most frequent observation.
    * Possible to have 0, 1, 2, or more modes (unimodal, bimodal, trimodal).

    Example: [30%, 28%, 25%, 23%, 28%, 15%, 5%] ⇒ Mode = 28%.
  - **Dealing with Outliers:**
    * **Trimmed mean**: Remove a fixed % of extreme values from each end.
    * **Winsorized mean**: Replace extremes with nearest remaining values.
  - **Measures of Location (Quantiles):**

  * Quartiles: 4 equal parts.
  * Quintiles: 5 equal parts.
  * Deciles: 10 equal parts.
  * Percentiles: 100 equal parts.
  * *Interquartile range*:

$$IQR = Q_3 - Q_1$$

    Example: $Q_3 = 15$, $Q_1 = 5 \Rightarrow IQR = 10$.

- **Box and Whisker Plot:**
  * Box = Interquartile range (IQR).
  * Whiskers = Full range.
  * Detects skewness and outliers visually.

• **LOS 3.b: Measures of Dispersion**

- **Range:**

$$\text{Range} = \max(X) - \min(X)$$

  Example: [30%, 12%, 25%, 20%, 23%] $\Rightarrow$ Range $= 30 - 12 = 18\%$.

- **Mean Absolute Deviation (MAD):**

$$\text{MAD} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

  Example: Returns [30%, 12%, 25%, 20%, 23%], $\bar{X} = 22\%$:

$$\text{MAD} = \frac{|30-22| + |12-22| + |25-22| + |20-22| + |23-22|}{5} = \frac{8+10+3+2+1}{5} = 4.8\%$$

- **Sample Variance:**

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

  Example: Same data, $\bar{X} = 22$:

$$s^2 = \frac{(8^2 + (-10)^2 + 3^2 + (-2)^2 + 1^2)}{4} = \frac{64+100+9+4+1}{4} = \frac{178}{4} = 44.5$$

  Units = $(\%)^2$.

- **Sample Standard Deviation:**

$$s = \sqrt{s^2}$$

  Example: $s = \sqrt{44.5} \approx 6.67\%$.

- **Coefficient of Variation (CV):**

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

  Example: T-bills: $\frac{0.36}{0.25} = 1.44$, S&P 500: $\frac{7.30}{1.09} = 6.70$.

- **Target Downside Deviation (TDD):**

$$TDD = \sqrt{\frac{\sum_{i=1}^{n} \min(0, X_i - B)^2}{n-1}}$$

Example: Target $B = 22\%$, returns [30, 12, 25, 20, 23]:

  * Deviations from target: [8, -10, 3, -2, 1].
  * Only negatives: [-10, -2] $\Rightarrow$ squares: [100, 4].
  * $TDD = \sqrt{\frac{100+4}{4}} = \sqrt{26} \approx 5.10\%$.

- **Module Quiz 3.1 — Key Answers**

1. B — Both trimmed mean and winsorized mean use denominator n.
2. A — Sample standard deviation = 9.8%.
3. B — Target downside deviation = 12.10%.

3.2 Skewness, Kurtosis, and Correlation

- **LOS 3.c: Interpret and evaluate measures of skewness and kurtosis**

  - **Symmetry in Distributions:**
    * A distribution is *symmetrical* if it is identical on both sides of its mean.
    * Symmetry $\Rightarrow$ losses and gains of equal magnitude occur with the same frequency.
    * Example: Mean return = 0%, frequency of losses in $[-6\%, -4\%]$ equals frequency of gains in $[+4\%, +6\%]$.
  - **Skewness:**
    * *Definition:* Degree to which a distribution is not symmetrical.
    * Caused by *outliers* (observations far from the mean).
    * *Positive skew* (right skew): Long right tail, outliers > mean.
    * *Negative skew* (left skew): Long left tail, outliers < mean.
    * **Effect on Mean, Median, Mode:**
      · Symmetrical: Mean = Median = Mode.
      · Positive skew: Mode < Median < Mean (mean pulled right).
      · Negative skew: Mean < Median < Mode (mean pulled left).
    * **Example (Positive Skew):** 100 homes, 99 priced at \$100,000, 1 at \$1,000,000: Median = Mode = \$100,000, Mean = \$109,000.
  - **Sample Skewness (large n):**

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

where s = sample standard deviation.

    - Denominator always > 0.
    - Sign of skewness determined by numerator:
        - > 0: Positive skew (right tail longer).
        - < 0: Negative skew (left tail longer).
    - $|\text{Skewness}| > 0.5 \Rightarrow$ significant skew.
  - **Kurtosis:**
    - Measures *peakedness* and tail thickness.
    - *Mesokurtic*: Same kurtosis as normal distribution ($= 3$).
    - *Leptokurtic*: More peaked, fatter tails (higher risk of extreme outcomes).
    - *Platykurtic*: Flatter, thinner tails.
    - **Excess kurtosis**:

$$\text{Excess Kurtosis} = \text{Kurtosis} - 3$$

        - Normal distribution: 0.
        - Leptokurtic: > 0.
        - Platykurtic: < 0.
  - **Risk Management Note:**
    - Security returns often exhibit skewness *and* excess kurtosis.
    - Greater excess kurtosis & negative skew $\Rightarrow$ higher risk of extreme negative returns.
    - Risk managers focus on tail behaviour, not just mean and standard deviation.
  - **Sample Kurtosis (large n):**

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

- **LOS 3.d: Interpret correlation between two variables**
  - **Scatter Plots:**
    - Plot one variable on x-axis, the other on y-axis.
    - Useful for detecting both linear and nonlinear relationships.
    - Can reveal patterns even if correlation coefficient is near zero.
  - **Covariance:**

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

    - Positive: Variables move in same direction.
    - Negative: Variables move in opposite directions.
    - Units: Product of units of X and Y.
    - Hard to interpret magnitude directly $\Rightarrow$ use correlation.

- **Correlation Coefficient:**

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{s_X s_Y}$$

  * Ranges: $-1 \leq \rho \leq +1$.
  * $\rho = 1$: Perfect positive linear relationship.
  * $\rho = -1$: Perfect negative linear relationship.
  * $\rho = 0$: No linear relationship.

- **Example (Correlation):**
  * Given:

$$\text{Var}(A) = 0.0028, \quad \text{Var}(B) = 0.0124, \quad \text{Cov}(A,B) = 0.0058$$

  * Step 1: Standard deviations:

$$s_A = \sqrt{0.0028} \approx 0.0529, \quad s_B = \sqrt{0.0124} \approx 0.1114$$

  * Step 2: Correlation:

$$\rho_{AB} = \frac{0.0058}{0.0529 \times 0.1114} \approx 0.981$$

    Interpretation: Very strong positive linear relationship.

- **Important Notes:**
  * Correlation $\neq$ causation.
  * Outliers can distort correlation.
  * Spurious correlation: High correlation due to association with a third variable or time trends.
  * Example: U.S. spending on science vs. suicides by hanging (1999–2009), correlation = 0.9987 — clearly unrelated causally.

- **Module Quiz 3.2 — Key Answers:**

  1. A — Mean > Median $\Rightarrow$ Positively skewed.
  2. B — Greater
  3. B — $\rho = +0.25$: When one variable is above its mean, the other tends to be above its mean too (weak positive association).

Key Concepts — Modul 3

- **LOS 3.a: Measures of Central Tendency**

  - **Arithmetic Mean** (average):

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

    - For a sample: $\bar{X}_{\text{sample}}$ is the arithmetic mean of sample values.
    - Example: Returns $= \{5\%, 7\%, 8\%\}$, Mean $= \frac{0.05+0.07+0.08}{3} \approx 0.0667$ or 6.67%.
  - **Median**:
    - The middle value when data are sorted from largest to smallest.
    - Example: $\{3, 5, 8, 12, 20\} \rightarrow$ Median $= 8$.
  - **Mode**:
    - Value that occurs most frequently in the dataset.
    - For continuous data, the *modal interval* is used.
    - Example: $\{2, 4, 4, 5, 6\} \rightarrow$ Mode $= 4$.
  - **Trimmed Mean**:
    - Omits outliers before calculating the mean.
    - Example: Returns $= \{5\%, 6\%, 100\%\}$, trimmed mean (removing 100%) $=$ Mean of $\{5\%, 6\%\} = 5.5\%$.
  - **Winsorized Mean**:
    - Replaces outliers with specified values before computing mean.
    - Example: Replace 100% return with 10% in above dataset before averaging.
  - **Quantiles**:
    - General term for a value below which lies a stated proportion of the data.
    - Examples:
      - Quartile: 4 equal parts.
      - Quintile: 5 equal parts.
      - Decile: 10 equal parts.
      - Percentile: 100 equal parts.

- **LOS 3.b: Measures of Dispersion**
  - **Range**:

$$\text{Range} = X_{\max} - X_{\min}$$

    Example: Prices $= \{20, 25, 18, 30\} \rightarrow$ Range $= 30 - 18 = 12$.
  - **Mean Absolute Deviation (MAD)**:

$$\text{MAD} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

    Example: Data $= \{2, 4, 6\}$, Mean $= 4$, MAD $= \frac{|2-4|+|4-4|+|6-4|}{3} = \frac{2+0+2}{3} \approx 1.33$.
  - **Variance**:

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

  - **Standard Deviation (SD)**:

$$s = \sqrt{s^2}$$

    Often used as a quantitative measure of risk (volatility).

- **Coefficient of Variation (CV)**:

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

  Lower CV $\rightarrow$ more return per unit of risk.

- **Target Downside Deviation (Semideviation)**:

$$\text{Semideviation} = \sqrt{\frac{\sum_{i=1}^{n} \min(X_i - B, 0)^2}{n}}$$

  where B = target return (often 0 or risk-free rate).

- **LOS 3.c: Skewness and Kurtosis**
  - **Skewness**:
    - Positive skew (right tail longer): Mean $>$ Median $>$ Mode.
    - Negative skew (left tail longer): Mean $<$ Median $<$ Mode.
    - Example: Housing prices with one very expensive property $\rightarrow$ positive skew.
  - **Kurtosis**:
    - Measures peakedness and probability of extreme outcomes.
    - Excess kurtosis:

$$\text{Excess Kurtosis} = \text{Kurtosis} - 3$$

    - Positive excess kurtosis (> 0): Leptokurtic — fat tails, more peaked.
    - Negative excess kurtosis (< 0): Platykurtic — thin tails, less peaked.

- **LOS 3.d: Correlation**
  - **Definition:** Standardized measure of linear association between two random variables:

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{s_X s_Y}$$

  - Range: $-1 \leq \rho \leq 1$.
    - $+1$: Perfect positive correlation.
    - -1: Perfect negative correlation.
    - 0: No linear relationship.
  - Scatter plots can reveal nonlinear patterns missed by correlation.
  - **Important:** Correlation does *not* imply causation.
  - **Spurious correlation:** Apparent correlation due to chance or common relationship with a third variable.
  - **Example:** U.S. spending on science vs. suicides by hanging (1999–2009), correlation = 0.9987, clearly not causal.

4 PROBABILITY TREES AND CONDITIONAL EXPECTATIONS

4.1: Probability Models, Expected Values, and Bayes' Formula

- **LOS 4.a: Expected Values, Variances, and Standard Deviations**
  - **Expected Value** of a discrete random variable X with outcomes $x_1, x_2, \ldots, x_n$ and probabilities $P(x_i)$:

$$E[X] = \sum_{i=1}^{n} P(x_i) \cdot x_i$$

    It represents the probability-weighted average of all possible outcomes.
  - **Example: Expected EPS**

$$E[\text{EPS}] = 0.10(1.80) + 0.20(1.60) + 0.40(1.20) + 0.30(1.00) = 1.28$$

EPS	Probability
1.80	0.10
1.60	0.20
1.20	0.40
1.00	0.30

Table 1: EPS Probability Distribution for Ron's Stores

  - **Variance and Standard Deviation** (from a probability model):

$$\text{Var}(X) = \sum_{i=1}^{n} P(x_i) \cdot (x_i - E[X])^2$$

$$\text{SD}(X) = \sqrt{\text{Var}(X)}$$

  - **Example: Stock A returns**

Scenario	Probability	Return	$R - E[R]$	$(R - E[R])^2$	$P \cdot (R - E[R])^2$
Good	0.30	0.20	0.07	0.0049	0.00147
Normal	0.50	0.12	-0.01	0.0001	0.00005
Poor	0.20	0.05	-0.08	0.0064	0.00128
				Variance	0.00280
				SD	5.29%

Table 2: Expected Return, Variance, and SD for Stock A

Expected return:

$$E[R_A] = (0.30)(0.20) + (0.50)(0.12) + (0.20)(0.05) = 0.13 \text{ (or 13\%)}$$

- **LOS 4.b: Probability Trees and Conditional Expectations**
  - **Probability Tree:** A graphical representation of possible outcomes, their probabilities, and resulting payoffs.
  - Probabilities at each branch multiply to give the joint probability of a final outcome.
  - Conditional expectations are computed by conditioning on a known event.
  - **Example: Expected EPS from probability tree**

Economy	Company Result	Joint Probability	EPS
Good	Good	0.18	1.80
Good	Poor	0.42	1.70
Poor	Good	0.24	1.30
Poor	Poor	0.16	1.00

Table 3: EPS Outcomes from Probability Tree

Expected EPS:

$$E[\text{EPS}] = (0.18)(1.80) + (0.42)(1.70) + (0.24)(1.30) + (0.16)(1.00) = 1.51$$

  - **Investment Application:** If a tariff is imposed, the conditional expected return for a domestic steel stock may increase relative to the no-tariff scenario.

- **LOS 4.c: Bayes' Formula for Updating Probabilities**
  - **Bayes' Formula:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)}$$

where:

    * $P(A)$ = prior probability of event A
    * $P(B|A)$ = probability of B given A
    * $P(A|B)$ = updated (posterior) probability after observing B
  - **Example: Economic Outperformance and Stock Gains**

Economy	Prob(Economy)	Prob(Gain—Economy)	Joint Probability
Outperform	0.60	0.70	0.42
Underperform	0.40	0.20	0.08
Outperform	0.60	0.30 (Loss)	0.18
Underperform	0.40	0.80 (Loss)	0.32

Table 4: Joint Probabilities for Economic Performance and Stock Gains

Total probability of gains:

$$P(\text{Gain}) = 0.42 + 0.08 = 0.50$$

Updated probability of outperforming economy given gains:

$$P(\text{Outperform}|\text{Gain}) = \frac{0.42}{0.50} = 0.84 \text{ (or 84\%)}$$

Interpretation: Observing stock gains increases the likelihood we assign to economic outperformance.

Key Concepts — Module 4

- **LOS 4.a: Expected Value, Variance, and Standard Deviation**
  - **Expected Value** of a discrete random variable X with outcomes $x_1, x_2, \ldots, x_n$ and probabilities $P(x_i)$:

    $$E[X] = \sum_{i=1}^{n} P(x_i) \cdot x_i$$

    Interpretation: The probability-weighted average of all possible outcomes.
  - **Variance** (probability model):

    $$\text{Var}(X) = \sum_{i=1}^{n} P(x_i) \cdot (x_i - E[X])^2$$

  - **Standard Deviation**:

    $$\sigma_X = \sqrt{\text{Var}(X)}$$

  - **Example: Expected EPS and Risk**

EPS	Probability
1.80	0.10
1.60	0.20
1.20	0.40
1.00	0.30

Table 5: EPS Probability Distribution

$$E[\text{EPS}] = 0.10(1.80) + 0.20(1.60) + 0.40(1.20) + 0.30(1.00) = 1.28$$
$$\text{Var(EPS)} = \sum P_i(x_i - 1.28)^2, \quad \sigma_{\text{EPS}} = \sqrt{\text{Var}}$$

- **LOS 4.b: Probability Trees and Conditional Expectations**

- **Probability Tree:** Graphical tool showing:
  * Probabilities of initial events
  * Conditional probabilities of subsequent events
  * Joint probabilities (product of branch probabilities)
- **Conditional Expected Value:** The expected value of a variable given the outcome of another event:

$$E[X \mid A] = \sum_i P(x_i \mid A) \cdot x_i$$

  This updates forecasts when new information arrives.
- **Example: EPS from Probability Tree**

Economy	Company Result	Joint Probability	EPS
Good	Good	0.18	1.80
Good	Poor	0.42	1.70
Poor	Good	0.24	1.30
Poor	Poor	0.16	1.00

Table 6: EPS Outcomes from Probability Tree

$$E[\text{EPS}] = (0.18)(1.80) + (0.42)(1.70) + (0.24)(1.30) + (0.16)(1.00) = 1.51$$

- **Investment Application:** If a new policy (e.g., tariff) occurs, recalculate expected returns using only the relevant conditional probabilities.

- **LOS 4.c: Bayes' Formula for Updating Probabilities**
  - **Formula:**

$$P(A|O) = \frac{P(O|A) \cdot P(A)}{P(O|A) \cdot P(A) + P(O|A^c) \cdot P(A^c)}$$

    where:
    * $P(A)$ = prior probability of event A
    * $P(O|A)$ = probability of observing O given A
    * $P(A|O)$ = updated (posterior) probability after observing O
  - **Example: Economic Outperformance and Stock Gains**

Economy	P(Economy)	P(Gain | Economy)	Joint Probability
Outperform	0.60	0.70	0.42
Underperform	0.40	0.20	0.08

Table 7: Joint Probabilities for Economic State and Stock Gains

$$P(\text{Gain}) = 0.42 + 0.08 = 0.50$$

$$P(\text{Outperform} \mid \text{Gain}) = \frac{0.42}{0.50} = 0.84 \text{ (or 84\%)}$$

- **Interpretation:** Observing stock gains significantly increases the probability estimate for an outperforming economy.

5 PORTFOLIO MATHEMATICS

5.1: Probability Models for Portfolio Return and Risk

LOS 5.a: Expected Value, Variance, Standard Deviation, Covariances, and Correlations of Portfolio Returns

- **Expected Portfolio Return:**

$$E(R_p) = \sum_{i=1}^{n} w_i \cdot E(R_i)$$

where:

  - w_i = weight of asset i in the portfolio
  - $E(R_i)$ = expected return of asset i

- **Portfolio Weights:**

$$w_i = \frac{\text{Market value invested in asset } i}{\text{Total market value of portfolio}}$$

- **Covariance:** Measures how two assets move together:

$$\text{Cov}(R_i, R_j) = E\left[(R_i - E(R_i))(R_j - E(R_j))\right]$$

Properties:

  - Covariance of a variable with itself = variance: $\text{Cov}(R_A, R_A) = \text{Var}(R_A)$
  - Positive covariance $\rightarrow$ assets move in the same direction
  - Negative covariance $\rightarrow$ assets move in opposite directions

- **Sample Covariance:**

$$\text{Cov}(R_i, R_j) = \frac{\sum_{k=1}^{n}(R_{i,k} - \bar{R}_i)(R_{j,k} - \bar{R}_j)}{n-1}$$

- **Portfolio Variance (Two Assets):**

$$\sigma_p^2 = w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_A w_B \cdot \text{Cov}(R_A, R_B)$$

  Alternative using correlation:

$$\sigma_p^2 = w_A^2\sigma_A^2 + w_B^2\sigma_B^2 + 2w_A w_B \rho_{A,B}\sigma_A\sigma_B$$

- **Portfolio Variance (Three Assets):**

$$\sigma_p^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} w_i w_j \cdot \text{Cov}(R_i, R_j)$$

- **Example: Three-Asset Portfolio**

Asset	Weight (%)	Var (%^{2})	Covariances (%^{2})
Domestic Stocks	60	2.25	Cov(S,B)=0.18, Cov(S,IE)=0.15
Domestic Bonds	30	0.64	Cov(B,IE)=0.12
Intl Equities	10	3.24	

Table 8: Covariance Matrix Data Example

$$\sigma_p^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} w_i w_j \cdot \text{Cov}(R_i, R_j)$$

  A lower covariance between assets leads to lower portfolio variance.

LOS 5.b: Covariance and Correlation from a Joint Probability Function

- Given joint probabilities, we can calculate:

$$E(R_A) = \sum_s P(s) \cdot R_A(s), \quad E(R_B) = \sum_s P(s) \cdot R_B(s)$$

$$\text{Cov}(R_A, R_B) = \sum_s P(s) \cdot [R_A(s) - E(R_A)]\,[R_B(s) - E(R_B)]$$

- **Example: Covariance from Economic States**

State	Probability	R_A	R_B
Boom	0.30	0.20	0.30
Normal	0.50	0.12	0.10
Slow	0.20	0.05	0.00

Table 9: Joint Probability Model

Expected returns:

$$E(R_A) = 0.13, \quad E(R_B) = 0.14$$

Covariance:

$$\text{Cov} = 0.30(0.07 \cdot 0.16) + 0.50(-0.01 \cdot -0.04) + 0.20(-0.08 \cdot -0.14)$$

$$\text{Cov} = 0.00336 + 0.00020 + 0.00224 = 0.005784$$

LOS 5.c: Shortfall Risk and Roy's Safety-First Criterion

- **Shortfall Risk:** Probability that $R_p < R_L$ over a period.
- **Roy's Safety-First Ratio:**

$$SFR = \frac{E(R_p) - R_L}{\sigma_p}$$

For normally distributed returns, the portfolio with the highest SFR has the lowest probability of shortfall.

- **Example: Safety-First Decision**

Portfolio	$E(R_p)$	σ_p
A	12%	18%
B	10%	12%

Table 10: Portfolios for Shortfall Analysis

For $R_L = 0\%$:

$$SFR_A = \frac{0.12 - 0}{0.18} = 0.667, \quad SFR_B = \frac{0.10 - 0}{0.12} = 0.833$$

Portfolio B has higher SFR → lower probability of negative returns.

- **Example: Endowment Fund Decision**

Portfolio	$E(R_p)$	σ_p	SFR ($R_L = 3\%$)
A	9%	12%	$(0.09 - 0.03)/0.12 = 0.50$
B	8%	10%	$(0.08 - 0.03)/0.10 = 0.50$
C	7%	8%	$(0.07 - 0.03)/0.08 = 0.50$

Table 11: Safety-First Ratios for Endowment Portfolios

Higher SFR → better choice for minimizing shortfall risk.

6 SIMULATION METHODS

6.1: Lognormal Distributions and Simulation Techniques

LOS 6.a: Relationship Between Normal and Lognormal Distributions

- A **lognormal distribution** is generated by the function e^x, where x is normally distributed.
- If Y is lognormally distributed, then $\ln(Y)$ is normally distributed.
- **Key Property:** Lognormal variables are strictly positive ($P(Y < 0) = 0$).
- Asset prices are modeled as lognormal because:

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

  where:

  - P_T = asset price at time T
  - P_0 = asset price today
  - $r_{0,T}$ = continuously compounded return from 0 to T
- If returns $r_{0,T}$ are normally distributed, then prices P_T are lognormally distributed.
- **Assumptions:**
  - Returns are **i.i.d.** (independently and identically distributed).
  - Stationarity: mean and variance of returns do not change over time.

LOS 6.b: Monte Carlo Simulation

- **Definition:** Repeated generation of random values for one or more risk factors to simulate possible asset values.
- **Procedure:**
  1. Define the probability distribution(s) and parameters (mean, variance, skewness) of each risk factor.
  2. Generate random values for each risk factor.
  3. Value the security for each generated set of inputs.
  4. Repeat the process many times ($\geq 1,000$ simulations).
  5. Analyze the simulated distribution of asset values.
- **Applications:**
  - Valuation of complex securities (e.g., path-dependent options).

- Estimating profits/losses from a trading strategy.
- Calculating **Value at Risk (VaR)**.
- Modeling pension fund assets and liabilities.
- Valuing portfolios with nonnormal return distributions.

- **Advantages:**
  - Can test scenarios not present in historical data.

- **Limitations:**
  - Results depend on assumptions of input distributions.
  - Statistical method — does not yield closed-form solutions.
  - Complexity of implementation.

Example: Monte Carlo Simulation for Option Valuation

1. Assume the stock price follows $S_T = S_0 e^{(\mu - 0.5\sigma^2)T + \sigma\epsilon\sqrt{T}}$, where $\epsilon \sim N(0, 1)$.
2. Generate random stock prices for many paths.
3. Apply option payoff formula for each simulated price.
4. Average the payoffs and discount to present value.

LOS 6.c: Bootstrap Resampling

- **Definition:** Statistical technique for generating simulated datasets by repeatedly sampling (with replacement) from observed historical data.

- **Procedure:**
  1. Start with observed historical returns (n data points).
  2. Draw a sample of size n **with replacement**.
  3. Compute statistics (mean, variance, etc.).
  4. Repeat many times to build a distribution of statistics.

- **Advantages:**
  - Preserves statistical properties of observed data.
  - Useful when the true population distribution is unknown.

- **Limitations:**
  - Inputs are limited to the range of actual historical outcomes.

Comparison: Monte Carlo vs. Bootstrap

Aspect	Monte Carlo	Bootstrap Resampling
Data Source	Theoretical probability distributions	Historical observed data
Flexibility	Can simulate unseen scenarios	Limited to observed outcomes
Complexity	Requires model specification	Computational but simpler
Use Case	Pricing models, risk management	Estimating sampling variability

7 ESTIMATION AND INFERENCE

7.1: Sampling Techniques and the Central Limit Theorem

LOS 7.a: Sampling Methods and Implications for Sampling Error

Probability Sampling:

- **Simple Random Sampling:** Each member of the population has an equal chance of selection.
  - Example: Selecting 5 bonds out of 50 by numbering each and randomly drawing.
  - Tools: Random number generator or random number table.
  - **Systematic Sampling:** Select every n-th member from a population (approximate random sampling).
- **Stratified Random Sampling:** Population divided into homogeneous subgroups (strata), and random samples are taken from each.
  - Useful for **bond indexing**: Stratify by duration, maturity, coupon rate.
  - Ensures proportional representation from each stratum.
  - Example: Municipal bond index of $1,000$ bonds. A cell with 50 bonds (2–4 years maturity, coupon ¡ 5%) $\Rightarrow$ select $(50/1,000) \times 100 = 5$ bonds.
- **Cluster Sampling:** Population divided into heterogeneous subgroups (clusters), each assumed to represent the overall population.
  - **One-stage:** Select clusters, include all items in them.
  - **Two-stage:** Select clusters, then random sample from within each cluster.
  - Higher sampling error than simple random sampling if clusters are not truly representative.
  - Example: Sampling counties to estimate state average income.

Nonprobability Sampling:

- **Convenience Sampling:** Select based on ease of access (greater sampling error).
- **Judgmental Sampling:** Researcher selects based on expertise/judgment.
  - Example: Auditing firms most likely to violate accounting standards.
  - Risk: Researcher bias may distort representativeness.

Key Point: Ensure population distribution is constant over sampling period — avoid combining data from periods with structural changes.

Comparison of Sampling Methods

Method	Advantages	Disadvantages
Simple Random	Easy to understand, unbiased	May not represent subgroups proportionally
Stratified Random	Ensures representation from each subgroup	More complex; needs classification
Cluster	Lower cost, quicker	Greater sampling error if clusters not representative
Convenience	Very easy, low cost	High sampling error, low representativeness
Judgmental	Can focus on relevant data	Risk of bias, higher error if judgment is poor

LOS 7.b: Central Limit Theorem (CLT)

Statement: For simple random samples of size n from a population with mean μ and finite variance σ^2, the sampling distribution of the sample mean:

- Approaches normal distribution as $n \to \infty$ (typically $n \geq 30$ is sufficient).
- Mean of sampling distribution = μ.
- Variance of sampling distribution = σ^2/n.

Standard Error of Sample Mean:

- If σ known:
$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$
- If σ unknown:
$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$
  where s = sample standard deviation.

Example 1: $n = 30$

$$SE_{\bar{x}} = \frac{0.20}{\sqrt{30}} \approx 0.036 \quad (3.6\%)$$

Interpretation: Mean = 2%, standard error = 3.6%.

Example 2: $n = 200$

$$SE_{\bar{x}} = \frac{0.20}{\sqrt{200}} \approx 0.014 \quad (1.4\%)$$

Observation: Increasing sample size reduces standard error $\Rightarrow$ more precise mean estimate.

LOS 7.c: Resampling Methods (Bootstrap, Jackknife)

Jackknife:

- Remove one observation at a time, compute sample mean for each subsample.
- Standard deviation of these means $\Rightarrow$ estimate of standard error.
- Useful for small datasets, computationally simple.
- Removes bias from some estimators.

Bootstrap:

- Draw repeated samples of size n **with replacement** from the dataset.
- Compute statistic (mean, median, etc.) for each sample.
- Standard deviation of these statistics $\Rightarrow$ estimate of standard error.
- Can estimate distribution of complex statistics and build confidence intervals.

Aspect	**Jackknife**	**Bootstrap**
Computation	Simple, low cost	More computationally intensive
Data Use	Omits one observation at a time	Resamples with replacement
Best For	Small datasets, bias correction	Estimating sampling distribution, complex statistics

Comparison of Jackknife vs Bootstrap

8 HYPOTHESIS TESTING

8.1: Hypothesis Testing Basics

LOS 8.a: Hypothesis Testing Concepts and Components

Definition:

- **Hypothesis:** Statement about the value of a population parameter (e.g., μ) developed for testing a theory or belief.
- Example: μ = mean daily return on stock options. Hypothesis: mean daily return > 0.
- Hypotheses are tested using sample statistics and probability theory to decide whether to reject or fail to reject the statement.

Steps in Hypothesis Testing (Figure 8.1):

1. State the **null hypothesis** (H_0) and the **alternative hypothesis** (H_a).
2. Identify the appropriate test statistic and its distribution.
3. Select the significance level α.
4. State the decision rule (critical value approach or p-value approach).
5. Collect sample data and compute the test statistic.
6. Make a decision: reject H_0 or fail to reject H_0.

Null vs. Alternative Hypothesis:

- **Null Hypothesis** (H_0): Hypothesis to be tested; includes the "equal to" condition.

$$H_0 : \mu = \mu_0$$

- **Alternative Hypothesis** (H_a): Conclusion if H_0 is rejected; mutually exclusive and exhaustive with H_0.

$$H_a : \mu \neq \mu_0$$

Two-Tailed Z-Test Example ($\alpha = 0.05$):

- Critical z-values: ± 1.96 (since $\alpha/2 = 0.025$ per tail).
- **Decision Rule:** Reject H_0 if $z < -1.96$ or $z > 1.96$.
- If $|z| \leq 1.96$, fail to reject H_0.

Test Statistic Formula:

$$\text{Test Statistic} = \frac{\text{Sample Statistic} - \text{Hypothesized Value}}{\text{Standard Error of the Statistic}}$$

Standard error for sample mean:

$$SE_{\bar{x}} = \begin{cases} \frac{\sigma}{\sqrt{n}}, & \text{if population } \sigma \text{ known} \\ \frac{s}{\sqrt{n}}, & \text{if } \sigma \text{ unknown} \end{cases}$$

Type I and Type II Errors

Definitions:

- **Type I Error:** Reject H_0 when it is true. Probability = α.
- **Type II Error:** Fail to reject H_0 when it is false. Probability = β.
- **Power of a Test:** $1 - \beta$ = probability of correctly rejecting a false H_0.

	H_0 **True**	H_0 **False**
Reject H_0	Type I Error (Prob = α)	Correct Decision (Power = $1 - \beta$)
Fail to Reject H_0	Correct Decision (Prob = $1 - \alpha$)	Type II Error (Prob = β)

Relationships:

- Lowering α reduces Type I error but increases β (reduces power).
- Increasing sample size reduces β and increases power.
- Trade-off between α and β for a given n.

Example:

- Given: $\alpha = 0.05$, $\beta = 0.60$.
- Power = $1 - \beta = 0.40$.
- Interpretation: 5% probability of rejecting a true H_0, 40% probability of rejecting H_0 when false.

One-Tailed vs. Two-Tailed Tests

Test Type	**Alternative Hypothesis**	**Critical Region**
One-Tailed (Right)	$H_a : \mu > \mu_0$	Upper tail beyond $z_{1-\alpha}$
One-Tailed (Left)	$H_a : \mu < \mu_0$	Lower tail below $-z_{1-\alpha}$
Two-Tailed	$H_a : \mu \neq \mu_0$	Both tails beyond $\pm z_{\alpha/2}$

p-Value Approach

- **p-value:** Smallest α at which H_0 can be rejected.
- Decision rule:

$$\text{If } p\text{-value} \leq \alpha, \text{ reject } H_0; \quad \text{else fail to reject.}$$

Module Quiz 8.1 Examples

Q1: $\beta = 0.60$, $\alpha = 0.05$.

$$\text{Power} = 1 - \beta = 0.40$$

Correct statement: 5% chance of rejecting a true H_0, 40% chance of rejecting when false.

Q2: $\alpha = 0.05$, $\beta = 0.15$.

$$\text{Power} = 1 - 0.15 = 0.85$$

Answer: 0.850.

8.2 Types of Hypothesis Tests

- **Overview:** Hypothesis tests differ based on:
  - Parameter being tested (mean, variance, difference in means, difference in variances)
  - Sample type (independent vs dependent)
  - Underlying distributional assumptions (parametric vs nonparametric)
- **Tests for Population Mean**
  - Use **t-test** if σ unknown and sample small; **z-test** if σ known or n large.
  - **Example:** $n = 250$, $\bar{x} = 0.1\%$, $s = 0.25\%$, $H_0 : \mu = 0$, $\alpha = 0.05$

$$SE = \frac{0.25\%}{\sqrt{250}} = 0.0158\%$$
$$z = \frac{0.1\% - 0}{0.0158\%} \approx 6.33$$

Decision: $|z| > 1.96 \Rightarrow$ reject H_0.

- **Difference Between Means (Independent Samples)**
  - Applicable if samples are independent, populations normally distributed.
  - If variances are equal but unknown, use pooled variance:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

    where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

  - **Example:** Horizontal mergers: $\bar{x}_1 = 1.0\%$, $s_1 = 1.0\%$, $n_1 = 61$; Vertical mergers: $\bar{x}_2 = 2.5\%$, $s_2 = 2.0\%$, $n_2 = 61$.

$$t = -5.474, \; df = 120, \; t_{crit} = \pm 1.980$$

    Decision: $t < -1.980 \Rightarrow$ reject H_0.

- **Paired Comparisons Test (Dependent Samples)**
  - Used when same units measured twice or observations are paired.
  - t-statistic:

$$t = \frac{\bar{d} - \mu_{d0}}{s_d/\sqrt{n}}, \quad df = n - 1$$

  - **Example:** Change in betas before vs after deregulation: $t = 10.26$, $n = 39$, $t_{crit} = \pm 2.024 \Rightarrow$ reject H_0.

- **Test for a Single Population Variance**
  - Use **Chi-square test**:

$$\chi^2 = \frac{(n - 1)s^2}{\sigma_0^2}, \quad df = n - 1$$

  - **Example:** $\sigma_0 = 4\%$, $s = 3.8\%$, $n = 24$, $\chi^2 = 20.76$, $CV_{lower} = 11.689$, $CV_{upper} = 38.076$. Decision: statistic inside bounds $\Rightarrow$ fail to reject H_0.

- **Equality of Two Population Variances**
  - Use **F-test**:

$$F = \frac{s_1^2}{s_2^2}, \quad df_1 = n_1 - 1, \; df_2 = n_2 - 1$$

  - Convention: put larger variance in numerator so only upper-tail CV is needed.
  - **Example:** Textile industry $s = 4.30$, Paper industry $s = 3.80$, $F = 1.2805$, $CV_{upper} = 1.94 \Rightarrow$ fail to reject H_0.

- **Parametric vs Nonparametric Tests**

	Parametric	**Nonparametric**
Assumptions	Strong (e.g., normality)	Few / none
Parameters tested	Mean, variance, etc.	Median, rank, randomness
Example tests	t-test, z-test, F-test	Runs test, sign test
Use cases	Interval/ratio data, CLT holds	Ordinal data, non-normal small samples

Key Decision Guidelines:

- Independent samples $\Rightarrow$ Difference in means t-test.
- Dependent samples $\Rightarrow$ Paired comparisons t-test.
- Single variance $\Rightarrow$ Chi-square test.
- Variance equality $\Rightarrow$ F-test.
- If assumptions fail or data ordinal $\Rightarrow$ Nonparametric test.

9 PARAMETRIC AND NON-PARAMETRIC TESTS OF INDEPENDENCE

9.1: Tests for Independence

- **LOS 9.a: Parametric and Nonparametric Tests of** $\rho = 0$
  - **Goal:** Test whether the **population correlation coefficient** ρ is equal to zero (no linear relationship).
  - **Parametric test: Pearson correlation coefficient**
    * Applicable when both variables are **normally distributed**.
    * **Test statistic:**

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

      where:
      · r: sample correlation coefficient
      · n: sample size
    * Degrees of freedom: $df = n - 2$
    * **Decision rule:** Reject $H_0 : \rho = 0$ if $|t| > t_{\alpha/2, df}$
  - **Example (Parametric):**
    * $r = 0.35$, $n = 42$

$$t = \frac{0.35\sqrt{42-2}}{\sqrt{1-(0.35)^2}} = 2.363$$

    - Critical value ($df = 40$, $\alpha = 0.05$ two-tailed) = 2.021
    - Since $2.363 > 2.021$, reject H_0: correlation is significantly different from zero.
  - **Nonparametric test: Spearman rank correlation**
    - Used when data are ordinal (ranks) or assumptions of normality are violated.
    - Ranks are assigned to values; ties share average rank.
    - **Formula (integer ranks, no ties):**

$$r_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

      where d_i = difference in ranks for observation i.
    - Test statistic (for $n > 30$):

$$t = \frac{r_S \sqrt{n-2}}{\sqrt{1 - r_S^2}}$$

      with $df = n - 2$.

- **LOS 9.b: Tests of Independence via Contingency Tables**
  - **Purpose:** Test whether two categorical variables are independent.
  - Data are summarized in a **contingency table** (cross-tab).
  - **Hypotheses:**

$$H_0 : \text{Variables are independent} \qquad H_a : \text{Variables are dependent}$$

  - **Test statistic: Chi-square (χ^2) test:**

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

    where:
    - O_{ij} = observed frequency in cell (i, j)
    - E_{ij} = expected frequency in cell (i, j) under independence:

$$E_{ij} = \frac{(\text{row total}_i)(\text{column total}_j)}{\text{grand total}}$$

    - r = number of rows, c = number of columns
    - Degrees of freedom: $(r - 1)(c - 1)$
  - **Example (Independence Test):**

Table 12: Observed Frequencies: Earnings Growth vs Dividend Yield

	Low DY	Med DY	High DY	Row Total
Low EG	28	14	7	49
Med EG	20	10	10	40
High EG	12	25	24	61
Col Total	60	49	41	150

- **Expected Frequencies Example:**

$$E_{1,1} = \frac{(49)(60)}{150} = 19.6, \quad E_{3,2} = \frac{(61)(49)}{150} = 19.9$$

- Sum over all cells:

$$\chi^2_{\text{calc}} = 27.43$$

- $df = (3-1)(3-1) = 4$, $\chi^2_{0.05,4} = 9.488$
- Decision: Since $27.43 > 9.488$, reject H_0: Earnings growth and dividend yield are dependent.

Summary Table

Test	**When Used**	**Test Statistic**	**DF**
Pearson t-test	$\rho = 0$, normal vars	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	$n-2$
Spearman rank	Ranks or non-normal vars	Same t formula with r_S	$n-2$
Chi-square	Categorical var independence	$\sum \frac{(O-E)^2}{E}$	$(r-1)(c-1)$

10 SIMPLE LINEAR REGRESSION

10.1: Linear Regression Basics

- **LOS 10.a: Simple Linear Regression Model**
  - **Purpose:** Explain variation in a **dependent variable** Y in terms of variation in a single **independent variable** X.
  - **Terminology:**
    - Dependent variable: explained variable, endogenous variable, predicted variable.
    - Independent variable: explanatory variable, exogenous variable, predicting variable.
  - **Example: Dependent vs. Independent Variables**
    - Predict stock returns with GDP growth:
      Y (dependent) = stock returns, X (independent) = GDP growth.
  - **Model Form:**

$$Y_i = b_0 + b_1 X_i + \varepsilon_i$$

  where:

    - b_0: intercept (value of Y when $X = 0$)
    - b_1: slope (change in Y for a 1-unit change in X)
    - ε_i: error term (unexplained variation)

- **Estimation Method: Ordinary Least Squares (OLS)**
  - Finds b_0, b_1 that minimize:

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  - $\hat{Y}_i = b_0 + b_1 X_i$ are the **fitted values**.
- **Formulas:**

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$b_0 = \bar{Y} - b_1 \bar{X}$$

- **Interpretation:**
  - $b_1 > 0$: positive relationship between X and Y.
  - $b_1 < 0$: negative relationship.
  - b_0: predicted Y when $X = 0$.
- **Example (ABC vs. S&P 500 Excess Returns):**
  - Estimated regression: $\hat{Y} = -0.023 + 0.64X$
  - If $X = -0.078$, predicted $Y = -0.023 + 0.64(-0.078) = -0.073$ (-7.3%).
  - Residual for May 20X4 = Actual Y − Predicted $Y = 0.011 - (-0.073) = 0.084$ (8.4%).
- **Special Case: Beta in CAPM**
  - In a regression of excess security returns on market returns:

$$b_1 = \beta = \text{systematic risk of the security}$$

  - $\beta < 1$: less risky than market, $\beta > 1$: more risky.

- **LOS 10.b: Assumptions of the Simple Linear Regression Model**
  1. **Linearity:** Relationship between X and Y is linear.
  2. **Homoskedasticity:** Variance of residuals is constant across all X values.
  3. **Independence:** Residuals are uncorrelated across observations.
  4. **Normality:** Residuals are normally distributed.

- **Diagnostics via Residual Plots:**
  - **Linearity Violation:** Residuals follow a pattern (positive → negative → positive).
  - **Heteroskedasticity:** Residual variance increases with X (funnel shape) or over time.
  - **Autocorrelation:** Residuals show systematic patterns over time (seasonality).

  - **Non-normality:** Outliers or skewed residual distribution.

- **Example Residual Patterns:**

Assumption	**Violation Pattern in Residual Plot**	**Implication**
Linearity	Curved pattern	Model form incorrect
Homoskedasticity	Funnel shape	Std. errors unreliable
Independence	Periodic pattern over time	Autocorrelation present
Normality	Heavy tails / skew	Invalid t-tests for small n

- **Example: Heteroskedasticity Over Time**
  - Plot residuals vs. time index.
  - Increasing spread over time $\Rightarrow$ volatility clustering.

Key Formulas

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}, \quad b_0 = \bar{Y} - b_1\bar{X}$$

$$SSE = \sum(Y_i - \hat{Y}_i)^2$$

$$\hat{Y}_i = b_0 + b_1 X_i$$

Summary Table

Term	**Meaning**
b_0	Intercept: predicted Y when $X = 0$
b_1	Slope: change in Y per unit change in X
ε	Error term: unexplained variation in Y
SSE	Sum of squared errors
OLS	Estimation method: minimizes SSE

10.2: Analysis of Variance (ANOVA) and Goodness of Fit

- **Purpose of ANOVA:** Statistical procedure for analyzing the total variability of the dependent variable into explained and unexplained components.
- **Key Definitions:**
  - **Total Sum of Squares (SST):**

$$SST = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Measures total variation in the dependent variable.

- **Sum of Squares Regression (SSR):**

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

  Variation explained by the independent variable.

- **Sum of Squared Errors (SSE):**

$$SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

  Unexplained variation (residuals).

- **Mean Square Regression (MSR):**

$$MSR = \frac{SSR}{k}$$

  For simple regression $k = 1 \Rightarrow MSR = SSR$.

- **Mean Squared Error (MSE):**

$$MSE = \frac{SSE}{n - k - 1}$$

  For simple regression: $MSE = \frac{SSE}{n-2}$.

- **Relationship:**

$$SST = SSR + SSE$$

  Total variation = explained variation + unexplained variation.

- **ANOVA Table (Simple Regression):**

Source	**df**	**Sum of Squares**	**Mean Squares**
Regression (explained)	1	SSR	$MSR = \frac{SSR}{1}$
Error (unexplained)	$n - 2$	SSE	$MSE = \frac{SSE}{n-2}$
Total	$n - 1$	SST	–

- **Standard Error of Estimate (SEE):**

$$SEE = \sqrt{MSE}$$

  Smaller SEE $\Rightarrow$ better model fit.

- **Coefficient of Determination (R^2):**

$$R^2 = \frac{SSR}{SST}$$

  Proportion of total variation explained by the independent variable. For simple regression: $R^2 = r^2$ where r is correlation.

- **Example (Given Data):**

$$SSR = 0.0076, \quad SST = 0.0482, \quad MSE = 0.0012$$

$$R^2 = \frac{0.0076}{0.0482} = 0.158\ (15.8\%)$$

$$SEE = \sqrt{0.0012} = 0.035$$

- **F-Statistic:**

$$F = \frac{MSR}{MSE} = \frac{SSR/k}{SSE/(n-k-1)}$$

For $k = 1$, $n = 36$:

$$F = \frac{0.0076}{0.0012} = 6.33$$

Compare with critical F_c at $\alpha = 5\%$: $F_c \approx 4.1$ Since $F > F_c$, reject H_0 (slope is significant).

- **Hypothesis Test for Regression Coefficient:**
  - Null Hypothesis: $H_0 : b_1 = 0$
  - Alternative: $H_a : b_1 \neq 0$
  - Test Statistic:

$$t = \frac{\hat{b}_1 - b_1}{s_{\hat{b}_1}}$$

  - Example:

$$\hat{b}_1 = 0.64, \quad s_{\hat{b}_1} = 0.26$$

$$t = \frac{0.64 - 0}{0.26} = 2.46$$

Critical $t_{0.025,34} = \pm 2.03 \Rightarrow t > t_c$, reject H_0.

- **Key Takeaways:**
  - SST splits into SSR (explained) + SSE (unexplained).
  - R^2 measures model explanatory power.
  - SEE is the standard deviation of residuals.
  - F-test checks if regression as a whole is significant.
  - t-test checks significance of individual coefficients.

10.3: Predicted Values and Functional Forms of Regression

LOS 10.e: Calculate and interpret the predicted value for the dependent variable, and a prediction interval for it

- **Predicted values** are obtained from the regression equation using estimated coefficients and a given forecast of the independent variable.
- For a simple linear regression:

$$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_p$$

  where:

  - $\hat{Y}$ = predicted value of the dependent variable
  - X_p = forecasted value of the independent variable

- **Example: Predicting the dependent variable**

$$\text{ABC Excess Returns} = -2.3\% + 0.64 \times (\text{S\&P 500 Excess Returns})$$

$$\text{Given:} \quad X_p = 10\%$$

$$\hat{Y} = -2.3\% + 0.64 \times 10\% = 4.1\%$$

  Interpretation: If S&P 500 excess returns are forecast at 10%, ABC's excess returns are predicted to be 4.1%.

- **Confidence Interval for Predicted Value:**

$$\hat{Y} \pm (t_c \times s_f)$$

  where:

  - t_c = two-tailed critical t-value (df = $n - 2$)
  - s_f = standard error of the forecast

- **Variance of forecast:**

$$s_f^2 = \text{SEE}^2 \left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{(n-1)s_X^2} \right]$$

  where:

  - SEE^2 = variance of residuals
  - s_X^2 = variance of the independent variable
  - X = forecast value of independent variable

- **Example: Prediction Interval**

$$\hat{Y} = 4.1\%, \quad s_f = 3.67, \quad t_c(5\%, df = 34) = 2.03$$
$$\begin{aligned} 95\% \text{ CI} &= 4.1\% \pm (2.03 \times 3.67\%) \\ &= 4.1\% \pm 7.5\% \\ &\Rightarrow [-3.4\%, 11.6\%] \end{aligned}$$

Interpretation: We are 95% confident ABC excess returns will be between −3.4% and 11.6% when S&P 500 excess returns are forecast at 10%.

LOS 10.f: Functional Forms of Simple Linear Regression

- Assumption: Relationship between X and Y is linear.
- If violated (e.g., exponential growth), transformations can linearize the relationship.
- Common transformations:
  1. **Log-lin model:**
  $$\ln Y_i = b_0 + b_1 X_i + \varepsilon_i$$
  Interpretation: b_1 = relative change in Y for an absolute change in X.
  2. **Lin-log model:**
  $$Y_i = b_0 + b_1 \ln(X_i) + \varepsilon_i$$
  Interpretation: b_1 = absolute change in Y for a relative change in X.
  3. **Log-log model:**
  $$\ln Y_i = b_0 + b_1 \ln(X_i) + \varepsilon_i$$
  Interpretation: b_1 = relative change in Y for a relative change in X.
- **Example Table: Functional Forms and Interpretations**

Model	**Equation**	**Interpretation of b_1**
Log-Lin	$\ln Y = b_0 + b_1 X$	% change in Y per unit change in X
Lin-Log	$Y = b_0 + b_1 \ln X$	Unit change in Y per % change in X
Log-Log	$\ln Y = b_0 + b_1 \ln X$	% change in Y per % change in X

- Selection of model depends on:
  - Nature of data
  - Goodness-of-fit metrics (R^2, SEE, F-statistic)

11 INTRODUCTION TO BIG DATA TECHNIQUES

11.1: Introduction to Fintech

LOS 11.a: Aspects of Fintech in Gathering and Analyzing Financial Data

- **Fintech** = Developments in technology applied to the financial services industry.
- Companies developing such technologies are called *fintech companies*.
- Primary areas of fintech development:
  1. Increased functionality to handle large, diverse datasets.
  2. Use of AI techniques for analyzing very large datasets.

LOS 11.b: Big Data, Artificial Intelligence, and Machine Learning

- **Big Data** = All potentially useful information generated in the economy.
- **Sources of Big Data:**
  1. **Traditional:** Financial markets, company reports, government statistics.
  2. **Alternative:**
     - **Individuals:** Social media posts, online reviews, emails, website visits.
     - **Businesses:** Bank records, retail scanner data (*corporate exhaust*).
     - **Sensors / IoT:** Smartphones, smart buildings, RFID chips.
- **Characteristics of Big Data (3 V's):**
  1. **Volume:** Measured in TB, PB, etc., growing exponentially.
  2. **Velocity:** Speed of data communication.
     - Low latency = real-time data (e.g., stock price feeds).
     - High latency = delayed data updates.
  3. **Variety:** Degree of structure in data.
     - Structured: Databases, spreadsheets.
     - Semi-structured: Photos, web code.
     - Unstructured: Video, audio.
- **Data Science:** Methods to process and visualize data.
  1. **Processing:**
     - Capture → collect/transform data.
     - Curation → adjust for bad/missing data.
     - Storage → archive/access.

    - Search → retrieve information.
    - Transfer → move data to needed location.
  2. **Visualization:**
    - Structured: charts, graphs.
    - Unstructured: word clouds, mind maps.

- **Challenges:** Ensuring data quality, avoiding sampling bias, processing unstructured data.

- **Artificial Intelligence (AI):** Systems programmed to simulate human cognition.
  - Example: Neural networks process data like the human brain.

- **Machine Learning (ML):** Algorithms that learn from input data without explicit programming assumptions.
  1. Uses training, validation, and test datasets.
  2. **Types:**
    - Supervised → input/output labeled, model learns mapping.
    - Unsupervised → no labels, model learns structure.
    - Deep Learning → multiple neural network layers; can be supervised or unsupervised.
  3. **Applications:** Image/speech recognition, fraud detection.
  4. **Risks:**
    - Overfitting → too complex, captures noise as signal.
    - Underfitting → too simple, misses true patterns.
    - Black box problem → hard to explain outputs.

LOS 11.c: Applications of Big Data and Data Science to Investment Management

- **Text Analytics:** Analysis of unstructured text/voice data.
  - Example: Detect sentiment from earnings call transcripts.
  - Use: Automating regulatory filing analysis.

- **Natural Language Processing (NLP):** AI interpreting human language.
  - Examples: Speech recognition, translation.
  - Use: Check regulatory compliance in emails; detect subtle sentiment shifts in analyst reports.

- **Risk Governance:** Monitor exposures, perform stress testing.

  - Use: Real-time risk monitoring via ML models.

- **Algorithmic Trading:** Automated trading based on rules.
  - Uses: Optimal order execution, high-frequency trading, intraday arbitrage.

Summary Table: Key Concepts in Fintech

Concept	**Definition**	**Finance Example**
Fintech	Tech applied to financial services	Robo-advisors, mobile banking apps
Big Data	Large, complex datasets from multiple sources	Combining stock market feeds with social media sentiment
AI	Simulating human cognition	Fraud detection via neural networks
Machine Learning	Algorithm learns from data	Predicting stock prices from historical patterns
NLP	AI interpreting human language	Automating sentiment analysis of analyst reports
Algorithmic Trading	Automated trade execution based on rules	Splitting large orders to minimize price impact

Example: Applying Fintech to Equity Trading

- Step 1: Use Big Data to aggregate market prices, order book data, and Twitter sentiment.
- Step 2: Apply NLP to interpret sentiment from news headlines.
- Step 3: Feed sentiment and market variables into ML model.
- Step 4: Generate buy/sell signals for algorithmic execution.

Module Quiz 11.1 (Answers):

1. **A** — Application of technology to the financial services industry.
2. **A** — Machine learning is most relevant for analyzing Big Data.